ConQuR-Bio

Consensus ranking with Query Reformulation for biological data

Bryan Brancotte[1], Bastien Rance[2], Alain Denise[1,3], Sarah Cohen-Boulakia[1]

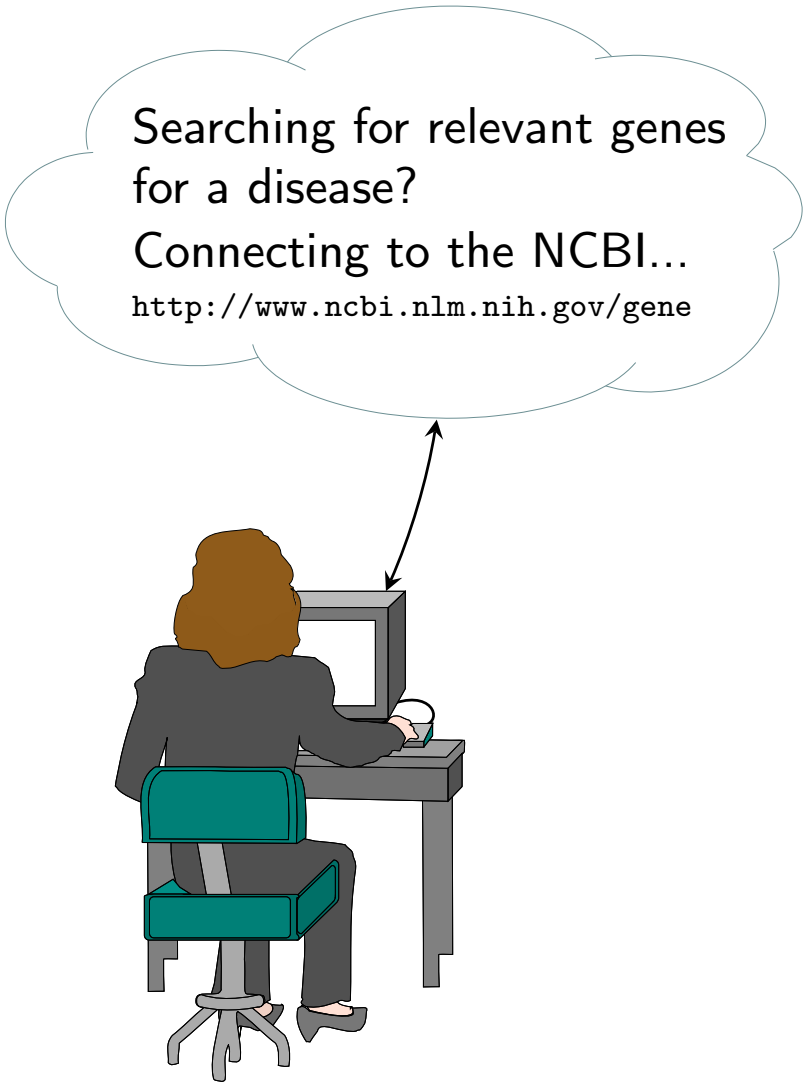
(1) Laboratoire de Recherche en Informatique (LRI) & INRIA, Université Paris-Sud, France

(2) Biomedical Informatics and Public Health Department, University Hospital Georges Pompidou, AP-HP, Paris, France & INSERM, Université Paris Descartes, Sorbonne Paris Cité, Faculté de médecine, Paris, France

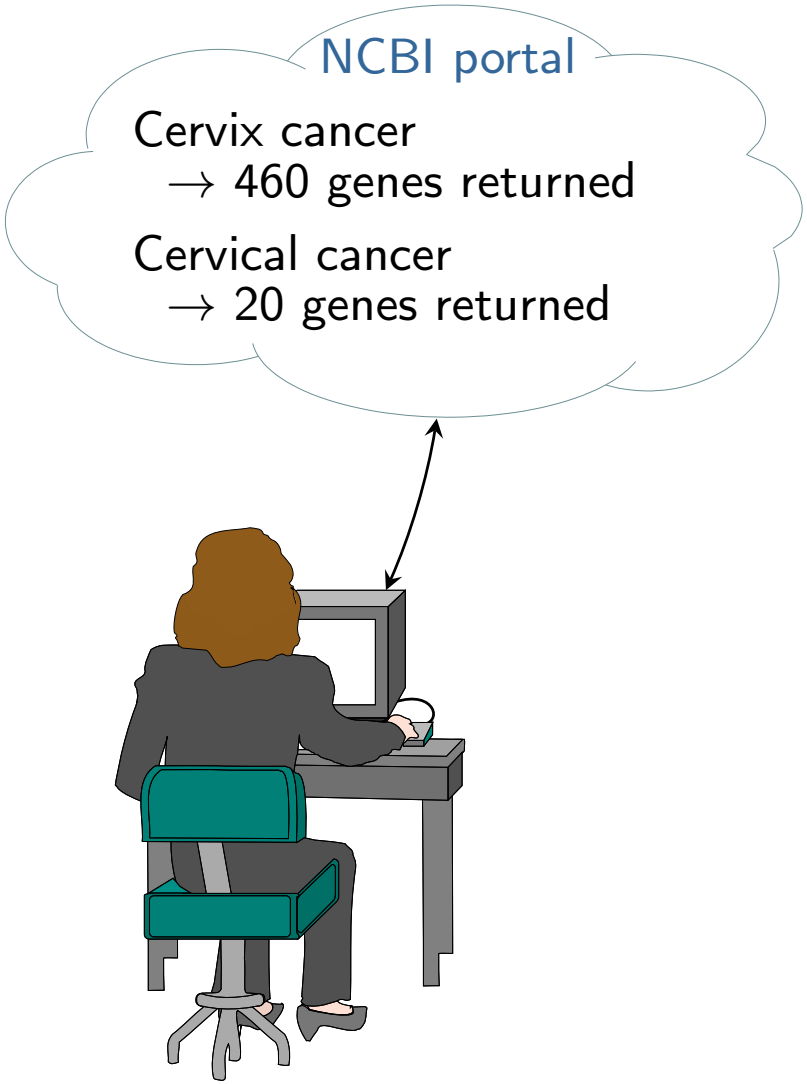
(3) Institut de Génétique et de Microbiologie (IGM), CNRS UMR 8621 Université Paris-Sud - France

