# Effect of Data Standardization on Chemical Clustering and Similarity Searching

Chia-Wei Chu, John D. Holliday, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, U.K.

Received July 7, 2008

Standardization is used to ensure that the variables in a similarity calculation make an equal contribution to the computed similarity value. This paper compares the use of seven different methods that have been suggested previously for the standardization of integer-valued or real-valued data, comparing the results with unstandardized data. Sets of structures from the MDL Drug Data Report and IDAlert databases and represented by Pipeline Pilot physicochemical parameters, molecular holograms and Molconn-Z parameters are clustered using the k-means and Ward's clustering methods. The resulting classifications are evaluated in terms of the degree of clustering of active compounds selected from eleven different biological activity classes, with these classes also being used in similarity searches. It is shown that there is no consistent pattern when the various standardization methods are ranked in order of decreasing effectiveness and that there is no obvious performance benefit (when compared to unstandardized data) that is likely to be obtained from the use of any particular standardization method.

## INTRODUCTION

The measurement of similarity plays an important role in several areas of chemoinformatics,[1,2] with a similarity measure possessing three major components: the representation that is used to characterize the molecules that are being compared; the standardization method that is applied to the various components of the chosen representation; and the similarity coefficient that is used to provide a quantitative measure of the degree of structural relatedness between a pair of representations.[3,4] There have been several discussions of the contributions that different types of representation and of similarity coefficient make to the measurement of molecular similarity;[5−7] here, we focus on the different ways in which a given representation can be standardized.

Standardization is the application of a mathematical transformation to each of the individual components of a multivariate representation so that all of the variables make a comparable contribution to the measurement of similarity. For example, a common structural representation is a set of computed physicochemical parameters such as logP, molecular weight, number of rotatable bonds, numbers of donors and acceptors, polar surface area, etc. Each of these has its own range of values, a range that may be very different from that of some or all of the other variables comprising a representation, and there is thus the possibility that one, or a small number, of the variables can dominate the calculation of a similarity coefficient value. Standardization ensures that no such bias occurs (although, as Kettenring notes[8] this lack of bias may not be desirable, even though removing it may appear to be a reasonable thing to do).

Standardization is a common operation in multivariate statistics and normally involves the Z standardization method (often called autoscaling), which uses the mean and the standard deviation of each variable under study.[9] While standardization has been found helpful in applications as diverse as the estimation of software costs,[10] the comparison of pattern recognition methods based on Minkowski metrics,[11] and the analysis of educational-test results[12] and of public-library performance,[13] there have been few comparative studies. The most important contribution is that of Milligan and Cooper,[9] who reported a detailed comparison of standardization methods when applied to the cluster analysis of artificial data sets, and we have used the methods that they studied in the work reported here. Their study suggested that methods involving the division of a variable's value by the range of that variable—often called range scaling—were more effective than the other methods tested (including autoscaling) at recovering the underlying cluster structure. While the Milligan-Cooper study is extensively cited in the cluster-analysis literature, the only subsequent detailed comparison is the work of Gnanadesikan et al.[14] They found that both autoscaling and range scaling were less effective than more complex schemes based on estimates of intracluster and intercluster variability, but their experiments were on a very small scale, e.g., five multivariate normal clusters of 75 artificial data points. Previous comparative studies in chemoinformatics have focused on the standardization of 2D fingerprints[15] and of 3D molecular fields.[16] Here, we consider the effectiveness of different standardization methods on the effect of chemical clustering and similarity searching using three further types of structure representation: physicochemical properties, molecular holograms, and topological indices.

## METHODS

**Standardization Methods.** The Milligan-Cooper study[9] involved seven different (but often closely related) standardization methods that had been reported by previous workers, and we have used these for the current work. In what follows,

* Corrresponding author phone: +44-114-2222633; e-mail: p.willett@sheffield.ac.uk.

