Appendix III:

Causal Explorer Software Library

Causal Explorer: A Matlab Library of Algorithms for Causal Discovery and Variable Selection for Classification

Alexander Statnikov[4], Ioannis Tsamardinos[1,2], Laura E. Brown[1], Constantin F. Aliferis[1,3,4]

[1] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[2] Department of Computer Science, University of Crete, Iraklio, Greece
[3] Department of Biostatistics, Vanderbilt University, Nashville, TN, USA
[4] Center of Health Informatics and Bioinformatics, New York University, New York, NY, USA

Abstract: *Causal Explorer* is a Matlab library of computational causal discovery and variable selection algorithms. *Causal Explorer* offers a wide variety of major prototypical and state-of-the-art algorithms in the field and a unified and easy-to-learn programming interface. *Causal Explorer* is designed for all researchers performing data analysis with the desire to gain an understanding in the underlying causal mechanisms that generated their data. In addition to the causal discovery methods, *Causal Explorer* contains related variable selection techniques. The variable selection algorithms in *Causal Explorer* are based on theories of causal discovery and the selected variables have specific causal interpretation. The *Causal Explorer* code emphasizes efficiency, scalability, and quality of discovery. The implementations of previously published algorithms included in *Causal Explorer* are more efficient than their original implementations. A unique advantage of *Causal Explorer* is the inclusion of very large scale and high quality algorithms developed by the authors of this chapter. The first version of *Causal Explorer* was introduced several years ago to the biomedical community. The purposes of this chapter are to re-introduce the library to a broader audience, to describe new functionality of the library, and to provide information on the use of *Causal Explorer* in community as a whole and in the Causation and Prediction Challenge.

1. Introduction

Discovery of causal knowledge is crucial for advancing research, developing new technology, and making sound policy, financial, and marketing decisions. Biologists need to know the factors that cause a disease to devise new therapeutic procedures. Public health policy makers need to know the factors that cause an increase in the number of medical errors in order to reduce them. Epidemiologists seek the factors causing disease in order to prevent it. Launching a new advertisement campaign requires knowing the factors that affect consumer behavior regarding the product. Increasing the number of visitors to a web site requires knowledge of what attracts them to the site.

Classically-trained statisticians often quote the maxim “association is not causation” to indicate that causal discovery is impossible without experiments. For example, simply observing a high occurrence of yellow stains on the fingers in patients with lung cancer relative to normal subjects does not imply a causal relation between cancer and staining (in reality heavy smoking is causing both to co-occur often). Similarly, observing that two items tend to be purchased together in high

frequency does not necessarily imply that increasing the sales of the first item will be followed by an increase of the sales of the second item.

Unfortunately, discovering causal relations strictly by randomized experimentation is inefficient and often impractical, unethical, or simply impossible. Recent advances in computational causal discovery theory and algorithm research and development mathematically prove and experimentally show respectively the feasibility of causal discovery from observational data alone under broad conditions. The acceptance and application of causal discovery methods are steadily gaining ground. The following are just a few of important references in this emerging and exciting branch of science and technology: (Neapolitan, 2004; Spirtes et al., 2000; Pearl, 2000; Glymour and Cooper, 1999; Neapolitan, 1990; Pearl, 1988).

A large body of causal discovery algorithm research and development relies on the formalisms of graphical models such as Bayesian Networks (BNs) and Causal Probabilistic Networks (CPNs). 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 (Neapolitan, 1990). 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* (Glymour and Cooper, 1999). This leads naturally to Causal Probabilistic Networks (CPNs), i.e., a special class of Bayesian Networks (BNs) in which edges between any two variables in the graph denote direct causal relationships between the two variables (Spirtes et al., 2000). A review of applications of CPNs and BNs is outside the scope of this chapter; however we do note that CPNs and BNs although introduced a mere 20 years ago have already led to a long series of pioneering applications in various scientific disciplines (Neapolitan, 2009; Taroni, 2006; Popp and Yen, 2006; Gámez et al., 2004; Friedman et al., 2000; Heckerman et al., 1992; Heckerman and Nathwani, 1992).

