A guide through the labyrinth of gene prioritization tools

Léon-Charles Tranchevent[1*], Francisco Bonachela Capdevila[2*], Daniela Nitsch[1*], Bart de Moor[1], Patrick De Causmaecker[2], Yves Moreau[1§]

[1] Department of electrical engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium
[2] Department of Informatics, Katholieke Universiteit Leuven, Campus Kortrijk, Belgium

[*] These authors contributed equally to this work
[§] Corresponding authors

Emails:

Léon-Charles Tranchevent Leon-Charles.Tranchevent@esat.kuleuven.be
Francisco Bonachela Capdevila Francisco.BonachelaCapdevila@kuleuven-kortrijk.be
Daniela Nitsch Daniela.Nitsch@esat.kuleuven.be
Bart De Moor Bart.DeMoor@esat.kuleuven.be
Patrick De Causmaecker Patrick.DeCausmaecker@kuleuven-kortrijk.be
Yves Moreau Yves.Moreau@esat.kuleuven.be

Introduction

One of the major challenge in human genetics is to find the genetic cause underlying disorders in order to unravel the molecular basis of these diseases and eventually elaborate medical treatments. In the past decades, the use of high-throughput technologies such as linkage analysis and association studies have permitted major discoveries in that field (Redon et al.[17122850], Marazita et al.[15185170]). These technologies are usually able to associate a chromosomal region with a genetic condition. One can also use expression arrays and obtain a list of transcripts differentially expressed in a disease sample with respect to a reference sample. A common characteristic of these methods is usually the large number of genes returned: for instance hundreds of differentially expressed genes are often reported (ref needed). The working hypothesis is often that only one or a few candidates are really of primary interest. Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task. Typically, a biologist would have to manually go through the list of candidates, check what is currently known about each gene and assess whether it is a promising candidate or not. The bioinformatics community has therefore introduced the concept of gene prioritization that takes advantage of both the progress made in computational biology and the large amount of genomic data available. It was first introduced in 2002 by Perez-Iratxeta et al. who already described the first approach to tackle this problem. Since then, many different strategies have been developed, some of which have been implemented into web applications and eventually experimentally validated. A similarity between the different strategies is their use of the so called guilt-by-association concept. That is the most promising candidates will be the ones that are similar to the genes already known to be linked to the biological process of interest. For example, when studying type 2 diabetes (T2D), KCNJ5 appears as a good candidate through its potassium channel activity (Iizuka et al.), an important pathway for diabetes ([11868613]) and because it is known to interact with ADRB2 (Lavine et al.), a key player in diabetes and obesity. This notion of similarity is not restricted to pathway or interaction data and can be extended to any kind of genomic data. Lately, important efforts have been made to experimentally validate these approaches. For instance, in 2006, two independent studies have used multiple tools in conjunction to propose new meaningful candidates for type 2 diabetes and obesity (Tiffin et al., Elbers et al.). More recently, Aerts et al. have developed a computationally supported genetic screen whose computational part is based on gene prioritization (ref). With this review, we aim at describing the current options for a biologist who needs to select the most promising genes from large gene lists. We have first selected strategies for which a web application was available, and we have described how they differ from each other and how they were successfully applied to real biological questions. Secondly, since it is very likely that novel methods will be proposed in the near future, we have also developed a website termed ‘XXX’ that represents an up-to-date electronic review in that field.

Results

With this study, we review 18 gene prioritization tools that fulfil the two following criteria. First, the strategy should have been developed for human candidate disease gene prioritization. Notice that predicting the function of a gene or its implication in a genetic condition are two closely related problems. Moreover, several gene function prediction methods have indeed been applied to disease candidate prioritization with reasonable performance (ref GFSST). However, it has been shown that gene prioritization is more challenging than gene function prediction since diseases often implicate a complex set of cascades covering different pathways and functions ([16869964]). Besides, to our knowledge, none of the existing gene function prediction methods includes disease specific data. Thus, these methods were excluded from the present study. For gene function prediction methods, readers are referred to the reviews by Troyanskaya et al. (ref) and Punta et al. (ref). Our second criterion is that there should exist a living web application associated with the strategy presented. Since the end users of these tools are not expert in informatics, approaches only providing a set of scripts, or some code to download have been discarded. Furthermore, we focus our analysis on the non commercial solutions and require the web tools to be freely accessible. Using these criteria, we were able to retain a total of 18 applications that still differ by (i) the inputs they need from the user, (ii) the computational methods they implement, (iii) the data sources they use, and (iv) the output they present to the user. The thorough discussion of these characteristics has allowed us to create a decision tree (figure 3) that support users in their decision process.

