CHEMMEDCHEM

DOI: 10.1002/cmdc.200700153

Analysis of System Structure–Function Relationships

Anton F. Fliri,[a] William T. Loging,[b] and Robert A. Volkmann*[a]

Preclinical pharmacology studies conducted with experimental medicines currently focus on assessments of drug effects attributed to a drug's putative mechanism of action. The high failure rate of medicines in clinical trials, however, underscores that the information gathered from these studies is insufficient for forecasting drug effect profiles actually observed in patients. Improving drug effect predictions and increasing success rates of new medicines in clinical trials are some of the key challenges currently faced by the pharmaceutical industry. Addressing these challenges requires development of new methods for capturing and comparing "system-wide" structure–effect information for medicines at the cellular and organism levels. The current investigation describes a strategy for moving in this direction by using six different descriptor sets for examining the relationship between molecular structure and broad effect information of 1064 medicines at the cellular and the organism level. To compare broad drug effect information between different medicines, information spectra for each of the 1064 medicines were created, and the similarity between information spectra was determined through hierarchical clustering. The structure–effect relationships ascertained through these comparisons indicate that information spectra similarity obtained through preclinical ligand binding experiments using a model proteome provide useful estimates for the broad drug effect profiles of these 1064 medicines in organisms. This premise is illustrated using the ligand binding profiles of selected medicines in the dataset as biomarkers for forecasting system-wide effect observations of medicines that were not included in the incipient 1064-medicine analysis.

Introduction

Relating the molecular structure of medicines to broad effect observations in organisms remains one of the greatest challenges in the discovery of new medicines.[1] Most contemporary methods for identifying these structure–effect relationships rely on the analysis of drug interactions with specific proteins for predicting drug effects in vivo.[2,3] Accordingly, most preclinical pharmacology studies conducted with experimental medicines focus on effects attributed to a drug's mechanism of action. The high attrition rate of experimental medicines in clinical trials, however, indicates that this approach does not provide an accurate assessment of the therapeutic utilities of medicines.[4] The growing realization that the pleiotropic nature of drug effects in organisms is determined by the protein network connectivity involved in system-wide information processing suggests that improving the success rates of current drug-discovery paradigms requires new methods for "system" structure–function analysis.[5–9]

Herein we describe a strategy for capturing "system-wide" drug effect information of 1064 medicines and the use of this information for system-wide structure–effect analysis by correlating pattern similarities between broad drug-induced effects with the chemical structure design of these medicines (Figure 1). The potential application of this method in drug discovery is illustrated by using the broad structure–effect information of a group of medicines (Supporting Information figure 1 a) for forecasting drug effect patterns of medicines that were not included in the original analysis.

Methods for predicting broad drug effects of medicines require the translation of molecular structure descriptors into descriptors of the drug effect patterns that are observed at the cellular and at the organism level in response to drug treatment. The accuracy of this translation depends on the quality of the structure–response descriptors employed. A particular challenge in this exercise is the gathering and comparison of broad and heterogeneous drug effect information from disparate data sources. Text mining of published scientific literature provides an attractive means to access the worldwide information base and provides a process by which an entire knowledge base is searched for logistic placement of words and phrases. Such approaches have been reviewed extensively.[10]

For identifying broad structure–effect information for 1064 medicines, we sought to identify the number of publications that contained search terms (drugs, proteins, clinical) in any number of combinations. For example, for the number of times a list of A was identified with a list of B, we captured this information as co-occurrence frequency data.

To capture the structural features of these medicines in a comparative way, two standard binary 2D fragment-based de-

[a] *Dr. A. F. Fliri, Dr. R. A. Volkmann*
Neuroscience Medicinal Chemistry, Pfizer Global Research and Development
Eastern Point Road, Groton, CT 06340 (USA)
Fax: (+1) 860-715-7569
E-mail: robert.a.volkmann@pfizer.com

[b] *Dr. W. T. Loging*
Computational Biology, Pfizer Global Research and Development
Eastern Point Road, Groton, CT 06340 (USA)

