ANTI-INFECTIOUS DRUG REPURPOSING USING AN INTEGRATED CHEMICAL GENOMICS AND STRUCTURAL SYSTEMS BIOLOGY APPROACH

CLARA NG

Department of Computer Science, Hunter College, the City University of New York,
695 Park Avenue, New York City, NY 10065, U. S. A.
Email: cng0003@hunter.cuny.edu

RUTH HAUPTMAN

Department of Computer Science, Hunter College, the City University of New York,
695 Park Avenue, New York City, NY 10065, U. S. A.
Email: rhauptma@hunter.cuny.edu

YINLIANG ZHANG

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, U. S. A.
Email: yiz071@ucsd.edu

PHILIP E. BOURNE

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, U. S. A
Email: pbourne@ucsd.edu

LEI XIE

Department of Computer Science, Hunter College, The Graduate Center, the City University of New York
695 Park Avenue, New York City, NY 10065, U. S. A.
Email: lei.xie@hunter.cuny.edu

The emergence of multi-drug and extensive drug resistance of microbes to antibiotics poses a great threat to human health. Although drug repurposing is a promising solution for accelerating the drug development process, its application to anti-infectious drug discovery is limited by the scope of existing phenotype-, ligand-, or target-based methods. In this paper we introduce a new computational strategy to determine the genome-wide molecular targets of bioactive compounds in both human and bacterial genomes. Our method is based on the use of a novel algorithm, ligand Enrichment of Network Topological Similarity (ligENTS), to map the chemical universe to its global pharmacological space. ligENTS outperforms the state-of-the-art algorithms in identifying novel drug-target relationships. Furthermore, we integrate ligENTS with our structural systems biology platform to identify drug repurposing opportunities via target similarity profiling. Using this integrated strategy, we have identified novel *P. falciparum* targets of drug-like active compounds from the Malaria Box, and suggest that a number of approved drugs may be active against malaria. This study demonstrates the potential of an integrative chemical genomics and structural systems biology approach to drug repurposing.

1. Introduction

Treatment of infectious diseases is under threat. The emergence of multi-drug resistance and extensively drug resistant microbes to antibiotics calls for new treatment regimes.[1] Yet, at the same time, the drug discovery process, characterized by a one-drug-one-gene-one-disease paradigm, has yielded few successes in combating drug resistance and is hampered by a high failure rate leading to soaring costs.[2] Fortunately, the cause of that failure is also cause for optimism. Since the failure is due, in part, to drug promiscuity there is also the opportunity to repurpose existing drugs to treat infectious diseases.[3] However, there are several unique challenges in anti-infectious drug repurposing. First, successful phenotype-based methods which compare the genome-wide molecular signature of repositioned drugs to a disease-induced phenotype,[4] have limitations when applied to anti-infectious drug discovery. Second, recent efforts in cell-based antibiotics screening produce thousands of active compounds, but gives few hints as to their molecular targets as well as their *in vivo* activities and toxicities.[5-6] Finally, due to the bias in high-throughput screening, existing chemical genomics databases only collect several thousand targets, most of which are from human and model organisms, not pathogens. Taken together these limitations hinder the application of state-of-the-art computational methods to anti-infectious drug repurposing.

These limitations can be addressed through *chemical genomics* - the construction of genome-scale drug-target interaction networks. Creating such networks requires that we address the question, given a chemical entity, how do we accurately identify its targets on a genome scale based on its structural similarity with known ligands and reliably determine the significance of those putative targets? Several data mining techniques have recently been developed to predict drug-target interactions.[7-15] However, few of them can assess the statistical significance of ranked targets. A notable advance was the development of Similarity Ensemble Approach (SEA) statistical model,[16-17] which is comparable to the state-of-the-art machine learning algorithms.[18] However, SEA and most of the existing machine learning techniques only consider local neighborhoods for relevance between chemicals.[19] Thus it remains a big challenge to find the global relationships between chemicals so that an expanded target space can be established.[20-27] In this paper, we introduce a fundamentally new methodology, ligand Enrichment of Network Topological Similarity (ligENTS), which integrates graph mining algorithms and random set theory to begin to address the above challenges. ligENTS considerably improves the performance of existing methods for drug-target prediction. Thus, ligENTS may open new doors to the next generation of chemical genomics algorithms.