10th International Conference on Data Integration in the Life Sciences
July 18, 2014

An approach driven by use cases

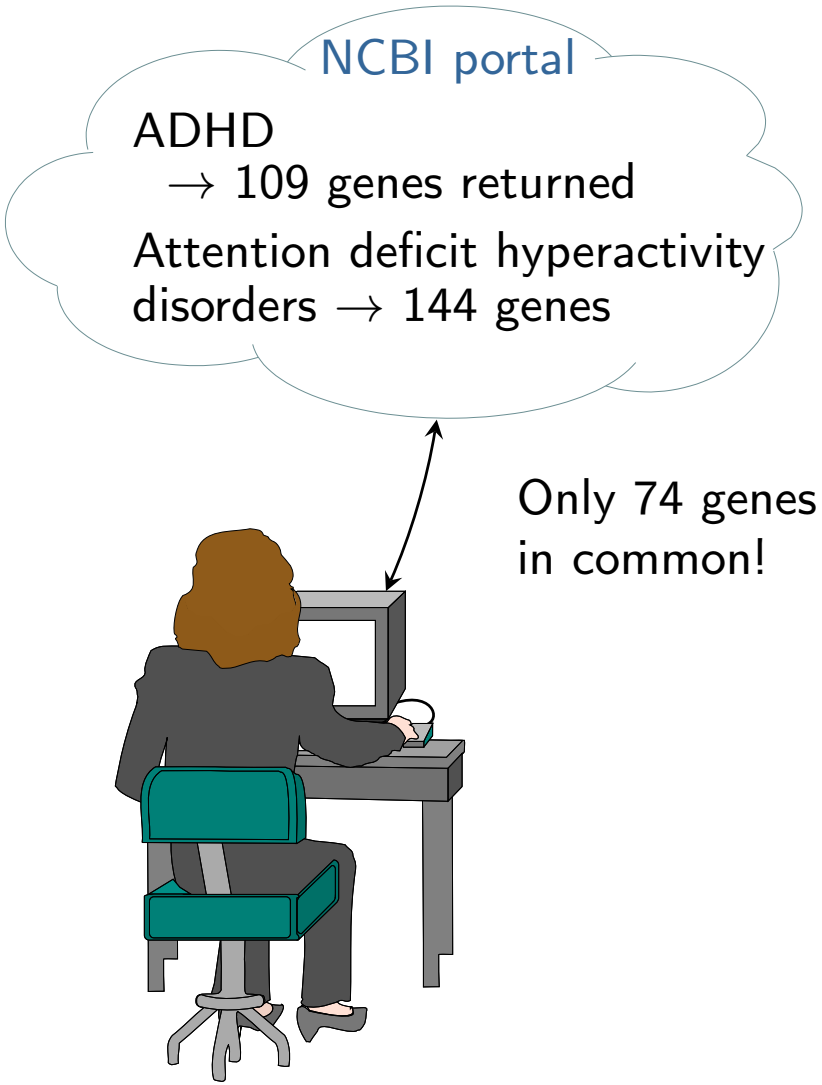


An approach driven by use cases



Equivalent reformulations: *cervix cancer* vs *cervical cancer*

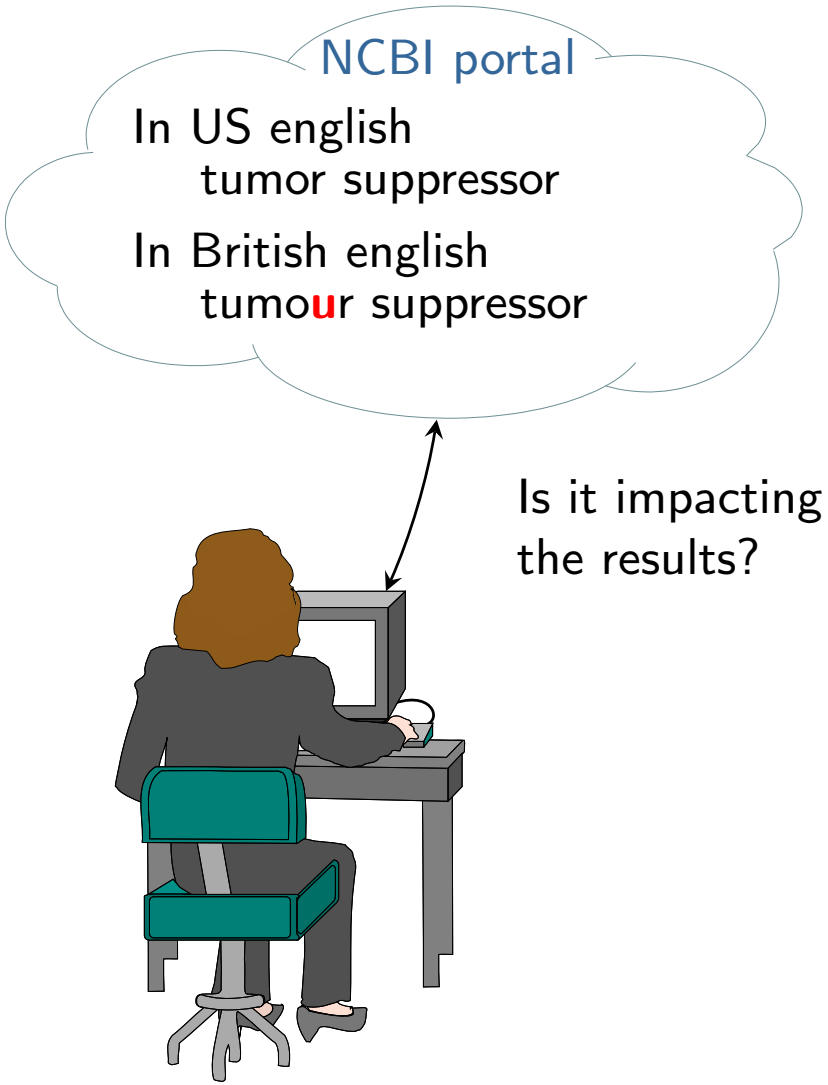
An approach driven by use cases



Equivalent reformulations: *cervix cancer* vs *cervical cancer*

Abbreviations: *Attention deficit hyperactivity disorders* vs *ADHD*

An approach driven by use cases



Equivalent reformulations: *cervix cancer* vs *cervical cancer*

Abbreviations: *Attention deficit hyperactivity disorders* vs *ADHD*

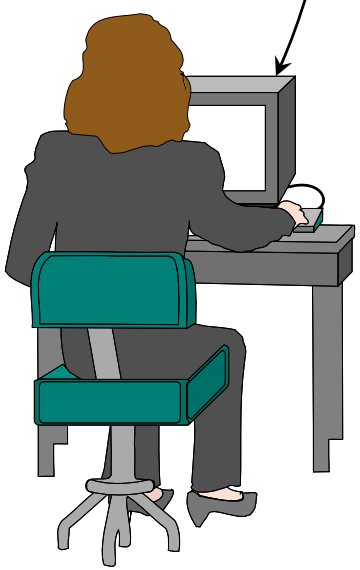
Lexical-based reformulations: *tumor suppressor* vs *tumour suppressor*

An approach driven by use cases

NCBI portal

Breast cancer tumor suppressor: → 681 genes

Breast cancer tumo**u**r suppressor: → 291 genes



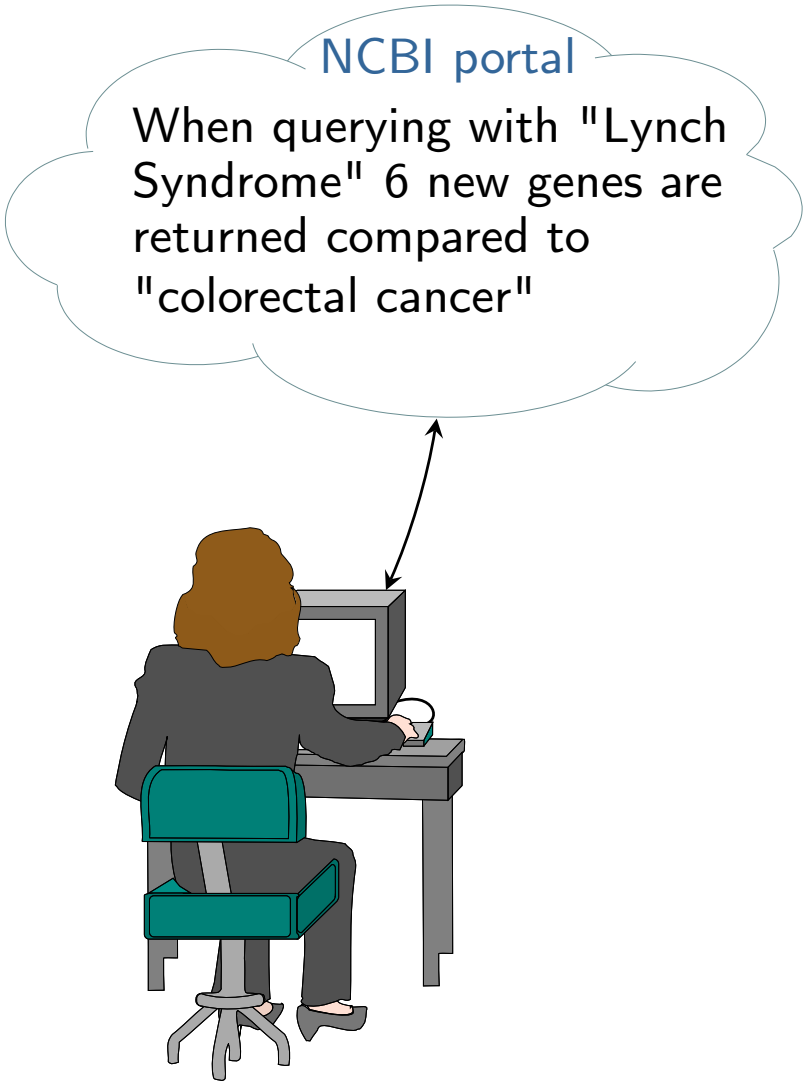
Regional settings are also important

Equivalent reformulations: *cervix cancer* vs *cervical cancer*

Abbreviations: *Attention deficit hyperactivity disorders* vs *ADHD*

Lexical-based reformulations: *tumor suppressor* vs *tumo**u**r suppressor*

An approach driven by use cases



Equivalent reformulations: *cervix cancer* vs *cervical cancer*

Abbreviations: *Attention deficit hyperactivity disorders* vs *ADHD*

Lexical-based reformulations: *tumor suppressor* vs *tumour suppressor*

Narrower-term-based reformulations: The *colorectal cancer* versus a subtype: the *Lynch Syndrome*

An approach driven by use cases

Problems we aim to address

- Querying without considering reformulations implies getting incomplete sets of answers
- Querying many reformulations is a time consuming task for the user
- Querying many reformulations provides possible huge amounts of answers

Equivalent reformulations: *cervix cancer* vs *cervical cancer*

Abbreviations: *Attention deficit hyperactivity disorders* vs *ADHD*

Lexical-based reformulations: *tumor suppressor* vs *tumour suppressor*

Narrower-term-based reformulations: The *colorectal cancer* versus a subtype: the *Lynch Syndrome*

How to address all the problem induced by use cases?

How to obtain a **relevant** list of genes taking into account **all reformulations**?

