bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Simulation and reconstruction of metabolite-metabolite association networks using a metabolic dynamic model and correlation based-algorithms

Sanjeevan Jahagirdar, Maria Suarez-Diez, and Edoardo Saccenti*

Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708WE Wageningen, the Netherlands

E-mail: esaccenti@gmail.com

Abstract

Biological networks play a paramount role in our understanding of complex biological phenomena and metabolite-metabolite association networks are now commonly used in metabolomics applications. In this study we evaluate the performance of several network inference algorithms (PCLRC, MRNET, GENIE3, TIGRESS and modifications of the MRNET algorithm, together with standard Pearson's and Spearman's correlation) using as a test case data generated using a dynamic metabolic model describing the metabolism of arachidonic acid (consisting of 83 metabolites and 131 reactions) and simulation individual metabolic profiles of 550 subjects. The quality of the reconstructed metabolite-metabolite association networks was assessed against the original metabolic network taking into account different degrees of association among the metabolites and different sample size and

*To whom correspondence should be addressed

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

noise levels. We found that inference algorithms based on resampling and bootstrapping to perform better when correlations are used as indexes to measure the strength of metabolite-metabolite associations. We also advocate for the use of data generated using dynamic models to test the performance of algorithms for network inference since they produce correlation patterns which are more similar to those observed in real metabolomics data.

INTRODUCTION

Biological networks and network analysis are rapidly becoming fundamental tools for the investigation and the interpretation of metabolomics data within a systems biology context.[1] A biological network can be defined as a system of interconnected biological sub-units or nodes: in metabolomics the nodes are metabolites measured in biological samples, and their association is defined on the base of their similarity, which is characterized by a similarity measure.[2] Correlation is usually used as an index to measure the similarity of metabolite profiles and biological information can be derived from both the sign and the magnitude of the correlation: strong positive correlation can indicate an equilibrium condition or enzyme dominance, while strong negative correlation can indicate the presence of a conserved moiety.[3]

Metabolite-metabolite association networks can be used to describe comprehensibly a given status of a biological system or to compare the systems across multiple conditions; in fact, the metabolites association patterns can change upon perturbation of the system, induced under experimental control, or because of the onset of pathological conditions, under the assumption that differences and commonalities in the biological processes are reflected in the characteristics of the reconstructed networks.[4]

In recent times, thanks to the knowledge flow between the metabolomics and the systems biology communities, methods originally developed for the inference of gene regulatory networks have been proposed, extended and adapted to investigate the patterns of association among metabolites.[5–9]

In a previous study[10] we briefly compared the performance of these algorithms on data with

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

known correlation structures and generated using standard multivariate approaches available within computational environments like Matlab[11] and R.[12] However, we noted that generating synthetic data representing the patterns of associations between metabolites and their distributions as observed in real biofluid samples is a non trivial problem. Consider for instance blood that collects molecules from all different tissues in the body: a modeling framework to simulate concentrations of blood metabolites would need to account for reactions and metabolic pathways together with their interconnections and for the uptake and release in and from different organs, tissues, and cell types, forming a system governed by highly non linear dynamics.

For these reasons, generating data to test network inference algorithms using standard multivariate approaches, although straightforward, may not be the most appropriate strategy. In the present study we used a (simplified) model of metabolism to generate metabolite concentration profiles with physiologically plausible distributional and correlation (association) patterns.

This metabolic model is a network structure of interdependent variables (*i.e.* metabolites) that can be described using a system of ordinary differential equations (ODE's). In this way the time course of metabolite concentrations are described as a function of rate laws that account for both enzymatic and non enzymatic reactions, implementing Michaelis-Menten[13] and mass action[14] kinetics. As a test case we reconstructed the metabolic pathway for the degradation of arachidonic acid (AA) making use of previous knowledge available in literature and biological databases. We used this dynamic model, consisting of 83 metabolites and 131 reactions, to generate physiologically plausible metabolite concentration profiles simulating data for 550 individuals, introducing inter-individual variability by randomly varying the kinetic constants of the model.

We used this data to test the performance of several algorithms to reconstruct metabolite-metabolite association networks, such as: PCLRC (Probabilistic Context Likelihood of Relatedness on Correlation),[7] MRNET (Multicast REduction NETwork),[15,16] GENIE3 (Gene Network Inference with Ensemble trees),[17] TIGRESS (TRustfull Inference of Gene REgulation using Stability Selection)[18] and modification of the MRNET algorithm, together with standard Pearson's

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and Spearman's correlation, by comparing the reconstructed networks with the structure of the original AA metabolic model.

Our results show the validity of such computational approach and indicates that methods based on resampling or non linear decision tress like Random Forest are be better suited for the reconstruction of metabolite-metabolite association networks than methods based on linear regression. Moreover, our approach also allows to evaluate to what extent high (partial) correlation coefficients correspond to known metabolic reactions, as it has also been shown,[3,5] and we will also show, that metabolites directly linked in a reaction step may not exhibit any correlation.

MATERIALS AND METHODS

Dynamic model of Arachidonic Acid metabolism

Arachidonic acid (AA) and its derivatives (eicosanoids) have been extensively studied in the last 30 years due to their activity as signaling molecules and their role in the cross-regulation of the inflammatory response in the human body.[19–21]

Here we have developed a dynamic model of AA degradation by retrieving information about metabolites and eiconasoids involved in AA metabolism and associated reactions from the following databases (see Figure 1 for an overview and the Supporting material (Tables S1 and S2) for a complete list of metabolites): Recon2.2,[22] the Human Metabolome Datababase (HMDB),[23–26] the Kyoto Encyclopedia of Genes and Genomes (KEGG)[27–29] and the Chemical Entities of Biological Interest database (ChEBI).[30–33] The data retrieved from these databases included metabolite names, identifiers and IUPAC names, metabolite concentrations in normal physiological conditions in humans, reaction descriptions (substrates, products, cofactors and stoichiometry) and list of associated enzymes. Kinetic data, such as $K_m$values for enzymes and k_i values as reaction/equilibrium constants, were collected from BRENDA[34] and SABIO-RK.[35]

The reconstructed AA metabolic pathway is depicted in Figure 2. In this model, the starting point is activation of phosphatidylcholine (PC) that leads to AA synthesis and formation of AA-

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

derived eicosanids such as prostaglandins, leukotrienes, thromboxanes and prostacyclines.[36,37] The model accounts for concentrations of 83 metabolites and includes 131 reactions. An ODE was written for each metabolite in the model to account for synthesis, conversion and (when needed) degradation. Enzymatic reactions were described using Michaelis-Menten kinetics:[13]

$$v = \frac{V_{max} \times [S]}{K_M + [S]}, \tag{1}$$

where v is the rate of reaction, V_{max} is the maximum reaction rate achieved by the system, $[S]$ is substrate concentration, and K_M is the Michaelis-Menten constant. Non-enzymatic reactions were described by simple mass action laws, where the rate of the reaction, v is proportional to substrate concentration, $[S]$, ($v = k_i \times [S]$.) and with the proportionality constant, k_i, being the equilibrium constant of the reaction.

It should be noted that we have greatly simplified the reactions in this pathway and we have not explicitly described co-factors such as ATP or NADP, instead their impact is implicitly incorporated in the values of unknown kinetic constants through the fitting to measured data (see below).

A full description of the model (and ODEs) together with the description and values (when available) of all kinetic constants[25,38–75] are given in the Supplementary Material S1 (Tables S1 and S2); the Supporting material S2 contains the metabolic model in SBML format.

As an example we provide here description of formation and degradation of AA, with the conversion of PC into AA: this is modelled as an ODE coupling the concentrations of each

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

metabolite ([AA] and [PC] respectively):

$$\begin{aligned}\frac{\mathrm{d[AA]}}{\mathrm{dt}} = & \left(\frac{V_2 \times [PC]}{Km_2 + [PC]}\right) - \left(\frac{V_{100} \times [AA]}{Km_{100} + [AA]}\right) - \left(\frac{V_{103} \times [AA]}{Km_{103} + [AA]}\right) - \left(\frac{V_{106} \times [AA]}{Km_{106} + [AA]}\right) - \\ & - \left(\frac{V_{109} \times [AA]}{Km_{109} + [AA]}\right) - (k_{112} \times [AA]) - (k_{114} \times [AA]) - (k_{116} \times [AA]) - \\ & - \left(\frac{V_{118} \times [AA]}{Km_{118} + [AA]}\right) - \left(\frac{V_{119} \times [AA]}{Km_{119} + [AA]}\right) - \left(\frac{V_{120} \times [AA]}{Km_{120} + [AA]}\right) - \\ & - \left(\frac{V_{121} \times [AA]}{Km_{121} + [AA]}\right) - \left(\frac{V_{122} \times [AA]}{Km_{122} + [AA]}\right) - (k_{129} \times [AA]) - \left(\frac{V_3 \times [AA]}{Km_3 + [AA]}\right) - \\ & - \left(\frac{V_4 \times [AA]}{Km_4 + [AA]}\right) - \left(\frac{V_{43} \times [AA]}{Km_{43} + [AA]}\right) - \left(\frac{V_{50} \times [AA]}{Km_{50} + [AA]}\right) - \left(\frac{V_{90} \times [AA]}{Km_{90} + [AA]}\right) - \\ & - \left(\frac{V_{94} \times [AA]}{Km_{94} + [AA]}\right) - \left(\frac{V_{95} \times [AA]}{Km_{95} + [AA]}\right).\end{aligned} \tag{2}$$

The model was constructed in a single compartment with a 1L volume so that concentration changes in metabolites (in nmol/L) due to small volume changes could be considered negligible.

Optimization of parameters in the model.

Not all reaction parameters for the reconstructed AA metabolic model (k_i, K_m and reaction rates) were available in literature. Missing parameters where estimated by identify sets of parameters maximizing the agreement between simulated concentrations (at steady state) and metabolite concentrations derived from literature (see Table S1 and S2 in the Supporting material S1 for a list of steady state concentrations and associated references). The optimization was done using the Genetic algorithm approach implemented in the simulation software COPASI[76] using 2000 generations and each time choosing the 20 best performing sets of parameters.

Simulation of individual metabolite concentration profiles

The metabolic model of AA metabolism was used to generate physiologically plausible metabolite concentration profiles for a pool of N individuals. Initial concentration of all metabolites was set to 0 except for phosphatidylcholine. Inter-individual variability (i.e. different profiles) was

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

simulated by varying the Km and the k_i constants and the reaction rates v for all reactions in the model. Values for Km, k_i and v were sampled from an uniform distribution with lower and upper bounds set to the reference values $\pm 10\%$. For instance, the j-th individual, the values of k_i, Km and v for a particular reaction were defined as

$$k_i^j \approx U(0.9 \times k_i, 1.1 \times k_i)$$

$$Km_i^j \approx U(0.9 \times Km_i, 1.1 \times Km_i)$$

$$V_i^j \approx U(0.9 \times V_i, 1.1 \times V_i)$$

The reference values used for k_i, Km_i and V_i, together with the associated references are given in Table S1 in the Supporting material.

