Use of prior knowledge in biological systems modelling

Reshetova, P.V.

Publication date
2017
Document Version
Final published version
License
Other

Polina Reshetova

Use of prior knowledge in biological systems modelling

P. Reshetova

Use of prior knowledge in biological systems modelling

2017

Use of prior knowledge in biological systems modelling

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel

op donderdag 2 maart 2017, te 10.00 uur

door

Polina Vladimirovna Reshetova
geboren te Terentyevskoye, Rusland

Promotiecommissie:

Promotor:

- Prof. dr. A. H. C. van Kampen — Universiteit van Amsterdam
- Prof. dr. A. K. Smilde — Universiteit van Amsterdam

Copromotor:

- dr. J. A. Westerhuis — Universiteit van Amsterdam

Overige leden:

- Prof. dr. S. Brul — Universiteit van Amsterdam
- Prof. dr. A. H. Zwinderman — Universiteit van Amsterdam
- Dr. P. L. Klarenbeek — Universiteit van Amsterdam
- Dr. J. E. Guikema — Universiteit van Amsterdam
- Prof. dr. R. M. H. Merks — Leiden Universiteit
- Prof. dr. J. Heringa — Vrije Universiteit Amsterdam
- Prof. dr. H. V. Westerhoff — Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research reported in this thesis was carried out at the Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam (Science Park 904, 1098 XH Amsterdam, The Netherlands) and was supported by the BioRange programme of the Netherlands Bioinformatics Centre (NBIC).

Contents

Chapter 1
Introduction

An enormous amount of biological knowledge has been generated by the scientific community and is available from a large number of biological databases, scientific literature, and domain experts. This knowledge is actively used to define new hypotheses and to validate new findings, but it may also be included in computational modelling and high-throughput data analysis as prior knowledge in order to improve the analysis or guide it towards meaningful solutions. However, the use of prior knowledge is not always straightforward and may additionally be hampered by its incompleteness. Moreover, the use of prior knowledge may bias the results towards known biology thereby preventing new findings. In this thesis we explored the use of prior knowledge in data-driven and knowledge-driven modelling approaches for high-throughput data analysis and biological systems modelling. In the first part of this thesis we reviewed methods that incorporate prior knowledge in statistical models for the analysis of high-throughput transcriptomics and metabolomics data (Chapter 2). We highlighted characteristics and differences of this methods and the type of prior knowledge that was used. In the second part of this thesis we used prior knowledge to model two biological systems. First, we used sparse prior knowledge to build a network-based model of a multi-organ genistein elimination pathway that can assist in the design of new experiments (Chapter 3). Secondly, we developed a mathematical model of B-cell affinity maturation based on incomplete prior knowledge about the selection of high affinity B cells. Here, we used prior knowledge to simplify the mathematical description of the B-cell selection process to avoid excessive model complexity. We showed that despite this simplification the model generated valuable insights in the affinity distribution among (un)expanded subclones (Chapter 4). Further, the model was used to identify changes in B-cell lineage trees during affinity maturation (Chapter 5). In summary, we explored possibilities to facilitate high-throughput data analysis with prior knowledge , and demonstrated the use of prior knowledge in biological systems modelling.

1.1. Modelling approaches

The modelling of biological systems is an essential part of nowadays research in systems biology. It aims to abstract a biological system in a statistical or mathematical modelling framework and to subsequently apply computational methods to determine (emerging) properties of the system. The work in this thesis aims to show how the modelling of biological systems can benefit from prior knowledge, and also how prior knowledge can be incorporated in the modelling procedure. We considered three types of models that either incorporate prior knowledge directly in the model computation or have been constructed by using prior knowledge:

- Statistical models for high-throughput data analysis;

- Network-based models of biological systems;
- Mathematical models.

Statistical models aim to find and quantify relationships between the variables in a dataset. These models may or may not assume a distribution of the variables under investigation. An example of a statistical model is a (linear) regression model that describes a relationship between one or more explanatory variables and a dependent variable. Other examples of statistical models include component models such as principal component analysis [1] or cluster methods such as k-means clustering [2] that also aim to find associations between objects (e.g. samples) and variables (e.g. genes). However, these statistical models generally do not explain the precise nature of relationships between variables in terms of biological processes. Hence, they are phenomenological. In contrast, mathematical models, such as differential equations, use equations to specify a mechanistic model, that is, the nature of the relationships between, for example, genes is explicitly specified. Moreover, the parameters in such model have biological definitions. Network-models such as Petri nets [3], Bayesian networks [4], and Boolean networks [5, 6] live between the phenomenological and mechanistic models. These models do not necessarily include a full mechanistic description of the biological system but specify relationships between objects in the model more explicitly then is the case in statistical models. Models reviewed or used in our research involve various statistical models, Petri nets, and ordinary differential equations.

Modelling approaches may further be divided in two categories: data-driven and knowledge-driven (Figure 1.1). Data-driven approaches include statistical models and network-based models to analyse high-throughput experimental data such as coming from transcriptomics and metabolomics studies in which many genes and metabolites, respectively, are measured. In this type of modelling one is generally interested in identifying (linear) relationships or correlations between the variables and, therefore, the models do not use any *a priori* known facts about the modelled biological system. However, due to the high data dimensionality (many variables are measured compared to the number of samples), these models may reveal chance correlations (i.e. spurious correlations, which are found for a specific data set but have no biological relevance). As a solution, incorporation of prior knowledge has been suggested, which may also improve the interpretability of the results by focusing them on known biology. For example, prior knowledge has been used to softly penalise the minimization function in a principal component based method in order to find principal components that are partially defined by already known information [7]. We reviewed a range of methods that followed this strategy in Chapter 2 [8].

Knowledge-driven approaches comprise methods that use known biological facts to define the model structure (e.g. equations and parameters in a mathematical model, or the topology in a network-based model) and, therefore, depend on literature, information from public biological databases, and/or knowledge provided by domain experts. Knowledge-driven approaches include mathematical models based on ordinary differential equations, and network-based models such as Petri nets, Boolean, and Bayesian networks. Both mathematical and network-based models are widely used to model relations between molecules or cells. Usually, network-based methods are preferred if quan-

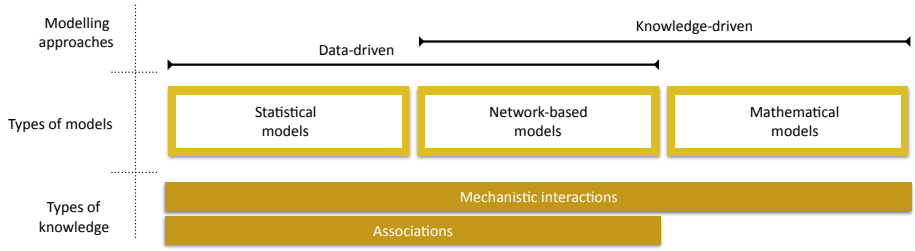


Figure 1.1: Conceptual representation of data-driven and knowledge-driven modelling approaches. Three types of models are considered, which are shown in the relation to two types of prior knowledge. Among data-driven modelling approaches we consider only models that incorporate prior knowledge.

titative values for parameters, such as required for differential equations, are difficult to obtain. While knowledge-driven approaches construct a model solely based on *a priori* available knowledge they generally use (public)experimental data at a second phase for parametrization and validation.

It is important to note that network-based models are successfully employed in both knowledge and data-driven approaches, which is indicated by the overlap of these approaches in Figure 1.1. For example, using a knowledge-driven approach Schlatter and co-authors have built a literature-based Boolean model to study a set of pathways involved in apoptosis [9]. In contrast, Sahoo and co-authors used a data-driven approach to construct a genome-wide Boolean model of gene pairs from a comprehensive set of microarray experiments [10].

The reviewed types of models have different requirements to the used prior knowledge (Figure 1.1). Mathematical models require detailed specifications of biological mechanisms (mechanistic interactions) to construct equations. In addition, some parameters in these models also have to be based on *a priori* biological knowledge (while other may be estimated from experimental data). For network-based models the same is true although, in general, less detailed information is required and the model also contains fewer parameters that do not necessarily have to reflect biological values. Statistical models and network-based models (used in the data-driven approaches) are less demanding and have capability to incorporate a wide variety of prior knowledge. Prior knowledge may come from associations between biological entities such as the regulation of gene expression by transcription factors, or the co-expression of genes. Associations may also come from pre-defined groups of biological entities such as gene products (defined in the gene ontology; GO)[11] that participate in the same pathway. Associations do not present mechanistic details and can not be incorporated in mathematical models.

Statistical models for high-throughput data analysis

High-throughput experimental techniques characterize each samples by thousands of variables (e.g., genes, proteins, metabolites) and potentially allow to reconstruct the underlying gene, protein, and metabolic networks involved in biological processes. This attractive property secured a place for high-throughput experiments in biomedi-

cal research. Consequently, a new subfield of bioinformatics has emerged that applies network-based and statistical models to the data. However, these methods have to handle specific features of the high-throughput data. Particularly, these methods do not always distinguish between biologically significant correlations and chance correlations and consequently lead to spurious findings. Moreover, biological differences among technical and biological replicates in high-throughput experimental data may be not the primary interest but may also significantly challenge the data analysis [8]. To prevent spurious findings, the use of prior knowledge has been suggested to restrict or guide the statistical modelling. We dedicated a chapter of this thesis to give a fairly broad overview of such methods in transcriptomics and metabolomics (Chapter 2).

Challenges using prior knowledge in high-throughput data analysis

Incorporation of prior knowledge in statistical analysis of high-throughput data guides the analysis towards known biological relationships and thereby reduces the detection of spurious relationships among variables. The improved separation of biologically relevant variation from the noise in the data could potentially lead to enhanced discovery of new biology. However, the amount, nature and quality of prior knowledge may drastically influence the quality of the resulting models and their ability to assist in exploring the system. Moreover, prior knowledge incorporated in the analysis may even bias the results towards the expected biology and thus leading to false positive detection of the expected relationships suggested by the prior knowledge, but not present in the data. Thus, there is a delicate balance between data-driven and knowledge-driven analysis. Unfortunately, there are no guidelines or a credible unified indicator that can help with this. We will discuss this topic in more details in Chapter 2 of this thesis.

Another important issue with prior knowledge is so-called negative prior knowledge [12]. While positive prior knowledge refers to known interactions (of any nature) between two biological entities, negative prior knowledge would reflect truly non-existing interaction between two entities. However, such non-existing interactions are generally not explicitly specified in literature or public databases making it hard to distinguish between interactions that truly do not exist and interactions that remain to be discovered and described. Non-existing interactions may be included in modelling approaches by defining which entities do not interact and therefore any identified correlations between them can be considered as spurious.

Another issue with prior knowledge and methods that incorporate prior knowledge is the evaluation of its added value. Currently, this problem has been addressed only by few authors. For example, in the work of Tian and co-authors “random” knowledge has been considered to evaluate the prior knowledge influence on the inference of gene interaction networks from high-throughput genomic data [13]. This allowed them to conclude that using real prior knowledge in their method was robust to false positive interactions between genes. Other research forwarded a general framework to investigate the relevance of different prior knowledge sources (such as databases, literature, and *a priori* gene co-expression experiments) for inferring gene interaction networks [14]. In our opinion, the evaluation of the added value of prior knowledge and the evaluation of the relevance of different prior knowledge sources requires a well-thought-of universal test framework including appropriate (synthetic) datasets. This would also greatly improve

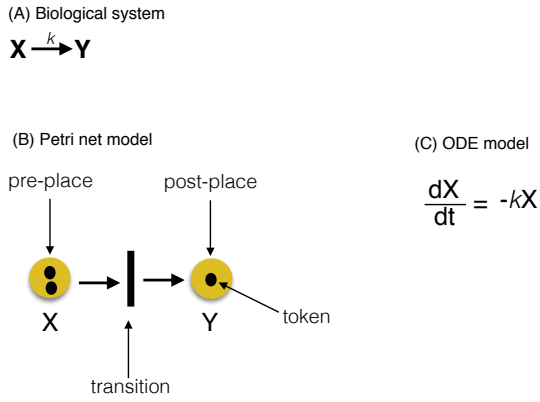


Figure 1.2: Example of (A) a biological system (chemical reaction): molecule X is converted to molecule Y with rate k. (B) Petri net model describing this chemical reaction. This Petri net contains two places and one transition. Firing of the transition moves token(s) from "pre-place" X to "post-place" Y. (C) ODEs model corresponding to the chemical reaction. The concentration change dX (and therefore dY) in time is proportional to the concentration X.

application and development of statistical approaches that include prior knowledge [8].

Network-based models of biological systems

A variety of network-based models have been developed including Petri nets [3], Boolean [5, 6] and Bayesian networks [4]. These models represent biological systems as a graph and allow to naturally resemble biological pathways. Network-based approaches allow to model a system in a range of abstraction levels depending on available prior knowledge or on the model purpose. In Chapter 3 of this thesis Petri nets were used to study the dynamics of a multi-organ genistein elimination pathway and to assist in the design of new experiments. Petri nets are bipartite graphs with two types of nodes: places and transitions [15, 3]. In Chapter 3 Petri net places represent molecules (metabolites) and transitions represent interactions between connected nodes (enzymatic reactions or metabolite transport). In addition, places contain tokens that represent the relative concentration of the corresponding molecule (Figure 1.2). During a simulation several steps are executed. At each step a transition 'fires' and tokens are moved from pre-transition places to post-transition places. A set of firing rules (heuristics) define which transition fires at each step and, in combination with topology of the network, allows to reproduce the dynamic behaviour of a real system. The unique feature of Petri nets is that various firing rules may be assigned to transitions to represent system dynamics in a desirable level of details and abstraction. This resulted in a wide range of Petri nets based methods such as Stochastic Petri nets [16], Time Petri Nets [17], Hybrid Functional Petri Nets [18] and Petri nets that incorporate Fuzzy logic [19] and even ODEs [20] providing a wide modelling capability.

Together with Petri nets, Boolean and Bayesian networks are widely used to model biological systems. Although Petri nets were used in this thesis, Boolean and Bayesian network-based approaches also play an important role in biological systems modelling.

To model biological pathways both approaches represent the biological system as a graph in which every node represents a molecule (e.g., protein, gene, and metabolite) and every edge represents a defined interaction between two molecules. A variety of interaction types may be represented: directional and undirectional, signed (inhibition/activation) and unsigned, a physical binding (e.g., binding of regulatory molecules), or correlations of gene expression. The simplest representation of biological systems is provided by Boolean network models [6]. They represent qualitative behaviour and may be used to model biological systems with no or sparse prior knowledge about quantitative parameters. Despite their simplicity Boolean networks have been widely used to study various signalling pathways and their properties [21, 22, 9, 23]. In Boolean models, each node has a Boolean state (0 or 1). Each edge holds a rule from the set of AND/OR/NOT values and this defines the interaction between two nodes. By changing initial node states or edge rules different hypotheses may be investigated. However, because Boolean networks only hold binary values they are strongly limited in the representation of continuous values and time. If a lower abstraction level is needed to model a system then Bayesian networks may be employed. Instead of a set of AND/OR/NOT rules Bayesian networks assign probabilities to node interactions. The probabilistic representation of relationships in a model is believed to be suitable to handle biological and experimental noise in high-throughput data and allows to combine Bayesian networks with analysis of high-throughput microarray data [24, 25, 4, 26, 27]. Further research led to the development of discrete multi-state models, which allow to assign multiple states to nodes or values between 0 and 1 and, therefore, allow to model sensitivities or concentrations of molecules [28]. Finally, discrete statements are transformed to continuous values of input and output in fuzzy logic models [29]. The time limitations of Boolean networks have been overcome by advanced approaches such as continuous or mixed discrete continuous Boolean networks [30, 31], and probabilistic Boolean networks [32, 33]. However, a lower abstraction level (i.e., more detail) and therefore a higher complexity of models require higher computational power as well as more parameters to be estimated.

Mathematical models of biological systems

Ordinary differential equations (ODEs) provide one often used framework to specify mathematical models of biological systems. To specify ODEs prior knowledge about the biological system is used to define the dynamics and relations between all biological entities involved (e.g. genes, metabolites, cells). Relations in an ODE may, for example, represent chemical reactions or cell differentiations (Figure 1.2). Solutions of the resulting set of equations describe or predict the temporal or spatial dynamic behaviour of the system. Because differential equations allow to represent non-linear behaviour they are widely used to model nontrivial dynamics like limit-cycle oscillations and multi-stability [34]. Moreover, while most sets of differential equations cannot be solved analytically, numerical methods are well developed and supported by various computational tools. Together the possibility to model a broad spectrum of biological systems behaviour and availability of various computational tools promote the use of differential equations. Therefore, we choose differential equations to model the complex behaviour of B-cell affinity maturation in Chapter 4. The resulting model allowed us to follow a large set of

B-cell subclones individually and, consequently, to follow affinity change in the context of B-cell lineage trees (Chapter 5). As a result, the model expanded our understanding of B-cell repertoire sequencing experimental data.

Several types of differential equations have been developed and applied in biological systems modelling. Widely used are mathematical models based on ordinary differential equations (ODEs) [35, 36]. They can be used to describe time dependent concentration or signal changes. To also model spatial dynamics partial differential equations (PDEs) may be used [37]. For example, Smith and co-authors used PDEs to reflect protein diffusion through the cytosol [38]. Another feature of biological systems is that stochastic processes inherent or external to the system may affect their behaviour. Individual differences in hormone levels or diet of subjects as well as small differences in experimental procedures like temperature may produce variations in experimental results. To address this dynamic variations stochastic differential equations (SDEs) have been applied [39]. For example, Chen and co-authors used SDEs to address stochasticity in gene regulations by transcription factors [39].

Despite their capability to model and analyse the dynamic behaviour of biological systems differential equations are not always the first choice. Particularly, the model output may strongly depend on its topology and corresponding parameter values and, therefore, a good knowledge about the biological system is required, as well as precise quantitative measurements for all parameters involved. Sometimes parameter values can be obtained from scientific literature or public databases. However, these parameter values may have been obtained under different experimental conditions that do not exactly match requirements for the new model. Alternatively, experiments to directly measure the parameter values may be conducted but this is usually time consuming if many parameters are needed, or may even be impossible. In such cases parameters may sometimes be estimated from experimental (time series) data that match the output of the model. However, parameters may be non-identifiable if the experimental data is insufficient to unambiguously determine all parameters [40, 41]. Consequently, for large and complex biological systems the parameterization of the model may become challenging and computationally expensive.

1.2. Scope and outline of the thesis

This thesis explored the use of prior knowledge in modelling approaches for high-throughput data analysis and biological systems modelling. Because the main interest was in how prior knowledge might be used in the analysis of biological systems in general, no restrictions on the specific types of models were specified. Therefore, three types of models for various applications were used. Firstly, the use of prior knowledge in data driven statistical models of high-throughput data was reviewed. Secondly, a Petri net model of human genistein elimination pathway was build based on sparse prior knowledge. Finally, ODEs were used to model B cells affinity maturation during an immune response using prior knowledge about affinity maturation to construct and eventually simplify the mathematical model.

Use of prior knowledge in the analysis of high-throughput data is an emerging field which is hoped to boost our understanding of biological systems on a big scale. The research started by a review of more than twenty high-throughput data analysis methods

in Chapter 2. The set includes methods of high-throughput data analysis in transcriptomics and metabolomics that use prior knowledge to define or to estimate statistical model parameters. Because prior knowledge may be incomplete, incorrect, or may hide new discoveries, specific attention was paid to the problem of balancing experimental data and prior knowledge. It was concluded that for further understanding of the influence of prior knowledge on the analysis, and for comparing different methods to incorporate prior knowledge, a well-defined test framework is required.

In Chapter 3 of this thesis very scarce and incomplete knowledge about a complex multi-organ genistein elimination pathway that comprises several concurrent routes was used. A Petri net based model was suggested that relied on topology alone to reconstruct metabolite concentration profiles. Furthermore, this model was used as an experimental design tool to propose new metabolite measurements to more precisely infer the relative contributions of concurrent elimination routes, and to improve the reconstruction of concentration profiles of all metabolites in this pathway. Overall it was shown that a Petri net model based on scarce prior knowledge may be used to assist in the design of future experiments to complete missing knowledge.

The second application of biological systems modelling aimed for better understanding of high-throughput B-cell repertoire RNA sequencing data. A mathematical model was developed, based on ordinary differential equations, to simulate B-cell affinity maturation during an immune response to determine the affinity distribution in (un)expanded B-cell subclones (Chapter 4) and to study the evolution of B-cell lineage trees during affinity maturation (Chapter 5). In this case available prior knowledge was used to define a simplified representation of the B-cell competition process (survival and positive selection). The simplification avoids an overcomplicated model but sufficient realistic to allow further interpretation of repertoire sequencing experiments.

This thesis is closed with Chapter 6 where some open issues are discussed and future research opportunities are suggested that could move forward the analysis of biological systems and experimental data with the use of prior knowledge.

Chapter 2
Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data[1]

High-throughput omics technologies have enabled the measurement of many genes or metabolites simultaneously. The resulting high dimensional experimental data poses significant challenges to transcriptomics and metabolomics data analysis methods, which may lead to spurious instead of biologically relevant results. One strategy to improve the results is the incorporation of prior biological knowledge in the analysis. This strategy is used to reduce the solution space and/or to focus the analysis on biological meaningful regions. In this article, we review a selection of these methods used in transcriptomics and metabolomics. We combine the reviewed methods in three groups based on the underlying mathematical model: exploratory methods, supervised methods and estimation of the covariance matrix. We discuss which prior knowledge has been used, how it is incorporated and how it modifies the mathematical properties of the underlying methods.

2.1. Background

High-throughput technologies such as DNA microarrays in transcriptomics and mass spectrometry in metabolomics produce large amounts of experimental data, where each sample is characterized by the expression levels of thousands of genes, or concentration levels of hundreds to thousands of metabolites respectively. This high number of variables gives a unique chance to catch a broad range of biological processes but, at the same time, poses significant challenges to statistical methods of analysis. First of all, traditional statistical methods highlight relationships among variables based only on mathematical criteria (e.g. maximizing variance or correlation among variables) and thereby do not always distinguish between correlations from biological origin and chance correlations that may arise because of the high dimensionality of the data and measurement noise. Secondly, biological differences of the subjects in the study produce variations in gene expression values and metabolite concentrations in experiments. Very often such biological variation is not of primary interest and not under control of the researcher.

[1] This chapter is published as: P. Reshetova, A. K. Smilde, A. H. C. van Kampen and J. A. Westerhuis 2014 *BMC Systems Biology*; 8(Suppl 2):S2. DOI: 10.1186/1752-0509-8-S2-S2

Therefore, a challenge of statistical methods in transcriptomics and metabolomics is to distinguish between the different variation sources.

Recently, new methods have appeared that use prior knowledge of the biological system to guide the statistical analysis to enhance discovery of new biology while reducing the detection of spurious relationships. In addition, prior knowledge may be used to check consistency of the available knowledge and experimental data to fill in possible gaps or add more detail. We focus our review on approaches that incorporate prior knowledge about the relationship between the genes or between metabolites to achieve an optimal balance between mathematical criteria and known biology. The relationships among variables (genes or metabolites) can be determined, for example, from public databases that contain results of previous experimental data analysis. For example, the KEGG [42] database contains information about metabolic pathways, GO [11] contains annotation of gene products, the TRANSFAC [43] database contains information about transcription factors, their binding sites, and target genes. We are specifically interested in how each method manages the balance between the discovery of new biology and, using prior knowledge, forcing the results towards existing biology.

In this review, we focus on high dimensional supervised and unsupervised data analysis methods that include prior knowledge into the mathematical model used for the analysis of metabolomics or transcriptomics data. Methods for genomics and proteomics also have been developed (see for example [44, 45]) but they are out of the scope of our review. To the best of our knowledge, our review is the first that provides a comprehensive overview of strategies using prior biological knowledge in metabolomics and transcriptomics data analysis. In the remainder of this text we will refer to metabolomics and transcriptomics data as "omics" data.

To structure this review, we classified the reviewed methods into three groups, based on the mathematical approach and whether the method is unsupervised or supervised. We distinguish three groups of mathematical approaches. The first group comprises component models that reduce the dimensionality of the data by constructing latent variables from the observed genes or metabolites. The second group comprises cluster models that use similarity measures to group related genes or metabolites, and the third group comprises covariance methods that primarily aim to estimate variances/correlations among the genes or metabolites. We wittingly have not grouped the methods based on the used type of prior knowledge. As can be seen from our review, the same type of prior knowledge is utilized by a range of mathematical methods. In section Exploratory methods we discuss unsupervised methods (component based models and clustering methods). These methods explore data and describe the major drivers underlying the observed data structure. In section Supervised classification methods we discuss supervised methods for finding a classification function that predicts class labels. In section Covariance matrices we discuss methods that estimate the covariance matrix. For each section we provide additional figures [see Supplementary Material 2.7].

2.2. Two phases of the analysis of high dimensional data

The term model has many interpretations in the bioinformatics literature. In the context of this review we define a "model" as a statistical or mathematical representation of omics data. Each model has specific estimated parameters, for example, principal

components in component models or coefficients in a regression model. Omics data is written as a two-way matrix $\mathbf{X}$ with I rows representing genes or metabolites, and J columns representing samples (e.g. subjects, tissues, treatments, diseases). J is usually much smaller than I. In this review we will focus on data analysis methods that handle two-way data. To facilitate the discussion and comparison of data analysis methods that use prior knowledge, we consider two phases in the analysis of omics data:

1. Definition of a model and estimation of the model parameters.

2. Interpretation of the model parameters in terms of biological knowledge.

Prior biological knowledge can be incorporated in each of these two phases. In the second phase the prior information is used to facilitate or even enable interpretation of the data analysis result. Examples of such methods are gene set enrichment analysis [46] and metabolite set enrichment analysis methods [47]. The enrichment methods have been extensively reviewed by others [48, 49]. In this paper we focus on the first phase in which the model parameters are estimated.

The inclusion of prior knowledge in data analysis implies that we have to weight the importance of the data against the importance of the biological knowledge. This is visualized by the slider in Figure 2.1. The methods discussed in this review implicitly or explicitly (using a weight factor) deal with this balance. Inclusion of prior knowledge aims to emphasize the known relationships between the genes or metabolites while eliminating spurious variation among these variables. However, it may also limit the possibility to make new discoveries if we put too much emphasis on the already known biology. The main challenge is to find an optimal position for the slider such that new discoveries can be made from the data that are in agreement with current biological knowledge. The methods reviewed in this paper follow different strategies to set this balance.

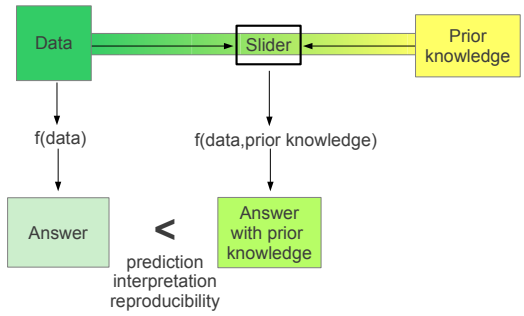


Figure 2.1: A general scheme of data analysis methods. *f(data)* is a function that does not include prior knowledge. *f(data, prior knowledge)* is a function that includes prior knowledge. "Answer with prior knowledge" gives a better predictive model, is easier interpretable and/or more reproducible than "answer" without prior knowledge. The slider controls the strength of the influence of prior knowledge on the result.

2.3. Exploratory methods

Component models

Component models are used in omics data analysis to extract and represent the most informative changes in the experimental data among different conditions or samples. Recently, new approaches have appeared that use prior information such as regulation networks, protein networks, and metabolic networks to highlight the most valuable changes that are related to the question of the study. The basic equation of component models is

$$\mathbf{X} = \mathbf{A}\mathbf{F}^T + \mathbf{E} \tag{2.1}$$

where $\mathbf{X}$ is an I by J data matrix that consists of I variables (genes or metabolites) and J samples. Matrix $\mathbf{A}$ contains the linear components (summarizers) of the data $\mathbf{X}$. Matrix $\mathbf{F}$ contains the weights of these components in each sample. The residual matrix $\mathbf{E}$ contains the part of the data not explained by the model. Matrices $\mathbf{A}$ and $\mathbf{F}^T$ are estimated to maximize the fraction of data variation that is explained by the components. The combination of columns in $\mathbf{A}$ and $\mathbf{F}^T$ are called principal components and are required to be orthogonal (Principal Component Analysis) or independent (Independent Component Analysis). In PCA matrices $\mathbf{A}$ and $\mathbf{F}^T$ are estimated in a least squares sense to minimize the sum of squares of the residuals (Equation 2.2).

$$\min_{A,F} \left[\left\| \mathbf{X} - \mathbf{A}\mathbf{F}^T \right\|^2 \right] \tag{2.2}$$

The prior information is translated into the mathematical model by applying various restrictions on the elements of $\mathbf{A}$ and $\mathbf{F}^T$. The restrictions are a predefined range of certain elements in $\mathbf{A}$ and $\mathbf{F}^T$ or dependencies of some elements on other elements in $\mathbf{A}$ or $\mathbf{F}^T$. Because of the restrictions, the new principal components are no longer forced to be orthogonal and may deviate from their standard requirements to reflect better the underlying biological processes. We define two concepts of incorporation of prior knowledge in a component model.

The first concept is based on a relatively simple idea. Metabolites or genes are split into two groups with the ones on which the focus will be in one group (X2) and the remainder in another (X1). The analysis also shows metabolites or genes from the first group, which follow profile patterns of the second group. Van den Berg *et al.* [50] adjusted consensus PCA to implement the concept. Equation 2.3 and Figure 2.2 (Supplementary Material 2.7) demonstrate the method.

$$\begin{bmatrix} w_1 \cdot \mathbf{X}_1 \\ w_2 \cdot \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \mathbf{F}^T + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix} \tag{2.3}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are two parts of $\mathbf{X}$. $\mathbf{X}_2$ contains a small group of metabolites that are thought to be important for the problem under the study. Initially, weights w_1 and w_2 were used to compensate the small size of matrix $\mathbf{X}_2$. This weights were set to the square root of the sum of squares of the corresponding matrix. Consequently, the total variation in each matrix became "1" and equally important for explaining variation in the data. We suggest the weights may also be used as the slider (Figure 1) to put more emphasis on known relationships among metabolites in $\mathbf{X}_2$. The method was applied

on experimental data from a phenylalanine overproducing strain and wild type strain of *E. coli* that contained measurements of metabolites under 28 conditions. The authors selected the phenylalanine biosynthesis pathway for the subset of genes in the matrix $\mathbf{X}_2$ and showed that the method was able to successfully identify large common effects between the metabolome in the matrix $\mathbf{X}_1$ and the specific metabolites in the matrix $\mathbf{X}_2$.

The second concept to include prior knowledge predefines variations, which must be described. For example, Network Component Analysis (NCA) of Liao and colleagues is forced to catch variation within a gene regulatory network [51]. For that the method specifically searches for changes in expression level of genes that are known to be regulated by a transcription factor. Variation caught by the method is interpreted as the activity of the transcription factor in the network during the experiment. Therefore, the method uses known qualitative information about regulatory networks topology to generate quantitative network information on the connection strength between genes and transcription factors while decomposing experimental data. The formula of the method is

$$\mathbf{X} = \mathbf{A}_{TF} \cdot \mathbf{F}^T + \mathbf{E} \tag{2.4}$$

where each column in $\mathbf{A}_{TF}$ is forced to represent the effect on genes of a single transcription factor by putting zeros representing that the specific gene is not regulated by that specific transcription factor (Figure 2.3 in Supplementary Material 2.7). The further estimation of the matrix $\mathbf{A}_{TF}$ is made only for elements that are not restricted to be 0. Thus only the genes that are regulated by the corresponding transcription factor will have parameters in $\mathbf{A}_{TF}$ estimated and the values in the matrix $\mathbf{F}^T$ are considered as the activity of that transcription factor in each sample.

The authors analyzed experimental data of a cell cycle regulation in *S. cerevisiae* and focused the analysis on 11 transcription factors that are known as regulators in the cell-cycle. NCA successfully revealed the role of each transcription factor; in contrast, the gene expression ratios of the transcription factors do not suggest their important role.

Later it was shown that NCA suffered from many false connections between genes and transcription factors in the prior knowledge. Therefore a realization of the slider that would control inputs of biological and mathematical constraints was needed. Yu and Li proposed to use the high-confident part of the prior knowledge to build a model [52]. The model is finalized later through iterations between an estimation of components on experimental data and an estimation on the low-confident part of the prior knowledge. The authors argued that the iteration process allowed to reduce the influence of false connections on the model. As proof the authors report two regulatory networks under different growth conditions for *S. cerevisiae*. For the same purpose, Tran *et al.* suggested to combine stepwise regression and NCA in an iterative approach [53]. The authors argue that their algorithm overcomes the problems of NCA in the analysis of large networks where multiple transcription factors regulate a single gene. Moreover, the authors argued that NCA could not be used in the case when the number of experiments was very small and their method overcame this limitation. The method was demonstrated on a network that contains 70 transcription factors, 778 genes, and 1423 edges between the transcription factors and genes.

Grey Component Analysis (GCA) is the first implementation of the slider that actually

allows to choose how much trust is given to the prior knowledge [7]. The topology of gene regulatory networks is used in the same way as in NCA. But how strict the analysis has to follow the prior knowledge is defined by a soft penalty approach. The penalty approach allows using the GCA method for two purposes. If the penalty is strict the decomposition is biased towards prior knowledge. If the penalty is soft the method analyzes the consistency of the data and the prior knowledge. By varying λ it is possible to show how well the data follows the prior knowledge and where it does not follow it anymore.

To implement the idea, GCA minimizes the combined sum of squares of the model residual and the penalty

$$\min_{A,F} [\left\|\mathbf{X}-\mathbf{A}\mathbf{F}^T\right\|^2 + \lambda\left\|\mathbf{W}\circ(\mathbf{A}-\mathbf{A}^{true})^2\right\|]; \lambda \geq 0 \tag{2.5}$$

where matrix **A** is defined according to the given prior knowledge, but the zeros applied to the matrix **A** in NCA method are allowed to be small values in GCA. The authors argued that in noisy data such as omics data, enforcing real zeros might lead to the mis-estimation of the nonzero values. The added part $\lambda\left\|\mathbf{W}\circ(\mathbf{A}-\mathbf{A}^{true})^2\right\|$ is the penalty. Matrix $\mathbf{A}^{true}$ is the structure as applied in NCA, **A** is the estimated matrix and **W** is an indicator matrix which assures that the penalty is only active on the positions in **A** where $\mathbf{A}^{true}$ has zeros. The parameter λ determines how much emphasis the method puts to fit the data and how much to follow the prior knowledge in $\mathbf{A}^{true}$.

Above we discussed various component models. Table 2.1 in Supplementary Material 2.8 summarizes our overview.

Cluster models

Cluster analysis aims to construct groups of genes or metabolites that share a biological factor such as a common function or co-regulation by transcription factors. Traditional cluster algorithms base their similarity score only on measured data (gene expression values or metabolite concentrations) and discard known relationships between genes or metabolites. This may, for example, result in clusters of genes/metabolites that exhibit similar profiles across samples but are not necessarily co-regulated, do not have similar functions, or do not participate in the same pathway. Here we describe three concepts of including prior knowledge into clustering:

1. Adjusting the distance measure by including prior knowledge.
2. Improving K-means clustering for variables with similar profiles within one regulatory pathway.
3. Extending model-based clustering by increasing the probability of grouping variables with similar prior knowledge.

These concepts are discussed in more detail below.

The first concept adjusts the clustering distance measure between variables. A distance between variables based on prior knowledge is calculated and added to the data based distance. Based on the combined score the hierarchical tree will cluster variables with both similar experimental profiles and prior knowledge. Figure 2.4 in Supplementary Material 2.7 shows the first concept where similarity between GO annotation is used as the prior knowledge distance measure. The slide ruler naturally fits the concept.

The first implementation of the concept was done in a publication of Cheng *et al.* [54]. To achieve the goal the method uses similarity in the GO annotation between genes. GO has a hierarchical structure where more general functional terms are located closer to the root, while more specific terms are located closer to leafs. The authors assumed that the first common ancestor of two terms that is closer to leafs reflects a larger functional similarity of the corresponding genes. The formula of the method is

$$d_{ii'} = s_{ii'} + g_{ii'} \tag{2.6}$$

where $s_{ii'}$ is the gene expression similarity score between gene *i* and *i'*, calculated as Euclidian distance between the expression profiles of gene *i* and gene *i'*, and $g_{ii'}$ is the annotation similarity between GO terms of two genes that is based on the GO terms common ancestor. The authors showed that a strong correlation between biological functions and expression profiles led to a cluster. Genes that had close expression patterns but did not have similar annotation were separated.

