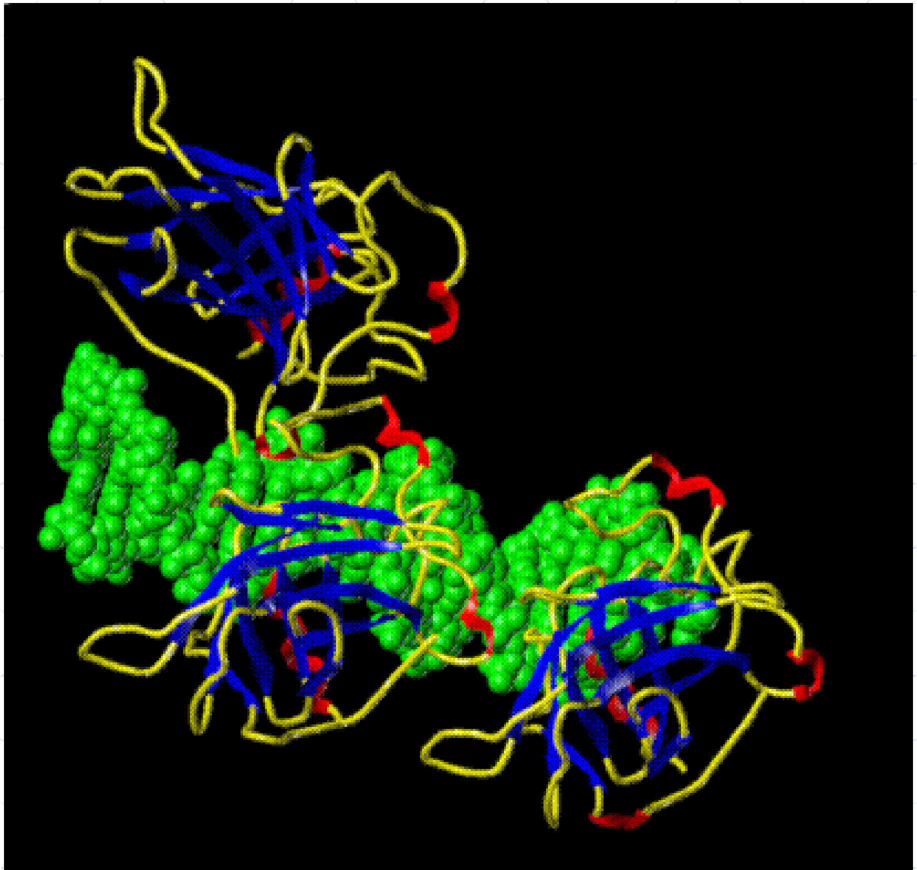
A critical view of spectral analysis

Harry B. Burke, MD, PhD
Associate Professor of Medicine
Director of Bioinformatics and Biostatistics
McCormack Genomics Center
George Washington University School of Medicine

Proteins



P53 CORE DOMAIN IN COMPLEX WITH DNA (1TSR)

- Very small quantities of a protein are amplified by enzymes
- Proteins do not act alone, they interact with other proteins
- Proteins complex
- Degradation products circulate briefly in the serum

Therefore, proteins are nonlinear and interactional

It’s a tough field

- The is the most complex and difficult area in medical research
- It is easy to go wrong and very hard to do it correctly
- The technology is so sensitive that what would not have been problems with other data is a problem here
- Although it is similar to gene microarrays in its use of a large number of continuous variables, its analysis differs in several significant ways from gene microarrays.

Burke H. Discovering patterns in microarray data. Molecular Diagnosis 2000;5:349-357.

What we can learn from Petricoin et al.

