Asymmetric Clustering Index in a case study of 5-HT$_{1A}$ receptor ligands

M. Śmieja[1,*], D. Warszycki[2], J. Tabor[1], A. J. Bojarski[2]

1 Faculty of Mathematics and Computer Science, Jagiellonian University, 6 Lojasiewicza Street, 30-348 Kraków, Poland

2 Institute of Pharmacology, Polish Academy of Sciences, 12 Smetna Street, 31-343 Kraków, Poland

∗ **E-mail: marek.smieja@uj.edu.pl**

Abstract

The automatic clustering of chemical compounds is an important branch of chemoinformatics. In this paper the Asymmetric Clustering Index (Aci) is proposed to assess how well an automatically created partition reflects the reference. The asymmetry allows for a distinction between the fixed reference and the numerically constructed partition. The introduced index is applied to evaluate the quality of hierarchical clustering procedures for 5-HT$_{1A}$ receptor ligands. We find that the most appropriate combination of parameters for the hierarchical clustering of compounds with a determined activity for this biological target is the Klekota Roth fingerprint combined with the complete linkage function and the Buser similarity metric.

Introduction

The rapidly growing number of compounds with a determined activity for a given molecular target leads to difficulties in using full, previously explored chemical spaces in virtual screening campaigns. Indeed, the use of a large number of ligands (e.g., the D_2 receptor has 9180 different ligands in ChEMBL database v. 16 [1]) in predictive model development usually generates substantial computational costs. Moreover, for active compounds of any protein target, large groups of similar ligands may significantly disrupt the search results, limiting virtual hits to close analogs of over-representative input structures [2, 3]. As a consequence, an appropriate clustering of the ligands' chemical space is of primary importance [4].

Manual (knowledge-based) clustering is usually the first choice for small groups of ligands because it provides the most natural partitions. However, for more abundant sets, this approach is time-consuming and requires extensive chemical knowledge (e.g., the manual clustering of 3616 5-HT$_{1A}$ receptor ligands performed by Warszycki et al. [5] took a couple of weeks). Therefore, automatic clustering algorithms are frequently used for categorizing chemical compounds. Consequently, it is crucial to employ indices that can verify how similar a numerically constructed partition is to the reference created by experts.

Unlike experts, who intuitively recognize and classify chemical structure, automatic clustering algorithms require molecule to be translated into an appropriate form. This is usually achieved by application of fingerprints which transform chemical structure on a bitstring, where "1" and "0" correspond to a presence or absence of a particular chemical pattern, respectively [6, 7]. Next, fingerprints can be compared using a similarity metric evaluating how much the compounds are similar [8]. Moreover, hierarchical clustering procedures require, the linkage function which determines the "distance" between two groups of compounds. Since there are a lot of available fingerprints, metrics and linkage functions, the number of their combinations is indeed quite high, which makes finding the most appropriate one, for a particular task, relatively difficult.

Several methods have been proposed to compare clusterings [9]. The most popular techniques are based on counting pairs of elements classified in the same way in both partitions, such as the rand index [10] and its modifications [11, 12]. Another group of methods uses normalized mutual information to quantify the information shared by the clusterings [13, 14]. An interesting approach for comparing

partitions relies on measuring the distance between clusterings with the use of information theory [15]. The main feature of these indices is their symmetry, which makes them suitable for finding the similarities between clusterings.

In the present study, we introduce the Asymmetric Clustering Index (Aci) for comparing two partitions. The asymmetry allows the index to distinguish between the fixed reference[1] $\mathcal{R}$ and the numerically constructed partition $\mathcal{C}$. As a consequence, the Aci is capable of measuring how well a given partition reflects the reference (not conversely). This index is defined as the ratio of the mutual information $\mathrm{MI}(\mathcal{R},\mathcal{C})$ to the entropy $\mathrm{SE}(\mathcal{R})$:

$$\mathrm{Aci}_{\mathcal{R}}(\mathcal{C}) = \frac{\mathrm{MI}(\mathcal{R},\mathcal{C})}{\mathrm{SE}(\mathcal{R})}.$$

The Aci is reminiscent of the indices proposed in [13, 14] but, due to its different normalization factor, has an asymmetry feature.