R. Kustra and A. Zagdanski improved this approach by including a weight factor that balances the contribution of profile distances and the prior knowledge [55]. The overall distance between two genes i and i' was defined as

$$d_{ii'} = \lambda s_{ii'} + (1-\lambda) g_{ii'}; 0 \leq \lambda \leq 1 \tag{2.7}$$

where the gene expression similarity $s_{ii'}$ is given by the Pearson's correlation between gene profiles; $g_{ii'}$ is the GO annotation similarity; λ represents the slider. This method also utilizes the hierarchical structure of the GO tree, but in contrast to the method of Cheng who used the common ancestor, this method uses the Information Content of each node. The authors suggested that the GO similarity measure would diminish spurious perturbations in gene expression levels and would lead to more meaningful clusters by focusing the analysis on the known biology. To study the influence of prior knowledge the authors clustered 3224 yeast genes from 424 microarray experiments. Specifically, the authors proposed to use a protein-protein interaction based measure to asses the biological relevance of clusters for λ = 0.0, 0.25, 0.5, 0.75, 1.0. However, since the protein-protein interactions also reflect functional relationships between the genes, it can not be used as an unbiased measure to evaluate the incorporation of GO annotations as prior knowledge. As expected the protein-interaction score increased for smaller λ, i.e., stronger influence of the GO annotations. Consequently, it was not possible to suggest a good value of λ. An additional stage of validation conducted by a biologist or a new measure of biological relevance of clusters were required.

Whereas the cluster methods we have discussed so far use GO information, Hanish *et al.* incorporated metrics on metabolic and regulatory pathways from KEGG into the distance function [56]. The distance function assigns small values to pairs of genes, which are close in a network and show similar expression patterns. Genes which are far apart in the network and are not co-regulated or even oppositely regulated are assigned large values. The distance function emphasizes genes that are co-regulated within pathways. The proposed model for the distance between two genes is

$$d_{ii'} = 1 - 0.5 * (g_{ii'} + s_{ii'}) \tag{2.8}$$

where $s_{ii'}$ is a Pearson correlation based measure and $g_{ii'}$ is a measure based on the 'minimal degree' of a path between two genes in a metabolic pathway. Both measurements were adapted in order to combine the Pearson correlation and the minimal degree to one joint function that would emphasize genes with a high expression profile correlation and which are tightly linked within a pathway. Compared to a distance measure based on either the correlation or minimal degree, this compound distance compensates for biased results due to, for example, very high profile correlations or missing pieces of prior knowledge. The degree of path is calculated as the sum of incident edges of all nodes between two genes. Note that the authors took a minimal path without hubs, because the hubs are considered to be unspecific or ubiquitous molecules and thus unimportant or misleading for the method. The method does not implement the slide ruler with an explicit weight factor but gives an equal importance for both expression data and prior knowledge.

In hierarchical clustering, the final clusters are defined by horizontally cutting the branches of the tree at a certain level. This may also be a non-trivial process. Dotan-Cohen and co-authors proposed a tree snipping algorithm that construct clusters by cutting selected edges at different levels [57]. This method uses GO terms to annotate each node and provides a novel partitioning of the cluster tree in order to have genes with similar GO annotation in one or closely related clusters. More specifically, during the procedure a GO label list of each leaf of a subtree is compared to the annotation of the corresponding cluster. A leaf with the most dissimilar list of labels will be excluded from the subtree while nodes from close subtrees and similar labels will be included. In the first step, the method builds the hierarchical tree without using the prior knowledge. Subsequently, the method changes the original grouping by incorporating the prior knowledge into the partitioning function. We note that the authors assumed any types of labels and not specifically GO annotation. For example, the transcription factors known to regulate genes can be used. For that reason the tree snipping algorithm does not utilize the hierarchical graph information that is specific for GO. Considering an improvement of other methods by including the graph information (as in [54]) we expect that it might give a better result for the tree partitioning algorithm as well. We also note that the method can be directly used in the field of metabolomics where the partitioning may be improved by metabolic networks, or biological annotation.

The second concept of incorporation of prior knowledge in clustering does not explicitly use prior knowledge as a similarity measure. Instead, in a first step genes are grouped according to prior knowledge and, subsequently, similarity among gene expression profiles within a single group is used to improve the clustering. Following this concept, Tseng *et al.* proposed a clustering method PW-Kmeans (Penalized and Weighted K-means) that extends the K-means method by incorporating GO functional annotations [58] (Figure 2.5 in Supplementary Material 2.7). The method groups genes according to known functional annotation from GO and then assigns a weight to each gene. The weight reflects how well the gene expression profile conforms to expression profiles of all others genes in its *a priori* defined group. High expression profile similarity among genes with common functional annotation results in small values of their weights and consequently in tight clusters. In addition, the method introduces a noise cluster that contains all scattered genes, which do not follow expression profiles of other genes with

similar GO annotation.

The method adapts the loss function in the following way

$$W(C;k,\lambda) = \sum_{k=1}^{K} \sum_{x_i \in C_k} w(x_i;L)d(x_i,C_k) + \lambda|S| \tag{2.9}$$

where K is the number of clusters; $d(x_i,C_k)$ is the distance between gene i expression profile x_i and the mean of the cluster C_k; $w(x_i;L)$ is the weight that codes the prior knowledge; λ is a penalty term that forces scattered variables in a separate cluster C_s; $|S|$ is the number of scattered genes in the noise cluster; $C = \{C_1,...,C_k,C_s\}$ is the resulting clustering assignment. Minimization of Equation 2.9 produces a clustering solution. Intuitively, a smaller λ will produce tighter clusters, but more genes will be assigned to the noise set.

The prior knowledge comes in the form of L known pathways in which gene i participates. If each pathway l (1,...,L) contains N_l genes then x_{nl} is the expression profile of gene n in pathway l. The value of the weight function $w(x_i;L)$ is directly proportional to the distance between the expression vector of gene i (i.e. x_i) and one of the l pathways. Thus, the value of $w(x_i;L)$ is small for genes whose expression vector x_i closely follows at least one of pathways in the set L. How well a gene follows pathways in the set L is defined by formula

$$\min_l \frac{1}{N_l} \sum_{n=1}^{N_l} \left\| x_i - x_{nl} \right\| \tag{2.10}$$

Shen and co-authors observed that the parameter $w(x_i;L)$ in PW-Kmeans algorithm is gene-specific and remains the same no matter which cluster the gene is assigned to [59]. Therefore, while weighting does help identifying the scattered genes, it does not enhance the clustering of genes with similar functions. To overcome this limitation, Shen proposed a novel weighted clustering method, Dynamically Weighted Clustering with Noise set (DWCN) that considers the same weight for all genes within one cluster. Instead of the parameter $w(x_i;L)$ in the original equation (2.9) Shen uses the smallest p-value of over representation of all possible GO terms for the genes in the cluster. Consequently, the method separates scattered genes and makes use of functional annotation data to enhance the clustering of genes with similar functions. The authors showed that DWCN outperforms both the original K-means and PW-Kmeans methods on simulated data and gave clusters with strong biological explanation.

The third concept also uses grouping of genes according to prior knowledge in purpose of better clustering. It extends model-based clustering by using the assumption that genes with similar GO annotation have the same probability to belong to one cluster. The concept was realized by Pan [60] in stratified model-based clustering method (Figure 2.6 in Supplementary Material 2.7). Model-based clustering methods build a gene probability distribution function to belong to all possible clusters and use the similarity among the functions to cluster genes. The initial probability distribution function for each gene to belong to the clusters is

$$f(x_i;\Theta) = \pi \sum_{c=1}^{C} f_c(x_i;\theta_c) \tag{2.11}$$

where x_i is the expression vector of gene i, C is the number of clusters, f_c is a probability distribution function with parameters θ_c ($\theta_c = \{\mu_c, \delta_c^2\}$ where μ_c is a gene expression mean and δ_c^2 is a gene expression variance in probability distribution c). The parameter π is the prior probability that a gene originates from each distribution (in other words, the prior probability that a gene belongs to each cluster). The parameter Θ is a set of unknown parameters (π, θ_c) that will be maximized in the procedure. Originally, π is assumed to be the same for all genes. Pan suggested to take an advantage of known grouping of genes and assign to all genes in each group a prior probability of belonging to a cluster. He replaces π by a cluster and gene group specific probability π_h. For that all genes are grouped to H_1,...H *subscripth* groups. Then, the same prior probability π_h to end up in one cluster c is assigned to all genes in a group h. The initial probabilistic function for any gene i in functional group h became

$$f_h(x_i;\Theta_h) = \sum_{c=1}^{C} \pi_h f_c(x_i;\theta_c) \tag{2.12}$$

where $\Theta_h = \{\pi_h, \theta_c\}$. Pan argued that the probability component of model-based methods fits very well the highly variable nature of biological data and gives a broad range of possibilities to include biological prior knowledge. As an example, Pan tested the probability of genes with the same GO labeling to comprise one cluster.

The discussed implementations of the second and third concepts do not realize the slider and do not allow to change the ratio between influence of prior knowledge and experimental data. Considering incompleteness and shortcomings of secondary databases that are used in the methods, new realizations of the slider are of interest.

We have summarized cluster methods that include prior knowledge in calculation of the similarity score in Table 2.2 Supplementary Material 2.8. All the methods described are from transcriptomics studies. We are not aware of any implementation of cluster models in metabolomics that includes prior knowledge. However, clustering of metabolomics data is a helpful and popular approach. No doubt it is worthwhile to implement clustering methods in metabolomics that are driven by prior knowledge. The functional annotation of metabolites is available and could potentially be used for knowledge guided clustering.

2.4. Supervised classification methods

The main goal of supervised methods is to infer a classification function from a labeled training dataset. The classification function should be able to correctly predict labels of new samples. Examples of such algorithms include regression analysis, support vector machine and decision trees. We define three concepts of including prior knowledge that are used to adjust supervised methods. The first concept separates all variables to groups according to prior knowledge and builds a classification model for each group independently. The second concept forces genes or metabolites that are connected in a network to have close coefficients in the classification function. The third concept uses prior knowledge to predefine the topology of a decision tree.

To reduce the multiple testing problem and to improve the sensitivity and specificity of the classification, the first concept uses a group of related variables to classify the

samples. A group may represent a pathway or a set of genes with similar GO annotation. The concept does not require the changes to be in the same direction (only up or only down) but it gives a larger score to a group where changes among more variables are found (Figure 2.7 in Supplementary Material 2.7).

The idea was suggested by Goeman *et al.* and implemented in the global test [61]. The authors employed the logistic regression model and rewrote it for *J* samples and *I'* genes as follow

$$E(Y_j|\beta) = h^{-1}(\alpha + \sum_{i=1}^{I'} x_{ij}\beta_i) \tag{2.13}$$

where α is the intercept, β_i the regression coefficient for gene *i*, *h* the logit function, x_{ij} is the gene *i* expression profile and *j* is the index for the samples (j =1, ...,J). Note, that the model is built for *I'* variables (genes), which belong to the same group (or pathway). For each group of genes, defined by the prior knowledge, a separate model will be build. The authors suggested a "gene influence" plot to uncover the influence of a single gene. As an example, the authors demonstrated the new method using gene expression data for a cell line treated and untreated with heat shock. While the overall expression profile was not notably different between two groups, the global test showed significant differences for groups of genes known to function in heat shock response according to GO database.

A possibility to use the global test for a metabolomics application was shown by Hendrickx *et al.* [62]. The authors successfully tested a selection of pathway metabolites from KEGG on metabolites profiles of *E. coli* and *S. cerevisiae*. Specifically, they showed that glycolysis pathway and the TCA cycle pathway are significantly different when aerobic conditions are compared to anaerobic conditions for *S. cerevisiae*. The authors concluded that the results of the global test correspond with the physiology of studied organisms and therefore can be used in metabolomics.

The idea of the global test was further developed by several authors including the highly cited method of Chuang *et al.* [63]. They interpret groups of genes as subnetworks and assume that proteins that are close in protein-protein interaction networks have a similar gene expression vector. While the idea to test a group of genes simultaneously instead of multiple testing for each single gene remains the same in Chuang's work, we would like to put attention on how the groups were defined. The authors score subnetworks of protein-protein interaction network in gene expression data of metastatic and non-metastatic breast tumors. To find subnetworks a greedy search algorithm is used. A score function is calculated for a set of genes that is combined based on topology of the network. In each iteration step a next closest gene is added and a new score is checked for increasing. The score function *O(l)* for a particular subnetwork *l* is calculated by the formula

$$O(l) = \sum_{x \in l} \sum_{y \in z} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{2.14}$$

where *p(x,y)* is the joint probability density function of subnetwork *l* and a set of output labels *z* (metastatic or non-metastatic); *p(x)* and *p(y)* are marginal density functions.

The score function *O(l)* represents the mutual information between gene expression vector x_i from subnetwork l over samples and a corresponding vector of sample labels z. All significantly different subnetworks represented as gene groups were used to train the logistic regression model. The authors showed that subnetwork markers are more reproducible and achieve higher accuracy in the classification than individual marker genes selected without network information.

The first concept incorporates prior information as a vector with binary labels that show whether or not a gene belongs to a group. This concept does not allow implementation of the slide ruler.

The second concept is based on the idea that genes in the same regulatory network will have similar expression profiles. Following this idea Rapoport *et al.* [64] suggested to consider a gene expression profile from one microarray experiment as a function and to apply the Fourier transform to it. Genes were arranged according to the topology of metabolic networks from the KEGG database. The Fourier transform was used to decompose the expression function into "low-frequency" and "high-frequency" components. The authors argued that the "high-frequency" components contained expression profiles of unimportant genes and measurement errors while the "low-frequency" components reflected properties of the system. The low-frequency part of the decomposed data was successfully used for PCA analysis and to train support vector machine classifiers to distinguish between irradiated and non-irradiated samples of *S. cerevisiae* strains.

Another attempt to implement the second concept was done by Li *et al.* in a network-constrained regularization procedure for a linear regression model [65]. The method requires a smoothness of the regression coefficients across the network. The smoothness means that two variables that are connected in the network must have close weights in the classification function (Figure 2.8 in Supplementary Material 2.7). The regularization is based on the normalized Laplacian of the network and similar to L_1 and L_2 penalties on the regression coefficients called the LASSO or elastic net [26] (Figure 2.9 in Supplementary Material 2.7 shows an example of Laplacian matrix). For nonnegative penalty coefficients λ_1 and λ_2 the network constrained regularization criterion is defined as follows:

$$\min_{\beta} \left[(y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda_1 |\beta| + \lambda_2 \beta^T \mathbf{L}\mathbf{p}\beta \right] \tag{2.15}$$

In the minimization procedure of $\beta^T \mathbf{L}\mathbf{p}\beta$ only coefficients of connected genes are important and coefficients of not connected genes are neglected by 0 in the Laplacian. Moreover, because the sum of each row of the Laplacian is zero, absolute values in β which are close in the network are forced to be similar. This is how the network-constrained coefficient $\beta^T \mathbf{L}\mathbf{p}\beta$ induces a smooth solution of β on the known network.

The third concept employs a pathway topology to build an easy interpretable decision tree (Figure 2.10 in Supplementary Material 2.7). Each inner node corresponds to a gene; each edge corresponds to either up regulation or down regulation of the gene. Finally, each leaf corresponds to a class in the classification problem. By the idea, each path from the root to the leafs can be analyzed for biological interpretation of the system. Consequently, it is possible to analyze the final decision tree and identify up and down regulated genes in each of discriminated classes. The concept was implemented

by Dutkowski and Ideker in the method Network-guided forest [66]. It is important to mention that the method is not forced to use all information about the network; only the important for studied experiment and classification problem part will be used. It is of the interest to implement the method in metabolomics, because the network guided forest method uses the network topology but does not assume a similar concentration of neighboring metabolites. The concentration freehold can be used as the decision value.

We summarize supervised methods that include prior knowledge to guide the analysis in Table 2.3 Supplementary Material 2.8.

2.5. Covariance matrices

This section provides a separate discussion of the covariance matrix because it plays a central role in many multivariate data analysis methods, as discussed in the previous sections.

Estimation of the covariance matrix from omics data with a low number of samples and a high number of variables is notoriously difficult. A solution is to regularize the estimation by a structured so-called target matrix. Schafer and Strimmer first gave an overview of the most widely used target matrix **Tt** for analysis of high dimensional genomics data that, however, did not incorporate prior knowledge [67]. The authors suggested that the covariance matrix **T** can be estimated as

$$\mathbf{T} = \lambda \mathbf{Tt} + (1 - \lambda)\mathbf{Tu} \tag{2.16}$$

where **Tu** is unstructured covariance matrix estimated from data; **Tt** is the structured covariance target matrix. Later, several authors suggested to use prior knowledge to define **Tt** to allow the regularization of all variables in one biological group together rather then individual regularization for each variable [68, 69]. We note that if **Tu** represents experimental data and **Tt** represents prior knowledge then λ provides an implementation of the slide ruler. The main concept of covariance matrix optimization by prior knowledge is to push the structure of the matrix towards known biology. For example, the confidence that a covariation between two genes in experimental data is not due to the high dimensionality of the data is higher when there is also evidence of a connection between these genes from the prior knowledge (Figure 2.11 in Supplementary Material 2.7). We discuss two methods of defining mymatrixTt by prior knowledge below.

Guillemont *et al.* presented a method called graph constrained discriminant analysis (gCDA) that regularized estimation of the gene covariances by the Laplacian matrix **Lp** of a known gene regulation network [69]. The authors defined the target matrix **Tt** as

$$\mathbf{Tt} = (\mathbf{Lp} + \mathbf{U})^{-1} \tag{2.17}$$

where **U** represents the unit (identity) matrix that stabilizes the covariance matrix **Tt**. We give an example of matrix **Tt** in Figure 2.12 in Supplementary Material 2.7. The authors compared performance of the method with gene regulation networks inferred from microarray data (other than the analyzed) and with gene regulation networks obtained from KEGG database. Interestingly, gene regulation networks inferred from microarray data always outperformed gene regulation networks from KEGG.

Tai and Pan also used gene regulation networks to construct a matrix **Tt** with block-diagonal structure [68]. All genes were combined in groups (according to pathways in which genes participate) and represented by matrices on the diagonal of **Tt**. The diagonal values are obtained from the covariance matrix **Tu**. The off-diagonals of genes that are not related are set to 0 while the off-diagonals of related genes are calculated by the formula

$$tt_{ii'} = t_h \sqrt{tu_{i'i'} tu_{ii}} \tag{2.18}$$

where t_h is the covariance mean of all genes in the group h; $t_{i'i'}$ and t_{ii} are diagonal values obtained from the covariance matrix **Tu**. The block-diagonal covariance matrix constructed this way mathematically represents the idea that genes from the same functional group will have more close covariances than genes from different functional groups. The final covariance matrix **T** was used in classification of simulated and real tumor data by linear discriminant analysis. The classification function based on the new covariance matrix showed a better performance compared to classification functions that were based on covariance matrices regularized by mathematical criteria along. Moreover, the interpretation of the result was improved because the classification function was guided by groups of genes with biological meaningful connection.

The final covariance matrix defined by Tai and Pan was further studied by Jelizarow *et al.* [70]. Specifically, they showed that an arbitrary solution to solve prior knowledge ambiguity affected the classification result. The prior knowledge ambiguity included genes that were in no functional group or genes that were in more then one functional group. The authors compared performance of ten structured matrices **Tt** that solved the ambiguity in ten different ways.

2.6. Discussions and Conclusions

In this work, we reviewed data analysis methods that incorporate prior biological knowledge in the definition of the model and the estimation of its parameters. Most of the reviewed methods are developed in the field of transcriptomic and only few are available for metabolomics data. It might reflect the problem of metabolite identification in metabolomics data. It remains hard to assign metabolite names to peaks what leaves us with only a limited number of variables which are known and those for which prior knowledge can be incorporated.

Authors of the methods claim that prior knowledge forces and guides the analysis towards the underlying biology and give more reproducible and reliable result. However, to promote a more wide-spread use of these approaches, much more validation of the results is required. Another factor that limits the further use and development of these methods is the lack of easy accessible implementations of the methods. Most of the algorithms are not available as commercial or open-source software (e.g., as an R package), nor are they available as a web-application or web-service. Since these algorithms are generally complex, it will not be easy for a biologist without mathematical and programming skills to implement any of these methods and use it to analyze the data.

One way forward is to define a common and accepted framework to test methods using prior knowledge. Currently, authors use their own set of data and validation pro-

cedures, which makes it very hard to compare the performance of such methods. Such a validation framework is important since recent evidence shows that prior knowledge does not always help to improve the result. For example, the probability model based clustering approach of Pan did not show an improvement when including prior knowledge on a set of 300 gene expression microarrays [60]. However, the method seems to give an improvement when applied to a smaller dataset. The authors suggest that in the large dataset there was enough information in the data itself. Staiger *et al.* showed that a simple aggregation of the expression levels of several genes did not outperform a single gene set to train prognostic classifiers in breast cancer [71]. Four methods were compared, including the method of Chuang *et al.* [63], which is discussed in our paper. The authors specifically evaluated a framework to compare performance of four cluster methods that used prior knowledge. First, protein-protein interaction networks and gene regulatory networks were used as prior knowledge to group genes with each of the four methods. Subsequently, the groups were used as features to train three classification methods (nearest mean classifier, logistic regression, 3 NN classifier). While authors of the four methods claimed to increase the stability of features chosen with prior knowledge and/or to increase classification accuracy, Staiger *et al.* showed that they did not perform better than single-gene based methods. To our knowledge, this is the first attempt to develop such a framework and in our opinion, the development of frameworks for correct comparison of different approaches desperately needs more attention.

In general, we can conclude that more research is needed to understand if and how to optimally apply prior knowledge in data analysis methods. A critical assumption is that the prior knowledge is correct and valid for the data being analyzed. If this assumption does not hold, prior knowledge might produce erroneous results. Moreover, it is necessary to study a role of prior knowledge in the analysis of pathological states when main metabolic and regulatory pathways undergo essential changes and no longer are in agreement with mainstream prior knowledge. For example, changes in metabolic pathways [72], gene regulatory pathways [73], and even massive genomic rearrangements [74] are well known for cancer cells. The question is, does knowledge about normal states of a system is appropriate or helpful for exploration of pathological states of the system?

We reviewed more than twenty methods that represent the current state of high-throughput data analysis by incorporating prior knowledge in transcriptomics and metabolomics. We highlighted features and differences of the methods and the type of prior knowledge that was used. We showed that there is a need for a proper framework which would allow a fair comparison of different methods and would help further understanding of how prior knowledge influences results.

Authors contributions

PR was responsible for the overall planning and coordination of the review as well as writing the paper; AHCK and JAW equally contributed to the elaboration of the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by The Netherlands Bioinformatics Centre (www.nbic.nl). JAW acknowledges funding from STATegra the Seventh Framework Programme [FP7/2007-2013] under grant agreement 306000. The publication costs for this article were funded by STATegra the Seventh Framework Programme [FP7/2007-2013] under grant agreement 306000 and COST-BMBS, Action BM1006 "Next Generation Sequencing Data Analysis Network", SeqAhead.

2.7. Supplementary Material 1. Additional figures.

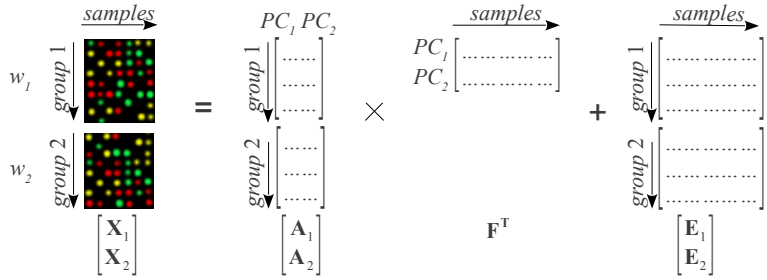


Figure 2.2: Consensus PCA. The matrix $\mathbf{X}$ with I metabolites and J samples is divided in two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Result of the decomposition is the matrix $\mathbf{A}$ that has two parts $\mathbf{A}_1$ and $\mathbf{A}_2$. Each part of the matrix $\mathbf{A}$ describes variations only in the respective part of the matrix $\mathbf{X}$.

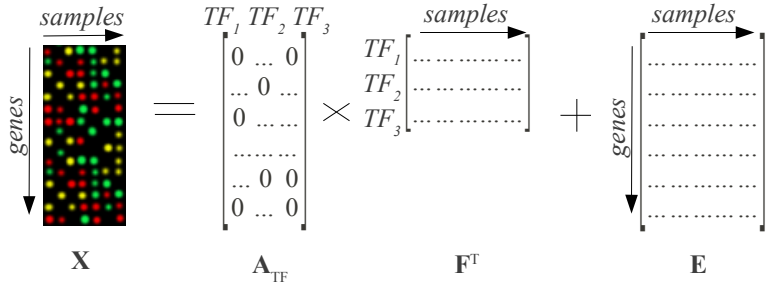


Figure 2.3: The Network Component Analysis. The matrix $\mathbf{A}_{TF}$ is predefined to represent transcription factors by columns. A column in $\mathbf{A}_{TF}$ has zeros for genes which are not regulated by a specific transcription factor. Values of the regulated genes are estimated. The values in the matrix $\mathbf{F}^T$ are considered as the transcription factors' activity in each sample.

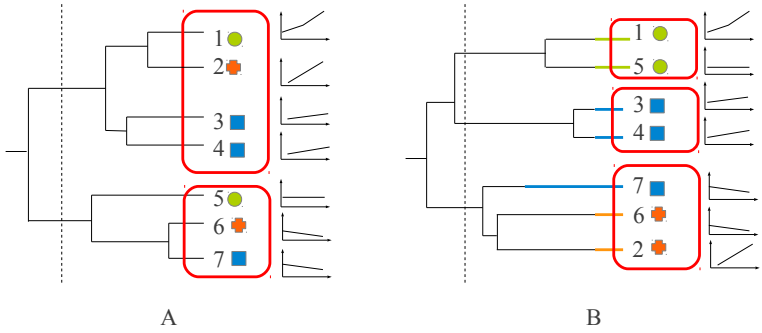


Figure 2.4: Extending hierarchical clustering by prior knowledge. The difference between hierarchical clustering without (A) and with (B) prior knowledge. Colored figures (circle, cross, and square) indicate GO labels. Red squares indicate clusters. The result of clustering with prior knowledge (B) combines genes with similar expression profiles and similar GO labeling in one cluster.

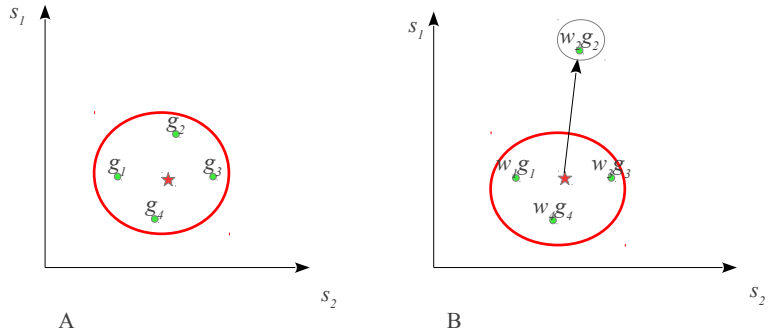


Figure 2.5: Extending the K-means clustering by pathway information. (A) shows the k-means clustering of four genes based on the gene expression profile similarities s in two samples s_1 and s_2. (B) shows k-means clustering that is extended by the prior knowledge. Distances between gene i and the cluster mean are multiplied by the gene specific weight w_i. w_1, w_3 and w_4 are close to 1 and do not change the distance to the cluster mean greatly. w_2 is big enough to push g_2 from the red cluster to the cluster of scattered genes c_s (c_s is showed by gray color).

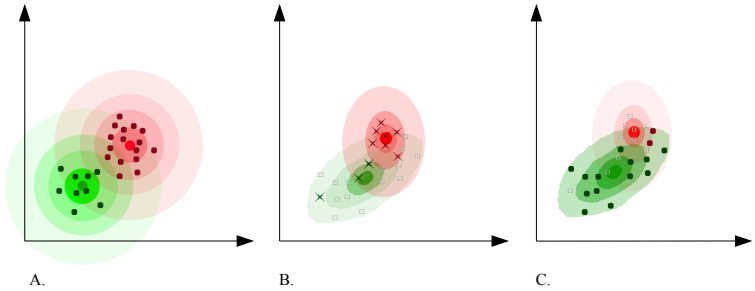


Figure 2.6: Extending model-based clustering by prior knowledge. Three mixture models based on two components (red and green circles). Intensity of the colors shows a combination of the weight of each component in a model and posterior probability for a particular variable. Model (A) treats all variables equally and simultaneously. By color of each dot the assigned cluster is shown. For models (B) and (C) variables are split into two groups G_1 (crosses) and G_2(circles). Model (B) is built for group of variables G_1 and the red component has a larger weight in the model. Model (C) is build for group G_2 and the green component has a larger weight in the model. Note that components parameters (the mean and dispersion) in both models stay the same

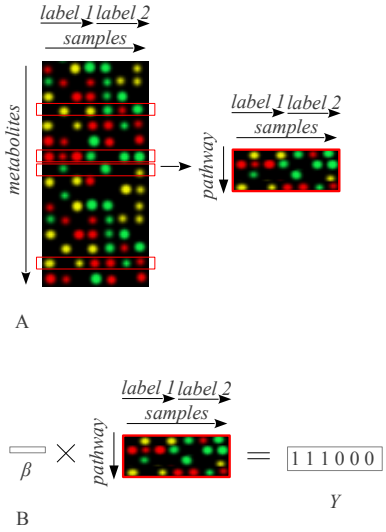


Figure 2.7: The global test. (A) The first concept selects a group of metabolites according to prior knowledge. (B) Next, a regression model is built for the metabolites in the group. β is a vector of the regression coefficients for each metabolite in the group and it is checked for an association with outcome labels.

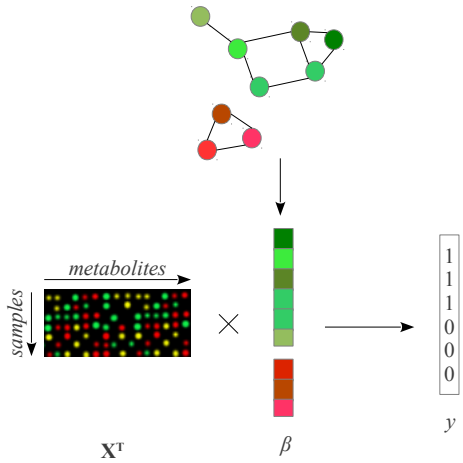
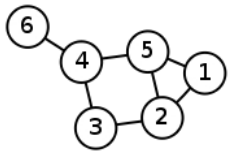


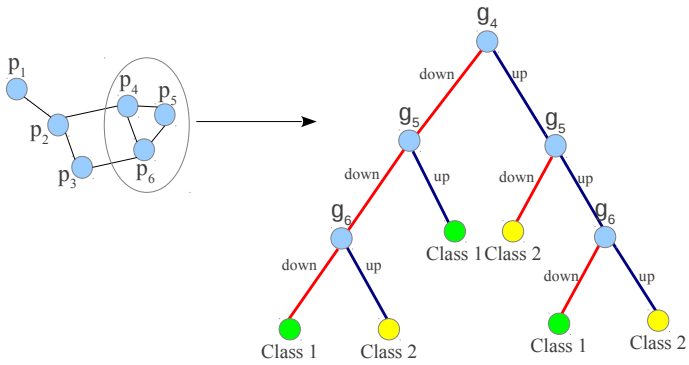
Figure 2.8: Extending linear regression model by prior knowledge. β is a vector of regression coefficients and it is optimized to be smooth along networks N_1 and N_2.



$$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

A B

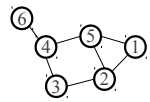
Figure 2.9: Example of Laplacian matrix. (A) shows a pathway, B shows the corresponding Laplacian matrix. On the diagonal of this matrix the number of links from a specific node can be found. Existence of a link between two nodes coded as -1 in the corresponding place in the Laplacian matrix **Lp**. This makes the sum of each row and each column equal to 0.



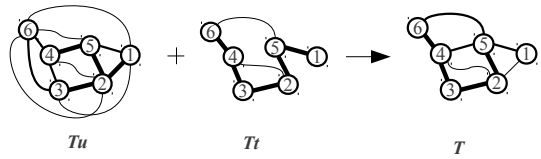
A

B

Figure 2.10: Extending decision tree method by prior knowledge. (A) is a priori known protein interaction network. The method searches for connected network modules and based on them builds decision trees. The gray circle shows an example of such module. (B) is a decision tree that is built based on the network module. For that each protein is assigned to the correspondent gene. Each inner node corresponds to a gene; each edge corresponds to either up regulation or down regulation of the gene. Each leaf corresponds to a class in the classification problem.



A. Real network



B. Inferred networks

Figure 2.11: Example of the covariance matrix. Graphs are inferred from different covariance matrices. (A) is a real pathway. (B) shows networks that are inferred from unstructured covariance matrix **Tu** (based on experimental data), structured target covariance matrix **Tt** (based on prior knowledge), and final covariance matrix **T** (based on combination of gene expression values and prior knowledge). Prior knowledge removes false positive links and emphasize known *a priori* links.

$$\left(\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right)^{-1} = \begin{pmatrix} 0.45 & 0.18 & 0.08 & 0.07 & 0.17 & 0.03 \\ 0.18 & 0.37 & 0.15 & 0.09 & 0.16 & 0.05 \\ 0.08 & 0.15 & 0.43 & 0.15 & 0.10 & 0.08 \\ 0.07 & 0.09 & 0.15 & 0.37 & 0.13 & 0.18 \\ 0.17 & 0.16 & 0.10 & 0.13 & 0.37 & 0.07 \\ 0.04 & 0.04 & 0.08 & 0.18 & 0.07 & 0.59 \end{pmatrix}$$

Lp U Tt

Figure 2.12: Example of the structured target covariance matrix **Tt** in graph constracted discriminant analysis of Guillemont *et al.*. This example is constructed for the network shown in Figure 2.9. When two nodes are connected in the graph, the covariance (off diagonals) is expected to be higher than the not connected nodes. For the variances in the covariance matrix (diagonal elements) we see that a node connected to many other nodes (e.g. a hub) is expected to have a lower variance than the nodes with few connections. It is a mathematical representation of the biological idea that hub nodes are tightly regulated and thus expected not to vary much in a particular situation.

2.8. Supplementary Material 2. Tables.

Table 2.1: Overview of methods that are based on PCA and include prior knowledge.

Method	Applied on / Prior knowledge	Principle
Consensus PCA [50]	Metabolom of *P.putida S12* and *E. coli* / set of important metabolites	**A** is divided on two parts, one of which explains variations in a selection of important metabolites and the other explains variations in the rest of metabolites.
NCA [51]	DNA microarrays of *S. cerevisiae* / transcription factors activities	**A** represents a transcription factor by each column and has zeros representing that the specific gene is not regulated by that specific transcription factor. Only elements in **A** that are not restricted to be 0 will be estimated to minimize the sum of the squared residuals in **E**
GCA [7]	DNA microarrays of *S. cerevisiae* / transcription factors targets	**A** represents a transcription factor by each column similar to NCA but the zeros are allowed to be small values. There is also a penalty where $\mathbf{A}^{true}$ is the structure as applied in NCA and the method allow **A** be different from $\mathbf{A}^{true}$ according to the penalty

Table 2.2: Overview of clustering methods that include prior knowledge.

Method	Applied on / Prior knowledge	Principle
Cheng et al [54].	The top 80 ranked genes in DNA microarrays according to F-scores in Leukocyte differentiation time-series experiment on mouse. / Similarity between two GO classes according to the topology of GO tree.	The similarity score between two genes is the sum of the GO annotation similarity and gene expression profile similarity.
R. Kustra, A. Zagdański [55].	3224 yeast genes from 424 microarray experiments / Information Content between two GO terms	The similarity score between two genes is a sum of the GO annotation similarity and the gene expression profile similarity. But the contribution of each part is specifically defined by the λ parameter
Penalized and Weighted K-means clustering (PK-means clustering) [58].	Mass spectrometry data of 2856 peptides of 22 amino acids long. DNA microarrays from *S. cerevisiae* cell-cycle dataset (from Spellman) / Gene functional annotations (GO)	Combine genes with a similar annotation and an expression profile in one cluster and create a cluster of scattered genes
Dynamically Weighted Clustering with Noise [59].	DNA microarrays from *S. cerevisiae* cell-cycle dataset and 112 segregants in a cross between two parental strains BY and RM / Gene functional annotations (GO)	Combine genes with similar annotation and expression profile in one cluster and and create a cluster of scattered genes. As opposed to PK-means method, each cluster has its own set of terms in the annotation.
Probability model-based clustering [60].	300 microarray experiments with gene deletions and drug treatments for *S. cerevisiae*. / GO functional annotations	Assign same prior probability of belonging to one cluster to all genes which are labeled by the same GO term.
Co-clustering of genes and vertices in the network [56].	DNA microarrays of seven time points for *S. cerevisiae*. After mapping to KEGG database 1571 genes and proteins were clustered / Metabolic pathways	Assign a similarity value to pairs of genes based on their distance in a network and expression the profile similarity
Hierarchical tree snipping [57].	DNA microarrays for *S. cerevisiae* cell-cycle experiment / GO annotations	Put genes which are close in the cluster tree and with similar GO annotation in one cluster by allowing cut clusters in different tree levels.

Table 2.3: Overview of supervised methods that include prior knowledge to guide the analysis.