10.1021/ci800224h CCC: $40.75 
Published on Web 12/23/2008

**Table 1.** Eleven Activity Classes Used for the Evaluation of the Standardization Methods

| activity class | number of MDDR actives | number of IDAlert actives |
|---|---|---|
| 5HT3 antagonists | 84 | 99 |
| 5HT1A agonists | 77 | 61 |
| 5HT reuptake inhibitors | 39 | 41 |
| D2 antagonists | 45 | 20 |
| renin inhibitors | 115 | 123 |
| angiotensin II AT1 antagonists | 100 | 12 |
| thrombin inhibitors | 65 | 76 |
| substance P antagonists | 122 | 66 |
| HIV-1 protease inhibitors | 79 | 32 |
| cyclooxygenase inhibitors | 75 | 87 |
| protein kinase C inhibitors | 40 | 51 |

$X$ denotes the observation value of the variable, $\mu$ and $\sigma$ denote the mean and the standard deviation of the observations for the variable, and *MAX(X)* and *MIN(X)* denote its maximum and minimum values, respectively.

The most common and traditional standardization method is the Z-Score, where the variable is transformed to zero mean and unit variance

$$S_1 = \frac{X - \mu}{\sigma}$$

The second form of standardization is similar to the Z-Score

$$S_2 = \frac{X}{\sigma}$$

The third method uses *MAX(X)*, rather than $\sigma$, as the divisor, i.e.

$$S_3 = \frac{X}{MAX(X)}$$

The next two methods involve the range of values for the variable

$$S_4 = \frac{X}{MAX(X) - MIN(X)}$$

$$S_5 = \frac{X - MIN(X)}{MAX(X) - MIN(X)}$$

The sum of the observations for a variable provides yet another divisor, i.e.

$$S_6 = \frac{X}{\sum X}$$

the final method uses the rank (or the average rank if there are tied values), rather than the value, of the variable, i.e.

$$S_7 = Rank(X)$$

We use $S_0$ to denote the original, unstandardized data.

**Data Sets.** Our experiments used two data sets: 10,191 molecules from the MDL Drug Data Report (MDDR) database and 11,607 molecules from the IDAlert database. Molecules were noted as being active or inactive (more probably, not tested) in eleven activity classes that had been studied previously by Hert et al.[17] The activity classes and numbers of molecules in the MDDR and IDAlert data sets are summarized in Table 1. The two data sets are very different in nature, having just a single molecule in common.

The molecules in the two data sets were represented in three different ways: 12 physicochemical properties (e.g., AlogP, logD, molecular weight, volume, and solubility) generated using the Pipeline Pilot software;[18] 523 topological indices (e.g., molecular connectivity, kappa shape, and electrotopological state indices) generated using the Molconn-Z software;[19] and 997-element molecular holograms generated using the Unity software with the default parameter settings.[20] The first two representations involve continuous data, while the last involves integer fragment occurrence data. Each of the three representations was standardized using the eight methods $S_0$-$S_7$ as defined above, giving a total of 24 different standardized representations of each of the two data sets. The various representations of the two data sets have been used in two important applications in chemoinformatics: similarity searching and clustering.[3,5,21]

**Table 2.** Mean Percentage of Actives Retrieved in the Top-500 Compounds in the Ranked Database[a]

| | Pipeline Pilot | Molconn-Z | Holograms |
|---|---|---|---|
| | (a) MDDR Database | | |
| $S_0$ | 18.8 | 16.0 | 23.0 |
| $S_1$ | 22.5 | 27.7 | 20.9 |
| $S_2$ | 22.5 | 27.7 | 20.9 |
| $S_3$ | 14.9 | 9.8 | 20.3 |
| $S_4$ | 22.0 | 25.3 | 20.3 |
| $S_5$ | 22.0 | 25.3 | 20.3 |
| $S_6$ | 19.3 | 21.6 | 20.4 |
| $S_7$ | 20.4 | 29.8 | 26.6 |
| | (b) IDAlert Database | | |
| $S_0$ | 19.2 | 15.7 | 23.6 |
| $S_1$ | 21.1 | 26.2 | 18.3 |
| $S_2$ | 21.1 | 26.2 | 18.3 |
| $S_3$ | 14.6 | 9.7 | 18.0 |
| $S_4$ | 21.0 | 23.8 | 18.0 |
| $S_5$ | 21.0 | 23.8 | 18.0 |
| $S_6$ | 20.7 | 20.3 | 16.8 |
| $S_7$ | 21.4 | 30.9 | 23.9 |

[a] The figures are mean values averaged over eleven activity classes and over ten reference structures for each activity class.

**Similarity Searching and Clustering Experiments.** A similarity search was conducted by selecting a reference structure, i.e., a molecule with a specific biological activity, and then ranking all the molecules in the database in order of increasing Euclidean distance from the reference structure.

