C²Maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships

Hui Huang1,2†, Xiaogang Wu1,2†, Ragini Pandey2, Jiao Li2, Guoling Zhao3, Sara Ibrahim4, Jake Y Chen1,2*

From IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS) 2011
San Antonio, TX, USA. 4-6 December 2011

Abstract

Background: Network pharmacology has emerged as a new topic of study in recent years. It aims to study the myriad relationships among proteins, drugs, and disease phenotypes. The concept of molecular connectivity maps has been proposed to establish comprehensive knowledge links between molecules of interest in a given biological context. Molecular connectivity maps between drugs and genes/proteins in specific disease contexts can be particularly valuable, since the functional approach with these maps helps researchers gain global perspectives on both the therapeutic profiles and toxicological profiles of candidate drugs.

Methods: To assess drug pharmacological effect, we assume that “ideal” drugs for a patient can treat or prevent the disease by modulating gene expression profiles of this patient to the similar level with those in healthy people. Starting from this hypothesis, we build comprehensive disease-gene-drug connectivity relationships with drug-protein directionality (inhibit/activate) information based on a computational connectivity maps (C²Maps) platform. An interactive interface for directionality annotation of drug-protein pairs with literature evidences from PubMed has been added to the new version of C²Maps. We also upload the curated directionality information of drug-protein pairs specific for three complex diseases - breast cancer, colorectal cancer and Alzheimer disease.

Results: For relevant drug-protein pairs with directionality information, we use breast cancer as a case study to demonstrate the functionality of disease-specific searching. Based on the results obtained from searching, we perform pharmacological effect evaluation for two important breast cancer drugs on treating patients diagnosed with different breast cancer subtypes. The evaluation is performed on a well-studied breast cancer gene expression microarray dataset to portray how useful the updated C²Maps is in assessing drug efficacy and toxicity information.

Conclusions: The C²Maps platform is an online bioinformatics resource that provides biologists with directional relationships between drugs and genes/proteins in specific disease contexts based on network mining, literature mining, and drug effect annotating. A new insight to assess overall drug efficacy and toxicity can be provided by using the C²Maps platform to identify disease relevant proteins and drugs. The case study on breast cancer correlates very well with the existing pharmacology of the two breast cancer drugs and highlights the significance of C²Maps database.

* Correspondence: jakechen@iupui.edu
† Contributed equally
1School of Informatics, Indiana University, Indianapolis, USA
Full list of author information is available at the end of the article

© 2012 Huang et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Background

Screening millions of chemical compounds to identify “hit” compounds for specific disease gene/protein targets has been a mainstream paradigm for modern drug discovery [1]. While the conventional “One disease, One gene, and One drug” paradigm [2] works effectively for simple genetic disorders, it fails to produce effective drugs for complex diseases such as cancer [3]. In complex diseases, many genes may be contributing to the disease’s phenotype; therefore, identifying a “magic bullet” drug compound can be quite elusive.

Polypharmacology, which focuses on multi-target drugs, has become a new paradigm in drug discovery. Polypharmacology drugs have conventionally been viewed to have undesirable ‘promiscuity’. However, recent research studies show, in the case of both older psychiatric drugs and modern anticancer therapies, that this promiscuity is intrinsic to the drug’s therapeutic efficacy [4]. Although there are over 40 drug-target (protein-compound interaction) databases according to Pathguide [5], (e.g. DrugBank [6], STITCH [7], CTT [8], CTD [9] and BindingDB [10], et al), a disease-specific searching platform is still needed to fully understand drug effects on the human body.

A new cancer systems biology approach to drug discovery has emerged in recent years. The primary focus of this paradigm is to understand the actions of drugs by considering targets in the context of the biological networks. By focusing on a systems level, it provides a better way to examine complicated diseases that can be caused by several gene mutations, such as cancer [11]. However, most methods published so far focus on modeling the structure of the drug target network qualitatively [12]. To examine a drug’s effect on a molecular network representative of the disease, more quantitative and accurate modeling techniques need to be developed by utilizing the concept of network pharmacology [11] or network medicine [13].

In post-genome biology, molecular connectivity maps have been proposed to establish comprehensive knowledge links between molecules of interest in a given biological context [14]. Molecular connectivity maps between drugs and genes/proteins in a disease-specific context can be particularly valuable because they allow researchers to evaluate drugs against each other using their unique gene/protein-drug association profiles. The functional approach to drug comparisons helps researchers gain global perspectives on both the toxicological profiles and therapeutic profiles of candidate drugs. Furthermore, the time it takes to develop high quality drugs in new therapeutic areas can also be reduced by using this method.

