LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities

David Vidal,[†] Michael Thormann,*[,‡] and Miquel Pons*[,†,§]

Laboratory of Biomolecular NMR, Parc Científic de Barcelona, Josep Samitier, 1-5 08028 Barcelona, Spain, Morphochem AG, Gmunder Strasse 37-37a, 81379 München, Germany, and Departament de Química Orgànica, Universitat de Barcelona, Martí i Franquès, 1-11, 08028 Barcelona, Spain

Received October 21, 2004

SMILES strings are the most compact text based molecular representations. Implicitly they contain the information needed to compute all kinds of molecular structures and, thus, molecular properties derived from these structures. We show that this implicit information can be accessed directly at SMILES string level without the need to apply explicit time-consuming conversion of the SMILES strings into molecular graphs or 3D structures with subsequent 2D or 3D QSPR calculations. Our method is based on the fragmentation of SMILES strings into overlapping substrings of a defined size that we call LINGOs. The integral set of LINGOs derived from a given SMILES string, the LINGO profile, is a hologram of the SMILES representation of the molecule described. LINGO profiles provide input for QSPR models and the calculation of intermolecular similarities at very low computational cost. The octanol/water partition coefficient (LlogP) QSPR model achieved a correlation coefficient R^2=0.93, a root-mean-square error RRMS=0.49 log units, a goodness of prediction correlation coefficient Q^2=0.89 and a QRMS=0.61 log units. The intrinsic aqueous solubility (LlogS) QSPR model achieved correlation coefficient values of R^2=0.91, Q^2=0.82, and RRMS=0.60 and QRMS=0.89 log units. Integral Tanimoto coefficients computed from LINGO profiles provided sharp discrimination between random and bioisoster pairs extracted from Accelrys Bioster Database. Average similarities (LINGOsim) were 0.07 for the random pairs and 0.36 for the bioisosteric pairs.

INTRODUCTION

The search for compounds similar to a given target ligand structure and compounds with defined biophysical profiles are two important principles of the modern drug discovery process. Both tasks make use of molecular descriptors with differing complexity (atomic, topographic, substructural fingerprints, 3D, biophysical properties, etc.) leading to different representations of the same molecule.[1] Such descriptors can then be used as input for QSPR models and intermolecular distance calculations.[2] The development, implementation, and application of molecular descriptors and the subsequent mathematical treatment of the information contained in these descriptors have become an area of intense theoretical and practical interest in recent years.[2−5]

Molecular databases are searched for molecules similar to those with known bioactivities in the hope that they will exhibit similar biological profiles. This concept is commonly termed bioisosterism. The intermolecular similarity value depends hereby on the molecular description and the distance calculation employed, and relatively small structural changes, especially in ring systems, can cause large deviations in the similarity values. The biological similarity metrics depends finally on the problem given which means that the same small structural changes introduced in a given molecule might have different effects in the biological activities depending on the target. In our experience different methods work better for different targets, and various methods should be employed independently instead of using consensus scoring.

Furthermore, the drug discovery process aims at selecting bioactive compounds with favorable ADME (absorption, distribution, metabolism and excretion) properties.[6] Unfavorable ADME properties are still the major cause for attrition through the drug discovery pathway. Experimental ADME studies are, however, generally low-throughput and require the synthesis of the compound. A paradigm change in drug discovery that prefers evaluation of ADME properties prior to synthesis requires, thus, methods to predict ADME properties of novel compounds yet to be synthesized. This allows a multiparametric target-focused drug design process in which compounds with undesired ADME properties are circumvented from early on. A variety of ADME properties (absorption, distribution) can be more or less directly linked to one or a combination of biophysical properties (solubility, lipophilicity, hydrophilicity, phase partition). Fortunately, sufficient experimental data are available to allow explanatory and predictive QSPR or neural network models to be established from suitable structural descriptors.[7,8]

A wide range of structural descriptors are based on two-dimensional or three-dimensional structures.[9,10] To process large databases two-dimensional descriptors are preferred for computation speed reasons. Such descriptors, for example MDL fingerprints,[11] are built on a predefined set of substructures, whose presence or absence is recorded as one bit in a bit-string. The building of such keys is a computationally

* Corresponding authors e-mail: mpons@ub.edu (M.P.) and michael.thormann@morphochem.de (M.T.).
[†] Parc Científic de Barcelona.
[‡] Morphochem AG.
[§] Universitat de Barcelona.

