Learning Causal Bayesian Networks from Literature Data

Péter Antal, András Millinghoffer
Department of Measurement and Information Systems
Budapest University of Technology and Economics
e-mail: {antal,milli}@mit.bme.hu

April 26, 2006

Abstract

In biomedical domains free text electronic literature is an important resource for knowledge discovery and acquisition. It is particularly true in the context of data analysis, where it provides a priori components to enhance learning, or references for evaluation. The biomedical literature contains the rapidly accumulating, voluminous collection of scientific observations boosted by the new high-throughput measurement technologies.

The broader context of our work is to support statistical inference about the structural properties of the domain model. This is a two-step process, which consists of (1) the reconstruction of the beliefs over mechanisms from the literature by learning generative models and (2) their usage in a subsequent learning phase. To automate the extraction of this prior knowledge we discuss the types of uncertainties in a domain with respect to causal mechanisms and introduce a hypothesis about certain structural faithfulness between the causal Bayesian network model of the domain and a binary Bayesian network representing occurrences (i.e. causal relevance) of domain entities in publications describing causal relations. Based on this hypothesis, we propose various generative probabilistic models for the occurrences of biomedical concepts in scientific papers. Finally, we investigate how Bayesian network learning with minimal linguistic analysis support can be applied to discover and extract causal dependency domain models from the domain literature.

Keywords Bayesian network learning, text mining.

1 Introduction

The rapid accumulation of biological data and the corresponding knowledge posed new challenges for knowledge engineering to make accessible the voluminous, uncertain and frequently inconsistent knowledge. In machine learning we

have to cope with high-dimensional, noisy and relatively "small sample" data and to incorporate a priori knowledge in various learning and discovery algorithms. In natural language processing it is essential to retrieve relevant raw information (i.e. publications) and to extract relevant information from it. Despite recent trends aiming to broaden the scope of formal knowledge bases in biomedical domains, free text electronic literature is still the central repository of the domain knowledge. This central role will probably be retained in the near future, because of the rapidly expanding frontiers ([3, 14, 32, 26, 13, 34]).

The extraction of explicitly stated or the discovery of implicitly present latent knowledge requires various techniques ranging from purely linguistic approaches to machine learning methods. In the paper we investigate a shallow-statistical, domain-model based approach to statistical inferences about dependency and causal relations. We use Bayesian networks as the causal domain models to introduce generative models of causal papers, then we examine the relation between the probabilistic models of the domain and of the corresponding domain literature, and evaluate this approach in the ovarian cancer domain.

The broader context of our work is to support statistical inference about the structural properties of the domain model. This is a two-step process, which consists of (1) the reconstruction of the beliefs over mechanisms from the literature by learning generative models and (2) their usage in a subsequent learning phase. Earlier applications of text mining focused on providing results for the domain experts or data analysts, whereas our aim is to go one step further and use the results of these methods automatically in the statistical learning of the domain models. For this, the Bayesian framework is an obvious choice. The first step consists of reconstructing collective beliefs from the literature as parameters of generative models. Actually it can be conceived as an a posteriori belief given the literature data. In the second phase the Bayesian inference about the a posteriori probabilities of structural properties of the domain model given the clinical or biological data is the practical choice. Finally the link between these two steps can be formalized using the principled probabilistic semantics, i.e. our goal is to provide the a priori probabilities on the structural properties of the domain model derived from the literature (see Fig. 1).

The paper is organized as follows. In Section 2 we review the types of uncertainties in biomedical domain from the causal, mechanism oriented point of view. Also here we present the Bayesian framework of our approach. The framework is based on Bayesian belief networks. It fits the proposed generative model of the publications to the domain literature and uses these results as a priori elements to support Bayesian analysis of domain data. In Section 3 we summarize recent approaches to the information extraction and the literature mining based on natural language processing (NLP) and "local" analysis of occurrence patterns. In Section 4 we formulate a new hypothesis about the relation of causal mechanisms in the domain and the causal mechanisms governing the occurrences of concepts in the domain literature. We conjecture certain structural faithfulness between the causal Bayesian network model of the domain and the binary Bayesian network representing occurrences (i.e. causal relevance) of domain entities in causal publications. Based on this hypothe-

sis, we propose various generative probabilistic models for the occurrences of biomedical concepts in scientific papers. Finally, we investigate how the uncertainties over causal mechanism enter (as parameters) the generative models of the publications.

Section 5 presents the application domain, the diagnosis of ovarian cancer. We investigate how Bayesian network learning with minimal linguistic analysis support can be applied to discover and to extract causal dependency domain models from the domain literature. Section 6 reports a causal evaluation of a maximum a posteriori Bayesian network based on the literature data with respect to the expert's references. Section 7 presents the conclusion.

2 Uncertainty of causal domain model

A biomedical domain frequently can be characterized by a dominant type of uncertainty with respect to the causal mechanisms. Such types of uncertainty show certain sequential dependency, related to the process of biomedical knowledge extraction and formulation, though a strictly sequential view is clearly an oversimplification.

1. *Conceptual phase:* Uncertainty over the domain ontology, i.e. the relevant entities and concepts. This is of fundamental importance, considering that an effective (probabilistic) decomposition and causal modeling is partly the consequence of properly constructed domain concepts, so the feedback from later phases to guide this phase is crucial [25].

2. *Associative phase:* Uncertainty over the association of entities. These are reported in the literature as undirected and indirect, correlational hypotheses, frequently as clusters of associated entities. Though we accept the general assumption of causal relations behind the associations, we assume that the exact causal functions and direct relations are not known in this phase.

3. *Causal (relevance) phase:* Uncertainty over causal relations between the entities (i.e. over mechanisms). Typically direct causal relations are theoretized as processes and mechanisms.

4. *Parametric causal phase:* Uncertainty over the analytic forms of the autonomous mechanisms embodying the causal relations.

5. *Intervention phase:* Uncertainty over the effects of the interventions.

In this paper we assume that the target domain is already in its Associative and Causal phase, i.e. we assume that the entities are more or less agreed, but their causal relations are in the discovery phase. The direct dependencies and the functions of the entities are not known in the reported associations. This assumption holds in many biomedical domains, particularly in domains linking biological and clinical levels. In such domains, the Associative phase is a crucial

and lengthy knowledge accumulation phase, in which wide range of research methods is used to report associated pairs or clusters of the domain entities. These methods admittedly produce causally oriented associative relations which are partial, biased and noisy (c.f. various "-omics" levels [42]).

