Lingos, Finite State Machines, and Fast Similarity Searching

J. Andrew Grant*

AstraZeneca Pharmaceuticals, Mereside, Macclesfield, Cheshire SK10 4TF, England

James A. Haigh and Barry T. Pickup

Department of Chemistry, The University of Sheffield, Sheffield S3 7HF, England

Anthony Nicholls and Roger A. Sayle

OpenEye Scientific Software Inc., 3600 Cerrillos Road, Suite 1107, Santa Fe, New Mexico 87507

Received May 25, 2006

We apply a recently published method of text-based molecular similarity searching (LINGO) to standard data sets for the purpose of quantifying the accuracy of the approach. Our implementation is based on a pattern-matching finite state machine (FSM) which results in fast search times. The accuracy of LINGO is demonstrated to be comparable to that of a path-based fingerprint and offers a simple yet effective method for similarity searching.

INTRODUCTION

Searching and matching patterns of text is a broad problem in information retrieval. Algorithms for string-matching in computational biology have been widely applied to problems of identifying relationships in DNA or protein sequences.[1,2] The well-defined alphabets of DNA and proteins make searching for similarities amenable to text-based approaches. By contrast, chemoinformatics approaches tend to focus on graph-based methods for describing molecules and identifying similarities between them. However, recent work has investigated simplified molecular representations to obtain similarity measures.[3−6] The Lingo method[5,6] elegantly demonstrates a text only approach for the efficient computation of molecular similarity.

The basis of the Lingo method is the encoding of molecular structure into simple linear strings. An example of such encoding is the SMILES chemical notation,[7] capable of representing molecular information about atoms, bonds, aromaticity, charge, stereochemistry, and isotopic substitution. Canonicalization[8] ensures that a given molecule possesses a unique SMILES string. The Lingo approach decomposes a canonical SMILES string into a set of character strings referred to as "Lingos". These substrings can be used as parameters for fitting linear models of experimental data such as LogP. Assessing the similarity of a pair of molecules is simply a matter of comparing the Lingos that occur in each molecule.

The premise that similar molecules exhibiting neighborhood behavior[9] are represented by similar canonical SMILES strings is a very approximate approach to describing molecular similarity. Recent methodological developments aim to improve the effectiveness of graph-based similarity measures by including pharmacophore features[10−12] or by applying data fusion techniques[13−16] to multiple similarity measures. These have a significant impact by increasing computational complexity. In contrast, the Lingo approach has the considerable advantage that it can be implemented without requiring the computation and storage of objects representing information about atoms and bonds. Implementations can also take advantage of established methods for string comparison such as finite state machines (FSM), also known as deterministic finite automata.[17] Although it has been observed that there is "a bewildering variety of approaches"[18] to calculating chemical similarity, the text-based nature of the Lingo approach allows computationally efficient implementations of well developed string-matching techniques taken from computer science. The efficiency offers the potential for fast neighborhood calculations, requiring only little computer memory, so that applications including real time navigation (browsing) of very large molecular databases and fast clustering of data become feasible. The focus of this paper is to establish whether the inherent simplicity of the Lingo assumption leads to a useful measure of similarity.

METHODS

The Lingo method is closely related to the d^2 algorithm[19,20] for the comparison of biological sequences. This alignment-free algorithm characterizes sequences as a set of substrings. Substring frequency is used as the basis for assessing sequence similarity.[21] Lingo character substrings are generated by stepwise fragmentation of a canonical SMILES string. Prior to fragmentation all ring closure digits are set to zero, but unlike the methodology described in the original reference[5] element names such as Cl and Br remain unaltered. Such a decomposition describes a molecule with canonical SMILES string of length n, in terms of the frequency distribution of a set of $(n-q+1)$ Lingos (of length q). Molecules are similar if they have similar Lingo distributions. This can be quantified for two molecules A and B using the Tanimoto coefficient, T^{AB}defined thus

* Corresponding author e-mail: andrew.grant@astrazeneca.com.

10.1021/ci6002152 CCC: $33.50
Published on Web 09/08/2006

$$T^{AB} = \frac{\sum_{i}^{s} \min(n_i^A, n_i^B)}{\sum_{i}^{s} n_i^A + \sum_{i}^{s} n_i^B - \sum_{i}^{s} \min(n_i^A, n_i^B)}$$

where n_i^A is the count of the ith Lingo type, belonging to molecule A, and s is the number of Lingo types in molecule A (or B). For decomposition into Lingos of length q, the total number of Lingos in molecule A is given by

$$\sum_{i}^{s} n_i^A = |A| - q + 1$$

where $|A|$ is the length of the SMILES string of molecule A (similarly for B).

The computation of T^{AB} is the problem of matching the substrings that occur in molecules A and B. This is trivial to implement using naïve (brute force) character comparisons over all pairs of Lingos in molecules A and B, but the approach has a high computational overhead in which in the worst case the number of character comparisons is of order $\Theta(|A| \times |B|)$

Applying Lingos for similarity searching or clustering will compare many (perhaps millions) of molecules to a given query, so a precomputational step exploiting the internal structure of the query SMILES string is useful. We can reduce the number of required character comparisons by constructing in $\Theta(|A|)$time, a pattern-matching FSM to represent the structure of the query SMILES string. With the FSM, it is possible to traverse the SMILES string of an arbitrary molecule and compute T^{AB} in a linear search time.

