

Sequential Superparamagnetic Clustering for Unbiased Classification of High-Dimensional Chemical Data

Thomas Ott,[†] Albert Kern,[†] Ausgar Schuffenhauer,[‡] Maxim Popov,[‡] Pierre Acklin,[‡]
Edgar Jacoby,[‡] and Ruedi Stoop^{*,†}

Institute for Neuroinformatics, University/ETH Zürich, Zurich, Switzerland, and Novartis Institutes for Biomedical Research, Discovery Technologies, Compound Logistics and Properties, Basel, Switzerland

Received March 16, 2004

For the clustering of chemical structures that are described by the Similog, ISIS count, and ISIS binary fingerprints, we propose a sequential superparamagnetic clustering approach. To appropriately handle nonbinary feature keys, we introduce an extension of the binary Tanimoto similarity measure. In our applications, data sets composed of structures from seven chemically distinct compound classes are evaluated and correctly clustered. The comparison, with results from leading methods, indicates the superiority of our sequential superparamagnetic clustering approach.

I. INTRODUCTION

Clustering is a fundamental process inherent to human cognition and perception. In the scene analysis paradigm, humans classify (group together) perceptual information, to develop a consistent picture of the environment. A priori, however, no information about the number of classes or class sizes is available. Classes or clusters are seldom clear-cut entities, a fact that has been exploited in fuzzy logic implementations. However, clustering has an inherent hierarchical branching nature, as is shown by the simple example of the partitioning of animals into species, subspecies, etc. In our approach, we consider clustering to be a self-organized learning process. Although the partitioning into classes, as we will show, can be achieved in an unbiased way, it is the **optimal choice of the resolution** that needs to be learned. In the latter aspect, we deal with a supervised learning approach. In technical applications, clustering is often applied to inhomogeneous sets of items that are represented as data points in a high-dimensional space. Given an appropriate measure of similarity in the form of a distance measure between points, the clustering approach then aims at identifying the clusters made up of similar data points. The rationale for classification and clustering lies in the similar-property principle. It states that structurally similar objects tend to have similar properties.

In chemistry, the clustering of substances is therefore of central interest.¹ Given a molecule of known properties, compounds that are structurally similar are likely to exhibit similar properties.² An important scientific question is to what extent criteria used to describe substances are “natural” and “efficient” and “pave the way towards new discoveries”. Economically, particularly in drug testing, the identification of “natural” classes is of paramount importance. If the classes are optimally chosen, efficient testing can be organized. This can be achieved by imposing a clustering resolution that is

in compliance with both the testing resources available and the need to distribute these tests over the ensemble. For an appropriate distribution of tests, it is not the chemical space but the relevant compound property space that is of importance. Clustering is able to unbiasedly determine the latter.

Clustering based on this insight has been applied for the prediction of physicochemical and biological properties of chemical compounds,^{3,4} the selection of diverse/representative compounds or reagent sets for the design of combinatorial libraries,⁵ compound acquisition selection,^{6,7} and the compilation of screening sets.⁸ A recent application deals with the analysis of high-throughput and the profiling of screening data.^{8–10} For clustering to be successful, an appropriate description of molecule structures and an adequate clustering algorithm are both essential. Among the variety of methods, the **hierarchical Ward**,¹¹ the **nonhierarchical Jarvis-Patrick**,¹² and the **K-means relocation methods**¹³ have become popular, mostly based on **fingerprint descriptors**.¹⁴ In chemical applications (separation of known actives and inactives, property prediction, and diversity selection), Ward’s method clearly outperforms the other two.^{1,8,14,15} Structural classes with very different shapes, densities, or sizes, however, present a problem to Ward’s method. Due to these inhomogeneities, the algorithm fails to properly discriminate classes. We introduce a modified superparamagnetic clustering approach that is not affected by these problems. **Superparamagnetic clustering**^{16,17} is based on the statistical, thermodynamical, description of the interactions between ferromagnetic Potts spin particles. Our comparison with Ward clustering will be based on a model system composed of well-defined chemical structural classes. ISIS count, ISIS binary keys,^{18,19} and Similog 2D pharmacophore triplets hologram descriptors²⁰ are the inputs for the similarity measure. One advantage of the new method is that it intrinsically provides the possibility to determine the optimal number of clusters, i.e., a natural clustering level. With the hierarchical Ward method the best level can only be determined a posteriori (using, for instance, the **Kelley**

* Corresponding author phone: +0041 1 635 3063; fax: +0041 1 635 3025; e-mail: ruedi@ini.phys.ethz.ch.

