Rank Aggregation for Big Biological Data

Bryan Brancotte[1], Sarah Cohen-Boulakia[1], Alain Denise[1] et Stéphane Vialette[2]

[1] Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623,
Université Paris-Saclay, Bâtiment Ada Lovelace, 91405 Orsay, Cedex

Laboratoire d'Informatique Gaspard-Monge, Université Marne-la-Vallée, UMR 8049,
Bâtiment Copernic - 5, bd Descartes, Champs sur Marne 77454 Marne-la-Vallée Cedex 2

The aim of biological data ranking is to help users face with huge amounts of data and choose between alternative pieces of information. This is particularly important in the context of querying biological data, where very simple queries to the huge repositories of biological data can return thousands of answers. Ranking biological data is a difficult task: data may be associated with various degrees of confidence; data are not independent of each other's; various ranking criteria can be considered (the most well-known data ranked first, or the freshest, or the most surprising, etc.). Rank aggregation techniques, which consists in aggregating several rankings into one consensual ranking, are very promising in this context. However, such approaches are intrinsically complex. A plethora of algorithm approximations and heuristics have thus been designed, making the choice of the approach to follow very difficult to make for the user. In the RankaBio project we have carefully studied the problem of rank aggregation for biological data, both practically and fundamentally by (i) performing a comparative study of rank aggregation algorithms, (ii) providing new results on the complexity of the problem and (iii) designing concrete tools able to efficiently rank biological data obtained as answer to a query on major biological data repositories (http://conqur-bio.lri.fr/).

This work has been supported in part by the CNRS PEPS "Fascido" (project RankaBio).