The clustering experiments used two of the most widely used methods of cluster analysis, specifically the K-Means and Ward's methods as implemented in Digital Chemistry's clustering toolkit.[22] Classifications were generated that contained 25, 50, and 100 clusters.

**Evaluation of Performance.** There is an extensive literature on the use of bioactivity data for the evaluation of similarity methods (e.g., refs 23 and 24). Here, we have used a simple approach based on the numbers of active molecules identified at the top of a ranked database. Specifically, a cutoff was applied to retrieve the top 500 compounds in the ranked database resulting from a similarity search, and a note then made of the percentage of the active molecules that occurred within these sets of nearest neighbors. Ten different reference structures were chosen for each activity class using a MaxMin diversity selection routine, and the results then averaged over the ten searches for each activity class. These mean results were then averaged again, over the eleven activity classes, to quantify the effectiveness of a given standardization method.

**Table 3.** MDDR Data Set: The Average (a) Shannon Entropy and (b) Probability of Correct Prediction Using Different Representations, Clustering Methods, and Numbers of Clusters[a]

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| | | | (a) Shannon Entropy | | | |
| 100 | 4.27 | 4.00 | 3.02 | 3.94 | 3.54 | 2.63 |
| 50 | 3.54 | 3.40 | 2.58 | 3.27 | 2.83 | 2.08 |
| 25 | 2.77 | 2.59 | 2.14 | 2.42 | 2.17 | 1.64 |
| | | | (b) Probability of Correct Prediction | | | |
| 100 | 0.84 | 0.83 | 0.70 | 0.77 | 0.75 | 0.68 |
| 50 | 0.64 | 0.64 | 0.60 | 0.61 | 0.67 | 0.62 |
| 25 | 0.55 | 0.54 | 0.54 | 0.56 | 0.58 | 0.55 |

[a] The results in each case are averaged over the eleven activity classes.

**Table 4.** IDAlert Data Set: The Average (a) Shannon Entropy and (b) Probability of Correct Prediction Using Different Representations, Clustering Methods, and Numbers of Clusters[a]

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| | | | (a) Shannon Entropy | | | |
| 100 | 4.62 | 4.71 | 4.30 | 4.29 | 4.05 | 4.18 |
| 50 | 4.02 | 3.93 | 3.77 | 3.69 | 3.11 | 3.57 |
| 25 | 3.24 | 2.97 | 3.17 | 2.99 | 2.37 | 2.78 |
| | | | (b) Probability of Correct Prediction | | | |
| 100 | 0.77 | 0.71 | 0.72 | 0.71 | 0.69 | 0.67 |
| 50 | 0.60 | 0.59 | 0.60 | 0.57 | 0.61 | 0.60 |
| 25 | 0.46 | 0.51 | 0.55 | 0.47 | 0.54 | 0.54 |

[a] The results in each case are averaged over the eleven activity classes.

There has been less discussion of the evaluation of clustering, as against similarity, methods. Here, we have used two, complementary approaches to compare the effectiveness of the various standardization methods in grouping active molecules together: a measure based on the distribution of actives across the clusters (with a good standardization method being one that minimized the spread of actives) and a measure based on the numbers of inactive molecules occurring in clusters that contained active molecules (with a good standardization method being one that minimized the numbers of inactives in such clusters). Each measure was calculated for each activity class, and the results were averaged over the eleven classes to quantify the effectiveness of a given standardization method.

The first, entropy-based approach assumes that the best possible classification is one in which all of the actives for some particular activity class are located in the same cluster; conversely, the worst possible classification is one in which they are distributed equally across the available clusters. The distribution of the actives was quantified using the Shannon Entropy (SE), which is defined[25] as

$$SE = -\sum_i p_i \log_2(p_i)$$

where $p_i$ is the fraction of the total number of active molecules that occur in the $i$-th cluster and where the summation is over all of the clusters. For example, if 4 of the 100 members of an activity class occur in some cluster A then $p_i = 0.04$, yielding a contribution to SE of 0.19. The performance measure is then the calculated entropy, with the results being averaged over all of the eleven activity classes.

