Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names

Michael Thormann*[,a], David Vidal[b], Michael Almstetter[a] and Miquel Pons[*,b,c]

[a]*Origenis GmbH, Am Klopferspitz 19a, 82152 Martinsried Germany*

[b]*Laboratory of Biomolecular NMR, Institute for Research in Biomedicine, PCB. Josep Samitier 1-5, 08028 Barcelona, Spain*

[c]*Departament de Química Orgànica, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*

Abstract: The International Union of Pure and Applied Chemistry (IUPAC) was formed in 1919 by chemists from industry and academia [1]. Over nearly nine decades the Union has succeeded in fostering worldwide communications in the chemical sciences and in uniting chemistry - academic, industrial and government - in a common language. As one of the results of the Union, IUPAC names nowadays serve as a commonly agreed text representation of chemical structures in patents, publications and databases. In public databases of chemical compounds, like PubChem with more than 12 million entries, chemical structures are identified by default using their IUPAC names [2]. We report a very fast linguistic method to extract the implicit information contained in IUPAC names to statistically predict pharmacologically relevant properties. This provides an efficient annotation tool that can be used to assess the likelihood of a given compound as a drug candidate and renders the entire chemical literature a searchable database for virtual screening experiments and data mining.

INTRODUCTION

The International Union of Pure and Applied Chemistry (IUPAC) was formed in 1919 by chemists from industry and academia [1]. Over nearly nine decades the Union has succeeded in fostering worldwide communications in the chemical sciences and in uniting chemistry - academic, industrial and government - in a common language. As one of the results of the Union, IUPAC names nowadays serve as a commonly agreed text representation of chemical structures in patents, publications and databases. In public databases of chemical compounds, like PubChem with more than 12 million entries, chemical structures are identified by default using their IUPAC names [2]. We report a very fast linguistic method to extract the implicit information contained in IUPAC names to statistically predict pharmacologically relevant properties. This provides an efficient annotation tool that can be used to assess the likelihood of a given compound as a drug candidate and renders the entire chemical literature a searchable database for virtual screening experiments and data mining.

The recently developed LINGO method provides an efficient way to extract the implicit information contained in SMILES representations of chemical structures [3]. Applications to virtual screening of large databases and molecular similarity searching have been reported [4,5]. Analogous to SMILES, IUPAC names are strings of characters describing a chemical structure by using a set of standard rules to provide a unique molecular representation. LINGO profiles are the complete set of fixed length overlapping strings of characters derived from fragmentation of the textual representation of chemical structures. Each of the fixed length strings is called a LINGO. LINGO generation is illustrated in Fig. (**1**).

IUPAC naming rules were developed as a standard procedure to assign a unique identifier to a chemical compound based on its structure [1]. These rules have the effect of encapsulating and describing the structure of the compound in an extremely compact format, its name. In practice, every compound has a preferred IUPAC name and may have alternatives. Both contain the same descriptive information which therefore allow the development of property prediction models that directly relate names and properties.

