FINAL

Send Orders for Reprints to reprints@benthamscience.ae

Mini-Reviews in Medicinal Chemistry, 2015, *15*, 000-000

Multiscale Modelling of Relationships between Protein Classes and Drug Behavior Across all Diseases Using the CANDO Platform

Geetika Sethi[1,#], Gaurav Chopra[1,2,3,#] and Ram Samudrala[1,3,*]

[1]*University of Washington, Department of Microbiology, Seattle, WA 98109, United States;* [2]*Diabetes Center, University of California, San Francisco (UCSF), San Francisco, CA 94143, United States;* [3]*Department of Biomedical Informatics, School of Medicine and Biomedical Sciences, State University of New York (SUNY), Buffalo, NY 14203, United States*


Ram Samudrala

Abstract: We have examined the effect of eight different protein classes (channels, GPCRs, kinases, ligases, nuclear receptors, proteases, phosphatases, transporters) on the benchmarking performance of the CANDO drug discovery and repurposing platform (http://protinfo.org/cando). The first version of the CANDO platform utilizes a matrix of predicted interactions between 48278 proteins and 3733 human ingestible compounds (including FDA approved drugs and supplements) that map to 2030 indications/diseases using a hierarchical chem and bio-informatic fragment based docking with dynamics protocol (> one billion predicted interactions considered). The platform uses similarity of compound-proteome interaction signatures as indicative of similar functional behavior and benchmarking accuracy is calculated across 1439 indications/diseases with more than one approved drug. The CANDO platform yields a significant correlation (0.99, p-value < 0.0001) between the number of proteins considered and benchmarking accuracy obtained indicating the importance of multitargeting for drug discovery. Average benchmarking accuracies range from 6.2 % to 7.6 % for the eight classes when the top 10 ranked compounds are considered, in contrast to a range of 5.5 % to 11.7 % obtained for the comparison/control sets consisting of 10, 100, 1000, and 10000 single best performing proteins. These results are generally two orders of magnitude better than the average accuracy of 0.2% obtained when randomly generated (fully scrambled) matrices are used. Different indications perform well when different classes are used but the best accuracies (up to 11.7% for the top 10 ranked compounds) are achieved when a combination of classes are used containing the broadest distribution of protein folds. Our results illustrate the utility of the CANDO approach and the consideration of different protein classes for devising indication specific protocols for drug repurposing as well as drug discovery.

Keywords: Druggable proteins, protein drug interactions, protein folds, protein classes, proteome drug discovery, drug discovery benchmark, multiscale modeling, polypharmacology.

INTRODUCTION

The discovery of a new drug targeting a specific disease or indication is usually initiated by finding ”hits” against ”target” proteins of interest using experimental high throughput screening (HTS) against a large chemical compound library. The *in vitro* hits are then assessed *in vivo* whereupon active compounds can then proceed to the lengthy FDA approval process. It is an iterative process which takes 10-15 years [1] and the cost of developing a new drug to market is on the order of $1.5 billion dollars including the cost of failures [2]. The traditional approach to drug discovery does not take into account the promiscuous interactions between the small molecule compounds and other proteins in an indication specific manner. The compounds being assayed for *in vitro* activity is done in a high-throughput manner, but with only a single or a few target(s) identified by biochemistry and molecular biology [3]. This traditional approach to drug discovery is based on designing a strong inhibitor against an essential protein as a “target”. Functional screens are performed to identify that such an essential protein to be targeted is responsible for pathogenesis, and the goal becomes to identify a small molecule compound which inhibits this target [4]. Most approved drugs currently in use have been developed by this approach [5-7]. Finding novel pathogenic targets is essential both for discovering new biology and for drug discovery, but the traditional single target approach is inherently flawed for drug discovery and therefore may be the reason for the reduction in the number of novel drugs discovered each year [1]. To this end, the high attrition rates of drug development projects [8], which leads to failure of a compound to become a drug during the development pipeline, further contributes to the reduction of novel drugs for existing and emerging diseases. The traditional approach for drug discovery goes against the evolutionary fact that protein structure is more conserved than its function, which provides a logical rationale for one compound being an excellent initial candidate for many

*Address correspondence to this author at the Department of Biomedical Informatics, School of Medicine and Biomedical Sciences, State University of New York (SUNY), 923 Main Street, Buffalo, NY 14203, USA; Tel: 206-251-8852; E-mail: ram@compbio.org

#These authors contributed equally to this work.

1389-5575/15 $58.00+.00

different protein targets. This promiscuity of small molecule drugs presents both a problem for novel drug development by traditional methods, giving rise to toxicity issues, as well as an opportunity for repurposing existing drugs for different indications/diseases. Commonly known as "drug repositioning" [9-12], finding new uses of existing drugs with known toxicity and safety profiles, enables pharmaceutical companies and researchers to accelerate drug development by 15-20%, resulting in a reduction of time and cost by a minimum of 2-3 years in drug development program duration [12]. There have been several examples of drugs that have been successfully repositioned, such as Amphotericin B [13], Aspirin® [14], and Bromocriptine [15]. Such drug repositioning strategies [16] not only promise reduction in cost and time but also result in higher chances of success at the clinical level [9]. Such successes are accompanied by the academic sector taking keen interest in developing drug leads by developing large drug discovery HTS activities, which has traditionally been done by the pharmaceutical industry [17]. These research activities are also accompanied by several National Institutes of Health (NIH) roadmap initiatives developing molecular libraries for drug development [17], as well as the National Center for Advancing Translational Science (NCATS). Such initiatives provide large amounts of data to the public sector which traditionally has been inaccessible to scientists at large, and selectively distributed by the private sector. This provides a launch pad for integrating high throughput screening data with novel computational methods to achieve higher successes rates for drug discovery.

Computational structural biology algorithms, including molecular docking, chemoinformatics, structural informatics, network inference, and bioinformatics data integration from genomic and proteomic databases, provide prospective insights into the complex relationships among drugs, genomic/proteomic targets, and diseases that may form the basis for successful drug repositioning and drug discovery. There are several methods to predict drug–protein interactions using graph-based molecular information of the interaction components and artificial neural networks as model systems [18, 19]. Development of novel computational methods to predict novel biomolecular interactions, and integration with indication specific data mined from existing knowledgebases, is essential for high-throughput drug repositioning. There have been several computational methods that have been developed previously for drug repositioning. Examples include ranking gene expression profiles to explore drug repositioning opportunities [20-22], and mining existing drug side effects information to deduce novel drug-target relationships to identify novel uses of existing drugs [23-24], among many others. We have developed the CANDO (Computational Analysis of Novel Drug Opportunities) drug discovery and repurposing platform that determines interactions between protein structures from all organisms (currently, 48278 protein structures) and all human ingestible compounds (currently a library of 3733 FDA approved drugs, supplements, and other compounds indicated for human use) to infer homology of compound/drug at the proteomic scale. Similar proteomic interaction signatures of compounds are indicative of similar functional behavior, which are used to repurpose existing drugs for particular indications. Most drug discovery methods depend on "similarities" to known drugs or on simulating the binding between a set of drug candidates and a protein target. CANDO is a departure from existing structural and chemical comparison methods of comparing compounds for similarity, as well as the well-established SAR (Structure-activity relationship) and QSAR (Quantitative SAR) approaches in drug discovery that extrapolate properties of known inhibitors to compounds of similar structure [25]. The CANDO compound-compound proteomic similarity implicitly integrates multitargeted interactions with proteomes from different organisms. This multiorganism proteomic profile is tolerant of missing dimensions in data variability (e.g. choice of structural and chemical comparison methods and docking algorithms, limitation of a singular targeted therapy etc.) resulting in a better signal-to-noise to identify compounds with similar phenotype for repositioning [26]. CANDO predicts interaction between all human approved compounds and all protein structures. We hypothesize that dissimilar proteomic interaction signatures (or regions of signatures) are indicative of off- and anti-target (side) effects, which can be used as a side effect prediction tool for any compound. Such compound-proteome relationship networks can also be used to predict novel drug targets in new and existing biological pathways, as well as predict safer drug combinations better than known singular therapies. Specifically, the effect of inhibition of known targets of an indication on the protein-compound relationship network that results in a set of co-expressed or interacting proteins known for a specific indication may find novel non-obvious targets. These targets can be used to find new drugs both for mono-therapy, and may also be useful to identify potentially synergistic multi-drug therapies. These hypotheses need to be benchmarked and validated, which is beyond the scope of this manuscript.

In this study, we use the CANDO platform to analyze the interaction signatures between human approved compounds and the druggable proteome to determine drug behavior for different indications/diseases. A druggable proteome consists of a set of proteins that interact with existing drugs, ideally with a therapeutic benefit to patients [27]. The pharmacology guided approach that focuses on the interaction of druggable proteins and compounds has previously proven to be successful [28]. We implemented a compound-centric benchmarking approach that uses an "all" vs. "all" methodology ("all" drugs × "druggable" proteins × "all" indications) to compare and contrast the interaction signatures of eight druggable protein classes (channels, GPCRs, kinases, ligases, nuclear receptors, proteases, phosphatases, transporters) in the context of indication specific drug discovery. We find that the predictive accuracy of the CANDO platform across all indications is enhanced when all the protein classes are considered. On average, across all indications, all druggable classes are important for a compound to become a drug for human use indicating that diversity of protein classes is essential to capture drug efficacy. The synergistic effect of the underlying network that differentiates a non-drug from a drug, suggests that the "network is the target". Nonetheless, there are specific classes that contribute significantly to CANDO accuracy for different indications. This information is useful to discover

novel pathways and/or may also be used to understand the underlying pathways responsible for pathogenesis. We have used our predictive bioanalytical benchmarking and prediction methodology for eight druggable protein classes but it is extendable to other protein classes as well. Our methodology enables "virtual surgery" to understand the relationships between atoms, molecules, pathways and physiological effectiveness of small molecules.