[†] Institute for Neuroinformatics.

[‡] Novartis Institutes for Biomedical Research.

measure^{15,21}). We will show that sequential superparamagnetic clustering clearly outperforms Ward's method.

II. MATERIALS AND METHOD

A. Similarity Measure. Chemical fingerprint descriptors lead to a representation of chemical substances by high-dimensional, binary or integer-valued, feature vectors. For instance, ISIS keys provide information about the presence or absence of 166 predefined structural features. Other fingerprint types, such as ISIS count and Similog keys, also include information on the abundance of these features. Nontrivial questions include assessing to what extent the chosen fingerprints are able to represent the structure and chemical properties. As a measure for the similarity of two fingerprint vectors, the Euclidian distance seems natural and is, in fact, widely used. The Euclidian distance measure, however, has a few shortcomings that, in connection with particular fingerprints, can lead to a complete failure of any clustering algorithm: Assume that from a set of substances described by binary ISIS keys, two chemical structures differ by only two structural features. This means that the associated binary fingerprint vectors differ in exactly two components. Two extreme cases can be imagined: (a) The two structures each have exactly one distinct structural feature (implying that only one component of each vector is nonzero and the places of these nonzero entries are distinct). In this case, the two substances are structurally distinct, and their Euclidian separation is $\sqrt{2}$. (b) Both structures exhibit all key features but one, where the respective exceptions occur for different features. In this case, both substances share many structural features; the Euclidian separation, however, still evaluates to $\sqrt{2}$: A good distance measure, however, should be susceptible to the number of shared structural features, i.e., to the number of nonzero vector components. For binary vectors, the **binary Tanimoto coefficient**¹ takes this into account. To also account for integer-valued vectors, we propose the following modification: For two integer-valued vectors a and b , our similarity measure is

$$T(a, b) = 100 \left(1 - \frac{\sum_i \min(a_i, b_i)}{\sum_i (a_i + b_i) - \sum_i \min(a_i, b_i)} \right) \quad (1)$$

This modification is essential in the presence of nonbinary substance representations.

B. The Superparamagnetic Clustering Algorithm. The basic idea of superparamagnetic clustering is to consider clustering as a self-organized process acting within an inhomogeneous Potts spin system.¹⁶ A Potts spin system is a set of sites that can interact with each other via Potts spins. The sites are given by the data set to be clustered, and their arrangement is determined by the similarity measure but otherwise not specified. We assume that the interaction strength decreases with increasing dissimilarity distance between two sites. The interaction among sites is as follows: Spins that belong to strongly coupled sites, i.e., sites separated by a small distance, tend to, or have a high chance of, being aligned. Thus, clusters of aligned spins may emerge during the system evolution, reflecting classes of

similar points. A parameter T , the temperature, controls the desired resolution of the clustering. It provides us the possibility to choose between different levels in a cluster hierarchy.

Potts spin systems are described within the framework of statistical mechanics. There the system is embedded in a heat reservoir, thus the formalism of canonical ensembles can be applied in order to determine a system's possible states and their respective probabilities. The temperature T is a control parameter, determining the average energy per degree of freedom of the system. Aligned spins minimize the energy and, thus, naturally reflect small temperatures. Accordingly, for small temperatures, the system is in the ferromagnetic phase, where all spins are likely to be aligned. Upon a slow increase of the temperature, the tendency of spins to be aligned decreases, which reflects the higher energy contained in the system. This, however, is not achieved in a smooth manner. Instead, transition points occur, where the general state of the system (the phase) changes abruptly. For homogeneous systems, where all spin-spin couplings are equal, a single transition from the ordered ferromagnetic to the disordered paramagnetic phase is observed. In the paramagnetic phase, single spins behave almost independently, and spin alignment becomes a random process. For inhomogeneous systems, the picture can be much more complicated: Between the ferromagnetic and paramagnetic phases, a superparamagnetic phase may occur, where local clusters of aligned spins emerge. These clusters correspond to regions with strong spin coupling. Upon a further increase in the temperature, local clusters may break up into smaller clusters, in cascades of transitions.