The integration of chemical genomics and structural genomics is needed since current chemical genomics methods have only identified targets for a small portion of the human (<10%) and pathogen genomes (often <1%), respectively.[28] In other words the molecular targets of a large number of active compounds against bacteria are still unknown. Complementary to the knowledge of existing drug targets, the structural information of proteins has increased rapidly.[29] Previously, starting from a known drug-target, we have developed a *structural systems biology* approach for linking drug molecules to pathogen structural genomes through target binding site similarity, thereby reconstructing high-resolution 3D drug-target physical interaction models.[30] However, these structural systems biology methods are not scalable to millions of chemicals. To address these limitations, we combine ligENTS with the structural systems biology approach to link entire bioactive chemical space to the pathogen structural genome. The innovative integration of chemical genomics with structural systems biology will not only greatly expand the scope of both ligand- and target-based methods, but also considerably improve the quality of predicted drug-target interaction models. Consequently, it may provide new opportunities for drug discovery.

To demonstrate the utility of this integrated approach, we apply it to identify molecular targets of drug-like compounds from the Malaria Box, and suggest drug repurposing opportunities for anti-malaria chemotherapies. Malaria is one of the most devastating and widespread tropical parasitic diseases and is the most prevalent in developing countries.[31] The Malaria Box includes 200 drug-like and 200 probe-like compounds that are active against the blood stage of *P. falciparum*, one of the most dangerous pathogen causing malaria. Although the compounds have desirable ADMET properties, their molecular targets in bacteria and human, as well as *in vivo* activity and toxicity, remain unknown. We use ligENTS to identify their target profiles in the chemical genomics databases, and their mapping to the *P. falciparum genome*. Using the target profile of active compounds as a proxy, we link approved drugs with active compounds against *P. falciparum.* Our results provide abundant testable hypothesis for further experimental validation.

2. Results and Discussion

2.1. *Ligand Enrichment of Network Topological Similarity (ligENTS) method*

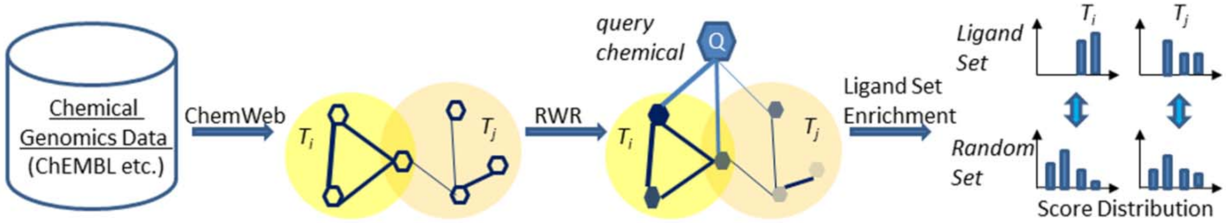

Fig. 1. Scheme of ligENTS. Hexagons represent chemicals. Two similar chemicals are connected. The more similar a chemical is to the query, the darker the hexagon. The chemicals in the colored sphere bind to corresponding targets T_i and T_j.

We have developed a new algorithm, ligand Enrichment of Network Topological Similarity (ligENTS), to assess the statistical significance of chemical-target associations based on the network topological similarity. As shown in Fig. 1, ligENTS consists of three key steps. **(1)** We connect around half a million chemicals in ChEMBL[32] into a chemical similarity network (termed ChemWeb). **(2)** Given a query, we apply a Random Walk with Restart (RWR) algorithm to define the network topological similarity between the query and other chemicals in ChemWeb. **(3)** To assess the statistical significance of the topological rank derived from the RWR, we apply random set theory to estimate the enrichment of a ligand set that is associated with a protein target in terms of the distribution of its network topological similarity scores. The final output of ligENTS is the false discovery rate (FDR) of a list of targets in the database, which may interact with the query chemical.