METHODS

Human Approved Compounds and Diseases/Indications Mapping Database

A database of human approved drugs and supplements were assembled from many different publically available databases including, DrugBank [29], NCGC Pharmaceutical Collection (NPC) [30], Wikipedia, and PubChem [31]. Each compound is first converted to a 3D structure using ChemAxon's MarvinBeans molconverter v.5.11.3 [32] to ensure that the input conformation does not bias the results. To remove redundancy between compounds from different databases, Xemistry's Cactvs Chemoinformatics Toolkit [33-34] was used to generate InChiKeys from the pre-processesed compounds. This resulted in a set of 3,733 unique human approved compounds, including all clinically approved drugs from the U.S. FDA, Europe, Canada and Japan, that map to 2030 indications/diseases. We obtained the disease-compound associations from the Comparative Toxicogenomics database [35] and mapped this dataset to our set of human approved compounds to obtain 1439 indications with at least two associated drugs, which were used for benchmarking.

Proteome Structure Prediction and Binding Site Identification

We used many different protein structure prediction methodologies to develop an integrated pipeline. The selection of tools was based on consistent performance of these methods at the Critical Assessment for protein Structure Prediction (CASP) experiments over several years. Protein structure prediction methods are assessed in a blind manner every two years at CASP. The results of the methods for the most recent CASP experiment (CASP10) can be viewed at http://www.predictioncenter.org/casp10/results.cgi. We have set up an integrated modeling pipeline using in-house benchmarking and HHBLITS [36], I-TASSER [37-38] and KoBaMIN [39-41] for protein modeling and COFACTOR [42] for identification of ligand binding sites. Moreover, we selected parameters and integrated these methods for our protein structural modeling pipeline based on several internal benchmarks [43-44]. As an example, we modeled the known structures for *M. tuberculosis* proteome, after the known structures and homologous structures with >30% sequence identity were removed from our library. This resulted in selection of parameters which were used to conservatively model different proteomes ~~for accurate modeling~~. A complete, full length model of each protein is generated using the I-TASSER [38, 45] package, which involves: (i) HHBLITS and LOMETS [46] to select templates for modeling; (ii) threading of protein sequences and gleaning of threading aligned region from templates as structural fragments; (iii) fragment assembly using replica-exchange Monte Carlo simulations; (iv) clustering of simulation decoys using SPICKER [47]; (v) generation of full atomic refined model from SPICKER cluster centroids using ModRefiner [37]; and (vi) using KobaMIN as an end step to refine the final models. This modeling pipeline has been applied to all human proteins that can be modeled with high accuracy (~50% of *Homo sapiens* proteome consisting of 14595 known and modeled proteins), *M. tuberculosis* (~70% of the proteome consisting of known and modeled proteins), *P. aeruginosa*, and a large number of viral proteomes. Ligand binding site locations and potential template ligands are predicted using the COFACTOR algorithm, which scans the known or modeled 3D protein structure against a representative template library of experimentally determined protein structures with bound ligand in the Protein Data Bank (PDB) [48]. The template proteins are scanned based on a global structure similarity search algorithm, followed by a local structure similarity refinement search on selected hits with the purpose of filtering out template proteins that do not share binding site similarity with the query protein. During both global and local structure similarity searching, the template protein is scored against the query protein using a custom structure-sequence similarity measure, which captures both the chemical and structural similarity of the ligand binding pocket between the query and the template proteins, namely, the BSscore. New structures are updated routinely with non-redundant protein structures from the PDB (currently, 31135 proteins) and other proteomes are added to the current list as soon as they are modeled. Currently, a total of 48278 protein structures with multiple binding sites are used to generate multiorganism compound-proteome signatures.

Compound-Proteome Interaction Signature – CANDO Matrix

The compound proteome interaction signature metrics for each of the eight druggable protein classes was derived using a hierarchical chem and bio-informatic fragment based docking with dynamics protocol to predict interactions between the protein structures and all the small molecule compounds. The CANDO v1 matrix is a set of predicted binding score values that have resulted from integrating a cheminformatic algorithm, OpenEye ROCS [49] and a structural bioinformatics based algorithm, COFACTOR [42]. Each protein can have multiple binding sites and each binding site prediction has a BSscore. There are multiple template ligands predicted for each binding site based on the chemical and structural similarity of the ligand binding pocket between the query and the template proteins. Chemical and structural similarity of 3,733 compounds with all predicted template ligands from all binding sites of all proteins are analyzed using the OpenEye ROCS 3D-similarity search algorithm, which uses atom-centered Gaussians to evaluate 3D chemical and structural overlap. This results in multiple ROCS scores for each binding site. The objective of this analysis is to identify those predictions in which the approved compounds are highly similar to high confidence template ligand predictions by COFACTOR and

may also bind at the same location. Different conformations of the human approved compounds are generated and compared to the template ligands predicted in the binding pocket to compute the ROCS score to account for conformational entropy of the compounds. If the BSscore is greater than a cut-off value of 1.1, and the chemical and structural similarity score for the template ligands in these binding pockets with respective human approved compounds (ROCS score) is also greater than the cut-off value of 1.1, then an interaction is considered to have occurred. These cutoffs are selected based on a benchmarking set of 1100 non-natural ligand pairs obtained from the PDB which bind to the same proteins around the same set of residues in the active site but where the ligands are chemically different [43]. Based on these two conditions, we assign the protein with the BSscore as the real value of each compound-protein interaction to populate the CANDO v1 interaction matrix.

Compound Proteome Homology

A compound-compound similarity matrix is generated using the CANDO v1 interaction matrix. To compute the similarity of compounds based on their interaction signatures, an "all vs. all" approach was adopted such that each compound-proteome interaction signature (vector of real numbers of interaction of compounds with multiple proteins) of the CANDO matrix was compared with every other compound-proteome signature in the CANDO matrix using the root mean square distance (RMSD) measure to derive a sorted compound-compound similarity matrix. Any two compounds with low RMSD values are proteomic homologues and should have similar function. For computational structural biology, metrics to measure homology (sequence identity) or structure similarity (RMSD/ TMscore) are used in an analogous fashion to infer the function of uncharacterized proteins. The same ideas are used here to infer homology between compounds using the proteomic interaction signature. This results in a better signal-to-noise to identify compounds with similar functional phenotype for repositioning.

Compound Ranking and CANDO Percentage Accuracy

Each of the 3733 compounds is compared to each of the other 3733 compounds and sorted according to similarity (highest to lowest). The similarity metric used is the RMSD of the interaction scores that comprise the compound-proteome interaction signature across 48278 protein structures. There are 1439 indications with more than one approved compound and nine compounds on average per indication (a compound may also be approved for more than one indication). The leave-one-out benchmark calculates the accuracy of finding two compounds approved for the same indication within the top10 ranked compounds relative to the total number of compounds approved for that indication expressed as a percent ("top10 accuracy"). The percentage accuracy for each indication is calculated using the formula:

CANDO 'top k' percentage accuracy for an indication j

$$= \frac{c_j^k}{d_j} \times 100$$

where c_j^k is the number of approved drugs for an indications within a certain rank k and d_j is the total number of drugs approved for an indication j. This is averaged over the 1439 indications with two or more approved compounds to produce comprehensive average percentage accuracies. Other rank cutoffs such as top25, top1% (top37 for 3733 compounds), top50, and top100, were also evaluated. However, in this study the accuracy of our benchmarking was done using the top10 criterion to show the accuracy of validating 10 compounds when predictions are made with CANDO for any particular disease/indication. This is a manageable task for any laboratory working on a particular indication to produce preliminary validation data. In addition, the top10 criteria is the stringent and most rigorous cutoff one can use for comparing benchmarking accuracy, and comparing relative performance at other cutoffs is confounding due to double counting. We have experimentally validated up to 40 compounds per indication (162 compounds total) to identify one or more leads for 11 diseases, resulting in an overall hit rate of ~35% that is comparable or better than the corresponding treatment for the disease [44].

Controls and Comparison Protein Sets for Benchmarking

We use several protein sets in order to identify the contribution of proteomes (or pathways) towards CANDO benchmarking accuracy. Appropriate random controls were also performed to measure the robustness of the integrated platform of algorithms for drug repurposing. Table **1** shows the details of known and modeled structures for different protein sets used for benchmarking. Specifically, we use the following protein sets for benchmarking: (i) All proteins. To challenge our computational experiments of benchmarking different druggable protein classes, we compared the accuracy levels of the protein classes to the 'All proteins' set. This set includes all protein classes and represents a "universe" of protein structures, with a total of 48278 proteins with multiple binding sites mapped to 3733 compounds. (ii) Best single sets. We performed a rigorous comparison of CANDO accuracy by incorporating multiple proteome sets comprised of different numbers of proteins (i.e. 10, 100, 1000, 10000) corresponding to each of the eight druggable protein classes. Additionally, a stringent test is designed where four control protein sets of 10, 100, 1000, and 10000 proteins were obtained by taking the averages of single best performing proteins that were obtained by analyzing the benchmarking accuracy results for all 48278 proteins individually. CANDO accuracies of 10 single best proteins, 100 single best proteins, 1000 single best proteins and 10000 single best proteins (across all five categories: top10, top1%, top50, top100) were averaged, and the resulting sets were named as 10 best single, 100 best single, 1000 best single and 10000 best single protein set respectively. (iii) Random controls. In order to quantify the chance event of obtaining a particular CANDO accuracy over all indications, we used two random controls for our benchmarking study. Our first control is calculated random

Table 1. Distribution of number of experimentally solved and computationally modeled protein structures for the eight druggable protein classes used to generate proteomic signatures.

