Reviewer's report

Title: GenDrux: A guided system for supporting prescription by gene expression-based sensitivity

Version: 2 Date: 10 September 2010

Reviewer: Adrien Coulet

Reviewer's report:

- Major Compulsory Revisions

1 Even if the aim of the presented work is honorable, the paper never refers to relative works.

Other groups worked or are working on gene-drug relationships extraction from text and on grouping those in public repository for guiding personalized medicine. For instance authors should compare their results to what is done by PharmGKB. PSB 2010 and 2011 workshops on the subject are probably good starting points.

The paper does not provide any evaluation of presented results.

In addition presented results are below the current state of the art in relation extraction.

2 "In 2008, more than 700 publications related to breast cancer and drug sensitivity were recorded. More than 60 papers have been published during 2009 alone, and are increasing."

Authors must explain the origin of these numbers.

3 "This information is available in the non-annotated, free text of the biomedical literature, abstracts and full length articles for which can be accessed at the web pages of Medline."

According to this sentence it can be understood that Medline displays full text article.

4 Authors claim that GenDrux can help directly physicians in prescribing drugs. This claim deserves at least to be discussed by authors.

For instance can GenDrux extract and present negated relationships or hypothetical ones?

If not, it is of importance to aware physicians of these limitations.

5 Authors describe their use of e-utils but do not give any reference to this tool set.

6 E-search returns batches of 500 hits. Authors describe a common methods to
overcome this limitation.
This part is not very informative.

7 "breast+cancer+gene+drug"
Author must make more clear that the pattern contains a constant "breast+cancer" and two variables, gene and drug.

8 E-utils has been chosen in this study to access titles and abstracts.
It would be of interest to explain why it has been preferred to other tools (such as lingpipe).
Is there any constraints associated with e-utils? Can you expend your study to full (free) text?

9 It would be of interest to mention how named entities are recognized in titles+abstracts. Is it string matching?
If so can GenDrux be improved by using existing tools such as BANNER?

10 In the section "Main data sources and databases" it is really hard to understand precisely how existing/public resources have been compiled.
For instance "from the Food and Drug Administration" is not really informative.
"the major annotation databases including NCBI for genes" Please precise which NCBI database.
The paper will gain in clarifying this section. Insights on the size of the knowledge base would help the reader.
Is it possible to map entities of the knowledge base to any reference ontology?

11 Can authors give the time needed to populate the knowledge base and the one requiered for the system to answer an usual query?

12 "The performance of processing was not affected by the size of the lists"
If authors double the number of named entities do the processing time is still the same? With the same quality?

13 Precision, recall measures would be helpful to judge on the quality of the extracted information.

14 Are the resulting gene and drug name (and synonym) lists are planned to be shared with the community?
Same question for the index gene X pmid and drug X pmid that has probably been generated.

15 Table 1 mentions a sensitivity ("increase" or "decrease"). It is not clear where this sensitivity is coming from.
It is even more important to develop since sensitivity is mentioned in the title of the article.
16 Authors claim "The system is scalable".
Please justify.

- Minor Essential Revisions

17 GenDrux and GenDRux are used alternatively by the authors. Please homogenize.
18 "five to 10 years" => five to ten
19 "This is considered the first step "to bring molecular-based medicine into current practice.""
   It is unclear where this citation is from.
20 "gene chips/arrays will be widely available for purposes of diagnosis, correlation, outcome prediction, and prescription guidance."
   I disagree about "correlation"; microarrays are already frequently used to propose correlations.
21 "The results of e-fetch are abstracts"
   Is it only abstract or title+abstracts?
22 "XML (eXtended Markup Language)"
   If authors want to expand XML abbreviation, it must be precised the first time XML is mentioned.
23 Authors are using impact factors of journals to score extracted relationships. It would be of interest to precise where these impact factor are taking form.
24 Title and running title are different.

- Discretionary Revisions

25 "The only parameters that need to be changed are a medication database and a disease domain."
   The notion of disease domain is confusing. Is it only a word to change when querying the knowledge base or is it a new knowledge base to build?
26 Figure 4 is useless.
27 Ref 7 is to format.
28 Figure 1 says "Web interface (add later)"
   Is it about the interface shown in Figure 2?
Level of interest: An article of limited interest

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.