The basic properties of the Aci are presented in Figure 1 and are listed below:

- it takes on values between 0 and 1,
- the reference can be recovered from the partition by merging selected groups if $\mathrm{Aci} = 1$,
- for the partitions that do not share any information, $\mathrm{Aci} = 0$.

Therefore, for successively subdivided partitions, the Aci converges to 1, in contrast to symmetric indices. Figure 2 presents the values of the Aci and other two similarity indices based on mutual information for a conducted experiment. When the number of clusters obtained in the hierarchical clustering is greater, the reference is better reflected by the partition. As a result, the Aci takes gradually higher values in contrast to the other indices. This behavior allows for a straightforward interpretation of the Aci – values close to 1 indicate that the numerically constructed partition contains much information about the reference.

To determine the optimal conditions reaching the maximum Aci values, 8 fingerprint types, 4 similarity metrics and 4 linkage functions were applied to a hierarchical clustering of the full chemical space of 5-$\mathrm{HT}_{1\mathrm{A}}$ receptor ligands (see Supplementary Information). As a reference, the manually constructed partition of Warszycki et al. [5] was taken, which generally follows the classification of 5-$\mathrm{HT}_{1\mathrm{A}}$R described in the literature [16,17]. The best clustering was achieved for a combination of the Klekota Roth fingerprint, the Buser similarity metric and the complete linkage function, which was then verified in an additional clustering experiment on a collection of compounds belonging to two explicitly different chemical classes. Thus, in further studies, automatic clustering should be performed with these parameters.

Materials and Methods

The Aci measures how well the automatically performed partition $\mathcal{C} = \{C_1, \ldots, C_n\}$ reflects the reference $\mathcal{R} = \{R_1, \ldots, R_m\}$. This index is obtained by normalizing the mutual information $\mathrm{MI}(\mathcal{R},\mathcal{C})$ by the entropy $\mathrm{SE}(\mathcal{R})$:

$$\mathrm{Aci}_{\mathcal{R}}(\mathcal{C}) = \frac{\mathrm{MI}(\mathcal{R},\mathcal{C})}{\mathrm{SE}(\mathcal{R})} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} P(R_i \cap C_j)\log_2 \frac{P(R_i \cap C_j)}{P(R_i)P(C_j)}}{-\sum_{i=1}^{m} P(R_i)\log_2 P(R_i)}, \tag{1}$$

where $P(A)$ denotes the probability that an element belongs to set A. The above metric quantifies the percent of information that $\mathcal{R}$ delivers about $\mathcal{C}$.

[1] which, by default, denotes the expert manual partition

The Aci attains a maximal value of 1 if the reference and the numerically constructed partitions are identical. However, as shown in Figure 3, we also obtain $\text{Aci}_{\mathcal{R}}(\mathcal{C}_1) = 1$ when the reference is subdivided into smaller clusters; clearly, this automatically constructed clustering contains at least as much information as the reference. Consequently, the reference can be reconstructed from the numerically obtained partition by merging selected groups. In contrast, if the partition $\mathcal{C}_2$ is random with respect to $\mathcal{R}$, then the clusterings are completely different, which results in $\text{Aci}_{\mathcal{R}}(\mathcal{C}_2) = 0$. This case holds, for example, when every cluster of $\mathcal{C}_2$ contains an equal number of elements in comparison to each cluster of $\mathcal{R}$. One can also consider a composition of these two examples.

