DOI: 10.1002/cmdc.200900317

Classification of Organic Molecules by Molecular Quantum Numbers

Kong T. Nguyen, Lorenz C. Blum, Ruud van Deursen, and Jean-Louis Reymond*[a]

The periodic table classifies elements by increasing atomic number in periods following the principal quantum number, and allows their physicochemical properties to be rationalized.[1] Herein, we propose a related system for organic molecules based on 42 molecular quantum numbers (MQNs), defined here as counts for simple structural features such as atom, bond and ring types, creating a multidimensional grid called MQN space. In analogy to the elements and their isotopes grouped in each entry of the periodic table, MQN isomers have identical MQNs and occupy the same position in MQN space. The MQN system is able to analyze large molecular databases and clusters compounds with similar structure, physicochemical properties and bioactivities, as illustrated for the databases ZINC[2] and GDB-11.[3]

Organic molecules can be named by systematic nomenclature,[4] or coded in line notations such as SMILES[5] or InChI.[6] These methods achieve an exact description of the molecular structure, but only provide a unidimensional classification, which is of limited use for analyzing molecular diversity. More recently, chemical space has emerged as a concept to classify large molecular databases.[7] Chemical space is most often represented as a property space whose dimensions measure a combination of structural parameters and predicted physicochemical properties, allowing useful data mining, such as the quest for natural products analogues.[8] While many descriptors of molecular structures and properties of varying complexity are known and may be used for defining chemical space,[9] we set out to test whether a system based only on counts for simple structural features (MQNs) might produce an easily accessible and logical classification system for organic molecules.

MQNs were considered counting atoms and bonds, polarity and topology (Table 1). MQNs were determined for the ZINC database, listing 8.4 million organic molecules,[2] and the GDB-11 database, listing 26.4 million possible molecules with up to 11 atoms of C, N, O, F,[3] giving a total of 6 501 005 MQN combinations, or MQN bins (Table 2).[10] Molecules with identical MQNs (MQN isomers) were strongly related structural isomers (Figure 1).

Principal component analysis (PCA) showed that MQN space organizes molecules by structural types. For ZINC, 73 % of the variability is visible in the PC1/PC2 plane. PC1 mostly represents molecular size (Figure 2a, and figure S1 in the Supporting Information). Molecules appear in elongated clusters distributed along the ascending diagonal with increasing number of rings (Figure 2b).

Table 1. Molecular quantum numbers.

Category 1: Atom counts		**Category 3: Polarity counts**	
1.	c (carbon)	20.	hbam (H-bond acceptor sites)
2.	f (fluorine)	21.	hba (H-bond acceptor atoms)
3.	cl (chlorine)	22.	hbdm (H-bond donor sites)
4.	br (bromine)	23.	hbd (H-bond donor atoms)
5.	i (iodine)	24.	negc (negative charges)[b]
6.	s (sulfur)	25.	posc (positive charges)[b]
7.	p (phosphorous)		
8.	an (acyclic nitrogen)	**Category 4: Topology counts**[c]	
9.	cn (cyclic nitrogen)	26.	asv (acyclic single valent nodes)
10.	ao (acyclic oxygen)	27.	adv (acyclic divalent nodes)
11.	co (cyclic oxygen)	28.	atv (acyclic trivalent nodes)
12.	hac (heavy atoms)[a]	29.	aqv (acyclic tetravalent nodes)
		30.	cdv (cyclic divalent nodes)
Category 2: Bond counts		31.	ctv (cyclic trivalent nodes)
13.	asb (acyclic single bonds)	32.	cqv (cyclic tetravalent nodes)
14.	adb (acyclic double bonds)	33.	r3 (3-membered rings)
15.	atb (acyclic triple bonds)	34.	r4 (4-membered rings)
16.	csb (cyclic single bonds)	35.	r5 (5-membered rings)
17.	cdb (cyclic double bonds)	36.	r6 (6-membered rings)
18.	ctb (cyclic triple bonds)	37.	r7 (7-membered rings)
19.	rbc (rotatable bonds)	38.	r8 (8-membered rings)
		39.	r9 (9-membered rings)
		40.	rg10 (≥ 10-membered rings)
		41.	afrc (nodes shared by ≥ 2 rings)
		42.	bfrc (edges shared by ≥ 2 rings)

[a] All non-H atoms. [b] Predicted charges at pH 7. [c] For parent graph without H's.

