Molecular modelling in compounds series with descriptors families

Lorentz JÄNTSCHI

&

Sorana D. BOLBOACĂ

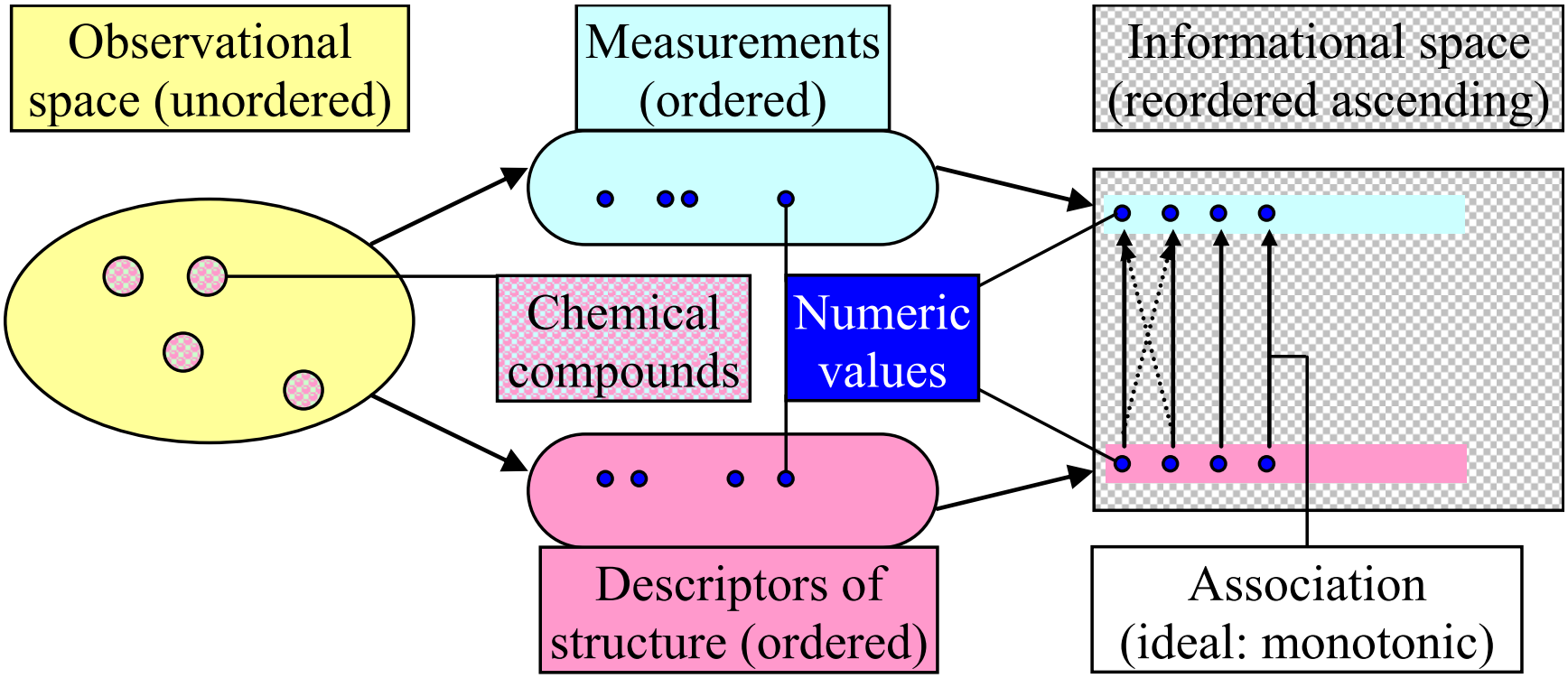
UT & UMF Cluj-Napoca

May 13, 2016

Reference materials

- Jäntschi L, 2000. Predicția proprietăților fizico-chimice şi biologice cu ajutorul descriptorilor matematici. Teză de doctorat (Chimie) - coordonator Prof. Diudea MV. Universitatea "Babeş-Bolyai" din Cluj-Napoca, Cluj-Napoca, Ian. 2000.
- Jäntschi L, 2004. MDF - A New QSAR/QSPR Molecular Descriptors Family. Leonardo Journal of Sciences 3(4): 68-85.
- Jäntschi L, 2005. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. Leonardo Electronic Journal of Practices and Technologies 4(6): 76-98.
- Jäntschi L, 2012. Structure vs. Properties: Algorithms and Models. Habilitation thesis in Chemistry. Bucharest: CNATDCU (defended in 2013 in Cluj-Napoca: Babeş-Bolyai University). 175 p.
- Jäntschi L, 2014. Szeged Matrix Property Indices. Online calculation software. URL: http://l.academicdirect.org/Chemistry/SARs/SMPI/
- Jäntschi L, Bolboacă SD, 2016. Families of molecular descriptors In: Explicative dictionary of nanochemistry (Ed. M.V. Putz). Oakville(ON): AAP & CRC Press, to appear.

Compounds series associations



Challenges:

- Measurement scale
- Value-based encoding of the structure
- Training of the association function