Method	Applied on / Prior knowledge	Principle
Global test [61]	microarray data of 3571 genes from 27 patients with Acute Lymphoic Leukemia and 11 patients with Acute Myeloid Leukemia. In-house 20160 oligonucleotides array for a cell line treated/untreated with a heat shock. / Groups of variables	Test if the mean of all variables in a group is related to different experimental conditions.
Global test in metabolomics [62]	metabolome of E. coli measured by LC-MS, GC-MS; LC–MS data of *S. cerevisiae* / Metabolic pathways	Test if the mean of all variables in a group is related to different experimental conditions.
Network-based classification [63]	Microarrays of metastatic and non-metastatic breast tumor tissues. / Protein-protein interaction network.	Define distinguishable for an outcome subnetworks, by testing the mean of expression of all genes in the subnetworks. Use the distinguishable subnetworks to train a classifier.
Network based decomposition of gene expression data [64]	Microarrays of irradiated and non-irradiated *S. cerevisiae* strains / metabolic pathways	Remove the high frequent component from gene expression profiles according to the topology of gene regulation pathways.
Li et al [65]	DNA microarrays of glioblastoma samples / Gene regulation networks	Define a network-constrained penalty function for linear regression model to make the coefficients smooth on the network Network-guided forest
Network-guided forest [66]	DNA microarrays of of germ samples, breast and brain cancer samples / Protein-protein interaction networks.	Build a classifier as classification tree based on a protein-protein interaction network topology.

Table 2.4: Table 4 - List of symbols.

Symbol	Meaning
X (I x J)	Data matrix
I, $i = 1,...,I$	Number of genes or metabolites
J, $j = 1,...,J$	Number of samples
x_i (1 x J)	Gene expression vector
A (IxR)	Score matrix in data decomposition methods
R , $r = 1,...,R$	Number of components in decomposition methods
F (J x R)	Loading matrix in data decomposition methods
E (I x J)	Residuales matrix in data decomposition methods
w	Weights in consensus PCA
W (I x R)	Indicator matrix in GCA
$\mathbf{A}^{true}$ (I x R)	Matrix predefined by a priory known transcription factors regulation for each gene.
S (I x I)	Matrix of similarity scores between genes based on experimental data
G (I x I)	Matrix of similarity scores between genes based on prior knowledge
D (I x I)	Matrix of similarity scores between genes based on combination of experimental data and prior knowledge
C (I x K)	Cluster matrix
K, $k = 1,...,K$	Number of clusters
C_s	Cluster that contains scattered variables
$\|S\|$	Number of scattered variables in cluster C_s
L (1 x L), $l = 1, ..., L$	Pathways
N_l, $n = 1, ..., N_l$	Number of genes in pathway l
x_{nl}	expression profile vector of gene n in pathway l
H, $h = 1, ..., H$	gene groups defined by prior knowledge
Lp	Laplacian matrix
Tu	Covariance matrix based on experimental data
Tt	Covariance matrix based on prior knowledge
T	Covariance matrix based on experimental data and prior knowledge
t_h	mean of covariances between genes in group h
U	Unit (identity) matrix

Chapter 3
Using Petri nets for experimental design in a multi-organ elimination pathway[2]

Genistein is a soy metabolite with estrogenic activity that may result in (un)favorable effects on human health. The elucidation of the mechanisms through which food additives such as genistein exert their beneficiary effects is a major challenge for the food industry. A better understanding of the genistein elimination pathway could shed light on such mechanisms. We developed a Petri net model that represents this multi-organ elimination pathway and which assists in the design of future experiments. Using this model we show that metabolic profiles solely measured in venous blood are not sufficient to uniquely parameterize the model. Based on simulations we suggest two solutions that provide better results: parameterize the model using gut epithelium profiles or add additional biological constraints in the model.

3.1. Introduction

Genistein is a soy metabolite with estrogenic activity that may result in (un)favorable effects on human health (for a review see [75]). The elucidation of the mechanisms through which food additives such as genistein exert their beneficiary effects is a major challenge for the food industry. A better understanding of the genistein elimination pathway could shed light on such mechanisms. Parts of this pathway are hosted by specific organs (including the small intestine, gut, liver, and kidney). Metabolite degradation products travel between these organs and eventually are secreted through the gut or kidney. Although many nutrikinetics studies have been conducted to explore the genistein multi-compartment elimination pathway in human and animal models, relatively few of its details are known and, consequently, the precise metabolic pathways and routes remain to be established. In this work we therefore do not consider detailed metabolic reactions involved in the elimination pathway (which are largely unknown) but focus on the routes of degradation products through the network of compartments (the involved organs and blood).

Mathematical modelling helps to gain a more detailed understanding of the genistein elimination pathway but this requires a model describing this system in sufficient

[2]This chapter is published as: P. Reshetova, A. K. Smilde, J. A. Westerhuis and A. H. C. van Kampen 2015 *Computers in Biology and Medicine*; 63:19-27. DOI: 10.1016/j.compbiomed.2015.05.001

detail. However, Petri nets are able to use incomplete and/or imprecise information to reconstruct system's behaviour. Petri nets developed by Carl Adam Petri provide a generic approach for modelling of concurrent systems [15]. A Petri net is a bipartite graph with two types of nodes - places and transitions. In biological applications, places generally represent biological entities such as molecules, genes and enzymes. Places contain tokens that reflect, for example, metabolite concentrations or gene expression levels. Transitions represent relations between biological entities such as enzymatic reactions or metabolite transport. A Petri net simulation results in a time-dependent redistribution of tokens reflecting system dynamics. Simulation of the Petri net model implies that we select and "fire" a specific transition resulting in tokens being moved from one place to the next. The firing rules define which transition fires and the number of tokens subsequently transferred. This, together with the topology of a Petri net model, results in a qualitative representation of the system's dynamics.

Petri nets have become a popular tool for studying biological networks such as metabolic networks [76]. The review paper of Baldan *et al* explains how metabolic pathways have been represented and modelled with Petri nets [3]. The authors also discuss various ways to use Petri nets for modelling network topology structures such as negative feed-back loops and inhibition. Modelling network topology with Petri nets has been shown to give qualitative biologically relevant insights about the dynamics of biological systems [77, 78]. A quantitative analysis of biological network dynamics required further extension of Petri nets in a way similar to mathematical modelling using ordinary differential equations (e.g., [20]). Such Petri nets require knowledge of kinetic parameters. However, it has been shown that network dynamics might be determined by using only network topology [79, 80]. For example, Ruths and co-workers [81] assumed that network connectivity is the most significant determinant of the signal propagation and that discarding kinetic parameters from the model still results in model outcome that agrees with experimental data in the majority of cases. A similar example, involved the use of Fuzzy Logic to reconstruct the topology of a cell signalling network from gene expression data *in silico* [19].

Petri net models can also assist in designing new wet-lab experiments to further characterize the system under investigation. In this paper we demonstrate how a Petri net model representing the human genistein multi-organ elimination network fits this purpose (Figure 3.1). This model describes various routes involved in the elimination of genistein after dietary exposure to this compound. Each transition of metabolites within or between organs is associated with a fraction (F) that indirectly represents its relative flux. The network topology and fractions define a model configuration and allow the simulation of time-resolved metabolite relative concentration profiles for different organs given an amount of genistein input G(I) administered to an individual (Figure 3.2A). The challenge, however, is to estimate fractions from measured profiles. In this work we use concentration profiles from LC-MS venous blood measurements obtained in a nutrikinetics study in which healthy volunteers were exposed to dietary genistein [82]. The estimation of fractions is challenging because given current domain knowledge and available data, the genistein elimination pathway is insufficiently constrained and, therefore, ambiguous in terms of model configurations (i.e., sets of fractions) being in agreement with experimental data. Therefore, we used our model as an experimental

design tool to investigate which additional information (data or prior knowledge) would be required to further constrain the system to allow accurate estimations of the fractions. In particular, we investigated if additional metabolite profiles and additional constrains (e.g., fixing fractions associated with excretion transitions) would result in better parameter estimates.

To answer these experimental design questions, we used simulated annealing (SA) to estimate fractions from experimental or simulated metabolite profiles.

3.2. Results

Fraction estimation from simulated reference profiles for all places.

We first explored if fractions can be correctly estimated based on concentration reference profiles simulated for all places in our model (Figure 3.1). Consequently, we configured the Petri net with all fractions arbitrarily set to 0.5 except for fractions F2, F30 and F31 which were set to 0.34, 0.33 and 0.33 respectively (Figure 3.2 (A)). This fulfills the requirement that the fractions corresponding to outgoing transitions of a single place sum to one, and no preference for a specific transition is assumed (see Materials and Methods section). Then, we executed ten simulations with 1000 input tokens for place G(I). A simulation is terminated when all tokens left the Petri net. During simulations we recorded the number of tokens in each place to obtain the concentration reference profiles. Subsequently, using these simulated reference profiles we reconstruct the fractions through simulated annealing (Figure 3.2 (B)). The results from these reconstructions show that some fractions are precisely estimated (e.g., F1 and F7 were estimated within 2% of their true value), while other fractions showed large variability (e.g., F17, F16 deviated 90% of their true value; Appendix Figure 3.12). Note that all boxplots presented in this paper are sorted according to their estimation variances.

We defined three classes of relative estimation errors, which only can be calculated for estimates based on simulated reference profiles since the true fractions underlying the experimental data are unknown. A transition was classified as “determinan” if the relative estimation error was less than 10%. The “moderate” and “flexible” classes correspond to estimation errors of 10-25%, and >25% respectively. Although simulated annealing may converge to sub-optimal solutions (fractions), we decided not to remove one or more runs with high(er) estimation errors (i.e. outliers in the boxplots) since this may lead to a biased result and, moreover, since this is also not possible for results based on experimental data due to unknown estimation errors.

To facilitate comparison with experimental data we defined three variance classes (low, medium, high, Appendix Figure 3.10). These classes are based on visual inspection of estimation variances observed from experimental data (Appendix Figure 3.11). Figure 3.3 shows a visual representation of the estimation errors and variances in the context of our model. This simulation experiment shows that even if concentration profiles are available for all places, only a part of the fractions can be determined with sufficient accuracy. For example, fractions F17/F16, F25/F26, F15/F18, and F24/F27 have the estimation error up to 90% in 10 optimization runs, while fractions F1/F7 have the maximum estimation error of 2% (Appendix Figure 3.12). Inspecting the concentration profiles that are produced from the model we observe that these are close to the simulated

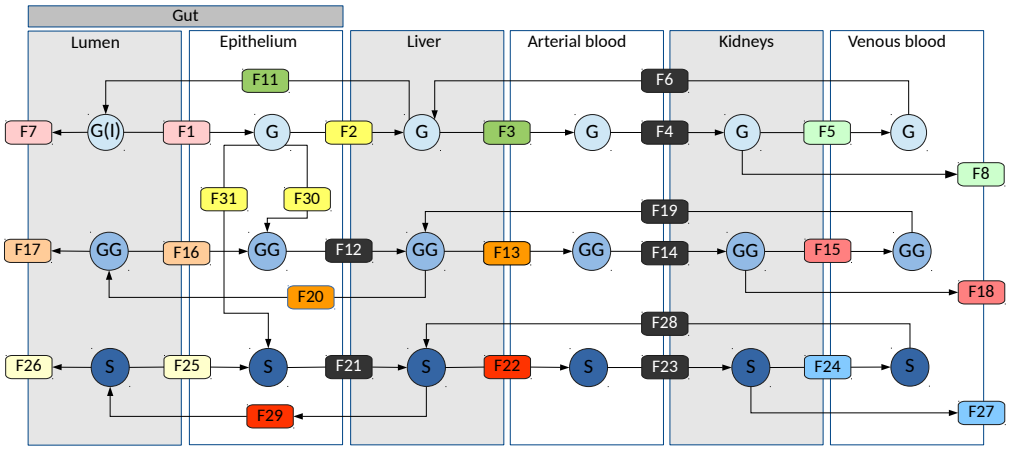


Figure 3.1: **Petri net model of the human genistein multi-compartment elimination pathway.** This model includes three metabolites (G - genistein, GG; genistein-7-glucuronide; S - genistein-7-glucuronide-4-sulphate) that travel within and between six compartments (organs, blood). G(I) is the input place, which is set to 1000 tokens at the start of a simulation and represents the amount of genistein administered to an individual. Each transition is associated with a fraction (F) indirectly representing a relative flux. Associated fractions corresponding to outgoing transitions originating from the same place are shown in the same color. Outgoing transitions associated with places without other outgoing transitions are shown in black. Six transitions (F7, F17, F26, F8, F18 and F27) represent the excretion of metabolites from the gut or kidney.

reference profile (Figure 3.4). Compared to the venous blood profiles the gut epithelium profiles are closer to the reference profiles due to intrinsic constraints between the various gut epithelium transitions. However, overall we conclude that additional model constraints are required to improve the estimation of the fractions.

Fraction estimation from simulated and experimental reference profiles for venous blood places.

Our experimental data comprised only three metabolite profiles (genistein, genistein-7-glucuronide, genistein-7-glucuronide-4-sulphate) measured in venous blood. Given our previous simulations, these profiles are not expected to provide sufficient information to estimate all fractions in the network with high accuracy. To confirm this, we simulated reference profiles for these three metabolites only. Fractions in the model were set either as 0.5 or 0.33 as previously. Subsequently, we estimated all fractions in the Petri net model from these three profiles. Indeed, the simulation shows that the fractions for none of the transitions could be precisely determined (Figure 3.5 and Appendix Figure 3.13). Compared to the previous results there is a clear shift to higher estimation errors (Appendix Figure 3.11). Despite the low accuracy of the estimated fractions, the model metabolite profiles for the venous blood places show a good approximation to the reference profile (Appendix Figure 3.14). The model profiles for, for example, the gut places now show much higher variability since these were not constrained by gut

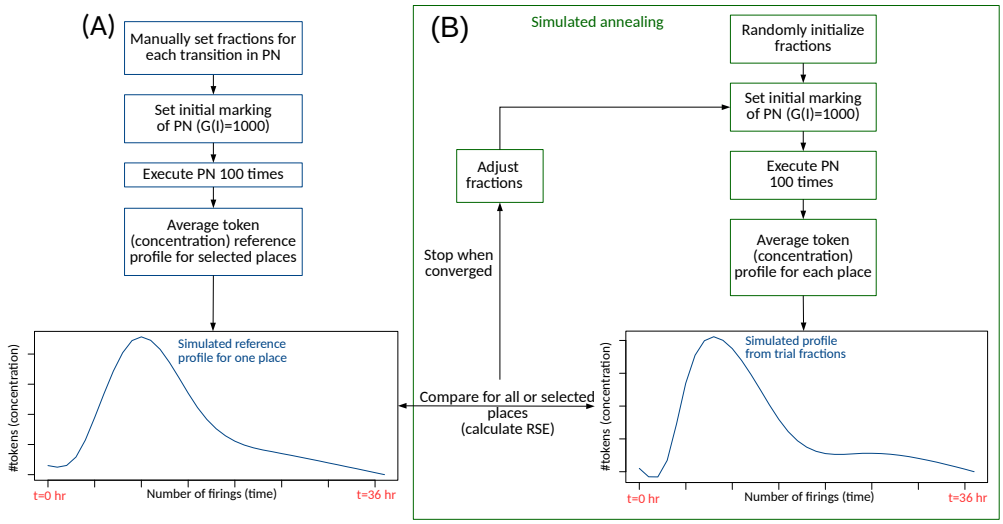


Figure 3.2: **Simulation of reference profiles and estimation of fractions from reference profiles.** (A) The Petri net model is used to simulate reference concentration profiles by manually defining a model configuration (set of fractions) and, subsequently, executing the Petri Net. The number of transitions that is required to remove all tokens from the model is set to 36 hours. (B) Simulated annealing (SA) is used to find fractions that generate model profiles that reproduce the reference profiles (simulated or experimental) from selected places. The Petri net is configured by randomly initializing the fractions. Subsequently the model is executed to generate simulated model profiles from the trial fractions. These profiles are compared to selected reference profiles by calculating of the root square error (RSE). Based on the comparison, the trial fractions are adjusted. This process continues until convergence to a set of optimal fractions, representing a specific network configuration, is reached.

reference profiles.

Next we estimated all fractions in the model from the experimental metabolite profiles of genistein, genistein-7-glucuronide, and genistein-7-glucuronide-4-sulphate measured in venous blood. In agreement with the results from the simulated data most of the fraction estimations show high or moderate variance (Figure 3.6, Appendix Figure 3.10). Results are even slightly worse compared to the previous results based on simulated profiles as a result of more noisy data. Even fractions F5 and F15 associated with the blood places could not be determined with high accuracy. Inspecting the model profiles we observe that the model profiles approximate the measured genistein and genistein-7-glucuronide to a lesser extent than the experimental profile for genistein-7-glucuronide-4-sulphate (Appendix Figure 3.15). Surprisingly, however, the estimation variance associated with transition F24 and F27 is low. Although we do not know the true fractions underlying our experimental data it seems that these two transitions take on more extreme fraction values (median values of 0.95 and 0.06 respectively; Appendix Figure 3.16) providing a more stringent constraint for our model compared to fractions that are closer to 0.5.

To test if a model initialized with more extreme fractions than the values we used so

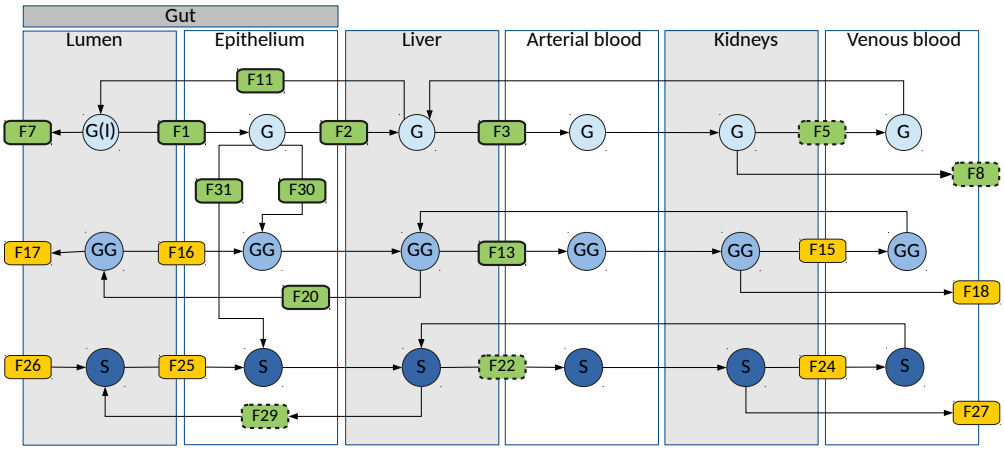


Figure 3.3: **Presentation of relative estimation errors (box outline thickness) and estimation variances (box colour) based on simulated reference profiles for all places (filled blue circles).** Green and orange correspond to low and medium estimation variances respectively. Boxes with thick outline represent determinant fractions (errors < 10%). Dashed boxes represent moderate fractions (errors between 10 and 25%). Boxes with thin outline represent flexible fractions (errors > 25%). Note that most low variance estimates correspond to 'determinant' fractions.

far (0.5 and 0.33) would provide more accurate estimates we performed another simulation. We initialized the fractions according to the median values obtained from the experimental data (Appendix Table 3.2). Figure 3.7 shows that there is no improvement of the estimations (Appendix Figure 3.17).

Fraction estimation from simulated reference profiles for gut and liver places.

Simulations based on reference profiles of all places showed that fractions associated with gut epithelium and liver places were determined with higher accuracy (Figure 3.3). We therefore asked if gut epithelium and liver measurements would provide more accurate fraction estimates compared to using three reference profiles from venous blood. Since gut and liver biopsy data were not available we generated reference profiles for three metabolites associated with determinant transitions, i.e., genistein in gut epithelium, and genistein and genistein glucuronide in liver. Fractions in our model were set either as 0.5 or 0.33 as previously. Results from the subsequent simulations indeed show that, in contrast to results obtained from simulated venous blood reference profiles, estimations from gut/liver profiles do not only allow to accurately determine several fractions associated with the reference profiles but also result in a better estimation of other fractions in the network (Figure 3.8, Appendix Figure 3.18). The overall distribution of estimation errors (Appendix Figure 3.11) confirms this improvement.

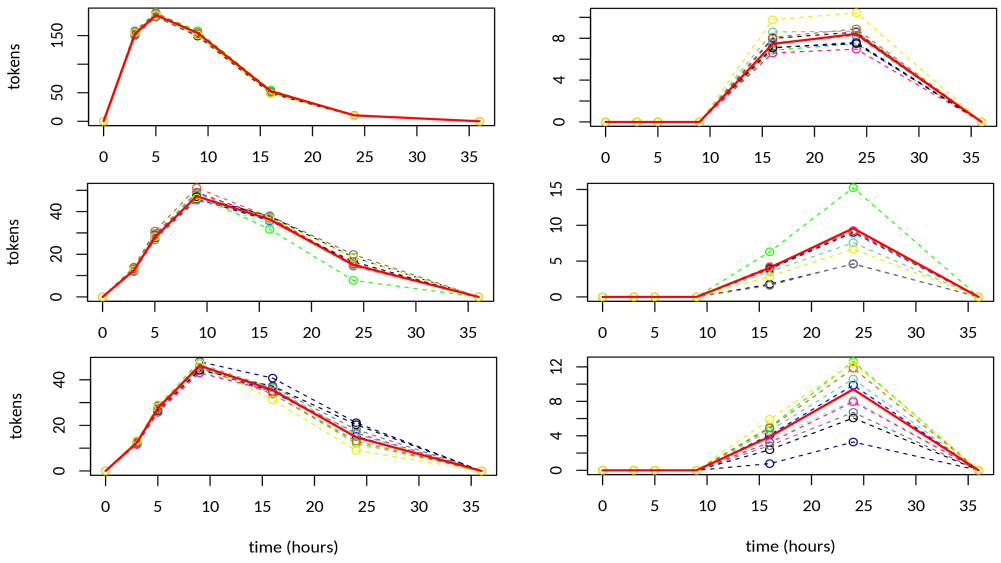


Figure 3.4: **Selected metabolite profiles generated from model shown in Figure 3.3.** The thick red line represents the simulated reference profile. Overall, the profiles generated from the 10 models produced by simulated annealing runs resemble the reference. Compared to the venous blood profiles, the profiles for gut epithelium are closer to the reference profile.

Inclusion of other constraints.

We performed a simulation to investigate if additional biological constraints would improve fraction estimations. Since it is possible to measure metabolites in urine it might be possible to constrain kidney fluxes which correspond to fractions F8, F18, F27 in our model. To test that we fixed the fractions F5/F8, F15/F18, F24/F27 to 0.5 and did not optimize them with simulated annealing. The simulation shows that the additional constraints essentially improve the estimation of the rest of the fractions (Figure 3.9 and Appendix Figures 3.19 and 3.20). Three fractions F2/F30/F31 have low variance and F31 is classified as moderate fraction. Variance also has decreased for most of the fractions and is classified as medium variance class.

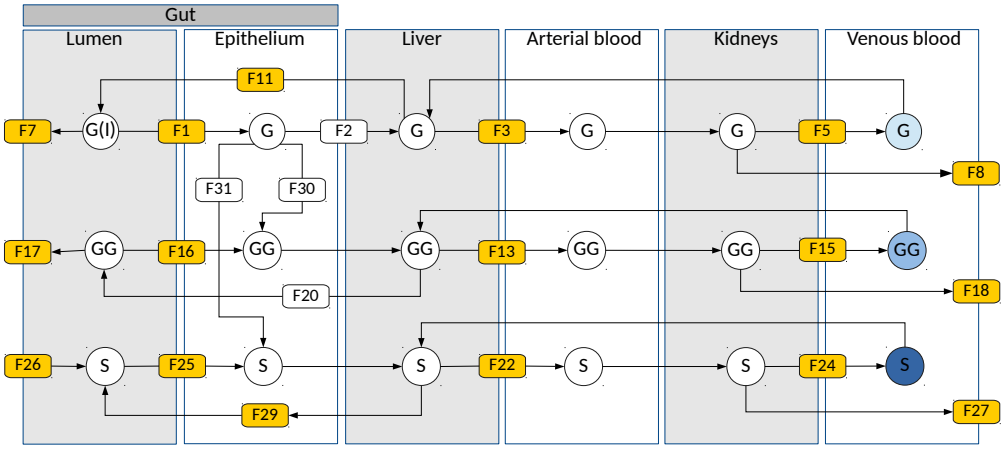


Figure 3.5: **Presentation of relative estimation errors and estimation variances (box colour) based on simulated reference profiles from only three venous blood places ((G, GG, S; filled blue circles).** Orange boxes correspond to medium estimation variances (no low and high variance class estimates). Based on the relative estimation errors all transitions were classified as 'flexible' (errors > 25%).

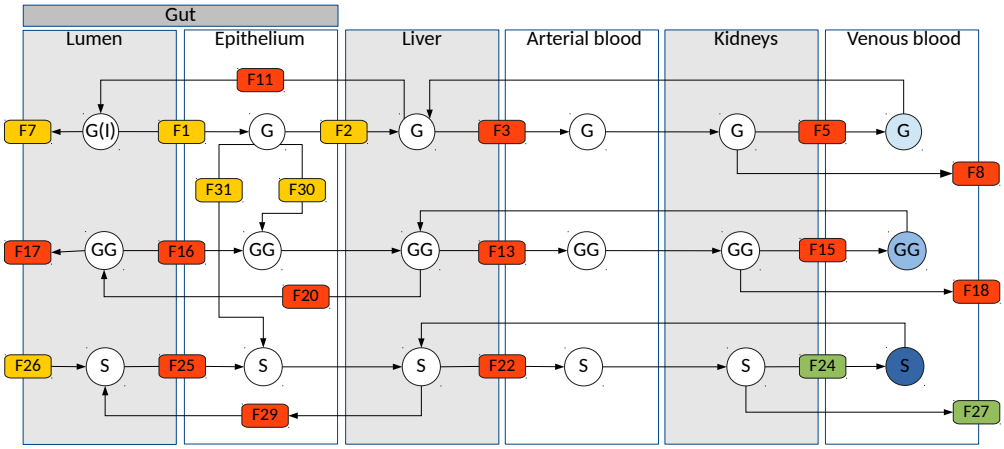


Figure 3.6: **Presentation of estimation variances (box colour) based on experimental reference profiles for three venous blood places (G, GG, S; filled circles).** Green, orange and red correspond to low, medium and high estimation variances respectively.

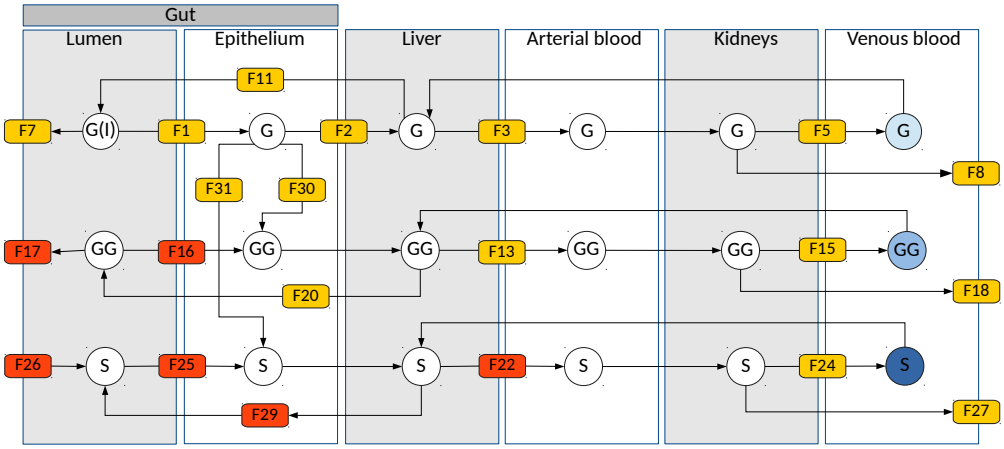


Figure 3.7: Estimated fractions from **Presentation of estimation variances (box colour) based on simulated reference profiles initialized with extreme fractions.** Reference profiles for three venous blood places were used (G, GG, S; filled circles). Orange and red correspond to medium and high estimation variances respectively.

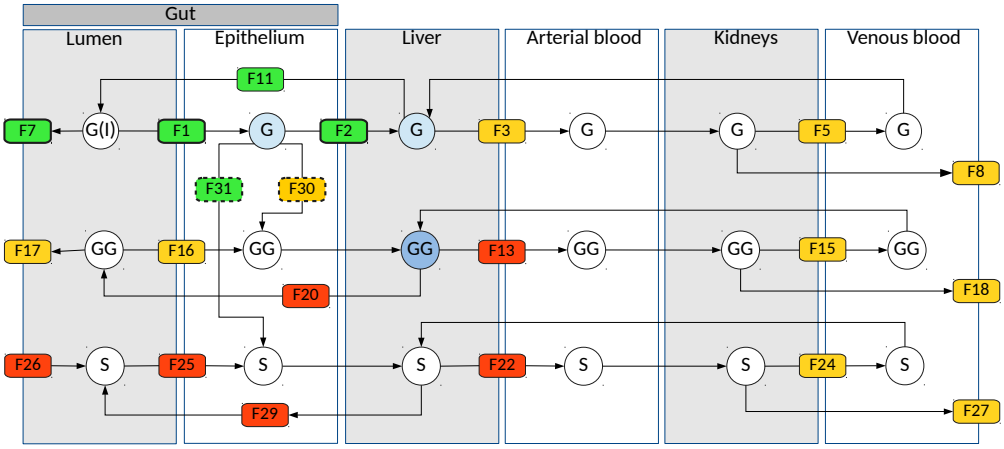


Figure 3.8: **Presentation of relative estimation errors (box outline thickness) and estimation variances (box colour) based on simulated reference profiles for three places (gut epithelium G, liver G, liver GG; filled blue circles).** Green, orange and red correspond to low, medium and high estimation variances respectively. Boxes with thick outline represent determinant fractions (errors < 10%). Dashed boxes represent moderate fractions (errors between 10 and 25%). Boxes with thin outline represent flexible fractions (errors > 25%).

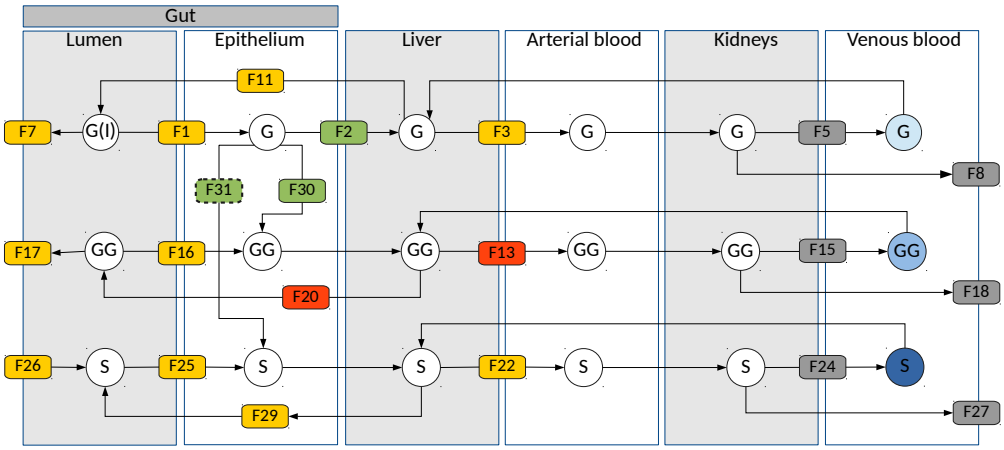


Figure 3.9: **Presentation of estimation variances (box colour) based on simulated reference profiles for all three venous blood places (G, GG, S; filled circles) and constrained outgoing routs.** Green, orange and red correspond to low, medium and high estimation variances respectively. Dashed boxes represent moderate fractions (errors between 10 and 25%). Boxes with thin outline represent flexible fractions (errors > 25%).

3.3. Discussion

Petri nets have developed into a popular tool to model biological networks to gain qualitative and quantitative understanding of its properties in the absence of information about kinetic parameters and stoichiometry. Lack of this information excludes use of methods such as kinetic models (e.g., [83]) or Flux Balance Analysis (FBA)[84], which assumes steady state while our model represents a perturbation of the steady state, i.e., healthy volunteers take genistein which is then secreted from the body. Dynamic FBA [85] would be an alternative for non-steady state systems but still requires network stoichiometry as well as concentration profiles for all metabolites in organs considered. In the current work we showed how network models can be used as an experimental design approach. In particular, we focused on the human multi-organ genistein elimination pathway which is of importance to gain more understanding of the (beneficial) health effects of genistein breakdown products. We used a standard Petri net approach but implemented a specific firing to ensure that experimental concentration profiles can be modelled. We integrated the Petri net model with simulated annealing to estimate the relative contributions (fractions) of the various paths in this network. Implicitly, our Petri net assumes first order kinetics. Each transition depends on only a single first-order metabolite. Alternative or future descriptions of the elimination pathway may therefore require changes in the Petri net model. However, the principles (and likely the results) shown in this paper remain valid.

The experimental profiles show a slight delay in onset (Appendix Figure 3.15). This is modelled in a natural way by the current Petri net. The parameterisation of the Petri net model imposes a similar delay on the tokens through the network. In principle, this network could also be modelled by delay differential equations but this requires the inclusion of additional parameters to account for these delays. Consequently, the estimation of the fractions from simulated or experimental data becomes more difficult (data not shown).

In the current work we used simulations to investigate to what extent the fractions of the genistein elimination pathway could be estimated from metabolic reference profiles from all or selected places. This was motivated by the observation that nutrikinetic studies are often based on metabolic measurements of easily accessible fluids such as blood and urine to gain additional insights in the underlying pathways. All presented results are based on ten optimization runs (i.e., ten sets of estimated fractions) and provide a worst case scenario since we did not leave out runs with high estimation errors and/or large variability between the estimations, which occasionally happens in global optimization problems. Overall, we observe that fractions in our model are mostly non-identifiable due to lack of sufficient constraints from data or prior knowledge. Increasing the number of SA iterations or using a local optimization method [86] following SA did not improve the results. This is a common problem in such studies [87, 40, 41].

To investigate if the estimation accuracy of a fraction correlates with the distance to the Petri net input G(I) or the distance to the closest output place we plotted this distances against the three accuracy classes (determinant, moderate, flexible; Appendix Figure 3.21). The figure reveals a relation between the distance to the input place and the assigned class. Particularly, places directly near the input place G(I) are all determinant. This indicates that fluxes in the beginning of mass distribution influence the con-

centration profiles the most. Also the figure shows that all fractions with low estimation accuracy (Flexible; F7, F17, F26, F8, F18, F27) are next to an output place or further away from the input G(I). In contrast, moderate fractions do not show an obvious relation to the distance. Although it is interesting to see this behavior it does not directly provide guidelines to improve the estimation. Since the input place is fixed this excludes the possibility to change the distance between the input and the fraction to be estimated. This is also true for the output places. However, as we already argued, additional constraints on the outgoing transitions may improve the estimation accuracy (e.g., Figure 3.9).

As expected our results show that none of the fractions could precisely be estimated from simulated blood profiles of genistein, genistein-7-glucuronide, and genistein-7-glucuronide-4-sulphate. Fractions were non-identifiable (errors > 10%). In addition, there was considerable variability between the SA runs. Estimations from experimental blood profiles showed even worse results as a result more noisy data although, surprisingly, two fractions with more extreme values were estimated with higher precision. Based on simulations using metabolite profiles for all places we showed that still only 9 of 21 fractions could be correctly estimated (error < 10%). These estimates, as well as few others, also showed lower variability between the ten optimization runs. Most of the low-error estimates were associated with gut epithelium and liver genistein and genistein-7-glucuronide. Subsequent simulations showed that these three profiles are sufficient to estimate some of the gut/liver associated fractions within an error of 10%. Moreover, the results also revealed an overall decrease in estimation variances and estimation errors of all fractions in the network compared to using venous blood profiles.

Another way to improve results would be to incorporate additional biological constraints into the model. We showed that constraints on outgoing routes (presumably it is possible to know elimination flux of the compounds through kidney) would significantly decrease estimation variance overall and would allow to obtain more reliable estimates for route F31. In addition to use the Petri net as an experimental design tool we hope that our approach inspires new research towards the modelling of multi-compartment networks.

3.4. Conclusion

We demonstrated how Petri net simulations help gaining insight in the definition of new experiments required to allow more precise fraction estimations for the genistein elimination pathway. Unfortunately, the results from our simulation are not easily validated since this will require data from liver and other organs that is difficult to obtain from healthy individuals in nurtikinetic studies. However, future data from model organisms may come to rescue. Our approach can be used for experimental design for similar (multi-organ) metabolite elimination pathways. Although other types of pathways (e.g., signalling) require different (sometimes existing) Petri net implementations, the presented strategy is generic.

3.5. Materials and Methods

Experimental data

Venous blood metabolic profiles from 12 healthy individuals were obtained in a recent nutrikinetics study [82]. One week before and during the experiment these volunteers followed a diet with low level of polyphenols. Each individual took one genistein tablet, containing 30mg of genistein (G(I) in Figure 3.1). Plasma samples were collected at seven time points: 30 minutes before the intake of the genistein tablet, and 3, 5, 9, 16, 24, 36 hours after the intake. Genistein, genistein-7-glucuronid, and genistein-7-glucuronid-4-sulphate (included in our model) were measured in venous blood through targeted mass spectrometry. Due to large variability between the measured profiles of these individuals we used data from only a single individual (Table 3.1).

