Topic Models with Relational Features for Drug Design*

A. Srinivasan[1], T.A. Faruquie[2], I. Bhattacharya[3], and R.D. King[4]

[1] Department of Computer Science, South Asian University, New Delhi, India.
[2] IBM Research—India, Block 4-C, Vasant Institutional Area, New Delhi, India.
[3] Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India.
[4] School of Computer Science, University of Manchester, United Kingdom.

Abstract. To date, ILP models in drug design have largely focussed on models in first-order logic that relate two- or three-dimensional molecular structure of a potential drug (a ligand) to its activity (for example, inhibition of some protein). In modelling terms: (a) the models have largely been logic-based (although there have been some attempts at probabilistic models); (b) the models have been mostly of a discriminatory nature (they have mainly used for classification tasks); and (c) data for concepts to be learned are usually provided explicitly: "hidden" or latent concept learning is rare. Each of these aspects imposes certain limitations on the use of such models for drug design. Here, we propose the use of "topic models"—correctly, hierarchical Bayesian models—as a general and powerful modelling technique for drug design. We retain the principal advantages of ILP by the incorporation of interesting relational features that are discovered automatically by an ILP engine. Specifically, we use: (1) The feature-construction abilities of a general-purpose ILP system to incorporate complex relational information about data that consists of drug-like molecules; and (2) A Bayesian modelling technique to construct a generative model for the data in this relational feature-space. The model consists of a set of latent (hidden) "topics"—loosely, concepts—along with probability distributions that relate molecules with topics, and topics with the specified relational features. Since the model is generative it can be used for a variety of tasks. Our main interest in this paper is to describe computational tools to assist the discovery of drugs for malaria. To this end, we describe the construction of topic models using the GlaxoSmithKline Tres Cantos Antimalarial TCAMS dataset. This consists of about 13,000 inhibitors of the 3D7 strain of *P. falciparum* in human erythrocytes, obtained by screening of approximately 2 million compounds. We consider three kinds of tasks that arise when identifying potential drugs. First, the discrimination of molecules into groups (for example, "more active" and "less active"). This is the usual problem of classification. Second, the retrieval of molecules "similar" to molecules of interest, in our case similar to ones with known targets.

* We are aware that this paper is longer than the length allowed for a short submission. Several aspects here have needed elaboration, and we seek the referees' indulgence in this matter.

Finally, the generation of new molecules for potential synthesis. In this shorter version of the paper, we only present results related to the first task, namely molecule classification. For this task, we present evidence that suggests that when it is important to maximise the detection of molecules with high activity ("hits"), topic-based classifiers may be better than those that operate directly on the feature-space representation of the molecules.

1 Introduction

Malaria is one of the world's worst diseases. The WHO's *World Malaria Report 2011* estimates that there were about 216 million cases of malaria in 2010 with about 655,000 fatalities.[5] Despite some decrease in the incidence of malaria over the last decade (about 17% since 2000), the disease remains widespread in tropical regions of Africa, Asia and the Americas. The primary parasite involved in the most severe cases is *Plasmodium falciparum*, and the drugs currently possessing the greatest efficacy against this parasite are artemisinin (empirical formula $C_{15}H_{22}O_5$) and its derivatives. The primary reason for the bioactivity for this group of drugs and the actual molecular targets that they attack is still not completely known [10], but the peroxide (two oxygen atoms bonded together) in a seven-membered ring is believed to play a significant role (see Fig. 1).

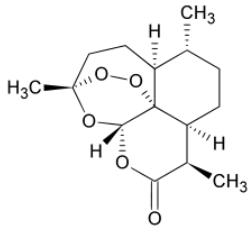

Fig. 1. The molecular structure of artemisinin (from `en.wikipedia.org/wiki/Artemisinin`). Artemisinin and its derivatives are currently the most effective treatment against the malaria parasite *P. falciparum*. The principal structural feature responsible for the activity of the molecule is the peroxide "bridge" in the 7-membered ring. This ring in artemisinin is connected to a lactone ring that contains 2 additional oxygen atoms.

