Solved and Unsolved Problems in Chemoinformatics

Johann Gasteiger
Computer-Chemie-Centrum
University of Erlangen-Nürnberg
D-91052 Erlangen, Germany
Johann.Gasteiger@fau.de
Overview

• objectives of lecture
• achievements
• challenges
• unsolved problems
• summary
Objectives

• many problems have been solved
  → let us be proud!

• there is still a lot to be done
  → chemoinformatics is a field of its own, is attractive for students
Chemoinformatics – An Old Discipline

- Structure activity relationships
  1963, Hantsch & Fujita
- structure representation
  1965, Morgan
- structure elucidation
  1965, Sasaki, Munk, DENDRAL
- synthesis design
  1970, Corey & Wipke, Ugi+Gasteiger, Hendrickson
- molecular modeling
  1970, Langridge, Marshall
- data analysis / chemometrics
  1970, Kowalski, Wold
Achievements

• access to chemical information
• learning from chemical information
• applications
Databases

- Chemical Abstracts Service (1975)
- Beilstein (1990)
- Gmelin (1990)
- ChemInformRX (1991)
- SpecInfo (1991)
- etc.
Number of Compounds in Chemistry

compounds published in CAS

102 million compounds, 66 million sequences (June 2015)
Search for Cancerostatic Drugs

protein/substrate complex

similar substrates
Chemical Structures

• computers have learnt the language of a chemist
  communicating by structure diagrams
• molecules are stored with atomic resolution providing access to each atom and bond
  enabling substructure search
• the complex structures of molecules can be visualized
  new insights can be gained
Summary

• databases have strongly contributed to the progress in chemistry and related fields

• without databases modern research in chemistry would be inconceivable
Learning from Chemical Information

• learning from data
• QSAR/QSPR
• representation of chemical structures
Problem: Not Enough Information

102,000,000 chemical compounds

1,000,000 3D structures in Cambridge Crystallographic Data File

we only have data on the 3D structure for less than 1% of the known compounds
Problem: Not Enough Information

102,000,000 chemical compounds

1,000,000 3D structures in Cambridge Crystallographic Data File

we only have data on the 3D structure for less than 1% of the known compounds

can we learn the rules from the known 3D structures
Rule-Based Learning: CORINA

- connection table & stereo descriptors
  - initialization of internal coordinates
    - rings:
      - ring perception
      - ring template search
      - ring assembly
  - acyclic systems:
    - removal of steric crowding
- 3D coordinates

generates 3D coordinates for >99.5% of all organic compounds
Data-Based Learning
Structure-Property Relationships (QSPR, QSAR)

- Molecular structure
- Property
- Representation
- Structure descriptors
- Model building
Representation of Chemical Structures

- topological indices
- fragment codes
- fingerprints
- ....

Several thousands of different types of chemical descriptors have been developed

Representation of Chemical Structures

Of Molecules and Humans
J. Gasteiger,
Prediction of Properties

- physical, e.g.
  - aqueous solubility
  - $^{13}$C NMR shifts

- chemical, e.g.
  - acidity

- biological, e.g.
  - toxicity
Applications of Chemoinformatics: Drug Design

Chemoinformatics has become an integral part of the drug design process

• lead discovery
  • virtual screening
  • pharmacophore searching

• lead optimization
  • QSAR
  • molecular docking

• prediction of ADME properties
  • solubility, adsorption, distribution, metabolism, excretion….
Handbook of Chemoinformatics
From Data to Knowledge

J. Gasteiger (Editor)

65 authors
73 contributions
4 volumes
1900 pages
Wiley-VCH, Weinheim
(August 2003)
Chemoinformatics - A Textbook -

J. Gasteiger, T. Engel (Editors)

650 pages

Wiley-VCH, Weinheim (September 2003)
Challenges

- essence of chemistry
- environmental impact of chemicals
- understanding chemistry
- understanding biological systems
- human health

applications in all fields of chemistry and related scientific disciplines
Chemical Industry is not selling compounds but is selling properties i.e., compounds with desired properties (drugs, paints, fibers, plastics, pesticides, catalysts, ceramics, etc.)
Fundamental Questions in Chemistry

What structure do I need for a certain property?
structure-activity relationships

How do I make this structure?
synthesis design

What is the product of my reaction?
reaction prediction
structure elucidation

In all those areas the use of chemoinformatics could help!
Risk Assessment of Chemicals

REACH – Registration, Evaluation, Authorization and restriction of Chemicals

- for those chemicals used with more than 10 tons/year manufactured or imported into the European Union a Chemical Safety Report is needed
- law since June 1, 2007; registration until Dec 1, 2013
- applies to about 35,000 chemicals
- testing on harmful effects on human health or environment, determination of persistence, bioaccumulation and toxicity
- testing is time-consuming, expensive and might need many animals

Use chemoinformatics methods for ranking of chemicals
Cosmetics Directive

• for chemicals used in cosmetics products
  • no compounds tested on animals are allowed in cosmetics in Europe since 2009.
  • all animal tested cosmetics will eventually be banned on the European market.

Use chemoinformatics methods for developing alternatives to animal testing
Problems Still to be Solved

• access to chemical information
• learning from chemical information
• applications
Improved Access to Chemical Information

• input of chemical structures (hand writing, voice)
• representation of neglected compounds (boranes, organo-metallic structures (ferrocene, etc.), Markush structures (patents), polymers)
• text mining (optical character recognition)
• publishing chemical information (3D structures, spectra)
• publishing and searching on the internet
Neede for Better Databases

- on compounds
  - store all available information (all properties)
  - store all spectra
- on reactions
  - give the entire stoichiometry of a reaction
  - store all reaction conditions (solvent, temperature, reaction time)
  - kinetic data
From Experiment to Reaction Database
(Presently)

- information producer
  - many breaks in information processing
- information consumer
- experiment
  - lab journal
  - report
    - manuscript
      - patent
        - secondary literature
          - abstract
            - input into reaction database
              - printed reaction scheme
              - reaction database

written on paper
electronically
From Experiment to Reaction Database
(Future: Electronic Lab Notebooks)

information producer

information consumer

experiment

e lab notebook → report

manuscript

patent

publication

abstract

input into reaction database

secondary literature

a direct information flow

written

electronically

on

paper

reaction scheme

printed reaction database

reaction database

C³ © Gasteiger
Learning from Chemical Information

- **Note**: standard QSAR studies have difficulties to get published if they do not also
  - provide new experimental results or
  - increase our knowledge and insight

- represent structures by descriptors that can be interpreted
  - combine substructures with physicochemical effects

- use data analysis methods that are not a black box
  - a forest of decision trees gives more exact predictions but cannot be interpreted; a single decision tree can be interpreted!

- combine a QSAR model with e.g. docking studies

From models to interpretation
Applications

- all fields of chemistry
- drug design
- analytical chemistry
- organic chemistry
- biochemistry
- toxicology
- material science
Drug Design

• conformational flexibility of drugs and proteins
• docking into proteins
  • no consensus scoring (science cannot be predicted by voting)
  • try to model the physicochemistry of the process
• protein-protein and protein-DNA interactions
• prediction of ADME-Tox properties
• model the various organs of a human
Analytical Chemistry

• origin of olive oils
• simulation of infrared spectra
• 3D structure from infrared spectrum
Italian Olive Oils

9 regions

572 olive oil samples

contents of 8 fatty acids:

- palmitic acid
- palmitoleic acid
- stearic acid
- oleic acid
- linoleic acid
- arachidic acid
- linoleic acid
- eicosenoic acid

Self-Organizing Map of Olive Oils

572 samples
9 regions
8 fatty acids

250 training
322 test

10 prediction errors

Self-Organizing Map of Olive Oils

The merits of unsupervised learning

Computer-Assisted Structure Elucidation (CASE)

- CASE was one of the roots of chemoinformatics
- various groups worked on it (Munk, Sasaki, Funatsu, Steinrück)
- not much done recently
- no useful general purpose system available
- much time spent by chemists on structure elucidation

structure elucidation should be done more efficiently, using all available information and using software
Simulation of Infrared Spectra

Predict the Infrared Spectrum from the structure of a compound

3D structure representation by radial distribution function

\[ g(r) = \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{-B(r-r_{ij})^2} \]

- \( N \) number of atoms
- \( A \) atomic properties
- \( B \) temperature or smoothing factor
- \( r_{ij} \) interatomic distance
Training of a Counterpropagation (CPG) Network

3D structure $\xrightarrow{\text{transformation}}$ Radialcode (128 values)

IR spectrum (128 absorbance values)
Simulation of IR Spectra

3D Structure Prediction from IR Spectrum

3D structure $\xrightarrow{\text{decoding}}$ Radialcode (128 values)

IR spectrum (128 absorbance values)
IR Spectrum of Unknown Compound
3D Structure Prediction from IR Spectra

Correct prediction!

Organic Chemistry

• chemical reactivity

• organic synthesis design
Chemical Reactivity

• needs better data in reaction databases
• reach out to theoretical chemists
• put the results of quantum mechanical calculations into databases
Calculation of Chemical Effects

charge distribution

inductive effect

resonance effect

polarizability effect

bond dissociation energy
Calculation of Proton Affinities (PA)

- alkyl amines only (49 cpds)
  \[
  \text{PA(kJ/mol)} = 205.6 + 2.82 \alpha_{d,N}
  \]

- alkyl amines + heteroatom substituted alkyl amines (80 cpds)
  \[
  \text{PA(kJ/mol)} = 1435.5 + 12.5 \alpha_{d,N} - 116.3 \chi_{r,N}
  \]
A \( pK_a \) Model of Aliphatic Carboxylic Acids

\[
pK_a = -37.54Q_{\sigma,o} + 12.27A_{2D,q} - 1.02\alpha_o + 0.11\chi_{\pi,\alpha-C} - 1.89I_{amino} + 19.10
\]

\[n = 1122, \ r^2 = 0.81, \ s = 0.42, \ F = 809\]

Computer-Assisted Organic Synthesis Design (CASD)

- CASD was one of the roots of chemoinformatics
- many products can be traced back to CASD work
  - MACCS, REACCS, Beilstein DB, ChemInform RX DB
- however CASD systems are not yet widely accepted by organic chemists
- but it is still true:

  “The amount of information to be processed and the decisions between many alternatives suggests the use of computers in synthesis design.”
  
  (H.Gelernter, 1973)

The design of organic syntheses should be done more efficiently, using all available information, by using software
Biochemistry and Bioinformatics

• Study of diseases
Gerhard Michal (Hrsg.)

Biochemical Pathways

Biochemie-Atlas
Application of BioPath.Database

• search for enzyme inhibitors
  M.Reitz, A.von Homeyer, J.Gasteiger,

• search for similar enzymes
  O.Sacher, M.Reitz, J.Gasteiger,
  X.Hu, A,Yan, T.Tan, O.Sacher, J.Gasteiger

• discover essential pathways of diseases
  G.Kastenmüller, J.Gasteiger, H.W.Mewes,
  Bioinformatics, 2008, 24, i56-i62
  G.Kastenmüller, M.E.Schenk, J.Gasteiger, H.W.Mewes,
  Genome Biology, 2009, 10, R28

Uncovering Metabolic Pathways to Phenotypic Traits

- PEDANT
  - Annotated genomes

- BioPath database

- Metabolic reconstruction

- Pathway profile

- Attribute selection + Pathway ranking

- Relevant pathways

- Phenotypic traits
Periodontal Disease

Human oral flora:
   > 700 species

PEDANT:
   15 fully sequenced oral genomes
   (incl. 4 of 6 periodontal pathogens)

BioPath
   68 global pathways
   306 smaller pathways
Relevant Pathways for Phenotype Periodontal Disease Causing

- Glutamate fermentation
- Biosynthesis of L-proline
- Biosynthesis of 5-formimino-THF
- Conversion of L-glutamate to L-proline
- Conversion of L-glutamate to L-ornithine
- Degradation of L-histidine to L-glutamate

Relevant Pathways for Phenotype

*Periodontal Disease Causing*

- Glutamate fermentation
- Biosynthesis of L-proline
- Biosynthesis of 5-formimino-THF
- Conversion of L-glutamate to L-proline
- Conversion of L-glutamate to L-ornithine
- Degradation of L-histidine to L-glutamate

Several of these pathways produce NH$_3$

Cytotoxic NH$_3$ plays a major role in periodontal disease

gene \rightarrow protein \leftarrow drug \rightarrow lead
Toxicity and Risk Assessment

• meeting the challenges posed by REACH, drug design and Cosmetics Directive
• projects funded by the European Union, Innovative Medicine Initiative and Cosmetics Europe
  • eTOX (11 academic groups, 6 SMEs, 13 pharma companies)
  • COSMOS (5 academic groups, 3 public institutions, 5 SMEs, 2 cosmetics companies)
Search Structures for Toxic Alerts

- **Chemotypes**
  > structure fragments with physico-chemical properties

- **ToxPrint**
  > alerting chemotypes for genotoxic carcinogens

- **Publicly available**
  - [www.chemotyper.org](http://www.chemotyper.org)
  - [www.toxprint.org](http://www.toxprint.org)

Material Sciences

- Diverse material properties
QSPR Modeling of Diverse Material Properties

David A. Winkler, CSIRO, Melbourne, Australia

- Nanomaterials
- Catalysts
- Biomaterials
- Polymers
- Ionic liquids
- Supercritical carbon dioxide
- Ceramics

Conclusions

• Chemoinformatics has become indispensable in drug design
• has many potential applications in all areas of chemistry and other scientific fields
• talk to colleagues about their problems
• build models with their data – and knowledge
• Society expects a lot from chemoinformatics
  (Nobel Prize in Chemistry 2013: Karplus, Levitt, Warshel)
Teaching

• several textbooks have been published
  • Gillet+Leach, Gasteiger+Engel, Bajorath, Wild
• define curriculum in chemoinformatics
  • various universities already teach chemoinformatics
  • the number is growing
• integrate chemoinformatics into regular chemistry curricula
Acknowledgements

• Many excellent and dedicated coworkers at Computer-Chemie-Centrum, University of Erlangen-Nuremberg, Germany
  
http://www2.chemie.uni-erlangen.de

• Excellent coworkers at
  
Molecular Networks
Inspiring Chemical Discovery

http://www.molecular-networks.com

• Dr. Ulrike Gasteiger
Atorvastatin