Methodological Aspects of Identifying Biomarkers:The Case of Microarray-Based Lung Cancer Diagnostics

Constantin F. Aliferis M.D., Ph.D.,
Discovery Systems Laboratory,
Department of Biomedical Informatics,
Vanderbilt University

Lung SPORE Research Conference 10-23-2003

Acknowledgments/Collaborators
(alphabetic order)

- D. Hardin Ph.D.
- P. Massion M.D.
- A. Statnikov M.S.
- I. Tsamardinos Ph.D.

Focus of this talk: Some Challenges in Using Array Gene Expression data and Data Analysis to Discover Potential Biomarkers

- Biomarkers = measurable molecular phenotypic parameters that characterize an organism's state of health or disease, or a response to a particular therapeutic intervention (Negm et al. Trends Mol Med, 2002)
- Biomarkers are sought as instruments to help in
  - disease risk assessment,
  - early disease detection, and
  - as surrogate endpoints in clinical trials (or in some cases as surrogates for environmental and other exogenous factors such as diet)

Biomarkers (cont'd)

- Phases in establishing/validating Biomarkers (EDRN, modified)
  - **Identify candidates**
  - **Clinical assays to diagnose known disease**
  - Detection of pre-clinical disease (pseudo-prospectively) & establishment of screen-positive rule
  - Prospective screening, establish extend and characteristics of identified disease as well as false referral rates
  - Quantification of overall impact on disease

- Many currently predominant assaying technologies for biomarker detection, e.g.:
  - Gene expression: SAGE, RT-PCR, NB, **MicroArray**, etc.
  - Proteomic: 2D PAGE – MALDI MS, SELDI-MS, LC-MSMS, Ab arrays, Tissue arrays, etc.

- Proposition 1: there is something fundamentally wrong with the way biomarker detection is being pursued from the data analysis perspective
- Proposition 2: there are better ways to think about candidate biomarker discovery

- …I will discuss evidence from previous and ongoing experiments…

BISTI Planning grant Pilot Project

Project title:

Computational Models of Lung Cancer:

Connecting Classification, Gene Selection, and Molecular Sub-typing

C.Aliferis M.D., Ph.D., Pierre Massion M.D.
I. Tsamardinos Ph.D., D. Hardin Ph.D.

Pilot Project: Goals Year #1

- **Specific Aim 1:** *"Construct computational models that distinguish between important cellular states related to lung cancer, e.g., (i) Cancerous vs Normal Cells; (ii) Metastatic vs Non-Metastatic cells; (iii) Adenocarcinomas vs Squamous carcinomas".*
- **Specific Aim 2:** *"Reduce the number of gene markers by application of biomarker (gene) selection algorithms such that small sets of genes can distinguish among the different states (and ideally reveal important genes in the pathophysiology of lung cancer)."*

Lung Cancer: Data & Methods

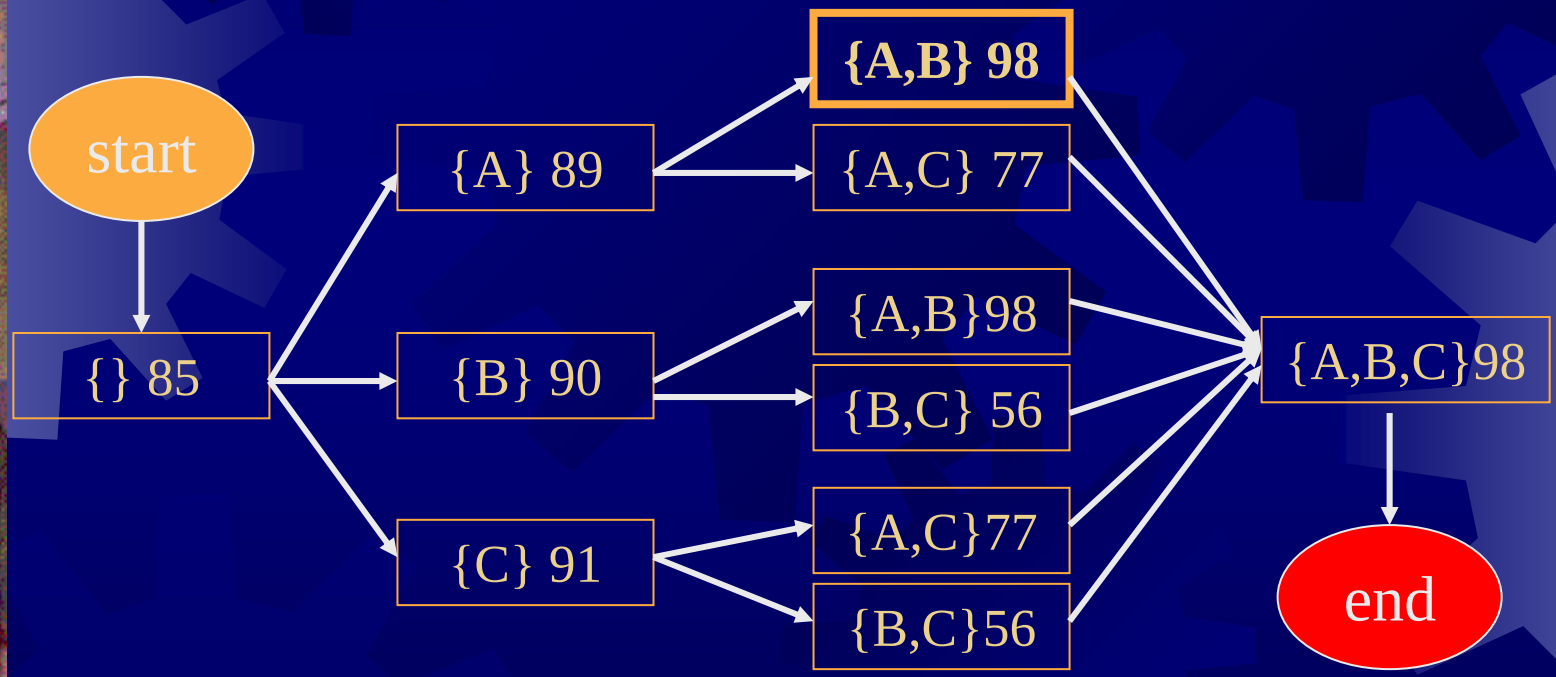
- Bhattacharjee et al. PNAS, 2001
- 12,600 gene expression measurements obtained using Affymetrix oligonucleotide arrays
- 203 patients and normal subjects, 5 disease types, (plus staging and survival information)
- Nested cross-validation: one level to optimize classifier parameters and one level to estimate performance

Brief detour:
Introduction to feature (a.k.a. biomarker) selection

Suppose we have predictors A, B, C and classifier *M*. We want to predict T given the smallest possible subset of {A,B,C}, while achieving maximal classification performance

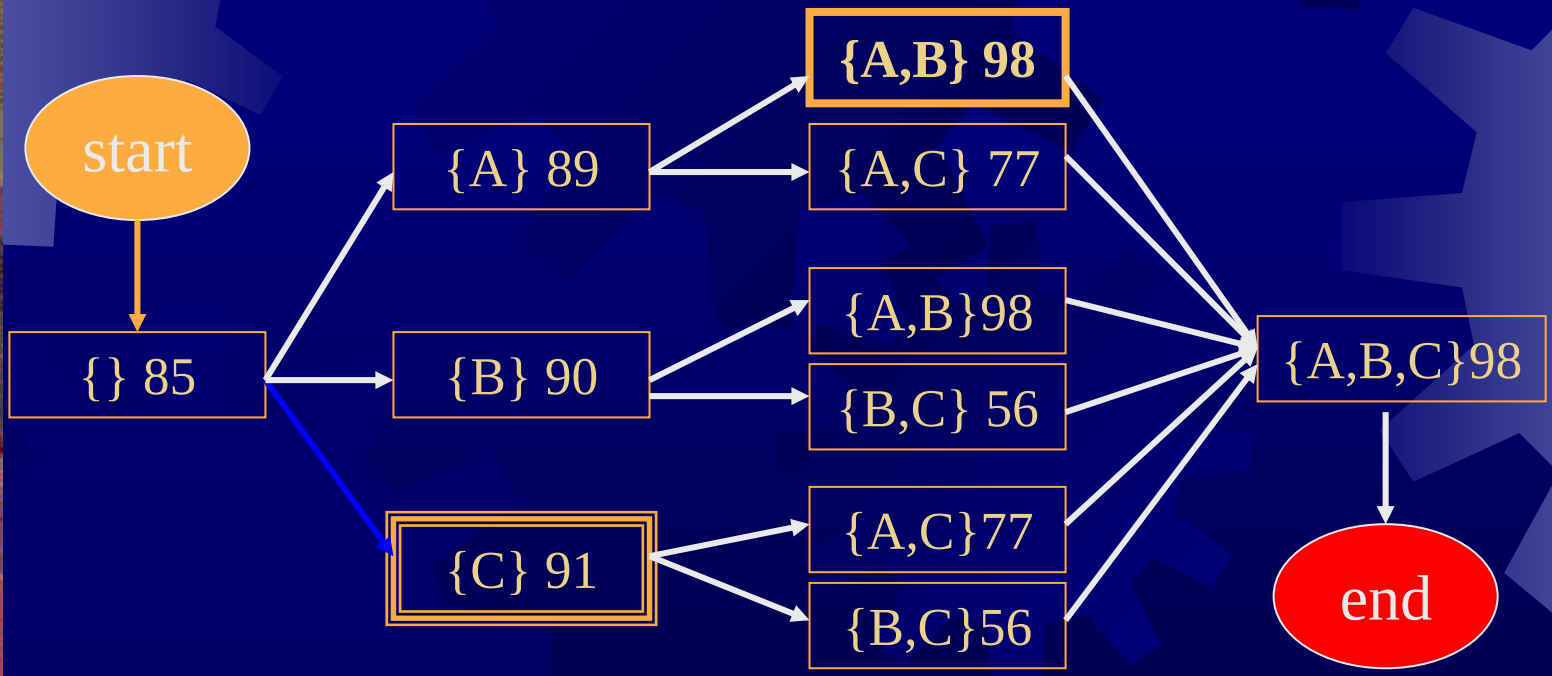
FEATURE SET	CLASSIFIER	PERFORMANCE
{A,B,C}	*M*	**98%**
{A,B}	*M*	**98%**
{A,C}	*M*	77%
{B,C}	*M*	56%
{A}	*M*	89%
{B}	*M*	90%
{C}	*M*	91%
{.}	*M*	85%

“Wrapper” approach: Search over all possible subsets



Search is expensive!