Table 3.1: **Experimental data used in the study.**

Metabolite	0h	3h	5h	9h	16h	24h	36h
genistein	7499	14615	47356	111118	55609	18977	0
genistein-7-glucuronid	0	1331	12464	2638	8241	2046	2160
genistein-7-glucuronid-4-sulphate	0	7540	47884	48228	59803	45840	11498

A Petri net model of genistein elimination pathway

In a basic Petri net a transition fires at the moment a token is present in the associated place. However, these basic (sometimes referred to as original or time-less) Petri nets cannot reproduce bell-shaped time-resolved metabolite profiles (Appendix Figure 3.15) because tokens (molecules) will be immediately transferred to subsequent places preventing molecules to accumulate in a specific place. To overcome this limitation various extensions have been developed in the past, such as Stochastic Petri nets [16], Time Petri Nets [17], Hybrid Functional Petri Nets [18]. However, these approaches are specifically tailored towards signaling networks and require kinetic or other physical/chemical parameters of the system, which are not available for the genistein elimination pathway.

In the current work we use a standard Petri net approach but defined a firing rule that allows places (representing metabolites in compartments) to reproduce metabolite profiles measured by LC-MS experiments in venous blood. Consequently, we can also use the model to simulate such profiles. Our Petri net model (Figure 3.1) was designed with information from domain experts and literature. Places reflect metabolites in specific compartments (organs, blood), and tokens reflect the amount of metabolite molecules as a measure for relative concentration. Transitions reflect degradation (F30, F31), elimination (F7, F17, F26, F8, F18, F27) or transport (other transitions) fluxes within or between these compartments. Fluxes, represented as fractions, associated with these transitions are estimated from experimental or simulated metabolite reference profiles using a global optimization method (simulated annealing; [88]). A flux is not represented by an absolute quantity but as the fraction of molecules that flows from one place to the next. If a place is associated with multiple outgoing transitions (e.g., F3 and F11 for G in liver) then fractions associated to these transitions sum to one. These fractions are

used as probabilities (see 'firing rules' below) for the selection of a specific transition that will be 'fired' in a simulation step. Therefore, by definition, each fraction assumes a value between 0 and 1 and the sum of fractions of transitions that leave the same place should sum to one. Thus, fractions F3 and F11 for the 'G' place in liver will be F3=0.5 and F11=0.5 if no preference for a specific transition is assumed. Consequently, the fractions represent relative contributions of transitions in our compartment network. Eight places are associated with only a single outgoing transitions. The corresponding fractions (F4, F6, F12, F14, F19, F21, F23, F28) are by definition fixed to one and are not estimated in the simulated annealing runs. Our Petri net consists of 31 transitions and 19 places distributed over six compartments (organs and blood). The model represents an open and non steady-state system. Genistein G(I) enters the elimination pathway at the first day of our simulation and the simulation stops (return to steady-state) when all genistein products (tokens) have disappeared from the Petri net.

The distribution of tokens over all places at time point t is called the 'marking' of the Petri net model and is denoted by m_t. The marking of the initial state of simulation ($t = 0$) is called the initial marking and is denoted by m_0. Since the number of G(I) tokens for the initial marking affects the reproducibility of Petri net executions we first determined the optimal number of tokens. We executed 100 Petri net for 20, 50, 100, 1 000, 10 000, 50 000, 100 000 and 1 000 000 input tokens for G(I). Note that the Petri net is executed 100 times due to the probabilistic nature of the firing rules (see below). For each set of 100 models and each place we determined the mean profile and, subsequently, calculated the root square error (RSE) as a measure for the variance of 100 profiles. Since the RSE depends on the magnitude of the generated profiles we normalized the RSE by the initial marking. Based on the results (Appendix Figure 3.22) we choose 1000 tokens for the initial marking in our simulation experiments. A larger number of tokens did not significantly decrease the variability between Petri net executions but it significantly increased computation time.

Firing rules

A Petri net firing rule determines when and how many tokens are transferred from one place to the next. To allow the Petri net to generate bell-shaped metabolic profiles for each place we implemented a two-stage firing rule. The first stage comprises a probabilistic selection of a place based on the number of tokens in each place. The probability for a place being selected is $\frac{n}{n_{total}}$, where n is the number of tokens in a place and n_{total} is the total number of tokens still in the model (which may be less than the initial number of tokens in G(I)). This selection ensures that places with a low number of tokens have less chance of being selected and, consequently, have more time to acquire additional tokens. If a single transition is associated with the selected place then this transition will always fire. If multiple transitions are associated then, as a second stage, one of the outgoing transitions is probabilistically selected to be fired. This selection is made based on the fractions assigned to these transitions. Firing of a transition will always move one token to the next place.

Execution of the Petri net involves repeated rounds of selection and firing and stops when all input tokens (G(I)=1000) have left the model through one of the outgoing transitions (elimination through gut lumen or kidneys). The precise number (Nt) of transitions that will fire during a simulation depends on the probabilistic process of place

and transition selection. By defining the first firing event to occur at t=0 hours and event Nt to occur at 36 hours (the last time point from the experiment profiles), we define the correspondence between Petri net execution steps (firings) and the measurement time (Figure 3.2).

Parameterization of the Petri net model

Given experimental or simulated metabolite reference profiles the challenge is to find the Petri net configuration (set of fractions) that reproduces these profiles. We approached this as a global optimization problem [89] and used simulated annealing [88] as the optimization method of choice (Figure 3.2B). Simulated annealing is an iterative procedure dedicated to find a set of parameters (fractions) that minimizes a chosen error function. In our application the difference between reference profiles and profiles produced by simulated annealing are minimized.

Simulated annealing starts with randomly initializing the fractions constrained to the condition that the sum of the fractions associated with outgoing transitions of the same place sum to one. Subsequently, the Petri net is executed 20 times to generate 20 token profiles from which average profiles are calculated for each place. These simulated annealing trial profiles (S) are compared to the model reference profiles (R) of all or selected places through the calculation of an error value based on the root square error (RSE):

$$RSE = \sum_{p=1}^{P} \sum_{t=1}^{T} \sqrt{(S_{tp} - R_{tp})^2} \tag{3.1}$$

where the first sum runs over all selected P places and the second sum runs over seven time points T.

To compare simulated annealing trial profiles (S) to experimental data we have to modify the error function. The experimental data represents relative concentrations measured by mass spectroscopy. Therefore peak heights among metabolites are not directly comparable. Using this data we were more interested in reproducing peak positions and peak shapes than peak heights. Accordingly we modified Equation 3.1 in such a way that it used relative concentrations and compared the shape of simulated and experimental profiles. To be exact, S_{tp} and R_{tp} became relative concentrations $X_{tp_{rel}}$ and were calculated by Equation 3.2

$$X_{tp_{rel}} = \frac{X_{tp}}{\sum_{t=1}^{T} (X_{tp})} \tag{3.2}$$

Once an initial error (RSE) value is determined, a new set of trial fractions is generated according to:

$$f_{new_i} = f_{old_i} + T * N_i(0, 0.1) \tag{3.3}$$

where f_{new_i} and f_{old_i} are the new and the previous values of fraction i. T is the SA temperature (see below) and N_i is a random number drawn from a normal distribution with mean=0 and standard deviation=0.1. After generation of the new fractions, the fractions of outgoing transitions corresponding to the same place are re-scaled such that their sum is one.

Given the new set of trial fractions the Petri net is again executed and the error calculated. If, compared to the previous iteration, the error value decreased then the new set of fractions is an improvement and accepted. Otherwise, the new fractions are accepted with a probability calculated by the Boltzmann criterion.

$$e^{-\Delta E/T} > U(0,1) \tag{3.4}$$

where ΔE is the difference between the previous and current error value (positive if the error increased), T is the temperature, and U is a random number drawn from a uniform distribution. The probability for accepting fractions that increase the RSE depends on the magnitude ΔE of this increase, and the temperature. Temperature at the start of a simulation run is generally set to a high value such that trial fractions that increase the RSE can be accepted to allow escapes from a local minima. During the optimization procedure the temperature gradually decreases according to a geometric cooling scheme ($T_{new} = \alpha * T_{old}$) ($\alpha = 0.9$) decreasing the acceptation probability of trial fractions that increase the error. We chose the initial temperature such that 90% of new trial fractions were accepted. The temperature was lowered after generating 2500 new sets of fractions or after accepting 250 sets. This whole procedure of parameter generation, error function evaluation, and selection was repeated until convergence (less than 5% of the new parameter sets are accepted). The set of fractions corresponding to the lowest error value is reported and used to configure the final model.

We used SA Toolbox for Optimization (SATFO) [90]. Petri net simulation program was implemented in C++. SATFO and Petri net simulation program are publicly available at http://www.bdagroup.nl/content/Downloads/software/software.php.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

PR performed simulations. PR and AHCK performed data analysis and wrote the paper. All authors contributed to the study design, read and approved the final manuscript.

Acknowledgements

We are grateful to Prof. Dr. Stanley Brul, University of Amsterdam and Dr. Igor Bendik-Falconnier, DSM Nutritional Products for a constructive discussion on the model of genistein elimination pathway. We thank Dr. Igor Bendik-Falconnier and Dr. Ric de Vos (Plant Research International, Wageningen) for providing us the experimental data. We also would like to thank Prof. Jaap Heringa and Dr. Anton Feenstra from Centre for Integrative Bioinformatics, Free University, Amsterdam for a helpful discussion on Petri nets theory and assistance in the implementation of Petri net simulation program.

This research is funded by The Netherlands Bioinformatics Center (NBIC).

3.6. Appendix

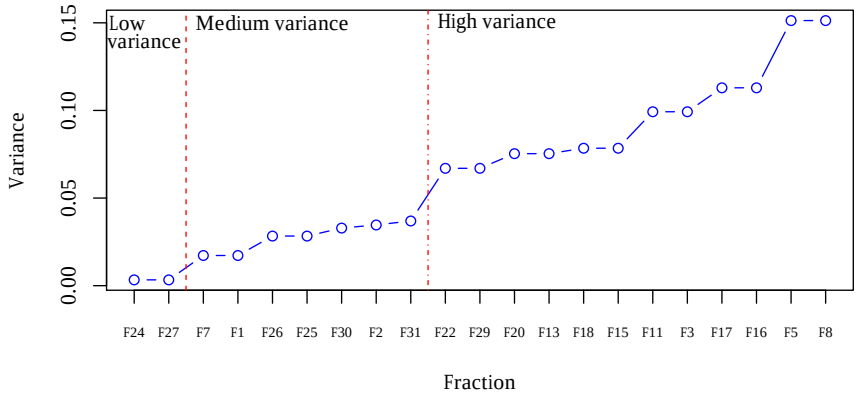


Figure 3.10: **Variance of estimated fractions from experimental reference profiles for three venous blood places (G, GG, S).** Variances are based on 10 repeated simulated annealing runs. The fractions associated with the 21 transitions (x-axes) are sorted according to estimation variances (y-axes). To facilitate comparison between different simulations we defined three broad variance classes based on visual inspection of this plot: low variance (<= 0.01), medium variance (> 0.01 and <= 0.06), and high variance (> 0.06).

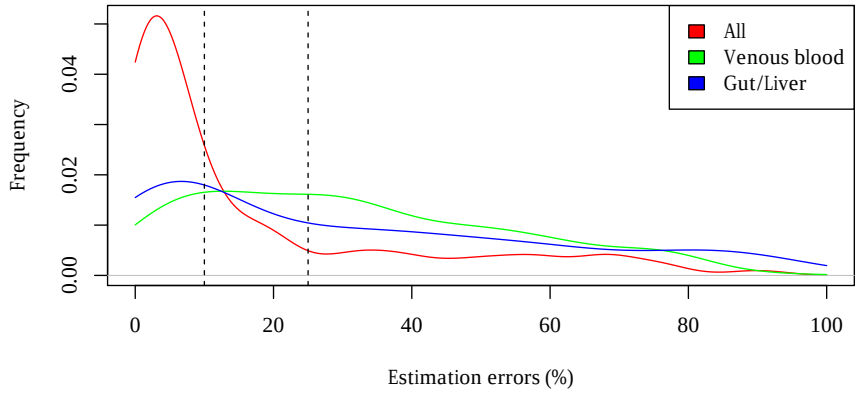


Figure 3.11: **Distribution of estimation errors of fractions for all transitions in our model.** Red: estimations errors based on simulated profiles for all places. Green: estimations errors based on simulated profiles three venous blood profiles. Blue: estimations errors based on simulated profiles associated with gut and liver metabolites (G gut epithelium, G liver, and GG liver). Dashed lines indicate 10% and 25% error defining three classes of transitions. The distributions show that availability of metabolite profiles for all places give a shift to smaller estimation errors. If only three reference profiles are available (3 venous blood places or 3 gut/liver profiles) then the gut/liver profiles give a slight advantage to estimate all fractions in the model.

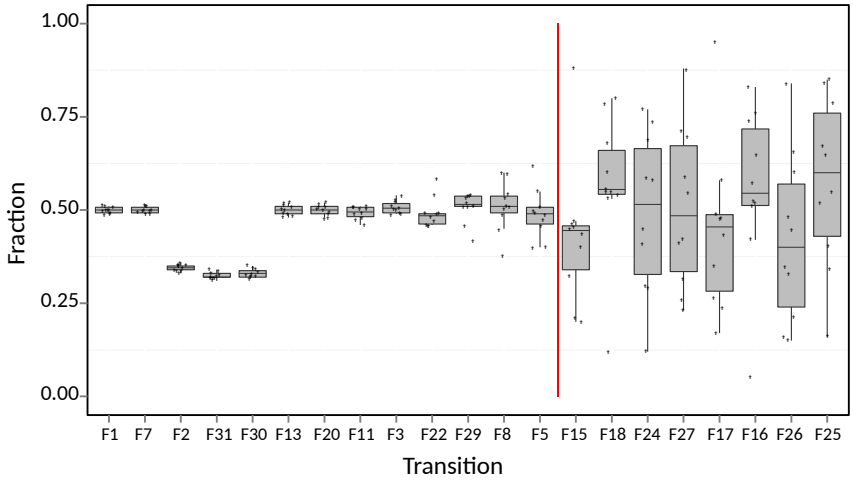


Figure 3.12: **Estimated fractions from simulated reference profiles for all places.** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Fractions in this model were set of 0.5 except for fraction F2, F31 and F30 which were set to 0.33. The boxes define the 25% and 75% quartiles. The box line is the median value. The whiskers represent the 1.5 inter-quartile range. The transitions are sorted according to the variance of the fractions. The 13 transitions left to the red line correspond to the low variance class. The other boxes correspond to the medium variance class. No transitions were classified in the high variance class.

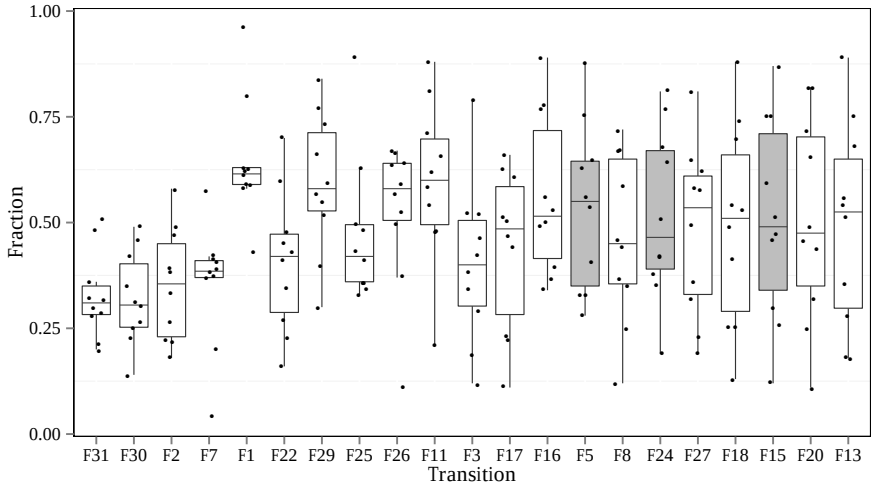


Figure 3.13: **Estimated fractions from simulated reference profiles for three venous blood places (G, GG, S).** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Grey boxes represent fractions corresponding to transitions associated with blood. Fractions in this model were set of 0.5 except for fraction F2, F31 and F30 which were set to 0.33. The boxes define the 25% and 75% quartiles. The middle box line is the median value. The whiskers represent the 1.5 inter-quartile range. The transitions are sorted according to the variance of the fractions. None of the transitions were classified to the low variance class.

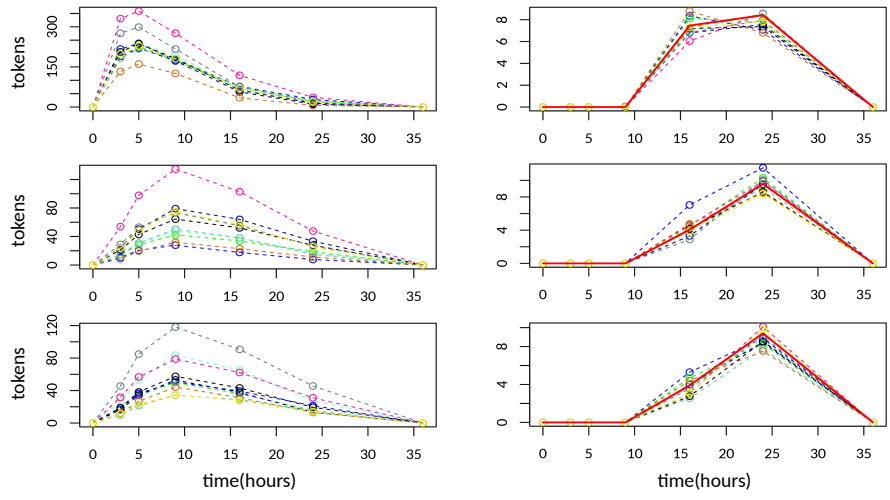


Figure 3.14: **Selected metabolite profiles generated from model shown in Figure 3.5.** The thick red line represents the simulated reference profiles for G, GG and S in venous blood. Overall, the profiles generated from the 10 models produced by simulated annealing runs resemble the reference. The profiles for gut epithelium show more variability since these were not constrained by reference profiles for gut places.

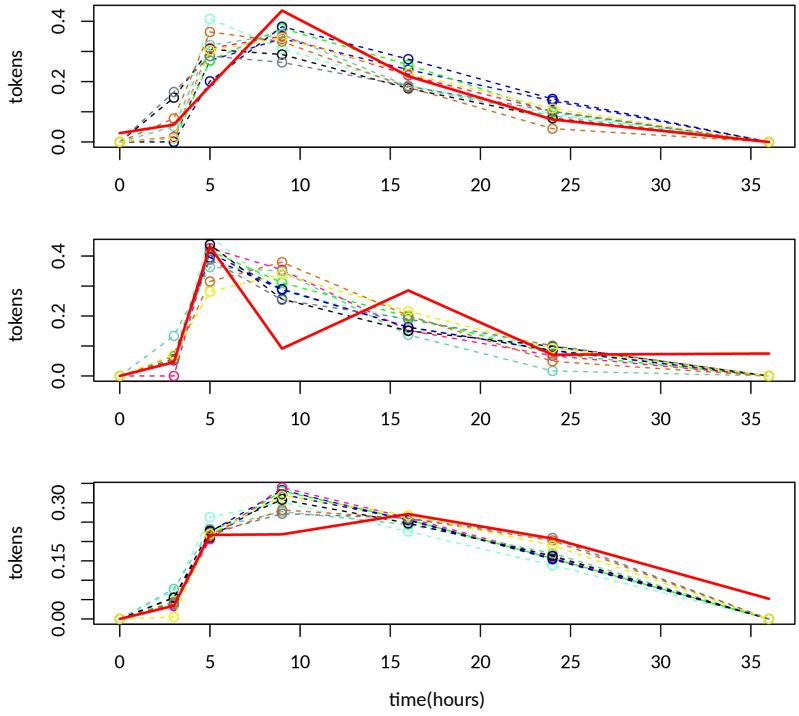


Figure 3.15: **Metabolite profiles generated for venous blood places (G, GG and S) from model shown in Figure 3.6.** The thick red line represents the experimental data for a single individual. Larger variance is observed between the genistein and genistein-7-glucuronide model profiles. Profiles for genistein-7-glucuronide-4-sulphate show less variability and approximate the experimental profile to a larger extent.

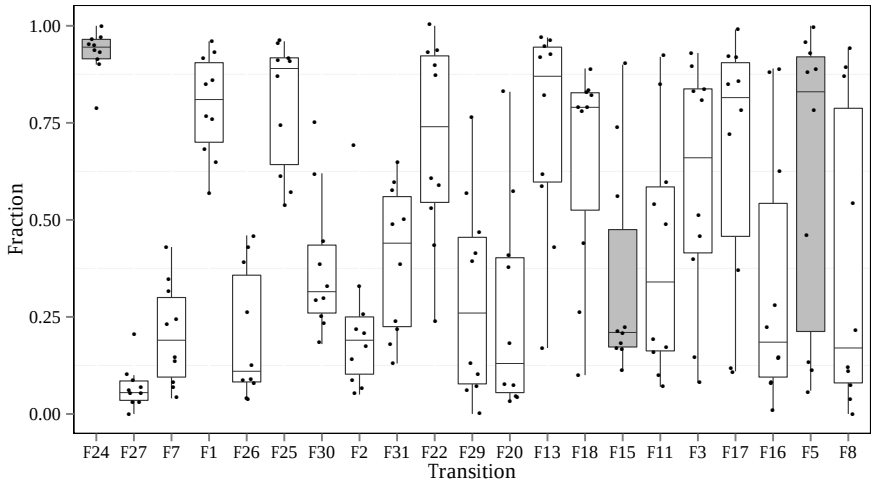


Figure 3.16: **Estimated fractions from experimental reference profiles measured for three venous blood places (G, GG, S).** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Grey boxes represent fractions corresponding to transitions associated with venous blood. The boxes define the 25% and 75% quartiles. The middle box line is the median value. The whiskers represent the 1.5 interquartile range. The transitions are sorted according to the variance of the fractions. Transition T24 and T27 were classified as low variance and are associated with more extreme fraction values (median values of 0.95 and 0.06 respectively).

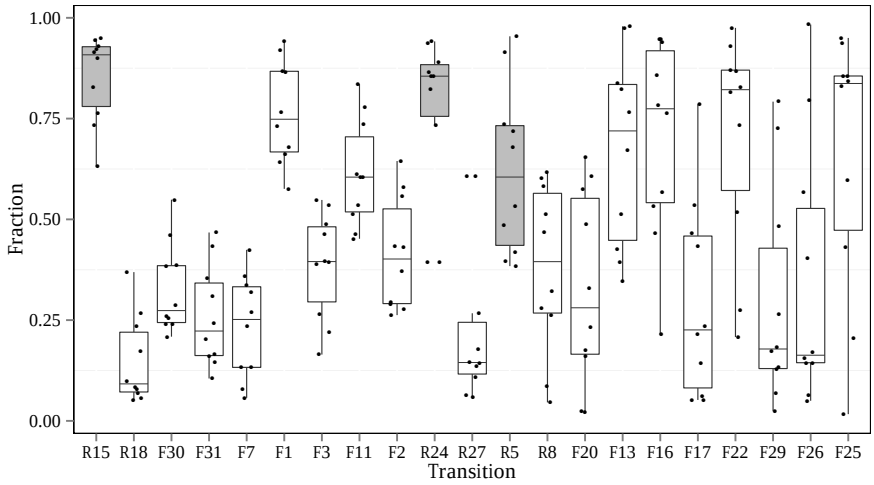


Figure 3.17: **Estimated fractions from simulated reference profiles initialized with extreme fractions and reference profiles measured for three venous blood places (G, GG, S).** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Grey boxes represent fractions corresponding to transitions associated with venous blood. The boxes define the 25% and 75% quartiles. The middle box line is the median value. The whiskers represent the 1.5 inter-quartile range. The transitions are sorted according to the variance of the fractions. The 6 transitions right to the red line correspond to the hight variance class. The other boxes correspond to the medium variance class. TODO: the red line should be drown between F13 and F16.

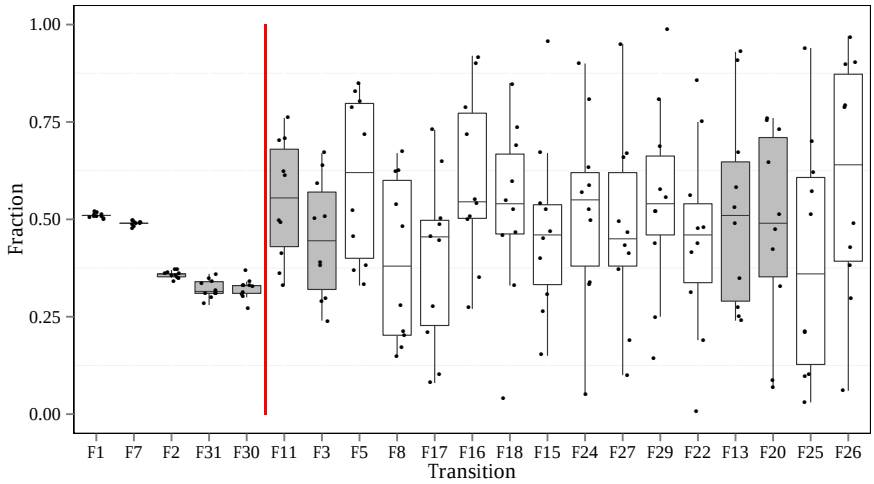


Figure 3.18: **Estimated fractions from simulated reference profiles for three places (gut epithelium G, liver G, liver GG).** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Grey boxes represent fractions (F1, F2, F31, F30, F11, F3, F13, F20) associated with the three gut/liver places. Fractions in this model were set of 0.5 except for fraction F2, F31 and F30 which were set to 0.33. The boxes define the 25% and 75% quartiles. The box line is the median value. The whiskers represent the 1.5 inter-quartile range. The transitions are sorted according to the variance of the fractions. The 5 transitions left to the red line correspond to the low variance class. The other boxes correspond to the medium or high variance class.

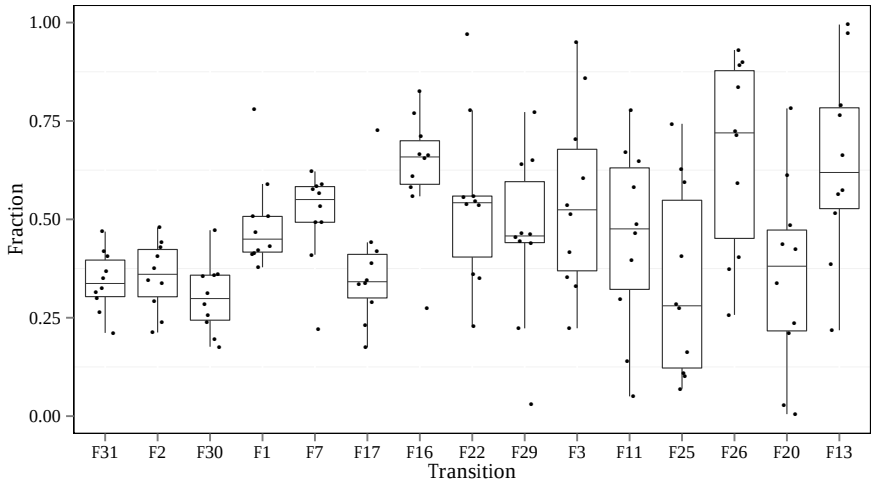


Figure 3.19: **Estimated fractions from simulated reference profiles for all three venous blood places (G, GG, S) and constrained outgoing routs.** Each box plot shows the fractions of 10 SA runs (dots) corresponding to the 21 transitions in our model. Fractions F5/F8, F15/F18, F24/F27 were set to 0.5. Other fractions were estimated by simulated annealing. The boxes define the 25% and 75% quartiles. The middle box line is the median value. The whiskers represent the 1.5 inter-quartile range. The transitions are sorted according to the variance of the fractions.

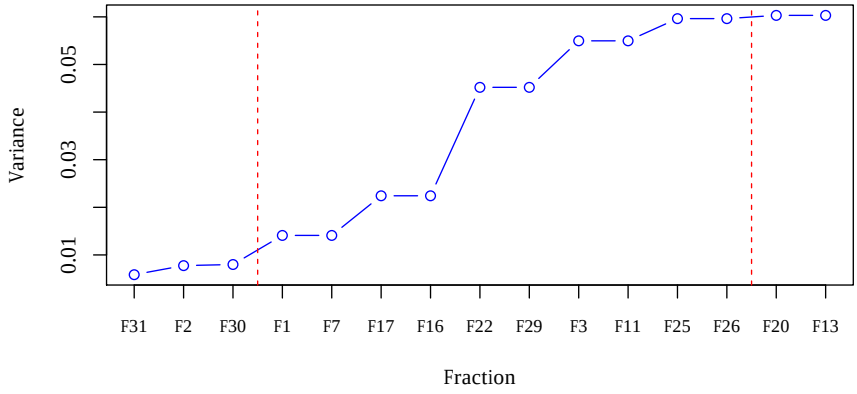


Figure 3.20: **Variance of estimated fractions from simulated reference profiles for all three venous blood places (G, GG, S) and constrained outgoing routs.** Variances are based on 10 repeated simulated annealing runs. Fractions F5/F8, F15/F18, F24/F27 were set to 0.5. Other fractions (x-axes) were estimated by simulated annealing and are sorted according to estimation variances (y-axes).

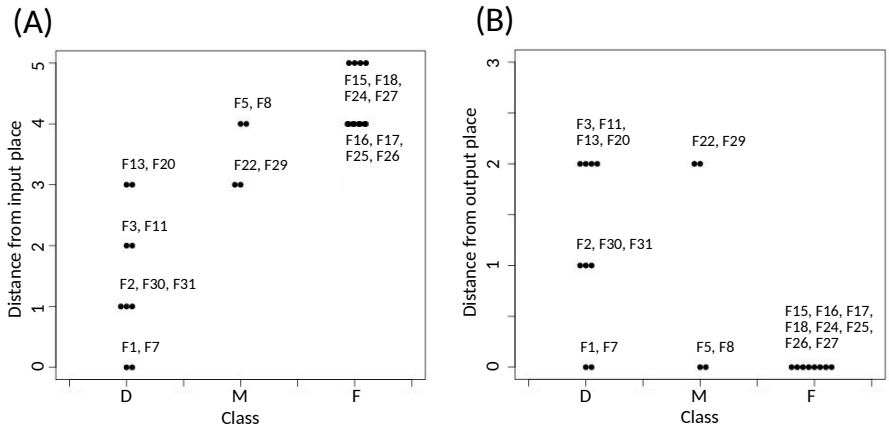


Figure 3.21: **Correlation between fraction estimation error classes (D: determinant, M: moderate, F: Flexible) and distance of corresponding transition to input place G(I) (A) or the closest output place (B).** The distance is defined as the minimum number of transitions between the input or the closest output place and the transition for which accuracy of the fraction was estimated. The figure shows the estimated accuracies for the 21 fractions in the Petri Net, which were estimated from the simulation represented in Figure 3.3.

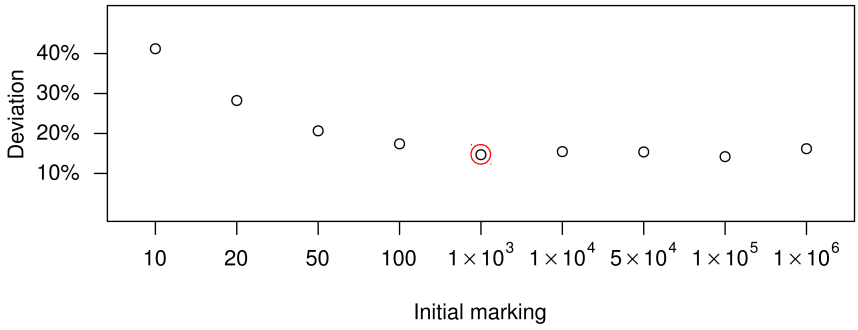


Figure 3.22: **Deviation of RSE mean.** Petri net simulations were performed 100 times for each initial marking. The deviation did not decrease from the initial marking of 1000 and therefore we chose it for further simulations.

Table 3.2: **Median fraction values.** Fractions were obtained from estimations based on experimental reference profiles for three venous blood places (corresponding to Figure 3.6 and Appendix Figure 3.15). Rounded values were used to initialize the model and generate reference profiles to test a set of extreme fractions (optimization correspond to Figure 3.7).

Fraction	Median
F1	0.81
F7	0.19
F2	0.19
F30	0.31
F31	0.50
F11	0.34
F3	0.66
F5	0.83
F8	0.17
F17	0.81
F16	0.19
F13	0.87
F20	0.13
F18	0.79
F15	0.21
F26	0.11
F25	0.89
F22	0.74
F29	0.26
F24	0.95
F27	0.05

Chapter 4

Computational model reveals limited correlation between germinal centre B-cell subclone abundancy and affinity: implications for repertoire sequencing[3]

Immunoglobulin repertoire sequencing strategies are successfully applied to identify expanded antigen-activated B-cell clones that play a role in the pathogenesis of immune disorders. These clones comprise lineages of subclones comprising variants within a VJ family produced by somatic hypermutation. Their B-cell receptor binding affinities for the antigen are higher than affinities of the naïve B cells in the background population, which is a direct consequence of the higher initial affinity of the activated B-cell(s) and the subsequent affinity maturation process in the germinal centre (GC). However, repertoire sequencing only provides information about subclone abundancies and not about their affinities. Consequently, although repertoire sequencing successfully identifies (sub)clones involved in disease, the selection of the most abundant (i.e., expanded) subclone(s) within of a clonal family may not necessarily be the highest affinity subclone. Unfortunately, the determination of affinities of many subclones within a clonal family is virtually impossible and, consequently, the relation between abundancy and affinity remains to be established. Knowledge about affinities within clonal families would likely improve the selection of relevant subclones for further characterization and antigen screening. Therefore, to gain insight in the putative affinity distribution among (un)expanded subclones we developed a computational model that simulates affinity maturation in a single GC while tracking individual subclones in terms of abundancy and affinity. We show that the model correctly captures the overall GC dynamics, and that the amount of expansion is qualitatively comparable to expansion observed from B cells isolated from a normal human lymph node. Analysis of the fraction of high- and low-affinity subclones among of the unexpanded and expanded subclones reveals a only partial correlation between abundancy and affinity and that the low abundant subclones are of highest affinity. Thus, our

[3] P. Reshetova, B.D.C. van Schaik, P.L. Klarenbeek, M.E. Doorenspleet, R.E.E. Esveldt, P.P. Tak, J.E.J. Guikema, N. de Vries, A.H.C. van Kampen. Manuscript in preparation.

model suggests that selecting highly abundant subclones from repertoire sequencing does not automatically provides us the highest affinity B cell. We conclude that although selection of highly abundant subclones from B cell repertoires provides us with B cells involved in (auto)immune disorders, the utility of repertoire sequencing might be even further improved by following selection strategies that do not merely consider subclone abundance.

4.1. Introduction

The adaptive immune system is a key component of our defense against pathogens and comprises highly specialized cells and processes. Its humoral component is responsible for memory B-cell formation and high-affinity antibody (Ab) production resulting from affinity maturation in germinal centers (GCs) [91, 92]. During this process GC B cells undergo multiple rounds of proliferation, somatic hypermutation (SHM), and selection to improve their affinity for the given antigen (Ag). This results in a dynamic ensemble of low and high affinity B-cell subclones comprising variants of a clone within a VJ family produced by SHM. Higher affinity cells have increased chance to be positively selected for further rounds of proliferation and SHM, or for differentiation to memory and plasma cells.

Repertoire sequencing using high-throughput sequencing enables the determination of T- and B-cell repertoires in (clinical) samples by sequencing the expressed V, D, and J gene segments [93, 94, 95, 96]. Immune responses typically involve the initiation and coexistence of up to several hundreds of GCs, which emerge over an extended period of time [97, 98, 99]. Consequently, B-cell repertoire sequencing of clinical samples typically identifies (sub)clones originating from a multitude of Ag-activated B cells and GCs, or even from responses to multiple Ags. Despite this complexity we and others successfully used repertoire sequencing for the identification of (autoreactive) B cells involved in immune disorders while relying on the assumption that expanded clones play a key role in the pathogenesis of the disease [100, 101]. An expanded clone comprises a B-cell lineage of related subclones varying in abundance and number of acquired somatic mutations. B-cell receptor binding affinities for the antigen of these subclones are expected to be higher than affinities of the naive B cells in the background population, which is a direct consequence of the higher initial affinity for the Ag of the activated B-cell(s) and the subsequent affinity maturation process. However, repertoire sequencing only provides information about subclone abundancies and not about their affinities. Consequently, although repertoire sequencing successfully identifies clones involved in disease, the selection of the most abundant (i.e., expanded) subclone(s) within a clonal family may not correspond to the highest affinity subclone. Unfortunately, the determination of affinities of the many subclones within a clonal family is virtually impossible with current experimental technologies. Since cells are destructed in the sequencing experiment, affinity analysis requires either labor-intensive selective cloning of the individual B cells, or equally labor-intensive expression of single BCRs in cloning systems. Currently, these requirements prohibit high-throughput analysis of affinity of sequenced cells. Moreover, measurement of affinities also require knowledge of the Ag which is often not available for clinical samples. Only part of the subclones within a clonal lineage reaches high cell counts and these are typically selected for further characterization and Ag screening [100]. Given the nature of affinity maturation one would expect that high

abundant subclones are of highest affinity, which would argue for the selection of the most abundant subclone as a viable strategy. However, since the relation between abundancy and affinity is unknown, it cannot be excluded that a fraction of the large subclones are of low affinity and *vice versa*. Therefore, alternative (affinity-based) selection strategies might be of added value for downstream analysis.

