
Causal Discovery in Bioinformatics

Ioannis Tsamardinos

Assistant Professor

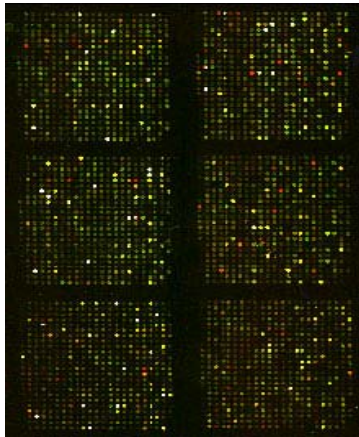
Computer Science Department, University of Crete

Inst. Comp. Sc., Foundation for Research and Technology -
Hellas

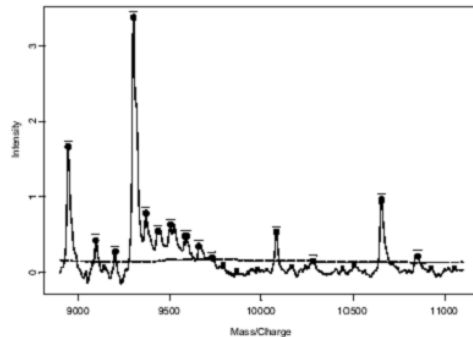
Acknowledgements

- The Biomedical Informatics Laboratory, ICS-FORTH
- Computer Science Department, University of Crete
- The Discovery Systems Laboratory, DBMI, Vanderbilt University
 - Constantin F. Aliferis, Director DSL, Assistant Prof. DBMI
 - Douglas Hardin, Associate Prof. Mathematics
 - Students/ Programmers
 - Laura E. Brown
 - Alexander Statnikov
 - Yerbolat Dosbayev
 - Support
 - NIH
 - Vanderbilt University