We consider two types of uncertainty of causal mechanisms. The first, called 'inherent', is the consequence of the subjective, partial understanding of the full mechanism and the objective, parallel presence of mechanisms. Uncertainty over the possible mechanisms can be modeled with another layer of uncertainty above the uncertain domain model by introducing new hidden variables that serve as selectors of the causal mechanisms. It is similar to the modeling of uncertainties over parameters with hyperparameters (see [24, 35], but this kind of uncertainty is conceptually different from the recent dualistic deterministic-probabilistic models of mechanisms in causal networks [25]. The second type of mechanism uncertainty, called 'contextual', corresponds to the contextual (in)dependencies [5, 16]. In this case the relevance of certain variables depends on the values of other variables (i.e. the relevance of a mechanism depends on the values of triggering variables). It can be modeled similarly with the introduction of a new hyperlayer with hidden variables, which serve as selectors of the causal mechanisms, though in this case hypervariables depend on the domain variables. Nonetheless, we treat the contextual uncertainties as inherent uncertainties, i.e. we assume that there is an independent belief for each variable over its corresponding potential mechanisms.

The central assumption to our work is that the beliefs over the mechanisms are important factors influencing the publications. They exert their effects as building blocks in generative models of the occurrences of domain entities in publications. Fig. 1 illustrates our assumptions about (1) the mechanism uncertainty in the domain in the Associative and Causal-relevance phases, (2) the corresponding literature data, (3) the reconstructed generative probabilistic model and (4) the application of reconstructed mechanism uncertainty as prior in statistical inferences about domain models.

3 Information extraction and literature mining

Causal relations (mechanisms) or related uncertainties are reconstructed from free text publications, mainly from abstracts. Abstracts report either causally associated domain entities (Associative phase) or report the explicit, direct causal relations with the causal functions of the entities (Causal-relevance phase). Our goal in this section is to highlight the differences between the knowledge discovery and information extraction methods and between the top-down and bottom-up methods. We will also illustrate the qualitative and quantitative relation between the domain model and its corresponding generative literature model.

The following list demonstrates the focus and characteristics of the approaches that have mainly influenced our work.

1. *Entity relationship extraction by linguistic approach* In the linguistic ap-

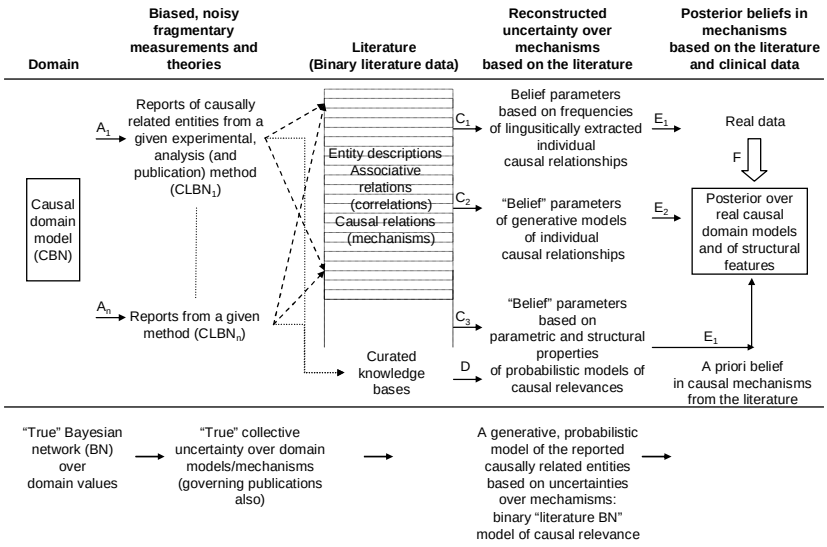

Figure 1: The reconstruction of fragmented prior knowledge in a biomedical domain from literature data and its incorporation in learning causal domain models.
Columns represent the phases of transformations of information concerning the domain. Headlines in the first row indicate the context, the second row contains the manifestations, and the third their possible representations.

proach explicitly stated relations are extracted from free text [11, 28, 29, 18], possibly with qualitative rating and negation, applying simplified grammars together with heuristic domain specific techniques such as POS taggers and frames (see e.g. the SUISEKI system [4]).

2. *Entity relationship extraction by co-occurrence frequency analysis* These methods are based on *name co-occurrence* quantifying the pairwise relation of two domain variables by the relative frequency of the co-occurrence of their names (and possibly synonyms) in documents from a domain specific corpus. In genomics, Stapley and Benoit [37] summarized the biological rationale for the relation between the biological relevance and the co-occurrence and performed a quantitative manual analysis for the model organism *Saccharomyces cerevisiae*, which indicated the usefulness of this approach for knowledge discovery in genomics. For human genes, Jensen et al. [21] performed an extensive quantitative manual check of such pair wise scorings based on the co-occurrence and concluded that the name co-occurrence in MEDLINE abstracts reflects biologically meaningful relationships with a practically acceptable reliability.

3. *Entity relationship (cluster) extraction by kernel similarity analysis* Methods based on *kernel similarity* quantify the relation of two domain variables

based on the vector representations of their textual descriptions (called kernels). The relation of two variables can be based on either direct similarity (if their descriptions are similar) or on indirect similarity (if the patterns of their descriptive documents are similar) [33].

4. *Entity relationship extraction by citation and temporal analysis* Friedman [22] suggested and tested a probabilistic generative model for individual relations that basically relies on a "true" (collective) belief of the relationship and then models the pattern of citations (corroborations and refutations).

5. *Relationship discovery by heuristic analysis of patterns of citation and co-occurrence* An early biomedical application from Swanson and Smalheiser [39] targeted relationship discovery by heuristic analysis of patterns of citation and co-occurrence, mainly relying on transitivity considerations.

6. *Relationship discovery by joint statistical analysis of patterns of co-occurrence* de Campos et al. [12] used the occurrence patterns of words to learn a restricted Bayesian network thesaurus from the literature.

