autoHGPEC: Automated prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network [version 1; peer review: 2 approved with reservations]

Duc-Hau Le, Trang T.H. Tran
School of Computer Science and Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam

Abstract
Identification of novel disease-gene and disease-disease associations is an important task in biomedical research. Recently, we have developed a Cytoscape app, namely HGPEC, using a state-of-the-art network-based method for such task. This paper describes an upgrading version of HGPEC, namely autoHGPEC, with added automation features. By adding these functions, autoHGPEC can be used as a component of other complex analysis pipelines as well as make use of other data resources. We demonstrated the use of autoHGPEC by predicting novel breast cancer-associated genes and diseases. Further investigation by visualizing and collecting evidences for associations between top 20 ranked genes/diseases and breast cancer has shown the ability of autoHGPEC.

Keywords
Cytoscape app, Automation features, CyREST, R, Disease-gene association, Disease-disease association, Random walk with restart algorithm, Heterogeneous network, Gene prioritization, Disease prioritization

This article is included in the Cytoscape Apps gateway.
Corresponding author: Duc-Hau Le (duchaule@tlu.edu.vn)

Author roles: Le DH: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Tran TTH: Software, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2018 Le DH and Tran TTH. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Le DH and Tran TTH. autoHGPEC: Automated prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network [version 1; peer review: 2 approved with reservations] F1000Research 2018, 7:658 (https://doi.org/10.12688/f1000research.14810.1)

Introduction
One of the challenging tasks in biomedicine is to prioritize candidate genes and diseases by the degree of their relevance to a disease of interest. This is the starting point to identify novel disease-gene and disease-disease associations. A large number of computational methods including network- and machine learning-based ones have been proposed for such a task. State-of-the-art network-based methods often integrate diseases and genes together to form a heterogeneous network, then a propagation algorithm is applied to exploit the similarity between diseases/genes and known disease-gene associations to predict novel associations. Some tools have been also developed to facilitate the use of the state-of-the-art methods. However, most of them only focus on predicting novel disease-gene associations, including some tools which were developed as apps of Cytoscape. Recently, we have developed a Cytoscape app, HGPEC, to predict both disease-gene and disease-disease associations based on a state-of-the-art method on a heterogeneous network of diseases and genes. HGPEC was shown to be better than two other network-based Cytoscape apps for prediction of novel disease-gene associations, GPEC and PRINCIPILE in terms of prediction performance. In addition, HGPEC can prioritize candidate genes of diseases without known molecular basis and collect evidence to support novel predictions from various data resources such asGene Ontology, Disease Ontology, KEGG pathway, GeneRIF, PubMed, protein complexes and OMIM. Being developed as an app of Cytoscape, HGPEC can exploit advanced features of Cytoscape such as data visualization and integration. However, Cytoscape is a desktop-based tool, thus HGPEC cannot link to other analysis tools such as R and Python flexibly. Therefore, this also limits the use of HGPEC because it cannot be used automatically as a component of a complex analysis pipeline in these tools. In addition, this prevents Cytoscape from integrating data from other data resources. Recently, automation features have been added to Cytoscape to facilitate those tasks.

In this study, we upgrade HGPEC by adding automation features into it and name the new app as autoHGPEC. Basically, autoHGPEC has the same functions as HGPEC. However, these functions can be called by both CyREST functions and commands, thus can be called from external environments. To use autoHGPEC, a heterogeneous network of diseases and genes composing of a disease similarity network, a gene/protein network and known disease-gene associations has to be given. Then, a disease of interest must be selected from the disease similarity network. After that, the disease and its known associated genes (if any) are used as training/seed data. A set of candidate genes then has to be defined by selecting from the gene network or chromosome. These candidate genes and all remaining diseases are then ranked by a RWRH-based method (see the Methods section). Finally, users can select top ranked genes/diseases for further analyses such as visualization and evidence collection. We show the ability of autoHGPEC in predicting novel genes and diseases associated with breast cancer.

Methods
RWRH-based method
autoHGPEC was implemented using a ranking algorithm, random walk with restart on a heterogeneous network (RWRH).

