WAVE TOP-K RANDOM-D FAMILY SEARCH

Interactive exploration of a space of structured patterns

Etienne Lehembre[1], Bruno Cremilleux[1], Bertrand Cuissart[1], Abdelkader Ouali[1] et Albrecht Zimmermann[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

etienne.lehembre@unicaen.fr

GREYC Laboratoire de recherche en sciences du numérique

Normandie Université

UNICAEN

ENSI CAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs

PROJET FINANCÉ PAR L'ANR PROJECT FUNDED BY THE ANR

Plan de l'exposé

Subgraph mining and chemoinformatics

Data :

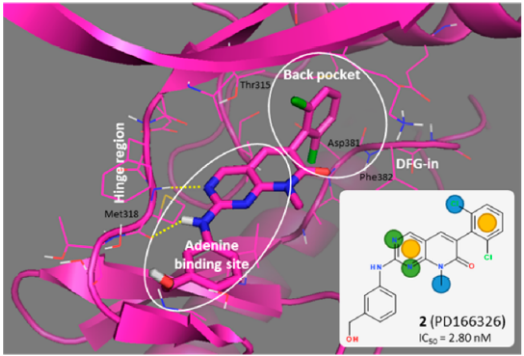
- 1 receptor : BCR-ABL
- 1485 molecules

Structured pattern :

- 112 363 pharmacophores
- 15 533 equivalence classes

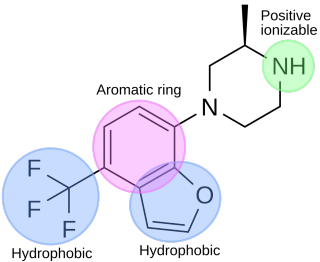
Method goal :

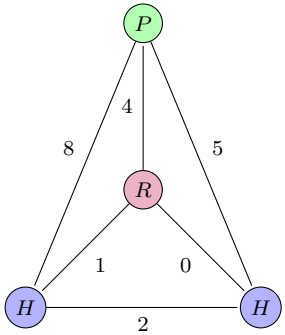
Guide an experts' exploration of the pharmacophores set.



J.-P. METIVIER, B. CUISSART, R. BUREAU, and A. LEPAILLEUR.

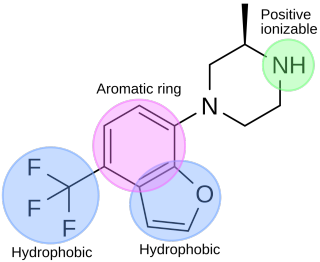
From molecules to pharmacophores (1)

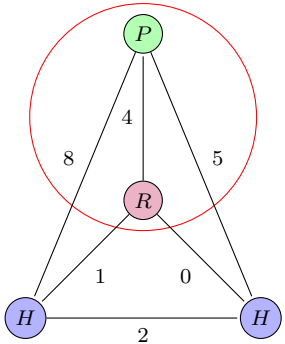




- Pharmacophoric features = pharmacophore vertices.
- Distances between features = labels of the pharmacophore edges.

From molecules to pharmacophores (2)





- The subgraph of a pharmacophore is a pharmacophore.
- We can build a partial ordered graph from the pharmacophore set.

Plan de l'exposé

1. Context and stakes

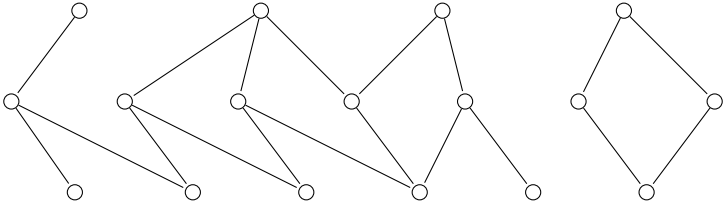
2. Method presentation

3. Experiences and evaluation

Proposed interactions

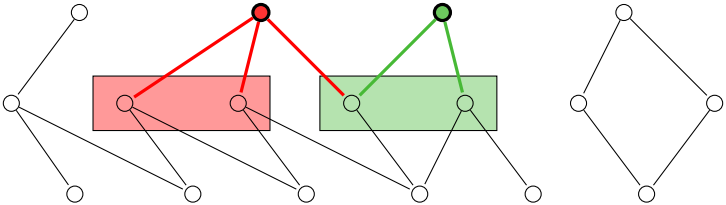
Interaction	Consequence·s	Color
Accepted	Prioritized area & Weights augmentation	Green
Interested	Weights augmentation	Blue
Uncertain	–	Purple
Not-interested	Weights diminution	Orange
Rejected	Exclusion area & Weights diminution	Red

