News Article

iSMART: An Integrated Cloud Computing Web Server for Traditional Chinese Medicine for Online Virtual Screening, *de novo* Evolution and Drug Design

Introduction

Traditional medicines and herbal medicines, while having different approaches from conventional western medicine, have been recognized worldwide for their therapeutic values. The 2003 World Health Assembly (WHA) has enlisted traditional medicine as one of the major health care issues. The World Health Organization's (WHO) 2002-2005 Traditional Medicine Strategy reported that the traditional medicine industry worldwide has a net value of over $60 billion US dollars, and was still increasing. Particularly, Traditional Chinese medicine (TCM), which is a vast knowledge bank of 5000-year medical practices, played an important role among traditional medicine. In Eastern World, TCM never ceased its popularity and influence among the society. Recent global TCM and herbal medicine have market values of approximately $20 billion in US dollar, have annual growth rate of 10-20%. Moreover, TCM user population is also increasing at 20% annual growth rate. In addition, the "WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region" has published standard translation of over 3,000 TCM terminologies for medical practice and scientific researches. Despite the increased documented importance of TCM, studies in TCM drug discovery remained a minor part of pharmaceutical researches.

With the progress of computer and information technology, recent development of pharmacology has adapted the *in silico* analysis, such as virtual screening, computer-aided drug design, and molecular dynamics, as critical leading steps for drug discovery. This *in silico* protocol has also been employed for drug design (1-13). However, there was no strong evidence for significant correlation between TCM and systems biology. The concept of bridging TCM and systems biology has been proposed as early as year 2005. Although this concept could be beneficial to the development of TCM, many issues were raised while joining the two medical systems. Nevertheless researchers, predominantly in the Eastern World, have found correlations between TCM studies to systems biology in certain medical aspects (14-16). The increasing popularity of TCM studies is becoming an unneglectible trend worldwide. For the globalization of TCM studies and user-friendly online analysis server, we introduced a multi-purpose, cloud computing-based web server, the "**i**ntegrated **S**yste**M**s Biology **A**ssociated **R**esearch with **T**CM" (iSMART) (Figure 1).

iSMART was developed for online TCM analysis focused on drug design. This server system was further implemented with genetic research features aimed to

Kai-Wei Chang[1]
Tsung-Ying Tsai[1]
Kuan-Chung Chen[1]
Shun-Chieh Yang[1]
Hung-Jin Huang[1]
Tung-Ti Chang[1]
Mao-Feng Sun[1,2]
Hsin-Yi Chen[3]
Fuu-Jen Tsai[2,3]
Calvin Yu-Chian Chen[1,3,5,6*]

[1]Laboratory of Computational and Systems Biology, School of Chinese Medicine, China Medical University, Taichung, 40402, Taiwan

[2] Department of Acupuncture, China Medical University Hospital, Taichung, Taiwan

[3]Department of Bioinformatics, Asia University, Taichung 41354, Taiwan

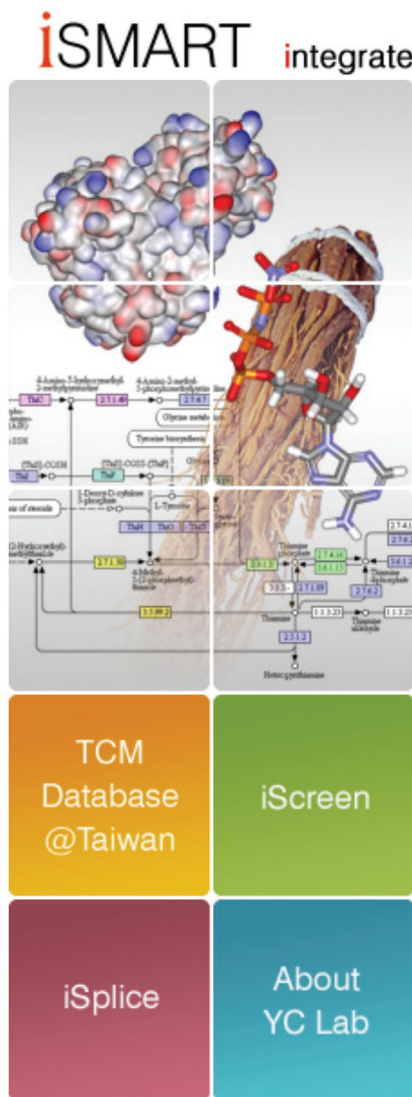[4]Department of Medical Genetics, China Medical University, Taichung 40402, Taiwan