Causal graphical models such as CPNs are also 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 (Neapolitan, 2009; Baldi and Hatfield, 2002). The appeal of CPNs is that, contrary to the 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 (Li et al., 2001; Eisen et al., 1998; Spellman et al., 1998)) the recently-developed theory of *causal induction* using graphical models and related distributions, provides guarantees for highly sensitive and specific discovery of causal relationships (Spirtes et al., 2000). 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 (Spirtes et al., 2000); continuous linear Gaussian systems with feedback loops in equilibria (Spirtes et al., 2000); dynamic systems outside equilibrium sampled at discrete time points (Friedman et al., 1998); and linear or non-linear systems of discrete variables in equilibria (Pearl and Dechter, 1996).

It has also been shown that under certain broad conditions, a Markov blanket which is the minimal set of predictors needed for the classification of a response variable of interest is the set

of direct causes, direct effects, and direct causes of the direct effects of the response variable in a CPN (Tsamardinos and Aliferis, 2003). Thus, causal discovery algorithms that find the Markov blanket by necessity solve the variable selection problem.

We have recently introduced to the biomedical audience the powerful technology of causal discovery and variable selection encapsulated in the *Causal Explorer* library (Aliferis et al., 2003b). Over the years, we have added more algorithms and new functionality to the library. The purposes of this chapter are to re-introduce *Causal Explorer* to a broader audience, to describe new functionality of the library, and to provide information on the use of *Causal Explorer* in community as a whole and in the Causation and Prediction Challenge. In addition, we wish to stimulate research with the set of causal discovery and variable selection algorithms that we have developed for datasets with very large numbers of variables (Aliferis et al., 2009a; Aliferis et al., 2009b; Tsamardinos et al., 2006a; Brown et al., 2005; Brown et al., 2004; Tsamardinos and Aliferis, 2003; Aliferis et al., 2003a; Tsamardinos et al., 2003a; Tsamardinos et al., 2003b; Aliferis et al., 2002).

2. The *Causal Explorer* Library

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 causal graph models from data is available to researchers (for a comprehensive collection of software tools see: *http://www.ai.mit.edu/~murphyk/Software/bnsoft.html*).

Causal graph induction algorithms come in three flavors: Bayesian (or search-and-score) approaches, constraint-based conditional independence approaches, and hybrid approaches. When a researcher is interested in a specific region of the causal graph (e.g., to find causes and effects of the response variable or to find a pathway), there is no need to induce the entire causal graph (i.e., perform "*global causal discovery*"), instead one can induce that specific region of interest (i.e., perform "*local causal discovery*") which is typically much more computationally efficient (Aliferis et al., 2009a; Aliferis et al., 2009b; Chickering et al., 1994). In our experience, local causal discovery methods can be applied to datasets with hundreds of thousands variables where global causal discovery methods may not be practical. Also, the so-called Markov blanket induction methods (which is a sub-family of local causal discovery techniques) provably solve the variable selection problem under the assumptions about the learner and loss function (Tsamardinos and Aliferis, 2003).

We describe here a software library (which we call *Causal Explorer*) that provides researchers with code that can be used for causal discovery (global and local) and variable selection. *Causal Explorer* can be used primarily to:

1. Discover the direct causal or probabilistic relations around a response variable of interest (e.g., disease is directly caused by and directly causes a set of variables/observed quantities).
2. Discover the set of all direct causal or probabilistic relations among the variables.

3. Discover the Markov blanket of a response variable of interest, i.e., the minimal subset of variables that contains all necessary information to optimally predict the response variable.

The selection of algorithms in *Causal Explorer* (see next section) emphasizes highly-scalable causal discovery, reliable and fast implementations and convenient integration to custom code. Such algorithms have been frequently employed in analysis of data in psychology, medicine, biology, weather forecasting, animal breeding, agriculture, financial modeling, information retrieval, natural language processing, and other fields. They can be used to automatically construct decision support systems from data (e.g., for medical diagnosis), or to generate plausible causal hypotheses (e.g., which gene regulates which).