The implementation of this approach proceeds along the following lines: To each data vector x_i , a Potts-spin variable s_i is assigned, which may take values $s_i \in \{1, \dots, q\}$. The number of Potts states q is largely arbitrary and in no way connected with the vector length of x_i . The choice of q does not affect the occurrence of clusters. It mostly affects the *extent* of the phases of stable clusters over the temperature range. For a large q , the superparamagnetic phase shrinks and the transitions sharpen,¹⁶ which may be of advantage for quick temperature sweeps. However, the computational costs for one temperature step increase. For the results reported in this contribution, a number of $q = 10$ Potts-spin states was found to yield optimal results.

Each site can interact via its spin with either (a) the k nearest neighbor sites (knN -coupling) or (b) all sites within a distance r (ball-coupling). The coupling strength for an interaction is given by

$$J_{ij} = J_{ji} = \frac{1}{\hat{K}} \exp\left(-\frac{\|x_i - x_j\|^2}{2a^2}\right) \quad (2)$$

where \hat{K} is the average number of couplings per site and $\|x_i - x_j\|$ denotes the similarity or distance between the two coupled sites. a is a local length scale, for which we take the average distance of coupled sites. Alternatively, for the coupling strength, similarly decreasing functions of the distance could be used. Note that the similarity measures between the feature vectors x are transferred into their respective coupling strengths. On the Potts spin q -space, this coupling strength then mediates the formation of the clusters.

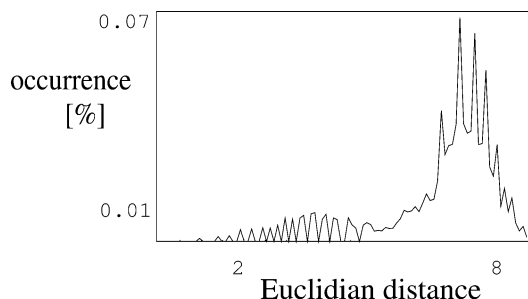


Figure 1. Euclidian distances distribution of test set (ISIS binary keys).

A particular realization of Potts-spin variables s_i is called a system *configuration* S . The probability of finding a configuration S is given by the canonical probability

$$p(S) = Z^{-1} e^{-H(S)/T} \quad (3)$$

where the partition function Z serves as a normalization factor. The Hamiltonian H is calculated according to

$$H(S) = \sum_{(ij)} J_{ij} (1 - \delta_{s_i s_j}) \quad (4)$$

where the sum runs over all pairs of coupled sites. H measures the energy of a configuration S and, therefore, determines its probability, since low energies are preferred. The attribution of points to a cluster is done via the pair-correlation criterion: Two points x_i, x_j belong to the same cluster, if the pair correlation G_{ij} exceeds a threshold Θ

$$G_{ij} = \langle \delta_{s_i s_j} \rangle = \sum_S p(S) \delta_{s_i s_j}(S) > \Theta \quad (5)$$

where the sum is over all possible configurations S . For $q = 10$, a convenient threshold is $\Theta = 0.2$. Due to the immense number of possible configurations (q^N , where N is the number of data points), G_{ij} cannot be calculated directly from eq 5. Instead, a Monte Carlo approach is used, according which the condition on G_{ij} is evaluated as

$$G_{ij} = \frac{1}{M} \sum_{t=1}^M \delta_{s_i s_j}(S(t)) > \Theta \quad (6)$$

where $S(t)$ is the configuration at step t and M is the number of Monte Carlo steps. We used the standard^{16,17} **Swendsen-Wang algorithm**²² to calculate (6).

