nature
chemical biology

# Analysis of drug-induced effect patterns to link structure and side effects of medicines

Anton F Fliri, William T Loging, Peter F Thadeio & Robert A Volkmann

**The high failure rate of experimental medicines in clinical trials accentuates inefficiencies of current drug discovery processes caused by a lack of tools for translating the information exchange between protein and organ system networks. Recently, we reported that biological activity spectra (biospectra), derived from *in vitro* protein binding assays, provide a mechanism for assessing a molecule's capacity to modulate the function of protein-network components. Herein we describe the translation of adverse effect data derived from 1,045 prescription drug labels into effect spectra and show their utility for diagnosing drug-induced effects of medicines. In addition, notwithstanding the limitation imposed by the quality of drug label information, we show that biospectrum analysis, in concert with effect spectrum analysis, provides an alignment between preclinical and clinical drug-induced effects. The identification of this alignment provides a mechanism for forecasting clinical effect profiles of medicines.**

One of the key functions of preclinical drug discovery is the fine-tuning of experimental medicines for modulating the information flow in cellular protein networks and relating these changes to disease intervention. The high failure rate of drug candidates in clinical trials, however, accentuates inefficiencies of current processes and implicates as main cause the incomplete translation of drug-induced effects on proteins into medically useful effects on organisms[1]. The misalignment between preclinical and clinical drug-induced effects is due, in part, to the remarkable ability of organisms to compensate for the loss or decline in function of specific proteins by rerouting the information flow in protein networks[2]. At the organism level, protein network perturbations, caused by inhibition or stimulation of the function of individual network components, become visible as a pattern of physical symptoms; it does not matter whether protein network perturbations are caused by a disease or by the administration of a medicine[3–6].

In spite of these complexities, the high costs associated with failure of experimental medicines in clinical trials underscore the need to improve methods for translating drug-induced effects on proteins into drug-induced effects on whole organisms[7,8]. Achieving this goal is a formidable challenge because preclinical methods for structure-function analysis focus on determination of structure-effect relationships of single protein network components and not on information exchange between protein and organ systems networks. In addition, no precise methods exist for comparing drug-induced effect information of medicines obtained in clinical trails. Hence, the induction of drug effects in clinical trials is highly variable and depends on age, sex, physical condition, genetic variance in drug targets, regulation of disease pathways, differences in metabolizing enzymes, dosage forms, and routes of drug administration. In fact, different dosages of drugs and routes of administration have been shown to not only affect the magnitude of a clinical response but also the specific nature of a response[9]. Complicating the translation of drug-effect observations between different organisms, drug effects may vary from organism to organism[10], and drugs differentiated on the basis of *in vitro* information may produce similar *in vivo* effects, but through entirely different mechanisms[11,12]. Thus, quantitative comparisons of drug effects between different medicines can generally only be made in clinical trials where drug exposure and methods for ascertaining biological effects have been defined. Notwithstanding these constraints, we explore herein whether identification of specific drug-inducible effect patterns using pattern recognition tools can provide information on whole-organism, structure-response relationships of medicines[13,14]. The experimental design of these investigations is shown in the graphical abstract and in **Supplementary Figure 1** online.

Recently we described the utility of preclinical drug-induced effect patterns for investigating broad structure-response relationships. This analysis method is based on the use of percentage inhibition values, which we determined for medicines at a single, high concentration (10 μM) for ninety-two proteins, representing a cross-section of the ligand-accessible ('druggable') section of the proteome (**Supplementary Table 1** and **Supplementary Fig. 2** online)[15,16]. The translation of percentage inhibition values into biological activity spectra (biospectra) is used for expressing a medicine's probability to induce a certain pattern of protein network perturbations. Molecular property descriptors generated from biospectra take advantage of the principle of neighborhood behavior; measurements of the interaction of medicines with individual screening targets also identify the probability that these medicines will interact with other members of the gene families represented by the ninety-two protein assays[16]. This neighborhood

Pfizer Global Research and Development, Eastern Point Road, Groton, Connecticut 06340, USA. Correspondence should be addressed to A.F.F. (anton.fliri@pfizer.com) or R.A.V. (robert.a.volkmann@pfizer.com).

Received 28 June; accepted 19 October; published online 6 November 2005; doi:10.1038/nchembio747

information is encoded in the shape of the spectrum (biospectrum). Using spectral representations of proteome-centered molecular property descriptors allows quantification of similarities between biospectra (biospectral analysis) by applying principles of spectroscopy[15]. We have previously shown that the comparison of biospectra yields precise chemical structure information[15] and identifies pharmacological similarities between medicines[16]. Accurate assessment of structure-function similarities between medicines[17] does not depend on information from putative drug targets but rather on the discriminative properties of molecular property descriptors[16]. Although the exact cause of this perplexing observation is not certain[18], it is notable that biospectra, derived from percentage inhibition values determined for 92 proteins at 10-μM drug concentrations, have proven effective for identifying agonist and antagonist effect profiles of medicinal agents even in the absence of information on putative drug targets[16].

## RESULTS

### Comparison of clinical drug-induced effect spectra

Starting from the premise that the combined pharmacology of currently prescribed medicines targets virtually every organ system in the body, we examined if placebo-controlled drug-induced effect (side effect) information appearing on drug labels could be used to construct molecular property descriptors that encode chemical structure and organism response information. Exploration of the feasibility of this approach relies on clinical side-effect data (extracted from commercial drug labels) listed in the CEREP BioPrint database[19] in the form of COSTART (coding symbols for thesaurus of adverse reaction terms) codes[20]. The scope of this investigation was limited to side-effect information on 1,045 compounds, selected purely on the basis of the availability of clinical effect data for these medicines. Because of the potential for classification error in hierarchical clustering due to reporting bias of frequency and severity information on drug labels, we converted side-effect data listed in the BioPrint database into binary effect descriptor sets. These binary drug effect representations allowed inclusion of seemingly contradictory effects or effects noted in the 'Frequency unknown' or 'Post-marketing reports' categories of drug labels. Five hundred ninety-one effect categories were selected. A value of 1 was assigned to COSTART fields listing placebo-controlled frequency and severity information on side effects. A value of 0 was assigned if a particular effect was not documented in any one of the 591 effect categories[21]. These descriptor sets were entered into hierarchical clustering using Ward's method and 'row average' as ordering function[22]. The classification results obtained with 1,045 compounds and a binary effect descriptor set consisting of 591 side effect categories (dataset I) are shown in **Figure 1**. The similarity values obtained in the formation of the *y*- and *x*-axis dendrograms (**Fig. 1**) are the products of hierarchical clustering methods (see Methods). Thus, the *y*-axis dendrogram (**Fig. 1a**) disperses 1,045 binary effect spectra into smaller groups, each containing medicines with similar effect pattern. The similarity between effect patterns is measured using node similarity values separating branches of the dendrogram. The scale of these node similarity values varies from 0 to 4,000, with 0 being most similar. Derived from the binary descriptor sets, these node similarity values depend on the number of side effects reported on the label of a medicine (information density). On the *x*-axis (**Fig. 1a**) appear clusters of symptoms that characterize drug-induced effects on the human body. Again, decreasing values in node similarity indicate increasing confidence that a particular symptom group is associated with the human body's response to medications.