10.1021/ci0496797 CCC: $30.25
Published on Web 02/15/2005

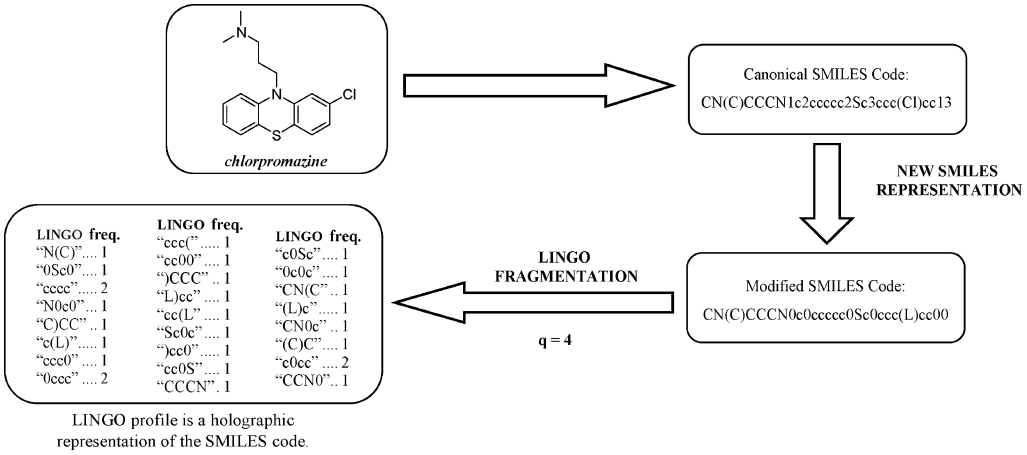


Figure 1. LINGO generation process.

expensive graph-theoretical NP-complete problem. However, the resulting keys allow very rapid molecular comparisons. It must be underlined that structural information is lost due to the limited number of predefined substructures and the binary representation. Hashed fingerprints overcome the latter problem by generating an exhaustive list of structural fragments according to a certain rule set.[12] The very large number of fragments potentially generated does not allow the assignment of an individual bit to each fragment, and, instead, several different substructures are represented by one bit using a pseudorandomizing algorithm. This leads to a new problem described as fragment collision.[13,14] This process reduces the accuracy of the fingerprint but allows the use of a much larger number of fragments.

Holographic structural representations in which the same atom can be part of several fragments have been shown to provide a computationally very efficient approximation to 3D-QSAR appropriate for handling large databases.[15] HQSAR still suffers, though, from the fragment collision problem which can complicate the interpretation of the HQSAR models.

The requirements for structural similarity calculation and property prediction are conceptually similar. Still, method development seems to diverge rather than to converge in the recent years. Molecular fingerprinting provides efficient ways for pairwise comparison of molecules and clustering, but it has not found much use on property prediction.[16] Recently, a method based on the selection of an atom type dictionary that optimizes a predictive model for logP has been shown to provide accurate predictions for other ADME properties.[17]

SMILES strings have become a standard way to store molecular connectivity in the form of a very compact text string.[18,19] A SMILES string represents univocally a molecule. The implicit molecular information contained in the SMILES string can be used to derive two- and three-dimensional structures by applying some reconstructing algorithm which unfolds the implicit information contained in atom types and their connectivities, like bond lengths, bond angles, protonation sites and by conformational sampling. The quality of the resulting three-dimensional structures does not depend on the SMILES string but on the quality of the unfolding algorithm (rule based, force field, ab initio, etc.) and off-SMILES environmental conditions (solvent, temperature, pH, etc.). Hence, a one-dimensional SMILES string (or an alternative representation like SLN[20]) implicitly contains information on the three-dimensional structure of a molecule and the properties derived from it. The one-to-one correspondence between one-dimensional representations and three-dimensional structures suggests that molecular descriptors and the associated structure based properties could be derived directly from one-dimensional representations, without the need to derive explicit two- or three-dimensional structures. To achieve this goal a suitable meta-descriptor of the one-dimensional structure is needed.

In this report we demonstrate that the ensemble of one-dimensional substrings obtained from a simple fragmentation of SMILES strings provides a holographic meta-description of the SMILES representation of a molecule and can be used directly to calculate molecular similarities and to predict structure related properties. We have coined the term LINGO to refer to each of the SMILES substrings. LINGO generation is a linear problem in contrast to the NP-complete problem associated with the generation of fragments from structures. LINGOs are text strings, and all the required manipulations can be carried out using efficient and universally available text processing software tools.

METHODS

LINGO Generation. A q-LINGO is a q-character string, including letters, numbers and symbols, such as "(", ")", "[", "]", "#", etc. obtained by stepwise fragmentation of a canonical SMILES molecular representation. For a given molecule, different SMILES representations are possible. The use of canonical SMILES ensures a one-to-one correspondence between structures and SMILES strings. Alternative canonization schemes would provide different LINGOs but would yield very similar or identical results for property or similarity calculations provided they are used consistently in the training and problem sets. A total number of (n-(q-1)) substrings of length q are extracted from a SMILES string of length n. In this work we use $q = 4$ unless indicated otherwise and the q- prefix is omitted. The molecule-specific LINGO profile is defined as the ensemble of LINGOs and their corresponding number of occurrences and does not depend on the order of appearance of LINGOs in the SMILES string. The LINGO generation process is summarized in Figure 1.