Protein class	Number of total structures	Number of solved structures	Number of modeled structures
Channel	156	68	88
GPCR	885	6	879
Kinase	1915	823	1092
Ligase	579	177	402
Nuclear receptor	105	20	85
Phosphatase	703	298	405
Protease	496	204	292
Transporter	584	117	467
10 best single	34	10	24
100 best single	325	133	192
1000 best single	2606	1271	1335
10000 best single	19381	10101	9280
All	48278	24958	23320

accuracy, i.e. calculated accuracy that is expected by chance. Since there are 9 associated drugs per indication on average, and a total of 3733 human approved compounds for all indications, we would expect to obtain 0.2% (9/3733*100) accuracy by random selection. For the second random control, we use the best average accuracy of ~1000 randomly generated matrices (by swapping/scrambling) all rows and columns to random positions, which also resulted in an average accuracy of 0.2% for the top10 ranking category.

RESULTS AND DISCUSSION

CANDO Platform Benchmarks

We benchmarked our CANDO drug discovery platform to retrieve known drugs for 1439 diseases for which there are at least two approved drugs. An overview of the CANDO platform is shown in (Fig. **1**), which describes the compound-centric benchmarking methodology using similarity of compound-proteome interaction signatures. (Figs. **2-4**) show the contribution of different protein classes to the drug indication benchmark using the CANDO platform. We use eight druggable protein classes, the set of all 48278 protein structures ("All proteins"), best performing single protein structures ("best single") and random controls to evaluate the benchmarking of the CANDO platform. Overall, 4/8 protein classes show similar performance compared to the best single protein sets (Figs. **2** and **3**) but the CANDO benchmarking accuracy is higher for individual protein classes for particular indications (Figs. **3A** and **4**). Proteomic signatures using "All proteins" performed much better than individual protein classes for all 1439 indications, on average, to retrieve known drugs ~~for indications~~.

CANDO Benchmarking Accuracy vs. Random Controls

We performed randomized controls to test the accuracy of our CANDO benchmarking platform. There is an average of 9 approved drugs for each indication from a total of 3733 approved drugs for all indications. Thus, the chance of picking the correct drug for an indication on average is ~0.2% (9/3733 *100). A more rigorous randomized control was also done by randomizing the CANDO interaction signature matrix values by changing the positions of all rows and columns and taking the best average accuracy of ~1000 randomly generated matrices. The accuracy for the random control is ~0.2% suggesting that there is true signal from the CANDO platform which is two orders of magnitude higher in performance relatively, with a lot of potential for improvement in the future. This is best described in the context of HTS assays developed for single protein targets to be tested against a compound library. If we wanted to do HTS to retrieve a known drug for one disease/indication, the hit rate, on average, is close to our benchmarking randomized control of ~0.2% (test 414 compounds to retrieve 1 known drug), compared to ~12% benchmarking accuracy with our methodology where the top10 predicted compounds can retrieve 1 (1.2) drug for any indication, on average, and up to 4 drugs for ~800 indications if the top 100 predicted compounds are tested. A comparison between traditional virtual screening and HTS has been done previously [50]. It was based on the Merck chemical collection against the tuberculosis target dihydrodipicolinate reductase resulting in a hit rate of < 0.2% for HTS, and 6% hit rate for traditional virtual screening using

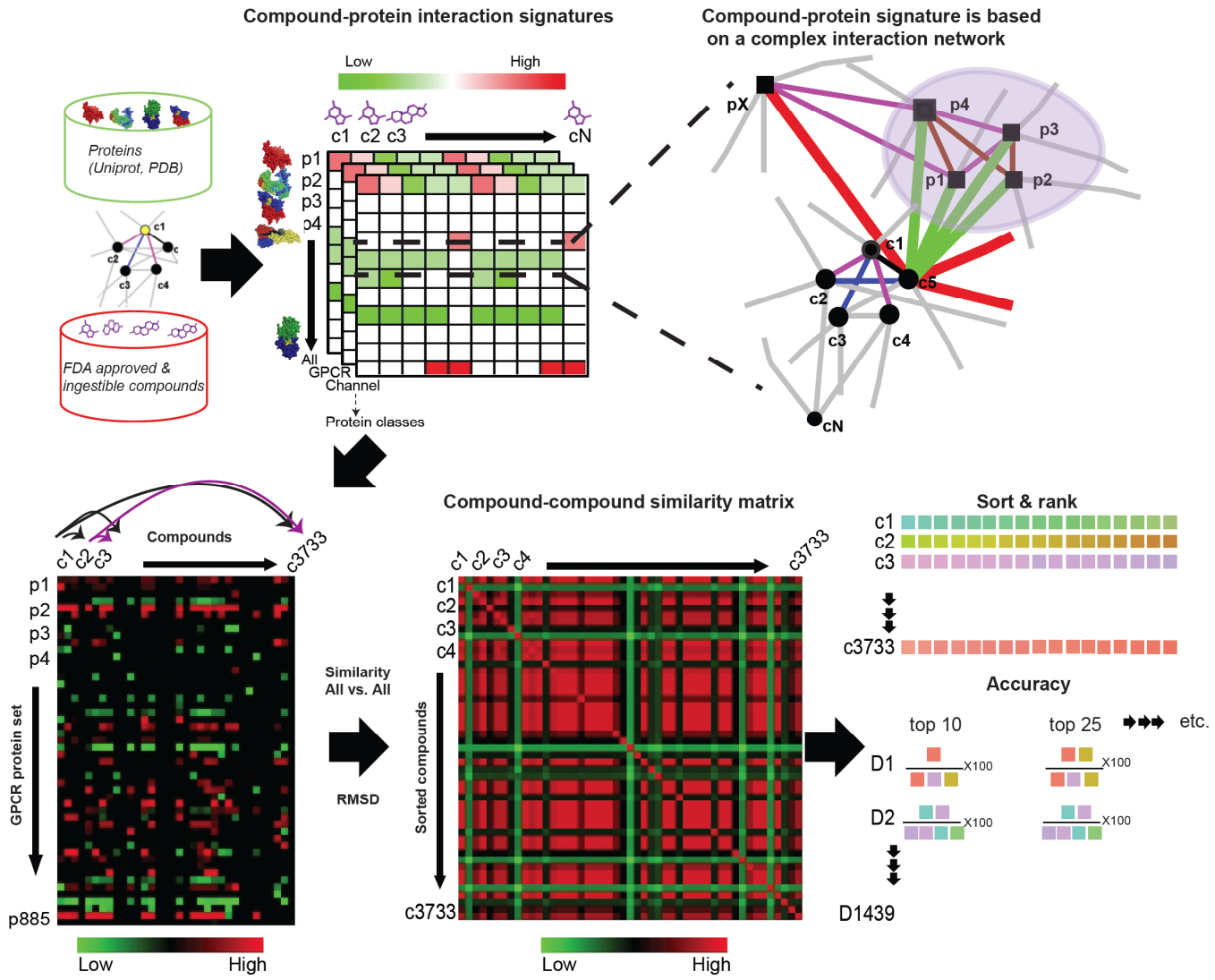

Fig. (1). Compound-centric benchmarking workflow. As a first step, the compound-protein interaction signatures are generated by mapping 3733 compounds (used to treat a total of 1439 indications) to a protein set of interest (8 druggable protein sets that include channels, GPCRs, kinases, ligases, nuclear receptors, phosphatases, proteases and transporters) using chemoinformatics and structural bioinformatics based approaches to obtain a compound-protein interaction matrix (aka CANDO matrix) of real values. These signatures are based on a complex interaction network of binding site comparisons with the PDB. Protein-protein interaction networks are not used for the benchmarking study and are shown to indicate an added layer of complexity for indication specific CANDO predictions, which have been used for prospective experimental validations. The signatures are then scored using an "all vs. all" similarity method (all compounds × all proteins × all indications) to obtain a sorted compound - compound similarity matrix. Interaction signatures of every single compound are compared to all other compound-proteome interaction signatures in a pairwise manner using the root mean square deviation (RMSD) method to derive a compound-compound similarity matrix shown using a heat map. Next, these scores are sorted and ranked so that all compounds are ranked relative to each other in an indication specific manner. Finally, average accuracy (using the formula: number of compounds approved for an indication within a certain rank/total number of compounds approved for that indication * 100) that reflects the recovery rate of related compounds (in each of the five categories top 10, top 25, top 37, top 50 and top 100) is calculated for each of the 1439 indications/diseases. This enables us to adopt a compound - centric leave one out procedure to accurately identify related compounds approved for the same indication.

docking with single or few targets [50]. Thus, a proteomic-based approach with a diverse distribution of protein fold space has two orders of magnitude higher accuracy compared to a random control, suggesting that the network of proteins is a more relevant target for the disease.

