REVIEW

Bayesian methods for proteomics

Gil Alterovitz[1, 2, 3, 4], Jonathan Liu[5], Ehsan Afkhami[4] and Marco F. Ramoni[1, 3, 4]

[1] Division of Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Boston, MA, USA
[2] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Children's Hospital Informatics Program, Boston, MA, USA
[4] Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, MA, USA
[5] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

Received: February 11, 2007
Revised: April 29, 2007
Accepted: April 30, 2007

Biological and medical data have been growing exponentially over the past several years [1, 2]. In particular, proteomics has seen automation dramatically change the rate at which data are generated [3]. Analysis that systemically incorporates prior information is becoming essential to making inferences about the myriad, complex data [4–6]. A Bayesian approach can help capture such information and incorporate it seamlessly through a rigorous, probabilistic framework. This paper starts with a review of the background mathematics behind the Bayesian methodology: from parameter estimation to Bayesian networks. The article then goes on to discuss how emerging Bayesian approaches have already been successfully applied to research across proteomics, a field for which Bayesian methods are particularly well suited [7–9]. After reviewing the literature on the subject of Bayesian methods in biological contexts, the article discusses some of the recent applications in proteomics and emerging directions in the field.

Keywords:
Bioinformatics / Data fusion / Proteomics methods / Statistical models

1 Introduction

Classical statistics provides methods to analyze data, from simple descriptive measures to complex and sophisticated models. The available data, assumed to be a representative sample, are processed and conclusions are drawn about this hypothetical population.

It is not hard to imagine situations, however, in which sampled data are not the only available source of information about the population. Suppose, for example, one needs to guess the outcome of an experiment that consists of tossing a coin. Given that one has likely not seen many biased coins before, one might be more is ready to believe that the coin is fair and that the outcome of the experiment can be either head or tail with the same probability. On the other hand, the situation might change if one learns that the coin has been altered so heads are more likely. This additional information can be taken into account to improve predictive power. In fact, this issue becomes critical when one is considering data for which knowledge corpora have previously been developed- as occurs in many cases in proteomics: from homology-based knowledge to known experimental biases [10].

Bayesian methods provide a principled way to incorporate this external information into the data analysis process. In a Bayesian approach, the data analysis process starts already with a previously established probability distribution, called the prior distribution. In the previous example, one would represent the fairness of the coin as a uniform prior probability distribution, assigning probability of one half to both sides of the coin. On the other hand, if one learns, from

Correspondence: Dr. Gil Alterovitz, Bioinformatics Core, Harvard Medical School, New Research Building, Room 250, 77 Ave Louis Pasteur, Boston, MA 02115, USA
E-mail: gil@mit.edu
Fax: +1-617-525-4488

Abbreviations: MNA, multilevel neighborhoods of atom; **PIE,** probabilistic interactome experimental; **PIP,** probabilistic interactome predicted

some external source of information, that the coin is biased toward heads, then one can model this *via* a prior probability distribution which assigns a higher probability to the event that the coin lands head. The Bayesian data analysis process consists of using the sample data to update this prior distribution into a posterior distribution. The basic tool for this updating process is the Bayes' theorem, named after Thomas Bayes, an 18th century clergyman.

2 Methods

2.1 Foundations

The notions of conditional probability and independence are fundamental to the Bayesian approach. Informally, independence refers to events that have no influence on each other. For example the result of a coin toss and catching the flu are totally independent events. Two events are said to be dependent if they are not independent. Conditioning refers to gauging the probability of an event occurring by observing the realization of another event. Thus, if two events are independent, then conditioning on one of them gives no more information about the other. However, if two events are dependent, then conditioning may provide substantial additional information. For example, since a coin toss and catching the flu may be considered independent events, by observing whether a dime lands heads, one cannot discern more about the possibility of flu. However, if one observes that others have had the flu recently, the likeliness of catching the flu has increased- compared to not having made this observation.

Suppose one was to be tested for a disease, one is interested in the probability of having the disease given the test results (evidence), the sensitivity of the test as well as the probability of the disease (prior). Let P (Test is positive|Disease is present) = 0.95, while the probability of missing the disease is P (Test is negative|Disease is present) = 0.05. Also let P (Test is positive|Disease is NOT present) = 0.05 consequently (probabilities of various states of a random variable must add up to 1), P (Test is negative|Disease is NOT present) = 0.95. Suppose the probability of the disease in the general public is P (Disease) = 0.01 so that only 1% of the population is diagnosed with it. P (Disease is present|Test is positive) is now of interest and for this Bayes' theorem can be used (see appendix). Let D and T denote the state of the disease and test respectively.

$$P(\mathrm{D} = \mathrm{present}|\mathrm{T} = \mathrm{positive}) = \\ = \frac{P(\mathrm{T} = \mathrm{positive}|\mathrm{D} = \mathrm{present})\,P(\mathrm{D} = \mathrm{present})}{P(\mathrm{T} = \mathrm{positive})}$$

where P (T = positive) = P(T = positive|D = present) P(D = present) + P(T = positive|D = NOT present) P(D = NOT present)

The result is P (T = positive|D = present) =

$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.161.$$

This may seem a low probability, however one must consider that the probability of having the disease is on 0.01 in the general population; thus making it a relatively unlikely event, explaining the low probability. A better approach would be to conduct a second independent test and calculate P (D = present|T_1 = positive $\cap$ T_2 = positive) given by:

$$\frac{\mathrm{P}(\mathrm{T}_1 = \mathrm{positive}|\mathrm{D} = \mathrm{present})\mathrm{P}(\mathrm{T}_2 = \mathrm{positive}|\mathrm{D} = \mathrm{present})\mathrm{P}(\mathrm{D} = \mathrm{present})}{\mathrm{P}(\mathrm{T}_1 = \mathrm{positive} \cap \mathrm{T}_2 = \mathrm{positive})}.$$

This results in a probability of 0.7826, substantially higher than the original 0.161.

Now consider the case where there is many events being conditioned upon, for example P (A|B, C, D, E) where B, C, D, E may not be independent events. It can be seen that as the pieces of supporting evidence increases the computational complexity of calculating Bayes' rule increases. Bayesian networks graphically model the interdependencies between such events while providing tools for tractable updates in beliefs given relevant observations as will be seen in the coming sections.

2.2 Bayesian data analysis

Bayes' theorem plays a central role in the two critical tasks of data analysis: parameter estimation and model selection. The problem of estimating or inferring the values of the unknown quantities (parameters) based on observation of some related random quantities is a common one. The general setting involves estimating some quantity, say x, which can be random, or deterministic, based on an observation on x, say y(x). Estimation then becomes a problem of extracting x from y(x). Therefore, one can think of the parameter estimation problem as a rule or a mapping from y(x) to x. This rule itself can be deterministic, or random, as well.

An example of parameter estimation would be the prediction of the amino acids in a specific protein using incomplete data. There is a parameter(s) as well as observations in terms of experimental data. The goal is to come up with an estimator (a rule), which *best* suits the objective. The definition of best can, of course, be very different according to the specific application- but there are some common metric that are widely used such as least squares estimation or maximum *a posteriori* (MAP) [11].

Figure 1 illustrates the general estimation procedure. Here, the leftmost circle is the set from which x takes on values (parameter), the next circle is the set where y(x) takes its values from (observation). The p(y|x) is the conditional probability of observing y(x) given that x was generated (by either an experiment or nature). Generally, p(y|x) is known. Then, one can design the estimator according to the given specifications and obtain z, called the estimate of x, with the goal of having z = x. Unfortunately, this is almost never the

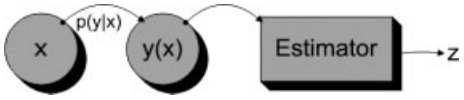

Figure 1. An overview of the estimation procedure. Here, x represents the parameter, y(x) represents the observation and after an estimator, the estimate z. In addition, p(y|x) is the conditional probability that y is generated given the value of x.

case and one generally designs the estimator as to minimize the cost function J(z,x). This function can simply be related to the distance between z and x, such as $|z-x|$ or $(z-x)^2$.