Some changes in the original SMILES code are introduced. Thus, "Cl" and "Br" are substituted by "L" and "R",

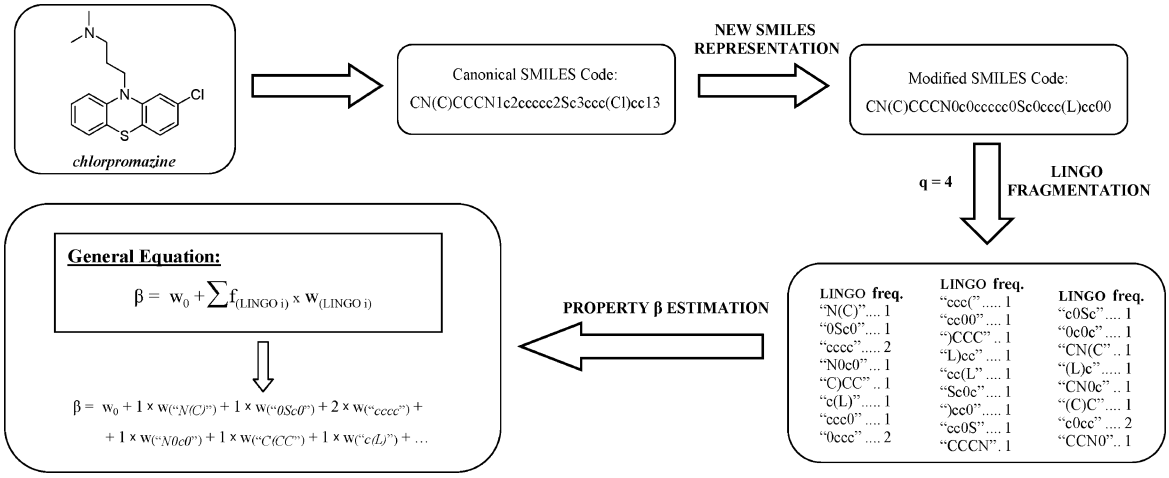


Figure 2. Application of the models predicting logP and logS.

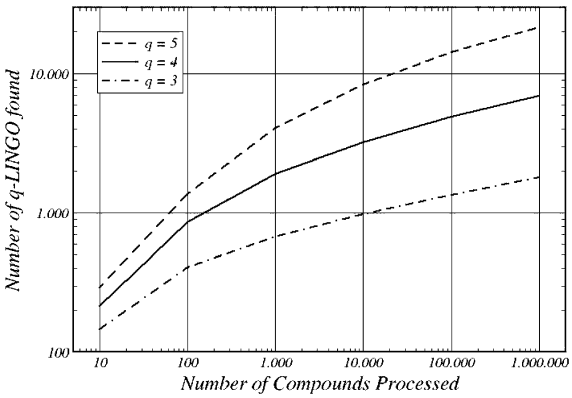


Figure 3. Number of q-LINGOs found as a function of the number of compounds processed.

respectively, and ring numbers are set to "0", for example "c1ccccc1" or "c2ccccc2" become "c0ccccc0". This normalization reduces the number of possible LINGOs and improves statistical sampling in the QSPR models, although it prevents, in general, a unique reconstruction of the original SMILES code from its holographic representation.

QSPR Modeling. LINGO profiles and experimental molecular properties were used to build QSPR models. Partial least squares (PLS)[21] was chosen as the regression method using the LINGO profile, i.e., f_L occurrences of LINGO L as X-variables and experimental property values for logP and logS as Y-variables, respectively, yielding the weights w_L corresponding to the set of LINGOs plus offset w_0. PLS analysis was performed by using the statistical package pls.pcr (version 0.2.4) in version 1.9.0 of R[22] a system for statistical computation and graphics, on a Linux PC. Latent variable PLS was performed using the kernel-PLS method with n-fold cross-validation (n=10). In the pls.pcr implementation the entire data set is scrambled randomly and divided into n nonoverlapping groups of equal size. Each group is used in turn as a test set, while the remaining n-1 groups are used as training data in separate PLS models. The final statistical parameters are then averaged over the n models. While R^2 and RRMS refer to the explanatory correlation coefficient and root-mean-square, Q^2 and QRMS refer to the average predictive correlation coefficient and root-mean-square. PLS models depend on the number of latent variables (LV) chosen. All models with LV ranging from 1 to, at least, 50 were calculated as outlined above. The model yielding the highest Q^2 and the lowest QRMS was finally chosen.

LlogP and LlogS QSPR Model Setup. Models predicting logP and logS were built from LINGO profiles. We call these models LlogP and LlogS, respectively. PLS regression was applied to derive both models. Compounds containing elements other than C, N, S, O, H, P, and halides were not included in the respective data sets. Application of these models to estimate molecular properties is presented graphically in Figure 2.

Experimental intrinsic aqueous solubility (logS) values were obtained from the widely used Huuskonen database, containing 1318 organic compounds extracted from the AQUASOL database of the University of Arizona and the PHYSPROP[23] database. After removing any duplicated compounds, a total of 1309 compounds were included in the data set. The logarithms of the aqueous solubility expressed in mol/L, were taken in the interval 20−25 °C and ranged from −11.62 to 1.58, with an average value of −2.76.

Experimental values of logP were obtained from the PHYSPROP database (version 2004) containing 25734 compounds. After removing compounds containing salts and undesired atoms, a total number of 12831 compounds, having experimental logP values, were selected for the data set. LogP values varied from −5.08 to 11.29, with an average value of 2.03.

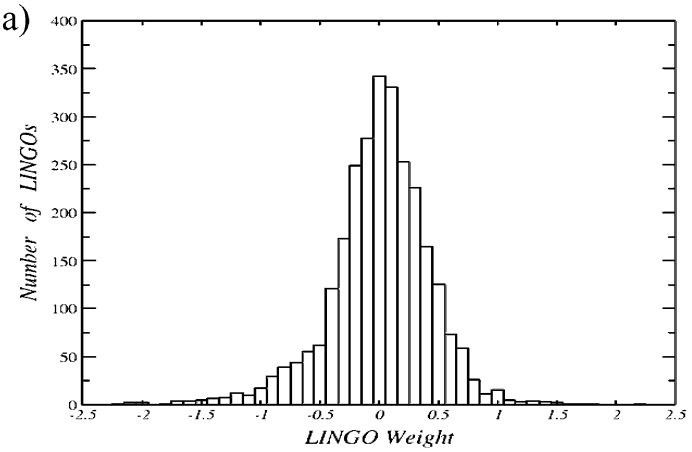
Intermolecular Similarity Calculation. Based on the LINGO profiles of the two molecules A and B to be compared, intermolecular similarities were computed using the integral Tanimoto coefficient of the form[5]

$$T_c = \frac{\sum_{i=1}^{l} 1 - \frac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l} \quad (1)$$

where $N_{A,i}$ is the number of LINGOs of type i in molecule A, $N_{B,i}$ is the number of LINGOs of type i in B, and l is the number of LINGOs contained in either molecule A or B. Molecules having the same types and numbers of LINGOs will show T_c=1. We call the LINGO profile based T_c "LINGOsim".

The BIOSTER[24] database was used for our similarity evaluation study as described in a similar way by Shuffen-

a)



b)

LINGO Weights - LlogP			
CCCR	1.797	OCCC	-1.025
CCCI	1.709	NCCS	-1.156
CCCL	1.574	OCCO	-1.167
=NNC	1.365	OCCS	-1.221
ICCI	1.069	OCc0	-1.222
CCN=	0.981	NCCO	-1.625
CCCF	0.913	NCCN	-2.019
C=CL	0.823	NCCc	-2.083
c(I)	0.822	NCCC	-2.164

Figure 4. Distribution of LINGO weights for LlogP. a) Frequency distribution of the weights of 2768 LINGO required for predicting LlogP. b) Individual LINGOs contributing with a higher absolute weight to LlogP predictions.

hauer et al.[10] Bioisosteric pairs containing salts, prodrugs, and ambiguous pairs were removed from the data set resulting in a two-column table of 7161 molecular pairs representing experimentally confirmed bioisosteric transformations. The same number of random pairs was produced by putting the compounds of the right column of this table into random order. The integral T_c values were calculated for every pair in the bioisoster and random pair lists for subsequent statistical analysis.

RESULTS AND DISCUSSION

LINGO Ensembles. Selection of the length of LINGOs is central for the success of the method. An n-character SMILES string provides an ensemble of n-(q-1) LINGOs of length q. The percentage of different LINGOs of lengths 3, 4 and 5 contained in a database containing over one million compounds from different commercial suppliers are plotted in Figure 3.

Figure 3 clearly shows that the number of different LINGOs obtained increases with q. If q is too small, most of the connectivity information will be lost, on the other hand, if q is too large the probability of finding the same string in two different SMILES strings would approach zero, preventing the use of LINGOs for similarity calculations or prediction of properties. Following trial calculations (results not shown) we have chosen a length of four as a compromise between the need to capture complex three-dimensional features implicit in the SMILES codes and the requirement of a statistically acceptable representation of each LINGO in databases.

LogP Prediction. The logarithm of the partitioning coefficient between *n*-octanol and water (logP) is an ADME related property besides being one of the key parameters in quantitative structure−activity relationship (QSAR) studies. LogP has been widely used as a measure of lipophilicity, and it is critical for both the pharmacokinetic and pharmacodynamic behavior of a molecule. For this reason, many different approaches have been developed for logP prediction based on nonexperimental structural parameters.[25−28] Fragment-based methods provide the best predictive accuracy according to the comparison of 14 different methods performed by Mannhold and Dross.[29] The number of fragments in such methods can vary from several hundred in clogP [30] to several thousand in the ACD/logP method.[31]