Relationship Between Protein Classes and Drug Behavior

In this study, we focused on eight druggable protein sets (channel, kinase, ligase, GPCR, nuclear receptor, protease, phosphatase, transporter) and evaluated their performance with respect to benchmarking accuracy in contrast to the

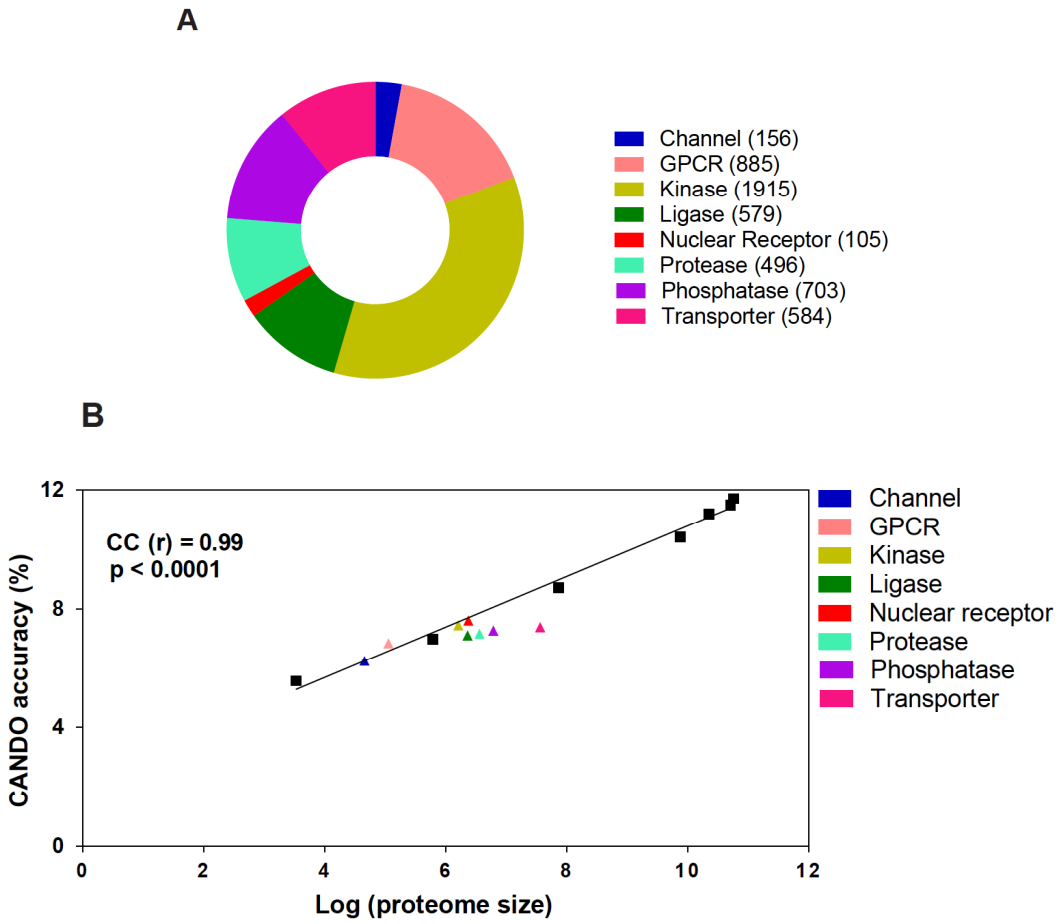

Fig. (2). Benchmarking accuracy as a function of protein class and size. **A**) Shown is the distribution of the number of proteins in each of the eight druggable protein classes (channels, GPCRs, kinases, ligases, nuclear receptors, proteases, phosphatases, and transporters) evaluated in this study. **B**) Shown are the benchmarking accuracy results for all 1439 indications retrieved for each of the seven reference/control/comparison sets fitted by linear regression (line fitted through black squares). A correlation (CC(r)) between accuracy and size of a protein set was found to be equal to 0.99 (p value < 0.0001). Next, each of the eight druggable protein classes (represented by colored triangles) is plotted against their respective number of proteins. Our results indicate that the diversity of protein classes is needed to capture drug efficacy and their potential human use.

reference/control/comparison sets. (Fig. **2A**) shows the distribution of number of proteins across each of the eight druggable protein sets. This number includes both solved and modeled structures. The number of solved and modeled structures in each of the eight druggable protein sets and the reference/control/comparison sets is shown in Table **1**. CANDO accuracy was greater with an increase in the number of proteins for each of the protein sets (Fig. **2B**). Four druggable protein classes (channel, GPCR, kinase and nuclear receptor) performed equally well as our reference protein set. The ligase, phosphatase and the protease sets showed minor deviation and the transporter class exhibited maximal deviation from the reference sets suggesting that GPCRs, kinases and channel proteins are important for drug discovery, as observed with the majority of currently marketed pharmaceuticals which target these proteins. Moreover, a study at Pfizer (U.K.) that aimed to evaluate protein classes that bind to low-molecular-weight compounds with binding affinities below 10 μM showed that GPCRs, kinases and channel proteins constitute a high percentage of druggable classes [51].

Various compound-protein interaction sets can also be incorporated into the CANDO platform. The observation of increased accuracy with increase in size of the protein sets can be explained by a statistical multiplication effect, as observed in shotgun sequencing methods. The synergistic effect of the underlying network that differentiates a non-drug from a drug suggests that the "network is the target". Moreover, there are specific druggable protein classes that contribute significantly to CANDO accuracy for different indications. These proteins represent important drug targets and evaluating the behavior of drugs in the context of druggable proteins is useful to discover novel pathways and/or to understand the underlying pathways for the pathogenesis of the disease. Therefore, it is important to characterize the universe of druggable proteins and understand their performance using predictive bioanalytical tools for drug discovery.

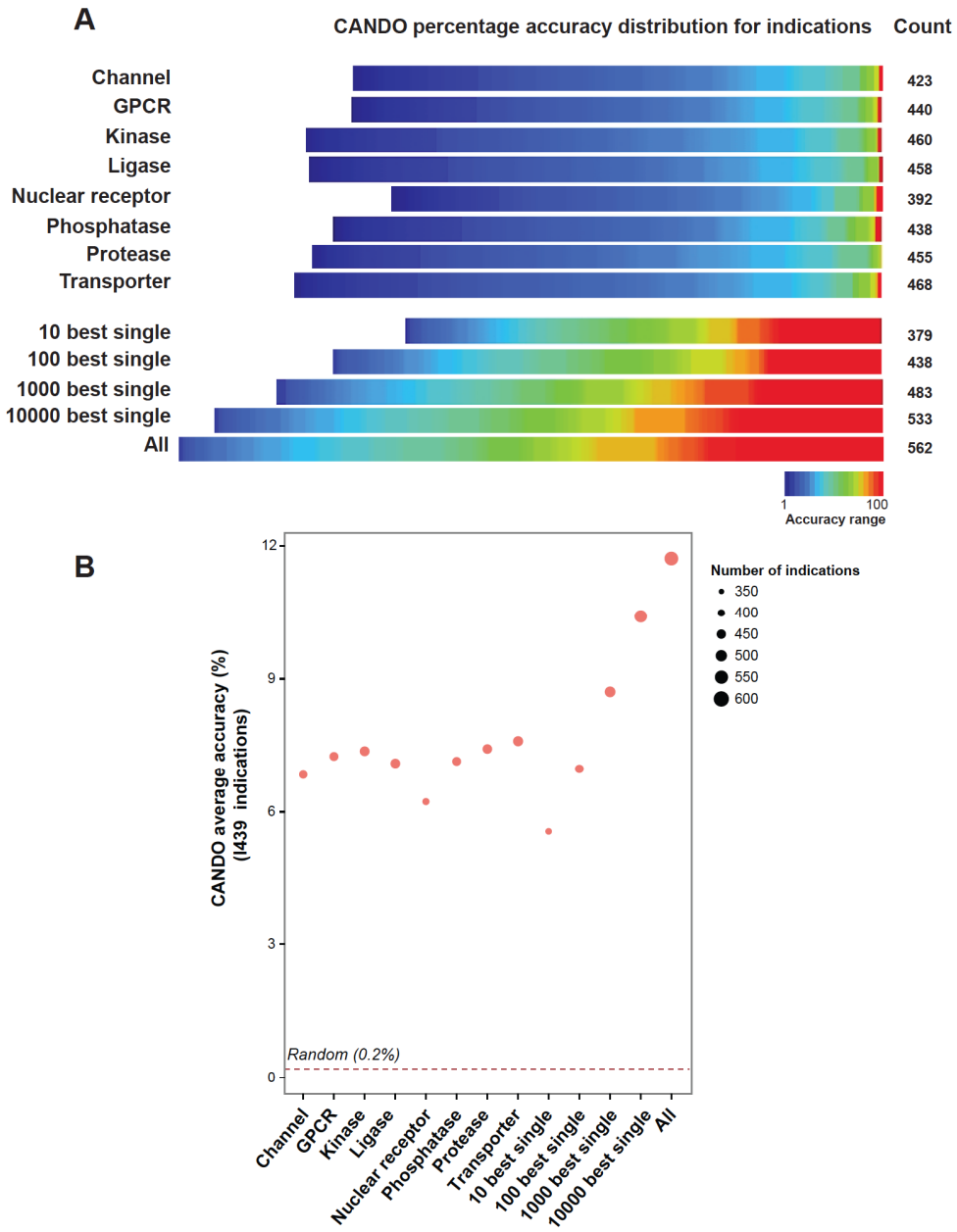

Fig. (3). Benchmarking accuracy distributions and average accuracies for eight druggable classes of proteins. The indications (out of 1439) that yielded non-zero accuracy values (accuracy > 0%) from benchmarking were considered successful indications and ranged from 379 to 563 for the eight druggable classes and the five reference/control sets (10 best single, 100 best single, 1000 best single, 10000 best single and the All set, see Methods for details) evaluated. A) Shown are the heat bars displaying the distribution of accuracy values in the top10 category (1% to 100%) for successful indications across the eight druggable protein sets and the five reference/control sets. The accuracy values range from 1.8% to 100%. B) Shown is a bubble plot representing the average accuracy (all 1439 indications) for each of the eight protein classes. The results were contrasted to the accuracies of single best proteins (1, 10, 100, 1000, 10000), multiple protein sets evaluated together (All set, see Methods section) and to the accuracy for a given compound obtained by chance (Random, see section ~~Choice~~ of controls/comparison protein sets for details). The eight classes yield similar results in terms of their accuracies for successful indications (standard deviation = 0.003) and perform well on different sets of indications. However, the best accuracies are obtained when all protein structures are used for signature comparison to determine compound similarity. Our results emphasize that interactions with all classes play an important role for a drug indicated for human use.

