BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 2 2011, pages 232–237
doi:10.1093/bioinformatics/btq629

Systems biology Advance Access publication November 12, 2010

Inference of gene networks—application to *Bifidobacterium*

Darong Lai[1,2,†], Xinyi Yang[1,†], Gang Wu[1], Yuanhua Liu[1] and Christine Nardini[1,*]

[1]Key Laboratory of Computational Biology, Chinese Academy of Sciences Max Planck Institute Partner Institute for Computational Biology (CAS-MPG), Yue Yang Road 320, Shanghai and [2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The reliable and reproducible identification of gene interaction networks represents one of the grand challenges of both modern molecular biology and computational sciences. Approaches based on careful collection of literature data and network topological analysis, applied to unicellular organisms, have proven to offer results applicable to medical therapies. However, when little *a priori* knowledge is available, other approaches, not relying so strongly on previous literature, must be used. We propose here a novel algorithm (based on ordinary differential equations) able to infer the interactions occurring among genes, starting from gene expression steady state data.

Results: The algorithm was first validated on synthetic and real benchmarks. It was then applied to the reconstruction of the core of the amino acids metabolism in *Bifidobacterium longum*, an essential, yet poorly known player in the human gut intestinal microbiome, known to be related to the onset of important diseases, such as metabolic syndromes. Our results show how computational approaches can offer effective tools for applications with the identification of potential new biological information.

Availability: The software is available at *www.bioconductor.org* and at *www.picb.ac.cn/ClinicalGenomicNTW/temp2.html*.

Contact: christine@picb.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 19, 2010; revised on October 18, 2010; accepted on November 7, 2010

1 INTRODUCTION

Recent advances in (bio)technology have provided researchers with several powerful data platforms, able to collect biological molecules' activity at the genome-wide level (Guiducci and Nardini, 2008). Nowadays, such an abundance of data potentially allows reconstruction of the complex interplay among molecules, responsible for diverse cellular functions. Network reconstruction, which is a broad area in modern biology research, focuses on two main directions (Bansal *et al.*, 2007): reconstruction of *physical* networks, where the relationships inferred represent actual connections among molecules of a diverse nature (i.e. metabolic networks), and reconstruction of interactions called *influence* networks. Physical networks allow to infer precise information, but require *a priori* knowledge to support initial assumptions. An interesting and successful example of physical network reconstruction application is given in Chavali *et al.* (2008), where the authors were able to describe in detail several mechanisms of action occurring in *Leishmania major*, a protozoan that causes cutaneous leishmaniasis in mammalian hosts. The network was reconstructed with extensive data collection, manual curation and finally automated topological analysis. Very interestingly, the whole approach generated hypotheses that were then experimentally validated to identify therapeutic targets. Therefore, the application of computational tools to simpler organisms that are in tight relation with the evolving health state of more complex organisms, like humans, can bear fruitful results. Conversely to physical networks, influence networks infer generic connections among genes, but require very little preliminary evidence and can therefore be applied to systems for which little *a priori* knowledge is available.

In this work, we explore a gene network related to *Bifidobacterium*, a crucial player in the gut intestinal (GI) microbiome, the importance of which is becoming more and more central in the onset of several types of diseases, despite little knowledge available today. The GI tract with its microbiome, in fact, is an important organ responsible for digestion, absorption and metabolism of dietary nutrients. It contributes to 9–12% of whole-body protein synthesis and is the most important route of entry for foreign antigens (i.e. natural toxins, invading pathogens, (Wang *et al.*, 2009). In particular, the human gut is the home of a complex and vast array of bacterial cells, which are estimated to be 10 times more numerous than the total number of cells in the human body (Palmer *et al.*, 2007), and with a collective genome (*metagenome*) that contains at least 100 times as many genes as our own genome (Gill *et al.*, 2006). Comparative metagenomics has uncovered functional attributes of the microbiome involved in the metabolism of glycans and amino acids, the production of methane and the biosynthesis of vitamins (Gill *et al.*, 2006). In particular, the amino acid (AA) metabolism appears to be a core functional group, highly consistent across different human microbiome samples. Processes performed by the microbiome are strictly intertwined with the host metabolism, to the point that life without this contribution may not be possible. This concept is summarized in the definition of *sym-xenobiotic* metabolites formed by both gut microbial and host metabolism that cannot be synthesized by either in isolation, and which forms no part of the energy metabolism, or biosynthetic machinery of either system (Nicholson *et al.*, 2005). In particular, *Bifidobacterium longum* (a subspecies of *Bifidobacteria*) has been reported to be crucial in a number of functions and is protective against the onset of obesity, prodrome of metabolic syndromes [see only as an example (Zhang *et al.*, 2010) and our recent validation