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

WILEY InterScience DISCOVER SOMETHING GREAT

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

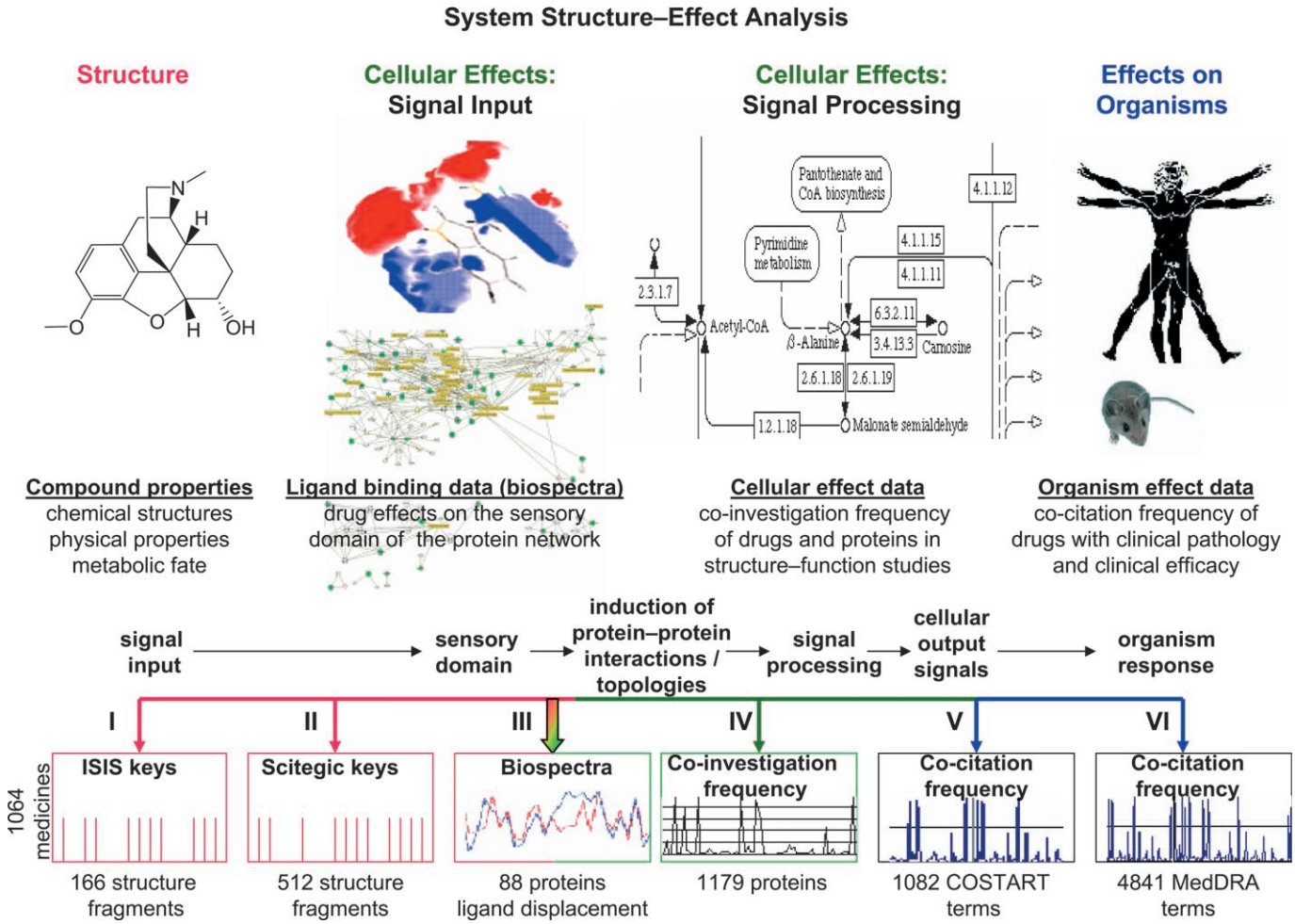

Figure 1. Comparing system-wide structure–function information using: 2D binary ISIS structure keys (I), 2D binary Scitegic structure keys (II), interaction profiles of drugs with a section of the sensory domain of the protein network (III), assessment of drug effects on protein functions using co-investigation frequencies of drugs with 1179 proteins as measurements (IV), assessment of in vivo drug effects of medicines using co-citation frequencies of medicines with 1082 COSTART terms as measurements (V), and assessment of in vivo drug effects using co-citation frequencies of drugs with 4841 MedDRA terms as measurements (VI).

scriptors were selected. For comparing effects of the same medicines at the cellular level, two protein-centered descriptors were examined. The first of these descriptor sets provides measurements of the interaction capacity of these medicines with the sensory domain of a model proteome by determining the ability of each of these medicines to displace ligands from a bioassay array consisting of 88 structurally diverse proteins. The second protein-centered descriptor set, which is derived from text mining of the Medline database, provides measurements of how often each of these medicines has been co-investigated (co-occurrence frequency) with a much larger set of 1179 proteins. This second descriptor set is anticipated to provide broad drug effect information for each of the 1064 medicines at the cellular level by assessing their interaction potential with 1179 protein network components. Lastly, for capturing drug effect information at the organism level, text mining of the Medline database was used for determining the co-occurrence frequency of the 1064 medicines with search terms associated with adverse effects (COSTART and MedDRA) vocabularies (Figure 1).

The identification of system-wide structure–effect relationships for these 1064 medicines requires a mechanism for quantifying similarity relationships between each of these descriptor sets. This comparison is enabled by translating the structure and the effect information for each of the 1064 medicines into spectra form and by determining information spectra similarity between these 1064 medicines using hierarchical clustering algorithms. This sorting strategy identifies groups of medicines with similar structure and effect information. Because hierarchical clustering identifies spectra similarity measures for all medicines in each descriptor set, the spectra similarity measures derived from each of the structure and response descriptors can be compared. In doing so, this methodology captures the entire knowledge spectrum of medicines and determines the alignment between preclinical and clinical structure–effect information based on information spectra similarity. For example, using the structure–effect descriptors shown in Figure 1 for 1064 medicines, over 0.5 million pairwise comparisons in each of the six datasets are obtained (information spectra; Figure 1). For determining correlations between structure and effect in-

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

formation spectra similarity, the spectra similarity measures for each of the medicines using all six descriptor sets were compared. Because all of these pairwise comparisons are treated as equal, examination of spectra similarity correlations between these datasets can be used for determining which of the preclinical structure and effect descriptor sets provides the best alignment with information spectra describing the broad system-wide effects at the cellular and the organism level. Once established, the utility of that particular preclinical descriptor set can be evaluated for predicting the broad and heterogeneous structure–effect observations in organisms by using medicines with similar structure–effect spectra as comparative standard. This methodology does not aim at predicting the probability that two medicines have the same individual effects but rather the likelihood that two medicines will exhibit similar system-wide effect observations.