[5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[6]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*Phone: +1-617-353-7123
E-mail: ycc@mail.cmu.edu.tw
       ycc929@MIT.EDU

Key words: AIDS: Traditional Chinese medicine (TCM); Cloud-computing; Docking; Screening; Systems biology.

243

identify erroneous splicing events. Functions with regard to the TCM and disease analysis in systems biology aspects, including items such as disease-pathway network and TCM drug network would be brought online in near future. We aim iSMART to become a comprehensive web system that bridges TCM and computer-based drug design to systems biology studies. iSMART is freely available and can be accessed through http://iSMART.cmu.edu.tw/.

iSMART integrated SysteMs Biology Associated Research with TCM

iSMART is a multi-funcitonal cloud-computing web system that comprised with three components:
(1) TCM Database@Taiwan, (2) iScreen, and (3) iSplice.

TCM Database@Taiwan
The TCM database is currently the world largest and most comprehensive free down small molecular database on traditional Chinese medicine for virtual screening.

iScreen
To take initiative in the next generation of drug development, we constructed a docking and screening web-server, iScreen, based on our world's largest TCM database, the TCM Database@Taiwan.

iSplice
iSplice is an online tool combining gapped-dinucleotide pattern probability with logarithmic odds (GO algorithm) to assess the likelihood of the activation of a cryptic 5' splice site, which competes with its paired authentic 5' splice site.

Future Works
The framework of iSMART system has been built and relevant development has been conducted to fabricate multiple analysis components prior the pre-clinical trial. More comprehensive analysis would be conducted with the establishment of complementary functions in iSMART.

Figure 1: The title page of iSMART. iSMART is built as a pioneering system for realizing the concept of bridging TCM studies to systems biology researches.

TCM Database@Taiwan

Following the description of the in-house developed TCM Database@Taiwan (17), this database maintains and organizes over 20,000 compounds originated from TCM substances. The source, basic molecular properties, 3D compound structure, and references were recorded. Relevant search and browse modules were implemented.

iScreen

iScreen is an online virtual screening web server focused on identifying compatible TCM compounds as well as their derivatives to the target protein. This web server was implemented with multiple molecular docking algorithms from the PLANTS

docking package (18) and the *de novo* compound evolution function from LEA3D ligand-design software (19). In addition, iScreen was also able to perform protein preparation services to utilize the molecular docking process.

iSplice

iSplice is a web server designed to identify the activation and the strength of cryptic 5' splice site on that may lead to disease development. An input DNA sequence would be evaluated by novel gapped-dinucleotide pattern probability and logarithmic odds algorithm (GOA), which have been validated within our research team. A total of 189, 249 5' splice sites from human genome (20) were employed for training the prediction model implemented in iSplice.

TCM Drug Network

The TCM drug network identified pair-wise TCM drug relationships based on their association in metabolic pathways. Drugs under the same TCM classification were grouped for cluster analysis. TCM linked Gene Ontology (GO) analysis (21), Kyoto Encyclopedia of Genes and Genomes (KEGG) networks (22), and Online Mendelian Inheritance in Man (OMIM) database (23) were employed for data mining and network construction. The drug-drug correlation is calculated based on the traditional mutual information (*MI*) entropy (24), which was defined as equation [1] for drug *x* and drug *y*.

