Scientific paper

Towards Identification of Gene Interaction Networks of Human Cholesterol Biosynthesis

Peter Juvan,[1,*] Tadeja Režen,[1] Damjana Rozman,[1] Katalin Monostory,[3] Jean-Marc Pascussi[4] and Aleš Belič[2]

[1] *Center for Functional Genomics and Bio-chips, Faculty of Medicine, University of Ljubljana, Zaloška cesta 4, SI-1000 Ljubljana, Slovenia*

[2] *Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia*

[3] *Chemical Research Center, Hungarian Academy of Sciences, Pusztaszeri 59–67, H-1025 Budapest, Hungary*

[4] *INSERM UMR-U632, 1919 route de Mende, F-34293 Montpellier, France; University Montpellier 1, F34000, Montpellier, France*

* *Corresponding author: E-mail: peter.juvan@fri.uni-lj.si*

Received: 30-01-2008

Abstract

It has long been demonstrated that the level of cholesterol in cells regulates the cholesterol biosynthesis through SREBF transcription factors, but lately it has been shown that other factors are also important. To study the system we employed Bayesian network inference and combined it with mathematical modeling and simulation. We constructed a mathematical model of cholesterol biosynthesis and studied its properties through simulation. We measured transcriptional changes of cholesterogenic genes using the Steroltalk microarray and treated human hepatocyte samples. We employed Bayesian inference to identify gene-to-gene interactions from both microarray measurements and simulated data. The inferred networks show that the expression of cholesterogenic genes can be predicted from the expression of 4 key genes, one of them being *SREBF2*. Networks also indicate a strong interaction between *SREBF2* and *CYP51A1*, but not between *SREBF2* and *HMGCR*, the rate-limiting enzyme of cholesterol biosynthesis. The expression of *HMGCR* seems to be regulated by other factor(s). Computer simulations of the mathematical model of cholesterol biosynthesis exposed that a large number of perturbations of the system is critical for identification of gene-to-gene interactions, and that differences between human individuals (biological variability) and measurement noise (technical variability) pose a serious problem for their automatic inference from DNA microarray data.

Keywords: Functional genomics, systems biology, gene interaction network, Bayesian inference, mathematical modeling and simulation, human cholesterol biosynthesis

1. Introduction

Cholesterol biosynthesis is an anabolic pathway in which a cholesterol molecule is built from acetyl-CoA through more then 20 reactions (Figure 1). It takes place in almost all cell types and involves numerous enzymes from different protein families. It is composed of two consecutive phases, the isoprenoid biosynthesis forming squalene, and the post-squalene phase resulting in cholesterol. It has long been demonstrated that the level of cholesterol in cells regulates the cholesterol biosynthesis by the negative feedback loop involving SREBF (sterol regulatory element binding factor) in signaling pathway.[1,2] SREBFs are membrane bound transcription factors, which are cleaved when cholesterol is limited. The DNA-binding portion of the SREBF protein is then transported to the nucleus, where it binds to sterol regulatory DNA elements in promoters of responsive genes and activates their transcription. The level of cholesterol in the cell determines the fate of SREBF: more cholesterol blocks the cleavage, less cholesterol results in SREBF cleavage and up-regulation of genes involved in cholesterol biosynthesis. However, SREBFs may not be the only transcription factors involved in control of biosynthesis. For example, the cAMP signaling pathway has been documented to control cholesterol biosynthesis in certain physiological conditions.[3,4] Additionally, TNFα might also have an effect on the cholesterol homeostasis.[5]

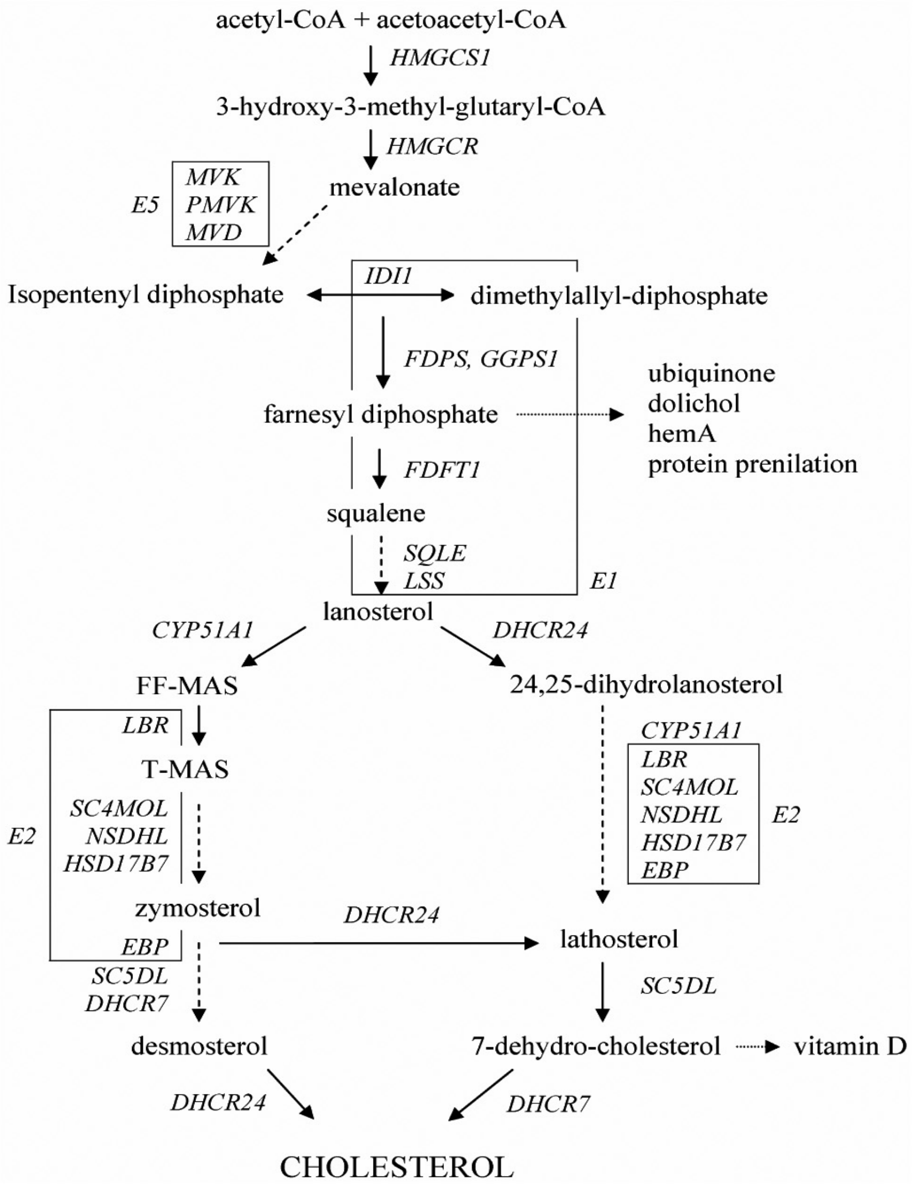

Figure 1. Schematic view of cholesterol biosynthesis in human (adapted from BioCyc Database Collection, http://biocyc.org). Gene names are listed in Table 1 in Supplement. FF-MAS: follicular fluid-meiosis activating sterol, T-MAS: testis meiosis-activating sterol.

The aim of this study was to combine tools from functional genomics and systems biology for identification of control mechanisms involved in expression of genes encoding enzymes of the cholesterol biosynthesis. The expression of cholesterogenic genes has been measured in primary hepatocytes of seven human individuals using the Steroltalk v2 microarray.[6,7] The hepatocytes of each individual were treated with three xenobiotic substances. Automatic network inference approach was employed for discovery of interaction between genes. In parallel, a mathematical model of cholesterol biosynthesis was constructed and used for simulation of the expression of cholesterogenic genes under various experimental conditions. Gene interaction network identified from the measured data was compared to the networks identified from the simulated data and effects of various experimental conditions were studied *in silico*. The structure of the mathematical model was adjusted according to data from microarray measurements.