These approaches can be further classified into information extraction or discovery methods. Roughly speaking, linguistic approaches assume that the individual relationships are sufficiently known, formulated and reported for automated detection methods, i.e. the linguistic approaches are applicable in the Causal-relevance phase or later. Whereas discovery methods assume that mainly causally associated entities are reported without or with tentative relations and direct structural knowledge. Consequently their linguistic formulation is highly variable, not conforming to simple grammatical characterization, i.e. these methods are applicable in the Associative phase. Therefore, linguistic approaches concentrate on the identification of individual relationships. The domain literature is analyzed piece-by-piece (by scientific papers or frequently by separated sentences) applying significant grammatical support. The integration is left to the domain expert who is supported by the raw summary of the individual relationships (such as e.g. pair wise literature networks). Statistical approaches on the contrary, after a simple grammatical and semantic preprocessing, concentrate on the identification of consistent domain models by analyzing jointly the numeric representation of the domain literature. These two groups rely on fundamentally different assumptions and consequently can be embedded differently in the knowledge engineering and the literature mining process. They require different preprocessing, their computational complexity, scalability with respect to the corpus, number of entities and relationships and sensitivity to the noise and bias in the scientific literature are different. Note that in the discovery methods the statistical inference proceeds from occurrence patterns of the entities to the probabilities of the entity relationships, whereas in Natural Language Processing (NLP) based information extraction it proceeds

from the reported entity relationships to the probabilities of the entity relationships. The NLP-based information extraction methods, useful to extract causal statements, can be applied prior to the *Causal* phase, whereas in the *Associative* phase only the causal discovery methods could deliver results.

If a domain theory does not exist yet or there is no consensus about an overall consistent causal domain theory, the NLP methods can identify the reported relations in the domain, whereas the discovery methods can discover new relations and autonomously prune the redundant, inconsistent, indirect relations by providing a unified consistent domain model. Interestingly, the discovery methods can also be applied to the Conceptual phase, because the identification of a consistent domain model may invoke new concepts ("hidden variables") as byproduct. This is consistent with the view that causation and the causal concepts that make it possible are the result of an active, constructive process [24, 25].

In practice, the Associative and Causal phases are never separated. It shows the necessity of a dual approach to the extraction and reconstruction of the uncertainties over mechanism. In the Associative phase uncertainties are present in the literature implicitly through patterns of occurrences of (causally) related entities, depending on the contemporary measurement technique, experimental methods, analysis of the experiment, publication policy and style, economic and social, ethical consequences. In the Causal phase uncertainties are present in the literature in explicitly stated forms, possibly explicitly naming the mechanism, and depend additionally on intentional, subjective, conscious beliefs over the mechanisms. We later consider in Sections 4.4 and 4.5 generative models that cover both aspects. For these models, the first, 'latent mechanisms' phase suggests a more experiment guided 'exploratory' interpretation of the model, whilst the second more intentional 'known mechanisms' phases suggest an 'explanatory' interpretations for the model, as closer to the intentional scientific investigation and explanation.

The construction of informative and faithful a priori probabilities over domain mechanisms or models from free text research papers is further complicated by:

1. *Uncertainty.* Usually there are multiple aspects such as uncertainty about the existence, scope (conditions of validity), strength, causality (direction), robustness for perturbation and relevance of mechanism. A related phenomenon is the overall inconsistency of the reported knowledge fragments.

2. *Incompleteness.* Certain relations are not reported because they are assumed to be well-known parts of *common sense* knowledge or of the *paradigmatic* knowledge of the community. Certain (*implicit*) relations are not reported purposely to decrease redundancy because they can be inferred from the usually reported knowledge items. Certain (*latent*) relations are not reported because they are unknown, though that they could be inferred from the already reported knowledge and finally there are objectively unknown dependencies, that are not reported.

3. *Consistency.* The extracted beliefs have to correspond to consistent domain models, possibly not decomposable to beliefs in individual mechanisms, e.g. to all possible (direct and indirect) pairwise relations.

4. *Scientific publication bias.* The information extraction method has to cope with the cognitive and publication constraints [30, 44], e.g. that new findings are not accompanied by a corresponding updated full survey of the domain, or the historical (temporal) and funding aspects of scientific publication.

Our approach is closest to the approach in [22], which investigates a generative model for the temporal sequence of occurrences of an individual relation incorporating the "true" (collective) belief in the relation and to the work reported in [39], which focused on the discovery of latent knowledge by consistency considerations, though in our case we focus and exploit the advantages of learning an overall, consistent domain model instead of the citational and temporal aspects.

4 Bayesian network models for literature mining

We will construct now a series of generative probabilistic models of publications mainly from the Associative phase, but also from the Conceptual and Causal-relevance phases. Of course, serious simplifications have to be made, because a probabilistic or causal model over these roles of the domain variables means a generative model of scientific explanation in publications, with certain implications to scientific research itself. Furthermore, beside the 'description', we should model the transitive associative nature of causal explanation over mechanisms, e.g. that causal mechanisms with a common cause or with a common effects are surveyed in an article, or that chains of causal mechanisms are tracked to demonstrate a causal path. On the other hand, we have to model the lack of transitivity, i.e. the incompleteness of causal explanations, e.g. that certain variables are assumed as explanatory, others as potentially explained, except for survey articles that describe an overall domain model. We use the belief network representation for the generative probabilistic models of publications.

4.1 Bayesian belief networks

A belief network represents a joint probability distribution over a set of variables [24]. We assume that these are discrete variables. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). The vertices V_i of the graph represent the random variables X_i and the edges define the direct dependencies (each variable is probabilistically independent of its non-descendants given its parents [24]). There is a probabilistic dependency model for each variable that describes its dependency on its parents.

Beside providing an efficient representation of high dimensional joint distributions, the Bayesian network representation has further advantages with respect to the structure of the domain variables. It provides an efficient and graphical representation of the conditional independencies with standard probabilistic semantics and enables inferences on conditional independencies [24]. It also provides a representation of causal domain models and enables causal inferences [25].

In the Bayesian framework the uncertainty over the structure of the domain model is represented by a distribution over the allowed Directed Acyclic Graph (DAG) structures. Assuming structure independence [6, 9], the probability of a domain model for a fixed ordering of the domain variables can be decomposed into the product of probabilities of the dependencies in the domain, which fits in the causal interpretation of the structure. Another frequent assumption, the so called edge independence, is that the belief in the substructures (i.e. in the parental sets) can be further decomposed into a product of probabilities corresponding to the belief that an individual parent is a member of the parental set, i.e. it is a direct cause of the investigated variable.