Example on a partially ordered graph



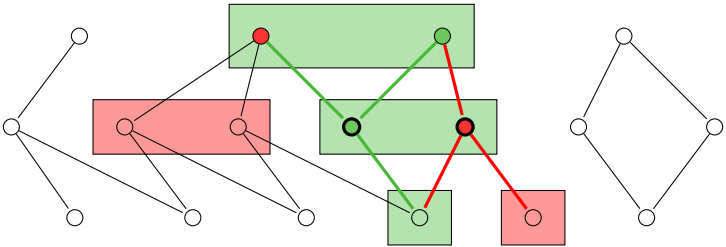
- Partially ordered graph as an input of the method.
- Each vertex is a pharmacophore.

Definition of the search areas (1)



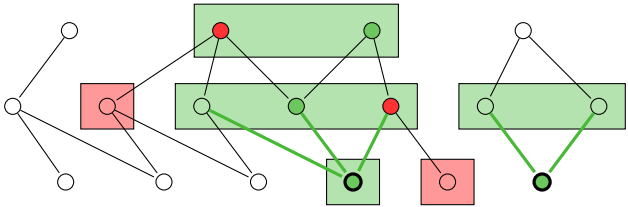
- **Interactions** : rejected (red) and accepted (green).
- Exclusion area in red, prioritized area in green.
- In case of **conflict**, the prioritized areas win.
- Areas are only defined on the **directs descendants**.

Definition of the search areas (2)



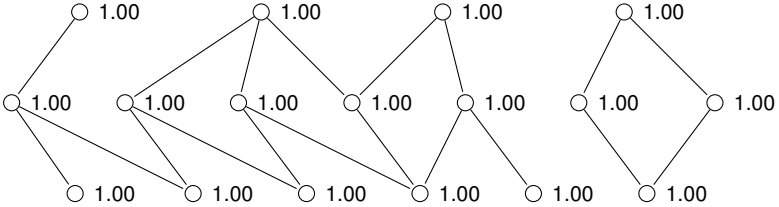
- Excluded and prioritized areas are defined on direct ancestors and descendants.
- There is not enough pharmacophores available in prioritized areas, we must offer a pharmacophore from an unbiased area.

Definition of the search areas (3)



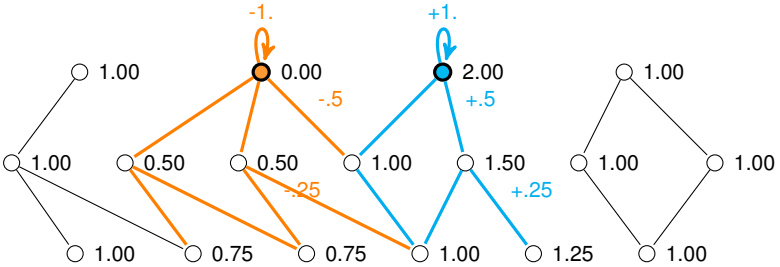
- Prioritized areas are defined on the direct ancestors.
- The search has reached the last layer. It goes up to the more generic patterns.
- **Question** : how to we choose the pharmacophores to propose ?

Weights diffusion (1)



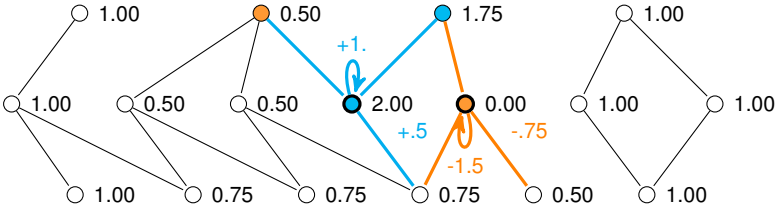
- The POG carries no interaction, every pharmacophores have the same weight.

Weight diffusion (2)



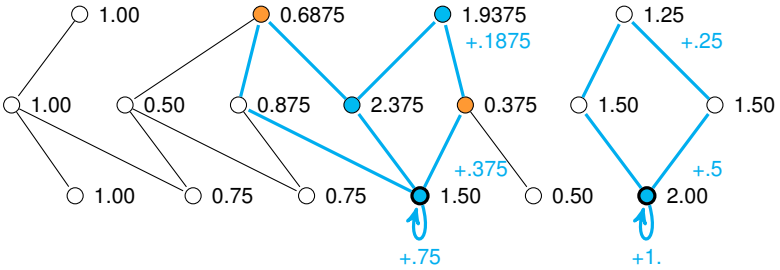
- **Interactions** : rejected or not-interested (orange) and accepted or interested (blue).
- Diffusing to the descendants : -1 for the negative interest and $+1$ for the positive interest.
- The raw interest diffused decreased with the increase of the distance from the origin of the interaction.