Data Analysis in Bioinformatics: Diagnosis and Prediction

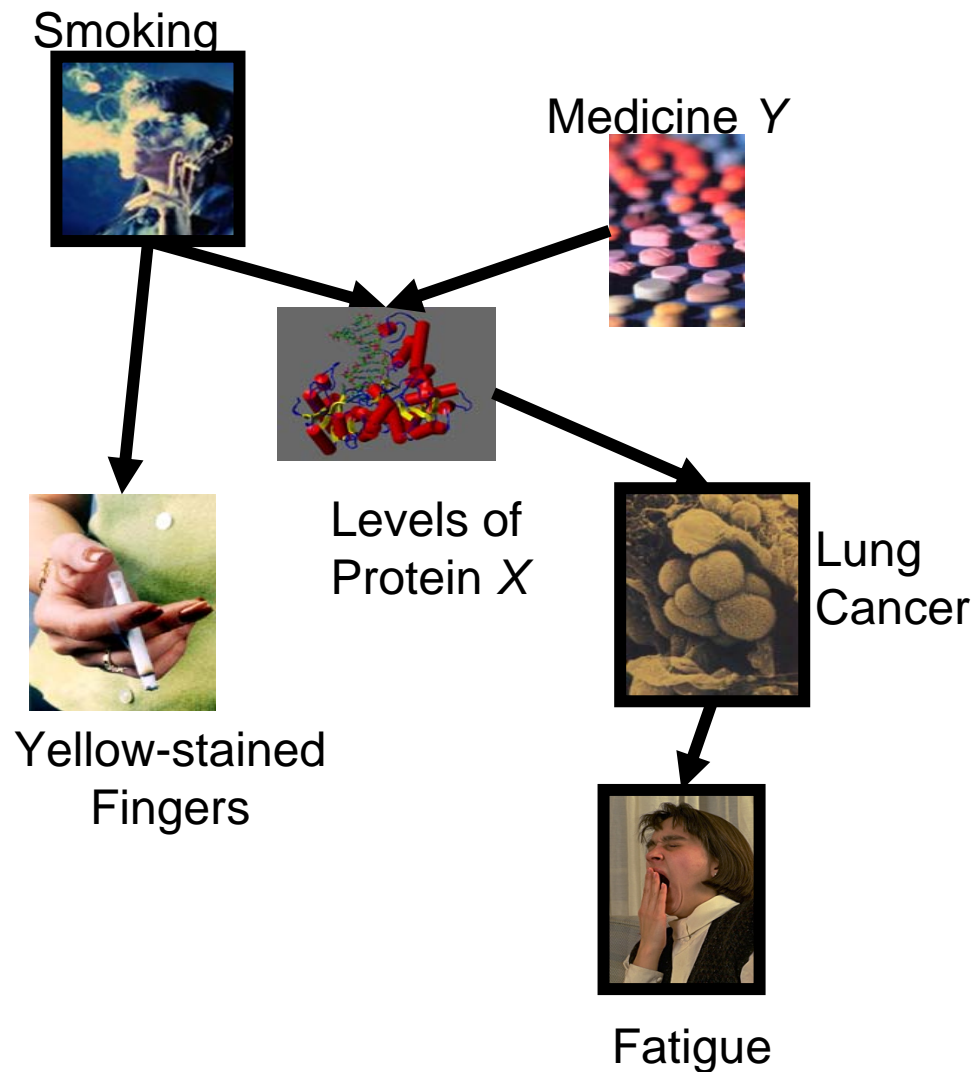


→ What is the diagnosis?



→ Is this an early cancer?

Causation



How to Discovery Causality

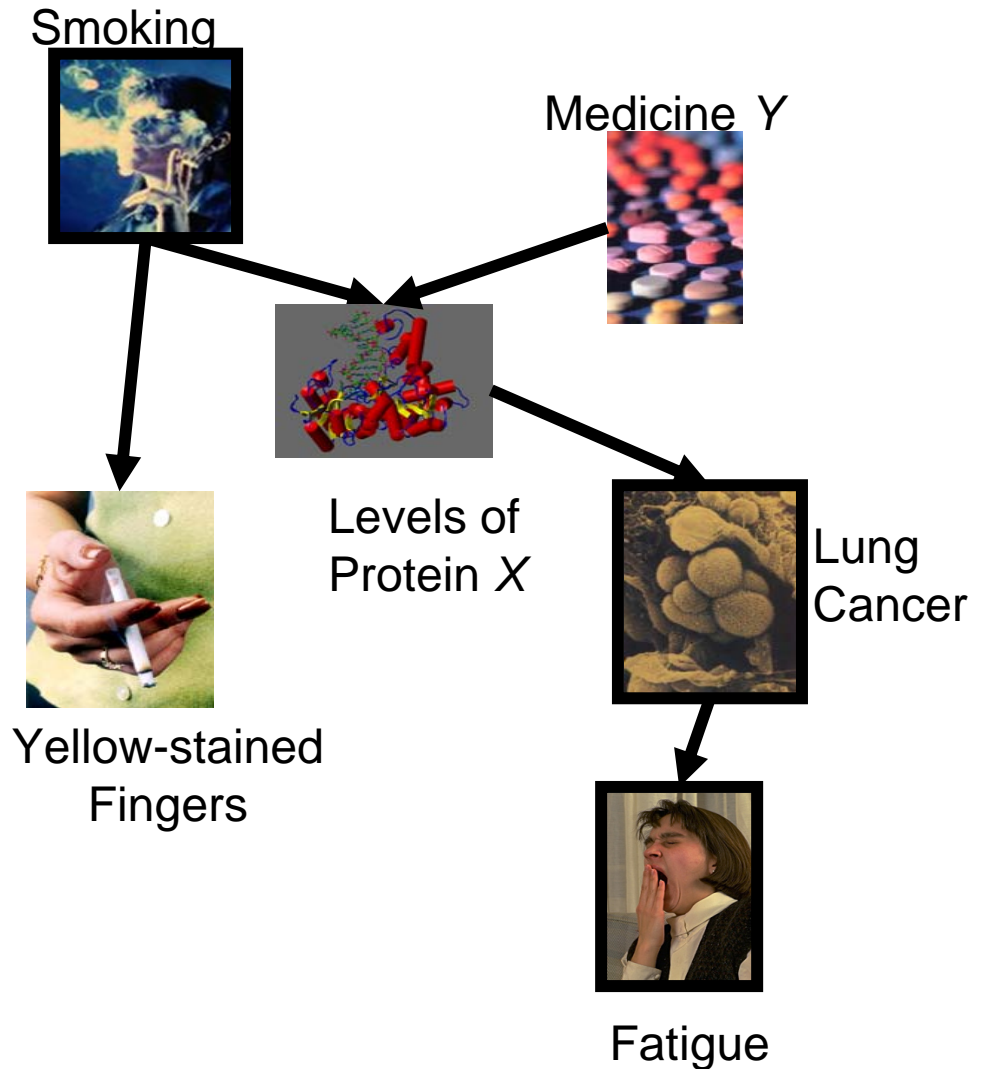
- Controlled experiments has been the typical scientific instrument for discovery of causality
- But ...
 - Unethical, costly, impossible, ...
- We need to be able to learn from observational data