Briefly, this network-based algorithm propagates the disease information embedded in a disease of interest and its known associated genes (also known as seed/training nodes) to other diseases and genes in the heterogeneous network. This propagation is performed by random walking from the seed nodes. At each node, the random walker goes to adjacency nodes or goes back to the seed nodes with a prior probability. This process is repeated iteratively until a steady-state is reached. A score assigned to each node at this state represents the degree of relevance to the seed nodes, thus relevance to the disease of interest. Finally, candidate genes and diseases are ranked by the scores and top ranked candidates can be selected as promising genes and diseases for further investigation.

Implementation
autoHGPEC is an upgrading version of HGPEC with added automation features. Therefore, main functions such as prioritization, visualization and evidence collection of HGPEC were kept. In addition, as in HGPEC, a number of databases were pre-installed in autoHGPEC to facilitate the use of this app. These include disease similarity networks, gene/protein networks and known disease-gene associations as well as annotation data such as Gene Ontology, Disease Ontology, KEGG pathways, GeneRIF, and protein complexes. However, users can also select other networks by themselves. In order to provide automation features for HGPEC, we first refactor source code of HGPEC to implement Cytoscape Tunable annotations to replace control panels of HGPEC in the west by a menu system. Therefore, all the functions of HGPEC are accessed through the menu system. In addition, the workflow of HGPEC is exposed to the users by using CyREST Command API (which can be followed in Swagger UI under the menu autoHGPEC). The CyREST API is developed with appropriated functions as well. Thus, the result of each step in the workflow can be passed on to the caller for further analysis in R or Python in JSON format.

Operation
autoHGPEC is designed to predict novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network with added automation features. Therefore, it operates in the same workflow as in HGPEC. However, in addition to desktop-based Cytoscape though the menu system, its functions can be called using CyREST Command API and from other analysis tools such as R. Figure 1 show the workflow of autoHGPEC in three running environments (see user manual in Supplementary File 1). As an app of Cytoscape with automation features, autoHGPEC can be run on any computer which satisfies the minimal requirements to run Cytoscape.

Use cases
To demonstrate functions of autoHGPEC with automation features, we showed its ability in predicting novel genes and diseases associated with breast cancer (OMIM ID: 114480). Here, we briefly describe this case study by following the 5-step workflow in Figure 1 (see user manual in Supplementary File 1 for more detail):

- First, a heterogeneous network of genes and diseases was constructed by connecting a preinstalled disease similarity.
<table>
<thead>
<tr>
<th>Step</th>
<th>Environment</th>
<th>Functions/Operations/Commands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Construct a heterogeneous network</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Desktop-based Cytoscape</td>
<td>Step 1: Construct a Heterogeneous Network</td>
</tr>
<tr>
<td></td>
<td>CyREST Command API</td>
<td>POST /v1.commands/autoHGPEC/step1_construct_network Step 1: Construct a Heterogeneous Network</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>&gt;commandRun('autoHGPEC step1_construct_network DiseaseGene=&quot;&quot; diseaseNetwork=&quot;&quot; geneNetwork=&quot;&quot;')</td>
</tr>
<tr>
<td></td>
<td>Select a disease of interest</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Desktop-based Cytoscape</td>
<td>Step 2: Select a disease of interest</td>
</tr>
<tr>
<td></td>
<td>CyREST Command API</td>
<td>POST /v1.commands/autoHGPEC/step2_1_select_disease Step 2.1: Select disease</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>&gt;commandRun('autoHGPEC step2_1_select_disease diseaseName=&quot;&quot;') &gt;commandRun('autoHGPEC step2_2_create_training_list diseaseTraining=&quot;&quot;')</td>
</tr>
<tr>
<td></td>
<td>Select candidate sets</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Desktop-based Cytoscape</td>
<td>Step 3: Provide Candidate Gene Set</td>
</tr>
<tr>
<td></td>
<td>CyREST Command API</td>
<td>POST /v1.commands/autoHGPEC/step3_PCG_allRemaining Step 3: Select candidate sets - All remaining genes in Gene Network</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>&gt;commandRun('autoHGPEC step3_PCG_allRemaining') &gt;commandRun('autoHGPEC step3_PCG_NBChromosome') &gt;commandRun('autoHGPEC step3_PCG_NBNetwork') &gt;commandRun('autoHGPEC step3_PCG_suscepChromo')</td>
</tr>
<tr>
<td></td>
<td>Prioritize</td>
<td>Step 4: Prioritize candidate genes and diseases</td>
</tr>
<tr>
<td>4</td>
<td>Desktop-based Cytoscape</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CyREST Command API</td>
<td>POST /v1.commands/autoHGPEC/step4_prioritize Step 4: Prioritize candidate genes and diseases</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>&gt;commandRun('autoHGPEC step4_prioritize backProb=0.5 jumpProb=0.6 subnetWeight=0.7')</td>
</tr>
<tr>
<td></td>
<td>Examine ranked genes and diseases</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Desktop-based Cytoscape</td>
<td>Step 5: Examine ranked Genes and Diseases</td>
</tr>
<tr>
<td></td>
<td>CyREST Command API</td>
<td>POST /v1.commands/autoHGPEC/step5_1_search_evidences Step 5.1: Evidence Search</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>&gt;commandRun('autoHGPEC step5_1_search_evidences') &gt;commandRun('autoHGPEC step5_2_visualize')</td>
</tr>
</tbody>
</table>