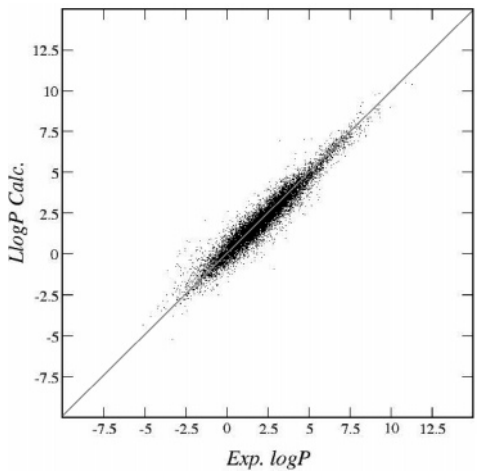
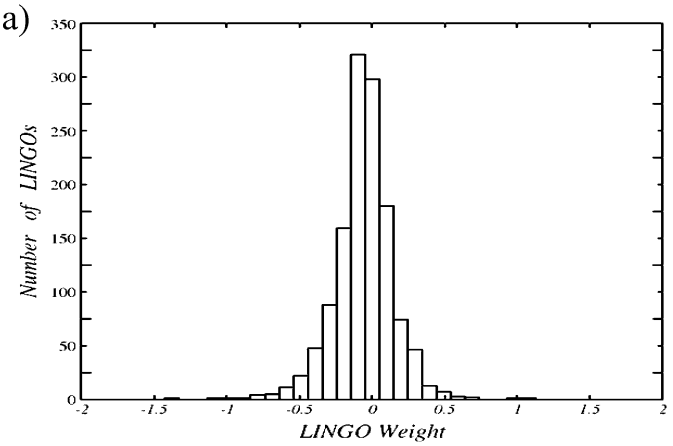


Figure 5. Relationship between experimental and calculated logP for 12831 compounds in PHYSPROP database (R^2=0.93, RRMS=0.49, Q^2=0.89, QRMS=0.61)

LINGO decomposition provides an unbiased analysis of SMILES representations. The occurrence of particular LINGOs may be associated in a statistical way to measurable properties, and a predictive model may be extracted from the analysis of a proper training set. We have found a robust LINGO based model to predict logP, that we have called LlogP. We have used the PHYSPROP database with 12831 compounds. These contained 2768 different LINGOs. The final model used 60 latent variables, and each LINGO had its own weight in the final equation estimating logP. Figure 4a shows the frequency distribution of weights of different LINGOs in the final model. In the tails of the distribution are those LINGOs that have a higher weight in the model. They are listed in Figure 4b. The actual weight of LINGOs is biased by their frequency. Thus, rare LINGOs present in molecules displaying extreme values for a given property will show higher weights. The frequency distribution in Figure 4a show that most of the LINGOs have small weights, thus even molecules containing LINGOs that are not present in the training set have a reasonable probability to be predicted correctly.

The predictions of the best 60 latent variable LlogP model gave a R^2 of 0.93 and RMS error of 0.49 for the training set, and with a Q^2 of 0.89 and QRMS of 0.61 for the cross-validation set (Figure 5). The Q^2 value of 0.89 indicates the

a)



b)

LINGO Weights - LlogS			
CCCR	-1.367	=O)O	0.460
Lc0c	-1.099	cncc	0.493
Rc0c	-0.909	N0CC	0.508
CC=C	-0.827	OCc0	0.543
CCCL	-0.798	Oc0c	0.570
CCCI	-0.785	CN0C	0.616
CC#C	-0.595	C)CO	0.693
CCCc	-0.560	CC=O	0.959
CCCC	-0.520	(C)O	1.127

Figure 6. Distribution of LINGO weights for LlogS. a) 1287 LINGO required for predicting LlogS model. b) Some of the LINGO weights yielded by PLS analysis.

model is robust and highly predictive. An analysis of compounds with absolute errors higher than 2 log units showed that ca. one-third are zwitterions, containing both an amino group and a carboxylic acid, for which logP is not properly defined.

For comparison, logP was also estimated for the same data set by using clogP, considered one of the standard methods for predicting logP. We obtained a R^2 correlation of 0.84 and a RMS Error of 0.78 log units. In a comparison of thirteen logP calculators presented by Duban et al. R^2 values ranged from 0.95 to 0.74 and the mean absolute errors, from 0.36 to 0.98.[32] Thus, LlogP has a performance that is comparable to the best existing programs at a considerably lower computational cost.

LogS Prediction. Aqueous solubility plays an important role in various physical and biological processes. Closely related to the ADME profile, aqueous solubility can affect the transport, release, and absorption of a drug. Many different methods have been developed for estimating aqueous solubility. A widely used method is the Yalkowsky general solubility equation (GSE)[33,34]

$$\log S_w = 0.5 - 0.01(\mathrm{MP}-25)-\log P \qquad (2)$$

where $\log S_w$ is the logarithm of the molar aqueous solubility, MP is the melting point and logP is the logarithm of the octanol−water partition coefficient. Unfortunately, MP is a complex property not easily accessible by computational means. Thus, other methods have been published employing structural parameters calculated directly from the molecular structure such as topological indices, molecular volume, surface area, and even quantum-chemical calculations.[35−37] LogS is a much more complex biophysical property than logP as it involves intermolecular interactions in the solid state, state transition, etc. In addition, for some compounds, other intermolecular interactions such as self-association or micelle formation in solution may be involved.