Bayesian estimation is one common and important estimation technique. Here x is regarded as random and has three elements which are crucial to Bayesian estimation:

(i) Parameter model: Since x is random, it has a distribution (*e.g.*, Gaussian distribution) and it is called the prior distribution on x.

(ii) Observation model: The observation model embodies the dependency between x and y(x). The relevancy is captured by the conditional probability distribution p(y|x).

(iii) Estimation rule: Let the cost of estimating x as z be represented J(z,x). This is called the cost function. In a Bayesian estimation problem the goal is to minimize the average of J(z,x), where the average is taken over all possible values of x. The average of J(z,x) is called the *Bayes risk*; common choices of the cost function include least squares error, $(z - x)^2$ and MAP. In MAP estimation, J(z,x) = 1 if $z = x$ and J(z,x) = 0 otherwise.

An example of an estimation problem is the measurement of protein microarray dye intensities. The intensity is not perfectly observed and noise is added, which has some intensity as well. The intensity (or more accurately, the log of the intensity) and the noise intensity can be modeled as a Gaussian distribution with respective means and covariances. Thus, the model is $y(x) = x + n$, where n denotes the noise. The p(y|x) is Gaussian for any given log of intensity x, due to the properties of a Gaussian. This other piece of information needed is the specification of a cost function so that the Bayes risk can be calculated. In this particular example, both least squares and MAP estimates yield the same result, namely the mean value of x.

The Bayesian approach is also useful in model selection. The intuition here is that the world, as one observes it, is defined by a set of processes. For instance, the process of aging affects a variety of other factors, such as physical height or marital status, but does not affect others, such as the basic eye color. When coupled with other features, such as gender, it can affect the baldness pattern or the physical shape.

One can regard a database as generated by these processes. This generation is stochastic, as these influences are often not deterministic. For example, not every male becomes balder as age increases and not every individual gets married. But in general, one expects to see such biases to show up in the data observed, unless the data have been collected with a particular intrinsic bias. For instance, a database of high school students will not be expected to show a bald pattern increase with respect to age.

The task of model selection is to reverse this generation process and, ideally, to discover the set of processes (called the model) responsible for the observed data. However, since the data observed are a stochastic manifestation of these processes, it is useful to first look for the most probable model given the data.

By considering that all alternative models constitute a set of mutually exclusive explanations for the data, one could create a probability distribution over them. Without explicit prior knowledge, one can start by assuming that this distribution is uniform, to encode the fact that all models are equally likely. One can then update this uniform prior probability on the basis of the observed data to obtain the posterior probability of the model given the data $P(M|D)$, where M is the model and D are the available data. By Bayes' theorem: $P(M|D) = (P(D|M) \times P(M))/P(D)$. Since this quantity is being used to compare models and all models are compared over the same data, $P(D)$ turns out to be a constant and can be removed. If it is assumed that all models are equally likely, then the quantity $P(M|D)$ becomes proportional to the quantity $P(D|M)$, known as marginal likelihood, which in many cases can be efficiently computed in closed-form.

While the marginal likelihood itself can be difficult to interpret, it can be used to compute a more intuitive measure known as *Bayes factor*, to assess the strength of evidence of a model M_1 against another model M_2. Since the Bayes factor is the ratio $P(M_1|D)/P(M_2|D)$ between the posterior probability of two alternative models M_1 and M_2, under the same assumptions used in the reduction above, Bayes factor can be computed as the ratio of the marginal likelihood: $P(D|M_1)/P(D|M_2)$.This more intuitive measure will tell the analyst how many times the model M_1 is more probable than M_2, and the confidence in the selected model will be a function of the distance between the selected model and the most probable alternative models [12]. The concept of the Bayes factor is similar to the LOD score commonly used in calculating genetic linkage.

2.3 Bayesian networks

Complex problems require the generalization of the conditioning procedure and more rigorous tools. Graphical models can help in terms of breaking down complex systems to simpler parts and allowing for the analysis of one variable at a time. Interestingly, the first attempt to use graphs as a means of inferring dates back to 1921 by a geneticist, Sewal Wright, who developed the method of path analysis [13]. More recently, this line of research has been developed in the artificial intelligence community [14, 15]. A parallel line of research is going on in graphical models area within statistics [16, 17]. Current Bayesian techniques to learn the graphical structure of a Bayesian network from data are based

on the evaluation of the posterior probability of network structure. That is, they use the probability of the graphical model conditional on the data.

At the core of every graphical model is a graph in which the nodes represent random variables. A graph G is an abstraction denoting relations between entities using vertex and edge sets. The vertex set V represents the entities themselves and the edge set E constitutes the relationships between the vertices. Specifically, if two entities represented by two respective vertices are related through some equivalence relationship, then an edge exists between these two vertices. Alternatively, the lack of an edge signifies a conditional independence assumption. The graphs not only help in making these equivalence relations rigorous, they also provide a pictorial representation of the abstract probabilistic notions.

A Bayesian network is a directed graph G with vertex set V and directed edge set E. It also includes an embedded conditional probability distribution. In a Bayesian network, a directed edge is typically used to represent the direction of induction. For example using a directed edge from X_1 to X_2 would mean X_1 induces X_2.

A simple Bayesian network example is illustrated in Fig. 2. Here the so called Markovianity property holds, such that conditioning on X_2 and X_3, the random quantities X_1, X_4, and X_5 are independent. The direction between the vertices also introduces new terminology. If an edge is directed from X to Y then X is a parent of Y and Y is child of X. Grandchildren and grandparents are also defined similarly. For example X_4 is a child of X_3 and X_1 is a grandparent of X_5 in Fig. 2. Directed edges can be used to determine causality, though such conclusions can only be derived under certain situations (see [18] for more details).

Bayesian networks also encode the parameters that link the vertices. An example is shown in Fig. 2. Here, X_1 represents whether or not a virus exists in a human body, X_2 and X_3 represent two possible symptoms, *e.g.*, coughing and nausea, while X_4 represents whether the person is sick or not. It possible to define the conditional probability distribution for each vertex given its parent, *i.e.*, $P(X_2|X_1)$ as shown in Fig. 3.

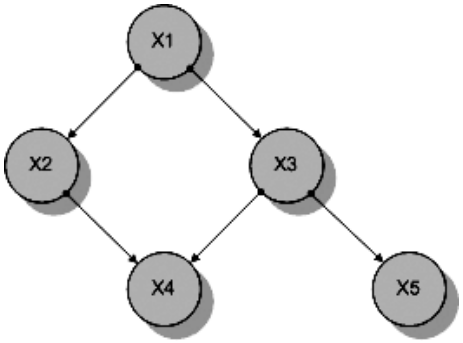

Figure 2. An example of a Bayesian network with 5 nodes.

***P*(X2|X1)**	P(X2=True)	P(X2=False)
X1=True	0.2	0.8
X1=False	0.05	0.95

***P*(X3|X1)**	P(X3=True)	P(X3=False)
X1=True	0.1	0.9
X1=False	0.02	0.98

***P*(X4|X2∩X3)**		P(X4=True)	P(X4=False)
X2	**X3**		
X2=False	X3=False	0.05	0.95
X2=True	X3=False	0.4	0.6
X2=False	X3=True	0.3	0.7
X2=True	X3=True	0.9	0.1

***P*(X5|X3)**	P(X5=True)	P(X5=False)
X3=True	0.6	0.4
X3=False	0.2	0.8

Figure 3. Conditional probability tables.