A source of great concern is that a resistance to artemisinins has been reported in a growing number of countries [16]. While this resistance has been controlled to some degree by the use of therapies that combine artemisinins with other drugs, there is now an urgency to develop new anti-malarial drugs

[5] The Institute for Health Medicine and Evaluation (IHME) at the University of Washington, Seattle (Vogel, G. (2012) How do you count the dead?, Science, 336. 1372-1374.) estimate a death toll twice as large.

that can match the efficacy of artemisinin-based drugs. The development of such drugs is a key recommendation in the WHO's *Global Plan on Artemisinin Resistance Containment* [15]. It is probable that compounds that share significant structural similarities to artemisinin may not be effective against artemisinin-resistant parasites. Developing new anti-malarials therefore necessarily has to account for the presence or absence of certain kinds of structural features, and it is important that any tool assisting this development also be capable of representing such features, and construct models that account for their importance in some principled manner.

Inductive Logic Programming (ILP) systems have shown in the past that they are capable of representing and constructing models using structures such as those shown in Fig. 1 [8]. It is conceivable, with appropriate background knowledge and data, an ILP system could have found the logical expression of the following rule:

A molecule m is active if:
m has a 7-membered ring r_1 and
r_1 has a peroxide bridge and
m has a lactone ring r_2 and
r_1 and r_2 are connected

To find such a rule an ILP system would need at least general definitions of ring-structures, peroxide bridges, and predicates to decide when one or more rings are connected to each other. It is evident that this particular rule could be used to confirm that the artemisinin molecule is active. More could be done: the rule could be used to identify all molecules in a database that satisfy these constraints. Also, with some slight effort, a "molecular generator" could be constructed that was guaranteed to satisfy these constraints. But there are clear limitations. The rule requires the joint presence of the peroxide bridge and the lactone ring. It is therefore not possible to consider these separately in the search or generation of new molecules, although it might in fact be the case that the two structural features may be responsible for different aspects of the behaviour of artemisinin. That is, if we want to look for molecules that contain at least one of the two structural features:

A molecule m is active if:
m has a 7-membered ring r_1 and
r_1 has a peroxide bridge

A molecule m is active if:
m has a 7-membered ring r_1 and
m has a lactone ring r_2 and
r_1 and r_2 are connected

then this cannot be inferred directly from molecules that satisfy the original rule. The rule of inference used by ILP systems also do not allow for the generation of molecules using a stochastic mechanism that "weights" these two features differently.[6] Yet these are all clearly important tasks in the search for new drugs.

[6] Although it is conceivable that recent work on probabilistic ILP, or PILP, systems will allow this.

Therefore what we would like is to combine the usual advantages of an ILP system: the use of background knowledge and the discovery of complex features first-order logic—with that of modelling technique that discovers automatically partitions on the features to represent sub-concepts, and allows the search and generation of new molecules that uses these concepts. We propose the probabilistic setting of hierarchical Bayesian modelling, sometimes called "topic" models, as the appropriate one for fulfilling the modelling requirements outlined here.

2 Topic Models for Molecules

Topic modelling, originally used in the analysis of text documents, is concerned with three principal entities: documents, topics, and words. Documents consist of one or more topics, which in turn consist of one or more words. We adapt the original topic modelling to modelling molecules in the following way. Molecules ("documents") will be taken to consist of one or more concepts ("topics"). Examples of concepts relevant to drug-design are "activity" and "toxicity". Concepts will be taken to consist of one or more features ("words", although more like phrases in the molecular setting). For example, an active molecule may consist of the following features: the presence of a 7-membered ring, the presence of a lactone ring connected to a 7-membered ring, and the presence of a peroxide bridge in a 7-membered ring. It is understood that some mechanism exists for deciding whether such features are present or absent in any molecule. The question of how the features are to be obtained in the first place is discussed in the next section.

Tools for constructing topic models for molecules accept as input feature-vector descriptions of the molecules. In this paper, we will assume these descriptions to be Boolean vectors, in which a "1" signifies the presence of the corresponding feature in a molecule, and the phrase "molecule m has feature f" should be taken to mean that feature f has the value 1 for molecule m. The output is a set of concepts, each characterised by a probability distribution over the features. Each molecule in turn is characterised by a probability distribution over the concepts. We elaborate on these further, using a standard technique for topic modelling (Latent Dirichlet Allocation, or LDA [2]).