In the case of hierarchical clustering, for every two partitions obtained by cutting at different levels, one partition is a subdivision of the second. Furthermore, when a partition has as many groups as the number of data-set elements (every cluster is a one-element set), then it contains information about every possible partition. Clearly, for a high number of clusters, practically all information about the reference partition can be deduced from the partition numerically constructed by an arbitrary clustering algorithm. In contrast, a partition cannot fully reflect the reference if it has fewer elements. Consequently, one of the possible methods for determining the optimal number of clusters is to maximize a selected measure of dispersion, e.g., the standard deviation or entropy. In other words, a given number of clusters is optimal for the Aci if it maximally distinguishes among the partitions (with respect to the corresponding Aci values). Numerical examples indicate that reasonable results are obtained when approximately twice the number of groups are taken in comparison to the reference division (see the next section for more details).

The idea of the Aci is based on information theory; in particular, this index involves the notions of entropy and mutual information content. The Shannon entropy, introduced as a measure of channel capacity in digital communications [18], is also used to quantify the information contained in the clustering [19]. Formally, the Shannon entropy (SE) of an n-element partition $\mathcal{C} = \{C_1, \ldots, C_n\}$ is defined by

$$\text{SE}(\mathcal{C}) = -\sum_{i=1}^{n} P(C_i) \log_2 P(C_i).$$

In the case of a one-element partition, the cluster of each element is known; therefore the SE equals 0. In contrast, if no information about the position of any element is provided (every cluster is equally probable), then the SE attains a maximum.

To compare two clusterings, the basic idea of the SE needs to be extended by defining the mutual information (MI). The MI determines the amount of information shared between partitions and is defined by [20]

$$\text{MI}(\mathcal{R}, \mathcal{C}) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(R_i \cap C_j) \log_2 \frac{P(R_i \cap C_j)}{P(R_i) P(C_j)}.$$

The relations between the introduced quantities are presented in Figure 4.

It is straightforward to demonstrate that the mutual information is symmetric [20], i.e.,

$$\text{MI}(\mathcal{R}, \mathcal{C}) = \text{MI}(\mathcal{C}, \mathcal{R}).$$

As mentioned in the Introduction, this property allows only one to evaluate the similarity between partitions. To define an asymmetric index that measures how well the reference can be recovered from the numerically created partition, the normalization by the entropy of reference partition is used, giving the following formula:

$$\text{Aci}_{\mathcal{R}}(\mathcal{C}) = \frac{\text{MI}(\mathcal{R}, \mathcal{C})}{\text{SE}(\mathcal{R})}.$$

By [20], we have $0 \leq \text{MI}(\mathcal{R}, \mathcal{C}) \leq \text{SE}(\mathcal{R})$, which leads to:

$$0 \leq \text{Aci}_{\mathcal{R}}(\mathcal{C}) \leq 1.$$

Results

One of the most popular techniques used to divide chemical compounds is hierarchical clustering [21]. The strength of this approach lies in the deterministic nature of the algorithm and the constructed hierarchical structure of clusters. This method requires the specification of several input parameters, but there is no unified methodology for determining which parameters will provide the best results. The ACI will be applied to determine the combination of parameters that best reflect the reference partition of 5-HT_{1A} receptor ligands.

As a reference, the manually constructed partition of Warszycki [5] was utilized. All ligands (retrieved from approximately 520 published papers) used for this clustering were extracted from ChEMBL database version 5 (August 2010) [1]. Ligands with an inhibition constant (K_i) of less than or equal to 100 nM were considered active; only these ligands were used for this clustering study.

The manual clustering generally follows the classification of 5-HT_{1A} ligands described in the literature (9 basic classes) [16,22,23]; however, some additional subgroups were then created, e.g., for arylpiperazines [17]. In the case of alkylamines (714 compounds), indole derivatives were first extracted and, with the exception of the tetrahydropyridoindoles, were divided depending on the distance between two crucial pharmacophore features: an aromatic system and a basic nitrogen atom. The entire procedure resulted in 28 clusters, each containing 17 to 605 compounds [5] (see Figure 5).