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

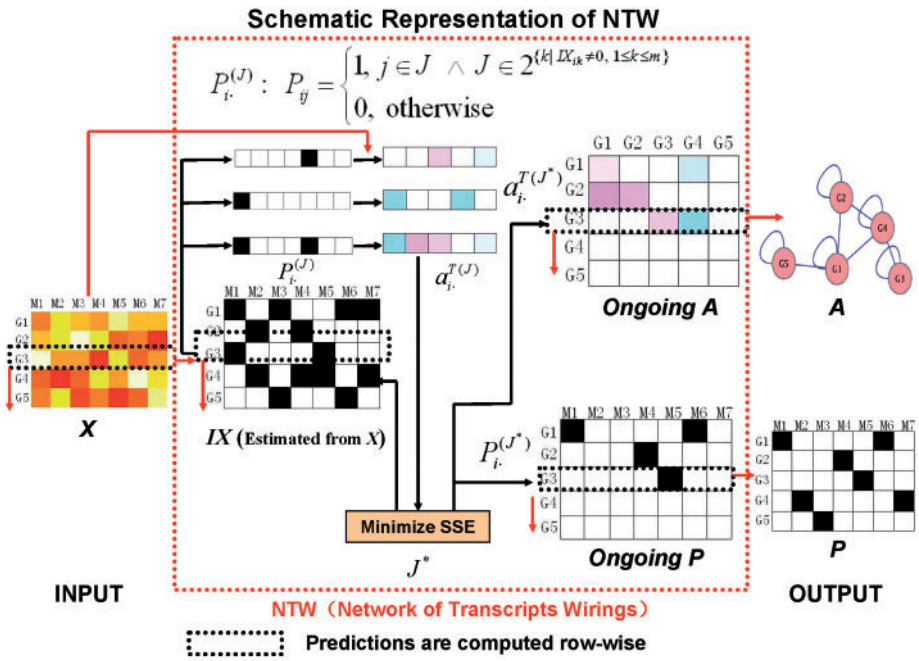

Fig. 1. Schematic illustration of the main procedure of NTW. NTW first evaluates the most likely perturbation triggers IX from X. In the main loop (large red dashed box), it estimates row wise all candidate perturbation vectors $P_{i\cdot}^{(J)}$ for gene i and produces the corresponding gene relations $a_{i\cdot}^{T(J)}$ between gene i and other genes via multiple regression. The relations $a_{i\cdot}^{T(J^*)}$ and the perturbation vector $P_{i\cdot}^{(J^*)}$ for gene i are optimal under the SSE, i.e. $J_i^* = \underset{J}{\operatorname{argmin}} \| X_i^{(J)} - X_i \|^2$ and $X_i^{(J)} = (tmpA^{-1} * tmpP)_i$, $1 \leq i \leq n$. See Section 2.2 and the Supplementary Materials for more details on NTW.

in (Liu *et al.*, 2010)]. Given these fact, we have chosen to apply our gene network reconstruction method to the AA metabolism of *B.longum* NCC2705, a particular strain of *B.longum*, isolated from infant feces, and which genome sequence was the first completed within the *Bifidobacterium* genus (Schell *et al.*, 2002) and revised in 2005 (GenBank database accession number no. AE014295).

Our method, network of transcripts wirings (NTW), is an ordinary differential equations (ODE) based model. Despite prior work showing good performances (Bansal *et al.*, 2007), the superiority of ODE models is far from assessed. However, ODE-based algorithms frame the reverse engineering problem with the quite flexible identification of a function that describes the variation of gene expression vector $\underline{x}$ over time $\underline{x}' = f(\underline{x}, \underline{p})$, where $\underline{x}$ represents the expressions of the n genes involved in the network across m experiments, f is the function that models how the transcriptional perturbations $\underline{p}$ lead to the new equilibria in $\underline{x}$.

As pointed out in (Marbach *et al.*, 2010), an extensive comparison with previous work is not easy, due to the abundance of algorithms produced and the partial availability of data and software, and would represent a work in itself. Nevertheless, we describe here some recent or popular approaches we will compare to. NIR (Gardner *et al.*, 2003) uses a linearized version of this equation, leading to the matricial form of the linearized equation $AX = -P$, where A represents the interaction network (adjacency matrix, $n \times n$), X the steady state expression values (expression matrix, $n \times m$) and P the transcriptional perturbations (a matrix of the same size as X, $n \times m$). Under the assumption that transcriptional perturbations triggering the new equilibria are known, NIR assumes that only a certain number of genes have interactions and minimizes the sum of squared errors (SSEs) to get the approximation of gene network matrix. Recently, a parallelized version of NIR has been published (Gregoretti *et al.*, 2010). NIRest (Lauria *et al.*, 2009) builds upon the original NIR and extends its use to cases in which the perturbations are not known, assuming only a few genes are perturbed in each experiment. NIRest first produces a rough estimation of the gene interaction matrix A, it then infers P based on the estimated A and X.

