Development of an internet based system for modeling biotin metabolism using Bayesian networks

Jinglei Zhou,
Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Dong Wang,
Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA, Tel: +1 4024724921 fax: +1 4024720736 dwang3@unl.edu

Vicki Schlegel, and
Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Janos Zempleni
Department of Nutrition and Health Sciences, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Abstract

Biotin is an essential water-soluble vitamin crucial for maintaining normal body functions. The importance of biotin for human health has been under-appreciated but there is plenty of opportunity for future research with great importance for human health. Currently, carrying out predictions of biotin metabolism involves tedious manual manipulations. In this paper, we report the development of BiotinNet, an internet based program that uses Bayesian networks to integrate published data on various aspects of biotin metabolism. Users can provide a combination of values on the levels of biotin related metabolites to obtain the predictions on other metabolites that are not specified. As an inherent feature of Bayesian networks, the uncertainty of the prediction is also quantified and reported to the user. This program enables convenient *in silico* experiments regarding biotin metabolism, which can help researchers design future experiments while new data can be continuously incorporated.

Keywords

Bayesian networks; Biotin; BiotinNet; Internet based

1. Introduction

Biotin is an essential water-soluble vitamin and the adequate intake (AI) for adults is 30 µg/ day. It serves as a coenzyme for acetyl-CoA carboxylases (ACC) 1 and 2, propionyl-CoA

Conflict of interest
None declared.

carboxylase (PCC), 3-methylcrotonyl-CoA carboxylase (MCC) and pyruvate carboxylase (PC). The biotin-dependent carboxylases catalyze pathways involved in fatty acid biosynthesis, gluconeogenesis, tricarboxylic acid cycle anaplerosis and pleiotropic gene regulation, particularly for genes in carbohydrate metabolism. Therefore, biotin homeostasis is crucial for maintaining normal body functions, in particular, in the heart muscle and brain [1,2].

Recent evidence has emerged showing that biotin plays a role in chromatin structure, mediated by its binding to distinct lysine residues in several classes of histones. Abnormally low biotinylation of histones appears to impair gene repression and repression of transposable elements, thereby decreasing genome stability [3].

In another important area, Mock and coworkers reported that approximately half of the pregnant women in the USA are marginally biotin-deficient, despite a normal dietary biotin intake [4]. Considering the potential link between marginal biotin deficiency and fetal malformations in humans, the findings by Mock and coworkers would have important implications for health policies and intake recommendations. Currently, this link is somewhat uncertain. While animal studies have clearly demonstrated that biotin deficiency is teratogenic, the severity of deficiency in these animal studies typically exceeded what was observed in pregnant women.

Biotin deficiency has also been reported in individuals treated with anticonvulsants. Other potential causes of biotin deficiency are intestinal malabsorption in individuals with short bowel syndrome, long-term use of drugs such as antibiotics, certain antiseizure medications and lipoic acid, excessive alcohol consumption and continuous consumption of raw egg white [5]. Importantly, biotin deficiency has been reported in severely malnourished children, creating a global public health problem [6].

In general, the importance of biotin for human health has been under-appreciated for many years. Evidence has been provided that marginal biotin deficiency might be more common than widely believed, particularly in certain subgroups of the general population, such as pregnant women, patients treated with certain drugs and severely malnourished children. There is plenty of opportunity for future research with great importance for human health.

A body of literature has accumulated on various quantitative aspects of biotin metabolism, including human feeding studies involving temporal biotin deficiency induced by egg white diet, as well as experiments using animal models and cultured cell lines. With these data sources, it is now feasible to carry out predictions or reasoning under various premises. For example, one might ask if an adult has urine biotin excretion level of 20 nmol/day, what is the likely range of biotin sulfoxide in the serum. Currently, this type of prediction involves reviewing literature on different biotin deficiency studies in humans, building models for urine and serum separately for biotin related metabolites, and then combining the models for prediction. Thus significant manual manipulation is required. The availability of a computer program that integrates various data from the literature and automatically performs the modeling process with input on selected variables will greatly expedite the prediction task of researchers. On the other hand, many important aspects of biotin metabolism are still unclear and warrant further study. By constructing an internet based computer system encoding quantitative information in published literature, BiotinNet, we provide a convenient tool to facilitate the reasoning task for researchers in biotin metabolism, to enable them to perform *in silico* experiments with different biological factors and to more efficiently plan for additional experiments.

