Ranking Big Data Sets using Rank Aggregation Techniques

Sarah Cohen-Boulakia

Laboratoire de Recherche en Informatique

CNRS UMR 8623, Université Paris-Sud
Université Paris-Saclay

General context

- Large set of data items can be obtained
  - as answer to a query, as produced by a tool…
- **Ranking** data items is crucial
- Ranking data may be difficult: **how** to rank?
- Alternative **ranking criteria** may be considered
  - (Quality-like) Reliability, Completeness…
- Alternative **methods** may rank the same set of data items
  - Google vs Yahoo vs …
  - Classifiers…

Several rankings taken as input to produce one ***consensus ranking*** minimizing their disagreements

$$\pi_1 = [A, D, C, B]$$
$$\pi_2 = [B, A, D, C]$$
$$\pi_3 = [D, A, B, C]$$

$$\pi^* = [A, D, B, C]$$

➔ **Rank Aggregation Techniques**

More formally…

- Rank aggreagation is based on a ***distance*** *to produce the closest ranking from a set of input rankings*
  - **The Kendall−τ D(π,σ) distance**

#pairs of elements **inversed between** two rankings

$$D(\pi,\sigma) = \left| \left\{ \begin{array}{ll} (i,j) : i < j & \wedge \\ \left(\begin{array}{ll} & \pi[i] < \pi[j] \ \wedge \ \sigma[i] > \sigma[j] \\ \vee & \pi[i] > \pi[j] \ \wedge \ \sigma[i] < \sigma[j] \end{array} \right) & \end{array} \right\} \right|$$

$$\begin{aligned} \pi_1 &:= [A, D, C, B] \\ \pi_2 &:= [B, A, D, C] \\ D(\pi_1, \pi_2) &= 1_{A>B} \\ &\quad +1_{B>D} \\ &\quad +1_{B>C} \\ &= 3 \end{aligned}$$

More formally…

- Rank aggreagation is based on a ***distance*** *to produce the closest ranking from a set of input rankings*

  - **The Kendall−τ D(π,σ) distance**

#pairs of elements **inversed between** two rankings

$$D(\pi,\sigma) = \left| \left\{ \begin{array}{l} (i,j) : i < j \ \wedge \\ \left(\begin{array}{rl} & \pi[i] < \pi[j] \ \wedge \ \sigma[i] > \sigma[j] \\ \vee & \pi[i] > \pi[j] \ \wedge \ \sigma[i] < \sigma[j] \end{array} \right) \end{array} \right\} \right|$$

  - **Kemeny Score**

$$S(\pi,\mathcal{P}) = \sum_{\sigma \in \mathcal{P}} D(\pi,\sigma)$$

  - **Optimal Consensus (median)**

$$\forall \pi \in \mathcal{S}_n : S(\pi^*,\mathcal{P}) \leq S(\pi,\mathcal{P})$$

Complexity [Dwork *et al* 2001, Biedl *et al.* 2009, *Bachmeier et al. 2017*] **NP-Difficult → Numerous heuristics**

$$\begin{aligned} \pi_1 &:= [A,D,C,B] \\ \pi_2 &:= [B,A,D,C] \\ D(\pi_1,\pi_2) &= 1_{A>B} \\ &\quad +1_{B>D} \\ &\quad +1_{B>C} \\ &= 3 \end{aligned}$$

$$\mathcal{P} \left\{ \begin{array}{l} \pi_1 = [A,D,C,B] \\ \pi_2 = [B,A,D,C] \\ \pi_3 = [D,A,B,C] \end{array} \right.$$

$$\pi^* = [A,D,B,C]$$

$$\begin{aligned} S(\pi^*,\mathcal{P}) &= 1_{A>B@\pi_2} \\ &\quad +1_{A>D@\pi_3} \\ &\quad +1_{B>C@\pi_1} \\ &\quad +1_{B>D@\pi_2} \\ &= 4 \end{aligned}$$

Our expertise @LRI (Data Science)

- Comparison of 15+ algorithms (exact, approx, heuristics) able to provide consensus rankings

  → Rank-and-ties platform

- Design (or tuning) of efficient rank aggregation algorithms
- Considering alternative distances to compute consensus rankings taking into account user needs, context
  - Candidates to an election → one single position per candidate

$$\pi_1 = [A, D, C, B]$$
$$\pi_2 = [B, A, D, C] \quad \pi^* = [A, D, B, C]$$
$$\pi_3 = [D, A, B, C]$$

