ConQuR-Bio: Consensus Ranking with Query Reformulation for Biological Data

Bryan Brancotte, Bastien Rance, Alain Denise, Sarah Cohen-Boulakia

ConQuR-Bio: Consensus ranking with Query Reformulation for Biological data

Bryan Brancotte[1,2], Bastien Rance[4,5], Alain Denise[1,2,3], and Sarah Cohen-Boulakia[1,2]

[1] Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623, Université Paris-Sud, 91405 Orsay Cedex, France
[2] AMIB Group, INRIA Saclay Ile-de-France - France
[3] Institut de Génétique et de Microbiologie (IGM), CNRS UMR 8621 Université Paris-Sud - France
[4] Biomedical Informatics and Public Health Department, University Hospital Georges Pompidou, AP-HP, Paris, France
[5] INSERM Centre de Recherche des Cordeliers, team 22: Information Sciences to support Personalized Medicine, Université Paris Descartes, Sorbonne Paris Cité, Faculté de médecine, Paris, France

Abstract. This paper introduces ConQuR-Bio which aims at assisting scientists when they query public biological databases. Various reformulations of the user query are generated using medical terminologies. Such alternative reformulations are then used to rank the query results using a new consensus ranking strategy. The originality of our approach thus lies in using *consensus ranking* techniques within the context of *query reformulation*. The ConQuR-Bio system is able to query the Entrez-Gene NCBI database. Our experiments demonstrate the benefit of using ConQuR-Bio compared to what is currently provided to users. ConQuR-Bio is available to the bioinformatics community at `http://conqur-bio.lri.fr`.

1 Introduction

In Biological research, findings are derived from the proper analysis of experiments which involves comparing at various scales new results obtained to existing data. Over the last three decades, scientists have had to face with an avalanche of data, of different kinds, and reported in a myriad of databases. Public biological databases thus contain more biological data than ever, all available to the scientific community. Large amounts of data can be easily obtained using portals such as Entrez NCBI[6] [14] daily used by the bioinformatics community by submitting *key-phrase queries* (list of keywords). However, properly querying such portals is not as easy as one may think. Two very similar queries may provide different sets of answers leading to the need for users to try various reformulations of their questions, considering synonymous terms, alternative spellings,

[6] http://www.ncbi.nlm.nih.gov/Entrez

various levels of granularity in the concepts involved in their queries (making use or not of the terminologies available such as MeSH [13] or SNOMED CT [16]). Results obtained should then be gathered, compared, and redundancies filtered out... Each set of results is ranked by the portal usually using the *relevance* as a ranking criteria (number of occurrences of the key-phrase in each piece of results instance). However, when several reformulations are considered, it is not clear how to rank the set of all the collected results, which may involve hundreds of elements. The expected ranking should be able to emphasize answers provided by various reformulations while putting less importance on elements classified as "good" by only a few.

The need for on-the-fly solutions both able to reformulate automatically queries exploiting the various terminologies available and rank answers provided to the user is thus of paramount importance.

In this paper, we introduce the ConQuR-Bio approach, which allows users to query public databases from NCBI while generating automatically all the possible reformulations and provides ranked answers using consensus ranking techniques.

The remainder of this paper is organized as follows. After a description of a set of use cases which have driven the design of our solution (Section 2), Section 3 introduces the architecture of our system. We present the original consensus ranking strategy we follow in Section 4. Section 5 introduces the interface and the main functionalities of the system we have implemented based on the ConQuR-Bio approach (available for use to the community at: `http://conqur-bio.lri.fr`). Section 6 provides the results obtained by ConQuR-Bio on several biological queries while Section 7 concludes the paper.

2 Use cases

Our approach is based on one of the most popular tool for querying biological sources, namely, the Entrez portal [15] from the *National Center for Biotechnology Information* (NCBI). More specifically, the kind of queries we consider consists in searching the gene names associated to a given disease by consulting the EntrezGene database [14] and focusing answers to human genes. We describe here-after a set of four use cases that we want to consider.

Use Case 1 (equivalent reformulations): Let us consider the case of a single user interested in genes involved in the cervical cancer. To express her query, she may type *cervix cancer* in the search field of EntrezGene. As a result, 460 genes are obtained. Interestingly, her query could have been expressed in two other ways, namely using *cervical cancer* and *cancer of the cervix*, leading respectively to 20 and 2 results, with 9 new genes of interest obtained (compared to the original query).

Use Case 2 (abbreviations): Another use case is related to the use of abbreviations in queries. Consider searching for genes associated to *Attention deficit hyperactivity disorders* also known as ADHD. While the full name of the disease returns 144 genes, its abbreviation provides 109 genes with only 74 in common.