In this study, three types of hierarchical clustering parameters were examined. The study focused on determining the optimal ACI values from a combination of eight fingerprint representations (Table 1), four linkage functions (Table 2) and four similarity metrics (Table 3). Both recently published works [8,24] and our experience, supported by preliminary studies, indicate that these four metrics are the most relevant for clustering purposes.

To determine the optimal number of clusters for the ACI, an additional experiment was conducted. The ACI was evaluated for all combinations of linkage functions, fingerprint representations and similarity metrics (total of 128 cases). The corresponding standard deviations for each number of clusters were calculated, as shown in Figure 6. Because this study focuses on selecting the optimal parameters, standard deviations were also computed for 12 combinations that provided the highest mean ACI values (averaged over all possible numbers of groups). This restriction reduced the number of clusters for which the maximal discrimination was attained (Figure 7). As a consequence, a total of 50 groups was chosen as a reasonable compromise between accuracy and complexity for this model.

The results (Table 4) shows that the choice of linkage function has the most significant impact on the clustering results, regardless of the fingerprint representation or similarity metric (clearly, this holds only for the types of metrics employed herein). The mean ACI values calculated for the clusterings for particular linkage functions indicate that optimal performance is obtained with the complete linkage function.

An analysis of the ACI values for partitions with the complete linkage function and various fingerprint representations and similarity metrics (Figure 8) points out the superiority of the KRFP fingerprint for all four metrics. The impact of the similarity metrics was then assessed by varying the number of clusters from 28 to 100 in series of experiments with the complete linkage function and the KRFP molecular representation. This investigation (Figure 9) demonstrated the superiority of the Buser similarity metric over the remaining three types for almost all cluster numbers.

Next, the ability of the optimally designed hierarchical clustering to separate compounds belonging to different chemical classes was additionally evaluated. For this purpose, three partitioning experiments were performed: the separation of (a) arylpiperazines with a sulfona(i)mide fragment from aporphines, (b) benzodioxans from benzylpiperazines and (c) N4-alkyl and N4-unsubstituted arylpiperazines from arylalkilamines with a three-atom linker. In the first two cases, the automatic process perfectly or very closely (ACI = 1.00 and ACI = 0.93, respectively) reflected the reference clustering. In the third case the obtained result was highly unsatisfactory (ACI = 0.006); however, increasing the number of clusters up to three significantly improve the quality of the separation (ACI = 0.57). Fixing the number of clusters

to 6 resulted in ACI = 0.75, while ACI = 0.86 was obtained for eight clusters. These results confirm the need to enforce a greater number of groups in the clustering process than expected.

In conclusion, the experiments demonstrate that the automatic hierarchical clustering of 5-HT_{1A} receptor ligands provides the best results when implemented with the complete linkage function, the KRFP fingerprint representation and the Buser similarity metric. It is worth mentioning that satisfactory results are also obtained with the use of three other metrics – the Tanimoto, Yule and Dice metrics.

Conclusion

This paper introduces a straightforward asymmetric index, the ACI, which allows one to evaluate how well a numerically constructed partition reflects the reference. The highest ACI was consistently obtained for hierarchical clustering based on the complete linkage function, the Klekota-Roth fingerprint and the Buser similarity metric, suggesting the application of these parameters for other groups of biologically active compounds. This approach was verified using a manually constructed partition of active 5-HT_{1A} ligands [5].

Supplementary Information

An SDF file containing the full collection of 3616 compounds is available free of charge via the Internet at `http://skandal.if-pan.krakow.pl/5-HT1A_ligands.sdf`. To obtain a hierarchical clustering of the considered chemical space, the `hclust` function of `R` software was used. A sample `R` code used for the ACI calculation is available free of charge at `http://skandal.if-pan.krakow.pl/aci.R`.

