Author's response to reviews

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

Authors:

Chiquito Crasto (chiquito.crasto@gmail.com)
Feliciano Yu (FYu@peds.uab.edu)
Andres Forero (mate@uab.edu)
Dongquan Chen (dchen@hsc.wvu.edu)

Version: 4 Date: 16 October 2010

Author's response to reviews: see over
Title: GenDrux: A guided system for supporting prescription by gene expression-based sensitivity

Version: 2 Date: 7 September 2010

We are grateful to the reviewers for their astute comments and criticisms regarding this paper. We have endeavored to make several changes, including title, to address these criticisms. The primary change is in the philosophy of this paper. It is now presented as a tool that will allow researchers and clinicians to query a large corpus of literature to access focus instances of information. This is different from presenting GenDrux as a tool that will necessarily help physicians make informed decisions related to prescriptions in chemotherapy. Not being able to do this is not a short-coming of the tool, but rather because of the lack of data and pharmaceutical product development represented in the literature—which one hopes, will change in the future.

Because this tool will undergo continuous updates since the archiving process is fully automated, we project that as reports from drug development and gene-based profiles from high throughput studies become available, we will likely see personalized-medicine based decision making.

Our responses to the reviewers' comments will be addressed in red.

Reviewer: Miguel Vazquez

Reviewer's report: 
The authors have developed a system that connects genes, drugs and diseases by co-publication in scientific articles. The applications stores these articles, offers a query interface to retrieve them, and shows them on a HTML table that lists the gene, and drug entities mentioned in the article, and other information such as journal and impact factor.

After carefully revising the manuscript and the application I would have to lean towards a *Major Revision* of the document. The reasons have to do both with scientific relevance of the work and with the crafting of the manuscript itself.
Following is a justification of this decision that will be further summarised into major, minor and discretionary revisions.

The document is structured into three sections, introduction, implementation and results and discussion, all three failing to completely achieve their purpose. The introduction presents the problem of predicting drug sensitivity from gene expression. The discussion in this section is not sufficiently sourced so that neither the importance of the problem nor the applicability of the GenDrux system are correctly established.

We have increased the number of references to software related to name-entity recognition and describe how GenDrux differs from these. We have shown how GenDrux can be advantageous. GenDrux is a supervised system where the name-entities are filtered from universally validated sources of drug names like FDA and resources of genes such as Genbank and Affymetrix annotation.

The authors claim that the application can help clinicians prescribe drugs taken drug sensitivity into account as assessed by gene expression profiles; as a non-expert in personalised medicine, it still remains unclear to me the extent to which information gathered from the tool can be actually be applied in practice.

In particular with regards to the profiling tools available; how likely is it that the technology is able to address the findings derived from the application? Direct applicability of the system to the clinical practice need not be a requirement, even though the systems is presented in that way, but it at least merits some comments. As has been indicated in the paragraph at the beginning of this document, the lack of sufficient data has caused the authors to reassess the main thrust of this paper. We now present this as a tool that will allow efficacious retrieval of information related to drugs and genes from the breast cancer literature. The basic algorithmic development will be successful in helping investigators and physicians make prescription based decisions when sufficient data and information about development becomes available.

The implementation section delves amply into technical details such as the use of E-utils or the database system used, which are best left for additional material or
documentation of the source code, while not specifying how critical steps of the application are addressed. In particular, for the detection of named entities (drugs, genes and diseases) only the source lexicons are listed but not the matching strategy used.

This is now addressed. The authors refer to exact matches in the literature because that is the preferred manner in which investigators present drug and gene names in the literature. In addition, the mapping of a synonym and other nomenclature of a gene to a gene is also described. Therefore, even a synonym for a gene will be identified in GenDrux.

The criteria for querying relevant articles in PubMed is unclear. It seems like it uses two annual queries since 1975 and takes only 500 for each, to a total of 8000; the authors could explain better how these queries are performed in PubMed (the search filters used in the query for example) and why they followed this particular strategy instead of just taking the same total number of documents from the most recent or relevant to the query.