The *Causal Explorer* library is provided as a collection of Matlab functions. The reasons for this choice are fourfold: (a) Matlab is a versatile and wide-spread environment for experimentation with data mining and modeling tasks; (b) Matlab codes 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 library can be made very quickly (e.g., compared to the much slower process of re-writing the new algorithms in C/C++); (d) Matlab code if written correctly (i.e., in "vectorized" form) is very efficient and in our experiments it often outperforms native implementations of the algorithms written in C/C++ and other languages.

The *Causal Explorer* library is provided free of charge for non-commercial research. Code, example data, and documentation are available online at: *http://www.dsl-lab.org/causal_explorer*.

3. Causal Discovery and Variable Selection Algorithms

In this section, we describe the algorithms implemented in *Causal Explorer*. Most constraint-based 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 (Anderson, 2003; Cover et al., 1991). In most cases this extends the functionality of the algorithms from their original published form. We also note that the algorithms HITON-PC, HITON-MB, MMHC, MMPC, and MMMB were not included in the first version of *Causal Explorer* (Aliferis et al., 2003b). The detailed information on running algorithms, their inputs, and outputs can be found in the user's manual that is included in the installation package of *Causal Explorer*.

3.1. ***PC***

PC is a prototypical global causal discovery constraint-based algorithm with well-developed theory and many applications (Spirtes et al., 2000). 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)*

TPDA is also a global causal discovery algorithm that achieves polynomial-time execution if a constraint on the distribution of variables is enforced (Cheng et al., 2002). 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 in BN PowerConstructor software (*http://www.cs.ualberta.ca/~jcheng/bnpc.htm*). It is also easily callable from other code.

3.3. ***SCA*** *(Sparse Candidate Algorithm)*

This is a fast search-and-score global causal discovery algorithm designed for sparsely connected domains, e.g., gene networks (Friedman et al., 1999).

3.4. ***MMHC*** *(Max-Min Hill Climbing)*

MMHC is a highly scalable hybrid global causal discovery algorithm that has been shown to outperform in speed and quality several state-of-the-art algorithms including techniques mentioned above (Tsamardinos et al., 2006a). MMHC first uses a local discovery algorithm MMPC to learn a skeleton of the network and then it uses search-and-score method for its orientation.

3.5. ***KS*** *(Koller-Sahami)*

The Koller-Sahami algorithm returns a heuristic approximation to the Markov blanket of the response variable (Koller and Sahami, 1996). A very fast implementation of expected cross entropy is used in the algorithm implementation.

3.6. ***LCD2***

The LCD2 algorithm is a local causal discovery algorithm that requires knowledge of one or more instrumental variables (i.e., variables that have no parents within the studied set of variables) (Cooper, 1997).

3.7. ***GS*** *(Grow-Shrink)*

The Grow-Shrink algorithm returns the Markov blanket of a variable (Margaritis and Thrun, 1999). In multinomial distributions, this algorithm requires sample size exponential to the number of variables in the Markov blanket.

3.8-3.11. ***IAMB*** *(Incremental Association Markov Blanket),* ***IAMBnPC***, ***InterIAMB***, ***interIAMBnPC***

These are algorithms that return the Markov blanket of a variable (Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003a). *IAMBnPC, InterIAMB, interIAMBnPC* either use the PC algorithm (Spirtes et al., 2000) or interleaved pruning to reduce the number of returned false positives relative to IAMB (trading off sample for speed) (Tsamardinos et al., 2003a). In

multinomial distributions, all these algorithms require sample size exponential to the number of variables in the Markov blanket.

3.12-3.13. ***HITON-PC*** *(HITON Parents and Children)* and ***MMPC*** *(Max-Min Parents and Children)*

HITON-PC and MMPC are local causal discovery algorithms that return the set of direct causes and effects of the response variable (Aliferis et al., 2009a; Aliferis et al., 2009b; Tsamardinos et al., 2006a; Aliferis et al., 2003a; Tsamardinos et al., 2003b). HITON-PC uses univariate heuristic for prioritization of variables, while MMPC uses max-min association heuristic. These are highly sample efficient discovery techniques.