Finite state machines are widely studied models of computation. Their domain of application is to problems that can be formulated into terms of a finite set of outcomes (states) and a finite number of inputs. The only effect that input to the machine can have is to change it from one state to some new state. Special states, designated "acceptor" states, recognize a subset of sequences of the input. In the context of Lingos, a FSM is built for the query in which the states are all the unique q-length Lingos and their character prefixes. Input to the machine is the characters of the SMILES language, and the acceptor states are the q-length Lingos of the query. Specifically, q-length Lingos are initially inserted into a trie data structure.[22] This data structure represents the set of relevant Lingos as a rooted n-ary tree, where examining successive characters selects the child branch to follow. For equal length Lingos, all leaf nodes appear at the same depth, and record information such as the number of occurrences of the Lingo in the query. As a second phase, this trie is converted into a deterministic finite state machine (or automata) by adding failure transitions, to each node for each character index that does not have a child. The nodes of the trie become the states of the finite state machine.[23,24] The important result is that in scanning a database each character is only ever visited once, to identify Lingos matching those of the query.

Computational Performance. The initial implementation of the FSM algorithm enables calculation of the Lingo pairwise similarity (T^{AB}) at a mean rate of 550000 per second. This performance was measured on a standard desktop PC (2.8 GHz, 2 G RAM, Windows 2000 operating system). The FSM of the query SMILES is represented as a lookup table, which requires little memory, and the algorithm scales linearly with respect to database size.

CALCULATIONS AND RESULTS

There are different ways to assess the performance of a similarity measure. An objective measure of a similarity method is the ability to distinguish compounds active in a certain biological assay from a collection of randomly selected molecules. Typically, a molecule from an active series is used as a query and neighbors to this molecule obtained from a database, based on the similarity measure. Such evaluations can depend strongly on the choice of query, database, and criteria used to assess successful outcomes of the search. To avoid introducing a design basis favorable to the Lingo method we adopt previously published protocols.[25,26] This allows a limited degree of comparison to what are considered more sophisticated approaches to similarity searching, as described in the references. These benchmarks use the MDL Drug Data Report (MDDR),[27] a compilation of druglike molecules associated with a target (or therapeutic) class. The benchmarks determine the extent to which molecules of a specified activity class can be retrieved from a set of molecules considered inactive and chosen at random from the MDDR database.

The Lingo similarity method has a dependence on the length of substrings generated from decomposition of a canonical SMILES string. There is also a more subtle dependence on the algorithm used to generate canonical SMILES. The current Daylight canonicalization algorithm[28] differs from previous versions and is in turn different than that implemented in the OEChem toolkit.[29] SMILES strings can optionally encode stereochemical information, although not all molecules in the MDDR have their stereochemistry defined. We use "isomeric" to designate canonical SMILES with stereochemistry descriptors and "nonisomeric" to those from which this information is omitted. The influence of using isomeric or nonisomeric SMILES, the dependence on the length of Lingo substrings, and the choice of canonicalization algorithm are considered in the first set of calculations.

Briem and Lessel Benchmark. Briem and Lessel[25] designed a protocol intended to simulate the identification of active compounds from an arbitrary collection. The role of "active" molecules is modeled by selecting MDDR ligands from five activity classes, a total of 383 molecules. The activity classes and their abbreviations are as follows: ACE inhibitors (ACE), TXA2 antagonists (TXA2), HMG-CoA reductase inhibitors (HMG-CoA), PAF antagonists (PAF), and 5HT3 antagonists (5HT3).

A set of 574 randomly selected compounds known not to belong to any of the five activity classes comprises an inactive set. Calculations using this protocol are simplified by the availability of the relevant reference codes (available at http://cheminformatics.org/datasets/). For a query belonging to a certain activity class a similarity search of the entire database of 957 molecules with the Lingo approach is carried out to obtain the 10 nearest neighbors. A retrieval rate is defined as the percentage of the neighbors that are in the same activity class as the query molecule. This is used as a

Table 1. Mean Lingo Retrieval Rates for Each Briem and Lessel Activity Class[a]

			isomeric		nonisomeric	
class	*N*	mean SMILES length	Daylight	OEChem	Daylight	OEChem
ACE	40	54	84.5	83.3	77.8	77.3
TXA2	49	50	72.7	66.1	63.9	64.5
HMG-CoA	111	60	83.8	84.0	84.2	84.5
PAF	134	58	67.2	65.5	68.3	65.9
5HT3	49	41	62.9	57.8	65.9	60.6
mean of all classes			74.2	71.3	72.0	70.5

[a] These are computed using isomeric and nonisomeric SMILES strings, canonicalized using the Daylight and OEChem algorithms. The Lingo length is 4; *N* is the number of ligands in each activity class.

measure of success in finding active molecules. The procedure is repeated for all molecules in a class to obtain a mean retrieval rate. This in turn is applied to each of the five activity classes, using different lengths of Lingo substrings and Daylight and OEChem canonicalization algorithms, and applied to both isomeric and nonisomeric SMILES strings. Results of these searches for a Lingo length of four are given in Table 1.