Long term simulations showed the system (with optimal parameters) to achieve a steady state for all metabolites at approximately 70 hours after activation of the AA conversion. This time is consistent with previously reported experimental data.[77,78] For each of the N individuals, a time course of 100h was simulated and data was collected at two time points, 10 and 90 hours after induction of AA conversions. These correspond to non-steady state (NS) and steady state (SS) conditions.

Noise addition

Gaussian uncorrelated noise with zero mean and variance-covariance matrix $\mathbf{\Sigma} = \mathbf{I}$ (with $\mathbf{I}$ a 83×83 identity matrix was added to data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ to simulated experimental noise:

$$\mathbf{X}^{noised} = \mathbf{X} + w \cdot N(0, \mathbf{I}), \tag{3}$$

where the noise intensity levels w correspond to signal-to-noise ratios (SNR) of 1%, 5%, 10% and 15%. SNR is defined as the ratio of median peak intensity to the intensity of additive noise. $\mathbf{X}$ represents the data matrices containing the metabolite concentration profiles.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data overview

Overall, we generated $N = 550$ metabolites profiles, from which we isolated $N_1 = 50$ and $N_2 = 500$ profiles to mimic a small and a large metabolomic study, respectively. In each case we generated data corresponding to steady state (SS) and non-steady state (NS) solutions. Data were stored in four data matrix $\mathbf{X}_1^{SS}$, $\mathbf{X}_1^{NS}$ (with dimensions 50×83) and $\mathbf{X}_2^{SS}$ and $\mathbf{X}_2^{NS}$ (dimensions 500×83. For each of the four data configurations we considered noise levels (SNR) 1%, 5%, 10% and 15%.

Model representation in matrix form

A network represents relationships among a set of objects, such as the molecular components of a biological system, like genes, proteins or metabolites. The components are represented as as nodes while relationships are represented as links (or edges) between the nodes.

A network can be represented as a square matrix called the adjacency (or also connectivity) matrix. Rows and columns in the matrices represent nodes (in this case, metabolites), whereas non-null entries represent links. In a *weighted* adjacency matrix the entries are real numbers that indicate the strength of the interaction. Values in the *weighted* adjacency can vary, for instance, in the $[-1, 1]$ range for correlation, in the $[0, +\infty)$ range for mutual information, or in the [0,1] range for probability. A weighted adjacency matrix $\mathbf{W}$ can be transformed into a binary adjacency matrix $\mathbf{A}$ by imposing a threshold on the (absolute) value on its entries W_ij:

$$A_{ij} = \begin{cases} 1 \text{ if } \ |W_{ij}| > \lambda \\ 0 \text{ otherwise} \end{cases}. \tag{4}$$

The threshold λ depends on the method considered and on the nature of the adjacency weights; for instance, if correlation is used, λ can be set to 0.6 as proposed in the metabolomics literature;[3] for mutual information a threshold $\lambda = 0$ is usually taken, while for PCLRC, a probability threshold $\lambda = 0.95$ has been proposed.[7]

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

From the network representation of the model (shown in Figure 1), we define three adjacency matrices $\mathbf{A1}$, $\mathbf{A2}$ and $\mathbf{A3}$ accounting for different degree of association among metabolites, *i.e.* different degree of neighborhood in the AA metabolic pathway.

The three adjacency matrices exemplified in the cartoon shown in Figure 3 and are defined as follows.

i) Primary adjacency matrix $\mathbf{A1}$: two metabolites are associated if they connected by a single reaction, i.e they are products or substrates in the same reaction. Examples are arachidonic acid and its direct derivative prostaglandin-H2 (PGH2) and PGH2 and its derivative prostaglandin-H2 (PGE2)

$$\mathrm{AA} \rightarrow \mathrm{PGH2}$$

ii) Secondary adjacency matrix $\mathbf{A2}$: two metabolites are associated if they are second neighbors in the metabolic map, *i.e.* if they can be inter-converted through two reaction steps via an intermediate metabolite. Examples are arachidonic acid and prostaglandin-H2 (PGE2) connected via PGH2:

$$\mathrm{AA} \rightarrow \mathrm{PGH2} \rightarrow \mathrm{PGE2}$$

It follows that any two metabolites connected within $\mathbf{A1}$ are also connected in $\mathbf{A2}$.

iii) Tertiary adjacency matrix $\mathbf{A3}$: two metabolites are associated if they are third neighbors in the metabolic map, *i.e.* if they can be inter-converted by three reaction steps via two intermediate metabolites. Examples are arachidonic acid and prostaglandin-A2 (PGA2) connected via PGH2 and PGE2:

$$\mathrm{AA} \rightarrow \mathrm{PGH2} \rightarrow \mathrm{PGE2} \rightarrow \mathrm{PGA2}$$

If any two metabolites are connected within $\mathbf{A1}$ and $\mathbf{A2}$ they are also connected in $\mathbf{A3}$.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Algorithms for network reconstruction

In our previous work[10] we have investigated the performance of some algorithm for the reconstruction of metabolite-metabolite associations networks, namely the CLR,[79] ARACNE,[80] PCLRC[7] algorithms together with standard Pearson's correlations. Here we considered further methods selecting several algorithms from those proposed in the comparative paper by Marbach and co-workers[81] where over 30 network inference methods were assessed for the reconstruction of gene regulatory networks from high-throughput data. The rationale for the selection was that the methods covered different kinds of approaches and that the algorithms were easily available, easy to use and could be used taking correlation as input metric. It should be noted that, apart from PCLRC, none of the proposed methods was developed with the goal of reconstructing metabolite-metabolite association networks but were developed to study gene regulation. Since in early work we found the resampling approach implemented in PCLRC rather effective, we propose here modifications of existing algorithms by introducing a resampling step. Hereafter, the algorithms considered are briefly explained, mostly following the concise descriptions reported in the supplementary material of Marbach's paper.[81] We refer the reader to the original publications for more details.

CORR - Correlation

This algorithm relies on computing the correlation coefficient among all possible metabolite pairs to infer an association network, although, as previously discussed, correlation may not necessarily translate into association and *viceversa*. The main drawback is the inclusion of chance correlations (especially in the case of small sample size) which may complicate the interpretation of the resulting association networks. Here, we will use both Pearson's and Spearman's correlation coefficient.[82,83] The methods are indicated with CORR-p and CORR-s, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PCRLC - Probabilistic Context Likelihood of Relatedness on Correlation

The PCLRC algorithm is based on the CLR algorithm[79] and has been first introduced to reconstruct metabolite correlation networks.[7] It combines the CLR approach replacing mutual information with correlation with iterative sampling. In each iteration, 70% of the samples from the full data set are randomly chosen to calculate the correlation matrix. Then, CLR is used to estimate possible associations and reconstruct a network and then dynamical threshold is chosen so that the highest 30% of interactions are kept. The complete procedure (sampling and network generation) is iterated K times. The final network is constructed by assigning to every edge a weight defined the frequency of the times that a particular edge was selected in the iterative process. This weight can be viewed as a probabilistic measurement of edge likeliness and can be interpreted as a confidence level on which to accept or reject the correlation between two pairs of metabolites. We set here $K = 10^5$. The algorithm output a weighted adjacency matrix whose entries are the probability of association for all possible X, Y metabolites pairs. We implemented the algorithm with both Pearson's and Spearman's correlation: the tow variants are indicated with PCLRC-p and PCLRC-s, respectively.

MRNET - Multicast Reduction NETwork algorithm

The algorithm works by evaluating the correlation matrix by the Minimum Redundancy Maximum Relevance (MRMR) selection process[84] for every variable.[15,16] The logic underlying the MRMR algorithm is that immediate edges are scored higher while distant edges are scored lower. This selection process works by choosing a target metabolite Y, and then finding the metabolite, X_i, with the largest correlation with Y (maximal relevance). The second selected variable X_j is the one having highest correlation with the target variable Y $\mathrm{corr}(X_i, Y)$ and at the same time, lowest correlation $\mathrm{corr}(X_i, X_j)$ with the previously selected variable X_i (minimal redundancy). This set S of variables is enlarged by selecting the X_k variables that maximize the MRMR score:

$$S(X, Y) = corr(X, Y) - < corr(X, X_i) > \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for all X_i already in S. The procedure stops when the score becomes negative. The MRMR score represents a trade-off between relevance and redundancy. In practice, for each target variable Y, MRNET identifies those variables X having maximal pairwise relevance with Y and, at the same time, maximal pairwise independence among them.[85] The algorithm returns a weighted adjacency matrix of the inferred metabolite-metabolite association network: the weight of each pair of metabolites X and Y is the maximum score between the one computed when Y is the target and the one computed when X is the target:

$$W_{XY} = \max\left(S(X,Y), S(Y,X)\right) \tag{6}$$

The original MRNET algorithm implemented mutual information instead of correlation. Here we replaced the mutual information matrix with the correlation matrix containing the pairwise correlation among all metabolites in the data set. We implemented the MRNET algorithm with both Pearson's and Spearman's correlation, indicated as MRNET-p and MRNET-s, respectively.

PMRNET - Probabilistic Multicast Reduction NETwork algorithm

The PMRNET algorithm is based on the combination of the MRNETp and MRNETs algorithms and iterative sampling of the data similarly to what done in the PCLRC algorithm. In each iteration, 70% of the samples from the original data set are randomly chosen and MRNETPp or MRNETs is used to calculate the weights matrix W_{XY} as in Eq. 6). Then dynamical threshold is chosen so that the 30% of the interactions with the highest weights are kept. The complete procedure (sampling and network generation) is iterated $K = 10^5$ times. The final network is constructed by assigning to every edge a weight defined the frequency of the times that a particular edge was selected in the iterative process. The two versions of this algorithm are indicated as PMRNET-p and PMRNET-s, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GENIE3 - GEne Network Inference with Ensemble trees

This algorithm represents prediction of interactions among p metabolites as p regression problems.[17,86] Random Forest is then used to predict the concentration of every target metabolite using the concentrations of all the other metabolites. The basic idea of tree-based regression methods is to recursively split the learning sample with binary tests based each on one input metabolite at a time; the tests are optimized to reduce the variance of the output variable (here, the concentration of the target metabolite) in the resulting subsets of samples. At each test metabolite, K attributes are selected at random among all candidate attributes before determining the best split. We used $K = \sqrt{p}$ and grow ensembles of 1000 trees.[81]

To each metabolite pair X, Y a weight W_{XY} is assigned that exploits the variable importance measure of the expression of metabolite X as derived from the tree-based model learned for metabolite Y. For each test metabolite, X the total reduction of the variance of the output metabolite due to the split is considered:

$$w(X) = \#S.\mathrm{var}(S) - \#S_t.\mathrm{var.}(S_t) - \#S_f.\mathrm{var}(S_f) \tag{7}$$

where $\#$ is the cardinality of a set of metabolites, S is the set of metabolites reaching metabolite X, S^t and S^f are the subsets for which test was true and false respectively and $var()$ is the variance of the output metabolite in the subsets. The importance measure for all metabolites was calculated as the average of all the scores from all the trees calculated for each metabolites.