From the application point of view, the dependence of cluster formation upon a variation of the internal parameters, and their optimal values, is of interest. We find a largely insensitive dependence of the results on q , and, similarly, on the threshold Θ , as long as the condition $1/q < \Theta < (1-2/q)$ is satisfied.¹⁶ This robustness is due to the fact that the pair correlation distribution typically displays two peaks, corresponding to similar and to dissimilar points, respectively. In contrast, the similarity measure distribution itself usually shows many peaks, see Figures 1 and 2. Already at $M = 220$ Monte Carlo steps, stable results are obtained. The optimal choice of the coupling range k is more delicate. We have found that ball-coupling tends to yield slightly better results than knN -coupling, provided a good estimation for the ball radius r can be found. The results presented are nevertheless based on knN -coupling, since the sequential

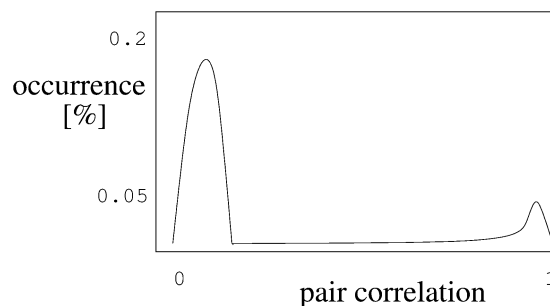


Figure 2. Corresponding distribution of pair correlations G_{ij} .

clustering introduced in the next section is able to cope with this aspect, by providing stability of results over a whole region $k \in [5, \dots, 15]$.

C. Sequential Clustering. The challenges in superparamagnetic clustering are to properly deal with inhomogeneities in shape, density, and size of different clusters. Some clusters may be stable over a long temperature range, while others are not. The reason for this is that different densities are dominant in different temperature ranges: sparse clusters generally only exist at small temperatures, whereas dense (sub-)clusters sometimes only emerge after cluster breakups at higher temperatures. This feature can be exploited, by considering that the most natural clustering level is not given by a global clustering resolution. In fact, there are regions where a higher resolution (higher T) and regions where a smaller resolution (smaller T) is advantageous, to identify the appropriate subclusters. This translates to a difficulty of finding the “right” coupling: Too weak couplings, i.e., too few nearest neighbors, may prevent cluster emergence in sparse regions, whereas too strong couplings, i.e., many nearest neighbors, hold the danger that superclusters, that are a conglomeration of many classes, immediately break up into small pieces. To overcome such difficulties, we introduced a second level clustering procedure. Basically, the most stable clusters are identified and extracted from the data set. The extracted sets and the residual set are reclustered separately. As a result, we obtain sequences of sets of increasing homogeneity, and cluster detection becomes easier.

This motivates the following automated implementation: We limit ourselves to the extraction of the most stable cluster, i.e., the one that is stable over the broadest temperature range. For this purpose, the algorithm is run over a predefined temperature range which is determined in the first clustering step. For a data set S , the following steps are performed:

- (1) apply superparamagnetic clustering to S ,
- (2) while (stop criterion not fulfilled for S):
extract the most stable cluster S_1 and the residual set $S_2 = S \setminus S_1$.
go back to 1 with $S = S_i$ ($i = 1, 2$);
- (3) identify classes:
case 1:
if $S = S_1$ in last loop : identify S as a natural class;
case 2:
if $S = S_2$ in last loop or $S =$ initial set: identify S as a set of unclustered points.

As a consequence of the repeated splitting of sets S into sets S_1, S_2 , we obtain a binary tree structure, where a stop criterion defines the end of each branch. We stop the procedure, if the biggest cluster after the ferromagnetic phase is smaller than threshold Θ_1 . This is reasonable because for

