Let's develop together an exchange virtual compounds computational platform

Francesco Ortuso

University "Magna Græcia" of Catanzaro
Italy

1st annual meeting of MuTaLig COST Action CA15135, Lugano 22 July 2016

Some (trivial) data

Industrial research

(a quasi-failure story)

- the 11.8% of under development drugs will arrive to the market[1]
- ~25-30 new drugs/year [2]
- 200-250 candidates/new drug fails due to ADME-Tox reasons
- molecules/candidate = hundreds
- 50K-75K failed candidates/year corresponding to 500K-750K discarded molecules/year

Academic research

(a full-failure story)

- more than 100,000 scientific papers/year in Chemistry Research field (ISI)
- tens of candidate drugs/year
- any drug on the market

1 DiMasi, J.A.; Grabowski, H.G; Hansen, R.W. *N. Engl. J. Med.* **2015**, *20*, 1972
2 Kinch, M.S.; Haynesworth, A.; Kinch, S.L.; Hoyer, D. *Drug Discov Today*. **2014**, *19*, 1033-1039

What about failed compounds?

Pharmaceutical Industries

- try drug-repurposing
- publish some scientific paper
- build commercial databases (to sold failed molecules)
- forget failed molecules in storage areas

Universities

- completely forget failed molecules
- try drug-repurposing, only recently and in very few cases

Some already available database

- ZINC (zinc.docking.org)
- ChEMBL (www.ebi.ac.uk/chembl)
- PubChem (pubchem.ncbi.nlm.nih.gov)
- NCI (dtp.cancer.gov)
- ChemSpider (www.chemspider.com)
- DrugBank (www.drugbank.ca)
- WOMBAT (dud.docking.org/wombat)
- DUD (dud.docking.org)
- CSD (www.ccdc.cam.ac.uk/solutions/csd-system/components/csd)
- eMolecules (www.emolecules.com)
- ChemBank (chembank.broadinstitute.org)
- GILDA (pharminfo.pharm.kyoto-u.ac.jp/services/glida)
- ……..

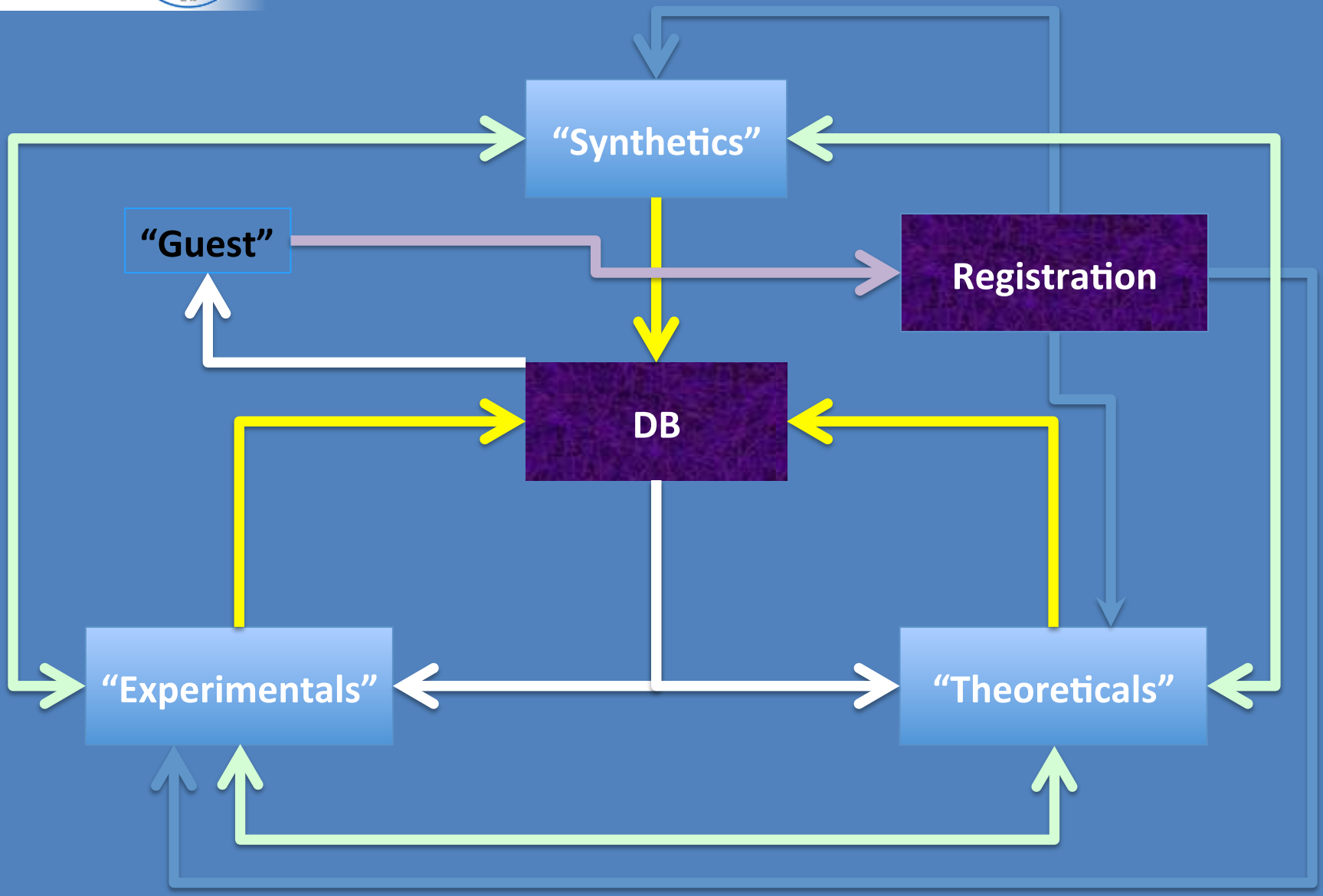
Does it make sense yet another DB?