### Hierarchical clustering of binary drug effect descriptors

The *x*-axis dendrogram (**Fig. 1**) identifies symptom association characterizing the human body's response to medication. The two side-effect clusters (indicated by the red numbers 1 and 2 in **Fig. 1a**,**b**) represent a list of the most prevalent side effects appearing on the 1,045 drug labels. Symptom cluster 1 (far right of **Fig. 1b**) describes drug effects on the gastrointestinal tract (GI). For example, nausea is listed on 651 out of 1,045 drug labels. The frequency of side effects associated with nausea in this grouping is shown in parenthesis: emesis (526), headache (613), dizziness (550), asthenia (451), rash (650), diarrhea (524) and abdominal pain (402). Combining associations of these symptoms creates a diagnostic pattern. For example, 492 drug labels list both nausea and emesis as most common side effects (47% of the 1,045), and 130 drug labels list all eight symptoms as drug-induced effects (12% of the 1,045). The next most prevalent symptom cluster (red number 2; the adjacent *x*-axis grouping in **Fig. 1b**) describes drug-induced effects on the immune system: allergic reactions (471), urticaria (411), anaphylactic reactions (274), pruritus (380), parasthesia (354), dyspepsia (345), edema (305), fever (339), myalgia (270), arthralgia (220), leukopenia (360), thrombocytopenia (339), anorexia (345), malaise (261), alopecia (249), anemia (234) and liver function abnormalities (234). Again, these symptom clusters create a diagnostic pattern. For example, 76 drug labels (7% of the 1,045) list all eight GI symptoms in combination with at least one of the symptoms describing effects on the immune system.

### Effect of information density on *x*-axis classifications

To investigate the effect of information density on *x*-axis classifications of binary effect descriptors, we created six different effect descriptor sets (I–VI) by changing descriptor length and data density (see Methods). To assess if symptom classifications have diagnostic utility and provide meaningful (reproducible) portrayals of whole-organism responses, we monitored whether symptom associations would remain intact using each of the six binary effect descriptors in independent classifications. For example, symptom associations such as palpitation, tachycardia, sweat, hypertension, vasodilatation and hypotension, representing a section of the *x*-axis dendrogram (**Fig. 1a**), maintained coherence using descriptor sets I–VI. Similarly, other symptom associations characterizing organ-system specific effects maintained coherence in each of these independent classifications. For example, the medicines known as tricyclic antidepressants (trimipramine (**5**), nortriptyline (**6**), protriptyline (**7**) and imipramine (**8**)) show a characteristic side-effect pattern and group together (section YY in **Fig. 1b**). Only a portion of the effect profile is shown (**Fig. 1b**). The *y*-axis dendrogram section XX (**Fig. 1b**) shows a cluster of medicines containing the muscarinic antagonist homatropine (**2**). As indicated previously, this section of the *y*-axis dendrogram (**Fig. 1b**) lists medicines with side-effect profiles most similar to homatropine (**2**)**.** These medicines include the two muscarinic antagonists methylhomatropine (**1**) and cyclopentolate (**4**) along with the pharmacologically distinct α adrenoreceptor agonist, dipivefrin (**3**)**.** All four of these are used in ophthalmic preparations[23] and list conjunctivitis on the drug label as a drug treatment–related effect (**Fig. 1b**). The ophthalmic α adrenoreceptor agonist dipivefrin (**3**), which is a prodrug cleaved by acetylcholinesterases into epinephrine, has a side-effect profile similar to ophthalmic muscarinic antagonists (compounds **1**, **2** and **4**)[23], which is unexpected considering its short half-life *in vivo*[23]. In light of observations indicating that dipivefrin (**3**) is rapidly systemically absorbed upon topical ocular administration, ophthalmologic preparations of it would be expected to have a side-effect profile similar to that of epinephrine, which