2. Computational methods and theory

At the core of BiotinNet are Bayesian networks [7, 8] encoding quantitative knowledge with related uncertainty from published results on biotin metabolism. Bayesian networks are an attractive approach for integrating results from various independent studies by means of utilizing directed acyclic graphs. This graphical representation makes clear statements of probabilistic relationships between variables, which are easily accessible to researchers with little training in computational modeling. Figure 1 demonstrates a directed acyclic graph that represents part of the Bayesian networks of BiotinNet. In this graph, the circles (nodes) represent variables related to biotin metabolisms while arcs indicate probabilistic relationships between variables, which are prescribed by the corresponding conditional probability distribution (more details later).

Bayesian networks utilize the formal calculus of probability while relying heavily on the application of Bayes theorem. As a result, it deals with the propagation of evidence throughout the network in a consistent manner making it possible to also represent uncertainty naturally. In addition to a rich history of using Bayesian networks for medical diagnosis and prognosis [9,10], researchers are currently showing considerable interest in applying the Bayesian networks approach to problems in genetics and bioinformatics, especially on data regarding gene expression [11], protein interactions [12], pedigree analysis [13], and genetic basis of complex traits [14]. Bayesian networks have proved to be very successful for reasoning under uncertainly for these and many other applications.

Bayesian networks also provide a convenient approach to calculate the predictions and predictive distributions of interest even for large scale networks. Once the structure and parameters have been specified for the Bayesian networks, the prediction (reasoning) can be applied for all variables (nodes) using available information. For example, if the user has measured the urinary biotin and bisnorbiotin levels for a nonsmoking healthy adult, the likely range can be calculated for the levels of urine biotin sulfoxide, plasma bisnorbiotin and other variables of interest. Information flows from any combination of nodes with measurement data (instantiated nodes) to all other nodes in the network, thus removing the need to construct a new model whenever a new problem needs to be considered. Note that parameters of the conditional probability distribution are specified by experts using information from literature, thus the BiotinNet is a tool for synthesizing and presenting knowledge regarding biotin metabolism. Admittedly, certain aspects of the knowledge base for BiotinNet are sparse due to lack of experimental data. However, because Bayesian networks automatically calculate a measure of uncertainty as part of the reasoning process, the need to collect more data on certain variables is emphasized

In Figure 1, nodes and arcs in black represent the Bayesian networks that have been implemented in BiotinNet, our internet based program on biotin mechanism. Nodes and arcs in gray are planned extensions. In most cases, parameters have been omitted from Figure 1 with the understanding that parameters are required for the probability distribution regarding all variables. The exception is the parameters related to urine biotin levels (shaded node) with which we intend to show that connections could be made to experiments in other species (mouse, rat, etc.) with certain assumptions on parameters by utilizing cross-design synthesis techniques [15]. This feature might make superfluous costly studies in human subjects and also reflects another advantage of Bayesian networks, i.e., the network can be grown gradually with availability of new data sources or expanded interest of researchers.

3. Technical Background and System Description

The BiotinNet system is placed on a server with Linux operation system, MySQL relational database, and Apache web server (Figure 2). The Bayesian networks are implemented with

the JAGS software (http://calvin.iarc.fr/~martyn/software/jags/), which is an open source alternative of the popular WinBUGS software [16] for Bayesian inferences and is oriented to Unix type operation systems. JAGS is a very general statistical software that can accommodate a broad range of statistical models using the simulation approach for Bayesian graphical modeling. In Figure 1, each arc represents a dependence relationship (e.g., there is dependence between urine biotin levels and dietary intake). To specify the relationships regarding each arc, extensive literature reviews have been carried out with the experimental data and other information stored in a relational database. Statistical analysis was then performed to derive the parameters of the Bayesian networks. At this point, the variable (node) at the head of the arc is assumed to have a conditional normal distribution given the variable at the tail of the arc, with the conditional mean specified by either linear or nonlinear regression, though more complex formulation can be used if necessary. The parameters are estimated from experimental data in the literature.