TIGRESS - TRustfull Inference of Gene REgulation using Stability Selection

The TIGRESS algorithm is based on Lasso regression[18] Similarly to the GENIE3 algorithm every target metabolite is predicted using the concentration of all other metabolites, thus turning the network inference problem into p regression problems. Specifically, for each metabolite Y, the

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

following regression problem is stated

$$Y = \mathrm{reg}(X_S) + \epsilon_Y \tag{8}$$

where Y is the concentration of the target metabolite, X_S is the subset of candidates metabolites associated with metabolite Y and $\mathrm{reg}()$ is any (linear or not linear) regression function. A scoring function $g(X)$ is defined for each candidate metabolite X using Least angle regression (LARS)[87] for feature selection. LARS uses a linear regression model, thus $\mathrm{reg}()$ in Equation (8) has the functional form

$$\mathrm{reg}(X) = \sum_X \beta_X X \tag{9}$$

Starting from a constant model where no candidates metabolites are used, metabolites are iteratively added to the model to refine the prediction of the target metabolite Y: after $K = 100$ steps of the LARS iteration, a ranked list of K metabolites, selected for their ability to predict the target metabolite is obtained. Candidate metabolites are not scored directly using LARS but a stability selection procedure is applied by running the feature selection method many times on randomly perturbed data, and score each metabolite by the number of times it was selected. The input data is perturbed by multiplying metabolite concentrations by a number uniformly sampled in the interval $[\alpha, 1]$ for some $0 \leq \alpha \leq 1$. Here we used $\alpha = 0.2$ as suggested in the original publication.[88]

This approach has been shown to reduce the sensitivity of LARS and Lasso to correlated features, and improve their ability to select correct features.[88,89] For each metabolite a final score between 0 and 1 is obtained: a 1 indicates that metabolite X is always selected by LARS in the top features to predict the concentration of metabolite Y, and 0 indicates it is never selected.[18]

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Metrics for performance assessment

To assess the performance of the different algorithms for the reconstruction of metabolite-metabolite association networks we used the area under the Precision-Recall (PR) curve (also called the Receiver Operating Characteristic curve) indicated with AUC (sometimes also referred as AUROC), comparing the inferred association against the golden standard given by the "true" association network for the AA metabolic pathway as defined by the adjacency matrices $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ previously defined.

For a binary classifier, the ROC curve is a plot of the recall ($i.e$ the $\mathrm{sensitivity}$) against $\mathrm{precision}$ ($i.e.$ $1 - \mathrm{specificity}$):

$$\mathrm{precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\mathrm{recall} = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{11}$$

where P is the number of positives, *i.e.* the number of interactions in the adjacency matrix representing the model (either $\mathbf{A}_1$, $\mathbf{A}_2$, or $\mathbf{A}_3$); TP is number of true positives *i.e.* the number of associations predicted by the algorithm that correspond to interactions in the adjacency matrix representing the model; FP the number of false positives: number of predicted associations not found in the adjacency matrix; and FN the number of false negatives, the number of associations in the adjacency matrix the algorithm failed to recover.

The methods for network inference considered here produce weighted adjacency matrices, $\mathbf{W}$. To compute precision and recall it is then necessary to select a threshold to obtain a binary adjacency matrix (Equation 4). The PR curve is constructed by varying the threshold to take values between the lowest and highest entries in the weighted adjacency matrix. Once the PR curve is plotted, the area under the curve, AUC, is calculated. The limiting value $AUC = 1$ indicates perfect classification, $i.e.$ equivalence of $\mathbf{W}$ with $\mathbf{A1}$ (or $\mathbf{A2}$ or $\mathbf{A3}$) while $AUC = 0.5$

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

indicates a random classification.

Experimental data

Two supporting experimental metabolomics data sets were also considered. The first data set contains 133 quantified metabolites measured on 2139 individuals from the Swedish TwinGene study[90] using Mass spectrometry. Data were downloaded from the MetaboLights public repository[91] (www.ebi.ac.uk/metabolights, accession no. MTBLS93). The second data set contains 29 blood metabolites measured using Nuclear Magnetic Resonance on 864 adult healthy volunteers.[7] We refer the reader to the original publications for more details on the study design, sample collection, and experimental protocols.

Software

Calculations were performed using R,[12] Matlab[11] and Python.[92] The R code for PCLRC is available at www.systemsbiology.nl under the software tab. We used the MRNET implementation available in the MINET R package;[93] and the Bioconductor package GENIE3.[17] The R code for TIGRESS was downloaded from http://cbio.ensmp.fr/tigress. Simulation and analysis of the AA biochemical network and tits dynamics was performed using COPASI (a COmplex PAthway SImulator).[76]

RESULTS AND DISCUSSION

In silico generation of metabolites profiles

In this study we wanted to create physiologically plausible metabolite concentration profiles showing correlation patterns similar to those that can be expected to be measured in a standard metabolomic experiment using analytical platforms like Nuclear magnetic resonance (NMR) or Mass spectroscopy (such as LC-MS/MS). Therefore we developed a dynamic model of AA

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

metabolism able to simulate the metabolite concentrations measured in the blood samples of different subjects. The presented model has been parametrized using literature information for kinetic constants and is able to reproduce, at steady state (SS), metabolite concentrations derived from literature.

AA conversion and the formation and degradation of downstream metabolites in the white blood cells and the dynamics of AA derived metabolites in blood (and their sampling) happens at very different scale: hours for blood dynamics and seconds for AA conversion. For this reason, the variation of the downstream products can be considered to be negligible, that is the systems can be considered to be in a steady state. Another reason to investigate the system in steady state is that metabolite-metabolite associations networks are usually build starting from correlations which, in principle, should be calculated starting for samples acquired under the same conditions. If the system(s), in this case, different subjects, are not in the same state, or if they are not in a steady state, the samples (*i.e.* the metabolite concentration profiles) will likely not align, possibly introducing bias in the correlations.[3]

Assuming a complex system like the human body to be in a steady state at the moment of sampling is not always realistic. This assumption may hold true in the case of a few metabolites which are measured under well-defined experimental conditions once the dynamic properties of the system are well characterized, but this is seldom the case in metabolomic studies which usually aim to provide a comprehensive snapshot of the metabolome under given (patho)physiological conditions often without any prior knowledge of the dynamic nature of the systems investigated. For this reason we considered also metabolite profiles sampled in non-steady state (NS), 10 hours after the activation of AA degradation. As expected the individual metabolic profiles acquired when the system is at steady state are more homogenous than those measured in the non-steady state, as indicate by the PCA analysis shown in Figure 4 panel A where the SS profiles are tightly clustered while NS profiles are dispersed in the PCA space.

Figure 4 panel B offers a comparison of the metabolite correlation patterns in SS and NS data, calculated using both Pearson's and Spearman's indexes. Pearson's correlations are lower than

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spearman's one, indicating the presence of non-linear relationships among the metabolites: this is what is usually observed in biochemical data because the dynamic governing the relationships among metabolites (see for instance AA model Equation (2)) are highly non-linear.

Correlations in the steady state are mostly below 0.6 (76.8% for Spearman's correlation and 90.5% for Pearson's correlations) and this is consistent to what observed in experimental practice and it is supported by chemical considerations: this can happen because the variance in the enzymes that control the reaction involving two metabolites affects them in equal amounts and different directions:[3] this is what happens to the majority of metabolite pairs in the dynamic model, and is a consequence of the systemic nature of metabolic control.[3]

When correlations are measured on samples acquired at non-steady state we observed a strong increase in the magnitude of correlations: 82.8% of Spearman's correlation are above 0.8. This may seem counter-intuitive since one can expect that if metabolite concentrations are not aligned the correlation can be biased downwards: however it should be noted that the median variance of metabolite concentrations in the non-steady state is much larger than the median variance in the steady state (3.7×10^5 versus 2.5) and this can bias correlation upwards. Another explanation is that sampling far away from the steady state could results in stronger non-linear relationships, a behavior well described by the Spearman's index. Another hypothesis is that at the particular sampling point chosen for the non-steady state most part of metabolite pairs are at chemical equilibrium and this usually originates very strong correlations.[3]

Overall, it seems that data simulated in the steady state shows a correlative structure which is more similar to real metabolomics data: as a term of comparison 93.4% and 91.4% of correlation values (Spearman) observed in two blood metabolomics data sets (MS and NMR, see Material and Methods for more details) are below 0.6.

Model representation

Three adjacency matrices $\mathbf{A1}$, $\mathbf{A2}$, and $\mathbf{A3}$, were used to represent the original network underlying the AA dynamic. These matrices are schematically represented in Figure 2 and provide standards

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

against which to compare the reconstructed networks. Each of these matrices account for different degree of association (direct and indirect) among the metabolites in the AA metabolic pathway, where association is here defined in terms of distance in the metabolic map.

The classical approach to test algorithms for network reconstruction in metabolomics applications is to generate multivariate data with a given distribution and known covariance-correlation matrix Σ and then reconstruct the correlation patterns which are then compared with the original known correlations. This can be done by directly comparing the actual values of the true and reconstructed correlations (*i.e.* comparing the weighted adjacency matrices) or by comparing the adjacency matrices after that a threshold as been imposed on the correlation values (*i.e.* the patterns of 0 and 1.)