We have automatized the extraction of:

- query reformulations from different biomedical terminologies (MeSH, OMIM, ICD10CM, ICD9CM, SNOMED CT)
- lists of genes sorted with the "relevance" criteria of the NCBI for each reformulation

We want to:

- take into account **all the genes** returned by each reformulation
- exploit the fact that some genes appear in the results of several reformulations while others don't
- provide **one single list** of genes answering the user's query ranked by relevance

$\Rightarrow$ The Median Ranking Problem can answer the needs!

From alternative rankings to one consensus ranking

Notions for the Median Ranking Problem

A ranking with ties is an ordering of buckets (set of elements) where two elements are ranked differently iff they are in different buckets and are tied otherwise.

$r = [\{A\}, \{C, B\}]$
B and C are tied, both ranked after A

$\rightarrow$ to compare rankings with ties we need a distance

The generalized Kendall$-\tau$ distance [Fag+04] denoted $K^{(p)}$
Counts the pairwise disagreements between two rankings with ties:

- counts **1** when two elements are **inversed**
- counts **p** $\in]0; 1]$ when two elements are **tied** in only one ranking

Example: $K^{(p)}\left([\{C\}, \{A\}, \{B\}] \ , \ [\{A\}, \{C, B\}] \right) = 0_{A-B} + 1_{A-C} + p_{B-C}$

$\rightarrow$ What is the median ranking problem?

From alternative rankings to one consensus ranking

The Median Ranking Problem

Formally, the median ranking problem is to find, for a set of input rankings R, a *median* ranking c^* such that: $K^{(p)}(R, c^*) \leq K^{(p)}(R, r), \forall r \in \mathcal{R}$;

$K^{(p)}$ is the generalized Kendall$-\tau$ distance [Fag+04]:

- counts 1 when two elements are inversed
- counts $p \in]0; 1]$ when two elements are tied in only one ranking

Example

Let us consider the set of input rankings $R = \{r_1, r_2, r_3\}$, the median under the generalized Kendall-τ distance is c

$$R = \left\{ \begin{array}{l} r_1 = [\{A\}, \{D\}, \{B, C\}] \\ r_2 = [\{B\}, \{A\}, \{D\}, \{C\}] \\ r_3 = [\{A, D\}, \{B, C\}] \end{array} \right\} \qquad \begin{array}{l} c = [\{A\}, \{D\}, \{B, C\}] \\ K^{(p)}(R, c) = 1_{A-B@r_2} + 1_{B-D@r_2} + \\ p_{A-D@r_3} + p_{B-C@r_2} = 2 + 2p \end{array}$$

$\rightarrow$ NP-Hard problem[Dwo+01], how to compute a solution?

$\rightarrow$ Real case rankings are not over the same elements, how to deal with it?

From alternative rankings to one consensus ranking

Dealing with missing elements

When alternative rankings (results of reformulations) produce different elements, how to apply the distance and construct a complete ranking?

Solutions

1. Induced Kendall-τ distance [Dwo+01] — KO
   Ignores disagreements related to missing elements

2. Projection process [BBN13] — KO
   Considers only elements appearing in all rankings (it **removes** the others)

3. Unification process [CBDH11] — OK, but adaptation to do
   Appends missing elements at the end in a **unification bucket**:
   $$\left\{ \begin{array}{l} r_1 = [\{A\},\{D\}] \\ r_2 = [\{B\},\{A,D\},\{C\}] \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} r'_1 = [\{A\},\{D\},\{B,C\}_u] \\ r'_2 = [\{B\},\{A,D\},\{C\}] \end{array} \right\}.$$

B and C are indeed **less relevant** than A and D in r_1
B and C are **not equally relevant** in r_1 and should not be considered so in r'_1!

Extending the generalized Kendall$-\tau$ distance to a pseudometrics $\mathcal{M}(r'_1, r'_2)$

Counts the pairwise disagreements between two rankings with ties:

- counts 1 when two elements are inversed
- **counts 0 when two elements are tied in at least one unifying bucket**
- counts $p \in]0; 1]$ when two elements are tied in only one ranking

Example

Let us consider the set of input rankings $R = \{r_1, r_2, r_3\}$, the median under the generalized Kendall-τ distance is c and the median under the pseudometrics is c'

$$R = \left\{ \begin{array}{l} r_1 = [\{A\}, \{D\}, \{B, C\}_u] \\ r_2 = [\{B\}, \{A\}, \{D\}, \{C\}] \\ r_3 = [\{A, D\}, \{B, C\}_u] \end{array} \right\} \quad \begin{array}{ll} c = & [\{A\}, \{D\}, \{B, C\}] \\ c' = & [\{A\}, \{D\}, \{B\}, \{C\}] \\ & \mathcal{M}(R, c) = 2 + p > \mathcal{M}(R, c') = 2 \end{array}$$

How can we actually compute a consensus?