Formal Computational Causal Discovery from Observational Data

- Formal algorithms exist for learning causal relations from observational data! (Spirtes, Scheines, Glymour, Pearl and others)
- Most are based on a graphical-probabilistic language called Causal Bayesian Networks
- Well-characterized properties of
 - What types of causal relations they can learn
 - Under which conditions
 - What kind of errors they may make

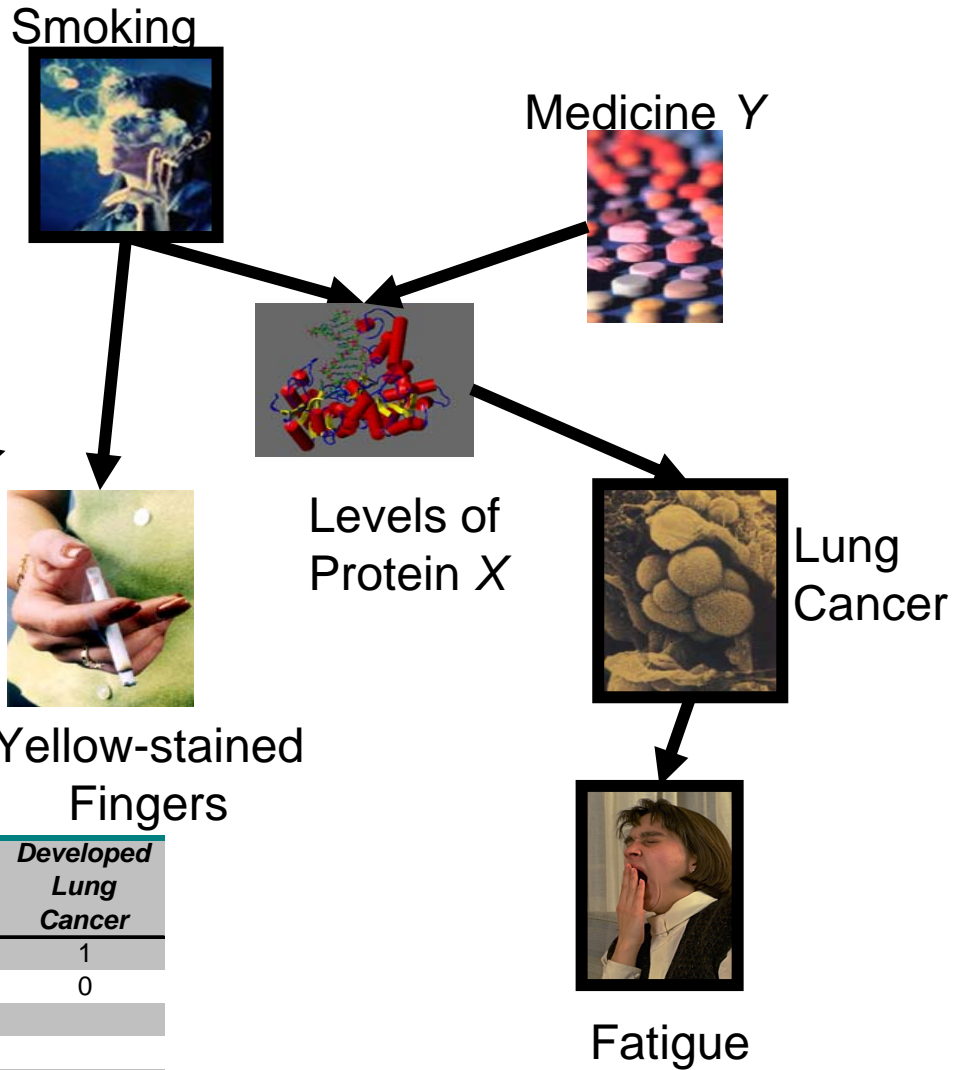
Causal Bayesian Networks

- Edges represent direct causal effects
- Probabilistic reasoning + causal inferences



Learning Bayesian Networks from Observational Data

Given the data below, identify the graph that best fits the (independencies in the) data



Patient #	Smokes	Has Yellow-stained fingers	Levels of Protein X	Takes Medicine Y	Feels Fatigue	Developed Lung Cancer
1	1	1	3.44	0	1	1
2	0	0	2.33	0	1	0
...						
m	0	0	1.92	0	0	0