As an example, if the probability of a person visiting Asia $P(X_1 = \text{True}) = 0.5$, the probability that the person has a fever given that he has caught strain A of the bird flu ($X_2 = \text{True}$) but has not caught strain B ($X_3 = \text{False}$) after a visit to Asia is given by $P(X_4 = \text{True}|X_2 = \text{True} \cap X_3 = \text{False}) = 0.4$. Ordinarily the joint probability distribution of all the variables is given by product rule: $P(X_1, X_2, X_3, X_4, X_5) = P(X_1) \times P(X_2|X_1) \times P(X_3|X_1,X_2) \times P(X_4| X_1, X_2, X_3) \times P(X_5|X_1, X_2, X_3, X_4)$. However, due to the conditional independence defined by the Bayesian network, the prerequisite calculations can be reduced to: $P(X_1, X_2, X_3, X_4, X_5) = P(X_1) \times P(X_2|X_1) \times P(X_3|X_1) \times P(X_4| X_2,X_3) \times P(X_5|X_3)$.

Inference allows one to make predictions based on the encoded network structure and conditional probabilities. Specifically, inference refers to the ability to estimate the value of an unobserved node(s) (representing events) given the values of observed nodes.

Looking at the previous example in Fig. 2, where the binary node values represent the occurrence or non-occurrence of an event, Bayes' rule can be used to infer quantities of interest such as the probability of having caught strain B of the flu- given the person has a cough. That is, $P(X_3 = \text{True}|X_5 = \text{True})$ which, by Bayes' rule can be computed as:

$$\frac{P(X_5 = \text{True}|X_3 = \text{True})P(X_3 = \text{True})}{P(X_5 = \text{True})} = \\ = \frac{0.6 \times 0.06}{0.6 \times 0.06 + 0.2 \times 0.94} = 0.16$$

As one might expect, the probability is relatively low 16% here. This type of inference is usually referred to as bottom-

up reasoning. That is, an effect has been observed and the likelihood of a cause needs to be explained. Conversely top-down reasoning is where one observes a cause and tries to predict the possibility of an effect. For example, the probability of having a fever (X_5 = True) given the person has recently visited Asia (X_1 = True) can be predicted, that is $P(X_5 = \text{True}|X_1 = \text{True})$.

Another interesting observation is that events X_2 and X_3 become conditionally dependent once their common child X_4 is observed. That is, certain inferences can be made about the state of X_2 and X_3 if the state of one of them and their common child X_4 is known. For example one may ask what the probability of having strain A of the flu is if the person is known to have strain B and is coughing, namely

$$P(X_2 = \text{True}|X_3 = \text{True} \cap X_4 = \text{True}) =$$
$$= \frac{P(X_3 = \text{True} \cap X_4 = \text{True}|X_2 = \text{True})\ P(X_2 = \text{True})}{P(X_3 = \text{True} \cap X_4 = \text{True})} =$$
$$= \frac{P(X_3 = \text{True})\ P(X_4 = \text{True}|X_2 = \text{True} \cap X_3 = \text{True})\ P(X_2 = \text{True})}{P(X_3 = \text{True} \cap X_4 = \text{True})}$$

Notice that the probability would be different had one only observed coughing (X_4 = True). This would result in

$$P(X_2 = \text{True}|X_4 = \text{True}) =$$
$$= \frac{P(X_4 = \text{True}|X_2 = \text{True})P(X_2 = \text{True})}{P(X_4 = \text{True})}$$

So far it has been illustrated how Bayesian networks can be intuitive structures for the representation of cause and effect. It can easily be seen that as the number of nodes increases, the computation complexity of performing inference quickly grows. Much research has focused on the development of exact and approximate algorithms that minimize this complexity.

The belief propagation (BP) algorithm [1] is an example of this type of algorithm over Bayesian networks. It is based on the exchange of information between parents and their children. If the Bayesian network is loop-free, the algorithm is guaranteed to give the correct inference result.

In order to illustrate the algorithm, it is useful to first to emphasize what the problem is and what the form of the solution looks like. Previously, it was stated that the vertices of the graph are associated with random variables. These random quantities carry information and the goal of inference is to gather information in order to make predictions. To understand the form of the solution, it is useful to note that conditioning may reduce the amount of information one has to store (specifically at each vertex). Hence, it also reduces the amount of data needed in order to make inference. Therefore, by using the Markov property, one may be able to efficiently infer- even in the presence of large graphs.

If there is a directed edge between vertex X_i and vertex X_j, X_i is called the neighbor of X_j (and X_j is called the neighbor if X_i). The intuitive idea behind the BP is that once one vertex knows the exact values of its neighbors it no longer needs the values of other nodes because its Venn diagram (see Fig. 4) does not overlap with non-neighbor nodes (by Markov property). Hence, BP exploits the Markovian property of Bayesian networks in order to achieve inference accurately and efficiently.

BP works in time steps and the inference algorithm runs by exchanging data values between neighboring vertices. The exchanged data values (messages) are initialized at each vertex by assigning these initial messages to the data present in this node. Then, at the very first time step, these initial messages are sent to neighbors. Before the next time step, the received messages from all neighbors are merged at every node. Then, at the next time step, these merged messages are sent to all the neighbors. The data messages can thus spread over the entire network. For more technical details on the BP algorithm, see ref. [1].

An example of BP is illustrated in Fig. 5. Here, X_1, X_2, and X_3 represent some random variables. In addition, $M_t(X_i \rightarrow X_j)$ denotes the message sent from X_i to X_j at time step t. For example, $M_1(X_1 \rightarrow X_3)$ denotes the message at the first time step sent from X_1 to X_3. As shown in Fig. 5, X_3 can be used to make exact inference by just getting messages from X_1 but not from X_2. For simplicity, only the messages that originated for X_3 are shown. The algorithm starts at time $t = 1$. At the first time step, $M_1(X_2 \rightarrow X_1)$ is initialized to the data value at X_2 and $M_1(X_1 \rightarrow X_3)$ is initialized to the data value at X_1. Then, these messages are sent to X_1 and X_3, respectively. Before the second time step, X_1 merges the messages $M_1(X_2 \rightarrow X_1)$ into its outgoing message

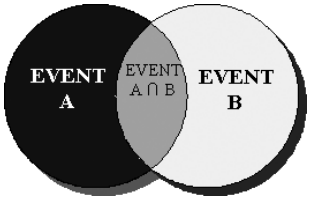

Figure 4. Two dependent events are shown in a Venn diagram. Note the overlapping of the sets.

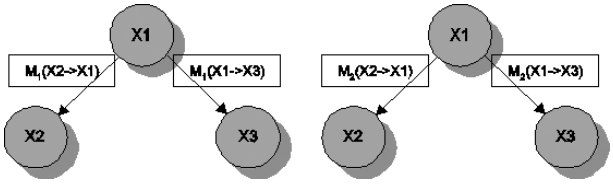

Figure 5. Two phases of the BP algorithm. The left figure shows the first message passing phase of the algorithm. X_3 does not still have the total information disseminated in the Bayesian network. However in the second phase of the algorithm X_1 sends the aggregate information, adding the message it received from X_2. Therefore X_3 has now enough information in order to complete the inference procedure. Note in both figures, X3 has no knowledge of X2.

$M_1(X_1 \rightarrow X_3)$. At time step 2, the message $M_2(X_1 \rightarrow X_3)$ is basically a superposition of $M_1(X_1 \rightarrow X_3)$ and $M_1(X_2 \rightarrow X_1)$. At the end of the second time step, X_3 is aware of all data originally disseminated in the Bayesian net and hence has enough information in order to do inference.

In addition to methods for deterministic approaches for inference, stochastic approaches, such as Gibbs sampling, are also commonly used for inference when closed-form solutions are difficult to obtain [4].

The foundations for a variety of Bayesian methods have been discussed here. For additional information on Bayesian theory and Bayesian data analysis, there are a number of relevant texts [5, 7]. In the next section, applications of these methods in the biological context, specifically in the proteomics domain, are illustrated from recent work in the area.