The dynamic metabolic model enables a more exhaustive analysis on the nature of the associations. When metabolite profiles are generated using a dynamic metabolic model, correlations observed in the data do not always reflect the structure of the metabolic network: two metabolites can be direct neighbors in the metabolic network but not correlated; conversely two metabolites can be very distant in the metabolic network but showing high correlation. This a well-known phenomenon and also in the case of this *in silico* simulated data we observe this typical behavior of the dynamic system. For instance, the correlation between PGD2 (prostaglandin D2) and dhkPGD2 (11,15-Dioxo-9S-Hydroxy-5z-Prostenoic Acid) is 0.94 but the two metabolites are separated by two enzymatic reactions and one intermediate metabolite (15kPGD2, 15-dehydro Prostaglandin D2). On the contrary, the correlation among AA and LTA4 (leukotriene A4) is only 0.00015 despite the fact that they are direct neighbors in the metabolic pathway. Another striking example are 6k-PGE1 (6-ketoprostaglandin E1) and 5-oxoETE (5-ketoeicosatetraenoic acid) whose correlation is 0.6 even if they are separated by 8 reaction steps and 7 intermediate metabolites.

From a practical point of view this poses the problem of defining a target to which compare the reconstructed association networks: when are two metabolites to be considered associated in the original metabolic network? Stated otherwise, what is the adjacency matrix corresponding to

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the AA metabolic pathway shown in Figure 2? This is why we defined three adjacency matrices to represent different degree of association among metabolites in the pathway and this is also the reason we prefer to talk about reconstructed metabolite-metabolite association networks rather than correlation network.

Performance of network inference algorithms

The performance of the different methods are summarized in Figure 5 for sample size $n = 500$ and in Figure 6 for sample size $n = 50$. Numerical results are presented in the Supporting material S1 (Tables S4-S13).

The PCLRC algorithm was, overall, the best performer in the comparison, and showed to be robust respect to both sample size and noise: these results are in line with previous results.[10] The use of simple correlation (either Pearson's or Spearman's) is surprisingly good in all situations, which is probably due to the relative larger sample size used, although we note, as expected, a reduction in performance when noise is added to the data.

The performance of the MRNET algorithm was suboptimal in all situations, irrespective of the sample size, data type (SS or NS), noise level or target adjacency used the AUC is constantly below 0.63, indicating that the PR curve is close to the diagonal in the P-R plane, indicating that the solution of the method is insensitive to the particular threshold imposed on the weighted adjacency matrix ($i.e$ the λ parameter in Equation (4)) to binarize its entries. This means that the false positive rate is consistently low and the algorithm seems to be able to produce only two types of solutions, namely, with low FPR or the trivial full connected solutions with FPR equal to 1. However, when a resampling approach is implemented (*i.e.* PMRNET-s and PRMNET-p) like in the PCLRC algorithm , the quality of the prediction increase significantly, with increased robustness towards noise.

The TIGRESS algorithm was the most affected by sample size: the AUC values vary between 0.65 and 0.83 for $n = 500$ but dropped to 0.5 for $n = 50$. This result is quite surprising: this algorithm implements a stability selection procedure by randomly perturbing the experimental

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

data by multiplying the metabolite concentration with random numbers. The low performance may be due to the interaction between the small sample size and the perturbation, since a larger sample size is usually needed to compensate for high(er) noise level or to an unfortunate choice of α. The algorithm indeed has been found to be very sensitive to this parameter,[18] however, when perfectly tuned, its performance was found to be equivalent to the GENIE3 algorithm[17,86] for the reconstruction of gene regulatory networks. Another possible explanation is that the algorithm implements variable selection and it may happen that such procedure is not stable when a small sample size is considered. For the case $n = 500$ the performance of TIGRESS was good although less robust against noise than GENIE3.

The GENIE3 algorithm was the overall top performer in the DREAM3 challenge for gene network inference.[81] In the present case it also showed a good performance and was minimally affected by changes in sample size. GENIE3 and TIGRESS follow a similar conceptual approach, however in the GENIE3 algorithm regression is performed using Random Forest, that is randomized decision trees, while in TIGRESS linear regression was implemented (see Equation (9)): we have seen that dynamic data exhibits non-linear behavior and Random Forest is particularly well-suited to model non-linear relationships and this may explain why GENIE3 outperformed TIGRESS.

With respect to the use of the primary ($\mathbf{A1}$), secondary ($\mathbf{A2}$) and tertiary ($\mathbf{A3}$) adjacency matrices as target for the network construction, we found the performance of CORR and PLCRC algorithms to increase when considering higher order metabolite-metabolite associations, with PCLRC performing slightly better. The use of PCLRC over the standard correlation approach (CORR) has the advantage that it is not necessary to set a threshold on the correlation value; the weights given to each metabolite-metabolite association do not depend on the magnitude of the correlation itself but on the number of times they appear in the top $Q\%$ (similarly to what is done in the stability step of the TIGRESS algorithm) and this has the double advantage of *i*) filtering out spurious correlations which may arise from sampling effects and *ii*) maintaining metabolite-metabolite associations regardless from the associated correlation.

This is particularly relevant considering that there isn't always a correspondence between the

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

correlation among two metabolites and their proximity in the metabolic map. It is of course possible to choose an optimal threshold that simultaneously maximizes precision and recall (see Equations (10) and (11)) but in real life the target adjacency matrix is not known. Other approaches have been suggested, like observing the topology of the network obtained using different thresholds.[94] Here a transition (from small to large) was observed in the number of connected metabolites at $\lambda = 0.5$, a value not far from 0.6, indicated as a lower bound for low/weak correlations in metabolomics.[3]

The performance of the TIGRESS algorithm based on Lasso regression heavily depended on the choice of the target adjacency matrix: when $\mathbf{A1}$ is used, *i.e.* when only direct neighbors metabolites are considered to be associated, the AUC is 0.836, but it drops to 0.614 when $\mathbf{A3}$ is considered. This behavior can be explained when considering that this algorithm was developed to infer gene regulatory networks, whose properties are different from metabolite-metabolite association networks. In gene regulatory networks there is asymmetry between the regulators (transcription factors) and their gene targets. This causes genes regulated by the same transcription factors to exhibit similar expression patterns, resulting in a large number of false positives in the reconstructed gene regulatory networks. For these reason algorithms for gene regulatory network reconstruction are often designed to downgrade the importance of indirect associations: this may explain the suboptimal performance of TIGRESS when used to infer indirect associations[1]. However, also GENIE3 has been originally developed to infer gene regulatory networks and, likewise PCLRC, its performance increase when considering higher order metabolite-metabolite associations.

The performance of MRNET is suboptimal in all cases, and it decreases when considering indirect associations, for the same reasons discussed for TIGRESS. However, when a resampling approach is implemented (*i.e.* PMRNET-s and PRMNET-p), the quality of the prediction increases when also considering indirect associations, as observed for PCLRC and TIGRESS.

Regarding the use of Pearson's or Spearman's correlations we found in general better perfor-

[1]In a previous study[10] we observed the same behavior for another algorithm ARACNE[80] which was not included in the present comparison because of its poor performance.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mance when using Spearman's correlation which is reasonable given the non-linear relationships existing among metabolite concentrations. Overall the performance of all algorithms is reduced in presence of noise: under the additive noise model considered (see Equation (3)) correlations are always attenuated[83,95] and this phenomenon is reduced when considering more replicates.

Dependence of the quality of the inferred network on the the use of data acquired in a steady or non-steady state is more complicate to discuss. When a limited sample size is used $n = 50$, we found all methods to perform better on steady state data than on non-steady state and this is reasonable since data is more homogenous as previously observed (see Figure 4.) We also see that for small sample size, the presence of noise has marginally less impact on steady state data than on non-steady state data. When a large sample size is considered ($n = 500$) we do not observe significant differences in the performance of the methods regarding the use of the two types of data. CORR-s performed well in several cases, but it showed the largest decrease in performance depending on sample size and noise level, which may render it sue not optimal for limited sample size.

CONCLUSIONS

We have explored the ability of several algorithms (and adaptations thereof) to reconstruct metabolite-metabolite association networks using correlation (Pearson's and Spearman's) as similarity metrics. We have tested the performance of the algorithms on data generated using a dynamic metabolic model of arachidonic acid metabolism. The model describes 131 reactions interconverting 83 metabolites. The developed model has been parametrized using literature knowledge and is able to reproduce measured concentrations; however it has been greatly simplified and effects such as enzyme or allosteric control have not been explicitly taken into account. This simplification is partly taken into account when considering different degrees of neighborhood in the adjacency matrices representing the model.

While correlation (CORR) is an almost native metric to investigate the association among

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

metabolites in metabolomics studies, all methods but one (PCLRC) were developed to infer gene regulatory networks that have markedly different properties from metabolite-metabolite association networks. We found methods based on resampling like PCLRC (and our modifications of the MRNET algorithm) to perform better together with regression methods based on Random Forest in which resampling is intrinsically implemented.

Except CORR-s in a few cases, none of the methods achieved consistently AUC above 0.8: considering that different adjacency matrices were used to represent the AA metabolic model implementing different levels of neighboring among metabolites and that in general the performance of the methods increased when indirect association are considered we can conclude that not all direct associations (*i.e.* metabolites connected by one single reaction step) are recovered. Rather than from a limitation of the reconstruction algorithms this is an intrinsic property of the systems studied:[3] due to the systemic nature of metabolic control, the concentration of metabolites connected by one single reaction step may not be correlated and this information cannot be recovered in absence of truly dynamic data or metrics able to capture such behavior.

We are aware that reconstructing and implementing a dynamic metabolic model to generate *in silico* data may not be an easy task and that such approach does not offer the operative flexibility given by standard multivariate approaches, not to mention the lack of statistical characterization of such dynamic data. A possible solution to partially overcome these limitations could be to generate multivariate data with known correlation structure derived from dynamic data and to model data distribution using more advanced data generation techniques.[96,97]

However, it should be remembered that metabolite-metabolite associations are not a property of metabolites and enzymes. Such associations may arise due variation of a single enzyme, or through differential amplification of the variance of a single enzyme but overall metabolite-metabolite association patterns are a property of the whole system. Hence the interest in analyzing biological systems at the network level.

On the basis of this comparative study we recommend the use of inference algorithms based on resampling and bootstrapping when correlations are used as indexes to measure metabolite-

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

metabolite associations. We also advocate for the use of data generated using dynamic models to test the performance of algorithms for network inference since they produce correlation patterns which are more similar to those observed in real metabolomics data.

