Optimisation problems

encountered during our study of pharmacophores

Nhat-Vinh Vo[‡] Bertrand Cuissart[‡] Ronan Bureau[†] Alban Lepailleur[†]

‡ **GREYC – FR CNRS UMR 6072**
† **CERMN – UPRES EA 4258 – FR CNRS 3038 INC3M**
firstname.lastname@unicaen.fr

May 23, 2019

AGAC
Analyse de Graphes Appliquée à la Chémoinformatique
Financial support from the Normandy region

Plan

- Context, Objective and Principle
- Optimisation problems
  - Compute Graph Edit Distance
  - Build selected training set
- Conclusion and Perspective.

Context

Terminologies:

- Ligand: an organic molecule that binds with a biomolecule
- Target: a part of the organism to which a ligand binds, producing a change of behaviour. Proteins is a common class of targets.

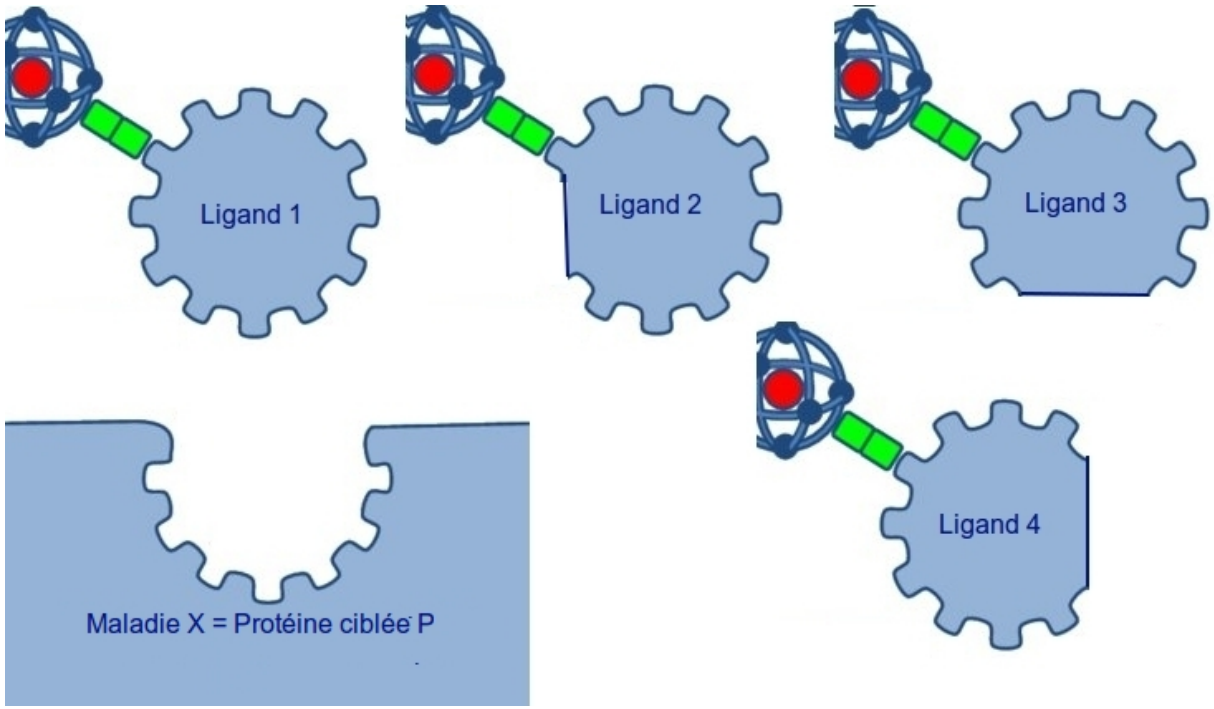


Figure : Ligand-Target Model.