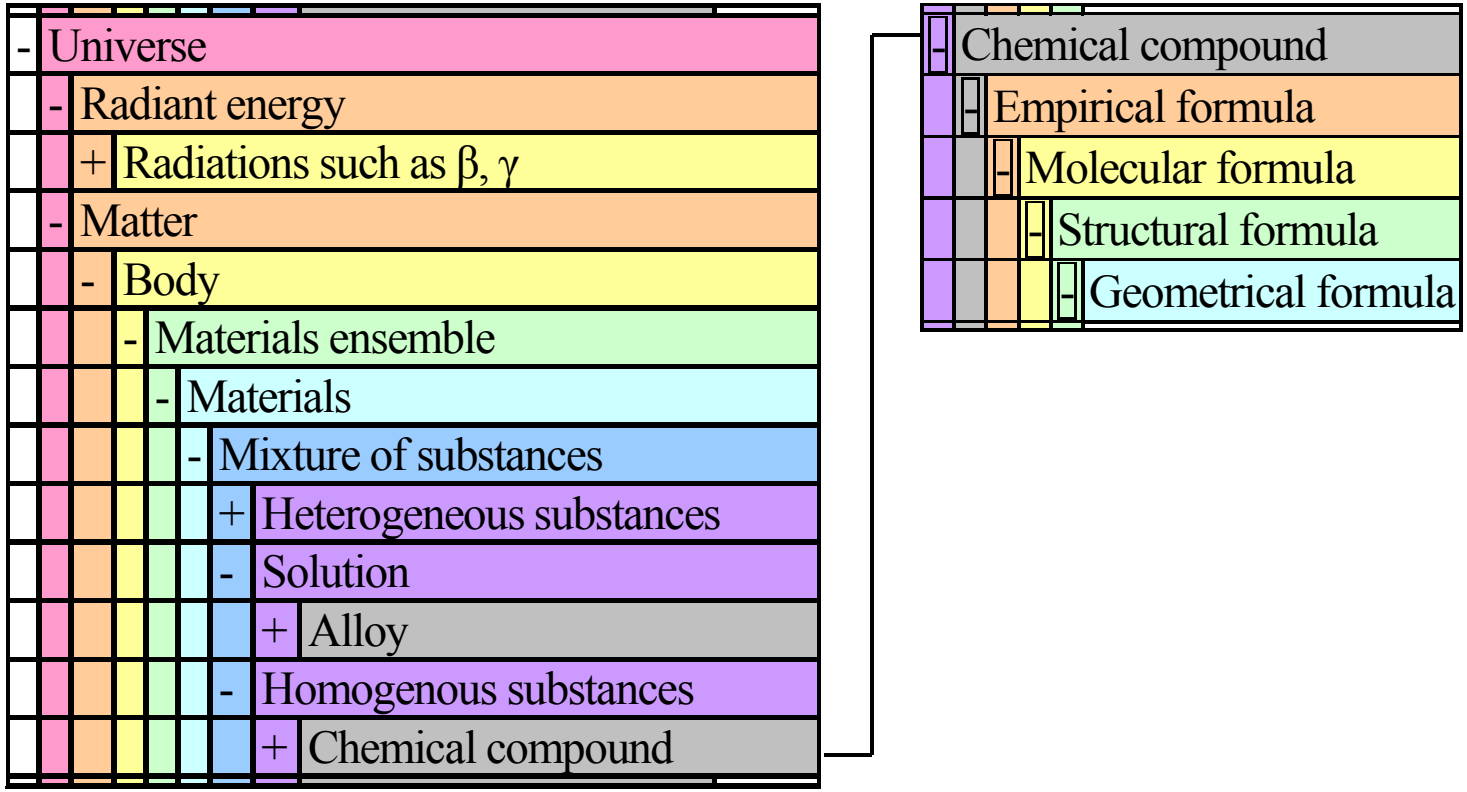
Measurement scales

Scale	Type	Operations	Structure	Statistics	Examples
Binomial	Logical	"=", "!"	Boolean algebra	Mode, Fisher Exact	Dead/Alive Sides of a coin
(multi) Nomi(n)al	Discrete	"="	Standard set	Mode, Chi squared	ABO blood group system Living organisms classification
Ordinal	Discrete	"=", "<"	Commutative algebra	Median, Ranking	Number of atoms in molecule
Interval	Continue	"≤", "-"	Affine space (one dimensional)	Mean, StDev, Correlation, Regression, ANOVA	Temperature scale Distance scale Time scale Energy scale
Ratio	Continue	"≤", "-", "*"	Vector space (one dimensional)	GeoMean, HarMean, CV, Logarithm	Sweetness relative to sucrose pH

Challenges:

- Resolution of the scale
- Degeneration (lost of the unicity for the observed)

Value-based encoding of the structure



Challenges:

- Translation of the topology (chemical bonds)
- Translation of the geometry (molecular arrangement)
- Treating of the different levels of isomerism

Training of the association function

- Model of association
  - Linear vs. nonlinear or user-defined
- Parameterizing of the model
  - Least squares vs. maximum likelihood