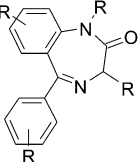
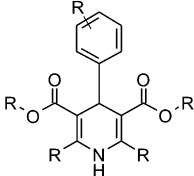
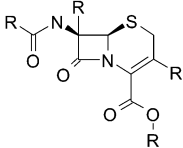
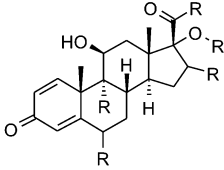
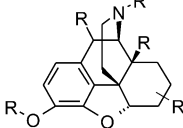
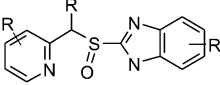
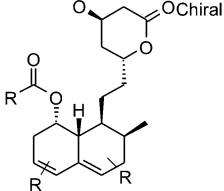
Class (MDDR activity_index)	Size	Structural scaffold	MDDR Numbers (EXTREG)		
Benzodiazepine (06210)	21		091031	091379	127118
			091122	091387	159474
			091323	091391	297743
			091339	091392	297744
			091375	091394	297745
			091376	091396	297746
			091377	091398	297747
Calcium Channel Blocker (31500)	24		090008	122198	162312
			090251	130598	163877
			090489	131335	174260
			091152	133799	226911
			108100	142671	271795
			115055	145448	291201
			115056	148863	306136
Cephalosporin (64200)	22		090261	112175	166459
			090370	127911	176375
			091106	129370	176379
			091132	138250	185023
			100078	141720	188550
			101081	144202	189332
			106459	153268	291858
Corticosteroid (02400)	18		090741	129983	166657
			091057	136805	167983
			091480	154424	168221
			107010	159710	170014
			111914	162123	188135
			117492	164968	188136
Opioid (01100)	20		091305	150526	226494
			091355	173614	263567
			091357	179330	263568
			144748	185269	278204
			145961	217944	279578
			145962	219842	283383
			147727	222693	
Proton Pump [H ⁺ /K ⁺ -ATPase] Inhibitor (54112)	21		090859	144868	147835
			123065	146030	147836
			136362	147725	147838
			139540	147831	147840
			141604	147832	149736
			143225	147833	149737
			144867	147834	152767
Statin [HMG-CoA Reductase (beta) Inhibitor] (52500)	27		138237	141090	141338
			138396	141091	141339
			139538	141092	142208
			140568	141093	149163
			141085	141332	151790
			141086	141333	151791
			141087	141334	152989
			141088	141335	155916
			141089	141336	158664

Figure 3. Classes, class sizes, defining structures, and MDDR numbers of the clustered substances.

sets without cluster structures, the paramagnetic phase immediately follows the ferromagnetic phase, and stable superparamagnetic clusters do not occur. Note, however, that the biggest cluster need not be the most stable one.

Care has to be taken when only one intrinsic cluster is present. In this case, we will have a direct transition from the ferromagnetic to the paramagnetic phase. The length of the ferromagnetic phase indicates the compactness of the cluster. For sets without a clear cluster structure, such as homogeneous residuals, the ferromagnetic phase occurs as well, but it is much shorter. A threshold Θ_2 for the transition temperature T_{Tr} from the ferromagnetic to the paramagnetic phase helps to determine whether we are dealing with an intrinsic cluster (exceptionally) or with a set of unclustered points (normally). Usually, the picture is obvious, since useful

values for Θ_2 can be chosen from a large interval. Taking this into account, case 2 is refined as follows:

case 2a: if $T_{Tr} > \Theta_2$: S is a natural class.

case 2b: else: S is a set of unclustered points.

Proceeding in this way, we found that over the range $k \in [5, \dots, 15]$, the results of our clustering approach did not change.

III. DATA SETS

The data set used for this study was taken from the **MDDR database**,²³ where 153 molecules from seven classes were selected. Each class shares a common substructure (see Figure 3) and a common, well-known mechanism of pharmacological action, described by the MDDR database activity index. From the MDDR database, the structures were

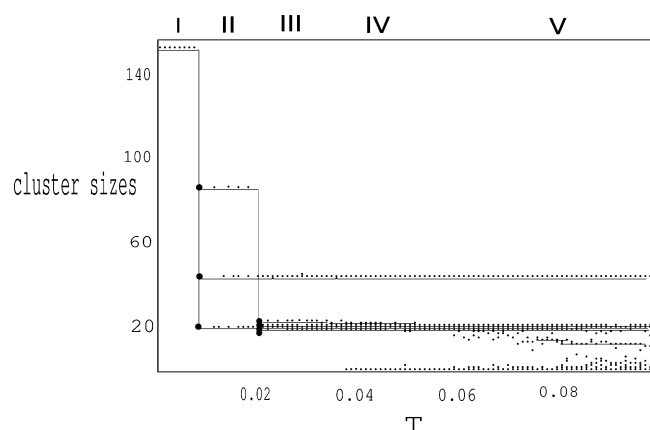


Figure 4. Output of the sequential superparamagnetic clustering, primary step. To guide the eye, lines are drawn connecting constant size clusters. Because of the four classes of almost equal size, the display is partially hampered. Temperature regions: I, one-cluster region (ferromagnetic phase); II, three-clusters region, III, most stable clusters region, showing correct number of clusters and cluster sizes; IV, region of relative stability of clusters; V, break-up region.