From alternative rankings to one consensus ranking

Computing a solution

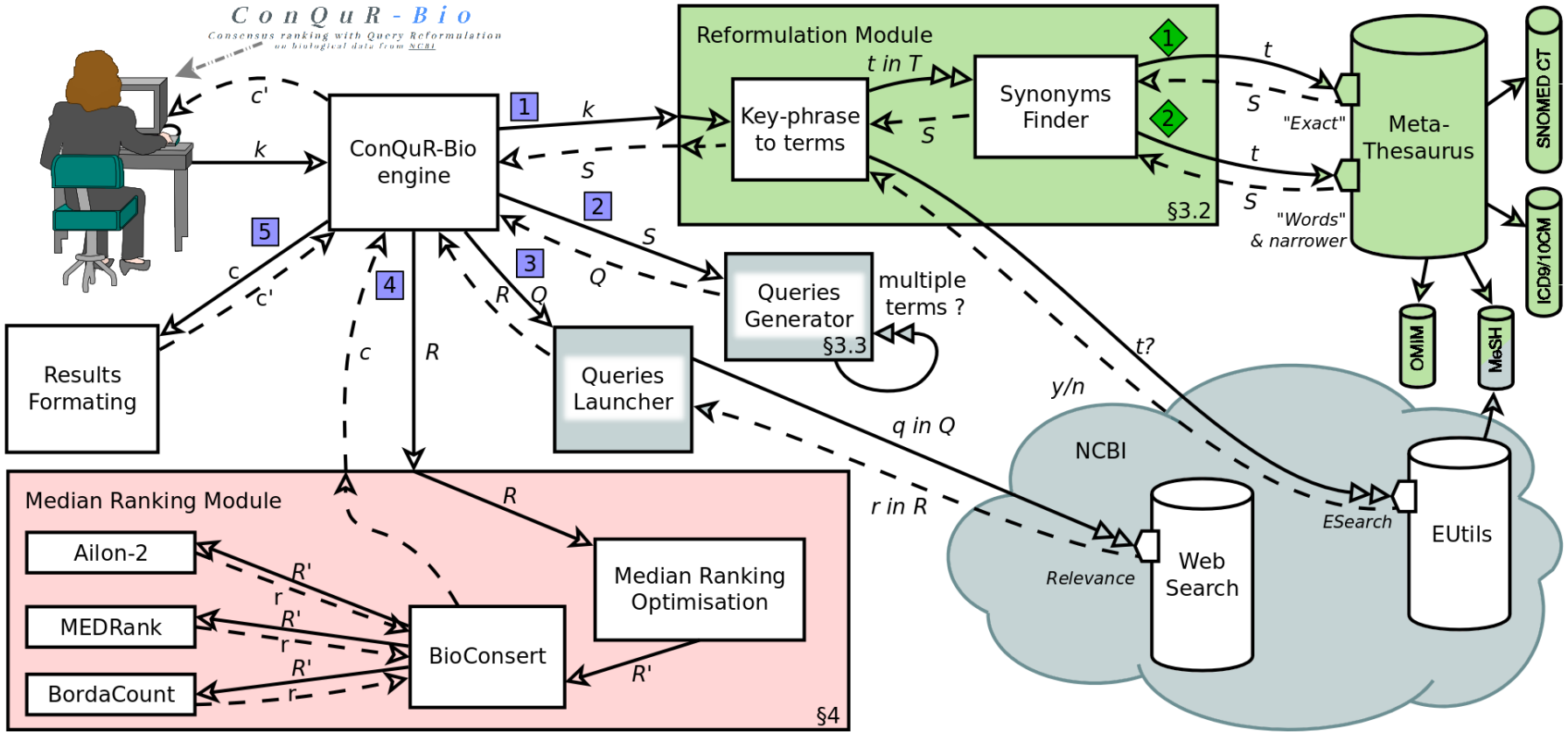
Plethora of algorithms, but either they are too time consuming [Fag+04; Mei+07; ACN08; Ail10] or they do not provide results "good" enough [Bor81; FKS03; Ail10]. Only two algorithms [CBDH11; Fag+04] can handle the pseudometrics $\mathcal{M}(r'_1, r'_2)$.

Main algorithm used: a greedy local search algorithm, BioConsert [CBDH11]: Starting with each input rankings and considering two edit operations (moving an element to a new bucket or an existing bucket), it applies operators as long as it reduces the distance between the current consensus obtained and the input rankings.

Figure: edit operations in BioConsert

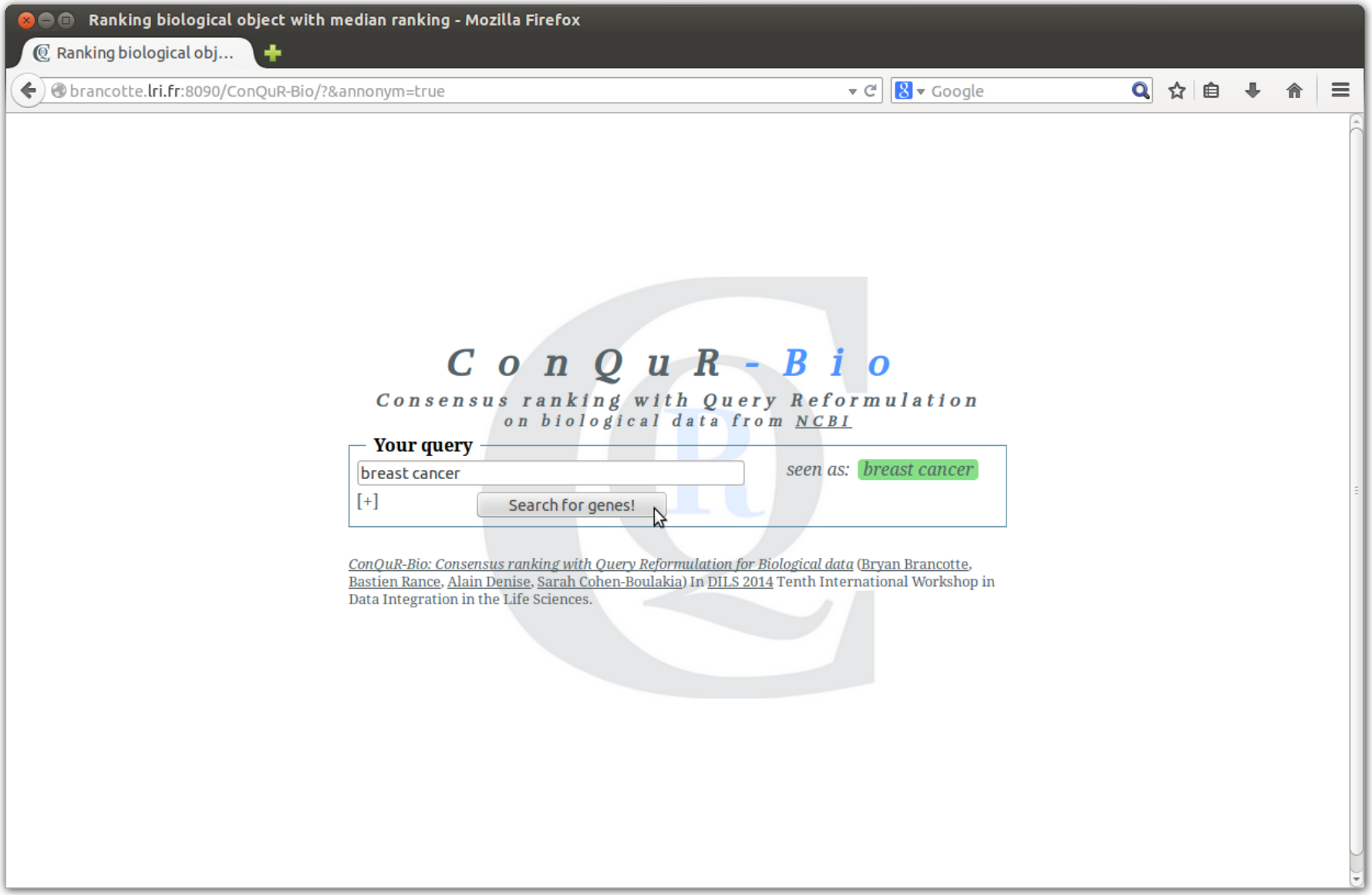
Tuning BioConsert: Starting from the best solution provided by BordaCount [Bor81], MEDRank [FKS03], and Ailon's 2-approximation [Ail10]