$$MI\left(x,y\right) = P\left(x,y\right) \cdot \log\left(\frac{P\left(x,y\right)}{P\left(x\right) \cdot P\left(y\right)}\right) \qquad [1]$$

The average distance (*d*) between TCM drugs was evaluated by equation [2] for drug *x* and drug *y*.

$$d\left(x,y\right) = \frac{\sum I(x,y,i) \cdot d(x,y,i)}{\sum I(x,y,i)} \qquad [2]$$

A scoring system (*sco*) is hence built for evaluating the relationships between a pair of TCM drugs by equation [3] in which larger value represents stronger correlation, and vice versa.

$$sco\left(x,y\right) = \frac{MI(x,y)}{d(x,y)} \qquad [3]$$

What Does iSMART Offer?

iSMART is an integrated multi-functional cloud computing web server aimed for performing comprehensive analysis for TCM-associated studies. Present system is comprised with three major components: (1) TCM Database@Taiwan, (2) iScreen, and (3) iSplice.

TCM Database@Taiwan offers both 2D and 3D structures of small molecules from TCM (17) (Figure 2). The TCM sources are arranged based on the TCM classification. Relevant molecular properties and references were documented for all compounds. In addition, the database is incorporated with advanced search and browsing functions that can extract data based on compound name, origin, molecular properties, and partial 3D structures. TCM Database@Taiwan is continuously expanding toward a more comprehensive database that comprised not only TCM compounds, but also compounds from natural sources. In addition, TCM Database@Taiwan is the core database component for relevant online components such as iScreen and the underdevelopment TCM-linked networks.

