Large-scale Drug Function Prediction by Integrating QIS D^2 and BioSpice

Ying Zhao, Charles Zhou, Ian Oglesby, Cliff Zhou
Quantum Intelligence, Inc.
3375 Scott Blvd Suite 100
Santa Clara CA 95054
Tel: 408-980-0090
charleszhou@alum.mit.edu

Abstract

Quantum Intelligence System for Drug Discovery (QIS D^2) is a unique adaptive learning system designed to predict potential large-scale drug characteristics such as toxicity and efficacy. BioSpice is a set of software tools designed to represent and simulate cellular processes funded by DARPA. We show a QIS D^2 model is successfully trained, tested and validated on experimental data sets for predicting the potential in vivo *effects of drug molecules in biological systems. QIS D^2 is interoperable with BioSpice. The workflow and visualization are built-in capabilities for easy-of-use. The integration of QIS D^2 and BioSpice draw on diversified technologies to deliver unique benefits for simulation and screening of potential drugs and their targets. We show that our approach leverages both structured and unstructured bioinformatics databases such as BioWarehouse and GeneWays in BioSpice to greatly enhance a QIS D^2 model. We show QIS D^2 models data from seven sources for 37,330 chemicals, performs an automatic sequence clustering using 1234 structure fragments, and accurately predict 1829 targets simultaneously.*

Keywords

Adaptive learning, data mining, text mining, large-scale prediction, drug discovery, efficacy, toxicity, *in silico* screening

1. Introduction

QIS D^2 (Quantum Intelligence System for Drug Discovery) is designed to predict large-scale drug characteristics such as toxicity and efficacy. BioSpice is a bioinformatics tool set produced by the DARPA biology community. The objective of this paper is to show how to use both tools in conjunction with publicly available data to demonstrate an innovative *in silico* screening procedure for predicting large-scale biochemical functions of potential drugs in various contexts.

Given known information of a chemical, for example, its 3D structure, it is important to predict its biochemical functions in various contexts. For example, in a drug discovery context, the predictive targets are often the specific defined therapeutic efficacies, as well as so-called drug-able or drug-like properties such as pharmacologic absorption, distribution, metabolism, excretion, toxicity (ADME/tox) properties. The prediction of efficacy and ADME/tox has been intensively studied in the area of bioinformatics for drug discovery, of particular, via the methodology of Structure Activity Relationship (SAR) modeling[1].

More recently, predicting a chemical's associations with biological targets, more interestingly, with molecular targets like proteins and genes, has become increasingly important for variety of applications. A drug-molecular target association can be either an interaction in a traditional biochemical sense or a relationship measured in a much broader way. For example, how often a drug and a protein might be mentioned in the same context (sentence, paragraph or article) across a large collection of publications can be defined a measure of association between the two. Obviously, understanding and predicting such associations can be very useful for screening innovative therapeutic agents for drug discovery. For environmental protection, predicting drug-molecular target associations can help understand long time toxic effects of chemicals to human and environment. For bio-threat countermeasures discovery, such predictive targets can be used for screening countermeasures against bio-threat like bacteria or virus.

Further more, the molecular targets can be grouped according to their functions and positions in a biochemical pathway. To predicting the cascading interactions and associations of a chemical along a pathway is also important for many applications.

One of the challenges for bioinformatics is to accurately screening the potentials for a large number of predictive targets. QIS D^2 is a unique adaptive learning system which is a capable of large scale association discovery and predictions. The focus of this paper is to apply QIS D^2 to predict a large number and variety of targets including traditional efficacy, ADME/tox properties as well as drug-molecular targets associations and drug-pathway associations. We also show how to integrate QIS D^2 with BioSpice.

2. QIS D^2 Methodology Overview

The current version of QIS D^2 is based on a previous version of QIS for predicting drug toxicity and efficacy funded by DARPA. QIS D^2 is an adaptive learning system, in other words, a QIS D^2 model can be successfully trained, tested and validated on experimental data sets for predicting the potential *in vitro* or *in vivo* effects of drug molecules in biological systems, with the predictive targets defined as interested for different applications. For the purpose, we first divide the samples of experimental data sets into the

- Training set: will be used to discover statistical predictive rules, patterns and associations between known information and targets
- Test set: will be held-out initially and used to validate the rules, patterns and associations

QIS D^2 is interoperable with DARPA BioSpice. The workflow and visualization are built-in capabilities for easy-of-use biochemical professionals as shown Figure 1.

3. Data Model

3.1 Chemical Structures

Structures of drugs in the training set and test set will be coded into structure fragment sequences using the QIS D^2 proprietary method. The method is similar to other chemical fragment based models used for structure-activity relationship (SAR) modeling [1] except the fragments are ordered in sequence. The input to the coding method is a Corina[2] formatted drug structure. Each fragment is a structural descriptor. The number of fragments of the method is in the order of thousands.