- available DBs contain million of compounds
- in terms of both computational and chemical resources, available DBs are well done
- they allow to get molecules + some “information” about them
- it is very hard to make something better, but ….
- available DBs are “closed”
  - Vendors furnish compounds (virtually and materially)
  - Researchers use compounds (and, usually, buy them from vendors)
  - Any feedback comes back from new tests (usually)
  - Any collaboration is promoted between vendors and researchers (different goals)
  - Any collaboration is promoted among researchers

What Mu.Ta.Lig Chemotheca could be

- A community based "open" DB based for scientific data sharing:
  - Web accessible both by registered and guest users (with different level of information allowed).
  - Registered users will be:
    a) "Synthetics", who synthetized and uploaded the molecule (i.e. WG1)
    b) "Experimentals", who retrieved the molecule and tested it (i.e. WG2)
    c) "Theoreticals", who retrieved the molecule and used it for molecular modelling investigation (i.e. WG4)
  - "Synthetics" can upload their own molecules and decide the policy to access them.
  - "Experimentals" interested on a certain molecule, will contact the "Synthetics" owner and will take accord with him for obtaining the hit.
  - Chemotheca will require to "Experimentals" to update activity information related to retrieved compounds.
  - "Theoreticals" will have access to the Chemotheca and they make a commitment to insert their theoretical information.