Based on the same ODE model, other algorithms have been designed. For example, in Julius *et al.* (2009) the smallest possible genetic network is obtained from the minimization of a convex cost function over a convex set of feasible solutions. Like NIR, this algorithm requires the transcriptional perturbations to be known.

Combinatorial perturbation-based interaction analysis (CoPIA; Nelander *et al.*, 2008) uses a non-linear function f and minimizes two errors: an SSE to measure the difference between the model predicted expression values and the corresponding observational values, and a structure error, which takes into account the number of interaction in the predicted gene network.

Finally in Madar *et al.* (2010), a mixed model is adopted, using information theory (mix-CLR method, Faith *et al.*, 2007) to extract preliminary knowledge and an ODE model (Inferelator, Efron *et al.*, 2004; Tibshirani, 1996) to refine it. We had also considered this type of approach and tested it using ARACNe (Margolin *et al.*, 2006) and NIR (Gardner *et al.*, 2003) with little success (data not published).

Our method, NTW, is part of this large family of ODE methods, which make use, like NIR and NIRest, of a linearized version of the original equation ($AX = -P$). The major advance offered by our approach compared with the seminal work in Gardner *et al.* (2003)

and its recent improvement in Gregoretti *et al.* (2010) lies in the ability to estimate A when P is unknown. Figure 1 schematically shows the basic principles on which the algorithm is based multiple regression is used to minimize both the error on the inferred A and P, see details in the Section 2).

2 METHODS

2.1 Culture conditions and real-time RT–PCR

Bifidobacteria longum NCC2705 was stored in Lactobacilli MRS Broth (BD Diagnostic Systems, Sparks, MD, USA) plus 50% glycerol at −80°C. Twenty microliter stock solution were transferred into 3 ml MRS broth supplemented with 0.05% L-cysteine (MRSC; Merck KGaA, Darmstadt, Germany) and cultured anaerobically using MGC anaerobic system (Mitsubishi Gas Chemical Company) at 37°C for 48 h. Then 5 μg/ml of puromycin, streptomycin, tetracyclines, chloramphenicol, cycloheximide and erythromycin were added into each MRSC tube separately as perturbations of the amino acid metabolism. *Bifidobacteria longum* NCC2705 growing in MRSC without any perturbations was used as control. After 40 h, *B.longum* NCC2705 was harvested for RNA isolation.

Total RNA was extracted using RNeasy Protect Bacteria Mini Kit (QIAGEN) following the manufacturer's instructions and then treated with RNase-Free DNase Set (QIAGEN) before cDNA synthesis. Reverse transcription (RT) was performed using PrimeScript™ RT reagent Kit (TaKaRa, Kyoto, Japan) following the manufacturer's instructions. Real-time PCR was performed using SYBR PrimeScript™ RT-PCR Kit II (TaKaRa, Kyoto, Japan) on the LightCycler 480 (Roche). 16S rDNA transcript (Matsuki *et al.*, 1999) was used as an endogenous control. The conditions of real-time polymerase chain reaction (PCR) were 95°C for 30 s, 40 cycles of 95°C for 5 s and 60°C for 20 s followed by the condition for melt curve analysis: 95°C for 0 s, 60°C for 15 s and 95°C for 0 s. Primers were designed with the Primer Premier 5.0 software. Forward and reverse primer sequences are listed in the Supplementary Materials. Cross-point threshold (C_t) and real-time fluorescence data were obtained using LightCycler 480 Software Version 1.5. Default software parameters were adopted to calculate C_t. The reaction efficiency E of each amplicon in each reaction was calculated using qpcR (Ritz and Spiess, 2008), an R package for quantitative real-time PCR analysis, with a five-parameter logistic model to fit onto qPCR data as described in Gardner *et al.* (2003). More details are also given in the Supplementary Materials.