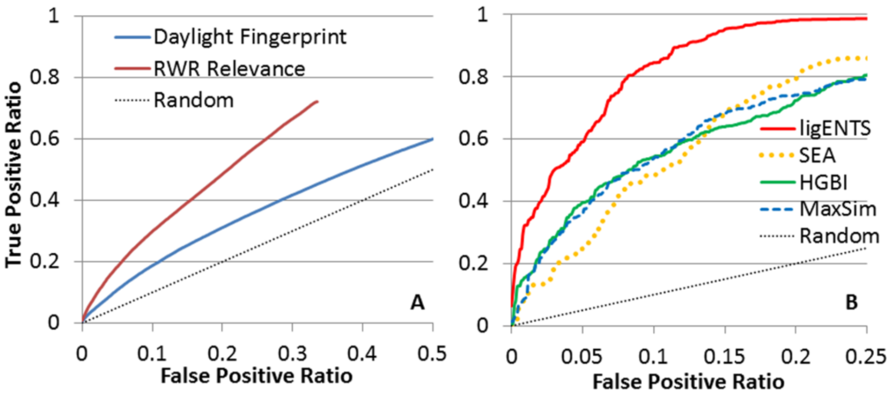

Fig. 2. Performance comparison of (A) global RWR relevance and Daylight fingerprint in detecting a pair of chemicals that share the same target, and (B) ligENTS (ligand Enrichment of Network Topological Similarity), SEA (Similarity Ensemble Approach)[16], HGBI (Heterogeneous Graph Based Inference)[15], and MaxSim (maximum similarity score in a set of ligands)[33] in ranking targets given a query chemical.

2.2. *Graph mining improves the performance of detecting pairwise chemical similarity*

State-of-the-art algorithms such as SEA, TurboSim,[34] MaxSim,[33] and IRV[19] only consider the similarity between the nearest neighbors, but ignores the global structure similarity relationships among all entities in a database. To overcome this limitation, we apply graph mining algorithms to define global relationships between chemicals. Given a query chemical, we first link the query to all nodes in ChemWeb, if edge weights between the query and any node are above a predefined threshold. Then, we use a Random Walk with Restart (RWR) algorithm to perform a probabilistic traversal of ChemWeb across all paths leading away from the query. The probability of choosing a path depends on the edge weight. The output of the algorithm is the list of all nodes (chemicals) in the network, ranked by the probability p_i for the query to reach node i. In this way, the query may detect related chemicals that are missed by the direct neighbors through intermediate nodes. As shown in Fig. 2A, a RWR transversal of ChemWeb improves the sensitivity and specificity of pair-wise chemical similarity search over a Daylight fingerprint similarity (http://www.daylight.com). When the Tanimoto Coefficient (TC) is 0.57 (approximately false positive ratio of 0.1), the Daylight fingerprint only identifies around 20% of all ligand pairs that bind to the same target. Using the same threshold to construct ChemWeb, the sensitivity of RWR is approximately 0.30, 50% more than that of the Daylight fingerprint. Thus the exploration of global community structures within the chemical similarity network allows us to detect novel protein-ligand interactions.

2.3. *Ligand Enrichment of Network Topological Similarity (ligENTS) considerably improves the performance of detecting novel drug-target associations*

Conventional ligand-based virtual screening focuses on ranking putative active compounds to a specific target. The issue that we need to address here is a reverse screening problem. Given a query chemical, how can we reliably rank all protein targets in a database by their likelihood to interact with the query chemical? To detect novel protein-chemical interactions, we developed a new algorithm, Ligand Enrichment of Network Topological Similarity (ligENTS). ligENTS combines RWR/ChemWeb with a ligand set enrichment framework. We compare the performance of ligENTS with three state-of-the-art algorithms: Similarity Ensemble Approach (SEA),[16] Heterogeneous Graph Based Inference (HGBI),[15] and the target assignment based on the most similar chemical in a ligand set (MaxSim).[33] SEA normalizes the sum of similarity scores between two sets of ligands known to bind to their targets, based on an empirical extreme value distribution model, and in an extensive benchmark study, SEA outperforms a state-of-the-art machine learning method.[18] SEA is the most relevant comparison to ligENTS in terms of statistical models for evaluating the chemical-target association. MaxSim is found to be the best performing method for ligand-based virtual screening when multiple ligands are used as a profile.[33] For comparison we modified the MaxSim algorithm to rank targets based on the maximum similarity score when comparing their ligands to the query. HGBI applies RWR on a heterogeneous drug-drug, drug-target, and target-target network to infer drug-target interactions, and outperforms other network inference algorithms for drug-target prediction.[15] As shown in Fig. 2B, HGBI is slightly better in the low false positive region than MaxSim. Consistent with a recent study in evaluating the performance of ligand profiles,[33] MaxSim outperforms SEA when the false positive rate is less than 0.15. Although HGBI is one of the best performers of the three of existing methods, HGBI does not provide a statistical significance assessment for predicted interactions.

