RECOGNITION OF PROTEIN FUNCTION USING THE LOCAL SIMILARITY

Kirill E. Alexandrov
Dmitry A. Filimonov
Boris N. Sobolev
Vladimir V. Poroikov

Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Moscow, Russia
Agenda

1. History of Problem
2. Sequence Local Similarity
3. Algorithm of Similarity Calculation
4. Local Similarity Approach Paradigm
5. Algorithm of Protein Function Recognition
6. Prediction Accuracy Estimation
7. Results of Local Similarity Approach Evaluation
8. Acknowledgements
The central dogma of SAR/QSAR/QSPR:

Property = Function ( Structure )

Continuity hypothesis:

the difference of structures is less,
the difference of properties is less

\[ y_{\text{pred}} = x_0 + \sum_i x_i F_i(S) \]

\( F_i(S) = \text{LogP, ..., (LogP)^2, ...} \) – traditional QSAR
\( F_i(S) = \text{Sim}(S,S_i) \) – similarity based QSAR

MLR – multiple linear regression
PLS – projections to latent structures
ANN – artificial neural network
SVM – support vector machine
The local similarity principle

Tripos' patented Comparative Molecular Field Analysis (CoMFA) has been used as the method of choice in hundreds of published QSAR studies.
Neighborhoods of atoms descriptors

MOLECULAR BIOLOGY
QUANTUM CHEMISTRY
QUANTUM FIELD THEORY:

\[ M = V + V_g M = V + V_g V + V_g V_g V + V_g V_g V_g + \ldots \]
\[ M_i = V_i + V_{ig} M = V_i + V_{ig} (M_1 + M_2 + \ldots + M_m) \]

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

MNA - multilevel neighborhoods of atoms
RMNA - reaction multilevel neighborhoods of atoms
QNA - quantitative neighborhoods of atoms
FNA - fuzzy neighborhoods of atoms

Multilevel neighborhoods of atoms descriptors – MNA

MNA/0: C

MNA/1: C(CN-H)

MNA/2: C(C(CC-H)N(CC)-H(C))

Multilevel neighborhoods of atoms descriptors – MNA

MNA/2

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

Prediction of activity spectra for organic compounds

According to the Bayes formula the probability $P(A|S)$ of that compound $S$ has activity $A$ is equal to:

$$P(A|S) = P(S|A)\cdot P(A)/P(S)$$

Let the descriptors of organic compound $D_1, ..., D_m$ are mutually independent, then:

$$P(S|A) = P(D_1, ..., D_m|A) = \prod_i P(D_i|A)$$

$P(A)$ and $P(A|D_i)$ are calculated as sums over all organic compounds of the training set:

Quatitative neighborhoods of atoms descriptors – QNA

\[ Q_i = a_i \sum_k [g(C)]_{ik} b_k \]

\( a_i \) and \( b_k \) are parameters of atoms \( i \) and \( k \)
\( g(C) \) is function of the connectivity matrix \( C \)

\[ P_i = B_i^{-\gamma_2} \sum_k (\text{Exp}(-\gamma_2 C))_{ik} B_k^{-\gamma_2} \]

\[ Q_i = B_i^{-\gamma_2} \sum_k (\text{Exp}(-\gamma_2 C))_{ik} B_k^{-\gamma_2} A_k \]

\( A = \frac{1}{2}(\text{IP} + \text{EA}), \quad B = \text{IP} - \text{EA}, \)

\( \text{IP} \) is the first ionization potential,
\( \text{EA} \) is the electron affinity.

Quatitative neighborhoods of atoms descriptors – QNA

ChemNavigator DataBase in QNA Space
976,545,026 QNA descriptors of 24,621,668 molecules

Initial QNA Space
Normalized QNA Space
Quatitative neighborhoods of atoms descriptors – QNA

Nicotinic Acid

Aspirin

Sulfathiazole
GUSAR – QNA based prediction of quantitative properties of organic compounds

\[ \text{property} = a + \sum_i \sum_j \text{feature}_i \text{feature}_j \]

GUSAR interface with prediction results:
- \( N = 65 \)
- \( R^2 = 0.885 \)
- \( F = 41.847 \)
- \( SD = 0.422 \)
- \( V = 10.000 \)
- \( Q^2 = 0.849 \)
GUSAR – QNA based prediction
of quantitative properties of organic compounds

CDK2 inhibitors

DHFR inhibitors

ACE inhibitors

Vibrio fischeri

Chlorella vulgaris

Tetrahymena pyriformis
GUSAR – QNA based prediction of quantitative properties of organic compounds

-0.10 -0.05 0.00 0.05 0.10 0.15 0.20

2D Cerius2
3D Cerius2
CoMSIA
CoMFA
EVA
HQSAR
GFA
MLR
PLS

delta R2 test
delta Q2
delta R2
OK. But, how local similarity can be used for recognition of protein function?...
Homology-derived annotation based on the pairwise sequence alignment was a general way to predict the protein function for a long time.
Sequence Local Similarity.
Frame 20, shift from -8 to +8

AANRDP SQFPDPHRF DVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALE 2
ANRDP SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALE 1
NRDPS SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1
RDPS SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 0
DP SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1
PS SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 2
SQFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1
QFPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1
FPDPHRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 2
PDPHRDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 0
DPHFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1
PHRDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 1

The best match

HRFDVTRDTRG HLSFGQGIHF CMGRPLAKLEG EVALEVR 9

Ri = 9

Query sequence

GTAIN KPLSEKMMLFG MGK RRC IGEVLAEFL FLAIL LLQ LEFSV 9
Sequence Local Similarity.
Algorithm of Similarity Calculation

\[ R_i = \text{primary similarity value} \]
\[ S_i = \text{local similarity value for position } i \text{ in the query sequence } A \text{ with sequence } B \]

\( i \) is position number in the query sequence \( A \)
\( a \) and \( b \) are aminoacid residuals in sequence \( A \) and sequence \( B \)
\( m \) is current shift between sequence \( A \) and sequence \( B \)
\( F \) is frame size
\( R_i \) is primary similarity value
\( S_i \) is the local similarity value for position \( i \) in the query sequence \( A \) with sequence \( B \)

About 1000 sequences per second.
Sequence Local Similarity.

CTR52A1 PRELIMINARY; PRT; 543 AA.
“If there exists correspondence between similarity of substrates and protein sequences in cytochrome P450 superfamily?”

The results of substrate-based clustering correspond to homology-based classification for families CYP 1, 2, 3, 4, 5, 6, 7, 11. For other families of P450 (CYP 8, 17, 19, 21, 24, 26, 27) substrate-based clustering brings to the contradictions with the traditional classification.

"Quantifying the Relationships among Drug Classes"

A subset of the MDDR database containing 65,367 compounds organized in 249 sets that associate with a specific biological target

“By multiple criteria, bioinformatics and chemoinformatics networks differed substantially, and only occasionally did a high sequence similarity correspond to a high ligand-set similarity.”

Protein function recognition based on learning by example

It is based on a data set of sequences with known properties. This data set must be subdivided into “positive” and “negative” examples – group A and its complement \( \neg A \).
Is there universal similarity reasonable?
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 $k$. 
Sequence Local Similarity. Algorithm of Classification

Belonging of the sequence under study to a class $A$ is calculated using statistical function $B(A)$:

$$B(A) = \frac{\sum S_{ik} [w_k(A) - w_k(\neg A)]}{\sum S_{ik} [w_k(A) + w_k(\neg A)]}$$

$$B^c(\neg A) = \frac{\sum S_{ik} [w_k(\neg A) - w_k(A)]}{\sum S_{ik} [w_k(\neg A) + w_k(A)]}$$

$i = 1, \ldots, n$ is position number in the sequence under study;
$k = 1, \ldots, N$ is the experimentally annotated sequence number;
$w_k(A), w_k(\neg A)$ are weights in class $A$ and its complement $\neg A$ of the experimentally annotated sequence $k$;
$S_{ik}$ is similarity for position $i$ of the sequence under study and the experimentally annotated sequence $k$. 
General Classification Problem

<table>
<thead>
<tr>
<th>Observed value</th>
<th>Calculated value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>TN</td>
<td>FN</td>
</tr>
</tbody>
</table>

Threshold
General Classification Problem.
Criteria of classification accuracy

Sensitivity = \frac{TP}{(TP+FN)}
Specificity = \frac{TN}{(TN+FP)}
Accuracy (Concordance) = \frac{(TP+TN)}{N}
Predictive value positive = \frac{TP}{(TP+FP)}
Predictive value negative = \frac{TN}{(TN+FN)}
False Negative Rate = \frac{FN}{(TP+FN)} = Error_1
False Positive Rate = \frac{FP}{(TN+FP)} = Error_2
Positive Likelihood = \frac{SENS}{(1-SPEC)}
Negative Likelihood = \frac{(1-SENS)}{SPEC}

http://www.intmed.mcw.edu/clincalc/bayes.html
General Classification Problem.
Independent Accuracy of Prediction (IAP)

IAP is calculated using Leave-One-Out Cross-Validation procedure.

$B_i$ is the estimation for sequence $i$ from the class $A$
$B_j$ is the estimation for sequence $j$ from its complement $\neg A$
$\Theta(x) = 1$ if $x > 0$, $\Theta(x) = \frac{1}{2}$ if $x = 0$, $\Theta(x) = 0$ if $x < 0$
$N_A$ is the number of sequences in the class $A$
$N_{\neg A}$ is the number of sequences in its complement $\neg A$

How sequence local similarity can be used?
Training sets used for the method evaluation

- **Serine proteases**
  EC 3.4.21 – 28 groups of 4th EC level, 623 sequences

- **Gold standard**, especially composed to test statistical learning methods
  5 enzyme superfamilies and 56 families, 832 sequences.

- **P450 superfamily** (CPD database)
  242 proteins classified by substrate specificity (579 compounds).
  163 proteins classified by inhibitor specificity (272 compounds).
Serine proteases

The average accuracy reached the maximum (close to one) at the maximal shift of 50 and frame of 50. 24 of 28 classes were recognized at this parameter values with 100% accuracy.
The average IAP exceeded 0.99.
4 superfamilies were recognized with 100% accuracy.
45 families were recognized with $\text{IAP} = 1$
and 11 families were recognized with $\text{IAP} > 0.96$.

The superfamilies seem to be clearly recognized by alignment-based methods; however the families of the same superfamily are worse recognized by the analysis of aligned sequences with phylogenomics methods.
Prediction of activity spectra for organic compounds

GUSAR – QNA based prediction of quantitative properties of organic compounds
Conclusions

Our approach revealed the high efficiency of function prediction with different sequence description types.

The high accuracy of prediction was obtained for different levels of protein functional classifications.

The projection method is useful both for functional specificity prediction and for sequences mapping, i.e. to reveal the local determinants of the functional specificity.

The approach “RECOGNITION OF PROTEIN FUNCTION USING THE LOCAL SIMILARITY” will be published in Journal of Bioinformatics and Computational Biology, 2008
Acknowledgements

This work was supported by Russian Federation of Basic Research (grant N 04-04-49390-a).
We are grateful to A.V. Lisitsa for providing the data on cytochrome P450 substrates and inducers specificity.