In the following section, we summarise the gene prioritization tools that we retain for this review. Their references and the URL of their web applications are presented in Table T1. Several approaches combine different data sources. SUSPECT (Adie et al.) ranks candidate genes by matching sequence features, gene expression data, interpro domains and GO terms, and also CANDID (Hutz et al.) uses several heterogeneous data sources, some of them chosen to overcome bias. Endeavour (Aerts et al.), however, uses training genes known to be involved in a biological process of interest and ranks candidate genes by applying several models based on various genomic data sources. Among the tools using different data sources, ToppGene (Chen et al.), SNPs3D (Yue et al.), GeneDistiller (Seelow et al.) and Posmed (Yoshida et al.) include mouse data within their algorithms, whereas ToppGene combines mouse phenotype data with human gene annotations and literature, SNPs3D (Yue et al.) identifies genes that are candidates for being involved in a specified disease based on literature, in GeneDistiller (Seelow et al.) mouse phenotype can be used to filter genes, and Posmed (Yoshida et al.) can use mouse data as well by XXXX. On the other hand, G2D (Perez-Iratxeta et al.) uses three algorithms based on different prioritization strategies to prioritize genes on a chromosomal region according to their possible relation to an inherited disease using a combination of data mining on biomedical databases and gene sequence analysis. G2D allows users to inspect any region of the human genome to find candidate genes related to a genetic disease of their interest. TOM (Rossi et al.) efficiently employes functional and mapping data and selects relevant candidate genes in from a defined chromosomal region.

Tools that mainly base on literature and text mining are PolySearch, MimMiner, BITOLA, aGeneApart, and GenePropector. PolySearch (Cheng et al.) extracts and analyses relationships between diseases, genes, mutations, drugs, pathways, tissues, organs and metabolites in human by using multiple biomedical text databases. MimMiner (Driel et al.) analyses the human phenome by text mining in order to rank phenotypes by their similarity to a given disease phenotype, and BITOLA (Hristovsky et al.) mines MEDLINE database to discover new relations between biomedical concepts. aGeneApart (Van Vooren et al.) creates a set of chromosomal aberration maps that associate genes to biomedical concepts by an extensive text mining of MEDLINE abstracts, using a variety of controlled vocabularies. GeneProspector (Yu et al.) searches for evidence about human genes in relation to diseases, other phenotypes, and risk factors, and selects and prioritizes candidate genes by using a literature database of genetic association studies.
Finding associations between genes and phenotypes is focused by Gentrepid and PGMapper, whereas Gentrepid (George et al.) predicts candidate disease genes based on their association to known disease genes of a related phenotype, and PGMapper (Xiong et al.) matches phenotype to genes from a defined genome region or a group of given genes by combining the mapping information from the Ensembl database and gene function information from the OMIM and PubMed databases.
Tools, such as GeneWanderer, Prioritizer, Posmed and PhenoPred make use of genome wide networks. GeneWanderer (Köhler et al.) is based on protein-protein interaction and uses a global network distance measure to define similarity in protein-protein interaction networks. PhenoPred (Radivojac et al.) uses a supervised algorithm for detecting gene - disease associations based on the human protein-protein interaction network, known gene-disease associations, protein sequence, and protein functional information at the molecular level. Instead of using a human protein-protein interaction network, Posmed (Yoshida et al.) is based on an artificial neural network-like inferential process in which each mined document becomes a neuron (documentron) in the first layer of the network and candidate genes populate the rest of layers.

We have defined a data source as a type of data that defines a particular view of the genes (see 'Gene view') and thus can correspond to several databases. We have defined 12 data sources: Text mining (co-occurrence and functional mining), Protein-protein interactions, Functional annotations, Pathways, Expression, Sequence, Phenotype, Conservation, Regulation, Disease probabilities and Chemical components. Using these categories, we have built a data source landscape, that is, for each data source, collecting in which tools it is used (see Table 2). The tools also differ in the inputs they require and the outputs they provide. Two types of inputs have been distinguished: the prior knowledge about the genetic disorder of interest and the candidate search space. Wu furthermore consider two possibilities for the prior knowledge as it can be defined by a set of genes or by a set of keywords. Similarly, the candidate search space is either a locus, or a set of candidate genes, or the genome. For the outputs, two types were considered, a

ranking and a selection of the candidate genes. In addition, we register which tools further give information about the statistical significance of their results. Table T3 shows an overview of the input data that is required by the tools as well as the output they produce. Also, a clustering of the tools regarding to their inputs and outputs is illustrated in Figure F2.

Discussion

We have reviewed 18 gene prioritization tools and organised the collected information to help users to decide which tools suit best their needs. We have defined a data source as a type of data, possibly encompassing multiple related databases. We have stressed that data sources are at the core of the gene prioritization problem since both high coverage and high quality data sources are needed in order to make accurate predictions. We have built the data source landscape map and have observed that text mining is by far the most widely used data source since 14 of the 18 tools are using either co-occurrence and/or functional text mining. Most of the approaches make use of a wide range of data sources covering many distinct views of the genes (see ‘Gene views’), but 4 tools are exclusively relying on text mining data (PGMapper, Bitola, aGeneApart and GeneProspector), however their use of advanced text mining techniques still allow them to make novel predictions. At the other end of the spectrum, conservation, regulation, disease probabilities and chemical components are poorly used by only 2 tools at maximum although they describe unique features that might not always be captured by the other data sources. However, we have also stressed that the rule is not to include as many data sources as possible but rather to reach a critical mass of data beyond which accurate predictions can be made.