- *There is no such thing as 100% sensitivity and 100% specificity in biology or medicine (even diagnostic biopsies are wrong sometimes)*
- There can be differences in patient characteristics – that can be result in unanticipated problems with controls and/or cases
- There can be differences in acquisition, storage, and processing of specimens – that can result in extreme bias
- There can be one dominant variable – and you don’t find it and report a pattern
- You should not generate many validated models and report the best one – you should report the distribution of model ROC scores
- Do not separate the ambiguous cases from the rest of the dataset

What we can learn from Petricoin et al.

- If you include many variables in a classifier where the classifier is based on a dataset with a low event rate – this not only results in model instability but usually generates a high accuracy that is artifactual
- Report the software settings in the Methods section of the publication
- Spectral analysis requires (1) attention to detail and (2) a high level of statistical expertise

Why are we analyzing serum protein peaks in early cancer or precancer?

- Detect a protein released from the few dying cancer cells that is spilled into the serum or that is secreted by the tumor.

  (A primary peak)
- Detect a protein released by the cancer tissue that is due to a secondary effect of the cancer

  (A secondary peak)
- Detect the effect of the tumor on other tissues, i.e., the amplification effect (A tertiary peak)

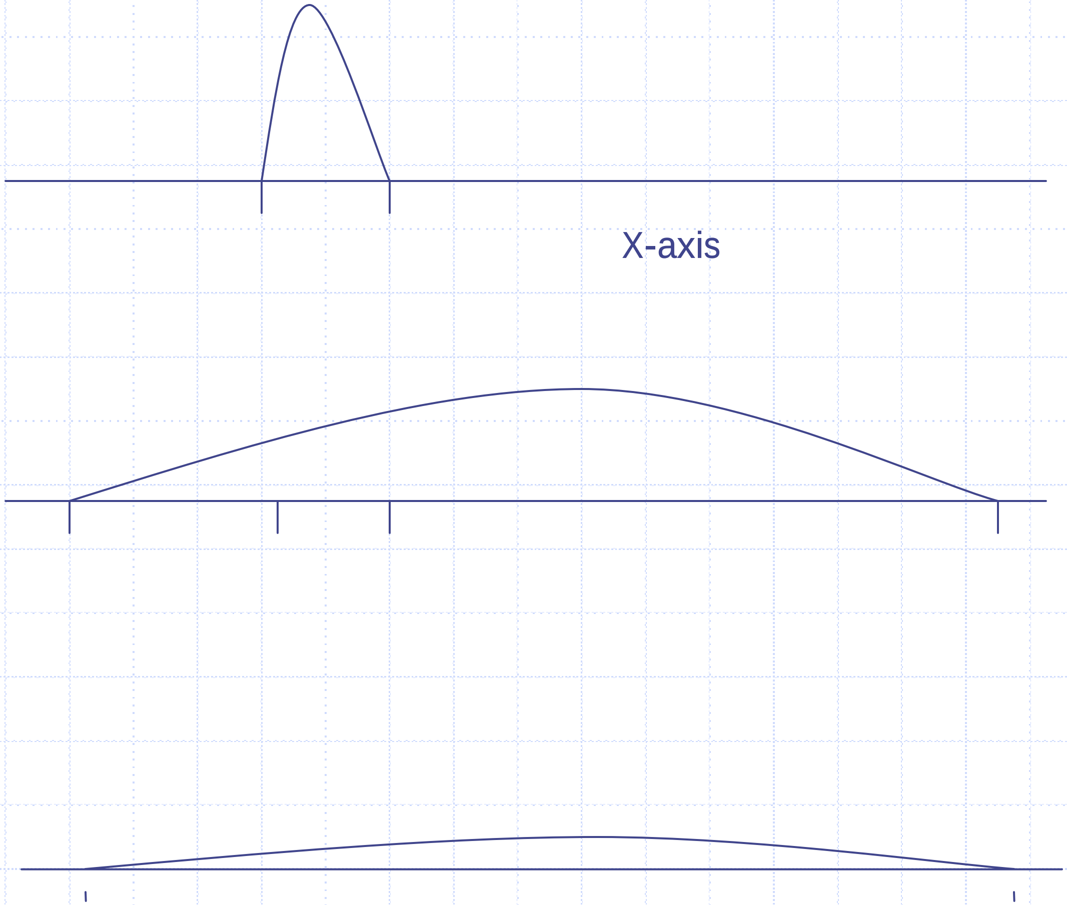
Seeing is believing

- “I know pornography when I see it.”
  - Supreme Court
- “I know a peak when I see it.”
- “Just because you see it , does not mean that it is there.”
  - Harry Burke