In this work we developed a computational model of a single GC to gain insight in the putative affinity distribution among expanded and unexpanded subclones identified by B-cell repertoire sequencing. Inspired by existing models of affinity maturation (e.g., [102, 103, 104, 105]), we implemented a mathematical model that comprises a large evolving set of ordinary differential equations (ODEs) providing information about the abundancy and affinity of individual subclones emerging during the GCR. We did not use one of the existing models since existing ODE models do not track individual subclones while agent based models (e.g., [105]) are faced with the additional complexity of GC spatial dynamics which we aimed to avoid. Moreover, most models are not available as a software implementation.

We show that our computational model is in agreement with the typical GC dynamics. We also show that the amount of expansion of selected B-cell lineages from repertoire data acquired from a human lymph node is qualitatively comparable to the level of expansion observed in the simulated data. Given this support for our model, we subsequently inspected the affinities and abundancies of the individual subclones from the simulations, and found that the expanded and unexpanded B-cell subclone compartments each comprise a mixture of high and low affinity cells, i.e., there is only partial correlation between affinity and abundancy of subclones within a clonal family. Moreover, the low abundant subclones were of highest affinity. Consequently, although repertoire sequencing enables the correct identification of expanded clones involved in immune disorders, the subsequent selection of high-abundant subclones within a clonal family does not necessarily result in the highest affinity subclone. Although at this stage these results cannot be experimentally validated, we argue that considering only subclone abundance might not be the most optimal strategy to select candidate B-cell clones for further characterization and Ag screening.

4.2. Material and Methods

Sample and experimental data

We selected a single sample for analysis and comparison to the simulation results. This sample represents leukocytes isolated from a lymph node from an otherwise healthy human individual, without ongoing infection (represented in biochemical parameters such as C-reactive protein). The sample was taken as described earlier [106]. Repertoire sequencing was performed as described in [93] using the Roche 454 sequencing platform to generate 7771 reads (6777 unique reads). Processing of the sequence data was performed as described in [93]. In brief, reads from a multiplexed sequence run were separated by their multiplex identifiers (MID) and aligned against the IMGT database [107] with BLAT [108] to identify the corresponding V and J segments. Subsequently, each read was translated to a peptide sequence and the CDR3 sequence was determined by identifying conserved motifs in the V and J segment that delineate the CDR3 [109]. Conse-

quently, only in-frame reads were used. Sequences with uncalled bases in their CDR3 region were excluded from analysis. This resulted in 4454 unique subclones (clones within a VJ family defined as a peptide with a unique V and J assignment, and unique CDR3 sequence). This amount of sequence reads is sufficient to represent capture most (expanded) subclones but may miss subclones occurring at very low frequencies. A full analysis and presentation of this and other lymph node samples we have will be part of future paper.

The mathematical model

We developed a mathematical model using ordinary differential equations (ODEs) to describe the dynamics of individual subclones during the GCR. This model is implemented in the R statistical environment version 3.2.2 [110] using R packages deSolve (version 1.12) [111], R6 (version 2.1.2), ggplot 2.0 and beeswarm 0.2.1. The software is available as open source (GPLv3) on request from the author.

Overall simulation setup

Our simulation framework represents a simplified but adequate model of the GCR [92, 91] (Figure 4.1). Briefly, prior to the GCR, B cells and T cells are activated by recognition of their cognate antigen in the primary follicle and T-cell zone respectively (day -2 in Figure 4.1). Activated B cells and T cells migrate to the interfollicular region and interact resulting in the full activation of B cells while the T cells differentiate to T follicular helper cells (Tfh). Two days after immunization the GCR is initiated (day 0 in our simulation) with the Tfh cells and activated B cells migrating into the follicle, which is characterized by a network of follicular dendritic cells (FDCs). Here, the B cells engage in a rapid monoclonal expansion to over 10,000 cells at day 4 forming the GC. During this expansion a dark zone comprising centroblasts (CBs) and a light zone comprising centrocytes (CC), FDCs and Tfh cells are established. The dark zone is the site of B-cell clonal expansion and BCR diversification through SHM producing novel subclones. The GC light zone is the site of positive B-cell selection through Ag and Tfh binding and signaling. Together, these processes are responsible for B-cell affinity maturation. SHM has been reported to start at day 7 post-immunization in mice [112]. Oprea and Perelson [102] assumed that the GC is initiated 3 days after immunization and, correspondingly, start SHM at day 4 of the GCR in their model. Others reported that SHM starts 2 days post-immunization [113], or even prior to GC formation [114]. Following Oprea and Perelson, we also start SHM at day 4 in our simulations. Following monoclonal expansion, memory cells and plasma cells are starting to be produced (day 4 in our simulation). Although the precise mechanisms and timing of the output cells are not well-understood [115], it has been proposed that initially mainly memory B cells are produced while at later stages the GCR is dedicated towards (higher affinity) long-lived plasma cells [116, 117]. In our model the production of memory and plasma cells starts at the same moment (day 4) but we made the rate of plasma cell differentiation dependent on the absolute affinity of the CCs resulting in a low plasma cell output during early stages of the GCR. Since we were not interested in the production of output cells, these are not further discussed in this paper. Our simulation starts at day 0 with three founder B cells (CBs) with different affinities, and terminates after 21 days, the life span of an average GC. Consequently, we

do not model GC shutdown since its mechanisms remain to be established. Our model does not explicitly includes the dark/light zones, Ags, FDCs, or Tfh cells since we are not interested in the spatial dynamics nor in the precise selection mechanisms but rather in modelling subclonal diversity, expansion, and affinity. Therefore, to avoid an overly complex model, we represent the Ag and Tfh survival signals with sigmoidal functions as explained below.

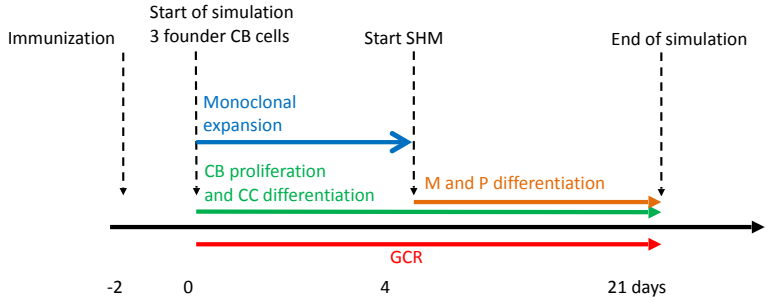


Figure 4.1: Simulation time line of the germinal center reaction. The GCR starts with 3 founder B cells (affinities 0.1, 0.2 and 0.3 μM) 2 days after immunization and continues for 21 days.

Somatic hypermutation, subclones, and affinity

The V, D, and J segments that make up the BCR cover four framework regions (FWRs) providing the Ab structural framework, and three Ag-binding complementary determining regions (CDRs) [118, 119]. Our model considers the FWR and CDR regions without an explicit nucleotide representation of the BCR but, instead, using a decision tree that decides on the fate of each individual SHM [103, 120, 121] (Figure 4.2). This tree involves probabilities for silent (synonymous mutations), lethal FWR, and affinity changing CDR mutations. The probabilities for replacement and silent mutations were determined from many mice germline sequences. The probability of the lethal mutations was based on studies that analyzed mutations patterns in real sequences. To determine the number of mutations during each CB cell division we defined the BCR to have a length of 600 nucleotides (i.e., one light and heavy chain). Given that the SHM rate is 10^{-3} per bp per division which this results in 0.6 mutations per division. We model this as a Poisson distribution $m = Poisson(\lambda = 0.6)$ and, consequently, each cell acquires 0, 1, or more mutations after each cell division. The mutation decision tree distinguishes the CDRs and their surrounding structural FWRs but does not differentiate between CDR1, CDR2 and CDR3 [119, 118].

In repertoire sequencing one is usually interested determining the population of (sub)clones in an immune response. Each of these subclones has its own binding affinity for the Ag. Since the CDR3 region is the main determinant in Ag-binding, one generally defines and discriminates these subclones on the basis of their unique CDR3 peptide sequence (within a VJ family). Alternatively, we can also define a subclone as having a unique BCR nucleotide sequences (i.e., V-CDR3-J). In the first situation, only non-synonymous SHMs in the CDR3 region produce new subclones, while in the second

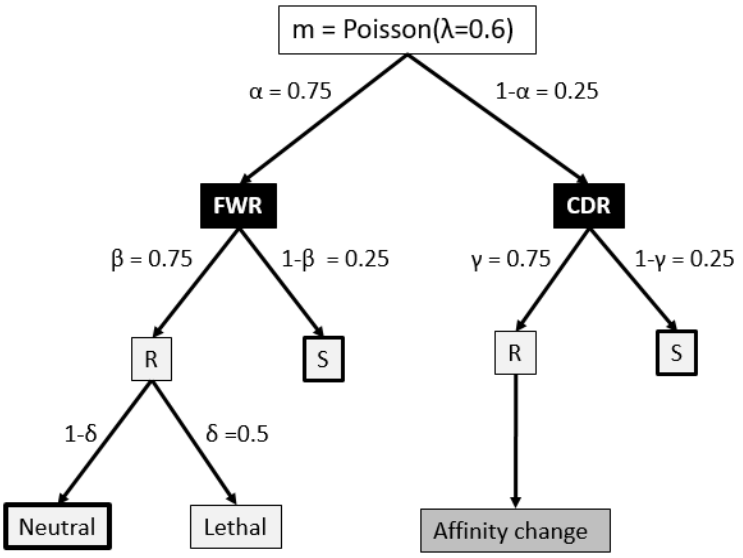


Figure 4.2: Fate of each somatic hypermutation. After each CB division the daughter cells are affected by $m \geq 0$ mutations affecting the framework region (FWR) with a probability of α or the complementary determining region (CDR). A mutation may replace (R) an amino acid of the Ig FWR or CDR region with probability β and γ respectively. A mutation in the FWR is lethal with probability δ. A replacement mutation in the CDR is neutral or changes the affinity of the subclone. Part of our simulations neglect mutations indicated by the thick boxes to produce subclones at the peptide level. Probabilities in this tree are according to [103].

situation each non-lethal SHM results in a new subclone. The mutation decision tree (Figure 4.2) is defined at the level of the nucleotide sequence and, consequently, in our simulation we implicitly define and track subclones at the nucleotide level throughout the GCR. Consequently, each SHM generates a new subclone that is initially represented as a single CB that subsequently proliferates and differentiates to co-exist as CB, CC, memory cell and plasma cell at succeeding time points. Alternatively, we may consider only CDR replacement mutations to define and track subclones at the peptide level. In this situation, each non-lethal replacement mutation in the CDR generates a new subclone. Since the tree does not specifically distinguish CDR3 from CDR1 and CDR2, our simulations at the peptide level effectively includes all three CDRs, which may give an overestimation of the number of unique clones compared to only considering the CDR3 as is done in repertoire sequencing experiments. However, since all three CDR regions are involved in Ag binding the simulation might be more realistic. Subclones with CB cell counts less than one (a result from using continuous differential equations; see below) are kept in our simulation but are not further be affected by SHM to avoid the generation of new subclones from these cells.

Each subclone in our model has a unique BCR with an absolute affinity σ that specifies the interaction strength with the Ag. The affinities of the three single cell founder CBs are set to arbitrary but different low affinity values (0.1, 0.3, and 0.5 μM). Three dif-

ferent values were chosen to establish an initial level competition between the founder cells. The magnitude of the initial affinities does not affect the dynamics of our model since this depends on relative affinities (see below). Only plasma cell output depends on absolute affinities. For each affinity changing mutation (Figure 4.2) the affinity of the affected subclone is updated according to $\sigma_{new\ subclone} = \sigma_{parent} + \Delta\sigma$ where $\Delta\sigma$ is drawn from a distribution $f(\sigma)$ with probability density function:

$$f(\sigma) = g(s,r) - \mu - (\sigma_{parent} * 0.1), \tag{4.1}$$

where $g(s,r)$ is the inverse gamma distribution with $s = 3$ and $r = 0.3$ representing the shape and rate parameter respectively. μ is the expected value of $g(s,r)$ and subtracted from $g(s,r)$ to center the distribution g around zero resulting in about equal chances for decreasing and increasing the affinity of mutated subclones. We used the gamma distribution because it is right skewed and, therefore, allows for a small chance for making larger affinity improvements representing key mutations [120, 122]. We do not distinguish between one or multiple affinity changing mutations. To account for the fact that mutations in higher affinity subclones have less chance to further improve affinity we shift distribution f to the left as a function of the parent cell affinity (Supplementary Figure S1). The distribution shape and rate parameters (3 and 0.3) and the affinity shift (0.1) were chosen by trial and error such to obtain the dynamics of a typical GC.

Positive and negative selection of subclones

Following cell division and SHM, the CBs differentiate to CCs which are programmed to undergo apoptosis (negative selection) unless they receive survival signals (positive selection) through interactions with the Ag (presented by FDCs) and Tfh cells [91]. These selection mechanisms impose competition between the B-cell subclones, which is assumed to be based on their relative BCR affinities σ_{rel} [91]. CCs bind Ag to acquire their first survival signal. Subsequently, the Ag is internalized and presented to Tfh cells. Higher-affinity B cells present more Ag and, therefore, compete favorably for the limited number of Tfh cells to acquire a second survival signal. Positively selected may CCs recycle to the dark zone for further rounds of division and SHM, or they differentiate into memory cells or plasma B cells.

To avoid an overly complex model, Ag and Tfh survival signals are modelled with a sigmoidal function:

$$S(\sigma_{rel,i}) = \frac{\sigma_{rel,i}^n}{k^n + \sigma_{rel,i}^n}, \tag{4.2}$$

where i denotes a subclone. This function converts relative affinities $\sigma_{rel,i}$ to a signal strength between 0 and 1. Relative affinities are obtained by scaling absolute affinities σ to values between 0 and 1. Signal S affects the CB to CC differentiation rate ($\eta_{CB\rightarrow CC}$) and the CC apoptosis rate (μ_{CC}) (Equations 5.2a and 5.2b). Recently, it was shown that higher affinity cells stay longer in the dark zone further facilitating their expansion and diversification resulting in less apoptosis [123]. This is accommodated by our model by multiplying the CB to CC differentiation rate with $(1-S)$ resulting in a rate between 0 and its maximum value $\eta_{CB\rightarrow CC}$ (Table 5.1). Similarly, a higher signal reduces the apoptosis

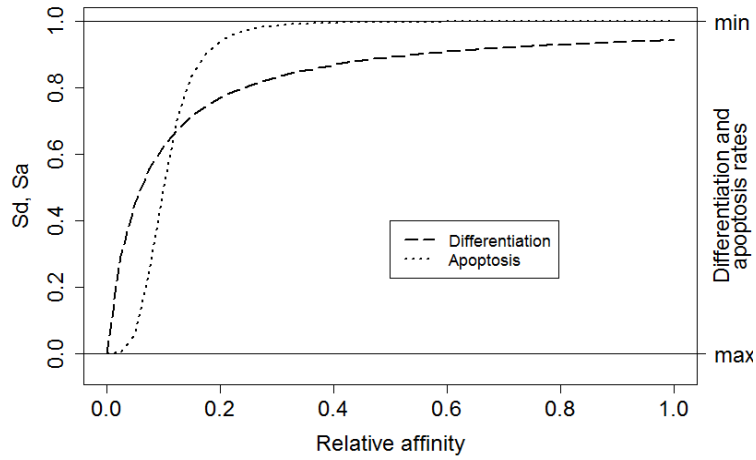


Figure 4.3: The relative affinity of each subclone determines the magnitude of the overall survival signal. Left axis: S_d corresponds to signal affecting the CB to CC differentiation rate (dashed line). S_a corresponds to the signal affecting CC apoptosis (dotted line). Right axis: effect of signal S on the differentiation and apoptosis rates. A high signal results in low differentiation and apoptosis rates.

rate. We assume that S does not affect these rates to the same extend and therefore we parameterized S differently for differentiation and apoptosis. We set $k = 0.06$ and $n = 1$ for differentiation (S_d), and $k = 0.1$ and $n = 4$ for apoptosis (S_a; Figure 4.3). The parameters k and n were chosen to obtain a typical GC response that attains a maximum number of cells during the first phase of the GCR. During our simulation the emergence of new subclones with higher absolute affinity will "push" existing subclones with lower affinities to lower relative affinities as result of the scaling and, hence, to smaller survival signals resulting the vanishing of these subclones.

Ordinary differential equations

Each subclone i assumes 4 phenotypes: centrocytes (CC_i), centroblasts (CB_i), memory cells (M_i), and plasma cells (P_i) (Figure 4.4). The temporal dynamics of each individual subclone is described by a set of ordinary differential equations (ODEs) representing these four phenotypes (Equations 5.2a to 5.2d).

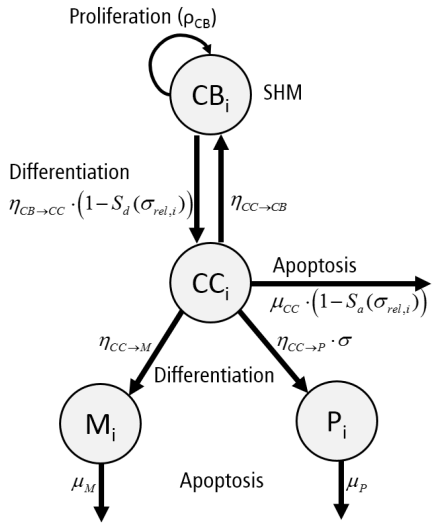


Figure 4.4: Graphical representation of the ordinary differential equations representing a single subclone i. Each subclone assumes four phenotypes: centrocytes (CC), centroblasts (CB), plasma cells (P) and memory cells (M). Cells proliferate (ρ_{CB}), differentiate ($\eta_{CB\to CC}, \eta_{CC\to CB}, \eta_{CC\to P}, \eta_{CC\to M}$), or go into apoptosis (μ_{CC}, μ_P, μ_M) with indicated rates. The apoptosis rate of CCs and differentiation rate of CBs depend on signal S_a and S_d respectively. Differentiation to plasma cells depend on the absolute affinity of the CCs.

$$\frac{dCB_i}{dt} = \rho_{CB}\cdot\left(\frac{A^h}{CB_{total}^h + A^h}\right)\cdot CB_i + \eta_{CC\to CB}\cdot CC_i - \left(1-S_d(\sigma_{rel,i})\right)\cdot\eta_{CB\to CC}\cdot CB_i \tag{4.3a}$$

$$\frac{dCC_i}{dt} = \left(1-S_d(\sigma_{rel,i})\right)\cdot\eta_{CB\to CC}\cdot CB_i - \eta_{CC\to CB}\cdot CC_i - \left(1-S_a(\sigma_{rel,i})\right)\cdot\mu_{CC}\cdot CC_i - \eta_{CC\to M}\cdot CC_i - \eta_{CC\to P}\cdot\sigma_i\cdot CC_i \tag{4.3b}$$

$$\frac{dM_i}{dt} = \eta_{CC\to M}\cdot CC_i - \mu_M\cdot M_i \tag{4.3c}$$

$$\frac{dP_i}{dt} = \eta_{CC\to P}\cdot\sigma_i\cdot CC_i - \mu_P\cdot P_i \tag{4.3d}$$

To allow the GC to grow to a sufficient number of cells during monoclonal expansion the signal $S_{\{d,a\}}$ is set to 0.9 for the first 4 days of the simulation to minimize differentiation of CBs to CCs and apoptosis of the initial CCs. The CB equation includes a density dependent expansion term defining nonspecific resource competition between the B cells, reducing their proliferation rate if the number of cells approaches A. The CC apoptosis rate and the CB to CC differentiation rate are multiplied by $\left(1-S_{\{d,a\}}(\sigma_{rel_i})\right)$ for reasons explained above. Plasma cell differentiation depends on the absolute affinity σ_i

to reduce their production at earlier stages of the GCR. During the simulation we calculate the differential equations for periods of six hours (the duration of one CB division). After each period we impose SHM and update the population of subclones as described above. For each non-lethal SHM a new subclone and an additional set of four ODEs is created. The CB cell count for new subclones is set to one, while the corresponding cell counts for the CCs, memory cells and plasma B cells are set to zero. The CB cell count of the parent subclone is reduced by one. If the sum of CC and CB counts for subclone i is less than 0.1 cells we remove the subclone and corresponding equations from the system. Since SHM is a stochastic process that affects the subclone population and their (relative) affinities, we repeated simulations 15 times with the same initial conditions (three founder B cells with initial affinities 0.1, 0.3, and 0.5).

Model parameters

Parameter values for proliferation, differentiation, and apoptosis were obtained from literature (Table 5.1). Parameters A and h were chosen to limit the maximum size of the GC. Values for parameters for k, n, s, r, and affinity shift were straightforwardly acquired by trail-and-error aiming to produce a typical GC response with a peak of at least 10,000 cells during the first phase of the GCR with our model. There is very limited (quantitative) data describing the GC response. We are not aware of any data obtained from human samples describing the dynamics of GC volume (number of cells) during the GCR. Consequently, the precise timing and magnitude of the maximum GC response, its decay, the biological variation of this response across samples and organisms, and the factors affecting this this response remain to be established. The canonical GC response has, for example, been observed by tracking follicle center volume as fraction of total splenic volume in mice [124] or as fraction of the total volume of the GC in rat [125], which may be used as GC cell count substitutes. These volumes showed a peak during the first phase of the GC. Such measurements have been used previously to validate a GC model [105]. However, other studies showed that there might not exist a typical GC in terms of size [126] and that GCs in a single immune response might not be synchronized [98]. The lack of precise quantitative data, current uncertainties in GC dynamics, and our decision not the model GC termination limits the possibilities and value of a compute-intensive parameter inference strategy to obtain values for the aforementioned parameters. However, instead of our trial-and-error approach, Approximate Bayesian Computation algorithms [127], MEANS [128], or other methods may be used to fit parameters on complex stochastic models such as ours.

Identification of expanded subclones

To determine a threshold that identifies expanded subclones we follow an approach that is similar to the method applied in our previous repertoire sequencing studies, e.g., [93, 101]. First, a histogram of counts c (cell counts for simulated data and read counts for experimental data) for all (un)expanded subclones is constructed to reflect their cell/read count frequencies $F(c)$ (Supplementary Figure S2). In general, subclones with low counts (e.g., $c = 1$) occur much more frequently (high F) than subclones with high count (e.g., $c = 100$). Next we define T as lowest count c for which $F(c) = 0$. That is, no subclones with c cells/reads are observed. We assume that $F(c \geq T) = 0$ for the un-

Table 1. Model parameters

B cell type	**Proliferation rate** (day^{-1})	**Differentiation rate** (day^{-1})	**Apoptosis rate** (day^{-1})
Centroblast (*CB*)	ρ_{CB} = 4 [129, 124, 130]	$\eta_{CB\rightarrow CC}$ = 6 [91]	
Centrocyte (*CC*)		$\eta_{CC\rightarrow M}$ = 1 [131] $\eta_{CC\rightarrow P}$ = 0.1 [131] $\eta_{CC\rightarrow CB}$ = 1 [91]	μ_{CC} = 4 [102]
Plasma cell (*P*)			μ_P = 0.25 [131]
Memory cell (*M*)			μ_M = 0.01 [131]
Other parameters			
Capacity A = 8000		k = 0.06, n = 1 (S_d)	s = 3.0
Number of founder cells: 3		k = 0.1, n = 4 (S_a)	r = 0.3
Initial affinities: 0.1, 0.3, 0.5 mol^{-l}		h = 20	affinity shift = 0.1

derlying but unknown null distribution of unexpanded subclones. We define subclones with $c > T$ ($F(c) \geq 1$) to be expanded. That is, subclones observed with cell/read counts $c > T$ are larger than expected based on the distribution of unexpanded subclones. The threshold T is stringent but could be relaxed by defining the threshold T as the lowest count c for which $F(T) < p$, with $p \geq 1$.

The expansion threshold T was estimated for each individual simulation. We assumed that repertoire sequencing experiments measure mainly CCs since CBs do not, or at very low levels, express BCRs. Consequently, for the simulated data we determine threshold T from CC cell counts only. CC cell counts were taken from the last time point of the simulation.

Comparison of simulated and experimental data

We qualitatively compare subclone cell counts from our simulations to read counts from a single sample repertoire sequencing experiment. Since our computational model does not explicitly represent the BCR as a nucleotide (or protein) sequence we do not consider multiple (back) mutations occurring at previously mutated positions. Consequently, the number of different mutations and, hence, subclones in our simulation is slightly overestimated.

Each unique nucleotide read obtained from repertoire sequencing (RNAseq) can be considered as a unique subclone representing a set of mutations acquired during affinity maturation. Statistics calculated for these subclones can be compared to statistics calculated for the nucleotide-level subclones generated in our simulations. Alternatively, we can define subclones measured in the sample at the peptide level as having unique combination of V and J segments (determined by alignment) together with a unique CDR3. The peptide level definition allows to compare the statistics from the experimental data to the peptide-level simulations (include all three CDRs). In contrast to subclones analyzed at the nucleotide-level, this definition considers any mutations in the CDR3 for the

Table 2. Selected B-cell subclones from sample LN25

Subclone	**Total**		**Largest cluster (lineage)**	
	Subclones	**Reads**	**Subclones**	**Reads**
V3.7 - J4 (nt)	171	334	84	232
V3.74 - J4 (nt)	125	249	23	56
V3.23 - J4 (nt)	37	60	5	12
Total (nt)	333	643	112	300
V3.7 - J4 (pep)	89	606	19	36
V3.74 - J4 (pep)	97	519	9	14
V3.23 - J4 (pep)	76	193	7	12
Total (pep)	262	1318	35	62
			Second largest cluster (lineage)	
V3.7 – J4 (pep)	89	606	9	417

V and J nomenclature following IGMT [118, 107]. Subclones are defined as unique nucleotide sequences (nt) or as peptides (pep) with unique V and J assignment and a unique CDR3 sequence. For each V-J family the number of subclones and corresponding number of sequence reads are shown. The selected clusters for the given V-J segments correspond to the largest cluster of subclones having ≤ 2 differences at nucleotide or peptide level. For V3.7-J4 the second largest cluster, which contains the most abundant subclone, is also included.

experimental data, and affinity changing mutations in CDR1,2,3 for the simulated data.

Repertoire sequencing experiments performed on tissue (e.g., lymph node) generally results in a representation of subclones from multiple GCs and, most likely, different Ag responses, while in our simulation we generate subclones from a single GCR initiated by three founder clones. We account for this by selecting subclones corresponding to three lineages from sample LN25. We first map all reads against reference sequences extracted from the IMGT database to determine their V and J segments. Subsequently, observed combinations of V and J are counted, and reads corresponding to the three most abundant V-J combinations (V3.7-J4, V3.74-J4, V3.23-J4) are selected. The resulting three groups of reads still comprise subclones from multiple lineages. Therefore, we subsequently aligned all pairs of reads within each V-J group to determine the number of nucleotide differences (mutations) between them. Each pair of reads with two or fewer differences are connected to form clusters of subclones that are assumed to belong to the same lineage. Finally, the largest cluster (lineage) for each V-J combination was selected. The same procedure was followed at the peptide level. Since these three clusters did not include the most abundant subclone we also selected the second largest cluster from the V3.7-J4 subclones. The results of this procedure are shown in Table 4.2. Note that the number of differences between pairs of not connected reads within a cluster (lineage) may be larger than 2. These clusters of reads could in principle be subjected to further phylogenetic analysis to determine a lineage tree establishing their relationships [132]

4.3. Results

First we confirm that the computational model produces the dynamics of a typical GC response. We performed 15 repeated simulations with subclones defined at the nucleotide level. In agreement with previous work the GC response peaks around day 8 (Figure 4.5A) [124, 125, 126]. The size of the GC reaches approximately 14,000 cells, which is in agreement with estimations from histological sections of two GCs [133]. The CB to CC ratio (not shown) after day 8 remains between 1.4 and 2.0 is in agreement with data obtained from intravital microscopy [134]. The maximum number of SHMs in subclones emerging from our simulation ranges from 4 (day 10) to 11 (day 21) and is in good agreement with the 9 somatic mutations found in a single Ab after affinity maturation [135], with the 8 to 18 mutations found in an analysis of BCR sequences obtained from cells from GC sections derived from human lymph nodes [133], and with the 4 to 9 mutations observed in B cells from single GCs obtained from mice lymph nodes [136]. Monoclonal expansion of the 3 founder cells results in many low affinity subclones at the initial GC stage, but gradually higher affinity clones start to appear and out-compete lower affinity subclones. As expected from affinity maturation, and in agreement with other computational models (e.g., [102, 105]), the subclone population evolves to higher affinities (Figure 4.5B). The drop in the CB cell count after day 4 is caused by the initiation of SHM and the subsequent differentiation to CCs that may go into apoptosis. Since we do not model GC shutdown the cell counts remain relatively stable after 14 days. These results show that our computational model adequately captures the dynamics of a typical GCR.

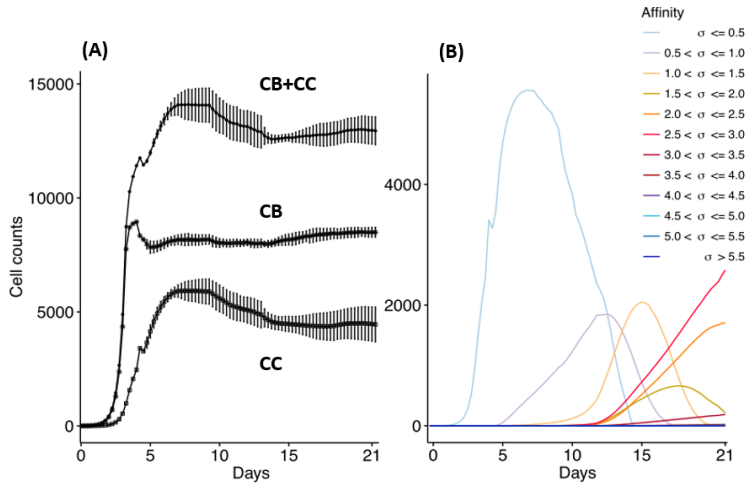


Figure 4.5: Overall GC dynamics emerging from the model. CB and CC with cell counts > 0 are plotted. (A) Dynamics of CB and CC cell counts during the GCR. Top curve shows the total cell count. Each point represents the average cell count of 15 simulations at time intervals of 6 hours (1 CB division). The vertical lines denote the standard deviations. (B) Evolution of absolute affinities during the GCR. Each colored line corresponds to an affinity class for which we summed the cell counts of the corresponding subclones.

Subclonal diversity

Figure 4.6 shows the dynamics of individual subclones during the GCR at the nucleotide and peptide level. Initially, 3 founder clones expand monoclonally until day 4 after which SHM is initiated and new subclones with higher affinity start to be produced. The three low-affinity founder subclones reach high cell counts since, during monoclonal expansion, no lethal SHM occurs and $S_{\{d,a\}}$ assumes a large value (0.9) resulting in a very low rate of CB differentiation and CC apoptosis. New (higher affinity) subclones realize much lower cell counts because they start as single proliferating cells but are also reduced in count due to new mutation events and apoptosis as a result of competition with higher affinity subclones. Interestingly, although the population of subclones evolves to higher affinities (Figure 4.5B) there is not a single nor a small set of subclones that dominates this population during the later stages of the GCR. In fact, the number of unique subclones (Figure 4.6A) remains around 550 during the second half of the GCR.

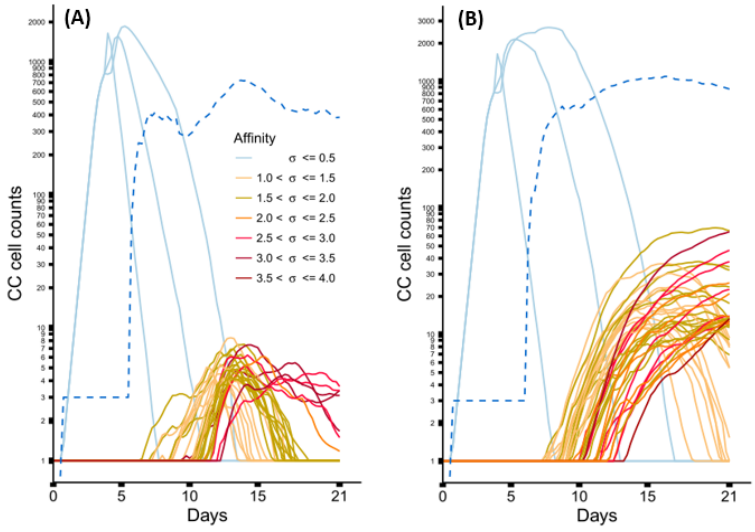


Figure 4.6: Dynamics of individual subclones from a representative simulations. (A) subclones defined at the nucleotide level. (B) subclones defined at the peptide level. Only subclones with (A) CCs cell counts ≥ 4 and (B) CCs cell counts ≥ 11 at any timepoint are shown. During the course of the GCR new subclones of higher affinity emerge (indicated by the colouring scheme). The light blue lines represent the 3 founder subclones of low affinity. The dotted blue line shows the number of unique subclones.

From sample LN25 we identified 112 nucleotide-level defined subclones (i.e. unique sequence reads) corresponding to 300 reads in the three largest lineages (Table 4.2). Since multiple sequence reads may originate from a single B-cell it is not possible to scale these numbers to 14,000 GC cells but obviously 300 reads do not represent this many GC cells. Therefore, these 112 subclones are an underestimation of the true number of subclones in a single GC. Although this number does not provide a validation for the 550 subclones observed in our simulations, it does show that the diversity of subclones in the experiment and the simulations is high. Using multiphoton microscopy and sequencing

it was recently shown that efficient affinity maturation can occur without homogenizing selection, and that loss of clonal diversity during the GCR varies widely from one GC to the other [136]. Note that when comparing Figure 4.6A (nucleotide level) to 4.6B (peptide level) the overall dynamic behavior is similar but the cell counts of higher affinity peptide-level subclones are about five times larger. An increase in cell count is expected since, in this scenario, neutral and synonymous somatic mutations do not result in new subclones and, hence, no reduction of cell counts. The number of unique subclones is still in the same order of magnitude as the previous simulation but counterintuitively increased compared to previous situation since a decrease is expected due to the fewer mutations imposed on these subclones. The observed increase is, however, a result of plotting and summing only the subclones with CC cell counts ≥ 1. Including cell counts < 1 shows that the number of subclones does indeed decrease (data not shown).

Subclonal expansion

Expanded subclones are derived from experimental data on the basis of their peptide-level definition and relative abundance. Basically, this definition neglects any mutation in the V and J region as well as synonymous mutations in the CDR3. We identified expanded subclones from the experimental data (Figure 4.7). First, the expansion threshold was determined using all subclones from the LN25 sample resulting in 34 expanded subclones. Using this threshold ($T = 14$), a total of 3 and 9 subclones from the V3.7-J4 and V3.23-J4 subclones respectively are expanded. For each V-J family, Figure 4.7 also shows the subclones corresponding to the largest cluster (read counts ranging from 1 to 11), and for V3.7-J4, the subclones corresponding to the second largest cluster (read counts ranging from 1 to 261). This shows that subclones within a B-cell lineage may exhibit a wide range of read counts, which is in agreement with our simulated data. It also shows that the most abundant subclones do not necessarily belong to the largest cluster within a V-J family.

The clonal size (number of reads of a subclone divided by total number of reads) of the expanded LN25 subclones varies from 0.2 to 3.4%. Together, these represent 0.8% (34 out of 4454) of all subclones. This is similar to the amount of expansion found in one of our previous studies where clonal sizes $\geq 0.5\%$ were found to represent expanded subclones representing 0.3% and 1.9% of the subclones in peripheral blood and synovial tissue of RA patients respectively [101]. Since our computational model does not explicitly consider V and J segments, and because we cannot distinguish CDR3 from CDR1 and CDR2 mutations, we cannot group subclones resulting from our simulation in a way similar to the experimental data. However, by neglecting neutral and silent FWR/CDR mutations we can simulate subclones at the peptide level. The resulting subclones differ only in their CDR regions. The expanded peptide-level subclones in our simulation represent clonal sizes ranging from 0.3 to 8.7% representing 0.3 to 1.0% of the subclones. This degree of expansion is in the same order of magnitude as expansion observed in our experimental data.

