Causal Discovery from Pediatric Infectious Disease Protein Biomarker Data

Subramani Mani MBBS PhD[1], Daniel Cannon MS[1], Robin Ohls MD[1], Douglas Perkins PhD[1], Karri Ballard PhD[2], Cristian Bologa PhD[1]

[1]University of New Mexico Health Sciences Center, Albuquerque, NM, 87131

[2]Myriad RBM, Austin, TX 78759

Abstract: Discovering cause and effect relationships from disease-linked proteomic biomarker data related to neonatal sepsis and other pediatric infections can shed light on the pathophysiology of these clinical manifestations and may lead to the identification of new drug targets and novel therapeutic advances. Finding causal biomarkers that aid in defining the immunological changes and inter-relationships among immune signaling molecules is an important step towards an improved understanding of the pathophysiology of pediatric infectious diseases. In this study, we applied the Bayesian Local Causal Discovery (BLCD) algorithm previously developed by us, for our causal discovery task in the domain of various pediatric infections. We demonstrate the potential of our causal modeling approach using proteomic biomarkers in the domain of neonatal sepsis obtained by enrolling 127 infants, and in the domain of other pediatric infections by enrolling 369 infants. The BLCD algorithm output nine potential causal relationships from the neonatal sepsis proteomic dataset and seven purported causal relationships from the other pediatric infections proteomic data. Two of the causal postulates for neonatal sepsis and four of the seven causal relationships for other pediatric infections appear plausible.

Introduction and background: Causal discovery is a challenging and important task. A significant part of human endeavor is concerned with the exploration of causes of various phenomena. In the domain of biomedicine, which lays the scientific foundation of healthcare, determining the cause of a disease helps in prevention and treatment. Researchers have mostly focused on predictive models, association rules and testing for dependence/independence between pairs of variables or sets of variables as part of data analytics using observational data. These modeling methods have advanced classification and data description tasks. However, there has also been a surge of interest in exploring further and develop methods to propose cause and effect relationships from passively collected data in various domains including healthcare settings [1-3]. Even when experimental studies in the form of randomized controlled studies or other types of interventional studies might be needed to confirm a causal postulate, methods to infer causal relationships from observational data could be used to narrow the experimental hypotheses search space and channel available resources efficiently. The aim here is not to replace experimental studies, which are extremely valuable in science, but to complement experimental studies when feasible by performing less invasive and less expensive studies using novel computational approaches. Based on meta-analysis of randomized (experimental) and nonrandomized (observational) studies in healthcare, researchers have found marked correlation between the observational and experimental studies [4, 5].

Though there were some concerns early on about the process of biomarker discovery, validation and clinical utility [6], recent advances in proteomic biomarker assay development [7-13] have considerably improved the utility and value for examining protein biomarkers in blood. Since changes in the expressed proteome impact directly on the disease manifestation and is downstream relative to the genome and transcriptome, assaying proteins holds considerable promise for discovering disease-linked biomarkers thereby opening the gates for the identification of potential molecular targets for newer therapies down the road. Identification of novel molecular targets and development of molecularly targeted therapies will enable physicians to select the most effective treatments for a patient's condition and skip protocols that are unlikely to improve outcomes [14].

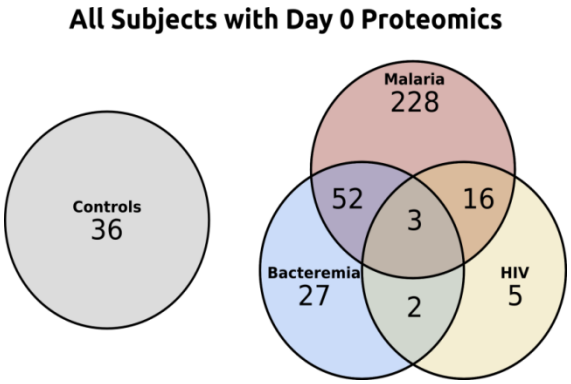

Figure 1: Disease conditions included in the 25-plex human inflammatory array.