3 Bayesian applications

3.1 Background

The Bayesian approach to analyzing biological data and incorporating prior information has increased the flexibility and capabilities of statistical models used in scientific research. By being able to create predictions with an associated probability rather than simply yielding a binary conclusion, more actionable information is given to researchers [19]. Furthermore, Bayesian methods have the ability to detect hidden variables which affect results, but would otherwise go unnoticed in experiments [20]. Bayesian methods have impacted a number of specialties within bioinformatics [6, 21] with applications in pathway modeling, inferring cellular networks, and phylogeny linkage analysis among many others [22–25]. Researchers have used the Bayesian approach in conjunction with transcriptome analysis in order to form genetic regulatory relationships between genes [26]. Dynamics of microarray-based gene expression have also been successfully quantified and clustered *via* Bayesian methods [27]. Bayesian networks have been used to find relationships between single nucleotide polymorphisms (SNP) and clinical outcomes [28]. In studying the evolutionary conservation of protein-coding DNA sequences, the Bayesian approach helps quantify the occurrence of nonsynonymous substitutions across sequences [29]. Bayesian methods have also been successfully employed in RNA secondary structure prediction by incorporating prior knowledge of RNA structure to compare sequence alignments with similar secondary structure [21, 30].

When fewer genes than expected were found in the human genome, researchers started to look elsewhere for answers. Proteomics is one such area of focus due to the wide diversity of protein function and interaction. Its implications extend beyond structure and function, but also into localization, protein–protein interactions (interactome), and PTM [31, 32]. As proteomics is a burgeoning field often involving large amounts of raw data and incomplete information, it lends itself well to the use of Bayesian analysis [33].

Though it is a relatively new field, proteomics already has a well established territory. It includes peptide sequencing, protein structure determination (from folding to full 3-D form), protein localization, function prediction, protein–protein interaction study (*e.g.*, using the popular yeast two-hybrid technique), and protein profiling (*i.e.*, *via* MS, 2-DE, and microarrays). Due to its versatility and ability to incorporate disparate and incomplete data into analysis, Bayesian methods have started to penetrate many aspects of proteomics.

In the following section, we wish to cast a wide net across many disciplines of proteomics which utilize Bayesian methods. Organization is based on type of research and the content outlines current modes of analysis which use the Bayes rule.

3.2 Peptide sequencing and phylogeny

One of the more prevalent methods of protein function determination is comparing the function of a characterized protein with an uncharacterized sequence, holding the assumption that favorable traits are evolutionarily conserved [34]. If similar sequences and secondary structures are found across a set of proteins, a similar function is likely to be conserved among them. Levy *et al*. recently employed univariate and multivariate Bayesian approaches that predict the probability that an uncharacterized peptide sequence has a particular function. They verified their method by comparing them with enzymes of known function and found that their approach yielded significantly better results than traditional statistical comparisons using BLAST [34].

A similar approach of predicting protein function from primary structure involves the use of structural multilevel neighborhoods of atom (MNA) descriptors. As described by Fomenko *et al*., MNA descriptors are strings depicting an atom and its neighbors. They are a new way of symbolizing a protein, containing information on protein structural features and thus are more informative than traditional amino acid sequences [35]. An MNA descriptor at level "0" is a single atom while one at level "1" is the atom and all atoms bonded to it by a single covalent bond. Fomenko *et al*. [35] showed that by using MNA descriptors and Bayesian methods together, accurate enzyme specificity predictions result.

Though current protein sequence is integral to understanding protein function, the analysis of ancient protein sequences may further such studies. In a study done at the Bellingham Research Institute, Barry Hall concluded that Bayesian methods as employed by the program MrBayes (version 3.11) were preferred over both parsimony and maximum likelihood approaches in deducing ancestral protein sequences [36]. In fact, the Bayesian framework was perfect in its correction of gaps in sequence in 25 out of 29 reconstructed sequences [36].

A prerequisite to deeply understanding protein function, however, is a firm grasp of protein space. This requires the formation of a similarity network. Unfortunately, many current networks rely on using a single similarity measure; selecting a similarity of variable quality can therefore greatly affect results [37]. Recently, a novel approach using Bayesian statistical analysis was developed which makes possible the development of multiattribute networks. The researchers created a Bayesian model to automatically classify new proteins based on structure. This Bayesian approach can classify an uncharacterized protein with greater accuracy than previous single attribute networks; in addition, this novel approach was shown to be more effective than the existing kernel-based information integration approach [37].

In order to differentiate between homologous and non-homologous proteins, the correct alignment between their sequences must be maintained. Thus another application of Bayesian techniques in sequencing involves peptide sequence alignment. Existing amino acid substitution matrices include BLOSUM and PAM. However, these matrices do not take into account the effect of structure on the conservation of amino acid sequence [38]. Sequence alignments can be created *via* the dynamic programming algorithm, but this method relies on a single, optimal solution for aligning sequences. While this solution has the highest score, it is not necessarily completely correct. Huang and Bystroff were interested in investigating the sub-optimal alignments as well. They found that the Bayesian adaptive alignment algorithm allowed them to model all the alignments. According to Huang, an additional 34% out of 147 reference alignments were improved using Bayesian adaptive methods compared to Dynamic programming. Using local structure predictions from HMMSTR (Hidden Markov Model for protein STRucture), researchers at the Rensselaer Polytechnic Institute (RPI) developed a novel model known as HMMSUM (HMMSTR-based SUbstitution Matrices) which is a set of 281 matrices, implementing the Bayesian Adaptive alignment method (see Fig. 6) [38]. According to the researchers at RPI, alignment predictions made by HMMSUM compare favorably to those made by BLOSUM [38].

In the same effort to deduce conserved peptide positions, the program *ConSurf* (http://consurf.tau.ac.il/) was developed to assist researchers by automatically calculating conservation scores as well as providing graphical mapping [39]. The development team used a maximum-likelihood method in a pervious version of the program. Consurf, a web-based application (see Fig. 7), now uses an empirical Bayesian method as it was found to be more accurate, especially when the number of sequences inputted was small [39, 40].

Looking at the rate of amino acid evolution in addition to position is important to our understanding of the proteome as well. Mayrose *et al.* [41] have developed a Bayesian approach using a Markov Chain Monte Carlo approach to predict all potential phylogenetic trees. This approach analyses all possible combinations of parameters and trees. This method does not rely on a single phylogenetic tree, but samples from a complete set. In their research they showed that this method performed better than methods involving a single tree [41].

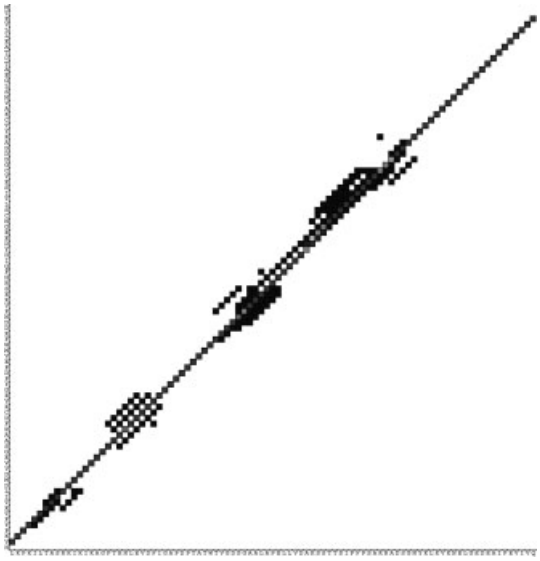

Figure 6. The histogram is generated by Bayesian adaptive alignment algorithm with the HMMSUM-D model. It represents the alignment of two remote homologous proteins with PDB code 1TLK and 1WIT. The color scheme shows the probability of aligning two positions between two sequences. Red and blue indicate high and low probabilities respectively, and white is zero probability. Courtesy of Yao-Ming Huang.