The rationale for the number of queries per year is now described in detail. The number of queries per year is based on manual validation of the number of relevant articles (breast cancer+gene+drug) for those years. It is not two queries per year since 1975: But two queries per year between 1975 and 1985; four queries a year between 1986 and 1995; and thereafter, one query a month. This has been clarified in the revised manuscript. We also show from a PubMed search that this rationale for how many queries per year was justified.

The results and discussion section provides a system architecture overview, more technical details such as security certificates, considerations about impact factors, etc. Most of these considerations belong to other sections or could be omitted. The manuscript would greatly benefit from including in this section the description of some example analysis performed using the application that could help discussed some of its characteristics.

We have omitted these sections. We have considered the reviewer’s recommendation and agree that it does not add to the main thrust of the documents.
With regards to writing, the manuscript needs much reworking. The language seems un-precise at times, and discourse does not follow a clear structure along the text, with paragraphs discussing different aspects, such as technical details, potential use cases or background knowledge, interlaced throughout the text.

We have endeavored to work on the language as well as the structure. This is reflected in the revised manuscript.

Overall, the relevance of the application presented could merit publication in this journal, but this can only be established from a much reworked manuscript that addresses the issues listed above. In particular, the basic characteristics and potential uses of the application should be clearly stated and easy to find. For example, the fact that it processes article abstracts and titles and not full text is only mentioned in passing.

We have revised the manuscript in response to this reviewer's comments.

# Major Compulsory Revisions
1-Reorganisation of the ideas presented in the article to better fit the sections.
3-Include some discussion on the application use, results, and applicability, both in quantitative and qualitative terms, that can help form an idea on what can be expected.

We have presented two user-case scenarios that show that not only can GenDrux instantly provide information related to genes, drugs (and both), but can also be used by researchers and knowledge seekers to make additional inferences. While we do not make those inferences ourselves, we show how an expert in the field might gain knowledge that is not immediately obvious in the results.

# Minor Essential Revisions
4-The diagram showing the system architecture is confusing, and it seems that not all the arrows have a clear interpretation. It also seems like there is one

This has been modified.

# Discretionary Revisions
5-The intention of the introductory text should be more obvious. If the main objective is to establish the relevance of the problem, or the appropriateness of the application, the arguments should be more explicitly laid out and sourced.

This has been modified.

6-The availability section states "The source code for GenDrux is available for download" but the application site does not seem to include the link. We are not making the source code available for download. That sentence has been removed. But the email contacts provided in the web-page for GenDrux will encourage users to contact us for help with using the program. We will also consider sharing the source code with potential collaborators.

**Level of interest:** An article of limited interest

**Quality of written English:** Not suitable for publication unless extensively edited

**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
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.

The revised manuscript now cites resources (including PharmGKB) that provide a service similar to GenDrux. We have also described how GenDrux is different from these resources.

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.
This has been explained.

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.
We have changed the text to reflect that entities with subscriptions to journals can avail of the full text. Our automated archiving system processes abstracts from Medline.

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.
Please see first paragraph where we show that based on reviewer comments, we have changed the basic thrust of this paper. To present negated relationships or hypothetical ones would cause us to use artificial intelligence methods. This is beyond the scope of this work. And since this tool when completely optimized will help physicians, we will leave it up to the end users to decide on possible impacts.

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

6 E-search returns batches of 500 hits. Authors describe a common methods to This part is not very informative.
This has been clarified.

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

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?
Our experience with E-utils has shown us that it is not the same as Lingpipe (which is cited in the revised manuscript). E-utils is created by NLM to allow users to automatically download abstracts in batches. While E-utils also provides some algorithmic support, we prefer to write our own “wrappers” because we have greater flexibility in pre-processing the abstract-text before archiving it.
Lingpipe is a resource of text-mining algorithms that can be used for different instances of text. Name-entity-recognition (presumably unsupervised) is one of those resources that is relevant to our work. But Lingpipe does not provide resources and methods to download abstracts from PubMed.
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? 