However, there is very limited work on mechanistic modeling of the biomarkers and identification of biomarkers that are likely to be druggable, that is, proteins which could be subject to manipulation by small molecules. Simply identifying the protein biomarkers differentially expressed in diseased versus healthy populations may be appropriate for early diagnosis and predicting prognostic outcomes, but this approach does not offer the ability to differentiate between

biomarkers that drive disease manifestations (causal influences) and those that result from the disease process (effects). Although biomarkers can play an effective role in early detection, diagnostic evaluation, and for assessment of prognosis, a cause-effect understanding is required if biomarkers are to be targeted for druggability and eventual use of the biomarkers for therapeutics.

Methods:

Study sample: For applying our causal discovery methodology we focused on two unique and comprehensive pediatric infectious disease cohorts: 1) a study on biomarkers for neonatal sepsis (from the Bio-signature Study of Neonatal Sepsis (NIH 1R44GM082038-01; PI, Ballard) conducted at the University of New Mexico Children's Hospital; and (2) a study on biomarkers for pediatric infectious diseases (for example, malaria, bacteremia, HIV-1, etc.; *Genetic Basis of Severe Malarial Anemia and Co-Morbidities,* AI051305; PI, Perkins) conducted at the Centers for Global Health at UNM and the Kenya Medical Research Institute (KEMRI, Kenya). For the neonatal sepsis bio-signature study, we enrolled 127 eligible very low birth weight (VLBW) infants (gestational age $\leq$ 32 weeks, birth-weight $\leq$ 1500 grams, postnatal age $\geq$ 120 hours) over a five-year period from 2007 to 2012. We had 39 cases (culture positive sepsis) with the remaining 88 serving as controls (sepsis negative). Serum samples were collected from each infant over a 21-day period to perform a focused proteomic assay of 90 potential biomarkers suspected to play a role in infection and/or inflammation. For the pediatric infections study we enrolled 1,655 children (< 14 months) and followed them post-enrollment for 36 months. A subset of the patients was selected for the day 0 (enrollment) visit to analyze protein biomarkers using a 25-plex human inflammatory array (n=369). The sample sizes for the mono- and co-infected children are shown in Figure 1.

Algorithmic methods: Our focus is on learning causal relationships from data using a causal Bayesian network (CBN) framework and the natural question is why we should try to learn these models from data. Initially Bayesian networks were built as knowledge-based systems and the structure and parameters were specified by experts [15-17]. However, it was challenging for domain experts to characterize the probabilistic dependencies and independencies among the variables in a domain and orient the edges between the variables to capture the probabilistic relationships even with the help of knowledge engineers. It was also very labor intensive and challenging for experts to specify precisely the prior and conditional probabilities that parameterized the models. For various domains, experts had problems assessing a full-fledged causal structure, or the parameters, and in some situations, both. Hence researchers started to focus on data—initially for parameter estimation, but subsequently to learn both the network structure as well as the associated marginal and conditional probabilities.

Broadly speaking there are two general approaches to learn CBNs from data—global and local. Global approaches try to build a causal model consisting of all the measured variables in a domain whereas the local approach explores subsets of variables to build local causal models. Both the global and local models can be generated from data using the constraint-based formalism or the Bayesian paradigm. Constraint-based approaches to causal discovery were put forward by Pearl and Verma [18] and by Spirtes, Glymour, and Scheines [19]. The PC and FCI algorithms [20], for instance, take a global approach to causal discovery and output a graph with different types of edges between all the variables to represent for example that X causes Y, X does not cause Y, or the causal direction is undetermined [2]. The FCI algorithm can also model latent variable patterns.