exported as a SD-file. Prior to the calculation of the descriptors, they were preprocessed in order to remove counterions and other disconnected low molecular weight fragments and to neutralize deprotonated acidic and protonated basic functional groups (using the PipelinePilot software²⁴). The preprocessed data set was then characterized by three descriptors: (1) the binary ISIS key descriptor, which is based on a curated dictionary of structural fragments,^{18,19} (2) the ISIS count descriptor, where the number of fragment occurrences is part of the descriptor, and (3) the pharmacophore Similog keys. Details on how these descriptors and the Similog keys are calculated can be found in ref 20.

IV. RESULTS AND DISCUSSION

To show the stark contrast to concurrent methods, we compared our results with those obtained by the **Ward approach**. We assumed the clusters induced by the chemical classes the data were taken from, as the ideal. To measure the difference between the clustering results and the ideal result, we computed the **Jaccard statistics**¹⁵ for the clusterings obtained with the different descriptors, relative to the ideal clustering. When using Ward's method, the optimal number of clusters was determined by applying the **Kelley measure**.^{15,21} The Kelley measure is an accepted cluster level selection criterion that balances between the number of clusters and their inherent densities. In the case of superparamagnetic clustering, the optimal number of clusters and the size of clusters emerge naturally from the approach.

Figure 4 shows a typical cluster size diagram obtained by applying superparamagnetic clustering to the ISIS binary keys model set. One can identify a short ferromagnetic phase, followed by a superparamagnetic phase. Clearly, several clusters that correspond to the chemical classes can be identified. As the temperature increases, they break up into subclusters or singletons. The temperature at which a cluster decays can be seen to depend on its consistency. Denser clusters break up later. One big cluster of 45 elements stretches across a broad temperature range. This supercluster, containing the classes *Statin* and *Corticosteroid*, is extracted

Table 1. Performance of Ward's Method, for Different Keys and Similarity Measures

keys	similarity measure	optimal no. of clusters (Kelley)	Jaccard coeff	best Jaccard coeff
ISIS, binary	Euclidian	19	0.495	0.769 for 6 clusters
ISIS, binary	Tanimoto	14	0.654	0.769 for 6 clusters
Similog	Euclidian	53	0.407	0.407 for 53 clusters
Similog	Tanimoto	15	0.534	0.540 for 16 clusters
ISIS count	Euclidian	15	0.674	0.827 for 13 clusters
ISIS count	Tanimoto	16	0.613	0.736 for 12 clusters

Table 2. Performance of Sequential Superparamagnetic Clustering

keys	similarity measure	no. of clusters (including singletons)	Jaccard coeff
ISIS, binary	Euclidian	9	0.956
ISIS, binary	Tanimoto	9	0.956
Similog	Euclidian	24	0.420
Similog	Tanimoto	9	0.947
ISIS count	Euclidian	9	0.941
ISIS count	Tanimoto	8	0.986

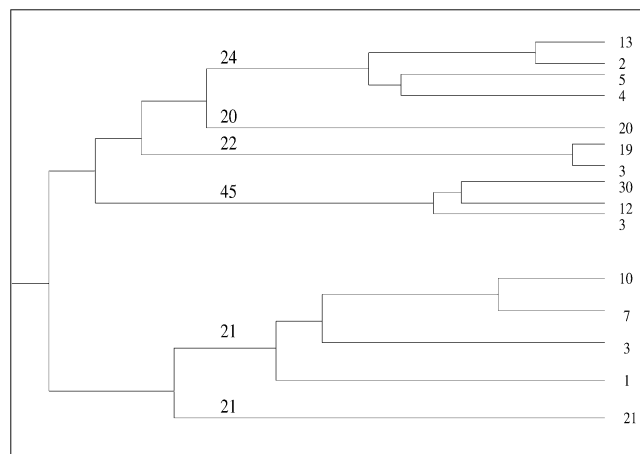


Figure 5. Dendrogram for Ward's method, using Tanimoto coefficients for ISIS binary keys (implementation by the authors). The number of clusters increases from left to right. The inserted numbers indicate the cluster sizes. The clustering works well only at the six-classes clustering level.

by sequential clustering and separately analyzed, to find the above-mentioned two classes.