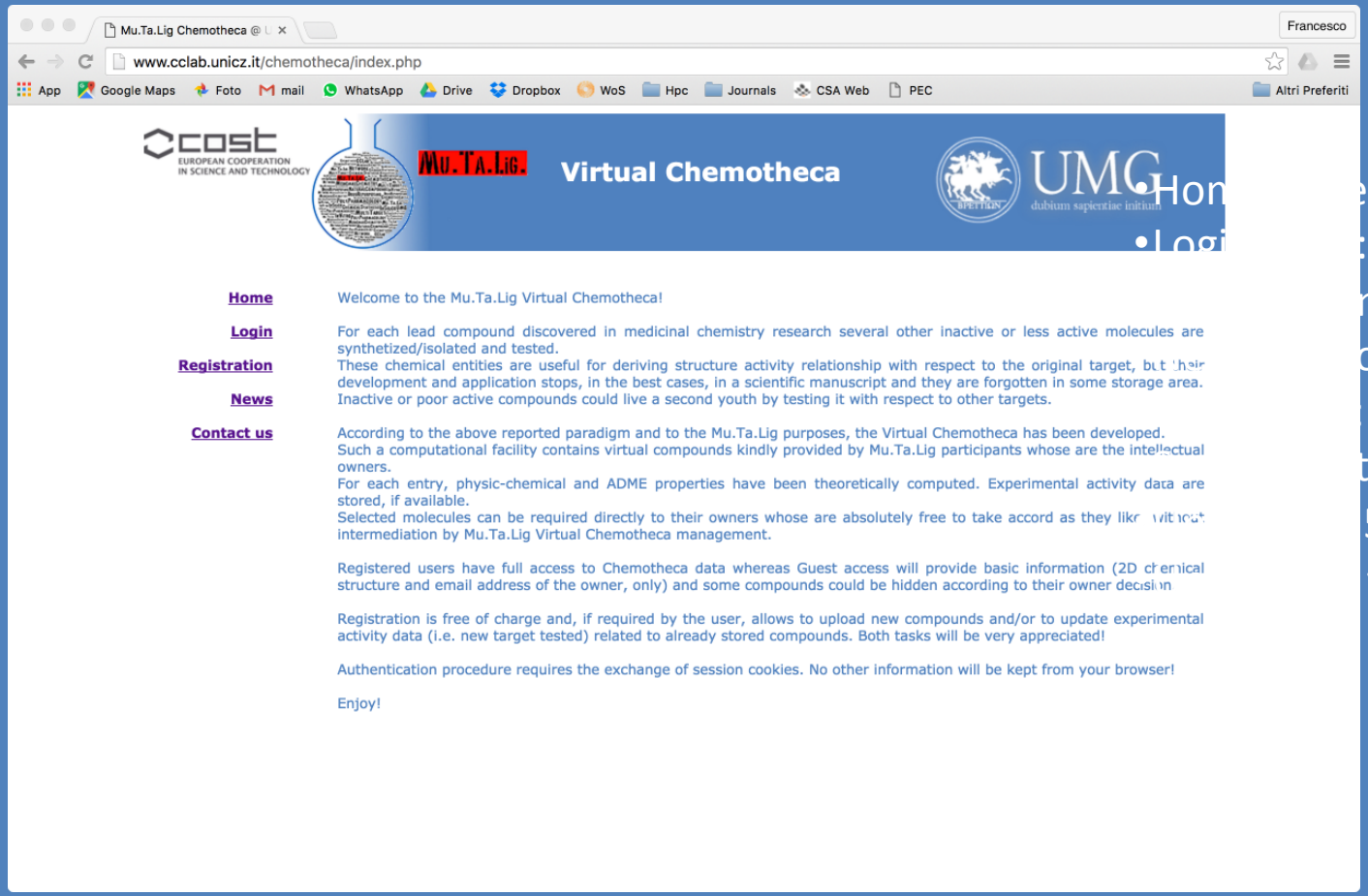
Chemotheca's network



How Chemotheca is and could be developed

- The entire tool is (and will be) based on open-source programming languages and architectures:
  - Multiprocessor Linux computer servers
  - Apache http demon (ver. 2.2)
  - Web user interface frontend is written in PHP (ver. 5.1)
  - Cookies will be adopted for session data
  - Database backend is MySQL (ver. 5)
  - JSME (or similar) molecule editor for searching and updating uploaded structures

What's the development state of Chemotheca?



: done!
: to do
n: 50% done
o
to do
to do
50% done
to do

The minus was done!

What information could be showed

Visible to Guest users

- Molecular structure (if allowed by the owner)

Visible to registered users only

1. Molecular structure and owner data
2. Biological information (i.e. already tested targets and corresponding activity)
3. Physico-chemical experimental information (i.e. MW, solubility, melting point)
4. Theoretical descriptors (i.e. Lipinski rule, LogP, LogD, tautomers, conformers)

Points 2, 3 and 4 will indicate “who” inserted the data and “when” it has been done

Few sharing rules

- Who virtually upload a molecule will be always its owner and he will be free to take accords for compounds delivery.

- Biological, physico-chemical and theoretical data become property of Chemotheca's registered users BUT experimental/computational details must be required to the data producers and they will be free to take accords for furnishing such supplementary information.

Nobody will lose the intellectual property on its own data

and

some molecule could live a second youth!

Chemotheca's espected results and requirements

- A potentially large number of available high quality compounds will be classified
- Each compound will include one or more activity data:
  - Improved knowledge related to polypharmacology and potential secondary/ side effects
- Scientific collaboration among Chemotheca's users will be promoted

Chemotheca's goals are strictly related to the users contribute!

What's the easier way for contributing

- "Synthetics": compounds upload
- "Experimentals": compounds activities update
- "Theoreticals": compounds descriptors, theoretical activities update