Earlier research on learning Bayesian networks from data using a Bayesian approach [21, 22] has simultaneously

Box 1: BLCD Algorithm steps

1. **Derive the Markov Blanket:** For each node $X \in \mathbf{X}$, heuristically derive the Markov Blanket of X using the *Procedure MB*. **X** denotes the set of all random observed variables in the dataset. Let **B** denote the MB of X.
2. **Update B:** Apply the MB *update* rule which states that if node A is in the MB of node C, add node C to the MB of node A if it is not already included.
3. **Pick** W_1, W_2, X and Z: Select sets of four variables from the set obtained by the union of **B** and X as follows. We refer to each set of four variables as a tetraset **T**. Since we are focusing on the MB of X, X is an essential element of **T**. Note that each tetraset can give rise to 3 "Y" patterns where the X variable is a cause and each of the other three are potential effects.
4. **Derive** $P(X \rightarrow Z \mid D)$: For each of the 3 "Y" patterns, the probability of $X \rightarrow Z$ is derived using Equation 1 rendered below.
5. **Generate output**: If $P(X \rightarrow Z \mid D) > t$, where t is a user-set threshold, then output $X \rightarrow Z$.

modeled all the causal relationships among the model variables. The optimal reinsertion (OR) algorithm is an algorithm for learning Bayesian networks using a score-based search method developed by Moore and Wong [23]. Tsamardinos et al. introduced the hybrid MMHC algorithm that combines constraint-based and Bayesian methods for Bayesian network structure learning [24]. These global approaches can require long search times when the number of variables is large. When they are used to model latent variables, these approaches can be extremely slow, even when modeling only a few variables.

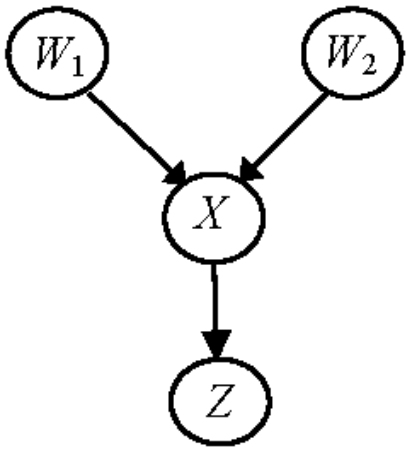

Figure 2: Y structure Bayesian network containing four nodes

We used the Bayesian local causal discovery algorithm (BLCD) [25, 26] that conjectures causal relationships between pairs of variables that have no common causes (confounders) for discovering pairwise cause and effect relationships. Instead of using constraint-based independence and dependence tests, we scored the models by a Bayesian method. BLCD can be implemented in an *anytime* framework, to output the "discovered" causes as they are found. To discover the effects of a node *X*, we require only the Markov blanket (MB) of *X* and data on *X* and the variables in the MB of *X*. The MB of a node *X* in a Bayesian network is the set of nodes obtained by the union of the parents of *X*, the children of *X* and the parents of the children of *X*. The steps of the BLCD algorithm are shown in Box 1.

BLCD is an efficient algorithm that uses the local causal discovery framework and a Bayesian approach. By making use of the "Y" structure for identifying unconfounded pairwise causal relationships and the Markov Blanket for defining the "locality" of a node, BLCD can output predominantly direct causal relationships, while keeping the number of false positives low. A Y structure Bayesian network is a Bayesian network containing four variables that has the structure shown in Figure 2, where the node labels are arbitrary. By developing BLCD we have formalized the task of causal discovery from observational data using a Bayesian approach and local search based on the identification of Y structures as *sufficient* structures for causal discovery. We have formally shown that discovering Y structures enables us to make causal claims for the relationship between *X* and *Z*, that is, to make the inference that *X* causally influences *Z* [26].

Equation 1: The probability of a causal relationship, that is, P ($X \rightarrow Z|\mathrm{D}$) can be approximated by $\frac{\mathrm{Score}(\mathrm{G}_1|\mathrm{D})}{\sum_{i=1}^{543}\mathrm{Score}(\mathrm{G}_i|\mathrm{D})}$ where G_i represents one of the 543 CBNs over $\mathbf{V} = \{W_1, W_2, X, Z\}$ and D denotes the dataset, using the BLCD algorithm (see Box 1).