2. Methods

2. 1. Microarray Experiments and Data Processing

Human hepatocytes isolated from 7 individuals were treated with rifampicin, a typical activator of the pregnane X receptor (PXR), rosuvastatin, a well known cholesterol lowering drug, and LK935, a novel cholesterol lowering drug-candidate (Lek Pharmaceuticals d. d., Ljubljana, Slovenia)[8] for 12, 24 and 48 hours (experimental details are to be published elsewhere). All cell cultures were grown for the same period of time in parallel with untreated cultures (control) and harvested simultaneously. RNA was isolated and hybridized to Steroltalk v2 microarray[6,7] using a common reference design. Arrays were scanned using a Tecan LS200 scanner (Tecan Group Ltd., Maennedorf, Switzerland) and images were analyzed using Array-Pro Analyzer v.4.5 software (Media Cybernetics, Bethesda, MD, USA).

Gene expression data was normalized using Orange[9] normalization widget as follows. First, spots with the intensity below 1.5 of the intensity of a local background or below 2 standard deviations of a local background were removed. Next, data from Lucidea Universal ScoreCard (Amersham Biosciences, GE Healthcare UK limited, Little Chalfont, UK) ratio spike-in controls were adjusted according to their expected ratios between Cy3 and Cy5 dyes. These controls, together with Lucidea Universal ScoreCard calibration controls, were used to fit LOWESS normalization curve.[10] The expression values of non-filtered genes were adjusted according to that normalization curve. Data from within-array replicated probes (3 per gene) were merged using median, thus removing potential outliers.

For each treatment a maximum treatment effect was assessed by comparing expression values from untreated samples to those from samples treated for 12, 24 and 48 hours. For Bayesian network inference, only the expression values from untreated samples and those representing maximum treatment effect were considered, thus removing potential statistical dependencies between consecutive expression measurements. Changes in gene expression as response to different compound administration were compared to those contributed to differences between individuals. As Spearman correlation coefficients between expression profiles of different individuals (0.408 ± 0.144, n = 21) was significantly lower (two-tailed t-test, p = 0.0013) than the correlation between various treatments (0.675 ± 0.183, n = 6), the differences between individuals were assumed large enough to represent divergent and independent instantiations of the observed biological system. Thus the expression profiles were considered individually. Missing values in data were estimated from the expression measurements of other genes using k-nearest neighbors algorithm[11] with *k* set to 10 and using a tricubic weighting scheme. The expression profiles of individual genes were discretized using three intervals of equal length, thus reducing noise in the data.

2. 2. Bayesian Network Inference

Using Banjo,[12] we employed steady-state Bayesian network inference[13] of interactions between genes involved in cholesterol biosynthesis that are present on the Steroltalk v2 microarray: *CYP51A1, DHCR24, DHCR7, EBP, FDFT1, FDPS, HMGCR, HMGCS1, HSD17B7, IDI1, LSS, MVD, MVK, NSDHL, PMVK, SC4MOL, SC5DL* and *SQLE* (for gene names see Table 1 in Supplement). Additionally, we considered expression of gene *SREBF2* which is most actively involved in the regulation of the above-listed genes. We used simulated annealing to search for the most probable network and evaluated the networks using Bayesian-Dirichlet scoring metrics.[14] Considering data from all of the above-listed genes, the straightforward inference approach resulted in many structurally-different networks with equal scores, thus making it impossible to select the most probable one. Dealing with too many variables ($n = 19$) for the given number of measurements ($m = 4 * 7 = 28$ considering 4 treatments, including control, and 7 individuals) this was an expected result.

We therefore employed model averaging approach[15] to estimate the significance of interactions between genes and constructed BNs considering only the most relevant interactions. The approach exploits the fact that the most probable BN can be found relatively quickly (within few seconds) if the number of variables (genes) is small enough. From the above-listed genes, we generated all possible subsets of 5 genes, thus considered each pair of genes within different data environments composed of all

Table 1. Experimental results (real) in compare with simulated data (sim) for several treatments. For simulated data, peak or trough value of the corresponding metabolite data is presented. Treatments were simulated by reducing the levels of targeted enzymes. Compounds 5d, 5j, 4c and 12a are described by Korošec et al., 2007.[8]

relative-to-normal values	cholesterol		desmosterol		7-dehydro cholesterol + zymosterol		lathosterol		FF-MAS		lanosterol	
treatment	**real**	**sim**	**real**	**sim**	**real**	**sim**	**real**	**sim**	**real**	**sim**	**real**	**sim**
atorvastatin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
rosuvastatin	0.3	0.3	0.6	0.4	0.5	0.4	0.4	0.3	0.4	0.3	0.0	0.1
comp. 5d 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.2	6.2	
comp. 5j 0.1	0.1	0.0	0.1	1.0	1.1	0.5	0.5	2.5	2.4	1.0	1.1	
comp. 4c 0.2	0.0	0.1	0.0	2.0	1.7	0.9	0.7	3.7	3.6	1.3	1.5	
comp. 12a	0.0	0.1	0.0	0.1	1.0	1.1	0.1	0.1	6.5	5.9	1.0	1.1
high-fat diet	–	3.0	–	0.1	–	0.1	–	0.1	–	0.1	–	1.0

possible combinations of three other genes. Note that the number of such subsets equals *n!/r!(n-r)!* with *n* and *r* corresponding to the number of genes and a subset size, respectively (considering 19 genes there are 11628 subsets of 5 genes). For each subset we searched for the most probable Bayesian subnetwork, which resulted in a single best-scoring solution. Considering all inferred subnetworks we analyzed the frequencies of discovered gene-to-gene interactions and their influence scores (IS), which, where assigned, indicate whether the influence is either positive or negative and its relative magnitude ranging from 0 (lowest) and 1 (highest). Cause-effect relationship between variables (genes) often cannot be determined unambiguously[16] resulting in interactions contradicting in the direction of the influence, but not in its sign; we therefore constructed undirected networks and computed relative frequencies of interactions irrespectively of directions of influences (EF: edge frequency) and corresponding frequencies of IS assignments, mean IS scores and their deviations from the mean. We selected only the most frequent interactions (EF filtering criterion) and presented them on edges of an undirected network.

We visually inspected the networks and compared them to each other on the basis of individual interactions and their structural complexity. We estimated structural complexity of the networks from the number of key genes, i.e. genes representing a minimal set of genes whose expression needs to be known in order to predict the expression of other genes in a network. To predict expression of a gene from the expression of all genes that are connected to that gene must be known. Although key genes often cannot be identified uniquely, they can be counted; their smallest number indicates the complexity of the considered expression data, thus implying on the number of the transcription factors involved in regulation of genes being considered. From structural complexity of the networks we reasoned about the complexity of the modeled system.

2. 3. Mathematical Model of Cholesterol Biosynthesis

A model of cholesterol biosynthesis was constructed using the data from literature[17,18] and structural information of the involved substances (KEGG, Biosynthesis of steroids – Reference pathway, http://www.genome.jp/kegg/pathway/map/map00100.html) with the purpose to study the interplay between metabolites, proteins, and genes that are involved in cholesterol biosynthesis as well as the effect of drugs on cholesterol biosynthesis. The model is based on non-linear differential equations that represent the dynamics of enzyme reactions, as well as biochemical and gene expression feedback mechanisms. Regulation of enzyme levels in the model is achieved in two ways. On the biochemical level, enzymes are regulated with reduction of active enzyme levels when cholesterol is raised above its physiological level.[19] Gene expression in the model is regulated by SREBF2 protein. The levels of active or cleaved SREBF2 are regulated by the levels of cholesterol and other sterols, such as desmosterol, and oxysterols, such as 25-hydroxycholesterol, 27-hydroxycholesterol, and 7-hydroxycholesterol.[20,21] The level of active SREBF2 in turn regulates the expression of cholesterogenic genes.