One approach for developing molecular connectivity map data is to generate disease-specific protein-drug association profiles computationally by mining bio molecular interaction networks and PubMed literature [15]. The Connectivity Maps (C²Maps) web server [16] is an online bioinformatics resource that provides biologists with potential relationships between drugs and genes/proteins in specific disease contexts based on network mining and literature mining. It’s based on the concept of network pharmacology by examining many drugs at the same time and studying the drug disease relationship based on the underlying protein interaction network instead of drugs’ direct target. C²Maps provides quantitative measurements of protein’s and drug’s relevance to a specific disease by applying networking mining and the statistical testing methods in text mining and thus offers new insight to assess overall drug efficacy and toxicity.

Occurrences between proteins and drugs from literature mining of C²Maps don’t necessarily tell research what type of relationships they have, therapeutic or toxic. To overcome these limitations, we further standardize the classifications between proteins and drugs and then perform literature curations to determine drugs’ effect on proteins on higher resolutions. Such valuable information is not readily available from the existing drug-target (protein-compound interaction) databases (e.g. DrugBank, STITCH, CTT, et al) though they may be scattered within a description or referenced text.

To assess drug pharmacological effect, such as drug efficacy and toxicity, we assume that “ideal” drugs for a patient diagnosed with a certain disease should modulate the gene expression profiles of this patient to the similar level with those in normal healthy people. Therefore, for those statistically over-expressed genes, drugs should be able to inhibit their expression level to the normal range. Similarly, for those statistically under-expressed genes, drugs should be able to activate their expression level to the normal range. In this way, drugs can treat or prevent the disease through reversing the gene expression level from disease status to the normal range, thus modulating cellular function as in normal cells.

By assuming that if the gene expression profiles of disease and drug are opposing, then the drug might be a potential treatment option of the disease, [14] identified novel drug indications in diet-induced obesity or Alzheimer’s disease. Another work by Atul [17] utilized the same gene expression data and algorithms with large scale gene expression data from GEO to study associations between 100 diseases and 164 drug molecules. They found candidate therapeutics for 53 of the diseases. These studies are proof of principle that how using public genomics database and similar hypothesis can benefit drug discovery. Though gene expression data are publicly available for more than 1000 compounds in the second release of [14], yet there are numerous compounds that are not part of the database. Another limitation of this overly simplified hypothesis lies in it doesn’t differentiate important
genes from unimportant ones. Ideally a biological meaningful scoring methods needs developed.

Drugs effect data from literatures could be complementary here. In this work, we focus on building comprehensive disease-gene-drug connectivity relationships with drug-protein directionality (inhibit/activate) information based on the C²Maps platform [16]. To show the feasibility of applying the data for computational drug discovery, we advanced previous hypothesis forward by assigning different weights to different genes. However this work aims to provide the data for the future network pharmacology research instead of developing a drug efficacy prediction method. This work has the following contributions:

1) It’s the first time that we have published this comprehensive C²Maps database server. Although [16] provides the underlying computational methodology, it only covers a small number of diseases such as Alzheimer’s Disease.

2) We create an interactive interface for directionality annotation of drug-protein pairs with literature evidences from PubMed.

3) We curate the directionality information of drug protein pairs for three disease phenotypes: breast cancer, colorectal cancer and Alzheimer disease from 5133, 19,569,563 PubMed abstracts, respectively. We also upload these curated directionality information into the C²Maps, and perform a statistical analysis on them. Curation of additional diseases, like pancreatic cancer and autism, is still on-going.

4) We enhance the functionality of disease-specific searching for relevant proteins and drugs with directionality information.

5) We update the comprehensive disease-gene-drug connectivity data in the C²Maps databases, including 19,569,563 PubMed abstracts in the current version and 142,523 unique 3 star protein interactions in the current version.

6) We also use breast cancer as a case study to demonstrate the functionality of disease-specific searching for relevant drug-protein pairs with directionality information.

7) Based on the searching result, we show the feasibility of performing drug pharmacological effect evaluation for two important breast cancer drugs to show the power of updated C²Maps in drug efficacy and toxicity assessment.

**Materials and methods**

**Data sources and systems design**

As shown in Figure 1, the C²Maps platform incorporates three major components in its systems design:

- **Network mining** component takes a query disease term as the input, and generates a ranked list of disease-relevant proteins as the output, through 1) MeSH term matching, 2) disease-associated gene searching from OMIM [18], 3) network expanding in HAPPI [19], and 4) network-based protein ranking.

  • **Text mining** component takes an input list of genes or proteins, and creates a list of enriched disease candidate drugs that are significantly associated with the disease-relevant proteins from the previous component as the output, through 1) gene/protein name mapping using UniProtKB, 2) article abstract retrieving from PubMed, 3) drug/chemical compound identification using MeSH term, and 4) disease-specific drug-protein pair ranking.