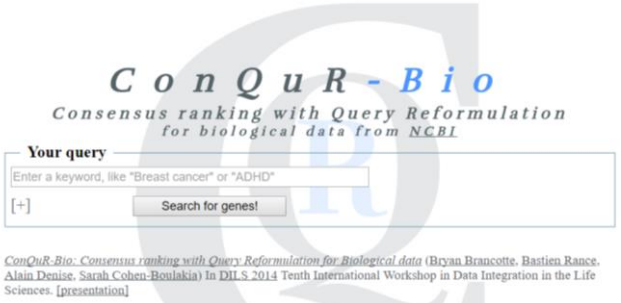
  - Participant to sport competitions → *ex-aequo* should be allowed π_1 = [A, {D,C}, B]…
  - Movies poeple like → not all the movies are ranked by anyone

    π_1 =[A, D, B], π_2= [A, C, B]…

Our expertise @LRI (Bioinfomatics)

- Using rank aggregation techniques to automatically rank results obtained to a query and *equivalent reformulations*

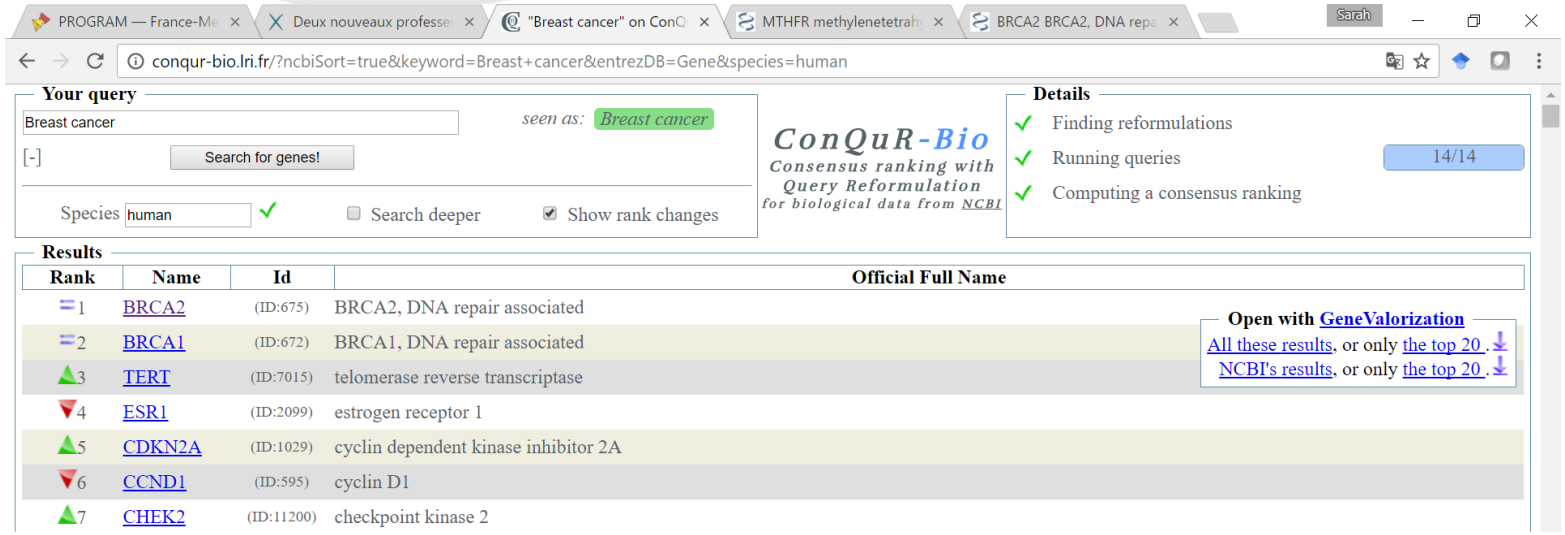
http://conqur-bio.lri.fr



Genes associated with *Breast cancer*? [G1, G2, G3, G4]

Genes associated with *Mamamilian Carcinoma*? [G1, G3, G4, G12]

…



Conclusion

- Rank aggregation techniques are interesting
  - When it is difficult to decide on which criteria to rank
  - When you have several rankings and want to highlight their common points
- We have expertise on
  - Design new rank aggregation algorithms
  - Help you choose and tune the right rank aggregation algorithm
- Current collab. with APHP P. Brousse on Conqur-Bio (Leukemia)
  - 6 months of M2 by CDS 2.0 for Pierre Andrieu (now PhD stud., MNRT)
  - A very efficient an promising heuristic has been designed!
- We are open to new collaborations
  - On the domain science side (new use cases)
  - On the data science side (evaluation of consensus rankings)

THANKS!