Weight diffusion (3)



- Diffusing interest to ancestors and descendants.
- Diffused interest is the **weight** of the pharmacophore carrying the interaction.

Weight diffusion (4)



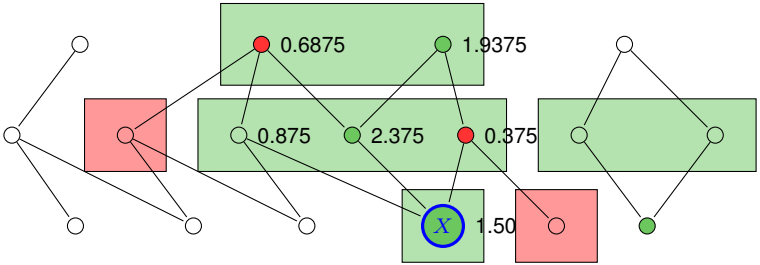
- Diffusing interest to the ancestors.
- The raw interest diffused decreased with the increase of the distance from the origin of the interaction.

Computing potential interest (1)

$$f_p(v,\mathbb{G}) = \sum_{i=0}^{k}(Weight(L_i^+(v,\mathbb{G})) + Weight(L_i^?(v,\mathbb{G}))+ \\ (Weight(L_i^\star(v,\mathbb{G})) - |Weight(L_i^+(v,\mathbb{G})) - Weight(L_i^-(v,\mathbb{G}))|)) * \frac{1}{2^i} \quad (1)$$

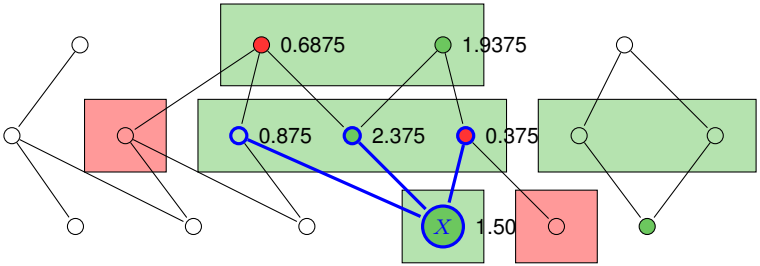
- $Weight(L_i^+(v,\mathbb{G}))$ is the **exploitation** heuristic.
- $Weight(L_i^?(v,\mathbb{G}))$ is the **exploration** heuristic.
- $(Weight(L_i^\star(v,\mathbb{G})) - |Weight(L_i^+(v,\mathbb{G})) - Weight(L_i^-(v,\mathbb{G}))|)$ is the **ambiguity** heuristic.
- $1/2^i$ is the distance modifier.

Computing potential interest (2)



$$f_p(X) = 1.50 \tag{2}$$

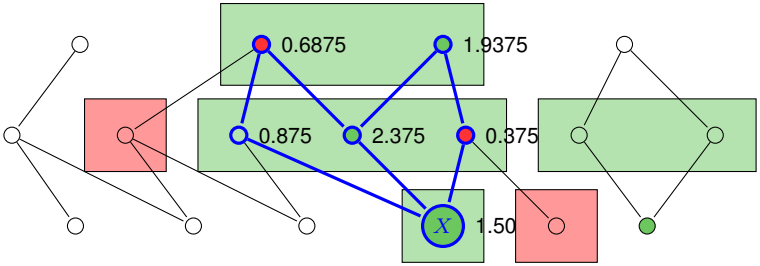
Computing potential interest (3)



$$f_p(X) = 1.50 \tag{3}$$

$$+ (2.375 + 0.875 + (2.375 + 0.375 - |2.375 - 0.375|)) * \frac{1}{2} \tag{4}$$

Computing potential interest (4)



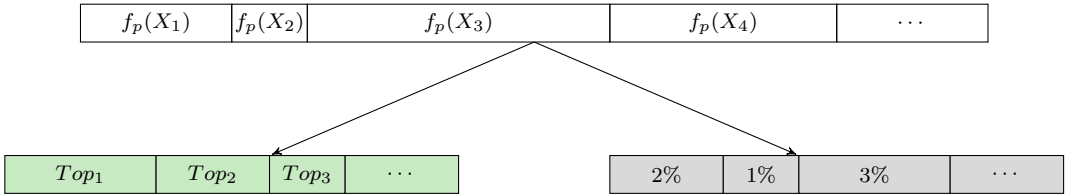
$$f_p(X) = 1.50 \tag{5}$$

$$+ (2.375 + 0.875 + (2.375 + 0.375 - |2.375 - 0.375|)) * \frac{1}{2} \tag{6}$$

$$+ (1.9375 + (1.9375 + 0.6875 - |1.9375 - 0.6875|)) * \frac{1}{4} \tag{7}$$

Potential interest and sampling

Unexplored layer of the partially ordered graph containing the pharmacophores.



