Activities of the KSU Bioinformatics and Medical Informatics (BMI) Working Group, 2001-2002

Technical Report
Department of Computing and Information Sciences
Kansas State University

William H. Hsu

1 Introduction to the Kansas State BMI Group

This group has explored the use of probabilistic representation, learning, and reasoning in:

Data mining for decision support: In cooperation with the **National Center for Supercomputing Applications (NCSA)** at the University of Illinois, where we retain a visiting appointment, we have assisted in development of three NSF REU programs and the *D2K* reference architecture for KDD [91]; locally, we developed *BNJ* [60] and *MLJ* [63] for use in our own and collaborators' courses. These have lead to published results on commercial KDD for decision support [55, 56, 61, 65] and to development of a graduate degree certificate, distance learning courses, public course tools, and software for PRISM [22]. We are also applying these tools to decision support in transportation engineering.

Bioinformatics: Over the past year and a half, we have entered into three fruitful collaborations with genome researchers. **First,** to facilitate progress towards our goal and projects that apply probabilistic learning and reasoning to gene expression modeling, we have established a bioinformatics working group with colleagues in biology, agronomy, and electrical and computer engineering. This group has outlined a project to develop temporal probabilistic models of gene regulatory dynamics in the model plant *Arabidopsis thaliana.* We believe this will contribute towards the *Project 2010* challenge problem of understanding the function of all *A. thaliana* genes, enabling construction of a "virtual plant" by the end of the decade [119]. Our working group has identified two **independent milestones** for this part of the project: (*i*) collection of microarray data from experimental treatments of *A. thaliana* and (*ii*) development and evaluation of robust learning algorithms for graphical probabilistic models using this data and existing repository data [117, 118]**.** To this we add (*iii*) development of data models and collection of content and historical user data and (*iv*) development, evaluation, and deployment of inference and decision support techniques for a collaborative filtering system. **Second,** we serve on a curriculum and faculty hiring committee in computational life sciences whose primary mission is to develop an interdisciplinary undergraduate specialization in bioinformatics [49]. **Third,** we have begun to work with bioinformatics groups at NCSA, **The Institute for Genomic Research (TIGR)**, and the University of Manchester, the lead institution of the ***myGrid*** initiative [121] and a partner of the **European Bioinformatics Institute (EBI)** [31].

2 Background

Because we will build on the existing XML implementation of the ***MAGE (Microarray Gene Expression)*** [83] data standard developed by the **Microarray Gene Expression Database (MGED)** group [84], we provide a short summary of *MAGE*, its ancestors, and an ontology, *TAMBIS* [8, 12, 14], that is important to the realization of our system. Other related work is discussed in other sections where relevant.

We seek to take existing ontologies and minimum information standards for computational genomics and create a refined and elaborated data model for decision support in retrieving data, metadata, and source codes to serve researchers. A typical collaborative filtering scenario using a domain-specific research index or portal [92, 109] is depicted in Figure 1. We now survey background material briefly to explain this scenario, then discuss the methodological basis of our research: development of learning and inference components that take records of use cases and queries (from web server logs and forms) and produce decision support models for the CF performance element.

2.1 Gene Expression Modeling using Microarrays

As a motivating example of a computational genomics experiments, we use gene expression modeling from microarray data. DNA hybridization *microarrays*, also referred to as *gene chips*, are experimental tools in the life sciences that make it possible to model interrelationships among genes, which encode instructions for production of proteins including the *transcription factors* of other genes.

Microarrays simultaneously measure the expression level of thousands of genes to provide a "snapshot" of protein production processes in the cell [17]. Computational biologists use them in order to compare snapshots taken from organisms under a control condition and an alternative (e.g., *pathogenic*) condition. A microarray is typically a glass or plastic slide, upon which DNA molecules are attached at up to tens of thousands of fixed locations, or *spots*. Microarray data (and source code for programs that operate upon them) proliferate rapidly [49] due to recent availability of chip makers and scanners.

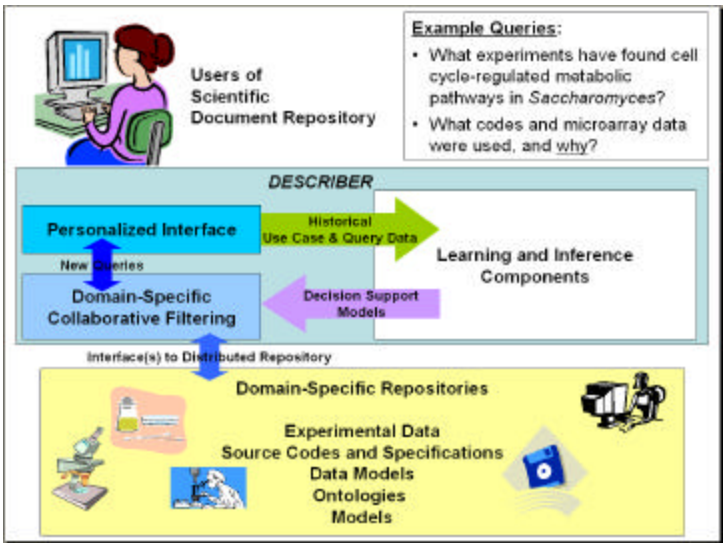

Figure 1. Overview of a typical collaborative filtering scenario using a scientific research index.

A major challenge in bioinformatics [17] is to discover gene/protein interactions and key features of a cellular system by analyzing these snapshots. [38] Our recent NSF EPSCoR First Award project [50] focuses on the problem of **automatically extracting gene regulatory dependencies from microarray data**, with the ultimate goal of building simulation models of an organism [119] under external conditions such as temperature, cell cycle timing (in the yeast cell), photoperiod (in plants), etc. Genomes of model organisms, such as *S. cerevisiae* (yeast) [26], *A. thaliana* (mouse ear cress or *weed*) [117, 118], *O. sativa* (rice) [123], *C. elegans* (nematode worm), and *D. melanogaster* (fruit fly), have been fully sequenced. These have also been annotated with the *promoter* regions that contain binding sites of *transcription factors* that regulate gene expression. Public repositories of microarray data such as the ***S**accaromyces* **G**enome **D**atabase (**SGD**) [26] for yeast have been used to develop a comprehensive catalog of genes that meet analytical criteria for certain characteristics of interest, such as *cell cycle regulation* in yeast. [113] We are using SGD data and a synthesis of existing and new algorithms for learning Bayesian networks from data to build robust models of regulatory relationships among genes from this catalog [49, 50].

2.2 Collaborative Filtering

Most data resources we plan to use in developing *DESCRIBER* are in the public domain [16, 26, 117, 118], while some are available under local and international research agreements (and based upon cooperative efforts with the labs of Leach *et al.*, Roe, Quackenbush, and Goble).

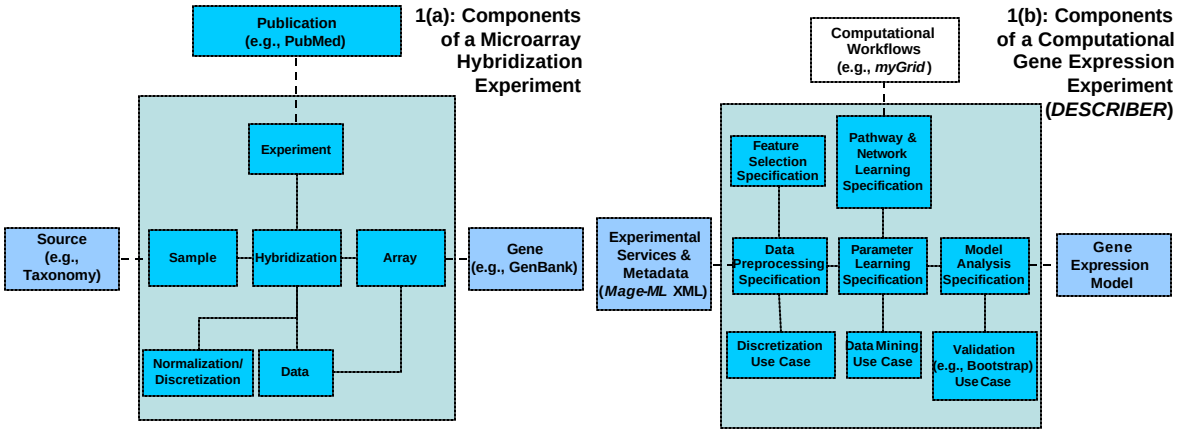

Figure 2. Schematics for (a) MGED [83, 85, 86] and (b) *DESCRIBER* [49] data models

The MGED group defines a microarray *experiment* [86] as "a set of one or more *hybridizations* [of labeled, single-stranded DNAs, synthesized from mRNA extracts of test organisms, with their complementary sequences on a microarray], each of which relates one or more *samples* to one or more *arrays*. The hybridized array is then scanned and the resulting image *analyzed*, relating each element on the array with a set of measurements. The data are *normalized* and combined with data from replicate hybridizations." Figure 2 (adapted from [86] and [83]) shows two schematic representations: on the left, MGED's block diagram describing a **data acquisition** experiment; on the right, our analogous diagram for **computational gene expression modeling from this data** [49, 53]. The diagrams sketch out relations among the experimental entities; MGED first specified the ones on the left in the ***Minimum Information in a Microarray Experiment (MIAME)*** data standard [86]. *MIAME* was implemented by the markup language *MAML* [85], which was unified with the *Gene Expression Markup Language (GEML)* [106] in November, 2001 to form *MAGE-ML* [83].