Accuracy Distributions and Average Accuracies

We identify successful predictions for indications with non-zero percent accuracy values using the CANDO benchmarking platform. (Fig. **3A**) shows the accuracy distribution for each of the successful indications (out of 1439) in top10 category for each of the eight druggable protein sets as well as the control protein sets. The number of successful indications across the eight druggable protein sets

and the five control sets ranged from 379 to 563, and the accuracy values ranged from 1.8% to 100%. Among the eight protein sets, the kinase and the transporter sets yielded the maximum number of successful indications. The nuclear receptor set yielded maximum number of diseases with 100% accuracy. (Fig. **3B**) shows the average accuracy across all 1439 indications with the number of successful indications represented by the bubble size. All the eight classes yield similar results in terms of accuracies for successful indications but perform much better on selected sets of indications. This may indicate that particular protein classes are responsible for pathogenesis of the disease and targeting them would lead to novel therapeutics. However, the best accuracies are obtained when all protein structures are used for interaction signature comparisons to determine compound similarity, suggesting the role of multiple networks working together in biology to achieve a certain phenotype.

Best Performing Indications for the Druggable Protein Classes

We evaluated the 10 best indications yielding the maximum percent accuracy for each of the eight druggable protein sets. (Fig. **4**) represents the distribution of accuracy percentage values for the best 10 indications retrieved for each of the eight classes across three percent accuracy divisions (40-60%, 60-80% and 80-100%). We identify selected indications that perform better for particular protein classes, to provide a rationale to target high accuracy druggable classes for such indications. There are overlapping indications among the best ones for different druggable protein classes suggesting the systems level interplay for these diseases. To verify our predictions of particular druggable classes being involved in the pathways for pathogenesis or treatment of selected indications with high CANDO accuracies, we collected independent evidence from existing literature [28, 52-94], which is summarized in (Supplementary Table **S1**). As an example, the indication congenital limb deformities has high CANDO accuracy with the kinases and ligases protein sets (Fig. **4**) with evidence from existing literature that these proteins are implicated in the disease [62, 69]. These findings strengthen the significance of druggable protein classes for selected indications.

CONCLUSIONS AND FUTURE WORK

The most effective drugs in humans (e.g. Aspirin® or Gleevec®) inevitably interact with and bind to multiple proteins, a feature that traditional drug discovery models based on single target approaches fail to take into account. Multitargeting is necessary because every drug has to be effective at its site of action (for example, HIV-1 protease inhibitors have to bind and inhibit the protease molecule) and has to be readily metabolized by the body (for example, by the cytochrome P450 (CYP450) enzymes, which are responsible for metabolizing the majority of drugs). Computational screening for multitarget binding and inhibition is effective because it exploits the evolutionary fact that protein structure is more conserved than sequence and function, providing logical evidence for one compound being an excellent initial candidate to inhibit many different protein targets. More directly, if we have a given compound that has gone through the FDA approval process for one indication, we may be able to reposition it for other indications more readily. We therefore have developed computational methods that account for protein/small molecule docking, network properties and integrate multiple data sources simultaneously as CANDO interaction signatures. In this study, the CANDO v1 compound-proteome interaction signatures are a real value set of vector of numbers created by mapping the interactions between protein structures and small molecule compounds via a complex predictive bioanalytic approach. Each real value is an evolutionary estimate of a protein structure physically binding to a small molecule compound. These interaction signatures are indicative of functional behavior of compounds at the proteomic level in terms of the clinical efficacy of the drug approved for their respective indication.

We show results from hold-one-out benchmarking experiments performed using 1439 indications with two or more approved compounds. The benchmarking determines the ability of the CANDO platform to accurately identify related compounds approved for the same indication. The criteria for a compound to be labeled approved for, or associated with, therapeutic use was determined based on US FDA approval as well as data obtained from the Comprehensive Toxicogenomics Database. Each compound is then ranked relative to every other compound based on the similarity between compound-proteome interaction signatures across 48278 proteins using the root mean square deviation (RMSD) of the interaction scores as the similarity detection metric. The accuracy of the ranking for a compound approved for an indication is evaluated based on whether another compound approved for the same indication falls within a particular cutoff in the ranked list of similar compounds. Average benchmarking accuracies range from 6.2 % to 7.6 % for the eight classes when the top 10 ranked compounds are considered, in contrast to ranges of 5.5 % to 11.7 % obtained for the comparison/control sets consisting of 10, 100, 1000 single best performing proteins. These results are generally an order or two of magnitude better than the average accuracy of 0.2% obtained when randomly generated interaction score matrices are used. Different indications perform well when different classes are used but the best accuracies (up to 11.7% in the top10 ranked compounds) are achieved when a combination of classes are used containing the broadest distribution of protein folds. Our results indicate that we are able to use the CANDO platform based on structural predictive bioanalytics to translate atomic level understanding of protein-small molecule interactions to the clinical behavior of drugs. The similarity of compound-proteome interaction signatures may thus be used more reliably than single molecule docking approaches to infer homology of drug behavior at the proteomic level for drug repositioning and discovery.

For benchmarking, have focused on all indications/ diseases with at least two approved drugs, which makes our study dependent on the accuracy of the drug-indication mappings. However, it would be valuable to assess the accuracy/confidence of the benchmarking program when