In the general case without a fixed ordering, either the features have to be selected carefully to ensure their independence (e.g. undirected edges) or their interaction can seriously distort the purported prior probabilities (for certain automated corrections see [7]).

The Bayesian update with complete data set $\underline{\mathbf{D}}$ can be performed using an analytic formula [9]. For a complete data set and fixed ordering, the posterior probability of a Bayesian network structure can be also decomposed into a product of independent parts, each expressing the a posteriori probability of the local dependency model conditioned on the data.

To summarize, in the Bayesian framework there are three layers of uncertainty related to Bayesian networks: uncertainty over the domain values in case of fixed structure and parameters ($P(X|\theta, S)$), uncertainty over the parameters ($P(\theta|S)$) and uncertainty over the structure ($P(S)$). Each of these can be used to represent uncertainty over mechanisms in a domain.

4.2 Occurrences of entities in causal publications

We start the construction of belief networks for the occurrences of the domain entities by considering possible interpretations, then the types of variables, the structure of the model and the local dependency models. We adopt the central role of causal understanding and explanation in scientific research [40, 41, 44]. We also assume the central role of causal explanations in scientific publications (for an overview of the relevance of non-causal relations, constraints, see [43]). Furthermore, we assume that the contemporary (collective) uncertainty over mechanisms is an important factor influencing the publications. In our formalization this mechanism uncertainty shows up in the publications as reports of the domain entities without specified direct relations (in the Associative phase) and as reports of the domain entities with specified direct mechanism (in publications from the Causal-relevance phase). We consider the interpretations of

the binary occurrences of the domain entities with respect to the Conceptual phase as independent descriptions (i.e. we neglect taxonomic publications) and with respect to the Associative and Causal-relevance phase as causally related and governed by mechanism uncertainty.

Wide range of interpretations can be obtained by considering the occurrences of domain entities in various types of the publications from all the phases, e.g. univariate, multivariate descriptive studies, taxonomic, bivariate cause-effect statistical studies, multivariate causal studies, surveys of the domain, diagnostic, therapeutic publications, etc. Corresponding interpretations, reflecting the pragmatic function of an occurring domain entity, can be the following (the interpretations for presence (positive-occurrence), absence (negative-occurrence) and missing status are given in parenthesis):

1. *Relevant:* unspecified relevance in discussing the domain (relevant / irrelevant / relevance unknown).

2. *Categorized:* investigated in domain taxonomy, logically relevant (categorized / not-in-domain-taxonomy / taxonomic status unknown).

3. *Observed / measured / known:* observed or known in the published study, or more specifically statistical data is collected about the variable (known / unknown / status unknown).

4. *(Independently) Described:* described without relation to other variables (described / nondescribed / description-unknown).

5. *Explanandum:* The variable is to be explained (explained / unexplained because of insufficient or incorrect explanation / epistemic status unknown).

6. *Explanans:* The variable is explanatory. (explanatory / nonexplanatory as unnecessary / epistemic status unknown).

7. *Explained / to be explained / understood / assumed (causally relevant): the merge of the explanandum (explained) and explanans (explanatory) interpretations.*

According to our causal stance, we accept the 'causal relevance' interpretation, more specifically the 'explained' (explanandum) and 'explanatory' (explanans), additionally, we allow the 'described' status. This is appealing because in the assumed causal publications both the name occurrence and the preprocessing kernel similarity method (see Section 5) expresses the presence or relevance of the concept corresponding to the respective variable. This implicitly means that we assume that publications contain either descriptions of domain concepts without considering their relations or occurrences of entities participating in known or unknown (latent) causal relations (c.f. Causal Markov Condition [36, 25, 15]). An in-depth analysis exceeds the scope of the paper, consequently we left it to the reader to consider the general "relevance" and "known" interpretation (for an overview of 'relevance' see [38]).

To model the occurrence pattern of the accepted three roles of the domain variables we continue with the types of variables, local dependency models and structures. According to our assumption about the dominance of the Associative phase, we assume that there is only one causal mechanism for each parental set (i.e. there is a one-to-one correspondence between the set of directly influencing variables and potential mechanism), so we will equate a given parental set and the mechanism based on this set (Assumption of *Single mechanism by relevance*). Theoretically this is not restrictive, but in later causal phases, such as in the Parametric-causal phase, there can be multiple alternative mechanisms for the same parental set.

4.3 The atomistic publication model

The simplest, atomistic approach is to assume that the reports of the causal mechanisms and the univariate descriptions are completely independent. Indeed, this is the currently prevailing assumption, because all the information extraction methods that extract, analyze and provide result separately for the individual relations rely on this assumption. These methods also assume that the individual reports of the causal mechanisms and the univariate descriptions can be sufficiently identified as shown in Fig. 2. Note that these methods are not intended to discover new latent mechanisms that are conjectured and loosely articulated or indicated only by associative patterns.

We assume that the belief in the hidden submechanism (HSM) is an important factor influencing the publication (other factors can be also mechanism specific such as e.g. social or financial factors). This factor establishes the link between the belief in the real world mechanism and the frequency of occurrence in the literature. It follows the approach in [22], constructing a generative model based on the belief in a pair-wise mechanism. Indeed, similar quantitative or qualitative hypotheses about the relations of real world properties of entities and relationships and publication properties are always analytically or qualitatively, tacitly assumed in the text mining applications (for an investigation of the relation of function and publication frequency of genes see [19]).

4.4 The intransitive publication model

Not known explanatory, explained and descriptive functions and mainly unstructured causal relevance associations or tentative relations cannot be identified sufficiently with linguistic methods. In such case the domain wide discovery methods can support the consistent identification of relations. In the construction of our first model, we assume that the reports of the causal mechanisms and descriptions are independent. In the explanatory interpretation it means that the subjective probabilities of the reports of causal mechanisms and descriptions are independent. In the exploratory interpretation it means that a fragmentary domain theory corresponding to a given experimental, analytical and publication method results in such independent causal relevance associations. We propose

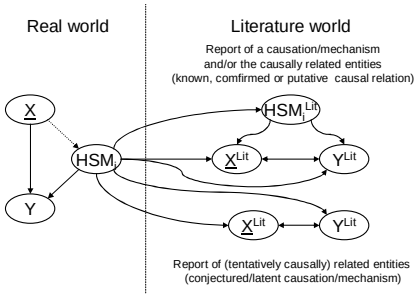

Figure 2: The separated extraction and analysis of the individual relations with the underlying assumption of complete independence of the report of the causal mechanisms and descriptions.