Results and Discussion

ISIS (I) and Scitegic keys (II) were selected as the two binary 2D structure fragment-based descriptors. These along with a proteome-centered descriptor set III (biospectra) were selected as preclinical descriptors to compare the structure designs between medicines.[7–9] Because biospectra (III) reflect the capacity of drugs to displace ligands from an in vitro panel of 88 cell surface and cytosolic proteins (Supporting Information figure 1 b), this descriptor set[11] was also used to compare drug-induced perturbations in the sensory domain of the cellular protein network. For capturing effect information on cellular systems and organisms associated with these medicines, three text-mining-derived descriptor sets IV–VI were constructed. The cellular effect descriptor set IV was derived through text mining of the Medline database and captures how often each of the 1064 medicines have been used in structure–function studies with 1179 structurally and functionally diverse proteins (Supporting Information figure 1 c).[12] This co-investigation frequency information was translated into information spectra and normalized by assigning a value of "100" to co-citation frequencies exceeding 100 literature citations based on the anticipation that co-investigation frequency measurements provide meaningful structure–effect information before this citation frequency is exceeded.[13] For ascertaining and comparing broad drug effects associated with these medicines at the organism level, text mining of the Medline database was again employed, and co-citation frequencies between each of the 1064 medicines and 1) 1082 COSTART (V) terms (Supporting Information figure 1 d) and 2) 4841 MedDRA (VI) terms (Supporting Information figure 1 e) were obtained. These two adverse effect descriptor sets capture broad clinical effects associated with these medicines. Again, both of these co-citation frequency datasets were translated into spectra form and "normalized" as described above for IV.[12]

System-wide structure–effect relationships between these 1064 medicines were examined by determining similarities between "structure spectra" [ISIS keys (I), Scitegic keys (II), biospectra (III)] and "effect spectra" [protein co-investigation (IV), COSTART co-citation (V), and MedDRA co-citation (VI) spectra] (Figure 1 and Figure 2). To ensure that structure–effect spectra similarity relationships were not unique to a particular similarity measure, pairwise distances between medicine spectra were determined using three spectra similarity clustering protocols: 1) Wards method algorithm[14] using half Euclidean distance, 2) UPGMA algorithm using Euclidean distance, and 3) UPGMA using cosine correlation as a similarity measure. In total, 18 da-

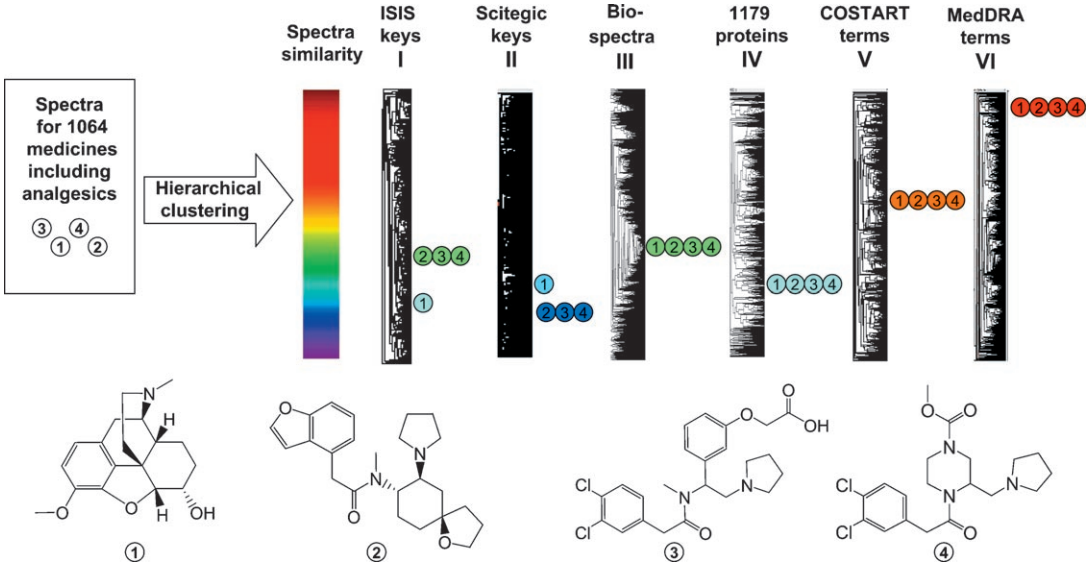

Figure 2. Comparison of information spectra I–VI though hierarchical clustering methods. Information spectra similarity is measured using dendrogram distance relationships (565 516) within each descriptor set. Depicted are spectra similarity relationships derived from UPGMA clustering using cosine correlation as similarity measure for 1064 medicines. The sorting of drugs using this metric is illustrated in the comparison of four analgesics medicines ①–④ shown. Structure–effect similarities between these medicines are ascertained by comparing structure spectra similarity (I–III) with effect spectra similarities (IV–VI). Determining these spectra similarity correlations enables structure–effect predictions by establishing definable distances between medicines using spectra similarity (I–VI) between medicines as starting point.

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

tasets were generated, each containing 565 516 pairwise distance measures derived by clustering each spectra dataset I–VI using the three clustering protocols. For investigating correlations between different spectra similarity measures, linear regression analysis was used for determining the statistical significance between distance measures provided by the three methods.[15] For clarity purposes, as this examination is based on the use of different clustering algorithms and similarity measures, references made herein to confidence in cluster similarity (CCS) values pertain to dendrogram relationships derived from UPGMA clustering with cosine correlations as similarity measure.