  • **Drug effect annotating** component can allow users to 1) retrieve disease-specific drug-protein association list, 2) curate drug-protein directionality information from PubMed abstract, 3) annotate these drug-protein directionality information interactively, and 4) browse disease-specific drug-protein directionality information online.

In specific, we apply the network mining method originally developed by Chen et al. [20] to fish cancer relevant proteins from the protein interaction network. We expand cancer related genes/proteins using PPIs recorded in the Human Annotated and Predicted Protein Interactions (HAPPI) database to construct cancer-specific PPI sub-network. A protein’s cancer relevance score \( r_p \) is calculated as the function (1).

\[
    r_p = k \ln \left( \frac{\sum_{q \in \text{NET}} \text{conf}(p, q)}{\sum_{q \in \text{NET}} \text{NET}^N(p, q)} \right) \tag{1}
\]

Here, \( p \) and \( q \) are indices for proteins in the cancer-related interaction network \( \text{PPI} \), \( k \) is an empirical constant (\( k=2 \) in this study), \( \text{conf}(p, q) \) is the confidence score assigned to each interaction between protein \( p \) and \( q \), and \( \text{NET}(p, q) \) holds the value of 1 if the protein \( p \) interacts with \( q \).

To retrieve the important drugs for cancers and to parse out drug terms in the articles, we acquire PubMed abstracts with the list of cancer-related genes/proteins derived earlier from PPI as queries. Drug term frequency is calculated and compared to term statistical distributions from the entire PubMed abstracts to get the p-value of drugs using function (2).

\[
    \Delta_j = \frac{df(d_j | T'_\text{NET}) - df(d_j | T'_\text{Random})}{\sqrt{\frac{\text{Var}(d_j | T'_\text{NET})}{N_{\text{NET}}} + \frac{\text{Var}(d_j | T'_\text{Random})}{N_{\text{Random}}}}} \tag{2}
\]

Here, \( T'_\text{NET} = \{T'_1, T'_2, \ldots\} \) is generated by sampling the entire collection of retrieved abstracts \( T'_\text{Random} = \{T'_1, T'_2, \ldots\} \). \( N_{\text{NET}} = |T'_\text{NET}| \) is the
size of each sample. $T_{\text{Random}} = \{T_{\text{Random}1}, T_{\text{Random}2}, \ldots\}$ refers to a random sample generated by randomly sampling the entire number of PubMed abstracts; the size of the random sample is $N_{\text{Random}}$. $df(d_j|T_{\text{NET}})$ and $df(d_j|T_{\text{Random}})$ refer to average document frequencies of $d_j$ in $T_{\text{NET}}$ and $T_{\text{Random}}$. $\text{Var}(d_j|T_{\text{NET}})$ and $\text{Var}(d_j|T_{\text{Random}})$ refer to document frequency variances of $d_j$ in $T_{\text{NET}}$ and in $T_{\text{Random}}$. A two-sided tails t-test was then performed to calculate the p-value. A thorough description of the computational components

Figure 1 CMap workflow for a given disease-specific study
and algorithms used, along with data sets and data processing parameters, is described in detail by Li et al. [16].

The C²Maps platform follows a multi-tier architecture design. The back end was implemented as PL/SQL packages in the Oracle 11 g database server, with the Oracle Text engine enabled, to ensure scalable querying of PubMed text documents. The C²Maps application middleware was implemented in the Oracle Application Express (APEX) server, which bridged between the Apache web server and the Oracle database server.

The current release of C²Maps uses the following data sets: 19,569,563 records in the PubMed/MEDLINE baseline database [21], 142,523 human protein-protein interactions above 3-star confidence ratings in the HAPPI database [19], 26,142 descriptors in the MeSH database (Category C for diseases and Category D for chemicals and drugs) [22], 20,331 entries for the curated human proteins in the UniProtKB database [23], 18,344 entities in the OMIM database [24], and 4,772 entities in the DrugBank. Current statistics for the included database records is also shown in Table 1. We manually curated the top 500 drug-protein pairs for ‘Alzheimer disease’, ‘Breast cancer’ and ‘Colorectal cancer’ from C²Maps by assigning the effects of drugs on proteins as defined in the next section. As a result, C²Maps platform contains 3928, 5133 and 4869 curated records for Alzheimer disease, Breast cancer and colorectal cancer, respectively. All data is warehoused in a local Oracle 11 g database.