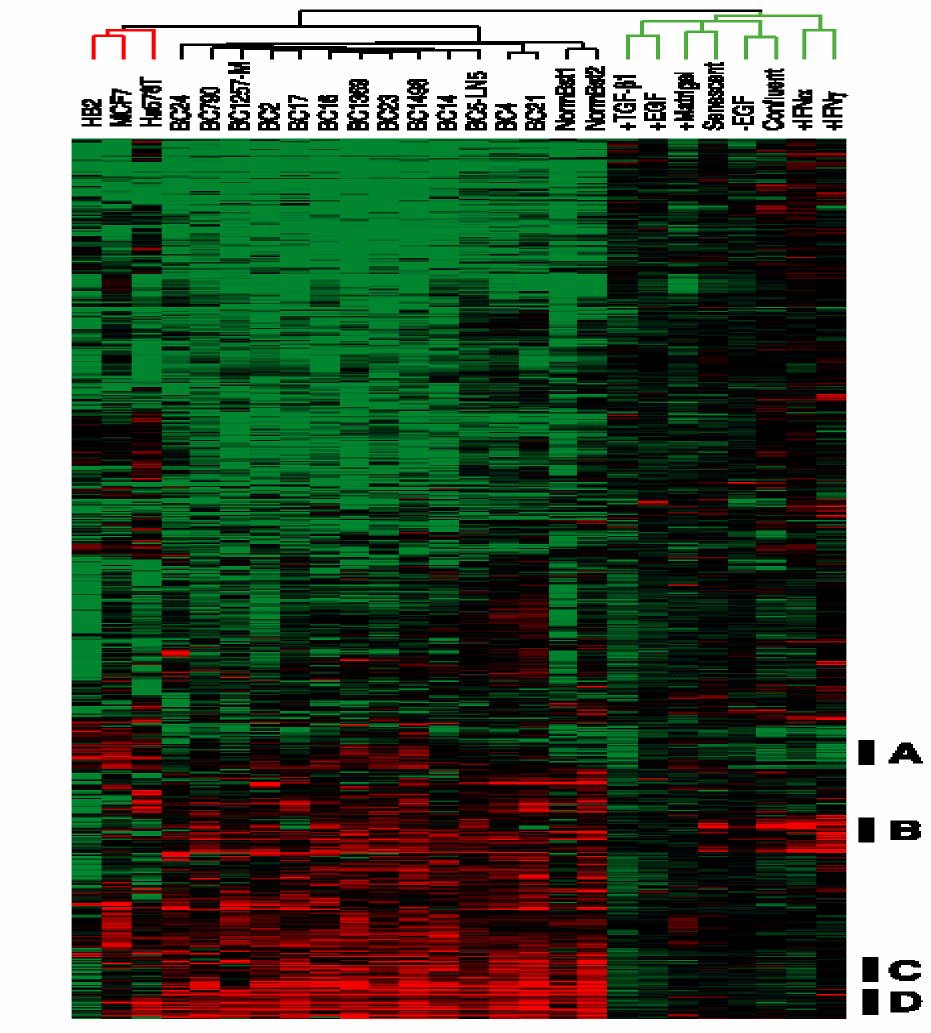
When is a peak really a hill?