Structure similarity assessments

Inspecting the results provided by the linear regression analysis of 565 516 pairwise distances measures derived from the clustering of descriptor sets I–III using the three methods indicates that the capacity of structure descriptor spectra I–III to identify concordant structure similarity relationships varies and depends on the clustering method (entries 1–3 of Table 1) and the confidence in spectra similarity value threshold (entries 4–7 of Table 1). For example, the structure similarity measures produced by the clustering of ISIS and Scitegic 2D binary structure keys (descriptor sets I and II) using different clustering methods and similarity measures provides concordant structure similarity rankings, with I, II, and III only in the high structure–spectra similarity ranges (entries 7, Table 1 a and 1 b). Consequently the uncertainty associated with the majority of structure similarity assessments derived from structure descriptor spectra I and II restricts system-wide structure–effect investigations to medicines that have the most similar structure design in the 1064-medicine database. In contrast, the clustering of the proteome-centered descriptor set (III) produces concordant spectra similarity assessments between the two clustering methods and similarity measures at much lower spectra similarity values, and enables structure–effect investigations even for structurally diverse datasets (entries 6 and 7, Table 1 c). For example, medicines known to exhibit bioequivalent properties such as compounds ①–④ (Figure 2) and **1**–**10** (Figure 3) cluster in proximate biospectra dendrogram positions, and hence are perceived by the model proteome as similar structure designs.

Table 1. Statistical significance[a] obtained (×) for pairwise distance comparisons of medicines using three structure descriptors and three system-wide drug effect descriptors.[b]

a) ISIS Key Comparisons

Entry	Structure			Effect		
	ISIS keys I[c]	Scitegic II	Biospectra III	1179 IV	COSTART V	MedDRA VI
1	ISIS Euclid. (565 516)				×	×
2	ISIS Ward (565 516)		×	×[d]	×	×
3	ISIS cosine (565 516)			×		×
4	CCS > 0.5 (214 537)				×	
5	CCS > 0.65 (41 792)		×[d]		×	×
6	CCS > 0.70 (18 471)		×		×	×
7	CCS > 0.85 (1512)	×	×	×	×	×

b) Scitegic Key Comparisons

Entry	Structure			Effect		
	Scitegic Keys II[c]	ISIS I	Biospectra III	1179 IV	COSTART V	MedDRA VI
1	Scitegic Euclid. (565 516)					
2	Scitegic Ward (565 516)		×			
3	Scitegic cosine (565 516)					
4	CCS > 0.2 (375 629)					
5	CCS > 0.3 (27 631)		×		×	
6	CCS > 0.35 (8187)		×	×	×	×
7	CCS > 0.40 (3834)	×	×	×	×	×

c) Biospectra Comparisons

Entry	Structure			Effect		
	Biospectra III[c]	ISIS I	Scitegic II	1179 IV	COSTART V	MedDRA VI
1	Biospectra Euclid. (565 516)			×		×
2	Biospectra Ward (565 516)	×	×	×	×	×
3	Biospectra cosine (565 516)			×	×	×
4	CCS > 0.1 (452 388)			×[d]	×	×
5	CCS > 0.3 (60 098)			×	×	×
6	CCS > 0.5 (25 862)	×	×	×	×	×
7	CCS > 0.7 (5076)	×	×	×	×	×

[a] $p < 0.05$. [b] Comparisons shown using each of the structure descriptor datasets (entries 1–3) were made by employing the same clustering algorithm and similarity measures for 1064 medicines for each of the six datasets. Confidence in cluster similarity ranges [CCS = 0 (least similar) to 1 (most similar)] were established (entries 4–7) using UPGMA/cosine correlation. [c] Number of spectra comparisons shown in parenthesis. [d] $p > 0.05$; p-values for this dataset are shown in Supporting Information figure 1.

Evaluating relationships between chemical structure descriptors I–III and effect descriptors IV–VI

The second step in this analysis probed if the pairwise distance measures produced by the clustering of structure descriptor sets (I–III) correlate with the pairwise distance measures produced by the clustering of the effect descriptor sets IV–VI. For comparing the six information spectra I–VI under the same conditions, spectra similarity measures in these databases were sampled by varying the confidence in spectra similarity thresholds produced by UMPGA clustering in combination with cosine correlation as similarity measure. Linear regression analysis at different sampling intervals was performed for determining correlations between the six different spectra similarity measurements I–VI associated with these intervals (entries 4–7, Table 1).