Algorithmic Advances in Bayesian Network Learning

Never Say Never

- “In our view, inferring *complete* causal models (i.e., causal Bayesian Networks) is essentially impossible in large-scale data mining applications with thousands of variables”, Silverstein, Brin, Motwani, Ullman 2000

Methods that Learn Bayesian Networks from Data

- The Max-Min Hill Climbing algorithm, Tsamardinos, Brown, Aliferis, Machine Learning Machine Learning 65(1): 31-78 (2006)
- One of the most efficient and accurate algorithms for learning Bayesian Networks with discrete data

Learning Large Bayesian Networks

- True model contains 10,000 variables and their causal relations
- Dataset simulated from the model
- Training size: 1000 instances
- Algorithm: ~MMHC

- Results
- Sensitivity of edge detection: 81%
- Specificity of edge detection: 99%
- Time: 62hours, 2.4GHz Pentium IV
- Largest BN ever reconstructed
- Nothing to compare with on such a large dataset

Empirical Evaluation of MMHC

- Considered thousands of causal models (~8,500) of various dimensionality
- Generated datasets of various sizes from the models
- Reconstructed the models from the data using all prototypical and state-of-the-art Bayesian Network learning algorithms
- Compared the learnt with the true model

Empirical Evaluation of MMHC

Algorithm	Average Relative Number of Errors	Average Relative Time	Description
MMHC	100%	100%	Tsamardinos, et. al.
OR1, k=5	137%	134%	Optimal Reinsertion, Moore et al. Carnegie Mellon
OR1, k=10	140%	126%	
OR1, k=20	144%	126%	
OR2, k=5	137%	233%	
OR2, k=10	133%	235%	
OR2, k=20	135%	232%	
SC, k=5	142%	843%	Sparse Candidate, Nir Friedman et al.
SC, k=10	139%	1047%	
GTS	193%	703%	Greedy Tabu Search
GS	170%	424%	Greedy Search, Heckerman et al. Microsoft
PC	516%	8332%	Spirtes et al. Carnegie Mellon
TPDA	285%	489%	Three Phase Dependency Analysis
GES	125%	159036%	Greedy Equivalent Search, Chickering et. al, Microsoft

Theoretical and Algorithmic Advances in Causal-Based Feature Selection

Challenges for Learning in Biomedicine

- Emergence of extremely high-dimensionality datasets
 - Gene expression microarray data: (range 10K-60K variables)
 - Mass-spectroscopy (60K-65K)
 - Chemical structural properties (140K)
 - Text-categorization (10K-20K)
 - Single Nucleotide Polymorphisms (>100K)
- Small number of training examples (often)

Feature Selection As a Solution to High-Dimensional Analysis

- Reduce the number of required observed quantities (variables/features) to build a predictive/diagnostic model
- Definition: Select the variable subset of minimal size with the maximal predictive or diagnostic, classification power for target variable T

Why Feature Selection

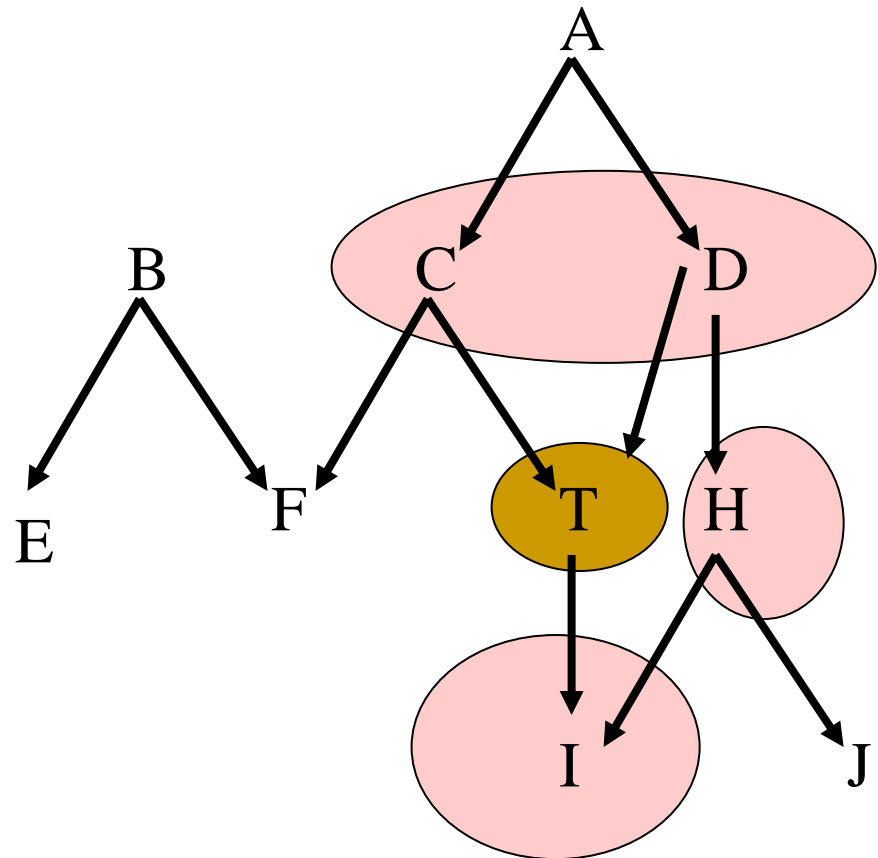
- Increases efficiency of learning
- Reduces cost and risk of observing the variables
- Increases time-efficiency of classification
- Increases human understanding of the model
- *Increases understanding of the domain*
 - *“identification of biomarkers”*
 - *or not? how?*