Pre-lanosterol pathway is generally a single-track pathway with one significant branch leading towards coenzyme Q. Thus the pathway can be reduced to only a few key metabolites and the rest can be grouped together as single metabolites without significantly affecting dynamical characteristics of the model. Post-lanosterol pathway, however, consists of many parallel tracks. In physiological conditions, only the most important track is active, therefore post-lanosterol pathway can also be simplified similarly to the pre-lanosterol one. Grouping of metabolites also implies grouping of proteins that catalyze reactions between them as well as of genes that encode these proteins. The model structure operates with the following protein and gene groups: *E1* (*IDI1, FDPS, GGPS1, FDFT1, SQLE* and *LSS*), *E2* (*LBR, SC4MOL, NSDHL, HSD17B7* and *EBP*), and *E5* (*MVK, PMVK* and *MVD*). For a single metabolic pathway with no branches, all genes of the pathway should be commonly regulated in order to prevent high accumulation of intermediates. Grouping of commonly regulated genes into single entities thus represents a simplification of the model that does not affect model dynamics significantly.

The model as composed in Dymola[22] is presented in Figure 2. It was validated against experimental data presented by Korošec et al., 2007[8] and adjusted accordingly, introducing negative biochemical feedback of lanosterol on HMGCR (described also by Song et al., 2005[23]), and negative biochemical feedback of 7-dehydrocholesterol on CYP51A1 activity (not confirmed by the literature). Validation results are shown in Table 1 together with high-fat diet simulation, which showed shut-down of cholesterol biosynthesis in the model.

3. Computer Simulation of Gene Expression Data

Computer simulation of the mathematical model was used to predict expression of cholesterogenic genes mimicking the experimental conditions of the conducted microarray measurements. Different experimental factors were considered and their effect on interactions between genes was studied. The following factors were considered: treatments with various compounds, measurement noise and differences between individuals. In simulation experiments, mRNA levels of genes/gene groups were sampled at 12 h, 24 h, and 48 h after xenobiotic administration in accordance with design of microarray experi-

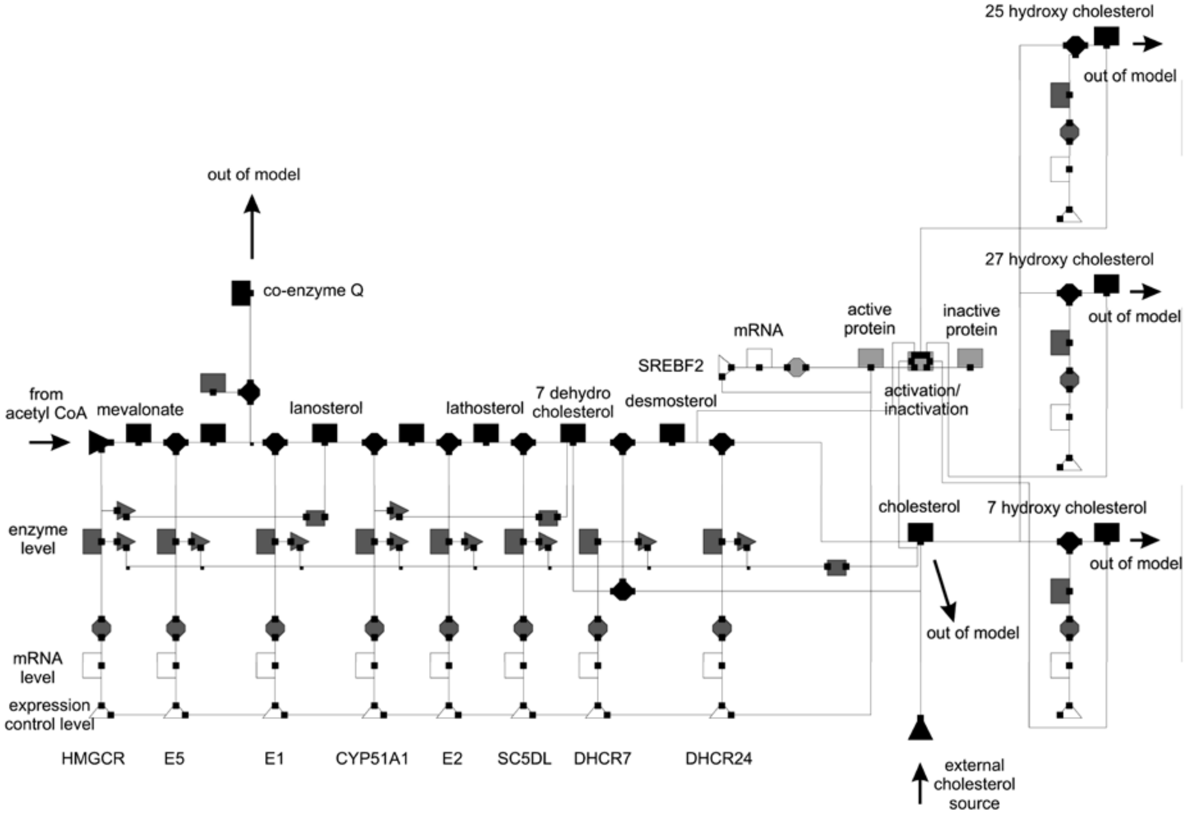

Figure 2. A simplified mathematical model of cholesterol biosynthesis. Black blocks represent metabolites (squares), enzyme reactions (circles) and metabolite sources (triangles). Dark grey blocks represent enzymes (large squares), their biochemical control (triangles and small squares), and enzyme production regarding gene expression (circles). White blocks represent gene expressions (squares) and their control (triangles). Light grey blocks represent transcription factors (squares), their production regarding gene expression (circles), and their activation/inactivation (squares). Gene names are listed in Table 1 in Supplement.

ments and the value with maximum deviation from the initial value (representing the starting time point) was selected to represent the simulated response of each gene.

Treatments were simulated by blocking or inducing the expression of proteins and transcription factors as a response to administration of different compounds. First, effects of the three compounds administered *ex vivo* were simulated: rifampicin by activating PXR[24] (in the model simulated as reduction of bile acid synthesis through inactivation of SREBF2 transcription factor), rosuvastatin by blocking HMGCR[25], and LK935 by inhibiting CYP51A1.[8] Next, hypothetic compounds were simulated by blocking E1, E2, E5, SC5DL, DHCR7, DHCR24 and active form of SREBF2, respectively. Thus simulated data from up to 10 different perturbations of the system was obtained. Measurement noise was simulated by adding Gaussian noise to the mRNA levels of genes with zero mean and relative standard deviation (RSD) up to 5% using pseudorandom number generator. The level of noise was selected in accordance to the technical variability in the performed microarray measurements. Simulations were repeated for 7, 20 and 100 times, thus producing data similar to performing 7, 20 and 100 technical replications of gene expression measurements.

Differences between individuals were simulated by perturbing parameters of the model describing responsiveness of genes to mature nuclear SREBF2 protein level changes ($mRNA_{max}$) and the rate of drug metabolism (k_e). Originally, all model parameters were set to arbitrary values that resulted in stable model responses, with time constants providing peak response at approximately 24 h. For each gene, pseudorandom number generator was used to vary values of $mRNA_{max}$ and k_e using normal distribution with mean equal to their original values and 1% RSD for $mRNA_{max}$ and up to 50% RSD for k_e. Other parameters of the model were kept constant since they showed limited or no effect on gene expression. Simulations were repeated for 7, 20 and 100 times, thus producing data similar to as performing 7, 20 and 100 biological replications of gene expression measurements. All simulations were repeated under identical conditions, but using a different seed to initialize a pseudorandom number generator. The reproducibility of the identified gene-to-gene interactions was judged by visual inspection of the inferred networks. The selected noise levels and inter-individual variability resulted in simulated data for which similar Spearman correlation coefficients between different individuals and treatments were observed as for the measured data.