- What is a peak? Necessary and sufficient criteria
- A peak depends on the x-axis resolution, the “shape” will change as the resolution increases – a peak becomes a hill
- How do you align peaks without losing truth value? (without changing the data itself)
- “Fill in” missing peaks (multiple imputation)
- High peaks have high variance because it is free to vary more than small peaks
- Adjust the height by its variance, peaks become hills
- The peak is a distribution and should be modeled as a distribution rather than an exact value

Peaks depend on scale



Replication:
Hierarchical clustering



Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 9212–9217, August 1999

Distinctive gene expression patterns in human mammary epithelial cells and breast cancers

CHARLES M. PEROU*, STEFANIE S. JEFFREY†, MATT VAN DE RIJN‡, CHRISTIAN A. REES*, MICHAEL B. EISEN*, DOUGLAS T. ROSS§, ALEXANDER PERGAMENSCHIKOV*, CHERYL F. WILLIAMS*, SHIRLEY X. ZHU‡, JEFFREY C. F. LEE¶, DEVAL LASHKARIi, DARI SHALON¶, PATRICK O. BROWN§**††, AND DAVID BOTSTEIN*††

Departments of *Genetics, †Surgery, ‡Pathology, and §Biochemistry and **Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; ¶Incyte Pharmaceuticals Inc., Fremont, CA 94555; and iGenometrix, The Woodlands, TX 77381

Variable and model validation and replication

- Cross-validation
  - Randomly split dataset
  - Leave-one-out
  - Bootstrap
- Two phases:
  - Phase I: Assessment
  - Split original dataset into: (1) training, (2) hold-out, (3) test data sets
  - Phase II: Validation
  - (4) Replication: An independent dataset analyzed by another investigator (what the EDRN was designed to do)

Variable and model accuracy (discrimination)

- Sensitivity and specificity require a threshold
- ROC is every sensitivity/specificity pair
- For one or more variables, report model ROC rather than the sensitivity and specificity for a particular threshold so others can compare their models to yours
- It is too early to use utility
- No one has ever agreed on a cost function
- Serum is low cost
- Assess variables by forcing all the variables into the model, asses the model's accuracy, remove one variable, assess model accuracy, replace the variable and remove another variable, assess model accuracy, etc.

Serial validation – an incorrect approach (it isn’t robustness)

- Select a set of variables from the first dataset
- Match a subset of the variables in the second dataset
- A certain number of variables will match due to chance – especially if there are many possible variables and the criteria for variable matching is not stringent

Prediction and biological knowledge

- Variables are predictive to the extent that they are related to the disease process
- We do not need to know about the disease process to use a variable as a predictive factor
- To the extent that the variable is predictive, it is a good target for further investigation
- We combine variables because no one variable is sufficiently powerful to accurately predict the outcomes of all the patients

Assumptions and consequences

1. There are at least 300,000 proteins
2. We can accurately and reliably detect the peak of each protein
3. We can quantify the relative amounts of each protein

If #2 and #3 are not currently true, then what we currently detect is ambiguous and unreliable

If our assumptions are true, how do we analyze these data?

- Top down processing
  - Analyze all 300,000 proteins as a pattern requires ~9,000,000 cases with a 50% event rate for model stability
  - Data reduction then analyze remaining proteins

An Information Theoretic Perspective

- Protein spectra are a technology that create an analog representation the relative quantity of a protein at a biologic moment in time
- The spectra is the ratio of the signal, the true amount of each protein, to the noise, the spurious and background activity level

Massively Parallel Information

- Spectra are massively parallel information
- The information is not the result of a conditional sequence of investigative events
- There are few examples of massively parallel information in science

Patterns of proteins

- We are interested in discovering patterns of proteins
- “Pattern” can be operationally defined as a set of elements that occur in a systematic and meaningful-for-the-task manner