The set of all subsets is the power set and its size is $2^{|V|}$. Hence for large V we cannot do this procedure exhaustively; instead such methods employ *heuristic search* of the space of all possible feature subsets. A common example of heuristic search is hill climbing: keep adding features one at a time until no further improvement can be achieved.



“Filter” approach: look at probability distribution

In the filter approach we do not rely on running a particular classifier and searching in the space of feature subsets; instead we select features on the basis of statistical properties. A classic example is univariate association filtering (UAF):

FEATURE	ASSOCIATION WITH TARGET	
{A}	89%	Threshold gives suboptimal solution
{B}	90%	Threshold gives optimal solution
{C}	91%	Threshold gives suboptimal solution

Characteristic Biomarker Selection Methods in Bioinformatics: Univariate Association Filtering (UAF)

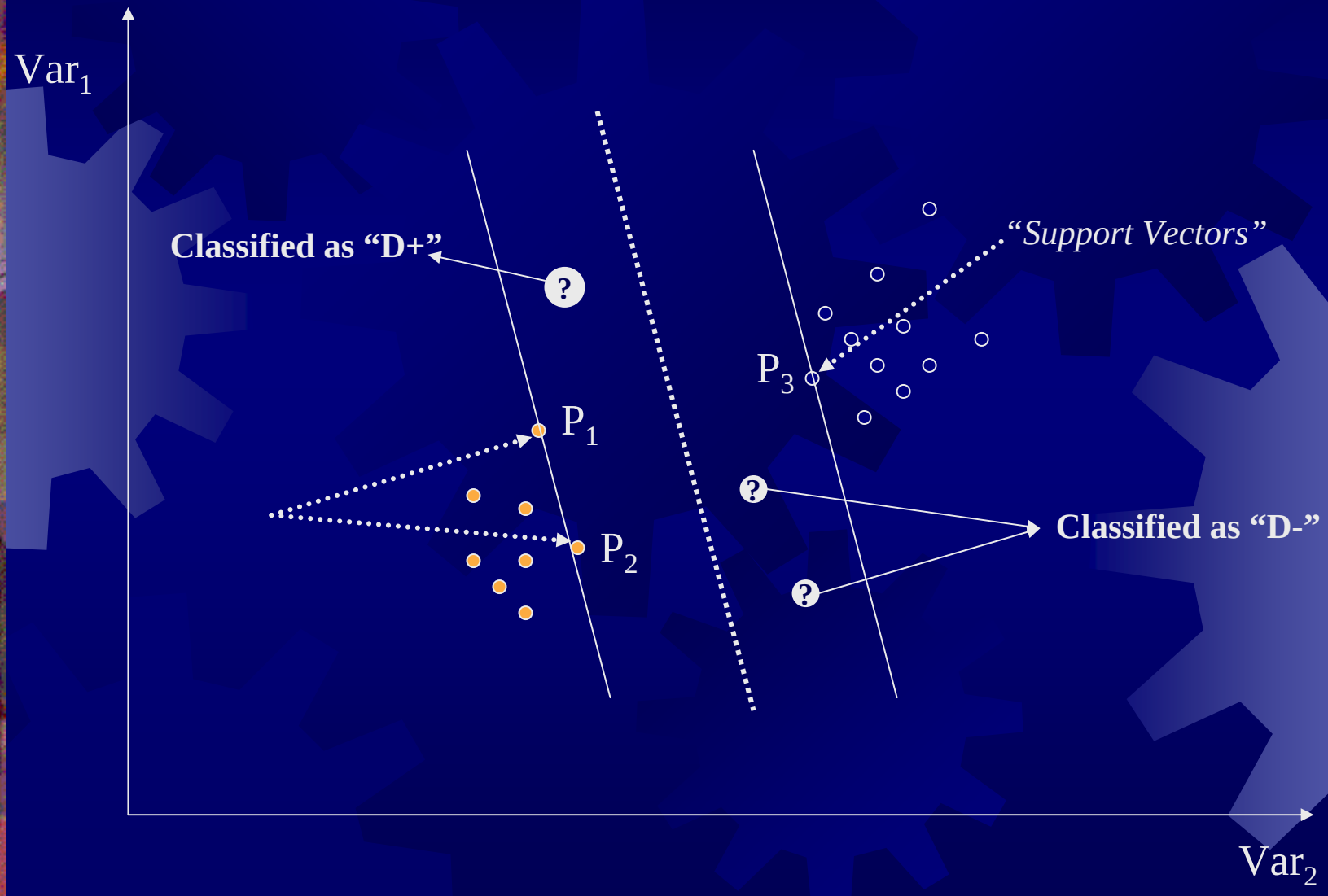
- Order all predictors according to strength of association with target
- Choose the first *k* predictors and feed them to the classifier
- Various measures of association may be used: X^2, G^2, Pearson r, Fisher Criterion Scoring, signal-to-noise ratio, etc.

Characteristic Biomarker Selection Methods in Bioinformatics: Recursive Feature Elimination (REF)

- Filter algorithm where feature selection is done as follows:

1. build linear Support Vector Machine classifiers using *V* features
2. compute weights of all features and choose the best *V/2*
3. repeat until 1 feature is left
4. choose the feature subset that gives the best performance
5. give best feature set to the classifier of choice.

Support Vector machines



Classification Performance

Cancer vs normal				Adenocarcinomas vs squamous carcinomas			Metastatic vs non-metastatic adenocarcinomas		
classifiers	RFE	UAF	All Features	RFE	UAF	All Features	RFE	UAF	All Features
LSVM	97.03%	99.26%	99.64%	98.57%	99.32%	98.98%	96.43%	95.63%	96.83%
PSVM	97.48%	99.26%	99.64%	98.57%	98.70%	99.07%	**97.62%**	**96.43%**	96.33%
KNN	87.83%	97.33%	98.11%	91.49%	95.57%	97.59%	92.46%	89.29%	92.56%
NN	97.57%	**99.80%**	N/A	98.70%	**99.63%**	N/A	96.83%	86.90%	N/A
Averages over classifier	94.97%	**98.91%**	99.13%	96.83%	**98.30%**	98.55%	**95.84%**	92.06%	95.24%

Gene Selection Analysis: Parsimony

Feature Selection Method	Number of features discovered		
	Cancer vs normal	Adenocarcinomas vs squamous carcinomas	Metastatic vs non-metastatic adenocarcinomas
RFE	6	12	**6**
UAF	**100**	**500**	500

Gene Selection Analysis: Novelty

	Cancer vs normal		Adenocarcinomas vs squamous carcinomas		Metastatic vs non-metastatic adenocarcinomas	
Contributed by method on the left compared with method on the right	RFE	UAF	RFE	UAF	RFE	UAF
RFE	0	2	0	5	0	2
UAF	96	0	493	0	496	0

Pilot Project: Goals Year #1

- **Specific Aim 3:** ***"Study how aspects of experimental design (including data set, measured genes, sample size, cross-validation methodology) determine the performance and stability of several machine learning (classifier and feature selection) methods used in the experiments".***

Explanatory Factors

- ***Overfitting***: we replace actual gene measurements by random values in the same range (while retaining the outcome variable values).
- ***Target class rarity***: we contrast performance in tasks with rare vs non-rare categories.
- ***Sample size***: we use samples from the set {40,80,120,160, 203} range (as applicable in each task).
- ***Predictor info redundancy***: we replace the full set of predictors by random subsets with sizes in the set {500, 1000, 5000, 12600}.

Explanatory Factors (CONT'D)

- ***Train-test split ratio****:* we use train-test ratios from the set {80/20, 60/40, 40/60} (for tasks II and III, while for task I modified ratios were used due to small number of positives, see Figure 1).
- ***Cross-validated fold construction***: we construct n-fold cross-validation samples retaining the proportion of the rarer target category to the more frequent one in folds with smaller sample, or, alternatively we ensure that all rare instances are included in the union of test sets (to maximize use of rare-case instances).
- ***Classifier type***: Kernel vs non-kernel and linear vs non-linear classifiers are contrasted. Specifically we compare linear and non-linear SVMs (a prototypical kernel method) to each other and to KNN (a robust and well-studied non-kernel classifier and density estimator).