Acknowledgments

This study was partially supported by the Polish-Norwegian Research Programme operated by the National Centre for Research and Development under the Norwegian Financial Mechanism 2009-2014 in the frame of Project PLATFORMex (Pol-Nor/198887/73/2013), the National Centre of Science from Poland (grant no. 2011/01/B/ST6/01887) and the Polish Ministry of Science and Higher Education from the budget for science in the years 2013–2015 (grant no. IP2012 055972).

References

1. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research 40: D1100–D1107.

2. Heikamp K, Bajorath J (2011) Large-scale similarity search profiling of ChEMBL compound data sets. Journal of Chemical Information and Modeling 51: 1831–1839.

3. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discovery Today 12: 225–233.

4. Brown N (2009) Chemoinformatics – an introduction for computer scientists. ACM Computing Surveys (CSUR) 41: 8.

5. Warszycki D, Mordalski S, Kristiansen K, Kafel R, Sylte I, et al. (2013) A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds – an application for 5-HT1A receptor ligands. PLoS ONE 8.

6. Willett P (2005) Searching techniques for databases of two-and three-dimensional chemical structures. Journal of Medicinal Chemistry 48: 4183–4199.

7. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today 11: 1046–1053.

8. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, et al. (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. Journal of Chemical Information and Modeling 52: 2884–2901.

9. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. Journal of Intelligent Information Systems 17: 107–145.

10. Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66: 846–850.

11. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218.

12. Yeung KY, Ruzzo WL (2001) Details of the adjusted rand index and clustering algorithms, supplement to the paper "an empirical study on principal component analysis for clustering gene expression data". Bioinformatics 17: 763–774.

13. Ana L, Jain AK (2003) Robust data clustering. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. IEEE, volume 2, pp. II–128.

14. Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3: 583–617.

15. Meilă M (2007) Comparing clusterings – an information based distance. Journal of Multivariate Analysis 98: 873–895.

16. Caliendo G, Santagada V, Perissutti E, Fiorino F (2005) Derivatives as 5HT1A receptor ligands-past and present. Current Medicinal Chemistry 12: 1721–1753.

17. Lopez-Rodriguez M, Ayala D, Benhamu B, Morcillo MJ, Viso A (2002) Arylpiperazine derivatives acting at 5-HT1A receptors. Current Medicinal Chemistry 9: 443–469.

18. Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5: 3–55.

19. Wagner S, Wagner D (2007) Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik.

20. Cover TM, Thomas JA (2012) Elements of information theory. John Wiley & Sons.

21. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32: 241–254.

22. Olivier B, Soudijn W, van Wijngaarden I (1999) The 5-HT1A receptor and its ligands: structure and function. In: Progress in Drug Research, Springer. pp. 103–165.

23. Jun OS, Ha HJ, Yoon CD, Kyung LH (2001) Serotonin receptor and transporter ligands-current status. Current Medicinal Chemistry 8: 999–1034.

24. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. Journal of Chemical Information and Modeling 50: 771–784.

25. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. Journal of Chemical Information and Computer Sciences 35: 1039–1045.

26. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, et al. (2003) The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. Journal of Chemical Information and Computer Sciences 43: 493–500.

27. Yap CW (2011) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 32: 1466–1474.

28. Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. Bioinformatics 24: 2518–2525.

29. Ewing T, Baber JC, Feher M (2006) Novel 2D fingerprints for ligand-based virtual screening. Journal of Chemical Information and Modeling 46: 2423–2431.

30. Murtagh F (1983) A survey of recent advances in hierarchical clustering algorithms. The Computer Journal 26: 354–359.