Pattern detection

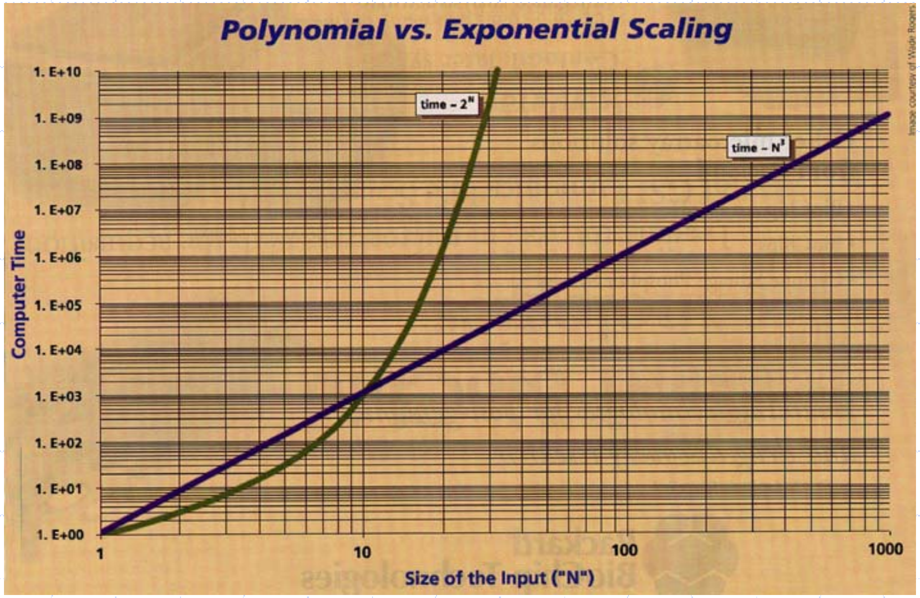
- In this context, there are two types of pattern tasks
- Pattern recognition is the recognition of a pattern when it occurs again, i.e., being able to identify a pattern as an instance of a known pattern
- Pattern detection is the detection (the discovery, or more correctly, the learning) of a pattern in the data

NP-hard problem

- Initially it should be assumed that every data element in a massively parallel information source has the potential to be a meaningful, i.e., to be a necessary but not sufficient part of the pattern
- The reason for this assumption is because if it were not true then massively parallel information would probably not necessary
- It is precisely because anything could be important that we are interested in, and willing to deal with the problems of, massively parallel information

NP-hard problem

- Arrays present an analytic problem that is NP-hard
- NP stands for “non-deterministic polynomial time”
- NP-hard problems are problems that are not known to be verifiable in polynomial time and may require exponential time in the worst case



The problem

- Every protein is a continuous variable
- There are 300,000 proteins
- There is intra-patient variation
- There is inter-patient variation
- There is disease variation (stage, subtype)
- There is error in the technology
- Few cases

Simplifying the problem

- The analysis can be simplified by:
- Thresholding each protein’s signal and considering it to be a binary variable
- Minimizing disease and inter-patient variation.
- In this the simplest of conditions there are 2^n possible patterns (where n is the number of proteins). 2^n is a very large number

High dimensional space

- This is very high dimensional space
- This space has its own characteristics, for example, the curse of multidimensionality
- In high dimensions the space becomes extremely large and the data points move to the edges of the dimensional space

Data reduction

- If you use two algorithms in a serial manner, the second algorithm can not be more accurate than the first algorithm
  - The accuracy of the data reduction algorithm determines the overall accuracy
- Can you *a priori* ignore or delete correlations between variables? No

Data reduction: unsupervised learning

- In unsupervised learning the final error metrics are not available during training, thus the algorithm is not guided by an outcome, this has been termed “blind separation” because there is no dependent variable
- It is based on how the variables fluctuate

Unsupervised learning algorithms

- The task is to reduce the data complexity with minimal loss in precision by discarding noise and revealing basic structures
- The algorithms accomplish this by optimizing a cost function which preserves the original data as completely as possible while simultaneously favoring prototypes with minimal complexity
- Unsupervised learning algorithms tend to focus on linear decorrelations or the maximization of signal-to-noise ratios usually assuming Gaussian sources
- They relate changes in the independent variables to each other – there is no necessary relationship between changes in these variables and changes in the dependent variable