Causality and Feature Selection

- The smallest subset with the optimal predictive power is the set of
 - Parents (direct causes)
 - Children (direct effects)
 - Spouses (direct causes of the direct effects)of the target variable (to predict) in the network that fits the data
- This set is called the Markov Blanket of the target
- Connections among Bayesian Networks, Markov Blanket, Feature Selection, Relevant Variable, and more
 - Tsamardinos, Aliferis, AI&Stats 2003

Causality and Feature Selection

- The Markov Blanket of T is:
 - Parents (direct causes)
 - Children (direct effects)
 - Spouses (direct causes of the direct effects)
- in the causal network
- Knowing the values of the Markov Blanket variables renders knowledge of the values of all other variables superfluous

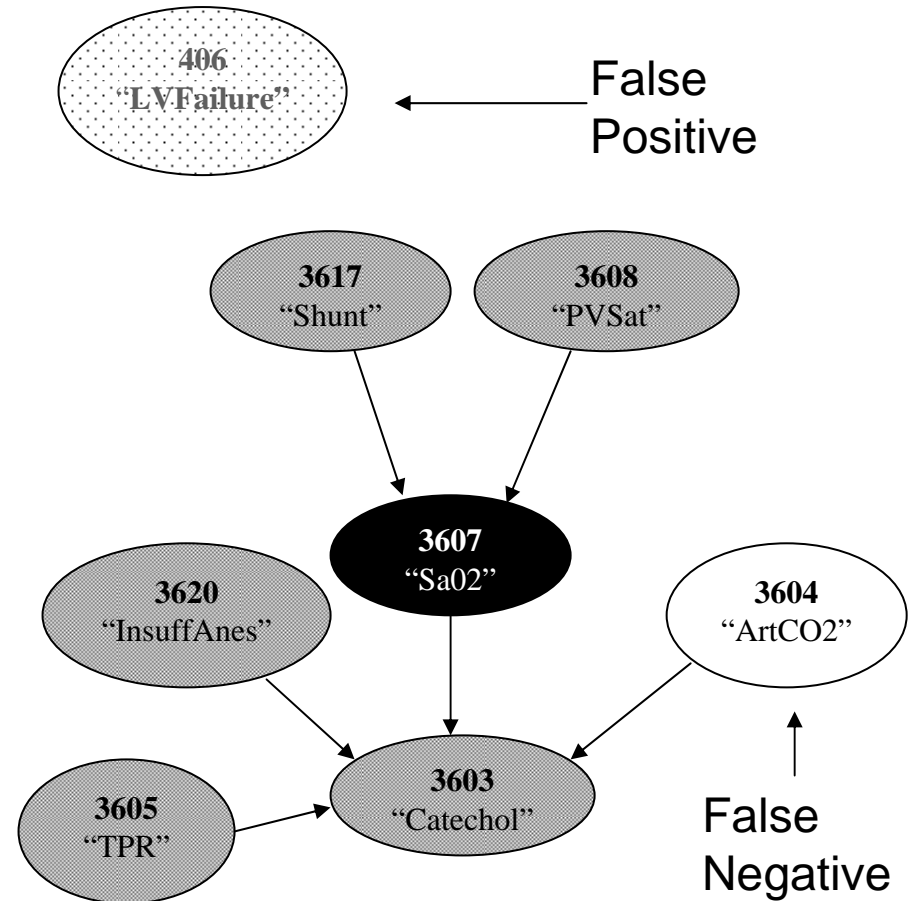


Causal-Based Feature Selection

- Identify the Markov Blanket of the target using methods based on causal theories
- Use the Markov Blanket variables to build the final predictive or diagnostic models
- Design efficient and accurate algorithms that identify the MB without having to learn the whole network
 - Max-Min Markov Blanket
 - Tsamardinos, Aliferis, Statnikov, KDD 2003
 - HITON
 - Aliferis, Tsamardinos, Statnikov, AMIA 2003
 - and many variants

