

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: CONQR-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!

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:

tumour vs tumor (& breast cancer) : 681 vs 291 genes

More precise reformulations:

colorectal cancer vs Lynch syndrom (+6 new genes)

Which are the genes associated to a given disease?



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 CONQR-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?...

$$\begin{aligned}\pi_1 &= [A, D, C, B] \\ \pi_2 &= [B, A, D, C] \\ \pi_3 &= [D, A, B, C]\end{aligned} \quad \pi^* = [A, D, B, C]$$

THE DATA SCIENCE SIDE OF CONQR-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)

$$\begin{aligned}\pi_1 &= [A, D, C, B] \\ \pi_2 &= [B, A, D, C] \\ \pi_3 &= [D, A, B, C]\end{aligned}\quad \pi^* = [A, D, B, C]$$

TOOL BUILDING SIDE OF CONQUR-BIO

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

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

ConQUR-Bio

Consensus ranking with Query Reformulation for biological data from NCBI

Your query

Enter a keyword, like "Breast cancer" or "ADHD"

[+]

Search for genes!

ConQUR-Bio: Consensus ranking with Query Reformulation for Biological data (Bryan Brancotte, Bastien Rance, Alain Denise, Sarah Cohen-Boulakia) In DILS 2014 Tenth International Workshop in Data Integration in the Life Sciences. [presentation]

The screenshot shows a web browser window with several tabs. The active tab is titled "Breast cancer" on ConQ. The address bar shows the URL: `conqur-bio.lri.fr/?ncbiSort=true&keyword=Breast+cancer&entrezDB=Gene&species=human`. The page content is divided into several sections:

- Your query:** A search box containing "Breast cancer" and a "Search for genes!" button. Below the search box, there are options for "Species" (set to "human"), "Search deeper" (unchecked), and "Show rank changes" (checked).
- seen as:** A green pill-shaped label containing the text "Breast cancer".
- ConQUR-Bio logo and description:** The logo and the text "Consensus ranking with Query Reformulation for biological data from NCBI".
- Details:** A box containing three green checkmarks and the following text: "Finding reformulations", "Running queries", and "Computing a consensus ranking". A blue button labeled "14/14" is also present.
- Results:** A table with columns "Rank", "Name", "Id", and "Official Full Name". The results are as follows:

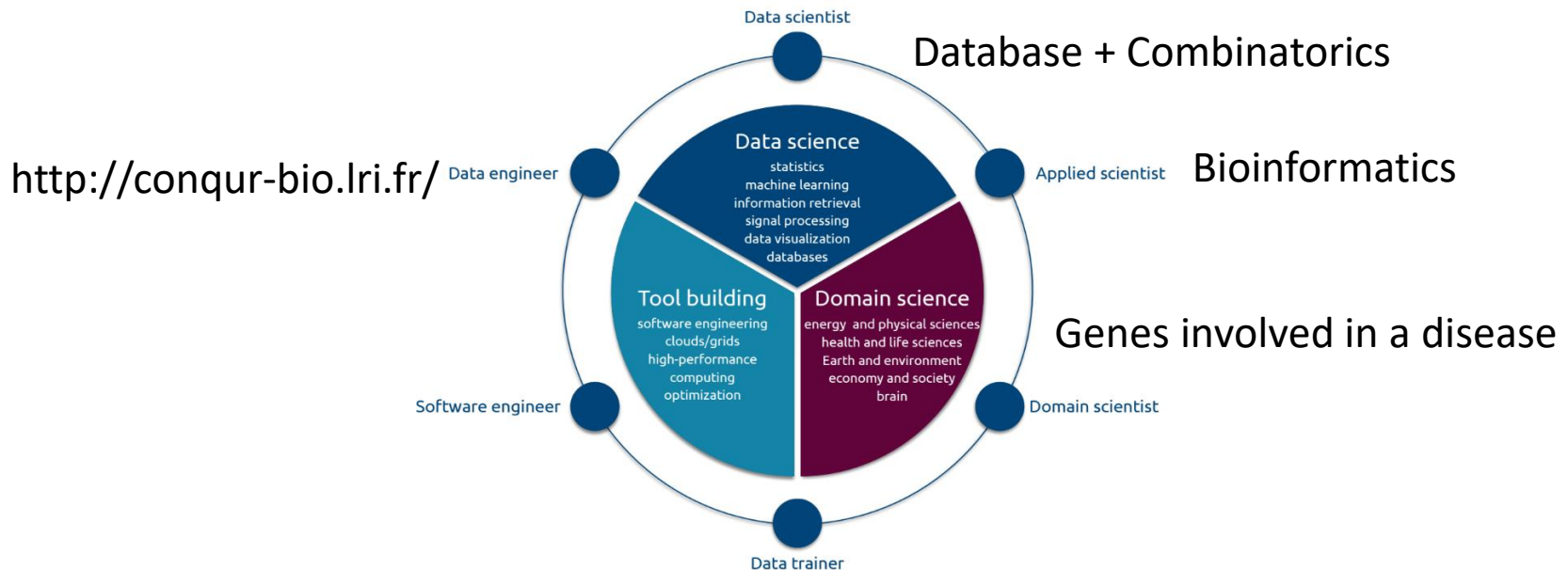
Rank	Name	Id	Official Full Name
1	BRCA2	(ID:675)	BRCA2, DNA repair associated
2	BRCA1	(ID:672)	BRCA1, DNA repair associated
3	TERT	(ID:7015)	telomerase reverse transcriptase
4	ESR1	(ID:2099)	estrogen receptor 1
5	CDKN2A	(ID:1029)	cyclin dependent kinase inhibitor 2A
6	CCND1	(ID:595)	cyclin D1
7	CHEK2	(ID:11200)	checkpoint kinase 2
- Open with GeneValorization:** A box containing the text "Open with GeneValorization" and two links: "All these results, or only the top 20." and "NCBI's results, or only the top 20.".