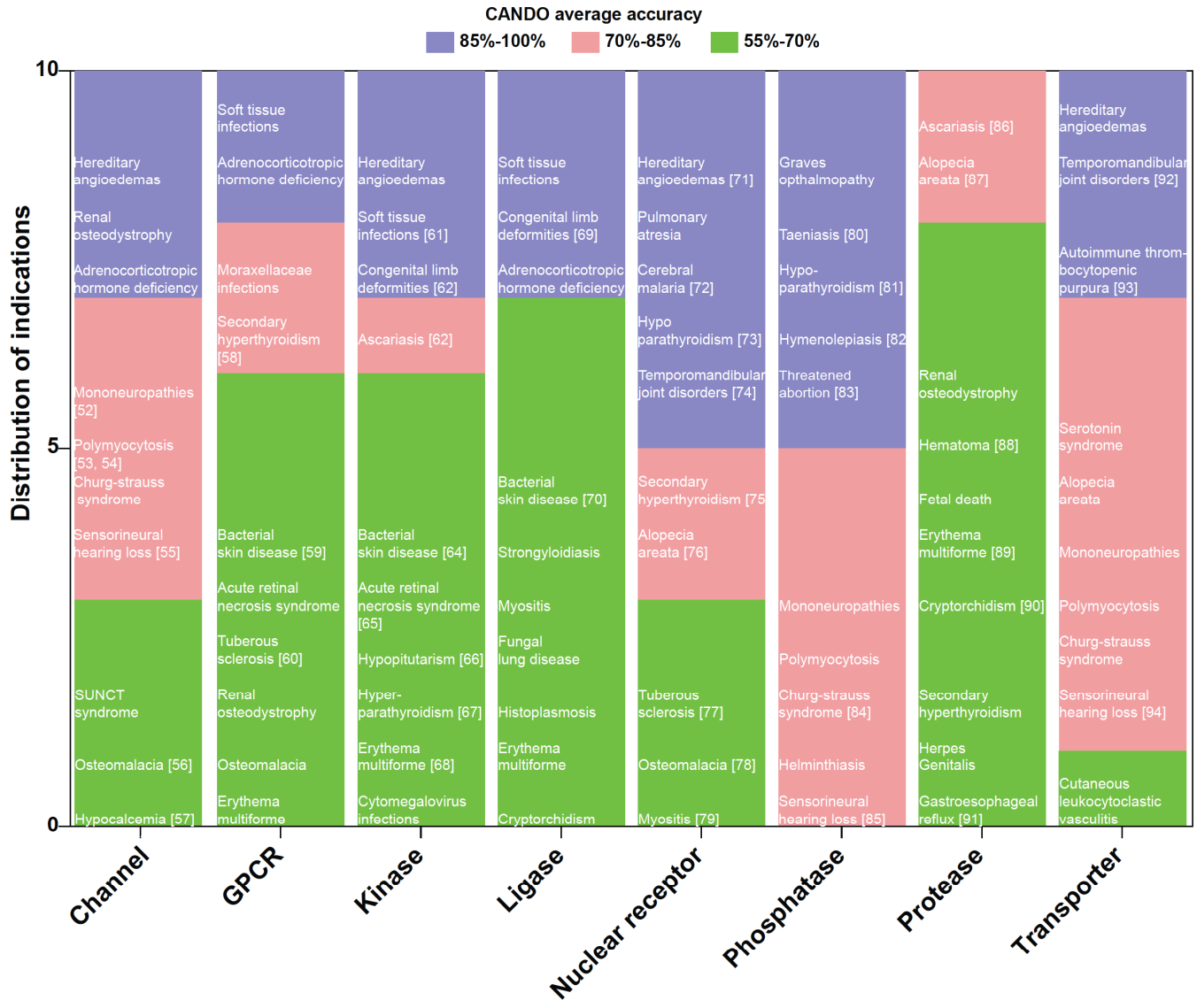

Fig. (4). Evaluation of best performing indications across all eight druggable classes. Shown are the bar plots representing the distribution of the top10 indications retrieved from benchmarking each of the eight classes according to their accuracy values. Indications were distributed across three categories for accuracy values (55-70% (bar area in green), 70-85% (bar area in pink) and 85-100% (bar area in purple)) to help analyze which indications perform better than others for a particular protein class. Seven out of eight classes retrieved indications with 100% accuracy. Additionally, independent evidence of protein class involvement (reference within square brackets) in these pathways based on a literature search (PubMed) is provided. These results indicate synergy between different protein classes for particular indications, emphasizing the value of a structure-based systems biology approach to drug discovery.

there is only one approved drug. To this end, have successfully used indication specific protocols to inhibit many viral pathogens *in vitro*, including dengue, where no current therapy exists [44], and herpes, where we inhibit all classes of herpes viruses with our prediction [95]. We have also use the same methodology to identify novel drugs using multiproteome interaction signatures of experimental compounds. To this end, we predicted novel compounds to treat beta-thalassemia based on integrating interaction signatures from experimental compounds and a set of compounds already used in clinical trials for the disease [96]; and for extreme drug resistance tuberculosis by using a weighted host-pathogen interaction network to identify multitargeted putative drugs [97]. We are in the process of following up these predictions with clinical studies. We are also testing many different interaction score metrics obtained by our hierarchical knowledge-based fragment docking with dynamics [98], and our shotgun evolutionary structural interaction network based docking with dynamics, methodologies [99]. The "compound-centric method" used in this study is based on the similarity of compounds to make predictions. The weighting of protein "targets" to scale these interactions for a specific disease is performed in an *ad hoc* manner. As part of our future work, we aim to adopt machine-learning approaches (neural networks, Support Vector Machines (SVM), and Bayesian probabilistic

networks) that will formalize the procedure for scaling interactions for known protein targets and networks in future studies. Multiple approaches will be evaluated to generate compound proteome interaction signatures and rank compounds in an indication-specific manner followed by iterative learning and integrating experimental data that will enable increasingly accurate predictions, predict adverse effects, and multi-drug therapies, resulting in a comprehensive drug discovery platform with higher efficiency, lowered costs and increased success rates.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We thank the members of the Samudrala research group and our numerous collaborators (http://protinfo.org/cando/collaborations). The study was supported by a NIH Director's Pioneer Award (1DP1OD006779-01) to Ram Samudrala and a JDRF fellowship to Gaurav Chopra. A free academic license graciously provided by OpenEye Software was used to carry out a portion of the interaction score calculations.

SUPPLEMENTARY MATERIALS

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Dimasi, J.A. New drug development in the United States from 1963 to 1999. *Clin. Pharmacol. Ther.,* **2001**, *69*(5), 286-296.

[2] DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G. The price of innovation: new estimates of drug development costs. *J. Health Econ.,* **2003**, *22*(2), 151-185.

[3] Huang, E.S.; Samudrala, R.; Ponder, J.W. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.,* **1999**, *290*(1), 267-281.

[4] Sethi, G.; Pathak, H.B.; Zhang, H.; Zhou, Y.; Einarson, M.B.; Vathipadiekal, V.; Gunewardena, S.; Birrer, M.J.; Godwin, A.K. An RNA interference lethality screen of the human druggable genome to identify molecular vulnerabilities in epithelial ovarian cancer. *PLoS One,* **2012**, *7*(10), e47086.

[5] Lombardino, J.G.; Lowe, J.A., 3rd. The role of the medicinal chemist in drug discovery--then and now. *Nat. Rev. Drug Discov.,* **2004**, *3*(10), 853-862.

[6] Cooper, M.A. Optical biosensors in drug discovery. *Nat. Rev. Drug Discov.,* **2002**, *1*(7), 515-528.

[7] Wang, Y.; Zeng, J. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics,* **2013**, *29*(13), 126-134.

[8] Wehling, M. Assessing the translatability of drug projects: what needs to be scored to predict success? *Nat. Rev. Drug Discov.,* **2009**, *8*(7), 541-546.

[9] Ashburn, T.T.; Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.,* **2004**, *3*(8), 673-683.

[10] Hurle, M.R.; Yang, L.; Xie, Q.; Rajpal, D.K.; Sanseau, P.; Agarwal, P. Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.,* **2013**, *93*(4), 335-341.

[11] Liu, Z.; Fang, H.; Reagan, K.; Xu, X.; Mendrick, D.L.; Slikker, W., Jr.; Tong, W. In silico drug repositioning: what we need to know. *Drug Discov. Today,* **2013**, *18*(3-4), 110-115.

[12] Reuters, T. Knowledge Based Drug Repositioning To Drive R&D Productivity. 2012.

[13] Delattin, N.; De Brucker, K.; Vandamme, K.; Meert, E.; Marchand, A.; Chaltin, P.; Cammue, B.P.; Thevissen, K. Repurposing as a means to increase the activity of amphotericin B and caspofungin against Candida albicans biofilms. *J. Antimicrob. Chemother.,* **2014**, *69*(4), 1035-1044.

[14] Fagan, S.C. Drug repurposing for drug development in stroke. *Pharmacotherapy,* **2010**, *30*(7 Pt 2), 51s-54s.

[15] Bellera, C.L.; Balcazar, D.E.; Alberca, L.; Labriola, C.A.; Talevi, A.; Carrillo, C. Application of computer-aided drug repurposing in the search of new cruzipain inhibitors: discovery of amiodarone and bromocriptine inhibitory effects. *J. Chem. Inf. Model.,* **2013**, *53*(9), 2402-2408.

[16] Guran, T.; Bircan, R.; Turan, S.; Bereket, A. Alopecia: association with resistance to thyroid hormones. *J. Pediatr. Endocrinol. Metab.,* **2009**, *22*(11), 1075-1081.

[17] O'Connor, K.A.; Roth, B.L. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.,* **2005**, *4*(12), 1005-1014.

[18] Gonzalez-Diaz, H.; Prado-Prado, F.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C.R.; Pazos, A.; Dea-Ayuela, M.A.; Gomez-Munoz, M.T.; Garijo, M.M.; Sansano, J.; Ubeira, F.M. MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. *J. Proteome Res.,* **2011**, *10*(4), 1698-1718.

[19] Gonzalez-Diaz, H.; Prado-Prado, F.; Sobarzo-Sanchez, E.; Haddad, M.; Maurel Chevalley, S.; Valentin, A.; Quetin-Leclercq, J.; Dea-Ayuela, M.A.; Teresa Gomez-Munos, M.; Munteanu, C.R.; Jose Torres-Labandeira, J.; Garcia-Mera, X.; Tapia, R.A.; Ubeira, F.M. NL MIND-BEST: a web server for ligands and proteins discovery--theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. *J. Theor. Biol.,* **2011**, *276*(1), 229-249.

[20] Iorio, F.; Bosotti, R.; Scacheri, E.; Belcastro, V.; Mithbaokar, P.; Ferriero, R.; Murino, L.; Tagliaferri, R.; Brunetti-Pierri, N.; Isacchi, A.; di Bernardo, D. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U S A,* **2010**, *107*(33), 14621-14626.

[21] Hu, G.; Agarwal, P. Human disease-drug network based on genomic expression profiles. *PLoS One,* **2009**, *4*(8), e6536.

[22] Wang, K.; Sun, J.; Zhou, S.; Wan, C.; Qin, S.; Li, C.; He, L.; Yang, L. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *PLoS Comput. Biol.,* **2013**, *9*(11), e1003315.

[23] Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J.; Bork, P. Drug target identification using side-effect similarity. *Science,* **2008**, *321*(5886), 263-266.

[24] Lounkine, E.; Keiser, M.J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J.L.; Lavan, P.; Weber, E.; Doak, A.K.; Cote, S.; Shoichet, B.K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature,* **2012**, *486*(7403), 361-367.

[25] Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in drug design--a review. *Curr. Top. Med. Chem.,* **2010**, *10*(1), 95-115.

[26] Minie, M.; Chopra, G.; Sethi, G.; Horst, J.; White, G.; Roy, A.; Hatti, K.; Samudrala, R. CANDO and the infinite drug discovery frontier. *Drug Discov. Today,* **2014**, *19*(9), 1353-1363.

[27] Sioud, M.; Leirdal, M. Druggable signaling proteins. *Methods Mol. Biol.,* **2007**, *361*, 1-24.

[28] Lee, H.S.; Bae, T.; Lee, J.H.; Kim, D.G.; Oh, Y.S.; Jang, Y.; Kim, J.T.; Lee, J.J.; Innocenti, A.; Supuran, C.T.; Chen, L.; Rho, K.; Kim, S. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Syst. Biol.,* **2012**, *6*, 80.