Example of Performance in Markov Blanket Discovery

- True Model contains 5000 variables
- Data generated from true model
- Algorithm MMMB estimated the Markov Blanket of all variables
- In the picture: node 3607 in black is the node with the average specificity and sensitivity in Markov Blanket identification



Variable selection experiments with real data and HITON

Dataset	Thrombin	Arrythmia	Ohsumed	Lung Cancer	Prostate Cancer
Problem Type	Drug Discovery	Clinical Diagnosis	Text Categorization	Gene Expression Diagnosis	Mass-Spec Diagnosis
Variable #	139,351	279	14,373	12,600	779
Variable Types	binary	nominal/ordinal /continuous	binary and continuous	continuous	continuous
Target	binary	nominal	binary	binary	binary
Sample	2,543	417	2000	160	326
Vars-to-Sample	54.8	0.67	7.2	60	2.4
Evaluation metric	ROC AUC	Accuracy	ROC AUC	ROC AUC	ROC AUC
Design	1-fold c.v.	10-fold c.v.	1-fold c.v.	5-fold c.v.	10-fold c.v.

Figure 2: Dataset Characteristics

- [C. F. Aliferis, I. Tsamardinos, A. Statnikov. “**HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection**”, AMIA 2003]

1. Drug Discovery (Thrombin)				
	UAF*	RFE	HITON	ALL
SVM	96.12%	93.29%	93.23%	93.69%
KNN	87.25%	89.71%	92.23%	88.21%
NN	<i>NA</i>	92.04%	92.65%	<i>NA</i>
Average	91.69%	91.68%	92.7%	90.95%
# of variables	34837	8709	32	139351
2. Clinical Diagnosis (Arrhythmia)				
	UAF*	B/F*	HITON*	ALL*
DTI	73.94%	72.85%	71.87%	73.94%
KNN	63.22%	63.45%	65.30%	63.22%
NN	58.29%	60.90%	60.38%	58.29%
Average	65.15%	65.73%	65.85%	65.15%
# of variables	279	96	63	279
3. Text Categorization (OHSUMED)				
	IG	X ²	HITON	ALL*
SVM	82.43%	85.91%	82.85%	90.50%
SBCtc	84.18%	86.23%	85.10%	84.25%
KNN	75.55%	81.76%	80.25%	77.56%
NN	82.47%	85.27%	83.97%	<i>NA</i>
Average	81.16%	84.79%	83.04%	84.10%
# of variables	224	112	34	14373

4. Gene Expression Diagnosis (Lung Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	99.32%	98.57%	97.83%	99.07%
NN	99.63%	98.70%	98.92%	N/A
KNN	95.57%	91.49%	96.06%	97.59%
Average	98.17%	96.25%	97.60%	98.33%
# of variables	330	19	16	12,600
5. Mass-Spectrometry Diagnosis (Prostate Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	98.50%	98.95%	99.10%	99.40%
NN	98.62%	98.78%	97.95%	99.27%
KNN	77.52%	86.53%	91.36%	76.94%
Average	91.55%	94.75%	96.14%	91.87%
# of variables	706	87	16	779
Averages Over All Tasks				
	Av. Over Baseline Algorithms	HITON	ALL	
Av. Perf. over classifiers	86.1%	87.1%	86.1%	
Av. variable #	4540	32.3	33,476	
Av. reduction	x 8	x 1124	x 1	

Figure 3: Task-specific and average model reduction performance (in bold, best performance per row; asterisks indicate that the corresponding algorithm yield the best model or a non-statistically significantly worse model than the best one).