Prioritized area (Drawing '$Top-k$' patterns)

Unbiaised area (Drawing '$Random-d$' pattern

- Prioritized area : we select patterns with the **highest potential interest**.
- Unbiased area : potential interest is converted into a **selection probability**.

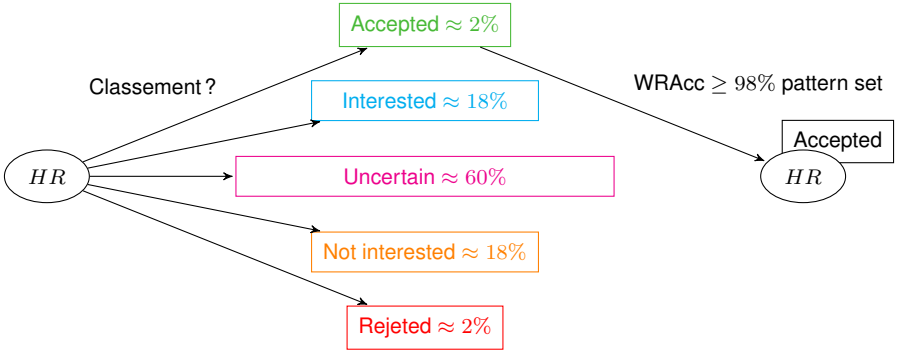
Plan de l'exposé

1. Context and stakes

2. Method presentation

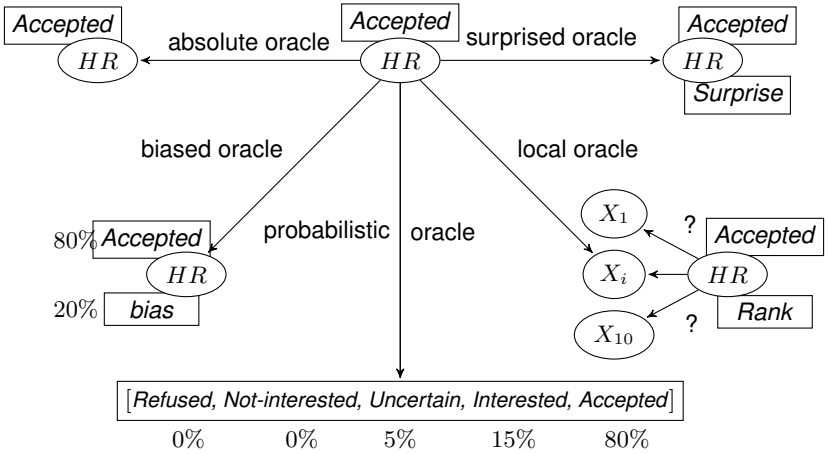
3. Experiences and evaluation

Oracles definition (1)



- The HR WRAcc is compared to the WRAcc of the studied pharmacophore set.

Oracles definition (2)