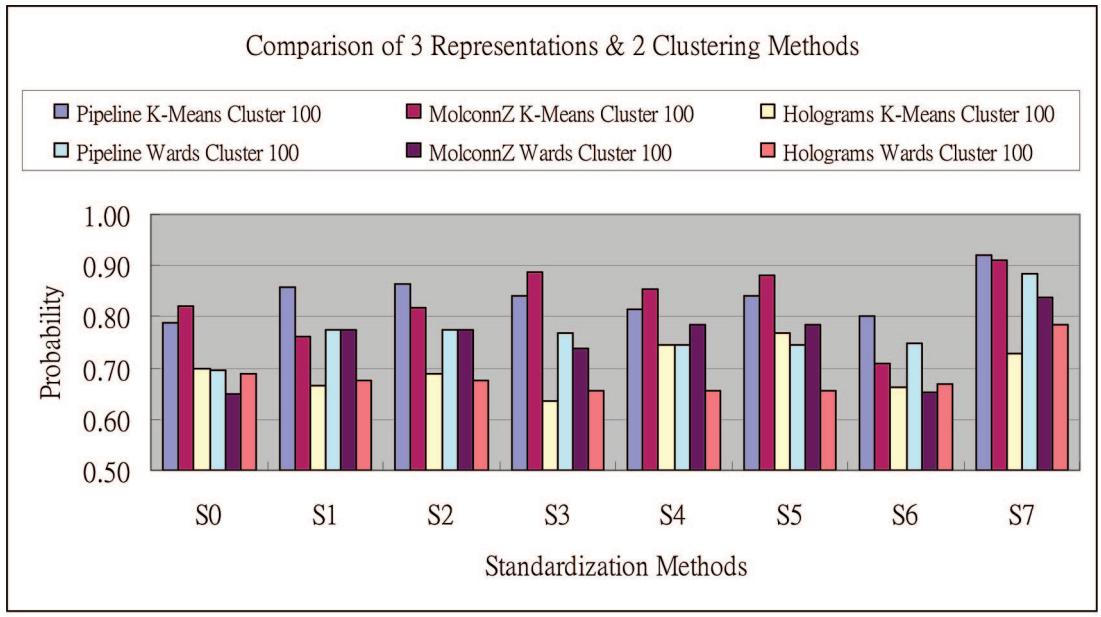
The entropy approach's focus on just the active molecules means that no account is taken of the actives' co-occurrence with inactives: the second approach hence takes account of both the actives and the inactives for some particular activity class. Let an *active cluster* be a cluster that contains at least one molecule from the chosen activity class. Define *P(active)* and *P(inactive)* for a particular cluster as

$$P(active) = \frac{a}{A} \quad \text{and} \quad P(inactive) = \frac{n-a}{N-A}$$

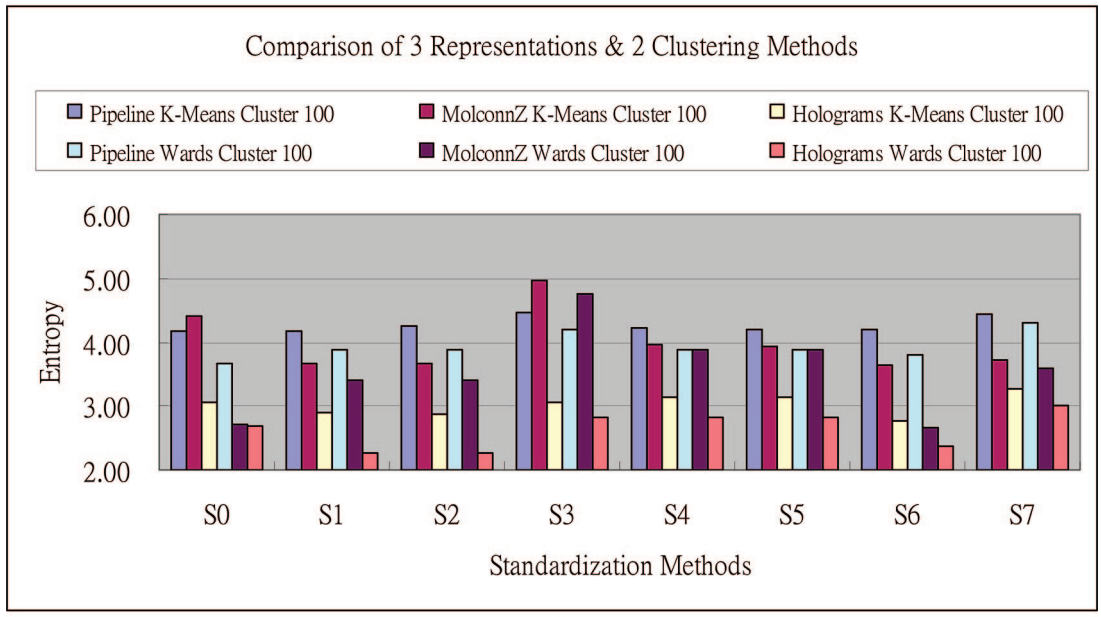
where $N$ is the total number of compounds in the data set, $n$ is the total number of molecules in the current active cluster, $a$ is the number of active molecules in that cluster, and $A$ is the total number of molecules exhibiting the chosen activity. The two values *P(active)* and *P(inactive)* hence describe the proportion of the actives and the proportion of the inactives that are present in the chosen cluster. We would hope that *P(active)* would be greater than *P(inactive)* in the case of an active cluster, i.e., that there is a greater concentration of active molecules present (whereas the converse would imply the presence of some small number of "stray" actives in a cluster composed predominantly of inactives). We then use the number of times when this is in fact the case as a measure of the effectiveness of clustering: the more frequently this happens, the greater the degree of concentration of the actives in the active clusters. For example, assume that $a = 2$ and $n = 10$ for some cluster and that $N = 820$ and $A = 20$ for the data set. Then the probabilities of activity and inactivity are