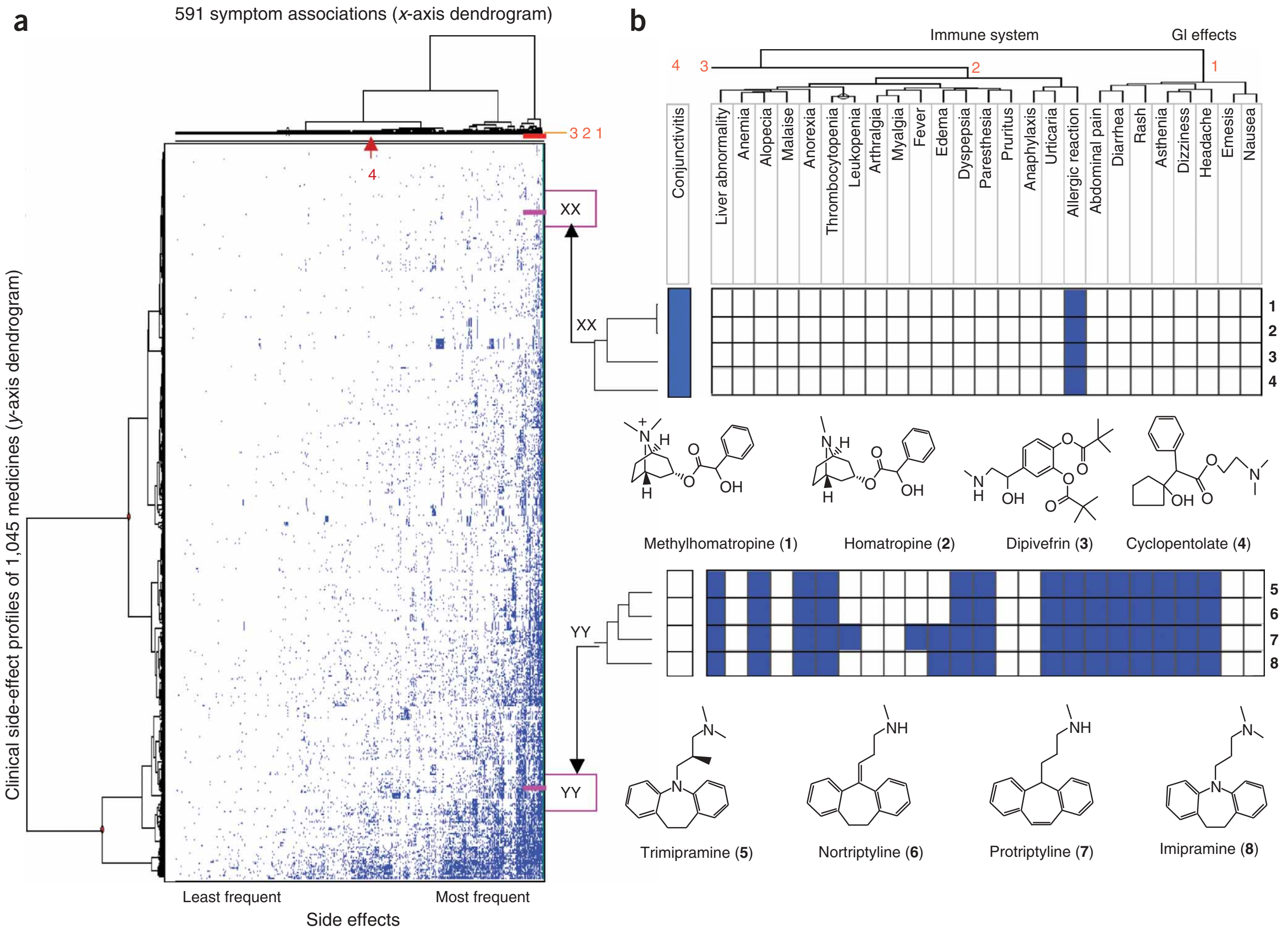

**Figure 1** Classification of side effect information for 1,045 medicines. (**a**) Hierarchical clustering of 1,045 binary-encoded clinical effect spectra, consisting of 591 side effects, using Ward's Method[22]. The dendrogram at left divides these 1,045 medicines into clusters using node similarity values to identify effect spectrum similarity. The branching of this dendrogram reflects similarity between classifications. This similarity scale ranges from 0–4,000; lower values translate into the greatest effect spectrum similarity. The colors indicate presence of side effects (populated COSTART fields are blue) or absence of side effects (white). The *y*-axis dendrogram layout indicates that these classifications depend on side-effect profile similarity and the frequency of side-effect information reported for medicines (fewest side effects at top left; most side effects at bottom right). The *x*-axis dendrogram identifies how often individual symptoms appear in associations. (**b**) A portion of the dendrogram (**Fig. 1a**) using 26 drug-induced effect categories (containing the most common side effects of medicines) to visualize the discrete side-effect pattern of anticholinergic (cluster XX, containing medicines **1**–**4**) and antidepressant medicines (cluster YY, containing medicines **5**–**8**).

resides with other adrenergic drugs in cluster A (**Fig. 2a**). Side-effect profile comparison demonstrates that local and systemic effects of dipivefrin (**3**) are differentiable from those elicited by the ocular administration of epinephrine.

## Information density effect on side effect classification

The *y*-axis dendrogram (**Fig. 1**) shows that medicines are classified according to similarity and frequency of side-effect information appearing on respective drug labels. At the top of the *y*-axis dendrogram (**Fig. 1a**) are drug clusters that have few side effects (with most of the 591 COSTART fields having values of 0) and at the bottom of the *y*-axis dendrogram are clusters of drugs with numerous reported effects. For example, anticholinergic compounds **1**–**4** residing in cluster XX (**Fig. 1a,b**), have similar (but few) side effects, whereas antidepressants **4**–**8** (cluster YY) have multiple side effects. Medicines on proximal branches of the *y*-axis dendrogram are structurally related and share similar pharmacology (**Figs. 2** and **3**).

## Preclinical and clinical drug-induced effect comparison

To determine whether *y*-axis dendrogram relationships in effect spectrum analysis provide a consistent portrait of the pharmacological similarity of medicines, we investigated how the *y*-axis sorting of medicines is affected by reporting bias, such as the listing of 'drug class'–associated effects on drug labels. Building on relationships between molecular structure, pharmacology and functional response of medicines established in previous studies[15,16], we investigated whether *y*-axis classifications obtained with binary effect descriptors (**Figs. 1**–**3**) mirror classifications obtained with preclinical data. This investigation required complete sets of preclinical and clinical data, thereby limiting the scope of this experiment to 872 medicines (**Supplementary Table 2** online). In addition, to reduce the effect of information density in binary effect descriptor classification, we used effect spectra containing 240 COSTART fields (**Supplementary Table 3** online, data set VI), because these constructs provided the best consensus between classifications produced by effect spectra sets

