Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB

Martin Theobald, Ph.D[1], Nigam Shah, M.B.B.S., Ph.D[2], and Jeff Shrager, Ph.D[3]
Departments of (1) Computer Science, (2) Biomedical Informatics,
and (3) Symbolic Systems (consulting)
Stanford University, Stanford, CA 94305 USA
Contact: jshrager@stanford.edu.

Abstract

Guided by curated associations between *genes*, *treatments* (i.e., drugs), and *diseases* in pharmGKB, we constructed n-way Bayesian networks based on conditional probability tables (cpt's) extracted from co-occurrence statistics over the entire Pubmed corpus, producing a broad-coverage analysis of the relationships between these biological entities. The networks suggest hypotheses regarding drug mechanisms, treatment biomarkers, and/or potential markers of genetic disease. The cpt's enable Trio, an inferential database, to query indirect (inferred) relationships via an SQL-like query language.

Goal

The goal of clinical research can be thought of as seeking the conditional probability (cp) of a cure given particular treatments and diseases; in terms of conditional probabilities: statistically quantified and directed relationships of the form p(cure|treatment,disease) [hereafter: p(c|t,d)]. Meta-analysis over clinical trials can obtain an improved value by combining evidence statistically across trials. Such meta-analyses extract a p(c|t,d) that is *statistically tacit* in the literature. In the present work we explore other potentially useful statistically tacit results available in the medical literature. Specifically, we compute conditional probabilities between treatments (usually drugs), diseases, and genes: p(t|d,g), by analyzing their co-occurrence in Pubmed (www.ncbi.nlm.nih.gov/pubmed). Although not as directly useful as p(c|t,d), these cp's and their algebraic co-forms may be interpreted in a number of useful ways, for example as personalized (e.g., genetically guided) treatment hypotheses [p(t|g,d)], as drug mechanism hypotheses [p(g|t)], as treatment-response predictive biomarkers [p(g|t,d)], or as potential markers of genetic diseases [p(g|d)].

Suppose, for example, a patient is given a particular diagnosis, and a genomic-analysis reveals a mutation in a gene for which a targeted treatment has been explored in the scientific literature, but for which there is no approved specific treatment. P(t|g,d) gives one a sense of which treatments have been explored the most in this circumstance. Once a treatment is chosen, one wants to know which gene expression responses to watch in this patient (often not be the same as the mutated gene). Here p(g|t,d) may offer a sense of what the literature suggests as likely treatment-response biomarkers.

Thus the conditional probabilities among these entities across the scientific literature may lead to practical new hypotheses, and support inference to p(c|t,d), or eventually even to the holy grail of personalized genetic medicine: p(c|t,d,g).

Background

Many researchers have extracted association-based knowledge from the medical literature. Zhu, et al. [1] computed co-occurrence of compounds and genes, and Jenssen et al. [2] computed a gene-to-gene co-citation network. These are relatively simple computations. Extracting conditional multivariate statistics is much more difficult because it requires computing all combinations of associations, plus background counts for normalization, and the potential vocabulary is very large. Wren [3] extracted a network of associations among genes, diseases, phenotypes, drugs, etc. using the mutual information of shared associations from Pubmed abstracts over a set of 10,000 common words. In order to control the computational complexity and avoid saturation (which is likely with common words), Wren restricted his calculations to only 100,000 abstracts. Similarly, Narayanasamy, et al. [4] mined co-occurrence in Pubmed to build an association graph and ranked

associations co-occurring with both the objects (equivalent to mutual information). Although most of these projects uncovered various suggestive associations, they have either used a small corpus, focused on only one kind of entity (e.g. gene-gene), focused only on co-occurrence (which is symmetric as opposed to conditional probability), or recognized concepts from only one (or few) ontologies. In the present work we mine quantified, directed (i.e., asymmetric) association statistics for all-way drug/gene/disease relationships over the entirety of more than 19 million Pubmed abstracts and using all UMLS ontologies.