Use Case 3 (lexical-based reformulation): Another typical use case consists in considering two users, one from the US the other from the UK, searching for *tumor suppressor genes* associated to the *breast cancer*. While the first one enters *breast cancer tumor suppressor*, the other enters *breast cancer tumo**u**r suppressor*. This orthographic variation leads to huge differences when querying the EntrezGene database: 681 genes are returned with *tumor* and 291 with *tumour*, and only 246 genes are common to both queries.
Use Case 4 (narrower-term-based reformulation): In a last use case, we consider the case of diseases presenting a variety of subtypes (usually corresponding to multiple phenotypes or a gradient of phenotypes associated with the disease). For example, when the *colorectal cancer* is hereditary and without polyposis it can be described by various names, including *Hereditary Nonpolyposis Colon Cancer*, also known as *Lynch syndrome*. Interestingly, querying the EntrezGene database with *Hereditary Nonpolyposis Colon Cancer*, and *Lynch syndrome*, allows to respectively find 1, and 6 genes which were not found when typing *colorectal cancer*.

From these use cases, the need for automatic reformulation of queries appears clearly as a necessity. Even more importantly, faced with the high number of answers obtained as result of each query (especially when several reformulations are considered), users should be guided in the order to which consider results. The originality of our approach lies in considering alternative reformulations of the user query and exploiting these reformulations to rank the results by order of interest (roughly, genes obtained by a large number of reformulations should be ranked before genes returned by only a few).

3 The ConQuR-Bio approach

In this section, we introduce ConQuR-Bio (Consensus ranking with Query Reformulation for Biological data) which aims at helping users finding genes associated to a given disease by considering various reformulations of each user query and exploiting such reformulations to rank the list of results. More precisely, our approach takes in several input rankings (several lists of genes, each provided by one reformulation) and outputs a *consensus ranking*, that is, a unified list considering all the input data ordered such that the disagreements between the list and the input rankings are minimized.

In the following, the main architecture of our approach is first presented, then two focuses are given, on the *reformulation module* and *queries generator module*.

3.1 General architecture

The standard use of ConQuR-Bio consists in the user providing a *key-phrase* k (i.e., a list of keywords). The key-phrase is sent (arrow 1 in Figure 1) to the *Reformulation Module* which decomposes k into a list T of terms and leverage various terminologies to generate the set S of synonyms (cf 3.2). S is then trans-

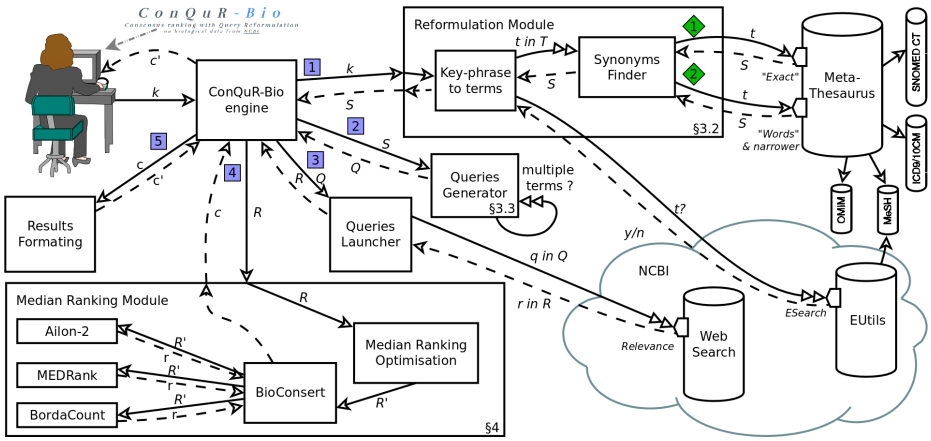

Fig. 1. Architecture of ConQuR-Bio. Solid arrows represent requests and dotted arrows responses. Two headed arrows represent possibly iterative requests. When several actions have to be done successively, their are numbered with a squared number. When alternative actions can be done, actions are represented with a diamonded number.

mitted (through arrow 2) to the *Queries Generator* to be expressed as a set Q of queries (cf 3.3). Q are run online (arrow 3) on the selected search engine (in our case, the NCBI web search engine for EntrezGene which provides sets of results ranked by *relevance*). When all the ranked results R of queries Q have been collected, they are sent 4 to the *Median Ranking Module* which is in charge of computing a unique consensus ranking, providing an ordering of all the answers (cf 4). Finally, 5 the *Results Formatting* module enriches the ranking of gene identifiers with names and descriptions.

A few parameters may be tuned by users, such as the selection of the species of interest (by default, *Human*) or the "Search deeper" option in which the *Reformulation Module* intends to find more reformulations for each term (details in 3.2). A default configuration is provided.

3.2 Reformulation Module

One of the two main modules of ConQuR-Bio is the *Reformulation Module*. It takes the user key-phrase as input, splits it into a list of terms and returns sets of reformulations for each term. The *Reformulation Module* leverages several medical terminologies within the UMLS® [3]. The terminologies are described here-after followed by the presentation of the process used to exploit such terminologies in ConQuR-Bio.