Suppose there is a set of features $\mathcal{F}$, and a set of molecules $\mathcal{M}$. Molecule m has a set features that is a subset of $\mathcal{F}$. We will take $|\mathcal{F}| = V$ and $|\mathcal{M}| = M$. We further assume that the the total number of concepts is K (usually, K is much smaller than V). Then, the LDA model asssumes the following:

1. Associated with each molecule m is a multinomial distribution over the entire set of K concepts. The parameter θ_m of this distribution gives the probability of observing the concepts for the molecule m. It is assumed that the molecule-specific concept-distributions are drawn using some prior distribution over multinomial parameters. It is convenient to assume that this prior distribution is a (K-dimensional) Dirichlet distribution with parameter α.
2. Similarly, associated with each concept k is a multinomial distribution over the entire set of V features. The parameter ϕ_k of this gives the probability of

observing the features for concept k. Again, it is assumed that these feature-specific distributions are in fact drawn using some prior distribution over V-dimensional multinomial parameters. A mathematically convenient prior distribution for the parameters is a (V-dimensional) Dirichlet distribution with parameter β.

Once the prior distributions, and the multinomials ϕ and θ are fully specified, it is possible to compute the (posterior) probability of observing the feature-values of any molecule. The difficulty however is this: the only observables are the molecule-specific feature values. All other quantities are hidden or latent. Estimating the parameters of the distributions from the observable data is a problem of Bayesian inference: the details of how it is accomplished here are not included in the short version of the paper.

Once parameters have been estimated the probabilistic model can be used in a number of different ways. Given any molecule it can compute the probability of each concept. The data are thus automatically reduced from the original V-dimensional representation to a smaller K-dimensional one. This reduced representation can then be used to construct models that classify molecules into one of several known categories, or to retrieve molecules that have similar concept-probabilities to a known ligand. More interesting to a synthetic chemist is that the model can be used to generate fragments of new molecules by drawing features from the underlying probability distributions (the details are not included in the short version of the paper).

2.1 Relational Features for Molecules

The probabilistic modelling technique requires that the objects being modelled be represented by feature-vectors. When modelling documents in a natural language, an obvious choice for features is the counts of words from a vocabulary. What about objects like molecules? The "characters" making up these objects are the usual chemical atoms (along perhaps with some "punctuation" symbols, to denote some special bonds and so on). These are connected together by bonds of various kinds to form various cyclic structures (rings), functional groups and the like. However a representation of molecules that simply counts certain kinds of atoms, rings, or functional groups is unlikely to be useful. A key assumption in the structure-activity modelling is that it is the relationships amongst these structures that gives rise to its chemical activity. Thus, it is not just that molecule has 3 benzene rings, but that it has 3 fused benzene rings that is actually important. This suggests a representation in which features are relations constructed from the basic molecular structural building blocks. This is somewhat akin to representing documents by phrases rather than words, and just as there may be a very large number of possible phrases, there may be a very large number of such relational features for molecules. A significant effort has been invested in ILP on how to obtain features for molecular data [13, 11]. We will not reproduce these techniques here, but simply show in Fig. 2 the process used to obtain structural fragments from a set of molecules using an ILP engine.

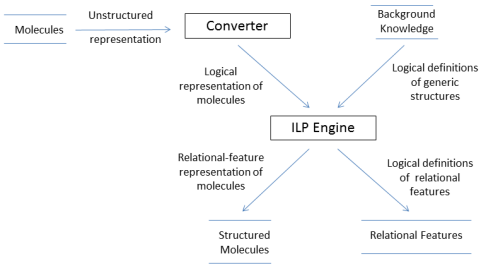

Fig. 2. The process of extracting relational features for molecules from data using an ILP engine. The ILP engine can use any of the "propositionalisation" techniques reviewed in [11]: the the specific one employed here is in the "Materials" section. The database of "structured molecules" will be used to construct topic models.

3 Empirical Evaluation

We intend to investigate the use of topic models for drug-design by using tasks relevant to the discovery of new anti-malarials. Within the broad area of structure-activity relations, we consider 3 tasks:

1. The discrimination of molecules into categories (like more or less active), based on their two-dimensional structure and some bulk properties (molecular weight and hydrophobicity);
2. The retrieval of molecules which are similar to an active molecule with a known target;
3. The generation of new "molecules" which share structural properties with a set of highly active molecules.

In the short version of this paper, we will only report on results from the first task, namely discrimination of molecules into categories. Specifically, we investigate the comparative performance of a classifier that uses the probabilities computed by an LDA model using relational features against some standard classifiers operating directly on the feature-space.