a two-layered Bayesian network structure as a corresponding probabilistic generative model. The upper layer contains variables corresponding to the possible causal functions of the entities, such as described, explained or explanatory (we treat explained as cause and explanatory as effect). In the explanatory interpretation these represent the authors' intentions, which induce the occurrences of the entities in the publication. In the exploratory interpretation these represent the bias and incompleteness of a given experimental technique. The lower layer contains variables representing the observable, external occurrences of the entities in the publications. An external variable depends only on the variables denoting the causal roles related to the corresponding causal mechanism (i.e. it is independent of other external variables, such as the number of reported domain entities in the paper and it is independent of other non-external variables of the neighboring causal mechanisms). The steps of the derivation from the first atomistic model to this more entity oriented model is shown in Fig. 3. This model extends the individual mechanism oriented information extraction by supporting the domain wide, consistent interpretation of causal roles, but still cannot model the dependencies (e.g. transitivity) between the reports of the mechanisms.

A further assumption, mainly motivated by the explanatory interpretation, is that the parental sets are composed of independent factors, i.e. that the belief in a mechanism is the product of the individual beliefs in the causes (see edge independence in Section 4.1). Consequently we use noisy-OR canonic distributions for the children in the lower layer and interpret the occurrence of a variable in a paper as described, explanatory or explained. In a noisy-OR local dependency [24], the edges can be labeled with an inhibitor parameter, inhibiting the OR function, which can be interpreted also structurally as the probability of an implicative edge (note the relation between the parametric and structural uncertainty). We set this parameter to zero for the 'explained to occurrence' edges, i.e. we assume that if a mechanism is explained, then the dependent variable is mentioned. In this generative model, these noise parameters represent the mechanism (structural) uncertainty over the domain model,

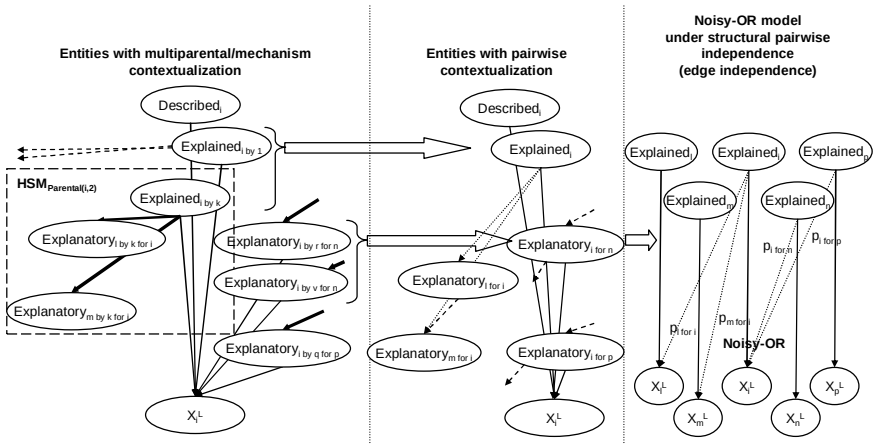

Figure 3: The derivation of the intransitive model with noisy-OR local dependencies from the first atomistic model.

i.e. we represented the mechanism uncertainty (structural uncertainty) over the domain model parametrically in the generative publication model. Though as noted above, because of the structural interpretability of the noisy-OR parameters, we can interpret in this special case that the mechanism (structural) uncertainty over the domain model is directly represented by the structural uncertainty over the generative publication model. In other words, the probability of an edge in a domain Bayesian network is equivalent to the probability of an edge in a corresponding Noisy-OR publication Bayesian network.

4.5 The transitive publication model

To devise a more advanced model with respect to the explanatory and exploratory interpretation, we relax the assumption of the independence between the variables in the upper layer representing causal functions, but maintain that an external variable depends only on the variables in the upper layer that participate within the same causal mechanism (Assumption of *'Sufficiency of causal explanation'*). First we consider if the reports of causal mechanisms are dependent in a causally transitive way, i.e. if we allow dependencies between the explained and the explanatory roles of the variables. In the explanatory interpretation this means that if a variable is explained, then it influences its explanatory role for other variables. If this transitivity dependency (explained to explanatory) is uniform in each pair-wise context, then a single explanatory variable can represent this role (Assumption of *'Uniform transitivity'*). The assumption of *'Full transitivity'*) means that this is an equivalence relation. In the explanatory interpretation it means, that if a variable is explained, then it can be explanatory for any other variable. In a full transitive case variables representing various causal roles such as the status of being explained and being

explanatory for another variable can be merged into one variable. Note that the transitivity of dependencies is satisfied in binary networks [24] conforming to an expectation about the transitivity of causal explanation. Furthermore we assume full transparency, i.e. the full observability of causal relevance (Assumption of *'Full transparency'*). Fig. 4 shows these steps.

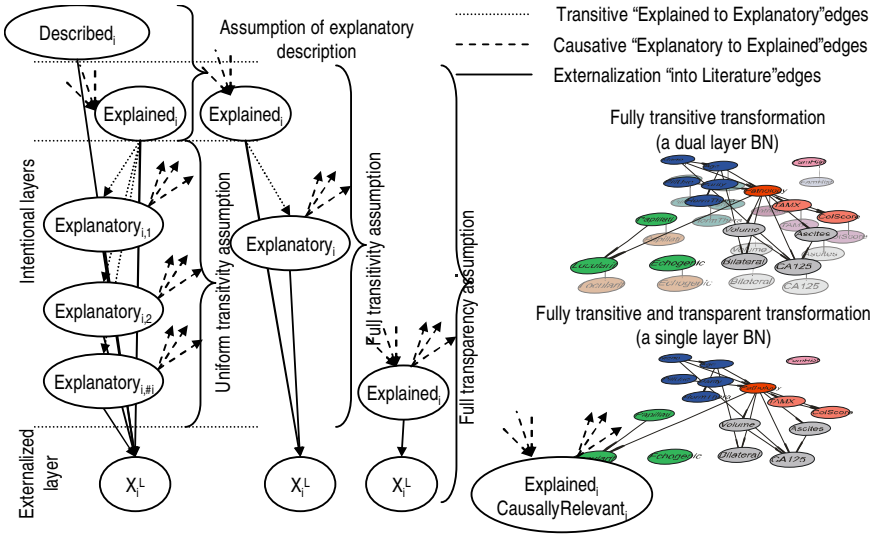

Figure 4: The derivation of the transitive model from the first atomistic model.