- Objective (constrains) for the association
  - Monotonic associated estimates vs. leverage
- Assessing of the estimating power
  - Leave-one-out vs. training-vs-test
- Assessing of the prediction power
  - Applicability domain vs. external set-test

Molecular geometry

Molecular modelling software

Name	Provider website
Abalone	http://biomolecular-modeling.com
ADF	http://scm.com
ChemBioOffice	http://cambridgesoft.com
Gaussian	http://gaussian.com
HyperChem	http://hyper.com
Materials Studio	http://accelrys.com
Q-Chem	http://q-chem.com
Spartan	http://wavefun.com
Others	wiki→ List_of_software_for_molecular_mechanics_modeling

Challenges:

- Starting geometry
- Theory level
- Vitro vs. vivo
- Convergence threshold

Typical information from modelling

The list of the atoms			
Label	Type	Coordinates (x, y, z)	Partial charge

The list of the bonds		
Atom Label	Atom Label	Bond type or order

Molecular descriptors families strategy

Structure	Feed	Genetic code	Breeding	Phenotypes
	$\rightarrow$ (+)	Gene | A | B | … | Z Genome: a_1 b_1 … z_1; a_2 b_2 … z_2; … … … …; a_{99} b_{99} … z_{99}	$\rightarrow$ (×)	$\{P_1, \ldots, P_{99\cdot 99\cdot \ldots \cdot 99}\}$

Set of compounds: $S_1, S_2, \ldots, S_m$

Activities from measurements

Structures	Activities	Phenotypes	Survival of the fittest				
S_1	A_1	$\{P_1(S_1), \ldots, P_{99\cdot\ldots\cdot 99}(S_1)\}$	$\{A_i\}\sim f(P_1(\{S_i\}))$	…	$\{A_i\}\sim f(P_j(\{S_i\}))$	…	$\{A_i\}\sim f(P_{99\cdot\ldots\cdot 99}(\{S_i\}))$
S_2	A_2	$\{P_1(S_2), \ldots, P_{99\cdot\ldots\cdot 99}(S_2)\}$					
…	…	…					
S_i	A_i	$\{P_1(S_i), \ldots, P_{99\cdot\ldots\cdot 99}(S_i)\}$					
…	…	…					
S_m	A_m	$\{P_1(S_m), \ldots, P_{99\cdot\ldots\cdot 99}(S_m)\}$					

FPIF: Fragmental Property Index Family

Code of FPIF descriptors

Gene	I_M	D_M	A_P	P_D	F_C	S_M	M_I	L_O
Genome	R	T	M	__p__	si	S	P_	I
	D	G	E	__d__	se	P	P2	R
			C	_1/p_	ji	A	E_	L
			Q	_1/d_	je	G	E2	
				_p*d_	fi	H		
				_p/d_	fe			
				_p/d2				
				p2/d2				

$FPIF = I_M \times D_M \times A_P \times P_D \times F_C \times S_M \times M_I \times L_O$

Ex.: RGseCp2/d2SE2, DGjeP_p/d2GP_

Ref: Jäntschi & Diudea, 2000

Calculation details:
- in the paper

Major advantage:
- fits on graph theory indices

Major disadvantage:
- complexity of calculation for j* and f* fragments

MDF: Molecular Descriptors Family

Code of MDF descriptors

Gene	D_M	A_P	I_D				I_M	F_C	S_M				L_O
Genome	t	C	D	Q	L	F	r	m	m	A	G	H	I
	g	H	d	q	l	f	R	M	M	a	g	h	i
		M	O	J	V	S	m	D	n	B	F	I	A
		E	o	j	E	s	M	P	N	b	f	i	a
		G	P	K	W	T	d		S	P	s		L
		Q	p	k	w	t	D						l

$MDF = D_M \times A_P \times P_D \times I_M \times F_C \times S_F \times L_O$

Ex: lsPRLGg, IhDDDCt

Ref: Jäntschi, 2004;

Calculation details:
- in the paper

Major advantage:
- fits on physical interactions

Major disadvantage:
- (high diversity in calculation)

MDFV: Molecular Descriptors Family - Vertex

Code of MDFV descriptors

Gene	D_O	A_P	I_D							S_F	S_M	I_T	E_U	L_O
Genome	T	C	J	R	N	Z	V	I	D	A	A	f	D	I
	G	H	j	r	n	z	v	i	d	a	a	F	d	R
		M	O	K	W	S	F	A	0	I	I	c		L
		E	o	k	w	s	f	a	1	i	i	C		
		Q	P	L	X	T	G	B	2	F	F	p		
		L	p	l	x	t	g	b	3	P	P	P		
		A	Q	M	Y	U	H	C	4	C	C	a		
			q	m	y	u	h	c	5			A		
									6			i		
									7			I		

$MDFV = D_O \times A_P \times I_D \times S_F \times S_M \times I_T \times E_U \times L_O$

Ex.: TEuIFFDL and GLbIAcDR

Ref: Bolboacă & Jäntschi 2009

Calculation details:

- in the paper

Major advantage:

- Descriptors derived for molecules losing one atom

Major disadvantage:

- Family too big

SAPF: Structural Atomic Property Family

Code of SAPF descriptors

Gene	C_F	D_O	A_P	D_P	P_P	O_M	M_P	L_O
Genome	D	T	C	I	I	S	I	I
	P	G	H	E	E	M	E	A
	C		M	H	H		H	S
			E	G	G		G	T
			A	A	A		A	Q
				Q	Q		Q	R
				S	S		S	L