Methods

We seek to extract all-way co-occurrence-based Bayesian networks among treatments (primarily drugs for this study), diseases, and genes. These can be estimated from subsets of conditional and non-conditional probabilities which are in turn derived from raw co-occurrence counts of drug/disease/gene entities in domain-specific corpora such as Pubmed. For non-conditional statistics, such a co-occurrence probability would simply be the number of documents (or abstracts) that mention these items together, divided by the total number of documents contained in the corpus. The desired conditional probabilities are: p(drug|gene), p(drug|disease), p(drug|gene,disease), etc. One can easily see how to compute such conditional probabilities over an appropriately annotated Pubmed database, simply by counting the single and combinational co-occurrences of all of these entities, and performing the obvious calculation, i.e., p(drug A|gene B, disease C) = (# distinct abstracts containing A and B and C)/(# distinct abstracts containing B and C). Notice that more general relationships are conceivable, i.e., considering many-to-many relationships between drugs, diseases, and genes. In the present experiment we limit our Bayesian network to a maximum of six conditional variables and a single target variable, thus extracting up to 2^6 conditional probabilities per net. This keeps the combinatorial complexity, and hence the number of co-occurrence queries issued against our underlying Pubmed corpus, reasonable.

The problem with this approach is that an enormous number of combinations must be computed. If there are, say, 20,000 genes, a thousand drugs or investigational drugs, and a thousand diseases, $2^{22,000}$ combinations would have to be calculated. In order to make headway in this endeavor, we need guidance on which combinations to explore. One source of guidance could be a user query about the relationships between particular treatments, diseases, and genes. It seems unlikely, however, that a user would come up with likely combinations a priori.

We found the desired guidance in the pharmGKB database (www.pharmgkb.org), which explicitly (although non-statistically) relates drugs, genes, and diseases (and other entities). PharmGKB offers relationships between drugs, diseases, and genes, based on specific papers and different types of evidence ranging from "clinical outcome" to simply "discussed". (We dropped any marked "not related".) Note that, although we use these relations in pharmGKB to guide our analysis, we do not prioritize the specific papers used in pharmGKB, but use the entire Pubmed database for our statistics. Thus no *quantitative* bias is introduced by the papers curated into pharmGKB.

Guided by the relations in pharmGKB[1], we combined information from a tagged Pubmed corpus created by processing all Medline abstracts[2] using the Mgrep tool (University of Michigan). Mgrep uses all of the alternative strings for UMLS concepts[3] and identifies their occurrence in the abstract using a radix tree based method that allows for very fast processing without sacrificing precision [5]. In our experience, this method has an average precision of about 85% for diseases [6]. (We have not evaluated precision for other entities.)

The tagged corpus used in the present experiment contains ~19 million articles and ~200 concepts assigned to each article resulting in ~3 billion unique Pubmed ID-to-UMLS Concept Unique Identifier (CUI) assignments. We combined these data with the highly reliable gene2pubmed database (ftp.ncbi.nih.gov/gene/DATA). Using only the relationships marked as "related" or "positively related", we extracted 1,730 disease/drug/gene relationships with up to 6 conditional variables, and extracted their respective conditional probability tables using co-occurrence statistics over the ~3 billion distinct Pubmed ID/CUI pairs, resulting in

[1] Late 2007 snapshot of the pharmGKB database.
[2] Medline version from August 2007
[3] UMLS 2007 AB and AA

19,092 conditional probabilities (again, compare with $\sim 2^{22,000}$ for the full-joint distribution). Although this is clearly still an offline process, requiring several days, once extracted, these tables serve as input for our Bayesian nets and allow for an efficient execution of arbitrary inferential queries; *any conditional probability of variables/entities expressed in a pharmGKB relationship can be directly computed from these*.

Results and Extensions

The result of our method is a miniature Bayesian network for each of the pharmGKB relationships. For example: p(antidepressants | affective disorders, GNB3) = 0.33 (abbreviated: p(an|af,g) = 0.33. (Values are rounded to 2 decimal places, and names are shortened to unambiguous abbreviations.) That is, out of all the documents that mention "affective disorder" and gene "GNB3", about one third also mention anti-depressants. The subordinate relationships in this set include: p(~an|af,g) =.67, p(an|~af,g) = 0.04, p(~an|~af,g) = 0.96, p(an|af,~g) = 0.11, p(~an|af,~g) = 0.89, p(an|~af, ~g) = 0.0, and p(~an|~af,~g) = 1.0. (Note that complementary conditional probabilities add to 1.0) Many of these subordinate relationships may, of course, be irrelevant. Another example: p(azidothymidine | HIV, ABCC4) = 0.6. That is, in 60% of the papers where HIV and ABCC4 are mentioned, azidothymidine is also mentioned.