Figure Legends

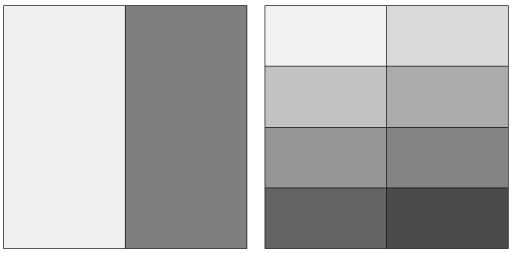

Figure 1. Presentation of the Aci. Partition $\mathcal{Q}$ contains more information than partition $\mathcal{P}$; thus, $\mathcal{P}$ can be restored from $\mathcal{Q}$ by merging four pairs of sets. In particular, $\textsc{Aci}_{\mathcal{P}}(\mathcal{Q}) = 1$ and $\textsc{Aci}_{\mathcal{Q}}(\mathcal{P}) = \frac{1}{3}$.

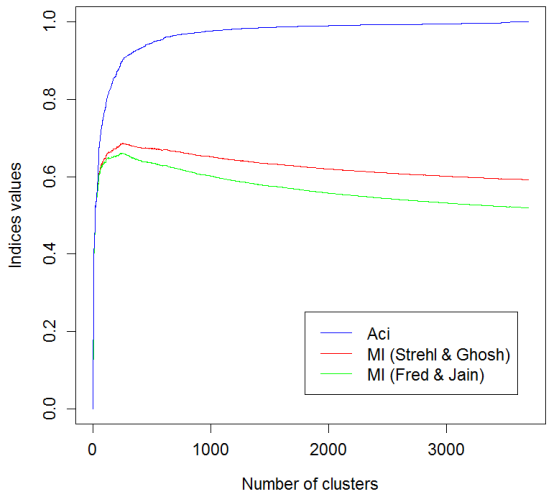

Figure 2. Comparison between the Aci and symmetric indices based on mutual information. These indices were evaluated based on the reference reported by Warszycki et al. [5], and the partitions were obtained from hierarchical clusterings performed with the Klekota Roth fingerprint combined with the Buser similarity metric and the complete linkage function.

Tables

Table 1. The characteristics of fingerprints, with the abbreviations used in this work.

Fingerprint	Abbreviation	Length of fingerprint
EState fingerprint [25]	estate	79
Fingerprint [26]	fingerprint	1024
Extended fingerprint [27]	extended	1024
Graph only fingerprint [27]	graph only	1024
Klekota Roth fingerprint [28]	KRFP	4860
MACCS fingerprint [29]	maccs	166
PubChem fingerprint [27]	pubchem	881
Substructure fingerprint [27]	substructure	308

All fingerprints were generated in PaDEL software [27].

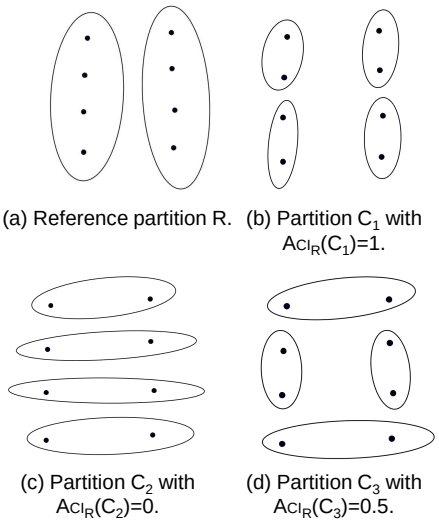

Figure 3. Illustration of the Aci. Partition $\mathcal{C}_1$ fully reflects the reference, $\mathcal{R}$ ($\textsc{Aci}_{\mathcal{R}}(\mathcal{C}_1) = 1$). In contrast, partition $\mathcal{C}_2$ is random with respect to the reference – the two results do not share any information ($\textsc{Aci}_{\mathcal{R}}(\mathcal{C}_2) = 0$). Partition $\mathcal{C}_3$ is a combination of the two previous situations – half of the reference can be recovered from this clustering ($\textsc{Aci}_{\mathcal{R}}(\mathcal{C}_3) = 0.5$).

Table 2. Linkage functions for two sets [30].