4. Results

4. 1. Gene-to-Gene Interactions From Measured Data

Figure 3 shows a BN of cholesterol biosynthesis constructed from the measured data using the model averaging approach. Only the most frequent interactions, i.e., interactions that appear within almost all subnetworks of 4 genes (EF = 0.99) are shown. Increasing/decreasing the subset size by one, only small differences in frequencies of less frequent interactions were observed, but no changes in the most frequent interactions (example not shown).

To enable comparison of the network inferred from the measured data to networks inferred from the simulated data we formed groups of genes in the same way as for the mathematical model. We used median of measured expression values of genes within each group to represent expression of individual groups. From that data we constructed a BN that is shown in Figure 4 (reduced model). Comparison of the reduced and the complete BN reveals the following correspondences between them: *HMGCR-SC5DL* and *SREBF2-CYP51A1* appear at 100% in both models; *SREBF2-E5* from the reduced model is represented by *SREBF2-PMVK* and *SREBF2-MVD* in the complete model; *SC5DL-E2* from the reduced model is represented by *SC5DL-SC4MOL* in the complete model; *HMGCR-E1* from the reduced model is represented by *HMGCR-SQLE* in the complete model; and *DHCR7-E1* from the reduced model is represented by *DHCR7-FDPS* and *DHCR7-LSS* in the complete model.

We examined the frequencies of interactions appearing only in the reduced model (see Table 2 for frequencies of interactions from the complete model which are not shown in Figure 3): *SREBF2-DHCR24* from the reduced model appears at 88% in the complete model; both *DHCR7-E2* and *SREBF2-E2* from the reduced model have their counterparts in the complete model, of which most frequent are *DHCR7-NSDHL* (90%) and *SREBF2-SC4MOL* (78%), respectively. We also examined the frequencies of interactions appearing only in the complete model (see Table 3 for frequencies of interactions from the reduced model which are not shown in Figure 4): *FDPS-EBP, LSS-EBP* and *IDI1-SC4MOL* from the complete model are represented by *E1-E2* in the reduced model at 57%; *HMGCR-HSD17B7* is represented by *HMGCR-E2* at 40%; *HSD17B7-MVD* is represented by *E2-E5* at 6%; *MVD-DHCR24* is represented by *E5-DHCR24* at 29%; and *MVK-SC5DL* is represented by *E5-SC5DL* at 40%.

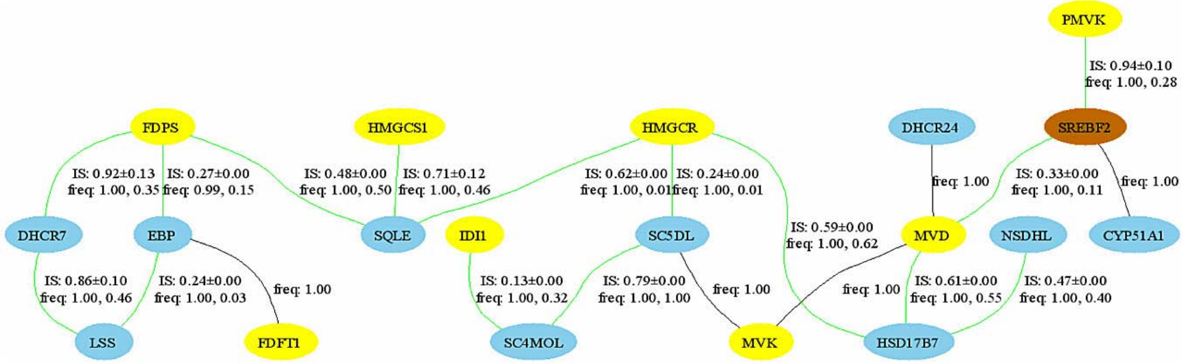

Figure 3. Complete BN of cholesterol biosynthesis inferred from measured data using model averaging approach. Only interactions which appear in all subnetworks of 4 genes are shown (filtering parameter EF = 0.99). Nodes represent genes (yellow for those involved in isoprenoid phase, blue for those involved in post-squalene phase) and edges represent gene-to-gene interactions (green for positive interactions and black for interactions of undeterminable sign). The numbers on edges show mean influence score (IS) and its deviation from the mean, and relative frequencies of interaction appearance (EF) and of IS assignment. For interactions of undeterminable sign only EF is shown. Gene names are listed in Table 1 in Supplement.

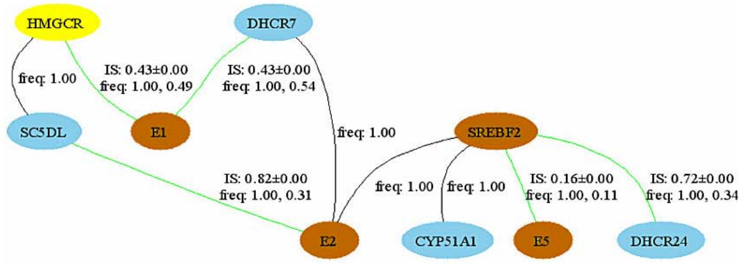

Figure 4. Reduced BN of cholesterol biosynthesis inferred from measured data using model averaging approach with 5 genes and filtering parameter EF = 1.0. Genes were grouped in agreement with the mathematical model. See Figure 3 for explanation of the symbols. Gene names are listed in Table 1 in Supplement.

Table 2. Gene-to-gene interactions from the complete BN of cholesterol biosynthesis inferred from measured data using model averaging approach. Columns represent interacting genes, mean influence score of their interaction (IS) and its deviation from the mean, frequencies of interaction appearance (EF) and IS assignment (IS freq). Only interactions that meet criterion 0.78 ≤ EF < 1.0 are listed.

Gene 1	Gene 2	mean IS ± st.dev.	EF	IS freq
CYP51A1	*FDFT1*	0.589 ± 0.094	0.852	0.426
CYP51A1	*MVD*	0.579 ± 0.269	0.86	0.559
DHCR24	*DHCR7*	0.474 ± 0.158	0.882	0.823
DHCR24	*EBP*	0.522 ± 0.000	0.986	0.978
DHCR24	*SREBF2*	0.721 ± 0.000	0.882	0.397
DHCR7	*SC5DL*	0.664 ± 0.000	0.86	0.566
EBP	*MVD*	0.195 ± 0.170	0.875	0.022
FDFT1	*SQLE*	0.516 ± 0.230	0.89	0.486
FDPS	*DHCR24*	0.362 ± 0.210	0.787	0.456
FDPS	*FDFT1*	0.576 ± 0.524	0.875	0.765
FDPS	*NSDHL*	0.556 ± 0.176	0.904	0.404
HMGCR	*FDFT1*	–	0.963	–
HMGCR	*HMGCS1*	0.111 ± 0.000	0.882	0.007
HMGCR	*IDI1*	0.472 ± 0.463	0.926	0.912
HMGCR	*LSS*	0.438 ± 0.173	0.868	0.559
HMGCR	*NSDHL*	0.739 ± 0.084	0.883	0.412
HMGCS1	*HSD17B7*	0.718 ± 0.000	0.897	0.537
HMGCS1	*IDI1*	0.672 ± 0.109	0.971	0.596
LSS	*FDFT1*	0.681 ± 0.283	0.882	0.875
LSS	*MVD*	0.658 ± 0.058	0.824	0.824
NSDHL	*DHCR7*	0.689 ± 0.153	0.904	0.714
NSDHL	*SC4MOL*	0.159 ± 0.057	0.786	0.036
NSDHL	*SC5DL*	0.683 ± 0.123	0.89	0.478
PMVK	*CYP51A1*	0.857 ± 0.000	0.875	0.235
PMVK	*MVD*	0.613 ± 0.127	0.801	0.265
SC4MOL	*HSD17B7*	0.223 ± 0.099	0.86	0.044
SC4MOL	*MVK*	0.775 ± 0.000	0.883	0.471
SC5DL	*CYP51A1*	–	0.875	–
SC5DL	*HSD17B7*	0.383 ± 0.000	0.875	0.404
SREBF2	*MVK*	0.786 ± 0.000	0.883	0.368
SREBF2	*SC4MOL*	0.158 ± 0.000	0.78	0.007
SREBF2	*SC5DL*	–	0.882	–