For comparison reasons we took the widely used Huuskonen database that contains experimental logS values for 1309 unique compounds yielding 1287 LINGOs in the QSPR model. The distribution of LINGO weights and a list of those in the tails of the distribution, and therefore having a stronger contribution to the LlogS values, are presented in Figure 6.

The best LlogS model based on the Huuskonen data set used 15 latent variables and yielded a R^2 of 0.91 and a RMS

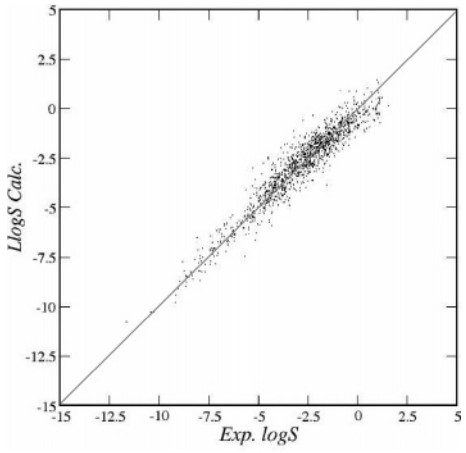


Figure 7. Relationship between experimental and calculated logS for 1309 compounds in Huuskonen database (R^2=0.91, RRMS=0.61, Q^2=0.82, QRMS=0.87).

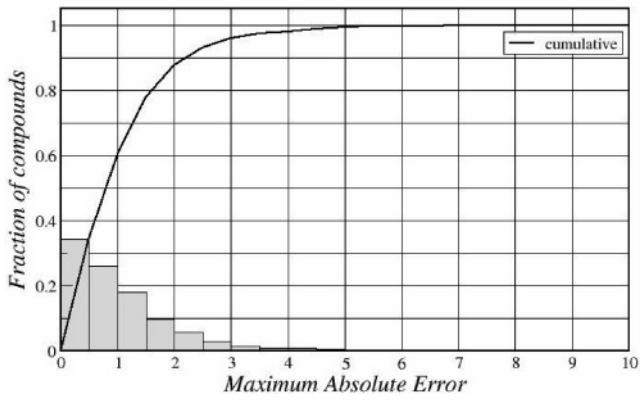


Figure 8. Distribution of absolute error between experimental and LINGO calculated logS values for a test set containing 3820 compounds from PHYSPROP database.

error of 0.60 for the test set, and a Q^2 of 0.82 and QRMS of 0.87 log units for the validation set (Figure 7). The major outliers in the set include mainly hydrocarbons and other highly hydrophobic compounds that have very low aqueous solubility or amphipathic molecules that are likely to form micelles in water.

Although the predictive power is lower than for logP, this model is useful and comparable to other existing methods.[17,34] 3820 unique compounds with measured logS values not

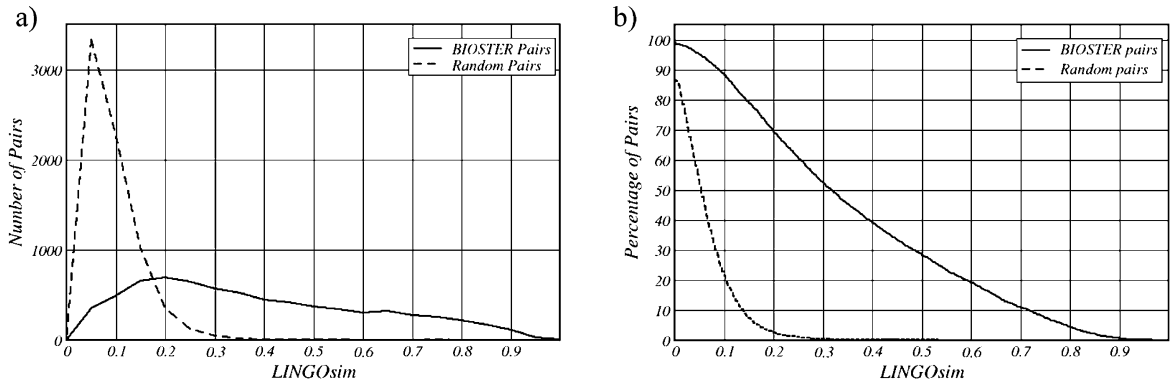


Figure 9. a) Frequency distributions of LINGO similarities for random and bioisoster pairs of compounds and b) cumulative distributions for the same pairs. Grey line refers to random pairs and black line to bioisosteric pairs.

present in the Huuskonen data set were selected from the PHYSPROP database and used to test the predictive power of the LlogS on a free prediction data set. The distribution of the absolute error between experimental and calculated values is plotted in Figure 8. The plot shows that LlogS for 60% of the compounds is predicted within less than 1.0 log units deviation. Regression coefficients of the distribution yielded a Q^2=0.64 and QRMS=1.42 log units. LlogS values were also estimated for the same data set of 3820 compounds by using the general solubility equation from Yalkowsky (equation 2) applied to the measured logP and MP values as taken from the PHYSPROP database. They yielded a distribution with a Q^2 of 0.58 and RMS error of 1.85 log units. Interestingly, most of the compounds showing the highest absolute errors using LlogS give also a similar absolute error using the Yalkowsky method. This might suggest problems with the experimental design or compound-inherent problems.

