EFFICIENTLY RANKING BIG BIOLOGICAL AND BIOMEDICAL DATA SETS USING RANK AGGREGATION TECHNIQUES

Sarah Cohen-Boulakia

Laboratoire de Recherche en Informatique

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

CONTEXT: CONQUR-BIO

- Original ConQuR-Bio partners
  - Computer science Lab (LRI), Bioinformatics group
  - APHP George Pompidou, Institut Curie
- *Problems met by domain scientists*
  - Get as much available information as possible on genes involved in a disease?
    - Several equivalent keywords to describe a disease
    - Each *keyword queried* provides a set of ranked answers…

    → How to combine such *ranked answers*?
- … Formalized as a *query reformulation* and *consensus ranking* problems for data scientists
- Data science
  - Databases, Knowledge representation, combinatorics
- Domain science
  - Genes possibly involved in diseases

QUERYING BIOLOGICAL DATABASES: ALTERNATIVE REFORMULATIONS!

Which are the genes associated to a given disease?

Synonyms:
Breast cancer vs *mammalian carcinoma (14 524 vs 766 genes, not all included)*

Abbreviations:
Attention deficit hyperactivity disorders vs *ADHD (109 vs 144 genes, 74 common)*

Linguistics variations:
*tumo**u**r* vs *tumor*
(& breast cancer) : 681 vs 291 genes

More precise reformulations:
colorectal cancer vs *Lynch syndrom*
(+6 new genes)

NCBI EntrezGene keywords queries

- Finding all **synonyms** is time-consuming
- Querying using all synonyms provide huge amounts of data sets which have to be **ranked**….

THE DATA SCIENCE SIDE OF CONQUR-BIO

I) Reformulations (synonyms) can be automatically generated

- Input: the user's Keyword
- Automatic search of synonyms in major biomedical terminologies (MeSH, OMIM, ICD10CM, ICD9CM, SNOMED CT)
- Output : set of Synonyms of the input keyword

II) DB querying is automated (based on keywords)

- Input : set of synonyms obtained in I)
- Querying gene databases with each synonym (#queries = # synonyms)
- Output : sets of rankings (ranked genes), one ranking per synonym

III) Need to **combine** the input rankings into one final ranking…

According to which ranking criteria?...

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

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

THE DATA SCIENCE SIDE OF CONQUR-BIO

I) Reformulations (synonyms) can be automatically generated

- Input: the user's Keyword
- Automatic search of synonyms in major biomedical terminologies
- Output : set of Synonyms of the input keyword

II) DB querying is automated (based on keywords)

- Input : set of synonyms obtained in I)
- Querying gene databases with each synonym (#queries = # synonyms)
- Output : sets of rankings (ranked genes), one ranking per synonym

III) Rank Aggregation is performed using consensus algorithms

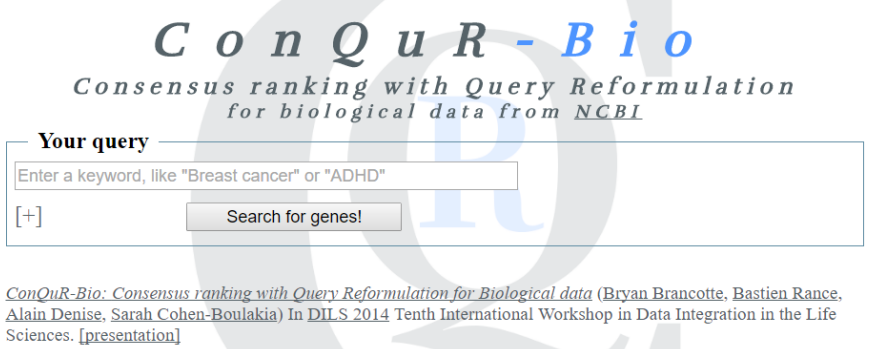
- Input : set of rankings
- Aggregation of rankings
- Output : one final *consensus ranking* (set of ranked genes) that is the closest of the input rankings (minimizing the disagreements)

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

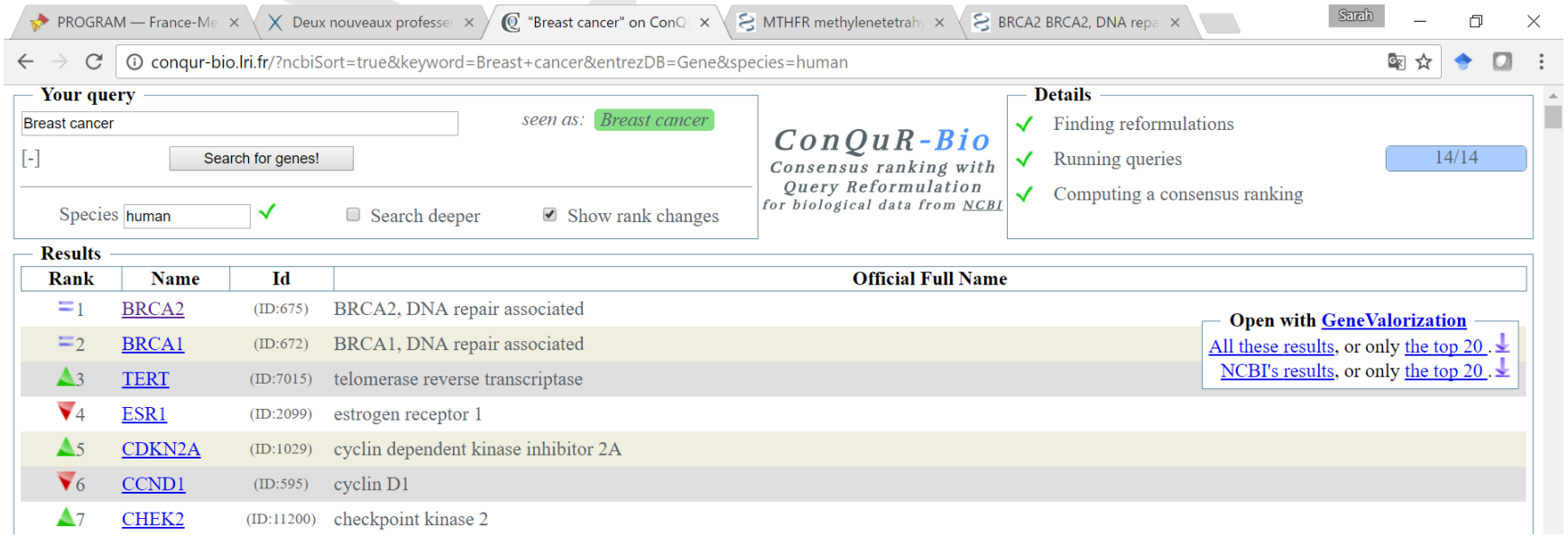
$$\pi^* = [A, D, B, C]$$

TOOL BUILDING SIDE OF CONQUR-BIO

http://conqur-bio.lri.fr/



Highly accessed by members of APHP (Hospitals), Institut Curie, Institut Pasteur…



NEW PLANS FOR CONQUR-BIO

- More reformulations needed → Increasing number of rankings
- …and increasing number of answers!

  → ConquR-Bio doesn't scale!

  → Need for a new heuristics
- In the meantime…
  - Very positive feedback from users at APHP Paul Brousse
  - Very promissing results on Blood cancers

  → Need to carefully evaluate the benefit of using ConQu-Bio

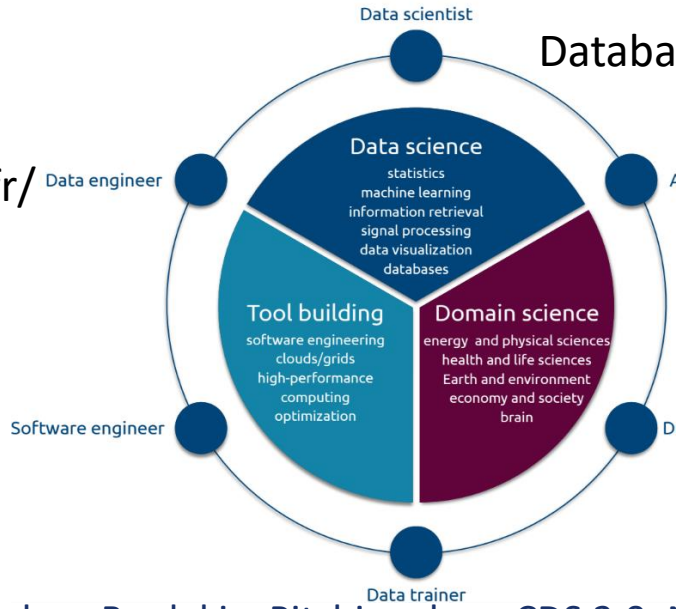
- **6 months internship (M2)**
  - New heuristics
  - bioevaluation of the results obtained in Leukemia
- To be followed by a PhD

CONCLUSION

- ConquR-Bio is concretely used by several domain users
- When data become bigger, need to improve the current heuristics
- Need to carrefully evaluate the benefit of using the tool in specific contexts (leukemia with APHP Paul Brousse)

→**6 months internship** for a bioinformatician master student (M2)

http://conqur-bio.lri.fr/



Database + Combinatorics

Bioinformatics

Genes involved in a disease

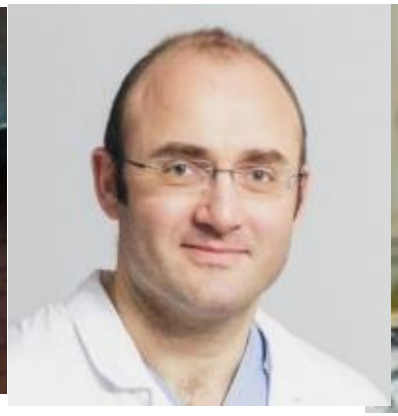
THANKS!

