Title: Efficiently Ranking big biological and biomedical data sets using rank aggregation techniques
Team: Bioinformatics group, Laboratoire de Recherche en Informatique, Université Paris-Sud (Bioinfo@LRI, Sarah Cohen-Boulakia, Alain Denise) + see the targeted collaboration below.

Context: The aim of biological data ranking is to help users faced with huge amount of data and choose between alternative pieces of information. This is particularly important in the context of querying biological data integration systems, where very simple queries can return thousands of answers. The need for ranking solutions, able to order answers, is crucial for helping scientists to organize their time and prioritize the new experiments to be possibly conducted. However, ranking biological data is a difficult task: biological data are usually annotation files reflecting expertise (with various degrees of confidence and quality); data are linked by cross-references and the network formed by these links plays a role in the popularity of the data; the need expressed by scientists varies, whether the most well-known data should be ranked first, or the freshest, or the "most surprising". As a consequence, although several ranking methods have been proposed in the last years within the bioinformatics community, none of them has been deployed on systems currently in use.

Originality: Our original approach is to rank biological data by considering rank aggregation techniques in a two steps process: (i) several ranking methods are applied to biological data, generating several input rankings (results are ordered using alternative ranking criteria and/or exploiting various ranking methods), (ii) rank aggregation techniques are used to reflect the input rankings' common points while not putting too much importance on elements classified as "good" by only one or a few rankings (i.e., minimizing their disagreements). The topic of rank aggregation has been of interest in various communities including information retrieval/database [FKM+04] [BYB+15], algorithmics [Ail10] and artificial intelligence [PHG00] to name a few. The problem is known to be *difficult* and while providing rank aggregation is a crucial need for big biological data sets, designing scalable algorithms is, by essence, highly challenging.

Expertise of the team, current situation: Bioinfo@LRI has a strong expertise in rank aggregation which is at the intersection of databases and combinatorics and in successfully applying such techniques to bio data (collaboration with the Children's Hospital of Philadelphia, Institut Curie and APHP George Pompidou). In particular, we have designed ConQuR-Bio [BRD+14] (http://conqur-bio.lri.fr/) to query biological databases and rank answers using rank aggregation techniques.

Limitations (Internship): ConQuR-Bio has to deal with an increasing volume of data sets to be ranked. Its core algorithm should be optimized to provide quick and high quality results in the first data items ranked (the 15 to 20 top-elements should be ranked as precisely as possible). The purpose of the (6 months) M2 internship (**3K€**) is to design and implement a series of evolutions for ConQuR-Bio to: (i) design new efficient consensus ranking algorithms able to provide excellent results in the top-20 elements, (ii) carefully evaluate the results obtained in dedicated biomedical queries.

Targeted collaboration in CDS 2.0: Our current users include members of the Hospital Paul Brousse, APHP (Christophe Desterke and Ivan Sloma). More specifically, we aim to use and ConQuR-Bio and its evolutions to evaluate and better priorize the importance of the genes involved in *leukaemogenesis*.

Bibliography

[FKM+04] R. Fagin, R. Kumar, M. Mahdian *et al.* Comparing and aggregating rankings with ties. Proc. of SIGMOD, pp. 47-58, 2004
[PHG00] D. M. Pennock, E. Horvitz, C. L. Giles, *et al.* Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering. In Proc of. AAAI, pp 729--734, 2000.
[Ail10] N. Ailon. Aggregation of partial rankings, p-ratings and top-m lists. Algorithmica, 57(2):284--300, 2010.
[BRD+14] B. Brancotte, B. Rance, A. Denise, S. Cohen-Boulakia. ConQuR-Bio: Consensus ranking with query reformulation for biological data. In Data Integration in the Life Sciences, pp 128--142.
[BYB+15] B. Brancotte, B. Yang, G. Blin, et al.: Rank aggregation with ties: Experiments and Analysis. In Proc. of VLDB 2015.