Table 2 vs Table 1 contrasts the results from sequential superparamagnetic clustering against those achieved by Ward's method. For ISIS count and Similog keys, our extension (1) of the Tanimoto coefficient was used. The results obtained by using the Kelley measure are shown together with the best Jaccard coefficient achieved. The higher this coefficient, the better the clustering. The comparison emphasizes that Ward clustering with the Kelley measure tends to favor more, but smaller, clusters and does not necessarily reflect the best choice of number of clusters. As a matter of fact, for the considered data, Ward's method struggles with inhomogeneities in the data sets. In the dendrogram Figure 5, we see that, due to its density, the big cluster containing *Statin* and *Corticosteroid* breaks up late (and not even properly). Some other classes, correctly indicated at the 6-cluster stage, break up earlier. Therefore, a simultaneous occurrence and detection of all chemical classes of the model set seems intrinsically impossible.

The results worked out by sequential superparamagnetic clustering are substantially better. The original chemical

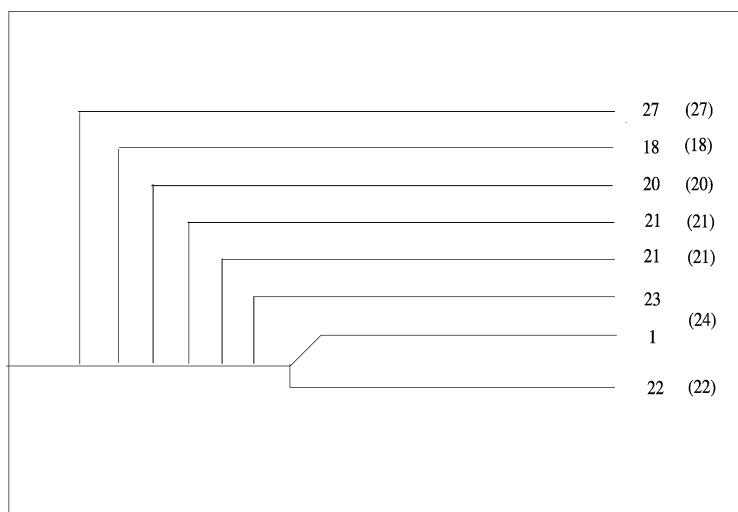


Figure 6. Dendrogram for sequential superparamagnetic clustering, using Tanimoto coefficients for ISIS count keys. Inserted numbers indicate the obtained cluster sizes. In brackets, the correct sizes are added.

classes are recognized almost perfectly. Exceptions range from a singleton in the case of ISIS count with Tanimoto (see Figure 6), up to at most three compounds for ISIS count using the Euclidean similarity measure. Only when using the Euclidean distance measure with Similog keys, satisfactory results could not be achieved. Since the Ward clustering results are similarly bad in this case, we think that this corroborates the claim that the **Euclidian distance should not be used as a similarity measure for the very high-dimensional Similog keys**. The overall-comparison demonstrates the superiority of the sequential paramagnetic clustering approach over Ward clustering.

For applications, a comparison of the computational costs raised by the two approaches is of interest. The core part of the superparamagnetic clustering consists of a Monte Carlo simulation. For this part, a time complexity of $O(N^2)$ is required, since the number of Monte Carlo steps can be held fixed. The number of natural classes within a data set, and therefore the number of clustering steps, is usually bounded. We therefore expect the time complexity to be characteristic for the whole procedure. Ward's method requires $O(N^2)$ time complexity as well.

V. CONCLUSIONS

Our results indicate that sequential superparamagnetic clustering easily outperforms traditional clustering methods. Our approach is particularly well forged for data sets with density differences between chemical classes, where some clusters only emerge in a subset of adequate homogeneity. Sequential clustering is effective, since it is not based on the simultaneous detection of all clusters. The number of classes emerges naturally and does not require any biased or a posteriori information. Sequential clustering implies that the most natural clustering level, i.e., the level where the most natural chemical classes occur, is not given by a globally fixed resolution. Instead, the optimal resolution depends on local densities. This principle is the key feature exploited by the sequential clustering procedure.

We found that in some cases (e.g. Similog keys), the choice of an appropriate similarity measure is crucial. The Euclidian distance measure, in particular, may introduce serious shortcomings. It is more advantageous to use our

novel similarity measure that can be considered an extension of the binary Tanimoto coefficient.