$$P(active) = \frac{2}{20} = 0.1 \quad \text{and} \quad P(inactive) = \frac{10-2}{820-20} = 0.01$$

with *P(active)* > *P(inactive)*, as would be predicted for an active cluster. The performance measure is then the fraction of active clusters that are indeed predicted to be active for




(a)




(b)

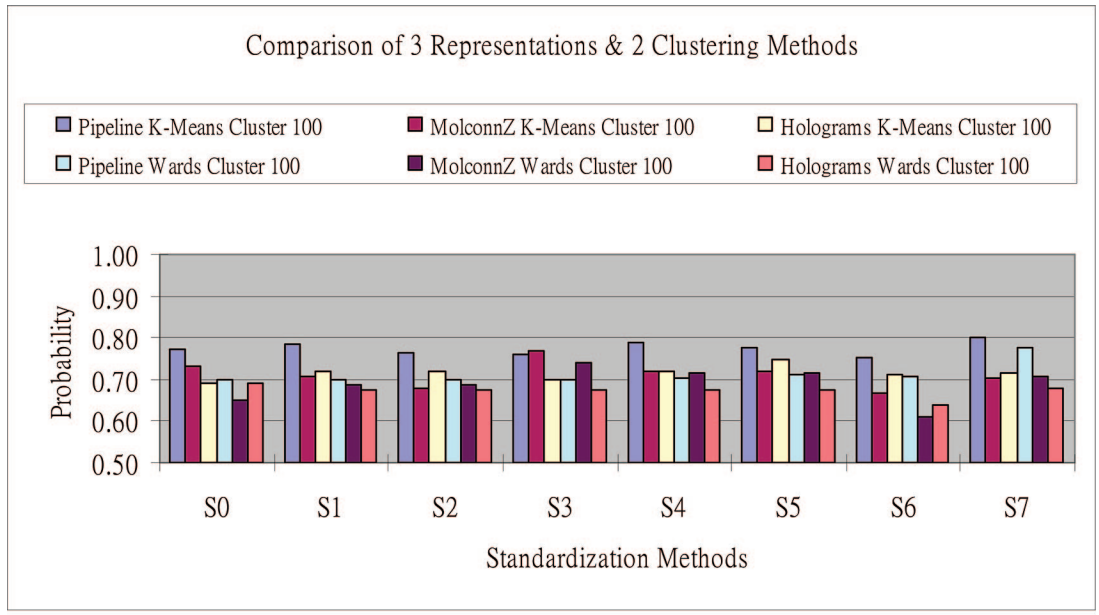
**Figure 1.** Comparison of representations and clustering methods (100 clusters) for the MDDR data set based on (a) the probability of correct prediction and (b) Shannon entropy.

the chosen activity class, with the results being averaged over all of the eleven activity classes.