It is clear that the Lingo method identifies molecules from a given activity class, with retrieval values ranging from 57 to 84% that are much higher than a random assignment (9%). The variation in retrieval rate is probably intrinsic to the number and type of structural motifs associated with the activity classes. The precise reasons for this variation were not explored beyond observing an approximate correlation with the mean Lingo pairwise similarity computed for each class. We are more concerned with describing the characteristics of the Lingo approach and the behavior with respect to existing similarity measures. A systematic feature observed in the retrieval data (Table 1) is that the Daylight canonicalization algorithm gives a slightly improved performance compared to what is currently implemented in OEChem. There is slightly better retrieval when stereochemistry information is retained, by using isomeric SMILES. Table 1 also shows that the retrieval rate averaged over all activity classes is nearly constant irrespective of the canonicalization details, or mean length, of the SMILES string.

This overall average was also computed for all Lingo lengths between 1 and 10, and the results were plotted in Figure 1. The simplest approach of matching character counts in SMILES strings (Lingo length of one) results in a reasonable average retrieval rate for this benchmark of 45−50%. However, using longer Lingo strings of 3−5 characters demonstrates considerable improvement in the overall retrieval rate, with an optimal performance at length 4. This is the Lingo length adopted in the original description of the method[5] and is used in all of the subsequent calculations.

Given the simplicity of Lingo it is useful to compare its performance relative to more complex approaches. Figure 2 compares the performance of Lingo relative to Daylight fingerprints (1024 bits). The Daylight retrieval rates were calculated using the same criteria as for the Lingos (i.e. the mean number of actives in the top 10 neighbors). For clarity, only the results using isomeric SMILES strings are plotted. The Lingo and Daylight fingerprint retrieval rates for each

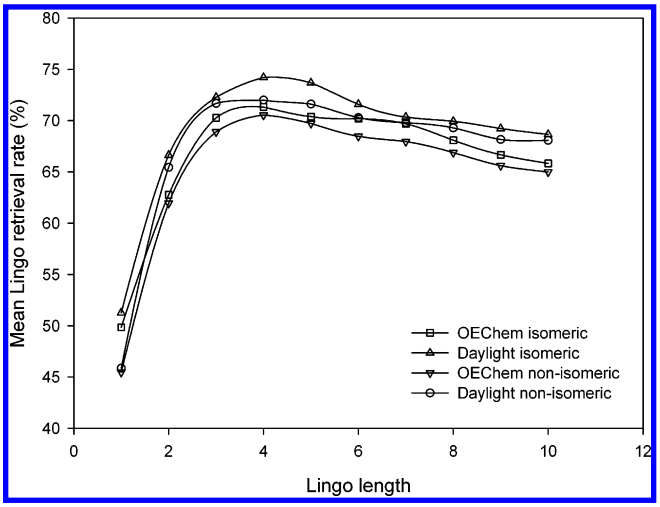

Figure 1. Variation in the Lingo retrieval rate with Lingo length. The retrieval rate is averaged over the five Briem and Lessel activity classes. Data are shown for isomeric and nonisomeric SMILES strings, canonicalized by Daylight or OEChem.

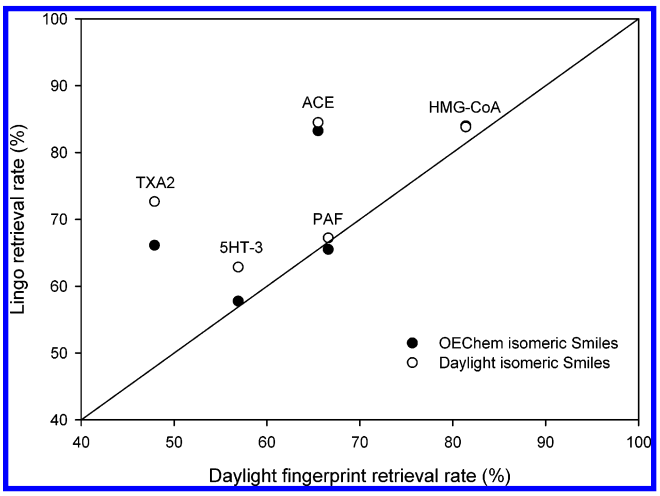

Figure 2. Comparison of Lingo and Daylight fingerprint retrieval rates, for the Briem and Lessel activity classes. The Lingos, of length 4, are obtained from Daylight and OEChem canonicalized isomeric SMILES strings. The Daylight fingerprints contain 1024 bits. The unit line is shown for clarity.