Figure 1. Workflow of autoHGPEC in three environments (i.e., Menu-based Cytoscape, CyREST Command API and R).

network (i.e., Disease_Similarity_Network_5) including 5,080 diseases and 19,729 interactions, a preinstalled human protein interaction network (i.e., Default_Human_PPI_Network) including 10,486 genes and 50,791 interactions, and known disease-gene associations collected from OMIM\textsuperscript{\textregistered}. This step can be accomplished by following commands from within R:

\texttt{>commandRun('autoHGPEC step1_construct_network DiseaseGene="Disease-gene from..."')}
Second, breast cancer (OMIM ID: 114480) was selected for investigation. This disease is known to be associated with 21 genes, which are also available in the human protein interaction network. Then, the training set was built with these genes and the disease of interest. We can run two following commands within R for this task:

```R
> commandRun('autoHGPEC step2_1_select_disease diseaseName="breast cancer"')
> commandRun('autoHGPEC step2_2_create_training_list diseaseTraining="MIM114480"')
```

Third, we selected all of 10,465 remaining genes in the protein interaction network as candidate genes. This option can be done by following command:

```R
> commandRun('autoHGPEC step3_PCG_allRemaining')
```

Fourth, all genes and diseases in the heterogeneous network are ranked by applying the RWRH-based method with back-probability, jumping probability and subnetwork importance weight were set to 0.5, 0.6 and 0.7, respectively. The following command can be used to accomplish this task:

```R
> commandRun('autoHGPEC step4_prioritize backProb=0.5 jumpProb=0.6 subnetWeight=0.7')
```

Finally, we visualized and collected evidence for the associations between 20 highly ranked candidate genes/diseases and breast cancer. The users must highlight the diseases and genes of their interest in the corresponding network. These tasks can be performed using two following commands, respectively:

```R
> commandRun('autoHGPEC step5_2_visualize')
> commandRun('autoHGPEC step5_1_search_evidences')
```

Visualization results (Figure 2a and b) show that most of the top ranked candidate genes are directly connected to known breast cancer-associated genes. In addition, highly ranked candidate diseases are directly connected to either known/training genes or the disease of interest. For evidence collection, we annotated and searched evidence for promising associations between the top ranked candidate genes/disease and breast cancer. Evidence collection results showed that each of the promising associations is supported by at least two data sources. More detail about interpretation on the results of visualization and evidence collection for these associations can be found in the HGPEC study.12

Therefore, users can easily call this CyREST API and use this result in their workflow as they need.