We propose to develop such data models by **building upon *MAGE-ML* and using ontology construction tools** such as the Ontology Inference Layer (*OIL*) [12, 48] to provide personalized **responses** to queries of the form shown in Figure 1 [122]. One ontology development tool designed for this purpose is *TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources)* [8].

3 Research Activities, July 2001 – July 2002

3.1 Learning gene expression networks from microarray data

We recently received an NSF EPSCoR First Award [50] (June, 2002 – August, 2003) to build probabilistic network models of cell cycle-regulated genes in yeast.

3.2 *DESCRIBER*: Intelligent Collaborative Filtering in Bioinformatics Repositories

In a 1998 invited talk at several co-located intelligent systems research conferences in Madison, Wisconsin, Kohavi described a technology adoption chasm between leading research in academic machine learning and development of commercial data mining products [73]. Our work seeks to bridge a similar gap [99]: between producers of repositories containing resources for computational biology and users who apply these resources to develop computational experiments. General-purpose technologies for intelligent search and document categorization have filled this gap only to a limited degree, leaving the

difficult challenge of matching user requirements to data, programs, and models in specialized domain areas.

PROBLEM STATEMENT: We consider the problem of *collaborative filtering* (CF) – analyzing the content of an information retrieval (IR) system and databases about actions of its users, to infer users' preferences and thereby predict topics or products that may be useful to them. To aid design and implementation of *gene expression modeling experiments*, we propose an integrated CF system for ***computational genomics***, based upon the **functional content** of repositories for bioinformatics research and development.

AIMS AND SCOPE: The focus of this integrative research and education program is on formalizing collaborative filtering environments and representing them using graphical models of probability such as Bayesian networks and relational and object-oriented extensions thereof [9, 68, 76]. The motivating application is to facilitate design of computational experiments in bioinformatics (especially genome analysis) and reuse [41, 77] in the development of software assemblies [33] that implement these experiments. Our central hypothesis is that probabilistic representation, learning, and reasoning are effective tools for developing the intelligent retrieval and indexing components of such a CF system. The primary contribution of this research shall be the novel combination of statistical algorithms for learning the structure of graphical models from data with existing techniques for constructing relational models of gene experiments, data, and programs. The desired outcome of applying our new technology is the first autonomous CF system for personalization of research in computational genomics. The technical objectives center around development of semi-structured data models [1] for collaborative filtering, algorithms for learning and reasoning using Bayesian networks, and statistical experiments to evaluate this approach.

Our research plan consists of three sections corresponding to aspects of our system design:

- ***DESCRIBER***: overall collaborative filtering system, built from our components
- **Learning structure from data**: automated use-case modeling, ontology-building, and indexing
- **The bioinformatics semantic web**: data models based upon *MAGE-ML* [83] and *TAMBIS* [8]

As documented by Karlin [70], many search engines use simple statistical models in order to rank pages. Collaborative filtering (CF) [45] is a *decision support* technique [101] that augments this type of modeling with reasoning about user goals, plans, actions, and preferences [108]. A simple application of collaborative filtering is to use previously-recorded associations [2, 102] among search targets to provide "related" content of plausible interest to the user. This approach is used in *market basket analysis* [19] – *Amazon.com* being one of the most well-known examples [109] – as well as cross-indexing systems such as that of *ResearchIndex* [80]. This *associational knowledge* can be compiled and used to make inferences in an efficient and scalable manner in IR systems [3, 115], but because of their representational simplicity, systems that use it have limited ability to generate explanations [97, 111].

Studying interactions among genes through the proteins that they can instruct a cell to produce gives biologists insight into many important cellular mechanisms. *DNA microarrays* or *gene chips* are data acquisition tools that can be used to build mathematical models of these interactions. [17] We use this as a motivating example and defer a technical synopsis of microarray technology to the research plan.

Suppose a developmental biologist working on functional mapping of the genome of the model plant *Arabidopsis thaliana* [119] has collected scans of DNA microarrays from her own *hybridization* experiments, and submits a query for **gene expression modeling** tools. The codes, annotated experiments, and resultant postprocessed data and models that apply to this data are diverse [5, 27, 87, 117, 118]. They may cover the following tasks, among others:

- **Image postprocessing** [6, 7]: chip alignment, quality assurance, feature extraction, image-based inspection, data normalization

? **Data analysis**: (a) single-gene profiling, clustering [11, 13, 20, 113]; (b) modeling of gene pathways and networks [15, 38, 40]

The above tasks fall into roughly sequential order, yet the identification of codes for performing a concrete instance of each task is still a challenging problem. The user must choose from a broad palette of tools and may waste considerable effort in integrating component codes into a usable software assembly [33] for her complete computational experiment. Even if these tools are organized in a proper taxonomy [8, 12] and have proper type annotations [105, 107], it is difficult to predict and explain in advance how they will function as a whole, and whether they fit the overall requirement. These obstacles to software reuse [41, 77] are clearly not specific to bioinformatics; in this proposal, however, we describe existing online resources for bioinformatics research that exemplify a broader class of CF problems: personalization and workflow development [44] over scientific code and data repositories.

Graphical models of probability [69, 89, 96] for machine learning and reasoning under uncertainty have been applied to automatically classify and build data models for searchable repositories [43, 82, 112]. The collection of user data from such digital libraries provides a test bed for the underlying intelligent CF technology. This proposal documents some functional relationships among the different resource types in a bioinformatics repository, as well as data and metadata standards [8, 14, 84], indexing methodology, and evaluation measures. It then addresses the problem of formulating tractable and efficient problem specifications for IR in bioinformatics, and describes a CF framework that applies existing and new algorithms for probabilistic learning and inference to these problems.

In summary, we hypothesize that developing constraint modeling and checking tools [122] for bioinformatics may reveal semantic gaps to be bridged by our work on extracting probabilistic models from CF data, leading to a general learning and inference architecture for *e-science* [103, 120].

3.2.1 Unifying objectives and strategy

The technology transfer chasm we referred to in section 3 and some of its technical causes have frequently been attested to in plenary and keynote addresses at conferences [42, 116]. In many computational science and engineering (CSE) domains, large collections of domain-specific tools (data, metadata, experimental documentation, and codes) are being built for use by researchers, but are underutilized due to deficiencies in categorizing and indexing the content of repositories and especially in modeling the structure of user requirements and the function of specialized software tools.

We now briefly review the foundational and supporting objectives of the project, and suggest concrete outcomes by which the success of the project can be measured. This is followed by a more detailed discussion of our strategy for realizing these objectives.

FOUNDATIONAL OBJECTIVE: The cornerstone of our project is the construction of the ***Data Entity and Source Code Repository Index for Bioinformatics Experimental Research (DESCRIBER)***. With this system, we can develop and evaluate new algorithms for probabilistic learning and inference in intelligent filtering and integrate them into tools for the bioinformatics Grid [35, 44, 120] and semantic web [12, 83]. This system shall serve as a foundation for experiments and refinement of our methodology.

SUPPORTING OBJECTIVES: Two other primary measures of success for this project, which we discuss in the rest of section 3, are the evaluation and validation of our *design approach* and *educational plan*.

Our plan includes five research tasks: (1) completion of our current NSF EPSCoR project [50] on graphical probabilistic models of gene expression, which meshes well with *DESCRIBER*; (2) continuing development of *BNJ* and *MLJ*; and development and application of our (3) XML data standard and use case modeling language, (4) intelligent CF system, and (5) ontology-building tool

3.2.2 Goal and rationale

The primary goal of our research, teaching, and mentoring over the next five years is to advance probabilistic learning and reasoning using graphical models as a fundamental methodology for building integrative repositories for computational science research and as a basis for developing models of data, experiments, software, and users to support intelligent filtering therein.