Alain Denise
LRI

Bryan Brancotte
(now @Pasteur)

Pierre Andrieu
LRI

Adeline Pierrot
LRI

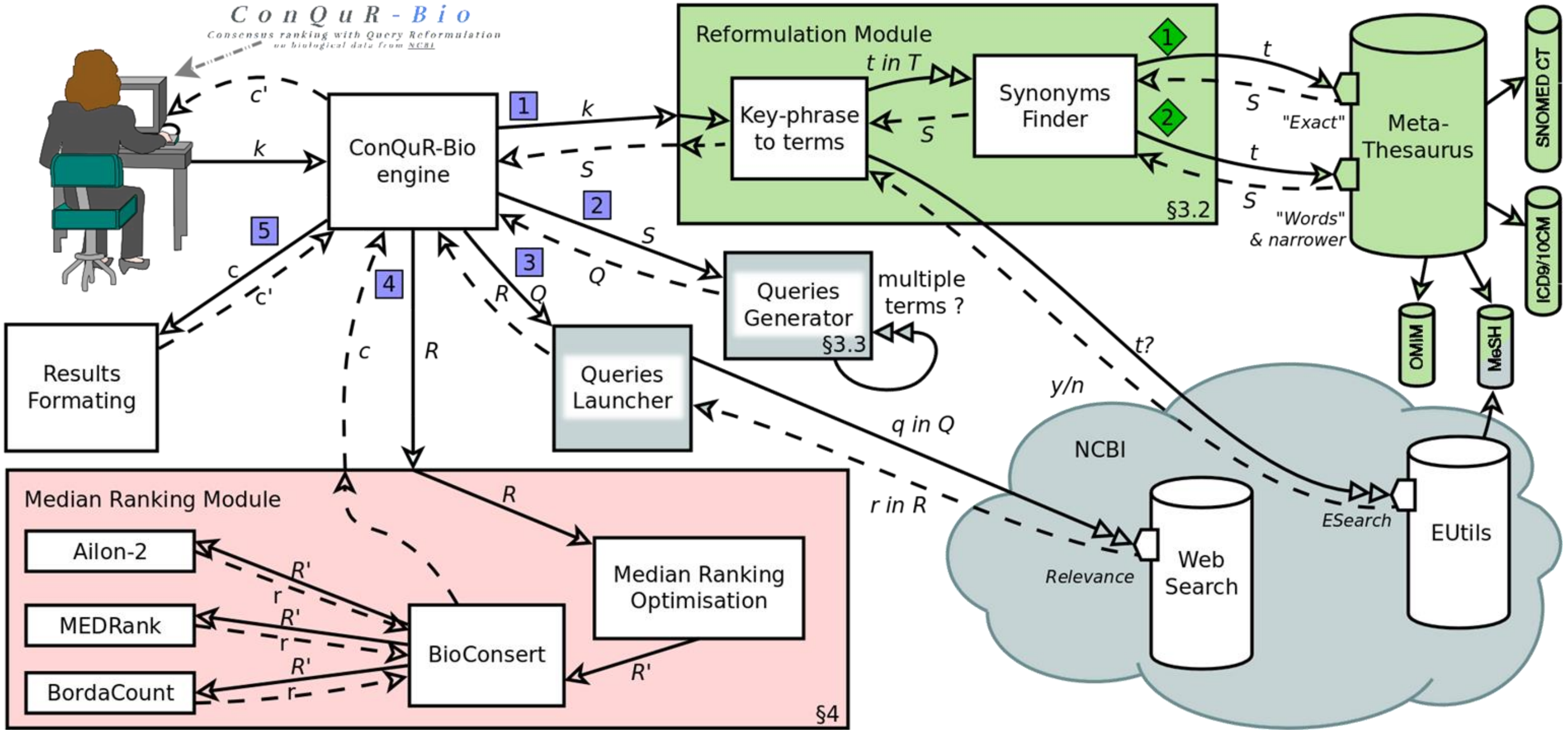
Robin Milosz
(Univ. Montréal &
LRI for 6 months)

Bastien Rance
APHP G. Pompidou

Ivan Sloma
APHP P. Brousse

Christophe Desterke
Inserm P. Brousse

Sylvie Hamel
(Univ. Montréal)



ConQuR-Bio
Consensus ranking with Query Reformulation on biological data from NCBI
c'
k
ConQuR-Bio engine
1
k
S
2
S
3
R
Q
Q
4
c
R
5
c
c'
Results Formating
Reformulation Module
t in T
Key-phrase to terms
S
Synonyms Finder
1
2
§3.2
t
S
"Exact"
t
S
"Words" & narrower
Meta-Thesaurus
SNOMED CT
ICD9/10CM
OMIM
MeSH
Queries Generator
§3.3
multiple terms ?
Queries Launcher
q in Q
r in R
y/n
t?
NCBI
Relevance
Web Search
ESearch
EUtils
Median Ranking Module
Ailon-2
MEDRank
BordaCount
R'
r
R'
r
R'
r
BioConsert
R
Median Ranking Optimisation
R'
§4