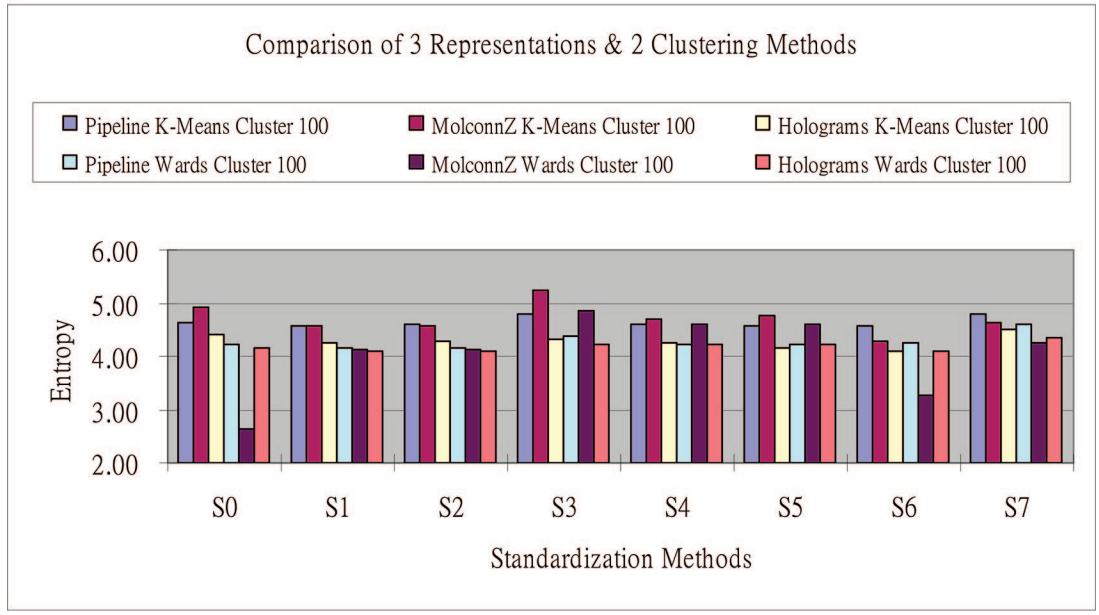
## RESULTS

**Similarity Searching.** The results for the MDDR and IDAlert databases are shown in Table 2. In each case, the results are mean percentage recalls (to one decimal place) when averaged over all of the activity classes.

The initial impression to be gained from these figures is that there is no single method that provides a consistently high level of performance across all of the three different types of representation. The raw data, i.e., $S_0$, is noticeably inferior to all of the other methods (with the exception of $S_3$ as discussed below) for the Pipeline Pilot and Molconn-Z searches; however, it does well with the molecular holograms, where $S_3$ also performs much better than for the other two sets of searches. The best overall level of performance is given by $S_7$, which gives the highest recall for all but the Pipeline Pilot searches of the MDDR database. An analogous pattern of behavior is observed if just the top-100, rather than the top-500, of the search outputs are analyzed (results not included here). Note that $S_3$, $S_4$, and $S_5$ give identical results for the hologram searches: the hologram representation has many zero-valued entries, and the $MIN(X)$ component of $S_4$ and $S_5$ is hence zero, meaning that the three standardized representations are identical. Also, the Pipeline Pilot and Molconn-Z representations contain some very large entries, which adversely affect the performance of $S_3$ for these two representations. The $S_1$ and $S_2$ results are the same because the $\mu$ component in the $S_1$ standardization cancels




(a)




(b)

**Figure 2.** Comparison of representations and clustering methods (100 clusters) for the IDAlert data set based on (a) the probability of correct prediction and (b) Shannon entropy.

