Title: Ranking biological and biomedical data: algorithms and applications
Supervisor (s): Sarah Cohen-Boulakia and Alain Denise
Contacts: cohen@lri.fr, denise@lri.fr
Location / Team: Team Bioinformatics LRI and INRIA AMIB, Bat 650 Université Paris-Sud
In collaboration with:

Ulf Leser (Professor for Knowledge Management in Bioinformatics at Humboldt Universität in Berlin)

The design and evaluation of the methods will be done in close collaboration with oncologists from the Institut Curie and the Children's Hospital of Philadelphia.

Context

The extremely rapid growth in the number of public biomedical data sources (more than 1,000 recorded in January 2012) associated with the exponential growth of data available in these sources makes the research phase of biomedical information increasingly complex and tedious. Several query systems of biomedical sources are now available and used by physicians (e.g., EntrezNCBI, SRS@EBI) but none of them offer to rank the data obtained in response to a query. This is particularly problematic in a context where simple queries can return thousands of results. For example, looking for *scientific research articles on the role of herceptin in breast cancer* returns (in January 2012) more than 1,600 responses.

Many *classical ranking* techniques can be considered for biomedical data, from very simple (eg, focus on the most recent data) to complex (e.g., techniques such as PageRank involving links between data, and / or combining the reliability of sources providing information [Birkland et al, 2006]). These algorithms mainly exploit two key aspects of bio data: (i) they are linked to each other's forming an intricate network (graph-based algorithms) and (ii) they are plain-text file (text mining like algorithms). Each of these techniques is based on an algorithm that has a given complexity. Accordingly, for the same query, depending on the technique chosen, the time cost of ranking data can vary from several seconds to several hours. The space needed for the computation of ranked data can also be a bottleneck facing the mass of information available.

As each ranking method returns a given ordering between data elements that is potentially different from one method to another, techniques able to calculate *consensus rankings* (aka *median*) have been proposed such as [Fagin et al, PODS 2004], [Ailon, Algorithmica 2010], and [Cohen-Boulakia *et al*, SSDBM 2011]. Instead of considering ranking methods as alternative approaches, consensus ranking methods propose to generate a new ranking able to highlight the common points of a set of rankings while minimizing their disagreements. The computation of the consensus ranking of a set of m rankings is then known to be NP-hard when m>4. Several (exact and approximated) algorithms and heuristics have been proposed.

Objectives: The aim of this thesis is to propose a uniform framework able to represent and compare (conceptually and on real data) both classical and consensual ranking methods and bring up new approaches for ranking large masses of biological data.

Work program:
The thesis will have 4 main tasks:

(1) **Formal study of the ranking algorithms**. After all the ranking algorithms possibly useful for ranking biological data identified, algorithms will be described in a uniform way so that the similarities and differences with each others can be underlined. In particular

- The time and space complexity of each algorithm will be evaluated. As for NP-hard problems, determining which assumptions and parameters cause the complexity will be a key point.
- On the classical algorithms, comparison criteria will include the kind of structure expected (e.g., some solutions may be able to work on acyclic graphs only), the kind of information expected as input (e.g., some solutions may be only based on the topology of the graph of data while other may need information on the presence of keywords in each data file to make use of text mining techniques)…
- On the consensus algorithms, comparison criteria may include the kind of data sets to be considered (e.g., ability to compare elements possibly partially or only fully ordered), the distance used, the operations considered to go from one rankings to another…

(2) **Evaluation of the ranking algorithms**. Data obtained by the various ranking algorithms on the one hand and expected or random data on the other hand will be compared. This comparison is a key aspect of the thesis as the criteria chosen for the comparison should allow to evaluate the relative *relevance* of the ranking algorithms. Several metrics will be considered (precision at k, average precision, ROC…). More precisely two kinds of data will be considered in the comparison process.

a) **Real data:** In close collaboration with physicians and biologists from the Curie Institute and the Children's Hospital in Philadelphia, study of the behavior of ranking algorithms against real data. In this step, indicators of relevance and explanation of the results obtained should be designed. At least two elements are to be taken into account:
    - *Gold standards:* Queries and expected ordered answers will be collected, forming the gold standards. Each gold standard will have the following form: query (e.g., *genes associated to the disease D*) and expected answers (e.g., *geneid1, geneid23, geneid45*). Results obtained by the algorithms studied in (1) will be compared with gold standards.
    - *Data provenance*: explanations on the presence of some information to a given place in a ranking should be provided to the user. The data provenance (where does the data come from?, from which source? etc.) is particularly important to be considered in this phase. Provenance information can be very large, it will thus be necessary to design methods (algorithms on graph structures) for providing the user with a subset but informative provenance information.

b) **Random data:** algorithms generating random rankings will be designed and their results compared with ranking methods.

(3) Reducing the search space (graph-topology). The ranking algorithms studied are inherently complex. The aim of this step is to study the size of the search space to be explored for the results to be still reliable while obtained in a reduced time. As an example, some ranking algorithms may use the length of the paths crossing a given piece of data to rank it. Studying the benefit of considering a path of length L+1 instead of L (that is, determining how many data are ranked in a better way when L+1 is considered instead of L) would help minimize the size of the subgraph to be considered.

(4) Mixing graphs-based and text mining-based techniques to rank biological data
This part of the thesis will propose to introduce complex text-mining techniques into ranking algorithms following research directions described in [Cohen-Boulakia and Leser, ICDE 2011].

Prerequisites: Knowledge of the structures of biological data sources and skills in object-oriented programming is expected.