BCR affinity of (un)expanded subclones

Repertoire sequencing only provides information about the relative abundance of B-cell subclones in a sample. In contrast, our computational model also provides information

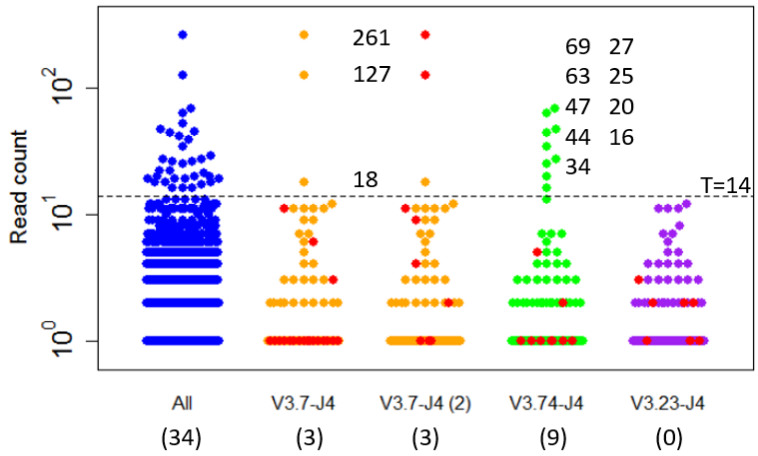


Figure 4.7: Subclones measured in a lymph node sample (LN25) from a healthy individual. The blue points show the read counts for all 4454 subclones measured in this sample (34 expanded subclones). The expansion threshold ($T = 14$) is determined from the all LN25 subclones, and indicated by the dashed line. Subclones of the three most abundant V-J combinations are shown in orange, green, and purple. The red dots indicate the subclones of the largest clusters and, for V3.7-J4, also the second largest cluster. Read counts of the expanded subclones are shown. The numbers in the parenthesis show the number of expanded subclones in the presented V-J subsets.

about the (relative) affinity of each subclone, which we use to gain insight in the affinity distributions among expanded and unexpanded subclones. High absolute affinity was defined by setting a threshold at the 75th percentile of absolute affinities of all subclones produced during the course of the GCR (range $1.53 - 10.6$; 75th percentile is 3.00). Figure 4.8 shows the number of high and low affinity subclones among (un)expanded subclones for 15 simulations with subclones defined at the peptide level. The number of low affinity subclones among expanded cells varies from 17 to 70%, while the number of high affinity subclones among the unexpanded cells is relatively constant at about 25%. In 14 out of 15 simulations the affinity of most abundant subclones belongs to the highest 25% of affinities (Figure 4.9A) but these subclones never assume the highest affinity (Figure 4.9B). Figure 4.9B shows that the affinity tends to increase with subclone abundance (spearman rank correlation is 0.6) but that the largest affinities correspond to low abundant subclones. Increasing the affinity threshold to 95% results in more low affinity subclones among the expanded subclones (data not shown).

Although the affinity distributions depends on the expansion and affinity thresholds, the results demonstrate that lower affinity cells will be among the expanded subclones and *vice versa*. However, in a repertoire sequencing experiment one might not detect the very low abundant (high affinity) subclones. The high-affinity cells in the unexpanded fraction are either new subclones that have undergone significant affinity improvement but did not yet have sufficient time to proliferate, or are high-affinity subclones previously expanded but now being outcompeted by new subclones.

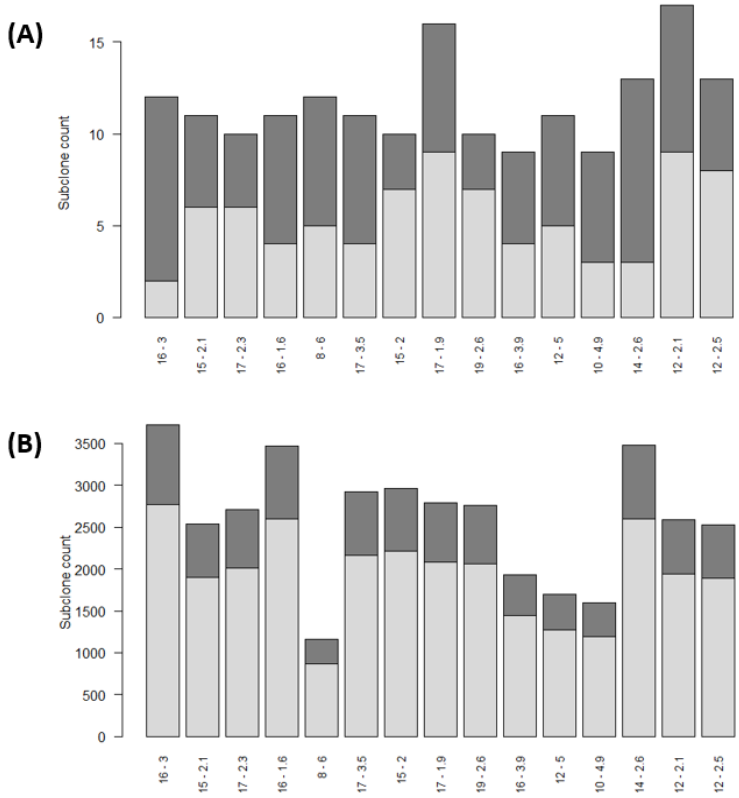


Figure 4.8: Numbers of high (dark gray) and low (light gray) affinity subclones among expanded (A) and unexpanded (B) subclones in 15 simulations (x-axis). Subclones were defined at the peptide level. There are many more unexpanded subclones compared to expanded subclones. Only subclones with CC cell counts > 0 were counts. The numbers at the x-axis denote the thresholds for expansion (T) and absolute affinity (75th percentile).

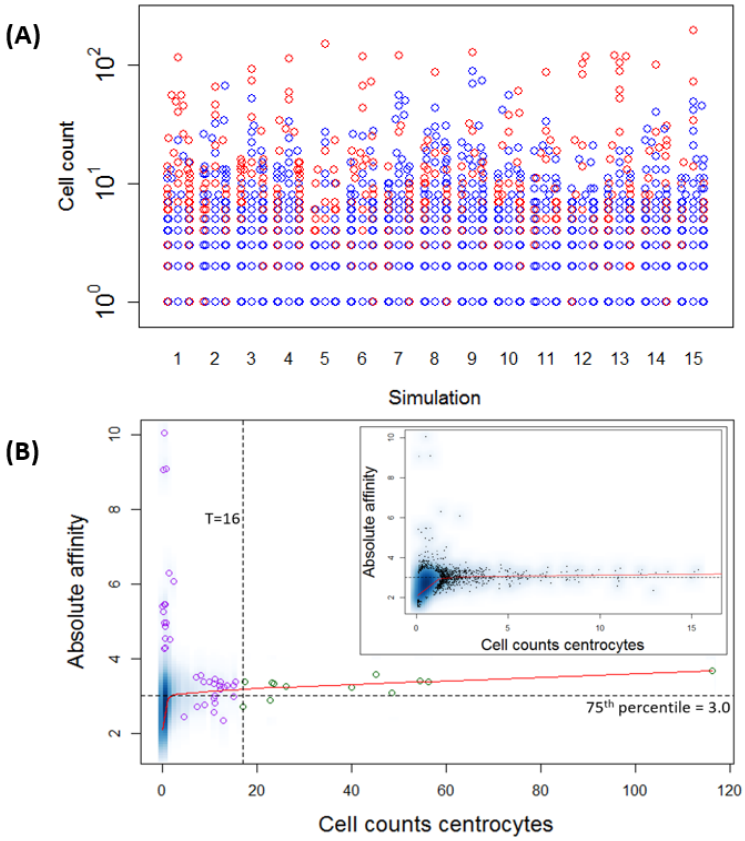


Figure 4.9: (A) distribution of high affinity subclones (red) among all subclones for 15 simulations. (B) Density plot of CC cell counts and absolute affinity for simulation 1. Inset shows only the low abundant subclones. Data points show a selection of subclones imposed on the density plot. Green points denote the expanded subclones. Purple points indicate a selection of low abundant subclones. The red line shows a lowess regression to indicate the overall relation between abundance and affinity.

4.4. Discussion

The identification of autoreactive B cells is important for understanding the pathogenesis of auto-immune diseases and developing therapies that target specific B cells to improve clinical outcome. However, for many autoimmune disorders the Ags are unknown which makes screening approaches challenging. Repertoire sequencing strategies have been developed as an alternative Ag-agnostic approach to identify autoreactive B cells relying on the assumption that expanded B cells measured in blood or tissue are involved in the pathogenesis of the disease. B-cell subclones identified by sequencing can be cloned and functionally characterized, and used to identity the autoantigen. In previous work we demonstrated that expanded clones identified by repertoire sequencing of synovium samples from RA patients point to putative autoreactive B cells [101]. This potentially provides the opportunity to develop novel therapeutic approaches targeting these cells.

(Deep) repertoire sequencing is successfully used for the identification of (autoreactive) B cells involved in immune disorders by relying on the assumption that expanded clones play a key role in the pathogenesis of the disease. However, no information is provided about the affinity of subclones measured with repertoire sequencing. It is reasonable to assume that expanded B cells have higher affinities than the background population of naive B-cells. However, since it is virtually impossible to measure affinity for many subclones detected in a sample, we developed a computational model to investigate the relation between subclone abundance and affinity. Although our computational model was not expected to provide precise quantitative results, we showed that the fraction of low affinity cells among expanded subclones, and the fraction of high affinity subclones among unexpanded B cells are substantial (Figure 4.8). Nevertheless, we showed a moderate positive correlation between subclone abundance and affinity. However, we also showed that the highest affinity subclones are of very low abundance. (Figure 4.9). We showed that the abundance of subclones within a lineage may vary widely. We conclude that repertoire sequencing is able to identify expanded Ag experienced clones but the most abundant subclones within these expanded clonal families are not necessarily the subclones with the highest affinity.

The abundancy-based selection of subclones is not a bad strategy since it leads to the identification of specific (sub)clones involved in (auto)immune disorders. The identified high abundant subclone can subsequently be characterized or used in Ag screening. Using the identified subclone together with phylogenetic analysis one could identify other subclone members of the same lineage and, subsequently, determine their affinities. The combination of abundancy and affinity might further guide the selection process. However, as explained, this is not feasible with current experimental approaches. There are, however, alternative selection strategies that can be used. For example, it has been shown that representative Abs selected from clonal families reconstructed by phylogenetic analysis neutralize influenza more effectively than “singleton” Abs that use heavy-chain V(D)J and/or light-chain VJ gene segments that are not used in any other Ab in the repertoire [95]. They showed that Abs from clonal families had significantly higher affinity than did singleton antibodies. Such strategy could be combined with subclone abundance. In previous work we have shown that the identification of pathogenic subclones in RA benefits from the selection of high-abundant subclones that are present in multi-

ple joints within a patient [101, 137]. It would be interesting to determine the affinity of these overlapping subclones in comparison to high abundant non-overlapping clones.

Our modelling efforts were motivated by the fact that it is currently infeasible to measure the affinities of large populations of subclones. We realize that at the same time this also prohibits direct experimental validation of the affinity distribution generated by our simulations. However, with evolving experimental technologies and approaches this may become feasible in the future. Using a tractable immunization mouse model and a well-defined Ag might be a first step towards validation. In this case a single cell strategy is required to sequence both the heavy and light Ig chains. Subsequently, the Igs must be cloned and expressed followed by measuring antibody-antigen binding kinetics using surface plasmon resonance [138]. However, it will remain difficult for clinical samples.

Surprisingly, our model shows that the number of unique subclones in a single GC remains remarkably constant throughout the GCR and does not evolve to a single or few high affinity dominating subclones although the affinity of the population as a whole increases as has been shown in previous studies [102, 105]. Moreover, the cell counts of individual subclones remain very low. Adding additional mechanistic detail (e.g., GC shutdown) is unlikely to change this observation. Moreover, this observation is in agreement with repertoire sequencing data and also seems in agreement with a recent study that showed that many clones may mature in parallel and sporadic clonal bursts generates many SHM variants of a clone [136].

Our model can be improved in several ways. Given the current results it would be interesting to investigate if our results would hold with more detailed GC models since with the model it is very difficult to control the amount of expansion by change the sigmoid functions without distorting the overall GC dynamics (although this might happen also *in vivo*. It would be interesting to investigate what exactly controls selection pressures and how this affects subclonal expansion and the BCR affinity distribution. Nevertheless, as we have shown, the current magnitude of expansion observed from the model is in the same order of magnitude as observed in experimental data. To allow a better comparison to the experimental data we plan to include an explicit representation of the BCR as a nucleotide sequence in our future model. This would allow to distinguish between the different CDR regions, to account for multiple (back) mutations at identical positions, and to more precisely specify subclones at both the nucleotide and protein level. In analogy to [139, 97], this would allow to explore the clonal composition and subclonal dynamics in a system where the best affinity BCR sequence is known and may be reached in few (key) mutations such as in the response against (4-hydroxy-3-nitrophenyl)acetyl [139, 97]. However, in general, the incorporation of realistic affinities in GC models will remain a challenge. Another interesting extension would include the egress of B cells to investigate the (sub)clonal composition in blood and to compare this to repertoire sequencing data obtained from blood samples.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

NdV and AvK designed the study. PR, AvK, JG, and PK defined the model. PR performed the simulations. PR and AvK analyzed the results from the simulations. PPT developed the protocol for acquiring LN samples, and provided the sample used in this study. RE conducted the repertoire sequencing experiment. PK, MD, BvS, AvK and NdV analyzed the experimental data. PR and AvK wrote the manuscript. All authors critically read, contributed, and approved the manuscript.

Funding

This work was carried out on the Dutch national e-infrastructure of SURFsara with the support of SURF Foundation. Research was supported by the Netherlands Bioinformatics Center (NBIC).

Acknowledgments

Prof. dr. Age Smilde (Biosystems Data Analysis Group) is acknowledged for critically reading and improving the manuscript. Dr. Huub Hoefsloot and Dr. Johan Westerhuis (Biosystems Data Analysis Group) are acknowledged for their suggestions during this research.

4.5. Supplementary Material

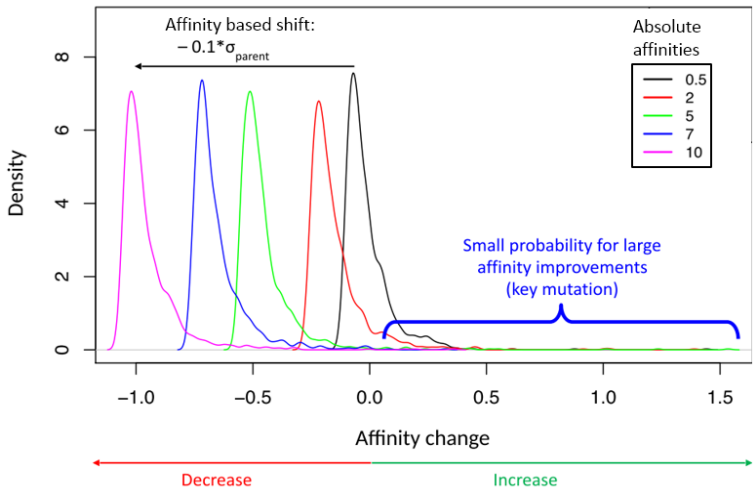


Figure 4.10: Distribution $f(\sigma)$ used to change affinity of mutated subclones. A mutation may decrease or increase the affinity of a B-cell. There is a small chance of making a large affinity improvements (representing key mutations). The distribution is shifted to the left with $0.1 * \sigma_{parent}$ for cells with higher affinities to decrease the chance for further improvements.

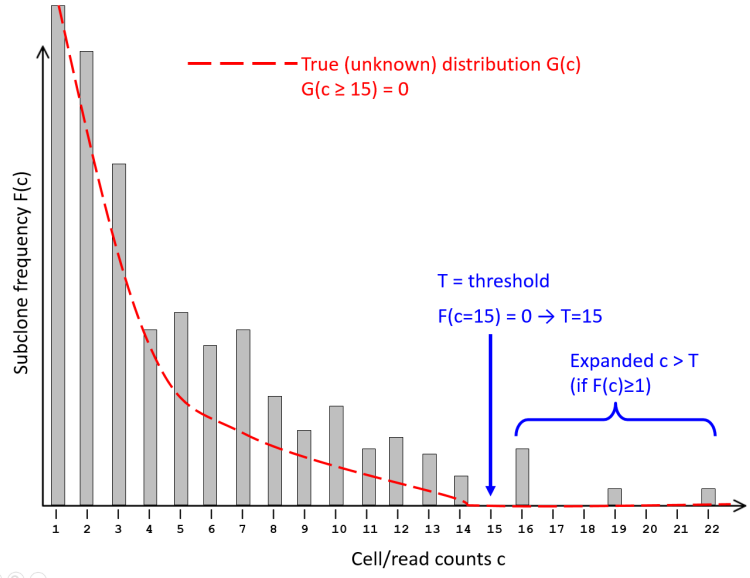


Figure 4.11: Determination of threshold for expanded subclones. See main text for further explanation.

Chapter 5
The evolution of B-cell lineage trees during affinity maturation[4]

B-cell affinity maturation in germinal centres (GCs) during an immune response is a crucial mechanism of our immune system, which is, however, hard to assess experimentally. Lineage trees reconstructed from B-cells subjected to affinity maturation have been reconstructed from experimental data. The resulting shape parameters of the tree have been shown to reflect the properties of affinity maturation. In this study we used a computational model to explore the evolution of B-cells lineage trees during affinity maturation in a single GC. We analyzed lineage tree parameters such as total number of nodes, node outgoing degrees and tree length and followed changes in these parameters as the immune response progressed. Since our model also provides information about subclonal abundance (cell counts) and affinity, we integrated those quantities in the lineage tree. This provides additional information about B-cell affinity maturation which is virtually impossible to obtain using experimental data.

5.1. Background

B-cell affinity maturation is a key defence mechanism of the adaptive immune system that improves the response against pathogens. Affinity maturation takes place in germinal centres (GCs), which are specialised structures in lymphoid organs where B cells proliferate with a high rate and undergo somatic hypermutation (SHM). SHM affects the genes coding for the B-cell receptor (BCR) which is responsible for antigen (Ag) binding. The evolution of B cells as result of affinity maturation can be visualized and analysed by constructing B-cell lineage trees from sequenced BCRs. In such tree every node represents a unique subclone (variant of a clone within VJ family produced by SHM [140]) and, consequently, SHM events can be followed. Tree edges represent the number of mutations between two connected subclones. The root of the tree generally represents the un-mutated germline sequence, while the leaves represent the subclones (at the end of affinity maturation) that were not further mutated. It has been suggested that the shape of the lineage tree reflects differences in the affinity maturation process in health and disease. Lineage tree shapes can be characterized by graph parameters such as number of nodes and average outgoing degree [141, 104]. Several tools have been developed to support the reconstruction and annotation of lineage trees from experimental data [142, 132, 143, 144].

Lineage trees have been used to confirm the role of GC as the location of SHM [145, 146, 139], to identify lineage relationships between cells from independent GCs

[4] P. Reshetova, B.D.C. van Schaik, A.H.C. van Kampen. Manuscript in preparation.

[147] or from different tissues [148], and to analyse broadly neutralizing HIV antibodies [149]. A general discussion about the molecular evolution of B-cell receptors may be found in [150]. The evolution of a lineage tree during the germinal centre reaction (GCR) from a limited number of experimentally derived sequences was shown by Jacob (1993) and further analysed by Dunn-Walters and co-authors [141]. They showed that during the GCR, the lineage trees gradually change to longer, more pruned, shapes due to positive and negative selection of B cells. In addition, they compared lineage trees constructed from immunoglobulin (Ig) sequences obtained from human spleen and Peyer's patches and concluded that B cells in spleen were subjected to stronger selective forces. Finally, they were able to show that a selection of tree graph parameters significantly correlated to parameters (mutation rate, selection threshold) of a simple computational model of affinity maturation. Similar analyses were conducted by Shahaf *et al.* who used lineage trees to compare primary and secondary immune responses *in silico* [104]. Based on their analysis, tree graph parameters such as the outgoing degree of the tree root could be related to the selection threshold and initial affinity. However, later it was argued that these correlation may not be significant since experimental factors were not accounted for and because some of the graph parameters measures may reflect differences in the overall population size between tested conditions [151]. Instead, Uduman *et. al* proposed to incorporate tree shape measures into statistical tests to detect selection of BCR sequences because shape parameters alone may not be very reliable.

Stern and co-workers analysed B-cell lineage trees to address the mechanism of B-cell maturation and trafficking between the central nervous system (CNS) and secondary lymphoid organs in multiple sclerosis (MS) [152]. It was suggested that MS CNS B cells encounter antigen and improve their affinity in the secondary lymphoid tissue. Part of these B cells then populate the CNS but continue trafficking between the CNS and periphery. Similarly, lineage tree analysis was used to study diversification of B cells found in inflamed intestinal tissue of two ulcerative colitis patients as well as B cells from mucosa-associated lymph nodes (LN) [153]. Tree shapes revealed active clonal diversification in ulcerative colitis patients. Moreover, B cells from intestinal tissues and the associated lymph nodes were shown to be clonally related, thus supplying evidence for B-cell trafficking between gut and associated lymph nodes. More recently, lineage trees were used to analyse two time-scales in affinity maturation of naive and reactivated B cells [96]. Naive B cells start with a germline state without mutations. In contrast B memory cells that are reactivated begin with a mutated, affinity-matured receptor, which is then further diversified during a GCR. B-cell lineage trees also revealed changes in affinity maturation with age supporting the observation that elderly produce increased levels of antibodies to autologous antigens and are less able to make high-affinity antibodies to foreign antigens [154]. This study showed also demonstrated tissue-specific differences between Peyer's patch GC and splenic GC with age. Lineage tree analysis applied to study autoimmune diseases showed that tree sizes in these diseases are larger compared to normal controls indicating that more mutations accumulated [155, 156]. This was expected from the chronic nature of autoimmune disorders. However, based on the analysis of the outgoing degree (see below) these studies also showed that autoimmune diseases and normal controls experienced similar selection pressure.

In this study we explore the evolution of B cells during affinity maturation in a sin-

gle GC by using a computational model which we developed in Chapter 4. In contrast to previous studies that investigated tree evolution from a limited set of experimentally-derived sequences [139, 141], our computational model enables the construction of trees including every subclone produced by SHM during affinity maturation. We analysed the change in graph parameters during the GCR and show that the total number of nodes and tree length behave in a similar fashion with experimentally derived trees. One advantage of our computational model is that it also provides information about subclonal abundance and affinity. We show how this can be integrated in the lineage tree to provide information that currently is virtually impossible to derive from experimental data. By repeating our simulations we also obtain insight in the variability of the tree shapes as result of SHM.

5.2. Methods

Software

We developed a mathematical model using ordinary differential equations (ODEs) to describe the dynamics of individual subclones during the GCR. This model is implemented in the R statistical environment version 3.2.2 [110] using R packages deSolve (version 1.12) [111], R6 (version 2.1.2), ggplot 2.0, igraph 1.0.1 and beeswarm 0.2.1. The software is freely available as open source (GPLv3) on request from the author.

Computational model

Here we describe the main aspects of our computational model. Further details can be found in Chapter 4. The GC and affinity maturation are reviewed in (Victora, 2012; Silva and Klein; 2015). Our model starts with a monoclonal expansion of B cells (centroblasts; CB) at day 0 to over 10,000 cells at day 4. During this expansion phase and the remainder of the GCR, the CBs differentiate to centrocytes (CC). SHM and the production of plasma B cells and memory B cells is initiated at day 4. At day 21 the GCR is terminated. The Ag and T follicular help (Tfh) survival signals are modelled with sigmoidal functions S_d and S_a that affect the rate of CB to CC differentiation, and the rate of CC apoptosis respectively. Effectively, these functions induce competition between the subclones through positive and negative selection. SHM with a rate of 10^{-3} per bp per division is modelled with a Poisson distribution $m = Poisson(\lambda = 0.6)$. The fate of each mutation is determined a decision tree involving silent (synonymous mutations), lethal FWR, and affinity changing CDR mutations [103, 120] (Figure 5.1). The lineage trees that we construct during the simulation are based on unique subclones that are defined as having a unique CDR protein sequence. Each new subclone created by SHM starts as a CB and, subsequently proliferates and differentiates to and co-exist as CB, CC, memory cell and plasma cell at succeeding time points. Subclones that with CB cell counts ≤ 1 are kept in our simulation be are not further be affected by SHM to avoid the generation of new clones from these cells. Each subclone in our model has a unique BCR with an absolute affinity σ for the Ag. The affinities of the three single cell founder CBs are set to arbitrary but different low affinity values (0.1, 0.3, and 0.5 μM). For each affinity changing mutation (Figure 5.1) the affinity of the affected subclone is updated according to $\sigma_{new\ subclone} = \sigma_{parent} + \Delta\sigma$ where $\Delta\sigma$ is drawn from a distribution $f(s, r, \sigma)$ with

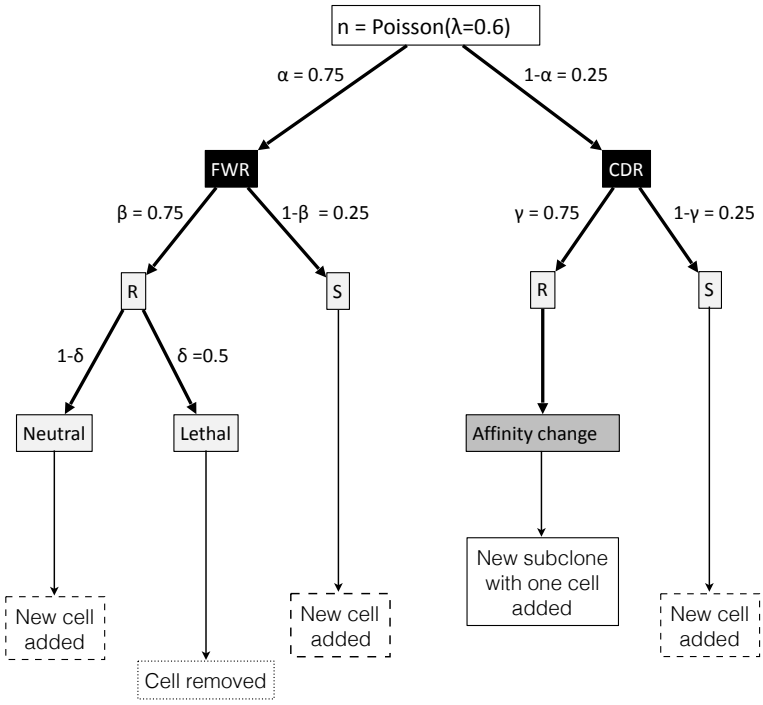


Figure 5.1: Fate of somatic hypermutations during the germinal centre reaction.

probability density function:

$$f(\sigma) = g(s = 3.0, r = 0.3) - \mu - (\sigma_{parent} * 0.1), \tag{5.1}$$

where $g(s,r)$ is the inverse gamma distribution. μ is the expected value of g and used to center the distribution around zero resulting in about equal chances for decreasing and increasing the affinity of mutated subclones. To decrease the chance for high affinity subclones to further improve their affinity we shift distribution f to the left as a function of the parent cell affinity.

Ordinary Differential Equations

Each subclone i can assume 4 phenotypes: centrocytes (CC_i), centroblasts (CB_i), memory cells (M_i), and plasma cells (P_i) (Figure 5.2). The temporal dynamics of each individual subclone is described by a set of ordinary differential equations (ODEs) representing these four phenotypes (Equations 5.2a to 5.2d; Table 5.1).

$$\begin{aligned}\frac{dCB_i}{dt} = {} & \rho_{CB} \cdot \left(\frac{A^h}{CB_{total}^h + A^h} \right) \cdot CB_i + \eta_{CC \to CB} \cdot CC_i \\ & - \left(1 - S_d(\sigma_{rel,i})\right) \cdot \eta_{CB \to CC} \cdot CB_i \end{aligned} \tag{5.2a}$$

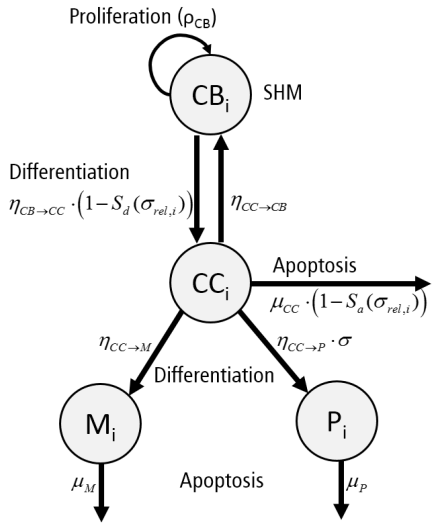


Figure 5.2: Representation of the GC model. Phenotypes of a single subclone are shown: centrocytes CC_i, centroblasts CB_i, plasma P_i and memory M_i cells, where i is a subclone index. Arrows represent cell proliferation rate (ρ_{cb}), differentiation rates ($\eta_{CB\to CC}$,$\eta_{CC\to CB}$,$\eta_{CC\to P}$,$\eta_{CC\to M}$,) and death rates (μ_{cc},μ_p,μ_m). S_d, and S_a represent the survival signals affecting differentiation and apoptosis. σ_i and $\sigma_{i,rel}$ represent the subclone absolute and relative affinities.

$$\begin{aligned}\frac{dCC_i}{dt} &= \left(1-S_d(\sigma_{rel,i})\right)\cdot\eta_{CB\to CC}\cdot CB_i-\eta_{CC\to CB}\cdot CC_i\\ &\quad-\left(1-S_a(\sigma_{rel,i})\right)\cdot\mu_{CC}\cdot CC_i-\eta_{CC\to M}\cdot CC_i\\ &\quad-\eta_{CC\to P}\cdot\sigma_i\cdot CC_i \qquad (5.2b)\\ \frac{dM_i}{dt} &= \eta_{CC\to M}\cdot CC_i-\mu_M\cdot M_i \qquad (5.2c)\\ \frac{dP_i}{dt} &= \eta_{CC\to P}\cdot\sigma_i\cdot CC_i-\mu_P\cdot P_i \qquad (5.2d)\end{aligned}$$

To facilitate monoclonal expansion to about 10,000 cells the signal S is set to 0.9 for the first 4 days of the simulation to minimize apoptosis of CCs and differentiation to of CBs to CCs. The CB equation includes a density dependent expansion term defining nonspecific resource competition between the B cells, reducing their proliferation rate if the number of cells approaches A. The CC apoptosis rate and the CB to CC differentiation rate are multiplied by $\left(1-S_{a,d}(\sigma_{rel_i})\right)$ to facilitate B-cell selection. Plasma B-cell differentiation depends on the absolute affinity σ_i to reduce their production at earlier stages of the GCR. During the simulation we calculate the differential equations for periods of six hours (the duration of one CB division). After each period we impose SHM and update the population of subclones. For each non-lethal SHM a new subclone and an additional set of four ODEs is created. The CB cell count for new subclones is set to one, while the corresponding cell counts for the CCs, memory cells and plasma B cells are set to zero. The CB cell count of the parent subclone is reduced by one. If the sum

Table 5.1: Model parameters.

B-cell type	**Proliferation rate** (day^{-1})	**Differentiation rate** (day^{-1})	**Apoptosis rate** (day^{-1})
Centroblast (*CB*)	ρ_{CB} = 4 [129, 124, 130]	$\eta_{CB\to CC}$ = 6 [91]	
Centrocyte (*CC*)		$\eta_{CC\to M}$ = 1 [131] $\eta_{CC\to P}$ = 0.1 [131] $\eta_{CC\to CB}$ = 1 [91]	μ_{CC} = 4 [102]
Plasma cell (*P*)			μ_P = 0.25 [131]
Memory cell (*M*)			μ_M = 0.01 [131]
Other parameters			
Capacity A = 8000		k = 0.06, n = 1 (S_d)	s = 0.3
Number of founder cells: 3		k = 0.1, n = 4 (S_a)	r = 0.3
Inititial affinities: 0.1, 0.3, 0.5 mol^{-l}		h = 20	affinity shift = 0.1

of CC and CB counts for subclone i is less than 0.1 cells we remove the subclone and corresponding equations from the system.

Lineage tree construction

Lineage trees reconstructed from experimentally derived sequence are mainly based on CCs (since CBs do not, or at very low levels express the BCR). Consequently, in our simulation we construct lineage trees from CCs only. We only used CCs with cell counts ≥ 1. In our simulation we incrementally build the lineage trees for each of the three founder B cells while new subclones are being produced by non-lethal SHM. Therefore, in contrast to tree reconstruction from experimental data we did not have to use any special tree reconstruction method. Moreover, we were able to construct lineage trees where the difference between any connected two nodes is a single mutation.

Calculation of graph parameters

We analysed lineage trees resulting from the founder cell with the highest initial affinity. We repeated the simulation 14 times to obtain information about the change in variability during the GCR for each individual graph parameter.

For every lineage tree we calculated and discussed a subset of nine graph parameters that changed during the GCR (Table 5.2, Figure 5.3):

The graph parameters can be interpreted in terms of affinity maturation. The total number of nodes (N) represents the number of subclones produced through the GCR. The total number of leaves (L) represents the number of distinct subclones that were not further affected by SHM. Internal nodes (IN) include all nodes except the root and leaves. Part of these internal nodes represent living subclones that still reside in the population of subclones. Pass through nodes (PTNs) represent subclones with exactly one child. Although a subclone (a PTN) could have produced descendants during the GCR, these subclones did not survive the selection process or the mutations producing these subclones were lethal. In contrast, split nodes (S) have two or more descendants. The outgoing degree represents the number child subclones produced by SHM for each non-leave

Table 5.2: Graph parameters.

	Graph parameter	Abbreviation	Range
1	Total number of nodes	N	[47,1235]
2	Total number of leaves	L	[33,867]
3	Number of internal nodes	IN	[13,483]
4	Number of pass through nodes	PTN	[7,361]
5	Average outgoing degree (except the root)	avgOD	[4, 267]
6	Root outgoing degree	rootOD	[6,711]
7	Average path length from the root to leave	avgRL	[2, 5]
8	Maximum path length from the root to the leave	maxRL	[4, 11]
9	Average distance between the root and any split node	avgRSN	[1,5]

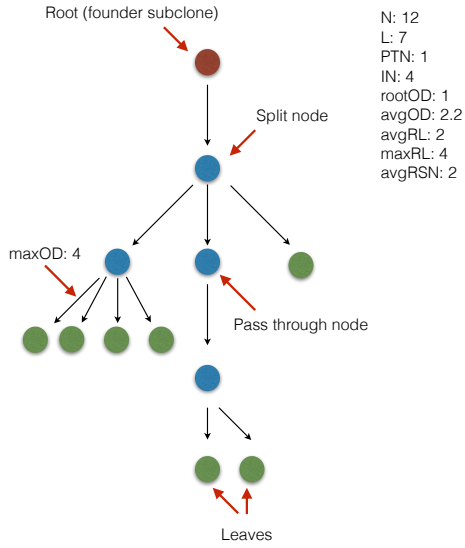


Figure 5.3: Graph parameters describing a B-cell lineage tree.

subclone. In the first phase of the GCR the founder subclone monoclonally expands to thousands of cells. Consequently, once SHM is initiated the founder subclone produces a large number of descendants, which leads to an exceptionally high root outgoing degree.

To avoid skewed results we consider the average outgoing degree of all split nodes except the root (avgOD) and the root outgoing degree (rootOD) separately. PTN, L, avgOD, and rootOD provide a measure for the "bushiness" of the tree which inversely correlates with the amount of selection pressure B-cell subclones experience. Decline of rootOD during affinity maturation as a result of competition between subclones is an indicator of the progression of the GCR. Also the average path length from root to leave (avgRL) and the maximum path length from root to leave (maxRL) represents the average and maximum number of mutations in single subclone. This parameter is influenced by the

mutation rate, initial affinity (lower affinity founder cells may acquire more mutations to obtain maximum possible affinity), and length of GCR reaction. Average distance between the root and any split node (avgRSN) indicates the progress of the GCR. While the GCR progresses, early low affinity root descendants are removed from the subclonal population as a result of the selection process and avgRSN is growing.

Lineage tree graph parameters were calculated for every day starting at day 10 and ending at day 21. Due to the nature of the differential equations the cell counts may become less than 1. In our model we allow subclones with cell counts < 1 to avoid too much interference with the simulation but remove a subclone (the corresponding set of equations) if cell counts are < 0.1. However, subclones with cell counts < 1 do not have a biological interpretation therefor we construct lineage trees for subclones with cell counts $\geqslant 1$. Trees before day 10 only contained very few nodes and, therefore, were excluded from the analysis. To gain insight in the (change in) variability of the graph parameters we repeated simulations 14 times resulting in 14 lineage trees at every time point. We report the values of the graph parameters as boxplots where the boxes indicate the 25th and 75th percentile, the horizontal line in the box represents the median, the whiskers indicate the 5th and 95th percentiles and the crosses indicate outliers (defined by extreme studentized deviate test [157]).