Principal components analysis

- The problem with PCA is that the reason we are using the spectra is because every protein is potentially informative. Therefore, what we don’t want to do is combine most of the proteins into a few categories

Self-organizing maps

- Self-organizing maps (SOM) were introduced by Kohonen in 1984 as a tool for visualization of high dimensional data spaces
- SOM can be said to do clustering/vector quantization (VC) and at the same time to preserve the spatial ordering of the input data reflected by an ordering of the code book vectors (cluster centroids) in a one or two dimensional output space, where the latter property is closely related to multidimensional scaling (MDS) in statistics

Self-organizing maps

- The issue is how good is SOM compared to either VQ or MDS techniques?
- In a series of multivariate normal clustering problems SOM was shown to perform significantly worse in terms of quantization error, in recovering the structure of clusters and preserving the topology compared to traditional MDS methods. (*Flexer A. In: Advances in Neural Information Processing Systems 9*.)

Clustering algorithms

- The object of cluster analysis is to determine a classification or taxonomic scheme that accounts for the variance among subjects.
- Clustering and related unsupervised learning techniques such as competitive learning and self-organizing maps traditionally rely on measures of similarity distance measures that operate on feature vectors, like Euclidean distance, which are generic across problem domains.

Clustering algorithms: Euclidean distance

$$\textit{For two observations } \mathrm{x}' = [x_1, \cdots, x_n] \textit{ and } y' = [y_1, \cdots, y_n],$$

$$d_{xy} = \sqrt{\sum_{i=1}^{n} (\mathrm{x_i} - \mathrm{y_i})^2}$$

Clustering algorithms

- The problem with clustering is that transformation invariance does not hold with generic distance metrics (nondeterministic)
- In other words, different generic distance measures produce different cluster results

Clustering algorithms

- What is necessary are domain-specific distance measures
- But this idea does not solve the problem, rather it shifts the problem to how to select the optimal domain-specific distance measure
- Unsupervised learning can be used to observe fluctuations in the data for QA purposes

Supervised learning

- The final error metrics are available during training
- For classifiers, the algorithm can directly reduce the number of misclassifications on the training data set
- It is usually the optimal strategy to turn an unsupervised learning problem into a supervised learning problem

Supervised learning

- The problem here is that regression model learning is best achieved when there are a few variables and many instances
- Massively parallel information data sets contain many variables and a few instances of all the variables
- Model instability – at least 15 events per independent variable
- Should be able to capture nonlinearity, interactions, and correlated variables

CART

- Splits on variations in variables – sensitive to small data sets
- Requires large data sets if there are many variables
- Uses an ad hoc splitting test (Chi-squared)
- Difficult to model nonlinearities and interactions
- We have always found it to be less accurate than other forms of regression (publication)

The use of multiple algorithms

- All algorithms are not equally efficient
- If we train several algorithms, how do you pick the “true” model, is it the most accurate?
- Mixture of experts

Solutions

- Refine the technology
- Minimize the intra and inter-patient variance
- Minimize the disease variance
- Careful data acquisition and sample preparation
- Acquire more cases
- Improve the statistical algorithms
- Restrict the domain, for example,
  - A small part of the spectra
  - A particular class of protein (e.g., in CAD we look at inflammation, coagulation, etc.)
- Reformulate the problem in a way that is not NP-hard

Reformulate the problem

- Do hypothesis testing (bottom-up processing)
- Perform data addition rather than data reduction
- Begin with a known pattern and add proteins in a motivated manner
- Add cases as the number of proteins increases

Bottom-up processing

- Start with what we know about proteins and build up models, where each addition of a variable is tested
- If a variable is added that is a variable for further research
- For example examine we could look at known serum carrier proteins and their associated proteins

How are predictive factors related to each other?