PLAN OVERVIEW: We plan to meet this goal by making significant contributions to:

Bioinformatics research: The formalization of collaborative filtering systems demands an understanding of relevant inference, pattern recognition, and machine learning problems and probabilistic representation [75] to support solutions to these problems. In our application domain of computational genomics, we will need to formalize data models for specification of (1) *molecular biology experiments* such as hybridizations using DNA microarrays, (2) *computational analyses* such as gene clustering and expression network discovery, and (3) cross-language application programmer interfaces for genome analysis software. One illustrative application of this data model is intelligent reasoning to map experimental data from microarray hybridization experiments (item 1 above) to the appropriate programs (3 above) for operating upon it, given a specification (2 above) for the desired computational analysis. We describe a system design that supports this intelligent reasoning and automated refinement of the data model, using a probabilistic representation.

Development of intelligent systems tools: In May, 2002, we released the first versions of two open source software packages [59, 62] for machine learning and probabilistic reasoning that we developed to serve in our research projects and courses in knowledge discovery in databases (KDD). These packages, called *Machine Learning in Java (MLJ)* [63] and *Bayesian Network Tools in Java (BNJ)* [60], currently contain a number of previously published algorithms as well as our own, to which we shall add via our own development, student projects, and collaborations with the research community [90].

Pedagogy: As practitioners of computational biology and bioinformatics address deeper challenges such as whole-genome alignment and expression modeling, there is an increasing synergy among the areas of data mining [124], software engineering, and high-performance distributed and parallel computing. The goal of our educational effort for the next five years is to develop a basic curriculum in **pattern recognition, intelligent systems, and data mining (PRISM)** methodologies [22] for problems in computational science and engineering, with our contribution focused on bioinformatics and collaborative filtering. In the long term, we would like to develop a new course on intelligent IR where key concepts such as clickstream analysis, document categorization, user modeling, and decision support are explained in a systematic way using graphical probabilistic models [69, 96] and information theory [29, 81].

RATIONALE: Recent systems such as *ResearchIndex* / *CiteSeer* [80] have succeeded in providing cross-indexing and search features for specialized but comprehensive **citation** indices of full documents. The indexing technologies used by such systems, as well as the general-purpose algorithms such as *Google PageRank* [18] and *HITS* [71], have several advantages: They use a *simple conceptual model* of document webs. They require little specialized knowledge to use, but organize and present hits in a way that allows a knowledgeable user to select relevant hits and build a collection of interrelated documents quickly. They are extremely popular, encouraging users to submit sites to be archived and corrections to citations, annotations, links, and other content. Finally, some of their content can be automatically maintained.

Despite these benefits, systems such as *ResearchIndex* have limitations [98] that hinder their direct application to IR from bioinformatics repositories:

? **Over-generality:** Citation indices and comprehensive web search engines are designed for the generic purpose of retrieving all individual documents of interest, rather than collections of data sets, program source codes, models, and metadata that meet common thematic or functional specifications.

? **Over-selectivity:** Conversely, IR systems based on keyword or key phrase search may return fewer (or no) hits because they check titles, keywords, and tags rather than semi-structured content [1, 122].

? **Lack of explanatory detail:** A typical user of an integrated collaborative filtering system has a specific experimental objective, whose requirements he or she may understand to varying degree depending upon his or her level of expertise. The system needs to be able to **explain relationships** among data, source codes, and models in the context of a bioinformatics experiment.

How can we achieve the appropriate balance of generality and selectivity? How can we represent inferred relationships among data entities and programs, and explain them to the user? Our thesis is:

Probabilistic representation, learning, and reasoning [75] are appropriate tools for providing domain-specific collaborative filtering capability to users of a scientific computing repository, such as one containing bioinformatics data, metadata, experimental documentation, and source codes.

We therefore present a plan to develop *DESCRIBER*, a research index for consolidated repositories of computational genomics resources, along with machine learning and probabilistic reasoning algorithms to refine its data models and implement collaborative filtering. The unifying goal of the work is to advance the automated extraction of graphical models of use cases for computational science resources, to serve a user base of researchers and developers who work with genome data and models. We present recent results from our own work and related research that suggest how this can be achieved through a novel combination of probabilistic representation, algorithms, and high-performance data mining not previously applied to collaborative filtering in bioinformatics. Our project shall also directly advance gene expression modeling and intelligent, search-driven reuse in distributed software libraries.

3.2.3 Development schedule

Our technical objectives can be divided among model *construction*, *evaluation*, and *application* phases.

CONSTRUCTION: The first 18-month phase shall produce entity-relational data models and a standardized, documented training corpus for learning probabilistic models from our own use cases [49, 53] and the yeast genome database [26], *MAGE* [83], and the EBI [31] and *OpenBio* [95] consortia. Meanwhile, we will also develop use cases using data and documentation archived at the *Stanford Microarray Database (SMD)* [16] and *The Arabidopsis Information Resource (TAIR)* [117] and produced by our collaborators at NCSA [91], TIGR [118], and the *myGrid*-EBI project [31, 121]. In an overlapping 42-month phase (steps 3-5), we will construct models and ontologies for CF by applying our learning and inference codes [60, 63] to this data.

EVALUATION: We will develop algorithms for both *learning* and *inference* of graphical models and compare them to existing ones for structure learning [28, 47] and exact [79] and approximate inference [24]. This project focuses on extending structure learning to semi-structured relational models and the evaluation of robustness by statistical validation of discovered models. We propose to generalize and refine existing evaluation methods [38, 43, 112] by (*i*) validating more model features; (*ii*) checking automatically extracted models against published ontologies. Because bootstrap methods for model evaluation are computationally intensive, we will use high-performance Grid applications [34, 35] that we are already developing with NCSA [50] to accomplish the indicated experiments.

APPLICATION: To validate the resultant models, we will use simulated or real observations of queries and IR sessions with real document repositories. As this section explains, model application is expected to lead to a process of iterative improvement of models, by refinement and redesign.

3.2.4 System design of *DESCRIBER*

Figure 3 depicts our design for *DESCRIBER*. 4(a), shown on the left, is the block diagram for the overall system, while 4(b) elaborates Module 1 as shown in the lower left hand corner of 4(a). Our current and continuing research focuses on algorithms that perform the learning, validation, and change of representation (inductive bias) denoted by Modules 2 and 4. We choose probabilistic relational models as a representation because they can express constraints (cf. Figure 1) and capture uncertainty about relations and entities [43] such as those elaborated from Figure 2 [49]. We hypothesize that this will provide more flexible generalization over use cases. We have recently developed a system for Bayesian network structure learning [58] that improves upon the *K2* [28] and *Sparse Candidate* [39] algorithms by using combinatorial optimization (by a genetic algorithm) to find good topological orderings of variables. Similar optimization wrappers have been used to adapt problem representation [25, 46, 65, 74, 100] in supervised inductive learning for classification, using decision trees and instance-based learning. We refer the interested reader to [64] and [49] for details of our system.

Other relevant work includes *BioIR* [94], a digital library for bioinformatics and medical informatics whose content is much broader than that of our work on genome analysis. *BioIR* emphasizes phrase browsing [93] and cross-indexing of text and data repositories rather than experimental metadata and source codes. Other systems such as *CANIS* [23], *SPIDER* [125], and *OBIWAN* [126] also address intelligent search and IR from bioinformatics digital libraries, emphasizing categorization of text documents. We view the technologies in these systems as complementary and orthogonal to our work because of this chief difference.