Terminologies used. ConQuR-Bio makes use of the Unified Medical Language System® (UMLS)[7], a terminology integration system developed at the U.S. Na-

[7] Version 2013AB of the UMLS is used in the current version of ConQuR-Bio and for the evaluation we provide in the next section

tional Library of Medicine (NLM). ConQuR-Bio uses the UMLS API to interact with the Metathesaurus® integrating more than 160 medical vocabularies. Our approach particularly benefits from the use of five terminologies covering a wide range of biomedical domains: (i) MeSH [13], developed at the U.S. NLM and designed for indexing PubMed; (ii) SNOMED CT [16], a worldwide used clinical terminology often used as a core for Electronic Health Records; (iii and iv) The two latest versions of the International Classification of Diseases (ICD 9 CM and ICD 10 CM), developed by the World Health Organization and used in hospitals; (v) The Online Mendelian Inheritance in Man (OMIM), cataloging all known genetic diseases in the human genome. Each UMLS concept is categorized with at least one *Semantic Type* (out of 150+) from the Semantic Network. The UMLS also provides a broad categorization of Semantic Types into 15 *Semantic Groups* (including Disorders). Using the Metathesaurus allows to access synonymous terms from the terminologies.

From key-phrase to MeSH terms. MeSH being de facto a *lingua franca* for biomedical literature querying, ConQuR-Bio starts with finding the largest recognized MeSH terms in the key-phrase provided by the user. More precisely, the key-phrase is decomposed into a list of terms where each term belongs to one terminology, but no concatenation of two or more consecutive terms belongs to any terminology. For example, the query "breast cancer oncogene" matches four MeSH terms "breast" "cancer", "oncogene" but also "breast cancer". The key-phrase is thus decomposed into the two terms "breast cancer" and "oncogene".

Reformulation modes. Once the MeSH terms in the query have been identified, ConQuR-Bio may follow two modes to find reformulated terms, leveraging the UMLS to identify synonyms of (1) the MeSH terms from the original query (default search mode), or (2) more precise (i.e. narrower) terms and their synonyms. In any case, alternative formulations of the query are generated. When only one reformulation is returned by the default search mode 1 (meaning that the term is recognized but has no synonym) then the second mode 2 (using narrower terms and their synonyms) is used. The second mode is also used in complement of the first mode when the *search deeper mode* is enabled.

Identifying synonyms (arrow 1). The default mode uses the UMLS API *exact match* search strategy to find UMLS concepts associated with each term. From these concepts, we extract all the synonymous terms from SNOMED CT, ICD9, ICD10 and MeSH, associated with this UMLS concept. For example, the term *cervix carcinoma* is mapped to the UMLS concept C0302592. This concept includes several synonyms, including *Cancer of cervix* (from SNOMED CT) and *Uterine Cervical Cancer* (from MeSH).

Identifying narrower terms (arrow 2). This alternative mode provides reformulations using narrower terms (in the sense of the organization of the hierarchy), which are thus more precise terms than the terms used in the original query. Synonyms of the narrower terms are also exploited. This mode corresponds to use UMLS API *word* search strategy. For example, using the "word" search strategy from with the term *Long QT syndrome* (UMLS concept C0023976) allows to identify several narrower concepts, including *Long QT syndrome type 1*

(UMLS concept C0035828, for which *Romano-Ward syndrome* is a synonym).
Semantic filtering. As searched terms are all expected to be diseases, only mappings to concepts from the UMLS semantic group *Disorders* are considered.

3.3 Queries generator module

The *Queries generator module* produces queries from the synonyms found for the terms identified in the user's key-phrase by the *Reformulation Module* (see 3.2). When the key-phrase has been split into multiple terms, we consider the Cartesian product of the reformulations of each term. Considering a key-phrase k composed of two terms a and b such as $k =$ "a b" and a, resp. b, is reformulated into $\{a, a'\}$, resp. $\{b, b'\}$. This module generates queries to search for "a b", "a' b", "a b'", "a' b'".

4 The Median ranking module

In this section we present the *Median Ranking module*, one of the major modules of ConQuR-Bio which provides a unique ranking to the user. This module takes in lists of elements (here, lists of genes), each list being obtained by a given reformulation. It outputs a *consensus ranking*, that is, a list of all the elements present in the inputs, ordered such that the disagreements between the consensus and the input rankings is minimized.

In the following, we first define the median ranking problem; we then show that a new metric is needed for our approach, and, for this purpose, we define a pseudometric for comparing rankings. Finally, we describe the heuristic that we have developed and tuned to compute consensus ranking, driven by the need to provide an on-the-fly solution.