Results:Area under the ROC curve with random predictor values

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.584 (139 cases)	0.583 (203 cases)	0.572 (160 cases)
KNN	0.581 (139 cases)	0.522 (203 cases)	0.559 (160 cases)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.968 (139 cases)	0.996 (203 cases)	0.990 (160 cases)
KNN	0.926 (139 cases)	0.981 (203 cases)	0.976 (160 cases)

Results:Area under the ROC curve when varying sample size

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.982 (40 cases) , 0,982 (80 cases), 0.969 (120 cases)	1 (40 cases) , 1 (80 cases), 1 (120 cases), 0.995 (160 cases)	0.981 (40 cases) , 0.988 (80 cases), 0.980 (120 cases)
KNN	0.893 (40 cases) , 0,832 (80 cases), 0.925 (120 cases)	1 (40 cases) , 1 (80 cases), 0.993 (120 cases), 0.970 (160 cases)	0.916 (40 cases) , 0.960 (80 cases), 0.965 (120 cases)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.968 (139 cases)	0.996 (203 cases)	0.990 (160 cases)
KNN	0.926 (139 cases)	0.981 (203 cases)	0.976 (160 cases)

Results:Area under the ROC curve with random gene set selection of varying size

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)-Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.944 (500 genes), 0.948 (1000 genes), 0.956 (5000 genes)	0.991 (500 genes), 0.989 (1000 genes), 0.995 (5000 genes)	0.982 (500 genes), 0.987 (1000 genes), 0.990 (5000 genes)
KNN	0.893 (500 genes), 0.893 (1000 genes), 0.941 (5000 genes)	0.959 (500 genes), 0.961 (1000 genes), 0.984 (5000 genes)	0.928 (500 genes), 0.955 (1000 genes), 0.965 (5000 genes)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)-Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.968 (139 cases)	0.996 (203 cases)	0.990 (160 cases)
KNN	0.926 (139 cases)	0.981 (203 cases)	0.976 (160 cases)

Results:Area under the ROC curve when varying train-test sample ratio

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.915 (30/70), 0.938 (43/57), 0.954 (57/43), 0.962 (70/30), 0.968 (85/15)	0.997 (40/60), 0.996 (60/40), 0.996 (80/20)	0.989 (40/60), 0.990 (60/40), 0.990 (80/20)
KNN	0.782 (30/70), 0.833 (43/57), 0.866 (57/43), 0.901 (70/30), 0.990 (85/15)	0.960 (40/60), 0.962 (60/40), 0.976 (80/20)	0.960 (40/60), 0.962 (60/40), 0.976 (80/20)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.968 (139 cases)	0.996 (203 cases)	0.990 (160 cases)
KNN	0.926 (139 cases)	0.981 (203 cases)	0.976 (160 cases)

Results:Area under the ROC curve with alternative strategy for constructing cross-validation splits (i.e., use of all rare-category instances)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.890 (40 cases) , 0,993 (80 cases), 0.965 (120 cases)	1 (40 cases) , 1 (80 cases), 1 (120 cases), 0.995 (160 cases)	1 (40 cases) , 1 (80 cases), 0.985 (120 cases)
KNN	0.918 (40 cases) , 0,849 (80 cases), 0.80 (120 cases)	1 (40 cases) , 0.96 (80 cases), 0.972 (120 cases), 0.982 (160 cases)	0.992 (40 cases) , 0.960 (80 cases), 0.990 (120 cases)

Classifier	Task I. Metastatic (7) – Nonmetastatic (132)	Task II. Cancer (186)- Normal (17)	Task III. Adenocarcinomas (139) - Squamous carcinomas (21)
SVMs	0.982 (40 cases) , 0,982 (80 cases), 0.969 (120 cases)	1 (40 cases) , 1 (80 cases), 1 (120 cases), 0.995 (160 cases)	0.981 (40 cases) , 0.988 (80 cases), 0.980 (120 cases)
KNN	0.893 (40 cases) , 0,832 (80 cases), 0.925 (120 cases)	1 (40 cases) , 1 (80 cases), 0.993 (120 cases), 0.970 (160 cases)	0.916 (40 cases) , 0.960 (80 cases), 0.965 (120 cases)

What is wrong with this picture?

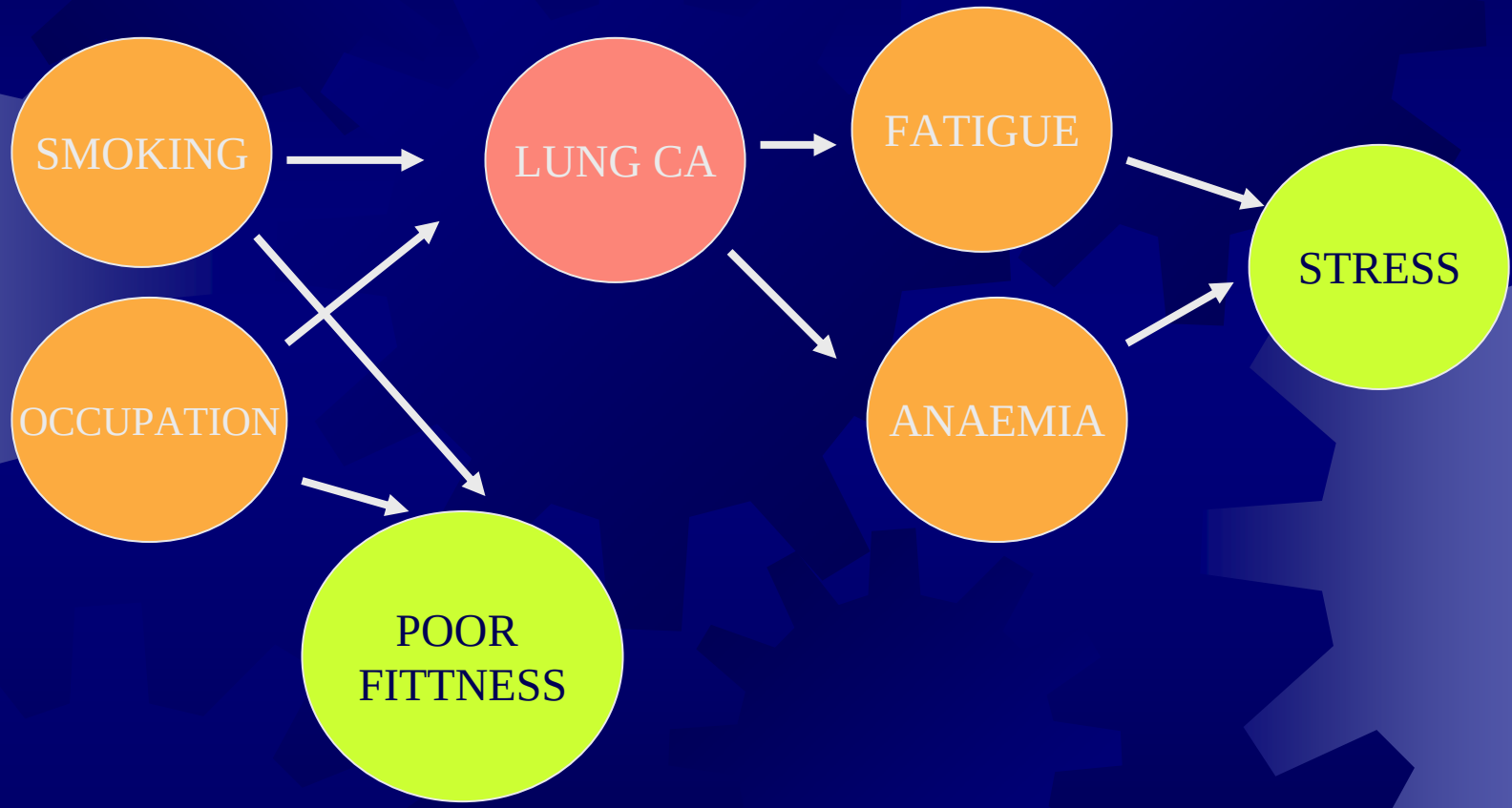
- If so many genes are excellent predictors of each disease type, which ones are the most promising genes to focus subsequent experimentation on for diagnostics and therapeutics?

…in other words, which are our best biomarker candidates?….

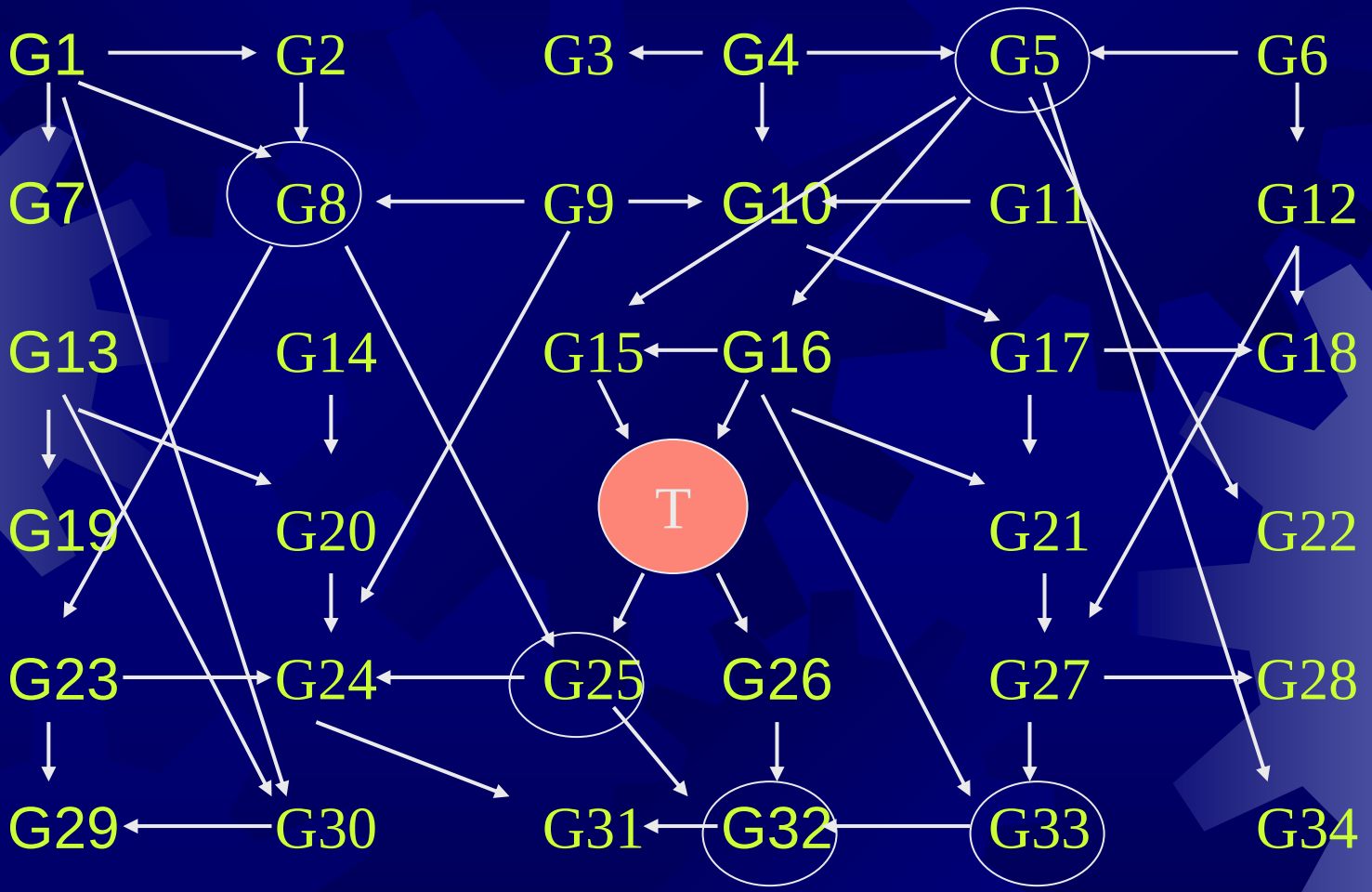
Questions

- How can it be that all these gene sets on one hand are excellent diagnostic predictors, and on the other hand are non-overlapping?
- Is it possible that are not directly related to the most interesting aspects of mechanism of disease?
- Different data analysis methods have different *inductive biases* (i.e., preferences for one type of model over the rest) and are optimized for different tasks. We need to interpret data analysis results in the context of the inductive biases of each method.