Name	Formula
Average	$\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d(a,b)$
Centroid	$d(c_A, c_B)$
Complete	$\max\{d(a,b) : a \in A, b \in B\}$
Single	$\min\{d(a,b) : a \in A, b \in B\}$

Used marks in the formula: d – metric, c_A – center of set A, $|A|$ – cardinality of set A .

Table 3. Similarity metrics [8].

Name	Formula
Buser	$\frac{\sqrt{(cd)}+c}{\sqrt{(cd)}+a+b-c}$
Dice	$\frac{2c}{a+b}$
Tanimoto	$\frac{c}{a+b-c}$
Yule	$\frac{cd-AB}{cd+AB}$

Used marks in the formula: a – on bits in structure 1, b – on bits in structure 2, c – on bits in both 1 and 2, d – off bits in both 1 and 2, $A = a - c$, $B = b - c$.

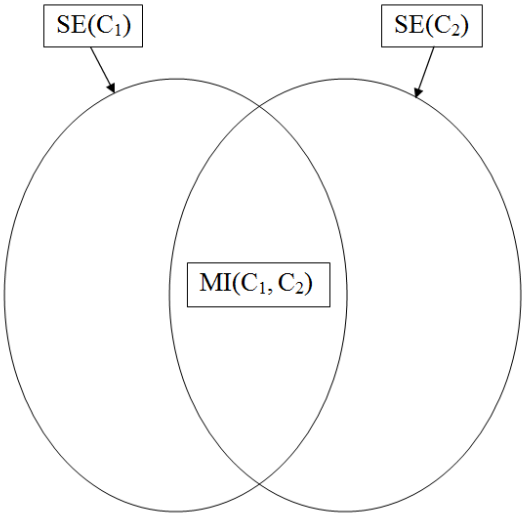

Figure 4. Comparison between entropy and mutual information. Each region describes the information provided by a particular clustering [20].

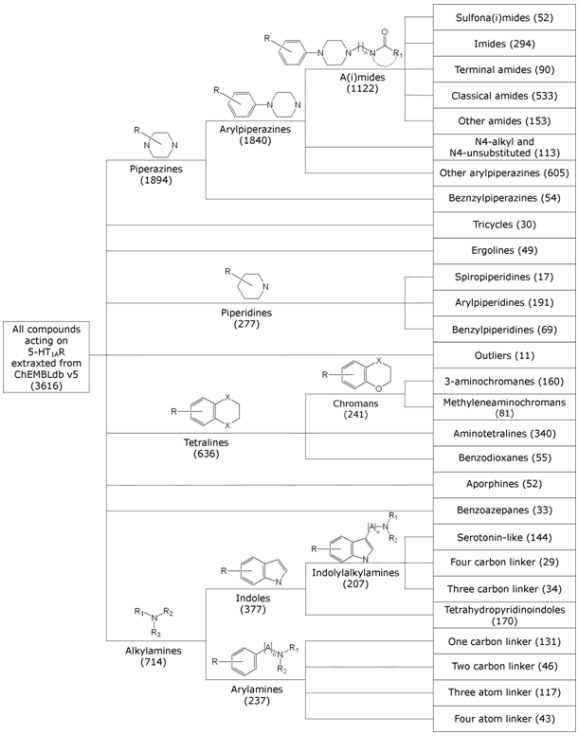

Figure 5. The results obtained by manual clustering of 5-HT$_{1A}$ receptor ligands. This process is described in Warszycki et al. [5].