**Discussion and conclusions**

Random walk with restart algorithm on heterogeneous network of diseases and genes was shown as a state-of-the-art method for predicting novel disease-gene and disease-disease associations compared to other network-based algorithms.3,12 However, its prediction performance highly depends on the used heterogeneous network, which is a combination of a disease similarity network and a gene/protein interaction network and known disease-gene associations. Indeed, a study showed that the prediction performance can be improved by using a gene ontology-based gene similarity network instead of using the
Figure 2. Visualization of highly ranked candidate genes and diseases in topological relationships with breast cancer. (a) Topological relationships between highly ranked candidate genes and known breast cancer-associated genes. (b) Topological relationships between highly ranked candidate diseases and breast cancer and its known associated genes. Note that: For diseases, nodes in rhombus and rectangle shapes are breast cancer and candidate diseases, respectively. For genes, nodes in triangle and octagon shapes are known breast cancer-associated genes and candidate genes, respectively. Nodes with high rankings are in red, relative high are in pink, medium are in white and light green, low are in green.

human protein interaction network\(^22\). In addition, we have recently shown that using the disease similarity network constructed by Human Phenotype Ontology\(^23\) improved the prediction performance of disease-associated genes\(^24\) as well as disease-associated non-coding RNAs\(^25,26\). Therefore, to facilitate the use of the similarity networks of diseases/genes, we enable user to provide these networks by themselves. For gene/protein network, user can import the network from various molecular interaction data sources or from other analysis pipelines. Similarly, disease similarity networks can be inputted from other analysis tools such as DOSim\(^27\) and HPOSim\(^28\). Moreover, the ranked candidate genes can be used as inputs of other annotation and enrichment toolkits to support more about their associations with the disease of interest such as DAVID\(^29\) and GSEA\(^30\). Taken together, with added automation features, autoHGPEC can be more useful and reached by a wider range of users.

**Summary**

Identification of novel disease-gene and disease-disease associations is an important task in biomedical research. Recently, we have developed a Cytoscape app, namely HGPEC, using a state-of-the-art network-based method for such task. This paper describes an upgrading version of HGPEC, namely autoHGPEC, with added automation features. By adding these functions, autoHGPEC can be used as a component of other complex analysis pipelines as well as make use of other data resources. We demonstrated the use of autoHGPEC by predicting novel breast cancer-associated genes and diseases. Further investigation by visualizing and collecting evidences for associations between top 20 ranked genes/diseases and breast cancer has shown the ability of autoHGPEC.

**Software and data availability**

1. autoHGPEC on Cytoscape Apps: http://apps.cytoscape.org/apps/autohgpec
2. User manual can be downloaded at https://sites.google.com/site/duchaule2011/bioinformatics-tools/autohgpec
4. Source code: https://github.com/trangtran86/autoHGPEC
5. Archived source code as at time of publication: http://doi.org/10.5281/zenodo.1228521
6. License: MIT

All prerequisite data are already included in the apps. Refer to the user manual (Supplementary File 1) for other additional annotation data such as Gene Ontology.

Supplementary material
Supplementary File 1: autoHGPEC user manual.

Click here to access the data.

References


Competing interests
No competing interests were disclosed.

Grant information
The author(s) declared that no grants were involved in supporting this work.


Open Peer Review

Current Peer Review Status:  ?  ?

Version 1

Reviewer Report 28 August 2018

https://doi.org/10.5256/f1000research.16119.r36046

© 2018 Le Van T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thanh Le Van
Janssen Pharmaceutica NV, Beerse, Belgium

Summary:

The paper "autoHGPEC: Automated prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network" presents an enhanced implementation of HGPEC, the previous work of one of the co-authors of the paper. The new implementation allows users to integrate the data analysis steps in Cytoscape with other data analysis pipelines in R or CyREST API. Indeed, this new feature would be useful as users now can take the advantages of the network-based data analysis and visualization in Cytoscape as well as the power of statistical data analysis of R, for example.

Below are my detail comments:

1. Is the rationale for developing the new software tool clearly explained?

In my opinion, the "automatic features" is not well explained in the paper. The first place where the authors introduce the concept of "automatic features" is the last sentence of the first paragraph in the Introduction section. However, there is no further explain of this concept. Hence, it is very easy for people in the machine learning community to be confused with the concept of automatic feature selection in the automated machine learning field.

To clear the possible confusion, we can do two things: 1) add a citation of the paper/website where Cytoscape orginally introduce this concept; 2) briefly explain how Cytoscape provides this type of feature and how HGPEC can leaverage the facilities provided by Cytoscape.