$SAPF = L_O \times G_M \times O_M \times P_P \times D_P \times A_P \times M_D \times C_F$

Ex.: SISHQEGC and TESHIMGP

Ref: Sestraş et al., 2012

Calculation details:
- in the paper

Major advantage:
- fits on physical interactions

Major disadvantage:
- low diversity in calculation

SMPI: Szeged Matrix Property Indices
FMPI: Fragments Matrix Property Indices

Code of SMPI descriptors

Gene	A_P	D_M	I_D	M_O	L_O
Genome	A	T	E	m	I
	B	G	U	M	R
	C	U	D	I	L
	D		P	J	
	E			E	
	F			F	
	G				

$SMPI = L_O \times M_O \times I_D \times D_M \times A_P$

Ex.: ImETA (first), LFPUG (last)

Ref: Bolboacă & Jäntschi 2016

Code of FMPI descriptors

Gene	F_C	A_P	D_M	I_D	M_O	L_O
Genome	S	A	T	E	m	I
	M	B	G	U	M	R
	N	C	U	D	I	L
		D		P	J	
		E			E	
		F			F	
		G				

$FMPI = L_O \times M_O \times I_D \times D_M \times A_P \times F_C$

Ex.: ImETAS (first), LFPUGN (last)

Ref: SMPI + $F_C \rightarrow$ FMPI

Calculation details: in the paper

Major advantage: fits on physical interactions

Major disadvantage: small sized families

Unique feature: freely online available for calculations

Software & data analysis (v.2016)

- [01_generate]
  - [mdf]
    - mdf_2004.php
    - mdf_2015.php
    - mdfv2008.php
  - [other]
    - chfp2015.pas
    - fmpi2015.php
    - sapf2011.pas
    - smpi2014.php
  - [compactize]
    - mdfx_compactize.php
    - otherscompactize.php
- [02_filter]
  - sort_all.php
  - families_join.php
- [03_properties]
  - gen_property_files.php
  - chi_distribution.php
- [04_regressions]
  - r1v_all.pas
    - $Y \sim aX + (b)$
  - r2f_all.pas
    - $Y \sim aX_1X_2 + (b)$
    - $Y \sim aX_1 + bX_2 + (c)$
    - $Y \sim aX_1 + bX_2 + cX_1X_2 + (d)$

Analysis steps

- Molecular modeling of structure with dedicated software
  - Topology, geometry & partial charges required from
- Generation of the molecular descriptors
  - [01_generate]Software from {[mdf] or [other]} & [compactize] → {fam_name}_r.asc file
- Filtering (no identical responses allowed)
  - [02_filter] → {fam_name}_t.asc file
- Structure-property association files
  - [03_properties] & properties.asc → {fam_name}_{prop_name}.txt file
- Regression
  - r1v_all.exe & {all}.txt → r1_{all}.txt
  - r2f_all.exe & r1_{all}.txt → rX_r1_{all}.txt

Features (2016)

- Revising of the molecular modeling strategy to import partial charges from post-HF and DFT calculations & to use Spartan i/o files (in preliminary stage)
- Standardization of the input & output for each family calculation (at stage [01_...])
- Possibility to join the pools of families descriptors (at stage [02…])
- Parallelization of all software (all stages) to use any arbitrary CPU number value (very useful for multi-core computers)

Conclusions – take away issues

- The estimating (& predicting) power of a model depends on the pool from which was extracted (‘the family’ here).
- Both diversity & size of the pool counts.
- Prerequisites (tests on the sampled data) and post-requisites (tests on the obtained association models) are both essential for the outcome of the analysis
- Applicability domain of a model is to be obtained from external sets by feeding the model with it and assessing the outcome

Thank you for your attention!

- Questions?