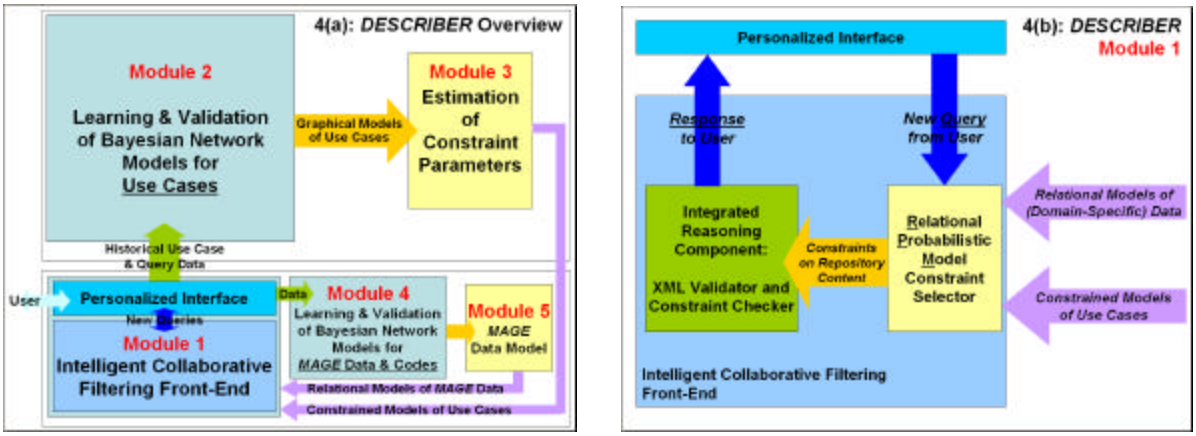

Figure 3. Design of (a) *DESCRIBER* and (b) its collaborative filtering front end.

We now continue with specific technical requirements generated by this goal, such as evaluating and developing computational techniques for **extracting novel, useful, and comprehensible information** [32] about microarray experiment records, data, and source codes from consolidated repositories.

3.2.5 The bioinformatics semantic web

We propose to develop an data model for intelligent filtering of bioinformatics repositories. Although we refer to *DESCRIBER* as a collaborative recommender system, strictly speaking, it also includes *structural recommendation* to model relations among documents [104]. As Figure 2(b) shows, we intend to use this data model to develop **ontologies for *computational* experiments using microarray data**. Figure 3(b) illustrates how *DESCRIBER* will use this to extract constraints that can be checked against semi-structured data. Producing a model to improve retrieval of source codes, experimental records, and data sets from our repository requires additional parameters to be learned or inferred from historical records. We will use traditional methods for Bayesian parameter estimation as well as new stochastic approximation algorithms that we are developing. The broad objectives of this line of research are to increase the robustness of inductive machine learning algorithms and develop learning systems that can

be automatically tuned to meet the requirements of a KDD performance element. In the past 30 months, our research group has produced *BNJ* [59, 60], a comprehensive Java-based software library for Bayesian network learning and inference, and *MLJ* [62, 63], an extendible Java port of the *MLC++* class library [72]. As *BNJ* development continues, we shall focus our research on issues of probabilistic representation, learning, and inference in decision support. One limitation of keyword-based search affects semistructured data models as well: they lack the *representation* of semantic structure to properly model the user's intentions and the relationships among data sets, program source codes, models, and metadata, which in bioinformatics research tends to greatly cut down the set of relevant documents.

EXPLANATORY CAPABILITY IN INTELLIGENT INFORMATION RETRIEVAL: A key benefit and anticipated new contribution of our approach is that it will allow *DESCRIBER* to produce explanations of why retrieved experimental documents and source codes match the user (e.g., use case) query. Pearl discusses the use of graphical models to *explain causal relationships* [97]. We will adapt existing machine learning techniques for personalization [82, 112], especially using relational probabilistic models [50], to produce these explanatory structures. During this five-year project, we will also develop new algorithms based upon probabilistic extensions of the relational and constraint models required to implement Figure 3(b).

3.2.6 Evaluation and validation of design approach

DESCRIBER serves a single goal: to benefit the bioinformatician in design and implementation of computational experiments. We will evaluate our system using two independent criteria: gain from collaborative filtering (CF) and experimental solution quality.

GAIN FROM COLLABORATIVE FILTERING: We plan to evaluate software reuse over repositories of bioinformatics data, metadata, and source code (effectively, distributed component libraries). The software engineering literature provides metrics for reuse and analytical and empirical evaluation methodology [10, 30] that we shall apply. This evaluation shall be carried out by experimenters, or "end users" of the CF system, in both our own group and our collaborators'. **Our hypothesis** that graphical relational models are the appropriate representation and probabilistic reasoning the appropriate techniques for building the constraint models for CF **shall be tested by comparing the level of reuse** (percent of new code implemented and development time) without intelligent CF, with the current *myGrid* CF technology, and with the *DESCRIBER* system. If our new algorithms for Bayesian network structure learning are also shown to be competitive (in robustness, inferential accuracy, and especially *utility gain* in CF) with current algorithms such as *TETRAD* [114] and the *Sparse Candidate* algorithm [39], they shall provide new tools for robust, automated extraction of data models in e-science [120].

SOLUTION QUALITY ASSESSMENT IN GENE EXPRESSION MODELING: Focusing on gene expression modeling, we will use both our synthetic expression data and biological data collected by our collaborators on microarray analysis of *S. cerevisiae* (yeast), *A. thaliana* (weed), *O. sativa* (rice), and *D. melanogaster* (fly) to assess solution quality.

Our recent technical report [49] and First Award project proposal [50] document how we will assess the ability of our algorithms to re-extract network information by measuring their robustness using *bootstrap sampling* cf. Friedman *et al.* [38, 40]. Simply put, suppose existing microarray data is randomly sampled many times and used to build multiple probabilistic models, and a link, pathway or parents-child relationship among two or more genes is frequently found using these many samples. One evaluation criterion is the frequency of trending relations in gene expression pathways: e.g., with *time*, *temperature*, *phenology*, or *constitutive* (invariant with time and temperature). If this frequency is very high, it will be a significant result, **illustrating the representational power of Bayesian networks** and giving further positive indications for dynamic probabilistic network [88] modeling of gene dynamics. The latter will refute it while raising the interesting question as to how the principles of chemical kinetics are being evaded. A middle result will let subsequent studies focus on subsystems that do or do not exhibit cell cycle dependence. Independent of the success of network analyses, gene expression

experiments developed using our collaborative filtering tool will provide further entry points for microarray hybridization experiments in *S. cerevisiae*. **If they do result in statistically robust results**, they may be further used to **suggest confirmatory hybridization experiments** for biologists. This is a significant benefit because of the **high cost of data acquisition**. Microarrays are still expensive at present, especially when used on the scale of *gene dynamics modeling* such as with temperature, metabolism. We further anticipate that this approach shall generalize to *A. thaliana* [119] and other organisms studied by our collaborators and pathways such as photoperiod-regulated ones in plants.

Our **preliminary experiments** [49, 53, 58] using *Asia*, *ALARM* [28], and the 250-gene data set studied by Friedman *et al.* [38] indicate that our system *BNJ* [60] should be able to **scale up to larger gene networks**. Several of our cooperative efforts are planned with plant biologists and therefore focus on simulation models of plant development [119] that produce time series of developmental state variables at daily or hourly granularity. These can be used as forcing functions to validate the expression models we will produce. In ***DESCRIBER***, we shall evaluate probabilistic representations in comparison with others such as non-probabilistic relational models developed from use case analysis [21] and entity-relational data modeling [110]. Our work on KDD [32] frequently compares inferential accuracy (or classification accuracy) in Bayesian networks to that of other representations. This project focuses on structure learning and evaluation of robustness by statistical validation of discovered models.

Finally, we will avail our project of **alternative sources of CF use cases and data** from proteomics and beyond bioinformatics (see the attached letter from Zollman and Bennett) both as a **safeguard against unexpected obstacles** in development of *DESCRIBER* and to **increase synergistic output**.

4 Teaching and Outreach Activities, July 2001 – July 2002

4.1 Outreach and Technology Transfer

4.1.1 Training experiences for students

HIGH SCHOOL: Our integrative research and education program includes planned demonstrations, workshops, and web learning materials for 8-12th grade outreach. We aim towards developing workshop activities for our university's science and technology program for teen women, and we have administered the summer science institute for high school juniors and seniors in our department for the past 2 years. The latter has about 40% female participation, significantly higher than the admissions or retention rates in our undergraduate program. We have noted a high level of interest among women in such summer programs who are our prospective majors in our programs or in the computational life sciences.

EARLY UNDERGRADUATE: We introduce undergraduates – as early as their second year – to AI, machine learning, simulation, and visualization algorithms that they help implement and use in experiments. Furthermore, we believe that early undergraduate research experiences are a key to retention of female students and other underrepresented groups of students in engineering, giving them exposure to both theoretical background and commercial and industrial applications. Our efforts are supported by collaboration with our university's Women in Engineering and Science Program. As a research scientist at NCSA, we were able to hire 40-60% women research programmers at the 11-12 through undergraduate level. To bring participation of women closer to this level in our programs and encourage retention through graduate study, we have regularly lectured at our undergraduate engineering honors and computers in society (ethics and applications) seminars. As departmental undergraduate honors chair, we have supervised numerous projects [51] and organized student seminars. We are also planning integrative research experiences for undergraduates, including but not limited to honors students.