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

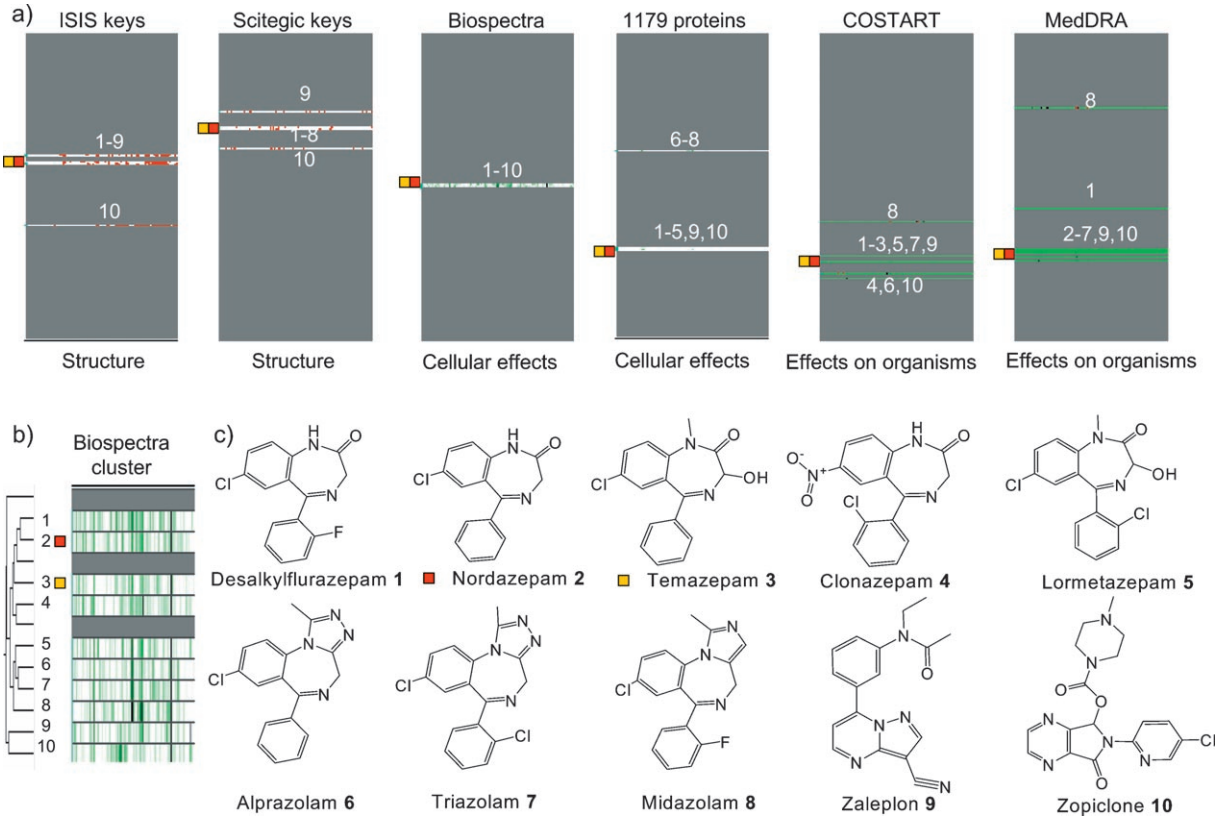

Figure 3. Dendrogram location of 10 medicines in datasets I–VI derived from UPGMA clustering using cosine correlation as similarity measure for 1064 medicines. a) Identification of two medicines residing in concordant "discrete event clusters" in each of the six information spectra (illustrated is the use of the average information for Nordazepam/Temazepam as comparative standard). Using the average information for both medicines ("standard profile") in each descriptor set for profile searching identifies the information similarity shared by medicines in each of the six discrete event clusters. Effect predictions are based on spectra similarity predictions for medicines sharing 60% protein co-investigation frequency (1179), MedDRA and COSTART co-citation frequency profile similarity with the standard profiles. b) The biospectra-derived discrete event cluster whose compounds share 60% effect profile similarity with the "standard profiles" in the IV, V, and VI effect spectra comparisons. c) Structures of the 10 medicines.

Inspection of Table 1 indicates that while the correlations between all spectra similarity measurements reach statistical significance at the highest confidence in spectra similarity values (entries 7, Tables 1a and 1b), the database sizes associated with these correlations varies. For example, the correlation between effect spectra similarity (IV–VI) and structure similarity provided by clustering each of the 2D chemical structure keys (I–II) reaches statistical significance for only a small percentage of medicines: ISIS (0.26%) and Scitegic keys (0.67% of the pairwise distances). In contrast, a much larger portion (4.5%) of the spectra similarity measurements provided by the proteome-centered structure descriptor set (III) correlates with effect spectra (IV–VI) similarity measurements (Table 1c). Notably, the spectra similarity measures provided by descriptor set III produce statistically significant structure and effect spectra similarity correlations irrespective of the clustering method or similarity measures used for these spectra comparisons (Supporting Information table 2). Moreover, these structure–effect spectra similarity correlations are observed over a broad range of spectra similarity measures.

Data density issues

Whereas cursory inspection of structure–function matrices I–VI indicates that cluster memberships of medicines established by descriptor set III (biospectra) parallels those produced by the clustering of co-investigation spectra (cellular effect descriptor set IV) and the clustering of co-citation frequency spectra (organism effect descriptor sets V and VI), medicines in single biospectra clusters (III) can be disbursed over several different "text-mining-derived" effect clusters (IV–VI) due to variations in reporting frequencies of structure–function or structure–effect information for older or more recently approved drugs (Figure 3). Therefore, for aligning "text-mining-derived" effect clusters with structure-descriptor-derived clusters, variations in information density in the text mining information must be considered. Fortunately, the clustering of ligand displacement data or structure spectra (descriptor sets I–III) is not affected by information density variations and, as a result, neighborhood boundaries for medicines produced by the clustering of structure spectra (I–III) can be used for identifying effect characteristics (IV–VI) shared by medicines residing within "discrete event" cluster boundaries.

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Discrete event clusters

To identify information characteristics shared by medicines that belong to a particular structure or effect spectra classification, a biomarker-based concept was used. Accordingly, for identifying medicines with aligned preclinical and system-wide effect information, medicines residing in proximate dendrogram positions in information spectra sets III–VI were used to construct average information spectra profiles, which serve as starting points for conducting profile searches in datasets III–VI. For identifying medicines with similar information characteristics in descriptor set III–VI, this average information spectrum (biomarker profile) was then used for locating other medicines with similar information characteristics in each of the datasets. While the identified medicines often reside in the same cluster as the medicines that were used to create the biomarker standard, they can also reside in different clusters because of data density variations which can obscure information spectra similarity ascertained through cluster analysis.