To conclude, we estimate that sequential superparamagnetic clustering will be of particular importance to chemical applications such as combinatorial library design and analysis of HTS data hit lists. We also anticipate a wide field of applications in other topical, technical and scientific, fields. These include, e.g., multisensor clustering and visual and auditory scene analysis. The results obtained in our contribution raise the hope that sequential superparamagnetic clustering will contribute to remove boundaries that currently obstruct progress in these fields.

ACKNOWLEDGMENT

The work was partially supported by the Swiss KTI research project *Information-based Approaches in Drug Design*.

APPENDIX

The *Kelley measure* for a cluster level i with k_i clusters is defined as

$$\text{Kelley}_i = (n - 2) \left(\frac{d_{wi} - \min(d_w)}{\max(d_w) - \min(d_w)} \right) + 1 + k_i \quad (7)$$

where d_{wi} is the mean of distances between points in the same cluster at level i , and $\min(d_w)$ and $\max(d_w)$ are the minimum and maximum of this value across all cluster levels (in Ward's clustering approach, the cluster levels are refined from N singleton to one all-encompassing cluster, where N equals the number of the data points in the sample).

The Jaccard coefficient measures how close a clustering C_1 is to an a priori known clustering C_2 (in our case, the known chemical classes). It is evaluated according to

$$\text{Jaccard}(C_1, C_2) = \frac{a}{a + b + c} \quad (8)$$

where a is the number of pairs of points that are clustered in both clusterings, b is the number of pairs that are clustered together in the first clustering but not the second, and c is

the number of pairs that are clustered together in the second clustering but not in the first.

REFERENCES AND NOTES

- (1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd.: Letchworth, 1987.
- (2) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Similarity*; Wiley: New York, 2000.
- (3) Brown, R. D.; Martin, Y. C. Use of structure activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (4) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (5) Herpin, T. F.; Van Kirk, K. G.; Salvino, J. M.; Yu, S. T.; Labaudiniere, R. F. Synthesis of a 10 000 member 1,5-benzodiazepine-2-one library by the direct sorting method. *J. Comb. Chem.* **2000**, *2*, 513–521.
- (6) Shemetsukis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided. Mol. Des.* **1995**, *9*, 407–416.
- (7) Menard, P. B.; Lewis, R. A.; Mason, J. S. Rational set design and compound selection: Cascaded clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497–505.
- (8) Engels, M. F. M.; Thielmans, T.; Verbinden, D.; Tollenaere, J. P.; Verbeeck, R. CerBeruS: A system supporting the sequential screening process. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 241–245.
- (9) Nicolaou, C. A.; Tamura, S. Y.; Kelly, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069–1079.
- (10) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.
- (11) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
- (12) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034.
- (13) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K., Boyd, D. B., Eds.; VCH Publishers: New York, 2002; Vol. 18, pp 1–40.
- (14) Forgy, E. Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics* **1965**, *21*, 768–780.
- (15) Wild, D. J.; Blankley, C. J. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155–162.
- (16) Domany, E. Superparamagnetic clustering of data – The definitive solution of an ill-posed problem. *Physica A* **1999**, *263*, 158–169.
- (17) Blatt, M.; Wiseman, S.; Domany, E. Superparamagnetic clustering of data. *Phys. Rev. Lett.* **1996**, *76*, 3251–3254.
- (18) The ISIS public keys are already used in MACCS-II and are also known as MACCS keys. ISIS/Base and MACCS are both products of MDL Information Systems, Inc., San Leandro, CA, <http://www.mdli.com>.
- (19) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- (20) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *2*, 391–405.
- (21) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR derived protein structures into conformationally related subfamilies. *Protein Eng.* **1996**, *9*, 1063–1065.
- (22) Wang, S.; Swendsen, R. H. Cluster Monte Carlo algorithms. *Physica A* **1990**, *167*, 565–579.
- (23) MDL Drug Data Report Version 2001.1, MDL ISIS/HOST software. MDL Information Systems Inc., San Leandro, CA, <http://www.mdli.com>.
- (24) SciTegic Inc., 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123 U.S.A.

CI049905C