2.2 NTW

The linearization of the $\underline{x}' = f(\underline{x}, \underline{p})$ equation modeling the network was performed based on the same assumptions in (Gardner *et al.*, 2003; Gregoretti *et al.*, 2010; Julius *et al.*, 2009; Lauria *et al.*, 2009), i.e. truncation of the Taylor series of f at an equilibrium point in regime of small perturbations. To free the network reconstruction from the necessity to know P, NTW processes X to rank genes in each experiment by their absolute expression values, and selects *TopK* genes with the highest values as potential target genes. It then produces a Boolean matrix IX where $ix_{ij} = 1$ if entry x_{ij} in matrix X is a potential transcriptional trigger, and $ix_{ij} = 0$ otherwise. In our experience, $TopK \leq 0.3 \times n$ is sufficient to identify the most likely perturbation triggers and store them in matrix IX, used to reconstruct perturbations P. Matrix A can then be computed row wise by using multiple regression of each row of P on the corresponding row of X, with the same sparsity assumptions as NIR and NIRest that the number of interactions associated with each gene is at most *restK* and that self-interaction on a gene is always present provided the gene is perturbed in the system. To compute the i-th row a_i^T of A, NTW iteratively searches an optimal vector P_i as the i-th row of P. The searching space of P_i consists of vectors determined by all non-empty subsets in the power set $S = 2^{\{k | IX_{ik} \neq 0, 1 \leq k \leq m\}}$. As a result, the cardinality of the searching space of P_i is $|S| - 1$, where $|S|$ represents the number of elements in set S. Given the upper value of $TopK$, the number of non-zero entries in each row of IX is small and so is the cardinality of S and the searching space of P_i. The identification of optimal P_i and thus a_i^T is then calculated through the minimization of the objective function SSE_i for regression of row i, and the reliabilities of interactions are given by the corresponding absolute interaction strengths.

For the inference of matrix P, NTW switches to compute SSE on the expression data matrix X. Namely, since P is a variable and not an input in NTW, it is not possible to use P as a reference to compute the error. Consequently, the row wise computation on SSE_i is performed on partial portions of the matrices P and A, i.e. $SSE_i = ||(tmpA^{-1} \times tmpP)_i - X_i||^2$, where $tmpA$ is an identity matrix with its i-th row being replaced by a_i^T; $tmpP$ is a matrix of all zeros with its i-th row being replaced by P_i, and $||\cdot||$ is the Euclidean norm. The rationale is that we assume self-connections (elements on the diagonal of A) to always be true (Gardner *et al.*, 2003). The matrix inversion necessary for calculating SSE can be avoided thanks to the peculiar structure of $tmpA$ (only one row of $tmpA$ differs from the same size identity matrix, more details provided in Supplementary Materials) and can therefore be obtained via simple column elimination. Thus, the computation cost of NTW is dominated by the cost of the multiple regression that can be done via QR decomposition in $O(n^2 m^2 |S|)$ for efficiency and numeric stability. More details on the computational complexity and the pseudo code for NTW are provided in the Supplementary Materials.

2.3 Networks performances

Global performances are computed as a function of the amount of information embedded in the network, and in particular in the edges between nodes. In general, this depends, with increasing difficulties of network reconstruction, on whether we are interested or not in the direction of the interaction between genes (*directed* and *undirected* networks, respectively, edges values 0, 1) and in the fact that regulation causes activation/inhibition of the target (*directed signed* or *directed unsigned* networks, edges values $0, +1, -1$).

Assessment of the performances of an algorithm are usually done by defining *positives* (P) and *negatives* (N) that label items according to the two values 0, 1 they assume in a reference Gold Standard (GS) and by defining *true* (T) and *false* (F), which represent the ability of the algorithm to classify a given item in agreement with the GS. In case of signed networks, the definitions of *T, F, N, P* require an extension due to the presence of positives (+1, −1) (Nardini *et al.*, 2008). For the computation of the performances, we used the R package *MultiClassTest* designed by Yuanhua Liu at CAS-MPG PICB (available at CRAN and at http://www.picb.ac.cn/ClinicalGenomicNTW/software.html). This allows the computation of the performances class by class (+1, −1, 0) as well as summary (binarized) performances (classes ±1 and 0). In this article, binarized results are displayed and performances are computed in terms of $PPV = \frac{TP}{TP+FP}$ (precision rate) and $Se = \frac{TP}{TP+FN}$ (recall rate). When not differently stated, results are computed for directed networks, never for undirected, and when the GS provides it, we also computed signed performances. The area under the curve (AUC) of the precision/sensitivity curve plots the couples (PPV, Se) for varying values of the threshold that defines the elements of A as True (1s) or False (0s), and it is used as a summary statistic. For comparison with other algorithms, the computation of AUC was only possible when the A matrix of continuous values was provided.