BANNER (cited in this paper) is an unsupervised name-entity-recognition program. Our system is supervised. We use exact string matching against validated resources of genes and pharmaceutical products. Therefore, while in terms of recognition, a system like BANNER is subject to issues of precision, our system is not. On the other hand, our system carries the disadvantage of recall in not extracting information that might not be listed in our resources. The authors decided that since the eventual aim was to help physicians, we had to go with validated resources—because our system’s eventual (not currently realized) aim is personalized medicine. Besides, since we will constantly update our archive, we are likely to have access to latest gene and drug information without resorting to the uncertainties of an unsupervised system.

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? We have clarified in the revised manuscripts that the pharmaceutical product names have been extracted from several resources (in the interest of being comprehensive), the FDA being one of them. Gene names on the other hand have been extracted from NCBI—Genbank.

We have taken the users recommendation and mapped the drug name to the page in the commercial web-site drugs.com and mapped the gene name to the Gene Ontology resource.

11 Can authors give the time needed to populate the knowledge base and the
one required for the system to answer an usual query?
The time required to populate the knowledgebase is a few hours. But this is run on a
weekend, when the NLM systems are less used. As for a query through the web-page,
the results are instantaneous. We have tested this using several combinations of
queries. This is reflected in the manuscript.

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?
Archiving is done in the background and is not likely to affect the users’ query time. We
also run our archiving automated system on weekends. This archive file is replaced by
the new one only when the archiving process is completed. If the archive doubles in
size, as an extreme example, it is conceivable that retrieval might take time, but it will not
be a discernible delay—as it currently stands, the results are instantaneous, irrespective
of how many results are returned.

13 Precision, recall measures would be helpful to judge on the quality of the
extracted information.
Because ours is a supervised system and we extract using exact string matches, if a
drug or gene name is in the title or abstract it is returned. We have verified this manually
taking several subsets of our archived system. A supervised system does not lend well
to precision and recall issues, especially from exact string (and synonym) matches.
Constantly updating our archive and the gene and drug name resources will ensure that
issues of recall will not seriously hamper retrieval results.
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.
All the resources that we used are publicly available. GenDrux is a query system that
helps users for arrive at focused results, without having to scroll to several thousand
articles to find a gene and/or drug name.
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.

Given the change in the thrust of this article, we have removed this section.
-Minor Essential Revisions

17 GenDrux and GenDRux are used alternatively by the authors. Please homogenize.
We have made this change.

18 "five to 10 years"=> five to ten
We have removed this sentence.

19 "This is considered the first step "to bring molecular-based medicine into current practice." " It is unclear where this citation is from.
We have added the reference or removed a sentence that is ambiguous.

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.
This sentence has been removed. The intent of the authors was to present correlations with respect to the impact of drug (sensitivity) and/or correlations among gene expressions (which I believe what the reviewer is referring to—and thank you for pointing this out, as it caused us to reassess the preciseness of the language)

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.
This has been corrected. This is also clarified at the end of the text in a list of abbreviations.

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.
We have removed any references to impact factors in the manuscript and also removed it from the GenDrux web resources. Our original intent was to allow users to judge the importance of a specific relationship based on impact factors, which may be controversial in its value in assessing importance of an article. We realize that there were several lists of impact factors and felt that it would lead to more
24 Title and running title are different. - Discretionary Revisions

The Running title is an abbreviated form of the Title and we have changed it to reflect this.

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?

We have removed this section as it was confusing.

26 Figure 4 is useless.

We have removed this figure.

27 Ref 7 is to format.

This has been reformatted.

28 Figure 1 says "Web interface (add later)"

Is it about the interface shown in Figure 2?

We have made this change.