3.14-3.15. ***HITON-MB*** *(HITON Markov Blanket)* and ***MMPC*** *(Max-Min Markov Blanket)*

These are Markov blanket induction algorithms that require much less sample compared to GS and IAMB family of Markov blanket inducers (Aliferis et al., 2009a; Aliferis et al., 2009b; Tsamardinos et al., 2006a; Aliferis et al., 2003a; Tsamardinos et al., 2003b). HITON-MB uses univariate heuristic for prioritization of variables, while MMMB uses max-min association heuristic.

4. Other Tools

In addition to the causal discovery algorithms mentioned in section 3, *Causal Explorer* also includes several tools that facilitate causal discovery experiments and development of new algorithms. These tools are outlined below. None of these tools were provided in the first version of *Causal Explorer* (Aliferis et al., 2003b). The detailed information on running these tools can be found in the user's manual that is included in the installation package of *Causal Explorer*.

4.1. Bayesian network tiling tool

It is well recognized in the field that the major technique for evaluating and comparing causal discovery algorithms is by simulation of data from a network of known structure. Then, it is easy to compare the reconstructed network as learnt by an algorithm with the true data-generating network to assess the quality of learning. For the results of the evaluations to carry to real-world data distributions the networks used for data simulations have to be representative of the real-world examples. Typically, the networks employed for the data simulation are extracted from real-world BN-based decision support systems. Unfortunately, the size of the existing known BNs is relatively small in the order of at most a few hundred variables. Thus, typically causal discovery algorithms were so far validated on relatively small networks (e.g., with less than 100 variables), such as the classical ALARM network (Beinlich et al., 1989) or other "toy-networks". Algorithms have also been developed to generate large random BNs. The *BNGenerator* system is one example for generating large random BNs from a uniform distribution (Ide and Cozman, 2002). However, the *BNGenerator* system and other algorithms of this type do not provide any guarantees that these networks resemble the networks of the distributions likely to be encountered in practice (Aliferis and Cooper, 1994). The emergence of datasets of very high-

dimensionality poses significant challenges to the development of new causal discovery algorithms.

To address this problem, *Causal Explorer* implements an algorithm for generating arbitrarily large discrete Bayesian networks by tiling smaller real-world known networks. The algorithm preserves the structural and probabilistic properties of the tiles so that the distribution of the resulting tiled network resembles the real-world distribution of the original tiles (Tsamardinos et al., 2006b).

4.2. Bayesian network data simulator

Causal explorer implements a procedure to simulate data from Bayesian networks using the Gibbs sampling algorithm (Russell and Norvig, 2003). Such data can be used for evaluation of existing causal discovery algorithms and development of new methods.

4.3. Utility for supervised discretization of continuous data

In order to discretize continuous data, *Causal Explorer* implements a supervised discretization method that works as follows:

1. Data is normalized so that each variable has mean 0 and standard deviation 1.
2. After normalization, association of each variable with the response variable is computed using either Wilcoxon rank sum test (for binary response variable) or Kruskal-Wallis non-parametric ANOVA (for multicategory response variable) at 0.05 alpha level (Hollander and Wolfe, 1999).
3. If a variable is not significantly associated with the response variable, it is discretized as follows:
    - *0* for values less than -1 standard deviation
    - *1* for values between -1 and 1 standard deviation
    - *2* for values greater than 1 standard deviation
4. If a variable is significantly associated with the response variable, it is discretized using sliding threshold (into binary) or using sliding window (into ternary). The discretization threshold(s) is determined by the Chi-squared test to maximize association with the response variable (Agresti, 2002).

The discretization procedure can be instructed to compute necessary statistics only using training samples of the data to ensure unbiased estimation of error metrics on the testing data.

5. 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; and (c) to derive a provably optimal minimal set of predictors for classification purposes.