LigENTS outperforms the above three methods in identifying novel chemical-target relationships, as shown in Fig. 2B. ligENTS identifies 200% and 50% of true positives more than that of HGBI at a false positive ratio of 0.01 and 0.05, respectively. The superior performance of ligENTS comes from its combination of the RWR search and global set statistics. The RWR captures the global structure of chemical space. However, conventional statistics models such as SEA fail when applied to global similarity problems. Global set statistics is more powerful than the fitted parametric statistical model. However, it is less useful when only the nearest neighbors are considered, as the scores of most ligands in the set are zeros, providing no information for the hypothesis testing. Enrichment of Network Topological Similarity (ENTS) by integrating RWR and global set statistics provides a general framework to enhance similarity search and association detection. Although this paper focuses on its application to chemical-target prediction, we have shown that ENTS improves the performance of protein fold recognition, RNA structure prediction, and disease gene identification. These results will be published elsewhere.

2.4. *Prediction of molecular targets of Malaria Box in the chemical genomics database*

To demonstrate the application of ligENTS to drug repurposing, we first use it to identify molecular targets of drug-like compounds from the Malaria Box, which are annotated in ChEMBL.

At a false discovery rate of 0.05, we associate 161 out of 200 drug-like active compounds from the Malaria Box with more than 577 proteins annotated in ChEMBL. The majority of these hits (~80%) are proteins from human and animal models. This reflects the screening and annotation bias in the chemical genomics databases. Nevertheless, enriched biological processes for these genes may provide valuable information on potential side effects (e.g., regulation of blood pressure, and muscle contraction) of these compounds, or their impact on pathogen-host interactions (e.g., response to molecule of bacterial origin), as shown in Table 1.

Table 1. Enriched biological processes of molecular targets of human and animal models for drug-like compounds from the Malaria Box.

Biological process	False Discovery Rate
second-messenger-mediated signaling	1.171e-58
positive regulation of lipase activity	3.028e-23
calcium ion homeostasis	1.923e-22
oxidoreductase activity	6.961e-17
regulation of blood pressure	1.117e-15
inflammatory response	1.593e-10
phosphoric diester hydrolase activity	3.745e-10
smooth muscle contraction	2.888e-07
regulation of apoptosis	2.353e-05
response to molecule of bacterial origin	3.870e-03

2.5 *Prediction of molecular targets of drug-like compounds in P. falciparum*

To identify the *P. falciparum* targets of drug-like compounds from the Malaria Box, we map the targets identified from the chemical genomics databases to the *P. falciparum* genome using both sequence similarity and ligand binding site similarity. Most of the mapped targets are essential genes in *P. falciparum*. Some of them (e.g., dihydroorotate dehydrogenase, beta-hydroxyacyl-ACP dehydratase, cysteine protease falcipain-3, and type II DNA topoisomerase) are novel targets under investigation.[35-38] When we rank the targets by the number of binding compounds, the top ranked targets include several proteins that bind to quinine, one of the most efficient drugs to treat malaria, providing support for our predictions. Other proteins include the JmjC domain containing protein, 3-oxoacyl-acyl-carrier protein reductase, and several putative transporters. The JmjC domain containing protein is particularly interesting. Twelve compounds are predicted to interact with JmjC. JmjC plays a key role in chromatin remodeling and histone posttranslational modifications that is fundamentally important in the developmental program of *P.*

falciparum.[39] However, this protein has not been explored as a drug target. Because the human homolog of JmjC exists, the detailed analysis of the drug binding site features may provide critical information on developing selective anti-malaria chemotherapy targeting JmjC. This analysis is ongoing.