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.
Reviewer's report:
The authors present GeneDruX, an online application that helps researchers find publications that discuss relationships between genes and drugs. Searches can be conducted starting with either a gene, a drug, or a disease. Each search leads to a table with titles for quick reference, and links to the PubMed citation of each article. I don't quite understand the ranking (sorting of results); it looks like a mixture of gene name (ascending) and impact (descending), at least for some of my queries; this should be explained in much more detail, as it largely influences the order in which a user finds -seemingly-relevant information. eUtils return results in descending order of publication date, and you’re method for re-ordering could help researchers significantly to find interesting papers. You could also provide an ordering by PubMed ID (=essentially by date), gene name, drug name, disease name, impact? The header of each column could be a clickable entry.

There is no hierarchical meaning to how the results are presented. They are presented in the same order as the information is archived, which is how the information was automatically downloaded from PubMed. The authors have tested this and agree that there is no way to meaningfully sort the web results for GenDruX. The notion of what constitutes an interesting paper is subjective and we leave it up to the user. Our attempts at categorizing based on impact factors also leads to confusion. This is addressed later in response to this reviewer's comment.

I tried a few searches and got results as expected; however, "Parkinson" (or spelling variations) yields no results.

This system is a pilot test system and the archive contains breast cancer related abstracts. While the user gets results by searching on other cancers (leukemia, etc.), it is likely that nothing related to Parkinson’s is archived. One of our user-cases depicts possible correlations between some cancers and not others.
A search for BRCA1/BRCA2, which you use an example in Background (breast cancer) and Table 1, also returns an empty result set ("No abstracts matched your specific requirements"). Probably that is due to the pre-selection of articles, but needs explanation in the paper.

The system allows searching only one gene at a time. We will, in due course modify the algorithm to perform multiple searches. The reviewer is right: the articles are pre-selected using an automated archiving system (see section on E-utils). In addition, we modified the system to search genes such as BRCA1 in a case insensitive way. The mentioned BRCA1, or brca1, now return more than 20 articles.

It appears that the selection of current content (8000 abstracts, from the last ten years) limits the user queries/results drastically, to the extend that for many queries, no results are returned.

The 8000 abstracts result from a search ‘breast+cancer+gene+drug) from 1975 to 2010. This was to focus this pilot system on breast cancer. We have briefly described how the automated procedures can allow extensibility in GenDrux to other disease.

What did not become clear is how you find a relationship between a gene and a drug, to select and present articles that contain at least one. Do you report any article that contains at least one gene and at least one drug? Do they have to occur in the title, or together in the same sentence? Are more sophisticated techniques used (see tools such as PolySearch, etc.)?

GenDrux as an information retrieval system, which will retrieve an article based on a query for one gene and one drug if references to those occur in the title and abstract (which is freely available to all users). We have not made any attempt to identify specific relationships, using natural language processing or artificial intelligence methods. GenDrux allows a filtering of information such that a user does not have to look at several thousand articles. We leave the identification of possible specific impacts of drugs on genes, or vice versa to the GenDrux user.

Where did you obtain the impact factors? If they come from Thompson-Reuters
ISI, from my own experience they will contact you shortly and ask to not publish them anymore. But as you point out, journal factors might not give too much of an indication about an individual article's impact. PLoS and other publishers started computing per-article scores, perhaps that could be a useful add-on

We agree about the confusion about impact factors. Our original intent was to allow users to judge the importance of a specific relationship based on impact factors although it is not a perfect solution. We realize that there were several lists of varying impact factors and felt that it would lead to more confusion. To remedy this, we have removed all references to impact factors from the manuscript as well as from the program.
Check spelling of the application's name in the paper vs website. GenDrux, GeneDRxUG?

We have made changes to be consistent.

**Level of interest:** An article whose findings are important to those with closely related research interests  
**Quality of written English:** Acceptable  
**Statistical review:** No, the manuscript does not need to be seen by a statistician.