ADVANCED UNDERGRADUATE AND POSTGRADUATE: We propose to leverage our existing research by incorporating our basic theoretical advances in machine learning and probabilistic reasoning into our intelligent systems courses, demonstrating their benefit to students through hands-on development experience with the software packages we have co-developed: *Machine Learning in Java (MLJ)* [63],

Bayesian Network Tools in Java (BNJ) [60], and *NCSA D2K* [91]. This leads to concurrent engineering of our research codes with the educational code base, offering undergraduates and new graduate students the chance to learn about state-of-the-field software tools from the original developers. We have found that providing visual explanation of technologies to students facilitates the process of active learning by allowing them to interact with models, data, or algorithms. We shall provide students with visual programming infrastructures [36, 91] for development of distributed, high-performance KDD and collaborative filtering systems and require them to develop user interfaces and visualizations of models. We expect benefits of this visualization approach to accrue to new interdisciplinary teaching programs in problem solving in the colleges of engineering, arts and sciences, agriculture, and architecture.

All but one of the upper-division courses offered in our KDD certificate program are open to undergraduates. The data mining practicum has been used for three years (2000-2002) as the partial basis of an NSF research experience for undergraduates (REU) at the National Center for Supercomputing Applications (NCSA), in which we participated as senior personnel.

4.1.2 Technology transfer and dissemination

The *DESCRIBER* project shall advance the capability to efficiently and flexibly retrieve data, documents, software tools, and models from gene research repositories. We discuss selected related work on bioinformatics data models and, in section **Error! Reference source not found.**, document a strategy to produce and evaluate new probabilistic reasoning and data modeling software tools to meet the four milestones listed in section 3. We shall then present our detailed plans for the integration of bioinformatics research results and intelligent systems tools into our **curriculum on applied intelligent systems**, and their dissemination to the user community.

DISSEMINATION: Dissemination of results shall be achieved by development and distribution of open-source software tools (augmenting our Java-based toolkits *MLJ* [63] and *BNJ* [60]), courseware, and models in an open, semi-structured data format (the XML Bayesian Network Interchange Format). The software modules shall be added to *Data to Knowledge (D2K)*, a suite of general-purpose, experimental codes managed by the National Center for Supercomputing Applications (NCSA).

SOFTWARE RELEASES: Our key dissemination effort is the development of **research codes** applicable to real-world learning and inference problems, including collaborative filtering and the specific bioinformatics applications we have discussed. This began over four years ago with the first ports of *SGI MLC++* to other platforms [57] and continued with our work at NCSA applying this port in research on commercial data mining [65]. Over 300 software developers employed by our industrial partners in NCSA's private sector program have been trained to use *D2K* [69] applications developed by the NCSA Automated Learning Group in conjunction with our group. *BNJ* has been downloaded over 300 times from *SourceForge* since 04 May 2002 [59] and has 50 registered users worldwide at the time of this writing [60], and *MLJ* has been downloaded 100 times [62] and has 30 registered users to date [63]. As **Error! Reference source not found.** illustrates, we anticipate that they shall be refined the next 2-5 years through interaction with our local and international collaborators and instructors of KDD-related courses at this and other universities.

LOCAL AND INTERNATIONAL MENTORING: In addition to funneling production-level research and development tools such as *BNJ* [60], *MLJ* [63], and *D2K* [69] back into the classroom [51], we have devoted focused effort to mentoring of students with potential to conduct research and become teachers in our subject area. This began in 1998 with our supervision of graduate research assistants and undergraduate programmers at NCSA and continued with our participation (1999-2000) in the Engineering Learning Enhancement Action/Resource Network (LEA/RN) [67]. Since 2001 we have served as our department's undergraduate honors advisor, organized spring seminars and summer workshops for this program, mentored a student who received a Goldwater scholarship for her contributions to *BNJ* [60], and begun serving as faculty advisor on a 2-student project in the Computing Research Association Collaborative Research Experience for Women (CREW) program (2002-2003)

whose proposal is currently under review. We will involve postgraduate and postdoctoral mentoring into synergistic activities such as interdepartmental seminars.

4.1.3 Outcomes and benefits of our education plan

Incipient demand for researchers and developers obliges both academia and (medical, pharmacological, and other biotechnological) industry to train bioinformaticians. To help meet this requirement, we have developed new courses and a formal interdisciplinary curriculum at both the graduate and undergraduate levels and integrated them into a research program that emphasizes rigorous specification, development, and assessment of KDD systems. We are delivering these to both traditional campus-based students and remote students, many of whom are in the information technology workforce. We are also working to develop programs for outreach at pre-collegiate levels. Our desired pedagogical benefits, concrete outcomes, and approaches are:

- **Curriculum improvement**: courses, degree programs, materials to facilitate active learning
- **Early educational outreach**: improved recruitment and retention of underrepresented groups
- **Undergraduate involvement in research**: collaborative experience**;** technology transfer
- **Increased competence**: mentoring from pre-collegiate through postdoctoral level

4.1.4 Curriculum development: courses and degree programs

Many curricula include a broad survey course in Artificial Intelligence (AI). This introduction is a valuable experience for students. Typically, it covers topics such as intelligent agent frameworks, problem solving, search, knowledge representation, logical semantics and inferential calculi, resolution theorem proving, basics of planning, probabilistic semantics and inferential techniques, machine learning, and assorted topics selected from among artificial neural networks, machine vision, robotics, and natural language processing. [108] Our experience with a mixed graduate-undergraduate sequence, starting with *Introduction to Artificial Intelligence* [51], and culminating in *Advanced Topics in Artificial Intelligence* [51], indicates that these alone do not provide enough time to cover machine learning in depth while retaining fundamental topics in planning and reasoning.

We therefore developed two new specialized courses that augment this core AI curriculum: a full-semester course that focuses on machine learning and pattern recognition and an intensive practicum (4-5 weeks) in implementation of high-performance data mining systems. We first taught the course on machine learning in Fall, 1999 and converted it from a special topics course into an annual course in Fall, 2001 [51]. We first taught the data mining applications course in Summer, 2000 and repeated it in 2001 and 2002. It consists of an integrated lecture and software engineering lab satisfying an implementation course requirement for our Master of Software Engineering (MSE) program.

Each student who completes the software portfolio requirement for our MSE degree must present three phases of work: first, methodology, requirements analysis, and design; second, formal specification, software quality assurance and testing plan, cost analysis, and third, an implementation synopsis, performance evaluation, and summary of experimental findings. Projects in the area of KDD fit this framework because each studies a methodology for information and data modeling, performs requirements analysis for the end use of the intelligent system, and produces a design for a KDD system. Specification of KDD algorithms and representations is crucial to our research, while the final phase demands careful evaluation of experimental results. We thus achieve a new synthesis of software engineering with intelligent systems, tailored to development of robust, practical, and general KDD.

In the machine learning, data mining, and advanced AI topics courses, students learn about new theoretical advances and state-of-the-practice implementation and empirical evaluation techniques. We survey new literature in the field and students are required to submit individual paper reviews on weekly reading, which we have found helps students develop their ability to conduct independent research of the literature, **critically assess** their own research, present experimental results more cogently, and prepare

publications. Meanwhile, we incorporate new programming tools that support KDD research [59, 60, 62, 63] into each offering of the courses. Students work on small individual programming projects in the machine learning course and team projects (in groups of 2-4) in the data mining and advanced AI courses. They are required to keep journals on implementation and experimental results and prepare a final 6-page project report in the style of a national workshop or conference (using the AAAI/IJCAI guidelines [4]). These reports are peer-reviewed and students are given the reviews at the end of the course.

Our new courses and new revisions of previously offered courses, are well integrated with our department's courses on theory of computation and software engineering, and with our university's curricula in mathematics, statistics, operations research and biologically-inspired and soft computing. Together, the five intelligent systems courses we have developed offer students a computational foundation for our integrative program of study in pattern recognition, intelligent systems, and data mining (PRISM) [22]. Degree certificates in PRISM shall be awarded to students in our MSE program and M.S., and Ph.D. programs university-wide who achieve proficiency relevant to KDD through electives in mathematics, statistics, and computational science and engineering, complete three of four core AI courses and at least one graduate-level database course, and achieve breadth in applications of learning and KDD systems. The latter includes graduate level courses in **computational molecular biology** and **computational genomics** that comprise our new bioinformatics program.