SUPPORTING INFORMATION

The following supporting information is available free of charge at ACS website http://pubs.acs.org:

Supporting Material S1. The file contains:

ODE's for the AA dynamic model

Table S1 : Kinetic values & Reaction rates

Table S2 : All metabolites & their details

Table S3 : All Reactions & their details

Table S4 : AUCs for 500 simulations with no noise

Table S5 : AUCs for 500 simulations with 1% noise

Table S6 : AUCs for 500 simulations with 5% noise

Table S7 : AUCs for 500 simulations with 10% noise

Table S8 : AUCs for 500 simulations with 15% noise

Table S9 : AUCs for 50 simulations with no noise

Table S10 : AUCs for 50 simulations with 1% noise

Table S11 : AUCs for 50 simulations with 5% noise

Table S12 : AUCs for 50 simulations with 10% noise

Table S13 : AUCs for 50 simulations with 15% noise

Supporting Material S2. Arachidonic acid metabolic model in SBML format

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

AUTHOR INFORMATION

Corresponding Author

Edoardo Saccenti: esaccenti@gmail.com; Tel: +31 (0) 3174 82018; Fax: +31 (0) 3174 83829.

Author Contributions

S. J analyzed data; M.S.-D. and E.S. designed the study; E.S. supervised the work; All authors wrote the manuscript and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partly supported by the European Commission funded FP7 project INFECT (contract no. 305340).

References

(1) Rosato, A.; Tenori, L.; Cascante, M.; Carulla, P. R. D. A.; dos Santos, V. A. M.; Saccenti, E. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics* **2018**, *14*, 37.

(2) Hendrickx, D. M. *Network inference from time-resolved metabolomics data*; Universiteit van Amsterdam [Host], 2013.

(3) Camacho, D.; de la Fuente, A.; Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **2005**, *1*, 53–63.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(4) Szymanski, J.; Jozefczuk, S.; Nikoloski, Z.; Selbig, J.; Nikiforova, V.; Catchpole, G.; Willmitzer, L. Stability of metabolic correlations under changing environmental conditions in Escherichia coli–a systems approach. *PLoS One* **2009**, *4*, e7441.

(5) Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology* **2011**, *5*, 21.

(6) Lusczek, E. R.; Lexcen, D. R.; Witowski, N. E.; Mulier, K. E.; Beilman, G. Urinary metabolic network analysis in trauma, hemorrhagic shock, and resuscitation. *Metabolomics* **2013**, *9*, 223–235.

(7) Saccenti, E.; Suarez-Diez, M.; Luchinat, C.; Santucci, C.; Tenori, L. Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. *Journal of Proteome Research* **2015**, *14*, 1101–1111.

(8) Saccenti, E.; Menichetti, G.; Ghini, V.; Remondini, D.; Tenori, L.; Luchinat, C. Entropy-based network representation of the individual metabolic phenotype. *Journal of proteome research* **2016**, *15*, 3298–3307.

(9) Vignoli, A.; Tenori, L.; Luchinat, C.; Saccenti, E. Age and sex effects on plasma metabolite association networks in healthy subjects. *Journal of proteome research* **2017**, *17*, 97–107.

(10) Suarez-Diez, M.; Saccenti, E. Effects of Sample Size and Dimensionality on the Performance of Four Algorithms for Inference of Association Networks in Metabonomics. *Journal of Proteome Research* **2015**, *14*, 5119–5130.

(11) MATLAB, *version 9.02.0 (R2017a)*; The MathWorks Inc.: Natick, Massachusetts, 2017.

(12) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2013.

(13) Menten, L.; Michaelis, M. Die kinetik der invertinwirkung. *Biochem Z* **1913**, *49*, 333–369.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(14) Waage, P.; Guldberg, C. Studier over affiniteten. *Forhandlinger I Videnskabs-selskabet I Christiania* **1864**, *1*, 35–45.

(15) Meyer, P. E.; Kontos, K.; Lafitte, F.; Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *Eurasip Journal on Bioinformatics and Systems Biology* **2007**, *2007*.

(16) Meyer, P. E.; Lafitte, F.; Bontempi, G. Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* **2008**, *9*.

(17) Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **2010**, *5*.

(18) Haury, A.-C.; Mordelet, F.; Vera-Licona, P.; Vert, J.-P. TIGRESS: trustful inference of gene regulation using stability selection. *BMC systems biology* **2012**, *6*, 145.

(19) Buczynski, M. W. An integrated omics analysis of eicosanoid biology. *Journal of Lipid Research* **2009**, *50*, 1015 – 1505.

(20) Needleman, P.; Truk, J.; Jakschik, B. A.; Morrison, A. R.; Lefkowith, J. B. Arachidonic Acid Metabolism. *Annual Review of Biochemistry* **1986**, *55*, 69–102.

(21) Vainio, P.; Gupta, S.; Ketola, K.; Mirtti, T.; Mpindi, J. P.; Kohonen, P.; Fey, V.; Perälä, M.; Smit, F.; Verhaegh, G.; Schalken, J.; Alanen, K. A.; Kallioniemi, O.; Iljin, K. Arachidonic acid pathway members PLA2G7, HPGD, EPHX2, and CYP4F8 identified as putative novel therapeutic targets in prostate cancer. *American Journal of Pathology* **2011**, *178*, 525–536.

(22) Swainston, N. et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **2016**, *12*.

(23) Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research* **2018**, *46*, D608–D617.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(24) Wishart, D. S. et al. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research* **2013**, *41*.

(25) Wishart, D. S. et al. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research* **2009**, *37*.

(26) Wishart, D. S. et al. HMDB: The human metabolome database. *Nucleic Acids Research* **2007**, *35*.

(27) Kanehisa, M.; Goto, S. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000**, *28*, 27–30.

(28) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **2016**, *44*, D457–D462.

(29) Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **2017**, *45*, D353–D361.

(30) Degtyarenko, K.; De matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; Mcnaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research* **2008**, *36*.

(31) Degtyarenko, K.; Hastings, J.; de Matos, P.; Ennis, M. ChEBI: An open bioinformatics and cheminformatics resource. 2009.

(32) Hastings, J.; De Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research* **2013**, *41*.

(33) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research* **2016**, *44*, D1214–D1219.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(34) Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New perspectives and new tools in BRENDA. *Nucleic Acids Research* **2017**, *45*, D380–D388.

(35) Wittig, U.; Kania, R.; Golebiewski, M.; Rey, M.; Shi, L.; Jong, L.; Algaa, E.; Weidemann, A.; Sauer-Danzwith, H.; Mir, S.; Krebs, O.; Bittkowski, M.; Wetsch, E.; Rojas, I.; Müller, W. SABIO-RK–database for biochemical reaction kinetics. *Nucleic acids research* **2012**, *40*, D790–6.

(36) Savchenko, T.; Walley, J. W.; Chehab, E. W.; Xiao, Y.; Kaspi, R.; Pye, M. F.; Mohamed, M. E.; Lazarus, C. M.; Bostock, R. M.; Dehesh, K. Arachidonic Acid: An Evolutionarily Conserved Signaling Molecule Modulates Plant Stress Signaling Networks. *The Plant Cell* **2010**, *22*, 3193–3205.

(37) Yui, K.; Imataka, G.; Nakamura, H.; Ohara, N.; Naito, Y. Eicosanoids Derived From Arachidonic Acid and Their Family Prostaglandins and Cyclooxygenase in Psychiatric Disorders. *Current neuropharmacology* **2015**, *13*, 776–785.

(38) Akimova, E. V. Prostacyclin and thromboxane A2 levels in children and adolescents with an inherited predisposition to coronary heart disease: a family study. *Coronary artery disease* **1994**, *5*, 761–5.

(39) Psychogios, N. et al. The human serum metabolome. *PLoS ONE* **2011**, *6*.

(40) Quehenberger, O. et al. Lipidomics reveals a remarkable diversity of lipids in human plasma,. *Journal of Lipid Research* **2010**, *51*, 3299–3305.

(41) Frolich, J. C.; Sweetman, B. J.; Carr, K.; Hollifield, J. W.; Oates, J. A. Assessment of the levels of PGA2in human plasma by gas chromatography-mass spectrometry. *Prostaglandins* **1975**, *10*, 185–195.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(42) Gordon-Wright, A. P.; Elder, M. G. The systemic absorption from the vagina of prostaglandin E2 administered for the induction of labour. *Prostaglandins* **1979**, *18*, 153–160.

(43) Walenga, R. W.; Boone, S.; Stuart, M. J. Analysis of blood hete levels by selected ion monitoring with ricinoleic acid as the internal standard. *Prostaglandins* **1987**, *34*, 733–748.

(44) Hall, W. L.; Sanders, K. a.; Sanders, T. a. B.; Chowienczyk, P. J. A high-fat meal enriched with eicosapentaenoic acid reduces postprandial arterial stiffness measured by digital volume pulse analysis in healthy men. *The Journal of nutrition* **2008**, *138*, 287–291.

(45) Dreiseitel, A.; Korte, G.; Schreier, P.; Oehme, A.; Locher, S.; Hajak, G.; Sand, P. G. SPhospholipase A2 is inhibited by anthocyanidins. *Journal of Neural Transmission* **2009**, *116*, 1071–1077.

(46) Watanabe, K.; Urade, Y.; Mäder, M.; Murphy, C.; Hayaishi, O. Identification of β-trace as prostaglandin D synthase. *Biochemical and Biophysical Research Communications* **1994**, *203*, 1110–1116.

(47) Rogge, C. E.; Ho, B.; Liu, W.; Kulmacz, R. J.; Tsai, A. L. Role of Tyr348 in Tyr385 radical dynamics and cyclooxygenase inhibitor interactions in prostaglandin H synthase-2. *Biochemistry* **2006**, *45*, 523–532.

(48) Torres Pazmiño, D. E.; Winkler, M.; Glieder, A.; Fraaije, M. W. Monooxygenases as biocatalysts: Classification, mechanistic aspects and biotechnological applications. 2010.

(49) Jones, D. A.; Fitzpatrick, F. A. "Suicide" inactivation of thromboxane A2 synthase: Characteristics of mechanism-based inactivation with isolated enzyme and intact platelets. *Journal of Biological Chemistry* **1990**, *265*, 20166–20171.

(50) Wada, M.; Yokoyama, C.; Hatae, T.; Shimonishi, M.; Nakamura, M.; Imai, Y.; Ullrich, V.; Tanabe, T. Purification and characterization of recombinant human prostacyclin synthase. *Journal of Biochemistry* **2004**, *135*, 455–463.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(51) Hamza, A.; Tong, M.; Abdulhameed, M. D. M.; Liu, J.; Goren, A. C.; Tai, H. H.; Zhan, C. G. Understanding microscopic binding of human microsomal prostaglandin e synthase-1 (mPGES-1) trimer with substrate PGH2 and cofactor GSH: Insights from computational alanine scanning and site-directed mutagenesis. *Journal of Physical Chemistry B* **2010**, *114*, 5605–5616.