In general, global causal discovery algorithms PC, TPDA, and SCA can be practically run 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 data-generating process is uniformly small. When the number of variables is of the order of thousands, MMHC algorithm will be most helpful as it is the most scalable method.

Local causal discovery algorithms will be most helpful when the number of variables is very large, or when the connectivity around the response variable 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 response variable (as a rule of thumb when several hundred samples are available for a Markov blanket with ~5 variables), GS and the IAMB variants will return excellent results. When the sample is smaller, HITON-MB and MMMB should be applied instead. Also, in many cases algorithms HITON-PC and MMHC that return the set of direct causes and effects of the response variable can be used for approximation of the Markov blanket.

6. Statistics of Registered Users

At the time of writing this chapter, *Causal Explorer* has 739 registered users in more than 50 countries all over the world. Based on provided information in the user registration form, 402 (54%) users are affiliated with educational, governmental, and non-profit organizations and 337 (46%) users are either from private or commercial sectors. Major commercial organizations that have registered users of *Causal Explorer* include IBM, Intel, SAS Institute, Texas Instruments, Siemens, GlaxoSmithKline, Merck, and Microsoft. Table 1 provides a list of major U.S. institutions that have registered users of *Causal Explorer*.

7. Use of *Causal Explorer* in the Causation and Prediction Challenge

Causal Explorer library was used both by participants and organizers of the Causation and Prediction Challenge.

7.1. Use of Causal Explorer by the Challenge participants

Here are major achievements enabled by the *Causal Explorer* library in the Causation and Prediction Challenge:

1. Gavin Cawley used *Causal Explorer* to become one of the Challenge winners. The software library allowed him to achieve the best prediction accuracy on SIDO and MARTI datasets (p12).
2. Jianxin Yin et al. used *Causal Explorer* to become the Challenge winners in the "*best overall contribution*" category. Specifically, they obtained the best position of the Pareto front of the Fscore vs. Tscore graph over all datasets (p12, p13).
3. Laura Brown and Ioannis Tsamardinos used *Causal Explorer* to achieve the top overall ranking on REGED dataset (p47).

Summary of the use of *Causal Explorer* by the Challenge participants is provided in Table 2.

Table 1: A list of major U.S. institutions that have registered users of *Causal Explorer.*

1. Boston University	30. University of California Berkley
2. Brandies University	31. University of California Los Angeles
3. Carnegie Mellon University	32. University of California San Diego
4. Case Western Reserve University	33. University of California Santa Cruz
5. Central Washington University	34. University of Cincinnati
6. College of William and Mary	35. University of Colorado Denver
7. Cornell University	36. University of Delaware
8. Duke University	37. University of Houston-Clear Lake
9. Harvard University	38. University of Idaho
10. Illinois Institute of Technology	39. University of Illinois at Chicago
11. Indiana University-Purdue University Indianapolis	40. University of Illinois at Urbana-Champaign
12. Johns Hopkins University	41. University of Kansas
13. Louisiana State University	42. University of Maryland Baltimore County
14. M. D. Anderson Cancer Center	43. University of Massachusetts Amherst
15. Massachusetts Institute of Technology	44. University of Michigan
16. Medical College of Wisconsin	45. University of New Mexico
17. Michigan State University	46. University of Pennsylvania
18. Naval Postgraduate School	47. University of Pittsburgh
19. New York University	48. University of Rochester
20. Northeastern University	49. University of Tennessee Chattanooga
21. Northwestern University	50. University of Texas at Austin
22. Oregon State University	51. University of Utah
23. Pennsylvania State University	52. University of Virginia
24. Princeton University	53. University of Washington
25. Rutgers University	54. University of Wisconsin-Madison
26. Stanford University	55. University of Wisconsin-Milwaukee
27. State University of New York	56. Vanderbilt University
28. Tufts University	57. Virginia Tech
29. University of Arkansas	58. Yale University

7.2. Use of Causal Explorer by the Challenge organizers

The *Causal Explorer* library was also used by the Challenge organizers. First, resimulation of REGED dataset (p166) employed HITON-PC algorithm (as a part of HITON-Bach method) that was implemented in *Causal Explorer*. In addition, HITON-PC and MMHC algorithm

implementations from *Causal Explorer* were used as the baseline methods to gain insight into the problem difficulty (p171, pp177-179, pp186-188). At the time of Challenge the use baselines was not disclosed to the participants.