**Drug effect annotation**

Since our hypothesis is that ideal drugs for a patient diagnosed with a certain disease should modulate the gene expression profile of this patient to the similar level with those in healthy people, we annotate a drug’s pharmacological effect on a protein using one of the following three categories (also illustrated in Figure 2):

- **Therapeutic**: if the drug activates the under-expressed protein or inhibits the over-expressed protein, we define that the drug has a therapeutic effect on that protein
- **Toxic**: if the drug activates the over-expressed protein or inhibits the under-expressed protein, we define that the drug has a toxic effect on that protein
- **Ambiguous**: if there is missing directionality information for either the nodes (i.e. proteins/drugs) or edges.

**Perturbation effects of drugs on proteins/genes**

We use breast cancer as an example to illustrate how we curate the directionality of a drug-protein pair retrieved from the C²Maps platform. The following are four categories of an annotated drug-protein relationship pair (also, refer to Table 2):

- **Activation**: “Subsequent injection of tamoxifen triggers the transient activation of Akt/PKB in mice.” (Tamoxifen and AKT1_HUMAN, PMID: 12640620).
- **Inhibition**: “Treatment of cells with Cycloheximide (CHX) prevented the activation of p53 in all phases of the cell cycle and its accumulation in G1/S and S.” (P53_HUMAN and Cycloheximide, PMID: 9484835).
- **Indirect Yes**: “Hydroxyurea-mediated DNA synthesis arrest of S phase MCF7 cells led to a loss of BRCA1 from these structures.” (BRCA1_HUMAN and Hydroxyurea, PMID: 9267023).
- **Ambiguous**: “GRalpha and GRbeta transcripts are coordinately upregulated in CEM-C7 cells and coordinately downregulated in IM-9 cells by dexamethasone.” (GCR_HUMAN and Dexamethasone, PMID: 12974663).

We can see that the literature does contain such information and thus provides basis for the directionality information retrieval. We focus on curating drug actions in the disease context rather than only in cell lines like in [14]. Different research works under specific contexts may produce conflicting conclusions regarding drug protein relationship. Take Tamoxifen and estrogen receptor as examples. As shown in Table 3, we successfully extracted 7 article abstracts which support the inhibitory effects of Tamoxifen on estrogen receptor and 2 PubMed abstracts which support the stimulatory effects of Tamoxifen on estrogen receptor. The pre-dominant evidence showing Tamoxifen’s inhibition on estrogen receptor [25] matches

| Dataset                          | Data Resource           | Record count |
|----------------------------------|-------------------------|--------------|
| Biomedical Literature            | PubMed                   | 19,569,563   |
| Human Protein-Protein Interaction | Unique HAPPI 3-star interactions | 142,523     |
| Disease and Drug Terminology     | MeSH descriptors         | 26,142       |
| Human Protein                    | UniProtKB                | 20,331       |
| Disease-Gene relationships       | OMIM                     | 18,344       |
| Drug Information                 | DrugBank                 | 4,772        |
well with the fact that Tamoxifen acts as an antagonist for estrogen receptor. Beside checking the majority vote of all the related papers, we also traced back the original references. For Tamoxifen, it inhibits estrogen receptor in the mammary tissue while activating estrogen receptor in bone density. In our breast cancer case study, we decided Tamoxifen inhibits estrogen receptor because the gene expression experiment was based on breast tissue. In the future, we plan to add additional contexts such as experimental conditions, disease subtypes and so on. In the current version, they are not added due to limited availability of those data in abstracts.

### Data access and website usage

The C²Maps online platform (http://bio.informatics.iupui.edu/cmaps) provides researchers a web-based bioinformatics user interface, following principles described in [26]. As shown in Figure 3, users can begin with a single disease term as a query and navigate to extract significant subsets of the disease-specific C²Maps. We also show snapshots of C²Maps output web pages in Figure 4, using “Breast Cancer” as a query example. The statistically significant relationships between proteins and drugs that are extracted from the literature are then displayed in tabular format (Figure 4). The proteins, drugs, and evidence numbers are further linked to protein, drug, and the article detail page, respectively. The search results can also be sorted by the protein ranking score (R-Score) and Chemical/Drug significance (P-Value). In addition, the page also lists Disease Context (disease name of user interest) and Disease Terminology (disease name containing query term in the controlled vocabulary of Medical Subject Headings). Users can also specify advanced search criteria for further biological/pharmacological analysis. Annotations for the extracted relations could be performed from the ‘Annotate Data’ tab.

### Browsing disease-specific drug-protein relationship information

Any public C²Maps database user can access the well curated drug-protein directionality data. The database will display disease, protein, drug directionality, and PubMed evidence for each record. Each column can be sorted or filtered. One can also display only drug-protein directionality belonging to certain disease by selecting it from the drop down list (Figure 5b). Furthermore, the user can also search the keywords, such as protein name or drug name, to retrieve only specific records. Currently, the C²Maps platform contains 3928 curated records for Alzheimer disease, 5133 curated records for breast cancer, and 4869 curated records for colorectal cancer. More curation information will be updated regularly.