When a user visits the web portal of BiotinNet, he is asked to provide information on any combination of variables that have been measured. This information is processed by Perl CGI scripts. Using the input information, necessary files for running JAGS are generated by Perl. When running JAGS, variables that are specified by the user are set at the provided value (instantiated) and Markov Chain Monte Carlo (MCMC) simulations are carried out. JAGS outputs of the MCMC simulation results are analyzed by Perl to obtain summary quantities such as mean values of metabolite levels and standard deviations. The results are then reported back to the user through CGI. Because JAGS is extremely capable of handling a wide range of Bayesian statistical models, our system is very flexible and expandable. More biological factors can be incorporated by adding new nodes and arcs. The model parameters can be updated when new literature becomes available.

The hardware requirements for this system are very moderate in relation to today's norms. The hosting server that we currently contains two quadcore Intel Xeon processors (2.33GHz), 8GB of RAM and two 120 GB hard drives. The physical server is also hosting several other services.

4. Samples of typical system runs

Users interested *in silico* modeling of biotin metabolism can access our program through the Nebraska Gateway to Nutrigenomics website http://nutrigenomics.unl.edu/software.shtml. By following the link provided, the user is provided with a form for data input (Figure 3), on which the user can select a combination of factors with known values. Currently, the list include days on biotin deficient diet, urine and serum biotin levels, as well as biotin related metabolites. After the selection is made, the user is presented with the second form (Figure 4) in which the values for the levels of measured metabolites can be entered. If the user entered values that are far out of the range of available experimental data, this fact will be explained to the user and further analysis will not be carried out. If acceptable values are entered, MCMC simulations are then performed using JAGS, and the relevant results are presented. As shown in Figure 5, the information submitted by the user is listed first in the result page. The predicted mean values and standard deviations from the MCMC simulation are then reported for all the metabolites in the Bayesian networks that are not instantiated (not supplied by the user). This provides both predictions for the levels of metabolites given the available information as well as a measure of uncertainty to the user. We have tested BiotinNet with a range of values and the results are consistent with available knowledge of biotin metabolism.

The requirements for users to access BiotinNet are a computer running a modern browser and an internet connection. The compatibility of the user's system can be found out simply

by visiting the BiotinNet website. The functionalities of BiotinNet are available free of charge to all users during the current pilot phase.

5. Discussion

The accumulation of experimental data on biotin mechanisms has make it possible to carry out *in silico* experiments on the levels various biotin related metabolites. The results can be used to design further experiments to flesh out the picture of biotin mechanism. Because many aspects of the biotin mechanism have not been well studied, it is important that the computer program to be capable of integrating the information dispersed in a number of publications with a variety of different experiments. The Bayesian networks are particular attractive for this task because of the flexibility of using Bayesian graphical models for knowledge synthesis. For this application, the quantification of uncertainty is as important as the prediction itself. A very large standard deviation suggests inadequacies of current knowledge and points to the need of more experiments on the corresponding metabolites. Here, Bayesian networks are also advantageous as they automatically provide the information on the range of possible levels of the metabolites that are not directly measured.

As an initial study, the current program incorporated experimental data on biotin related metabolites in urine and serum for human individuals on biotin deficient diet. We plan to expand BiotinNet according to Figure 1 and beyond as more experimental data become available. Zempleni's lab is a major source of research on biotin and experiments are under way to collect data on some important aspects of biotin mechanism.