- Current supervised models provide information regarding the relationship of the independent factors and the outcome, but provide no direct information about the relationship between the independent factors in the context of the outcome.
- For proteins related to the disease process, what proteins are related to each other (same mechanism) and which proteins are not related to each other (different mechanisms)?

Bayesian networks

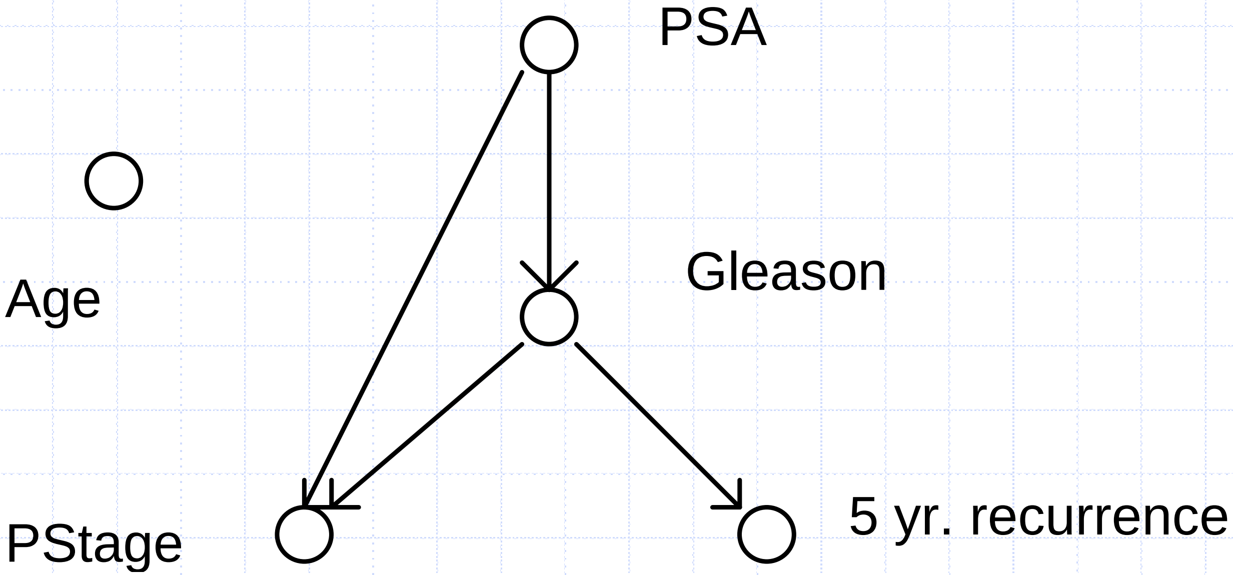
- Multivariate regression models have shown the Gleason score to be the strongest predictor of recurrence and disease specific survival in prostate cancer (results not shown)
- But these models do not show how the variables age, PSA and stage are related to each other and to the Gleason score, in the context of five-year recurrence
- We presented results that suggested that Bayesian networks can provide additional information not available in multivariate regression

Bayesian network

- A simplified Bayesian network representing the joint probability distribution over the variables: age, preop PSA, path stage, path Gleason score, and recurrence with the weakest arcs removed from the network

Hoang A, Burke HB. Bayesian network modeling of prostate cancer. Presented at National Cancer Institute Urologic Oncology Conference, Bethesda, MD, December 1-2, 2001.

The Bayesian Network (N = 1,961)



Hoang A, Burke HB. Bayesian network modeling of prostate cancer. Presented at National Cancer Institute Urologic Oncology Conference, Bethesda, MD, December 1-2, 2001.

What could be the next approaches to spectral analysis?

- Continue to improve sample acquisition, handing, and processing
- Continue to refine the spectral technology
- Continue to refine top-down processing including protein varification
- Begin to perform bottom-up processing
- Motivated analysis of groups of nonlinear, interactional, highly correlated proteins
- Bayesian networks, artificial neural networks