SIMILARITY ASSESSMENTS IN DRUG DISCOVERY

D. Filimonov, V. Poroikov

Institute of Biomedical Chemistry, Moscow, Russia

E-mail: dmitry.filimonov@ibmc.msk.ru  vladimir.poroikov@ibmc.msk.ru
Acknowledgements

Financial Support
The study was supported in the framework of the Russian Foundation for Basic Research grant No. 20-04-60285.

Alexey Lagunin,
Oleg Gomazkov,
Alexander Dmitriev,
Anastasia Rudik,
Boris Sobolev,
Dmitry Druzhilovskiy,
Olga Tarasova,
Pavel Pogodin,
Sergey Ivanov,
Tatyana Gloriozova,
Leonid Stolbov,
Dmitry Karasev,
Polina Savosina,
Nikita Ionov,
Nadezhda Biziukova,
Vladislav Sukhachev
Cortellis Drug Discovery Intelligence (CDDI) database, structural formula of Perindopril as a query. Structural similarity search results, only the most similar of 15 "similar" structures are presented.
Hierarchical clustering using similarity

Let’s start with $N$ clusters – each object is a separate cluster.

Next, we consistently combine clusters for which the distance between them is minimal.

The result of clustering depends on the selected method of determining the distance between clusters.

The distance between the "centers of mass":

$$D_{ij} = d \left( \frac{1}{n_i} \sum_{x \in G_i} x, \frac{1}{n_j} \sum_{x \in G_j} x \right)$$

where $D_{ij}$ is the distance between two clusters, $G_i$ и $G_j$.

Hausdorff distance:

$$D_{ij} = \min_{i,j} \max_{x',x''} \left( d(x', x'') \mid x' \in G_i, x'' \in G_j \right)$$

- two clusters for which the distance between the most distant points is minimal are combined into one cluster.
Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison

Figure 5. Pool of 1000 compounds predicted to inhibit the 3CL protease of the novel SARS-CoV-2, (red) mapped against the SARS-CoV compounds (blue), within the DrugBank reference frame.

The hypothesis that structurally similar compounds exhibit similar biological effects or some other properties is taken as an axiom.

Tyrosine kinase ABL inhibitors, reported pKi values. SAR continuity is observed where gradually changes in compound structure (tracked by horizontal arrows) are accompanied by moderate activity alterations.

Muratov EN et al. (2020) QSAR without borders, Chemical Society Reviews, 49, 3525-3564. https://doi.org/10.1039/d0cs00098a
Tyrosine kinase ABL inhibitors, reported pKi values. The inhibitors display SAR discontinuity - small structural modifications lead to large changes in activity. Vertical arrows indicate the formation of pairwise activity cliffs.

Muratov EN et al. (2020) QSAR without borders, Chemical Society Reviews, 49, 3525-3564. https://doi.org/10.1039/d0cs00098a
Molecular similarity, as a paradigm, contains many implicit and explicit assumptions.

One does not know a priori which properties of the molecular structure are essential for its biological activity; therefore, the description of the structure can be only heuristic.

The selection of molecular descriptors and the estimation of molecular similarity based on this selection crucially determine the final result of the study.

However, for novel pharmacological targets (like SARS-CoV-2 coronavirus proteins), when only limited number of antiviral agents that may be used as a “query” are known, similarity assessment is the method-of-the-choice.
Local Correspondence Concept
Sequence Local Similarity. Frame 20, shift from 0 to 17

AANRDPSQFPDPHRFDVTRDTRGHLSGFGQGIHF CMGRPLAKLEGEVA 2
ANRDPFSQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVAL 1
NRDPSQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALR 1
RDPSQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRA 0
DPSQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRAL 1
PSQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALF 2
SQFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFG 1
QFPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGR 1
FPDPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRP 0
DPHRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFP 1
PHRFVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPAL 0

<table>
<thead>
<tr>
<th>HRFDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALS</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDFVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSL</td>
<td>0</td>
</tr>
<tr>
<td>FDVTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLG</td>
<td>3</td>
</tr>
<tr>
<td>DTVTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSGI</td>
<td>1</td>
</tr>
<tr>
<td>VTRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGI</td>
<td>1</td>
</tr>
<tr>
<td>TRDTRGHLSGFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGID</td>
<td>2</td>
</tr>
</tbody>
</table>

GTAINKPLSEKMLFGMKRRCKIIEVLKWEIFLFLAILLQQLEFSV 9

R_q = 9

http://www.biomedcentral.com/1471-2105/11/313
Sequence Local Similarity. It is descriptor itself!