3 RESULTS

3.1 Benchmarks validation

Different synthetic and real datasets were used to test and validate our algorithm. The data used consist of the baseline (unperturbed) expression levels of genes and of their values reached at the new equilibrium (perturbed), as well as the known network of interaction GS. Performances are evaluated in terms of the ability of NTW

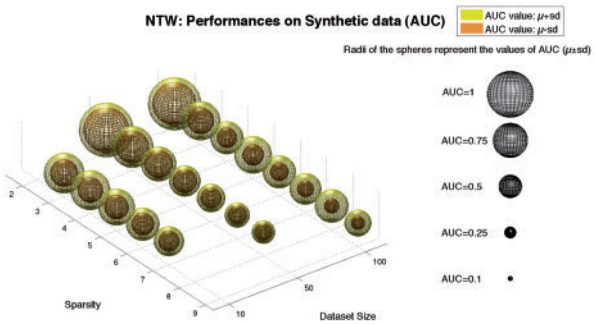

Fig. 2. Effect of the predicted sparsity of the network reconstructed by NTW and the datasets size on the algorithm performances in synthetic data. The radii of spheres represent the values of AUC averaged (mean) by size of the network: DREAM 2 (Size 50, 1 network), DREAM 3 (Size 10, 50 and 100, 10 networks each) and DREAM 4 (Size 10 and 100, 15 networks each).

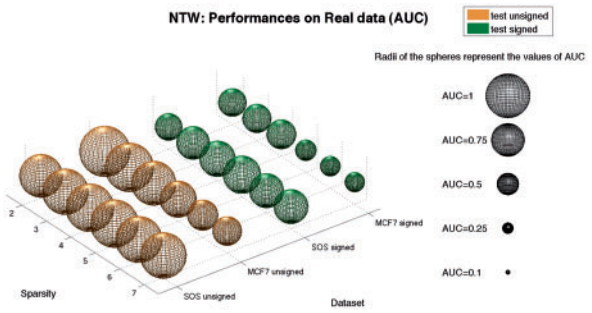

Fig. 3. Effect of the predicted sparsity of the network reconstructed by NTW and the datasets size on the algorithm performances in real data. The radii of spheres represent the values of AUC.

to reconstruct the network for varying sizes of the dataset and for changing values of the expected maximum connectivity among genes, a tunable parameter of NTW. This corresponds to the *sparsity* of matrix *A*, and although its definition implies some guess or *a priori* knowledge on genes interconnectivity, we will show that, importantly, performances do not depend crucially on varying values of this parameter.

Synthetic data: synthetic datasets were used as proposed by the Dialogue for Reverse Engineering Assessments and Methods (DREAM, wiki.c2b2.columbia.edu/dream) Consortium, an international effort to improve exchanges among researchers in the area of molecular networks reconstruction. In particular, from DREAM 2, the dataset 'Heterozygous InSilico 1 Challenge 4', which contains steady state levels of 50 genes of an hypothetical wild-type organism and 50 heterozygous knock-down strains were used; from DREAM 3, data of the same type but of three different sizes (10, 50 and 100 genes); from DREAM 4, data of both size 10 and size 100 each with three types of simulations: knock-out, knock-down and multifactorial perturbations (versus single perturbation) were used. Results in Figure 2 indicate how performances are stable (variations around the mean are small) with respect to the size of the dataset (x-axes, *Size*).

Although performances variations can be seen for different values of the maximum sparsity allowed by the reconstruction algorithm (*y*-axes, *Sparsity*)—this indicates that a (biological) knowledge of the maximum connectivity among genes can be relevant for a better reconstruction of the network—performances remain good across the whole range of variation.

Real data: We tested our approach on two real datasets. The first dataset, published in Gardner *et al.* (2003), has been re-used several times in gene network reconstruction and is therefore a known benchmark. This dataset represents a portion of the SOS pathway in *Escherichia coli*, and consists of RNA expression changes resulting from a set of nine steady state transcriptional perturbations. The second dataset was published in Nelander *et al.* (2008) to validate another ODE-based method, and consists of perturbations on MCF7 human breast carcinoma cells. As perturbants of the system, compounds targeting EGFR (ZD1839), mTOR (rapamycin), MEK (PD0325901), PKC-δ (rottlerin), PI3 kinase (LY294002) and IGF1R (A12 anti-IGF1R inhibitory antibody) were used. In this network, the targets of perturbants are unknown, i.e. *P* is unavailable. Results, shown in Figure 3, give rise to several interesting observations. First of all, as expected, prediction of the unsigned network appears to be easier than prediction of the signed network. We observe also that performances are more sensitive to *sparsity* for signed networks. This is consistent with the more challenging reconstruction offered by directed signed networks. Finally, although this may be an early generalization, we observe that NTW (and in general gene network reconstruction algorithms) appear to perform better on *E.coli* than on human cell lines, the most straightforward explanation for this fact is that unicellular organisms represents simpler systems, and are therefore easier to reconstruct.