[29] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.,* **2011**, *39*(Database issue), D1035-1041.

[30] Huang, R.; Southall, N.; Wang, Y.; Yasgar, A.; Shinn, P.; Jadhav, A.; Nguyen, D.T.; Austin, C.P. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.,* **2011**, *3*(80), 80ps16.

[31] Li, Q.; Cheng, T.; Wang, Y.; Bryant, S.H. PubChem as a public resource for drug discovery. *Drug Discov. Today,* **2010**, *15*(23-24), 1052-1057.

[32] Garcia-Cordoba, F.; Garcia-Santos, J.M.; Gonzalez Diaz, G.; Garcia-Geronimo, A.; Munoz Zambudio, F.; Penalver Hernandez, F.; Del Bano Aledo, L. [Decrease of unnecessary chest x-rays in Intensive Care Unit: application of a combined cycle of quality improvement]. *Med. Intensiva,* **2008**, *32*(2), 71-77.

[33] Ihlenfeldt, W.D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical-Properties in Cactvs - an Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*(1), 109-116.

[34] Xemistry chemoinformatics. http://www.xemistry.com (**2012**).

[35] Davis, A.P.; Murphy, C.G.; Johnson, R.; Lay, J.M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B.L.; Rosenstein, M.C.; Wiegers, T.C.; Mattingly, C.J. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.,* **2013**, *41*(D1), D1104-D1114.

[36] Remmert, M.; Biegert, A.; Hauser, A.; Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods,* **2011**, *9*(2), 173-175.

[37] Xu, D.; Zhang, J.; Roy, A.; Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins,* **2011**, *79 Suppl 10*, 147-160.

[38] Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,* **2008**, *9*, 40.

[39] Rodrigues, J.P.; Levitt, M.; Chopra, G. KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res.,* **2012**, *40*(Web Server issue), W323-328.

[40] Chopra, G.; Kalisman, N.; Levitt, M. Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins,* **2010**, *78*(12), 2668-2678.

[41] Chopra, G.; Summa, C.M.; Levitt, M. Solvent dramatically affects protein structure refinement. *Proc. Natl. Acad. Sci. U S A,* **2008**, *105*(51), 20239-20244.

[42] Roy, A.; Zhang, Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure,* **2012**, *20*(6), 987-997.

[43] Chopra, G.; Hatti, K.; White, G.; Roy, A.; Samudrala, R. BINDNET: Predict BINDing NETwork between proteomes and small molecules. *unpublished*.

[44] Chopra, G.; Sethi, G.; White, G.; Samudrala, R. Multiscale modeling of complex molecular and physiological systems with application to shotgun drug repurposing for myriad diseases. . *unpublished*.

[45] Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.,* **2010**, *5*(4), 725-738.

[46] Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.,* **2007**, *35*(10), 3375-3382.

[47] Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins,* **2004**, *57*(4), 702-710.

[48] Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.,* **2013**, *41*(Database issue), D1096-1103.

[49] Hawkins, P.C.D.; Skillman, A.G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.,* **2007**, *50*(1), 74-82.

[50] Paiva, A.M.; Vanderwall, D.E.; Blanchard, J.S.; Kozarich, J.W.; Williamson, J.M.; Kelly, T.M. Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. *Biochim. Biophys. Acta,* **2001**, *1545*(1-2), 67-77.

[51] Hopkins, A.L.; Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.,* **2002**, *1*(9), 727-730.

[52] Matsushita, Y.; Araki, K.; Omotuyi, O.; Mukae, T.; Ueda, H. HDAC inhibitors restore C-fibre sensitivity in experimental neuropathic pain model. *Br. J. Pharmacol.,* **2013**, *170*(5), 991-998.

[53] Kimura, T.; Takahashi, M.P.; Okuda, Y.; Kaido, M.; Fujimura, H.; Yanagihara, T.; Sakoda, S. The expression of ion channel mRNAs in skeletal muscles from patients with myotonic muscular dystrophy. *Neurosci. Lett.,* **2000**, *295*(3), 93-96.

[54] Suzuki, S.; Satoh, T.; Yasuoka, H.; Hamaguchi, Y.; Tanaka, K.; Kawakami, Y.; Suzuki, N.; Kuwana, M. Novel autoantibodies to a voltage-gated potassium channel Kv1.4 in a severe form of myasthenia gravis. *J. Neuroimmunol.,* **2005**, *170*(1-2), 141-149.

[55] Cross, J.H.; Arora, R.; Heckemann, R.A.; Gunny, R.; Chong, K.; Carr, L.; Baldeweg, T.; Differ, A.M.; Lench, N.; Varadkar, S.; Sirimanna, T.; Wassmer, E.; Hulton, S.A.; Ognjanovic, M.; Ramesh, V.; Feather, S.; Kleta, R.; Hammers, A.; Bockenhauer, D. Neurological features of epilepsy, ataxia, sensorineural deafness, tubulopathy syndrome. *Dev. Med. Child Neurol.,* **2013**, *55*(9), 846-856.

[56] Devuyst, O.; Thakker, R.V. Dent's disease. *Orphanet J. Rare Dis.,* **2010**, *5*, 28.

[57] Chubanov, V.; Gudermann, T. TRPM6. *Handb. Exp. Pharmacol.,* **2014**, *222*, 503-520.

[58] Nagano, N. Pharmacological and clinical properties of calcimimetics: calcium receptor activators that afford an innovative approach to controlling hyperparathyroidism. *Pharmacol. Ther.,* **2006**, *109*(3), 339-365.

[59] Balabanian, K.; Levoye, A.; Klemm, L.; Lagane, B.; Hermine, O.; Harriague, J.; Baleux, F.; Arenzana-Seisdedos, F.; Bachelerie, F. Leukocyte analysis from WHIM syndrome patients reveals a pivotal role for GRK3 in CXCR4 signaling. *J. Clin. Invest.,* **2008**, *118*(3), 1074-1084.

[60] Arvisais, E.W.; Romanelli, A.; Hou, X.; Davis, J.S. AKT-independent phosphorylation of TSC2 and activation of mTOR and ribosomal protein S6 kinase signaling by prostaglandin F2alpha. *J. Biol. Chem.,* **2006**, *281*(37), 26904-26913.

[61] Kanangat, S.; Postlethwaite, A.; Hasty, K.; Kang, A.; Smeltzer, M.; Appling, W.; Schaberg, D. Induction of multiple matrix metalloproteinases in human dermal and synovial fibroblasts by Staphylococcus aureus: implications in the pathogenesis of septic arthritis and other soft tissue infections. *Arthritis Res. Ther.,* **2006**, *8*(6), R176.

[62] Moon, H.; Song, J.; Shin, J.O.; Lee, H.; Kim, H.K.; Eggenschwiller, J.T.; Bok, J.; Ko, H.W. Intestinal cell kinase, a protein associated with endocrine-cerebro-osteodysplasia syndrome, is a key regulator of cilia length and Hedgehog signaling. *Proc. Natl. Acad. Sci. U.S.A,* **2014**, *111*(23), 8541-8546.

[63] Urban, J.F., Jr.; Hu, Y.; Miller, M.M.; Scheib, U.; Yiu, Y.Y.; Aroian, R.V. Bacillus thuringiensis-derived Cry5B has potent anthelmintic activity against Ascaris suum. *PLoS Negl. Trop. Dis.,* **2013**, *7*(6), e2263.

[64] Li, D.; Lei, H.; Li, Z.; Li, H.; Wang, Y.; Lai, Y. A novel lipopeptide from skin commensal activates TLR2/CD36-p38 MAPK signaling to increase antibacterial defense against bacterial infection. *PLoS One,* **2013**, *8*(3), e58288.

[65] Sato, M.; Abe, T.; Tamai, M. Expression of the varicella zoster virus thymidine kinase and cytokines in patients with acute retinal necrosis syndrome. *Nihon Ganka Gakkai Zasshi,* **2000**, *104*(5), 354-362.

[66] Lee, N.C.; Tsai, W.Y.; Peng, S.F.; Tung, Y.C.; Chien, Y.H.; Hwu, W.L. Congenital hypopituitarism due to POU1F1 gene mutation. *J. Formos. Med. Assoc.,* **2011**, *110*(1), 58-61.

[67] Hibi, Y.; Kambe, F.; Imai, T.; Ogawa, K.; Shimizu, Y.; Shibata, M.; Kagawa, C.; Mizuno, Y.; Ito, A.; Iwase, K. Increased protein kinase A type Ialpha regulatory subunit expression in parathyroid gland adenomas of patients with primary hyperparathyroidism. *Endocr. J.,* **2013**, *60*(2), 215-223.

[68] Ikeda, M.; Fujita, T.; Mii, S.; Tanabe, K.; Tabata, K.; Matsumoto, K.; Satoh, T.; Iwamura, M. Erythema multiforme induced by sorafenib for metastatic renal cell carcinoma. *Jpn. J. Clin. Oncol.,* **2012**, *42*(9), 820-824.

[69] Shieh, P.B.; Kudryashova, E.; Spencer, M.J. Limb-girdle muscular dystrophy 2H and the role of TRIM32. *Handb. Clin. Neurol.,* **2011**, *101*, 125-133.

[70] Streker, K.; Schafer, T.; Freiberg, C.; Brotz-Oesterhelt, H.; Hacker, J.; Labischinski, H.; Ohlsen, K. *In vitro* and *in vivo* validation of ligA and tarI as essential targets in Staphylococcus aureus. *Antimicrob. Agents Chemother.,* **2008**, *52*(12), 4470-4474.