Context

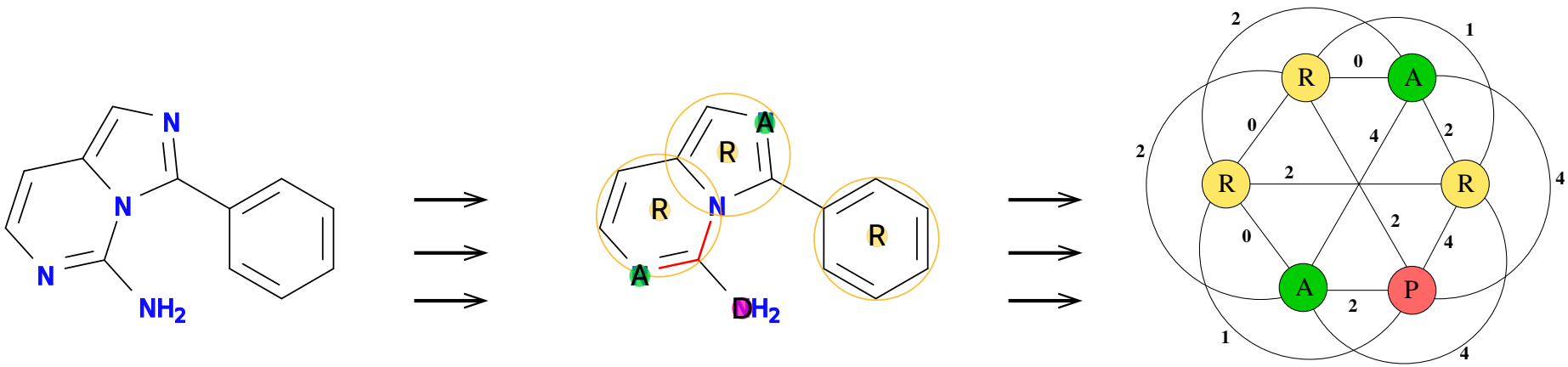
Terminologies:

- Pharmacophore: a part of a molecular structure that is responsible for a particular biological or pharmacological interaction that it undergoes.

Pharmacophore Graph includes the following features

- Hydrogen Bond **A**cceptor, Hydrogen Bond **D**onor, aromatic **R**ing, **H**ydrophobic area, **P**ositively ionizable group, **N**egatively ionizable group[a]
- an adjustable part of the input.

[a] OpenBabel, N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, J. Cheminformatics 2011, 3, 33.