Proposals for an undergraduate minor in bioinformatics and a graduate certificate program in PRISM are presently in preparation for submission to our university curriculum committees.

4.1.5 Textbook development

Integrating and complementing our efforts toward dissemination and outreach is the planned development of a textbook on implementation of probabilistic learning and reasoning systems using technical computing languages. It shall provide simple, working code (available over the web) for well-known linear algebra and symbolic math tools such as *MATLAB* and *Mathematica*, statistical computing packages such as *SAS*, common analytical processing platforms such as *Microsoft Excel*, and functional programming languages such as *Standard ML of New Jersey*. Rather than providing a heterogeneous, disorganized mixture of implementations of learning algorithms, our textbook project undertakes to document these algorithms in the simplest and clearest way by illustrating common data structures and representations of hypotheses using many common technical computing tools. Our goal is to bring probabilistic and reasoning and the tools supporting our research to as wide an audience as possible. Therefore we use a complementary approach of developing open-source software [59, 62] both in general-purpose imperative languages such as Java and in technical computing languages. This trades off portability and interoperability (Java) against readability and accessibility (*MATLAB*, etc.) to bioinformaticians who may be new to both programming and intelligent systems.

4.1.6 Assessment

To assess new course content, materials, and curricula (degree program requirements), we shall work with our **O**ffice of **E**ducational **I**nnovation and **E**valuation (OEIE) and the independent **I**ndividual **D**evelopment and **E**ducation **A**ssessment (IDEA) Center [66]. We are preparing a separate research and curriculum development proposal [22] to support work with both organizations to evaluate pedagogical outcomes. We shall also consult with our collaborators at national and industrial labs to make sure our program includes the requisite knowledge for professional practice of bioinformatics and PRISM.

5 References

[1] Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web: From Relations to Semistructured Data and XML*. San Francisco, CA: Morgan Kaufmann Publishers.

[2] Adomavicius, G. & Tuzhilin, A. (2001). Using data mining methods to build customer profiles. *IEEE Computer, 34*(2):74-82.

[3] Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. In Buneman, P., & Jajodia, S., editors, *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, Washington, DC, p. 207-216. Chapel Hill, NC: ACM Press.

[4] American Association for Artificial Intelligence. (2002). AAAI Press Editorial and Author Information. Distributed from URL: http://www.aaai.org/Press/Editorial/editorial.html.

[5] *Arabidopsis* Functional Genomics Consortium. (2002). *AFGC Website.* URL: http://afgc.stanford.edu.

[6] Bajcsy, P., & Liu, L. (2002). An image-based visualization of microarray features and classification results. In *Proceedings of the 10th Conference on Intelligent Systems for Molecular Biology (ISMB-2002)*, Edmonton, Alberta, CANADA (in press). Menlo Park, CA: AAAI Press.

[7] Bajcsy, P., Liu, Z., & Liu, L. (2002). Quality assurance methods for processing microarray imagery. In *Proceedings of the 10th Conference on Intelligent Systems for Molecular Biology (ISMB-2002)*, Edmonton, Alberta, CANADA (in press). Menlo Park, CA: AAAI Press.

[8] Baker, P. G., Brass, A., Bechhofer, S., Goble, C., Paton, N., & Stevens, R. (1998). TAMBIS: transparent access to multiple bioinformatics information sources. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB-98)*, Montreal, CANADA (p. 25-34). Menlo Park, CA: AAAI Press.

[9] Bangsø, O. & Wuillemin, P. (2000). Top-down construction and repetitive structures representation in Bayesian networks. In *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2000)*, Orlando, FL. Menlo Park, CA: AAAI Press.

[10] Banker, R. D., Kauffman, R. J., & Zweig, D. (1993). Repository Evaluation of Software Reuse. *IEEE Transactions on Software Engineering, 19*(4):379-389.

[11] Barash, Y. & Friedman, N. (2001). Context-specific Bayesian clustering for gene expression data, In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology (RECOMB-2001)*, Montréal, CANADA. Chapel Hill, NC: ACM Press.

[12] Bechhofer, S., Horrocks, I., Goble, C., & Stevens, R. (2001). *OilEd*: A reason-able ontology editor for the semantic web. In *Lecture Notes in Computer Science 2174:396-405*. Berlin: Springer-Verlag.

[13] Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology 6*(3-4):281-297.

[14] BioOntologies Consortium. (2002). *Portal to the World of Life Sciences.* URL: http://www.bioontology.org.

[15] Blazquez, M. A. (2000). Flower development pathways. *Journal of Cell Science, 113*:3547-3548.

[16] Botstein, D., Sherlock, G., Binkley, G., Brown, P. O. & Ball, C. (2002). Stanford Microarray Database. Stanford, CA: Stanford University. Distributed from URL: http://genome-www5.stanford.edu/MicroArray/SMD/.

[17] Brazma, A., Parkinson, H., Schlitt, T., & Shojatalab, M. (2001). *A quick introduction to elements of biology – cells, molecules, genes, functional genomics, microarrays.* Web tutorial, European Bioinformatics Institute. Distributed from URL: http://www.ebi.ac.uk/microarray/biology_intro.htm.

[18] Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1-7):107-117.

[19] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, Tucson, AZ, USA, p. 255-264. Chapel Hill, NC: ACM Press.

[20] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. Jr., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines, In *Proceedings of the National Academy of Science, 97*:262-267.

[21] Carlson, D. (2001). Modeling XML Applications with UML: Practical e-Business Applications. Reading, MA: Addison Wesley.

[22] Chang, S.-I., Das, S., and Hsu, W. H. (2002). *Graduate Degree Programs in Pattern Recognition, Intelligent Systems, and Data Mining (PRISM).* A proposal to the Kansas State University Graduate College. Distributed from URL: http://www.kddresearch.org/Groups/Data-Mining/PRISM/.

[23] Chen, H. C., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A., & Lin, C. T. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval - the illinois digital library initiative project. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*:771-782.

[24] Cheng, J. & Druzdzel, M. J. (2000). AIS-BN: an adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research, 13*:155-188.

[25] Cherkauer, K. J. & Shavlik, J. W. (1996). Growing simpler decision trees to facilitate knowledge discovery. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, p. 315-318. San Mateo, CA: AAAI Press.

[26] Cherry, J. M., Ball, C., Dolinski, K., Dwight, S., Harris, M., Matese, J. C., Sherlock, G., Binkley, G., Jin, H., Weng, S., & Botstein, D. (2002). *Saccharomyces Genome Database*. Stanford, CA: Stanford University. Distributed from URL: http://genome-www.stanford.edu/Saccharomyces/.

[27] Cold Spring Harbor Laboratory. (2002). *Arabidopsis genome analysis.* Distributed from URL: http://nucleus.cshl.org/protarab/.

[28] Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4):309-347.

[29] Cover, T. M. & Thomas. J. A. (1991). *Elements of Information Theory.* New York, NY: John Wiley and Sons.

[30] Devanbu, P., Karstu, S., Melo, W. L., & Thomas, W. (1996). Analytical and empirical evaluation of software reuse metrics. In *Proceedings of the 18th International Conference on Software Engineering*, Berlin, Germany. IEEE Press.

[31] European Bioinformatics Institute. (2002). *EBI home page*. URL: http://www.ebi.ac.uk. [76]

[32] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3):37-54.

[33] Firesmith, D. G. (1993). *Object-Oriented Requirements Analysis and Logical Design: A Software Engineering Approach.* New York, NY: John Wiley and Sons.

[34] Foster, I. (2001). *Access Grid*. Argonne National Laboratory. URL: http://www-fp.mcs.anl.gov/fl/accessgrid/.

[35] Foster, I. & Kesselman, C., editors. (1999). *The Grid: Blueprint for a New Computing Infrastructure*. San Mateo, CA: Morgan Kaufmann.

[36] Frank, E., Hall, M., Trigg, L., Kirkby, R., Schmidberger, G., Ware, M., Xu, X., Bouckaert, R., Wang, Y., Inglis, S., & Witten, I. H. (2002). Waikato Environment for Knowledge Analysis (WEKA) v.3: Machine Learning Software in Java. Distributed from URL: http://www.cs.waikato.ac.nz/~ml/weka/.

[37] Friedman, N., & Goldszmidt, M. (1998). *Learning Bayesian Networks From Data*. Tutorial, American National Conference on Artificial Intelligence (AAAI-98), Madison, WI. San Mateo, CA: AAAI Press.

[38] Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB-2000)*, Tokyo, JAPAN. Chapel Hill, NC: ACM Press.