For example, two benzodiazepines, Nordazepam (**2**) and Temazepam (**3**), reside in proximity in the clustering-derived dendrograms of all six descriptor sets (Figure 3). Four Nordazepam/Temazepam biomarker profiles (a biospectra III, a 1179 protein co-occurrence IV, a COSTART V, and a MedDRA co-citation profile) were created, representing the average information of both medicines in spectra sets III–VI. These biomarker profiles were then used in profile searches in each of the appropriate databases III–VI for identifying medicines sharing $>50\%$ information characteristics with this "biomarker" profile. For example, by using the biomarker profile representing the average biospectra III information for Nordazepam/Temazepam, 17 medicines sharing 55% biospectra profile similarity with this biomarker profile were identified from the entire 1064-medicine database. Of these 17 medicines, 13 reside with Nordazepam (**2**) and Temazepam (**3**) in a single biospectra (III) "discrete event" cluster (CCS >0.53). To determine if these medicines share similar effect characteristics, the biomarker profiles of **2**/**3** in descriptor set IV–VI were used to identify medicines that share $>60\%$ effect spectra characteristics. This profile comparison indicates that 10 of the 13 medicines residing in discrete event cluster III share $>60\%$ cellular effect spectra characteristics in spectra set IV. Likewise, the average information profile of **2**/**3** in descriptor sets V and VI (biomarker profiles V and VI) were generated and used to identify medicines that share $>60\%$ effect characteristics with these "biomarkers" in databases V and VI. Again, the result of these profile comparisons indicates that 10 of the 13 medicines in discrete event biospectra cluster III also share $>60\%$ effect characteristics in the COSTART and MedDRA spectra sets V and VI (Figure 3).

Forecasting of structure–effect relationships

The spectra similarity thresholds identifying statistically significant correlations between structure and effect spectra similarity (Table 1) enables the forecasting of effect characteristics shared by medicines that reside within structure-spectra derived "discrete event" cluster boundaries. For example, the range-finding experiments described above suggest that medicines sharing $>55\%$ biospectra biomarker profile III characteristics with Nordazepam and Temazepam are expected to share at least 60% of the effect spectra characteristics with the biomarker profiles derived from Nordazepam and Temazepam in datasets IV–VI. For evaluating the accuracy of this forecast, 141 medicines with complete sets of information spectra I–V, but which were omitted from the 1064-compound dataset because they lacked complete sets of MedDRA spectra (VI) data, were added to the 1064-compound database. The information spectra (I–V) for these 1205 compounds were then clustered by using the same similarity measures and clustering methods as before. Structure–effect spectra similarity relationships were then examined by using the biomarker profiles III–V of Nordazepam and Temazepam defined by the 1064-compound dataset as comparative standard. Accordingly, using biomarker profile III (biospectra) of Nordazepam/Temazepam for profile searching in the new 1205-spectra database, 21 medicines were identified which share $>55\%$ profile similarity with the biomarker III standard (Figure 4). All of these medicines reside within a single biospectra cluster with CCS >0.462, and 17 of these medicines have a biospectra similarity of CCS >0.5, which exceeds the threshold (Table 1) required for structure–effect forecasting. To determine if these medicines have indeed $>60\%$ effect similarity as predicted by the structure–effect spectra similarity correlations described above, the protein co-investigation (IV) and COSTART co-citation (V) biomarker profiles of Nordazepam/Temazepam established with the 1064-medicine dataset were used to identify compounds in the 1205-medicine database sets that share profile similarity with these biomarker standards. This effect profile comparison indicates that all of the 17 medicines that have a biospectra similarity of CCS >0.506 also have $>60\%$ profile similarity in effect spectra sets IV and V (Figure 4). This observation confirms the structure–effect predictions based on spectra similarity correlations (Table 1).

Moreover, this forecasting method is not only useful for relating structure to effects but also effects to structure, and depends on whether structure or effect biomarker profiles are used. For example, if one uses the effect biomarker profiles (IV and V) of Nordazepam and Temazepam, as defined by the 1064-medicine databases, to identify medicines that share $>60\%$ effect profile similarity in the 1205-compound database, 23 medicines are identified. Twenty of these 23 medicines share $>50\%$ similarity with biomarker profile III and are located in the CCS >0.464 biospectra cluster described above (Figure 5).

Conclusions

In summary, spectra representation of structure and effect information of medicines enables the quantitative comparison of vast amounts of data (for 1064 medicines, biospectra III contains ~90 000 data points, protein co-investigation frequency spectra IV captures data from ~11 million abstracts, and the COSTART (V) and MedDRA (VI) co-citation frequency spectra