2.6 *Repurposing approved drugs to target P. falciparum*

We apply ligENTS to 1,484 approved small molecule drugs in DrugBank to identify their molecular targets in ChEMBL. If the target profile of a drug is similar to that of the active compounds from the Malaria Box, we hypothesize that the drug is active against malaria. We term this strategy Target Similarity Profiling (TSP). Based on TSP, Table 2 shows the top ranked drugs that have the potential to treat malaria. The top hit sirolimus is a macrolide compound, targeting the FK506 binding protein. It has been used as an anti-fungal and an anti-neoplastic agent. FK506 binding protein in *P. falciparum* has been suggested as a novel target to fight malaria infection.[40] Several other drugs are predicted to target phosphodiesterase, dihydrofolate reductase, protease, carbonic anhydrase, somatostatin receptor, and ion channels. All these proteins are novel targets for anti-malaria therapeutics.[41-46] Doxycycline is a known anti-malaria agent, providing putative validation to TSP predictions. Thus, TSP provides abundant testable hypotheses for anti-malaria drug repurposing.

Table 2. Top 10 ranked drugs by TSP and their predicted target by ligENTS

Drug	Target(s)	Primary indication
Sirolimus	FK506 binding protein	anti-fungal and anti-neoplastic
Acitretin	Lyase, Nitric oxide synthase, DNA methyltransferase, Collagenase	treatment of psoriasis
Roflumilast	Phosphodiesterase (PDE)	chronic obstrtuctive pulmonary disease
Trimetrexate	dihydrofolate reductase (DHFR)	Antibiotics
Metaxalone	Protease	muscle relaxant
Piperazine	Carbonic anhydrase	Anthelmintic
Doxycycline	demethylase, hydrolase, dehydrogenase	Anti-malaria
Octreotide	Somatostatin receptor	treatment of acromegaly and reduction of side effects from cancer chemotherapy
Benazepril	Sodium channel subunit alpha, Voltage-dependent calcium channel subunit alpha	Hypertension

3. *Conclusion*

In this paper, we introduce a new chemical genomics algorithm, ligENTS, to map the chemical universe to its global pharmacological space, as well as an integrated chemical genomics and structural systems biology approach for anti-infectious drug repurposing. Although the detailed implementation of the algorithm needs to be improved, its prototype outperforms existing state-of-the-art methods, and demonstrates the potential for use in anti-infectious drug repurposing. The further development of this new strategy may consolidate phenotype-, ligand-, and target-based drug discovery, thereby facilitating the transformation of the conventional drug discovery process to a new paradigm of systems pharmacology.

4. Methods

4.1. *Benchmark*

We extract positive and negative cases from the bioactivity database ChEMBL[32]. To reduce the chance of including false positive hits, we only include those pairs with $IC50<10.0$ μM as positive cases. The negative cases include those pairs in which no binding is detected. We define the benchmark using the intersection of ligand sets in the positive and negative cases. After removing the chemical redundancy (Tanimoto Coefficients (TC) of 0.85, a common threshold in virtual screening), the final benchmark includes 390 chemicals, which involve 803 true and 1,336 false chemical-target interactions, respectively. We evaluate the sensitivity and specificity of the ranked target for a benchmark chemical when querying ChemWeb in which all benchmark chemicals are excluded.

4.2. *Construction of similarity matrix of ChemWeb*

Using a Daylight fingerprint representation of each chemical and TC as a similarity measure, we connect 415,975 chemicals that have high confidence annotation to targets in ChEMBL into a pairwise chemical similarity network. We represent ChemWeb as a weighted graph, in which nodes are chemicals. An edge is formed between two chemicals if they share the same activity and their chemical similarity is above a certain threshold. With a TC larger than 0.57, a threshold used by SEA but not optimized for ligENTS, ChemWeb consists of more than 10 million edges. We represent the ChemWeb weighted graph as a similarity matrix W.

4.3. *Implementation of Random Walk with Restart (RWR) algorithm*

We modified the RankProp algorithm,[47] a variant of RWR, and implemented it using a boost library (http://www.boost.org). The pseudo code of the algorithm is shown as follows.

```
Input: A graph representation of ChemWeb, with $i$ = 1, …, $N$ chemicals and their chemical similarity matrix
$W$ with the instance of $w_{ji}$; a diffusion vector $A$ with the instance of $a_i$, and a query chemical $q$.
Initialization: $p_q(0) = 1$; $p_i(0) = 0$
    while $t = 0, 1, 2, \ldots$ do
        for $i = 1$ to $N$ do
            $p_i(t+1) = w_{qi} + a_i \sum_{j=1}^{N} w_{ji} p_j(t)$
        end for
    until convergence $t = T^*$
output: a ranked list of $p_i(T^*)$
```

a_i corresponds to the restart probability in the RWR and determines how far the query will propagate through ChemWeb. In this study, a_i was set as a constant of 0.65.