Table 2. MQN-statistics for ZINC and GDB-11.

	ZINC	GDB-11
no. of compounds	8 436 272	26 434 567
no. of MQN-bins	3 654 836	2 859 938
no. of single occupied MQN-bins	1 832 566	660 851
no. of compounds in most occupied MQN-bin	300	1982
no. of shared MQN-bins	13 769	13 769
no. of compounds in shared MQN-bins	30 779	254 604

The number of rotatable bonds (rbc) representing molecular flexibility, the number of H-bond acceptor sites (hbam, counting nonbonding electron pairs on N- and O-atoms) and the topological surface area (TPSA in Å^2)[11] indicative of polarity, all increase along the descending diagonal (Figure 2c, d & e). The calculated water–octanol partition coefficient (clog *P*)[12] follows molecular size and, in part, rings and hbam (Figure 2f). PCA of GDB-11 shows similar patterns in the PC1/PC2 plane, containing 63 % of the variability (Supporting Information, figure S2/S3).

Interestingly, compounds with similar bioactivities form groups in MQN space. Ranking by MQN distance (calculated as

[a] *Dr. K. T. Nguyen, L. C. Blum, R. van Deursen, Prof. Dr. J.-L. Reymond*
Departement of Chemistry and Biochemistry, University of Berne
Freiestrasse 3, 3012 Berne (Switzerland)
Fax: (+41) 31-631-8057
E-mail: jean-louis.reymond@ioc.unibe.ch

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cmdc.200900317.

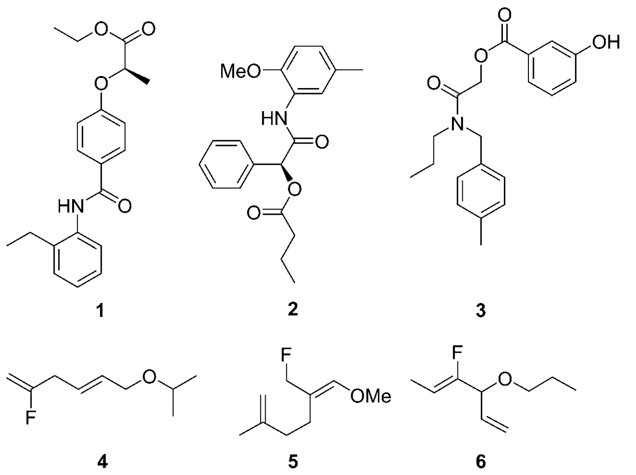

Figure 1. Examples of MQN isomers from the most occupied MQN bin in ZINC (**1**–**3**) (300 compounds) and GDB-11 (**4**–**6**) (1982 compounds).

city-block distance = sum of absolute differences between MQNs) to a reference bioactive ligand efficiently sorts other active compounds from ZINC (Figure 2 g, h & i, and figure S4 in the Supporting Information). Sorting is comparable to ranking by Tanimoto similarity coefficients for structural fingerprints,[13] although the two measures are only weakly correlated (Supporting Information, figure S5). Despite their simplicity, MQNs capture relevant structural features for bioactivity and might be useful in virtual screening for pharmaceutically relevant compounds.

In summary, MQNs provide a readily accessible classification system for organic molecules. MQNs can be determined manually from any structural formula. The method defines a universal chemical space in which organic molecules from any database of interest can be placed.

Acknowledgements

This work was financially supported by the University of Berne and the Swiss National Science Foundation. The authors gratefully acknowledge ChemAxon Ltd. for donation of the academic license to the JChem package.

