Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery

C. F. Aliferis
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

I. Tsamardinos
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

A. R. Statnikov
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

L. E. Brown
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

Abstract *Causal Probabilistic Networks (CPNs), (a.k.a. Bayesian Networks, or Belief Networks) are well-established representations in biomedical applications such as decision support systems and predictive modeling or mining of causal hypotheses. CPNs (a) have well-developed theory for induction of causal relationships, and (b) are suitable for creating sound and practical decision support systems. While several public domain and commercial tools exist for modeling and inference with CPNs, very few software tools and libraries exist currently that give access to algorithms for CPN induction. To that end, we have developed a software library, called Causal Explorer, that implements a suit of global, local and partial CPN induction algorithms. The toolkit emphasizes causal discovery algorithms. Approximately half of the algorithms are enhanced implementations of well-established algorithms, and the remaining ones are novel local and partial algorithms that scale to thousands of variables and thus are particularly suitable for modeling in massive datasets.*

Keywords*:* Software Tools for Bioinformatics Community, Data Mining and Bioinformatics, Bayesian Models in Medicine

1 Introduction and Goals

Bayesian Networks (BNs) are computational and mathematical objects that represent compactly joint probability distributions by means of a directed acyclic graph denoting dependencies and independencies among variables and conditional probability distributions of each variable given its parents in the graph [1]. The fundamental axiom of BNs is the *Markov Condition* that allows for a concise factorization of the joint distribution and captures the main characteristic of causation in macroscopic systems, namely that causation is *local* [2]. This leads naturally to Causal Probabilistic Networks (CPNs), i.e., a special class of Bayesian Networks (BNs) [3] in which edges between any two variables in the graph denote direct causal relationships between the two variables [3]. A review of applications of CPNs and BNs in biomedicine is outside the scope of this paper, however we do note that CPNs and BNs although introduced a mere 15 years ago have already led to a long series of pioneering biomedical applications in diagnostic, treatment selection, predictive modeling and causal hypothesis generation tasks [4-12].

CPNs are also increasingly recognized in bioinformatics and computational biology, as important representations for modeling causal relationships at a finer granularity than standard clustering or

regression methods, and as having sound statistical foundations for handling noise, missing data and doing inference [13]. The appeal of CPNs is that, contrary to the pioneering heuristic approaches for generation of causal hypotheses in bioinformatics and medical research, (e.g., methods that were based on clustering, regression, and variable selection as in [14,15,16]) the recently-developed theory of *causal induction* using graphical models and related distributions, provides guarantees for highly sensitive and specific discovery of causal relationships [3]. For example, it has been theoretically proven that such methods can be used to reliably infer causal relationships among variables in: distributions captured by acyclic graphs (i.e., when feedback loops are not present) [3]; continuous linear gaussian systems with feedback loops in equilibria [3]; dynamic systems outside equilibrium sampled at discrete time points [17]; and linear or non-linear systems of discrete variables in equilibria [18].

CPN induction algorihms have also been used to find the *minimal set of predictors needed for the classification of one or more variables of interest* [19,20], known as the "Markov Blanket" (set of direct causes, direct effects, and direct causes of the direct effects) of a variable. Domain-specific molecular biology applications of CPNs so far fall under the categories of prediction of bioactivity from structural molecular properties (e.g., [21]) and induction of regulatory networks and putative causal relationships from expression data (e.g., [22]). These applications are very recent and as such only indicative of the great promise these tools hold for biomedical discovery.

The goal of the present work is to make available the powerful technology of CPN learning to a wide range of biomedical researchers that otherwise would not have access to it due to lack of technical familiarity or resources (for implementation and testing). In addition, we wish to stimulate research with a novel set of CPN algorithms we have developed for datasets with very large numbers of variables [28].

2 The *Causal Explorer* Toolkit

Currently a rich variety of software is available for modeling and inference with BNs but only a limited amount of commercial and public domain software for learning CPNs from data is available to researchers (e.g., [23,24]; for a comprehensive collection, see: *http://www.ai.mit.edu/~murphyk/Software/bnsoft.html*).

CPN induction algorithms come in two flavors: Bayesian (search-and-score) approaches, and conditional independence approaches [2]. Even with distributional assumptions that reduce computational complexity significantly, all known algorithms are practically not practical for inferring networks with more than a few hundred variables [3,21,23]. This has led some researchers to pursue modified CPN learners that instead of the full network, learn a "local neighborhood" around one or more variables of interest (Local Causal Discovery). Such methods are Markov Blanket algorithms [20,28] that can efficiently be applied for tens of thousands of variables. Another highly scalable alternative is to learn only *some* of the causal relationships (this strategy is also referred to as "Local Causal Discovery" although more precisely it is a *Partial Causal Discovery* approach) [25].