3.1 Materials

Data The Tres Cantos Antimalarial TCAMS dataset is available at the ChEMBL Neglected Tropical Disease archive (www.ebi.ac.uk/chemblntd). This archive provides free access to screening and chemical data relevant to a number of tropical diseases affecting countries in Asia, Africa and the Americas. The TCAMS

database is a result of screening GlaxoSmithKline's library of approximately 2 million compounds. The database consists of 13,000 of the chemicals that were found, on screening, to inhibit significantly the growth of the 3D7 strain of *P. falciparum* in human erythrocytes [5]. Data made available include some bulk-properties of the compounds (like molecular weight and hydrophobicity), along with the SMILES representation of the structure of the molecules. The distribution of inhibition activities of the molecules is shown in Fig. 3

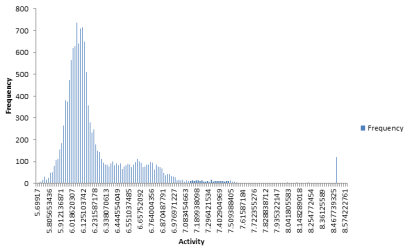

Fig. 3. Distribution of inhibition activity of the 3D7 strain of *P. falciparum* by molecules in the TCAMS database (pXC50_3D7 in the data).

Background Knowledge The ILP engine uses definitions of a number of standard cyclic structures and functional groups that have been used in structure-activity applications before [8, 9]. For reasons of space, we do not reproduce these here.

Algorithms and Machines The algorithms used are the following:

1. The conversion of the representation of molecules from their SMILES representation to a logical form is done by a combination of Open Babel [3], that converts from SMILES to a "mol2" format, and a converter that converts the "mol2" representation to a logical form.
2. The ILP engine used to generate relational features from the logical representation of the molecules is Aleph [12].
3. The topic models are constructed by a package provided with the R statistical package [6].
4. All classification is done by classifiers within the WEKA data mining software [7].

The topic models were constructed on a 2 GHz processor 4 core laptop with 4GB main memory running Mac OS v10.6.8. The ILP engine and classification were performed on a Pentium Core i5 laptop with access to 4GB of main memory running Fedora 13, and using the Yap Prolog engine.

3.2 Method

Our method is straightforward:

1. Two-dimensional structure information in terms of a logical representation of their atom and bond structure is obtained for the 13,403 molecules using the converters.
2. Partition the data into "training" and "test" subsets (the former for constructing models and the latter for testing them).
3. With the training data:
   (a) The ILP engine is used to construct relational features using the logical representation of the molecules and the background knowlege provided.
   (b) Each molecule is converted into a feature-vector representation in which an entry of "1" denotes that the corresponding relational feature is present in the molecule. This will be called the "feature-space representation" of the training data.
   (c) The topic-modeller is given molecules in the feature-vector representation to obtain a topic-model with exactly K topics
   (d) Molecules are re-expressed as K-dimensional probability vectors, in which the i^{th} entry for molecule m represents the probability that m contains topic i. This will be called the "topic-space representation of the training data"
4. With the test data:
   (a) Using the features constructed with the training data, obtain the feature-space representation of the molecules.
   (b) With the topic model constructed with the training data, obtain the topic-space representation of the molecules.
5. Obtain classification labels for all the molecules.
6. With both training and test data:
   (a) Add label information to the feature-space representation. We will call this the "labelled feature-space representation" of the data.
   (b) Add label information to the topic-space representation. We will call this the "labelled topic-space representation" of the data.
7. Construct the best classifier possible with the labelled feature-space representation of the training data. We will call this the "feature-based model": let its performance on the corresponding test data be P_F.
8. Construct the best classifier possible with the labelled topic-space representation of the training data. We will call this the "topic-based model": let its performance on the corresponding test data be P_T.
9. Compare P_F and P_T.

The following details are relevant:

1. 30% of the total number of molecules are set aside for testing the models obtained. This results in 9382 molecules in the training set and 4021 molecules in the test set.
2. The ILP engine Aleph is used to obtain features. We restrict these to be from at most $10,000$ definite clauses with no more than 5 literals. The features are constructed using Aleph's random-search strategy, independent of each other, and without access to any class labels. The minimum number of molecules to be "covered" by a feature is set very low to allow a large number of frequent features to be found.
3. There is no theoretically optimal number of topics. We construct models with $K = 10, 25$ and 50 topics. We note also that if classification is the only task of interest, then better variants of topic models exist. We would expect that an approach similar to that described in [14] would be better than the unsupervised topic-modelling we perform here.
4. In the classification task addressed here, we attempt to discriminate between molecules in the right-hand tail of the distribution in Fig.3. By "right-hand tail" we will mean the molecules in the 85-percentile or above, which translates to an activity threshold of approximately 6.6. We call these molecules "highly active", and the rest simply "active". The details of the resulting datasets are as follows:

Class	**Train**	**Test**
Highly Active	1486	666
Active	7896	3355
Total	9382	4021

5. All molecules in the data are inhibitors of *P. falciparum*. The focus on the right-hand tail results in a substantially skewed class distribution. There are differential costs associated with misclassification. Specifically, the cost of missing "highly active" molecules is substantially greater. Our experience suggests 10-fold difference. That is, denoting "highly active" as positives for learning and the rest as negatives, the cost of false negatives is likely to be 10 times that of false positives. A more complete picture is obtained by considering a distribution of cost values: in [1] a triangular distribution with a mode at the likely value is suggested, using a re-scaled version of costs in which the cost of false negatives (c_1 in [1]) is in the range $0 \ldots 1$ and the sum of costs of false negatives and false positives (c_0) is 1. We will consider a variation of cost ratios from 1:1 to 1:20. That is, c_0/c_1 is in the range $[0.05, 1]$ with the most likely value corresponding to $c_1 = 0.91$ (that is, a cost ratio of 1:10). Following [1], the distribution of c_1 values is taken to be a triangle with end-points at $c_1 = 0.67$ and $c_1 = 0.95$, with a peak at $c_1 = 0.91$. The height of the triangle is $2/(0.95 - 0.67) = 7.12$ (obtained so that the area of the triangle is 1). We will use this distribution to compute the LC-index.
6. The classifiers used here are implementations of a tree-builder (J48) and Naive Bayes in WEKA. A meta-classifier is used to minimise expected misclassification cost. This requires posterior probability estimates from the

classifiers, and it is known that heavily pruned trees give poor estimates of such probabilities [4]. We therefore use unpruned trees, with a Laplacian smoothing on the posterior probabilities. For Naive Bayes, we reiterate that feature-construction by the ILP engine is performed randomly, without using information from other features or the class label. We would expect conditional independence assumptions of naive Bayes to hold when using the feature-based representation. The extent to which the topic-space attributes satisfy these assumptions is however less clear.

3.3 Results

The ILP engine identifies about 2100 features. The results obtained with these features are shown in Fig. 4.

Fig. 4. A cost distribution and the performance of models using a tree-based (left) and naive Bayes classifier (right). In each case, "Feature" represents the classifier model in the feature-space, and "Topic" represents the model in the topic-space. For simplicity, the latter only shows the best model obtained with 10, 25 and 50 topics (what is meant by the "best" model will be apparent shortly). The horizontal axis (c_1) is the cost of false negatives, re-scaled to the interval $[0, 1]$ (the number in parentheses gives the actual costs of some points). The mode of the triangular distribution is at $c_1 = 0.91$ (unscaled value = 10). The bold segments against each classifier denotes the range of c_1 values for which that model has lower expected cost.

In this short version of the paper, we do not present any extensive discussion on the results. We note that for tree-based models, the best topic model exhibits superior performance as the cost of false negatives approaches the likely value of 10, or is greater. With naive Bayes, the best opic model performs better across all costs. Figure 5 tabulates the LC-index obtained by pairwise comparison of the different kinds of topic models against their feature-based counterparts.[7] This suggests that overall, the best tree-based topic model is still reasonably better than its feature-based counterpart. The evidence in favour of topic models with naive Bayes is evident.

[7] The LC-index ranges between -1 and $+1$, and is a (distribution-)weighted summation of the value of a function $L(c_1)$ that takes the value $+1$ when the feature-based model is better than the topic-model being compared against. A positive LC-index indicates the feature-based model performs better (the closer to $+1$, the better it is): see [1] for more details.

Topic Model	**Feature-based Model**	
	Tree-based	Naive Bayes
10-Topic	−0.71	−1.00
25-Topic	−0.42	−1.00
50-Topic	+1.00	−1.00

Fig. 5. A comparison of classifier models constructed using features and topics. The comparison tabulates the LC-index [1] obtained by comparing the performance of a feature-based model against a topic-based model, across a range of costs. A positive value of the index means that the feature-based model performs better than the topic model. A negative value means that the topic model is better. Index values lie between −1 and +1, and the closer the numbers are to +1 or −1 are a quantitative indicator of how much better or worse the feature-based model performs than its topic-model counterpart.