Up to a hundred times faster!

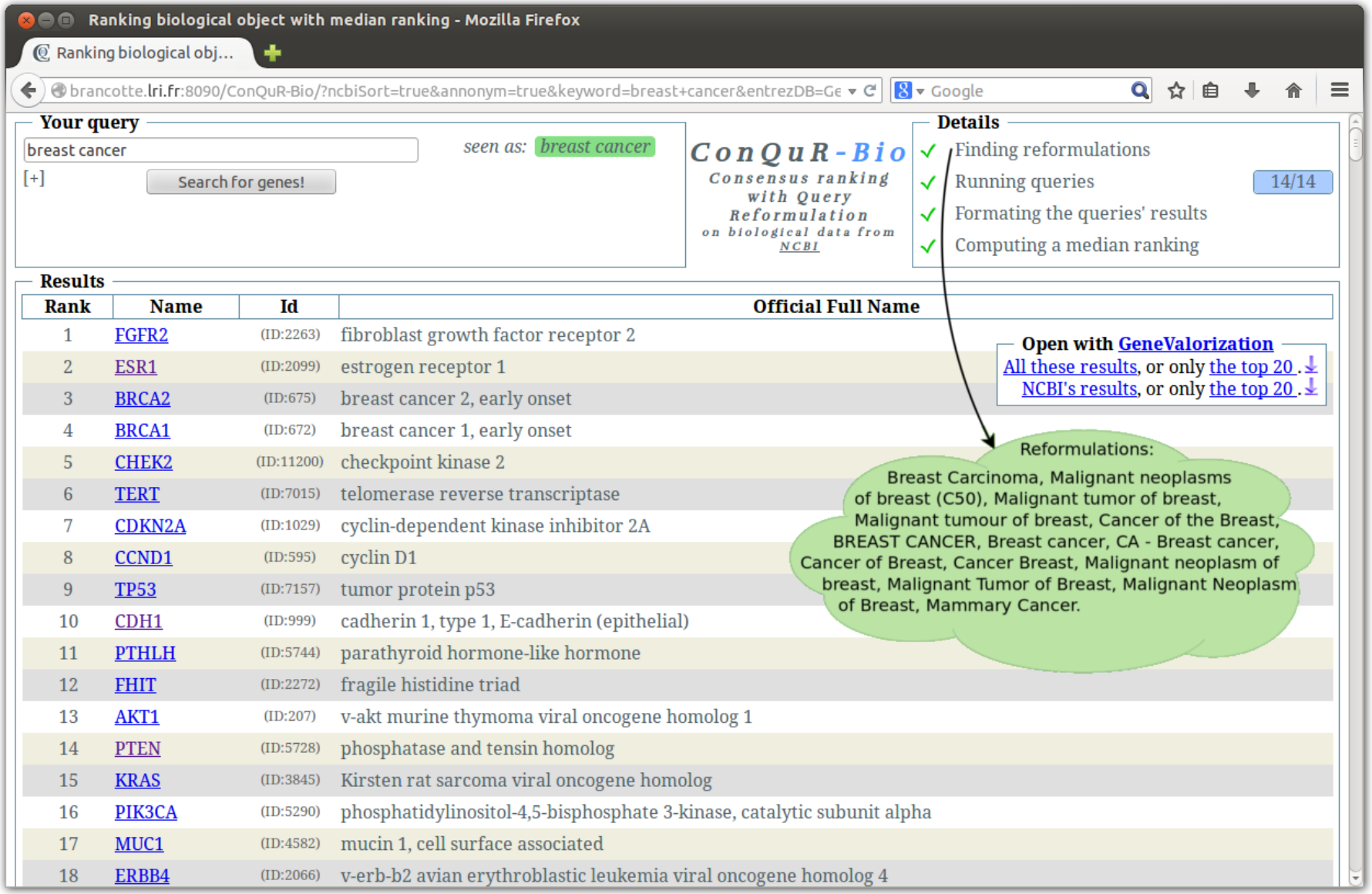


Architecture of ConQuR-Bio:

- Green area: identification of terms in the query and reformulations.
- Blue area: querying the NCBI to obtain lists of genes for each reformulation.
- Red area: computing a consensus.



Online at `http://conqur-bio.lri.fr/`



Online at `http://conqur-bio.lri.fr/`

Bibliometrics indicators

Focusing on the top 20 genes. For each gene we consider publications co-citing the gene name and the query, and compare to the top 20 genes returned by the NCBI portal.

Using the number of publications

Summing the **number of publications** co-citing the gene name and the query.
With ConQuR-Bio: **56% more publications** associated than with the NCBI.

Using publication "freshness"

Averaging the number of **days since the last publication** co-citing the gene name and the query.
With ConQuR-Bio: **25% less days** than with the results provided by the NCBI.