Considering the fact that gene groups were formed with no consideration of the measured expression levels, we found the simplified mathematical model to be an acceptable approximation of the *in vivo* system and also appropriate for simulation of the conducted microarray experiments. The structural complexity of the reduced network is 4, which is the number of key genes whose expression needs to be known in order to predict the expression of other genes within that network. There are 5 possible combinations of 4 key genes; though they cannot be identified uniquely, *SREBF2* appears in all of the combinations.

Table 3. Gene-to-gene interactions from the reduced BN of cholesterol biosynthesis inferred from measured data using model averaging approach. Columns represent interacting genes, mean influence score of their interaction (IS) and its deviation from the mean, frequencies of interaction appearance (EF) and IS assignment (IS freq). Only interactions that meet criterion 0.058 ≤ EF < 1.0 are listed.

Gene 1	Gene 2	mean IS ± st.dev.	EF	IS freq
CYP51A1	*DHCR24*	–	0.572	–
CYP51A1	*E1*	–	0.114	–
CYP51A1	*E2*	–	0.285	–
CYP51A1	*E5*	0.757 ± 0.000	0.572	0.143
DHCR7	*CYP51A1*	0.638 ± 0.000	0.114	0.114
DHCR7	*DHCR24*	0.485 ± 0.084	0.714	0.714
DHCR7	*SC5DL*	0.664 ± 0.000	0.371	0.114
DHCR7	*SREBF2*	0.111 ± 0.000	0.058	0.029
E1	*E2*	–	0.571	–
E5	*DHCR24*	0.652 ± 0.000	0.285	0.114
E5	*E2*	0.649 ± 0.000	0.058	0.029
HMGCR	*DHCR7*	0.124 ± 0.024	0.228	0.114
HMGCR	*E2*	0.437 ± 0.000	0.4	0.086
SC5DL	*CYP51A1*	0.167 ± 0.000	0.714	0.143
SC5DL	*E1*	0.572 ± 0.000	0.2	0.114
SC5DL	*E5*	–	0.4	–
SREBF2	*SC5DL*	0.329 ± 0.000	0.743	0.171

4. 2. Gene Interactions From Simulated Data

We constructed BNs from expression data generated by computer simulation of the cholesterol biosynthesis model shown in Figure 2. Within this model, the expression of cholesterogenic genes is regulated solely by active SREBF2 protein. Simulations of the following conditions were considered:

a. Ideal: 10 treatments of a single individual through simulation of administration of different (existing and hypothetical) compounds.
b. Noise: 7, 20 and 100 technical replications of gene expression measurements of a single individual treated with rosuvastatin, LK935 and rifampicin by adding Gaussian noise to mRNA levels (5% RSD).
c. Individuals: biological replications of gene expression measurements of 7, 20, and 100 individuals treated with rosuvastatin, LK935 and rifampicin by perturbing the parameters $mRNA_{max}$ (1% RSD) and k_e (50% RSD) of the model.
d. Realistic: biological replications of gene expression measurements of 7, 20, and 100 individuals treated with rosuvastatin, LK935 and rifampicin (similar to condition c, but using 30% RDS for k_e) and accounting for technical variability by adding Gaussian noise to mRNA levels (similar to condition b, but using 3% RDS).

Similarly as for the measured data we employed the model averaging approach and considered only the most frequent interactions (EF = 1.0). According to the structure of the mathematical model where SREBF2 protein is the only transcription factor and induces all genes of the cholesterol biosynthesis,[26] a BN of low structural complexity was expected. Simulating ideal experimental conditions (condition a: 10 treatments with no inter-individual differences or noise) resulted in complex network structures such as shown in Figure 5 with large number of key genes (6 for the network shown in Figure 5). This indicates that though a single transcription factor regulates the expression of all genes, the complexity of the modeled system is high. Tests with fewer treatments (3, simulating compounds administered *ex vivo*) resulted in networks of a similar complexity (though often split into several highly interconnected subnetworks), thus providing additional evidence of the complexity of the modeled system.

Simulations with random noise (condition b) typically resulted in reduced number of connections and network structures of low complexity such as the network shown in Figure 6. For that network, only expression of a single gene (*E1*) is required to predict the expression of other genes. Increasing noise above 5% of mean mRNA level resulted in networks reduced to a set of unconnected genes. The simulations indicate that measurement noise may hinder our ability to discover interactions between genes solely from the expression measurements.

Simulations of random individuals (condition c) typically resulted in structures similar to that shown in Figure 7 where expression of 4 key genes (*E1, E2, E5* and *SC5DL*) is required to predict the expression of other genes. Similarly to simulations with noise, increasing the number of individuals results in a reduced number of connections between genes compared to simulations of ideal conditions. However, the number of key genes is not reduced so drastically as in the case of simulations of noise. In case of exceeding differences between individuals, the network is reduced to a set of unconnected genes similarly to the case of exceeding noise. The simulations indicate that inter-individual differences affect interactions between genes, but the effect is different from that of noise. Yet it remains unclear whether the inter-individual differences may be exploited to substitute for a lack of sufficient number of independent perturbations of the system.

4. 3. Structural Adjustment of the Mathematical Model

BNs inferred from simulation of the mathematical model indicated that adaptation of its structure is required

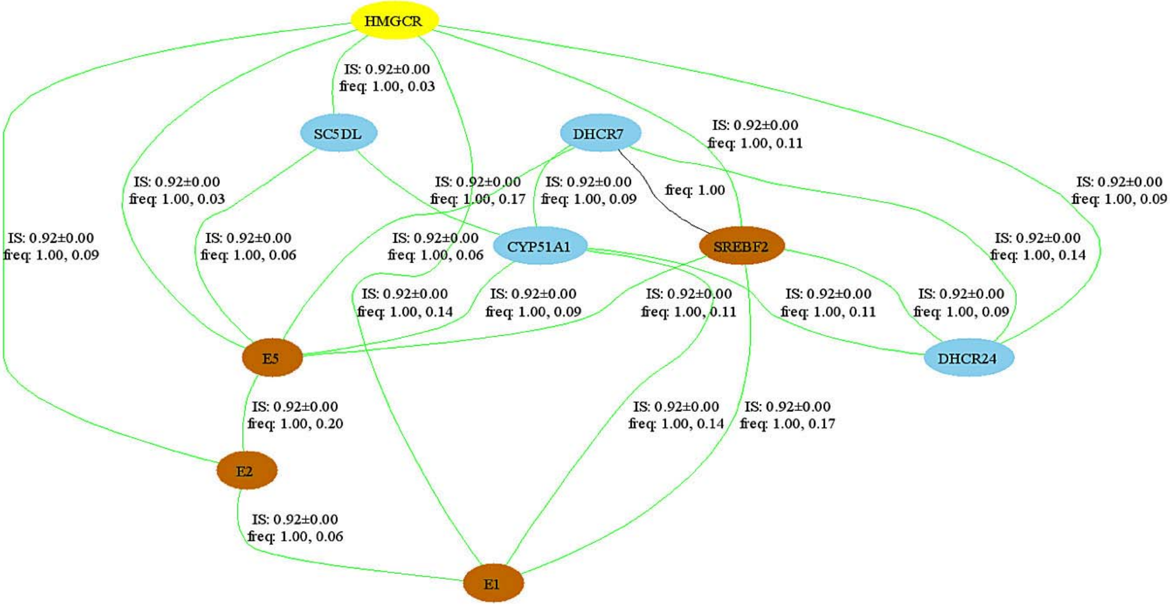

Figure 5. BN of cholesterol biosynthesis inferred from simulation of 10 treatments (condition a) using model averaging approach with 5 genes and filtering parameter EF = 1.0. See Figure 3 for explanation of the symbols. Gene names are listed in Table 1 in Supplement.