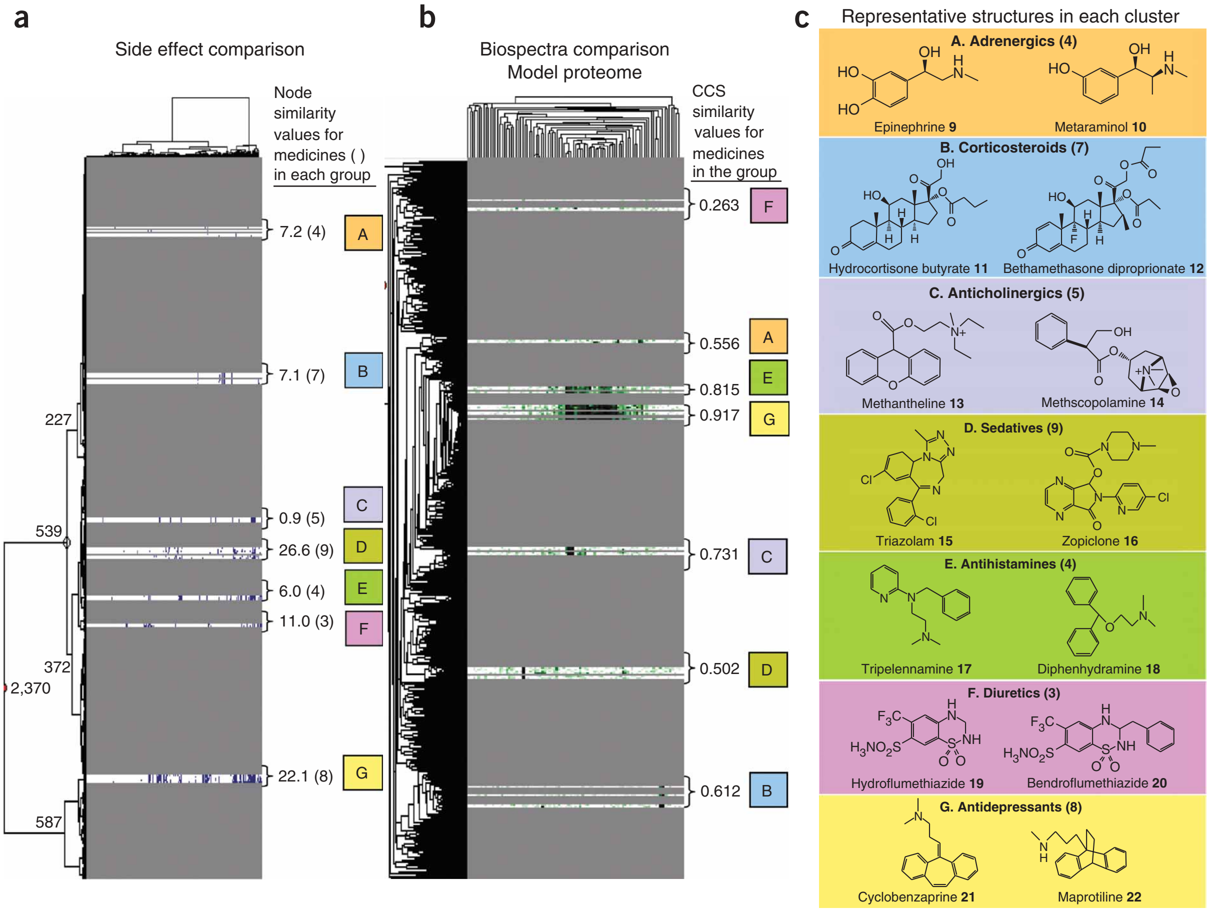

**Figure 2** Comparison of preclinical and clinical drug-induced effect similarity of 40 medicines. (**a**) Classification of 872 medicines' binary side effect spectra (240 side effects) using Ward's method[22] as hierarchical clustering method. Highlighted are seven clusters (A–G) arranging the effect spectrum similarity of 40 medicines in groups. Node similarity values measuring side-effect profile similarity between medicines in clusters A–G are indicated at right, and numbers in parentheses denote number of compounds in each effect cluster. Low node similarity values indicate high effect spectrum similarity. (**b**) Dendrogram relationship for 872 medicines using preclinical drug-induced effect data (biospectra) produced by the UPGMA clustering method. Confidence in cluster similarity values measuring biospectral similarity between medicines in cluster A–G (indicated at right) reflect proximity of medicines in the *y*-axis dendrogram. Compounds with the highest confidence in cluster similarity score (CCS) have the highest biospectral similarity[15]. (**c**) Illustration of two representative structures from each of clusters A–G.

I–VI. Thus, we performed two independent classifications: one assessing biospectral similarity using the unweighted pair group method with arithmetic mean (UPGMA) for clustering preclinical data (biospectra), and the other assessing effect spectrum similarity using Ward's method for the clustering of binary clinical effect spectra. Hierarchical clustering of these independent drug-induced effect descriptor sets is described in Methods.

Seven classes of medicines, denoted A–G (**Fig. 2a,b**), illustrate the substance of these classification results. A more detailed comparative analysis has been conducted for the group of sedative medicines in cluster D (**Fig. 2a,b**). For the purpose of clarity, **Figure 2a,b** omits data on 832 compounds and shows classification information on only forty medicines representing seven pharmacological classes: (i) adrenergics (group A; four medicines), including epinephrine (**9**) and metaraminol (**10**); (ii) antiinflammatory corticosteroids (group B; eight medicines), including hydrocortisone butyrate (**11**) and bethamethasone diproprionate (**12**); (iii) anticholinergics (group C; five medicines), including methantheline (**13**) and methscopolamine (**14**); (iv) hypnotic anxiolytics (group D; nine medicines), including triazolam (**15**) and zopiclone (**16**); (v) antihistamines (group E; four medicines), including tripelennamine (**17**) and diphenhydramine (**18**); (vi) diuretics (group F; three medicines), including hydroflumethiazide (**19**) and bendroflumethiazide (**20**) and (vii) antidepressants (group G; eight medicines), including cyclobenzaprine (**21**) and maprotiline (**22**). Pharmacological clusters A–G are evenly distributed across the entire *y*-axis side-effect dendrogram (**Fig. 2a**), thereby representing medicines with the least and the most reported side effect information. The effect spectrum similarity range for compounds in clusters A–G is shown at right in **Figure 2a**. For example, the node similarity measuring the effect spectrum similarity between triazolam (**15**) and zopiclone (**16**) in group D has a value of 26.6. We compared node similarity values for medicines in clusters A–G and found that the five medicines in group C (node similarity = 0.9) had the highest effect spectrum similarity of all the medicines in groups A–G. This assessment is based on the comparison of the node similarity values for all five medicines in group C. The node similarity value of these five compounds exceeds the node similarity value determined for methantheline **13** and methscopolamine **14** (node similarity = 0.9), which is the pair of medicines with the lowest effect spectrum similarity in this group. For comparison, the biospectral similarity between medicines in individual clusters A–G (preclinical effect spectra) is shown in **Figure 2b** and is expressed using confidence in cluster similarity values (CCS) produced by the UPGMA algorithm[15]. In biospectra comparison, identical biospectra have a CCS value of 1 (refs. 15,16). For example, the biospectral similarity between medicines appearing in clusters A–G (**Fig. 2b**) ranges from a CCS value of ≥0.263, for the biospectra of hydroflumethiazide (**19**) and bendroflumethiazide (**20**) (group F), to a CCS value of ≥0.917 for the biospectra of cyclobenzaprine (**21**) and maprotiline (**22**) in group G. We compared CCS values for medicines in clusters A–G (**Fig. 2b**) and found that medicines in group G had the highest biospectral similarity. Again, the neighborhood of medicines in cluster G is defined by comparing the CCS values of medicines in cluster G with the CCS value separating the pair of medicines with the lowest biospectra score (in group G, this CCS value is 0.917). We compared node similarity values between drug-induced effect spectra in individual clusters A–G (**Fig. 2a**) and found that the clinical effect profile similarity between compounds in clusters A–G (**Fig. 2a**) was mirrored in the preclinical drug-induced effect profile similarity of medicines in clusters A–G (**Fig. 2b**)**.** This observation indicates that compounds residing in each of the individual clusters A–G not only have the greatest clinical effect spectrum similarity (**Fig. 2a**) but also have the greatest biospectral similarity (**Fig. 2b**).

## Identification of diagnostic side effect patterns

Of the seven medicine classes (**Fig. 2**), the hypnotic-anxiolytic medicines shown at the center of the *y*-axis dendrogram in group D (**Fig. 2a**) showed the lowest effect spectrum similarity among medicines in clusters A–G (node similarity value for D was 26.6). From the biospectral similarity of compounds in cluster D (**Fig. 2b**), we identified a corresponding CCS value of ≥0.502. Previous observations using biospectral analysis indicate that biospectra comparison yielding similarity values below a threshold (CCS ≥ 0.8) may not provide reliable structure-function information[15,16]. Thus, a strong association between structure and effect should not necessarily be expected for compounds in cluster D. Accordingly, the dendrogram section (**Fig. 2b**) with CCS ≥0.502 contains ten additional medicines omitted in the illustration (**Fig. 2**) because the effect spectra associated with these ten additional medicines fell outside the node similarity criterion (node similarity = 26.6) used to illustrate the neighborhood relationship of medicines in effect spectra dendrograms (**Fig. 2a**) and biospectra dendrograms (**Fig. 2b**). To examine the relevance of the relationship between biospectral similarity (CCS values) and effect spectrum similarity (node similarity values), we investigated the entire cohort of medicines in the biospectral similarity range CCS ≥ 0.502, containing the medicines shown in group D (**Fig. 2b**). This particular dendrogram segment is shown in **Figure 3a**.