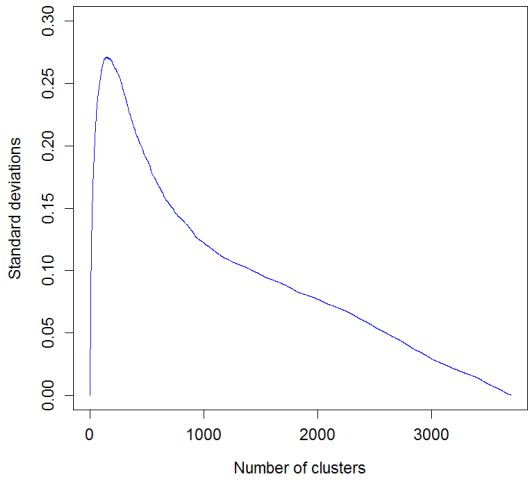

Figure 6. Standard deviations of Aci values collected for the 128 combinations of hierarchical clustering parameters.

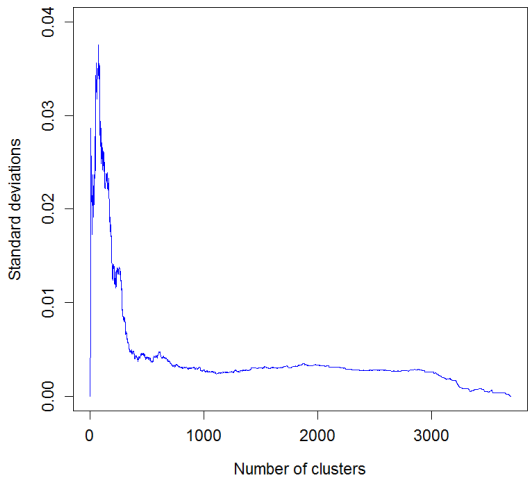

Figure 7. Standard deviations of Aci values collected for the 12 best combinations of hierarchical clustering parameters. These combinations correspond to the highest mean Aci values over all possible cluster numbers. The maximum occurs for the cluster numbers between 50 and 80.

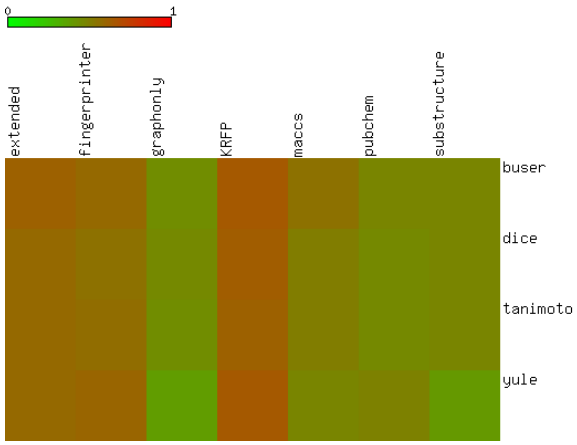

Figure 8. Aci values for hierarchical clusterings with the complete linkage function.

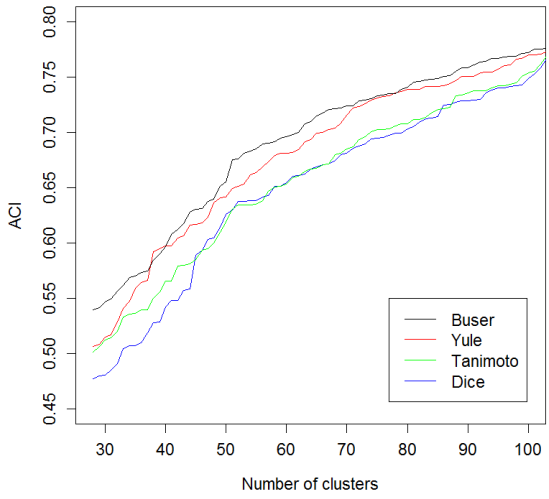

Figure 9. Aci values for hierarchical clusterings. The number of groups ranged from 28 to 100. Results are presented for the complete linkage function, the Klekota Roth fingerprint and four different similarity metrics.

Table 4. Complete linkage function rankings.

Linkage function	Aci
Complete	0.51
Average	0.40
Centroid	0.09
Single	0.04

Mean Aci values obtained for fixed four types of linkage functions and various types of fingerprints and similarity metrics.