Expertise based indicator

Gold-standards

Clinicians of the *Institut Curie (France)* and the *Children's Hospital of Philadelphia (PA, USA)* **provided gold-standards**, list of expected genes, **for 9 different diseases**: **7 cancers** (*bladder, breast, cervical, colorectal, neuroblastoma, prostate, retinoblastoma*), one **heart disease** (the *Long QT Syndrome*), and one **psychiatric disorder** (the *attention deficit (with) hyperactivity disorder)*. Diseases are often combined with terms *tumor suppressor* and *oncogene*.

Measure: the **AUC**

The Area Under the ROC Curve [Bra97] is closely related to *precision* and *recall* measures, and allows to highlight the **presence of elements of the gold standard in the top results** . It provides a number in $[0, 1]$, 1 being the highest score.

Examples with two expected results (●●)

AUC(●●●●●)=0.50
AUC(●●●●●)=0.66
AUC(●●●●●)=0.66
AUC(●●●●●)=0.83

Results: using expertise of clinician collaborators

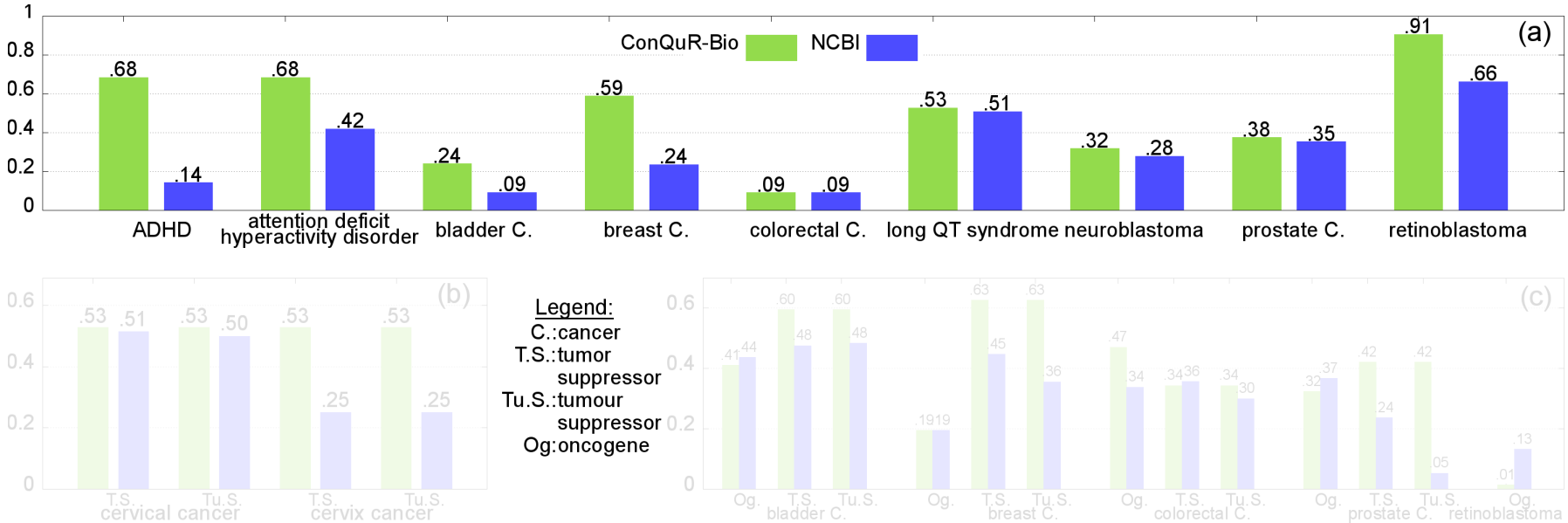


Figure: AUC for the 20 first genes returned by ConQuR-Bio and the NCBI WebSearch for **single-term queries.**

Average AUC increased of 58%

Results: using expertise of clinician collaborators

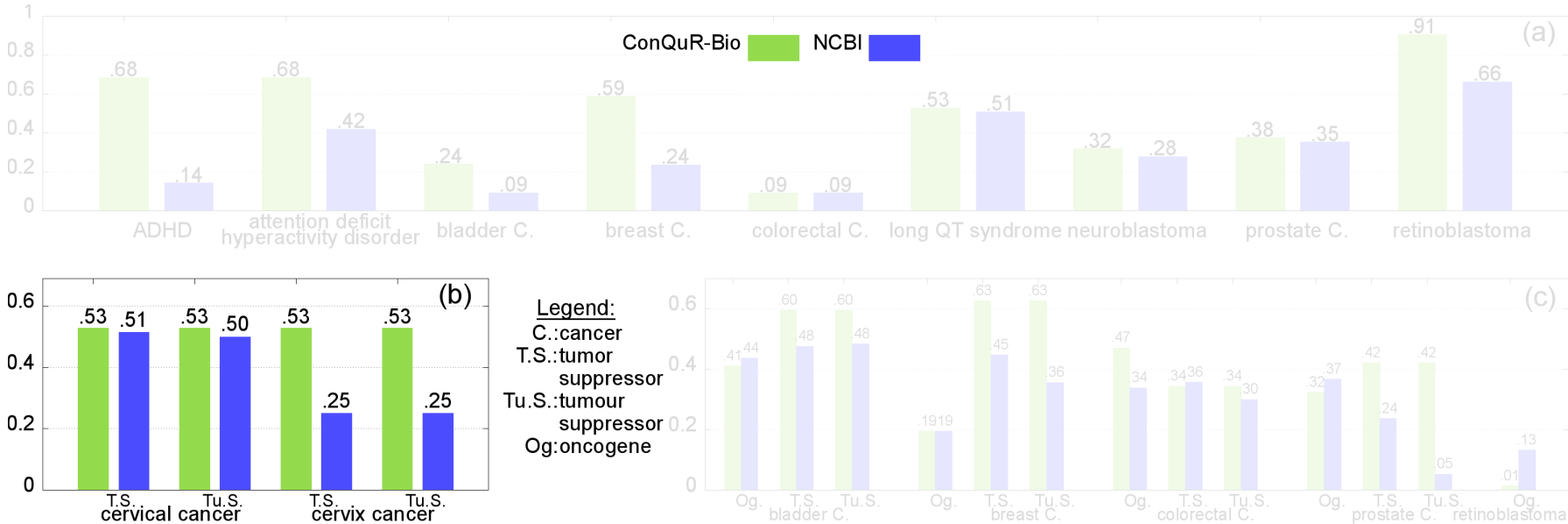


Figure: AUC for the 20 first genes returned by ConQuR-Bio and the NCBI WebSearch for **lexical variation around *cervix cancer tumor suppressor*.**