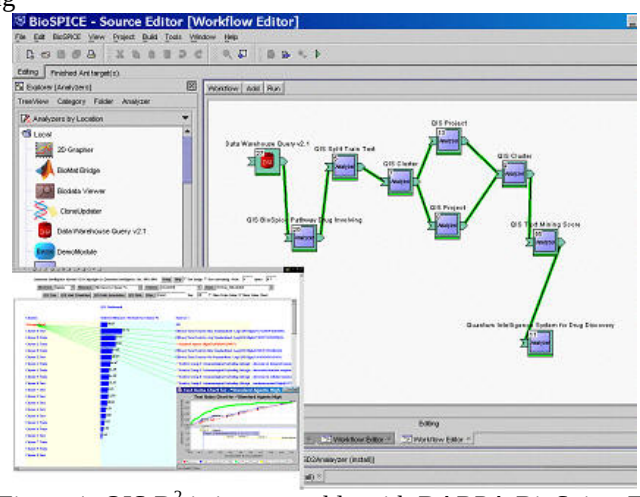

Figure 1: QIS D^2 is interoperable with DARPA BioSpice. The workflow and visualization are built-in capabilities for easy-of-use

3.2 Efficacy

We use the NCI anti-cancer databases, which include the evidence of anti-cancer efficacy measures for 41,000 compounds, which are the concentrations required to inhibit growth by 50% for 60 cancer cell lines. It is the only public database that contains data of a large number of molecular target (gene and protein) profiles against a large set of compounds. We have selected 208 proteins and 1000 genes as anti-cancer molecular targets in this paper.

3.3 Toxicity

We use the toxicity data from Registry of Toxic Effects of Chemical Substances (RTECS). About 500 toxic unique effects across a wide range of categories, including primary irritation, mutagenic effects, reproductive effects, and tumorigenic effects, have been collected by The National Institute for Occupational Safety and Health (NIOSH) from 70s' for 150,000 chemicals.

3.4 Associations with Molecular Targets

The molecular targets considered in the paper include proteins and genes. We have selected 208 molecular targets and 1000 genes from the NCI anti-cancer databases. The association of a drug with a

molecular target is defined as their correlation along the 60 cancer lines[3].

3.4.1 BioSpice BioWarehouse. BioWarehouse in BioSpice has extracted, transformed and loaded (ETL) the contents of popular biomedical databases such as KEGG [4], HumanCyc [5] and GenBank [6] using a uniform schema. Therefore, it is very convenient for BioSpice interoperable tools to use these databases with little work on ETL. We have focused pathway related databases KEGG and HumanCyc in BioWarehouse for pathway scoring in QIS D^2

3.4.2 BioSpice GeneWays. BioSpice GeneWays [7] is a bio-text mining tool which is capable of extracting specific gene and gene interaction from a large collection of literature. For example, given a gene name, tnf-receptor, it generates "tnf-receptor phosphorylate glo1". GeneWays is based natural language rules to extract the gene and gene interaction from publications, which can be viewed as discovering associations in a logic level. We use the tool to extract gene-gene associations and then use them to induce new pathways that are not previously directly discovered from the experimental data.

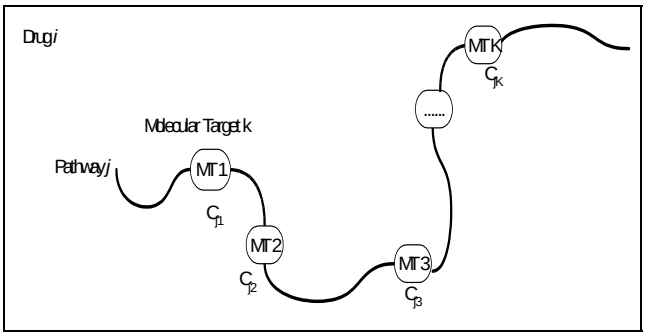

Figure 2. Illustration of Pathway *j* for Drug *i*, where Drug *i* is correlated with molecular target labeled as "MT *k*" with the correlation value as c*ik* where k=1, 2, …, K. The drug-pathway association is the sum of the absolute value of the correlations.

3.4.3 Pathway Scoring. This is a unique QIS D^2 capability that takes the output of the drug-molecular targets correlation from the above and adds up the absolute value of the drug-molecular target correlation along a pathway found by BioWarehouse and GeneWays. **Error! Reference source not found.** illustrates Pathway *j* for Drug *i* and molecular targets k=*1,2,…,K* found along a pathway. Suppose that Pathway *j* contains *K* molecular targets labeled as MT 1, MT 2, MT 3, …and MT K, and that the correlation value for each *Molecular Target* k with *Drug i* is c_{ik}, where *k=1,2,…,K.* The association between Drug *i* and Pathway *j* can be computed as follows:

$$\textit{Association between Drug i and Pathway } j = \sum_{k=1}^{K} \left\| c_{ik} \right\|$$

In a summary, Table 1 shows the total number of features after the integration of data described above. The structure sequence information is used as the input for predictive modeling, all the others are used predictive targets.