For utilizing the uses of the TCM database, we implemented iScreen for online TCM-based drug design (Figure 3). The web server is able to identify sets of TCM compounds that fit into the proteins of interest. For utilizing the molecular docking procedure, iScreen provides protein preparation service that cleans up additional components from the original protein data. For customized molecular docking and screening, iScreen simulates multiple docking conditions, including standard, with simulated water, with specific pH condition, and with flexible protein residues. A set of TCM compounds could be identified, relevant docking scores
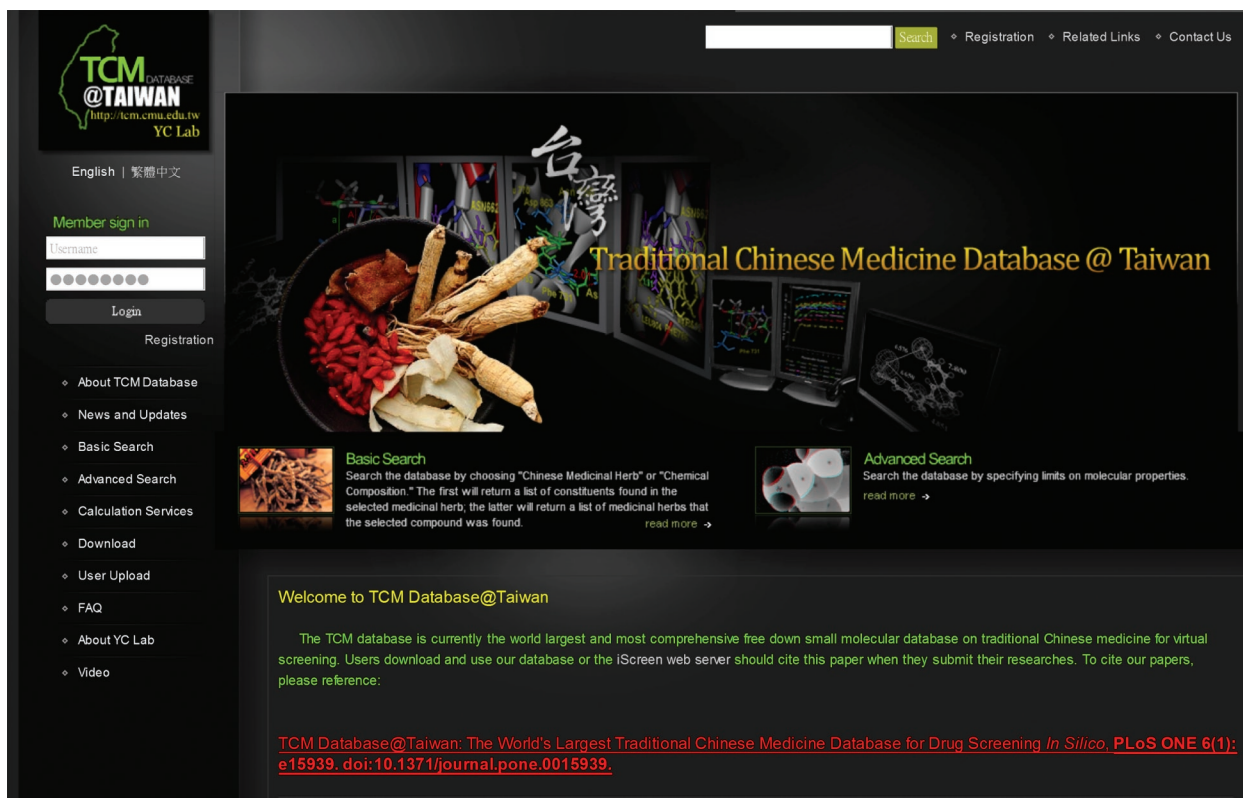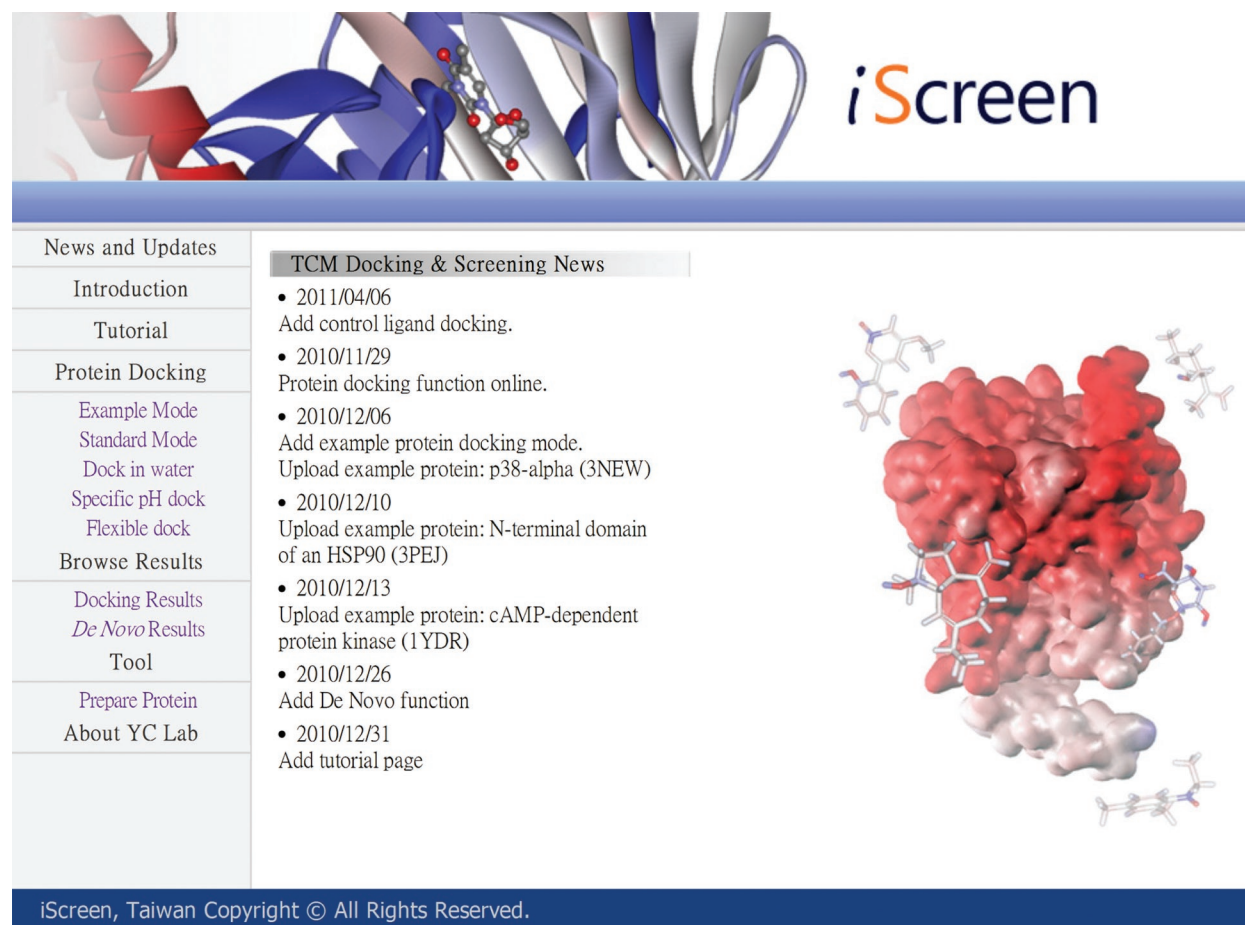
Figure 2: TCM Database@Taiwan, the world's largest traditional Chinese medicine database for drug screening *in silico*. This database maintains over 20,000 TCM compounds based on TCM classification. Simple molecular properties and molecular structure of each compound were recorded and ready for search and browse.