Descriptor is defined as the similarity value \( S_{ik} \) for position \( i \) of sequence under study and experimentally annotated sequence \( k \).
Local Correspondence Concept. Neighborhoods of atoms descriptors

The most biological activities of organic compounds are the result of molecular recognition, which in turn depends on the correspondence between particular atoms of the ligand and the target.

\[
M_i = V_i + V_i g M = V_i + V_i g(M_1 + M_2 + \ldots + M_m)
\]

All descriptors are based on the concept of atoms’ of molecule description subject to the neighborhood of them:

- **MNA** - Multilevel Neighborhoods of Atoms
- **LMNA** - Labeled Multilevel Neighborhoods of Atoms
- **QNA** - Quantitative Neighborhoods of Atoms

Multilevel neighborhoods of atoms descriptors – MNA

**MNA/2**

\[
\begin{align*}
&\text{C(C(CC-H)C(CC-C)-H(C))} \\
&\text{C(C(CC-H)C(CN-H)-H(C))} \\
&\text{C(C(CC-H)C(CN-H)-C(C-O-O))} \\
&\text{C(CCC-H)N(CC)-H(C)} \\
&\text{N(C(CN-H)C(CN-H))} \\
&\text{-H(CCC-H))} \\
&\text{-H(CCN-H))} \\
&\text{-H(-O(-H-C))} \\
&\text{-C(CCC-O(-H-C)-O(-C))} \\
&\text{-O(-H(-O)-C(C-O-O))} \\
&\text{-O(-C(C-O-O))}
\end{align*}
\]

PASS (Prediction of Activity Spectra for Substances) software

Theoretical basis:
The Local Correspondence Concept
The original Bayesian Classifier
High-quality training dataset
Knowledge base "structure-mechanism-effect"

1992
The first version of PASS: 9,314 BAS, 114 A, 0.85 APA

1998
New version of PASS: 31,000 BAS, 435 A, 0.94 APA

1999
The first PASS Online program on the Internet

2009
New version of PASS: 206,000 BAS, 3,750 A, 0.95 APA

2022
New version of PASS: 1,614,066 BAS, 8,387 A, 0.94 APA

2021: 20 web services
PASS Online Statistics:
1 137 046 predictions;
>1 200 publications;
30 758 users from 104 countries of the world

Russia (27,6%)
India (12,4%)
USA (8%)
China (5,3%)

BAS is the number of substances in the training set; A is the number of predicted activity types; APA is the average prediction accuracy
Quantitative neighborhoods of atoms descriptors – QNA

In fact, interatomic and intermolecular forces are electrical in nature according to the Hellman-Feynman theorem.


\[
P_i = B_i \sum_k (\exp(-\frac{1}{2}C))_{ik} B_k
\]

\[
Q_i = B_i \sum_k (\exp(-\frac{1}{2}C))_{ik} B_k A_k
\]

\[
A = \frac{1}{2}(IP + EA)
\]

\[
B = (IP - EA)^{-\frac{1}{2}}
\]

\[
C\text{ is the connectivity matrix of a molecule,}
\]

\[
IP\text{ is the first ionization potential,}
\]

\[
EA\text{ is the electron affinity.}
\]

Normalization:

\[
P' = (P - E(P))/D(P)
\]

\[
Q'' = (Q - E(Q))/D(Q)
\]

\[
Q' = (\mu P' - Q'')/D(PQ)
\]

\[
E(P') = 0, \quad D(P') = 1
\]

\[
E(Q') = 0, \quad D(Q') = 1
\]

\[
\text{Cov}(P'Q') = 0
\]


QNA descriptors’ space

Nicotinic Acid

Acetylsalicylic acid

Sulfathiazole

QNA descriptors’ space
Similarity estimation using the QNA descriptors:

\[ F(A, B) = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)} \]

\[ n(A \cap B) = \frac{1}{2} \left( \sum_A \max_{b \in B} [s_{ab}] + \sum_B \max_{a \in A} [s_{ba}] \right) \]

where \( s_{ab} \) and \( s_{ba} \) are the pairwise similarities of QNA descriptor of atom \( a \) in a molecule \( A \) and QNA descriptor of atom \( b \) in a molecule \( B \):

\[ s_{ab} = \exp \left( -4N_B \left( (P_a - P_b)^2 + (Q_a - Q_b)^2 \right) \right) \]

\[ s_{ba} = \exp \left( -4N_A \left( (P_a - P_b)^2 + (Q_a - Q_b)^2 \right) \right) \]

where \( P_a \) and \( Q_a \) are values of QNA descriptor of atom \( a \) in a molecule \( A \), \( P_b \) and \( Q_b \) are values of QNA descriptor of atom \( b \) in a molecule \( B \).
Similarity of compounds – HIV-1 integrase inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.
Similarity of compounds – HIV-1 protease inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.
Similarity of compounds – HIV-1 reverse transcriptase inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.
**Big Chemical Data**

Many different sources are used – thus, there is a problem of duplicating the molecule structures.