AUC of ConQuR-Bio results are **stable** and **superior** to the AUC of NCBI results

Results: using expertise of clinician collaborators

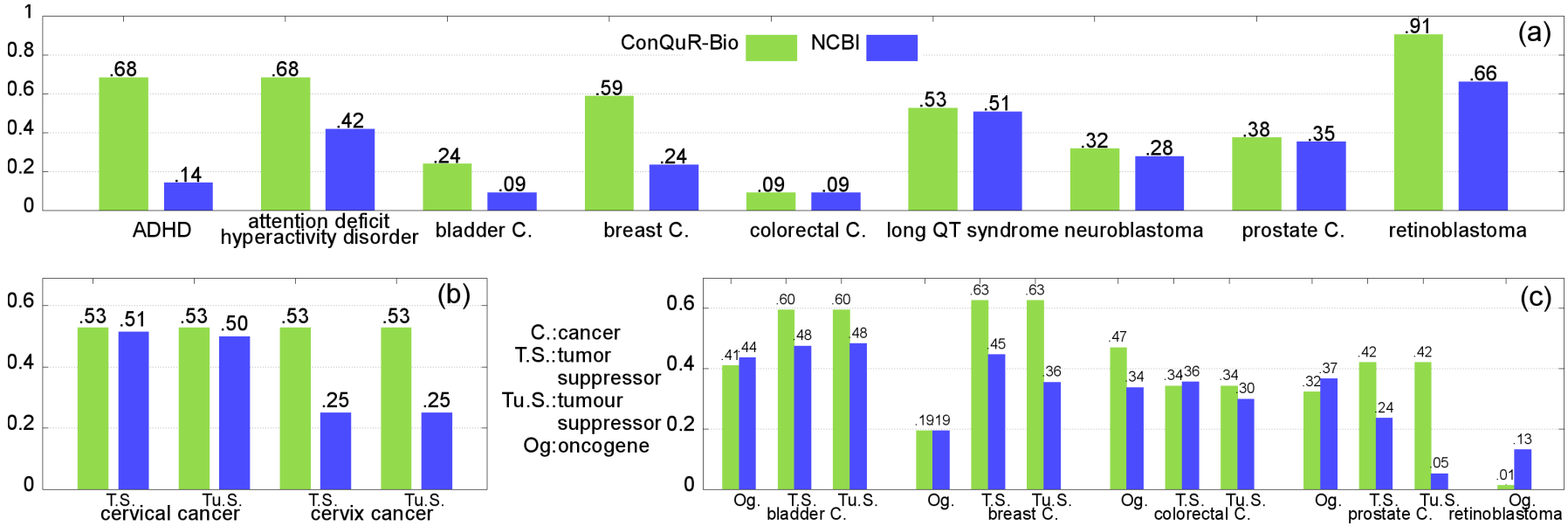


Figure: AUC for the 20 first genes returned by ConQuR-Bio and the NCBI WebSearch for (a) Single-term queries, (b) lexical variation around *cervix cancer tumor suppressor*, and (c) the remaining query for which we have gold standards.

Average AUC of ConQuR-Bio results are **increased** of 44%

Conclusion

ConQuR-Bio...

- exploits **biomedical terminologies** to reformulate the user query
- proposes a **consensus emphasizing agreements** between reformulations results
  - introducing a new pseudometrics answering problematics brought by the data
  - leveraging state-of-the-art algorithms to efficiently propose a relevant consensus
- increases the results **AUC** of $+$**44%** compared to the NCBI
- follows an **on-the-fly** approach
- is **free to use** at `http://conqur-bio.lri.fr`

Ongoing work

- Consider fine grain recognition of terms in users queries.
- Extend to larger and customizable amounts of biomedical terminologies.

References - I

[ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. "Aggregating inconsistent information: ranking and clustering". In: Journal of the ACM (JACM) 55.5 (2008), p. 23.

[Ail10] Nir Ailon. "Aggregation of partial rankings, p-ratings and top-m lists". In: Algorithmica 57.2 (2010), pp. 284–300.

[BBN13] Nadja Betzler, Robert Bredereck, and Rolf Niedermeier. "Theoretical and empirical evaluation of data reduction for exact Kemeny Rank Aggregation". In: Autonomous Agents and Multi-Agent Systems (2013), pp. 1–28.

[Bor81] J.C.de Borda. "Mémoire sur les élection au scrutin". In: Histoire de l'académie royal des sciences (1781), pp. 657 –664.

[Bra97] Andrew P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: Pattern Recognition 30 (1997), pp. 1145–1159.

[CBDH11] Sarah Cohen-Boulakia, Alain Denise, and Sylvie Hamel. "Using medians to generate consensus rankings for biological data". In: Proc. SSDBM: Scientific and Statistical Database Management Conference. LNCS 6809. Portland. Springer, 2011, pp. 73–90.

[Dwo+01] Cynthia Dwork et al. "Rank aggregation methods for the web". In: Proceedings of the 10th international conference on World Wide Web. ACM. 2001, pp. 613–622.

[Fag+04] Ronald Fagin et al. "Comparing and aggregating rankings with ties". In: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. PODS '04. Paris, France: ACM, 2004, pp. 47–58. URL: http://doi.acm.org/10.1145/1055558.1055568.

[FKS03] R. Fagin, R. Kumar, and D. Sivakumar. "Efficient similarity search and classification via rank aggregation". In: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM. 2003, pp. 301–312.

[Mei+07] Marina Meila et al. "Consensus ranking under the exponential model". In: (2007).