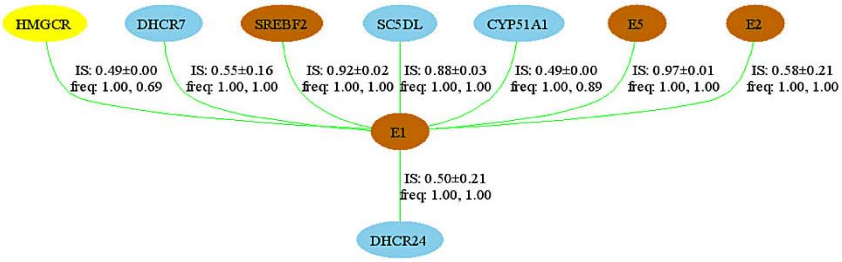

Figure 6. BN of cholesterol biosynthesis inferred from simulation of 3 treatments of a single individual and 100 technical replications (condition b) using model averaging approach with 5 genes and filtering parameter EF = 1.0. See Figure 3 for explanation of the symbols. Gene names are listed in Table 1 in Supplement.

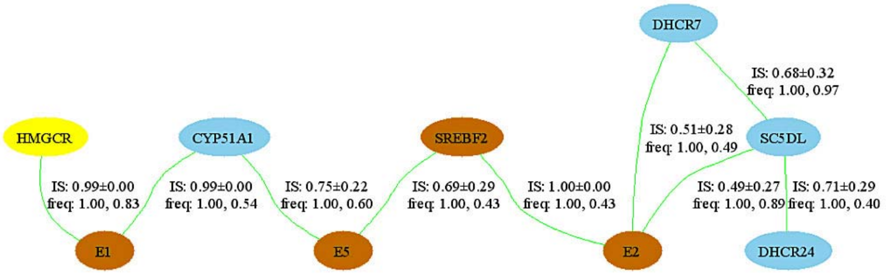

Figure 7. BN of cholesterol biosynthesis inferred from simulation of 3 treatments of 100 individuals (condition c) using model averaging approach with 5 genes and filtering parameter EF = 1.0. See Figure 3 for explanation of the symbols. Gene names are listed in Table 1 in Supplement.

in order to generate BNs similar to the one inferred from the measured data. As SREBF2 may not be the only transcription factor interfering with gene expression of cholesterol biosynthesis, we considered expanding the model with other biological regulatory mechanisms. Correlation coefficients (Spearman r) between the measured expressions levels of the observed genes (data not shown) indicate that *HMGCR, E1* and *E2* may be regulated differently from the remaining genes. We therefore presumed that an additional transcription factor regulates the three above-listed genes/gene groups and that this factor was activated during the conducted biological experiments. Simulation experiments showed that this factor is not directly influenced by cholesterol, but possibly by some of the intermediates or other signaling pathways related to cholesterol biosynthesis. Thus, a hypothetical transcription factor T1 regulating expression of *HMGCR*, *E1* and *E2* was introduced to the model.

Simulations of 100 random individuals with noise (condition d) using structurally adjusted model resulted in a network shown in Figure 8. Comparing the structure of that network to the network inferred from the measured data (Figure 4), both networks have similar connections between genes and the same number of key genes. *SREBF2* is the most inter-connected gene (4 and 5 connections in the measured and the simulated network, respectively). Expression of *CYP51A1* can be predicted solely from the expression of *SREBF2*. *SREBF2, CYP51A1, E2* and *E5* are interconnected in the same way; they are also connected to both *SC5DL* and *DHCR7*, though the pathway is mediated by *E2* in the measured and by *SREBF2* in the simulated network, respectively. In the latter model, an additional connection between *SC5DL* and *E5* appears. Similar is also the interaction between *HMGCR* and *E1*. Among the differences the most noticeable is the connection between *HMGCR* on one hand and *SREBF2*/*CYP51A1* on the other. In the measured network, that pathway is mediated by *E2*, while in the simulated network it is mediated by *E1*.

Structural complexity of both networks is similar: expression of 4 key genes is needed to predict expression of all other genes, and there are 5 possible combinations of 4 key genes with *SREBF2* appearing in all of them. In the measured network, *E2, E1, HMGCR, SC5DL* and *DHCR7* appear in three of the combinations of key genes, while in the simulated network, *E1* in appears in four, *E5* and *DHCR24* in three, *DHCR7* in two, and *HMGCR* and *SC5DL* in one combination of key genes.

5. Discussion

To study the system of cholesterol biosynthesis, we used primary human hepatocytes and treated them with different compounds to perturb the system in different ways. We measured the expression of cholesterogenic genes using the Steroltalk v2 microarray[6,7] and used Bayesian inference approach to infer interactions between

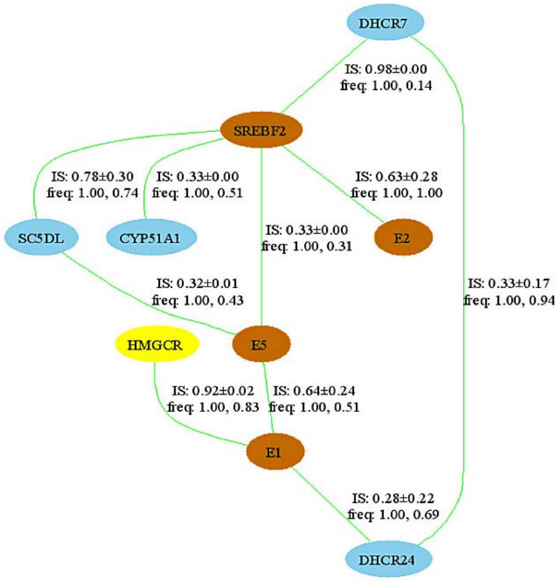

Figure 8. BN of cholesterol biosynthesis inferred from simulation of 3 treatments of 100 individuals (condition d) using the structurally adjusted model. Inference was made using model averaging approach with 5 genes and filtering parameter EF = 1.0. See Figure 3 for explanation of the symbols. Gene names are listed in Table 1 in Supplement.

genes. Experimental data only allowed for steady-state BN inference due to measurements at only few time points at which microarray measurements were performed and slow sampling rate.[27] Low number of the performed perturbations represented an additional obstacle for inference of a reasonable model of cholesterol biosynthesis from the data alone. We therefore took a more knowledgeable approach and explored the possibility of adjusting a literature-based mathematical model of cholesterol biosynthesis according to the gene expression data. We identified gene interaction networks from the measured as well as from the simulated data using Bayesian inference and the model averaging approach. We compared the two networks in order to study effects of different experimental conditions *in silico* and to structurally adjust the mathematical model according to the gene expression measurements. In simulations, the design of the conducted microarray measurements was considered and different parameters of the model were perturbed to mimic the real experimental conditions.