4.4. *Implementation of set statistics*

Inspired by Gene Set Enrichment Analysis, we adapted the random set method[48] to estimate the enrichment of a ligand set that is associated with a protein target. For the RWR output $p_i(T^*)$, $i = 1, \ldots, N$, an unnormalized score for a ligand set S consisting of m chemicals is calculated as the average of the RWR outputs of these chemicals

$$\bar{X} = \frac{\sum_{p_j \in S} p_j}{m}$$

To compare the enrichment in a ligand set S with that of all other (N, m) distinct randomly drawn ligand sets of size m, the ligand set S is now considered as a random collection of m ligands whose score p_j are fixed. The exact distribution of $\bar{X}$ is intractable, but can be approximated with the normal distribution with mean and variance as follows:

$$\mu = \frac{1}{N} \sum_{j=1}^{N} s_j$$

$$\sigma^2 = \frac{1}{m}\left(\frac{N-m}{N-1}\right)\left[\left(\frac{1}{N}\sum_{j=1}^{N} {s_j}^2\right) - \left(\frac{1}{N}\sum_{j=1}^{N} s_j\right)^2\right]$$

The enrichment score is then normalized with

$$Z = \frac{\bar{X} - \mu}{\sigma}.$$

The false discovery rate (FDR) is estimated by fitting the enrichment score Z with the false positive ratio from the benchmark.

4.5. *Target identification of active compounds from the Malaria Box in the ChEMBL database and P. falciparum genomes*

LigENTS was first used to identify potential molecular targets of active compounds from the Malaria Box found in the ChEMBL database. Because most of the targets in ChEMBL are from human or model organisms, SMAP[49-51] and PSI-Blast[52] are applied to map the targets identified by ligENTS, which are not from *P. falciparum* genome, to *P. falciparum* proteins.

4.6. *Functional Enrichment Analysis*

Functional Enrichment Analysis of human targets is carried out using the DAVID functional annotation tool (http://david.abcc.ncifcrf.gov/). The whole genome of Homo sapiens is used as background.

Acknowledgments

This research was supported, in part, under National Science Foundation Grants CNS-0958379 and CNS-0855217 and the City University of New York High Performance Computing Center at the College of Staten Island**.** C.N. and R.H. were supported by the John P. McNulty Scholars Program.

References

1. World Health Organization. Fact sheet N°194 (2012).
2. S. Nwaka and A. Hudson. *Nat Rev Drug Discov* **5**, 941-955 (2006).
3. A. P. Chiang and A. J. Butte. *Clin Pharmacol Ther* **86**, 507-510 (2009).
4. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, *et al. Sci Transl Med* **3**, 96ra76 (2011).
5. W. A. Guiguemde, A. A. Shelat, D. Bouck, S. Duffy, *et al. Nature* **465**, 311-315 (2010).
6. F. J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, *et al. Nature* **465**, 305-310 (2010).
7. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, *et al. Bioinformatics* **24**, i232-240 (2008).
8. N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, *et al. PLoS Comput Biol* **5**, e1000397 (2009).
9. D. Vina, E. Uriarte, F. Orallo and H. Gonzalez-Diaz. *Mol Pharm* **6**, 825-835 (2009).
10. A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan. *Mol Syst Biol* **7**, 496 (2011).
11. F. Cheng, C. Liu, J. Jiang, W. Lu, *et al. PLoS Comput Biol* **8**, e1002503 (2012).
12. J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li, *et al. Bioinformatics* **29**, 238-245 (2013).
13. T. van Laarhoven and E. Marchiori. *PLoS One* **8**, e66952 (2013).
14. S. Alaimo, A. Pulvirenti, R. Giugno and A. Ferro. *Bioinformatics* **29**, 2004-2008 (2013).
15. W. Wang, S. Yang and J. Li. *Pac Symp Biocomput*, 53-64 (2013).
16. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, *et al. Nat Biotechnol* **25**, 197-206 (2007).
17. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, *et al. Nature* **462**, 175-181 (2009).
18. J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea, *et al. J Chem Inf Model* **48**, 755-765 (2008).