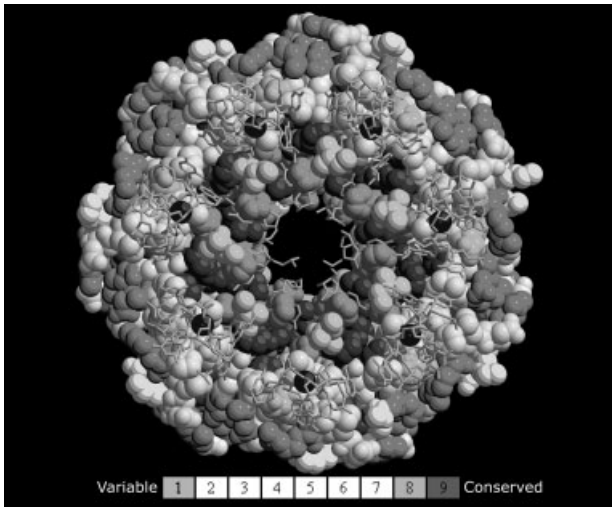

Figure 7. A ConSurf analysis of the evolutionary conservation profile of the RNA-binding protein Sm. The heptameric Sm ring (PDB code 1m8v) is presented using a space-filled model. The sticks mark the uridine heptamer, surrounding the Ca^{2+} ions. The amino acids are categorized by their conservation grades using a gradient-coding bar. The analysis demonstrates that the uridine-specific binding pocket in the internal cavity is highly conserved. Courtesy of Nir Ben-Tal.

Wickstead and Gull also contributed to the advance in protein phylogeny. In order to map phylogenetic trees, one needs a versatile model. This model must take into account that while there are areas of conservation across proteins, proteins evolve at different speeds as a result of different stimuli. Wickstead and Gull considered the Bayesian approach using priori statistical information to estimate a posterior distribution to be more robust than maximum-likelihood approaches which make one estimate of parameters from data. According to Wickstead, there has not been definitive proof showing one method being greater than another, but from personal experience, he finds that Bayesian approaches lead to trees that make more biological sense. Therefore, utilizing a Bayesian framework, they made inferences of the topology of phylogenic trees, identifying three new families and two new phylum-specific groups [42]. They used the package "MrBayes," created by John Huelsenbeck and Fredrik Ronquist, a tool that has recently become wildly popular in phylogenetic study. Wickstead and Gull used MrBayes to infer a tree from a distribution of Bayesian posterior probabilities for tree topologies. The authors found that this method was perfect for parallel processing and attribute their success of building such complicated trees in a practical timeframe to Bayesian methods. The Bayesian approach employed by Wickstead and Gull was extremely successful and has great implications for the further understanding protein evolution.

It is important to note that while Bayesian methods have helped the development of phylogenetic analysis, along with it comes several limitations. Sawa *et al.* in their comprehensive review of modern approaches to phylogenetic analysis cite that by treating each parameter as random, each with a probability distribution, it makes calculations overly complicated. They further say that due to this complication, methods such as Monte Carlo simulations are necessary to unravel and solve the Bayesian outputs [43].

In terms of peptide alignment, Webb *et al.* developed a novel method of sequence alignment determination which exceeded the capabilities of existing algorithms. They created the Bayesian algorithm for local sequence alignment (BALSA) which offers all potential alignments and compensates for unknown variables [44]. By Bayesian application, the posterior distribution for a particular alignment A* is described by the equation [44]:

$$P(\mathrm{A}^*|\mathrm{R}^{(1)},\mathrm{R}^{(2)}) = \frac{\sum_{\Theta,\Lambda} P(\mathrm{R}^{(1)},\mathrm{R}^{(2)}|\mathrm{A}^*,\Theta)P(\mathrm{A}^*|\lambda_0,\lambda_e)}{\sum_{\mathrm{A}}\sum_{\Theta,\Lambda} P(\mathrm{R}^{(1)},\mathrm{R}^{(2)}|\mathrm{A},\Theta)P(\mathrm{A}|\lambda_0\lambda_e)}$$

Here, it is defined that for a pair of sequences of interest, the data obtained is organized as vectors $\mathrm{R}^{(1)}$ and $\mathrm{R}^{(2)}$, A defines the alignment matrix, and $\Lambda = (\lambda_o,\lambda_e)$ denotes a set of previously determined gap odds ratios. Θ describes a set of sequence alignment scoring matrices (*e.g.*, BLOSUM) [44]. When compared to the popular alignment algorithm, SSEARCH, which employs a local dynamic programming algorithm, BALSA (http://www.wadsworth.org/resnres/bioinfo/) performed with a higher detection of homologs: 38 to 41.3%, respectively [44].

Sequence alignment can also be used for other applications. For example, Siddharthan *et al.* recently used a Bayesian approach to locate the binding sites of regulatory proteins. Current methods of sampling over related species present some problems, including the phylogenetic relationship and peptide conservation across specie as well as multiple alignments [45]. Their Bayesian methods technique was an attempt to get around these prior limitations. In developing PhyloGibbs (http://www.imsc.res.in/~rsidd/phylogibbs/), they found that on average it finds over 50% of the binding sites in *Saccharomyces cerevisiae* [45].

Bayesian analysis has had a tremendous impact in facilitating the link between peptide sequencing and protein function. It has also proved fruitful in the investigation of conservation of protein sequence and peptide alignment. Future steps will include the refinement and automation of these Bayesian methods while working to lessen their limitations.

3.3 Protein localization

Knowing where proteins are physically located within a cell has many potential implications in research. For example, it gives information as to which proteins can interact with each other, shows how particular cell signaling pathways work, and allows better prediction of function for proteins of unknown characterization.

Drawid and Gerstein used a Bayesian system to predict subcellular localization of proteins in yeast. This system utilized 30 features that ranged from genome data to specific characteristics of a sequence [46]. By using Bayesian logic, these features allowed for a constant update of a protein's probability of a particular localization. The Bayes' rule that they followed is described as:

$$p_{\mathrm{m}}(\mathrm{L}|\mathrm{feature}) = p_{\mathrm{m}}(\mathrm{L}) \times p(\mathrm{feature}|\mathrm{L})/\mathrm{Z}$$

Here, *L* denotes the location of the protein in question, *Z*, the normalization factor, and "feature" relates a feature. In trials of 1300 proteins, their method was accurate in 75% of protein localization predictions [46]. A year later, using Bayesian reasoning, Kumar *et al.* (including Drawid and Gerstein) created a proteome localizome (localization of proteins) for yeast. They predicted that 47% of proteins were cytoplasmic, 13% mitochondrial, 13% exocytic, and 27% were nuclear [47].

Researchers have since been trying novel methods to fully realize the localizome of yeast. One such method by Scott *et al.*, is a Bayesian network predictor named PSLT2. PSLT2 takes into account possible variable characteristics among proteins and outputs localization predictions into nine possible locales: the ER, Golgi apparatus, cytosol, nucleus, peroxisome, plasma membrane, lysosome, mito-

chondrion, and extracellular space [48]. When PSLT2 was compared to experimental data, the only differences were those proteins involved in secretion [48].

To assist researchers in such endeavors, a new program called Proteome Analyst (PA) was recently developed by researchers at the University of Alberta. A web-based program (http://www.cs.ualberta.ca/~bioinfo/PA/), PA predicts characteristics of proteins across a proteome [49]. Its authors report that it is currently the most accurate program in predicting subcellular localization. Using a Bayesian framework, however, it boasts two additional features: the ability to create user-defined classifications without programming as well as providing an intuitive explanation as to why one hypothesis was picked over another (see Fig. 8) [49].

The frontier of protein localization now involves developing methods or complement of methods that can fully determine multiple eukaryotic localizomes. By taking advantage of the Bayesian ability to continuously update localization data, this goal is becoming increasingly realizable.

3.4 Protein–Protein interactions and cell signaling

Similar to the thrust to complete the localizome, there has been a strong push in research of protein–protein interactions (*i.e.*, interactome). The interactome is unique because it relies on all levels of protein traits, from sequence to active site, localization to structure.