A second LlogS model based on the PHYSPROP data set was calculated including 4117 compounds with experimental logS values obtained at 20−26 °C. This model based on 2240 LINGOs yielded R^2=0.84, RRMS=0.91, Q^2= 0.75 and QRMS = 1.13 using 21 latent variables. This model performs worse than the Huuskonen-based LlogS. This is related to the worse statistical sampling of the distinct LINGO types in the data set (data not shown).

Intermolecular Similarity. A measure of similarity can be obtained by calculating the metric distance between LINGO profiles derived from SMILES representations of two molecules using the integral Tanimoto coefficient (equation 1). We have calculated the LINGOsim values for the set of 7161 bioisosteric pairs and the same number of random pairs. Distributions of similarities are plotted in Figure 9a.

This method recognizes similarities between bioisoster pairs and discriminates them from random pairs. This conclusion is supported by the cumulative frequency shown in Figure 9b, where less than 5% of random pairs have similarities higher than 0.17 while 75% of bioisoster pairs have higher similarities. The marked separation between bioisosteres and random pairs demonstrates that biologically and chemically relevant structural information is indeed captured by simple LINGO profiles. This method requires only simple string manipulation and scales only linearly with SMILES length.

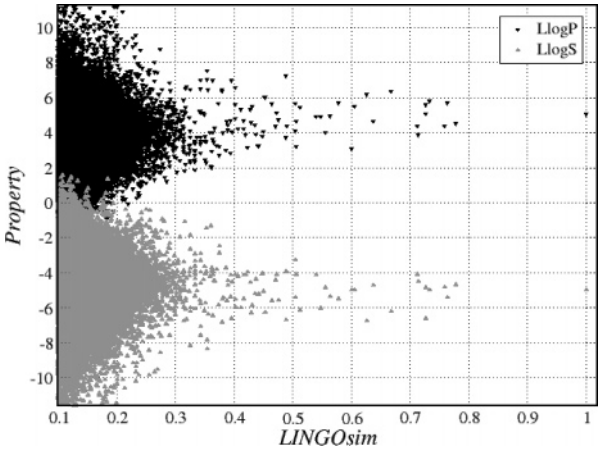


Figure 10. Relationship between logP and logS predictions and similarities based on LINGO profiles. The reference compound is chlorpromazine, and the analysis is of the NCI database.

The average similarity (± standard deviation) for the bioisoster pairs set is 0.36±0.23 using LINGOsim and 0.54±0.21 using UNITY 2D. The average and standard deviation of integral Tanimoto coefficients between random pairs is 0.07±0.06 for LINGOsim and 0.22±0.09 for UNITY 2D. Thus, LINGOsim and UNITY 2D show similar discrimination between bioisosters and random pairs.

Bioisosterism and Biophysical Properties. Holographic LINGO profiles provide a straightforward combined approach to biophysical properties and intermolecular similarity and can, thus, be applied to drug discovery and lead optimization projects. Medicinal chemistry introduces commonly small changes into a lead molecule that are hopefully tolerated by the target receptor. At the same time the biophysical properties of these bioisosters change in respect to the initial lead. It can, thus, be expected that bioisosters show also similar biophysical properties. To address this question with an example, tentative chlorpromazin bioisosters have been retrieved from the NCI database and their biophysical properties have been subsequently computed using LINGOsim and LlogP and LlogS, respectively. Figure 10 shows impressively that all chlorpromazine analogues with LINGOsim > 0.4 differ by less than 2 log units in their predicted logP and logS values. Hence, LINGOsim is indeed correlated with structural similarities that give rise to molecular descriptors directly related to ADME properties.

CONCLUDING REMARKS

LINGO profiles provide a straightforward method to derive structure-related properties and to compute similarities directly from one-dimensional SMILES text strings, without the need to derive 2D or 3D structures. Models for logP and logS have performances that are comparable to the best models presently used and similarity evaluations show a very good discrimination between bioisosteric and random molecular pairs.

Concerning computational performance, LlogP calculated 247160 compounds of the NCI database in 2.05 s (120566 compounds per second) using a Pentium 4 Xeon 3.0 GHz and 2GB RAM. Using the same database and computer, clogP (version 4.71) takes 252.5 s (979 compounds per second). LINGOsim performs over 75000 comparisons per second in the same system.