All 19 medicines with CCS ≥ 0.502 (**Fig. 3a**) have sedative-hypnotic and anxiolytic pharmacology, and all have benzodiazepine-like structures (**Supplementary Fig. 3** online), with the exception of the sedative-hypnotic zopiclone (**16**). Consistent with previous observations[15,16], medicines with the highest biospectral similarity (medicines **28** and **29**, **30** and **31**, and **33** and **34**) have the highest structure similarity (**Supplementary Fig. 3**). The classification produced by the clustering of 872 drug-induced effect spectra containing these 19 medicines is shown in **Figure 3b**,**c**. The *y*-axis placement of these 19 medicines (**Fig. 3b**,**c**) provides information on the clinical effect spectrum similarity of these medicines and the corresponding *x*-axis dendrogram (**Fig. 3b**,**c**) lists the effect of these 19 drugs on organ systems. Cursory examination of the *y*-axis dendrogram (**Fig. 3b** and **Supplementary Fig. 3**) indicates that data density and side effect similarity drive effect spectra classifications for these medicines. The effect of data density is particularly evident in the *y*-axis (**Fig. 3b**), dispersing these 19 medicines into subgroups with low (L), medium (M) and high (H) data density. The side-effect similarity between these 19 medicines is easily recognized if one inspects the corresponding *x*-axis dendrogram regions, denoted as $C_1$–$C_4$ (**Fig. 3b**) and $C_1$–$C_3$ (**Fig. 3c**). Thus, diazepam (**27**), midazolam (**32**) and flumazenil (**37**), residing in *y*-axis dendrogram section H (**Fig. 3b**), have an average of 16 symptoms reported in *x*-axis dendrogram sections $C_1$, $C_2$ and $C_3$ (**Fig. 3c**). In contrast, clordiazepoxide (**24**), oxazepam (**26**), lormetazepam (**36**) and lorazepam (**28**), residing in the *y*-axis dendrogram section L (**Fig. 3b**), have on average only five symptoms reported in the *x*-axis dendrogram sections $C_1$, $C_2$ and $C_3$ (**Fig. 3c**). Conversely, zopiclone (**16**) and 11 of its neighbors (**15**, **23**, **25**, **29**, **30**, **31**, **33**, **34**, **35**, **38** and **39**) residing in *y*-axis dendrogram section M (**Fig. 3b**) have on average ten symptoms reported in *x*-axis dendrogram sections $C_1$, $C_2$ and $C_3$**. Figure 3c** indicates that irrespective of *y*-axis dendrogram placement in group L, M and H, each of these 19 medicines exhibits one or more of the symptoms identified in *x*-axis clusters $C_1$, $C_2$ and $C_3$ (**Fig. 3b**,**c**).

A detailed representation of these symptom associations (enlarged *x*-axis dendrogram section $C_1$–$C_3$) is shown (**Fig. 3c**). Accordingly, all 19 drug labels list somnolence as a side effect, 12 list somnolence in combination with nervousness, and 13 drug labels list agitation and confusion in combination with somnolence. These symptom associations create a diagnostic pattern. Presence of somnolence in each of these characteristic symptom patterns indicates an association between side-effect pattern similarity and primary pharmacology[24,25]. Notably, this diagnostic pattern is recognizable even in cluster L (**Fig. 3b**), which contains medicines with only five symptoms in dendrogram sections $C_1$, $C_2$ and $C_3$**.** This observation indicates that these 19 medicines, identified by biospectral similarity CCS ≥ 0.502, not only have the same primary pharmacology but also produce very similar effects on organ system networks, as indicated by the shared characteristic side-effect pattern. Notably, the presence of drug target information is not necessary for assessing pharmacologically relevant neighborhood behavior. For example, deletion of the GABA A benzodiazepine receptor from the biospectra of 872 medicines and repeating UPGMA clustering with truncated biospectra does not affect the close biospectra association between medicines **27**, **28**, **29**, **30**, **31** and **32**, although the CCS value determining biospectral similarity for these compounds decreases from a CCS value of >0.682 (value for the biospectral similarity between these compounds (**Fig. 3a** and **Supplementary Fig. 3**)) to a value of CCS > 0.531 for the truncated form. All six of these medicines have somnolence and agitation as common side effects, and five of them list nervousness, GI disturbance and confusion as additional symptoms. Again, identification of biospectral similarity aids in the identification of diagnostic symptom patterns and illustrates the close relationship between primary pharmacology and side-effect similarity of these six medicines. Although this experiment points out that low confidence in cluster similarity values diminishes the reach of biospectral analysis in neighborhood assessment, it also shows that the identification of diagnostic side effect patterns shared by compounds in specific biospectra clusters greatly aids assessment of meaningful neighborhood behavior by providing independent information on neighborhood properties.

## Assessing the relevance of biospectra classifications

The biospectra deletion experiment described above indicates that the identification of diagnostic clinical effect patterns coinciding with the determination of biospectral similarity provides a mechanism for calibrating CCS values and the relevance of biospectra classifications.