(52) Kabututu, Z.; Manin, M.; Pointud, J. C.; Maruyama, T.; Nagata, N.; Lambert, S.; Lefrançois-Martinez, A. M.; Martinez, A.; Urade, Y. Prostaglandin F2αsynthase activities of aldo-keto reductase 1B1, 1B3 and 1B7. *Journal of Biochemistry* **2009**, *145*, 161–168.

(53) Jung, A.; Schlegel, W.; Jackisch, R.; Friedrich, E. J.; Wendel, A.; Ruckrich, M. F. [15-Hydroxyprostaglandin dehydrogenase from human placenta. 1. Isolation and characterization]. *Hoppe Seylers.Z.Physiol Chem.* **1975**, *356*, 787–798.

(54) Moschini, R.; Rotondo, R.; Renzone, G.; Balestri, F.; Cappiello, M.; Scaloni, A.; Mura, U.; Del-Corso, A. Kinetic features of carbonyl reductase 1 acting on glutathionylated aldehydes. *Chemico-Biological Interactions* **2017**, *276*, 127–132.

(55) Jacquot, C.; Wecksler, A. T.; McGinley, C. M.; Segraves, E. N.; Holman, T. R.; Van Der Donk, W. A. Isotope sensitive branching and kinetic isotope effects in the reaction of deuterated arachidonic acids with human 12- and 15-lipoxygenases. *Biochemistry* **2008**, *47*, 7295–7303.

(56) Hargreaves, K. M.; Clandinin, M. T. Phosphatidylethanolamine methyltransferase: evidence for influence of diet fat on selectivity of substrate for methylation in rat brain synaptic plasma membranes. *Biochimica et Biophysica Acta (BBA)/Lipids and Lipid Metabolism* **1987**, *918*, 97–105.

(57) Yoichi, K.; Akihiko, H.; Toshiyuki, A.; Hiroshi, K. Role of stearoyl-CoA desaturase and 1-acylglycerophosphorylcholine acyltransferase in the regulation of the acyl composition of

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

phosphatidylcholine in rat liver. *Biochimica et Biophysica Acta (BBA)/Lipids and Lipid Metabolism* **1985**, *837*, 222–229.

(58) Itasaka, O.; Kosuga, M.; Okayama, M.; Hori, T. Characterization of a novel ceramide octasaccharide isolated from whole tissue of a fresh-water bivalve, Corbicula sandai. *Biochimica et Biophysica Acta (BBA)/Lipids and Lipid Metabolism* **1983**, *750*, 440–446.

(59) Shimizu, T.; Rådmark, O.; Samuelsson, B. Enzyme with dual lipoxygenase activities catalyzes leukotriene A4 synthesis from arachidonic acid. *Proceedings of the National Academy of Sciences of the United States of America* **1984**, *81*, 689–693.

(60) Carmagnol, F.; Sinet, P. M.; Jerome, H. Selenium-dependent and non-selenium-dependent glutathione peroxidases in human tissue extracts. *BBA - General Subjects* **1983**, *759*, 49–57.

(61) Dai-Eun, S.; Kang, J. B.; Hong Dae Shin, 15-hydroxyeicosatetraenoic acid dehydrogenase activity in microsomal fraction of mouse liver homogenate. *Biochemical and Biophysical Research Communications* **1988**, *156*, 524–529.

(62) Pongdontri, P.; Hills, M. Characterization of a novel plant acyl-CoA synthetase that is expressed in lipogenic tissues of Brassica napus L. *Plant Molecular Biology* **2001**, *47*, 717–726.

(63) Shimizu, T.; Izumi, T.; Honda, Z.; Seyama, Y.; Kurachi, Y.; Sugimoto, T. Biosynthesis and functions of leukotriene C4. *Adv Prostaglandin Thromboxane Leukot Res* **1990**, *20*, 46–53.

(64) Habib, G. M.; Shi, Z. Z.; Cuevas, A. A.; Lieberman, M. W. Identification of two additional members of the membrane-bound dipeptidase family. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **2003**, *17*, 1313–1315.

(65) Wickham, S.; West, M. B.; Cook, P. F.; Hanigan, M. H. Gamma-glutamyl compounds:

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Substrate specificity of gamma-glutamyl transpeptidase enzymes. *Analytical Biochemistry* **2011**, *414*, 208–214.

(66) Orning, L.; Krivi, G.; Fitzpatrick, F. A. Leukotriene A4 hydrolase. Inhibition by bestatin and intrinsic aminopeptidase activity establish its functional resemblance to metallohydrolase enzymes. *Journal of Biological Chemistry* **1991**, *266*, 1375–1378.

(67) Cho, H.; Tai, H. H. Threonine 11 of human NAD+dependent 15-hydroxyprostaglandin dehydrogenase may interact with NAD+during catalysis. *Prostaglandins Leukotrienes and Essential Fatty Acids* **2002**, *66*, 505–509.

(68) Cho, H.; Tai, H. H. Inhibition of NAD+-dependent 15-hydroxyprostaglandin dehydrogenase (15-PGDH) by cyclooxygenase inhibitors and chemopreventive agents. *Prostaglandins Leukotrienes and Essential Fatty Acids* **2002**, *67*, 461–465.

(69) Aleem, A. M.; Jankun, J.; Dignam, J. D.; Walther, M.; Kühn, H.; Svergun, D. I.; Skrzypczak-Jankun, E. Human Platelet 12-Lipoxygenase, New Findings about Its Activity, Membrane Binding and Low-resolution Structure. *Journal of Molecular Biology* **2008**, *376*, 193–209.

(70) Strömberg-Kull, F.; Haeggström, J. Z. Purification and characterization of leukotriene A4hydrolase from Xenopus laevis oocytes. *FEBS Letters* **1998**, *433*, 219–222.

(71) Morisseau, C.; Wecksler, A. T.; Deng, C.; Dong, H.; Yang, J.; Lee, K. S. S.; Kodani, S. D.; Hammock, B. D. Effect of soluble epoxide hydrolase polymorphism on substrate and inhibitor selectivity and dimer formation. *Journal of Lipid Research* **2014**, *55*, 1131–1138.

(72) Torres, V.; Barra, L.; Garcés, F.; Ordenes, K.; Leal-Ortiz, S.; Garner, C. C.; Fernandez, F.; Zamorano, P. A bicistronic lentiviral vector based on the 1D/2A sequence of foot-and-mouth disease virus expresses proteins stoichiometrically. *Journal of biotechnology* **2010**, *146*, 138–42.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(73) Nottbohm, C. T.; Sopher, R.; Heilemann, M.; Sauer, M.; Gölzhäuser, A. Fluorescently labeled 1 nm thin nanomembranes. *Journal of biotechnology* **2010**, *149*, 267–71.

(74) Zehentgruber, D.; Dragan, C. A.; Bureik, M.; Lütz, S. Challenges of steroid biotransformation with human cytochrome P450 monooxygenase CYP21 using resting cells of recombinant Schizosaccharomyces pombe. *Journal of Biotechnology* **2010**, *146*, 179–185.

(75) Zhang, H.; Wan, H.; Song, L.; Jiang, H.; Wang, H.; Qiao, C. Development of an autofluorescent Pseudomonas nitroreducens with dehydrochlorinase activity for efficient mineralization of γ-hexachlorocyclohexane (γ-HCH). *Journal of Biotechnology* **2010**, *146*, 114–119.

(76) Hoops, S.; Sahle, S.; Gauges, R.; Lee, C.; Pahle, J.; Simus, N.; Singhal, M.; Xu, L.; Mendes, P.; Kummer, U. COPASI—a complex pathway simulator. *Bioinformatics* **2006**, *22*, 3067–3074.

(77) Raetz, C. R. H. et al. Kdo ₂ -Lipid A of Escherichia coli , a defined endotoxin that activates macrophages via TLR-4. *Journal of Lipid Research* **2006**,

(78) Dennis, E. A. et al. A mouse macrophage lipidome. *Journal of Biological Chemistry* **2010**,

(79) Faith, J. J.; Hayete, B.; Thaden, J. T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J. J.; Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology* **2007**, *5*, 0054–0066.

(80) Margolin, A. A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R. D.; Califano, A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **2006**, *7*.

(81) Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nature Methods* **2012**, *9*, 796–804.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(82) Pearson, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **1895**, *58*, 240–242.

(83) Spearman, C. The proof and measurement of association between two things. *The American journal of psychology* **1904**, *15*, 72–101.

(84) Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **2005**, *27*, 1226–1238.

(85) Meyer, P.; Marbach, D.; Roy, S.; Kellis, M. Information-Theoretic Inference of Gene Networks Using Backward Elimination. BioComp. 2010; pp 700–705.

(86) Ruyssinck, J.; Huynh-Thu, V. A.; Geurts, P.; Dhaene, T.; Demeester, P.; Saeys, Y. NIMEFI: Gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS ONE* **2014**, *9*.

(87) Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *The Annals of statistics* **2004**, *32*, 407–499.

(88) Bach, F. R. Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of the 25th international conference on Machine learning. 2008; pp 33–40.

(89) Meinshausen, N.; Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2010**, *72*, 417–473.

(90) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research* **2012**, *41*, D781–D786.

(91) Magnusson, P. K.; Almqvist, C.; Rahman, I.; Ganna, A.; Viktorin, A.; Walum, H.; Halldner, L.; Lundström, S.; Ullén, F.; Långström, N. The Swedish Twin Registry: establishment

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of a biobank and other recent developments. *Twin Research and Human Genetics* **2013**, *16*, 317–329.

(92) RPython Core Team, Python: A dynamic, open source programming language. Python Software Foundation., 2015.

(93) Meyer, P. E.; Lafitte, F.; Bontempi, G. Minet: an open source R/Bioconductor package for mutual information based network inference. *BMC bioinformatics* **2008**, *9*.

(94) Fukushima, A.; Kusano, M.; Redestig, H.; Arita, M.; Saito, K. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC systems biology* **2011**, *5*, 1.

(95) Adolph, S. C.; Hardin, J. S. Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Functional Ecology* **2007**, *21*, 178–184.

(96) Camacho, J.; Padilla, P.; Díaz-Verdejo, J.; Smith, K.; Lovett, D. Least-squares approximation of a space distribution for a given covariance and latent sub-space. *Chemometrics and Intelligent Laboratory Systems* **2011**, *105*, 171–180.