could be determined and downloaded for further analysis. In addition, iScreen also allows advanced users to define the TCM compound derivatives through *de novo* evolution. The compatibilities of the TCM derivatives to the targets would be evaluated. iScreen is a direct application of the in-house TCM Database. This web server provides an overview of potential TCM functions in molecular level. The identified compounds as well as their TCM origins could be tested for the therapeutic values.

Nevertheless, iSMART is not limited to TCM based drug design and relevant researches directly from the given protein targets. For the identification of potential disease-causing targets, iSplice was developed to evaluate genetic components that may lead to diseases. More specifically, iSplice identifies the activation and the activity strength of cryptic 5' splice sites due to mutation (Figure 4). Follow-up evaluation of the results from iSplice could lead to the identification of disease-causing targets due to genetic aberration. iSplice is also one of the components that may lead the TCM studies toward genetic components and regulations.

iSMART is not limited to the three components described above, this system is also expandable for adapting addition components and creating associative links for providing more comprehensive analysis services. The relationships between the TCM components have been established based on the TCM drugs networks. For example, Figure 5 illustrates a relationship network between yang-tonifying TCM substances. The association strengths between TCM substances could be determined

Figure 3: iScreen, the world's first cloud computing web server for TCM drug screening and De Novo design. For molecular screening/docking, iScreen provides 4 docking modes: standard, in water, in specified pH, and with flexible ligands, as well as 1 demonstration docking module. The compounds from the screening results can be further derived based on the De Novo drug design function. Both docking results and De Novo results are visible in the queuing system. iScreen is also implemented with protein preparation tool to reduce errors caused by excessive information in the raw input data.

in pairwise analysis based on such a network. Similar protocol would be further applied to build the relationship networks between TCM formulae. With more functions established in the foreseeable future, iSMART would be able to establish networks between diseases, signaling pathways, and metabolic pathways. Associative TCM compounds could hence be identified accordingly.

Conclusion

iSMART is developed in the hope to become a central hub for TCM studies and systems biology research. Although this web-based system is at its infant stage, its components, TCM Database@Taiwan, iScreen, and iSplice, have been established and used by researchers interested in TCM studies. The TCM Database@Taiwan is the central component for storing data of TCM small molecule. This database has been adapted by iScreen as well as other network system components currently not established. The iScreen is a user-friendly web server for TCM-based

molecular screening that could increase the efficiency in TCM-related drug design. The iSplice server is the first iSMART component for genetic analysis. At present, iSMART is a multi-purpose cloud computing web system for TCM-associated analysis.