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

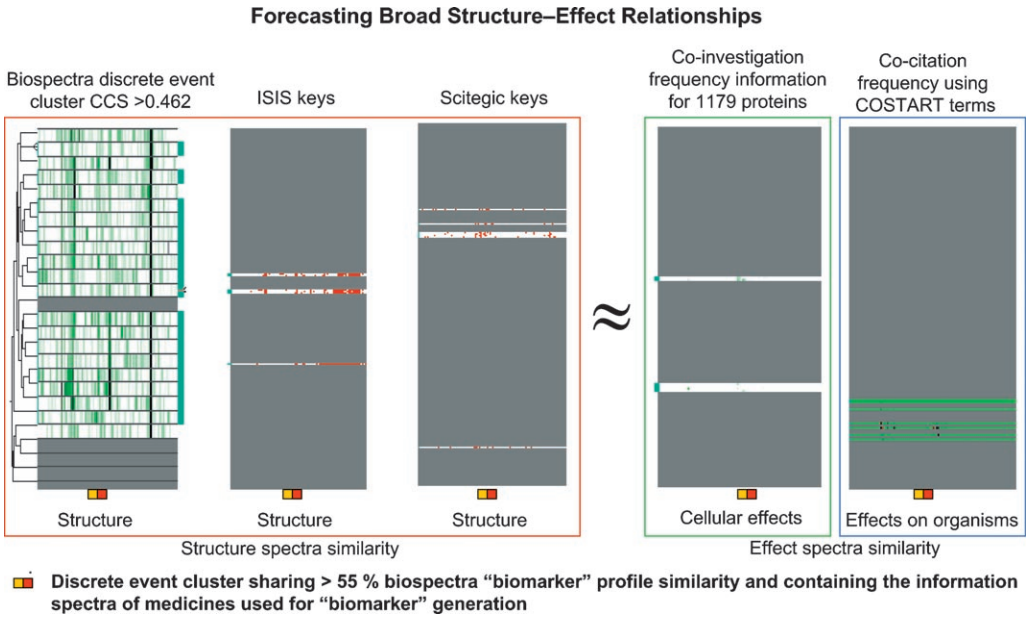

Figure 4. Twenty-one medicines (CCS > 0.462 cluster) identified from a 1205-medicine database meet the biomarker threshold (0.555) of the Nordazepam/Temazepam biospectra standard and reside in one biospectra neighborhood. Seventeen in CCS > 0.506 cluster (blue line) have > 60% effect profile (1179/COSTART) similarity to the standard.

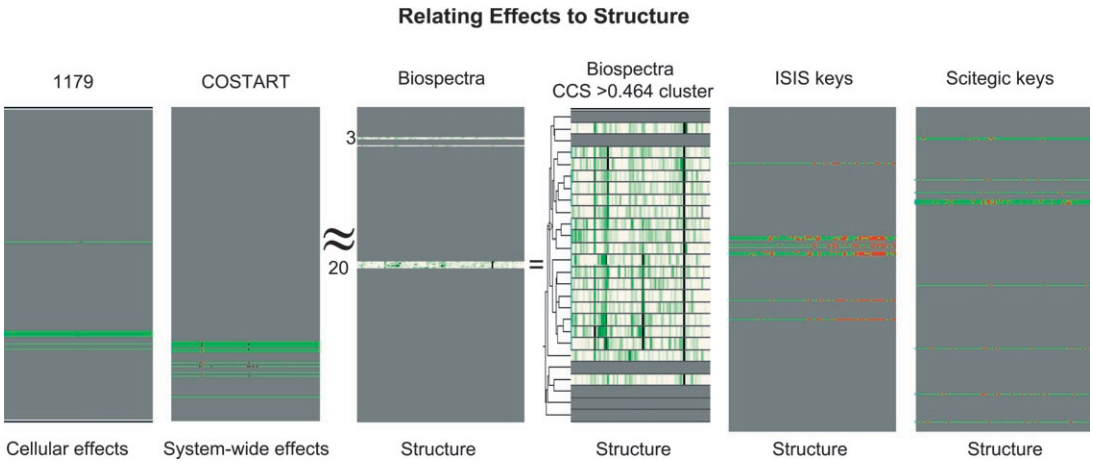

Figure 5. Twenty-three medicines identified from a 1205-medicine database that have > 60% effect (COSTART and 1179) profile similarity to the Nordazepam/Temazepam standard. Twenty are located in the CCS > 0.464 biospectra cluster.

capture over 100 million data points). Information spectra comparisons identify system-wide structure–effect relationships and enable structure–effect prediction by capturing information characteristics shared by medicines in discrete event clusters. The correlation between structure and effect spectra similarities shown in Table 1 indicates that the proteome-centered structure descriptor set III outperforms 2D ISIS and Scitegic binary chemical structure descriptors (I and II) in systems-wide structure–effect relationship analysis.

These observations also indicate that the properties of medicines will likely be most similar if they belong to groups that share a high degree of spectra similarity. This observation is consistent with general structure–activity relationships concepts indicating that structure–effect correlations become quantifiable only within certain structural boundaries, meaning that subsets of data fall into groups with similar structure–function relationships when the structure similarity associations strengthen. In the case of text-mining-based approaches, ascertaining these relationships depends on the information density, that is, the range of publications and the structural diversity of the chemicals and proteins investigated. However, unlike traditional SAR methods, spectra similarity-based SAR methods enable the comparison of data from a broad range of sources. As additional publications and compounds/proteins are incorporated (whether they be positive or negative research reports) and more medicines are examined, this approach will likely yield a better resolution of the details embedded in individual structure–function relationships. Notwith-

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

standing these current limitations, the ability to compare broad drug effects on cellular function with broad drug effects on organisms is a key to the design of medicines.

The observations described above indicate that the determination of preclinical biomarker profile similarity (biospectra similarity) enables the forecasting of broad effect characteristics of medicines by establishing correlations between structure and effect spectra similarities for medicines that share discrete event cluster boundaries. Anticipating that the expression of effect patterns at the organism level is, at the cellular level, regulated by protein network circuits, the observed similarity between broad effect patterns induced by pharmacologically related medicines indicates that the information provided by this systems approach is useful for investigating the structure of large-scale protein networks involved in the translation of drug-induced cellular signals into organism responses. The capability to align chemical structure and pharmacology similarity using vast and heterogeneous information sources is a key component in deciphering this translation process. In this regard, interaction networks, which are much larger than the network perimeter required for regulating discrete biochemical functions associated with the activities of single proteins, are likely responsible for observed broad cause-and-effect relationships. This topic is currently of great interest to those involved in systems biology and systems medicine research.

Experimental Section

Biological activity spectra. A portion of the BioPrint database[11] of Cerep (Rueil-Malamaison, France) was used for our investigation. A total of 1064 compounds in the first set and 1205 compounds in the second set (Supporting Information figure 1a) and 88 ligand binding assays (Supporting Information figure 1b) were used for constructing the biological spectra datasets containing complete percent inhibition values at 10 μM ligand concentration. Primary screening at 10 μM was carried out in duplicate (additional screening occurs if results vary by $>20\%$). The 88 assays were selected to represent a cross-section of the drugable proteome. Hierarchical clustering (HAC) of biological spectra was performed with the 1) UPGMA algorithm and cosine correlation as similarity measurement,[7–9] 2) UPGMA algorithm using Euclidean distance, 3) Wards method algorithm[14] with half Euclidean distance using Spotfire Decision Site 8.1. (Spotfire, Somerville, MA, USA).