Table 2. Mean Tanimoto and Z-Score Values for the Briem and Lessel Activity Classes

	$\langle T \rangle_{10}$[a]		$\langle T \rangle_{db}$[b]		$<$Z-score$_{10}$[c]	
class	Lingo	DyFP	Lingo	DyFP	Lingo	DyFP
ACE	0.44	0.58	0.13	0.28	4.1	4.1
TXA2	0.41	0.51	0.12	0.25	4.1	4.3
HMG-CoA	0.47	0.61	0.11	0.25	4.6	4.9
PAF	0.40	0.59	0.10	0.27	4.9	4.6
5HT3	0.33	0.54	0.09	0.27	4.3	4.1

[a] Class averaged mean Tanimoto values for each active relative to its top 10 neighbors. [b] Class averaged mean Tanimoto value for each active relative to the entire database. [c] Class averaged mean Z-score of the top 10 neighbors for each active.

activity class roughly track each other, though for two classes (ACE and TXA2) active compounds are better identified using Lingos.

Table 2 gives explicit Tanimoto values for Lingo and Daylight fingerprints (DyFP). For each active, the mean Tanimoto value relative to the top 10 neighbors and the entire

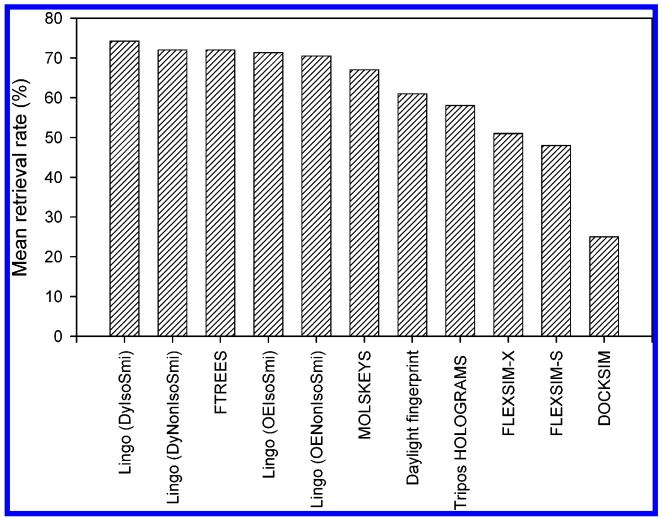

Figure 3. Variation in the retrieval rate using different similarity methods. The retrieval rate is identically computed for all methods and averaged over the Briem and Lessel activity classes. Data for the methods other than Lingos are taken from the original publication of the benchmark.[25]

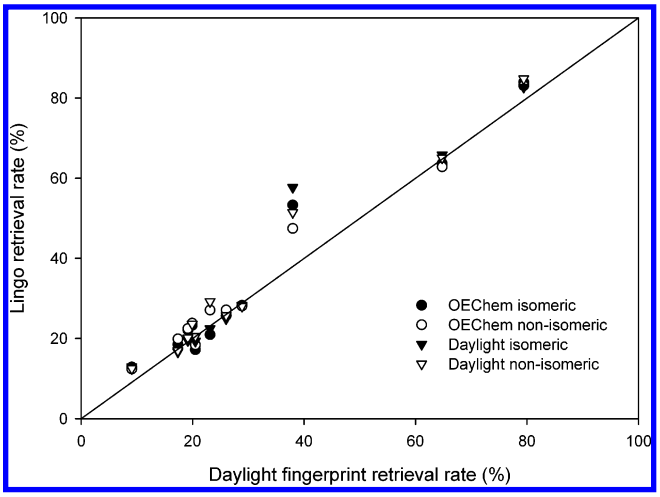

Figure 4. Comparison of Lingo and Daylight fingerprint retrieval rates, for the Hert et al. activity classes. The Lingos, of length 4, are obtained from Daylight and OEChem canonicalized isomeric and nonisomeric SMILES strings. The Daylight fingerprints contain 1024 bits. The unit line is shown for clarity.

database is computed. These means are averaged over all of the actives for each class to give the values $\langle T \rangle_{10}$ and $\langle T \rangle_{db}$, respectively. It can be seen that for either method there is little variation dependent on activity class. The mean Z-scores[30] obtained from the Tanimoto values of the top 10 neighbors show almost no difference between classes or methods. Hence, while Lingo Tanimoto values are systematically smaller than their DyFP counterparts, the underlying distributions are practically identical, which is reflected in the comparable overall retrieval rates.

To further exemplify the overall performance of the Lingo approach Figure 3 reports the retrieval rate averaged over each activity class, relative to the methods reported originally by Briem and Lessel. These results indicate that the Lingo method behaves similarly to other 2D-based approaches to similarity (Daylight fingerprints, MOLSKEYS, FTREES, Tripos Holograms).[25] The non-Lingo data in Figure 3 are taken from the original work describing the benchmark.[25] The overall mean Daylight fingerprint retrieval rate was determined in this work to be 64%, which is in good agreement with the published value of ~61%. Finally, we analyze the average overlap neighbors lists computed by Lingo and Daylight fingerprints. These were computed by counting the number of compounds in common in the 10 nearest neighbors for each active compound. Average overlaps for each activity class were as follows: ACE 49%, TXA2 40%, HMG-CoA 44%, PAF 54%, and 5HT3 40%. This level of overlap suggests a shared characteristic between the Lingo and Daylight fingerprint approach.