Future Development

The framework of iSMART system has been built and relevant development (Figure 6) has been conducted. We define iSMART as a TCM-based online resource that fabricate multiple analysis components prior the pre-clinical trial in a standard drug development procedure. With the establishment of the comple-

Introduction
Search
Analyze
Submit
Contact
Help

iSplice

Although many splice sites have been identified by computational and in vitro splicing techniques, a comprehensive analysis tool for predicting the cryptic 5' splice site activation in gene mutation was not available. iSplice is an online tool combining gapped-dinucleotide pattern probability with logarithmic odds (GO algorithm) to assess the likelihood of the activation of a cryptic 5' splice site, which competes with its paired authentic 5' splice site. This study showed that iSplice has better prediction rate than two other scoring methods, namely information contents (Ri) and free energy (Δ G). iSplice was a web-based tool that could calculate the possibility of various splicing events in designed mutagenesis experiments.

Figure 4: iSplice, a web server for predicting the activation of cryptic 5' splice site. It is a unique component in iSMART for a specific class of genetic analysis. Users can identify both authentic and cryptic 5' splice sites, and their relevant activation strength based on the input gene sequence and point of mutation.
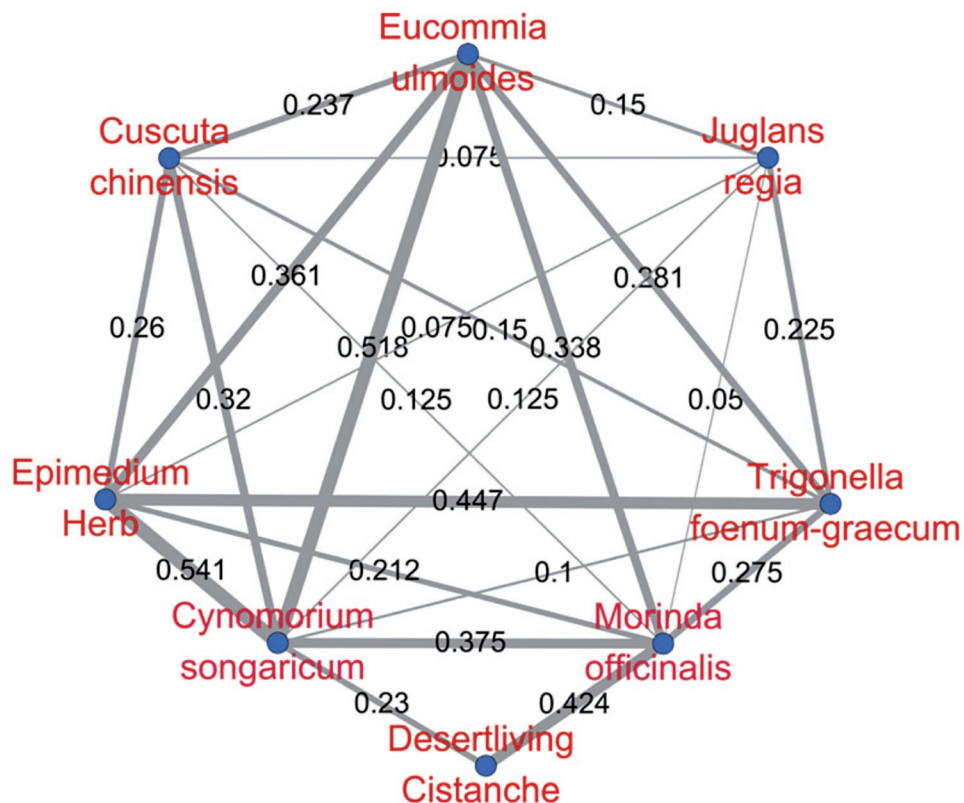

Figure 5: Correlation network of 8 TCM drugs categorized under yang-tonifying class. 8 TCM drugs were included in this group. Each line represents the correlation between a pair of TCM drugs. Line width and the attached value indicated the correlation strength between a pair of TCM drugs.

mentary functions as indicated in the framework diagram, more comprehensive analysis could be conducted through iSMART. iSMART is built as a pioneering system for realizing the concept of bridging TCM studies to systems biology researches. The development of iSMART could be a milestone for the globalization of TCM development.