A consequence of the assumption of full transparency is that under this interpretation the lack of occurrence of an entity in a paper means causal irrelevance and not a neutral omission, i.e. there are no missing values. With full transitivity this would also imply that we model only full survey papers, but the general, unconstrained multinomial dependency model used in the transitive Bayesian network provides enough freedom to avoid this as discussed below. A possible semantics of the parameters of a binary, transitive literature Bayesian network $P(X_i|Parents(X_i))$ can be derived from causal stance that the presence of an entity X_i is influenced only by the presence of its potential explanatory entities, i.e. its parents. Consequently, $P(X_i = 1|Parents(X_i) = \underline{x}_i)$ can be interpreted as the belief that the present parental variables can explain the entities X_i as causes. A more strict interpretation requires necessity beside sufficiency, where $P(X_i = 1|Parents(X_i) = \underline{x}_i)$ denotes the belief that the present parental variables are the sufficient and necessary causes. These interpretations are also related to constructing explanations for Bayesian networks ([8, 23]). The multinomial model allows that at each node there are entity specific constants combined into the parameters of the conditional probability table that are not dependent on other variables (i.e. unstructured noise). This permits the modeling of the description of the entities ($P(X_i^{\text{Described}})$), the initiation of the transitive scheme of the causal explanation ($P(X_i^{\text{Assumed}})$) and the reverse effect

of not continuing the transitive scheme ($P(X_i^{\text{EnabledExplanation}})$), as follows:

$$
\begin{aligned}
P(Y|\underline{X}) \quad &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1)\\
&\quad P(Y^{\text{Described}} \vee Y^{\text{Assumed}})\\
&+ \quad P(Y|\underline{X} \wedge \neg Y^{\text{Described}} \wedge \neg Y^{\text{Assumed}} \wedge Y^{\text{EnabledExplanation}})\\
&\quad P(\neg Y^{\text{Described}} \wedge \neg Y^{\text{Assumed}} \wedge Y^{\text{EnabledExplanation}})\\
&= \quad (1 - P(\neg Y^{\text{Described}})P(\neg Y^{\text{Assumed}}))\\
&+ \quad P(Y|\underline{X} \wedge \neg Y^{\text{Described}} \wedge \neg Y^{\text{Assumed}} \wedge Y^{\text{EnabledExplanation}})\\
&\quad P(\neg Y^{\text{Described}})P(\neg Y^{\text{Assumed}})P(Y^{\text{EnabledExplanation}})
\end{aligned}
$$

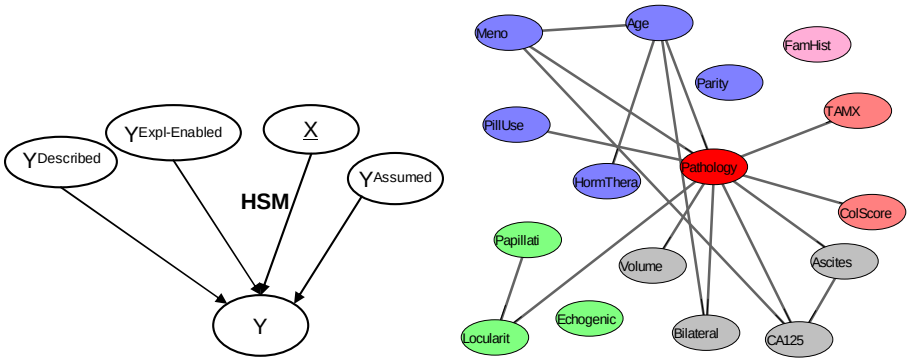

Figure 5: (Left) The auxiliary variables, which enrich the strictly causal transitive explanation with independent descriptions, unexplained assumptions and abruption of the explanation. (Right) The expert model: edges occurring in the highly relevant model are indicated by dashed lines, edges in the moderately relevant model are indicated by dotted lines

The effect of these auxiliary variables are illustrated in the left side of Fig. 4.5, demonstrating that this model allows partial explanations also. As the detailed discussion of related models is outside the scope of this paper, we stop here and note that a "backward" model using an effect-to-cause orientation is similarly an interesting model of the publications (c.f. means-ends analysis), in which the noisy-OR dependency model can be also used as in the intransitive model.

To summarize, the assumption of *'Sufficiency of causal explanation'*, *'Uniform transitivity'*, *'Full transitivity'* and *'Full transparency'* implies the structural faithfulness of a single layer generative probabilistic model of the publications to the real causal domain model. Furthermore, it is also capable to model the independent descriptions and the partial causal explanations with unrestricted (multinomial) local conditional probability models. The parameters of the Bayesian network encode the structure uncertainty over the domain, i.e. the mechanism uncertainty, because of our assumption of *Single mechanism*

by relevance. Of course, the merge of hidden variables, i.e. the incorporation of their effect distort it, but only as unstructured and partly analytically decomposable noise (see Eq. 1). Note, that the structural uncertainty over the domain model (i.e. a hyper level uncertainty) is represented parametrically in a generative Bayesian network, which can be conceived as a probabilistic model over relations. In the extreme case, a fully connected network can encode the beliefs of parental sets i.e. valid under the corresponding topological ordering (in this case the settings of the parameters are very similar to the frequency counting in the atomistic approach).

However, in the Bayesian framework there is a structural uncertainty also, i.e. uncertainty over the structure of the generative models (literature Bayesian networks) themselves. So to compute the probability of a parental set $Parents(Y) = \underline{X}$ given a literature data set D, which can be encoded in a literature Bayesian network with structure S as $P(Y = 1|Parents(Y) = \underline{X}, S)$, we have to average over the structures using the posterior given the literature data D as follows:

$$\begin{aligned} P(Y = 1|Parents(Y) &= \underline{X}, D) \\ &= \sum_S P(Y = 1|Parents(Y) = \underline{X}, S)P(S|D) \\ &= \sum_{S\, containing\, `\underline{X}\rightarrow Y'} P(Y = 1|Parents(Y) = \underline{X}, S)P(S|D) \\ &\approx \sum_{S\, containing\, `\underline{X}\rightarrow Y'} P(S|D) \\ &\approx I_{\{S^{MAP}\, contains\, `\underline{X}\rightarrow Y'\}} \end{aligned} \tag{2}$$

Consequently, the result of learning of Bayesian networks from literature data can be multiple, either using a maximum a posteriori network structure and the corresponding parameters or the a posteriori distribution over the structures. In the first parametric case, the special structural interpretation of the binary network guarantees that the parameters and the result of standard parametric inference in such a network can be interpreted structurally and can be converted into an a priori distribution for a subsequent learning. In the later case, we neglect the parametric information and focus on their structural constraints, we transform the a posteriori distribution over the structures of the literature networks into an a priori distribution over the structures of the real Bayesian networks with possibly multivalued or continuous variables. Finally, we can use only the structural features of a maximum a posteriori model for approximation.

