Chemoinformatics: Basic Concepts and Areas of Application

Alexandre Varnek
Laboratory of Chemoinformatics, University of Strasbourg
Double diploma UniStra/KFU

KAZAN (VOLGA REGION) FEDERAL UNIVERSITY
Chem(o)informatics
Cheminformatics
Chemical Informatics
Infochimie
Chémoinformatique
Хемоинформатика
**Chemoinformatics - definition**

*Chemoinformatics* is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.

*G. Paris, 1998*

*Chemoinformatics* is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.

*F.K. Brown, 1998*

*Chemoinformatics* is the application of informatics methods to solve chemical problems.

*J. Gasteiger, 2004*

*Chemoinformatics* is a field based on the representation of molecules as objects (graphs or vectors) in a chemical space.

*A. Varnek & I. Baskin, 2011*
Selected books in chemoinformatics
Gallium discovery: the first QSAR successful story

Predicted in 1869

Density\textsubscript{pred} \approx 6.0 \text{ g/cm}^3

Discovered in 1875

Density\textsubscript{exp} = 4.7 (initial)

Density\textsubscript{exp} = 5.935 (corrected)
Chemoinformatics:

new discilne combining several „old“ fields

• Chemical databases

• Structure-Activity modeling (QSAR)

• Structure-based drug design

• Computer-aided synthesis design
OUTLOOK

• Needs for chemoinformatics
• Fundamentals of chemoinformatics
• Chemical Space paradigm
• Virtual screening approaches
• Perspectives
Needs in Chemoinformatics
Chemical universe

> 100 M compounds are currently recorded

- How to select useful compounds from this huge dataset?
- How to design new compounds?
- How to synthesize these compounds?
Virtual Screening

Large libraries of molecules

Target Protein

Small Library of selected hits

High Throughout Screening
Virtual screening is inevitable to analyse a huge amount of protein-ligand combinations

**Chemical universe:**
- $>10^8$ compounds are currently available
- $10^{33}$ druglike molecules could potentially be synthesised

(see P. Polischuk, T. Madzidov et al., JCAMD, 2013)

**Human proteome:**
- 5000 druggable proteins

Virtual screening must be very fast and efficient!
Ionic Liquids

Ionic Liquids are composed of

large organic cations:

\[ \text{PF}_6^-, \text{Cl}^-, \text{BF}_4^-, \text{CF}_3\text{SO}_3^-, [\text{CF}_3\text{SO}_2]_2\text{N}^- \]
There exist $10^{18}$ combinations of ions that could lead to useful ionic liquids.
Virtual screening: finding the needle in the haystack

CHEMICAL DATABASE

~$10^6 - 10^9$ molecules

\[
\text{HN} \quad \text{HN}
\]

\[
\text{H}_3\text{C}
\]
**Chemoinformatics:** pattern recognition in chemistry

- Specific structural motifs,
- Selected molecular properties (shape, fields, ...),
- Interaction patterns,
- Mathematical equations

\[ \text{Activity} = F(\text{structure}) \]
Chemoinformatics: Virtual screening “funnel”

- Filters
- Similarity search
- (Q)SAR
- Pharmacophore
- Docking

Chemical Database

~10^6 – 10^9 molecules

Virtual Screening

~10^1 – 10^3 molecules

HITS

INACTIVES
Chemoinformatics: Virtual screening “funnel”

Filters

Similarity search

(Q)SAR

Pharmacophore

Docking

Ligand-based

Structure-based

CHEMICAL DATABASE

~10^6 – 10^9 molecules

VIRTUAL SCREENING

HITS

~10^1 – 10^3 molecules

INACTIVES
Chemoinformatics as a theoretical chemistry discipline
Chemoinformatics is defined as an individual discipline characterized by its own molecular model, basic concepts, major applications, and learning approach.
Theoretical chemistry

Quantum Chemistry

Force Field
Molecular Modelling

Chemoinformatics

- Molecular model
- Basic concepts
- Major applications
- Learning approaches
Chemoinformatics is a field based on the representation of molecules as objects (graphs or vectors) in a chemical space.
Chemoinformatics: From Data to Knowledge

Chemoinformatics learns from experimental data!
Basic concepts

Quantum Chemistry  \(\rightarrow\) wave/particle dualism

Force Field Molecular Modelling  \(\rightarrow\) classical mechanics

Chemoinformatics  \(\rightarrow\) chemical space
Chemical space paradigm
Chemical Space representations

graphs-based

SPACe = objects + metric

descriptors -based
Graph-based chemical space
Scaffold Tree

Natural Product Scaffold Tree

Cyclohexane branch containing cyclic terpenes, terpenoids and steroids

Courtesy of P. Ertl
Descriptors-based chemical space

vectorial space defined by molecular descriptors
Case study: Hansch Analysis

Biological Activity = \( f (\text{Physicochemical parameters}) + \text{constant} \)

\[ \text{Activity} = a (\log P)^2 + b \log P + \rho \sigma + \delta E_s + \text{cont} \]

3 types of physicochemical parameters are used:

- **Electronic** \((\sigma)\)
- **Steric** \((\delta E_s)\)
- **Hydrophobic** \((\log P)\)
Case study: Hansch Analysis
Molecular Descriptors:

ensemble of topological, electronic, geometry parameters calculated directly from molecular structure

Molecular graph

- Topological indices,
- Atomic charges,
- Inductive descriptors,
- Substructural fragments,
- Molecular volume and surface, …

Descriptor vector

> 5000 types of descriptors are reported
Chemography:

Design and visualization of chemical space
Dimensionality Reduction problem

Greenland 2.2 M km²

Arabian Peninsula 3.5 M km²

Australia 7.7 km²
Generative Topography Mapping (GTM)

Swiss Roll

• GTM relates the latent space with a 2D “rubber sheet” \((\text{manifold})\) injected into the high-dimensional data space.

• The visualization plot is obtained by projecting the data points onto the manifold and then letting the “rubber sheet” relax to its original form.

Similarity principle: similar molecules possess similar properties

GTM of a dataset containing 10 activities from DUD
Similar compounds possess similar properties
Chemical space representation: Activity Landscapes

\[ \bar{A}_k = \frac{\sum_i A_i R_{ik}}{\sum_i R_{ik}} \]

Expectation of activity \( \bar{A}_k \) in \( k \)-node for the training set
Activity landscape of lanthanides’ binders

Generative Topographic Mapping of the set of Ln binders

Contours correspond to different logK values

Weak binders

Strong binders

H. Gaspar, I. Baskin, G. Marcou, A. Varnek  unpublished results
Visualization of models’ Applicability Domain

*Case study*: classification models for BDDCS classes

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>High solubility</td>
<td>Low solubility</td>
</tr>
<tr>
<td>High metabolism</td>
<td>High metabolism</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>High solubility</td>
<td>Low solubility</td>
</tr>
<tr>
<td>Low metabolism</td>
<td>Low metabolism</td>
</tr>
</tbody>
</table>

**DATASET**: 893 drugs

**DESCRIPTIONS**: VolSurf
BDDCS classes probability distribution

Class Preference Factor \( CPF = \frac{\max_c P(k|C)}{P(k|C_i)} \), \( \forall C_i \neq C \)

\( CPF \leq 1 \), coverage = 100 %

\( CPF \leq 5 \), coverage = 47 %

Colored zones on the maps correspond to model’s applicability domain

H. Gaspar, G. Marcou, A. Varnek  
JCIM, 2013
Chemoinformatics: Properties predictions
Quantitative Structure-Activity Relationships (QSAR)

Activity = \( F(\text{structure}) \)
= \( F(\text{descriptors}) \)

machine-learning methods

• neural networks, support vector machine, random forest, naïve Bayes, PLS, …

predictions of > 20 physico-chemical properties and NMR spectra for each individual compound
<table>
<thead>
<tr>
<th>logP</th>
<th>VAR</th>
<th>TRUST</th>
<th>REASON</th>
</tr>
</thead>
</table>
| 1.59 | 0.546| NONE  | - None of the local models have applicability domains covering this compound  
- Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width |
Machine Learning Methods in Chemoinformatics: Quo Vadis?