### Table 2 Curation of drug-protein relations from Pub-Med abstracts

| Relation       | Protein    | Drug            | PMID     | Provenance                                                                 |
|----------------|------------|-----------------|----------|-----------------------------------------------------------------------------|
| Up-Regulated   | BRCA1      | Estradiol       | 7553629  | "BRCA1 mRNA and protein levels were significantly decreased in estrogen-depleted MCF-7 and BT20T cells and increased again after stimulation with beta-estradiol". |
| Down-Regulated | PS3        | Cycloheximide   | 9484835  | "Treatment of cells with cycloheximide (CHX) prevented the activation of p53 in all phases of the cell cycle and its accumulation in G1/S and S". |
| Indirect       | BRCA1      | Hydroxyurea     | 9267023  | "Hydroxyurea-mediated DNA synthesis arrest of S phase MCF7 cells led to a loss of BRCA1 from these structures". |
| Ambiguous      | GCR        | Dexamethasone   | 12974633 | "GRalpha and GRbeta transcripts are coordinately upregulated in CEM-C7 cells and coordinately downregulated in IM-9 cells by dexamethasone". |
| Unknown        | AKT1       | Phosphothreonine| 11087733 | The drug-protein relation is not mentioned in the text.                      |
Interactive interface for directionality annotation

An authorized C2Maps database user can also annotate selected C2Maps contents by performing manual curation from the ‘Annotate Data’ tab. The user may apply for an annotator’s account to edit protein-drug interactions suggested by the C2Maps automated recommendation system. This editing is provided through a separate user interface that enables the annotator to categorize protein-drug relationships as direct (including activation, inhibitory, ambiguous), indirect, or unknown (Figure 6f). The user will first select the assigned disease and the C2Maps web-server will populate disease relevant protein and drug pairs. All the PubMed abstracts mentioning both the relevant protein and drug will be pulled out and the curator can read the abstract to annotate the directionality between the drug and the protein. The user can also edit (Figure 6d) each record or delete it.

Results

Statistical analysis for reliability

The major computational components of the C2Maps platform were developed using validated computational techniques. In the network mining component, protein interaction network expansion was able to reduce the initial biases and low data coverage, which may have existed in the seed list of protein. We used the new HAPPI database instead of other protein interaction databases because of its overall better data quality (comparable or better than data in the HPRD database for quality star grades of 3 and above) and coverage (more than 280,000 human protein interactions with star grades of 3 and above).

Table 3 PubMed evidence for Tamoxifen’s effect on ESR1

| Drug   | Protein     | PMID     | Direction |
|--------|-------------|----------|-----------|
| Tamoxifen | ESR1_HUMAN | 14507640 | -1        |
| Tamoxifen | ESR1_HUMAN | 2359140  | -1        |
| Tamoxifen | ESR1_HUMAN | 14759988 | -1        |
| Tamoxifen | ESR1_HUMAN | 11774281 | -1        |
| Tamoxifen | ESR1_HUMAN | 2137212  | -1        |
| Tamoxifen | ESR1_HUMAN | 9328205  | -1        |
| Tamoxifen | ESR1_HUMAN | 11261829 | -1        |
| Tamoxifen | ESR1_HUMAN | 12767276 | 1         |
| Tamoxifen | ESR1_HUMAN | 11812086 | 1         |

1 represents that the drug up regulates the protein while -1 represents down regulation.

Figure 3 The navigational site map of the CMap platform. The numbers refer to: 0. Configure search preference; 1. Input interest disease; 2. Generate Protein-Drug connectivity; 3. Refine initially constructed connectivity map; 4. Link to protein page; 5. Link to drug page; 6. Link between protein and drug pages; 7. Link to evidence article pages; 8. Link to external data resources (MeSH, UniProt and PubMed); 9. Import enriched disease specific protein-drug associations for further annotation; 10. Authorized users (curators) set up profiles and login in; 11. Annotate effects of drugs on protein/genes; 12. Release annotation results; 13. Browse annotated disease-protein-drug connectivity relationships; 14. Filter or search for interested subset of connectivity relationships.
above), which was thoroughly described in Chen et al. [19]. In the text mining component, the PubMed abstract retrieval for each protein was shown to improve Information Retrieval (IR) recall performance without sacrificing precision. The quality of disease drug identifications was shown to outperform comparable systems with balanced sensitivity, specificity, and positive predictive values (for details, refer to Li et al. [16]).