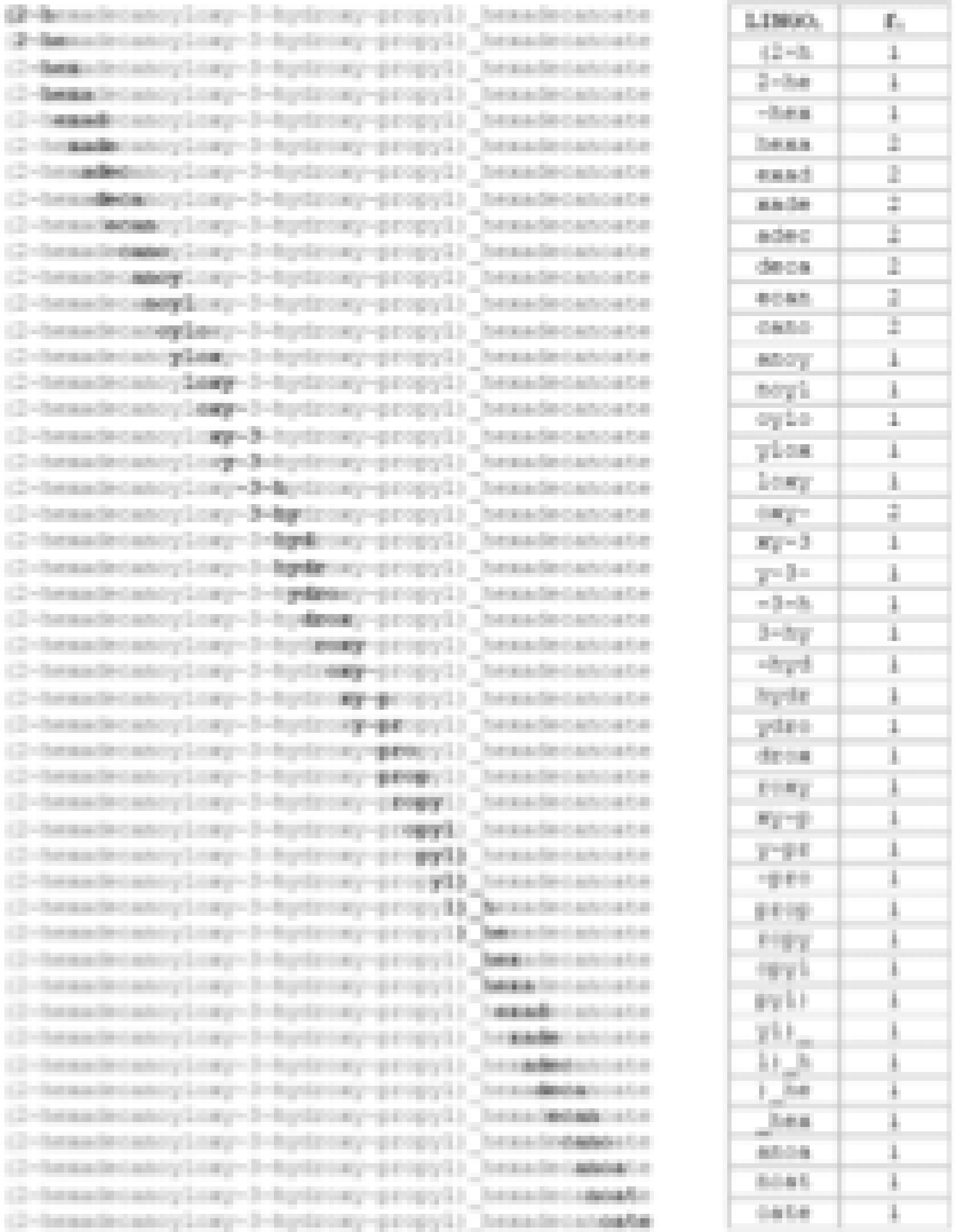
The conceptual connection between name, structure and property spaces is presented in Fig. (**2**). The quantitative structure property relationship is commonly termed QSPR. Here we show the direct, structure-free connection of name and property spaces and decided to term this quantitative name property relationship QNPR.

The analysis of text strings to predict complex properties is a well established practice in bioinformatics, where, for example, protein function can be statistically predicted from the analysis of DNA sequences. LINGO based analysis is related to alignment free sequence comparisons [6].

The application of the LINGO method to IUPAC names allows the computation of molecular properties such as logP, the logarithm of the partition coefficient between n-octanol and water, or logS, the logarithm of the intrinsic aqueous solubility. A set of experimentally measured values and the corresponding IUPAC names is used to train the model and to derive property and LINGO specific weights by statistical means.

*Address correspondence to these authors at the Origenis GmbH, Am Klopferspitz 19a, 82152 Martinsried, Germany;
E-mail: michael.thormann@origenis.de
Departament de Química Orgànica, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain; E-mail: mpons@ub.edu

1874-1363/07 2007 Bentham Science Publishers Ltd.



LINGO	f.
(2-h	1
2-he	1
-hex	1
hexa	2
exad	2
xade	2
adec	2
deca	2
ecan	2
cano	2
anoy	1
noyl	1
oylo	1
ylox	1
loxy	1
oxy-	2
xy-3	1
y-3-	1
-3-h	1
3-hy	1
-hyd	1
hydr	1
ydro	1
drox	1
roxy	1
xy-p	1
y-pr	1
-pro	1
prop	1
ropy	1
opyl	1
pyl)	1
yl)_	1
l)_h	1
)_he	1
_hex	1
anoa	1
noat	1
oate	1

Fig. (1). LINGO profile generation. The set of four-character LINGOs are shown in boldface, superimposed to the original name string to illustrate the LINGO generation process.

Fig. (**3a**) shows the accurate prediction of experimental logP (the logarithm of n-octanol/water partition coefficient) values using 4-character LINGOs derived from IUPAC names of 11,477 compounds present in the PHYSPROP database [7]. The model has a correlation coefficient of R^2=0.95 (RRMS=0.38) for the training set and of Q^2=0.84 (QRMS=0.75) for 10-fold cross-validation.

Many important properties can be derived accurately from 3D molecular structures. In order to test how much of the information contained in these 3D molecular structures is also accessible at IUPAC name space level, we computed a variety of molecular properties based on 3D structures [8-10] and used this information as training set for IUPAC-LINGO-QNPR models. We selected different types of surface areas, physicochemical measurable prop-

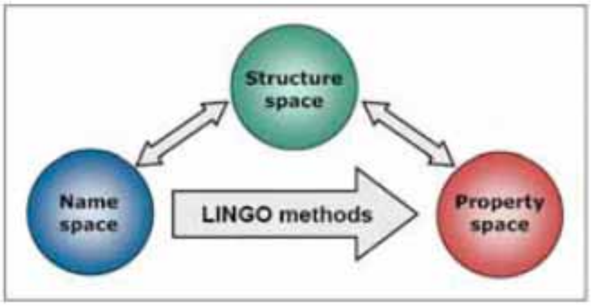


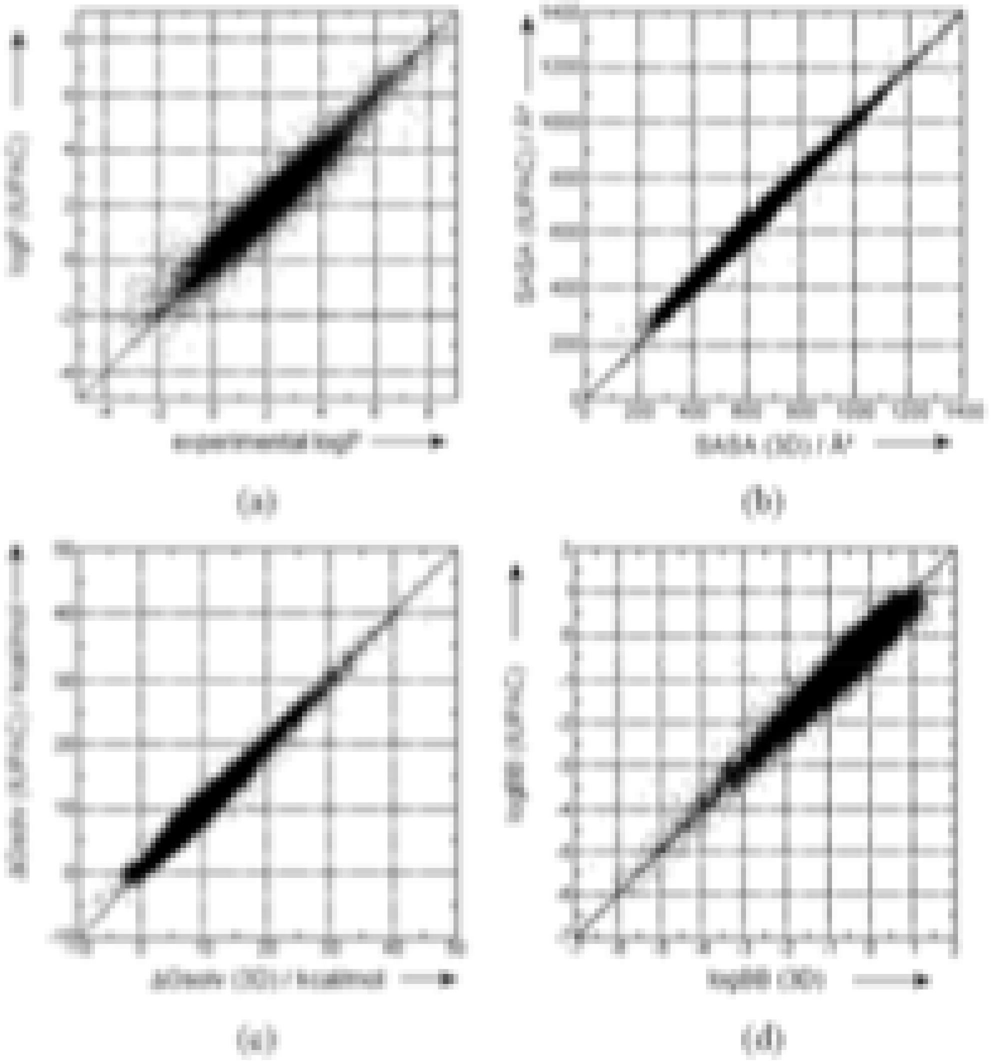
Fig. (2). Relationship between Name, Structure and Property spaces. While conventional methods in chemoinformatics use structure space as central element of chemical information, LINGO based methods provide direct access to such information at the name space level without invoking structure space.

erties such as solubility or partition coefficients of biphasic systems, and complex biopharmacological properties related to cell permeability or the interaction to specific receptors. The quality of the structure-based predictions depends on the complexity of the underlying models and of the experimental data in which these are based. In order to proof the principle of applicability of QNPR, we used QikProp to generate a large, diverse property data set.

IUPAC-LINGO-QNPR can thus be used with experimentally derived data or computed data derived from higher level, often more time-consuming prediction methods. Using the output of such programs as training set it is possible to develop LINGO-QNPR models that approximate the predictions of 3D-structure based methods directly from IUPAC names. The upper limit of the prediction capability of the LINGO-QNPR models will be obviously that of the parent models. It will, however, benefit from the much higher speed of the text based LINGO methods.

We extracted the IUPAC names and generated 3D structures for 445,085 molecules from PubChem (see Methods section), calculated 28 molecular properties with QikProp, and used them as pseudoexperimental values to construct IUPAC-LINGO-QNPR models. 20,000 IUPAC names with their 3D-structure derived properties were randomly selected from the PubChem database to train the IUPAC-LINGO-QNPR models. The resulting QNPR models were tested using 425,085 IUPAC names not used for the generation of the models. Fig. (**3b-d**) shows plots correlating the IUPAC-name based QNPR values and 3D-structure derived reference values for three of the calculated properties: surface accessible surface area (SASA), free energy of solvation in water (ΔGsolv), and logarithm of the blood-brain barrier partition coefficient.

It turns out that most molecular properties can be calculated accurately and directly from their IUPAC names using LINGO-QNPR models. Many of these are holistic properties related to molecular size, all kinds of surface areas, partition coefficients, free energies of solvation in



(a)

(b)

(c)

(d)

Fig. (3). **a**) Correlation between logP values calculated using IUPAC names and experimental ones. **b-d**) Correlation between calculated properties using IUPAC names and values obtained from the 3D structure, **b**) surface accessible surface area (SASA, R^2=0.99, RMS=22) , **c**) free energy of solvation in water (ΔGsolv, R^2=0.97, RMS=0.4), **d**) logarithm of the blood-brain barrier partition coefficient (logBB, R^2=0.95, RMS=0.3).