Keywords: classification · databases · drugs · quantum numbers · the periodic table

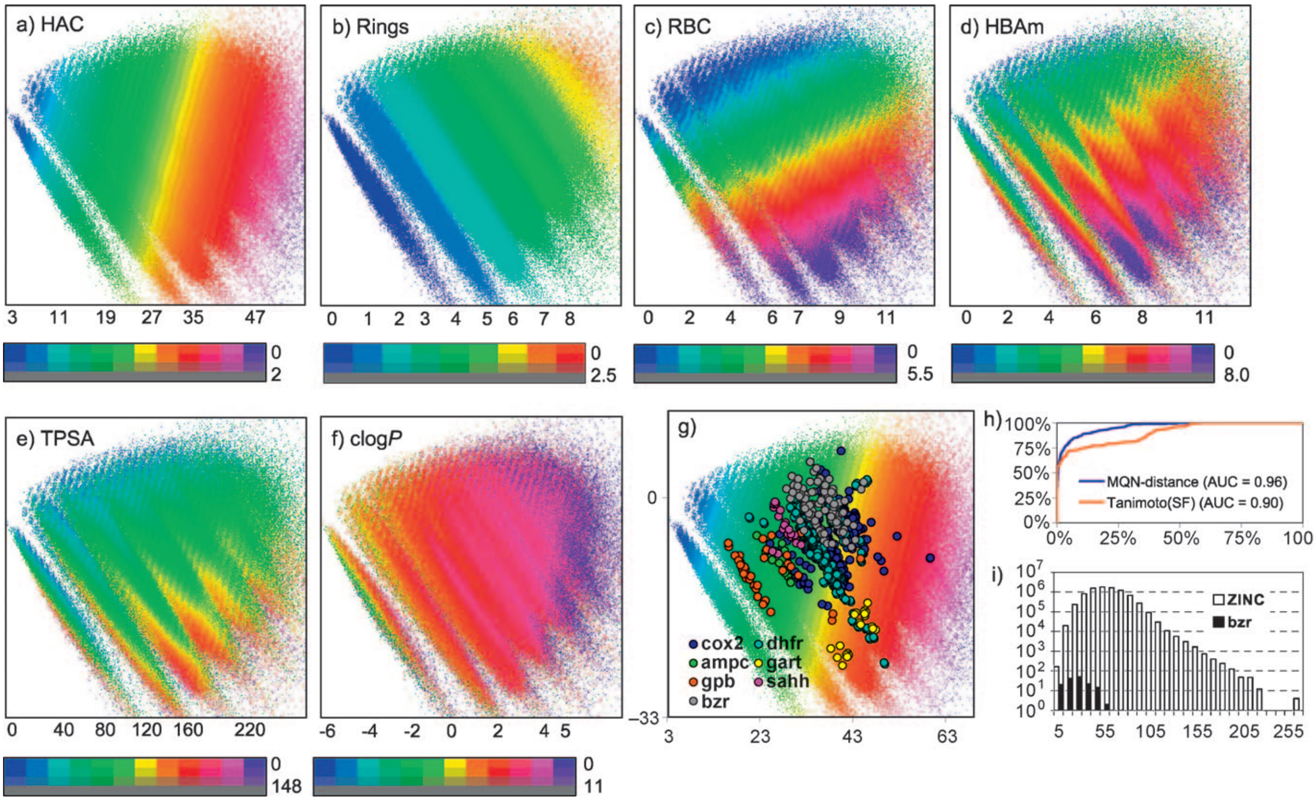

Figure 2. Principal component analysis (PCA) of the ZINC database in MQN space. PC1: horizontal axis, 49 % of total variability (range: +3.06 to +68.76); PC2: vertical axis, 24 % of total variability (range: −32.58 to +12.93). PCA was computed using non-normalized MQN values. a)–f) The surface is color coded in HSL code for average value (hue scale) and standard deviation (saturation to grey) according to the scales below each image for selected MQN (a–d) and computed molecular properties (e–f). g) Bioactivity classes in MQN space. Positions of known inhibitors in MQN space in the PC1/PC2 plane of ZINC (cox2 = cyclooxygenase 2; dhfr = dihydrofolate reductase; ampc = AmpC beta lactamase; gart = glycinamide ribonucleotide transformylase; gpb = glycogen phosphorylase β; sahh = S-adenosyl-homocysteine hydrolase;[14] bzr = benzodiazepine receptor[15]). h) Enrichment curve showing recovered bzr ligands (%) from ZINC as a function of the database screened (%) when ranked by MQN distance or by Tanimoto similarity coefficient to a reference nabenzodiazepine taken at the corresponding cluster center. i) Logarithmic scale histogram of the number of ZINC or bzr compounds as a function of MQN distances to same reference. See also Supporting Information, figure S4.

[1] S.-G. Wang, W. H. Schwarz, *Angew. Chem.* **2009**, *121*, 3456–3467; *Angew. Chem. Int. Ed.* **2009**, *48*, 3404–3415.

[2] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

[3] a) T. Fink, H. Bruggesser, J.-L. Reymond, *Angew. Chem.* **2005**, *117*, 1528–1532; *Angew. Chem. Int. Ed.* **2005**, *44*, 1504–1508; b) T. Fink, J.-L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342–353; c) L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

[4] H. A. Favre, K.-H. Hellwich, G. P. Moss, W. H. Powell, J. G. Traynham, *Pure Appl. Chem.* **1999**, *71*, 1327–1330.

[5] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[6] http://www.iupac.org/inchi/, (last accessed: July 2, 2009).