Recall graphs

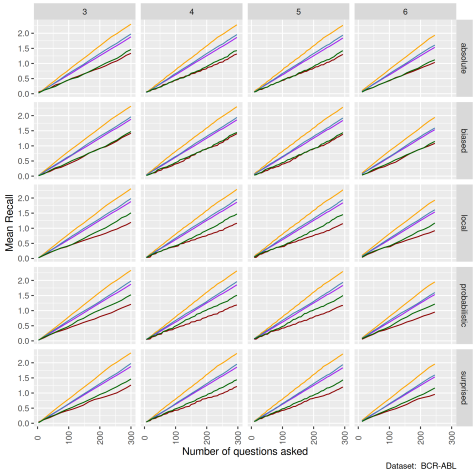


FIGURE – Random Sampling

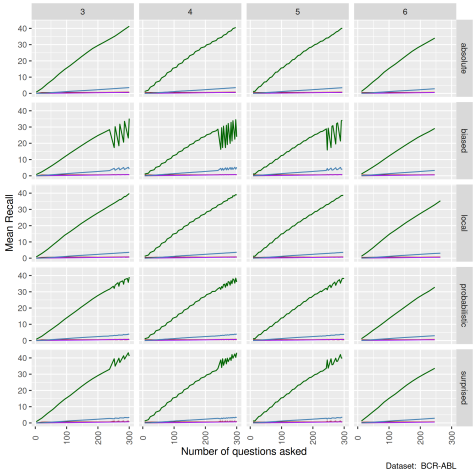
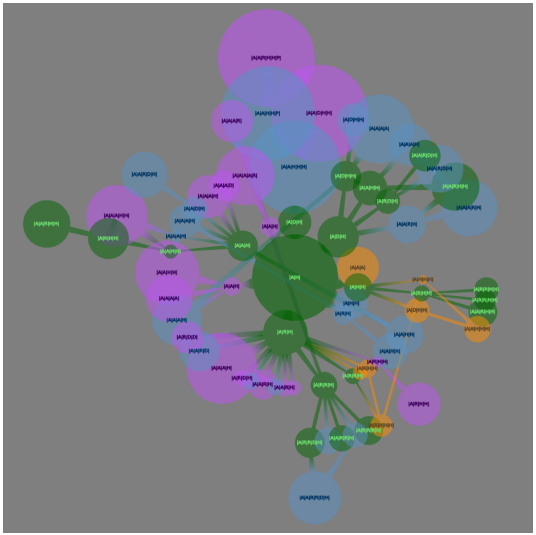


FIGURE – WTRFS

Interests' partially ordered graph



Conclusion and future works

Conclusion

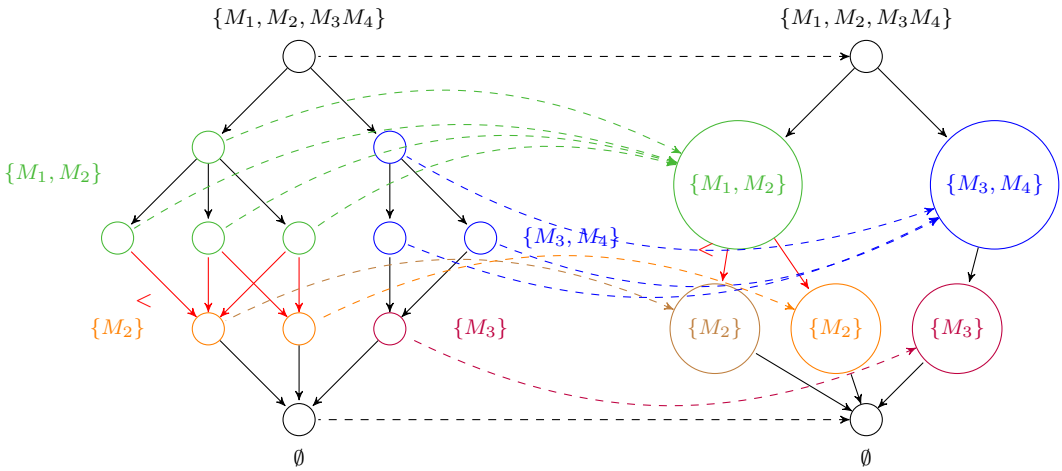
- Creation of the Interests' partial order graph.
- Interactive exploration without descriptor.
- New ideas for the evaluation process.

Future works

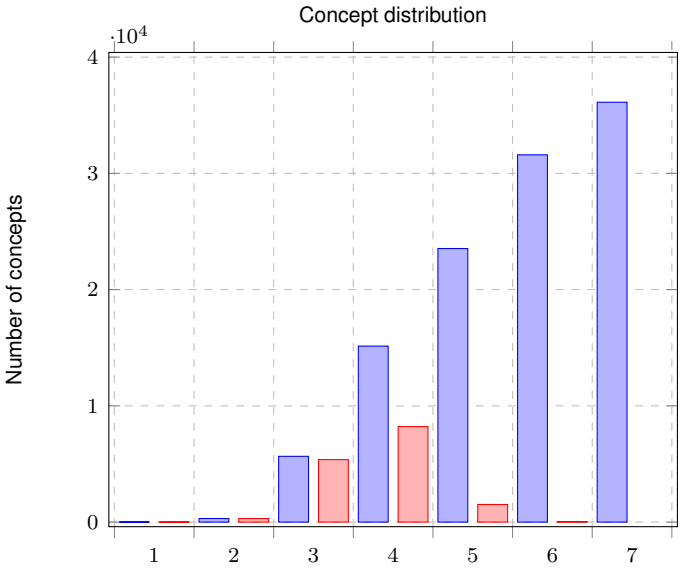
- Working on the visual representation (GUI and more).
- Adapting the algorithm to a mining task.
- Creation of new heuristics.

Annexes

Structured Equivalence Classes



BCR-ABL : patterns distribution by layers



Case study : selecting an outstanding pattern