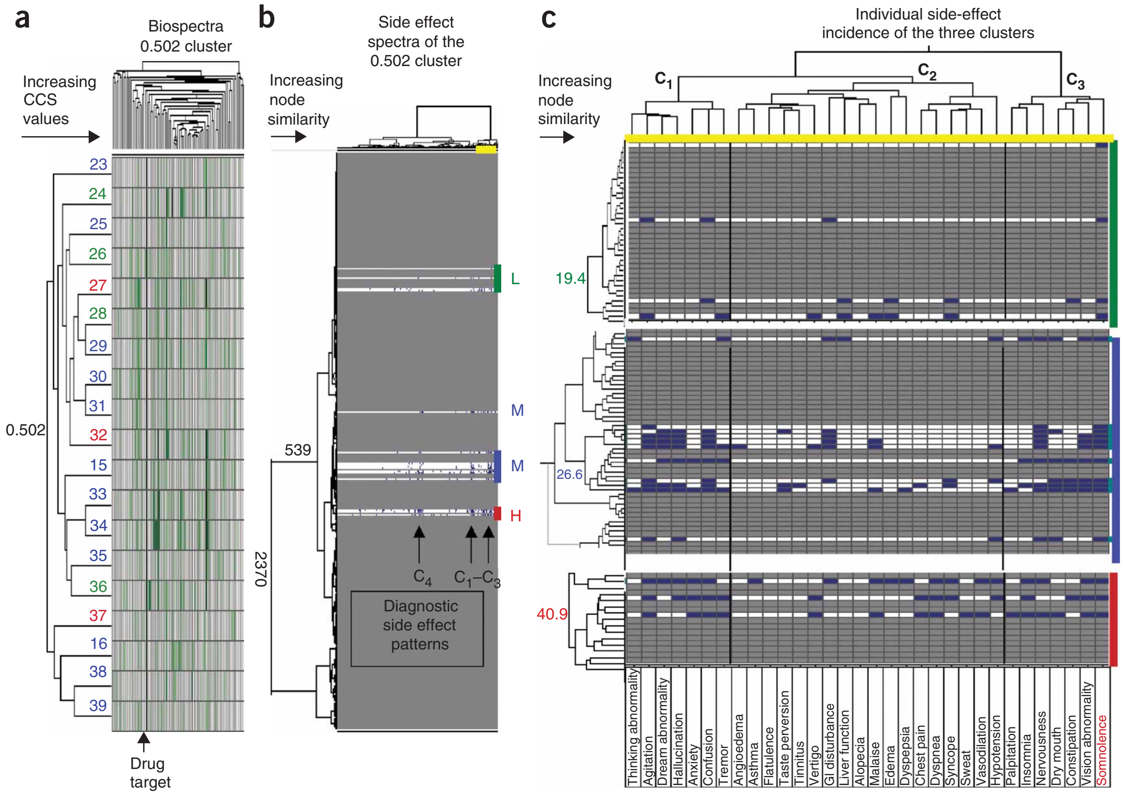

**Figure 3** Comparison of biospectra and effect spectra of 19 sedative-hypnotic medicines. (**a**) Hierarchical clustering using the biospectra of 872 medicines. Shown is a cluster containing 19 medicines with CCS > 0.502. Medicines are numbered and color-coded. The four medicines numbered in green have had the fewest side effects reported, medicines numbered in red have had the most side effects reported, and those in blue have had an average number of side effects reported on drug labels. The structures of these medicines are shown in **Supplementary Figure 3**. (**b**) Effect spectrum classification of 872 medicines using binary side effect spectra (240 side effects) and Ward's method[22]. Effect spectrum similarity values (node similarity values) obtained are indicated in the *y*-axis dendrogram at left. For clarity, only the dendrogram sections containing the 19 side effect spectra of interest are shown. Three clusters containing medicines with high, low and medium numbers of side effects reported are indicated by L, M and H. Medicines in section M are shown in blue, those in dendrogram section L are shown in green and those in section H are shown in red. (**c**) A subset of the characteristic side-effect pattern shown in **b** containing 32 side effects (somnolence is attributed to the primary pharmacology of these medicines and is highlighted in red).

This neighborhood property determination is an entirely empirical and unbiased operation. The identification of characteristic drug-induced symptom patterns provides a medical diagnosis for medicines residing within a certain biospectral similarity range (neighborhood) and identifies the pharmacological relevance of biospectra classifications. This relationship, quantified in biospectrum and effect spectrum analysis, allows the translation of drug-induced effects on protein networks into drug-induced effects on organ systems. This capability has application in the drug design process. For example, if one were interested in targeting the GABA A benzodiazepine receptor to discover sedative-hypnotic medicines with reduced side-effect liabilities, modern drug design would go about this task by identifying chemical structures that have high affinity for the GABA A receptor[25]. Currently, there are no guiding principles that subsequently would help drug discovery scientists differentiate between drug design choices on the basis of clinical effect predictions. Structural diversity proponents, for example, might predict that zopiclone-like compounds would differ from those of benzodiazepines (**Supplementary Fig. 3**) and hence favor selection of a zopiclone-like structure on the grounds of structural dissimilarity with benzodiazepines. Biospectrum analysis, on the other hand, projects that zopiclone (**16**) would produce clinical effects similar to those of benzodiazepines (those in the 0.502 biospectra cluster). The retrospective analyses (**Figs. 2** and **3**), affirmed by recent clinical experience, suggest that zopiclone (**16**) indeed shares pharmacology and side-effect similarity with benzodiazepines[25,26].

## Comparing neighborhood behavior over a pharmacology range

To investigate whether unbiased neighborhood assessments could be extended over the entire pharmacological range, we identified the distance matrices defining the similarity between each combination of biospectra pairs represented in the 872-compound database. We then repeated the same process using the clinical effect spectra to determine effect spectrum similarity matrices (distance matrices) between the same medicines. These calculations using R statistical software and Manhattan (or City Block) distance are described in Methods[27,28]. We then randomly sampled compounds in the 872-compound dataset and identified each sample's nearest neighbor using (i) the biospectra distance matrix and (ii) the effect spectra distance matrix. The result of this sampling experiment is shown in **Figure 4a**.

Accordingly, the *y*-axis (**Fig. 4a**) demonstrates the effect spectrum similarity (distance) between any pair of randomly sampled medicines, and the *x*-axis mirrors the biospectral similarity (distance) for the same pair of medicines. Increasing *y*-axis values (**Fig. 4a**) therefore translate into a decrease in effect spectrum similarity between two medicines. Likewise, increasing *x*-axis values (**Fig. 4a**) translate into a decrease in biospectral similarity between two medicines. The

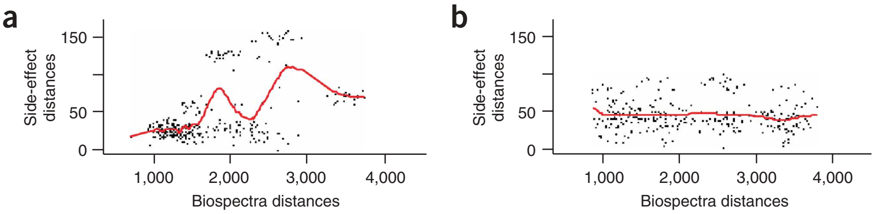

**Figure 4** Association between biospectral and side effect similarity for 25 medicines: an independent assessment of similarity distance relationships between side-effect and biospectral profile similarity for compounds in the 872-medicine database. (**a**) Using dataset VI (872 compounds/240 binary effect descriptors), the distance matrix determining the effect spectrum similarity between each medicine in the dataset was produced using R statistical software and Manhattan distance as similarity measure[27,28]. The distance matrix determining the biospectral similarity between each medicine in the entire dataset was created in the same manner. Twenty-five medicines were randomly selected, and the similarity (distance) between their respective side effect profiles was plotted against the similarity (distance) between the same medicine's biospectra. The linear interpolation (red line) was then plotted in order to show a correlation between the drug biospectra and side effects ($R = 0.79$). (**b**) In order to assess the possibility that the correlation in **a** could be due to random chance, a duplicate experiment was performed using the same approach. However, in this case, both similarity distances were randomly selected through independent (uncoupled) sampling. These results ($R = 0.2$) show that the correlation observed in **a** is not due to random chance.

correlation between these two independent similarity assessments (**Fig. 4a**) using a trend line (in red) indicates that preclinical effect spectrum similarity is mirrored in clinical effect spectrum similarity. Inspecting the trend line indicates existence of neighborhood behavior wherein medicines within a certain biospectral similarity distance range also reside in neighborhoods with similar effect spectra distance ranges. Accordingly, compound pairs that have the highest biospectral similarity (highest confidence in cluster similarity values) also have the greatest side effect spectrum similarity.

In order to assess if the correlation (**Fig. 4a**) was due to random chance, we conducted a control experiment (**Fig. 4b**) in which distances between medicines were randomly sampled in the effect spectrum similarity matrix and plotted on the *y*-axis (**Fig. 4b**). Next, a second round of random sampling was used to independently identify a different pair of compounds in the biospectral similarity matrix, thereby neglecting the coupling of information associated with neighborhood behavior. The distance between this second pair of compounds was again plotted on the *x*-axis (**Fig. 4b**). The corresponding trend line (**Fig. 4b**) shows that, in this case, we did not observe any correlation between distance parameters, indicating that the correlation between neighborhood behaviors (**Fig. 4a**) is not the product of random chance. We repeated the random sampling 100 times, and the outcome of both of these experiments was not affected (**Supplementary Fig. 4** online). These results indicate that that the correlation between neighborhood behaviors extends throughout all pharmacological classes represented in our 872-medicine database.