Explanation #1: Frequent Confusion of causality with prediction



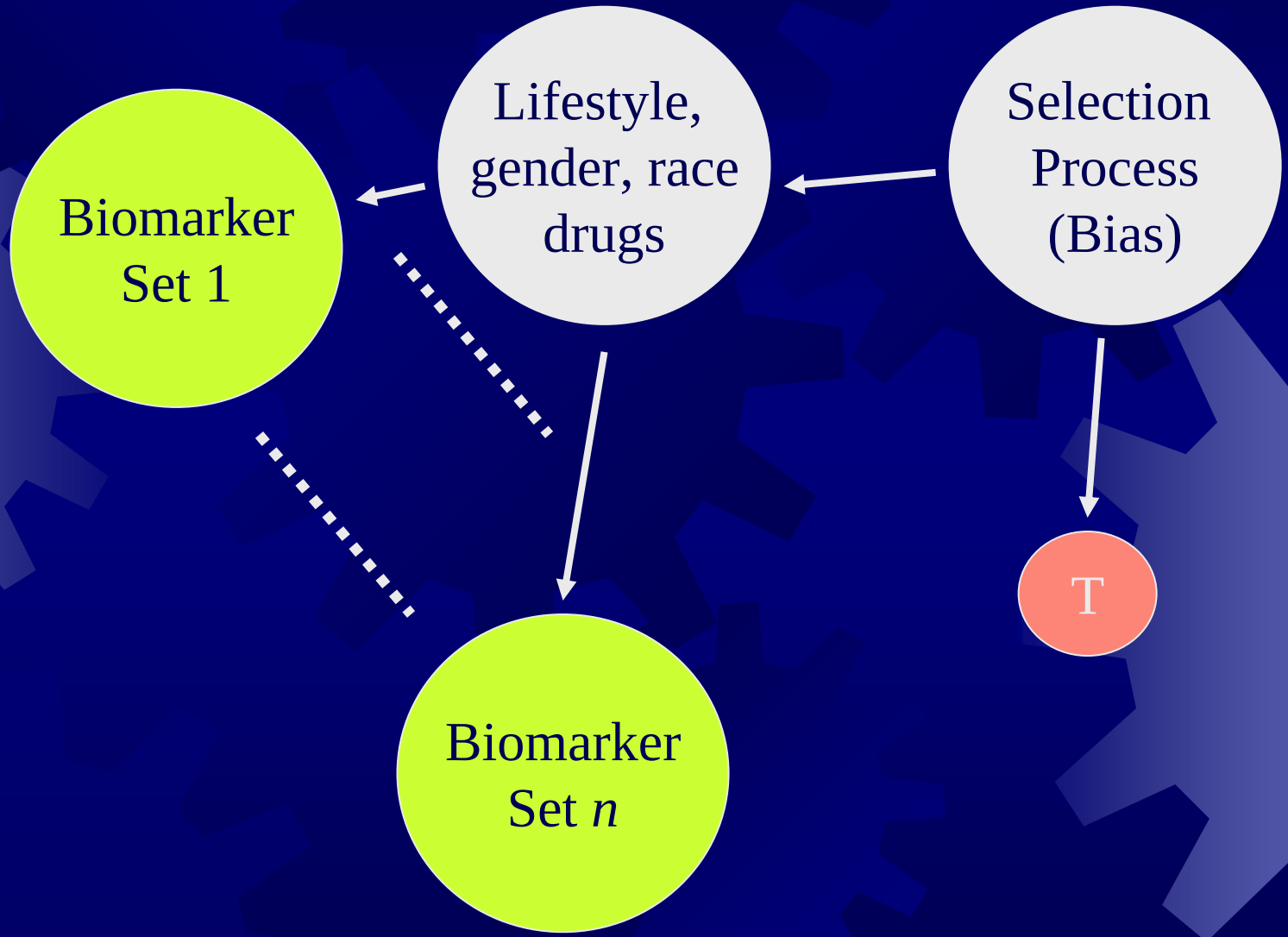
Combine causality/prediction confusion with massive network connectivity effect:



Massive network connectivity effect

- In human networks this is manifested as the “6-degree of separation” phenomenon (X is *somehow* related to Y most often indirectly)
- In bioinformatics it can lead to a certain lack of precision in conclusions like these:
  - Li et al 2000: *causal* interpretations of the strongest and most consistent gene predictors of leukemia histological subtypes: *“…The CD2…plays an important role in mediating the interactions between human T lymphocytes and accessory cells…Blk may play an important role in B cell proliferation…Immunoglobulin-associated beta (B29), …, belongs to family of surface adhesion molecules.”*
  - Eisen et al 1998: *“Genes of similar function cluster together. …strong tendency for these genes to share common roles in cellular processes”*.
  - Zhou et al, 2002: *“It is clear that in a biological pathway a gene is likely to show strong correlations with its neighbour genes, but not with genes that lie far apart in the pathway”.*

Explanation #2: insufficient pre-analytic controls

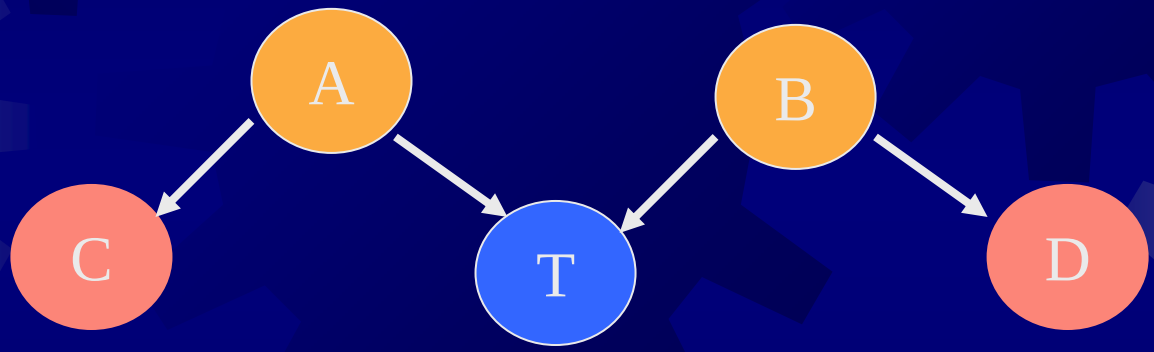


Explanation #3: linearity

- Empirically the most powerful classifiers in microarray gene expression diagnosis are linear; this is because either gene regulatory relationships are linear OR because sample is very small and thus classifiers with high-bias (i.e., very restricted hypothesis spaces) will perform better with such sample (as predicted by the theory of “bias-variance” decomposition of classification error)

Explanation #3: linearity

- If the actual relationships are not linear here is a simplified representation of what may be happening:



$T = \text{sign}\,(a*A^2 - B*Y^3 + c)$

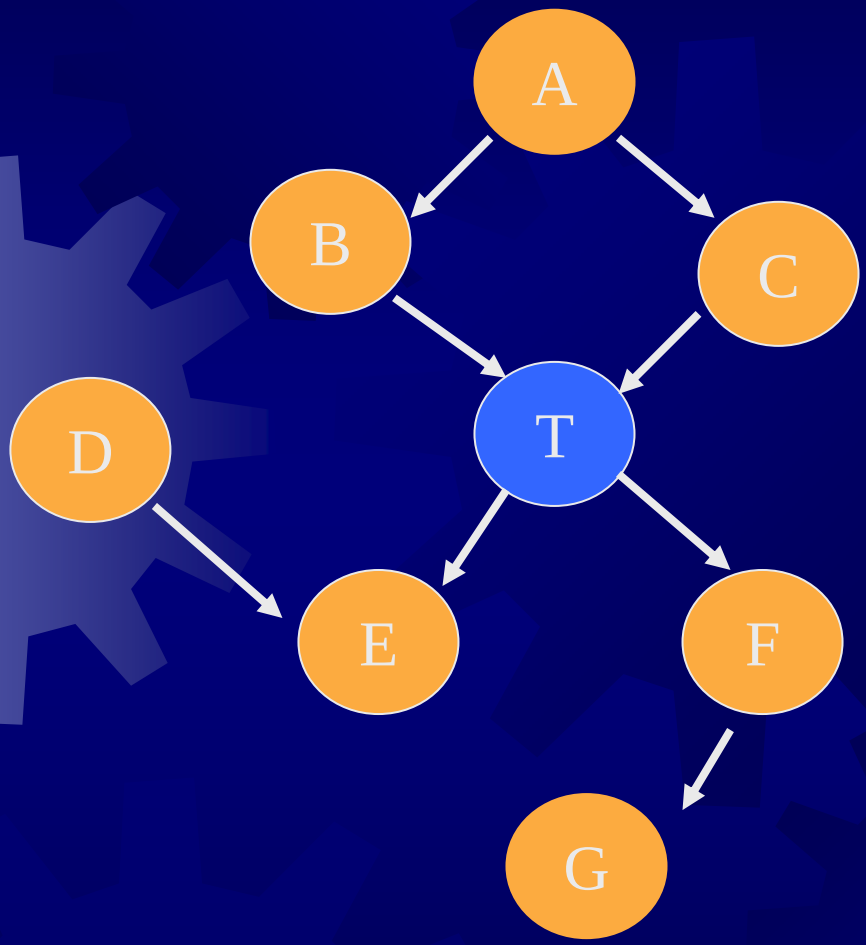
$C = d*A^2$

$D = g*B^3$

$T = \text{sign}\,(a'*C - b'*D + c)$

A new way to think about biomarker discovery

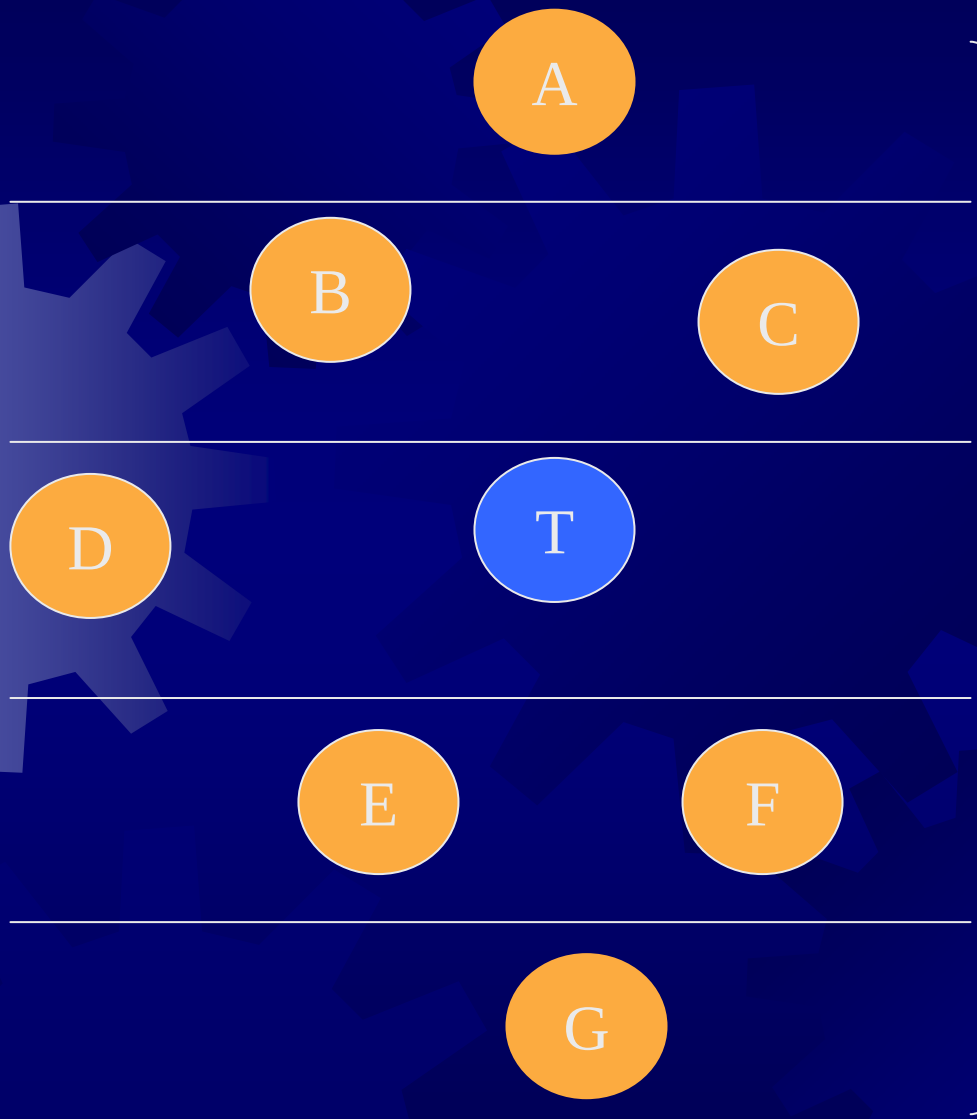
- Bio-medical researchers need to be thinking about what types of rigorously-defined relationships they want to discover and work with data analysis experts to construct appropriate methods
- In other words: we need to make sure that data analysis methods for identifying candidate biomarkers are designed such that the discovered candidates are highly likely to be on the causal pathway of interest
- We can design our data analysis methods to answer a variety of specific questions:



Example 1:

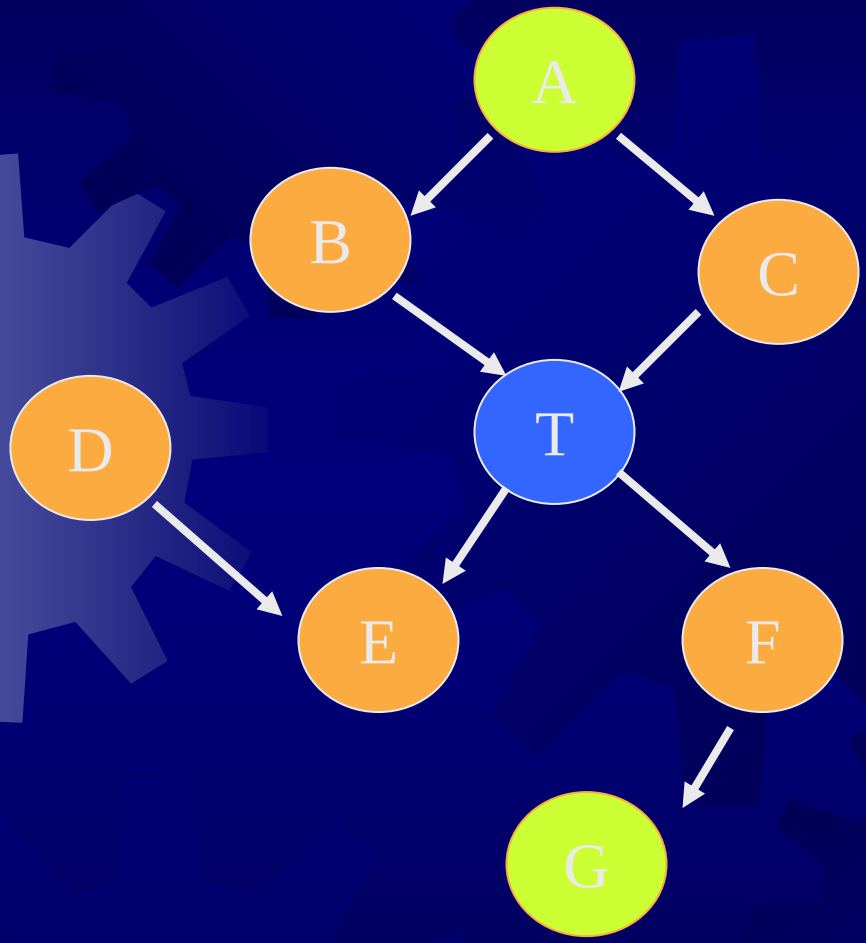
Goal is to find full network

of (causal) regulatory relationships



Example 2:

Goal is to find causal order



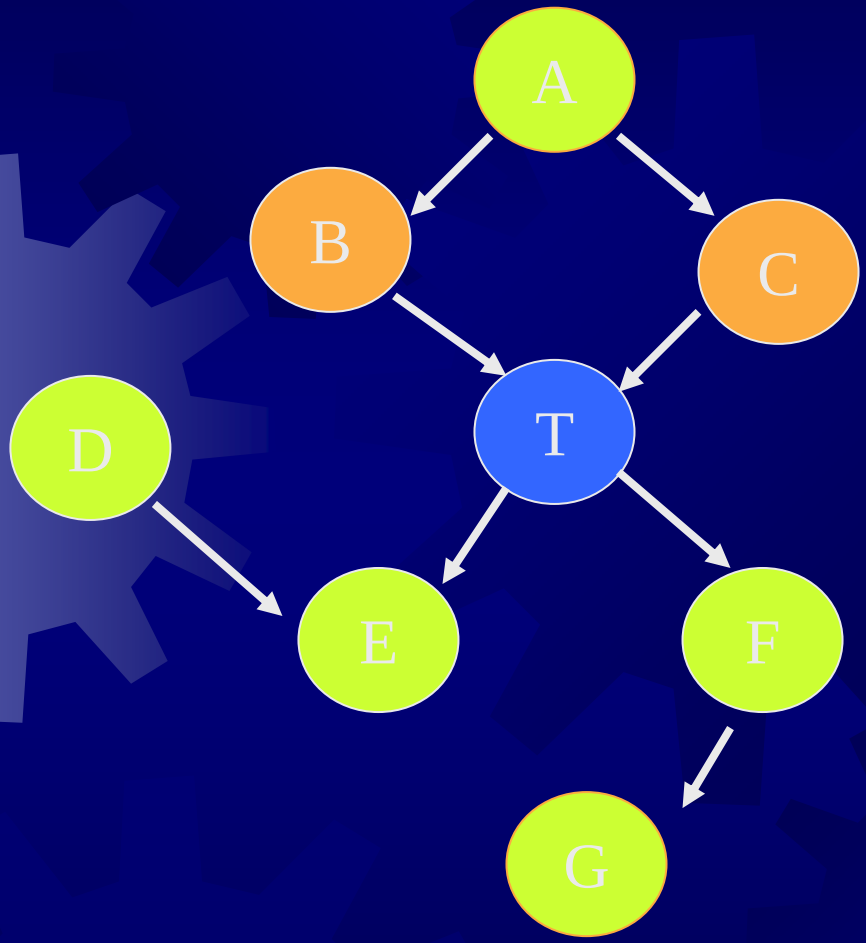
Example 3:

Goal is to find

Markov Blanket

(=> optimal prediction,

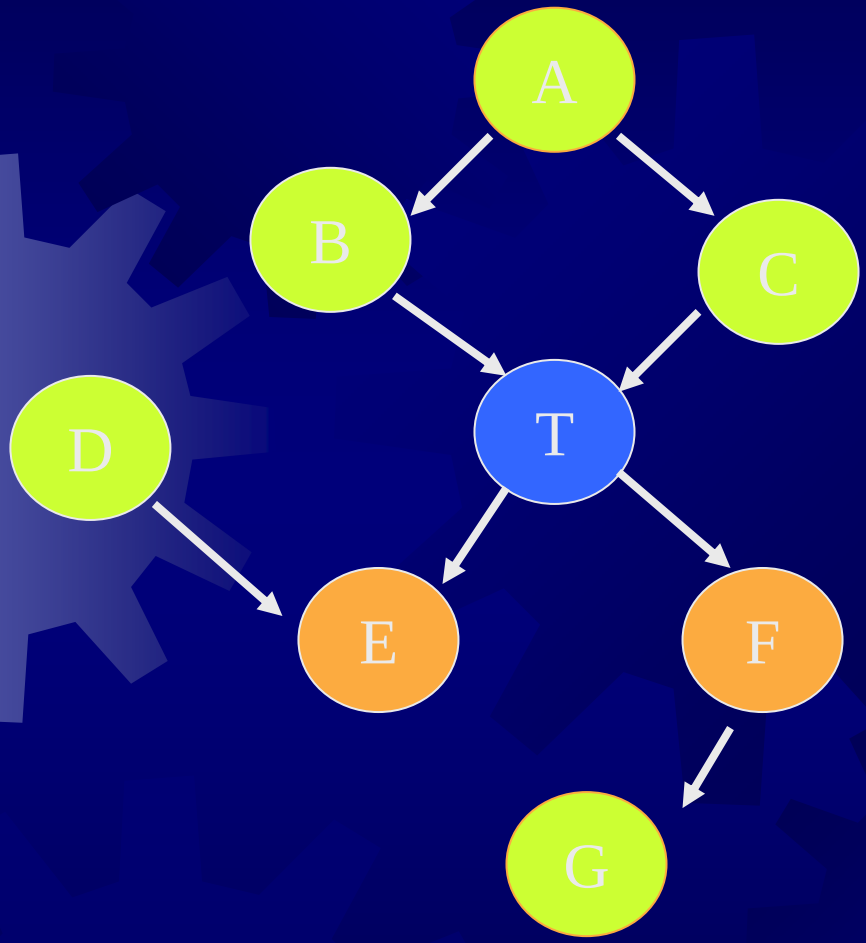
And tight superset of direct causes and direct effects)



Example 4:

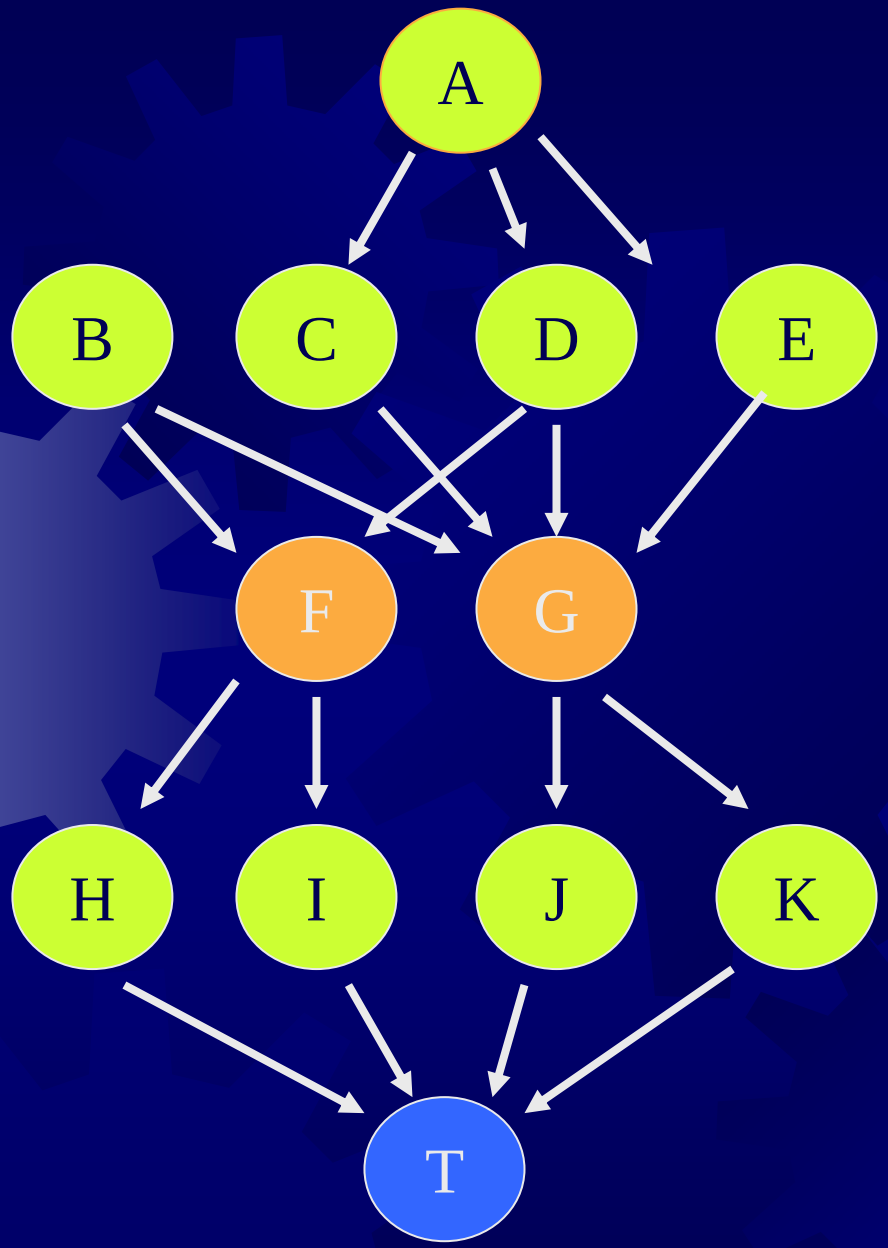
Goal is to find

direct causes



Example 5:

Goal is to find direct effects



Example 6:

Goal is to find "chokepoints" suitable of being drug targets

Technical Challenges

- Although a rich theoretical framework does exist in computer science and other fields (statistics, economics, philosophy) for computational causal discovery (examples of which we just saw) many challenges remain to be solved:
  - Aggregation effects (temporal and cellular)
  - Dealing with small sample sizes
  - Integrating existing knowledge
  - Scaling up suitable algorithms to genomic scale
  - Verifying methods experimentally or otherwise
  - In proteomics, additional complications exist since proteins are mostly unknown and signals distorted
  - In epidemiology, most factors are unmeasured, etc.

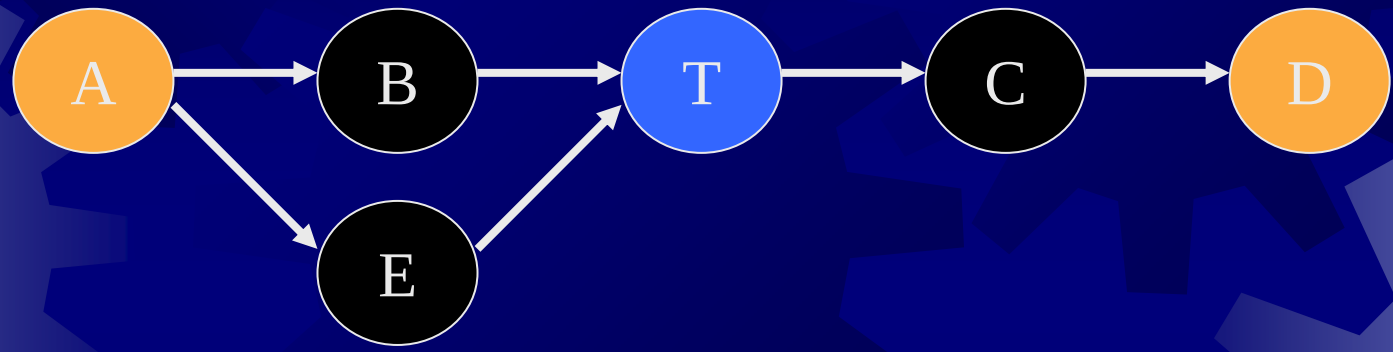
Focused causal biomarker discovery needs more sample than simple classification (or classification-oriented biomarker selection)

- **Why?**
- Statistical Reminder: Conditional independence
- Two variables *X* and *Y* are conditionally independent given ***Z***, denoted as **I(***X; Y*| ***Z*****)**, iff the probability distribution of *X* is the same for all values of *Y* and this holds for each value of *Z:*
- Intuitive Meaning: given that I know *Z ("conditioned on Z"), X* does not give me information about *Y (X is "uninformative", "non-predictive", "independent" of Y)*

$$\frac{p(X+,Y+)}{p(X-,Y+)} = \frac{p(X+,Y-)}{p(X-,Y-)} \quad [Z+]$$

$$\frac{P(X+,Y+)'}{p(X-,Y+)'} = \frac{p(X+,Y-)'}{p(X-,Y-)'} \quad [Z+]$$

- Conditional independence is the basis for every known causal discovery method!



- Simplified example (barring confounding detection details): *A* cannot cause *T* directly because it is independent of *T* given *{B, E}.*

  C must be a direct cause or direct effect of *T* because there is no variable subset that renders *C conditionally* independent of *T*.
- But: as the conditioning set grows, the required sample grows very fast
- => Conditional independence is sample-intensive!!