**Table 5.** MDDR Kendall $W$ and $\chi^2$ Values Based on (a) Shannon Entropy and (b) Probability of Correct Prediction

| | K-Means | | Ward's | |
|---|---|---|---|---|
| clusters | $W$ | $\chi^2$ | $W$ | $\chi^2$ |
| | | (a) | | |
| 100 | 0.60 | 12.67 | 0.66 | 13.91 |
| 50 | 0.58 | 12.11 | 0.66 | 13.91 |
| 25 | 0.67 | 14.00 | 0.66 | 13.91 |
| | | (b) | | |
| 100 | 0.49 | 10.33 | 0.51 | 10.69 |
| 50 | 0.33 | 6.94 | 0.14 | 2.97 |
| 25 | 0.17 | 3.64 | 0.22 | 4.58 |

**Table 6.** IDAlert Kendall $W$ and $\chi^2$ Values Based on (a) Shannon Entropy and (b) Probability of Correct Prediction

| | K-Means | | Ward's | |
|---|---|---|---|---|
| clusters | $W$ | $\chi^2$ | $W$ | $\chi^2$ |
| | | (a) | | |
| 100 | 0.74 | 15.56 | 0.83 | 17.37 |
| 50 | 0.66 | 13.89 | 0.88 | 18.41 |
| 25 | 0.59 | 12.44 | 0.63 | 13.22 |
| | | (b) | | |
| 100 | 0.35 | 7.33 | 0.57 | 12.07 |
| 50 | 0.25 | 5.33 | 0.12 | 2.62 |
| 25 | 0.28 | 5.89 | 0.51 | 10.80 |

out when the Euclidean distance is computed (and similarly so for the $MIN(X)$ component in the $S_5$ standardization).

Visual inspection of the results in Table 2 suggests that $S_7$ is the most consistently effective of the standardization

methods. We have sought to obtain a more quantitative view of the effectiveness of the methods using Kendall's $W$ test of statistical significance, which is used to evaluate the consistency of $k$ different sets of ranked judgements of the same set of $N$ different objects.[26] Here, we have considered each of the representations (Pipeline Pilot, Molconn-Z and molecular holograms) as a judge ranking the different standardization methods in order of decreasing effectiveness (as measured by the prediction or entropy criteria), i.e., $k = 3$ and $N = 8$. The significance of the computed $W$ values can be tested using the $\chi^2$ distribution since (for $N > 7$)

$$\chi^2 = k(N-1)W$$

with $N$-1 degrees of freedom.

The computed values of $W$ and $\chi^2$ for the MDDR searches are 0.62 and 13.11. The critical value for $\chi^2$ at the $\alpha$=0.01 level of statistical significance is 18.48 for seven degrees of freedom, and it will hence be seen that the computed value is not significant. It is also not significant at the $\alpha$=0.05 level, where the critical value is 14.07. The computed values of $W$ and $\chi^2$ for the IDAlert searches are 0.68 and 14.37; these values are again not significant at the 0.01 level (but they are significant at the 0.05 level of statistical significance). Taken together, these results would suggest that there is no obvious ranking of the eight standardization methods.

**Clustering Experiments.** The overall results for the MDDR and IDAlert data sets are shown in Tables 3 and 4, respectively. In each case, the results are Shannon Entropy or mean probability values (to two decimal places) when averaged over all of the eight standardization methods.

Analysis of the individual experiments that have been averaged in Tables 3 and 4 reveals a high degree of variability, with no obvious "best" method (i.e., maximal value for the prediction runs and minimal values for the entropy runs). For example, Figures 1 and 2 show the 100-cluster runs using prediction and entropy, for the MDDR and IDAlert data sets, respectively. Focusing in Figure 1, $S_7$ does well in the prediction experiments across the three descriptors and two clustering methods but poorly in the entropy experiments, with the notable exception of the two Molconn-Z runs. Also, Pipeline Pilot $S_1$ does better than $S_0$ in Figure 1a for both clustering methods; however, for Molconn-Z, $S_0$ is better for K-means and $S_1$ better for Ward's method. Indeed, there are many occasions where the unstandardized representation is comparable or superior to one of the methods $S_1$-$S_7$. Finally, consider the Molconn-Z $S_6$ results: those in Figure 1a are poor (probability criterion), but those in Figure 1b are among the best (entropy criterion); and this combination gives the best prediction results of all for both of the performance criteria in the 50-cluster and 25-cluster runs (data not shown). A similar lack of consistency is evident in the IDAlert experiments summarized in Figure 2.