(97) Camacho, J. On the generation of random multivariate data. *Chemometrics and Intelligent Laboratory Systems* **2017**, *160*, 40–51.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURES

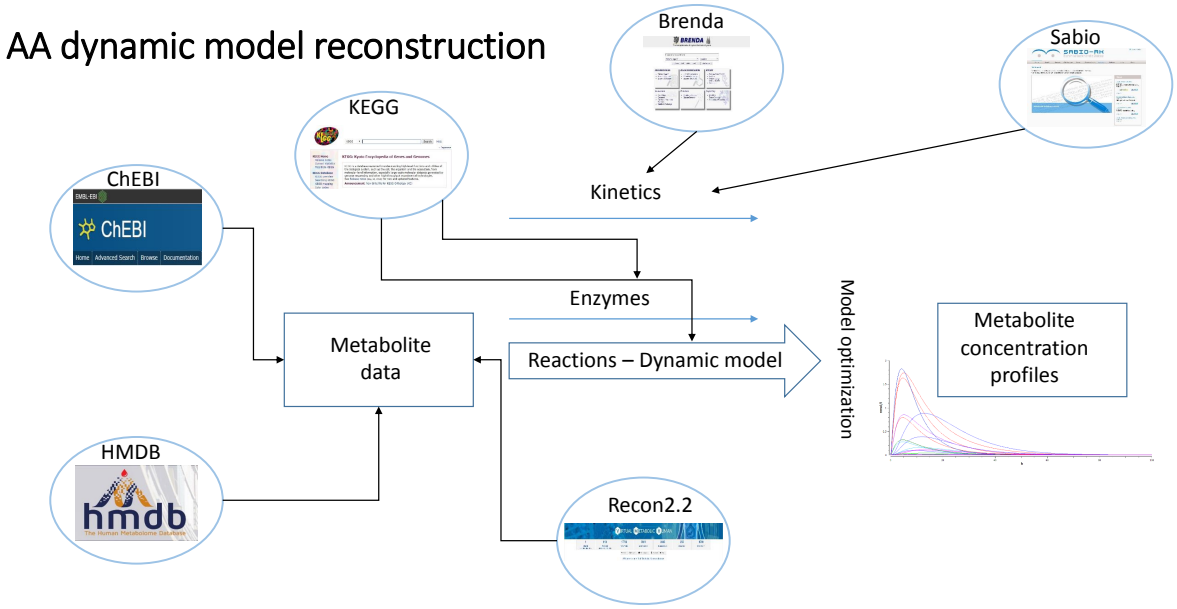

Figure 1: Overview of the strategy for the reconstruction of the the dynamic model for arachidonic acid (AA) metabolism. The databases used are indicated together with the type of information retrieved.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

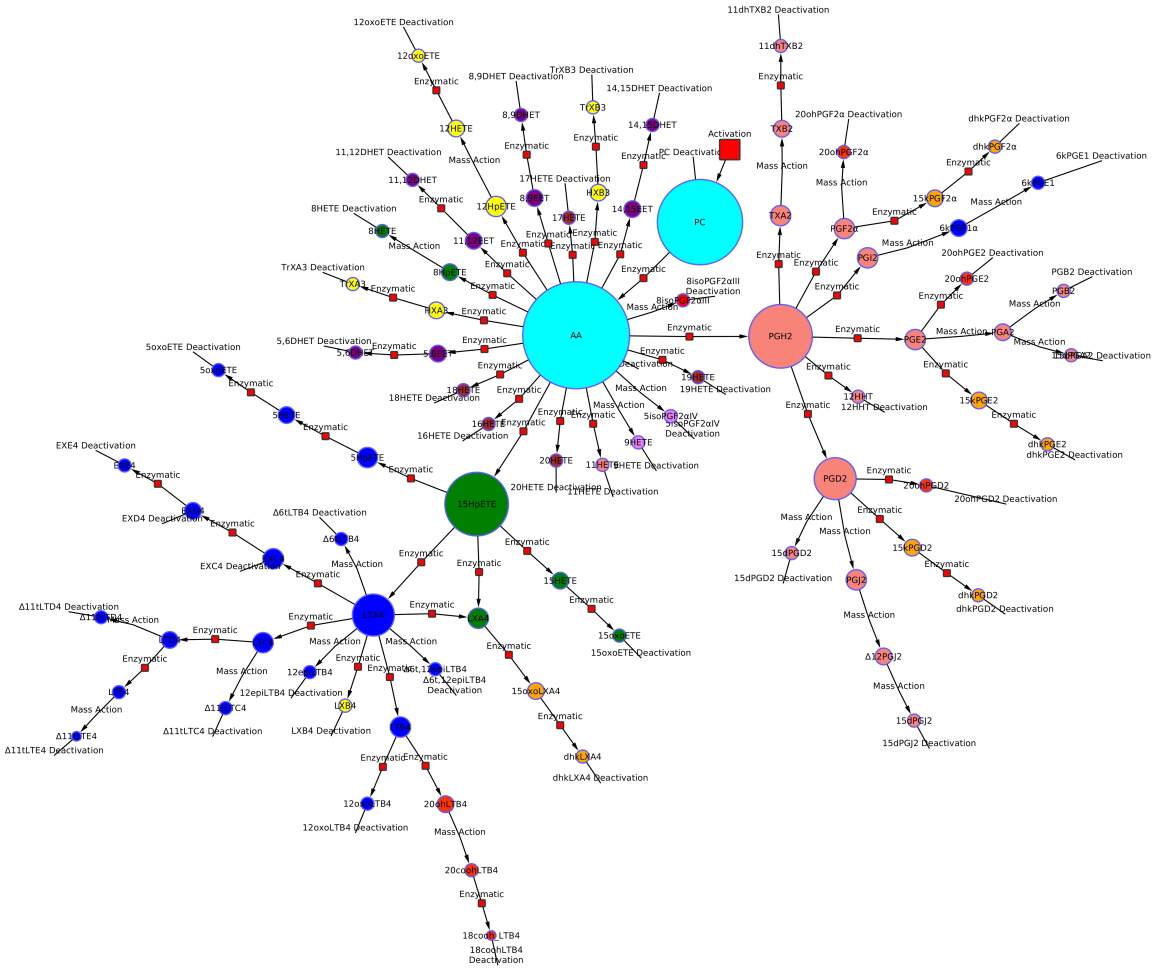

Figure 2: Graphical representation of the AA metabolic pathway. The enzymatic or non-enzymatic (mass action) nature of the reaction kinetic is indicated. The model comprises 131 reactions and 83 metabolites. The ODEs governing the dynamic model, metabolite desscription and reaction details are given in the supplementary material.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

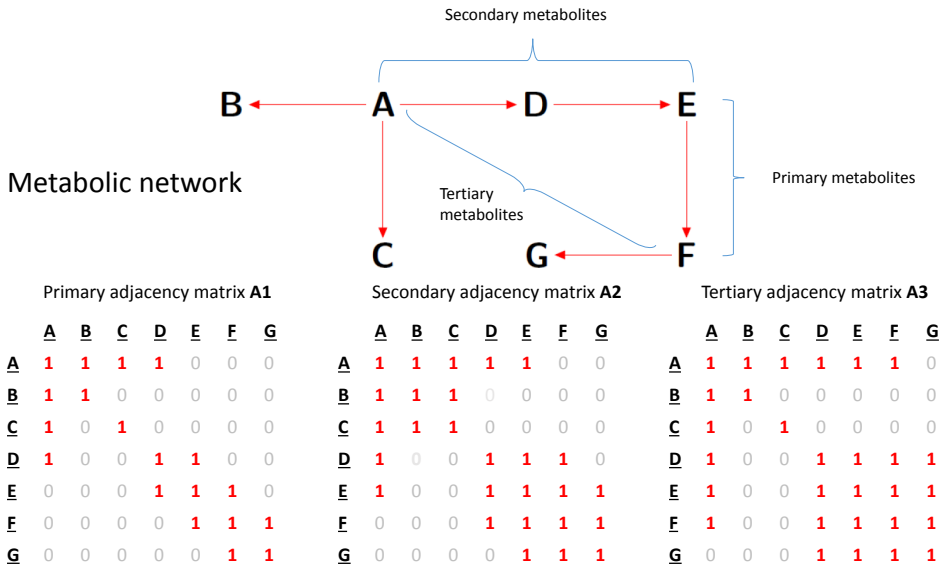

Figure 3: Schematic overivew of the defintion of the adjacency matrices for a metabolic pathway consisting of five metabolites (A, B, C, D, E, F and G) and six reaction (represented by arrows →). Example of primary, secondary and tertiary metabolites are indicated together the corresponding primary (**A1**), secondary (**A2**) and tertiary (**A3**) adjacency (connectivity) matrices. Note that the definition of the matrices takes into account reaction directionality. In this example, **B** and **D** can not be inter-coverted, and as a result they are not connected in any of the adjacency matrices.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