Compounds upload

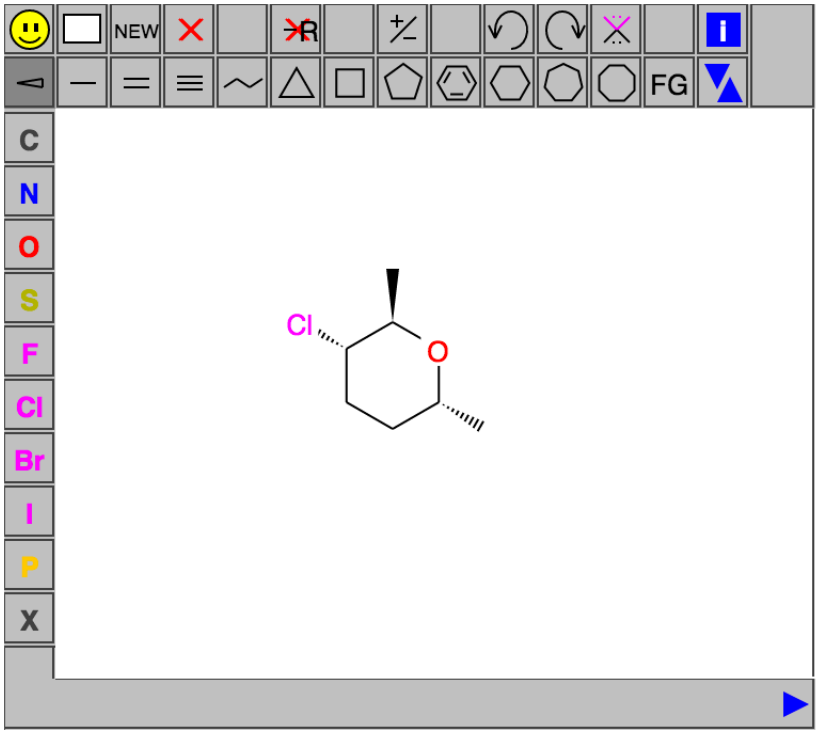
- Mandatory information:
  - ID (decided by the owner for its easier compound identification)
  - Molecular structure (including stereochemistry)
  - Stereochemistry
  - Is it a salt? y/n
- Other information (if available):
  - Melting point (°C)
  - Water solubility (g/L)
  - DMSO solubility (g/L)
  - Ethanol solubility (g/L)

- Upload ways
  - Web form
  - Spreadsheet (xlsx, ods or csv)
  - ….

Web form

Compound's upload service

*required field!



*Chiral centers? ---------

*E and Z isomers? ---------

*Do you have designed the right chirality and/or double bond configuration? ---------

*Will be it available as a purified stereoisomer or as a mixture? ---------

*Is it a salt? ---------

Melting point °C

Water solubility g/L

DMSO solubility g/L

Ethanol solubility g/L

Upload compound

Using spreadsheet

- one row -> one compound
- 1st row contains column labels
- information start at row n. 2
- Column A: compound uniq ID (not blank space are allowed)
- Column B: compound InChI key
- Column C: compound structure as SMILE string (be careful to the stereochemistry)
- Column D: is it salt (y/n)
- Column >= E: other information (if available)

C13

	A	B	C	D	E	F	G	H
1	ID	InChI key	SMILE	Salt	Melting (°C)	Water Sol (g/L)	Ethanol Sol (g/L)	DMSO Sol (g/L)
2	abc123	YBDQLHBVNXARAU-ZCFIWIBFSA-N	C[C@H]1OCCCC1	n				
3	aaa2	WPYMKLBDIGXBTP-UHFFFAOYSA-N	O=C(O)C1=CC=CC=C1	n	122	2,9	58,4	
4	qwe	YGSDEFSMJLZEOE-UHFFFAOYSA-N	O=C(O)C1=C(O)C=CC=C1	n	159	2,24	348,7	
5	449	BSYNRYMUTXBXSQ-UHFFFAOYSA-N	O=C(O)C1=C(OC(C)=O)C=CC=C1	n	135	4,6		
6	Abt321	JZLOKWGVGHYBKD-UHFFFAOYSA-M	O=C([O-])C1=C(OC(C)=O)C=CC=C1.[Na+]	y				
7								
8								

After spreadsheet upload check

- A table reporting uploaded information and 2D compound structures will appear
- Please verify the stereochemistry and, in case of error, will be possible to edit or delete the entry
- If the compound is a salt, don't care about the ions relative position
- When everything sounds….. Go ahead!

The code has to be written…

Compounds activity data editing

- By means of search UI, registered users will select molecule(s) of interest
- Registered users will can add new information (i.e. Target IC_{50} in nM)
  - It will be mandatory to specify: (a) target, (b) kind of measurement and (c) measure unit (i.e. nM)
- Added information will be available for all registered users but can be edited by their owner only
- If a certain compound will be retired by its owner, it will remain into the Chemotheca but it will marked as “unavailable”

Chemotheca's usage

- Users will be allowed to search the DB by means of:
  - Textual web form (TWF), using the compound properties and boolean operators
    - Compound ID, InChI key, Molecular weight, Lipinski rule (or other available descriptors), target, kind of activity (IC_{50}, Ki, ED_{50})
  - A molecular editor (ME), using exact or substructure search algorithm
- TWF and ME could be used together in advanced search