4.1 The Median Ranking Problem

Starting with multiple rankings called *input rankings*, the Median Ranking Problem consists in finding one *ranking* able to minimize the distance to the input rankings. When the Kendall-τ distance is considered [12], the input rankings must be over the same elements and the problem of finding an optimal solution is known to be NP-Hard when more than 3 rankings are considered [9]. Polynomial-time approximation algorithms and heuristics have thus been proposed (*e.g.* [11,1]). In this paper, we will call *consensus* the solutions proposed by consensus algorithms (including heuristics or approximation algorithms), while we will use the term *median rankings* to denote optimal solutions.

We consider here rankings with ties, that is, rankings where some elements may be grouped into one bucket and may thus not been compared to each others. More precisely, each bucket contains at least one element, and two elements have a different rank iff they are in two different buckets. For instance, in the ranking $r = [\{B, A\}, \{C\}, \{D\}]$, the elements A and B are tied in a bucket and thus equally good, they are also better than C and D, and C is better than D.

As underlined in the use cases introduced in section 2, two reformulations may not necessarily provide the same sets of data (i.e., sets of genes obtained may be different from one reformulation to another). Unifying the data sets taken as input is then the first step to achieve to compute the corresponding consensus ranking. ConQuR-Bio makes use of the *unification process* introduced by [7] to consider input rankings over different sets of elements. This treatment adds a single bucket at the end of each ranking and places in this bucket all the elements that appear in other rankings but not in the current one. We call such buckets *unifying buckets*. For example, consider $r' = [\{C\}, \{E\}]$ and the ranking r introduced above. The unifying process provides the two unified input rankings: $r'_{unified} = [\{C\}, \{E\}, \{A, B, D\}_u]$ and $r_{unified} = [\{B, A\}, \{C\}, \{D\}, \{E\}_u]$, leading to two input rankings over the same sets of elements (A to E). Note that unified buckets are suffixed: $\{...\}_u$.

When considering ranking with ties, the distance used in the median ranking problem is the generalized Kendall-τ distance [11,7] defined as follows:

Definition 1. *Let r and c be two ranking with ties over n elements, c being a consensus. Let $r[i]$ be the rank of i in ranking r. The generalized Kendall-τ distance is:*

$$\begin{aligned} K^{(p)}(r,c) = \#\{(i,j) : & r[i] < r[j] \text{ and } c[i] > c[j] \text{ or} \\ & r[i] > r[j] \text{ and } c[i] < c[j]\} \\ +p * \#\{(i,j) : & r[i] \neq r[j] \text{ and } c[i] = c[j] \text{ or} \\ & r[i] = r[j] \text{ and } c[i] \neq c[j]\} \qquad \text{where } 0 < p \leq 1 \end{aligned}$$

This distance counts 1 for each pair of elements when their order is inverted, and counts p when two elements are tied in one ranking and not in the other. The distance between a consensus c and a set of input rankings R is the sum of the distances between c and the rankings in R: $K^{(p)}(R, c) = \sum_{r \in R} K^{(p)}(r, c)$. A *median* of a set of input rankings is defined as follows:

Definition 2. *Let $\mathcal{R}$ be the set of all rankings with ties over n elements, and let $R \subseteq \mathcal{R}$ be a set of rankings. A ranking c^* is called a* median ranking *of R iff:*

$$K^{(p)}(R, c^*) \leq K^{(p)}(R, r), \forall r \in \mathcal{R};$$

Example 1. Let us consider the set of input rankings $R = \{r_1, r_2, r_3\}$ where $r_1 = r_2 = [\{A\}, \{D\}, \{B, C\}_u]$, $r_3 = [\{B\}, \{A, D\}, \{C\}]$. The median ranking is $c^* = [\{A\}, \{D\}, \{B, C\}]$. The disagreements are: the order inversion of B-A and B-D (+2) plus A-D untying (+p) plus B-C tying (+p) thus $K^{(p)}(R, c^*) = 2+2p$.

4.2 A new pseudometric to compare rankings

The intuition behind the need for a new metric can be illustrated on the above example. Two points should be emphasized. First, elements A and D are tied in r_3 because the search engine ranked them at the same position, they thus

should be considered as equally relevant. Second, elements B and C are tied in r_1 and r_2 due to the unification process, contrary to the previous situation, no search engine has ever indicated any rank between such two elements (neither one before the other, nor both at the same position).

The generalized Kendall-τ distance does not allow to make a distinction between elements tied in an unification bucket from those tied in a classical one. A new metric taking into account the nature of buckets has thus to be defined. In particular, the metric should consider true disagreements between elements ranked by several reformulations while not penalizing any difference between the relative positions of elements present in the unifying buckets: our aim is to consider that untying elements from the unifying bucket has no cost.