Within the space of all possible gene interaction networks of cholesterol biosynthesis, which is too large to be examined exhaustively, Bayesian inference relies upon a heuristic search,[15] thus resulting in a near-optimal solution consisting of interactions of which some are important, other spurious. The straightforward inference approach lacks a mechanism for estimation of confidence in interactions; we therefore resorted to the model averaging approach and estimated frequencies of their appearance considering different subsets of genes. Decreasing the number of genes considered at a time reduces the search-space of possible structures, which can consequently be examined exhaustively within a reasonable amount of time. The approach is different from the bootstrap method[28] employed by Friedman et al., 2000[15] – while their aim was to assess a reasonable model of the data, our was rather to focus on the most confident interactions in order to compare the networks from the simulated data to the one from the measured data. One might argue that the proposed approach is limited to interactions appearing within a context of a limited number of genes. A similar limitation exists within all practical implementations of BN inference algorithms: due to computational efficiency the inference is limited by the number of interactions a single gene may receive. However, the model averaging approach inherently overcomes this limitation by considering each gene within a context of all possible combinations of other genes, though for that reason the frequencies of interactions may be underestimated.

Simulations with the initial mathematical model resulted in many different networks of which some examples are shown in Figures 5–7. Simulation of a large number of treatments (condition a) resulted in highly-interconnected network (Figure 5) indicating high complexity of the modeled system. Tests with different number of treatments showed that their sufficient number (i.e., the number of independent perturbations of the system) is critical, especially if the underlying system is a complex one, and that even a small increase in their number is beneficial. Either adding noise (condition b) or introducing inter-individual differences (condition c) resulted in a reduced number of connections and reduced network complexity (see Figures 6–7), though the complexity was less affected by inter-individual differences than by noise. Overall, the simulations revealed several facts which can be used for planning future experiments. Measurement noise hinders our ability to discover interactions between genes, therefore technical variability of microarray measurements must be kept as low as possible. Though inter-individual differences affect interactions between genes, their effect is quite different from that of measurement noise. Further simulation studies would need to be performed to reveal whether inter-individual differences may be exploited in order to improve the inference of interactions between genes and potentially substitute for insufficient number of perturbations of a system.

Our inability to achieve an exact match between the simulated and the measured BNs indicates that the structure of the measured BN is influenced by measurement noise and large variability between the individuals involved in the study, which is obstructing our ability to infer the gene-to-gene interactions of the observed biological system. The differences between BNs from the measured and simulated data may also be due to simplification of the mathematical model by forming groups of genes *E1, E2* and *E5*. In case of genes within a group being regulated differently from each other, those with stronger regulation will prevail and consequently the gene group will connect to other genes differently as the genes themselves would, resulting in the observed differences between the networks. Therefore in future studies dealing with expression of individual genes instead of gene groups should be considered, accounting for the fact that increased number of variables will result in substantial increase of computational complexity.

Since active SREBF2 protein regulates expression of cholesterogenic genes as well its own expression,[29] our question was whether *SREBF2* mRNA levels could be a marker for the level of active protein and hence the expression of cholesterogenic genes. BNs of cholesterol biosynthesis constructed from the measured (Figure 4) and simulated data (Figure 8) indicate that knowing solely the expression of *SREBF2* is not sufficient for predicting the expression of other genes. In fact, expression of a minimum 4 key genes is needed in order to predict the expression of all remaining genes. However, in both networks *SREBF2* is always one of the 4 key genes. Looking at the measured and simulated networks we can generate 5 different combinations of key genes from which one combination is in common: *SREBF2, SC5DL, DHCR7* and *E1*. BNs from the measured and simulated data showed persistent connection between *SREBF2* on one side and *CYP51A1*, *E5* and *E2* on the other. Yet *HMGCR* and *E1* are never directly connected to *SREBF2*, but rather interconnected. This indicates that expressions of *CYP51A1*, *E5*, *E2,* and *SREBF2* are similarly regulated, while expressions of *HMGCR* and *E1* are regulated by other factors. This was also taken into account for structural adjustment of the mathematical model: a hypothetical transcription factor T1 regulating *HMGCR*, *E1* and *E2* had to be included in order to obtain a network more similar to the one inferred from the measured data. We were not able to determine which metabolite(s) may regulate the hypothetical factor except that it must be either an intermediate in cholesterol biosynthesis or a metabolite of cholesterol. In simulations, if the factor was regulated by metabolites that were not closely related to cholesterol biosynthesis, the resulting BN was split into two interconnected groups of genes, which was not the case if metabolites from the cholesterol biosynthesis pathway were involved in the regulation. It also remains unclear why the factor would affect expression of genes which are not consecutive enzymes in the biosynthetic pathway (i.e., *HMGCR, E1* and *E2*). Interpreting the results we must consider that the above conclusions hold for liver cells only and that the results may be biased towards the individuals involved in the study; therefore the proposed structural adjustment of the mathematical model needs to be further evaluated using microarray data from additional perturbations of the system and potentially samples from larger number of individuals.

In the study we also explored the possibility of finding key genes of cholesterol homeostasis. Similarly as it was shown for metabolites previously, intermediate lathosterol correlates well with HMGCR enzyme activity and was therefore selected as a marker of the rate of the cholesterol biosynthesis.[30] Using gene interaction networks we wished to find markers whose mRNA levels would indicate the expression of other cholesterogenic genes and also the rate of cholesterol biosynthesis. The results indicate that *SREBF2* is a strong candidate.

6. Conclusion

Bayesian networks have been proven to be a useful tool in deciphering interactions between genes and also in development of mathematical models of biosynthetical pathways. This study revealed that inter-individual differences and measurement noise may seriously hinder our ability to discover gene-to-gene interactions and therefore pose a serious problem if not properly managed. However, using a reasonable number of individuals, proper experimental design together with a stringent control of technical variability of microarray experiments, and large number of independent perturbations of the system enable us to discover interactions between genes and thus gain an important insight into the system under study on the level of mRNA. Within this study we demonstrated how such information may be used to improve the structure of a mathematical model representing comprehensive knowledge of the studied system.

Within the context of cholesterol biosynthesis we have demonstrated that SREBF2 protein may not be the only transcription factor important for regulation of cholesterogenic genes in liver, and proposed how an additional (hypothetical) transcription factor may be involved in the regulation.

7. Acknowledgements

We thank Dr. J. A. Contreras for performing hybridizations, Dr. D. Kuzman and Dr. K. Fon Tacer for the work on initial mathematical model, and Dr. R. Karba for valuable suggestions during mathematical modeling and interpretation of results. The work was supported by the Slovenian Research Agency, grants J1-9438, L1-9569, P1-0104 and Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00413/05). This paper and work it concerns was generated in the context of the Steroltalk project, funded by the European Community as contract No. LSHG-CT-2005-512096 under 6th Framework Programme for Research and Technological Development in the thematic area of Life sciences, genomics and biotechnology for health. This paper reflects only the authors' views and the European Community is not liable for any use that may be made of the information contained therein.

8. Abbreviations

BN – Bayesian network: a probabilistic graphical model that represents a set of variables and their probabilistic independencies

E1 – group of genes* *IDI1, FDPS, GGPS1, FDFT1, SQLE* and *LSS*

E2 – group of genes* *LBR, SC4MOL, NSDHL, HSD17B7* and *EBP*

E5 – group of genes* *MVK, PMVK* and *MVD*

EF – edge frequency: relative frequency of a gene-to-gene interaction appearance irrespectively of the direction of the influence

IS – influence score: a metric for representing the degree to which a gene-to-gene interaction is monotonic in nature, and if so, whether the influence is either positive or negative and its relative magnitude ranging from 0 (lowest) to 1 (highest)

RSD – absolute value of the coefficient of variation (ratio of the standard deviation to the mean) expressed as a percentage

* Abbreviations of gene names follow annotations from Table 1 in Supplement.