Skeletal formula | Feature occurrences | Pharmacophore graph

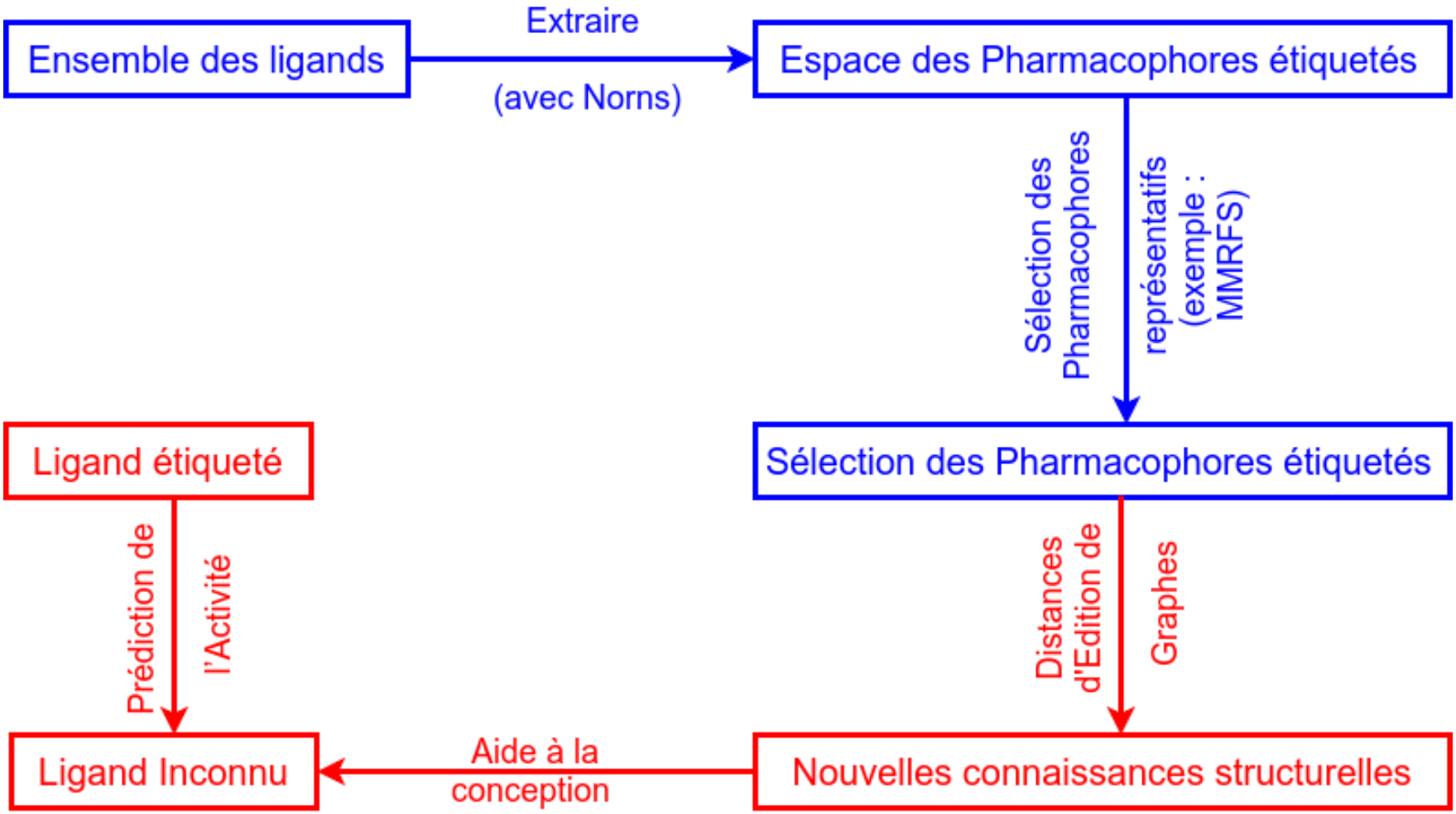
Objective

- Input:
  - Learning set of ligands whose structures and biological profiles are known
  - Ligand L whose structure is known but not its profile
- Output:
  - Prediction of the biological profile of L.

Principle

- Compute similarity between studied ligands
- Elaborate a mechanism for a supervised classification
- Apply the supervised classification to the candidate ligand of which the bioactivity has not been known yet.

Principle


Ensemble des ligands
Extraire
(avec Norns)
Espace des Pharmacophores étiquetés
Sélection des Pharmacophores représentatifs (exemple : MMRFS)
Sélection des Pharmacophores étiquetés
Distances d'Edition de Graphes
Nouvelles connaissances structurelles
Aide à la conception
Ligand Inconnu
Prédiction de l'Activité
Ligand étiqueté

Optimisation problems

- Compute Graph Edit Distance
- Build selected training set

Compute Graph Edit Distance

Graph edit distance

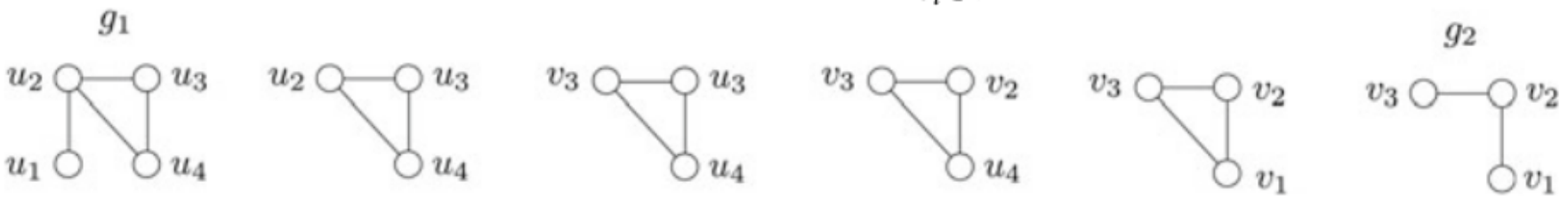
- measures distances between two graphs g_1 and g_2 by the amount of distortion that is needed to transform g_1 into g_2.[a]

[a] Graph Edit Distance, K. Riesen, Structural Pattern Recognition with Graph Edit Distance 2015, p29-45.