Definition 3. *Let r and c be two rankings with ties over n elements. Let $r[i]$ be the rank of i in ranking r. Let $unif(r)$ denote the unification bucket of r. ($unif(r) = \emptyset$ if r has no unification bucket.) Let us define $\mathcal{M}(r, c)$ as follows:*

$$\begin{aligned}\mathcal{M}(r,c) = \#\{(i,j) :& r[i] < r[j] \text{ and } c[i] > c[j] \text{ or}\\ & r[i] > r[j] \text{ and } c[i] < c[j]\}\\ +p\#\{(i,j) :& r[i] \neq r[j] \text{ and } c[i] = c[j] \text{ and } i \notin unif(c) \text{ or}\\ & r[i] = r[j] \text{ and } c[i] \neq c[j] \text{ and } j \notin unif(r)\}\end{aligned}$$

Clearly $\mathcal{M}$ is not a distance as it may not be always possible to distinguish two different rankings: $\mathcal{M}([\{A\},\{B\}],[\{A,B\}_u]) = 0$. However, it is a pseudometric [17] as the symmetry and triangular inequality properties are respected, and any element has a metric at zero compared to itself: $\mathcal{M}(r, r) = 0$. Similarly to the generalized Kendall-τ distance, when considering a consensus c and a set of input rankings R: $\mathcal{M}(R, c) = \sum_{r\in R} \mathcal{M}(r, c)$.

Example 2. Let us consider a set of input rankings $R = \{r_1, r_2, r_3\}$ where $r_1 = r_2 = [\{A\},\{D\},\{B,C\}_u]$, $r_3 = [\{B\},\{A,D\},\{C\}]$. Under the generalized Kendall-τ distance, the median ranking is $c = [\{A\},\{D\},\{B,C\}]$ (cf. *Example* 1) while under the pseudometric $\mathcal{M}$ the median ranking is $c' = [\{A\},\{D\},\{B\},\{C\}]$ as $\mathcal{M}(R, c) = 2 + p > \mathcal{M}(R, c') = 2$ (note that $K^{(p)}(R, c') = 2 + 3p$). From a user perspective, c' is a better median than c as it still promotes A and D, but also makes use of information provided by r_3 such as the fact that B is more relevant than C.

Other strategies have been developed in [9] in order to deal with sets of rankings which are not necessarily over the same elements: the *induced* Kendall-τ distance allows to compare a ranking c over all elements with a ranking r over a subset of these elements. The idea is to consider the projection of c onto r, by removing from c all elements that are missing in r. However, this distance is not relevant for our purpose as it does not allow to consider missing elements as being less relevant than the returned ones (the missing elements of r are completely removed from c and thus do not contribute to any (dis)agreement).

4.3 Median Ranking in the context of query reformulations

BioConsert [7] is an heuristic designed in the context of biological data and considers a distance between rankings with ties. It uses each input ranking as starting point, and refines them by iteratively applying two edit operators (moving an element to an existing/new bucket) as long as the distance between the current consensus obtained and the input rankings is reduced. Finally, it returns the best consensus computed. Our approach differs from [7] by using the pseudometric $\mathcal{M}$, presented in §4.2, instead of the generalized Kendall-τ distance. $\mathcal{M}$ is parametrized by $0 < p \leq 1$ which expresses the importance of tying and untying elements. In our setting, tying and untying elements should be penalized while when two elements have the same number of rankings placing one element before and after the other, the two elements should be tied. As a consequence, we have set $p = 0.5$ in ConQuR-Bio.

Tuning BioConsert. ConQuR-Bio is an on-the-fly system which intends to quickly provide a consensus ranking from the reformulations obtained. To do so, it requires to have a fast and good algorithm to produce the *consensus of answers*. The time complexity of BioConsert depends, among other parameters, on the number of input rankings m. In order to speed up the computation, we consider a smaller and constant amount of rankings to start the algorithm (and not all input rankings as in [7]). More precisely, we selected three state-of-the-art algorithms: BordaCount [4], MEDRank [10], and Ailon's 2-approximation [1] which do not provide as good results as BioConsert, but provide solution in at most complexities of $nm\ log(nm)$, where n is the number of elements to be ranked. Experiments (not shown here) performed to compare this new strategy to the default strategy of BioConsert show that the time to compute a consensus is reduced up to one hundred times while the quality of the results is not significantly altered.

5 The ConQuR-Bio system

The main interface of ConQuR-Bio is provided in Figure 2 and composed of three areas, the query area (top left panel), the running and progression details (top right panel), and the results (bottom).