9. References

1. L. J. Engelking, G. Liang, R. E. Hammer, K. Takaishi, H. Kuriyama, B. M. Evers, W. P. Li, J. D. Horton, J. L. Goldstein, M. S. Brown, *J Clin Invest* **2005,** *115,* 2489–2498.
2. J. D. Horton, J. L. Goldstein, M. S. Brown, *J Clin Invest* **2002,** *109,* 1125–1131.
3. D. Rozman, M. Fink, G. M. Fimia, P. Sassone-Corsi, M. R. Waterman, *Mol. Endocrinol.* **1999,** *13,* 1951–1962.
4. S. K. Halder, M. Fink, M. R. Waterman, D. Rozman, *Mol. Endocrinol.* **2002,** *16,* 1853–1863.
5. K. Fon Tacer, D. Kuzman, M. Seliškar, D. Pompon, D. Rozman, *Physiol Genomics* **2007,** *31,* 216–227.
6. T. Režen, P. Juvan, K. F. Tacer, D. Kuzman, A. Roth, D. Pompon, L. P. Aggerbeck, U. A. Meyer, D. Rozman, *BMC Genomics* **2008,** *9,* 76.
7. T. Režen, J. A. Contreras, D. Rozman, Drug Metab. *Rev.* **2007***, 39,* 389–399.
8. T. Korošec, J. Ačimovič, M. Seliškar, D. Kocjan, K. F. Tacer, D. Rozman, U. Urleb, *Bioorg. Med. Chem.* **2007**.
9. J. Demšar, B. Zupan, G. Leban, *Orange: From Experimental Machine Learning to Interactive Data Mining,* White Paper, http://www.ailab.si/orange, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, **2004**.
10. W. S. Cleveland, *J. Amer. Statistical Assoc.* **1979,** *74,* 829–836.
11. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, *Bioinformatics* **2001,** *17,* 520–525.
12. J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, E. D. Jarvis, *Bioinformatics* **2004,** *20,* 3594–3603.
13. J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc., **1988.**
14. D. Heckerman, *A Bayesian Approach to Learning Causal Networks,* Technical Report, Microsoft Corporation, Redmond, WA 98052, **1995.**
15. N. Friedman, M. Linial, I. Nachman, D. Pe'er, *J. Comput. Biol.* **2000,** 7, 601–620.
16. N. Friedman, K. P. Murphy, S. J. Russell, *Learning the Structure of Dynamic Probabilistic Networks,* in: G. F. Cooper, S. Moral (Eds.), Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), Morgan Kaufmann, University of Wisconsin Business School, Madison, Wisconsin, USA, **1998,** pp. 139–147.
17. A. Nikitin, S. Egorov, N. Daraselia, I. Mazo, *Bioinformatics* **2003,** *19,* 2155–2157.
18. E. A. Ananko, N. L. Podkolodny, I. L. Stepanenko, E. V. Ignatieva, O. A. Podkolodnaya, N. A. Kolchanov, Nucleic *Acids Res* ***2002,*** *30,* 398–401.
19. P. J. Espenshade, A. L. Hughes, Annu. *Rev. Genet.* **2007,** *41,* 401–427.
20. M. T. Bengoechea-Alonso, J. Ericsson, *Curr. Opin. Cell Biol.* **2007,** *19,* 215–222.
21. K. Schoonjans, C. Brendel, D. Mangelsdorf, J. Auwerx, *Biochim. Biophys. Acta* **2000,** *1529,* 114–125.
22. *Dymola, Dynamic Modeling Laboratory User's Manual,* Dynasim AB, Lund, Sweden, **2004.**
23. B. L. Song, N. B. Javitt, R. A. DeBose-Boyd, *Cell Metab* **2005,** *1,* 179–189.
24. T. Li, J. Y. Chiang, *Am J Physiol Gastrointest Liver Physiol* **2005,** *288,* G74–84.
25. A. G. Olsson, *Clin. Cardiol.* **2001,** *24,* III18–23.
26. K. F. Tacer, T. B. Haugen, M. Baltsen, N. Debeljak, D. Rozman, *J Lipid Res.* **2002,** *43,* 82–89.
27. S. D. Bay, L. Chrisman, A. Pohorille, J. Shrager, *J. Comput. Biol.* 2004, *11,* 971–985.
28. B. Efron, R. J. Tibshirani, An introduction to the bootstrap, Chapman and Hall, **1993**.
29. D. Eberle, B. Hegarty, P. Bossard, P. Ferre, F. Foufelle, *Biochimie* **2004,** *86,* 839–848.
30. I. Bjorkhem, T. Miettinen, E. Reihner, S. Ewerth, B. Angelin, K. Einarsson, *J. Lipid Res* **1987,** *28,* 1137–1143.

Povzetek

Že dolgo je poznano, da raven holesterola v celici uravnava sintezo holesterola preko transkripcijskih faktorjev SREBF, toda v zadnjem času vse več raziskav kaže na vpletenost tudi drugih dejavnikov. Da bi raziskali ta sistem, smo uporabili pristop avtomatske konstrukcije Bayesovih mrež in ga združili s pristopom matematičnega modeliranja in simulacije. Skonstruirali smo matematični model sinteze holesterola in s simulacijami proučevali njegove lastnosti. Z mikromrežo Steroltak smo izmerili spremembe v izražanju genov sinteze holesterola v tretiranih primarnih človeških hepatocitah. S pomočjo Bayesovih mrež, zgrajenih tako iz podatkov meritev z mikromrežami kot tudi iz simuliranih podatkov, smo določili interakcije med geni. Rezultati Bayesovega modeliranja kažejo, da je izražanje holesterogenih genov možno napovedati iz poznavanja izražanja 4 ključnih genov, med katerimi je tudi *SREBF2*. Mreže tudi kažejo na močno interakcijo med genoma *SREBF2* in *CYP51A1*, ne pa tudi med *SREBF2* in *HMGCR*, in da je izražanje *HMGCR* verjetno uravnavano z drugimi dejavniki. Računalniške simulacije matematičnega modela sinteze holesterola so pokazale, da je za določitev interakcij med geni ključno zadostno število perturbacij sistema, in da razlike med posameznimi osebki (biološka variabilnost) in meritvena napaka (tehnična variabilnost) predstavljajo resno oviro pri njihovem avtomatskem določevanju iz podatkov meritev DNA mikromrež.

Supplement

Table 1. Gene names with corresponding accession number and description.

Gene symbol	Gene description	GeneBank Acc. No.
CYP51A1	Lanosterol 14a-demethylase	NM_000786
DHCR24	24-dehydrocholesterol reductase	BC004375
DHCR7	7-dehydrocholesterol reductase	NM_001360
EBP	Emopamil binding protein (sterol C7,8-isomerase)	BC046501
FDFT1	Squalene synthase	BC009251
FDPS	Farnesyl diphosphate synthase	BC010004
HMGCR	HMG CoA reductase	BC033692
HMGCS1	HMG CoA synthase 1	NM_002130
HSD17B7	17-beta-hydroxysteroid dehydrogenase type 7	NM_016371
IDI1	Isopentenyl-diphosphate delta isomerase	BC005247
LBR	Lamin B receptor (potential sterol Δ14-reductase)	not present on Steroltalk array
LSS	Lanosterol synthase	BC035638
MVD	Mevalonate (diphospho) decarboxylase	NM_002461
MVK	Mevalonate kinase	NM_000431
NSDHL	NAD(P) dependent steroid dehydrogenase-like	BC007816
PMVK	Phosphomevalonate kinase	NM_006556
SC4MOL	Sterol-C4-methyl oxidase-like	BC010653
SC5DL	Sterol C5 desaturase	NM_006918
SQLE	Squalene epoxydase	BC017033
SREBF2	Sterol regulatory element binding transcription factor 2	NM_004599