Edwards *et al.*, has noted that a portion of published protein–protein interactions differed from the known 3-D structures of several proteins [50]. This was significant because it showed that prior techniques for interaction determination were not accurate enough and produced many false-positives. This realization prompted them to develop a Bayesian approach which significantly decreased error rates [50]. The Bayesian method used is described by the equation in which I corresponds to the likelihood of an interaction:

$$O(I|e_1, e_2 \ldots e_N) = L(e_1, e_2 \ldots e_k|I)O(I)$$

In this equation, input source k contain information in e_k which either strengthens or weakens a protein–protein interaction hypothesis. Using structural information for RNA Polymerase II, false-positives ranges from 41 to 67%, however, after incorporation of their Bayesian network, the error rate turned to 30% [50].

In other work, Mark Gerstein and his laboratory have combined genomic features that are not strongly linked with interaction, such as colocalization, into a Bayesian networks approach in order to predict protein–protein interactions across yeast [8]. Reasons they chose the Bayesian networks approach include its ability to accommodate disparate types of data and ability to work with missing data. Their method can often incorporate experimental interaction data into its analysis [8]. Gerstein created two networks, the first is the "probabilistic interactome experimental" (PIE), including protein interaction data from literature sources. The second is "probabilistic interactome predicted" (PIP), with information that is not strongly associated with protein interaction, such as information on biological function. The intention of PIE is to view interactions based on interaction data while PIP predicts interactions based on data not directly related. They found that PIP and PIE had similar output while PIP had wider coverage. Upon comparison, their method outperforms existing high-throughput analytical techniques.

There has also been a focus on cell signaling protein interactions such as G-proteins. The interaction between GTP-binding proteins and G-protein coupled receptors (GPCRs) has been intensely studied, as it is crucial to understanding how cells communicate with each other. Despite this study, there remains the inability to use peptide sequence to predict G-protein interaction. In an effort to do this, a Bayesian method was developed to predict such coupling in which each GPCR domain, out of 80 with known coupling, was treated as a unique random variable [51]. Out of 55 that were successfully tested, 72% were properly classified.

Sachs *et al.*, in an effort to understand protein-signaling, measured how phosphorylated proteins and phospholipids change in human immune cells while disturbing the environment with interfering molecules. These data were analyzed in with a Bayesian framework, allowing the prediction of a new protein-signaling network [20]. Reviewing the eloquent way Sachs *et al.* described the framework of their method, consider Fig. 9. The arc from A to B shows that A is the "parent" of B. We also know that if A were activated, it's possible that B and C would be observed. If inhibited, we might hypothesize an absence of B and C. If B were inhibi-

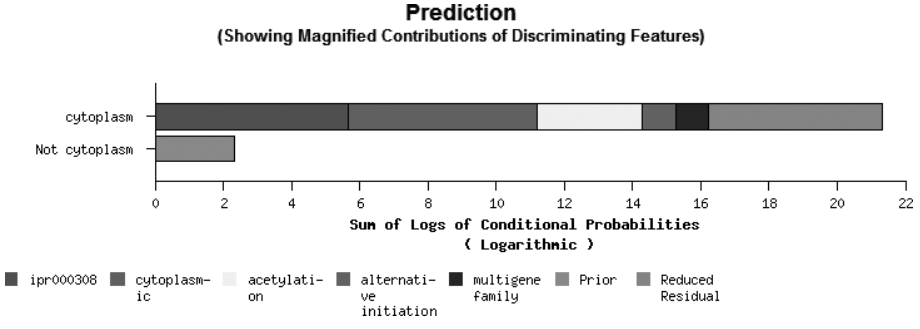

Figure 8. Prediction output screenshot of the Proteome Analyst. This figure shows the program's rationale for predicting that this protein is located in the cytoplasm and not elsewhere. Courtesy of Kurt McMillan.

ted, however, we would still see A. Flow cytometry allows the prediction of relationships between different proteins. With this method, there is no need to define a specific arc between A and C; the relationships between A and B as well as B and C are sufficient. If, however, B was not determined, this method would still observe a relationship between A and C anyway. Similarly, a relationship between C and D does not have to be explicitly addressed as they are linked *via* B following the same logic. Using single-cell flow cytometry in this way allowed Sachs *et al.* to measure the effects of the interfering molecules acting *in vivo*, thereby determining a causality influence map (see Fig. 10) [20].

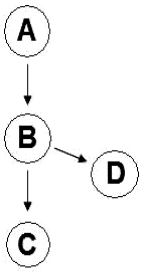

Figure 9. An example of how Bayesian methods apply to the experiment by Sachs *et al.* [20].

Another application lies in transcription factor interaction with genes. Creating a combined probability model of both transcription factor binding and gene expression levels allows researchers to better characterize and cluster genes. Barash and Friedman created such a model and developed a search method which uses the Bayesian approach. Among many other attributes, their method is unique in its ability to handle unrelated data to clustering into the analysis [52].

Perhaps the best applications of Bayesian methods are when they are performed in conjunction with other methods, augmenting one's weakness with another's strength. Nariai *et al.* describe an experiment in which they used protein–protein interaction information in combination with DNA microarray data and a Bayesian system. The rationale was that protein–protein interactions are fundamentally important to understanding networks and only using information from nucleic acid would provide an incomplete picture [53]. Nariai *et al.* maximize posterior probability of the graph G as described in the equation:

$$\pi(G|X) \propto \pi(G) \int \prod_{i=1}^{n} f(x_{i1}, \ldots, x_{ip}|\theta_G)\pi(\theta_G|\lambda)d\theta_G$$

Here, $\pi(G)$ denotes the graph's prior probability, $\pi(\theta_G|\lambda)$ describes the prior distribution on θ_G, the parameter, given hyperparameter vector λ. Analysis using data from the *S. cerevisiae* cell cycle showed that this approach improved predicted gene networks. The authors expect that their next steps include incorporating other information, such as DNA–protein interactions, in their investigation of gene maps.

3.5 Protein profiling using MS and microarrays

Two popular methods to characterizing and profiling proteins include MS and protein microarrays. MS is a technique used to identify the m/z of ions. In proteomics, MS is used to identify proteins *via* PMF or quantify them with the incorporation of heavy, but stable, isotopes. Protein microarrays have been developed that function by immobilizing proteins to glass and applying binding molecules to form complexes with these fixed proteins. This binding is illuminated by a detection system, a common one of which is a fluorescent marker.

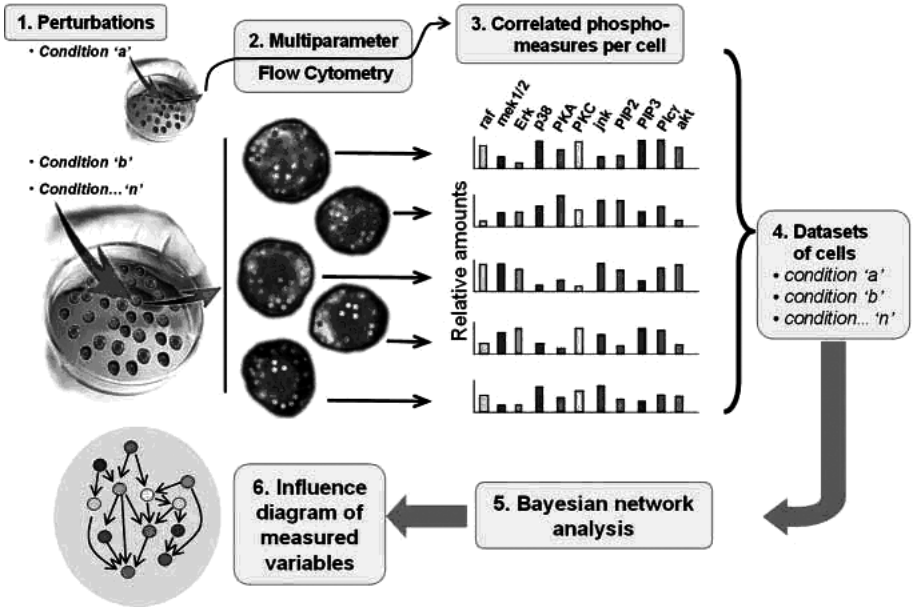

Figure 10. The Bayesian method inference protocol using multiparameter flow cytometry used by Sachs *et al.* [20]; reprinted with permission from AAAS.