Even if the presented "publication models" are simplistic, because of neglecting e.g. (1) general linguistic and pragmatist (publication-specific) constraints, (2) social, economic, historic and ethical factors and because (3) it is memoryless (consider that research and publication can be modeled as governed by the discrepancy between the published and believed "truth"), it is useful to test the resulting simple model to refine or relax the assumptions experimentally. To our knowledge such formal approach to investigate the assumptions behind struc-

ture oriented text mining applications has not been formalized earlier, though properties of these assumptions and the model was probably always tacitly assumed in the usage of the associative analysis of domain literature, such as in the co-occurrence analysis or in clustering [37, 20, 21, 2, 27].

5 The application domain: ovarian cancer

The experiments were performed in the ovarian cancer domain using sixteen clinical variables selected from a larger study and eighty genes [1]. We assume the existence of annotations for the Bayesian network variables (which include a textual name for the random variable, synonyms, a free text description (the kernel) and references to documents), collection of domain documents, and domain vocabularies (for an overview see [1]).

We have asked medical experts to select the *most relevant* journals for the domain and performed the query 'ovarian cancer' in the PubMed database[1] between 1998 and 2002 which resulted in 5000 papers. These publications were converted to a vector representation resulting in the literature data used in the paper (for the description of the domain, model construction and conversion steps of literature, see [1]).

Note that this preprocessing fits our assumption about the Associative phase, because the literature data contains only the binary occurrences (presence or absence) of the domain concepts corresponding to the domain variables.

6 Results

The structure learning of the transitive model is achieved by an exhaustive evaluation of parental sets using BD_{eu} score [17] up to maximum three parents using the ordering of the variables from the medical expert, which was a technical choice to be compatible with the learning of the intransitive model with hidden variables. The final network is shown in the left side of Fig. 6.

The structure learning of the two-layered model has a higher computational cost, because the evaluation of a structure requires the optimization of parameters, which can be performed e.g. by a gradient-descent algorithm. The possible (examined) structures have to satisfy that variables have less than a fixed number of parents, limited to four parents in this experiment, because of the computational complexity, only those variables in the upper layer can be the parents of an external variable that precede it in the causal order. Note that beside the optional three parental edges for the external variables, we always force a deterministic edge from the corresponding non-external variable. During the parameter learning of a fixed network structure the non-zero inhibitory parameters of the lower layer variables are adjusted according to a gradient descent method to maximize the likelihood of the data (see [31]). After the best structure is found, it has to be converted into the ordinary real world model by

[1] http://www.ncbi.nlm.nih.gov/PubMed/

merging the corresponding pairs of nodes in lower and upper layer. The final network is shown in the right side of Fig. 6.

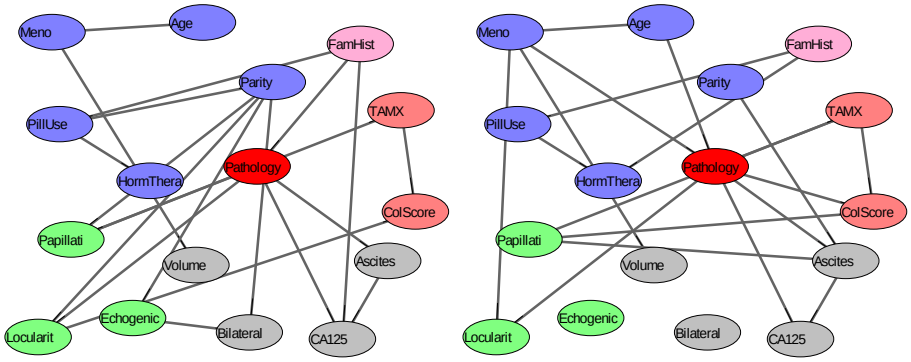

Figure 6: The transitive Bayesian network model with multinomial conditional tables and the intransitive Bayesian network with noisy-OR local conditional dependency models, to the left and to the right respectively (Note that the latter model is the conversion of the two-layered Bayesian network with hidden variables).

We compared the trained models to the expert model using a quantitative score that is based on the comparison of the types of the pairwise relations in the models. Exploiting the causal interpretation of the structure we use the following types of pairwise relations:

1. *Causal path (P):* There is a directed path from one of the nodes to the other.
2. *Causal edge (E):* There is an edge between the nodes.
3. *(Pure) Confounded (Conf):* The two nodes have a common ancestor. The relation is said to be pure, if there is no edge or path between the nodes.
4. *Independent (I):* None of the previous (i.e. there is no causal connection between the nodes).

The difference between two model structures can be represented in a matrix containing the number of relations with a fixed type in the expert model and in the trained model (the type of the relation in the expert model is the row index and the type in the trained model is the column index). E.g. the element $(I, Conf)$ shows the number of those pairs, which are independent in the reference model and are confounded in the examined. These matrices (i.e. the comparison of the transitive and the intransitive models to the expert's) are shown in Table 6.

Scalar scores to evaluate the goodness of the trained model can be derived from this matrix, e.g. a standard choice is to sum the elements with different

Table 1: Causal comparison of the transitive and the intransitive domain models (columns, to the left and to the right respectively) with the expert model (rows).

	I	Conf	P	E
I	14	14	12	12
Conf	6	14	0	2
P	44	48	24	14
E	14	6	4	12

	I	Conf	P	E
I	44	0	0	8
Conf	14	8	0	0
P	82	18	20	10
E	8	4	2	22

weights [10, 45]. One possibility e.g. if we take the sum of the diagonal elements as a measure of similarity. By this comparison, the intransitive model achieves 94 points, while the transitive only 64, so the intransitive preserves more faithfully the pair-wise relations. Particularly important is the (E, E) element according to which 22 of the 36 edges of the expert model remains in the two-layered model, on the contrary the transitive model preserves only 12 edges.