Table 1: The features in the QIS data model

Feature Dimensions	# of features
Structures (QIS)	1234
Toxicity (QIS)	~500
Molecular Targets (QIS)	208 proteins 1000 genes
Pathway Scores (Biowarehouse – HumanCyc & KEGG)	57
Pathway Scores (GeneWays)	3

4. Large Scale Prediction

QIS D^2 is designed to simultaneously predict large scale posterior probabilities based on a tied-mixture EM [11, 12] method. EM stands for Expectation and Maximization. It is a statistical method used to compute maximum likelihood estimates given incomplete samples[13]. Comparing to traditional statistical regression, pattern recognition and data mining algorithms such as logistic regression, decision trees or neural networks [14] for prediction, this method is especially capable of predicting large scale targets.

Using QIS D^2 sequence clustering, chemicals are first clustered into characteristic groups based on their known structures. Each group is quantitatively represented as a fingerprint *o (o=1,…,K)*. A drug is then measured for how much it assembles the fingerprints.

QIS D^2 also cluster predictive targets into characteristic groups group *j (j=1,…,M)* based on the similarity of the targets' properties. Each group also represents a target fingerprint or a target class. A new target will be compared quantitatively with the fingerprints and then classified into the right class where it belongs to.

For new and unseen chemical, QIS D2 predicts a target based on the statistical patterns discovered from the training and validation process. In other words, the

goal is to predict if a drug cluster o is associated with a target class j, defined as P(o|j)/P(o), where P(o|j) is the probability of a potential drug classified to cluster o given that it associates with a target class j; P(o) is the probability of a potential drug classified to cluster o in the population as a whole. This statistical association measures how tight an input class o is associated with a target class j. Each P(o|j)/P(o)(where o=1,…,K and j=1,…,M) could represents a mode of input-target association. The tied-mixture EM in QIS D2 is to accurately compute the association between an actual drug x and a predictive target y P(x|y)/P(x) by tying all of the modes P(o|j)/P(o) (where o=1,…,K and j=1,…,M) together.

5. Results

We have used the system to predict the efficacy, toxicity, association with molecular targets and pathways of 146 anti-cancer standard agents, which are either already FDA approved chemotherapy drugs or in the pipeline. Our system can successfully screen 30% of the candidates containing 80% of standard agents. The resultant system is in the process of being integrated in a DOD application of search for biothreat countermeasures.

6. Conclusions

In this paper, we show QIS D^2 is an adaptive learning system. A QIS D^2 model is successfully trained, tested and validated on experimental data sets for predicting the potential large-scale *in vivo* effects of drug molecules in biological systems. QIS D^2 is interoperable with DARPA BioSpice. The integration of QIS D^2 and BioSpice draw on diversified technologies to deliver unique benefits for simulation and screening of potential drugs and their targets with innovative use of structured and unstructured bioinformatics databases. We show a QIS D^2 models diversified data from seven sources for 37,330 chemicals, performs an automatic sequence clustering using 1234 structure fragments, and accurately predict 1829 targets simultaneously.

Acknowledgements

The project is partially supported by DARPA contract # W31P4Q-04-C-R197. We are very grateful for the collaboration and support of the BioSpice community (https://community.biospice.org/), especial thanks to Dr. Peter Karp, Mr. Tom Lee from Stanford Research for extracting data from BioWarehouse and Dr. Andrey Rzhetsky from Columbia University for extracting data from GeneWays server.

References

[1] Hert, J., et al., *Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.* J Chem Inf Comput Sci, 2004. 44(3): p. 1177-85.

[2] Miller, M.A., *Chemical database techniques in drug discovery.* Nature Reviews Drug Discovery, 2002(1): p. 220-227.

[3] Scherf, U., et al., *A gene expression database for the molecular pharmacology of cancer*. Nat Genet, 2000. 24(3): p. 236-44.

[4] Kanehisa, M., *A database for post-genome analysis.* Trends Genet., 1997. 13: p. 375-376.

[5] Romero, P., et al., *Computational prediction of human metabolic pathways from the complete human genome.* Genome Biology, 2004. 6(R2): p. 1-17.

[6] Benson, D.A., et al., *GenBank: update.* Nucleic Acids Res, 2004. 32(Database issue): p. D23-6.

[7] Rzhetsky, A., W.J. Wilbur, and M. Morris, *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.* Journal of Biomedical Informatics, 2004. 37(1): p. 43 - 53.

[8] van Rijsbergen, C.J., *Information Retrieval*. 1979, London: Butterworths.

[9] Letsche T.A., B., M. W., *Large-scale information retrieval ith latent semantic indexing.* Information Sciences, 1997. 100(1-4): p. 105-137.

[10] Dumais S. T., F.G.W., Landauer T. K., Deerwester, S. *Using latent semantic analysis to improve information retrieval.* in *In Proceedings of CHI'88: Conference on Human Factors in Computing*. 1988. New York: ACM.

[11] Huang, X.D., Y. Ariki, and M.A. Jack, *Hidden Markov Models for Speech Recognition*. 1990, Edinburgh: Edinburgh University Press.

[12] Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis.* 1988: Prentice Hall, 1.

[13] Dempster, A., N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, 1977. Series B, 39(1): p. 1-38.

[14] Ripley, *Pattern Recognition and Neural Networks.* 1996: Cambridge University Press.