Dissemination

Causal Explorer

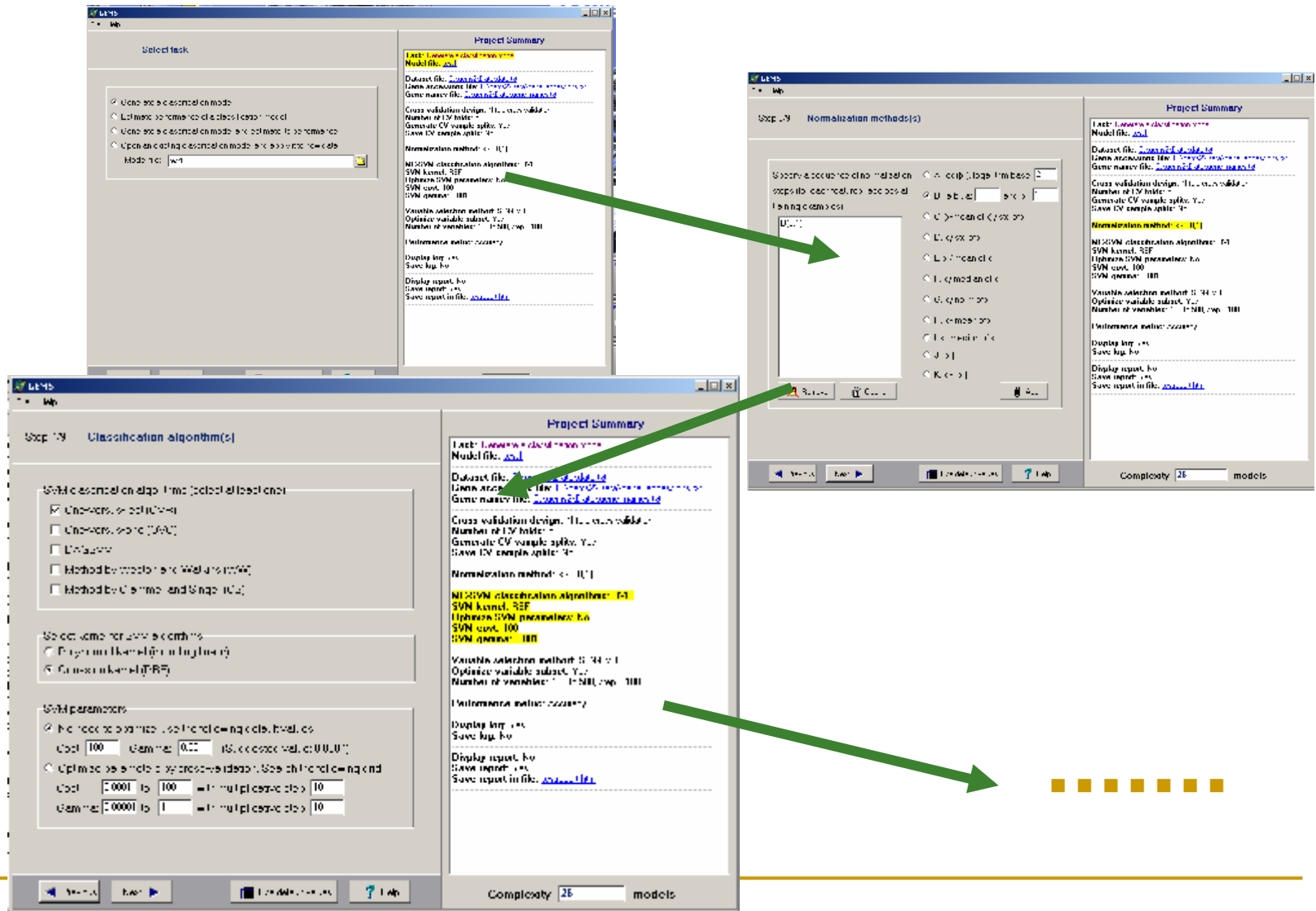
- A library of algorithms with a common, comprehensible, easy-to-learn API
- Algorithms for
 - Learning a Bayesian network that fits the data
 - Learning the graph that captures all variable (conditional) dependencies or independences
 - Learning the graph of all pair-wise plausible causal relations among the variables
 - Local network learning
 - Learning the local neighborhood of the BN capturing the data around a target variable of interest
 - Learning the direct causes and direct effects of a target variable of interest
 - Variable Selection
 - Selecting the variable subset of minimal size that can provide the most accurate predictive/diagnostic/classification model
- [Constantin F. Aliferis, Ioannis Tsamardinos, Alexander Statnikov, Laura E. Brown. “**Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery**”, *METMBS '03*]

GEMS[©]

Gene Expression Model Selector

- System that creates, evaluates, and applies models for array gene expression cancer diagnosis or outcome prediction and biomarker discovery in a fully-automated fashion.
- Wizard-like interface
- Numerous methods for classification and variable selection
- Automatic parameter optimization for the classifiers and optimization of variable selection method
- Enforces non-overfitting protocols (a major problem in bioinformatics)
- Nomination for Best Student Paper in Medinfo 2004 (Alex Statnikov)
- Best poster award, ISMB 2005