In the query area, the key-phrase provided by the user is split into MeSH terms on-the-fly (cf. 3.2) and displayed into colored boxes next to the key-phrase field. Colors indicate different status for a term: green when the term is recognized as a MeSH term, red when the term is not recognized, and orange when the term is matched with an existing MeSH term while the spelling is different. In addition to the orange semantics, when a term is matched with an alternative spelling, a check mark allows the user to accept the correction and update the key-phrase field, while a cross mark forces the system to use the given spelling. Several options are made available to the user, and are by default hidden. They can be displayed/hidden by clicking on "[+]"/"[-]" like in Figure 2). Options

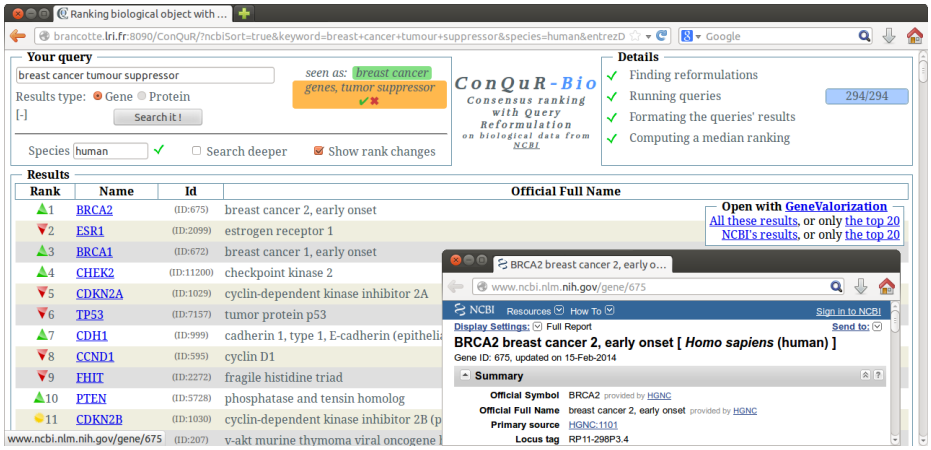

Fig. 2. ConQuR-Bio interface and the window open after clicking on BRCA2.

are the species considered, the "Search deeper" mode which allows to use reformulations with narrower terms (cf. in 3.2), and the type of biological object ranked.

The results area presents a ranking (with ties) of genes with their official descriptions as it can be found when browsing the NCBI website. Each gene is linked to its associated page in the NCBI Website, allowing the user to navigate in a familiar environment. Close to the rank of each gene, a symbol (hidden in the default mode) allows users to know whether the rank of the gene is raised (), equal (), lowered (), or new () in ConQuR-Bio compared to the results returned in the NCBI ranking.

Another interesting feature is the ability of ConQuR-Bio to provide users with information on the number of publications associated with each gene returned. This functionality is obtained by calling the GeneValorization[6] tool able to quickly browse PubMed.

6 Results on medical queries

We have tested our approach over a set of queries collected from collaborators of the *Institut Curie (France)* and the *Children's Hospital of Philadelphia (PA, USA)* and linked to their respective fields of expertise. The results presented considered 9 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*). For cancers, we searched for information on the name of the cancer while also using additional words (and reformulations of such words) to refine the query, namely *tumor suppressor* and *oncogene*. The exact list of words used are shown in Figure 3.

Evaluating such an approach is a difficult task as we face the users' perception of the results. We have chosen to consider three criteria of evaluation, focusing on the 20 first results returned for each key-phrase (top-20). The first criterion is based on *Gold Standards* and compares the results obtained to the list of expected genes according to our experts. We classically use the area under the ROC curve [5] in this series of experiments. The next two criteria are bibliometrics ones: the second criterion is the number of publications associated with each gene of the list and the key-phrase while the last criterion is a "freshness" indicator, measuring the average number of days since such an article has been published. The assumptions behind such measures is that well-studied genes are more likely to be relevant and experts can be interested in the latest, up-to-date, information.

6.1 Using expertise

We constructed with our clinician collaborators the list l_d of the most relevant genes known to be associated with each disease d. The "goodness" of a consensus ranking c_d provided by ConQuR-Bio thus relies on the presence of elements of l_d in the top-ranked elements of c_d. In order to compare the results returned by ConQuR-Bio and the EntrezGene NCBI Web search engine with respect to *Gold Standards*, we used the Area Under the ROC Curve [5] (ROC standing for *Receiver Operating Characteristic*) or *AUC* (closely related to precision and recall measures [5]). The AUC aims at differentiating the presence of expected data versus non expected data, taking into account the place of pieces of data (roughly, placing expected data before unexpected data increases the score of the AUC). AUC provides numbers ranged in $[0, 1]$, 1 being the highest score.

In Figure 3, we plot AUCs for the top-20 first results obtained for each key-phrases with both NCBI search engine and ConQuR-Bio. Globally, using ConQuR-Bio compared to NCBI allows to increase in average the AUC of 44.24%. More precisely, four points deserve attention.

First, when focusing on single term key-phrases (i.e., considering the name of the disease only without adding *oncogene* or *tumo[u]r suppressor*, corresponding to Figure 3.a and all use cases), ConQuR-Bio returns better results than the NCBI in 88.89% of the cases and always provides as good results as the NCBI. The average AUC is increased of 58.52% with ConQuR-Bio compared to NCBI.