Another penalizing score, which penalizes only the incorrect identification of independence (i.e. those and only those weights have a value of 1 which belong to the elements $(I, .)$ or $(., I)$, the others are zero), gives a score 102 and 112 for the transitive model and the intransitive respectively, suggesting that the intransitive model is too conservative and results overly sparse models.

7 Conclusion

We investigated the applicability of Bayesian network learning methods to discover a causal domain model. We proposed two machine learning methods based on Bayesian networks, the first method assumes that the reporting activity of causal mechanisms follows a transitive scheme, the second method assumes that the causal mechanisms in the domain are reported autonomously (i.e. more or less independently). We performed an evaluation of these methods in the ovarian cancer domain. The evaluation shows that the fully observable transitive model and the intransitive model with hidden variables performs comparable to the performance of a human expert and the second, computationally more complex method proved to be slightly better than the first one. In future, we plan to test more complex transitive models and extend these methods to incorporate more information extracted by linguistic techniques.

References

[1] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004. Special issue on Bayesian Models in Medicine.

[2] P. Antal, P. Glenisson, G. Fannes, J. Mathijs, Y. Moreau, and B. De Moor. On the potential of domain literature for clustering and Bayesian network learning. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM-KDD-2002)*, pages 405–414, 2002.

[3] T. Berners-Lee and J. Hendler. Publishing on the semantic web. *Nature*, 410:1023–1024, 2001.

[4] C. Blaschke and A. Valencia. The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems*, 17(2):14–20, 2002.

[5] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks, 1996.

[6] W. L. Buntine. Theory refinement of Bayesian networks. In Bruce D'Ambrosio and Philippe Smets, editors, *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.

[7] R. Castelo and A. Siebes. Priors on network structures. biasing the search for Bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

[8] U. Chajewska and D. L. Draper. Explaining predictions in Bayesian networks and influence diagrams. In *Proc. of the AAAI 1998 Spring Symposium Series: Interactive and Mixed-Initiative Decision-Theoretic Systems*, pages 23–32, 1998.

[9] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[10] G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 116–125. Morgan Kaufmann, 1999.

[11] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1-2):69–113, 2000.

[12] L. M. de Campos, J. M. Fernández, and J. F. Huete. Query expansion in information retrieval systems using a Bayesian network-based thesaurus. In Gregory Cooper and Serafin Moral, editors, *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence (UAI-1998)*, pages 53–60. Morgan Kaufmann, 1998.

[13] R. J. Roberts et al. Building a 'genbank' of the published literature. *Science*, 291:2318–2319, 2001.

[14] M. Gerstein and J. Junker. Blurring the boundaries between scientific 'papers' and biological databases, 2001. *Nature* (web debate, on-line 7 May 2001).

[15] C. Glymour and G. F. Cooper. *Computation, Causation, and Discovery.* AAAI Press, 1999.

[16] D. Heckerman and J. S. Breese. Causal independence for probability assesment and inference using bayesian networks, 1995.

[17] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[18] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561, 2002.

[19] R. Hoffmann and A. Valencia. Life cycles of successful genes. *Trends in genetics*, 18:1–3, 2003.

[20] I. Iliopoulos, A. J. Enright, and C. A. Ouzounis. Textquest: document clustering of MEDLINE abstracts for concept discovery in molecular biology. In *Proc. of Pacific Symposium on Biocomputing (PSB01), Hawaii*, volume 58(2-3), pages 384–395, 2001.

[21] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, 2001.

[22] M. Krauthammer, P. Kra, I. Iossifov, S. M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, and A. Rzhetstky. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, 18:249257, 2002.

[23] C. Lacave and F. J. Diez. A review of explanation methods for Bayesian networks. Technical Report Tech. Report IA-2000-01, Dept. Int. Art. UNED, Madrid, 2002.

[24] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, CA, 1988.

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

[26] H. Pearson. The future of the electronic scientific literature. *Nature*, 413:1–3, 2001.

[27] Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature*, 31:316–319, 2002.

[28] D. Proux, F. Rechenmann, and L. Julliard. A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proc. of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000), LaJolla, California*, pages 279–285, 2000.

[29] T. C. Rindflesch, L. Tanabe, and J. N. Weinstein. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 514–525, 2000.

[30] A. Rosenberg. *Philosophy of Science: A contemporary introduction*. Routledge, 2000.

[31] Stuart J. Russell, John Binder, Daphne Koller, and Keiji Kanazawa. Local learning in probabilistic networks with hidden variables. In *IJCAI*, pages 1146–1152, 1995.

[32] N. Shadbolt. What does the science in e-science. *IEEE Intelligent Systems*, 17(May/June):2–3, 2002.

[33] H. Shatkay, S. Edwards, and M. Boguski. Information retrieval meets gene analysis. *IEEE Intelligent Systems*, 17(2):45–53, 2002.

[34] Vanessa Speding. Xml to take science by storm. *Scientific Computing World*, Supplement (Autumn):15–18, 2001.

[35] D. J. Spiegelhalter, A. Dawid, S. Lawritzen, and R. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[36] P. Spirtes, C. Glymour, and R. Scheines. Constructing Bayesian network models of gene expression networks from microarray data. In *Proc. of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology*, 2000.

[37] B. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 529–540, 2000.

[38] D. Subramanian, R. Greiner, and J. Pearl. The relevance of relevance. *Artificial Intelligence*, 97:l–5, 1997.

[39] D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[40] P. Thagard. Explaining disease: Correlations, causes, and mechanisms. *Minds and Machines*, 8:61–78, 1998.

[41] P. Thagard. Pathways to biomedical discovery. *??*, 18:1–18, 2002.

[42] M. Vidal. A biological atlas of functional maps. *Cell*, 104:333–339, 2002.

[43] J. Williamson. *Foundations for Bayesian networks*, pages 11–140. Kluwer Academic Publ., 2001.

[44] J. Woodward. Scientific explanation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2003.

[45] Xiaofeng Wu, Peter Lucas, Susan Kerr, and Roelf Dijkhuizen. Learning bayesian-network topologies in realistic medical domains. In *Medical Data Analysis: Second International Symposium, ISMDA 2001*, pages 302–308. Springer-Verlag, Berlin, 2001.