2. Is the description of the software tool technically sound?

Yes

3. Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
The user manual is quite detailed. However, there are rooms for improvements of the presentation, for example, the space between pictures and paragraphs, and the indent of paragraph are not always consistent and pleasant to read. I highly recommend to use latex to produce the manual.

There is a Vietnamese sentence on page 10 of the manual, which should be removed.

4. Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

- Please add a citation when mentioning that breast cancer is known to be associated with 12 genes (first paragraph, page 5)
- Please briefly explain why the results of the demonstration make sense

5. Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Overall, the author should have further explain: what is automatic features and why it is worthy of investigation.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
This paper aims to address the challenging task of trying to prioritize candidate genes and diseases based on their degree of relevance within a known heterogeneous network of disease-gene and disease-disease associations. A previous iteration of this Cytoscape app called HGPEC employs a network learning-based method only for disease-gene associations however now it has been upgraded to autoHGPEC with added automation features, the flexibility to be integrated in more complex analysis pipelines and the ability to take input other data resources. Another addition is that autoHGPEC predicts not only disease-gene associations but disease-disease associations as well. This paper details the application of autoHGPEC in predicting novel breast cancer-associated genes and diseases with increased automation and flexibility.

I commend the authors in the presentation of their work. The paper was logical, concise and easy to read. Below are my comments. I hope it will benefit them well.

Major comments:
1. It is mentioned that the app is highly dependent on a used heterogenous network, gene/protein interaction networks and known disease-gene associations. Firstly, clarify how these networks are provided and what exactly they are. Secondly, perhaps using gene ontology-based gene similarity networks could be better. This is crucial, as there could be post translational modification effects to be explored. However, with this said, the paper claims HGPEC can prioritize candidate genes without knowing its molecular basis. Reconsider the reasoning behind this claim.

   How is the input and output data reliable if we don’t know the molecular basis? Perhaps this is a dangerous statement, as there’s now lots of computational tools providing evidence to the pertinence of post translational modification effects such as methylation, acetylation, etc. Be careful how this is stated. It has current and will have future ramifications. At least provide an explanation as to why knowledge of this isn’t needed.

2. If autoHGPEC performs better than GPEC and PRINCIPLE, present data on it. Where’s the comparison between the three and what is the criteria for ranking them?

3. Has a comparison of results for when a set of candidate genes are selected from a gene network or chromosome been done? This is a flexibility-based component of the app however testing these cases could give evidence for which is more reliable and in what circumstances.

Minor comments
1. “Recently, we have developed a Cytoscape app, HGPEC, to predict both disease-gene and disease-disease associations based on a state-of-the-art method on a heterogeneous network of diseases and genes.” Be specific. I understand what the paper is trying to say here, with “state of
the art” meaning either network or machine learning based approaches however it took a few times to get a clear understanding. Mention how pertinent these approaches are but afterwards, mention which state of the art approach is being used. It will get the point across quickly without introducing any confusion.

2. Technical and grammatical writing errors:

- “…we first refactor source code of HGPEC to implement Cytoscape Tunable annotations to replace control panels of HGPEC in the west” –
- “Beside the fact is that almost commands of autoHGPEC return results in JSON format…”
- “autoHGPEC is an upgradeD version of HGPEC…”
- “For gene/protein networkS, user can import the network from various molecular interaction”

There are more throughout the paper and especially the supplementary document. Read through and correct carefully. Pay attention to your font formatting and spacing to keep things consistent.

3. Change the node colors on page 8 of the supplementary document. It is a little confusing to see red as top ranked, green as bottom ranked and white and light-green as middle ranked. Maybe provide a legend.


Overall, good job. Two crucial benefits to the new app is that it can take data in from multiple sources as opposed to only one source – possibly lowering the chance of error and bias, and that it lets you integrate it with R and Python, allowing for integration as a component of more complex analysis pipelines. However, it seems this app can only be used if known disease-gene/protein associations and disease similarity networks are given. Why doesn’t the paper mention protein interactions any further? Specifically, what pertinent information is being taken away from these said protein interactions?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.
I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com