Table 1: Neonatal Sepsis Causal Output

Cause	Effect	Plausible
1. A2Macro	CK-MB	?
2. IL-15	A2Macro	?
3. IL-15	Thrombopoietin	?
4. IL-15	IL-5	?
5. CK-MB	Myoglobin	?
6. CD40-L	TBG	?
7. A2Macro	Fibrinogen	?
8. TNFR2	CRP	Yes
9. IL-15	IL-12p40	Yes

Table 2: Pediatric Infections Causal Output

Cause	Effect	Plausible
1. IL-4	L-2R	Yes
2. IL-4	Eotaxin	Yes
3. IL-4	MCP-1	Yes
4. IFN-a	IP-10	Yes
5. MIP-1B	IL-1Ra	?
6. MIP-1B	IL-5	?
7. MIP-1B	IL-7	?

Results: The BLCD algorithm output nine potential causal relationships from the neonatal sepsis proteomic dataset which are shown in Table 1. Causal relationship #8 (TNFR2 causally influencing CRP) is supported by [27] with TNFR2 becoming upregulated prior to CRP. Causal relationship #9 (IL-15 causally influencing IL-12p40) is supported by [28]. For "causal relationships" #1 - #7 additional evaluation is needed. The BLCD algorithm also output seven causal relationships from the pediatric infections proteomic data which are shown in Table 2. The first four relationships (#1 - #4, Table 2) seem to be biologically meaningful, have been substantiated in *in vivo* and *in vitro* findings, and converged upon our previous studies showing how inflammatory mediators influence disease severity in pediatric infectious diseases (for review see [29]). Additional relationships generated (#5 - #7) such as the causal relationships between MIP-1B and IL-1Ra, IL-5 and IL-7, although not previously established, warrant further investigation.

Discussion and Conclusion: Finding causal biomarkers that aid in defining the immunological changes and inter-relationships among immune signaling molecules is an important step towards an improved understanding of the pathophysiology of neonatal sepsis and other pediatric infectious diseases. For the neonatal sepsis causal discovery study, we used only data that were collected using samples drawn on and before the time of blood draw for culture testing and the causal interactions would thus conform to the causal model prevalent during that time. Such causal knowledge will enable exploration of new protein targets for drug development resulting in novel therapeutic advances. One common outcome that the three endemic conditions (i.e., malaria, HIV-1, and sepsis) share is the development of life-threatening anemia. We have found that anemia occurs because of mono-infection in the three diseases and is more profound in the context of co-infection and we plan to explore the causal influences for the different diseases and disease combinations as single entities as we assay additional samples. The results obtained from the causal output of BLCD are highly encouraging since many the cause and effect relationships that emerged between the inflammatory molecules are biologically meaningful but further exploration and evaluation are needed. Since the causal relationships are modeled using the formalism of DAGs, an assumption of acyclicity is inherent in the modeling and feedback loops among the biomarkers cannot be discovered using the method. However, by using protein measurements from a single time point the problem can be mitigated.

When experimental studies are contra-indicated, due to ethical, logistical, or cost considerations, causal discovery from observational data remains the only feasible approach. Moreover, in resource limited settings such methods can be initially used to generate plausible causal hypotheses that can then be tested using experimental methods resulting in better utilization of available resources.

In our previous work we performed mechanistic annotation of the top-ranked disease-linked protein biomarkers using a knowledge-based approach by developing an interactive druggability profiling algorithm. See [30] for additional details. In our current work, we use a data-driven approach to discover causal relationships.

Computationally, researchers have typically focused on generating global models, that is, Bayesian network models consisting of all the variables in a domain. However, this approach may not scale to very large datasets, and local causal discovery methods such as the BLCD approach we used hold significant promise when dealing with very large biomedical and clinical datasets which are becoming increasingly common in healthcare settings.

Future Work: We plan to extend the BLCD algorithm and create a temporal version, BLCDt, to propose cause and effect relationships from temporal data in the neonatal sepsis, pediatric infections and other domains. Even with non-temporal data consisting of a large number of variables and/or records, learning global causal models is problematic. Hence learning global CBNs from temporal data to ascertain temporal causal relationships will succeed only in very small domains. Temporal data typically consist of multiple values with time stamps for a subset or all the variables with varying granularities. Temporal causal discovery endeavors such as BLCDt are needed to navigate the enormous search spaces of non-trivial temporal domains.