We have developed several new biomarker candidate selection algorithms

- They are guaranteed to return direct causes-effects of a target gene or disease given broad assumptions
- They can be run with minimal computer equipment and reasonable sample sizes
- Let’s revisit the Lung Cancer biomarker selection tasks with the new algorithms

Lung Cancer: Distinguishing Normal vs Cancer Cells [Model Performance]

17 non-cancerous patients (186 cancerous), 5-fold cross-valiadation (3(5) non-cancerous patients per fold)						
	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	98.62%	99.28%	99.73%	97.03%	99.26%	99.64%
PSVM	99.55%	99.19%	99.46%	97.48%	99.26%	99.64%
KNN	93.05%	93.19%	92.74%	87.83%	97.33%	98.11%
NN	**100.00%**	**100.00%**	99.91%	97.57%	99.80%	N/A
Averages of FS Algorithms	97.80%	97.91%	97.96%	94.97%	98.91%	99.13%

Lung Cancer: Distinguishing Normal vs Cancer Cells [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	1	19	18
MMMB	84		9	103	101
MMMB+PC	76	0		94	92
RFE	6	6	6		2
UAF	99	98	98	96	

Lung Cancer: Distinguishing Normal vs Cancer Cells [Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	19	0
	MMMB	103	1
	MMMB+PC	94	1
	RFE	6	0
	UAF	100	1

- p53 and p63 a strong homolog to p53, - p16 (cyclin-dependent kinase inhibitor 2A that inhibits CDK4))
- k-ras
- akt (v-akt murine thymoma viral oncogene homolog 1; AKT1 , AKT2 ,)
- hTERT (telomere reverse transcriptase)
- c-myc (v-myc avian myelocytomatosis viral oncogene homolog)
- ornithine decarboxylase 1, - kallikrein 11,
- surfactant protein (surfactant protein A binding protein, surfactant, pulmonary-associated protein D, surfactant, pulmonary-associated protein C, surfactant, pulmonary-associated protein B)

Lung Cancer: Distinguishing AdenoCa vs Squamus Cancer [Model Performance]

21 squamous patients (139 adeno), 5-fold cross-validation (4(5) squamous patients per fold)						
	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	**99.09%**	98.49%	98.49%	98.57%	99.32%	98.98%
PSVM	96.91%	98.72%	98.72%	98.57%	98.70%	99.07%
KNN	98.26%	93.07%	93.07%	91.49%	95.57%	97.59%
NN	98.88%	99.32%	**99.38%**	98.70%	99.63%	N/A
Averages of FS Algorithms	98.28%	97.40%	97.41%	96.83%	98.30%	98.55%

Lung Cancer: Distinguishing AdenoCa vs Squamus Cancer [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	0	12	2
MMMB	52		0	64	37
MMMB+PC	52	0		64	37
RFE	11	11	11		5
UAF	489	472	472	493	

Lung Cancer: Distinguishing AdenoCa vs Squamus Cancer [Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	13	0
	MMMB	65	0
	MMMB+PC	65	0
	RFE	12	0
	UAF	500	2

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Model Performance]

7 metastatic patients (132 non-metastatic), 7-fold cross-validation (1 metastatic patient per fold)						
	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	90.87%	97.32%	93.35%	96.43%	95.63%	96.83%
PSVM	88.89%	**97.62%**	94.84%	**97.62%**	96.43%	96.33%
KNN	86.90%	84.92%	84.92%	92.46%	89.29%	92.56%
NN	90.08%	96.03%	96.83%	96.83%	86.90%	N/A
Averages of FS Algorithms	89.18%	93.97%	92.49%	95.84%	92.06%	95.24%

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	0	14	6
MMMB	62		5	75	55
MMMB+PC	57	0		71	53
RFE	6	5	6		2
UAF	492	479	482	496	

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	14	0
	MMMB	76	0
	MMMB+PC	71	0
	RFE	6	0
	UAF	500	3

Lung Cancer Models: Time Efficiency

- MMPC: 0.5 to 5 minutes (depending on the data split of the cross-validation)
- MMMB: 15 to 70 minutes (depending on the data split)
- Platform: Unoptimized, interpreted Matlab code on a Pentium 4, 2 GHz, Windows 2000

Next Steps

- Literature
- Cell-line experimental validation

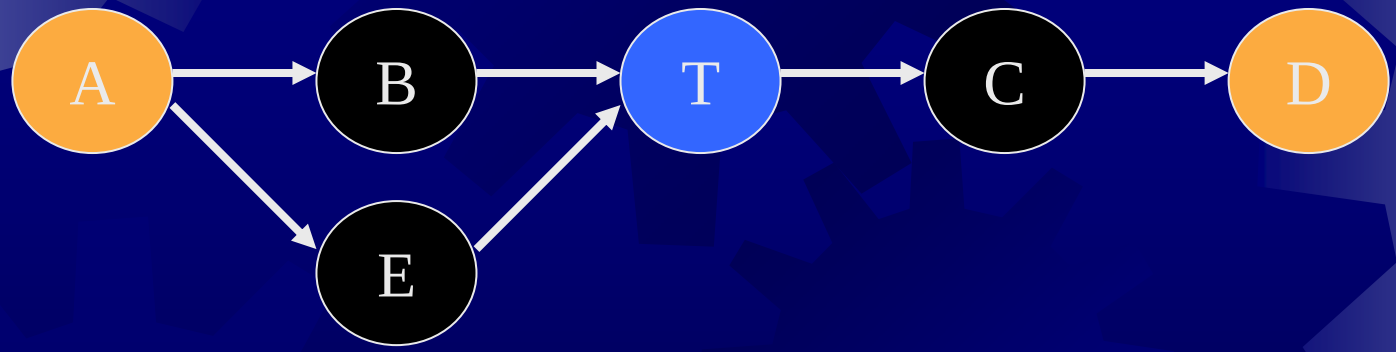
  (Recently NIH-funded Project:

  Principled Methods for Very-Large-Scale Causal Discovery (Aliferis, Tsamardinos, Massion, Carbone, DuPont)
- Further algorithmic improvements

A Method of to Characterize the Genes Selected By Various Methods

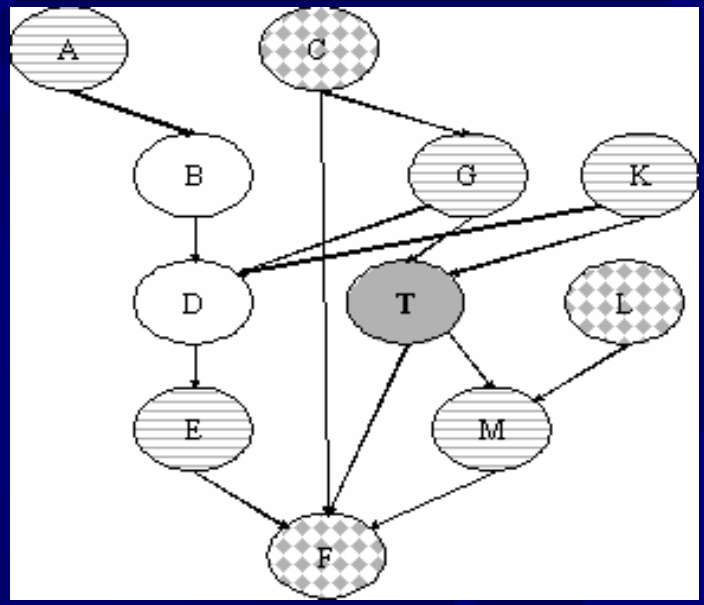
- **Relative Conditional Blocking**

(intuitive example: {A,D} may predict T as well as the larger {B,C,E} set, however {B,C,E} are closer to T than {A,D})



Note: A more technical way to describe this measure is “Relative Divergence from the Causal Markov Condition”; It answers the “causal proximity” question

Characterizing the Genes Selected By Various Methods



Suppose algorithm *Alg1* returns variables V_1 = {*A, G, K, E, M*} (textured with horizontal lines) and algorithm *Alg2*, variables V_2 = {*C, F, L*} (with checkered texture). Conditioned on some subset of V_1 variables *L* and *C* are independent of the target (66% of V_2 is "blocked" by V_1). Conditioned on some subset of V_2 variables *A* and *E* are independent of *T* (40% of V_1 is "blocked" by V_2). We conclude that *Alg1*'s output is closer to the direct causes and effects of *T*, than *Alg2*'s output.

Relative Conditional Blocking In Lung Cancer data

Percenage of features 1 (vertical) eliminated by features 2 (horizontal)	MMPC	RFE
Task 1: Non-Metastatic Adeno vs. Metastatic Adeno		
MMPC	-1	**42.86%**
RFE	**83.33%**	-1
Task 2: Cancerous vs. Non-Cancerous		
MMPC	-1	**10.53%**
RFE	**100.00%**	-1
Task 3: Adeno vs. Squamous		
MMPC	-1	**16.67%**
RFE	**100.00%**	-1

Predicting Single Genes: Performance
(Tumor protein 63 kDa with strong homology to p53)

69 patients with over-expressed gene (134 under-expressed cases), 6-fold cross-valiadation (11(14) over/expr cases/fold)						
	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	**94.21%**	91.96%	92.74%	80.42%	87.92%	90.32%
PSVM	89.67%	89.98%	91.64%	75.35%	87.93%	84.93%
KNN	89.48%	88.68%	92.43%	79.27%	88.29%	84.98%
NN	93.38%	93.78%	93.90%	80.39%	89.27%	N/A
Averages of FS A	91.69%	91.10%	92.68%	78.86%	88.35%	86.74%

Predicting Single Genes: Performance
(V-AKT MURINE THYMOMA VIRAL ONCOGENE HOMOLOG 1; AKT1)

69 patients with over-expressed gene (134 under-expressed cases), 6-fold cross-valiadation (11(14) over-expressed cases per fold

	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	68.15%	70.30%	79.71%	73.83%	73.22%	70.00%
PSVM	68.43%	72.81%	77.71%	71.53%	66.29%	68.39%
KNN	68.99%	72.74%	73.78%	72.58%	71.11%	76.29%
NN	71.86%	78.29%	80.70%	71.78%	71.52%	N/A
Averages of FS Algorithms	69.36%	73.53%	77.98%	72.43%	70.54%	71.56%

Predicting Single Genes: Stability, Novelty, Relative Blocking

- Results same when choosing different thresholds and when doing regression instead of classification
- Results similar in terms of non-intersection of gene lists among methods and with gene list
- Relative blocking same as with prediction of disease
- Size of direct causes/effects: 7 and 9 (compared to14, 19, and 13 for three diagnostic tasks)