Hert and Willett Benchmark. A more expansive benchmark, designed by Hert et al.,[26] quantifies the ability of similarity methods to select actives from a database. Active sets of molecules are selected from the MDDR, but unlike Briem and Lessel the entire MDDR database is used to model the inactive compounds. Table 3 shows the sets of active compounds used for the benchmark. Although the reference codes for the molecules belonging to these activity classes are available from http://cheminformatics.org/datasets/, it is difficult to match these codes with those in the present release of the MDDR. Consequently, we assembled the active sets of molecules by extracting them from the MDDR based on the activity class identifier. This resulted in slightly different numbers in each activity class compared to Hert et al. and is shown Table 3. Following removal of duplicates and molecules with heavy atom count greater than 100, the number of molecules in the database used to represent inactive decoys was 155 319, compared to 102 535 reported by Hert et al.

Table 3. Activity Classes Used in the Hert et al. Benchmark[26] [a]

activity class	activity class (MDDR code)	*N* (Hert et al.)	*N* (this work)
5HT3 antagonists	06233	752	858
5HT1A agonists	06235	827	1029
5HT reuptake inhibitors	06245	359	951
D2 antagonists	07701	395	510
renin inhibitors	31420	1130	1246
angiotensin AT1 antagonists	31432	943	2185
thrombin inhibitors	37110	803	1183
substance P inhibitors	42731	1246	1763
HIV protease inhibitors	71523	750	1026
cyclooxygenase inhibitors	78331	636	722
protein kinase C inhibitors	78374	453	565

[a] *N* is the number of ligands in each activity class and is shown for the original and present work.

A ranked list of neighbors based on a similarity measure is obtained for a query by searching the entire database. The retrieval rate is the fraction of the activity class of the query found in the top 5% of this list. This rate is computed by obtaining an average from all members of the class. This is slightly different to the Briem and Lessel benchmark that defines retrieval rate as a fraction of the neighbor list that belongs to the activity class of the query.

Figure 4 compares the average retrieval rate for each class computed using Lingos and Daylight fingerprints. There is little variation in the Lingo performance depending on the details of generation of the SMILES string, although as shown in our Briem and Lessel results there is some benefit from using the Daylight canonicalization algorithm. The graph shows that these results span a much greater range in retrieval rate than the previous benchmark. Importantly the performance of the Lingo approach is almost identical to that of Daylight fingerprints across the entire range.

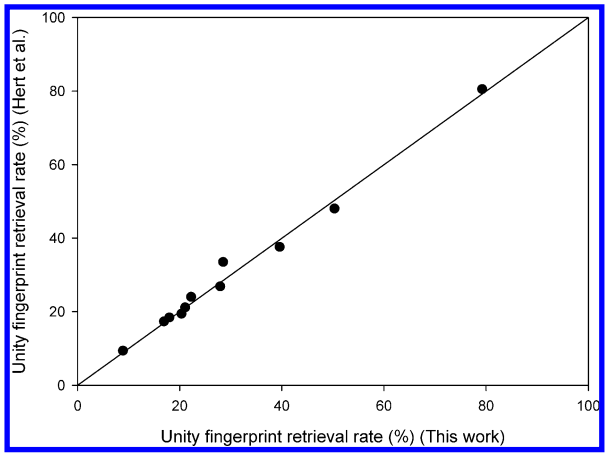

Figure 5. Comparison of the Unity fingerprint retrieval rate obtained in this work with that of Hert et al.[26] The unit line is shown for clarity.

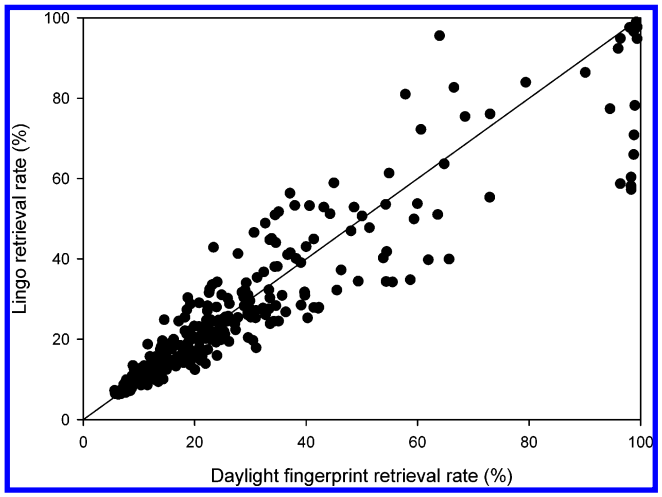

Figure 6. Comparison of Lingo and Daylight fingerprint retrieval rates, for all of the MDDR activity classes containing at least 100 compounds. Lingos of length 4 are obtained from OEChem canonicalized isomeric SMILES strings. The Daylight fingerprints contain 1024 bits. The unit line is shown for clarity.

Figure 5 compares the retrieval rate computed using Unity fingerprints[31] obtained from our data sets to that published in Table 10 in Hert et al.[26] Despite small variations in the size of each activity class, and the difference in the number of compounds used to model inactives, the agreement is excellent.