## DISCUSSION

Classification of drug-induced effect patterns can demonstrate meaningful relationships between preclinical and clinical drug-induced effect information on medicines. This is shown using independent preclinical and clinical data, different classification methods (UPGMA, Ward) and different similarity measures in drug-induced effect spectra classifications (cosine correlation, half-square euclidean distance, Manhattan distance). The information obtained in effect spectrum comparisons is consistent with primary pharmacology, structural similarity and preclinical effect observations of medicines. Binary representations of side effects, collected at early stages of clinical trials, circumvents problems associated with reporting of side effects, and classification of these effect spectra yields information on the drug-induced effect signatures (pharmacology) of medicines. Effects of classification error on analysis outcome (bias in effect spectrum analysis) can be investigated easily using bioassay deletion experiments, which assist in identifying meaningful relationships between preclinical and clinical drug-induced effect patterns. Quantification of this relationship, in turn, allows forecasting of clinical effects of medicines[29]. Refinements in analysis tools and increasing the quality of clinical (*in vivo*) and preclinical (*in vitro*) effect databases will undoubtedly improve the ability of this methodology to capture and organize the information necessary to translate drug-induced effects on proteins into medically useful effects on organisms.

Lack of tools to translate medicinally useful interactions between protein and organ systems networks into new drug designs limits the efficiency of drug discovery. Identification of medicinally useful molecular structures has hitherto been difficult, as molecular properties that ensure optimal drug target interactions are often not those that provide therapeutically useful effects. Here we report that the alignment of preclinical and clinical effect spectra provides a mechanism for linking side effects, molecular structure and primary pharmacology of medicines. By determining structure-effect relationships between interacting networks, this approach facilitates the translation of complex biological response information into drug designs. Although current data quality provides only rough estimates for the relationship between structure designs (encoded in biospectra) and broad clinical effects on medicines (encoded in effect spectra), the forecasting capability of this method is limited in principle only by size, completeness and availability of clinical and preclinical effect data. Despite obvious data quality concerns, binary side effect spectra are useful for estimating broad clinical effect profiles of medicines. This should provide incentives for clinical investigators to expedite the creation of public databases containing not only clinical, preclinical and drug safety information on marketed medicines, but also Phase I and Phase II data of failed experimental drugs. In addition, standardizing reporting requirements of clinical and preclinical effect data and creation of more comprehensive databases should improve the appreciation of the unique relationship between molecular structure and *in vivo* response of medicines. Quantifying these relationships will ultimately lead to a better understanding of the information exchange between protein and organ system networks and will assist in the generation of new medicines that affect organ system functions in a therapeutically useful way.

## METHODS

**Biological activity spectra.** A portion of the CEREP BioPrint database[19] was used for our investigation. A total of 1,045 medicines and 92 ligand-binding assays were used to construct the dataset of biospectra containing complete percentage inhibition values at a ligand concentration of 10 μM. Primary screening at 10 μM was carried out in duplicate. Additional screening was carried out at a ligand concentration of 10 μM if results varied by more than 20%. The 92 assays were selected to represent a cross-section of the druggable proteome[15,16]. The assays and compounds used for this investigation are shown in **Supplementary Tables 1** and **2**. Spotfire Decision Site 7.2 was used for hierarchical clustering. Hierarchical clustering of biological spectra was performed using the UPGMA algorithm and cosinus correlation as similarity measurement. The UPGMA method fragments data contained in the database into a hierarchical structure by representing the similarity between different database fragments in the form of a tree representation. Dendrograms are obtained by iteratively splitting the database into smaller subsets by placing each biospectrum initially into a unique cluster and then computing for each pair of clusters some value of dissimilarity or distance using cosinus correlation as a similarity measure. In every step, clusters with the minimum distance in the current clustering are merged until all biospectra are contained in one cluster. This process creates neighborhood groups of biospectra, in which biospectra with the highest similarity are most closely associated, based on local distance information. For example, the distance (dissimilarity) between three clusters $a1$, $a2$ and $a3$, which each contain $n1$, $n2$ and $n3$ number of records, upon association of clusters $a2$ and $a3$ into a new cluster termed $a4$ is calculated as $\mathrm{sim}(a1, a4) = (a \times \mathrm{sim}(a1, a2)) + (b \times \mathrm{sim}(a1, a3))$, where 'sim' is the similarity between two indexed clusters, $a = n2/(n2 + n3)$, and $b = n3/(n2 + n3)$. The cosine correlation ranges from +1 to –1, where +1 is the highest correlation. Complete opposite profiles have a correlation of –1. The dendrogram similarity between biospectra derived through UPGMA clustering is measured by using confidence in cluster similarity values, where 1 is the highest and 0 is the lowest confidence value. Biospectra classifications (**Figs. 2** and **3**) were produced using the biospectra of 872 medicines in UPGMA clustering. Biospectra deletion experiments were carried out by eliminating percentage inhibition values for the GABA A benzodiazepine receptor from the 92-bioassay array suite and repeating hierarchical clustering with truncated biospectra (91 bioassays) as described above.

**Effect spectra.** Clinical response data, which was derived from clinical response data reported from commercial drug labels listed in the CEREP BioPrint database[19], was used for our investigation. In total, 1,045 medicines and ~800 COSTART[20] codes were used for constructing six datasets. Side-effect data was converted into binary effect descriptor sets with a value of 1 assigned to populated COSTART fields. Those include reported (with or without frequency and/or severity), expected and post-marketing reports. A value of 0 was assigned if a particular effect was not documented. To investigate effects of information density and preclassification bias on classification, six separate datasets (I–VI) were generated by changing descriptor length and data density. Eliminating the most- and least-populated COSTART fields from set I, which had 591 field descriptors, generated descriptor sets II and III (**Supplementary Table 3**). Thus, binary drug effect descriptor set II was generated using 583 COSTART fields, eliminating the eight most common side effects from descriptor set I (nausea, emesis, headache, dizziness, asthenia, rash, diarrhea and abdominal pain, which appear at the far right of the $x$-axis dendrogram (**Fig. 1**) and represent ~4,000 data points). Binary drug effect descriptor set III (240 fields) was generated by eliminating, from the 583 fields of set II, the most sparsely populated COSTART fields (containing $<0.01\%$ of the data; far left on the $x$-axis dendrogram of **Fig. 1**). Elimination of compounds from the data set of 1,045 fields provided an additional means for assessing the effect of clinical data density on cluster analysis. Hence, by removing 173 compounds from the 1,045-field datasets (I, II and III), datasets IV, V and VI, containing 872 compounds (**Supplementary Table 3**), were generated. These 173 compounds were eliminated because they did not have complete preclinical datasets (biospectra), which were required for the analysis shown in **Figure 2a**. Hierarchical clustering using Ward's method and average ordering function provided classifications for each of these binary side effect descriptor sets I–VI (ref. 22). Again, Ward's hierarchical clustering method splits the database into a hierarchical structure in which the similarity between different database fragments is expressed in form of a tree representation. For each pair of clusters in this dendrogram, a value of dissimilarity or distance is computed using half-square euclidean distance as similarity measure. For example, the distance (dissimilarity) between three clusters $a1$, $a2$, $a3$, which each contain $n1$, $n2$ and $n3$ number of records upon association of clusters $a2$ and $a3$ into a new cluster termed $a4$, is calculated as $\mathrm{sim}(a1, a4) = (a \times \mathrm{sim}(a1, a2)) + (b \times \mathrm{sim}(a1, a3)) - (c \times \mathrm{sim}(a2, a3)$, where 'sim' is the similarity between two indexed clusters, $a = (n1 + n2)/(n1 + n2 + n3)$, $b = (n1 + n3)/(n1 + n2 + n3)$, and $c = n1/(n1 + n2 + n3)$. The half-square euclidean distance is always $\geq 0$. The measurement would be 0 for identical profiles and high for profiles that show little similarity.