Acknowledgments

References

1. Camporeale, G.; Zempleni, J. Biotin. In: Bowman, BA.; Russell, RM., editors. Present Knowledge in Nutrition. Washington, DC, USA: International Life Sciences Institute; 2006. p. 314-326.

2. Zempleni J, Hassan YI, Subhashinee SSK. Biotin and biotinidase deficiency, Expert. Rev. Endocrinol. Metab. 2008; 3:715–724.

3. Hassan YI, Zempleni J. A novel, enigmatic histone modification: biotinylation of histones by holocaroxylase synthetase. Nutr. Rev. 2008; 66:721–725. [PubMed: 19019041]

4. Mock DM. Marginal biotin deficiency is teratogenic in mice and perhaps humans: a review of biotin deficiency during human pregnancy and effects of biotin deficiency on gene expression and enzyme activities in mouse dam and fetus. J. Nutr. 2003; 133:2519–2525. [PubMed: 12888630]

5. Mock DM, Henrich-Shell CL, Carnell N, Stumbo P, Mock NI. 3-hydroxypropionic acid and methylcitric acid are not reliable indicators of marginal biotin deficiency in humans. J. Nutr. 2004; 134:317–320. [PubMed: 14747666]

6. Terran-Garcia M, Ibarra I, Velazquez A. Urinary organic acids in infant malnutrition. Pediatr. Res. 1998; 44:386–391. [PubMed: 9727718]

7. Cowell, RG.; Dawid, AP.; Lauritzen, SL.; Spiegelhalter, DJ. Probabilistic Networks and Expert Systems. New York, NY: Springer; 1999.

8. Kjaerulff, UB.; Madsen, AL. Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis. New York, NY: 2007.

9. Spiegelhalter DJ, Dawid AP, Lauritzen SL, CowellRG RG. Bayesian analysis in expert systems. Statistical Science. 1993; 8:219–247.

10. Onisko A, Drudzel MJ, Wasyluk H. Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. Int. J. Approx. Reasoning. 2001; 27:165–182.

11. Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004; 303:799–805. [PubMed: 14764868]

12. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302:449–453. [PubMed: 14564010]

13. Lauritzen SL, Sheehan NA. Graphical models for genetic analysis. Statistical Science. 2004; 18:489–514.

14. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH MH. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet. 2005; 37:435–440. [PubMed: 15778708]

15. DuMouchel W, Groer PG PG. A Bayesian methodology for scaling radiation studies from animals to man. Health Phys. 1989; 57 Suppl 1:411–418. [PubMed: 2606701]

16. Spiegelhalter, DJ.; Thomas, A.; Best, N.; Lunn, D. MRC Biostatistics Unit. Imperial College School of Medicine, UK: Institute of Public Health and Department of Epidemiology and Public Health; 2003. WinBUGS User Manual, Version 1.4. available at http://www.mrc-bsu.cam.ac.uk/bugs.