[39] Friedman, N., Nachman, I., & Pe'er, D, Learning Bayesian network structures from massive datasets: the sparse candidate algorithm, In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, p. 206-215. San Mateo, CA: Morgan Kaufmann.

[40] Friedman, N., Nachman, I., & Pe'er, D. (2002). *Using Bayesian Networks to Analyze Gene Expression Data*. Hebrew University of Jerusalem, ISRAEL. Distributed from URL: http://www.cs.huji.ac.il/labs/compbio/expression/.

[41] Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Reading, MA: Addison-Wesley.

[42] Gates, W. H. III. Keynote Address, *International Joint Conference on Artificial Intelligence (IJCAI-01)*, Seattle, WA, 07 Aug 2001. URL: http://www.microsoft.com/billgates/speeches/2001/08-07aiconference.asp.

[43] Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2001). Learning probabilistic models of relational structure. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, p. 170-177. San Francisco, CA: Morgan Kaufmann.

[44] Goble, C. (2002). myGrid: Personalised e-Science on the Grid. Presentation, *Official Opening of the National e-Science Centre*, University of Glasgow, SCOTLAND, 25 April 2002. Distributed from URL: http://umbriel.dcs.gla.ac.uk/nesc/general/esi/events/opening/lect_material.html.

[45] Goldberg, D., Nichols, D., Oki, B. & Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, *35*(12):61-70.

[46] Guerra-Salcedo, C., & Whitley, D. (1999). Genetic approach to feature selection for ensemble creation. In *Proceedings of the Genetic and Evolutionary Computation Conference(GECCO-99).* San Mateo, CA: Morgan Kaufmann.

[47] Heckerman, D. A. (1996). *A Tutorial on Learning With Bayesian Networks.* Microsoft Research Technical Report 95-06, Revised June 1996.

[48] Horrocks, I., Fensel, D., Broekstra, J., Decker, S., Erdmann, M., Goble, C., van Harmelen, F., Klein, M., Staab, S., Studer, R., & Motta, E. (2000). *OIL: The Ontology Inference Layer.* Technical Report IR-479, Vrije Universiteit Amsterdam, Faculty of Sciences.

[49] Hsu, W. H. (2002). *Activities of the bioinformatics and medical informatics working group.* Technical Report, Kansas State University Department of Computing and Information Sciences. Distributed from URL: http://www.kddresearch.org/Groups/Bioinformatics.

[50] Hsu, W. H. (2002). *Algorithms for discovery of Bayesian network models of gene regulation in Saccharomyces cerevisiae from microarray data*. NSF EPSCoR First Award proposal (funded 10 June 2002 – 09 August 2002). Distributed from URL: http://www.kddresearch.org/Groups/Bioinformatics/First-Award/.

[51] Hsu, W. H. (2002). *Laboratory for Knowledge Discovery in Databases home page*. Manhattan, KS: Kansas State University. URL: http://www.kddresearch.org.

[52] Hsu, W. H. & Chang, S.-I. (2002). Bayesian network structure learning. In *Proceedings of the 5th Military Applications Symposium*, Memphis, TN. Distributed from URL: http://www.kddresearch.org/Groups/Probabilistic-Reasoning.

[53] Hsu, W. H. & Guo, H. (2002). A relational probabilistic model for computational microarray experiments. In preparation.

[54] Hsu, W. H., & Gustafson, S. M. (2002). Genetic programming and multi-agent layered learning by reinforcements. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2002)*, New York, NY. San Francisco, CA: Morgan Kaufmann Publishers.

[55] Hsu, W. H., Auvil, L. S., Pottenger, W. M., Tcheng, D., & Welge, M. (1999). Self-organizing systems for knowledge discovery in databases. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-99)*, Washington, DC. IEEE Press.

[56] Hsu, W. H., Cheng, Y., Guo, H., & Gustafson, S. (2000). Genetic algorithms for reformulation of large-scale KDD problems with many irrelevant attributes. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, Las Vegas, NV. San Francisco, CA: Morgan Kaufmann Publishers.

[57] Hsu, W. H., Gettings, N. D., & Perry, M. (1999). *Linux port of MLC++*. Distributed from URL: http://www.kddresearch.org/Resources/.

[58] Hsu, W. H., Guo, H., Perry, B. B., & Stilson, J. A. (2002). A permutation genetic algorithm for variable ordering in learning Bayesian networks from data. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2002)*, New York, NY. San Francisco, CA: Morgan Kaufmann Publishers.

[59] Hsu, W. H., Guo, H., Perry, B. B., & Thornton, J. A. (2002). *Bayesian Network Tools in Java (BNJ) v1.0a*. Kansas State University Laboratory for Knowledge Discovery in Databases. Distributed from URL: http://sourceforge.net/projects/bndev.

[60] Hsu, W. H., Guo, H., Perry, B. B., & Thornton, J. A. (2002). *Bayesian Network Tools in Java (BNJ) Project Page*. Kansas State University Laboratory for Knowledge Discovery in Databases. Distributed from URL: http://www.kddresearch.org/Groups/Probabilistic-Reasoning/BNJ.

[61] Hsu, W. H., Kargupta, H., Liu, H., & Street, N., editors. (2001). *Working Notes of the Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases (ML-5),*

International Joint Conference on Artificial Intelligence (IJCAI-01). Seattle, WA, 04 August 2001. Distributed from URL: http://www.kddresearch.org/Workshops/IJCAI-2001.

[62] Hsu, W. H., Louis, J. A., & Plummer, J. W. (2002). *Machine Learning in Java (MLJ) v1.0a.* Kansas State University Laboratory for Knowledge Discovery in Databases. Distributed from URL: http://sourceforge.net/projects/mldev.

[63] Hsu, W. H., Louis, J. A., & Plummer, J. W. (2002). *Machine Learning in Java (MLJ) Project Page*. Kansas State University Laboratory for Knowledge Discovery in Databases. Distributed from URL: http://www.kddresearch.org/Groups/Machine-Learning/MLJ.

[64] Hsu, W. H., Ray, S. R., & Wilkins, D. C. (2000). A multistrategy approach to classifier learning from time series. *Machine Learning, 38*(1-2):213-236.

[65] Hsu, W. H., Welge, W., Redman, T., & Clutter, D. (2002). Constructive Induction Wrappers in High-Performance Commercial Data Mining and Decision Support Systems. *Data Mining and Knowledge Discovery*, to appear. Preprint URL: http://www.kddresearch.org/Publications/Journal/HWRC1.pdf.

[66] Individual Development Education Assessment (IDEA) Center. (2002). *IDEA Center home page.* URL: http://www.idea.ksu.edu.

[67] Iowa State University College of Education. (2002). *Project LEA/RN: Learning Enhancement and Action / Resource Network*. Distributed from URL: http://www.educ.iastate.edu/ess/learn.htm.

[68] Jaeger, M. (1997). Relational Bayesian networks. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97)*, Providence, RI (p. 266-273). San Mateo, CA: Morgan Kaufmann.

[69] Jordan, M. I., editor. (1998). *Learning in Graphical Models*. Cambridge, MA: MIT Press.

[70] Karlin, A. (2001). A biased view of web search algorithms. Invited talk, *17th Conference on Uncertainty in Artificial Intelligence (UAI-2001)*, Seattle, WA, 09 August 2001.

[71] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5):604-632.

[72] Kohavi, R. & Sommerfield, D. (1996). *MLC++: Machine Learning Library in C++, Utilities v2.0.* Mountain View, CA: Silicon Graphics, Incorported. Distributed from URL: http://www.sgi.com/Technology/mlc.

[73] Kohavi, R. (1998). Crossing the chasm: from academic machine learning to commercial data mining. Invited talk, *15th International Conference on Machine Learning (ICML-98)*, Madison, WI, 25 July 1998. Distributed from URL: http://robotics.stanford.edu/~ronnyk/chasm.pdf.

[74] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, Special Issue on Relevance, 97*(1-2):273-324, 1997.

[75] Koller, D. Representation, Reasoning, Learning. Computers and Thought Award Lecture, *17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, Seattle, WA, 07 Aug 2001. Distributed from URL: http://robotics.stanford.edu/~koller/CnT-web.htm.

[76] Koller, D. & Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97)*, Providence, RI (p. 302-313). San Mateo, CA: Morgan Kaufmann.

[77] Krueger, C. W. (1992). Software reuse. *ACM Computing Surveys, 24*(2):131-183.

[78] Larrañaga, P., Kuijpers, C. M. H., Murga, R. H., & Yurramendi, Y. (1996). Learning Bayesian networks structures by searching for the best ordering with genetic algorithms, *IEEE Transactions on Systems Man and Cybernetics, 26*(7)487-493.