GEMS2[©]: Wizard Interface



Current Limitations of Causal Discovery

- We can determine the best causal model up to statistical indistinguishability
- Assumes no cycles in causality
- Does not model time
 - Extensions to Dynamic Bayesian Networks (they can handle causal cycles too)
- Assumes no hidden variables
 - Extensions that discover causality with hidden variables exist but are inefficient
- Assumes Faithfulness
 - ... meaning that if A causes B, there should be pairwise association between them
- Assumes we have enough data to have enough statistical power to detect associations
- Research on the subject is picking up!

Conclusions

- The emergence of high-dimensional datasets presents new exciting possibilities and challenges to analyses methods
- Ultimate goal is causal knowledge so that one can intervene and manipulate a system

Conclusions

- Theoretical advances in variable selection connect the Markov Blanket, Bayesian Networks, causal formalisms, notions of relevancy and redundancy together
- Algorithmic advantages lead to efficient and accurate algorithms for Markov Blanket identification, variable selection and Causal Bayesian Network learning
 - MMHC one of the best algorithms for learning Bayesian Networks
 - MMMB and HITON some of the best feature selection algorithms
 - Variables selected have a causal interpretability

Conclusions

- Systems Causal Explorer and GEMS encapsulate the algorithms and automate machine learning experimentation
- Promising future directions in causal discovery, variable selection, and Bayesian Network learning in biomedicine from extremely high-dimensional datasets
- Interest is increasing: NIPS Workshop on Causal Discovery, JMRL special issue Causal Discovery, publications, etc.

Research in the Bioinformatics Laboratory, ICS-FORTH

1. Novel (biomedically motivated) methods in Machine Learning and Data Analysis
2. Theory of Machine Learning
3. Analyses for answering specific biomedical questions
4. Systems and tools to facilitate analysis

References

- Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, “**Challenges in the Analysis of Mass-Throughput Data: A Technical Commentary from the Perspective of Statistical Machine Learning**”, *Cancer Informatics*. 2006; 2: 133–162.
- I. Tsamardinos, L.E. Brown, C.F. Aliferis. “**The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm**”, *Machine Learning Journal*; 65: 31-78
- Laura E. Brown, Ioannis Tsamardinos, Constantin F. Aliferis, “**A Comparison of Novel and State-of-the-Art Polynomial Bayesian Network Learning Algorithms**”, in the *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*, pp. 739-745, 2005
- Alexander Statnikov, Ioannis Tsamardinos, Yerbolat Dosbayev, Constantin F. Aliferis, “**GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data**”, *International Journal of Medical Informatics*, 74(7-8):491-503, 2005
- Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardin, Shawn Levy, “**A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis**”, in *Bioinformatics* 21(5):631-643, 2005
- Douglas Hardin, Ioannis Tsamardinos, Constantin F. Aliferis, “**A Theoretical Characterization of Linear SVM-Based Feature Selection**”, in *The Twenty-First International Conference on Machine Learning (ICML 2004)*, 2004
- Laura E. Brown, Ioannis Tsamardinos, Constantin F. Aliferis, “**A Novel Algorithm for Scalable and Accurate Bayesian Network Learning**”, in *Proceedings of 11th World Congress in Medical Informatics (MEDINFO '04)*, 2004.
- Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, “**Methods for Multi-Category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development**”, *Proceedings of 11th World Congress in Medical Informatics (MEDINFO '04)*, 2004, Gold Medal in the Student Paper Competition
- C. F. Aliferis, I. Tsamardinos, A. Statnikov. “**HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection**”, in the *American Medical Informatics Association meeting 2003 (AMIA 2003)*
- I. Tsamardinos, C.F. Aliferis, A. Statnikov. “**Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations**”, in *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003)*, p. 673-678
- Constantin F. Aliferis, Ioannis Tsamardinos, Alexander Statnikov, Laura E. Brown. “**Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery**”, *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*, p. 371-376.
- C. F. Aliferis, I. Tsamardinos, P. Massion, A. Statnikov, D. Hardin. “**Why Classification Models Using Array Gene Expression Data Perform So Well: A Preliminary Investigation Of Explanatory Factors**”, *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*, 47-53.
- Ioannis Tsamardinos, Constantin F. Aliferis, “**Towards Principled Feature Selection: Relevancy, Filters, and Wrappers**”, *Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, January, 2003 (AI&Stats 2003)*.

End