In silico design of new molecules ("inverse QSAR")

Predictive performance

- small and diverse datasets
- large and diverse datasets
- applicability domain
- Training and test sets belonging to different data domains
- Construction of "optimal" training sets

Machine-Learning methods

Incompleteness of molecular descriptors

Accounting for multiple species (conformers, tautomers, ...)

Functional endpoints
Chemoinformatics: virtual screening in 3D
Virtual screening: finding the needle in the haystack

Chemical Database

~$10^6 - 10^9$ molecules

\[
\text{Chemical Structure: } \text{HNCH}_{2}\text{CHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCHCH
What is in common between these two molecules?

Arg-Gly-Asp-Phe

Tirofiban
Pharmacophore model of ligand complementary to integrin $\alpha_{IIb}\beta_3$

- Positive charge, H-donor
- Negative charge, H-acceptor

Hydrophobic interactions

Distance:
- 16 Å
- 15.5 Å
- 5 Å
Molecular Shape similarity analysis

\[ pK_i = 7.51 \]
\[ \text{TanimotoCombo} = 0.74 \]

\[ pK_i = 7.82 \]
\[ \text{TanimotoCombo} = 0.67 \]
Molecular fields

Interpretation of a field point pattern. The size of the point indicates the potential strength of the interaction.
Ligand-to-protein docking:

Lock-and-key paradigm

Hermann Emil Fischer

Ligand-Protein complex
Selected *in silico* designed compounds that were synthesized and successfully tested for bioactivity.
Chemoinformatics: areas of application

- Drug design (pharmacodynamics and pharmacokinetics),
- Prediction of physico-chemical properties,
- Materials design,
- Synthesis design,
- Molecular spectra simulations
Chemoinformatics: perspectives
Assessment of biological activity
Assessment of side effects

Pharmacological Profiling
Chemogenomic Profiling

<table>
<thead>
<tr>
<th>OH</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Molecule 1" /></td>
<td><img src="image2.png" alt="Molecule 2" /></td>
<td><img src="image3.png" alt="Molecule 3" /></td>
<td><img src="image4.png" alt="Molecule 4" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Molecule 5" /></td>
<td><img src="image6.png" alt="Molecule 6" /></td>
<td><img src="image7.png" alt="Molecule 7" /></td>
<td><img src="image8.png" alt="Molecule 8" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Molecule 9" /></td>
<td><img src="image10.png" alt="Molecule 10" /></td>
<td><img src="image11.png" alt="Molecule 11" /></td>
<td><img src="image12.png" alt="Molecule 12" /></td>
</tr>
</tbody>
</table>

Chemoinformatics: Complexity challenge

P. Csermely1 et al. *Pharmacology & Therapeutics*, 2012
Day 1: Databases

Chemical Databases: Encoding, Storage and Search of Chemical Structures

Timur Madzidov

SciFinder - The choice for chemistry research

Veli-Pekka Hyttinen

Tutorial with ChemAxon

Gilles Marcou
Dragos Horvath
Day 2: QSAR

Obtaining, Validation and Application of SAR/QSAR Models

SAR/QSAR Modelling: state of the art

ADMET Predictions

Tutorial with OChem
Day 3: virtual screening in 3D

Conformational Sampling

Dragos Horvath

Pharmacophore and Its Applications

Thierry Langer

Molecular Docking Methods

Gilles Marcou

Tutorial with LigandScoute

Sharon Bryant

Tutorial with LeadIt

Gilles Marcou  Dragos Horvath
Day 4: Drug Design applications

Computational Mapping Tools for Drug Discovery

Konstantin Balakin

Drug Design & Discovery in Academia

Vladimir Poroikov