Encoding molecular structures as ranks of models: A new, secure way for sharing chemical data and development of ADME/T models

Igor V. Tetko

IBPC, Ukrainian Academy of Sciences, Kyiv, Ukraine and Institute for Bioinformatics, Munich, Germany

March 14th, ACS
Encoding molecular structures as SHUFFLED ranks of models: A new, secure way for sharing chemical data and development of ADME/T models

Igor V. Tetko

IBPC, Ukrainian Academy of Sciences, Kyiv, Ukraine and Institute for Bioinformatics, Munich, Germany
Structure-Property correlations

- Require representation (description) of the molecule in a format that can be used for machine learning methods, i.e. MLRA, neural network, PLS
- Two major systems: topological and 3D based

- Fragment-based indices
  - topological indices
  - E-state indices
- Quantum-chemical parameters
  - VolfSurf descriptors
  - Molecular shape parameters
Three scenario for structure decoding

• Can we identify the molecule provided we have it in our portfolio?  -- the most difficult scenario

• Can we do the same in knowledge that the molecule can be originated from one of several chemical series?

• Can we identify the molecule provided we do not know anything about it?  -- the practical scenario
Can we identify the molecule provided we have it in our portfolio? Topological indices.

- The ability to unambiguously identify a molecule is limited to information content of indices.

- If the indices contain sufficient information, the identification is possible.

- Information content of a molecule:

  - CCCCC -- 11111 (5 bits)   C -- 1 bit
  - C1CCCC1N -- 12111123 (11 bits)  1 -- 2 bits
  - N -- 3 bits
Information content of molecules in set of 12908 molecules (PHYSPROP database)

<table>
<thead>
<tr>
<th>Element</th>
<th>Frequency</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>78777</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>76965</td>
<td>2</td>
</tr>
<tr>
<td>)</td>
<td>42336</td>
<td>3</td>
</tr>
<tr>
<td>(</td>
<td>42336</td>
<td>4</td>
</tr>
<tr>
<td>O</td>
<td>29349</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>23648</td>
<td>6</td>
</tr>
<tr>
<td>=</td>
<td>20610</td>
<td>7</td>
</tr>
<tr>
<td>N</td>
<td>16156</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>12658</td>
<td>9</td>
</tr>
</tbody>
</table>

not optimal -- Huffman, arithmetic coding, other algorithms: gz, zip -- 3.5 bits/atom, bzip2 -- 2.9 bits/atom
Information content of a molecule

- 30 -- 40 atoms -- 90 -- 110 bits
- 1 double value -- 32 bits, 3 -- 4 topological indices potentially contains sufficient information to unambiguously decode molecule with 40 atoms!
- In reality a larger number of indices can be required because of rounding effects, non-optimal storage of information
- Thus, the encoding of molecules using topological indices can be insecure.
When reverse engineering is impossible? A practical scenario.

- ALOGPS program:
  75 indices per molecule for logP
  33 indices per molecule for logS

- We use decreased resolution of data, i.e. to just 3 significant digits per index (7-10 bits instead of 32 bits)
- Additional bits are coming from range ~ 11 bits per index => 10-12 indices per molecule with 40 atoms

**The information encoded in the indices could be (theoretically) adequate to decode the molecules with < 50 heavy atoms.**

But, this can be too pessimistic conclusion. The theoretical possibility to decode does not propose a way how this can be done!
ALOGPS 2.1

• LogP: 75 input variables corresponding to electronic and topological properties of atoms (E-state indices), 12908 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.35, MAE=0.26, n=76 outliers (>1.5 log units)

• LogS: 33 input E-state indices, 1291 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.49, MAE=0.35, n=18 outliers (>1.5 log units)

• Tetko, Tanchuk & Villa, JCICS, 2001, 41, 1407-1421.
• Tetko, Tanchuk, Kasheva & Villa, JCICS, 2001, 41, 1488-1493.
• Tetko & Tanchuk, JCICS, 2002, 42, 1136-1145.
Welcome to the ALOGPS 2.1 program!

Provide CAS RN or SMILES of a molecule and press the "submit" button.

Clcccccl

Upload a file with molecule[s] in 48 formats

Benzene

CAS RN 71-43-2 formula C6H6 MW 78.11

SMILES clcccccl

logP (exp): 2.13

AOGPS 2.03 <-0.10>

IA logP

CLOGP 2.14 <+0.01>

milogP 2.13 <0.00>

KOWMN 1.99 <-0.14>

XLOGP 2.02 <-0.11>

PhysProp reference

Sangster reference

User's LogP-library (upload library)

The calculated results are available.

For more information click on a keyword or a calculated result or contact Igor V. Tetko.

You can also download a stand-alone version of the program.
Artificial Feed-Forward Back-propagation Neural Network (FBNN)
Early Stopping Over Ensemble (ESE)
ASNN: an example correction

---

1-kNN correction

*Morphinan-3-ol, 17-methyl-*

Calculated logP=3.65, δ=+0.54

\[ 3.65 - 0.76 = 2.89 \ (δ=+0.22) \]

*Levallorphan*

Calculated logP=4.24, δ=+0.76

\[ 4.24 - 0.54 = 3.70 \ (δ=+0.22) \]

-- both molecules are the nearest neighbors, r²=0.47, in space of residuals!
Associative Neural Network (ASNN)

A prediction of case $i$: $[x_i] \cdot [\text{ANNE}]_M = [z_i] = \begin{bmatrix} z_1^i \\ \vdots \\ z_k^i \\ \vdots \\ z_M^i \end{bmatrix}$

Ensemble approach:

$$\bar{z}_i = \frac{1}{M} \sum_{k=1,M} z_k^i$$

Pearson’s (Spearman) correlation coefficient $r_{ij}=R(z_i,z_j)>0$ in space of residuals

$$\bar{z}'_i = \bar{z}_i + \frac{1}{k} \sum_{j \in N_k(x_i)} \left( y_j - \bar{z}_j \right)$$

$\lll$ ASNN bias correction

The correction of neural network ensemble value is performed using errors (biases) calculated for the neighbor cases of analyzed case $x_i$ detected in space of neural network models.
Prediction Space of the model does not cover the “in house” compounds

Each new molecule is encoded as rank of models
Encoding of a molecule as rank of models

\[ \Delta \log P = \log P_{\text{exp}} - \log P_{\text{calc}} \]

- 64 values, ranks of NN

<table>
<thead>
<tr>
<th></th>
<th>0.89</th>
<th>0.88</th>
<th>0.86</th>
<th>0.90</th>
<th>0.91</th>
<th>0.85</th>
<th>.885</th>
<th>0.83</th>
<th>0.95</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.09</th>
<th>0.08</th>
<th>0.06</th>
<th>0.10</th>
<th>0.11</th>
<th>0.05</th>
<th>.085</th>
<th>0.03</th>
<th>0.15</th>
<th>0.14</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.59</th>
<th>0.38</th>
<th>0.26</th>
<th>0.60</th>
<th>0.71</th>
<th>0.15</th>
<th>.485</th>
<th>0.03</th>
<th>0.95</th>
<th>0.84</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Millions of solutions provide the same ranks of NN responses --> no way to decode -- previous name of the paper, but...
How selective is rank coding?

- 8x64 = 512 bits (comparable to MDL keys)
- 126 out of 121281 Asiprox (0.1%)
- 12 out of 12908 PHYSPROP (0.1%)
ALOGPS: Extrapolation vs Interpolation


ALOGPS logP (blind): $\text{MAE} = 1.27$, $\text{RMSE}=1.63$
ALOGPS logP (LIBRARY): $\text{MAE} = 0.49$, $\text{RMSE}=0.70$
Analysis of Pfizer data

ALOGPS prediction for ElogD set of 17,861 compounds

ALOGPS "as is" → ALOGPS LIBRARY

- Pallas PrologD: $MAE = 1.06$, $RMSE=1.41$
- ACDlogD (v. 7.19): $MAE = 0.97$, $RMSE=1.32$
- ALOGPS: $MAE = 0.92$, $RMSE=1.17$
- ALOGPS LIBRARY: $MAE = 0.43$, $RMSE=0.64$

PHYSPROP data set

Total: 12908

star set

CLOGP
9429

XLOGP
1873

nova set

3479

training “nova” --> prediction
star set
Prediction performance as function of similarity in space of models of “star” set

Blind prediction

max correlation coefficient of a test compound to training set compounds

MAE=0.43

LIBRARY mode

max correlation coefficient of a test compound to LIBRARY compounds

MAE=0.28 (0.26)
Reliability of new compound predictions

- NCI, 250,000
  - >0.7
  - ~0.6
  - ~0.5
  - ~0.4
  - <0.3

- http://asinx.com, 120,000
  - 0.2-0.4
  - 0.4-0.6
  - 0.6-0.8

- http://ambinter.com, 650,000
  - 0-0.2

PHYSPROP
Reliability of new compound predictions

- **NCI, 250,000**
  - http://ambinter.com
- **http://asinex.com**, 120,000
- **http://ambinter.com**, 650,000

- **Aurora data**
  - 0.8-1: <0.3
  - 0.6-0.8: ~0.4
  - 0.4-0.6: ~0.5
  - 0.2-0.4: ~0.6
  - 0-0.2: >0.7

- **PHYSPROP**
  - 0.8-1: <0.3
  - 0.6-0.8: ~0.4
  - 0.4-0.6: ~0.5
  - 0.2-0.4: ~0.6
  - 0-0.2: >0.7
Is identification possible?

PHYSPROP -- Asinex study

\[ r^2 = 0.97 \]

\[ r^2 = 0.85 \]

\[ r^2 = 0.82 \]
Is identification possible?

PHYSPROP -- Asinex study
Is identification possible?
PHYSPROP -- Asinex study

0.61

0.60

0.60
Securing the data -- shuffling ranks!

Shuffle $r^2=0.8$

Shuffle $r^2=0.6$
Rank shuffling

- Shuffled rank molecule is less similar to itself than the molecules from the other series will be pick-upped --> secure encoding
- Different molecules will have different distribution of neighbors as function of similarity=> lower level of security (e.g. 1 in $10^5$, 1 in $10^6$) can be determined individually for each single compound using an external library (e.g. complete enumeration, compilation of public libraries)
- Everything can be done in completely automatic mode
Possible approaches

Raw topological indices

• Development of new global models, after the development the data can be discarded
• There is a theoretical possibility to decode the structure, particular for smaller number of atoms in a molecule (not clear if such algorithm can be realized)
• One-to-one contract may be required...

Rank of models

• Allows to incorporate explicit structural parameters as feature elements
• No limitation on the number of indices
• The quality of local correction is comparable to retraining
• Very appealing to share on the WWW
• Security can be controlled by shuffling but will deteriorate prediction quality of model

Development of new models

• Develop new models in-house
• Provide them to be included in the set of models
• Predict new data using an ensemble of diverse models (ASNN in space of models of different companies)
• A complete set of automated tools to develop them can be provided
Acknowledgement

Part of this presentation was done thanks to Virtual Computational Chemistry Laboratory INTAS-INFO 00-0363 project (http://www.vcclab.org).

I thank Prof Hugo Kubinyi, Drs Pierre Bruneau and Gennadiy Poda for collaboration and Prof. Tudor Oprea for inviting me to participate in this conference.

Thank you for your attention!