When the number of structures in data set, $N$, is many millions, the pairwise comparison of structures requires an inaccessible resources to perform $N(N - 1)/2$ comparisons.

To solve such tasks, we use the $Q$ index which is the sum of the $Q$ values of QNA descriptors.

The resulting array of $N$ real values can be easily sorted, and then you need to compare only those structures whose $Q$ indexes differ by less than $1E-9$. 
The MetaTox 2.0 web portal provides an opportunity for a comprehensive analysis of the biological activity profiles of existing and developing drugs, taking into account their metabolism in the human.

http://www.way2drug.com/metatox
Three schemes of Diclofenac metabolism from different sources and structures of Diclofenac-like compounds with similarity estimates for MNA and QNA descriptors (top right).

More than 2000 biologically active compounds with known metabolic pathways, which were extracted from DrugBank, MetXBIODB and ChEMBL databases

http://www.way2drug.com/metatox
World Wide Approved Drugs (WWAD) contains information on more than 4,000 medicines approved by regulatory authorities in various countries.

For the initial compound, a table is given containing the structure of the drug and two similarity values calculated using the MNA and QNA descriptors.

http://www.way2drug.com/metatox
The search for precursor compounds by the similarity method for Diclofenac acyl glucuronide.

The method of evaluating possible precursors of compounds is to search for similar metabolites among the known ones.

http://www.way2drug.com/retro
Our selection of “hits” by virtual screening in the JEDI Grand Challenge

1 082 000 000 molecules

SimQNA → SimMNA

12 230 - 3CLpro
25 812 - PLpro
3 584 - RdRp
883 - TMPRSS2

42 509 hits

7 148 - 3CLpro
25 782 - PLpro
3 544 - RdRp
882 - TMPRSS2

6 655 - 3CLpro
6 981 - PLpro
3 387 - RdRp

Dock 6.5 & AutoDock Vina

*IC50 < 10μM

45 - 3CLpro
38 - PLpro

https://discord.com/channels/694851986042126366/694851987208011818
Results of synthesis and biological testing of 820 compounds (36 from us) in the framework of JEDI

### Details of hits

<table>
<thead>
<tr>
<th>Name</th>
<th>SMILES</th>
<th>Teams</th>
<th>List position</th>
<th>Predicted Target</th>
<th>Exp Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc2_247951907</td>
<td>CC(C)(O)CNC(=O)(C)(O)C2CCC(O)CCl(O)C(O)CClCC1</td>
<td>kyken</td>
<td>9147</td>
<td>Acc2</td>
<td>S</td>
</tr>
<tr>
<td>5_dmd55631224</td>
<td>CCOCN(C)(O)C2CCC(O)CCCl(O)C(O)CClCC1</td>
<td>deumpol</td>
<td>56</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>acc2_84865828</td>
<td>CN1C(C(O)C2CCC(O)CCl(O)C(O)CClCC1)C1</td>
<td>cytk</td>
<td>6844</td>
<td>Acc2</td>
<td>S</td>
</tr>
<tr>
<td>msps_15734005</td>
<td>CN1C(C)(O)C2CCC(O)CCl(O)C(O)CClCC1</td>
<td>JXU</td>
<td>212.8 &amp; 115.7</td>
<td>Nsp5</td>
<td>Nsp5</td>
</tr>
<tr>
<td>msp12s4614385</td>
<td>O=CNC1C2CC(C)C3CC(C)C2C1</td>
<td>covid19ddc</td>
<td>46</td>
<td>Nsp12</td>
<td>Nsp5</td>
</tr>
<tr>
<td>acc2_3484105</td>
<td>CN1C2CC(C)C3CC(C)C2C1 =O</td>
<td>cytk</td>
<td>34</td>
<td>Nsp12</td>
<td>Nsp5</td>
</tr>
<tr>
<td>msp12s8452892</td>
<td>O=CNC1C2CC(C)C3CC(C)C2C1</td>
<td>virtuflow</td>
<td>84.5 &amp; 84</td>
<td>Nsp12, TMPRSS2</td>
<td>Nsp5</td>
</tr>
<tr>
<td>trmrss_228210</td>
<td>O=CNC1C2CC(C)C3CC(C)C2C1</td>
<td>cermm</td>
<td>93</td>
<td>TMPRSS2</td>
<td>Nsp5</td>
</tr>
<tr>
<td>trmrss_6575826</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>aLsscience</td>
<td>56.8 &amp; 167</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp1_10432016</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>104</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_39492161</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>394</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_12425240</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>93.8 &amp; 124</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_54289358</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>41</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_48491882</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>849</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_53469020</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>34</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_56272940</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>52</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_48385913</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU, awinter</td>
<td>1011.2 &amp; 2438</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_35872038</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>4358</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_55272679</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>57</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_54641236</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>46</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_66580668</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>264</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_5489691</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>JXU</td>
<td>45.8 &amp; 48</td>
<td>Nsp5</td>
<td></td>
</tr>
<tr>
<td>msp5_34622522</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>way2drug</td>
<td>63.6 &amp; 64</td>
<td>Nsp3</td>
<td>Nsp3</td>
</tr>
<tr>
<td>msp5_31533607</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>kyken</td>
<td>515</td>
<td>Nsp3</td>
<td>Nsp3, macromdomain</td>
</tr>
<tr>
<td>msp5_31543051</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>kyken</td>
<td>1</td>
<td>Nsp3</td>
<td>Nsp3, macromdomain</td>
</tr>
<tr>
<td>msp12s5749081</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>covid19ddc</td>
<td>57</td>
<td>Nsp12</td>
<td>Nsp12</td>
</tr>
<tr>
<td>msp12s95277234</td>
<td>CN1C2CC(C)C3CC(C)C2C1</td>
<td>mlnch</td>
<td>4952</td>
<td>Nsp12</td>
<td>Nsp12</td>
</tr>
</tbody>
</table>

**PASS_MNA_COUNT:** 40  
**PASS_MNA_NEW_COUNT:** 0  
**PASS_RESULT_COUNT:** 4 of 8 Possible Activities at Pa>Pi

**PASS_ACTIVITY_SPECTRUM**

- **Pa:** 0.908 0.001 Papain-like Protease (SARS-CoV-2) Inhibitors
- **Pi:** 0.341 0.188 Spike Glycoprotein (S) (SARS-CoV-2)/ACE2 Interaction inhibitors
- **0.365 0.249 SARS-CoV-2 Infection Reduction in Cell-Based Assay**
- **0.136 0.090 3C-Like Protease (SARS-CoV) Inhibitors**

![5-amino-N-[1-(3,4-dimethoxyphenyl) ethyl]-2-methylbenzamide](attachment:image.png)
Informational-Computational System AntiCOVID-19
Drug repurposing: Molnupiravir as a query

How to do that?

- Identify the probable analogs of the antiviral agent among the more than four thousand world-wide approved drugs. [see more information]
- Draw your molecule with Marvin JS editor.
  - You may use the Marvin JS editor for drawing your molecule. Marvin JS runs in any HTML5-capable browser without any plugin. For IE, version 9 or above is needed for Marvin JS due to use of HTML5.
- Find structural analogs among over 4,000 approved drugs.
  - Click "Generate Hypothesis" button to identify new probability application among the world wide approved drugs. [see the example]
### Structural analogs of Molnupiravir among the launched drugs

<table>
<thead>
<tr>
<th>Structure</th>
<th>Name</th>
<th>Indication</th>
<th>Activity</th>
<th>Approval</th>
<th>Target</th>
<th>Similarity MNA</th>
<th>Similarity QNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image]</td>
<td>cytarabine</td>
<td>Myeloid Leukemia</td>
<td>Antineoplastic (my...)</td>
<td>1969-06-17 FDA</td>
<td>DNA polymerase...</td>
<td>0.576</td>
<td>0.602</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lymphoid leukemia</td>
<td>CYP3A4 substrate</td>
<td></td>
<td>Cytidine deaminase...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>enocitabine</td>
<td>Acute leukemia</td>
<td>Antileukemic</td>
<td>2009-09-01 PMDA</td>
<td></td>
<td>0.529</td>
<td>0.452</td>
</tr>
</tbody>
</table>
Thank you for your attention! We are open for collaboration.