Expanded subclones

Subclonal expansion was defined following Chapter 4. In brief, a histogram of CC cell counts c for all subclones at the end of the GCR is constructed to reflect their frequencies $F(c)$. Next we define T as lowest count c for which $F(c) = 0$. We assumed that $F(c \geq T) = 0$ for the underlying but unknown null distribution of unexpanded subclones. Consequently, we define all subclones with cell counts $> T$ and $F(T) \geq 1$ to be expanded. The expansion threshold T was estimated for each individual simulation at the end of the GCR.

5.3. Results

We first visualized lineage trees development at different time points during the GCR. Secondly, we analysed the tree graph parameters during the GCR. Finally, we analysed subclonal expansion and affinity in the context of a lineage tree.

Visualization of lineage tree development during the GCR

To visualize the lineage tree evolution during the GCR we selected a representative simulation and constructed lineage trees for days 10, 13, 17, and 21 (Figure 5.4). The length of these trees (maxPL) increases from 3 at day 10 to 6 at day 17.

Graph parameters N, L, PTN and IN

The median total number of subclones (N) stays relatively stable during the course of the GCR (Figure 5.5(A)) with a slight peak between days 12-14. This shows that subclonal diversity does not narrow to a single or few high affinity subclones that outcompete the other lower affinity subclones. The median total number of leaves (L) also stays relatively constant. Number of internal nodes (all nodes except the root and leaves) (IN) stays close to the number of pass through nodes (PTN) indicating that only a minor set of

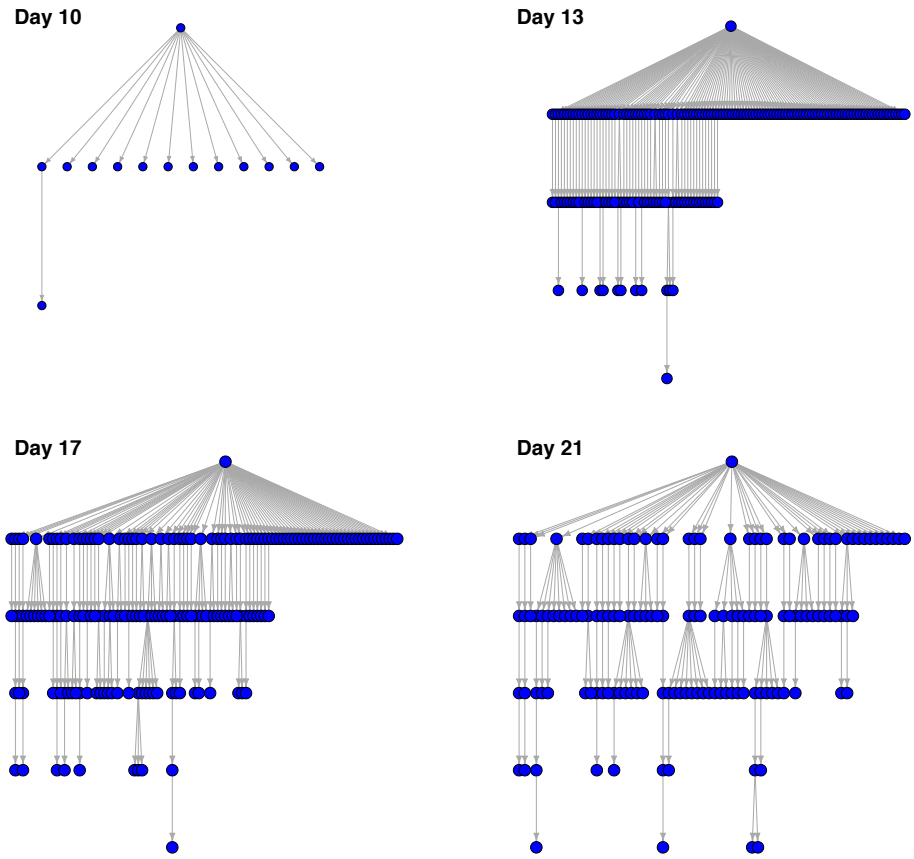


Figure 5.4: Evolution of a lineage tree during the GCR. CC with cell counts ≤ 3 and their ancestors are shown. Four time points were selected from a single simulation to visualize the tree evolution. Every node represents a subclone. Each edge represents a single mutation.

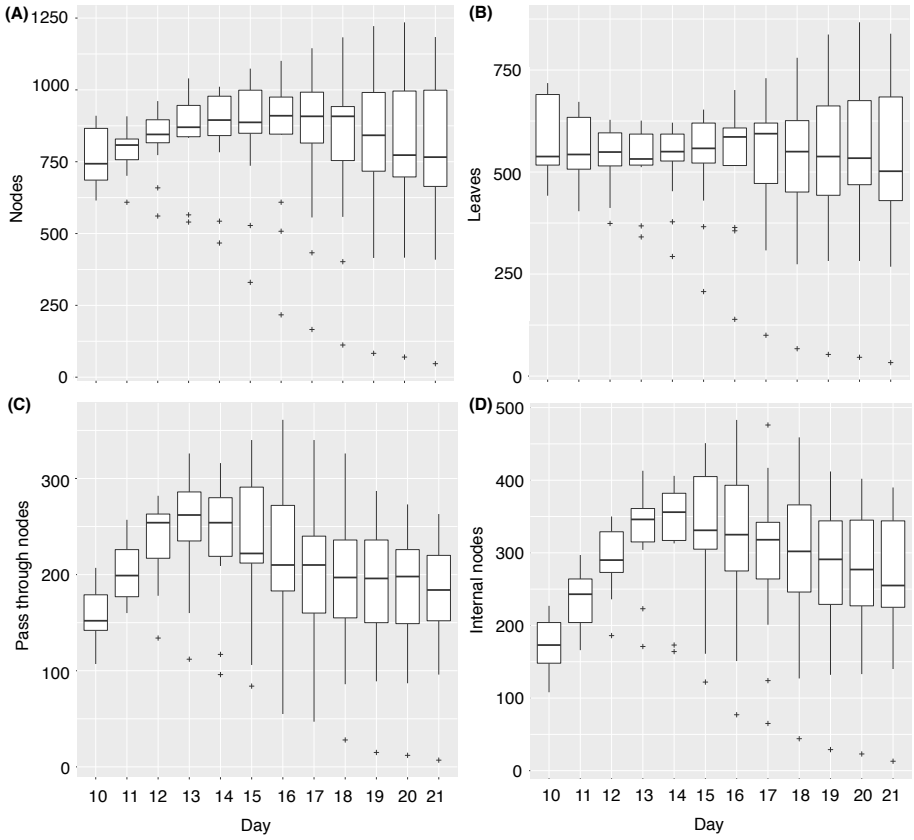


Figure 5.5: (A) Total number of nodes (N), (B) total number of leaves (L), (C) number of pass through nodes (PTN), and (D) number of internal nodes (I).

nodes is able to produce more than 1 descendants (and thus are not considered PTN). The median number of PTN shows a maximum at 13 days indicating that at that stage during affinity maturation we have the largest percentage of non-diversifying subclones (Figure 5.5(B)). Graph parameter boxplots also show a variability of the trees during the GCR caused by the random process of SHM. For example, N shows that the tree size at the end of the GCR varies between approximately 700 and 1000 subclones Figure 5.5. One simulation resulted in an outlier corresponding a very small tree containing only 47 subclones.

Graph parameters rootOD and avgOD

The outgoing degrees provide a measure for the bushiness of the lineage tree. The root outgoing degree (rootOD) and the average outgoing degree (avgOD) are decreasing while the affinity maturation is progressing (Figure 5.6).

Both outgoing degrees decrease with ongoing GCR and, therefore, are correlated to some extent. However, it remains to be established how these graph parameters pro-

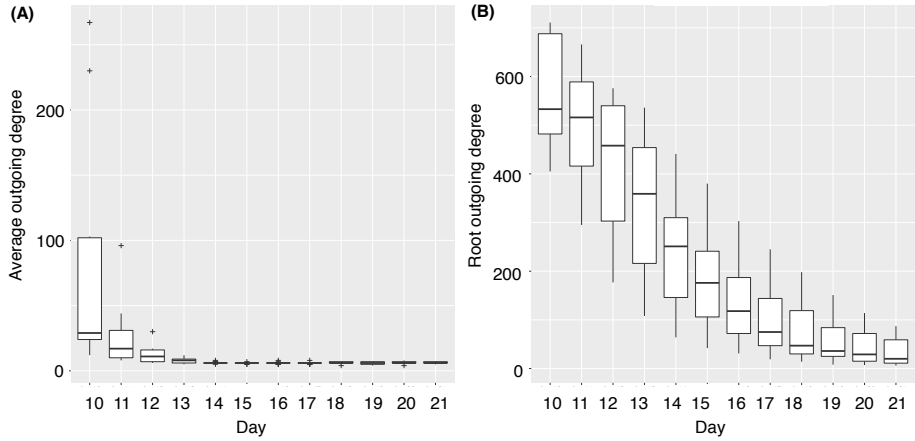


Figure 5.6: (A) The average outgoing degree (avgOD) and (B) root outgoing degree (rootOD).

vide information about the underlying process. According to [141] each of these outgoing degrees relate to the mutation rate and selection threshold. In contrast Shahaf et al [104] related only avgOD to the mutation rate, while rootOD was related to the selection threshold although with an opposite effect than by Dunn-Walters et al. Uduman et al did not find any correlation between these ODs and model parameters [151]. However, from their research we might conclude that increasing the mutation rate would lead to trees that are less bushy.

In contrast to the increasing variance for N, L, PTN, and IN the variance for the outgoing degrees decreases with progressing GCR. Thus, although the GCR may produce trees that show large variability in size, the number of different descendent subclones produced from a single subclone decreases leading to less bushy trees. This is mainly a consequence of subclone abundance which decreases in time (see Chapter 4). Initially the abundance of the founder clones is very high as a result of monoclonal expansion. Each of these cells can be mutated and produce a new subclone represented by a single cell. These new subclones do proliferate but due to new mutation event and competition with other subclones their abundances stay relatively low, resulting in fewer possibilities to produce new subclones. If we consider rootOD then we observe that a large number of new subclones originate from the founder clone during the initial phase of the GCR (Figure 5.4). At later stages the founder clone is reduced in abundance and many of its descendants did not survive.

Graph parameters avgRL, maxRL, avgRSN

Figure 5.7 shows various distances in terms of acquired mutations between nodes in the lineage tree. The maximum path length from root to any leave (maxRL) shows that the length of the tree increases over time but seems to stabilize at the end of the GC. maxRL corresponds to the maximum number of mutations acquired by specific subclones. Stabilization is expected since it becomes increasingly unlikely to have mutations that further increase the affinity once it has reached the maximum binding affinity for the Ag.

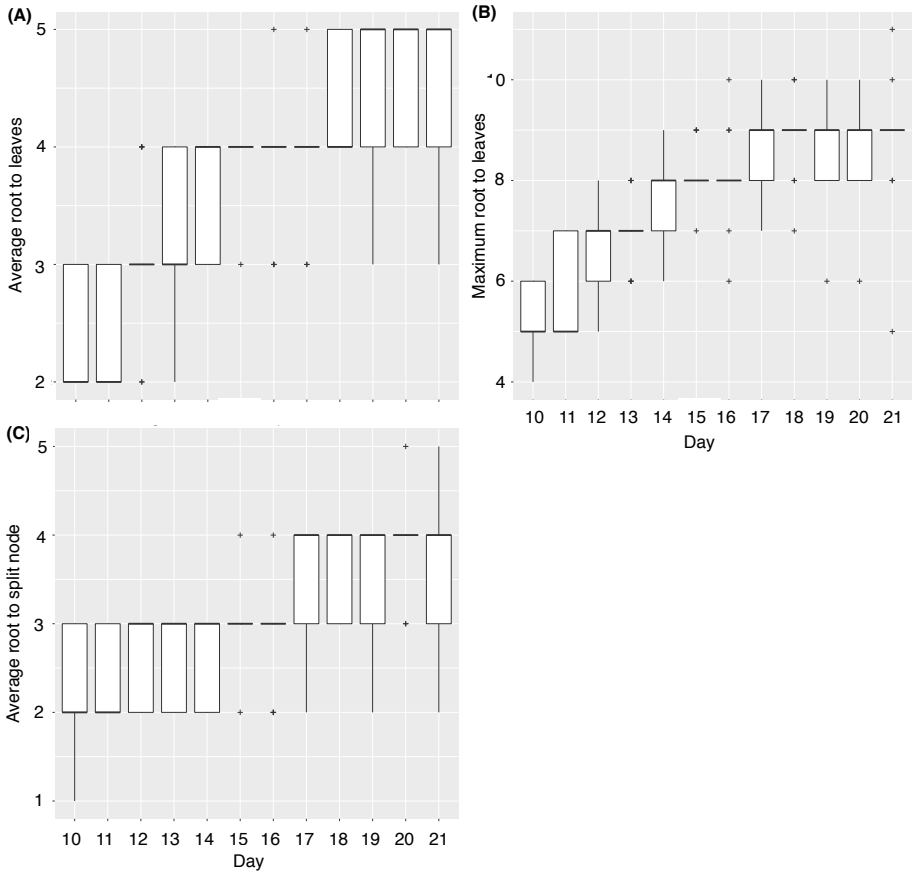


Figure 5.7: (A) Average path length from the root to leave (avgRL), (B) the maximum path length from the root to the leave (maxRL), and (C) average distance between the root and any split node (avgRSN).

Consequently, new mutations will decrease affinity and, therefore, the survival probability of this new subclone. The maximum number of mutations (maxRL) observed from the lineage tree is 11 and is in the same order of magnitude as the previously reported maximum number of mutations occurring during the GCR [133, 135]. Shahaf et al suggested association between maxRL and the selection threshold [104]. The average distance between root and any leave (avgRL) shows that the average number of mutations acquired by the subclones is about 5 at the end of the GCR. minRL (data not shown) is either 1, or 2 at the end of the GCR and therefore is not very informative. As expected, average distance between the root and any split node (avgRSN) is growing indicating the progress of the GCR.

Subclonal expansion and affinity in the context of lineage trees

Our simulations keeps track of the lineage tree but also of the abundance and affinity of all subclones at all time points during the GCR. This enables us to explore the dynam-

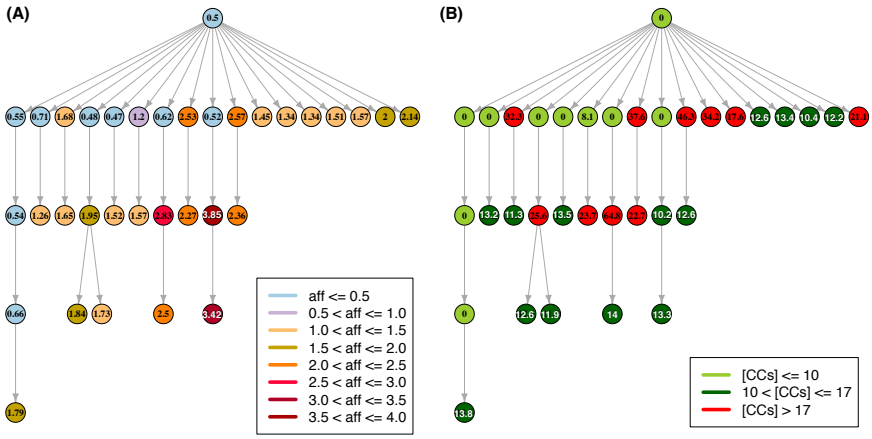


Figure 5.8: Lineage tree that includes subclones with CC cell counts > 10 at the end of the GCR and their ancestors. Every node is coloured according to the subclone (A) affinity class and (B) abundance. Expanded at the end of the GCR subclones (subclones with cell counts > 17) are coloured red.

ics of subclonal expansion and affinity in the context of a lineage tree. We selected the same simulation as used for Figure 5.4. To create a representative but not overcrowded lineage tree we select all subclones with CC cell counts $\geqslant 10$ at the end of the GCR. Subsequently, we constructed the lineage tree from these subclones and their (unexpanded) ancestors and coloured each node according to their affinity and expansion (Figure 5.8). Effectively, this lineage tree corresponds to the tree shown in Figure 5.4 day 21 but with a sub-selection of subclones. However, because of the sub-selection criterion the resulted tree is smaller than on Figure 5.4 and is only maximum of 5 nodes long.

To identify expanded subclones we determined the threshold T to be 17 cells. Consequently, all subclones with > 17 cells are expanded. Only ten subclones reached the expansion level at the end of the GCR. In general we observe less expansion at lower levels of the lineage tree.

Most of the branches of the tree show subclones with increasing affinity. However, few branches represent subclones with decreasing affinity as result of a SHM. As expected, subclones with low affinity extinct or have low cell counts at the end of the GCR and higher affinity subclones show increased cell counts or have been determined as expanded.

5.4. Discussion

In the current work we used a previously developed mathematical model of affinity maturation to investigate the evolution of B-cell lineage trees during affinity maturation. The model tracks individual subclones and, consequently, enables the construction of B-cell lineage trees. avgRL and maxRL demonstrate that lineage trees become longer during GCR while at the same time the outgoing degrees (rootOD, avgOD) decrease. Consequently, the lineage trees evolve from bushy trees to longer pruned trees. The total number of terminal subclones (L) and total subclones (N) stays relatively constant, around

800. This is in contrast to trees reconstructed and analysed by Dunn-Walters [141], based on experimental data from Jacob [139], where the number of leaves decreases. This could be a shortcoming of our computational model that does not incorporate the GC shutdown. Another possibility is that the previous study was based on a limited experimental data compare to currently available high-throughput repertoire sequencing data, which demonstrate a higher number of nodes per tree. Use of repertoire sequencing experimental data might provide a more fair comparison with our simulation results. Our simulations also demonstrate that the stochastic process of SHM is responsible for a lineage trees that largely vary in size. In contrast, the range of the observed outgoing degrees becomes much smaller with proceeding affinity maturation.

We also explored subclone expansion and affinity maturation in the context of a B-cell lineage tree (Figure 5.8). The lineage tree included subclones that overcame competitors and reached relatively high cell counts at the end of the GCR. As expected, the tree demonstrate the advantage of high affinity subclones, particularly, high abundance or expansion is achieved by high affinity subclones. Noteworthy, the tree also illustrates that the affinity maturation is not necessarily a linear process and high affinity subclones may originate from not necessarily the highest affinity branch. That may be a result of SHM stochasticity, when a low affinity cell may potentially originate a high affinity subclone. Moreover, high cell counts of a subclone also do not guaranty the production of high affinity or highly expanded descendants. Further, high affinity subclones may not develop further but instead lose cell counts due to the high chance of lethal mutations.

For the best of our knowledge, such a representation of subclonal expansion and subclonal affinity in the context of a lineage tree was demonstrated for the first time. This is particularly important because information about subclonal affinity is virtually impossible to obtain from experimental data with current experimental technologies. First, the measurement of BCRs of all subclones at a specific time point during the GCR with repertoire sequencing [96, 158] would require microdissection of the GC and, consequently, would destroy the GC prohibiting further measurements. Alternatively, one could select different GCs at different time points as was done for a limited set of Ig sequences by Jacob [139] and Dunn-Walters [141] but this assumes a strong relationship and synchronicity between these GCs, which may not be true [98]. Furthermore, with current technologies it is impossible to measure the affinity of the BCR for the Ag for all subclones, which would also require that the Ag is known. To overcome the current experimental limitations in our work we successfully used a GCR model to obtain a graph representation of subclonal expansion with corresponding subclonal affinity in the context of a B-cell lineage tree.

Chapter 6
Discussion

Systems biology is a multi-disciplinary rapidly developing research field that focuses on complex (dynamic) non-linear interactions within biological systems such as biological networks. It generally involves a combination of wet-lab experiments and computational approaches. Experimental data is integrated in statistical or mathematical models to generate testable hypotheses, to predict the system's behaviour, and to facilitate discovery and description of the system's properties. A range of approaches towards the modelling of biological networks have been developed and according to Stelling and co-authors may be divided into three categories [159]. The first category involves methods based on interactions between network components only. These methods are often explorative and based on statistical approaches applied to genome-wide omics data such as discussed in Chapter 2 of this thesis. Examples include the construction of co-expression networks [160] and protein-protein interaction networks [161]. A second category consists of constrained based methods that aim to include information such as reaction stoichiometry and reaction reversibility. Flux Balance Analysis is an example from the second category [84]. We did not consider this type of modelling in our research. Finally, there are methods that include detailed interaction mechanisms, for example, ordinary differential equations (ODEs). ODEs are typically used to model the kinetics of metabolic networks [83] or cellular mechanisms such as discussed in Chapter 4 and 5 of this thesis. Interaction-based and constrained-based methods are static methods that do not require detailed parameters of the modeled network. In contrast, dynamic models such as represented by ODEs generally require such information to be applied successfully. Moreover, static models provide a qualitative description of the system dynamics in comparison to the quantitative results of ODEs. In our research we also considered network-based models (Petri nets) such as discussed in Chapter 3. These models can either be implemented as static models or as dynamic models.

All the modelling frameworks mentioned here can make use of prior biological knowledge either to define the model's topology and parameters in knowledge-driven modelling approaches or to guide the modelling process in data-driven approaches by directly incorporating the prior knowledge in the modelling method. In our research we demonstrated several approaches to accomplish this.

6.1. Prior knowledge in statistical models

In chapter 2 we reviewed more than twenty high-throughput data analysis methods in transcriptomics and metabolomics that incorporating prior knowledge to restrict or guide the statistical modelling. We highlighted features and differences of the methods and the type of prior knowledge that was used. However, it is extremely difficult to compare different methods without a proper framework which would allow a fair compar-

ison and would further facilitate understanding of how prior knowledge influences the results. Such framework could be based on an appropriate synthetic datasets. For example, a prototype of such test framework can be based on the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project. DREAM aims to provide a framework for a fair and rigorous testing of various gene network inference methods. DREAM suggests to infer a gene network from simulated gene expression data without sharing the details of the kinetic model that was used to generate the data [162]. Various error tests allow to measure a performance of every method and to show the difficulties faced by each of the applied methods. For example, if two genes are co-regulated by a single transcription factor their transcription levels correlate. This correlation may lead to a false prediction of an interaction between the two co-regulated nodes. Analysis of such network motifs helps to reveal systematic prediction errors of the tested methods. Further, based on the used synthetic network, prior knowledge may be generated, perhaps with variable percentage of missing or false positive gene interaction links. This would allow both to compare different methods that incorporate prior knowledge in network inference and to assess the added value of the prior knowledge in each method.

Despite that synthetic data certainly may offer a valuable basis for a validation framework, wet-lab experiments remain highly desirable. Few projects present an example of a possible experimental system where an artificial network has been constructed and incorporated into a cell. For example, a synthetic gene network in yeast as a part of a test framework for systems biology approaches has been presented by Cantone and co-authors [163]. Another example is a synthetic gene network integrated in human kidney cells by Kang and co-authors [164]. These examples present gene networks with known interactions and thus suitable to generate wet-lab experimental data for further use in a validation framework.

6.2. Prior knowledge to model genistein elimination pathway with Petri nets

Petri nets are used to study the dynamics of biological systems (for a review see [165]). Similar to ODEs, Petri nets provide a formal mathematical framework for the analysis of biological systems. Various extensions for Petri nets have been developed to qualitatively or quantitatively model networks. Basic (also referred as original or time-less) Petri nets require only the topological structure and stoichiometry of the studied network. While they provide some insight in possible dynamics the result of the analysis is qualitative [77, 166]. Further two examples are based on the assumption that kinetic parameters are less important than network topology and therefore topology based models are able to provide information about system dynamics, thus providing quantitative insights. A method of Ruths and co-workers uses this assumption to develop a strategy for non-parametric Petri net modeling and execution that uses token distribution and sampling to reproduce the dynamics of cellular signaling networks [81]. A method of Kuffner and co-workers is based on fuzzy logic and provides a very good estimation of gene regulatory networks from gene expression data *in silico* [19]. The authors argued that their Petri net extension provides a simpler discrete modelling system compared to more detailed ODEs. Examples of Petri nets that aim to quantitatively model networks comprise

Stochastic Petri nets [16], Time Petri Nets [17], and Hybrid Functional Petri Nets [18]. Similar to ODEs they require kinetic parameters of the system.

However, so far, it seems to be ignored that even in the absence of kinetic parameters Petri nets may directly be converted to differential equations if all the kinetic parameters can be obtained through parameter estimation procedures. Using the same prior knowledge, our Petri net of the genistein elimination pathway can also be modelled with ODEs:

$$
\begin{aligned}
\frac{dG_{GL}}{dt} &= -(F1+F7)G_{GL} + F11G_L \\
\frac{dG_{GE}}{dt} &= -(F2+F30+F31)G_{GE} + F1G_{GL} \\
\frac{dG_L}{dt} &= -(F3+F11)G_L + F2G_{GE} + F6G_{VB} \\
\frac{dG_A}{dt} &= -F4G_A + F3G_L \\
\frac{dG_O}{dt} &= -(F5+F8)G_O + F4G_A \\
\frac{dG_{VB}}{dt} &= -F6G_{VB} + F5G_O \\
\frac{dGG_{GL}}{dt} &= -(F17+F16)GG_{GL} + F20GG_L \\
\frac{dGG_{GE}}{dt} &= -F12GG_{GE} + F16GG_{GL} + F30G_{GE} \\
\frac{dGG_L}{dt} &= -(F13+F20)GG_L + F12GG_{GE} + F19GG_{VB} \\
\frac{dGG_A}{dt} &= -F14GG_A + F13GG_L \\
\frac{dGG_O}{dt} &= -(F18+F15)GG_O + F14GG_A \\
\frac{dGG_{VB}}{dt} &= -F19GG_{VB} + F15GG_O \\
\frac{dS_{GL}}{dt} &= -(F25+F26)S_{GL} + F29S_L \\
\frac{dS_{GE}}{dt} &= -F21S_{GE} + F31G_{GE} + F25S_{GL} \\
\frac{dS_L}{dt} &= -(F22+F29)S_L + F12S_{GE} + F28S_{VB} \\
\frac{dS_A}{dt} &= -F23S_A + F22S_L \\
\frac{dS_O}{dt} &= -(F24+F27)S_O + F23S_A \\
\frac{dS_{VB}}{dt} &= -F28S_{VB} + F24S_O
\end{aligned}
$$

All parameters in this ODE model can then be estimated through parameter estimation based on experimental data as was done with the Petri net model (Chapter 3). Then the question is what would be the advantage of the Petri net model over the ODE based model? Also if both methods allow the dynamic analysis then what would be the difference between results of two methods? Additional work is required, which would compare genistein elimination Petri net and ODE based models. This work hopefully would lead to a grounded advice when to choose Petri nets over ODEs and vice versa in the situation of lack of the exact topology and kinetic parameters.

6.3. Prior knowledge to model B-cell affinity maturation with differential equations

The B-cell affinity maturation has been intensively studied for a few decades, however, its precise mechanism still remains to be elucidated. For modelling a Germinal Centre Reaction (GCR) in Chapter 4 and 5 the known details were not enough to set equations that would precisely describe the crucial B-cell selection mechanisms. For example, following cell division and somatic hypermutation, B-cells are programmed to undergo apoptosis unless they receive survival signals through interactions with the antigen and T follicular helper cells. These selection mechanisms impose competition between the B-cell subclones, which is assumed to be based on their relative BCR affinities. The precise mechanisms involved in B-cell competition is unknown. To avoid assumptions due to the lack of knowledge and to avoid an overly complex model, antigene and T follicular helper survival signals were modelled with a general sigmoidal function. This function converts relative B-cell affinities to a signal strength between 0 and 1. Despite this simplification, the model successfully generated valuable insights in B-cell affinity distribution after affinity maturation.

6.4. Databases as stores of prior knowledge

There are many sources of knowledge that may be used as prior knowledge in the analysis of biological data and systems. Perhaps the most widely used sources are expert domain knowledge and traditional (low-throughput) experiments that are very precise and reliable. However, knowledge from an expert can only be used in data analysis if we capture his knowledge in a computer accessible format, which is time consuming and requires significant effort for a stable collaboration and to ensure that the domain expert also benefits from such effort. A store of knowledge with easy access may facilitate collaboration with domain experts. Moreover, a database may help to structure knowledge from a large range of interdisciplinary studies, which otherwise would be too comprehensive and complex to be absorbed by one mind.

Several technologies have been suggested to support biological databases varying from saving data in a comma separated file to complex object-oriented databases [167, 168]. One of the most widely used technologies to store biological data is a relational database [169]. Some authors emphasize that relational databases provide a straightforward way to think about biological knowledge in the information way and allow to facilitate collaborations between experts and to cover large areas of knowledge

[170]. Many relational biological databases with a variety of content, purpose, and technical characteristics have been created [171].

Recently, a new promising approach, i.e., the Resource Description Framework (RDF) technology has been suggested [172]. A large effort to standardize RDF has been taken by W3 community (https://w3.org/RDF/), which greatly facilitates the use of the technology. Moreover, RDF does not require a fixed list of biological entities and relations as is required by other standards. The format flexibility combined with the standardization effort makes it easier to create biological data stores, to share and integrate data among various fields and to promote database evaluation over time [173, 174]. As a result, this technology has been successfully applied in systems biology to create databases that facilitate the decision making procedure and experimental data analysis. Venkatesan and co-authors created the Gene eXpression Knowledge Base (GeXKB) - a database that contains integrated knowledge about gene expression regulation [175]. The flexibility of the technology allowed to integrate the database and experimental gene expression data to explore new potential candidates for regulatory network extensions. Willemsen and co-authors used the technology to create a comprehensive knowledge base which focuses on four key peroxisome pathways and several related genetic disorders in humans [176]. The authors particularly have focused on creating a general framework to construct biomedical knowledge bases from scattered resources and on its visualization aspects.

Our research would greatly benefit from a knowledge base containing details of genistein elimination pathway (Chapter 3) or B-cell maturation process (Chapter 4). Considering the capacity of knowledge bases to organize and share data, serve as an expert communication platform, and to facilitate knowledge visualisation, the creation of such a knowledge base may be advised as a first step in biological systems modelling.

6.5. Biomedical text mining as a source of prior knowledge

Without doubts, scholarly documents provide the biggest source of scientific knowledge: millions of scholarly documents are currently available from literature databases (e.g. PubMed) and other repositories [177]. Such huge amount of data makes the manual extraction of relevant information extremely time consuming. Text mining provides a solution to this problem. In general, this scientific field is called natural language processing (NLP). The main idea behind NLP is to define and recognize terms used in the domain of interest, to create a collection of these terms together with annotation describing their meaning, and to use this collection to analyse a large corpus of text looking for co-occurrence among terms identified in the text. Further, co-occurrence is used in various statistical tests to find relations among terms and consequently suggest a relation among corresponding biological entities. Text mining may be used as a stand alone tool to discover relationships among biological terms or to create biological databases for various purposes [178, 179, 180, 181]. Moreover, text mining has been used to create advanced search engines that aim to speed up and facilitate literature search for specific topics [182, 183]. Use of such search engines would greatly improve time spend on knowledge gathering during modelling process. Worthy to notice that biomedical sources used in text mining are not limited by scientific literature in biology, medicine, and chemistry. Information from medical internet communities and patient records also

contain valuable knowledge as well and some interesting applications already have been suggested (e.g., [184, 185, 186, 187]).

Chapter 7
Summary

Use of prior knowledge to plan experiments or to compare results with already known details has always been a crucial step in scientific research. Quick development of high-throughput experimental techniques and information technologies allows to evolve the use of prior knowledge even further. This work explored the use of prior knowledge as a basis for biological systems modelling and analysis. We discussed (1) the incorporation of prior knowledge in the analysis of high-throughput data in transcriptomics and metabolomics, and (2) the use of incomplete prior knowledge to build models of biological systems.

Chapter 2 reviews methods in transcriptomics and metabolomics that incorporate prior knowledge in the analysis of high throughput data. We specifically focused on a collection of methods that incorporated prior knowledge to estimate model parameters; we excluded methods that used prior knowledge to verify or validate the final results of a model or analysis. We divided the reviewed methods into three groups based on the underlying mathematical model: exploratory methods, supervised methods, and estimation of covariance matrices. By defining relationships among variables in high throughput data based on known *a priori* knowledge the reviewed methods reduce the solution space and/or focus the analysis on biological meaningful results. In this way the methods lead the analysis towards underlying biology. Despite this advantage, incorporation of prior knowledge into a model is not widely used. We concluded that the definition and acceptance of a common test framework to test methods incorporating prior knowledge and to test the prior knowledge influence on results is missing and urgently needed. The test framework would help to understand when and how to optimally apply prior knowledge in data analysis methods. Moreover, it should help to understand when prior knowledge is not correct or not appropriate for the analysed system.

Next, in Chapter 3 we used incomplete and scattered knowledge about the human genistein elimination pathway to build a Petri net model. Scattered knowledge in conjunction with the complicated nature of alternative genistein elimination routes hampers building and parameterization of quantitative models. For this reason, we suggested that the network structure alone might contain enough information to study the system dynamics. Using the Petri net model we showed that widely used metabolic profiles solely measured in venous blood were not sufficient to uniquely parameterize the model. Additional simulations based on the model suggested that gut epithelium metabolite profiles would allow to infer the relative contributions of concurrent elimination routes with higher accuracy, and to improve the reconstruction of concentration profiles of all metabolites in this pathway. Overall, we showed that a Petri net model based on scarce prior knowledge may be used to explore the pathway properties and to assist in the design of future experiments to complete missing knowledge.

In Chapter 4 we built an ODE model to determine the affinity distribution among B-cell populations measured with RNA repertoire sequencing during an immune response. While a lot is known about B-cell maturation, many important details remain to be elucidated. Particularly, while the lack of specific details about B cells, T cells, and antigen interactions prohibit the implementation of a precise model, these interactions determine which B-cells survive and, therefore, determine the affinities observed in the B-cell population. To overcome this lack of knowledge but also to avoid an overly complex model we suggested a simplification through a general sigmoidal function that imposes competition between B cells without the implementation of mechanistic details of B cells, T cells and antigen interactions. Despite this simplification, the model successfully generated valuable insights in the B-cell affinity distribution during affinity maturation. The result is intriguing because we show that expanded clones, widely used for further downstream analysis, might not be the highest affinity cells. We hope that this result will get experimental validation in the near future and will have impact in future clinical strategies for selection and characterization of B cells.

In Chapter 5 we used the ODE model of B-cell maturation to explore and visualize the evolution of B-cell lineage trees during affinity maturation. We followed changes in lineage tree parameters such as total number of nodes, node outgoing degrees and tree length, with progression of the immune response. Our simulations showed that lineage tree sizes largely varied while the range of the observed outgoing degrees became much smaller with proceeding affinity maturation. Moreover, our model allowed to investigate the B-cell affinity maturation in a novel way that is currently virtually impossible to obtain using experimental data. Particularly, the model allowed to simultaneously follow affinity changes and subclonal abundance (cell counts) in the context of B-cell lineage trees. It showed that the affinity maturation is not necessarily a linear process and high cell counts of a subclone do not guarantee the production of high affinity or highly expanded descendants. In general, the work in Chapters 4 and 5 expanded our understanding of the B-cell maturation.

Chapter 8
Samenvatting

Het gebruik van voorkennis om nieuwe experimenten te plannen of om resultaten te vergelijken met bekende details, is altijd al een cruciale stap geweest in wetenschappelijk onderzoek. De snelle ontwikkeling van "high-throughput" experimentele technieken en informatietechnologieëen maken het mogelijk om het gebruik van voorkennis nog verder te laten groeien. Dit werk exploreert het gebruik van voorkennis als een basis voor het modelleren en analyseren van biologische systemen. We hebben laten zien (1) hoe voorkennis kan worden geïncorporeerd in de analyse van high-throughput data afkomstig van transcriptomics en metabolomics, en (2) hoe incomplete voorkennis kan worden gebruikt om modellen te bouwen van biologische systemen.

Hoofdstuk 2 geeft een overzicht van methoden die voorkennis incorporeren en die worden gebruikt voor de analyse van transcriptomics en metabolomics high-throughput data. We hebben ons specifiek gefocusseerd op methoden die voorkennis incorporeren om modelparameters te schatten; we hebben niet gekeken naar methoden die voorkennis gebruiken om eindresultaten van modellen of analyse te verifiëren of te valideren. We hebben de beschreven methoden op basis van het onderliggende mathematische model ingedeeld in drie groepen: explorerende methoden, gesuperviseerde methoden, en methoden voor de schatting van covariantie matrices.

Door relaties tussen variabelen die gemeten zijn met high-throughput technieken te definiëren op basis van *a priori* informatie, kunnen de beschreven methoden de oplossingsruimte reduceren en/of de analyse focussen op biologische relevante resultaten. Op deze manier kunnen deze methoden de analyse leiden naar de onderliggende biologie. Ondanks dit voordeel wordt voorkennis nog niet veel geïncorporeerd in data analyse methoden. We concluderen dat de definitie en acceptatie van een gemeenschappelijk raamwerk om deze klasse van methoden en hun resultaten te kunnen testen op dit moment ontbreekt maar wel hard nodig is. Zo'n test raamwerk kan ons helpen te begrijpen wanneer en hoe we optimaal gebruik kunnen maken van voorkennis in data-analyse methoden. Bovendien zou dit duidelijk moeten maken wanneer het gebruik van voorkennis niet geschikt of niet correct is voor het systeem dat wordt geanalyseerd.