In Table 4, we show a summary of C²Maps platform performance, by comparing its overall sensitivity, specificity, PPV (positive predictive value), F-score, and ACC (accuracy) measures among a number of cancers. The result confirmed that C²Maps performed well consistently across different disease studies.

A case study on breast cancer specific searching for relevant drug-protein pairs with directionality information

We evaluated breast cancer drugs from C²Maps based on our hypothesis. First, we obtained top 500 drug protein pairs for breast cancer from the C²Maps web server,
23 drugs and 103 proteins, respectively. The $r_p$ scores for those proteins range from 1.69 to 169.82 and the P-Values for those drugs are all below 0.05. Well known breast cancer related proteins, like BRCA1, or related drugs, like tamoxifen, were included in these 500 pairs. We then examined all supporting evidence for each drug protein relations, a total of 5,225 PubMed abstracts, and manually curated them to extract the relevant drug effect information. Out of those 500 pairs, 155 pairs contained information of how the drug affects the protein in the literature, yielding a total of 19 drugs and 52 proteins. After performing manual curation, 79 drug protein pairs contained only up-regulation information, 57 only down-regulation information, 11 primarily up-regulation information, and 8 primarily down-regulation information. The distribution of directionality categories for breast cancer is shown in Figure 7a. A subnetwork based on the directionality information specific for Tamoxifen can be constructed from C2Maps directionality data (shown in Figure 7b). Another subnetwork based on the directionality information specific for Plicamycin is also shown in Figure 7e.
Drug efficacy can be measured by the ability of a drug to produce the desired phenotypic effect or molecular effect. To evaluate drug efficacy in the molecular level based on our hypothesis illustrated in Figure 2, we need to know how drugs can affect the expression of its interacting genes and how those genes are expressed in disease conditions. We have got the former from the above case study of $C^2$Maps. To get the latter, we performed differential analysis on a well-studied microarray
dataset-GSE3191[27]. This experiment contains breast cancer subtype luminal A, basal-like and also normal breast tissues. We obtained the differential genes for both breast cancer subtypes - luminal A and basal-like when compared to normal. We identified 579 differential genes between luminal A and normal, 773 differential genes between basal like and normal. We used these two sets for the following case study.

Tamoxifen efficacy and toxicity assessment for the breast cancer subtype - luminal A
Tamoxifen is a standard drug clinically used for breast cancer and has 15 interacting proteins with directionality annotations from C2Maps (shown in Table 5). We intersected differential genes from luminal A microarray experiment with Tamoxifen’s interacting partners from C2Maps. Four proteins out of 15 are differentially expressed between luminal A and normal, including ESR1. In Figure 7c, drugs are represented as hexagons and proteins as circles. The size of a protein node is proportional to the rp score, an indication of importance of this protein to breast cancer. Red nodes stand for over-expressed proteins in breast cancer while green ones represent under-expressed proteins. For edges between drugs and proteins, red symbolizes that the drug activates the protein while green symbolizes inhibition. From the Figure, Tamoxifen has 3 therapeutic effects: it inhibits over-expressed ERBB2, which might help explain the clinical fact that basal or triple negative breast cancer patients fail to benefit from Tamoxifen treatment.

Plicamycin efficacy and toxicity assessment for the breast cancer subtype - luminal A
Plicamycin was an approved antineoplastic antibiotic for a variety of advanced forms of cancer. It has been withdrawn from market in 2000. In Figure 7f, we showed the drug protein interaction for Plicamycin in Luminal A patients. It has 2 interacting proteins with directionality annotations (shown in Table 5) and both are not significant in breast cancer with a low rp score. Only 1 protein out of these 2 is differentially expressed between luminal A and normal. Plicamycin has a toxic effect overall by inhibiting under-expressed MYC. This implied a neutral or toxic effect when using Plicamycin in Luminal A subtype breast cancer patients since it is not able to reverse its interacting proteins in basal condition (Figure 7d). This agrees well with the clinical fact that basal or triple negative breast cancer patients fail to benefit from Tamoxifen treatment.

Conclusions
In this study, we present an upgraded C2Maps platform to evaluate drug pharmacological effects based on the hypothesis that an ideal drug can reverse the gene expression level in a disease back to those in normal conditions. This online platform will enable users to query high-coverage protein-drug connectivity maps in real time. It enables users to research up-to-date knowledge of connectivity maps for a specific disease, explore therapeutic protein targets, design repurposed drug compounds, and assess toxicological impacts of drug compounds on disease-relevant genes/proteins. Three efficacy case studies prove the feasibility to apply the literature mined drug directionality data from C2Maps for drug efficacy study. It will be a major resource to biomedical researchers.
interested in developing disease-specific therapeutic and diagnostic applications based on progresses in network biology and network pharmacology.