Zhang and Chait, at Rockefeller University created the protein search engine *ProFound* [54]. They found that Bayesian methods were optimal for their purpose due to the open-ended capability to add restraints as they become available. This program, by means of a Bayesian approach, takes MS data to identify proteins from existing databases. These proteins are then ranked based on the probability of correspondence given the information known about the sample. The equation used to determine ranking is:

$$P(k|DI) \propto P(k|I)\frac{(N-r)!}{N!}\prod_{i=1}^{r}\left\{\sqrt{\frac{2}{\pi}}\frac{m_{\max}-m_{\min}}{\sigma_i}\times \right.$$
$$\left.\times\sum_{j=1}^{g_i}\exp\left[-\frac{\left(m_i-m_{ij0}\right)^2}{2\sigma_i^2}\right]\right\}F_{\text{pattern}}$$

The variable k represents our protein of interest, D, the data obtained from analysis and I, any background information available about k. Given experimental data and background protein information, the sum of all the different hypotheses is normalized to 1. Zhang and Chait concluded that *ProFound* identified the correct protein even when the data was of low quality and when the item was apart of a large protein mixture.

According to Ramakrishnan *et al.*, *ProFound* is a program utilizing a fine filter. This means a large amount calculations and comparisons are needed in its analysis. Ramakrishnan *et al.* recently explored the method of analysis using a coarse, but lossless, filter [55]. This coarse filter reduces search time, which would otherwise be quite significant in high-throughput analysis. This type of filter analysis has produced results consistent with other industry programs including SEQUEST [55].

SEQUEST is a program which finds associations between MS/MS data and protein sequence databases. Since its initial development, several statistical packages have been developed to verify the associations, noting the difficult balance between too much stringency and false positives [56]. One such package is PeptideProphet, by the Institute of System Biology, which uses the Fisher's Linear Discrimination Analysis, the Expectation-Maximization algorithm and the Bayes Rule to convert SEQUEST results into a probability score [57]. These packages have been shown to increase the accuracy of associations and are an asset to the shotgun proteomics field.

Researchers at the University of Texas Anderson Cancer Center have recently come up with a Bayesian wavelet-based functional mixed model to analyze MALDI-TOF MS data. Their method is promising because by treating data as functions, peak analysis can be avoided [9]. Furthermore, the generated output may be used to determine posterior probabilities, incorporating information related to both practical and statistical significance. Upon analysis, their method found many differentially expressed spectral regions. While these corresponded to peaks in many, some could not have been found by peak analysis alone [9]. The worth of such a method is therefore clearly visible.

One issue in MALDI or SELDI analysis is that the results often yield unidentified protein biomarkers. Recent work has developed Bayesian network-based methodologies with the goal of identifying such proteins [58]. It is based on the idea that protein network perturbations are relayed throughout constituent links in a manner that identifies the underlying nodes *via* their relationships.

Other developments in protein identification include the integration of Bayesian analysis for MS with LC. Chen *et al.* recently developed a LC-MS method using Bayesian scoring which does high-throughput analysis to identify characteristics of unknown proteins. These characteristics were then matched with previous databases created from LC-MS/MS tests and scored based on overlap [7]. In a single LC-MS analysis, this novel Bayesian method identified six times the number of proteins compared to previous nonBayesian analysis [7].

The development of microarrays gave researchers a large amount of raw data. A wealth of information could be potentially gleaned with the proper methods to parse through this data and extract meaning from it. Unfortunately, there are currently two main problems which limit microarray effectiveness. The first is that the number of trials is small; in order to perform a statistically valid experiment, there must be a large-enough sample size [59]. Secondly, the numerous quantity of genes in each microarray analysis leads to genes being analyzed more than once [59].

Developments in microarray technology using Bayesian logic have worked to correct these problems. Unfortunately, they have also largely been in the realm of DNA microarrays. These techniques, however, can easily be transferred to protein microarrays as the methods are analogous. One example of a Bayesian method in DNA microarrays that could be applied in proteomics with minor modifications is the Bayesian analysis of gene expression levels (BAGEL) method as described by Meiklejohn and Townsend. This method compensates for the errors commonly produced by spot effects and spot saturation [60]. One of the benefits of BAGEL lies in its ability analyze differences in gene expression levels, rather than simply looking at whether genes are expressed or not.

3.6 Bayesian applications in other disciplines

Bayesian methodology is a widely used tool not just restricted to proteomics. For example, in software engineering, The Lumiere Project of Microsoft (research.microsoft.com/~horvitz/lumiere.htm) develops Bayesian user models which incorporate information about the user's actions to develop better interface software. These concepts were behind the development of the Office Assistant in Microsoft Office '97. In the realm of clinical nursing, Bayesian reasoning is being investigated to help with nursing judgment and accepting uncertainty [61]. Also, with respect to transportation and railroad safety, Bayesian methods have been used to measure safety decision benefits while observing uncertainty [62].

4 Conclusion

While proteomics itself is a new science, Bayesian methods have been used for quite some time. In their prior use in computational biology and genomics, Bayesian approaches have proved invaluable. The same is now proving to be true in proteomics; significant progress is already being made. Using a Bayesian statistical approach to experimentation could save researchers both time and money in tedious and noisy experiments. While Bayesian methods may not be a panacea to all of the challenges of proteomics, future refinements and development of hybrid techniques may lead to further integration opportunities and applications that bridge proteomics with related fields.

This work was supported in part by the National Library of Medicine (NLM/NIH) under grant 5T15LM007092 and the National Human Genome Research Institute (NHGRI/NIH) under grant 1R01HG003354.

5 References

[1] Alterovitz, G., Afkhami, E., Barillari, J., Ramoni, M., in: Akay, M. (Ed.), *Encyclopedia of Biomedical Engineering*, John Wiley & Sons, New York 2006.

[2] Alterovitz, G., Staelin, D. H., Philip, J. H., *J. Clin. Monit. Comput.* 2002, *17*, 351–359.

[3] Alterovitz, G., Liu, J., Chow, J., Ramoni, M. F., *Proteomics* 2006, *6*, 4016–4022.

[4] Gelman, A., Carlin, J. C., Stern, H., Rubin, D. B., *Bayesian Data Analysis*, Chapman & Hall, New York 1995.

[5] Ramoni, M. F., Sebastiani, P., in: Hand, D. J. (Ed.), *Intelligent Data Analysis: An Introduction*, Springer, New York 2003, pp. 128–166.

[6] Alterovitz, G., Ramoni, M. F. (Eds.), *Systems Bioinformatics: An Engineering Case-Based Approach*, Artech House, Boston, MA 2007.

[7] Chen, S. S., Deutsch, E. W., Yi, E. C., Li, X. J. *et al.*, *J. Proteome Res.* 2005, *4*, 2174–2184.

[8] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y. *et al.*, *Science* 2003, *302*, 449–453.

[9] Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., Coombes, K. R., *UT MD Anderson Cancer Center Department of Biostatistics and Applied Mathematics Working Paper Series* 2006, Working Paper 22.

[10] Liebler, D. C., *Introduction to Proteomics: Tools for the New Biology*, Humana Press, Totowa, NJ 2002.

[11] Papoulis, A., Pillai, S. U., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York 2002.

[12] Kass, R. E., Raftery, A. E., *J. Am. Stat. Assoc.* 1995, *90*, 773–795.

[13] Wright, S., *Ann. Math. Stat.* 1934, *5*, 161–215.

[14] Lauritzen, S. L., Spiegelhalter, D. J., *J. R. Stat. Soc. B Met.* 1988, *50*, 157–224.

[15] Cooper, G. F., Herskovitz, G. F., *Mach Learn* 1992, *9*, 309–347.

[16] Whittaker, J., *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, New York 1990.

[17] Lauritzen, S. L., *Graphical Models*, Clarendon Press, Oxford 1996.

[18] Pearl, J., *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK 2000.

[19] Xia, Y., Yu, H., Jansen, R., Seringhaus, M. *et al.*, *Annu. Rev. Biochem.* 2004, *73*, 1051–1087.

[20] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., Nolan, G. P., *Science* 2005, *308*, 523–529.

[21] Ding, Y., *RNA* 2006, *12*, 323–331.

[22] Needham, C. J., Bradford, J. R., Bulpitt, A. J., Westhead, D. R., *Nat. Biotechnol.* 2006, *24*, 51–53.

[23] Woolf, P. J., Prudhomme, W., Daheron, L., Daley, G. Q., Lauffenburger, D. A., *Bioinformatics (Oxford, England)* 2005, *21*, 741–753.

[24] Jiang, Z. F., Huang, D. W., Zhu, C. D., Zhen, W. Q., *Mol. Phylogenet. Evol.* 2006, *38*, 306–315.

[25] Alfaro, M. E., Huelsenbeck, J. P., *Syst. Biol.* 2006, *55*, 89–96.

[26] Li, H., Lu, L., Manly, K. F., Chesler, E. J. *et al.*, *Hum. Mol. Genet.* 2005, *14*, 1119–1125.

[27] Ramoni, M. F., Sebastiani, P., Kohane, I. S., *Proc. Natl. Acad. Sci. USA* 2002, *99*, 9121–9126.

[28] Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T., Steinberg, M. H., *Nat. Genet.* 2005, *37*, 435–440.

[29] Huelsenbeck, J. P., Jain, S., Frost, S. W., Pond, S. L., *Proc. Natl. Acad. Sci. USA* 2006, *103*, 6263–6268.

[30] Knudsen, B., Hein, J., *Bioinformatics (Oxford, England)* 1999, *15*, 446–454.

[31] Davis, T. N., *Curr. Opin. Chem. Biol.* 2004, *8*, 49–53.

[32] Stults, J. T., Arnott, D., *Methods Enzymol.* 2005, *402*, 245–289.

[33] Eddy, S. R., *Nat. Biotechnol.* 2004, *22*, 1177–1178.

[34] Levy, E. D., Ouzounis, C. A., Gilks, W. R., Audit, B., *BMC Bioinformatics* 2005, *6*, 302.

[35] Fomenko, A., Filimonov, D., Sobolev, B., Poroikov, V., *OMICS* 2006, *10*, 56–65.

[36] Hall, B. G., *Proc. Natl. Acad. Sci. USA* 2006, *103*, 5431–5436.

[37] Camoglu, O., Can, T., Singh, A. K., *Bioinformatics (Oxford, England)* 2006.

[38] Huang, Y. M., Bystroff, C., *Bioinformatics (Oxford, England)*2006, *22*, 413–422.

[39] Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F. *et al.*, *Nucleic Acids Res.* 2005, *33*, W299–302.

[40] Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., *Mol. Biol. Evol.* 2004, *21*, 1781–1791.

[41] Mayrose, I.,Mitchell, A.,Pupko, T., *J. Mol. Evol.* 2005, *60*, 345–353.

[42] Wickstead, B., Gull, K., *Mol. Biol. Cell* 2006, *17*, 1734–1743.

[43] Sawa, G., Dicks, J., Roberts, I. N., *Brief. Bioinform.* 2003, *4*, 63–74.

[44] Webb, B. J., Liu, J. S., Lawrence, C. E., *Nucleic Acids Res.* 2002, *30*, 1268–1277.

[45] Siddharthan, R., Siggia, E. D., van Nimwegen, E., *PLoS Comput. Biology* 2005, *1*, e67.

[46] Drawid, A., Gerstein, M., *J. Mol. Biol.* 2000, *301*, 1059–1075.

[47] Kumar, A., Agarwal, S., Heyman, J. A., Matson, S. *et al.*, *Genes Dev.* 2002, *16*, 707–719.

[48] Scott, M. S., Calafell, S. J., Thomas, D. Y., Hallett, M. T., *PLoS Comput. Biol.* 2005, *1*, e66.

[49] Szafron, D., Lu, P., Greiner, R., Wishart, D. S. *et al.*, *Nucleic Acids Res.* 2004, *32*, W365–371.

[50] Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D. *et al.*, *Trends Genet.* 2002, *18*, 529–536.

[51] Cao, J., Panetta, R., Yue, S., Steyaert, A. *et al.*, *Bioinformatics (Oxford, England)* 2003, *19*, 234–240.

[52] Barash, Y., Friedman, N., *J. Comput. Biol.* 2002, *9*, 169–191.

[53] Nariai, N., Kim, S., Imoto, S., Miyano, S., *Pac. Symp. Biocomput.* 2004, 336–347.

[54] Zhang, W., Chait, B. T., *Anal. Chem.* 2000, *72*, 2482–2489.

[55] Ramakrishnan, S. R., Mao, R., Nakorchevskiy, A. A., Prince, J. T. *et al.*, *Bioinformatics (Oxford, England)* 2006.

[56] Nesvizhskii, A. I., Aebersold, R., *Drug Discov. Today* 2004, *9*, 173–181.

[57] Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A. *et al.*, *Proteomics* 2005, *5*, 3475–3490.

[58] Alterovitz, G., *Ph.D. Thesis*, Massachusetts Institute of Technology, Cambridge 2005, p. 89.

[59] Gottardo, R., Pannucci, J. A., Kuske, C. R., Brettin, T., *Biostatistics* 2003, *4*, 597–620.

[60] Meiklejohn, C. D., Townsend, J. P., *Brief. Bioinform.* 2005, *6*, 318–330.

[61] Harbison, J., *J. Clin. Nurs.* 2006, *15*, 1489–1497.

[62] Washington, S., Oh, J., *Accident; Analysis and Prevention* 2006, *38*, 234–247.

6 Appendix – Review of Bayes theorem

Consider an experiment where one randomly picks numbers between 1 and 20 out of a hat, so that the probability of picking any one number is 1/20 or 0.05. Let A be the event that the number is divisible by 2, this will be the set {2, 4, 6, 8, 10, 12, 14, 16, 18, 20}. Let B be the event that the number is divisible by 3, this will be the set {3, 6, 9, 12, 15, 18}. The Venn diagram representation of Events A and B is given in Fig. 4. Here, there are two dependent events represented by the blue and yellow sets. The common elements in the two sets are the numbers {6, 12, 18}. The probability of: event A is 10/20 = 0.5, event B is 6/20 = 0.3, and event A and B together occurring is 3/20 = 0.15.

This logic can be extended to arbitrary events A and B. Letting A and B be events such that the $P(\mathrm{A}) \neq 0$. Then, the conditional probability of B given A, denoted as $P(\mathrm{B}|\mathrm{A}) = P(\mathrm{A} \cap \mathrm{B})/P(\mathrm{A})$ where $P(\mathrm{A} \cap \mathrm{B})$ is the probability of both events occurring. Through a algebraic manipulation, Bayes' theorem can be ascertained:

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B})P(\mathbf{B})}{P(\mathbf{A})}$$

In Bayesian terminology $P(\mathrm{B}|\mathrm{A})$ is referred to as the posterior, denoting the probability of event B taking place given that event A has occurred. It is also customary to refer to the event being conditioned upon as the evidence, in this case event A. $P(\mathrm{A}|\mathrm{B})$ is the likelihood of the evidence, A, being observed if the event of interest, B, had taken place. $P(\mathrm{B})$ is the probability of event B, irrespective of any other observations and is called the prior. $P(\mathrm{A})$ is a normalizing constant that ensures the posterior adds up to 1.