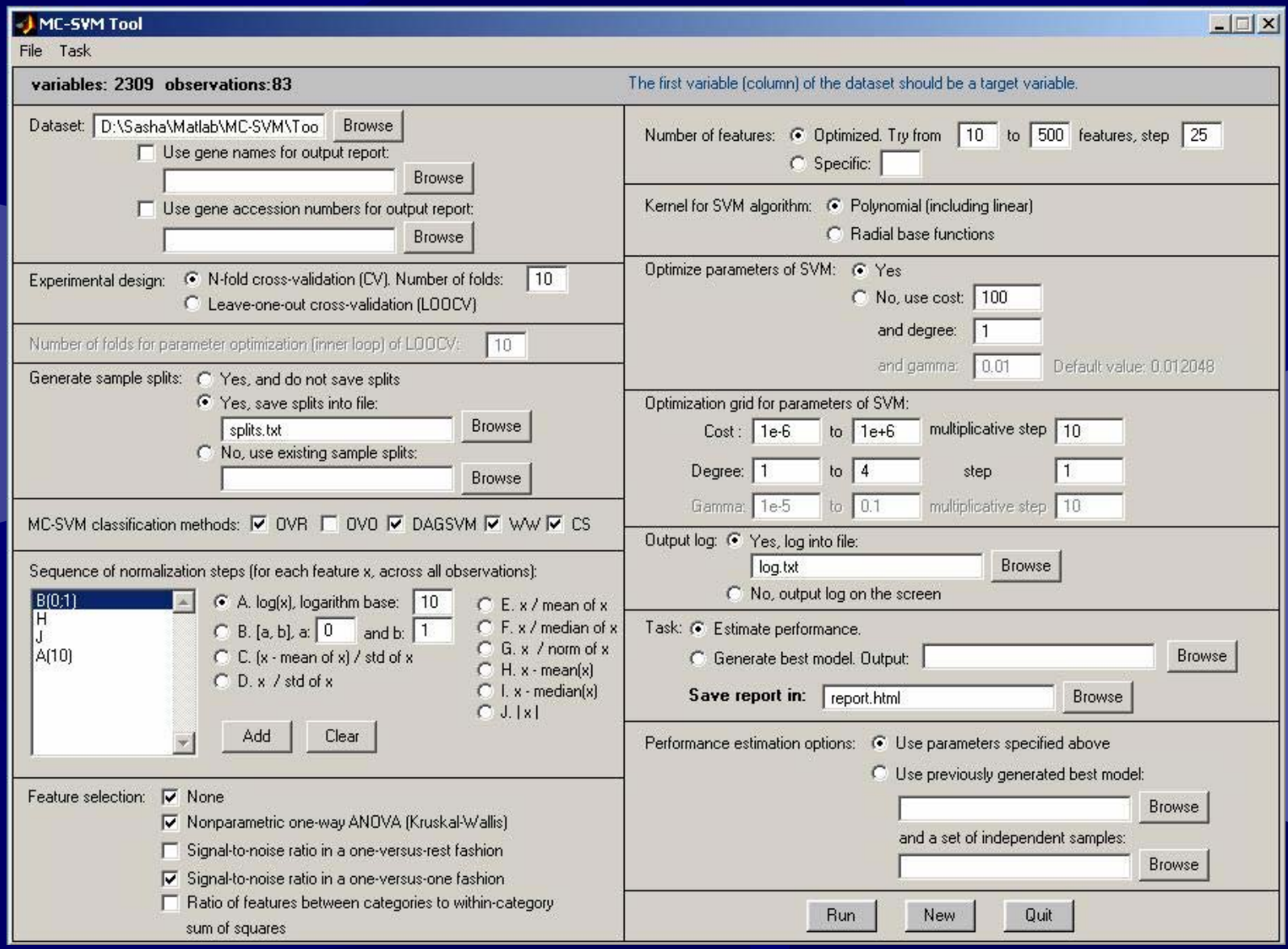
Conclusions

1. By using causally-oriented methods we can derive optimal or near-optimal prediction and diagnostic models in gene expression diagnostic tasks
2. The novel methods achieve excellent to outstanding reduction in the number of predictors
3. The novel methods are efficient in time and sample even in the presence of massive numbers of variables
4. They offer additional information than state-of-the art biomarker selection methods and by design have a causally-oriented interpretation
5. This work is fairly well-developed in its algorithmic aspects but is just beginning biologically. A lot remains to be learnt about the relative merits of the available approaches and methods.

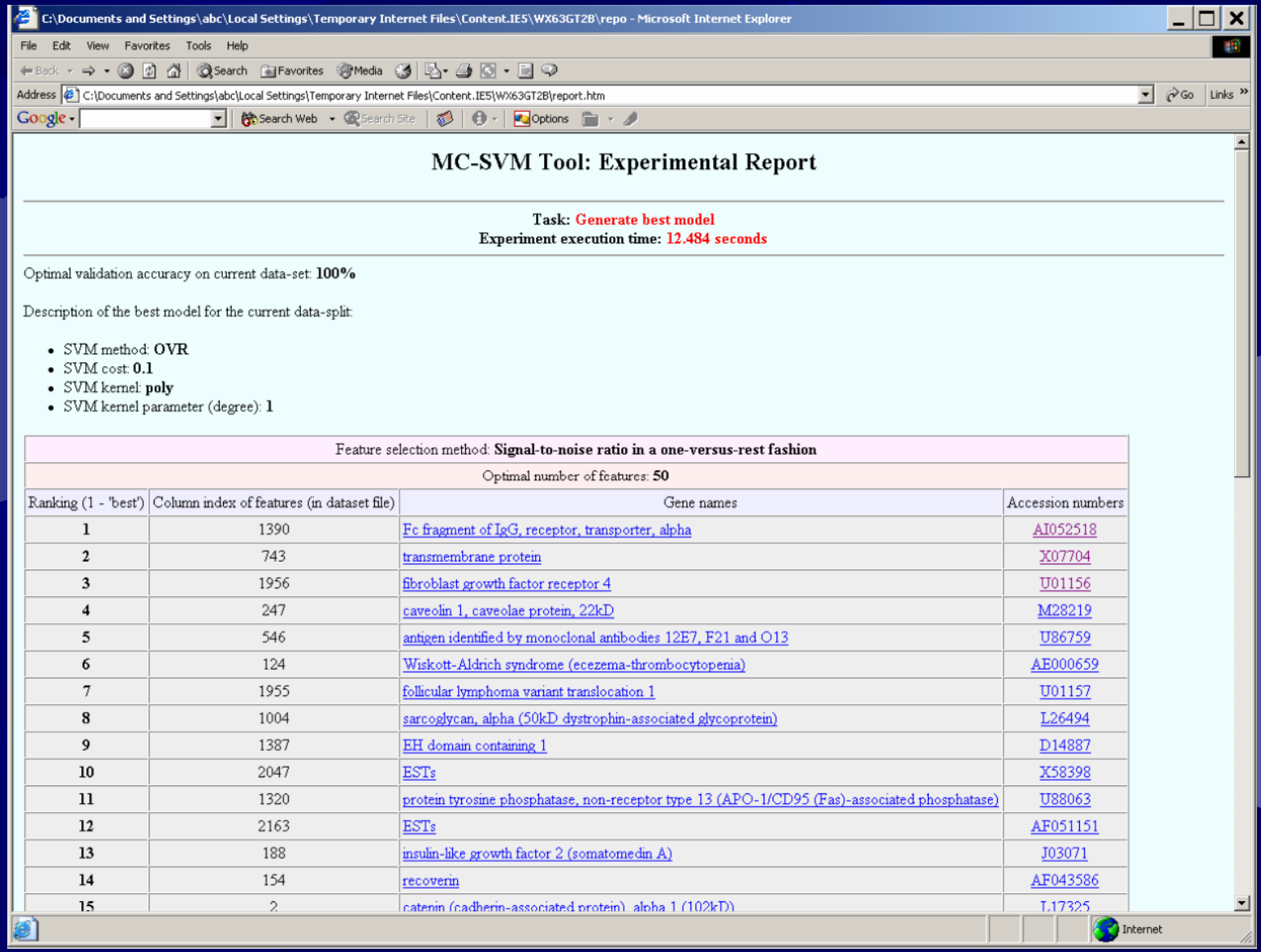
Bonus section: computer tools for biomarker discovery & diagnosis

- We have recently completed an extensive analysis of all multi-category gene expression-based cancer datasets in the public domain. The analysis spans >75 cancer types and >1,000 patients in 12 datasets.
- On the basis of this study we have created a tool that automatically analyzes data to create diagnostic systems and identify biomarker candidates using a variety of techniques.
- The present incarnation of the tool is oriented toward the computer-savvy researcher; a more biologist-friendly web-accessible version is under development.

MC-SVM: User Interface



MC-SVM: Sample Output



C:\Documents and Settings\abc\Local Settings\Temporary Internet Files\Content.IE5\WX63GT2B\repo - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\Documents and Settings\abc\Local Settings\Temporary Internet Files\Content.IE5\WX63GT2B\report.htm

MC-SVM Tool: Experimental Report

Task: Generate best model
Experiment execution time: 12.484 seconds

Optimal validation accuracy on current data-set: **100%**

Description of the best model for the current data-split:

- SVM method: **OVR**
- SVM cost: **0.1**
- SVM kernel: **poly**
- SVM kernel parameter (degree): **1**

Feature selection method: **Signal-to-noise ratio in a one-versus-rest fashion**

Optimal number of features: **50**

Ranking (1 - 'best')	Column index of features (in dataset file)	Gene names	Accession numbers
1	1390	Fc fragment of IgG, receptor, transporter, alpha	AI052518
2	743	transmembrane protein	X07704
3	1956	fibroblast growth factor receptor 4	U01156
4	247	caveolin 1, caveolae protein, 22kD	M28219
5	546	antigen identified by monoclonal antibodies 12E7, F21 and O13	U86759
6	124	Wiskott-Aldrich syndrome (eczema-thrombocytopenia)	AE000659
7	1955	follicular lymphoma variant translocation 1	U01157
8	1004	sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)	L26494
9	1387	EH domain containing 1	D14887
10	2047	ESTs	X58398
11	1320	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)	U88063
12	2163	ESTs	AF051151
13	188	insulin-like growth factor 2 (somatomedin A)	J03071
14	154	recoverin	AF043586
15	2	catenin (cadherin-associated protein), alpha 1 (102kD)	L17325

Discovery Systems Laboratory

For more Information (causal discovery tools, publications, contact information)

http://discover1.mc.vanderbilt.edu/discover/public/