Definition 2.1 (*Edit Path*) A set $\{e_1, \ldots, e_k\}$ of k edit operations e_i that transform g_1 completely into g_2 is called a *(complete) edit path* $\lambda(g_1, g_2)$ between g_1 and g_2. A *partial edit path*, i.e., a subset of $\{e_1, \ldots, e_k\}$, edits proper subsets of nodes and/or edges of the underlying graphs.

Definition 2.2 (*Graph Edit Distance*) Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ be the source and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ the target graph. The *graph edit distance* $d_{\lambda_{\min}}(g_1, g_2)$, or $d_{\lambda_{\min}}$ for short, between g_1 and g_2 is defined by

$$d_{\lambda_{\min}}(g_1, g_2) = \min_{\lambda \in \Upsilon(g_1, g_2)} \sum_{e_i \in \lambda} c(e_i), \tag{2.1}$$



Compute Graph Edit Distance

- Compute Graph Edit Distance using a Branch-and-Bound procedure *[Abu-Aisheh et al., 2017]*[1]
- Graph Edit Distance vs. Optimisation problems:
  - Graph matching problem
  - Load balancing problem
  - Branch-and-Bound technique: branching scheme, upper bound.

[1] Abu-Aisheh, Raveaux, Ramel, Martineau - 2017 - "Parallel Graph Edit Distance" - Expert Systems with Application, p41:57.

Compute Graph Edit Distance

- Compute similarity between ligands using Graph Edit Distance (GED)
- Apply GED to 1492 ligands (from 8 to 13 vertices) from a ligand set for ABL-target

  $\longrightarrow$ It exists instances without any optimality within 10 computation minutes.

Example: H-H-H-4-3-3 vs. A-H-H-H-4-3-3-4-7-8

$\longrightarrow$ Assumption: good cost configuration

$\longrightarrow$ 3 Node substitutions and 1 Node insertion: $3! \times 4 = 24$ possibilities

- 1 Node substitution: H $\rightarrow$ A
- 1 Node insertion: H
- 3 Edge insertions: A-H (3); H-H (7); H-H (8)
- 3 Edge substitutions: H-H (4) $\rightarrow$ A-H (4); H-H (3) $\rightarrow$ H-H (4); H-H (3) $\rightarrow$ A-H (3).

Compute Graph Edit Distance

- Apply **Norns** (by projet Minomics) to the dataset to extract a space of pharmacophores which satisfy the following constraints *[Métivier et al., 2018]*[2]:
  - Size of graph: from 1 to 6 vertices
  - Support/Frequency: at least 10
  - Growth Rate: used to classify pharmacophores
  - With and without MMRFS algorithm