Further Evaluation of Retrieval Rates of MDDR Activity Classes. The calculations in the previous sections involve searching for compounds in a few MDDR activity classes. These had been selected to construct standard benchmarks to test the performance of similarity measures and statistical methods. To investigate whether the selection of specific activity classes introduced bias into the retrieval data, we selected every activity class in the MDDR with more than 100 molecules, resulting in 291 classes. The Hert et al.[26] protocol was applied to these, and the results for Daylight fingerprints and Lingos are shown in Figure 6. A potential drawback is that not all these classes correspond to a well-defined biological target, such as a protein receptor. Some of the classes are described in very generic terms, such as being classified as analgesics. This introduces the caveat that not all of the points in Figure 6 represent searches for compounds exhibiting straightforward neighborhood behavior

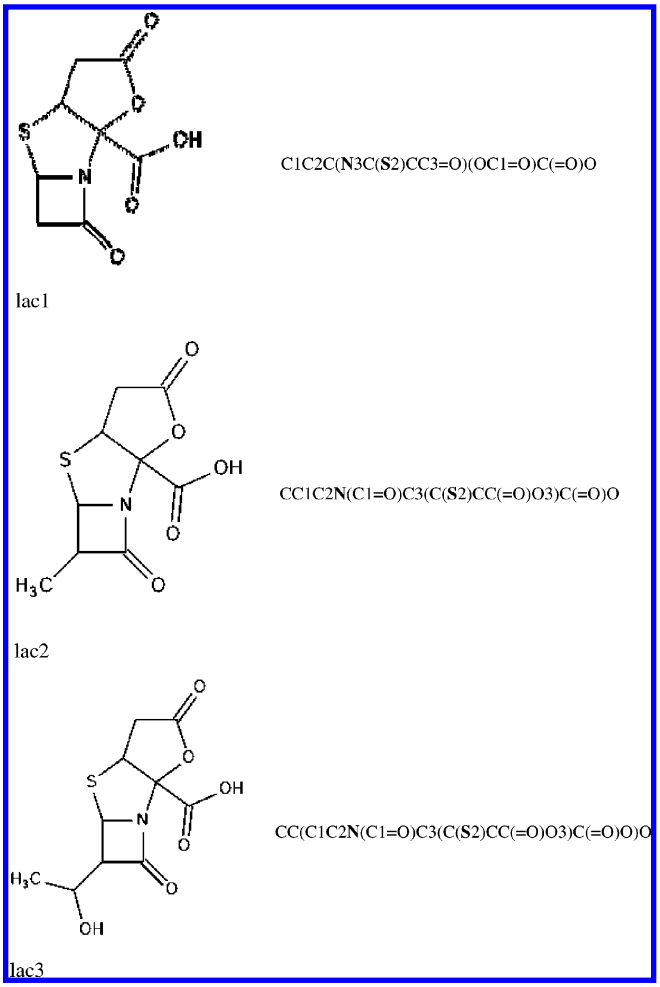

Figure 7. A set of antibiotic compounds from the MDDR activity class 64 300 and their associated canonical SMILES strings. The nitrogen and sulfur atoms are highlighted in bold type to illustrate their different positions in the SMILES string.

with respect to biological activity. However, the graph does show that overall the Lingo approach has a comparable performance relative to Daylight fingerprints.

The graph also reveals some significant differences in retrieval rates between the two approaches. A few of these outliers have been investigated. For example, the Lingo approach retrieves nearly all (96%) of the corticosteroidal activity class 02400, compared to only 64% by the Daylight fingerprint method. Of interest to us are the points for which there is nearly perfect retrieval using Daylight but markedly lower using the Lingo approach. Although the Lingo retrieval rate for these searches is still high (approximately 55−75%), these are much less effective in yielding active compounds compared to Daylight. An example is the antibiotic activity class 64 300, which has a 98% retrieval rate by Daylight but only 57% using Lingos. The lack of effectiveness becomes clear when considering the structures and their associated canonical SMILES strings in Figure 7.

These compounds have almost identical chemical structures. For example, lac1 and lac2 differ only by a methyl substitution on the beta lactam ring, but this produces a significant effect on the atom ordering within the canonical SMILES string. This can be seen in Figure 7 in which the relative positions of the nitrogen and sulfur atoms have been highlighted. Such reordering can change the pattern of Lingos for a given molecule. This results in the computed Lingo

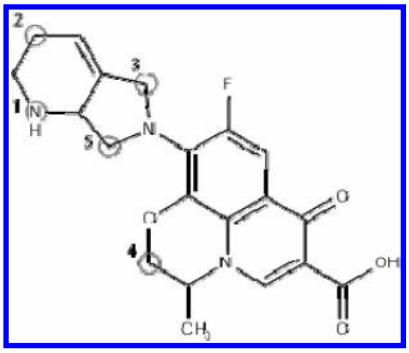

Figure 8. An antibacterial pyridonecarboxylic acid derivative. The circles denote points at which methyl groups have been introduced for the purpose of pairwise similarity calculations using Lingos and Daylight fingerprints.