A novel feature of our causal modeling approach is the focus on the use of efficient and scalable Bayesian local causal discovery approaches that can be used with very large datasets with tens of thousands of variables and millions of records. Researchers have recently taken an approach to parallelize global causal discovery algorithms to solve the computational time complexity problem and the fGES algorithm is a step in this direction [31]. Likewise, the proposed BLCDt algorithm will also scale to temporal data of this magnitude because it combines local search with a dynamic moving window for modeling temporal data to propose causal influences across the temporal dimension. When the numbers of variables and/or records are less we also propose to use traditional Bayesian network learning algorithms or causal discovery algorithms that assume that all the variables are measured. Then we plan to use a post-processing algorithmic method developed by us called the post processing Y arc (PPYA) algorithm (see Box 2 in the Appendix) to output cause and effect relationships that are not confounded by unmeasured (hidden) variables.

Appendix:

The steps of the PPYA algorithm are given in Box 2.

Box 2: PPYA Algorithm (pronounced "papaya")

Input: A BN or essential graph G and a set of nodes **X** in G.
Output: A set of Y arcs denoted as **Y**.

```
Initialize set of YA as Y:= {}.
For each X ∈ X
DO
   /* Pa(X): Set of parents of X */
   Determine Pa(X) for X.
   If |Pa(X)| ≤ 1
      Continue /* Next iteration of DO */

   /* Ch(X): Set of children of X */
   Determine Ch(X) for X.
   If |Ch(X)| = 0
      Continue /* Next iteration of DO */

   /* Look for Y structure */
   For each pair of parents W1, W2 of X
   DO
      If W1 and W2 are adjacent then Continue
      For each child Z ∈ Ch(X)
      DO
         If (W1, Z) or (W2, Z) adjacent then Continue
         If (X → Z) ∉ Y
            Y:= Y ∪ {X → Z}
      OD
   OD
OD
Return Y
```


References:

1. Pearl J: **Causality: models, reasoning and inference**. Cambridge: Cambridge University Press; 2000.
2. Spirtes P, Glymour C, Scheines R: **Causation, Prediction, and Search**. Cambridge, MA: MIT Press; 2000.
3. Glymour C, Cooper GF (eds.): **Computation, Causation, and Discovery**. Cambridge, MA: MIT Press; 1999.
4. Ioannidis JPA, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J: **Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies**. *JAMA* 2001, **286**(7):821-830.
5. Benson K, Hartz AJ: **A comparison of observational studies and randomized controlled trials**. *The New England Journal of Medicine* 2000, **342**(25):1878--1886.
6. Rifai N, Gillette MA, Carr SA: **Protein biomarker discovery and validation: the long and uncertain path to clinical utility**. *Nature biotechnology* 2006, **24**(8):971-983.
7. Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD, Springer DL, Pounds JG: **Toward a Human Blood Serum Proteome analysis by multidimensional separation coupled with mass spectrometry**. *Molecular & Cellular Proteomics* 2002, **1**(12):947-955.
8. Fan R, Vermesh O, Srivastava A, Yen BK, Qin L, Ahmad H, Kwong GA, Liu C-C, Gould J, Hood L: **Integrated barcode chips for rapid, multiplexed analysis of proteins in microliter quantities of blood**. *Nature biotechnology* 2008, **26**(12):1373-1378.
9. Piliarik M, Bocková M, Homola J: **Surface plasmon resonance biosensor for parallelized detection of protein biomarkers in diluted blood plasma**. *Biosensors and Bioelectronics* 2010, **26**(4):1656-1661.
10. Seibert V, Ebert MP, Buschmann T: **Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery**. *Briefings in functional genomics & proteomics* 2005, **4**(1):16-26.
11. Stern E, Vacic A, Rajan NK, Criscione JM, Park J, Ilic BR, Mooney DJ, Reed MA, Fahmy TM: **Label-free biomarker detection from whole blood**. *Nature nanotechnology* 2010, **5**(2):138-142.
12. Whiteaker JR, Zhao L, Anderson L, Paulovich AG: **An automated and multiplexed method for high throughput peptide immunoaffinity enrichment and multiple reaction monitoring mass spectrometry-based quantification of protein biomarkers**. *Molecular & Cellular Proteomics* 2010, **9**(1):184-196.
13. Zhang H, Liu AY, Loriaux P, Wollscheid B, Zhou Y, Watts JD, Aebersold R: **Mass spectrometric detection of tissue proteins in plasma**. *Molecular & Cellular Proteomics* 2007, **6**(1):64-71.