In order to give meaning to the performances shown in Figures 2 and 3, we present here comparisons with published algorithms, chosen based on the benchmarks and the underlying model (ODE) they use. Results are presented in Tables 1 and 2. Difficulties in comparison depend critically from the poor availability of algorithms. In fact, the results on DREAM 3 of CLR + Inferelator in Table 1 were obtained directly from (Madar *et al.*, 2010) and those on MCF7 of CoPIA in Table 2 were from (Nelander *et al.*, 2008). Since the authors reported in Madar *et al.* (2010) that CLR + Inferelator performs very poorly on DREAM 2, these results were neglected in Table 1. Another limiting factor lies in the lack of widespread usage of common benchmarks, and partially in the availability of reliable benchmarks. In fact, even the excellent DREAM initiative retains the limitation due to the use of synthetic networks. As shown in the tables, NTW shows interesting performances on a variety of benchmarks. In particular, NTW behaves excellently in both signed SOS and unsigned MCF7 networks. In other cases, NTW also shows comparable performance among all four ODE algorithms, especially compared with NIRest. To further compare the performances of NTW and NIRest, we used another interesting datasets recently published (Cantone *et al.*, 2009). We found that in such datasets, NTW performs much better than NIRest in all cases (see Supplementary Table III).

3.2 Application to *B.longum*

Bifidobacteria are natural commensals in the gastrointestinal tract of humans and mammals, which have been used as markers to indicate the presence of a healthy microbiota in humans and are included in a variety of commercial food products accompanied by health-related claims (Klijn *et al.*, 2005). Despite their increasing use, the impact of various probiotic preparations on resident members of the gut microbiota and on the host are generally lacking (Sonnenburg

Table 1. Performance comparisons on DREAM 2 and 3 synthetic data

Method	DREAM 3		DREAM 2	
	PPV	Se	PPV	Se
NTW	**0.59**	**0.43**	**0.65**	**0.41**
NIRest	0.52	0.53	0.55	0.45
CLR + Inferelator	0.30	0.39	–	–
Random	0.02	–	0.06	–

The results on DREAM 3 for CLR + Inferelator are obtained directly from Madar *et al.* (2010). Random refers to the expected performance of an algorithm that selects pairs of genes randomly and then infers an edge between them. The results from NTW are highlighted in bold.

Table 2. Performance comparisons on real data

.

Method	Signed			Unsigned		
	PPV	Se	AUC	PPV	Se	AUC
NTW (SOS)	**0.52**	**0.46**	**0.47**	**0.69**	**0.76**	**0.6**
NIRest (SOS)	0.42	0.43	0.42	0.73	0.27	0.67
NIR (SOS)	0.52	0.49	0.49	0.72	0.67	0.75
Random (SOS)	0.3	–	–	0.6	–	–
NTW (MCF7)	**0.31**	**0.41**	**0.37**	**0.61**	**0.41**	**0.63**
NIRest (MCF7)	0.5	0.33	0.45	0.5	0.33	0.45
Copia (MCF7)	0.38	0.38	–	0.38	0.38	–
Random (MCF7)	0.11	–	–	0.22	–	–

Details about the GS used for comparison with CoPIA are available in the Supplementary Materials. Random refers to the expected performance of an algorithm that selects pairs of genes randomly and then infers an edge between them. AUC values could be computed only in cases that the codes or real-value matrix *A* were made available. The results from NTW are highlighted in bold.

et al., 2006). Furthermore, the elaborate gene regulation network and metabolic pathway of bifidobacteria remains unclear. We therefore applied our method to *B.longum* NCC2705 to verify whether known data could be validated, and if novel information could be reasonably proposed to aid the reconstruction of part of the amino acid biosynthesis pathway, shown in Figure 4.

Valine, leucine and isoleucine are branched chain amino acids (BCAAs) whose carbon structure is marked by a branch point. BCAAs are considered to be essential amino acids because they cannot be synthesized in human beings. Hence, they must be ingested, usually as a component of proteins. BCAAs are synthesized in plants and microorganisms via several steps starting from pyruvic acid (Lehninger *et al.*, 2005). Although information on *Bifidobacteria* are far from complete, the valine, leucine and isoleucine biosynthesis pathway exist for *B.longum* NCC2705 on KEGG database (PATHWAY: blo00290) and can be used as GS.