[7] a) R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50; b) T. I. Oprea, *Curr. Opin. Chem. Biol.* **2002**, *6*, 384–389; c) M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272–17277; d) S. N. Pollock, E. A. Coutsias, M. J. Wester, T. I. Oprea, *J. Chem. Inf. Model.* **2008**, *48*, 1304–1310; e) J. L. Medina-Franco, K. Martinez-Mayorga, M. A. Giulianotti, R. A. Houghten, C. Pinilla, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322–333; f) N. Singh, R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten, J. L. Medina-Franco, *J. Chem. Inf. Model.* **2009**, *49*, 1010–1024.

[8] a) T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166; b) P. Ertl, S. Jelfs, J. Muehlbacher, A. Schuffenhauer, P. Selzer, *J. Med. Chem.* **2006**, *49*, 4568–4573; c) J. Larsson, J. Gottfries, S. Muresan, A. Backlund, *J. Nat. Prod.* **2007**, *70*, 789–794; d) A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, J. W. Davies, *J. Chem. Inf. Model.* **2009**, *49*, 108–119; e) J. Rosén, J. Gottfries, S. Muresan, A. Backlund, T. I. Oprea, *J. Med. Chem.* **2009**, *52*, 1953–1962.

[9] a) P. Kolb, A. Caflisch, *J. Med. Chem.* **2006**, *49*, 7384–7392; b) V. J. Sykora, D. E. Leahy, *J. Chem. Inf. Model.* **2008**, *48*, 1931–1942.

[10] The average processing rate for determining the MQN was 8.2 million SMILES h^{-1} on an Intel Core 2 Duo (2.13 GHz) machine with 3 GB of memory or 10.5 million SMILES h^{-1} on an AMD Athlon X2 6000+ (3.0 GHz) machine with 3 GB of memory.

[11] a) N. Huang, B. K. Shoichet, J. J. Irwin, *J. Med. Chem.* **2006**, *49*, 6789–6801; b) M. von Korff, J. Freyss, T. Sander, *J. Chem. Inf. Model.* **2009**, *49*, 209–231.

[12] J. J. Sutherland, L. A. O'Brien, D. F. Weaver, *J. Med. Chem.* **2004**, *47*, 5541–5554.

[13] P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.* **2000**, *43*, 3714–3717.

[14] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, R. K. Robins, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

[15] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

Received: July 30, 2009
Published online on September 22, 2009