Tables 5 and 6 (for MDDR) and 6 (for IDAlert) list the results of a Kendall's $W$ analysis, showing the $W$ and $\chi^2$ values for each number of clusters and each performance criterion, respectively. As noted previously, the critical value for $\chi^2$ at the $\alpha$=0.01 level of statistical significance is 18.48 for seven degrees of freedom, and it will hence be seen that none of the values in Table 5 are significant; they are also not significant at the $\alpha$=0.05 level. There would hence appear to no significant measure of agreement between the three characterizations of the MDDR data set in their ordering of the different methods using the two performance criteria, and it is thus not possible to recommend any particular standardization method as being of general applicability. Broadly comparable conclusions can be drawn from the IDAlert data set, where none of results are again significant at the $\alpha$=0.01 level (although there are three significant results at the 0.05 level).

Taken together, these results would again suggest that there is no obvious ranking of the eight standardization methods.

## CONCLUSIONS

Standardization is widely used in multivariate statistical analyses to ensure that all of the attributes in an object's representation contribute equally when two objects are compared to determine their similarity. In this paper, we have investigated the use of standardization methods for similarity searching and clustering in databases of chemical structures characterized using three common types of molecular representation. Our results suggest that, for chemical data of the sort considered here, there is no consistent performance benefit that is likely to be obtained from the use of any particular standardization method. The choice of method is hence not a critical component of procedures for chemical clustering and searching.

## ACKNOWLEDGMENT

We thank Digital Chemistry Ltd., IDBS Ltd., MDL Information Systems Inc., SciTegic Inc., and Tripos Inc. for data and software support.

## REFERENCES AND NOTES

(1) Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; Wiley-VCH: Weinheim, 2003.
(2) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, 2nd ed.; Kluwer: Dordrecht, 2007.
(3) Willett, P. Similarity Methods in Chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **2009**, *43*, 3–71.
(4) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(6) Sheridan, R. P.; Kearsley, S. K. Why do we Need so many Chemical Similarity Search Methods. *Drug Discovery Today* **2002**, *7*, 903–911.
(7) Glen, R. C.; Adams, S. E. Similarity Metrics and Descriptor Spaces - which Combinations to Choose. *QSAR Comb. Sci.* **2006**, *25*, 1133–1142.
(8) Kettenring, J. R. The Practice of Cluster Analysis. *J. Classif.* **2006**, *23*, 3–30.
(9) Milligan, G. W.; Cooper, M. C. A Study of Standardization of Variables in Cluster Analysis. *J. Classif.* **1988**, *5*, 181–204.
(10) Strike, K.; El-Emam, K.; Madhavji, N. Software Cost Estimation with Incomplete Data. *IEEE Trans. Software Eng.* **2001**, *27*, 890–907.
(11) Doherty, K. A. J.; Adams, R. G.; Davey, N. In *Non-Euclidean norms and data normalisation*, ESANN′ 2004 - European Symposium on Artificial Neural Networks; Bruges (Belgium), 2004; Bruges (Belgium), 2004; pp 181−186.
(12) Dorans, N. J.; Kulick, E. Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *J. Educ. Meas.* **1986**, *23*, 355–368.
(13) Creaser, C. Comparing Performance of Service Points in Public Libraries. *Performance Meas. Metrics* **2001**, *2*, 109–135.
(14) Gnanadesikan, R.; Tsao, S. L.; Kettenring, J. R. Weighting and Selection of Variables for Cluster Analysis. *J. Classif.* **1995**, *12*, 113–136.
(15) Bath, P. A.; Morris, C. A.; Willett, P. Effect of Standardisation on Fragment-Based Measures of Structural Similarity. *J. Chemom.* **1993**, *7*, 543–550.

(16) Turner, D. B.; Willett, P.; Ferguson, A.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-based Similarity Searching. *SAR QSAR Environ. Res.* **1995**, *3*, 101–130.
(17) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(18) *Pipeline Pilot*; Accelrys Software Inc.: San Diego CA, 2008.
(19) *Molconn-Z*; eduSoft LC: Ashland VA, 2008.
(20) *Unity*; Tripos Inc.: St Louis MO, 2008.
(21) Downs, G. M.; Barnard, J. M. Clustering Methods and their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1–40.
(22) *Clustering Toolkit*; Digital Chemistry Ltd.: Harewood, U.K., 2008.
(23) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343–357.
(24) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(25) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060–1066.
(26) Siegel, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioural Sciences*, 2nd ed.; McGraw-Hill: New York, 1988.

CI800224H