14. Lyman GH, Moses HL: **Biomarker tests for molecularly targeted therapies—the key to unlocking precision medicine**. *New England Journal of Medicine* 2016, **375**(1):4-6.
15. Beinlich IA, Suermondt HJ, Chavez RM, Cooper GF: **The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks**. In: *Proceedings of the Second European Conference on Artificial Intelligence in Medicine.* London: Chapman and Hall; 1990: 247--256.
16. Andreassen S, Woldbye M, Falck B, Andersen SK: **MUNIN --- A causal probabilistic network for interpretation of electromyographic findings**. In: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann; 1987: 366--372.
17. Heckerman D, Horvitz E, Nathwani B: **Towards Normative Expert Systems: Part I The Pathfinder Project**. *Methods of Information in Medicine* 1992, **31**:90--105.
18. Pearl J, Verma T: **A Theory of Inferred Causation**. In: *Principles of Knowledge Prepresentation and Reasoning: Proceedings of the Second International Conference.* San Francisco, CA: Morgan Kaufmann; 1991: 441--452.
19. Spirtes P, Glymour C, Scheines R: **An Algorithm for Fast Recovery of Sparse Causal Graphs**. *Social Science Computer Review* 1991, **9**(1):62--72.
20. Spirtes P, Glymour C, Scheines R: **Causation, Prediction, and Search**. New York: Springer-Verlag; 1993.
21. Cooper GF, Herskovits E: **A Bayesian method for the induction of probabilistic networks from data**. *Machine Learning* 1992, **9**:309--347.
22. Heckerman D, Geiger D, Chickering DM: **Learning Bayesian Networks: The Combination of Knowledge and Statistical Data**. *Machine Learning* 1995, **20**(3):197--243.
23. Moore A, Wong W-K: **Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning**. In: *ICML: 2003*. 552-559.
24. Tsamardinos I, Brown LE, Aliferis CF: **The max-min hill-climbing Bayesian network structure learning algorithm**. *Machine Learning* 2006, **65**(1):31-78.
25. Mani S, Cooper GF: **Causal discovery using a bayesian local causal discovery algorithm**. In: *Proceedings of MedInfo.* Amsterdam: IOS; 2004: 731-735.
26. Mani S, Spirtes P, Cooper GF: **A theoretical study of Y structures for causal discovery**. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 2006: 314–323.
27. Doellner H, Arntzen KJ, Haereid PE, Aag S, Brubakk A-M, Austgulen R: **Increased serum concentrations of soluble tumor necrosis factor receptors p55 and p75 in early onset neonatal sepsis**. *Early human development* 1998, **52**(3):251-261.
28. Parayath KE, Harrison TS, Levitz SM: **Effect of interleukin (IL)-15 priming on IL-12 and interferon-γ production by pathogen-stimulated peripheral blood mononuclear cells from human immunodeficiency virus-seropositive and-seronegative donors**. *Journal of Infectious Diseases* 2000, **181**(2):733-736.
29. Perkins DJ, Were T, Davenport GC, Kempaiah P, Hittner JB, Ong'echa JM: **Severe malarial anemia: innate immunity and pathogenesis**. *International journal of biological sciences* 2011, **7**(9):1427.
30. Mani S, Cannon D, Ohls R, Oprea T, Mathias S, Ballard K, Ursu O, Bologa C: **Protein biomarker druggability profiling**. *Journal of Biomedical Informatics* 2017, **66**:241-247.
31. Ramsey J, Glymour M, Sanchez-Romero R, Glymour C: **A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images**. *International journal of data science and analytics* 2017, **3**(2):121-129.