Co-investigation frequency data. Co-investigation matrix creation was based on previously published text-mining work,[12] whereby co-investigation was defined as the occurrence of both a compound term and a protein within the same Medline abstract. More than 5000 biomedical journals containing over 15 million citations from Medline 2006 were scanned for co-occurrence of query compounds and proteins, resulting in 10.9 million compound–protein associations across more than one million abstracts. A full matrix of 1064 compounds (also 1205 compounds) with co-investigation counts against 1179 proteins (Supporting Information figure 1c) was created using the Python coding language (http://www.python.org). This dataset was then normalized, that is, all co-investigation counts >100 were set to 100. Bootstrapping experiments on the co-investigation matrix were conducted as previously described.[16].

COSTART co-citation frequency data. The COSTART co-citation matrix creation was based on previously published text-mining work,[12] whereby co-citation was defined as the occurrence of both a compound term and COSTART medical terminology (*Coding Symbols for a Thesaurus of Adverse Reaction Terms*, developed by the United States Food and Drug Administration) within the same Medline abstract. More than 5000 biomedical journals containing over 15 million citations from Medline 2006 were scanned for co-occurrence of query compounds and COSTART terms. A full matrix of 1064 compounds (also 1205 compounds) with co-citation counts against 1082 COSTART terms (Supporting Information figure 1d) was created using the Python coding language (http://www.python.org). This dataset was then normalized, that is, all co-citation counts >100 were set to 100.

MedDRA co-citation frequency data. The MedDRA co-citation matrix creation was based on previously published text-mining work,[12] whereby co-citation was defined as the occurrence of both a compound term and MedDRA medical terminology (*Medical Dictionary for Regulatory Activities*, managed by the Maintenance and Support Services Organization) within the same Medline abstract. More than 5000 biomedical journals containing over 15 million citations from Medline 2006 were scanned for co-occurrence of query compounds and MedDRA term. A full matrix of 1064 compounds (also 1205 compounds) with co-citation counts against more than 4500 MedDRA terms (Supporting Information figure 1e) was created using the Python coding language (http://www.python.org). This dataset was then normalized, that is, all co-investigation counts >100 were set to 100.

Chemical structure keys. ISIS (MDL Drug Data Report Version 2001.1, MDL ISIS/HOST software, MDL Information Systems Inc., San Leandro, CA (USA) http://www.mdli.com) and Scitegic Structure Keys (SciTegic Inc., 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123 (USA), CI049905C) for each of the 1205 compounds were generated using Pipeline Pilot 5.1 (http://www.scitegic.com) software. The finalized matrices of biospectra, co-investigation frequency, and structure spectra were joined together into a single data frame and visualized using Spotfire 8.1 Decision Site software.

Statistics. Correlation calculations of ligand displacement spectra and structure keys were conducted for Table 1 by creating pairwise UPMGA cosine correlation, UPGMA Euclidean, and Wards method distance calculations using Spotfire. The CCS cluster distance pair information was placed into linear regression analysis for each dataset and an associated *p* value of correlation was obtained using Spotfire Decision Site 8.1. software.

Acknowledgements

W.T.L. thanks Rob Peitzsch for bootstrap methods as well as Lee Harland and Bryn Williams-Jones for useful discussions on Medline text mining.

Keywords: biospectra · cellular response · drug design · proteins · structure–function relationships

[1] G. Zybarth, N. Kley, *Curr. Drug. Targets* **2006**, *7*, 387 – 395.
[2] A. G. Maldonado, J. P. Doucet, M. Petitjean, B.-T. Fan, *Mol. Diversity* **2006**, *10*, 39 – 79.
[3] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[4] Y. C. Martin, J. L. Kofron, L. M. Traphagen, *J. Med. Chem.* **2002**, *45*, 4350 – 4358.
[5] S. Ekins, *Drug Discovery Today* **2004**, *9*, 276 – 285.
[6] N. Wang, R. K. DeLisle, D. J. Diller, *J. Med. Chem.* **2005**, *48*, 6980 – 6990.
[7] A. F. Fliri, W. T. Loging, P. F. Thadeio, R. A. Volkmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 261 – 266.
[8] A. F. Fliri, W. T. Loging, P. F. Thadeio, R. A. Volkmann, *J. Med. Chem.* **2005**, *48*, 6918 – 6925.
[9] A. F. Fliri, W. T. Loging, P. F. Thadeio, R. A. Volkmann, *Nat. Chem. Biol.* **2005**, *1*, 389 – 397.
[10] W. Loging, L. Harland, B. Williams-Jones, *Nat. Rev. Drug Discovery* **2007**, *6*, 220 – 230.
[11] C. M. Krejsa, D. Horvath, S. L. Rogalski, J. E. Penzotti, B. Mao, F. Barbosa, J. C. Migeon, *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470 – 480.
[12] P. Srinivasan, D. Hristovski, *Medinfo.* **2004**, *11*, 808 – 812.
[13] R. M. Losee, *J. Inf. Sci.* **1989**, *15*, 179 – 189.
[14] D. L. Wild, C. J. Blankley, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155 – 162.
[15] J. A. Rice, *Mathematical Statistics and Data Analysis, 2nd ed.*, Duxbury Press, Belmont, **1995**.
[16] A. C. Davison, D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University, Cambridge, **1997**.

Received: June 25, 2007
Revised: August 29, 2007
Published online on October 19, 2007

 © 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim