Bioorganic & Medicinal Chemistry 20 (2012) 5372–5378

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Cluster analysis of the DrugBank chemical space using molecular quantum numbers

Mahendra Awale, Jean-Louis Reymond *

Department of Chemistry and Biochemistry, Freiestrasse 3, University of Berne, 3012 Berne, Switzerland

ARTICLE INFO

Article history:
Available online 14 March 2012

Keywords:
Chemical space
Clustering
DrugBank
Virtual screening
Cheminformatics
Descriptors
Fingerprints

ABSTRACT

DrugBank (>6000 approved and experimental drugs) was analyzed using molecular quantum numbers (MQNs), which are 42 integer value descriptors of molecular structure counting atoms, bonds, polar groups and topological features. Principal component analysis of MQN-space showed that drugs differ mostly by size (PC1, 67% variance) and structural rigidity and polarity (PC2, 18% variance). Twenty-eight groups of target specific drugs were recovered by proximity sorting in MQN-space as efficiently as by substructure fingerprint (SF) similarity, but in different order allowing for lead-hopping relationships not seen in SF similarity. Clustering by MQN- or SF-similarity produced very different types of clusters. Each of the 28 drug groups spread over different clusters in both MQN- and SF-clustering, and most clusters contained drugs from different target specific groups, showing that structure-based classifications only partially overlap with bioactivity. An MQN-browsable version of DrugBank is available at www.gdb.unibe.ch.

1. Introduction

Organic molecules consist of groups of covalently bonded atoms of carbon, hydrogen, oxygen, nitrogen, sulfur, halogens and a few additional elements. Any molecule is defined by a particular set of atoms (elemental formula), their covalent connectivity pattern (2D-structure), and the relative positions of these atoms in physical space (3D-structure, including stereochemistry and conformations). These structural features define the molecule's identity and determine its properties including physical properties (e.g., solubility), its chemical and biochemical reactivity (e.g., stability and degradation), and its biological activity (e.g., interactions with biomolecules). In the case of drug molecules, one hopes to capture the key structural features determining the biological properties in theoretical models with predictive value. Quantitative structure–activity relationship (QSAR) studies address the detailed rationalization of the activity of focused compound series using statistical methods.[1]

The advent of combinatorial chemistry and high-throughput screening have led to the necessity to consider general models capable of predicting compound properties in the framework of collections of millions of molecules with highly diverse structural types.[2] This has led to the concept of the chemical space as an approach to classify these large datasets based on structural features. 'Chemical space' describes the ensemble of all organic molecules to be considered when searching for new drugs, as well as the property spaces in which these molecules are placed for the sake of predicting and understanding their properties. Property spaces can be constructed from various descriptors of molecular structure,[3] or from binary fingerprints.[4] Once placed in a chemical space, molecules can be compared with one another by measuring distances using similarity measures.[5]

A well conceived chemical space should group compounds with similar physico-chemical and biological properties together, such that similarity measures might be used for ligand-based virtual screening. Following Occam's razor principle, one should seek to find a chemical space with descriptors as simple as possible yet with reasonable predictive value. Recently we reported a chemical space based on 42 integer value descriptors of molecular structure called molecular quantum numbers (MQNs)[6] in analogy to the atomic and principal quantum numbers classifying the elements in the periodic system.[7] These descriptors include classical topological indexes such as atom and ring counts,[8] and a few additional counts such as cyclic and acyclic unsaturations, atoms and bonds in fused rings, and electrostatic charges predicted for neutral pH (Table 1). Since the MQNs only have integer values, MQN-space is composed of 'MQN-bins' to which molecules are assigned if they share the same MQN values. Such MQN-isomers may be compared to isotopes sharing the same atomic and principal quantum number in the periodic system of the elements. The 42 MQNs can be determined from the structural formula by anyone with basic training in organic chemistry. Although the information contained in the MQN is much less detailed than that of binary substructure fingerprints (SF), a commonly used method to describe molecules by the presence or absence of defined substructures,[4] the possibility to determine and understand MQNs without the aid of a computer is an important advantage. Furthermore, MQN-values are extremely rapidly computed and can be used to classify very large databases such as

* Corresponding author. Fax: +41 31 631 80 57.
E-mail address: jean-louis.reymond@ioc.unibe.ch (J.-L. Reymond).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.03.017

Table 1
The 42 molecular quantum numbers (MQNs)

Atom counts (12)		Bond counts (7)	
c	Carbon	asb	Acyclic single bonds
f	Fluorine	adb	Acyclic double bonds
cl	Chlorine	atb	Acyclic triple bonds
br	Bromine	csb	Cyclic single bonds
i	Iodine	cdb	Cyclic double bonds
s	Sulfur	ctb	Cyclic triple bonds
p	Phosphorous	rbc	Rotatable bond count
an	Acyclic nitrogen		
cn	Cyclic nitrogen	*Topology counts*[b] (17)	
ao	Acyclic oxygen	asv	Acyclic monovalent nodes
co	Cyclic oxygen	adv	Acyclic divalent nodes
hac	Heavy atom count	atv	Acyclic trivalent nodes
		aqv	Acyclic tetravalent nodes
Polarity counts[a] (6)		cdv	Cyclic divalent nodes
hbam	H-bond acceptor sites	ctv	Cyclic trivalent nodes
hba	H-bond acceptor atoms	cqv	Cyclic tetravalent nodes
hbdm	H-bond donor sites	r3	Three-membered rings
hbd	H-bond donor atoms	r4	Four-membered rings
neg	Negative charges	r5	Five-membered rings
pos	Positive charges	r6	Six-membered rings
		r7	Seven-membered rings
		r8	Eight-membered rings
		r9	Nine-membered rings
		rg10	$\geqslant$10 membered rings
		afr	Atoms shared by fused rings
		bfr	Bonds shared by fused rings

[a] Polarity counts consider the ionization state predicted for the physiological pH 7.4. hbam counts lone pairs on H-bond acceptor atoms and hbdm counts H-atoms on H-bond donating atoms.
[b] All topology counts refer to the smallest set of smallest rings. afr and bfr count atoms respectively bonds shared by at least two rings.

PubChem,[9] the fragment subset of PubChem,[10] or the entire chemical universe database GDB-13, which lists 977 million molecules up to 13 atoms of C, N, O, S and Cl.[11] The relevance of MQN-space is evidenced by the fact that strong enrichments are observed when using MQN-similarity to search for bioactive analogs of known drugs, a principle which can be applied for prospective drug discovery within the entire GDB-13 database.[12] The MQN-space of PubChem and GDB-13 are freely searchable via the website www.gdb.unibe.ch.[13]

Herein we describe the analysis of DrugBank in MQN-space. DrugBank is an open access database containing data on over six thousand experimental and approved small molecule drugs.[14] By analyzing the positions of 28 groups of 20–100 drugs with documented activity on 28 corresponding targets, we show that biosimilar drugs are relatively well grouped in MQN-space. Thus relatively high AUC (area under the curve) values are obtained in receiver operating characteristic (ROC) curves for virtual screening by MQN-similarity. The enrichments are comparable to those obtained by SF similarity. Interestingly, proximity in MQN-space allows for lead-hopping relationships between drugs that are not visible in the SF classification, in line with previous studies of the PubChem database.[9,10] Clustering of DrugBank by MQN- or SF-similarity shows that drugs form relatively well-defined clusters, however these clusters only partially overlap with groups of biosimilar compounds. Overall the analysis shows that MQN-classification clusters biosimilar drugs as efficiently as SF-classification, with the advantage that MQNs have a more direct and transparent relationship to molecular structure than SF.

2. Results and discussion

2.1. Dataset and fingerprint determination

The DrugBank was downloaded in sdf file format from http://drugbank.ca/downloads, and 6404 drug were successfully read into SMILES for further analysis with an average molecular weight MW = 335 ± 161 Da. The 42 MQN values and a 1024 bit Daylight type binary substructure fingerprint was determined for each compound. Principal component analysis (PCA) was performed with the MQN dataset and color-coded maps of the (PC1,PC2)-plane were produced to provide an overview of the compounds in DrugBank. In this MQN-map, the molecular size (heavy atom count HAC) mostly determined PC1, and the number of H-bond donor atoms (HBA), cycles (rings) and rotatable bonds (RBC) determined PC1 and PC2 in the MQN-map (Fig. 1).

2.2. Virtual screening

DrugBank was analyzed for groups of at least 20 drugs annotated with a common target for their mode of action. By allowing drugs for different isoforms of the same target to be in the same group, we obtained 28 groups of 20–100 drugs each. In addition, three randomly selected groups of 100 drugs each were taken as controls. Virtual screening was performed by calculating ROC curves for recovering each of the 31 groups by similarity to the molecule that was most similar to all other molecules within each group. The ROC curves were computed using the city-block distances CBD_{MQN} or CBD_{SF} and the Tanimoto similarity coefficients T_{MQN} and T_{SF} as similarity measures (Fig. S1). As measured by the area under the curve (AUC), significant enrichments (AUC >80%) were obtained for 16 of the 28 drug groups, while the three control groups gave no enrichment (Fig. 2).[15] The similarity measure CBD_{SF} gave significantly lower and often insignificant AUCs compared to the other three similarity measures CBD_{MQN}, T_{MQN} or T_{SF}. One of the strongest enrichments was obtained for glucocorticoid receptor ligands, which are all derived from the cortisone steroid skeleton. On the other hand acetylcholine esterase (AChE) inhibitors gave no significant enrichment, in agreement with the structurally highly diverse compound types encountered in this family (see Supplementary data smi files).

The occurrence of significant enrichments using both MQN and SF based similarity showed that biosimilar drugs had common features according to both classification principles. However the scores of single drugs obtained by MQN or SF were not correlated. Thus, high scoring compounds in term of MQN-similarity (low CBD_{MQN}, CBD is used preferentially for MQN data) had very variable substructure similarity to the reference compound in terms of SF (variable T_{SF}, Tanimoto similarity is used preferentially for substructure fingerprints), and vice versa (Fig. 3A/B). MQN-similar but SF-dissimilar compounds with the same bioactivity are of particular interest since they feature scaffold-hopping relationships that are typically difficult to identify,[16] as shown for here three examples (Fig. 3C).

The difference between MQN-based similarity and SF-based similarity stems from the very different nature of the molecular features encoded by each fingerprint. Thus, MQNs measure the number of atoms, bonds, polar groups and topological features in a molecule, and are primarily sensitive to molecular size, rigidity and polarity. The highest AUCs were observed using the city-block distance CBD_{MQN} as similarity measure, which therefore appears as the similarity measure of choice for MQN. On the other hand SF report the presence or absence of detailed substructural elements independent of their frequency and are not sensitive to molecular size, but rather to the details of the structural elements present in a molecule. The Tanimoto similarity coefficient T_{SF} appears well suited to quantify similarities between such binary fingerprints.

2.3. Clustering

The DrugBank compounds comprise marketed and investigational drugs as well as food additives and vitamins. Many drug ser-

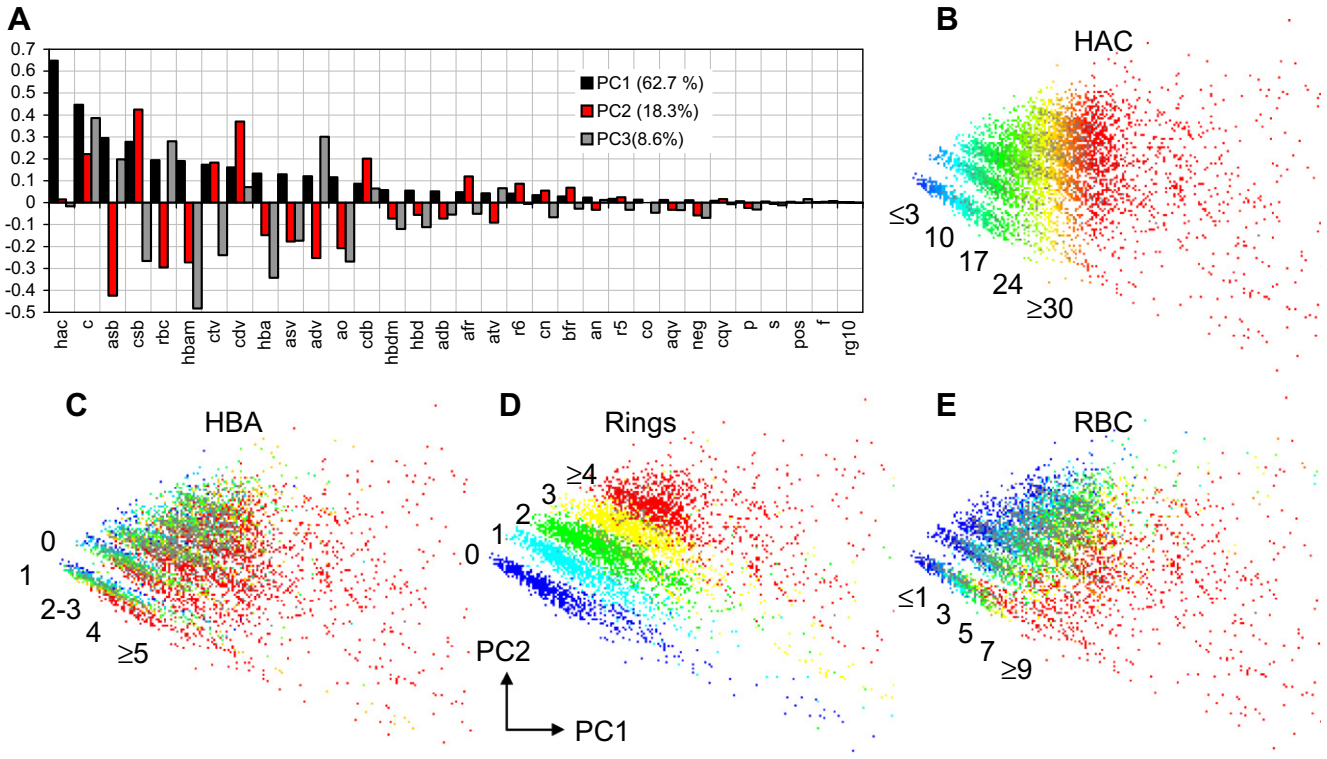

Figure 1. Principal component analysis of the MQN-space of DrugBank (6404 compds). (A) PC loadings. MQNs are ordered by decreasing PC1 loading. The color-coded maps of the (PC1,PC2)-plane for the MQN-space of DrugBank are shown for (B) the heavy atom count, (C) the hydrogen bond acceptor atom count, (D) the number of cycles, and (E) the rotatable bond count. The numbers next to each image indicate the average descriptor values for blue, cyan, green, yellow and red (5 values). A grey shading is added to code for the standard deviation of the values. See Table 1 for full list of MQNs.

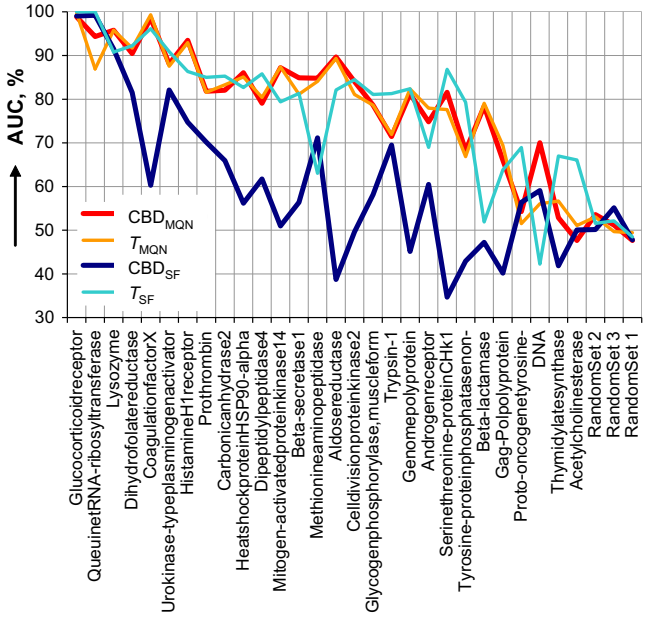

Figure 2. Area under curve values (AUC, in %) for receiver operating characteristic (ROC) curves for recovery of 28 groups of same target drugs and 3 control random sets by sorting DrugBank according to CBD_{MQN} (red thick line, average AUC = 77.3%), T_{MQN} (orange thin line, average AUC = 76.7%), CBD_{SF} (blue thick line, average AUC = 60.3%), T_{SF} (cyan thin line, average AUC = 76.7%), from the most average compound in the set according to the corresponding similarity measure. Groups are sorted by decreasing sum across the four AUCs. An AUC value of 50% corresponds to no enrichment (random recovery), while AUC = 100% corresponds to full prioritization (all actives from a set are the first to be found in the sorted database).

ies consist of structurally related compounds representing variations on a common theme, resulting in part from 'me-too' approaches to drug development, but also from the fact that certain drugs have been developed by optimizing side-activities of drugs through a limited number of modifications.[17] The occurrence of structurally related compound families within DrugBank can be automatically detected by clustering. The DrugBank was clustered on the basis of the similarity measures CBD_{MQN} and T_{SF}, which had produced the best enrichments in the virtual screening examples above. We used the recently reported affinity propagation as clustering algorithm because this method works particularly well in our hand for clustering compound datasets.[18] Clustering using CBD_{MQN} produced 429 clusters and clustering by T_{SF} produced 945 clusters. A second level clustering was then performed by clustering the cluster centers together, which produced 53 2nd level clusters from the CBD_{MQN} dataset and 139 2nd level clusters from the T_{SF} dataset. The clusters were relatively well defined at the 1st level of clustering but were more mixed at the 2nd level of clustering, as shown by the distance histograms for compound pairs within clusters and between clusters (Fig. 4A/B).

The size of clusters obtained by CBD_{MQN} decreased regularly in the cluster size-sorted list, with a relatively high fraction of DrugBank occupying well populated clusters (Fig. 4C/D). Thus, 89% of DrugBank appeared in 1st level MQN-clusters of $\geqslant$10 compounds each, and 90% of DrugBank appeared in 2nd level MQN-clusters of >110 compounds each. In the case of clustering by T_{SF} by contrast, there were generally more clusters containing relatively few compounds, and most of DrugBank compounds occupied relatively small clusters. Thus, only 37% of DrugBank appeared in 1st level SF-clusters of $\geqslant$10 compounds each, and only 65% of DrugBank appeared in 2nd level SF-clusters of >50 compounds each.

The composition of the individual clusters was very different between clustering by CBD_{MQN} and clustering by T_{SF}, as could be

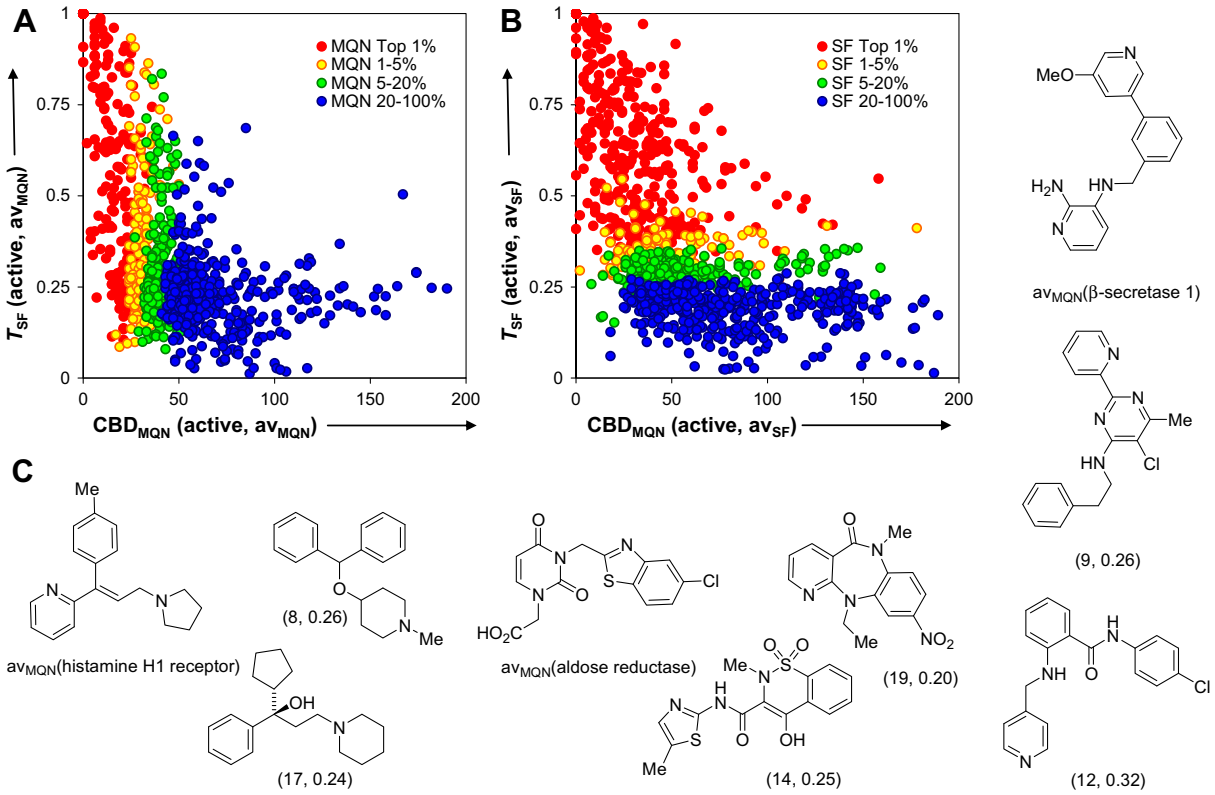

Figure 3. Lead-hopping relationships identified by MQN-similarity search. The scatter plot show the Drugbank compounds listed as active against the 28 targets in Figure 2 with their (CBD_{MQN}, T_{SF}) coordinates relative to the reference average compound in terms of CBD_{MQN} (A) or T_{SF} (B). The points are color-coded according to their position in the ROC curves for sorting DrugBank by CBD_{MQN} (A) or T_{SF} (B) (Fig. S1). (C). For three examples, the query molecule closest to the av_{MQN} is shown with two examples of known actives with relatively low CBD_{MQN} (1st number, high MQN-similarity) and low T_{SF} (2nd number, low substructure similarity).

expected from the very different nature of the similarity measures (Fig. 5A/B). One of the most visible difference concerned the molecular weight (MW) of compounds within clusters. In the MQN-clusters the average MW per cluster was relatively well defined and increased with decreasing cluster size from MW ~ 200 Da in the largest clusters to MW ~ 1100 Da in the clusters occupied by single compounds (Fig. 4E/F). In the SF-clusters by contrast, the MW was not as well defined in each cluster and did not depend on cluster size. The different nature of the clusters was also visible in the distinct distribution of 1st level and 2nd level cluster centers across the (PC1,PC2) MQN-map, in particular at the more even distribution of the 53 2nd level MQN-clusters (Fig. 5C).

2.4. Do clusters correspond to drug targets?

The average similarities of actives found after screening 20% of DrugBank by CBD_{MQN} or T_{SF} relative to the corresponding center of gravity, which is apparent as the separation between green and blue points in Figure 3A/B, corresponds approximately to the CBD_{MQN} and T_{SF} threshold values separating within-cluster from between-cluster compound pairs for 1st level clusters as shown in Figure 4A/B (CBD_{MQN} ~ 35, T_{SF} ~ 0.30). Is it possible that groups of compounds specific for a given target might correspond to a few well-defined clusters in MQN-space or SF-space? Do clusters generally contain only compounds specific of a given drug target?

The possible grouping of drugs specific for a given target in a limited number of clusters was analyzed by counting the number of 1st level and 2nd level clusters occupied by each of the 28 target specific drug groups. Each group of drugs was taken either as a whole, or considering only the drugs found within the first 20% screening of DrugBank in the ROC curves. Ten randomly selected groups of varying size were included as controls. The expected concentration of drugs in fewer clusters was indeed observed in the target-specific groups and not in the controls with both MQN- and SF-clustering, but the effect was relatively modest. Thus, there were in general twofold less 1st level clusters than drugs in the 28 target specific groups, while the controls distributed in as many clusters as drugs. At the 2nd level clustering, there were approximately threefold less 2nd level clusters than drugs in the target specific groups, compared to 1.7-fold less 2nd level clusters than drugs in the controls (Fig. S2 A/B). The effects were comparable when considering only the drugs found within the first 20% screening of DrugBank, showing that the grouping of biosimilar drugs in fewer clusters did not depend on their proximity to the center of gravity of the drug class (Fig. S2 C/D). Note that the centers of the 28 target specific groups were concentrated in a relatively limited are of the (PC1,PC2)-MQN map, which is also the most densely populated (Fig. 5C).

The number of different targets addressed by each cluster was analyzed focusing on the 28 groups of target-specific drugs (Fig. 5D). In the 1st level clustering 97 (30%) of the 320 1st level MQN-clusters and 349 (63%) of the 557 1st level SF-clusters occupied by any of the 28 drug groups contained drugs for only a single group. On average there were 2.5 targets/1st level MQN-cluster and 1.5 target/1st level SF-cluster. The specificity was lower at the level of the 2nd level clusters, with an average of 9.6 targets/2nd level MQN-cluster and 4.8 targets/2nd level SF-cluster. Although the target specificity of the MQN-clusters was a factor of two lower than that of SF-clustering, these values are actually comparable if one considers that MQN-clusters were approximately twice as large as SF-clusters (1st level: 15 drugs/MQN-cluster vs 7 drugs/SF-cluster; 2nd level: 119 drugs/MQN-cluster vs 46

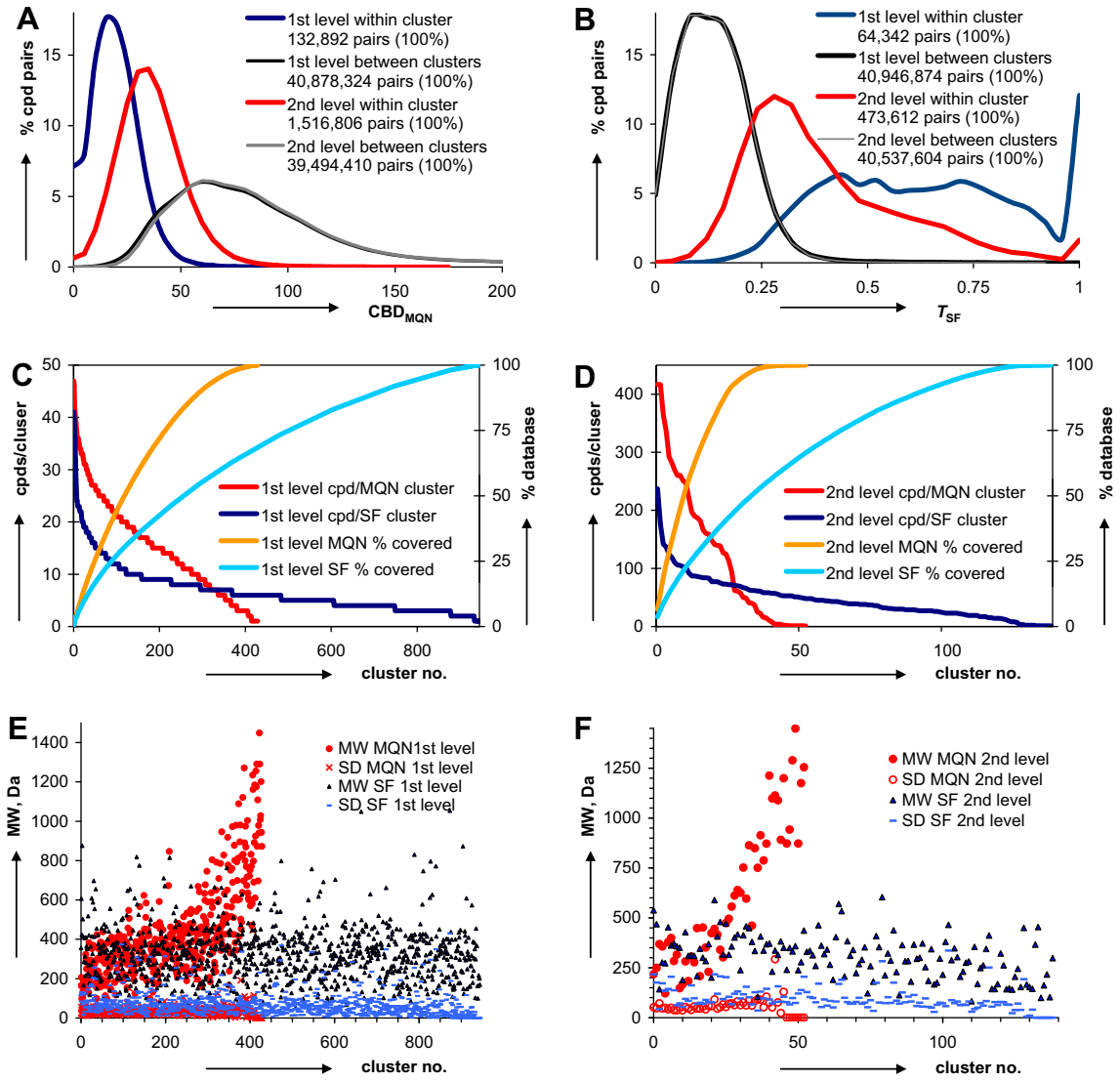

Figure 4. Clustering of DrugBank by affinity propagation. Histogram of (A) CBD_{MQN} distance and (B) T_{SF} similarity of compound pairs within and between clusters at 1st and 2nd level clustering The average similarity values of compound pairs within/between clusters are $CBD_{MQN}^{1stlevel} = 21 \pm 12/92 \pm 52$, $CBD_{MQN}^{2ndlevel} = 38 \pm 15/94 \pm 52$, $T_{SF}^{1stlevel} = 0.66 \pm 0.22/0.16 \pm 0.08$, and $T_{SF}^{2ndlevel} = 0.42 \pm 0.19/0.16 \pm 0.08$. Cluster occupancy and coverage of the DrugBank for clusters sorted by decreasing size for (C) 1st level and (D) 2nd level clustering. Average MW and standard deviation (SD) of compounds within clusters for (E) 1st level and (F) 2nd level clustering (cluster are ordered in decreasing size according to Fig. C and D).

drugs/SF-cluster). Overall the data showed that many clusters did not represent groups of biosimilar compounds but had mixed content.

3. Conclusion

The analysis of DrugBank using MQN descriptors provides a useful overview of its contents. MQN-similarity is comparable to SF-similarity in terms of AUC for recovering groups of biosimilar drugs from DrugBank, but allows for interesting lead-hopping relationships between actives not seen in the SF-similarity. These results confirm earlier virtual screening results with PubChem[9,10] and GDB-13,[11d,12] and fall in line with earlier comments on assigning compounds to common bioactivity classes from similarity between substructure fingerprints versus short fingerprints.[19]

Clustering of DrugBank by MQN- or SF-similarity showed that target-specific drug groups tend to focus on fewer clusters than randomly selected drug groups, and that these clusters may be quite selective of only few drug targets. However the correspondence between MQN- or SF-based clusters and target specific drug groups is far from perfect. This reflects at least in part the very approximative molecular description of MQN and SF, although a more sophisticated approach by shape-based similarity clustering on the DUD dataset gave somewhat similar results.[20] Indeed the structure–activity relationships of drugs is often quite complex, and structural modifications with only minor effects on descriptor sets such as the MQN or SF vector may entirely change the bioactivity profile of a molecule. Because they consist of only simple—and mostly classical—atom counts and topological indices, MQNs have a more direct and transparent relationship to molecules than SF. The MQN system offers a simple method for a first level structure-based classification of drugs, and is publicly available via the MQN-browsable version of DrugBank at www.gdb.unibe.ch.

4. Methods

4.1. Dataset preparation

DrugBank was downloaded as sdf files from http://drugbank.ca/downloads, and 6404 of the 6704 files were successfully converted

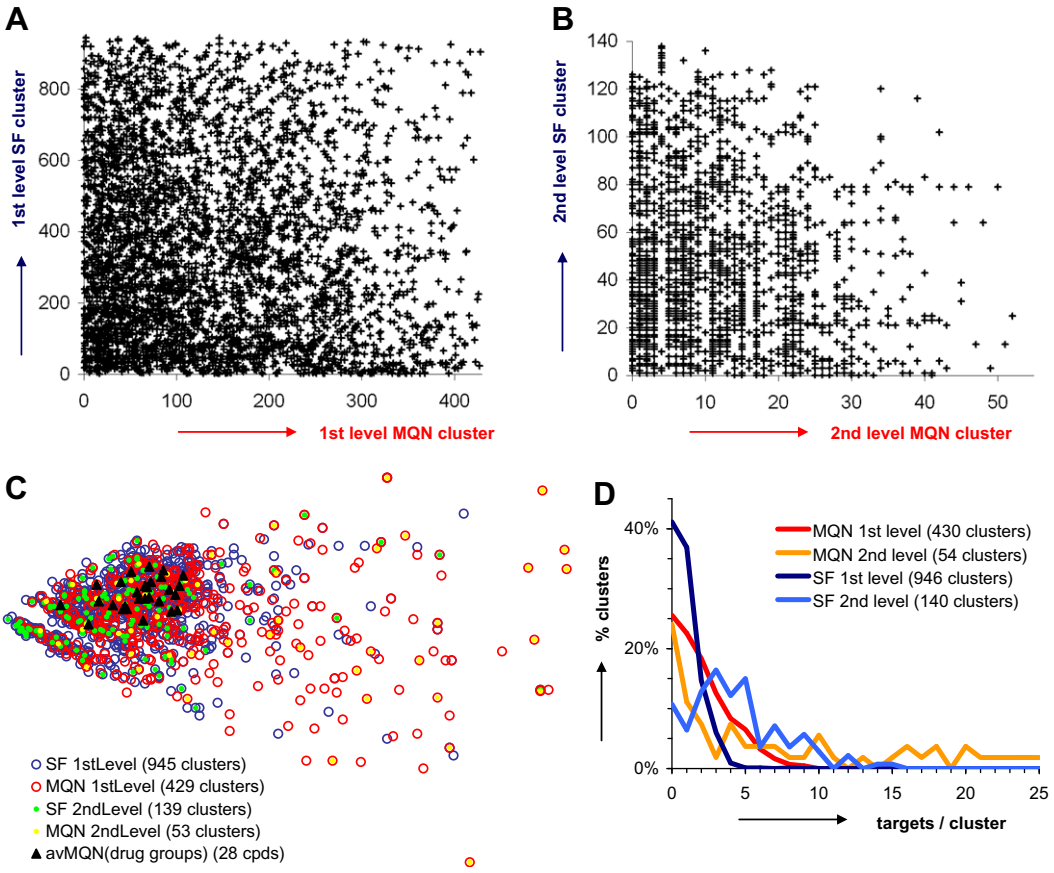

Figure 5. Assignment of DrugBank compounds to clusters from clustering by CBD_{MQN} or T_{SF} at (A) 1st level and (B) 2nd level clustering (cluster are ordered in decreasing size according to Fig. 4C and D). (C). Distribution of the MQN- and SF-cluster centers and drug group centers in the (PC1,PC2) map of MQN-space. (D). Target specificity of MQN- and SF-clusters analyzed for the drugs in the selected 28 target-specific groups in Figure 2.

to SMILES. Each molecule was annotated with its target property as given in DrugBank. For the formation of 28 target specific groups we considered only groups of at least 20 molecules with a single reported target property. MQNs were computed using the previously reported source code (Supplementary data in Ref. 6). The source code was written in Java using JChem from Chemaxon, Ltd as a starting library. For SF, a Daylight-type 1024-bit hashed fingerprint from ChemAxon were computed using the JChem library.

4.2. Similarity calculations

The city block distance (CBD), also known as *Manhattan* distance or *absolute value* distance, is the sum of the *absolute differences* between coordinates of a pair of objects. City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity. The City block distance between two points($CBD_{A,B}$), *A* and *B*, with *K* dimensions is calculated as:

$$CBD_{A,B} = \sum_{j=1}^{K} |A_j - B_j|$$

For the molecule *A* and *B* represented by vector X_A and X_B with length n and attributes *j*, their Tanimoto Similarity Coefficient ($T_{A,B}$)[21] is calculated as:

$$T_{A,B} = \frac{\sum_{j=1}^{n} X_{jA}X_{jB}}{\sum_{j=1}^{n}(X_{jA})^2 + \sum_{j=1}^{n}(X_{jB})^2 - \sum_{j=1}^{n} X_{jA}X_{jB}}$$

4.3. Principal Component Analysis (PCA)

PCA was performed using an in house program written in JAVA using the JSci science library.

4.4. ROC curves

Enrichment studies were carried out with 31 target groups (28 target specific groups and 3 random sets) using four methods CBD_{MQN}, T_{MQN}, CBD_{SF} and T_{SF} as similarity measures. The molecule that was most similar to all other molecules within each group (the molecule with highest T_{SF}, respectively lowest CBD_{MQN} to all other molecules within each group) was used as reference for this group. These four sets of 31 reference molecules were then used for the ROC by sorting DrugBank according to the shortest CBD_{MQN}/CBD_{SF} and highest T_{MQN}/T_{SF}.

4.5. Affinity propagation (AP)

The AP is a clustering algorithm also known as message-passing algorithm.[18] AP does not require the user to input the number of clusters in advance, rather it automatically comes up with the optimum number of clusters based upon input self-similarity value. AP shows a much lower error when compare with other clustering algorithms. The APCluster: affinity propagation algorithm from R-Statistical package[22] and JAVA programming language were used for the clustering of DrugBank and for further analysis. The initial clustering provided 429 and 945 clusters for MQN and SF respectively. These clusters are referred as 1st level clusters. The cluster

centers from 1st level clusters were then subjected to affinity propagation clustering which results in 53 and 139 clusters for MQN and SF respectively. Based upon these clusters of '1st level cluster centers' we then merged the 1st level clusters into bigger clusters called as 2nd level clusters.

Acknowledgments

This work was supported financially by the University of Berne, the Swiss National Science Foundation, and the NCCR TransCure.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmc.2012.03.017.

References and notes

1. (a) Peltason, L.; Bajorath, J. *Future Med. Chem.* **2009**, *1*, 451; (b) Guha, R. *Methods Mol. Biol.* **2011**, *672*, 101.
2. Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H. J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E. *Future Med. Chem.* **2011**, *3*, 751.
3. (a) Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Design* **1998**, *9–11*, 339; (b) Geppert, H.; Vogt, M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 205; (c) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. *Med. Chem. Commun.* **2010**, *1*, 30; (d) Akella, L. B.; DeCaprio, D. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325; (e) Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225; (f) Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28; (g) Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157; (h) Rosen, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. *J. Med. Chem.* **2009**, *52*, 1953.
4. Willett, P. *Drug Discovery Today* **2006**, *11*, 1046.
5. (a) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W. *J. Chem. Inf. Model.* **2009**, *49*, 108; (b) Khalifa, A. A.; Haranczyk, M.; Holliday, J. *J. Chem. Inf. Model.* **2009**, *49*, 1193.
6. Nguyen, K. T.; Blum, L. C.; Van Deursen, R.; Reymond, J.-L. *ChemMedChem* **2009**, *4*, 1803.
7. Wang, S. G.; Schwarz, W. H. *Angew. Chem., Int. Ed.* **2009**, *48*, 3404.
8. (a) Flower, D. R. *J. Mol. Graph. Model.* **1998**, *16*, 239; (b) Bender, A.; Glen, R. C. *J. Chem. Inf. Model.* **2005**, *45*, 1369.
9. van Deursen, R.; Blum, L. C.; Reymond, J. L. *J. Chem. Inf. Model.* **2010**, *50*, 1924.
10. van Deursen, R.; Blum, L. C.; Reymond, J. L. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 649.
11. (a) Fink, T.; Bruggesser, H.; Reymond, J. L. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504; (b) Fink, T.; Reymond, J. L. *J. Chem. Inf. Model.* **2007**, *47*, 342; (c) Blum, L. C.; Reymond, J. L. *J. Am. Chem. Soc.* **2009**, *131*, 8732; (d) Blum, L. C.; Vandeursen, R.; Reymond, J. L. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637.
12. Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. *J. Chem. Inf. Model.* **2011**, *51*, 3105.
13. Reymond, J. L.; Blum, L. C.; Van Deursen, R. *Chimia* **2011**, *65*, 863.
14. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. *Nucleic Acids Res.* **2011**, *39*, D1035.
15. Nicholl, A. *Methods Mol. Biol.* **2011**, *672*, 531.
16. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894.
17. Wermuth, C. G. *Drug Discovery Today* **2006**, *11*, 348.
18. Frey, B. J.; Dueck, D. *Science* **2007**, *315*, 972.
19. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.* **2002**, *45*, 4350.
20. Perez-Nueno, V. I.; Ritchie, D. W. *J. Chem. Inf. Model.* **2011**, *51*, 1233.
21. Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
22. Bodenhofer, U.; Kothmeier, A.; Hochreiter, S. *Bioinformatics* **2011**, *27*, 2463.