LINGO representations share similarities to both fingerprint and fragment based methods. LINGO profiles could be considered a fingerprint of the molecule since they are derived from a unique SMILES representation. However, they are not mapped to a fixed length representation, and its constituents are not addressed by their position on a bit string. The limited number of possible 4 character LINGOs allows exhaustive direct inspection of all of them using text manipulation tools. On the other hand, the number of LINGOs that occur with statistically significant frequency in a database and can be used to derive property prediction models is large enough. In that sense, LINGOs can be considered also as fragments. In property prediction, LINGOs are not used in a binary mode, but the number of occurrences of each LINGO is used for increase accuracy. It should be kept in mind; however, that LINGOs are conceptually not descriptors of molecules but of their one-dimensional SMILES representations.

It is interesting that molecular properties derived from 3D structures can be derived directly from one-dimensional representations, and the parallelism with the transfer of information from a one-dimensional DNA sequence is obvious. LINGO decomposition makes use of the concept of holographic decomposition previously shown to successfully approximate the performance of 3D-QSAR using 2D structures. The versatility of LINGOs for both property prediction and similarity calculations may prove to be advantageous for addressing pharmacologically significant similarity searches. Ensembles of computed molecular properties have been used to define similarities that provided a significant enrichment on active compounds from a collection of drug-like molecules. It is expected that LINGO similarities may similarly lead to an improvement in the efficiency of virtual screening and compound design. Our groups are actively proceeding along these lines.

ACKNOWLEDGMENT

We thank the Spanish Ministries of Ciencia y Tecnologia and Educación y Ciencia for financial support (BIO2001-3115, PTR1995-0795-OP and GEN2003-20642-C09-04). D.V. is the recipient of a predoctoral fellowship and a short term travel grant from the Ministerio de Ciencia y Tecnología.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Diversity* **1999**, *4*, 1−22.
(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(3) Flower, D. R. On the properties of bit-string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(4) Wild, D. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures. alignment of molecular electrostatic potential fields with a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159−167.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(6) Butina, D.; Segall M.D; Frankcombe, K. Predicting ADME Properties in Silico: Methods and Models. *Drug Discov. Today* **2002**, *7*, S83−S88.
(7) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.
(8) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(9) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug. Discovery Des.* **1997**, *7/8*, 31−49.
(10) Schuffenhauer, A.; Gillet, V. J.; Willet, P. Similarity Searching of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptor. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.
(11) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
(12) UNITY Reference Manual; Tripos Inc; St. Louis. MO. 1995.
(13) Winkler, D. A.; Burden, F. R. Holographic QSAR of benzodiazepine. *Quant. Struct.-Act. Relat.* **1998**, *17*, 224−231.
(14) HQSAR Software. Tripos Associates. http://www.tripos.com/sciTech/inSilicoDisc/strActRelationship/hqsar.html.
(15) Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR. *ACS Symp. Ser.* **1999**, *719*, 212−225.
(16) Wild, D. J.; Blankey, 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.
(17) Sun, H. A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.
(18) Weininger, D. SMILES: a Chemical Language and Information System. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(19) 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.
(20) Ash, S.; Cline, M. A.; Homer, W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.
(21) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Analysis. Principles and Applications.* Umetrics Academy: Kinnelon, NJ, 2001.
(22) R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.
(23) The PHYSPROP database is available from Syracuse Research Corporation at URL http://www.syrres.com.
(24) The BIOSTER database is available from Synopsys Scientific Systems at URL http://www.synopsys.co.uk/.
(25) Meylan, W. M.; Howard P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(26) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. Prediction of *n*-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407−1421.
(27) Leo, A.; Calculating LogP from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(28) Klopman, G.; Li, J. Y.; Wang, S.; Dimayuga, M. Computer Automated LogP Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(29) Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: A Comparative Study. *Quant. Struct.-Act. Relat.* **1996**, *15*, 403, 409.
(30) CLOGP Daylight Chemical Information Systems, Santa Fe, NM, http://www.daylight.com.

(31) Petrauskas, A. A.; Kolovanov, E. A. ACD/LogP Method Description. *Perspect. Drug Discov. Design* **2000**, *19*, 1−19.
(32) Duban, M. E.; Bures, M. G.; DeLazzer, J.; Martin, Y. C. Virtual Screening of Molecular Properties: A comparison of LogP Calculators. In: B. Testa, H. v Waterbeend, G. Folkers, R. Guy (Eds.) *Pharmacokinetic Optimization in Drug Research*, Wiley-VCH: Zurich, 2001, pp 485−497.
(33) Jain, N.; Yalkowsky, S. H. Estimation of Aqueous Solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.
(34) Peterson, D. L.; Yalkowsky S. H. Comparison of two Methods for Predicting Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1531−1534.
(35) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(36) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 439−445.
(37) Klamt, A.; Eckert F. COSMO−RS: A Novel Way from Quantum Chemistry to Free Energy, Solubility, and General QSAR-Descriptors for Partitioning. In: Hoeltje, D. and Sippl, W. (Eds.), *Rational Approaches to Drug Design*, Prous Science Publ., Barcelona, 2001, pp 195−205.

CI0496797