Table 4. Pairwise Lingo Similarity (Daylight in Parentheses) for the Antibiotic Compounds Shown in Figure 7

	lac1	lac2	lac3
lac1	1.00 (1.00)	0.24 (0.99)	0.23 (0.95)
lac2		1.00 (1.00)	0.84 (0.96)
lac3			1.00 (1.00)

similarity (Table 4) of lac1 relative to lac2 and lac3 being 0.24 and 0.23, respectively, which reduces performance of the search. The high similarity for all of these structures using Daylight fingerprints gives near-perfect retrieval of compounds in this activity class. This presumably arises from a special path or cycle characteristic of the molecular graphs of these compounds.

A similar example is the antibacterial pyridonecarboxylic acid derivative shown in Figure 8 from the MDDR activity class 68 210. Methyl substitution at any of the labeled positions generates a total of six structures. Computing their pairwise similarity using Daylight fingerprints results in a narrow range of high Tanimoto values, 0.95−0.99. Using Lingos the range is 0.13−0.86. As before this is a result of the very different atom ordering in the canonical SMILES strings.

Simplified Models. The results have been used to compare Lingos to more complex similarity measures, such as fingerprints. As a further assessment, we investigate simpler text-based methods. Pairwise similarity is computed from differences in lengths of SMILES strings and also differences in molecular formulas (using the city block distance[32]). These results are shown in Figures 9 and 10. Although these approaches gave retrieval rates that were fine for a small number of the activity classes in the MDDR, overall most performed little better than expected by random.

A potential simplification to the Lingo approach is to omit the canonicalization step. To explore the potential impact of this we generated 1760 different SMILES for one ACE inhibitor from the Briem and Lessel data set. The values in pairwise Lingo similarity for this set range from 0.15 to 1.0, with a mean and standard deviation of 0.45 and 0.11, respectively. Clearly, this variation points to the value of using canonicalized SMILES to obtain a consistent and transferable similarity measure. Repeating the Briem and Lessel analysis multiple times with randomly generated SMILES reduced the retrieval performance by approximately 10% on average. This smaller than expected drop in performance probably arises from the fact that although the computation

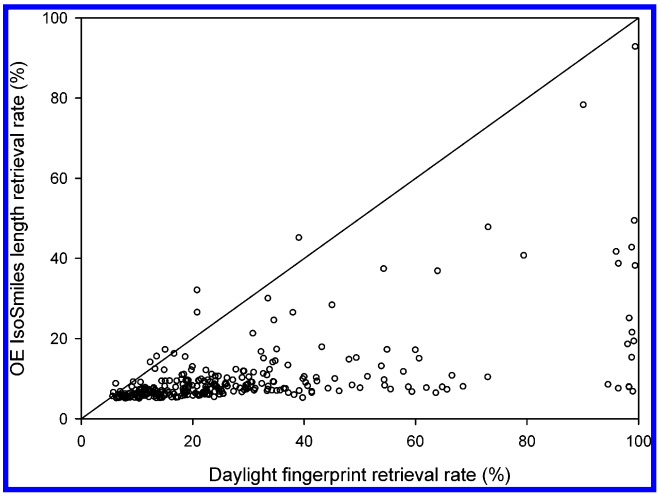

Figure 9. Comparison of difference in SMILES length and Daylight fingerprint retrieval rates, for all of the MDDR activity classes containing at least 100 compounds. The SMILES lengths are calculated from OEChem canonicalized isomeric SMILES strings. The Daylight fingerprints contain 1024 bits. The unit line is shown for clarity.

of self-similarity can have a significant range using randomly generated SMILES, mean values are still greater than similarity to random compounds.

Future Work. Analyzing results for the lactam and pyridonecarboxylic acid series suggest that future work could focus on developing a specialized canonicalization algorithm. It would attempt to preserve, where possible, atom orderings for core structures (such as the lactam ring template). The current benchmarks provide a basis for assessing whether such a modification would improve the retrieval signal. Other modifications include adaptations to Lingo generation, such as attempts to remove redundant information in the SMILES string. For instance, closed square brackets always partner an open square bracket and are not particularly representative of structure in the way parentheses denoting branching are.

This work established the optimal Lingo length to be four characters based on the analysis of molecules whose complexity is typical of that of lead or druglike molecules.[33] Further work is required to see if this is generally applicable, particularly for molecules of low complexity such as those used in fragment screening libraries.[34]

DISCUSSION

The Lingo approach is based on the insight that calculating molecular similarity can be treated in an analogous way to alignment-free biological sequence comparison.[21] The original publication emphasized the use of Lingos as parameters for linear fitting of physical properties. Although similarity searching was outlined, no systematic investigation was presented, which motivated us to examine such an approach. We were cogent of the observations by Sheridan and Kearsley:[18] "There is an unfortunate tendency in the literature for investigators to invent a new method, compare it to some standard method (usually Daylight fingerprints) for one or two activities, and then claim global superiority if the new method does better at selecting actives". Keeping this in mind, we attempted to establish the performance of Lingos relative to existing similarity methodologies. To do so, we utilized two well-established, independently designed bench-