We introduce here a software library (which we call *Causal Explorer*) that provides researchers with code that can be used for experimentation with CPN learning. The selection of algorithms

emphasizes highly-scalable causal discovery (via local and partial methods as explained above), reliable and fast implementations and convenient integration to custom code. The toolkit is provided as compiled Matlab [26] functions (in the form of DLLs) running on Wintel platforms. The reasons for this choice are fourfold: (a) Matlab is a versatile and wide-spread environment for experimentation with data mining and modeling tasks in mathematic and engineering; (b) Matlab executables can be interfaced with practically any standard language such as C++, Java, etc. (c) As newer versions of the contained algorithms are being developed, transfer to the toolkit can be made very quickly (compared to the much slower process of re-writing the new algorithms in C++ or Perl etc.); (d) Matlab code if written correctly (i.e., in "vectorized" form) is very efficient and in our experiments it often outperformes native implementations of the algorithms written in C/C++ etc.

The toolkit is provided free of charge for non-commercial research. Code, example data, and documentation are available at: *http://discover1.mc.vanderbilt.edu/discover/public/causal_explorer/*

3 Algorithms

In this section, we describe the algorithms in *Causal Explorer*. All algorithms currently support three statistical tests of independence (or measures of association depending on context): G^2 and thresholded mutual information for multinomial distributions, and Fisher's z-test for multivariate Gaussian distributions [3]. In most cases this extends the functionality of the algorithms from their original published form.

3.1 PC

PC is a prototypical global algorithm for causal discovery with well-developed theory and several applications [3]. The *Causal Explorer* implementation of PC does not impose limits on the number of variables or cases in the input, and is conveniently callable from other code via the provided API.

3.2 TPDA (Three Phase Dependency Analysis, a.k.a. BN PowerConstructor)

TPDA is also a global algorithm that achieves polynomial-time execution if a constraint on the variables distribution is enforced [24]. The *Causal Explorer* implementation of TPDA employs a very fast implementation of mutual information and does not restrict the number of input variables or cases unlike the version distributed by the TPDA inventors). It is also easily callable from other code.

3.3 Sparse Candidate Algorithm

This is a fast search-and-score algorithm designed for sparsely connected domains e.g., gene pathways [22].

3.4 KS

The Koller-Sahami algorithm [20] returns a heuristic approximation to the Markov Blanket of a target variable). A very fast implementation of expected cross entropy is used.

3.5 LCD2

The LCD2 algorithm [25] is a partial induction algorithm that requires knowledge of one or more instrumental variables (i.e., variables that have no parents within the studied set of variables).

3.6. GS

The Grow-Shrink algorithm returns the Markov Blanket of a variable [27].

3.7 IAMB (Iterative Associative Markov Blanket)

IAMB [19] is a novel algorithm that returns the Markov Blanket of a variable.

3.8-3.10 IAMBnPC, InterIAMB, interIAMBnPC

These are novel algorithms that return the Markov Blanket of a variable and either use the PC algorithm or interleaved pruning to reduce the number of returned false positives relative to IAMB (trading off sample for speed) [28].

3.11 pchIAMB

A novel parallel version of IAMB suitable for multiple-CPU machines running Unix.

4 General Guidelines and Context of Use

The algorithms in *Causal Explorer* can be used in several different experimental tasks and contexts: (a) to gain insight in the causal structure of the studied domain; (b) to locate promising variables for subsequent experimentation or detailed modeling (e.g., by detailed measurements and fitting PDEs). (c) To derive a provably optimal minimal set of predictors for classification purposes.

In general, global algorithms (PC, TPDA, SCA) will be most helpful when the number of variables is up to a few hundred and the connectivity (i.e., number of direct causes/effects around variables) of the generating process is uniformly small.

Local algorithms will be most helpful when the number of variables is very large, or when the connectivity around the target variables is small (relative to available sample) while around other variables it may be large.

In particular, when the sample is large relative to the size of the Markov Blanket of the target variables (as a rule of thumb when several hundred samples are available for Markov Blankets with ~ 10 variables), GS and the IAMB variants will return excellent results. When the sample is smaller, KS and LCD2 can be applied to provide an approximation to the Markov Blanket or a subset of the global network, respectively.

5 Discussion

CPNs are powerful mathematical formalisms that are useful for variable selection, dimentionality reduction, causal hypothesis generation, and automatic creation of predictive/classification tools and decision support systems. Unfortunately the complexity of most related algorithms prevents many researchers from employing them in experiments since proper implementation often requires extensive familiarity with CPN theory and a substantial investment of resources for proper coding and testing. In addition, the existing code in the public domain typically comes in stand-alone executable form, and contains hard-coded limitations in input data size.