NEW PLANS FOR CONQR-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 promising 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 **carefully evaluate the benefit of using the tool** in specific contexts (leukemia with APHP Paul Brousse)
- **6 months internship** for a bioinformatician master student (M2)



THANKS!

université
PARIS-SACLAY



Alain Denise



Bryan Brancotte



Ivan Sloma



Bastien Rance



Fabien Reyat

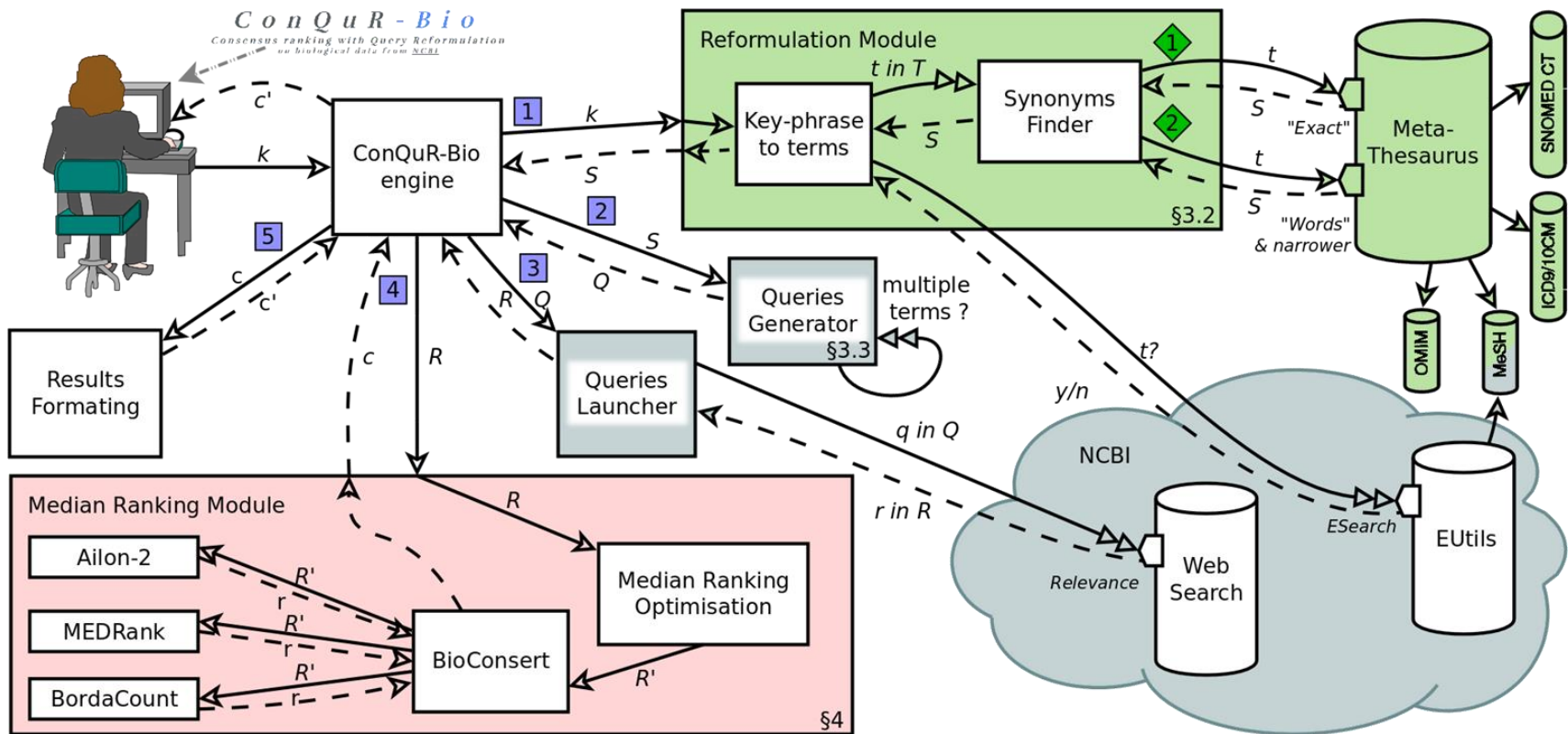


Francois Radvanyi



Christophe Desterke

Sarah Cohen-Boulakia, Pitching days, CDS 2.0, Nov. 9th 2016



La distance de Kendall-

« Trouver un consensus proche des classements en entrée »

τ [Kendall 1938]

La distance de Kendall- τ $D(\pi, \sigma)$ compte le nombre de paires d'éléments inversés entre deux classements

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

$$\begin{aligned} \pi_1 &:= [A, D, C, B] \\ \pi_2 &:= [B, A, D, C] \end{aligned}$$

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

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

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

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

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

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

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