[79] Lauritzen, S., & Spiegelhalter, D. J. (1988). Local Computations with Probabilities on Graphical Strutures and Their Application to Expert Systems. *Journal of the Royal Statistical Society*.

[80] Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6):67-71.

[81] MacKay, D. J. C. (2002). Information Theory, Pattern Recognition and Neural Networks. Course materials and textbook. Distributed from URL: http://www.inference.phy.cam.ac.uk/mackay/itprnn/.

[82] McCallum, A. K., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval, 3*(2):127-163.

[83] Microarray Gene Expression Data (MGED) Group. (2002). *MicroArray and Gene Expression Markup Language (MAGE-ML).* Distributed from URL: http://www.mged.org/Workgroups/MAGE/mage-ml.html.

[84] Microarray Gene Expression Data (MGED) Group. (2002). *Microarray Gene Expression Databases Group Home.* URL: http://www.mged.org.

[85] Microarray Gene Expression Data (MGED) Group. (2002). *MicroArray Markup Language.* The XML Cover Pages. Distributed from URL: http://www.oasis-open.org/cover/maml.html.

[86] Microarray Gene Expression Data (MGED) Group. (2002). *Minimum Information in A Microarray Experiment.* Distributed from URL: http://www.mged.org/Workgroups/MIAME/miame.html.

[87] Munich Information Center for Protein Sequences. (2002). *MIPS Arabidopsis thaliana database.* Distributed from URL: http://mips.gsf.de/proj/thal/db/.

[88] Murphy, K. & Mian, S. *Modelling Gene Expression Data using Dynamic Bayesian Networks.* Technical Report, Computer Science Division, University of California, Berkeley, 1999.

[89] Murphy, K. P. (2001). *A Brief Introduction to Graphical Models and Bayesian Networks.* Distributed from URL: http://www.cs.berkeley/~murphyk/Bayes/bayes.html.

[90] Murphy, K. P. (2002). Software Packages for Graphical Models / Bayesian Networks. Distributed from URL: http://www.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html.

[91] National Center for Supercomputing Applications Automated Learning Group. (2002). *About Data to Knowledge (D2K) v3.0*. Distributed from URL: http://archive.ncsa.uiuc.edu/STI/ALG/activities/.

[92] Neapolitan, R. E. (1990). *Probabilistic Reasoning in Expert Systems: Theory and Applications.* New York, NY: Wiley-Interscience.

[93] Nevill-Manning, C.G., Witten, I.H. & Paynter, G.W. (1997). Browsing in digital libraries: a phrase-based approach. In Allen, R.B., & Rasmussen, E., editors, *Proceedings of the 2nd ACM International Conference on Digital Libraries*, Philadelphia, PA, p. 230-236.

[94] Nevill-Manning, C.G., Witten, I.H. & Paynter, G.W. (1999). Lexically-generated subject hierarchies for browsing large collections. *International Journal of Digital Libraries, 2/3*:111-123.

[95] Open Bioinformatics Foundation. (2002). *OpenBio home page*. URL: http://www.open-bio.org.

[96] Pearl, J. (1998). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

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

[98] Pratt, W. (1999). *Dynamic Categorization: A Method for Decreasing Information Overload.* Ph.D. Dissertation, Stanford Medical Informatics, Stanford University, Stanford, CA.

[99] President's Information Technology Advisory Committee. (2001). Report to the President on *Transforming Health Care through Information Technology.* 09 February 2001. Distributed from URL: http://www.itrd.gov/pubs/pitac/pitac-hc-9feb01.pdf.

[100] Raymer, M., Punch, W., Goodman, E., Sanschagrin, P., & Kuhn, L. (1997). Simultaneous Feature Extraction and Selection using a Masking Genetic Algorithm, In *Proceedings of the 7th International Conference on Genetic Algorithms (ICGA-97)*, San Francisco, CA, p. 561-567.

[101] Resnick, P. & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3):56-58.

[102] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, p. 175-186. Chapel Hill, NC: ACM Press.

[103] Robinson, A. (2002). The myGrid Project: e-Science, the Grid, and Web Services. Distributed from URL: http://industry.ebi.ac.uk/~alan/NT/PowerPoint/myGrid/myGrid.ppt.

[104] Rocha, L. M. (2001). Adaptive Webs for Heterarchies with Diverse Communities of Users. Modeling, Algorithms, and Informatics Group (CCS-3), Los Alamos National Laboratories. Distributed from URL: http://www.c3.lanl.gov/~rocha/GB0/adapweb_GB0.html.

[105] Rollins, E. J. & Wing, J. M. (1991). Specifications as search keys for software libraries. In Koichi, F., editor, *Proceedings of the 8th International Conference on Logic Programming (ICLP-91)*, Paris, FRANCE, p. 173-187.

[106] Rosetta Biosoftware. (2002). Gene Expression Markup Laguage. Distributed from URL: http://www.rosettabio.com/products/conductor/geml/omg.htm.

[107] Runciman, C. & Toyn, I. (1989). Retrieving reusable software components by polymorphic type. In *Proceedings of the 4th International Conference on Functional Programming Languages and Computer Architecture (FPCA-89)*, London, UK, p. 166-173. Chapel Hill, NC: ACM Press.

[108] Russell, S. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* Englewood Cliffs, NJ: Prentice Hall.

[109] Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery, Special Issue on Applications of Data Mining to Electronic Commerce, 5(1-2)*:115-153.

[110] Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *Database System Concepts, 4th edition.* New York, NY: McGraw-Hill.

[111] Silverstein, C., Brin, S., Motwani, R., & Ullman, J. D. (2000). Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery, 4*(2-3):163-192.

[112] Slattery, S. & Mitchell, T. M. (2000). Discovering test set regularities in relational domains. In Langley, P., editor, *Proceedings of the 17th International Conference on Machine Learning (ICML-2000)*, Stanford, CA, p. 895-902. San Mateo, CA: Morgan Kaufmann.

[113] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccaromyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell 9:*3273-3297.

[114] Spirtes, P., Glymour, C., & Scheines, R. (2000). *The TETRAD Project: Causal Models and Statistical Data.* Pittsburgh, PA: Carnegie-Mellon University. Distributed from URL: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html.

[115] Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. In Heckerman, D., Mannila, H., Pregibon, D., & Uthurusamy, R., editors, *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, p. 67-73. Menlo Park, CA: AAAI Press.

[116] Syswerda, G. Successful Commercialization of Genetic Algorithms. Plenary address, *Genetic and Evolutionary Computation Conference (GECCO-2000)*, Las Vegas, NV, 11 July 2000.

[117] The Arabidopsis Information Resource. (2002). *TAIR Homepage.* URL: http://www.arabidopsis.org.

[118] The Institute for Genomic Research. (2002). *Arabidopsis functional genomics at TIGR.* Distributed from URL: http://atarrays.tigr.org.

[119] The Salk Institute for Biological Studies. (2000). *Functional Genomics and the Virtual Plant: A blueprint for understanding how plants are built and how to improve them.* NSF Workshop Report, The Arabidopsis Information Resource (TAIR). Distributed from URL: http://www.arabidopsis.org/workshop1.html.

[120] UK Research Councils. (2002). *UK e-Science (Grid) Core Programme*. URL: http://www.escience-grid.org.uk.

[121] University of Manchester myGrid Group. (2002). *myGrid: Directly Supporting the e-Scientist.* URL: http://mygrid.man.ac.uk.

[122] Wallentine, V. & Zhou, S. (2002). Validating XML document content with the object constraint language. In *Proceedings of the 2002 Internet Conference*, Las Vegas, NV.

[123] Whitfield, J. (2002). Rice genome unveiled. *Nature Science Update.* Distributed from URL: http://www.nature.com/nsu/020402/020402-6.html.

[124] Zaki, M. J. (2000). Data mining in bioinformatics. In *New Directions in Bioinformatics and Biotechnology Workshop*, Troy, NY. Distributed from URL: http://www.cs.rpi.edu/~zaki/PS/BIO00.ps.gz.

[125] Zaki, M. J. Scalable algorithms for association mining. (2000). *IEEE Transactions on Knowledge and Data Engineering*, *12*(3):372-390.

[126] Zhu, X., Gauch, S., Gerhard, L., Kral, N., & Pretschner, A. (1999). Ontology-Based Web Site Mapping for Information Exploration In *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM-99)*, Kansas City, MO, p. 188-194.