Discussion
From the case study on breast cancer drug effect evaluation, we can see that there is still room for improvement, although the two breast cancer drugs were well-evaluated. The information for judging whether a drug has global therapeutic effects on other diseases is limited due to the manual curation procedure. This information is very valuable for drug repurposing, however. In the current version of C²Maps, drug-protein pairs are mainly come from literature mining based on disease-gene

Figure 7 Breast cancer case study for drug pharmacological effect evaluation with C²Maps. (a) Distribution of directionality categories for breast cancer; (b) A drug-target subnetwork with directionality information specific for Tamoxifen; (c) Drug effect evaluation for Tamoxifen on breast cancer subtype - luminal A; (d) Drug effect evaluation for Tamoxifen on breast cancer subtype - basal-like; (e) A drug-target subnetwork with the directionality information specific for Plicamycin; (f) Drug effect evaluation for Plicamycin on breast cancer subtype - luminal A.
searching and network mining, which can be supplemented by plenty of publicly-available drug target databases. We will continue to update C2Maps and improve its usability through achieving the following functionalities in the near future.

1) We will increase the functionality of drug-orientated searching for relevant disease phenotypes and proteins in the C2Maps. It will allow users to input drug names, not just disease names. It should be able to retrieve all the disease names and genes/proteins related to this drug. This function will be very useful for drug repurposing.

2) We will increase the functionality of disease-orientated browsing for relevant proteins and drugs in the C2Maps by using disease phenotype trees. It will allow users to browse the database by clicking the disease name. The current version only supports the disease-specific searching function without any browsing function.

3) We will also enhance the functionality of interactive directionality information annotation for drug-protein pairs in the C2Maps by using natural language processing (NLP) techniques. The literature curation for breast cancer, colorectal cancer and Alzheimer disease took three experts nearly one year’s effort to complete. While this ensures the data quality, it’s time consuming. With those golden standard dataset from curation, we will NLP techniques to allow users to curate and annotate directionality information from PubMed abstracts more easily and semi-automatically.

Table 5 Tamoxifen relevant proteins and their directionality

| Drug     | Protein    | RpScore | Association | Direction |
|----------|------------|---------|-------------|-----------|
| Tamoxifen| AKT1_HUMAN | 82.99   | 1.77        | 1         |
| Tamoxifen| BRC2A2_HUMAN | 21.14   | 3.29        | 1         |
| Tamoxifen| CADH1_HUMAN | 17.57   | 0.95        | 1         |
| Tamoxifen| CDK2_HUMAN | 2.94    | 2.12        | 1         |
| Tamoxifen| E2F1_HUMAN | 2.95    | 1.59        | 1         |
| Tamoxifen| ESRB2_HUMAN | 2.07    | 3.19        | 1         |
| Tamoxifen| ESRI_HUMAN | 72.39   | 5.11        | -1        |
| Tamoxifen| FRS1_HUMAN | 2.51    | 1.29        | -1        |
| Tamoxifen| JUN_HUMAN | 2.91    | 1.51        | 1         |
| Tamoxifen| MYC_HUMAN | 3.49    | 2.5         | 1         |
| Tamoxifen| NCOA3_HUMAN | 2.61    | 3.01        | -1        |
| Tamoxifen| NCOR1_HUMAN | 2.81    | 4.05        | 1         |
| Tamoxifen| PS3_HUMAN | 169.82  | 0.8         | -1        |
| Tamoxifen| P53A_HUMAN | 2.92    | 1.57        | 1         |
| Tamoxifen| PTEN_HUMAN | 3.98    | 0.98        | 1         |
| Plicamycin| MYC_HUMAN | 3.49    | 4.16        | -1        |
| Plicamycin| SPI1_HUMAN | 3.32    | 6.24        | -1        |

1 represents that the drug up regulates the protein while -1 represents down regulation.

Acknowledgements

Based on “Predicting drug efficacy based on the integrated breast cancer pathway model”, by Hui Huang, Xiaorang Wu, Sara Ibrahim, Marianne McKenzie and Jake Y Chen which appeared in Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on. © 2011 IEEE [28].

We would like to thank the IUPUI solution center, MURI, UROP and Indiana Center for Systems Biology and Personalized Medicine for financial support. We thank Dr. Davide Bolchini for his guidance on web site usability. We also thank Sina Reinhard for her curation on Alzheimer’s disease, and Taiwo Ajumobi for her curation on autism.

This article has been published as part of BMC Genomics Volume 13 Supplement 6, 2012. Selected articles from the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS) 2011. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S6.