Second, multi-term key-phrases (Fig 3.b,c (use case 3)) have an AUC increased of 37.70% in average when using ConQuR-Bio compared to NCBI. This relatively less good results (37% vs. 58% of improvement) is actually due to the fact that the term *oncogene* has, in addition, one reformulation (*gene transforming*) less interesting (considered as "too vague" by our experts) than others.

Third, considering *ADHD* and its unabbreviated name (use case 2), the AUC is drastically increased using ConQuR-Bio. Also, as expected the complete name and its abbreviation have different AUCs with the NCBI while remaining the same with ConQuR-Bio (since all the reformulations are considered). In the same spirit, lexical variations around the *cervical cancer tumor suppressor* (Fig 3.b) show the importance of taking into account all lexical and orthographic

variations: ConQuR-Bio returns identical results for the four variants with an AUC of 0.53 while NCBI results have systematically inferior and variable AUCs.

Finally, there were a few key-phrases, namely *colorectal cancer* and *neuroblastoma*, for which only plural reformulations were actually available (no actual synonyms available). The results obtained for such queries are then less impressive than in the previous cases while some of their respective AUCs are still increased compared to NCBI.

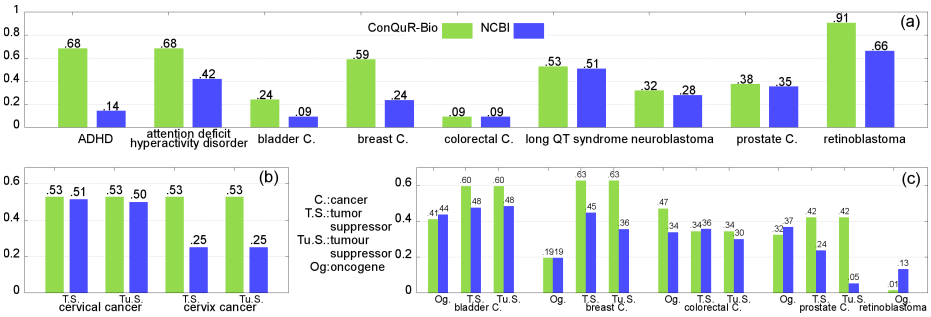

Fig. 3. The Area under the ROC curve (AUC) for the 20 first genes returned by ConQuR-Bio and the NCBI WebSearch for (a) Single-term key-phrases, (b) lexical variation around *cervix cancer tumor suppressor*, and (c) the remaining key-phrases.

Our experiments have shown that all the reformulations associated to use cases 1, 2, 3 were taken into account and that using our approach based on consensus ranking systematically improved the answers provided to the user. However, we have not yet provided specific information on the use case 4 which made use of lexical narrower terms. Two points should thus be mentioned.

First, interestingly, narrower terms have actually automatically been exploited in the previous results for the *long QT syndrome* as this term did not have any synonym. Specific forms of the disease, such as the *Romano Ward Syndrome*, raise the AUC from 0.51 to 0.53.

Second, the use of narrower terms can be done manually by selecting the "search deeper" option. Back to the example illustrating the use case 4, using narrower terms drastically change the results: among the 11 genes provided by our experts as being very relevant for *colorectal cancer*, only 2 are in the top-20 results of the NCBI (AUC=0.09) while 6 are in the ConQuR-Bio first 20 answers (AUC=0.43).

A last point that deserves attention is the time taken by ConQuR-Bio to provide answers: While the NCBI search engine provides a ranking in at most $2s$, ConQuR-Bio takes $41s$ in average for the 9 single term key-phrases listed in Figure 3.a. This difference lies in the fact that the average number of synonyms retrieved by ConQuR-Bio (and thus the average number of queries to be answered and which elements should be ranked) is 17.

6.2 Using the number of publications

The second measure considers the top-20 genes obtained and sums the number of publications co-citing each gene name and the query key-phrase. As an example, the numbers of publications associated with the top-20 first genes returned for *retinoblastoma* by the NCBI and ConQuR-Bio are represented in Figure 4. It clearly shows that the top-20 genes provided by ConQuR-Bio are associated to more publications than the top-20 genes provided by NCBI.

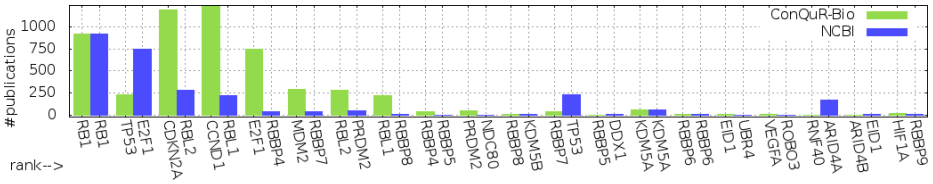

Fig. 4. #publications for each of the 20 first ranked genes for *retinoblastoma*

More generally, over the 28 key-phrases studied, 25 provide more (or, in 2 situation equal) publications than the NCBI. Overall in average, ConQuR-Bio returns top-20 results associated with 56% more publications.