[71] Kessel, A.; Peri, R.; Perricone, R.; Guarino, M.D.; Vadasz, Z.; Novak, R.; Haj, T.; Kivity, S.; Toubi, E. The autoreactivity of B cells in hereditary angioedema due to C1 inhibitor deficiency. *Clin. Exp. Immunol.,* **2012**, *167*(3), 422-428.

[72] Balachandar, S.; Katyal, A. Peroxisome proliferator activating receptor (PPAR) in cerebral malaria (CM): a novel target for an additional therapy. *Eur. J. Clin. Microbiol. Infect. Dis.,* **2011**, *30*(4), 483-498.

[73] Kim, B.S.; Kim, Y.K.; Yun, P.Y.; Lee, E.; Bae, J. The effects of estrogen receptor alpha polymorphism on the prevalence of symptomatic temporomandibular disorders. *J. Oral Maxillofac. Surg.,* **2010**, *68*(12), 2975-2979.

[74] Canaff, L.; Zhou, X.; Mosesova, I.; Cole, D.E.; Hendy, G.N. Glial cells missing-2 (GCM2) transactivates the calcium-sensing receptor gene: effect of a dominant-negative GCM2 mutant associated with autosomal dominant hypoparathyroidism. *Hum. Mutat.,* **2009**, *30*(1), 85-92.

[75] Latus, J.; Lehmann, R.; Roesel, M.; Fritz, P.; Braun, N.; Ulmer, C.; Steurer, W.; Biegger, D.; Ott, G.; Dippon, J.; Alscher, M.D.; Kimmel, M. Analysis of alpha-klotho, fibroblast growth factor-, vitamin-D and calcium-sensing receptor in 70 patients with secondary hyperparathyroidism. *Kidney Blood Press. Res.,* **2013**, *37*(1), 84-94.

[76] Gilhar, A.; Keren, A.; Shemer, A.; Ullmann, Y.; Paus, R. Blocking potassium channels (Kv1.3): a new treatment option for alopecia areata? *J. Invest. Dermatol.,* **2013**, *133*(8), 2088-2091.

[77] Shu, H.F.; Yu, S.X.; Zhang, C.Q.; Liu, S.Y.; Wu, K.F.; Zang, Z.L.; Yang, H.; Zhou, S.W.; Yin, Q. Expression of TRPV1 in cortical lesions from patients with tuberous sclerosis complex and focal cortical dysplasia type IIb. *Brain Dev.,* **2013**, *35*(3), 252-260.

[78] Ryan, J.W.; Anderson, P.H.; Turner, A.G.; Morris, H.A. Vitamin D activities and metabolic bone disease. *Clin. Chim. Acta,* **2013**, *425*, 148-152.

[79] Santiago, R.A.; Silva, C.A.; Caparbo, V.F.; Sallum, A.M.; Pereira, R.M. Bone mineral apparent density in juvenile dermatomyositis: the role of lean body mass and glucocorticoid use. *Scand. J. Rheumatol.,* **2008**, *37*(1), 40-47.

[80] Vatankhah, A.; Assmar, M.; Vatankhah, G.R.; Shokrgozar, M.A. Immunochemical characterization of alkaline phosphatase from the fluid of sterile and fertile Echinococcus granulosus cysts. *Parasitol. Res.,* **2003**, *90*(5), 372-376.

[81] Sikjaer, T.; Rejnmark, L.; Rolighed, L.; Heickendorff, L.; Mosekilde, L. The effect of adding PTH(1-84) to conventional treatment of hypoparathyroidism: a randomized, placebo-controlled study. *J. Bone Miner. Res.,* **2011**, *26*(10), 2358-2370.

[82] Baratov, R.D.; Rasulov Kh, U.; Rustamov, B.R.; Nasyrova, R.M. [Enterokinase and alkaline phosphatase activity in children with hymenolepiasis and lambliasis]. *Med. Parazitol. (Mosk),* **1984**, *5*, 23-26.

[83] Nikodem, W. [Serum alkaline phosphatase activity in threatened abortion, prolonged pregnancy and late pregnancy toxemia]. *Pol. Tyg. Lek.,* **1981**, *36*(37), 1429-1431.

[84] Martorana, D.; Maritati, F.; Malerba, G.; Bonatti, F.; Alberici, F.; Oliva, E.; Sebastio, P.; Manenti, L.; Brugnano, R.; Catanoso, M.G.; Fraticelli, P.; Guida, G.; Gregorini, G.; Possenti, S.; Moroni, G.; Leoni, A.; Pavone, L.; Pesci, A.; Sinico, R.A.; Di Toma, L.; D'Amico, M.; Tumiati, B.; D'Ippolito, R.; Buzio, C.; Neri, T.M.; Vaglio, A. PTPN22 R620W polymorphism in the ANCA-associated vasculitides. *Rheumatology (Oxford),* **2012**, *51*(5), 805-812.

[85] Akil, O.; Hall-Glenn, F.; Chang, J.; Li, A.; Chang, W.; Lustig, L.R.; Alliston, T.; Hsiao, E.C. Disrupted bone remodeling leads to cochlear overgrowth and hearing loss in a mouse model of fibrous dysplasia. *PLoS One,* **2014**, *9*(5), e94989.

[86] Zoltowska, K.; Jablonowski, Z.; Dziekonska-Rynko, J. Trypsin and alpha-amylase activity in the pancreas of guinea pigs. V. Effects of administration of vitamin A and B 2 during the larval stage of ascariasis. *Wiad. Parazytol.,* **1991**, *37*(3), 343-350.

[87] Heffler, L.C.; Kastman, A.L.; Jacobsson Ekman, G.; Scheynius, A.; Fransson, J. Langerhans cells that express matrix metalloproteinase 9 increase in human dermis during sensitization to diphenylcyclopropenone in patients with alopecia areata. *Br. J. Dermatol.,* **2002**, *147*(2), 222-229.

[88] Hirose, T.; Matsumoto, N.; Tasaki, O.; Nakamura, H.; Akagaki, F.; Shimazu, T. Delayed progression of edema formation around a hematoma expressing high levels of VEGF and mmp-9 in a patient with traumatic brain injury: case report. *Neurol. Med. Chir. (Tokyo),* **2013**, *53*(9), 609-612.

[89] Caproni, M.; Torchia, D.; Volpi, W.; Frezzolini, A.; Schena, D.; Marzano, A.; Quaglino, P.; De Simone, C.; Parodi, A.; Fabbri, P. Expression of matrix metalloproteinases 2, 9 and 11 in erythema multiforme: immunohistochemical comparison with Stevens-Johnson syndrome/toxic epidermal necrolysis. *Br. J. Dermatol.,* **2008**, *158*(5), 1163-1166.

[90] Hayashi, T.; Yoshida, S.; Yoshinaga, A.; Ohno, R.; Ishii, N.; Yamada, T. HtrA2 is up-regulated in the rat testis after experimental cryptorchidism. *Int. J. Urol.,* **2006**, *13*(2), 157-164.

[91] Soyer, T.; Soyer, O.U.; Birben, E.; Kisa, U.; Kalayci, O.; Cakmak, M. Pepsin levels and oxidative stress markers in exhaled breath condensate of patients with gastroesophageal reflux disease. *J. Pediatr. Surg.,* **2013**, *48*(11), 2247-2250.

[92] Huang, B.; Takahashi, K.; Sakata, T.; Kiso, H.; Sugai, M.; Fujimura, K.; Shimizu, A.; Kosugi, S.; Sato, T.; Bessho, K. Increased risk of temporomandibular joint closed lock: a case-control study of ANKH polymorphisms. *PLoS One,* **2011**, *6*(10), e25503.

[93] Ruiz-Soto, R.; Richaud-Patin, Y.; Lopez-Karpovitch, X.; Llorente, L. Multidrug resistance-1 (MDR-1) in autoimmune disorders III: increased P-glycoprotein activity in lymphocytes from immune thrombocytopenic purpura patients. *Exp. Hematol.,* **2003**, *31*(6), 483-487.

[94] Munnich, S.; Taft, M.H.; Manstein, D.J. Crystal structure of human myosin 1c--the motor in GLUT4 exocytosis: implications for Ca2+ regulation and 14-3-3 binding. *J. Mol. Biol.,* **2014**, *426*(10), 2070-2081.

[95] Chopra, G.; Jenwitheesuk, E.; Lagunoff, M.; Samudrala, R. Discovery and verification of a novel broad spectrum herpes virus inhibitor. *unpublished*.

[96] Chopra, G.; Samudrala, R. Integrating clinical trails data to repurpose human approved compounds for treatment of beta thalassemia. *unpublished*.

[97] Chopra, G.; Roy, A.; White, G.; Samudrala, R. Modeling host-pathogen interactome: *Homo sapiens* and *Mycobacterium tuberculosis* interaction network predicts treatment for extreme drug resistance tuberculosis. *unpublished*.

[98] Chopra, G.; Konc, J.; Bernard, B.; Samudrala, R. A knowledge-based fragment docking with dynamics protocol for protein-ligand docking. *unpublished*.

[99] Konc, J.; Samudrala, R.; Chopra, G. Candock: Computational analytics on evolutionary interaction network based docking. *unpublished*.

Received: ?????????, 2014 Revised: ?????????, 2014 Accepted: ?????????, 2014