Figure 1.
Illustration of the structure of a subset of Bayesian networks in BiotinNet. Nodes (circles) represent variables in biotin metabolism. Arcs represent relationships of dependence between variables. With certain assumptions regarding parameters (shaded node), information obtained through experiments in other species can also be incorporated. Abbreviations: BNB (bisnorbiotin), BSO (biotin sulfoxide), 3HIA (3-hydroxylsovaleric acid), lymph. (lymphocyte), PC (pyruvate carboxylase), PCC (propionyl-CoA carboxylase).
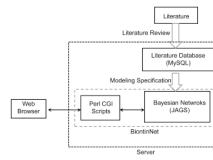
Figure 2.
The implementation of BiotinNet. Relevant literature is curated and deposited in a database system based on MySQL. Statistical analysis is carried out to obtain parameters for model specification of the Bayesian networks with inferences performed with JAGS. User inquiry will be handled by a web server (Apache).
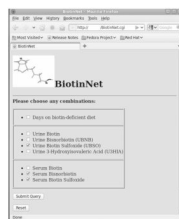
Figure 3.
A snapshot of the page for selecting factors with prespecified values.
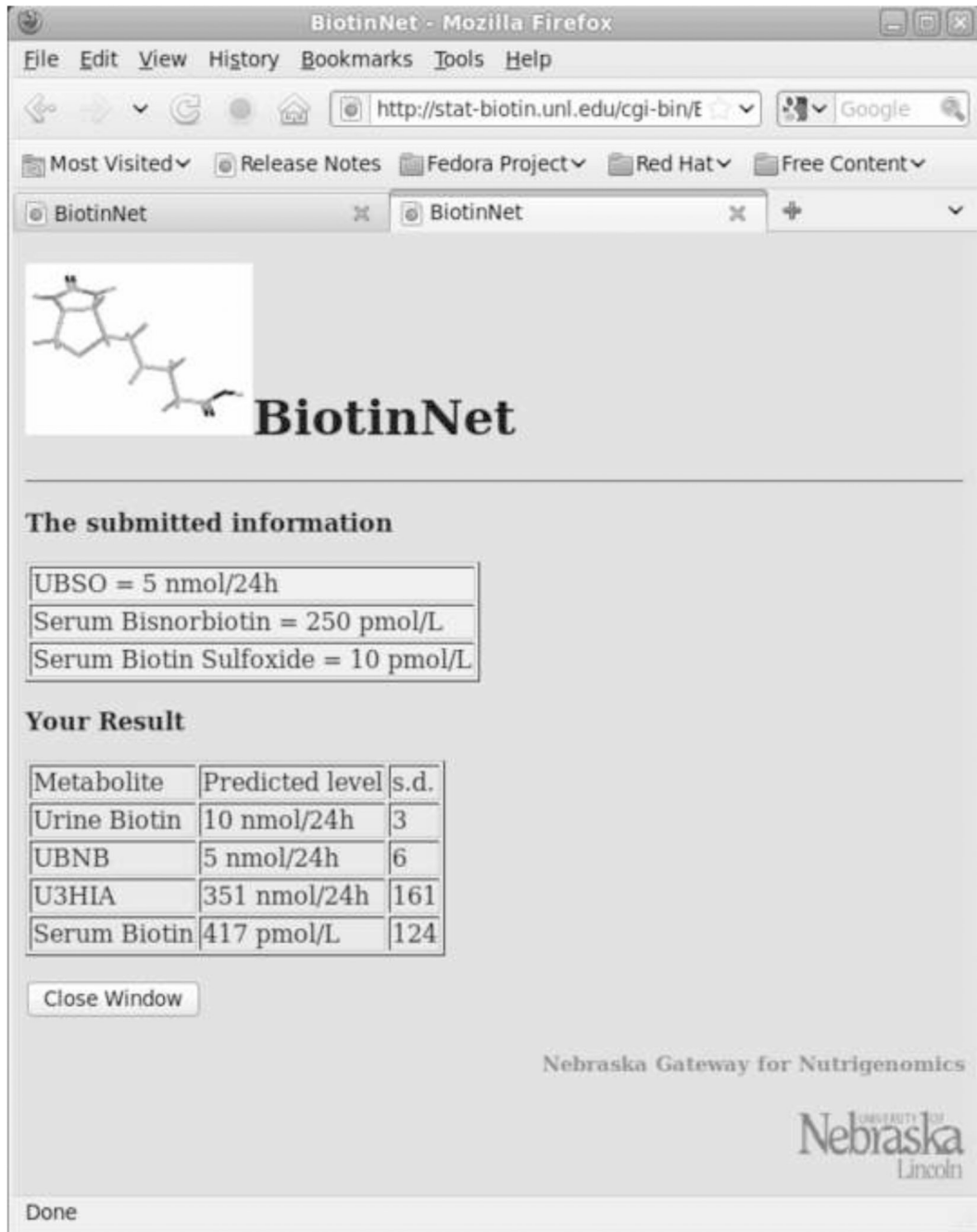
Figure 4.
A snapshot of the page for entering prespecified values.

Figure 5.
A snapshot of the results from BiotinNet.