  $\longrightarrow$ Without MMRFS (S_1): **87175** pharmacophores

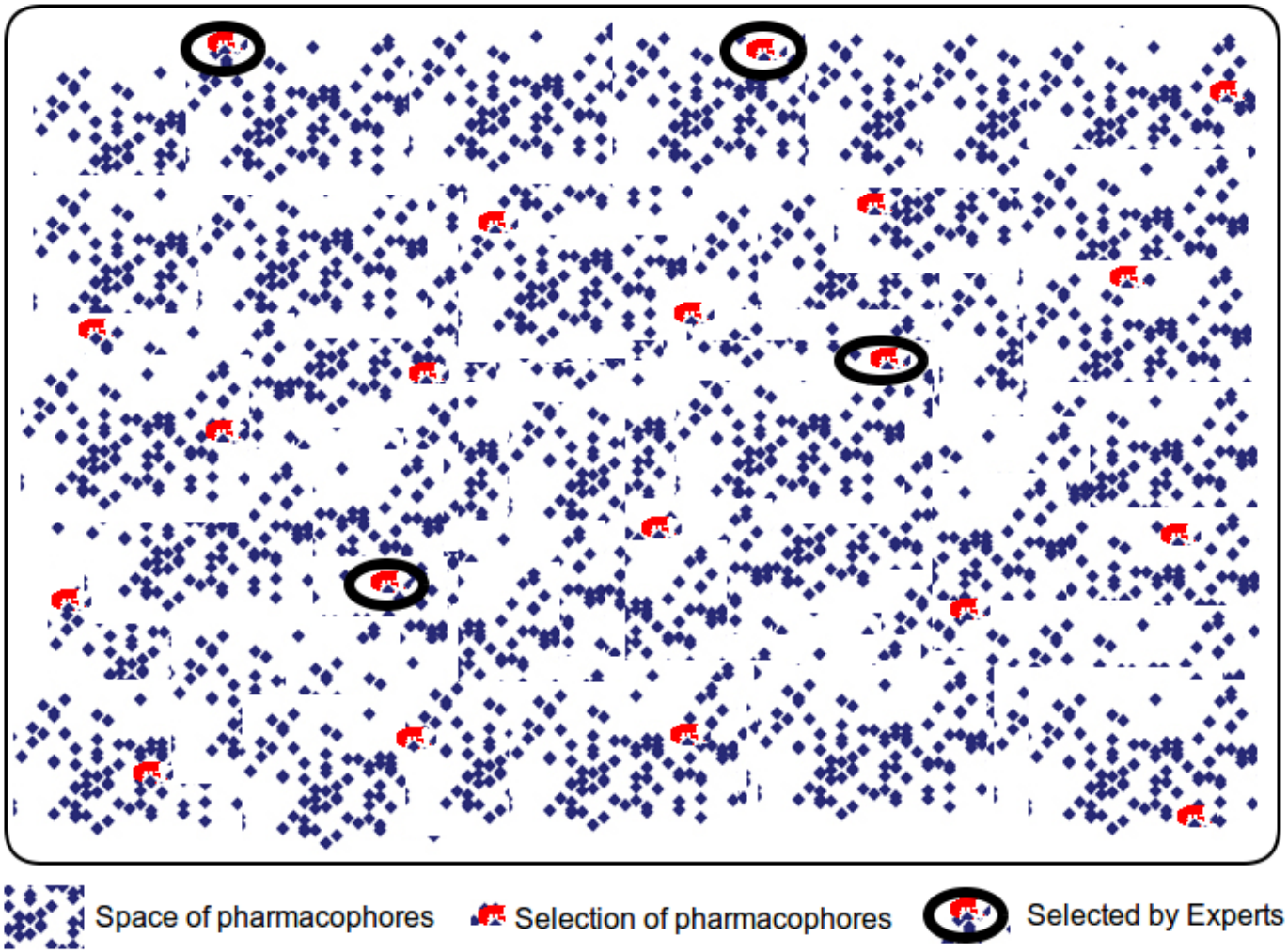
  $\longrightarrow$ With MMRFS (S_2): 137 pharmacophores
- Edit Cost Configuration:
  - Node Deletion/Insertion: 6
  - Node Substitution: 10
  - Edge Deletion/Insertion: 0
  - Edge Substitution: *difference*(edge1,edge2)

  $\longrightarrow$ Computation time for GED: **0.5 second**

[2] Métivier, Cuissart, Bureau, Lepailleur - 2018 - "The Pharmacophore Network: A Computational Method for Exploring Structure-Activity Relationships from a Large Chemical Data Set" - Journal of Medicinal Chemistry, Vol.61, p3551:3564.

- Compute Graph Edit Distance
- **Build selected training set**

Build selected training set



Objective: build selected training set including:

- Selection of pharmacophores
- Pharmacophores best relating to pharmacophores selected by experts.

Build selected training set

Compute generalised-median pharmacophores $\overline{P}$ to determine pharmacophores best relating to pharmacophores selected by experts:

- Space of all pharmacophores $\mathbb{G}$
- Selection of pharmacophores Sel
- Set of pharmacophores selected by experts $S = \{P_1, P_2, \ldots, P_{|S|}\}$

$$\overline{P} = ARGMIN_{P \in \mathbb{G} \setminus Sel} \sum_{P_i \in S} GED(P, P_i) \quad (1)$$

$\longrightarrow$ How to deal with this problem?

Conclusion and Perspective

- Use GED as similarity between ligands and between pharmacophores
- Apply a Branch-and-Bound procedure to GED computation

  $\longrightarrow$ Is there any better way (computation time) for an exact solution?
- Compute generalised-median pharmacophores to build selected training set

  $\longrightarrow$ How to deal with large search space?