Author details

1 School of Informatics, Indiana University, Indianapolis, USA. 2 Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, USA. 3 Department of Computer Science and Technology, Shandong College of Electronic Technology, Jinan, China. 4 Biology Department, School of Science, Indiana University, Indianapolis, USA.

Authors’ contributions

JYC conceived this work, guided the research team by providing ideas and feedback along the way, and revised the manuscript. HH involved the drug efficacy evaluation hypothesis development, carried out annotation web page development, designed the case studies and wrote the manuscript. XW participated in the drug efficacy hypothesis initiation, directionality concept development, case studies and manuscript writing. RP updated the Pubmed database and tables in the Apex application. JL developed the previous version of C2Maps and performed case studies about the web server in her paper. GZ helped with annotation web page development. SJ performed the directionality curations for breast cancer and colorectal cancer. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 26 October 2012

References

1. Mestres J: Computational chemogenomics approaches to systematic knowledge-based drug discovery. Curr Opin Drug Discov Devel 2004, 7(3):304-313.

2. Durrant JD, Amaro RE, Xie L, Urbaniak MD, Ferguson MA, Haapalainen A, Chen Z, Di Guilmi AM, Wunder F, Bourne PE, et al. A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. Pacs computational biology 2010, 6(1): e1000648.

3. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M: Drug-target network. Nat Biotechnol 2007, 25(10):1119-1126.

4. Hopkins AL: Drug discovery: Predicting promiscuity. Nature 2009, 462(7270):167-168.

5. [http://www.pathguide.org/].

6. Knox C, Law V, Jewison T, Liu P, Ly S, Folkis A, Pon A, Banco K, Mak C, Neveu V: DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs. Nucleic acids research 2011, 39(suppl 1):D1035.

7. Kuhn M, Saliarzcky D, Franceschini A, von Mering C, Jensen LJ, Bork P: STITCH 3: zooming in on protein-chemical interactions. Nucleic acids research 2012, 40(D1):D267-268.

8. Zhu F, Han BC, Kumar P, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ: Update of TTD: therapeutic target database. Nucleic acids research 2010, 38(suppl 1):D787-D791.
9. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenberg MC, Wiegens TC, Mattigny CJ. Comparative Toxicogenomics: Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic acids research 2009; 37(Database issue):D786-D792.
10. Liu T, Lin Y, Wien X, Jorsives RN, Gilson, MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic acids research 2007, 35(Database issue):D198-D201.
11. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nature chemical biology 2008, 4(11):682-690.
12. Bugrim A, Nikolaikaya T, Nikolaisky Y. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. Drug Discov Today 2004, 9(3):127-135.
13. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics 2011, 12(1):56-68.
14. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006, 313(S795):1929.
15. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc 2006, 349-353.
16. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS computational biology 2009, 1(7):e1000450.
17. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science translational medicine 2011, 3(116):36a77.
18. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research 2005, 33(Database issue):DS14-DS17.
19. Chen JY, Mamidipalli S, Huan T. HAPPI: an online database of comprehensive human annotated and predicted protein interactions. BMC genomics 2009, 10(Suppl 1):S16.
20. Chen JY, Shen C, Svachenko AV. Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pacific Symposium on Biocomputing 2006, 367-378.
21. Shi Z, Li H. Application of artificial neural network approach and remotely sensed imagery for regional eco-environmental quality evaluation. Environ Monit Assess 2007, 128(1-3):217-229.
22. Sakamoto M, Kondo A, Kawasaki K, Goto T, Sakamoto H, Miyake K, Koyamatsu Y, Akiya T, Iwabuchi H, Muroya T, et al. Analysis of gene expression profiles associated with cisplatin resistance in human ovarian cancer cell lines and tissues using cDNA microarray. Hum Cell 2001, 14(4):305-315.
23. Wu CH, Apweiler R, Barooshian C, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic acids research 2006, 34(Database issue):D187-191.
24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research 2005, 33(Database issue):D514-D517.
25. Blackwell KL, Haroon ZA, Shan S, Saito W, Broadwater G, Greenberg CS, Dewhurst MW. Tamoxifen inhibits angiogenesis in estrogen receptor-negative animal models. Clinical Cancer Research 2000, 6(11):4359.
26. Balchin D, Finkelstein A, Perrone V, Nagl S. Better bioinformatics through usability analysis. Bioinformatics 2009, 25(3):406-412.
27. Sortie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America 2001, 98(19):10869-10874.
28. Huang H, Wu X, Ibrahim S, McKenzie M, Chen JY. Predicting drug efficacy based on the integrated breast cancer pathway model. Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on: 4-6 December 2011, 42-45.