Table 2: Summary of the use of Causal Explorer by the Challenge participants.

Participant team	Algorithms used in *Causal Explorer*	Challenge ranking on Tscore				Reference
		REGED	SIDO	CINA	MARTI	
Gavin Cawley	• HITON-MB, • HITON-PC, • MMHC	2	1	3	1	p118
Jianxin Yin et al.	• MMPC, • Supervised discretization	3	5	4	2	p95, p100, p153
Cristian Grozea	• Markov blanket algorithm (details are not provided)	7	12	7	6	p129
H. Jair Escalante and Luis Enrique	• HITON-PC	6	8	9	5	p130
Ernest Mwebaze and John Quinn	• HITON-PC	9	7	8	-	p139
Marius Popescu	• HITON-MB, • TPDA	5	-	-	-	p145
Wu Zhili	• HITON-MB, • HITON-PC	13	13	14	11	p156
Laura Brown and Ioannis Tsamardinos	• MMPC, • MMMB, • MMHC, • HITON-MB	*Excluded from the Challenge ranking due to conflict of interest*				p35, p114

8. Discussion

CPNs and other causal graphical models are powerful mathematical formalisms that are useful for variable selection, dimensionality 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 the 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 may contain hard-coded limitations on input data size.

The first contribution of the present chapter is that it re-introduces the *Causal Explorer* library to a broader audience and describes new functionality compared to the previous version of the library (Aliferis et al., 2003b). The second contribution is that the library makes available to the research community a suite of algorithms designed by authors of this chapter for coping efficiently and reliably with thousands of variables. These algorithms have been recently tested with a variety of datasets with excellent results (Aliferis et al., 2009a; Aliferis et al., 2009b), however at this stage the potential of these methods is practically untapped. Finally, we also describe the use of *Causal Explorer* in community at a whole and in the context of Causation and

Prediction Challenge. It is our hope that the *Causal Explorer* library will stimulate interest in, and experimentation with this important class of mathematical and computational tools by the broader research community.

9. Acknowledgements

This work supported by NIH grants R56 LM007948-04A1, R01 LM007948-01, P20 LM007613-01, and Vanderbilt AVCF for the Discovery Systems Laboratory.

10. References

Agresti,A. (2002) *Categorical data analysis*. Wiley-Interscience, New York, NY, USA.

Aliferis,C.F. and Cooper,G.F. (1994) An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI).*

Aliferis,C.F. et al. (2009a) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. (In press) *Journal of Machine Learning Research*.

Aliferis,C.F. et al. (2009b) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions. (In press) *Journal of Machine Learning Research*.

Aliferis,C.F., Tsamardinos,I. and Statnikov,A. (2002) Large-scale feature selection using Markov blanket induction for the prediction of protein-drug binding. *Technical Report DSL 02-06*.

Aliferis,C.F., Tsamardinos,I. and Statnikov,A. (2003a) HITON: a novel Markov blanket algorithm for optimal variable selection. *AMIA 2003 Annual Symposium Proceedings*, 21-25.

Aliferis,C.F. et al. (2003b) Causal Explorer: a causal probabilistic network learning toolkit for biomedical discovery. *Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS).*

Anderson,T.W. (2003) *An introduction to multivariate statistical analysis*. Wiley-Interscience, Hoboken, N.J.

Baldi,P. and Hatfield,G.W. (2002) *DNA microarrays and gene expression*. Cambridge University Press, Cambridge, UK.

Beinlich,I. et al. (1989) The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*.

Brown,L.E., Tsamardinos,I. and Aliferis,C.F. (2004) A novel algorithm for scalable and accurate Bayesian network learning. *Medinfo 2004.*, 11, 711-715.