Fig. (4). a) Quantitative prediction of logP for a selected compound illustrating the IUPAC-LINGO-QNPR contributions. **b**) Qualitative visualisation of the predicted, backprojected molecular properties onto IUPAC names using the property-specific LINGO weights (top: computed n-octanol/water partition coefficient logP, mid: hydrophilic surface area FISA, bottom: carbon pi surface area PISA). Numbers denote IUPAC name-based predictions, numbers in brackets denote 3D-structure derived predictions.

different solvents or cell permeabilities. Interestingly, other properties for which a detailed description of the molecular graph would seem necessary, such as the number of hydrogen bond donors or acceptors or the number of rotatable bonds, are also well described. Indeed, most QNPR models have correlation coefficients R^2>.90 for the 20,000 compounds in the training set, and 18 models have cross-validated correlation coefficients Q^2>0.85 for >400,000 compounds in the test data set. Correlation plots and statistical parameters for the 28 properties studied are given as supplementary material. Not surprisingly, properties strongly dependent on the 3D structure and spatial electron distribution, such as dipole moments, electron affinities and ionization potentials, cannot be described well by LINGO-QNPR methods.

LINGO tools are orders of magnitude faster than the 3D structure based methods. QNPR property computations can be carried out at a speed of 120,000 molecules per second on a single CPU.

LINGO-QNPR models can be easily visualised and interpreted. The specific weights of IUPAC-LINGOs that are used to compute the molecular property can be backprojected character-wise onto the IUPAC name. In Fig. (**4b**) individual characters from IUPAC names were colour-coded according to their property- and LINGO-specific weights obtained from the QNPR statistics. The weight of each LINGO was evenly distributed over its four corresponding characters (Fig. **4a**). This procedure is completely automated and does not involve any chemical judgement. Reassuringly, a chemist can readily recognize that parts of the name have qualitatively expected contribution to different properties. For example, logP (a measure of polarity) correspond to parts of the name that denote the existence of polar or apolar groups. Likewise, the complementarity of the contributions to polar and non-polar fractions of the solvent accessible surface area becomes obvious in the IUPAC-LINGO-QNPR generated colour coding shown in Fig. (**4b**). However, using only chemical intuition and the IUPAC name it would not be possible for a scientist to perform quantitative property predictions. The visualisation capability of QNPR models is a great advantage to identify interactively parts of the molecule whose modification would have a strong effect on the predicted property.

As IUPAC names are used in all kind of scientific literature sources related to chemistry. IUPAC-LINGO-QNPR enables the direct access to these information sources and allows the assignment of property vectors to chemical entities without invoking chemical structure recognition and processing. IUPAC-LINGO-QNPR models allow a quick and accurate prediction of many important molecular properties direct out of articles, patents, compound databases and any kind of written reports.

Using LINGO-QNPR-based compound similarity, the different information sources can be extremely fast and automatically connected for a novel kind of knowledge base concept and are, thus, perfectly suited for chemical ontology applications [11]. We are actively pursuing this promising research line.

EXPERIMENTAL SECTION

LINGO-QNPR. Similar to quantitative structure property relationships (QSPR), the general equation used for quantitative name property relationships (QNPR) predicting molecular properties from a LINGO profile containing m different LINGOs is

$$LP_{\beta}\left(\xi\right) = w_{0,\beta} + \sum_{i=1}^{m} n_{i,\xi} \cdot w_{i,\beta} \qquad (1)$$

where $LP_{\beta}(\xi)$ is the LINGO-QNPR estimated value of property β for molecule $\xi, w_{0,\beta}$ is the offset, $w_{i,\beta}$ is the weight of LINGO i in the LINGO-QNPR model of property β and $n_{i,\xi}$ is the number of occurrences of LINGO i in the IUPAC name of molecule ξ. Weights are obtained by

Partial Least Squares (PLS) using a training set. Weights of LINGOs not present in the training set are set to zero. LINGO-QNPR can be used with any rule based descriptive naming system.

PLS analysis was performed by using the statistical package pls.pcr (version 0.2.4) in version 1.9.0 of R [12], a system for statistical computation and graphics, on a Linux PC. Latent variable PLS was performed using the kernel-PLS method with ten-fold cross-validation as implemented in pls.pcr.