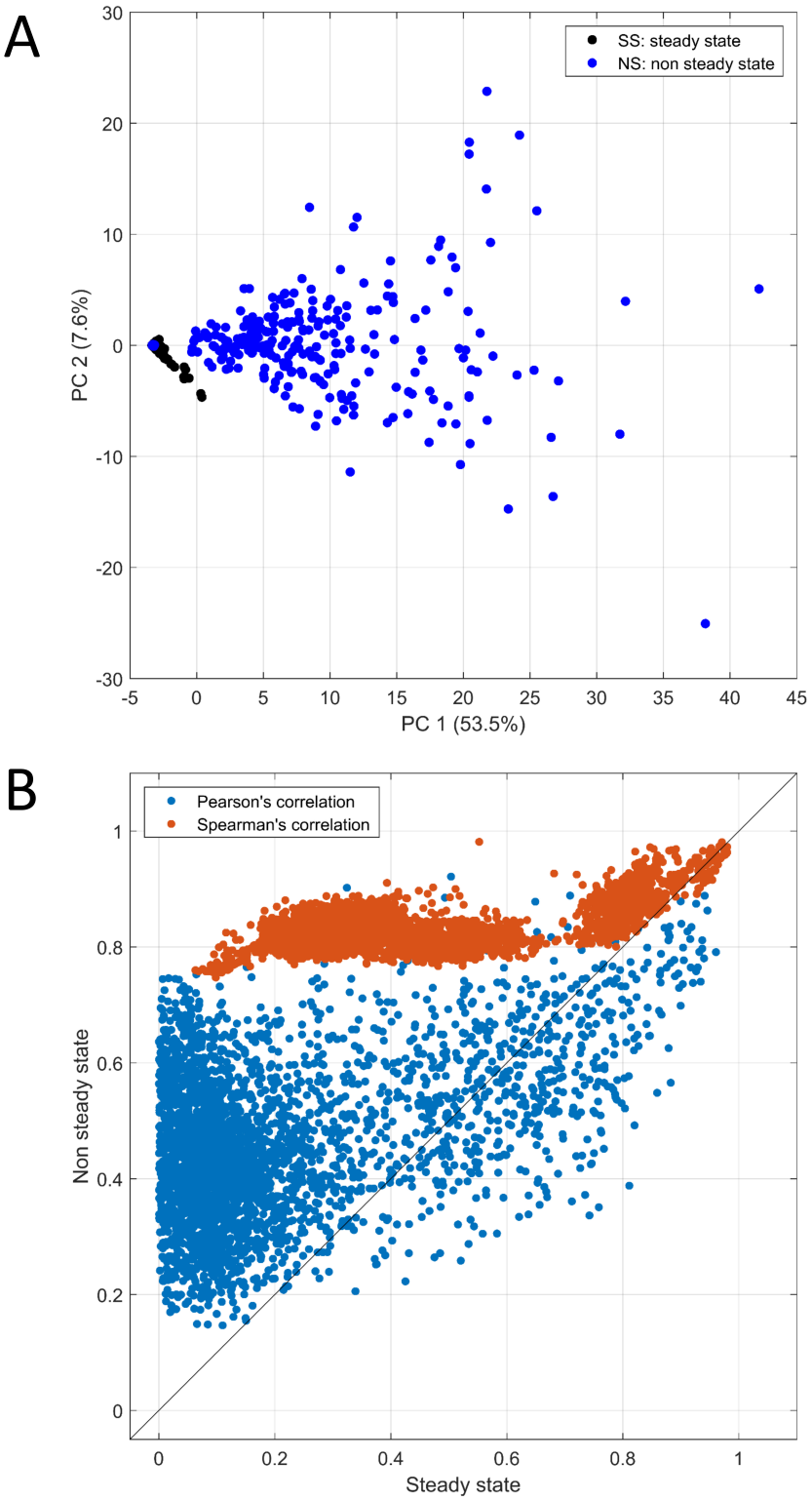

Figure 4: A) Scatter plot of the first two principal components of a PCA model on the data set formed by 500 individual metabolic profiles measure in the steady state (SS) and 500 profiles measured in non-steady state (NS). B) Scatter plot of the $(83^2 - 83)/2 = 3403$ pairwise correlations between metabolites for both SS and NS data calculated using both Pearson's and Spearman's indexes.

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

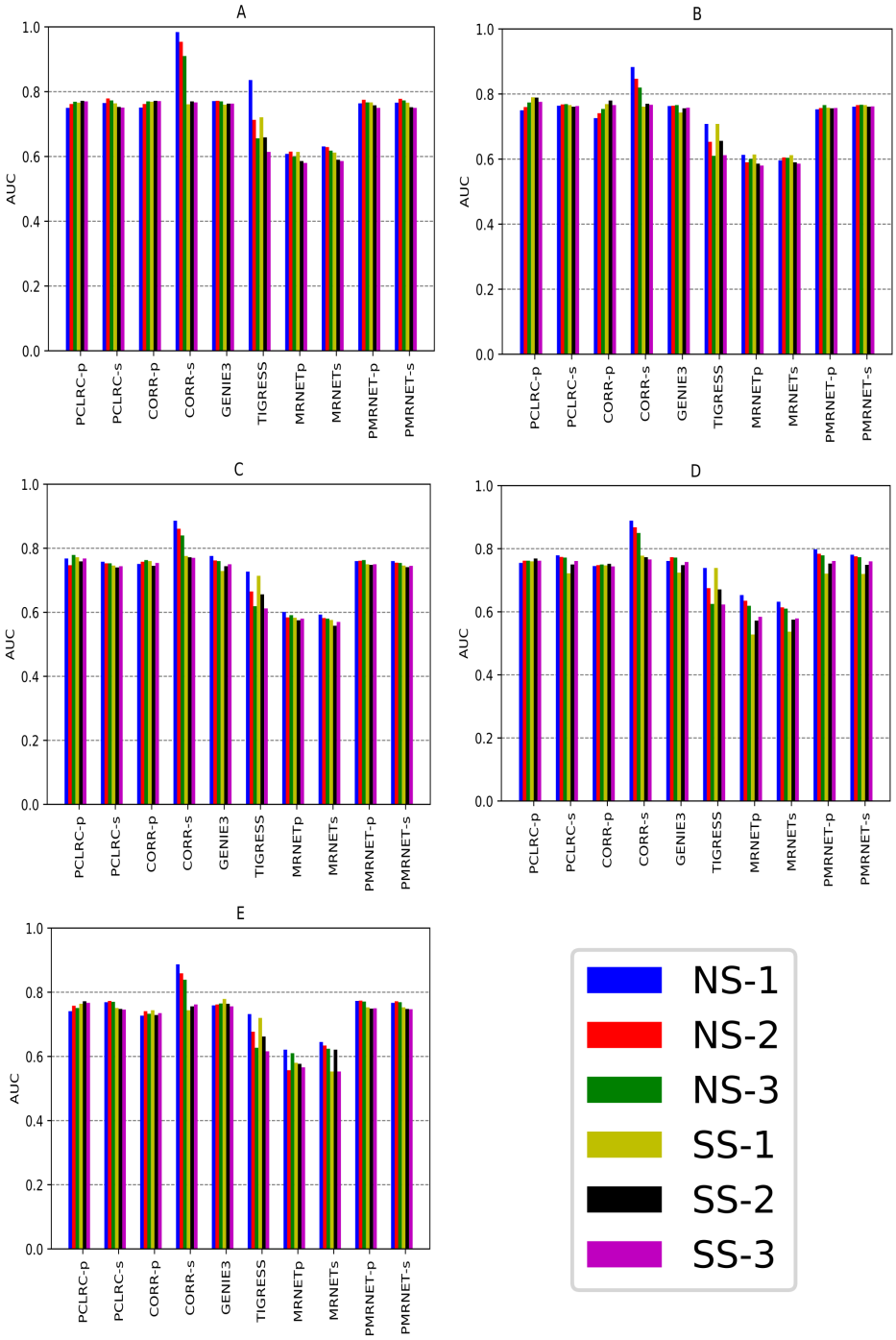

Figure 5: Performance of network reconstruction methods (AUC values) for sample size $n = 500$ A) no noise data; B) 1% SNR; C) 5% SNR; D) 10% SNR; D) 15% SNR. In color legend, SS and NS indicate that metabolites' profiles obtained at steady and non-steady state conditions have been used, whereas indices 1,2 and 3 refer to the comparison with the adjacency matrices **A1**, **A2**, and **A3** respectively

bioRxiv preprint doi: https://doi.org/10.1101/460519; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

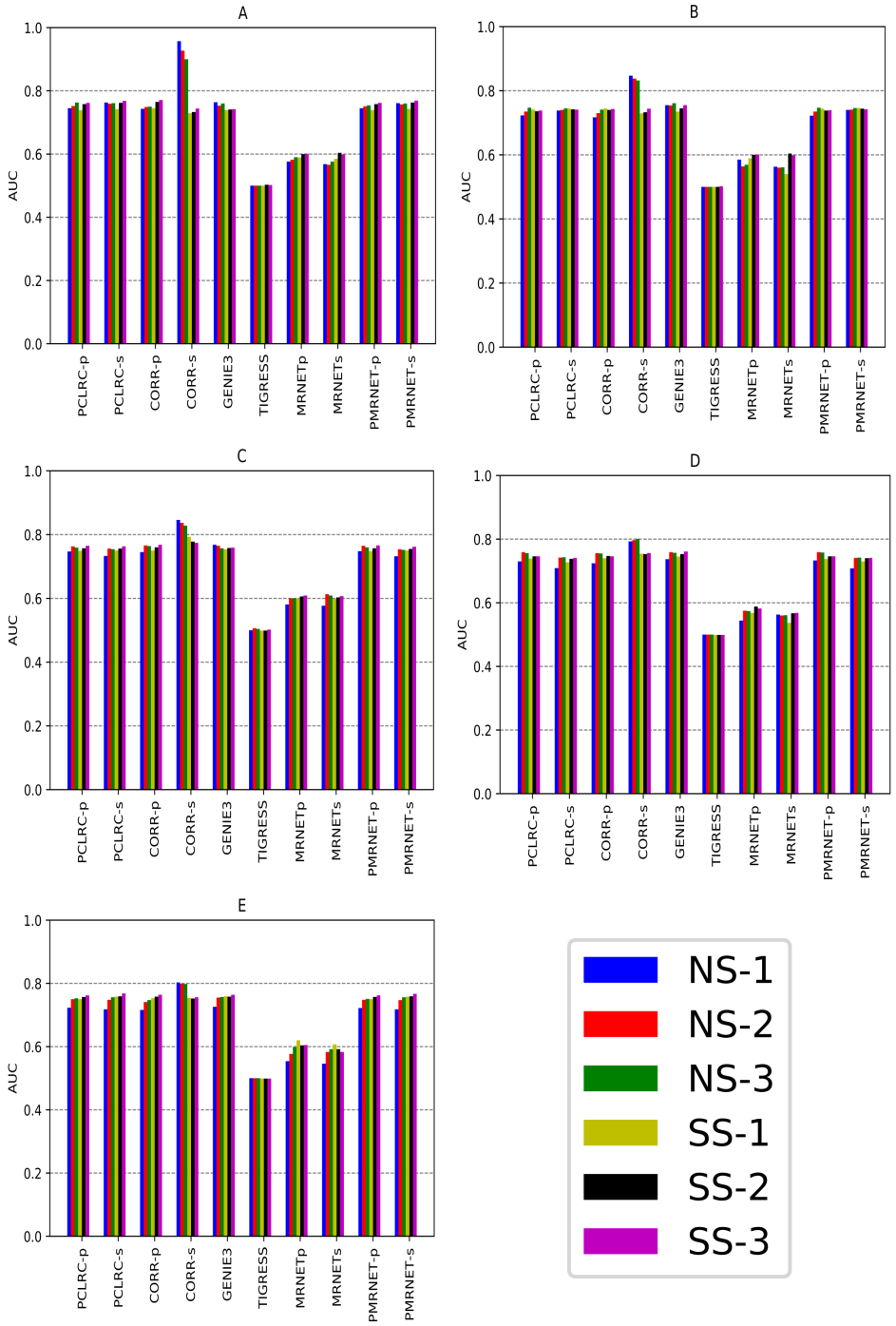

Figure 6: Performance of network reconstruction methods (AUC values) for sample size $n = 50$ A) no noise data; B) 1% SNR; C) 5% SNR; D) 10% SNR; D) 15% SNR.In color legend, SS and NS indicate that metabolites' profiles obtained at steady and non-steady state conditions have been used, whereas indices 1,2 and 3 refer to the comparison with the adjacency matrices **A1**, **A2**, and **A3** respectively