19. S. J. Swamidass, C. A. Azencott, T. W. Lin, H. Gramajo, *et al. J Chem Inf Model* **49**, 756 766 (2009).

20. V. Namasivayam, P. Iyer and J. Bajorath. *Chem Biol Drug Des* **79**, 22-29 (2012).

21. H. Sun, G. Tawa and A. Wallqvist. *Drug Discov Today* **17**, 310-324 (2012).

22. R. Guha. *J Chem Inf Model* **52**, 2181-2191 (2012).

23. J. J. Irwin. *Nat Chem Biol* **5**, 536-537 (2009).

24. S. Renner, W. A. van Otterlo, M. Dominguez Seoane, S. Mocklinghoff, *et al. Nat Chem Biol* **5**, 585-592 (2009).

25. R. D. Cramer. *J Comput Aided Mol Des* **25**, 197-201 (2011).

26. G. Schneider. *Nat Rev Drug Discov* **9**, 273-276 (2010).

27. G. Hu, G. Kuang, W. Xiao, W. Li, *et al. J Chem Inf Model* **52**, 1103-1113 (2012).

28. L. Xie, L. Xie and P. E. Bourne. *Curr Opin Struct Biol* **21**, 189-199 (2011).

29. H. M. Berman, B. Coimbatore Narayanan, L. D. Costanzo, S. Dutta, *et al. Febs Lett* (2013).

30. S. L. Kinnings, L. Xie, K. Fung, L. Xie, *et al. PLoS Comp Biol* **6**, e100976 (2010).

31. World Health Organization. *World malaria report 2008* (2008).

32. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, *et al. Nucleic Acids Res* **40**, D1100-1107 (2012).

33. R. J. Nasr, S. J. Swamidass and P. F. Baldi. *J Cheminform* **1**, 7 (2009).

34. J. Hert, P. Willett, D. J. Wilton, P. Acklin, *et al. J Med Chem* **48**, 7049-7054 (2005).

35. M. A. Phillips and P. K. Rathod. *Infectious disorders drug targets* **10**, 226-239 (2010).

36. D. Kostrewa, F. K. Winkler, G. Folkers, L. Scapozza, *et al. Protein Sci* **14**, 1570-1580 (2005).

37. C. Teixeira, J. R. Gomes and P. Gomes. *Curr Med Chem* **18**, 1555-1572 (2011).

38. C. Garcia-Estrada, C. F. Prada, C. Fernandez-Rubio, F. Rojo-Vazquez, *et al. Proc Biol Sci* **277**, 1777-1787 (2010).

39. L. Cui and J. Miao. *Eukaryot Cell* **9**, 1138-1149 (2010).

40. N. Bharatham, M. W. Chang and H. S. Yoon. *Curr Med Chem* **18**, 1874-1889 (2011).

41. K. Yuasa, F. Mi-Ichi, T. Kobayashi, M. Yamanouchi, *et al. Biochem J* **392**, 221-229 (2005).

42. Y. Yuthavong, B. Tarnchompoo, T. Vilaivan, P. Chitnumsub, *et al. Proc Natl Acad Sci U S A* **109**, 16823-16828 (2012).

43. C. Wegscheid-Gerlach, H. D. Gerber and W. E. Diederich. *Curr Top Med Chem* **10**, 346-367 (2010).

44. S. Reungprapavut, S. R. Krungkrai and J. Krungkrai. *J Enzyme Inhib Med Chem* **19**, 249-256 (2004).

45. J. X. Pan, R. B. Mikkelsen, D. F. Wallach and C. R. Asher. *Mol Biochem Parasitol* **25**, 107-

111 (1987).

46. S. A. Desai. *Curr Drug Targets Infect Disord* **4**, 79-86 (2004).

47. I. Melvin, J. Weston, C. Leslie and W. S. Noble. *Bioinformatics* **25**, 121-122 (2009).

48. M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, *et al. Ann. Appl. Stat.* **1**, 85-106 (2007).

49. L. Xie and P. E. Bourne. *BMC Bioinformatics* **8 Suppl 4**, S9 (2007).

50. L. Xie and P. E. Bourne. *Proc Natl Acad Sci U S A* **105**, 5441-5446 (2008).

51. L. Xie and P. E. Bourne. *Bioinformatics* **25**, i305-312 (2009).

52. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, *et al. Nucleic Acids Res.* **25**, 3389-3402 (1997).