6.3 Using publication freshness

While the number of publications is one important factor for determining the level of interest associated to a result, another complementary factor is the freshness of the associated publications (i.e. how recently studies based on a given gene have been published). The measure we consider in this subsection computes the average number of days since the last publication co-citing the gene name and the key-phrase has been published.

Over the 28 key-phrases studied, and when considering the top-20 genes, ConQuR-Bio returns genes with fresher results for 22 of them. In average, the top-20 genes returned by ConQuR-Bio have one associated article which was published within 25% less days that the NCBI ones.

7 Discussion

With ConQuR-Bio, we made the connection between the *query expansion* field and the *median ranking* field. We leveraged terminologies integration in the UMLS system (an approach and system shown to be effective [8]) to propose reformulations. From two UMLS search modes, we provided reformulations based on MeSH terms identified in the users key-phrases. To generate a consensus answer to the user emphasizing the agreements between the reformulations, we backed its computation on a new pseudometric, extending the state-of-the-art

generalized Kendall-τ distance. With this new pseudometric, we adapted and combined several median ranking algorithms, allowing the system to quickly compute a consensus. We compared our approach to the main portal used to browse gene-centric biological data, namely the EntrezGene database from the NCBI website and its ranking function based on relevance. We showed that when measuring the presence and order of expected results (based on Gold standards), ConQuR-Bio outperforms the NCBI with an AUC increased of 69.30%. When focusing on biometrics indicators and compared to the NCBI relevance sorting, ConQuR-Bio returned genes associated with 56% more publications, published in 25% less days. Last but not least, we made the system available and free to use at `http://conqur-bio.lri.fr` as a website.

We now provide a discussion and perspectives considering the various steps of our approach.

ConQuR-Bio starts with identifying MeSH terms from key-phrases. We have currently chosen to follow a greedy (and naive) process enabling a very fast answer rate, compatible with the on-the-fly feature of our approach. This strategy is entirely satisfactory on evaluated key-phrases. Future work will explore the detection of concepts from the users key-phrases by deploying concept recognition software such as MetaMap [2] or BioAnnotator, enabling advanced reformulation options (e.g. different levels of granularity). Providing results in a few seconds while augmenting their overall quality will be the most challenging point.

The reformulation module plays a major role in the quality of the results. This module is based on two components: the set of terminologies used and the way such terminologies are queried and exploited.

As for the terminologies, we currently use terminology sources from the UMLS which allowed us to have manageable and relevant amounts of reformulations. Ongoing work includes selecting a larger and customizable number of sources from the main two biological terminology integration systems (namely, the UMLS and the BioPortal [18]) to cover a broader scope of biological domains. To cope with the possibly too broad aspect of reformulations, we plan to allow (experienced) users to select the reformulations to be or not to be used by our system.

As for the way terminologies are exploited, in our current version, the "search deeper" mode provides narrower reformulations. However, work still have to be done as the semantics of this mode is very permissive and does not exploit the hierarchical feature of the links between concepts. The UMLS system provides typed links for broader and narrower concepts unified between terminologies, and their adequacy should be evaluated. Ongoing work consists in exploiting the hierarchical relations from the sources to improve the detection of concepts and their synonyms.

References

1. Nir Ailon. Aggregation of Partial Rankings, p-Ratings and Top-m Lists. *Algorithmica*, 57:284–300, 2010.

2. Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
3. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
4. J.C.de Borda. Mémoire sur les élection au scrutin. *Histoire de l'academie royal des sciences*, pages 657 – 664, 1781.
5. Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.
6. Bryan Brancotte, Anne Biton, Isabelle Bernard-Pierrot, François Radvanyi, Fabien Reyal, and Sarah Cohen-Boulakia. Gene List significance at-a-glance with GeneValorization. *Bioinformatics*, 27(8):1187–1189, 2011.
7. 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, pages 73–90. Springer, 2011.
8. Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F Loane, Bastien Rance, François-Michel Lang, Nicholas C Ide, Emilia Apostolova, and Alan R Aronson. A knowledge-based approach to medical records retrieval. In *TREC*, 2011.
9. C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th World Widw Web conference*, pages 613–622, New York, NY, USA, 2001. ACM.
10. 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*, pages 301–312. ACM, 2003.
11. Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '04, pages 47–58, New York, NY, USA, 2004. ACM.
12. M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.
13. Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.
14. Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 39(sp1):D52–D57, 2011.
15. Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011.
16. Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. Snomed clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*, page 662, 2001.
17. Lynn Arthur Steen, J Arthur Seebach, and Lynn A Steen. *Counterexamples in topology*. Springer, 1978.
18. Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545, 2011.