Note that the order of the contexts (following the vertical bar in the c.p.) is not relevant, but the targeted posterior (right side of the vertical bar) is relevant, and that the relationships are not symmetric across the conditional (vertical bar). Contrast, for example: p(mercaptopurine | azathioprine, thioguanine, TPMT) = 0.84, p(thioguanine | azathioprine, mercaptopurine, TPMT) = 0.73, and p(azathioprine | mercaptopurine, thioguanine, TPMT) = 0.89. And the subordinate relationships: p(azathioprine | thioguanine, TPMT) = 0.86, p(thioguanine | azathioprine, TPMT) = 0.44. The most clinically important results are, of course, the conditional probabilities of different treatments (drugs), for the same disease&gene combination. For example: p(salmeterol | Asthma, ADRB2) = 0.07 and p(salbutamol | Asthma, ADRB2) = 0.16.

Aside from direct relationships, one may want to assess indirect (i.e., *inferred*) relationships based on the very same precomputed nets. There are, of course, many more of these than the already burgeoning set of direct relationships. By directly extracting the conditional probabilities as input into a Bayesian net, our method allows for a more compact representation of the desired dependencies than it would be possible via capturing the full joint-distribution of all variables involved in such a relationship. Any conditional probability of variables involved in the net can be efficiently derived via Bayesian inference. For example, marginalized conditional probabilities of the form p(disease|gene) can be directly calculated from a conditional probability table initially extracted for p(drug|gene,disease) without going back to the source to extract more co-occurrence statistics. In this special setting, the obtained net always has a tree structure, permitting linear-time inference queries. This approach can be extended to extract arbitrary concepts beyond just drug/disease/gene relationships, by sampling co-occurrence statistics from pubmed for arbitrary text tokens.

Implementation

The present work is implemented as an extension of the Trio system [7], a database system for the integrated management of data, uncertainty and lineage. Trio uses an extended relational schema to capture data uncertainty (in the form of *alternative attribute values* and *confidences* associated with each of these attribute alternatives), as well as data lineage (i.e., pointers to internal or external sources of the data). For the specific inference setting explored here, it turns out that the notion of lineage can nicely be generalized to capturing arbitrary relationships between entities (or records in a database), thus providing pointers to other entities (again other records), which allows for a convenient way of encoding Bayesian nets directly on top of this extended relational setting. For the new inference component, each present/absent combination of variables in a pharmGKB relationship is encoded as a different alternative of such an "uncertain" record, along with a confidence value, which allows us to capture arbitrary, discrete probability distributions (including cpt's) for each record in the Trio schema.

Moreover, this affords a simple, declarative way of issuing true inference queries on top of the precomputed conditional and non-conditional probabilities. For example, the query:

```
SELECT mercaptopurine |
azathioprine , thioguanine
FROM DRUGS;
```

would simply select the conditional probabilities p(mercaptopurine|azathioprine,thioguanine) from the precomputed cpt's for DRUGS. Conversely (still based on the same input table DRUGS capturing the cpt's for the initial target variable *mercapturine*, but also pointers to the non-conditional priors of *azathioprine* and *thioguanine*), we can, for example, initiate an on-the-fly inference query asking for p(azathioprine|mercaptopurine), thus swapping the direction of the conditional probability and marginalizing the distribution (i.e., eliminating the conditional variable *thioguanine*) in a single, SQL-like query:

```
SELECT azathioprine |
mercaptopurine FROM DRUGS
COMPUTE INFERENCE;
```

The result of this inferential query is a new cpt for p(azathioprine|mercaptopurine) that had not been precomputed, and whose computation is triggered by the "COMPUTE INFERENCE" clause using the inferencing extension in Trio. Issuing such an inference query is much faster over these simple (in our case *tree-like*) Bayesian nets than going back to the entire Pubmed database and mining for the respective co-occurrence statistics at query processing time. Issuing an inference query in Trio over the precomputed cpt's takes less than a second, whereas extracting the raw co-occurrence statistics from the entire set of Pubmed abstracts for a single pharmGKB relation with up to 6 variables may take several minutes in our current, rather limited, computing environment.