Experimental values of logP were obtained from a 2004 release of PhysProp database containing 25,734 compounds. After removing compounds containing salts and elements other than C, N, S, O, H, P and halides, a total number of 12,831 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. IUPAC names were assigned using AutoNom 2000 [13] and compounds for which AutoNom failed were removed yielding a final set of 11,477 compounds.

The PubChem database (at May 31st 2005) was directly downloaded in 2D SDF format *via* FTP from the PubChem web page. Molecular structures including salts and the corresponding counter-ions were removed. Molecules containing elements other than C, N, S, O, H, P and halides, with a molecular weight higher than 750 Daltons, or with its IUPAC name not included, were also removed. The remaining database contained 445,085 compounds. The corresponding energy-optimized 3D conformations of the neutral forms with explicit hydrogen atoms, were generated using Moloc [14]. Finally, QikProp was used to calculate the properties for each molecule.

In order to prove that IUPAC-LINGO-QNPR models are indeed predictive, only 20,000 randomly selected IUPAC names were used for the training set while more than 425,000 IUPAC names were used as the test set. It should be noted that both name sets were generated using the same IUPAC name generator Lexichem [15] which introduces a sort of canonization and determines the model domain in the strict sense.

ACKNOWLEDGEMENTS

This work was partially supported by funds from the Spanish Ministerio de Educación y Ciencia to M.P. (BIO2004-5436 and GEN2003-20642-C09-04). D.V. was a recipient of a predoctoral fellowship from the Spanish Ministerio de Ciencia y Tecnología.

REFERENCES

[1] The International Union of Pure and Applied Chemistry, http://www.iupac.org.

[2] C. P. Austin, L. S. Brady, T. R. Insel, and F. S. Collins, “NIH Molecular libraries initiative”, *Science*, vol. 306, pp. 1138-1139 November 2004.

[3] D. Vidal, M. Thormann, and M. Pons, “LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities”, *J. Chem. Info. Model.* vol. 45, pp. 386-393, February 2005.

[4] D. Vidal, M. Thormann, and M. Pons, “A novel search engine for virtual screening of very large databases”, *J. Chem. Inf. Model.*, vol. 46, pp. 836-843, February 2006.

[5] J. A. Grant, J. A. Haigh, B. T. Pickup, A. Nicholls, and R. A. Sayle, “Lingos, Finite State Machines, and Fast Similarity Searching”, *J. Chem. Inf. Model.*, vol. 46, pp. 1912-1918, September 2006.

[6] S. Vinga, and J. Almeida, “Alignment-free sequence comparison - a review”, *Bioinformatics*, vol. 19, no. 4, pp. 513-523, 2003.

[7] Syracuse Research Corp. “The Physical Properties Database (PHYSPROP)”, May 2004, http://www.syrres.com.

[8] E. M. Duffy and W. L. Jorgensen, “Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol and water”, *J. Am. Chem. Soc.*, vol. 122, pp. 2878-2888, March 2000.

[9] W. L. Jorgensen, and E. M. Duffy “Prediction of drug solubility from Monte Carlo simulations”, *Bioorg. Med. Chem. Lett.*, vol. 10, pp. 1155-1158, June 2000.

[10] Schrödinger Inc., “QikProp, v. 2.1”, September 2003, http://www.schrodinger.com.

[11] J. B. L. Bard, and S. Y. Rhee, “Ontologies in biology: Design, applications and future challenges”, *Nat. Rev. Genet.*, vol 5, pp. 213-222, March 2004.

[12] R Development Core Team, "R: A language and environment for statistical computing”, April 2004, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org.

[13] Elsevier MDL, “AutoNom 2000”, May 2004, http://www.mdli.com.

[14] P. R. Gerber, “MOLOC - A molecular design software suite”, May 2004, http://www.moloc.ch.

[15] OpenEye Scientific Software, “Lexichem”, June 2005, http://www.eyesopen.com/products/toolkits/lexichem.html.

Received: August 31, 2007 Revised: October 05, 2007 Accepted: October 09, 2007