Acknowledgements

integrated SysteMs Biology Associated Research with TCM
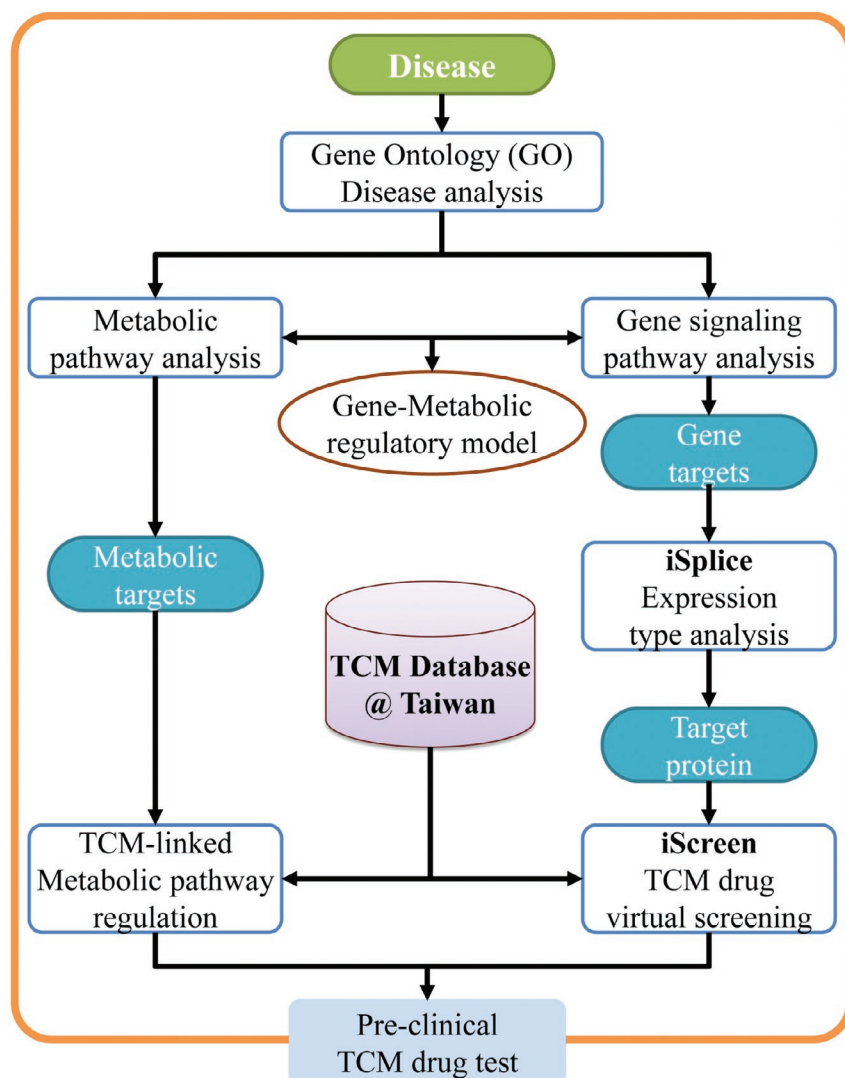
Figure 6: The schematic framework of iSMART's component relationships. Systems biology analysis functions were implicitly defined based on the establishment of the analysis components. For analysis, a disease would be categorized and its associative signaling and metabolic pathways would be defined accordingly. The genetic defects, such as cryptic 5' splice site activation, would be identified, and relevant targets would be extracted for drug design. On the other hand, the therapeutic targets along the metabolic pathway would be identified accordingly. iScreen would be applied to identify potential TCM compounds from TCM Database@Taiwan against the designated targets. The analysis results and their medical relevance would be validated through pre-clinical tests.