The first contribution of the present work is therefore that it makes available many state-of-the-art algorithms to biomedical researchers in a form that can be directly used for computational experiments. The toolkit also offers extensions by adding statistical tests that will be useful when the corresponding distributional assumptions hold. The second contribution is that the toolkit makes available to the research community *for the first time, a suite of novel algorithms especially designed for coping efficiently and reliably with thousands of variables*. These algorithms have been recently tested with a variety of datasets [28], however at this stage the potential of

these methods is practically untapped. It is our hope that the *Causal Explorer* toolkit will stimulate interest in, and experimentation with this important class of mathematical and computational tools by the broader biomedical research community.

6 Acknowledgments

Support for this research was provided in part by NIH grant LM 007613-01.

7 References

[1] Neapolitan, R.E. *Probabilistic Reasoning in Expert Systems.* 1990: John Wiley and Sons.

[2] Glymour, C. et al. *Computation, Causation, and Discovery*. 1999, AAAI Press/The MIT Press.

[3] Spirtes, P., et al. *Causation, Prediction, and Search*. 2000: The MIT Press.

[4] Heckerman DE, Horvitz EJ, Nathwani BN. *Toward normative expert systems: Part I. The Pathfinder project*. Methods Inf Med. 1992 Jun; 31(2):90-105.

[5] Heckerman DE, Nathwani BN. Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. Methods Inf Med. 1992 Jun; 31(2):106-16.

[6] Shwe M, Cooper G. An empirical analysis of likelihood-weighting simulation on a large ,multiply connected medical belief network. Comput Biomed Res. 1991 Oct; 24(5):453-75.

[7] Aliferis CF, Cooper GF. Temporal representation design principles: an assessment in the domain of liver transplantation. Proc AMIA Symp. 1998; 170-4.

[8] Ngo L, Haddawy P, Krieger RA, Helwig J. Efficient temporal probabilistic reasoning via context-sensitive model construction. Comput Biol Med. 1997 Sep;27(5):453-76.

[9] Haddawy P. *Generating Bayesian Networks from Probability Logic Knowledge Bases.* Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, July, 1994.

[10] Daphne Koller and Avi Pfeffer. *Object-Oriented Bayesian Networks*, UAI, 1997

[11] Aliferis C.F. *A Temporal representation and Reasoning Model for Medical Decision Support Systems*, Doctoral Thesis, 1998

[12] Fiszman M, Chapman WW, Evans SR, Haug PJ. *Automatic identification of pneumonia related concepts on chest x- ray reports*. Proc AMIA Symp. 1999; 67-71.

[13] Baldi, P. et al. *DNA Microarrays and Gene Expression.* 2002: Cambridge University Press.

[14] Eisen, M. et al. *Cluster analysis and display of genome-wide expression patterns PNAS* 95:1486314868, 1998.

[15] Spellman,P.T., et al. (1998) *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*. Mol. Biol. Cell, 9, 3273-3297.

[16] Li, et al. *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method*. Bioinformatics 2001 17: 1131-1142.

[17] Friedman N., et al. *Learning the structure of dynamic probabilistic networks.* In Fourteenth Conf. On Uncertainty in Artificial Intelligence (UAI) 1998.

[18] Pearl J., and R. Dechter. *Identifying Independencies in Causal Graphs with Feedback.* Twelfth Conference on Uncertainty in Artificial Intelligence, 1996.

[19] Tsamardinos I., Aliferis C.F. *Towards Principled Feature Selection: Relevancy, Filters, and Wrappers*, Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, January, 2003.

[20] Koller, D. and M. Sahami. *Toward Optimal Feature Selection.* In Thirteenth International Conference in Machine Learning. 1996.

[21] Cheng J. et al. *KDD Cup 2001 Report.* SIGKDD Explorations. 2002, 3 (2): P.1-18.

[22] Friedman, N., et al. *Using Bayesian networks to analyze expression data*. J. Comp. Bio., 7:601620, 2000.

[23] *http://www.phil.cmu.edu/projects/tetrad/*

[24] *http://www.cs.ualberta.ca/~jcheng/bnpc.htm*

[25] Cooper, G.F. *A simple algorithm for efficiently mining observational databases for causal relationships*. Journal of Data Mining and Knowledge Discovery 1 (1997) 203-224.

[26] Matlab. The Mathworks Inc.

[27] Margaritis, D., and Thrun, S. *Bayesian Network Induction via Local Neighborhoods*, Carnegie Mellon University, Technical Report CMU-CS- 99-134, August 1999.

[28] Tsamardinos I., Aliferis C.F. *Algorithms for Large Scale Markov Blanket Discovery*, The 16th International FLAIRS Conference, St. Augustine, Florida, USA, May 2003.