Alain Denise

Bryan Brancotte

Ivan Sloma

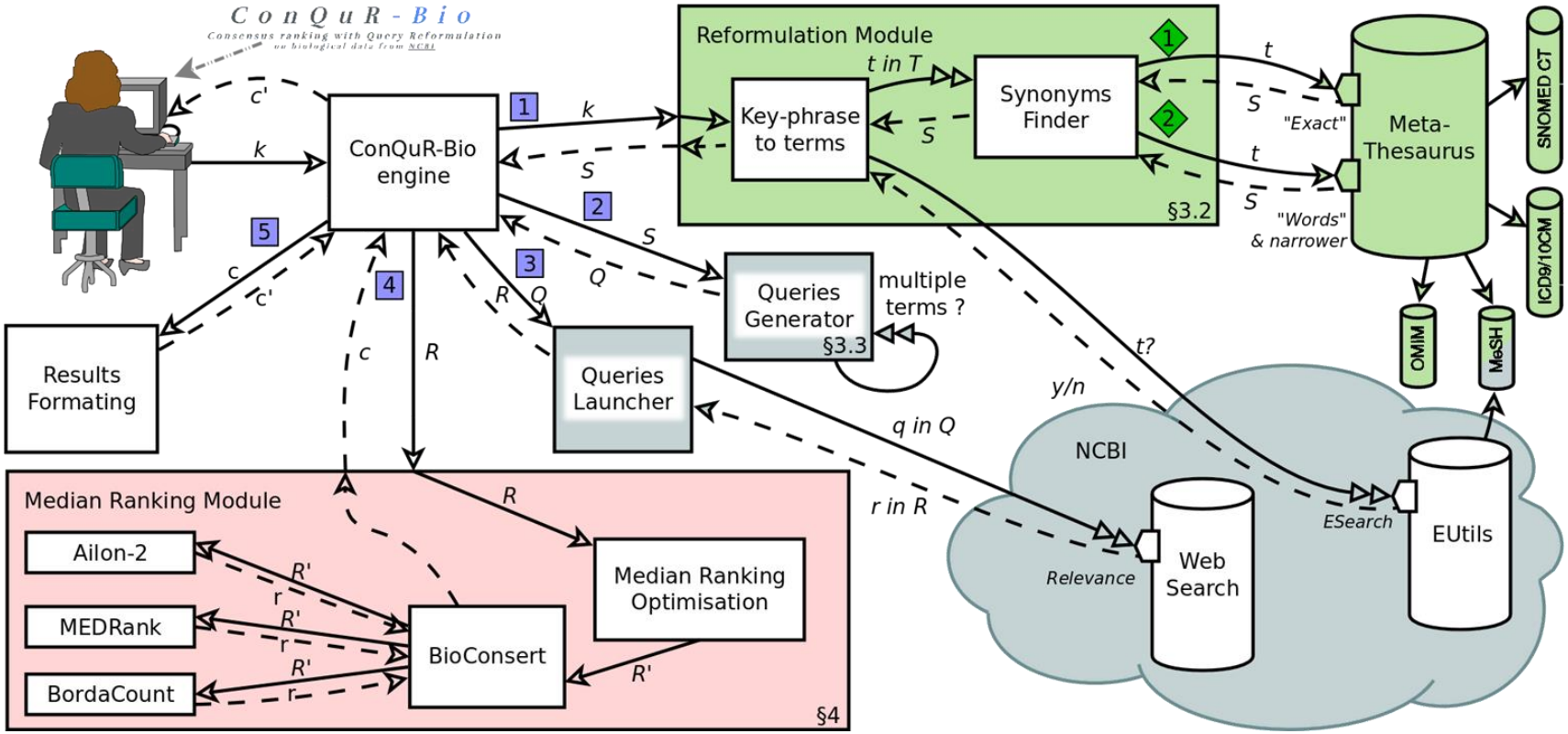
Bastien Rance



Fabien Reyal

Francois Radvanyi

Christophe Desterke


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

La distance de Kendall-τ

« Trouver un consensus proche des classements en entrée »

[Kendall 1938]

La distance de Kendall−τ D(π,σ) compte le nombre de paires d'éléments inversés entre deux classements

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

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

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

Définition d'un consensus optimal

« Trouver un consensus proche des classements en entrée »

Score de Kemeny

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

Consensus Optimal

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

$\mathcal{S}_n$: l'ensemble des classements de n éléments

classements de n éléments

Complexité [Dwork *et al* 2001, Biedl *et al.* 2009] :
NP-Difficile pour un nombre de permutations pair et ≥ 4

$$\mathcal{P} \begin{cases} \pi_1 = [A, D, C, B] \\ \pi_2 = [B, A, D, C] \\ \pi_3 = [D, A, B, C] \end{cases}$$

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

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