In Hoofdstuk 3 hebben we gebruik gemaakt van incomplete en verdeelde kennis over de humane genistein eliminatie route om een Petri net model te bouwen. De verspreidde kennis over en de complexiteit van alternatieve genistein eliminatie routes maken het moeilijk om een kwantitatief model te bouwen en te parameteriseren. Hierom hebben we gesuggereerd dat alleen het gebruik van de netwerkstructuur voldoende zou kunnen zijn om de dynamica van dit systeem te bestuderen. Door gebruik te maken van het Petri net model hebben we laten zien dat de veelgebruikte metabolietprofielen gemeten in aderlijk bloed niet voldoende zijn om het model uniek te parameteriseren. Verdere simulaties gebaseerd op dit model suggereren dat metabolietprofielen gemeten in dar-

mepitheelcellen het mogelijk zouden maken om de relatieve bijdragen van de parallelle eliminatieroutes met meer nauwkeurigheid in kaart te brengen, en om de concentratieprofielen van alle metabolieten in dit netwerk beter te kunnen reconstrueren. In het algemeen hebben we laten zien dat Petri net modellen gebaseerd op beperkte voorkennis kunnen worden gebruikt om netwerk eigenschappen in kaart te brengen en om te assisteren bij het opzetten van toekomstige experimenten om ontbrekende informatie aan te vullen.

In Hoofdstuk 4 presenteren we een ODE model dat kan helpen om de affiniteitsdistributie te bepalen van een populatie van B cellen die gemeten zijn met RNA repertoire sequensen tijdens een immuunrespons. Ondanks dat er veel bekend is over B-cel rijping zijn er veel details nog onduidelijk. In het bijzonder zijn nog veel details onbekend over de interactie tussen B cellen, T cellen, en de het antigen zodat een precies model nog niet kan worden geconstrueerd. Om dit gebrek aan kennis te omzeilen en om te voorkomen dat we een te complex model krijgen, hebben we voorgesteld om een algemene sigmoïde functie te gebruiken om competitie tussen B cellen te modelleren zonder de mechanistische details te implementeren. Ondanks deze simplificatie is het model succesvol gebleken in het genereren van waardevolle inzichten in de distributie van affiniteiten tijdens affiniteitsrijping. Het resultaat is intrigerend omdat we laten zien dat geëxpandeerde klonen, die gebruikt worden voor verdere analyse, niet *per se* de hoogste affiniteit hebben. We hopen dat dit resultaat in de toekomst experimenteel kan worden gevalideerd en een impact zal hebben op klinische strategieën voor de selectie en karakterisatie van B cellen.

In Hoofdstuk 5 hebben we gebruik gemaakt van het ODE model van affiniteitsrijping om de evolutie van B-cel “lineage trees” te exploreren en te visualiseren tijdens affiniteitsrijping. We hebben veranderingen in de parameters (zoals totaal aantal knopen, uitgaande graad van de knopen, en de boomlengte) van de lineage tree gevolgd tijdens de voortgang van de immuunrespons. Onze simulaties hebben laten zien dat de grootte van de lineage trees sterk varieerden terwijl het bereik van de uitgaande graden kleiner werd naarmate de affiniteitsrijping vorderde. Ons model maakte het ook mogelijk om B-cel affiniteitsrijping op een manier te onderzoeken die op dit moment nog niet mogelijk is met experimentele data. In het bijzonder maakte ons model het mogelijk om simultaan veranderingen in affiniteiten en sub-klonale abundantie (celaantallen) te volgen in de context van B-cel lineage trees. Dit liet zien dat affiniteitsrijping niet noodzakelijk een lineair proces is en dat hoge celaantallen van een sub-kloon niet de productie van hoge affiniteit of hoge abundante nakomelingen garandeert. In het algemeen draagt het werk in de Hoofdstukken 4 en 5 bij aan een verder begrip van B-cel affiniteitsrijping.

References

[1] Jackson, J. E. 1991. *A User's Guide to Principal Components*. Wiley Interscience, New York

[2] Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22: 281–5

[3] Baldan, P., N. Cocco, A. Marin, and M. Simeoni. 2010. Petri nets for modelling metabolic pathways: a survey. *Nat. Comput.* 9: 955–989

[4] Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303: 799–805

[5] Wang, R.-S., A. Saadatpour, and R. Albert. 2012. Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.* 9: 055001

[6] Morris, M. K., J. Saez-Rodriguez, P. K. Sorger, and D. A. Lauffenburger. 2010. Logic-based models for the analysis of cell signaling networks. *Biochemistry* 49: 3216–3224

[7] Westerhuis, J. A., E. P. P. A. Derks, H. C. J. Hoefsloot, and A. K. Smilde. 2007. Grey component analysis. *J. Chemom.* 21: 474–485

[8] Reshetova, P., A. K. Smilde, A. H. C. van Kampen, and J. A. Westerhuis. 2014. Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. *BMC Syst. Biol.* 8 Suppl 2: S2

[9] Schlatter, R., K. Schmich, I. A. Vizcarra, P. Scheurich, T. Sauter, C. Borner, M. Ederer, I. Merfort, and O. Sawodny. 2009. ON/OFF and beyond - A Boolean model of apoptosis. *PLoS Comput. Biol.* 5: e1000595

[10] Sahoo, D., D. L. Dill, A. J. Gentles, R. Tibshirani, and S. K. Plevritis. 2008. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* 9: R157

[11] The Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res.* 36: D440–444

[12] Linde, J., S. Schulze, S. G. Henkel, and R. Guthke. 2015. Data- and Knowledge-Based Modeling of Gene Regulatory Networks: an Update. *EXCLI J.* 14: 346–378

[13] Tian, Y., B. Zhang, E. P. Hoffman, R. Clarke, Z. Zhang, I.-M. Shih, J. Xuan, D. M. Herrington, and Y. Wang. 2014. Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks. *BMC Syst. Biol.* 8: 87

[14] Olsen, C., K. Fleming, N. Prendergast, R. Rubio, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains, and J. Quackenbush. 2014. Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* 103: 329–336

[15] Petri, C. A. 1962. *Kommunikation mit Automaten*. Ph.D. thesis

[16] Goss, P. J. and J. Peccoud. 1998. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc. Natl. Acad. Sci. U. S. A.* 95: 6750–6755

[17] Popova-Zeugmann, L., M. Heiner, and I. Koch. 2005. Time Petri Nets for Modelling and Analysis of Biochemical Networks. *Fundam. Informaticae* 67: 149–162

[18] Matsuno, H., Y. Tanaka, H. Aoshima, A. Doi, M. Matsui, and S. Miyano. 2011. Biopathways representation and simulation on hybrid functional petri net. *Stud. Health Technol. Inform.* 162: 77–91

[19] Küffner, R., T. Petri, L. Windhager, and R. Zimmer. 2010. Petri nets with fuzzy logic (PNFL): Reverse engineering and parametrization. *PLoS One* 5: 1–10

[20] Breitling, R., D. Gilbert, M. Heiner, and R. Orton. 2008. A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. *Brief. Bioinform.* 9: 404–421

[21] Saez-Rodriguez, J., L. Simeoni, J. A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U. U. Haus, R. Weismantel, E. D. Gilles, S. Klamt, and B. Schraven. 2007. A logical model provides insights into T cell receptor signaling. *PLoS Comput. Biol.* 3: 1580–1590

[22] Gupta, S., S. S. Bisht, R. Kukreti, S. Jain, and S. K. Brahmachari. 2007. Boolean network analysis of a neurotransmitter signaling pathway. *J. Theor. Biol.* 244: 463–469

[23] Li, F., T. Long, Y. Lu, Q. Ouyang, and C. Tang. 2004. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U. S. A.* 101: 4781–4786

[24] Husmeier, D. 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* 19: 2271–2282

[25] Kim, S. Y., S. Imoto, and S. Miyano. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* 4: 228–235

[26] Zou, M. and S. D. Conzen. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21: 71–9

[27] Dojer, N., A. Gambin, A. Mizera, B. Wilczyński, and J. Tiuryn. 2006. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics* 7: 249

[28] Wu, M., X. Yang, and C. Chan. 2009. A dynamic analysis of IRS-PKR signaling in liver cells: A discrete modeling approach. *PLoS One* 4: e8040

[29] Aldridge, B. B., J. Saez-Rodriguez, J. L. Muhlich, P. K. Sorger, and D. A. Lauffenburger. 2009. Fuzzy Logic Analysis of Kinase Pathway Crosstalk in TNF/EGF/Insulin-Induced Signaling. *PLoS Comput. Biol.* 5: e1000340

[30] Mendoza, L. and I. Xenarios. 2006. A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theor. Biol. Med. Model.* 3: 13

[31] Glass, L. and S. A. Kauffman. 1973. The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39: 103–129

[32] Shmulevich, I., E. R. Dougherty, S. Kim, and W. Zhang. 2002. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18: 261–274

[33] Trairatphisan, P., A. Mizera, J. Pang, A. A. Tantar, J. Schneider, and T. Sauter. 2013. Recent development and biomedical applications of probabilistic Boolean networks.. *Cell Commun. Signal.* 11: 46

[34] Murray, J. D. 2002. *Mathematical Biology I. An introduction.* Springer

[35] Eungdamrong, N. J. and R. Iyengar. 2004. Computational approaches for modeling regulatory cellular networks. *Trends Cell Biol.* 14: 661–669

[36] Tomita, M., K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. A. Hutchison. 1999. E-CELL: Software environment for whole-cell simulation. *Bioinformatics* 15: 72–84

[37] Slepchenko, B. M., J. C. Schaff, I. Macara, and L. M. Loew. 2003. Quantitative cell biology with the Virtual Cell. *Trends Cell Biol.* 13: 570–576

[38] Smith, A. E., B. M. Slepchenko, J. C. Schaff, L. M. Loew, and I. G. Macara. 2002. Systems analysis of Ran transport. *Science* 295: 488–491

[39] Chen, K.-C., T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao. 2005. A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. *Bioinformatics* 21: 2883–2890

[40] Daniels, B. C., Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers. 2008. Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotechnol.* 19: 389–395

[41] Gutenkunst, R. N., J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. 2007. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 3: 1871–1878

[42] Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40: D109–114

[43] Wingender, E. 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* 9: 326–332

[44] Namkung, J., P. Raska, J. Kang, Y. Liu, Q. Lu, and X. Zhu. 2011. Analysis of exome sequences with and without incorporating prior biological knowledge. *Genet Epidemiol.* 35: S48–55

[45] Ramakrishnan, S., C. Vogel, T. Kwon, L. Penalva, E. Marcotte, and D. Miranker. 2009. Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics* 25: 2955–2961

[46] Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102: 15545–15550

[47] Xia, J. and D. S. Wishart. 2010. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* 38: W71–77

[48] Ackermann, M. and K. Strimmer. 2009. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10: 47

[49] Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2008. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37: 1–13

[50] van den Berg, R. A., C. M. Rubingh, J. A. Westerhuis, M. J. van der Werf, and A. K. Smilde. 2009. Metabolomics data exploration guided by prior knowledge. *Anal. Chim. Acta* 651: 173–181

[51] Liao, J. C., R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* 100: 15522–15527

[52] Yu, T. and K.-C. Li. 2005. Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics* 21: 4033–4038

[53] Tran, L. M., D. R. Hyduke, and J. C. Liao. 2010. Trimming of mammalian transcriptional networks using network component analysis. *BMC Bioinformatics* 11: 511

[54] Cheng, J., M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M. A. Siani-Rose. 2004. A knowledge-based clustering algorithm driven by Gene Ontology. *J. Biopharm. Stat.* 14: 687–700

[55] Kustra, R. and A. Zagdanski. 2006. Incorporating Gene Ontology in Clustering Gene Expression Data. *Proc. 26th IEEE Int. Symp. Comput. Med. Syst.*. 555–563

[56] Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18: S145–154

[57] Dotan-Cohen, D., A. A. Melkman, and S. Kasif. 2007. Hierarchical tree snipping: clustering guided by prior knowledge. *Bioinformatics* 23: 3335–3342

[58] Tseng, G. C. 2007. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* 23: 2247–2255

[59] Shen, Y., W. Sun, and K.-C. Li. 2009. Dynamically weighted clustering with noise set. *Bioinformatics* 26: 341–347

[60] Pan, W. 2006. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics* 22: 795–801

[61] Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. 2003. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99

[62] Hendrickx, D. M., H. C. Hoefsloot, M. M. Hendriks, A. B. Canelas, and A. K. Smilde. 2012. Global test for metabolic pathway differences between conditions. *Anal. Chim. Acta* 719: 8–15

[63] Chuang, H., E. Lee, Y. Liu, D. Lee, and T. Ideker. 2007. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3: 140

[64] Rapaport, F., A. Zinovyev, M. Dutreix, E. Barillot, and J. Vert. 2007. Classification of microarray data using gene networks. *BMC Bioinformatics* 8: 35

[65] Li, C. and H. Li. 2008. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24: 1175–1182

[66] Dutkowski, J. and T. Ideker. 2011. Protein networks as logic functions in development and cancer. *PLoS Comput. Biol.* 7: e1002180

[67] Schäfer, J. and K. Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4: Article32

[68] Tai, F. and W. Pan. 2007. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics* 23: 3170–7

[69] Guillemot, V., A. Tenenhaus, L. Le Brusquet, and V. Frouin. 2011. Graph constrained discriminant analysis: a new method for the integration of a graph into a classification process. *PLoS One* 6: e26146

[70] Jelizarow, M., V. Guillemot, A. Tenenhaus, K. Strimmer, and A.-L. Boulesteix. 2010. Over-optimism in bioinformatics: an illustration. *Bioinformatics* 26: 1990–1998

[71] Staiger, C., S. Cadot, R. Kooter, M. Dittrich, T. Müller, G. W. Klau, and L. F. A. Wessels. 2012. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One* 7: e34796

[72] Zaidi, N., L. Lupien, N. B. Kuemmerle, W. B. Kinlaw, J. V. Swinnen, and K. Smans. 2013. Lipogenesis and lipolysis: The pathways exploited by the cancer cells to acquire fatty acids. *Prog. Lipid Res.* 52: 585–589

[73] Cui, Q., Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, E. O. Purisima, and E. Wang. 2007. A map of human cancer signaling. *Mol. Syst. Biol.* 3: 152

[74] Stephens, P. J., C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. 2011. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* 144: 27–40

[75] Vitale, D. C., C. Piazza, B. Melilli, F. Drago, and S. Salomone. 2013. Isoflavones: estrogenic activity, biological effect and bioavailability. *Eur. J. Drug Metab. Pharmacokinet.* 38: 15–25

[76] Reddy, V. N., M. L. Mavrovouniotis, and M. N. Liebman. 1993. Petri net representations in metabolic pathways. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB. Int. Conf. Intell. Syst. Mol. Biol.*. Chemical Engineering Department, University of Maryland, College Park 20742, USA.. volume 1. 328–336

[77] Koch, I., B. H. Junker, and M. Heiner. 2005. Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics* 21: 1219–1226

[78] Zevedei-Oancea, I. and S. Schuster. 2003. Topological analysis of metabolic networks based on Petri net theory. *In Silico Biol.* 3: 323–345

[79] Masoudi-Nejad, A., A. Moeini, I. Nassiri, and M. Jalili. 2012. Nonparametric Simulation of Signal Transduction Networks with Semi-Synchronized Update. *PLoS One* 7: e39643

[80] Hardy, S. and P. N. Robillard. 2008. Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways. *Bioinformatics* 24: 209–217

[81] Ruths, D., M. Muller, J.-T. Tseng, L. Nakhleh, and P. T. Ram. 2008. The Signaling Petri Net-Based Simulator: A Non-Parametric Strategy for Characterizing the Dynamics of Cell-Specific Signaling Networks. *PLoS Comput. Biol.* 4: 15

[82] Smit, S., E. Szymanska, I. Kunz, V. G. Roldan, M. VanTilborg, P. Weber, K. Prudence, F. van der Kloet, J. van Duynhoven, A. Smilde, R. de Vos, and I. Bendik. 2014. Nutrikinetic modeling reveals order of genistein phase II metabolites appearance in human plasma. *Mol. Nutr. Food Res.* 58: 2111–2121

[83] Peskov, K., E. Mogilevskaya, and O. Demin. 2012. Kinetic modelling of central carbon metabolism in Escherichia coli. *FEBS J.* 279: 3374–85

[84] Orth, J. D., I. Thiele, and B. Ø. Palsson. 2010. What is flux balance analysis?. *Nat. Biotechnol.* 28: 245–248

[85] Willemsen, A. M., D. M. Hendrickx, H. C. J. Hoefsloot, M. M. W. B. Hendriks, S. A. Wahl, B. Teusink, A. K. Smilde, and A. H. C. van Kampen. 2015. MetDFBA: incorporating time-resolved metabolomics measurements into dynamic flux balance analysis. *Mol Biosyst.* 11: 137–45

[86] Nelder, J. A. and R. Mead. 1965. A Simplex Method for Function Minimization. *Comput. J.* 7: 308–313

[87] Raue, A., C. Kreutz, T. Maiwald, U. Klingmuller, and J. Timmer. 2011. Addressing parameter identifiability by model-based experimentation. *IET Syst. Biol.* 5: 120–130

[88] Kirkpatrick, S., C. D. G. Jr., and M. P. Vecchi. 1983. Optimization by simmulated annealing. *Science* 220: 671–680

[89] Archetti, F. and F. Schoen. 1984. A survey on the global optimization problem: General theory and computational approaches. *Ann. Oper. Res.* 1: 87–110

[90] van Kampen, A. H. C. and L. M. C. Buydens. 1997. The Ineffectiveness of Recombination in a Genetic Algorithm for the Structure Elucidation of a Heptapeptide in Torsion Angle Space. A Comparison to Simulated Annealing. *Chemom. Intell. Lab. Syst.* 36: 141–52

[91] Victora, G. D. and M. C. Nussenzweig. 2012. Germinal Centers. *Annu. Rev. Immunol.* 30: 429–457

[92] De Silva, N. S. and U. Klein. 2015. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* 15: 137–148.

[93] Klarenbeek, P. L., P. P. Tak, B. D. C. van Schaik, A. H. Zwinderman, M. E. Jakobs, Z. Zhang, A. H. C. van Kampen, R. A. W. van Lier, F. Baas, and N. de Vries. 2010. Human T cell memory consists mainly of unexpanded clones. *Immunol. Lett.* 133: 42–48.

[94] Calis, J. J. A. and B. R. Rosenberg. 2014. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* 35: 581–590.

[95] Tan, Y. C., L. K. Blum, S. Kongpachith, C. H. Ju, X. Cai, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin. Immunol.* 151: 55–65.

[96] Yaari, G., J. I. C. Benichou, J. A. Vander Heiden, S. H. Kleinstein, and Y. Louzoun. 2015. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370: 20140242

[97] Kleinstein, S. H. and J. P. Singh. 2001. Toward quantitative simulation of germinal center dynamics: biological and modeling insights from experimental validation. *J. Theor. Biol.* 211: 253–275.

[98] Wittenbrink, N., T. S. Weber, A. Klein, A. A. Weiser, W. Zuschratter, M. Sibila, J. Schuchhardt, and M. Or-Guil. 2010. Broad volume distributions indicate non-synchronized growth and suggest sudden collapses of germinal center B cell populations. *J. Immunol.* 184: 1339–47

[99] Or-Guil, M., N. Wittenbrink, A. A. Weiser, and J. Schuchhardt. 2007. Recirculation of germinal center B cells: A multilevel selection strategy for antibody maturation. *Immunol. Rev.* 216: 130–141.

[100] Robinson, W. H. 2014. Sequencing the functional antibody repertoire — diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* 11: 1–12.

[101] Doorenspleet, M. E., P. L. Klarenbeek, M. J. H. de Hair, B. D. C. van Schaik, R. E. E. Esveldt, A. H. C. van Kampen, D. M. Gerlag, A. Musters, F. Baas, P. P. Tak, and N. de Vries. 2014. Rheumatoid arthritis synovial tissue harbours dominant B cell and plasma cell clones associated with autoreactivity. *Ann. Rheum. Dis.* 73: 756–762.

[102] Oprea, M. and A. S. Perelson. 1997. Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts. *J. Immunol.* 158: 5155–5162

[103] Shlomchik, M. J., P. Watts, M. G. Weigert, and S. Litwin. 1998. Clone: a Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation, and antigen-driven selection. *Curr. Top. Microbiol. Immunol.* 229: 173–197

[104] Shahaf, G., M. Barak, N. S. Zuckerman, N. Swerdlin, M. Gorfine, and R. Mehr. 2008. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: A large-scale simulation study. *J. Theor. Biol.* 255: 210–222

[105] Meyer-Hermann, M. 2002. A mathematical model for the germinal center morphology and affinity maturation. *J. Theor. Biol.* 216: 273–300.

[106] de Hair, M., I. Zijlstra, M. Boumans, M. van de Sande, M. Maas, D. Gerlag, and P. Tak. 2012. Hunting for the pathogenesis of rheumatoid arthritis: core-needle biopsy of inguinal lymph nodes as a new research tool. *Ann. Rheum. Dis.* 71: 1911–1912.

[107] Giudicelli, V., D. Chaume, and M.-P. Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic acids research* 33: D256–D261

[108] Kent, W. J. 2002. BLAT — The BLAST -Like Alignment Tool. *Genome research* 12: 656–664

[109] Lefranc, M. P. 2014. Immunoglobulin and T cell receptor genes: IMGT and the birth and rise of immunoinformatics. *Front. Immunol.* 5: 22

[110] R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* 1: 409

[111] Soetaert, K., T. Petzoldt, and R. W. Setzer. 2010. Solivng Differential Equations in R: Package deSolve. *J. Stat. Softw.* 33: 1–25

[112] Jacob, J., R. Kassir, and G. Kelsoe. 1991. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl . I . The architecture and dynamics of responding cell populations. *J. Exp. Med.* 173: 1165–1175

[113] Levy, N. S., U. V. Malipiero, S. G. Lebecque, and P. J. Gearhart. 1989. Early onset of somatic mutation in immunoglobulin VH genes during the primary immune response. *J. Exp. Med.* 169: 2007–2019

[114] Berek, C., A. Berger, and M. Apel. 1991. Maturation of the immune response in germinal centers. *Cell* 67: 1121–9

[115] Zotos, D. and D. M. Tarlinton. 2012. Determining germinal centre B cell fate. *Trends Immunol.* 33: 281–288

[116] Shlomchik, M. J. and F. Weisel. 2012. Germinal center selection and the development of memory B and plasma cells. *Immunol. Rev.* 247: 52–63.

[117] Weisel, F. J., G. V. Zuccarino-Catania, M. Chikina, and M. J. Shlomchik. 2016. A temporal switch in the Germinal Center determines differential output of memory B and plasma cells. *Immunity* 44: 116–130

[118] Lefranc, M., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, and V. L. G. Thouvenin-Contet. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27: 55–77.

[119] Lefranc, M. P., C. Pommie, Q. Kaas, E. Duprat, N. Bosc, D. Guiraudou, C. Jean, M. Ruiz, I. Da Piedade, M. Rouard, E. Foulquier, V. Thouvenin, and G. Lefranc. 2005. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.* 29: 185–203.

[120] Kleinstein, S. H. and J. P. Singh. 2003. Why are there so few key mutant clones? The influence of stochastic selection and blocking on affinity maturation in the germinal center. *Int. Immunol.* 15: 871–884

[121] Hershberg, U., M. Uduman, M. J. Shlomchik, and S. H. Kleinstein. 2008. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.* 20: 683–694

[122] Radmacher, M. D., G. Kelsoe, and T. B. Kepler. 1998. Predicted and inferred waiting times for key mutations in the germinal centre reaction: evidence for stochasticity in selection. *Immunol. Cell Biol.* 76: 373–381.

[123] Gitlin, A. D., Z. Shulman, and M. C. Nussenzweig. 2014. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature* 509: 637–640.

[124] Liu, Y. J., J. Zhang, P. J. Lane, E. Y. Chan, and I. C. MacLennan. 1991. Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur. J. Immunol.* 21: 2951–2962

[125] Hollowood, K. and J. Macartney. 1992. Cell kinetics of the germinal center reaction–a stathmokinetic study. *Eur. J. Immunol.* 22: 261–266

[126] Wittenbrink, N., A. Klein, A. A. Weiser, J. Schuchhardt, and M. Or-Guil. 2011. Is There a Typical Germinal Center? A Large-Scale Immunohistological Study on the Cellular Composition of Germinal Centers during the Hapten-Carrier-Driven Primary Immune Response in Mice. *J. Immunol.* 187: 6185–6196

[127] Luciani, F., M. T. Sanders, S. Oveissi, K. C. Pang, and W. Chen. 2013. Increasing viral dose causes a reversal in CD8+ T cell immunodominance during primary influenza infection due to differences in antigen presentation, T cell avidity, and precursor numbers. *J. Immunol.* 190: 36–47

[128] Fan, S., Q. Geissmann, E. Lakatos, S. Lukauskas, A. Ale, A. C. Babtie, P. D. W. Kirk, and M. P. H. Stumpf. 2016. MEANS: python package for Moment Expansion Approximation, iNference and Simulation. *Bioinformatics* : ahead of print

[129] Beyer, T., M. Meyer-Hermann, and G. Soff. 2002. A possible role of chemotaxis in germinal center formation. *Int. Immunol.* 14: 1369–1381

[130] MacLennan, I. C. 1994. Germinal centers. *Annu. Rev. Immunol.* 12: 117–39

[131] Shannon, M. and R. Mehr. 1999. Reconciling repertoire shift with affinity maturation: the role of deleterious mutations. *J. Immunol.* 162: 3950–3956

[132] Barak, M., N. S. Zuckerman, H. Edelman, R. Unger, and R. Mehr. 2008. IgTree: Creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods* 338: 67–74

[133] Küppers, R., M. Zhao, M. L. Hansmann, and K. Rajewsky. 1993. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *EMBO J.* 12: 4955–4967

[134] Victora, G. D., T. A. Schwickert, D. R. Fooksman, A. O. Kamphorst, M. Meyer-Hermann, M. L. Dustin, and M. C. Nussenzweig. 2010. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* 143: 592–605.

[135] Wedemayer, G. J., P. a. Patten, L. H. Wang, P. G. Schultz, and R. C. Stevens. 1997. Structural insights into the evolution of an antibody combining site. *Science* 276: 1665–1669

[136] Tas, J. M. J., L. Mesin, G. Pasqual, S. Targ, J. T. Jacobsen, Y. M. Mano, C. S. Chen, J.-C. Weill, C.-A. Reynaud, E. P. Browne, M. Meyer-Hermann, and G. D. Victora. 2016. Visualizing antibody affinity maturation in germinal centers. *Science* 351: 1048–1054.

[137] Klarenbeek, P. L., M. J. H. de Hair, M. E. Doorenspleet, B. D. C. van Schaik, R. E. E. Esveldt, M. G. H. van de Sande, T. Cantaert, D. M. Gerlag, D. Baeten, A. H. C. van Kampen, F. Baas, P. P. Tak, and N. de Vries. 2012. Inflamed target tissue provides a specific niche for highly expanded T cell clones in early human autoimmune disease. *Ann. Rheum. Dis.* 71: 1088–1093.

[138] Hearty, S., P. Leonard, and R. O'Kennedy. 2012. Measuring antibody-antigen binding kinetics using surface plasmon resonance. *Methods Mol. Biol.* 907: 411–442.

[139] Jacob, J., J. Przylepa, C. Miller, and G. Kelsoe. 1993. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. III. The kinetics of V region mutation and selection in germinal center B cells. *J. Exp. Med.* 178: 1293–1307

[140] Hershberg, U. and E. T. Luning Prak. 2015. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. B Biol. Sci.* 370: 20140239

[141] Dunn-Walters, D. K., A. Belelovsky, H. Edelman, M. Banerjee, and R. Mehr. 2002. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Dev. Immunol.* 9: 233–243

[142] Lees, W. D. and A. J. Shepherd. 2015. Utilities for High-Throughput Analysis of B-Cell Clonal Lineages. *J. Immunol. Res.* 2015: 323506

[143] Kepler, T. B. 2013. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Research* 2: 103

[144] Kepler, T. B., S. Munshaw, K. Wiehe, R. Zhang, J. S. Yu, C. W. Woods, T. N. Denny, G. D. Tomaras, S. M. Alam, M. A. Moody, G. Kelsoe, H. X. Liao, and B. F. Haynes. 2014. Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front. Immunol.* 5: 170

[145] Kocks, C. and K. Rajewsky. 1988. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. *Proc. Natl. Acad. Sci. U. S. A.* 85: 8206–10

[146] Manser, T., L. J. Wysocki, M. N. Margolies, and M. L. Gefter. 1987. Evolution of antibody variable region structure during the immune response. *Immunol. Rev.* 96: 141–162

[147] Vora, K. a., K. Tumas-Brundage, and T. Manser. 1999. Contrasting the in situ behavior of a memory B cell clone during primary and secondary immune responses. *J. Immunol.* 163: 4315–4327

[148] Dunn-Walters, D. K., L. Boursier, P. J. Ciclitira, and J. Spencer. 1997. Immunoglobulin genes from human duodenal and colonic plasma cells are mutated. *Biochem Soc Trans* 25: 324S

[149] Sok, D., U. Laserson, J. Laserson, Y. Liu, F. Vigneault, J. P. Julien, B. Briney, A. Ramos, K. F. Saye, K. Le, A. Mahan, S. Wang, M. Kardar, G. Yaari, L. M. Walker, B. B. Simen, E. P. St. John, P. Y. Chan-Hui, K. Swiderek, S. H. Kleinstein, G. Alter, M. S. Seaman, A. K. Chakraborty, D. Koller, I. A. Wilson, G. M. Church, D. R. Burton, and P. Poignard. 2013. The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies. *PLoS Pathog.* 9: e1003754

[150] Hoehn, K. B., A. Fowler, G. Lunter, and O. G. Pybus. 2016. The Diversity and Molecular Evolution of B-Cell Receptors during Infection. *Mol. Biol. Evol.* 33: 1147–1157

[151] Uduman, M., M. J. Shlomchik, F. Vigneault, G. M. Church, and S. H. Kleinstein. 2014. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J. Immunol.* 192: 867–74

[152] Stern, J., G. Yaari, J. Vander Heiden, G. Church, W. Donahue, R. Hintzen, A. Huttner, J. Laman, R. Nagra, A. Nylander, D. Pitt, S. Ramanan, B. Siddiqui, F. Vigneault, S. Kleinstein, D. Hafler, and K. O'Connor. 2014. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* 6: 248ra107–248ra107

[153] Tabibian-Keissar, H., N. S. Zuckerman, M. Barak, D. K. Dunn-Walters, A. Steiman-Shimony, Y. Chowers, E. Ofek, K. Rosenblatt, G. Schiby, R. Mehr, and I. Barshack. 2008. B-cell clonal diversification and gut-lymph node trafficking in ulcerative colitis revealed using lineage tree analysis. *Eur. J. Immunol.* 38: 2600–2609

[154] Banerjee, M., R. Mehr, A. Belelovsky, J. Spencer, and D. K. Dunn-Walters. 2002. Age- and tissue-specific differences in human germinal center B cell selection revealed by analysis of IgVH gene hypermutation and lineage trees. *Eur. J. Immunol.* 32: 1947–1957

[155] Steiman-Shimony, A., H. Edelman, M. Barak, G. Shahaf, D. Dunn-Walters, D. I. Stott, R. S. Abraham, and R. Mehr. 2006. Immunoglobulin variable-region gene mutational lineage tree analysis: Application to autoimmune diseases. *Autoimmun. Rev.* 5: 242–251

[156] Steiman-Shimony, A., H. Edelman, A. Hutzler, M. Barak, N. S. Zuckerman, G. Shahaf, D. Dunn-Walters, D. I. Stott, R. S. Abraham, and R. Mehr. 2006. Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: Chronic activation, normal selection. *Cell. Immunol.* 244: 130–136

[157] Rosner, B. 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* 25: 165–172

[158] Niklas, N., J. Pröll, J. Weinberger, A. Zopf, K. Wiesinger, K. Krismer, P. Bettelheim, and C. Gabriel. 2014. Qualifying high-throughput immune repertoire sequencing. *Cell. Immunol.* 288: 31–38

[159] Stelling, J. 2004. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.* 7: 513–518

[160] de Jong, S., M. P. M. Boks, T. F. Fuller, E. Strengman, E. Janson, C. G. F. de Kovel, A. P. S. Ori, N. Vi, F. Mulder, J. D. Blom, B. Glenthøj, C. D. Schubart, W. Cahn, R. S. Kahn, S. Horvath, and R. A. Ophoff. 2012. A gene co-expression network in whole blood of Schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One* 7: e39498

[161] von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417: 399–403

[162] Marbach, D., R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. 2010. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U. S. A.* 107: 6286–6291

[163] Cantone, I., L. Marucci, F. Iorio, M. A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. P. Cosma. 2009. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell* 137: 172–181

[164] Kang, T., J. T. White, Z. Xie, Y. Benenson, E. Sontag, and L. Bleris. 2013. Reverse engineering validation using a benchmark synthetic gene circuit in human cells. *ACS Synth. Biol.* 2: 255–262

[165] Chaouiya, C. 2007. Petri net modelling of biological networks. *Brief. Bioinform.* 8: 210–219

[166] Sackmann, A., D. Formanowicz, P. Formanowicz, I. Koch, and J. Blazewicz. 2007. An analysis of the Petri net based model of the human body iron homeostasis process. *Comput. Biol. Chem.* 31: 1–10

[167] Stein, L. D. and J. Thierry-Mieg. 1998. Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. *Genome Res.* 8: 1308–1315

[168] Kochut, K. J., J. Arnold, J. A. Miller, and W. D. Potter. 1993. Design of an object-oriented database for reverse genetics. *Proc. 1st Int. Conf. Intell. Syst. Mol. Biol.*. volume 1. 234–242

[169] O'Neil, E. and P. O'Neil. 1999. *Database principles, programming, performance*. Morgan Kaufmann Publishers

[170] Rice, M., W. Gladstone, and M. Weir. 2004. Relational databases: a transparent framework for encouraging biology students to think informatically. *Cell Biol. Educ.* 3: 241–52

[171] Zou, D., L. Ma, J. Yu, and Z. Zhang. 2015. Biological databases for human research. *Genomics, Proteomics Bioinforma.* 13: 55–63

[172] Klyne, G. and J. J. Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. *W3C Recomm.* 10: 1–20

[173] Wolstencroft, K., S. Owen, O. Krebs, Q. Nguyen, N. J. Stanford, M. Golebiewski, A. Weidemann, M. Bittkowski, L. An, D. Shockley, J. L. Snoep, W. Mueller, and C. Goble. 2015. SEEK: a systems biology data and model management platform. *BMC Syst. Biol.* 9: 33

[174] Antezana, E., W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, and M. Kuiper. 2009. BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics* 10 Suppl 1: S11

[175] Venkatesan, A., S. Tripathi, A. Sanz de Galdeano, W. Blondé, A. Lægreid, V. Mironov, and M. Kuiper. 2014. Finding gene regulatory network candidates using the gene expression knowledge base. *BMC Bioinformatics* 15: 386

[176] Willemsen, A. M., G. A. Jansen, J. C. Komen, S. van Hooff, H. R. Waterham, P. M. T. Brites, R. J. A. Wanders, and A. H. C. van Kampen. 2008. Organization and integration of biomedical knowledge with concept maps for key peroxisomal pathways. *Bioinformatics* 24: i21–27

[177] Khabsa, M. and C. L. Giles. 2014. The number of scholarly documents on the public web. *PLoS One* 9: e93949

[178] Ongenaert, M., L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge. 2008. PubMeth: A cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.* 36: D842–846

[179] Rodriguez-Esteban, R. 2009. Biomedical text mining and its applications. *PLoS Comput. Biol.* 5: e1000597

[180] Zhu, F., P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen. 2013. Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* 46: 200–211

[181] Gonzalez, G. H., T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene. 2016. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief. Bioinform.* 17: 33–42

[182] Thomas, P., J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser. 2012. GeneView: A comprehensive semantic search engine for PubMed. *Nucleic Acids Res.* 40: W585–591

[183] Bowes, J. B., K. A. Snyder, C. James-Zorn, V. G. Ponferrada, C. J. Jarabek, K. A. Burns, B. Bhattacharyya, A. M. Zorn, and P. D. Vize. 2013. The Xenbase literature curation process. *Database* 2013: bas046

[184] Warrer, P., E. H. Hansen, L. Juhl-Jensen, and L. Aagaard. 2012. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br. J. Clin. Pharmacol.* 73: 674–684

[185] Feldman, R., O. Netzer, A. Peretz, and B. Rosenfeld. 2015. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 8: 1779–1788

[186] Ernst, P., A. Siu, and G. Weikum. 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* 16: 157

[187] Torii, M., S. Tilak, S. Doan, D. Zisool, and J. Fan. 2016. Mining Health-Related Issues in Consumer Product Reviews by Using Scalable Text Analytics. *Biomed. Inform. Insights.* 8: 1–11