Limitations and Directions

The probabilities that we derive reflect only co-occurrence in the literature, and *not*, for example, recommendations, so one must be cautious in interpreting these results. What, then, are they telling us, and is what they are telling us useful? Because the literature is historical, these results are not telling us what to try, but *what has been tried*, or, possibly, *what has been suggested* (if not actually tried). Under this analysis one might regard a high conditional probability as a sort of ranking of hypotheses regarding *potential* treatments given the context of disease&gene combinations, and, symmetrically: hypotheses about the potential biomarkers (genes) given the context of a disease&treatment combination.

Interpretation aside, many aspects of this present analysis need improvement before this method can be applied. First, as with most statistical methods applied to natural language, we have ignored the specific relationships between the entities, both in pharmGKB and in Pubmed, and especially the possibility of negatively expressed relationships. Of course, we already know, from pharmGKB, that there is some positive correlation, because we filter out those that are marked as "not related" in that database. Moreover, given that our statistics include a huge number of papers, it is unlikely that a large fraction of them are telling us that "drug X does NOT have any effect on disease Y" (etc.), especially as the scientific literature does not often report negative results. Second, the particular tagger that we used does a poor job of dealing with gene synonyms. Synonym resolution is dependent on the UMLS CUIs. We use all synonymous strings for a CUI while doing the tagging, and the output only contains the CUI. This is not a particularly good solution to this complex issue.

We recognize that our use of Mgrep, as well as our use of co-occurrence as a substitute for actual relationships may introduce significant inaccuracies. In a project in progress we are generating parse trees of each sentence in the abstract and then only using the noun phrases for recognizing mentions of diseases and drugs. This should significantly increase accuracy.

More critically, because this method is focused by pharmGKB, we cannot discover direct relations that might be important, but which are not mentioned in pharmGKB. One way to resolve this might be to build the method into a search engine and use the combinations that are explicitly searched for as guidance (instead of using pharmGKB); indeed, this is explicitly enabled by the Trio infrastructure, and given the precalculated Pubmed/CUI database, seeking any given relationship set takes only a few minutes over the entire set of annotated Pubmed abstracts. There could be other sources of such guidance as well. For example, one could use the literature itself: co-mentions in specific abstracts, either all of them (only a few million computations vs. $2^{22,000}$), or perhaps a reduced hash that selects all unique co-mentions (certainly less than the whole literature).

Regardless, of these proposed extensions to the method that we have demonstrated, and approaches to its limitations, exhaustive evaluation is clearly needed to justify its utility.

Acknowledgements

MT's work was supported by NSF grants IIS-0324431 and IIS-0414762 and by grants from the Boeing and Hewlett-Packard Corporations. NS's work was supported by NIH grant U54 HG004028 and a gift from CommerceNet. JS's work was supported by CollabRx, Inc. The comments of several anonymous reviewers helped to improve the final paper.

References

1. Zhu, S., Okuno, Y., Tsujimoto G., Mamitsuka H. 2005. Mining literature co-occurrence data using a probabilistic model. IPSJ SIG Technical Reports, 99(BIO-2):9-16.

2. Jenssen, T-K, Lægreid, A, Komorowski, J., Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28, 21-28.

3. Wren, J.D. 2004. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*. **5:**145.

4. Narayanasamy V, Mukhopadhyay S, Palakal M, Potter DA. 2001. TransMiner: mining transitive associations among biological objects from text. *J Biomed Sci.* 11(6):864-73.

5. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B., Meng, F. 2008. An Efficient Solution for Mapping Free Text to Ontology Terms. Poster at AMIA Summit on Translational Bioinformatics, San Francisco.

6. Bhatia N., Shah N.H., Rubin D.L., Chiang A.P. and Musen M.A. 2009. Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. Paper accepted to the AMIA Summit on Translational Bioinformatics, San Francisco.

7. Benjelloun, O. Das Sarma, A., Halevy, A. Theobald, M., Widom, J. 2008. ULDBs: databases with uncertainty and lineage. *The International Journal on Very Large Data Bases*, Special Issue 2/08:243-264.