The application of our approach to this pathway was performed in two steps. First, we investigated how performances change in the cases in which different amounts of *a priori* information are available, i.e. some of the non-null values of *A* are known. This presents another feature of NTW, able to embed known information in the computation of the network. We observed that $AUC_0 = 0.58$ when no information is known, while for 1 known edge the mean

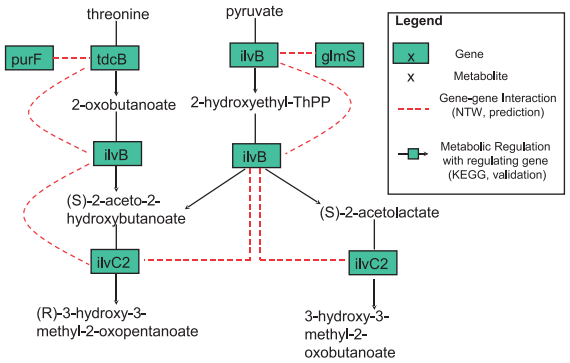

Fig. 4. Biosynthesis subnetwork of valine, leucine and isoleucine in *B.longum* NCC2705. The network composed of green boxes (genes) and dashed arrows (gene–gene interactions) is obtained from NTW. Validation of these interactions is obtained from KEGG, which gives the pathway composed of unboxed molecules (metabolites) connected by solid arrows (metabolic regulations) specified by green boxes (regulating genes).

value over all possible single edges is $AUC_1 = 0.56 \pm 0.04$ and for 2 edges $AUC_2 = 0.59 \pm 0.02$.

Several large projects—notably the International Human Microbiome Consortium (IHMC) launched in 2008 that includes the Human Microbiome Project (HMP) funded by NIH, and the metaHIT project funded by the EU—are currently contributing to elucidate the role of the GI microbiome as a whole, via the study of variations in its composition and also at the individual level with sequencing of hundreds of bacteria strains. Therefore, given the ongoing research on GI bacteria, it is likely that novel information becomes available. This could then efficiently be re-used to improve network prediction. As expected, the results show improving performances when more information is offered.

Second, we compared the results inferred by NTW without *a priori* knowledge, to investigate its ability not only to validate but also to possibly predict novel information, see Figure 4. In terms of validation, several key reaction steps of this pathway have been reconstructed by NTW: pyruvate is catalyzed by ilvB to produce 2-hydroxyethyl-ThPP, which is then catalyzed by ilvB to produce (S)-2-acetolactate. Next (S)-2-acetolactate is catalyzed by ilvC2 to produce 3-hydroxy-3-methyl-2-oxobutanoate, which is the precursor of valine; in parallel, 2-hydroxyethyl-ThPP together with 2-oxobutanoate (which is catalyzed by tdcB on the basis of threonine) produce (S)-aceto-2-hydroxybutanoate, which is then catalyzed by ilvC2 to produce (R)-3-hydroxy-3-methyl-2-oxopentanoate, the precursor of isoleucine. This result confirms that isoleucine and valine syntheses appear to be independently regulated despite the fact that the same enzyme system is used (Westfall *et al.*, 1983).

4 DISCUSSION

From the benchmarks validation, NTW shows some interesting advantages over other state-of-the-art methods. More concretely, comparing to NIR (Gardner *et al.*, 2003; Gregoretti *et al.*, 2010), NTW does not need to know *P*, which can be an important unknown of the system, to compute *A*. Comparing to NIRest (Lauria *et al.*, 2009), NTW directly produces a rough perturbation matrix without computing an estimate of *A* first, and infers *A* and *P* by

minimizing the SSEs for each candidate perturbation. This may have the advantage that the estimate of P depends directly and only on the experimental data (X), and not from previous assumptions on A. Although biological phenomena are highly non-linear, we observe that our algorithm is able to reconstruct gene networks with performances comparable or superior to CoPIA (Nelander *et al.*, 2008) (which uses a non-linear ODE model) at least on the dataset proposed. Given the simpler underlying model, NTW requires less tuning, which can be an advantage given the application of these approaches to systems where *a priori* data, necessary for an informed setting of the parameters, are poor.

In relation to the application to *B.longum*, as shown in Figure 4, two new interactions: glmS-ilvB and tdcB-purF have been predicted by NTW. In particular, tdcB is an amidophosphoribosyltransferase related to the alanine, aspartate and glutamate metabolism and purF is a L-threonine ammonia-lyase associated with valine, leucine and isoleucine biosynthesis pathway, both can be found in the KEGG database (PATHWAY: blo00290). However, due to the limited information currently available on these two genes, it is difficult to speculate on the validity of this connection predicted by NTW. Conversely, although there is no direct evidence to prove the interaction between ilvB and glmS, one study demonstrated that they are novel ClpCP targets in glucose-starved *Bacillus subtilis* and can be radioimmunoprecipitated together (Gerth *et al.*, 2008). Since ilvB initiates the biosynthesis of the branched-chain amino acids and glmS catalyzes the first step in hexosamine metabolism, converting Fructose-6P into Glucosamine-6P using glutamine as a nitrogen source, the interaction between ilvB and glmS could regulate the initiation of both of these two metabolic pathways. This also suggests that there could be an overlap between these two pathways in *B.longum* NCC2705.