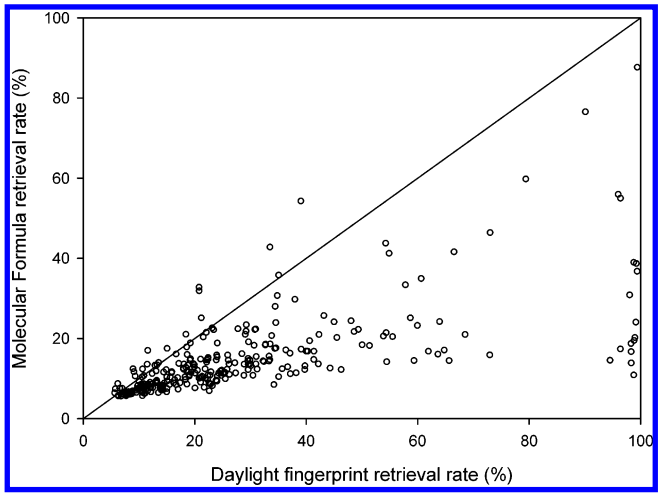

Figure 10. Comparison of the difference in molecular formula and Daylight fingerprint retrieval rates, for all of the MDDR activity classes containing at least 100 compounds. The molecular formula difference is calculated as the city block distance. The Daylight fingerprints contain 1024 bits. The unit line is shown for clarity.

marks using 16 different activity classes and a range of different methodologies for similarity searching. Despite the simplicity of the Lingo approach, the performance in these benchmarks is comparable to Daylight and in the case of the Briem and Lessel benchmark indistinguishable from several methods.

Although this claim of "global equality" might not seem interesting or significant, the superiority of the model lies in its simplicity, which has both practical and abstract consequences. From a practical point of view, the method is straightforward to implement and combines well with modern computer science techniques for text matching, leading to fast performance. The small memory requirement reduces the tendency for searches of large databases to become I/O bound. From a more abstract point of view, simpler models are preferable for several reasons. Unnecessary complexity wastes effort, for example computing fingerprints requires interpreting SMILES as molecular object, which is converted into a set of paths for hashing. Overly complex models are also difficult to interpret. The hashing associated with fingerprints often obscures the reasons molecules are considered to be similar, whereas similarities computed using Lingos are more straightforward to interpret. In general, simpler models provide good starting points to assess if complexity introduced into models is of value. Our experiments on very simplified models (Figures 9 and 10) not surprisingly showed them to be ineffective. Lingos appear to offer the simplest effective method for similarity searching, setting a performance benchmark for other methods.

REFERENCES AND NOTES

(1) Gusfield, D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*; Cambridge University Press: 1997.
(2) Navarro, G.; Raffinot, M. *Flexible Pattern Matching in Strings: Practical On-Line Search Algorithms for Texts and Biological Sequences*; Cambridge University Press: 2002.
(3) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.
(4) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21*, i359-i368.
(5) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386-393.
(6) Vidal, D.; Thormann, M.; Pons, M. A Novel Search Engine for Virtual Screening of Very Large Databases. *J. Chem. Inf. Model.* **2006**, *46*, 836−843.
(7) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(8) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(9) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(10) Bath, P. A.; Poirette, A. R.; Willett, P.; Allen, F. H. Similarity searching in files of three-dimensional chemical structures: Comparison of fragment-based measures of shape similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(11) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.
(12) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(13) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(14) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(15) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.
(16) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint Based Similarity Searches for Targeted Class Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201−1213.
(17) Cormen, T. H.; Clifford, S.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; The MIT Press: 2001.
(18) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(19) Hide, W.; Burke, J.; Davison, D. B. Biological evaluation of d^2, an algorithm for high-performance sequence comparison. *J. Comput. Biol.* **1994**, *1*, 199−215.
(20) Wu, T. J.; Burke, J. P.; Davison, D. B. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics* **1997**, *53*, 1431−1439.
(21) Vinga, S.; Almeida, J. Alignment-free sequence comparison—a review. *Bioinformatics* **2003**, *19*, 513−523.
(22) Knuth, D. *The Art of Computer Programming, Volume 3: Sorting and Searching*; Addison-Wesley: 1997.
(23) Sayle, R. A. Improved SMILES Substructure Searching. EuroMUG. http://www.daylight.com/meetings/emug00/Sayle/substruct.html. 2000.
(24) Sayle, R. A. SMILES Multigram Compression. MUG. http://www-.daylight.com/meetings/mug01/Sayle/SmiZip/index.htm. 2001.
(25) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231−244.
(26) 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. Model.* **2004**, *44*, 1177−1185.
(27) MDDR; MDL Information Systems, Inc.: San Leandro, CA.
(28) Daylight Toolkit 4.91; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA.
(29) OEChem; OpenEye Scientific Software: Santa Fe, NM.
(30) Bradshaw, J.; Sayle, R. A. Some thoughts on significant similarity and sufficient diversity. EuroMUG. http://www.daylight.com/meetings/emug97/Bradshaw/Significant_Similarity/Significant_Similarity.html. 1997.
(31) UNITY. [4.4.2]; Tripos Inc.: St. Louis, MO, 2004.
(32) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(33) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Model.* **2001**, *41*, 1308−1315.
(34) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660−672.

CI6002152