References

1. C. Y.-C. C. Kuan-Chung Chen. *Soft Matter 7*, 4001-4008 (2011).
2. A. M. Andrianov. *J Biomol Struct Dyn 26*, 445-454 (2009).
3. A. M. Andrianov and I. V. Anishchenko. *J Biomol Struct Dyn 27*, 179-193 (2009).
4. M. T. Cambria, D. Di Marino, M. Falconi, S. Garavaglia, and A. Cambria. *J Biomol Struct Dyn 27*, 501-509 (2010).
5. C. Y. Chen. *J Biomol Struct Dyn 27*, 627-640 (2010).
6. C. Y. Chen, Y. H. Chang, D. T. Bau, H. J. Huang, F. J. Tsai, C. H. Tsai, and C. Y. C. Chen. *J Biomol Struct Dyn 27*, 171-178 (2009).
7. E. F. F. da Cunha, E. F. Barbosa, A. A. Oliveira, and T. C. Ramalho. *J Biomol Struct Dyn 27*, 619-625 (2010).
8. L. I. D. Hage-Melim, C. H. T. D. da Silva, E. P. Semighini, C. A. Taft, and S. V. Sampaio. *J Biomol Struct Dyn 27*, 27-35 (2009).
9. H. J. Huang, K. J. Lee, H. W. Yu, C. Y. Chen, C. H. Hsu, H. Y. Chen, F. J. Tsai, and C. Y. C. Chen. *J Biomol Struct Dyn 28*, 23-37 (2010).
10. H. J. Huang, K. J. Lee, H. W. Yu, H. Y. Chen, F. J. Tsai, and C. Y. Chen. *J Biomol Struct Dyn 28*, 187-200 (2010).
11. A. K. Kahlon, S. Roy, and A. Sharma. *J Biomol Struct Dyn 28*, 201-210 (2010).
12. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
13. T. T. Chang, H. J. Huang, K. J. Lee, H. W. Yu, H. Y. Chen, F. J. Tsai, M. F. Sun, and C. Y. Chen. *J Biomol Struct Dyn 28*, 309-321 (2010).
14. M. Wang, R. J. Lamers, H. A. Korthout, J. H. van Nesselrooij, R. F. Witkamp, R. van der Heijden, P. J. Voshol, L. M. Havekes, R. Verpoorte, and J. van der Greef. *Phytother Res 19*, 173-182 (2005).
15. P. Li and L. P. Yang. *Journal of Chinese integrative medicine 6*, 454-457 (2008).
16. A. B. Burgin, O. T. Magnusson, J. Singh, P. Witte, B. L. Staker, J. M. Bjornsson, M. Thorsteinsdottir, S. Hrafnsdottir, T. Hagen, A. S. Kiselyov, L. J. Stewart, and M. E. Gurney. *Nat Biotechnol 28*, 63-70 (2010).
17. C. Y.-C. Chen. *PLoS ONE 6*, e15939 (2011).
18. O. Korb, T. Stutzle, and T. E. Exner. *J Chem Inf Model 49*, 84-96 (2009).
19. D. Douguet, H. Munier-Lehmann, G. Labesse, and S. Pochet. *J Med Chem 48*, 2457-2468 (2005).
20. K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang, Y. Takeshima, M. Matsuo, G. Sobue, and K. Ohno. *Nucleic Acids Res 35*, 5995-6003 (2007).
21. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. *Nat Genet 25*, 25-29 (2000).
22. M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. *Nucleic Acids Res 38*, D355-360 (2010).
23. V. A. McKusick. *Am J Hum Genet 80*, 588-604 (2007).
24. D. Beeferman, A. Berger, and J. Lafferty. *Mach Learn 34*, 177-210 (1999).

Date Received: March 26, 2010

Communicated by the Editor Ramaswamy H. Sarma