4 Concluding Remarks

In this paper, we have introduced the application of hierarchical Bayesian models—commonly called topic models—that use relational features to tasks in drug-design. Our goals are specifically to assist in the discovery of new antimalarials. To this end, we have proposed 3 kinds of tasks that could benefit from the combined use of topic models and ILP. In the short version of the paper, we have reported results on this simplest of these tasks, namely, one of discriminating amongst molecules with two different qualitative levels of activity. The results obtained are promising: if the costs of missing "hits" is high, then the use of topic-based modelling could perform better than the usual approach adopted in ILP-based propositionalisation (constructing models over a space of features identified by an ILP engine). The use of ILP to construct relational features for molecules retains some important aspects of those systems namely: the use of background knowledge and the benefits of a first-order representation to capture complex structural features. We believe that combining this with hierarchical probabilistic modelling allows us further advantages in terms of discovery of hidden concepts, handling uncertainty, and providing a generative model for molecules. In a full-length version of this paper, we will present results from investigating the use of this form of hybrid modelling for molecule retrieval and synthesis. A shortcoming of the approach should already be evident, namely, the need to specify the appropriate number of topics beforehand. Our results suggest that irrespective of the classification method, even a 10-topic model performs better than using all the features. In practice, we would usually not know this, and we may be forced to use techniques that determine the number of topics automatically. One possibility is the more sophisticated technique of non-parameteric hierarchical Bayesian modelling, in which the number of topics is determined automatically from the data. In this study, we are still able to ask: Are topic-models simply performing a kind of feature selection? A

comparative evaluation of topic models against models obtained with standard feature-selection methods could provide some answers to this question.

Acknowledgements. A.S. holds a Ramanujan Fellowship from the Government of India. He is a Visiting Professor at the Department of Computing Science, University of Oxford; and a Visiting Professorial Fellow at the University of New South Wales, Sydney.

References

1. N.M. Adams and D.J. Hand. Comparing classifiers when misallocation costs are uncertain. *Pattern Recognition*, 32:1139–1147, 1999.
2. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(3/1/2003):993–1022, 2003.
3. N.M. Boyle, M. Banck, C.A. James, C.Morley, T. Vandermeersch, and D.R. Hutchison. Open Babel: an open chemical toolbox. *Chemoinformatics*, 3:33, 2011. Available at: http://www.openbabel.org.
4. Charles Elkan. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.
5. F. Gamo and L.M. Sanz and J.Vidal and C. de Cozar and E.Alvarez and J. Lavandera and D.E. Vanderwall and D.V.S. Green and V.Kumar and S. Hasan and J.R. Brown and C.E. Peishoff and L. R. Cardon and J.F. Garcia-Bustos. Thousands of chemical starting points for antimalarial lead identification. *Nature*, 465(7296):305–310, 2010.
6. Bettina Grun and Kurt Hornik. topicmodels: An R Package for fitting Topic Models. *Journal of Statistical Software*, 40(13):1–30, 2011.
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.
8. R.D. King, S.H. Muggleton, A. Srinivasan, and M.J.E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. of the National Academy of Sciences*, 93:438–442, 1996.
9. R.D. King and A. Srinivasan. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives*, 104(5):1031–1040, 1996.
10. P.M. O'Neill and V.E. Barton and S.A. Ward. The Molecular Mechanism of Action of Artemisinin—The Debate Continues. *Molecules*, 15:1705–1721, 2010.
11. S.Kramer, N. Lavrac, and P. Flach. Propositionalization approaches to relational data mining. In *Relational Data Mining*, pages 262–286. Springer-Verlag, New York, 2001.
12. A. Srinivasan. The Aleph Manual. Available at http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/, 1999.
13. A. Srinivasan and R.D. King. Feature construction with Inductive Logic Programming: a study of quantitative predictions of biological activity aided by structural attributes. In S.H. Muggleton, editor, *Proceedings of the Sixth Inductive Logic Programming Workshop*, volume 1314 of *LNAI*, pages 89–104. Springer-Verlag, 1996.
14. C. Taranto, N. Di Mauro, and F. Esposito. rsLDA: A Bayesian heirarchical model for relational learning. In *ICDKE*, pages 68–74, 2011.

15. WHO. Global plan for artemisinin resistance containment (GPARC). Available at http://www.who.int/malaria/publications/atoz/ 9789241500838/en/index.html, 2011.
16. WHO. World Malaria Report 2011. Available at http://www.who.int/malaria/world_malaria_report_2011/en/, 2011.