**Statistics.** All hierarchical clustering of biospectra and adverse event data were conducted using Spotfire analysis tools. The clustering algorithms for each dataset were UPGMA (using cosine distance as similarity measurement) and Ward's method, respectively. All calculations and graphs (**Fig. 4**) were made using R statistical software. Distance matrices (Manhattan distance calculation[27,28]) were calculated for the selected medicine's biospectral profile and were plotted against the distance matrices of the same medicine's side-effect profile.

**Bioinformatics.** Protein domain analysis in the graphical abstract and in **Supplementary Figure 2** was based on the network topology of the mammalian proteome. Biological functions were assigned to each gene network using findings extracted from the scientific literature and were imported into the Ingenuity Pathway Analysis software (Ingenuity Systems). The biological functions assigned to each network are ranked according to the significance of that biological function to the network. The networks are shown graphically as nodes (genes/gene products) and edges (the biological relationships between the nodes). Human, mouse and rat orthologs of a gene, although stored as separate objects in the knowledge base, are represented as a single node in the network.

*Note: Supplementary information is available on the Nature Chemical Biology website.*

ACKNOWLEDGMENTS

The authors would like to thank D.M. Potter for useful discussions.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturechemicalbiology/

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Sole, R.V., Pastor-Satorras, R., Smith, E. & Kepler, T.B. A model of large-scale proteome evolution. Preprint at http://xxx.lanl.gov/cond-mat/0207311 (2002).
2. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
3. Petricoin, E.F., Zoon, K.C., Kohn, E.C., Barrett, J.C. & Liotta, L.A. Clinical proteomics:translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* **1**, 683–695 (2002).
4. Dimpfel, W. Preclinical data base of pharmaco-specific EEG fingerprints (Tele-Stereo-EEG). *Eur. J. Med. Res.* **8**, 199–207 (2003).
5. Zajchowski, D.A. *et al.* Identification of selective estrogen receptor modulators by their gene expression fingerprints. *J. Biol. Chem.* **275**, 15885–15894 (2000).
6. Shi, L.M. *et al.* Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. *Mol. Pharmacol.* **53**, 241–251 (1998).
7. Pellegrini, M. Defining interacting partners for drug discovery. *Expert Opin. Ther. Targets* **7**, 287–297 (2003).
8. Cunningham, M.L., Bogdanffy, M.S., Zacharewski, T.R. & Hines, R.N. Workshop overview: Use of genomic data in risk assessment. *Toxicol. Sci.* **73**, 209–215 (2003).
9. Padrini, R. *et al.* Pharmacogenetics. *N. Engl. J. Med.* **348**, 2041–2043 (2003).
10. Mathew, R.J., Weinman, M.L., Thapar, R., Reck, J.J. & Claghorn, J.L. Somatic symptoms in depression and antidepressants. *J. Clin. Psychiatry* **44**, 10–12 (1983).
11. Hamilton, L.W. & Timmons, C.R. Sex differences in response to taste and postingestive consequences of sugar solutions. *Physiol. Behav.* **17**, 221–225 (1976).
12. Antkowiak, B. How do general anaesthetics work? *Naturwissenschaften* **88**, 201–213 (2001).
13. Steiner, S. & Anderson, N.L. Expression profiling in toxicology-potentials and limitations. *Toxicol. Lett.* **112–113**, 467–471 (2000).
14. Blower, P.E. *et al.* Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J.* **2**, 259–271 (2002).
15. Fliri, A.F., Loging, W.T., Thadeio, P.F. & Volkmann, R.A. Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. USA* **102**, 261–266 (2005).
16. Fliri, A.F., Loging, W.T., Thadeio, P.F. & Volkmann, R.A. Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* **48**, 6918–6925 (2005).

17. Hamadeh, H.K., *et al.* Prediction of compound signature using high density gene expression profiling. *Toxicol. Sci.* **67**, 232–240 (2002).
18. Park, D. *et al.* Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map). *Bioinformatics* **21**, 3234–3240 (2005).
19. Krejsa, C.M. *et al.* Predicting ADME properties and side effects: The BioPrint approach. *Curr. Opin. Drug Discov. Devel.* **6**, 470–480 (2003).
20. Food and Drug Administration. *COSTART: Coding Symbols for Thesaurus of Adverse Reaction Terms* (3rd edn.) (Food and Drug Administration, Center for Drugs and Biologics, Division of Drug and Biological Products Experience, Rockville, Maryland, 1989).
21. Gao, H., Williams, C., Labute, P. & Bajorath, J. Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **39**, 164–168 (1999).
22. Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
23. German, E.J., Wood, D. & Hurst, M.A. Ocular effects of antimuscarinic compounds: Is clinical effect determined by binding affinity for muscarinic receptors or melanin pigment? *J. Ocul. Pharmacol. Ther.* **15**, 257–269 (1999).
24. Itil, T.M. The discovery of antidepressant drugs by computer-analyzed human cerebral bio-electrical potentials (CEEG). *Prog. Neurobiol.* **20**, 185–249 (1983).
25. Sanger, D.J. The pharmacology and mechanisms of action of new generation, non-benzodiazepine hypnotic agents. *CNS Drugs* **18** (Suppl.) 9–16 (2004).
26. Hindmarch, I. Myths, medicine and the media. *Hum Psychopharmacol. Clin. Exp.* **14**, 223–224 (1999).
27. Becker, R.A., Chambers, J.M. & Wilks, A.R. *The New S Language: a Programming Environment for Data Analysis and Graphics* (Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California, 1988.
28. Mardia, K.V., Kent, J.T. & Bibby, J.M. *Multivariate analysis*. Academic Press, (1979).
29. Labute, P., Nilar, S. & Williams, C. A probabilistic approach to high throughput drug discovery. *Comb. Chem. High Throughput Screen.* **5**, 135–145 (2002).