Brown,L.E., Tsamardinos,I. and Aliferis,C.F. (2005) A comparison of novel and state-of-the-art polynomial Bayesian network learning algorithms. *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI).*

Cheng,J. et al. (2002) Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 137, 43-90.

Chickering,D.M., Geiger,D. and Heckerman,D. (1994) Learning Bayesian networks is NP-hard. *Technical Report MSR-TR-94-17*.

Cooper,G.F. (1997) A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Mining and Knowledge Discovery*, 1, 203-224.

Cover,T.M. et al. (1991) *Elements of information theory*. Wiley New York.

Eisen,M.B. et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A*, 95, 14863-14868.

Friedman,N. et al. (2000) Using Bayesian networks to analyze expression data. *J Comput. Biol.*, 7, 601-620.

Friedman,N., Murphy,K. and Russell,S. (1998) Learning the structure of dynamic probabilistic networks. pp. 139-147.

Friedman,N., Nachman,I. and Pe'er,D. (1999) Learning Bayesian network structure from massive datasets: the "Sparse Candidate" algorithm. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI).*

Gámez,J., Moral,S. and Salmerón,A. (2004) *Advances in Bayesian networks*. Springer, Berlin.

Glymour,C.N. and Cooper,G.F. (1999) *Computation, causation, and discovery*. AAAI Press, Menlo Park, Calif.

Heckerman,D.E., Horvitz,E.J. and Nathwani,B.N. (1992) Toward normative expert systems: Part I. The Pathfinder project. *Methods Inf. Med*, 31, 90-105.

Heckerman,D.E. and Nathwani,B.N. (1992) Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. *Methods Inf. Med*, 31, 106-116.

Hollander,M. and Wolfe,D. (1999) *Nonparametric statistical methods*. Wiley, New York, NY, USA.

Ide,J.S. and Cozman,F.G. (2002) Random generation of Bayesian networks. *Lecture Notes in Computer Science*, 366-375.

Koller,D. and Sahami,M. (1996) Toward optimal feature selection. *Proceedings of the International Conference on Machine Learning*, 1996.

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

Margaritis,D. and Thrun,S. (1999) Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems*, 12, 505-511.

Neapolitan,R.E. (1990) *Probabilistic reasoning in expert systems: theory and algorithms*. Wiley, New York.

Neapolitan,R.E. (2004) *Learning Bayesian networks*. Pearson Prentice Hall, Upper Saddle River, NJ.

Neapolitan,R.E. (2009) *Probabilistic methods for bionformatics (with an introduction to Bayesian networks)*. Morgan Kaufmann Publishers, Burlington, MA.

Pearl,J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, California.

Pearl,J. (2000) *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K.

Pearl,J. and Dechter,R. (1996) Identifying independencies in causal graphs with feedback. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI)*, 420-426.

Popp,R.L. and Yen,J. (2006) *Emergent information technologies and enabling policies for counter-terrorism*. Wiley-Interscience, Hoboken, N.J.

Russell,S.J. and Norvig,P. (2003) *Artificial intelligence: a modern approach*. Prentice Hall/Pearson Education, Upper Saddle River, N.J.

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.

Spirtes,P., Glymour,C.N. and Scheines,R. (2000) *Causation, prediction, and search*. MIT Press, Cambridge, Mass.

Taroni,F. (2006) *Bayesian networks and probabilistic inference in forensic science*. Wiley, Chichester, England.

Tsamardinos,I. and Aliferis,C.F. (2003) Towards principled feature selection: relevancy, filters and wrappers. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AI & Stats)*.

Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003a) Algorithms for large scale Markov blanket discovery. *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 376-381.

Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003b) Time and sample efficient discovery of Markov blankets and direct causal relations. *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD)*, 673-678.

Tsamardinos,I., Brown,L.E. and Aliferis,C.F. (2006a) The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65, 31-78.

Tsamardinos,I. et al. (2006b) Generating Realistic Large Bayesian Networks by Tiling. *Proceedings of the 19th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*.