In consideration of the limited number of genes identified in *B.longum* NCC2705, our experiment has not covered all the genes likely to be associated with the amino acid metabolic pathway. However, the sub-network of the amino acid metabolism of *B.longum* NCC2705 inferred by NTW provides interesting insight into this little-known and yet important GI bacterium.

ACKNOWLEDGEMENTS

The authors would like to thank Yin Jin and Wei Xiao for their initial contribution to this research and Jennifer E. Dent for her important feedback. *B.longum* NCC2705 was kindly provided by Nestlé Research Center, Lausanne, Switzerland.

Funding: Sino-Swiss Science and Technology Cooperation Project (grant number: GJHZ0911).

Conflict of Interest: none declared.

REFERENCES

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**.

Cantone,I. *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**,172–181.

Chavali,A.K. *et al.* (2008) Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Mol. Syst. Biol.*, **4**.

Efron,B. *et al.* (2004) Least angle regression. *Annal. Stat.*, **32**, 407–451.

Faith,J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Gardner,T. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102.

Gerth,U. *et al.* (2008) Clp-dependent proteolysis down-regulates central metabolic pathways in glucose-starved Bacillus subtilis. *J. Bacteriol.*, **190**, 321.

Gill,S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355.

Gregoretti,F. *et al.* (2010) A parallel implementation of the network identification by multiple regression (NIR) algorithm to reverse-engineer regulatory gene networks. *PLoS ONE*, **5**, e10179.

Guiducci,C. and Nardini,C. (2008) High parallelism, portability, and broad accessibility: technologies for genomics. *ACM J. Emerg. Technol. Comput. Syst.*, **4**, 1–39.

Julius,A. *et al.* (2009) Genetic network identification using convex programming. *IET Syst. Biol.*, **3**, 155–166.

Klijn,A. *et al.* (2005) Lessons from the genomes of bifidobacteria. *FEMS Microbiol. Rev.*, **29**, 491–509.

Lauria,M. *et al.* (2009) NIRest: a tool for gene network and mode of action inference. *Annal. N Y Acad. Sci.*, **1158**, 257–264.

Lehninger,A.L. *et al.* (2005) *Lehninger Principles of Biochemistry*. Wh Freeman. Available at http://www.citeulike.org/user/swtimmer/article/3823091.

Liu,Y. *et al.* (2010) Adapting functional genomic tools to metagenomic analyses: investigating the role of gut bacteria in relation to obesity. *Brief. Funct. Genomics* [Epub ahead of print; doi: 10.1093/bfgp/elq011].

Madar,A. *et al.* (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE*, **5**, e9803.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sciences USA*, **107**, 6286–6291.

Margolin,A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Matsuki,T *et al.* (1999) Distribution of bifidobacterial species in human intestinal microflora examined with 16S rRNA-gene-targeted species-specific primers. *Appl. Environ. Microbiol.*, **65**, 4506–4512.

Nardini,C. *et al.* (2008) MM-Correction: Meta-analysis-based multiple hypotheses correction in omic studies. *Springer CCIS*, **25**, 242–255.

Nelander,S. *et al.* (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol. Syst. Biol.*, **4**.

Nicholson,J.K. *et al.* (2005) Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Microbiol.*, **3**, 431–438.

Palmer,C. *et al.* (2007) Development of the human infant intestinal microbiota. *PLoS Biol.*, **5**, e177.

Ritz,C. and Spiess,A.-N. (2008) qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics*, **24**, 1549–1551.

Schell,M. *et al.* (2002) The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proc. Natl Acad. Sci. USA*, **99**, 14422.

Sonnenburg,J.L. *et al.* (2006) Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. *PLoS Biol.*, **4**, e413.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.. Ser. B (Methodological)*, **58**, 267–288.

Wang,W.W. *et al.* (2009) Amino acids and gut function. *Amino Acids*, **37**, 105–110.

Westfall,H.N. *et al.* (1983) Multiple pathways for isoleucine biosynthesis in the spirochete Leptospira. *J. Bacteriol.*, **154**, 846.

Zhang,C. *et al.* (2010) Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME J.*, **4**, 232–241.