Our analysis reveals that the methods can be divided up according to the inputs they need, two possibilities are distinguished, a training set and a keyword set. The definition of a training set and a keyword set are problem specific meaning that no solution should be automatically preferred. The retrieval of a training set requires the knowledge of, at least, one disease causing gene, but preferably more than one. In addition, it should be homogeneous, meaning that it usually contains between 5 and 25 genes that, together, describe a precise biological process. When no disease gene can be found, members of the pathways disturbed by the diseases are also an option (ref). Alternatively, several tools accept text as input, text is either a disease name, selected from a list, or a set of user defined keywords that describe the disease under study. For the latter, the expert should define a complete set of keywords that covers most aspects of the disease (e.g., to obtain reliable results, 'diabetes' should be used in conjunction with 'insulin', 'islets', 'glucose' and others diabetes related keywords but not alone). A second important choice is the candidate search space definition, we have distinguished between QTL (set of neighbouring genes), eQTL (set of non neighbouring genes), and full genome. Although the two first options are similar, the distinction we made is important since several tools allow the definition of QTL but not of eQTL and vice-versa. Alternatively, 9 tools allow the exploration of the full genome, in case no candidate gene set can be defined. Last point, the output returned can either be a ranking, or a selection, or both. From the 18 tools, 4 perform a selection of the

candidates and 3 of these 4 first select candidates and then rank them. Of interest, a selection can be obtained from a ranking by using a threshold on the ranking. This extra-step is nevertheless not implemented in the tools we reviewed and should be done manually. In addition, we have created a decision tree to help users to choose the most suitable tools for their biological question. The tree is based on three basic questions that users should ask themselves before selecting one or several tools to use. By answering these questions, users define first, which genes are candidate, second, how the current knowledge is represented, and if necessary third, what is the desired output type.

Beside the data, the inputs and the outputs, what reveals to be critical for one tool to be used is its interface. Ideally, it has to be an intuitive interface that accepts simple input and provides detailed output. A past success and reference in bioinformatics is BLAST (ref) for which only a single sequence needs to be provided. In return, Blast provides the complete detailed alignments together with cross-links to sequence database so that the user can fully understand why the input sequence matches to a given database sequence. We, as a community, should develop tools that answer the end users' needs and that is probably corresponding to the simple input - detailed output paradigm described above.

Beside the benchmarks that are usually performed to estimate the real performance, several biological applications have been described in the literature. Table 5 gives an overview of these applications. Interestingly, three of them are analysing type 2 diabetes associated loci and are using several gene prioritization tools in conjunction (Teber et al., Tiffin et al., Elbers et al.). Elbers et al. analysed five loci previously reported to be linked with both type 2 diabetes and obesity that encompass more than 600 genes in total (ref). The authors used six gene prioritization tools in conjunction and reported 27 interesting candidates. Some of them were already known to be involved in either diabetes or obesity (e.g., TCF1 and HNF4A, responsible for maturity onset diabetes of the young, MODY) but some candidates were novel predictions. Among them, 5 genes were involved in immunity and defence (e.g., TLR2, FGB) and it is well known that low-grade inflammation in the visceral fat of obese individuals causes insulin resistance and subsequently T2D. Also, 10 candidate genes were so called 'thrifty genes' because of their involvement in metabolism, sloth and gluttony (e.g., AACS, PTGIS and the neuropeptide Y receptor family members). Using a similar strategy, Tiffin et al. prioritized type 2 diabetes and obesity associated loci and proposed another set of 164 promising candidates (ref). Of interest, 4 of the 27 candidates reported by Elbers et al. were also reported by Tiffin et al., (namely CPE, LAMA5, PPGB, and PTGIS). Although there is an overlap between the predictions, some important discrepancies remain and can be explained by the fact that the two studies do not focus on the same loci and do not use the same gene prioritization tools. This indicates that several gene prioritization tools can be applied in parallel to strengthen the results. Teber et al. compared the finding from recent genome-wide association studies (GWAS) to the predictions made by 8 gene prioritization methods (ref). Of the 11 genes associated with highly significant SNPs identified by the GWAS, eight were flagged as promising candidates by at least

one of the method. Another interesting validation is a computationally supported genetic screen performed by Aerts et al. in Drosophila melanogaster (ref). The aim of a genetic screen is to discover in vivo associations between genotypes and phenotypes. A forwards genetic screen is usually performed in two steps: in the first step, the loci associated to the phenotype under study are identified and in a second step, the genes from these loci are assayed individually. Aerts et al have introduced a computationally supported genetic screen in which the associated loci found in the first step are prioritized using Endeavour. Then, in the second screen, only the genes in the top 30% of every locus are assayed reducing then the cost of that step by 70%. Additionally, the authors have shown that 30% is a very conservative threshold since, in that case, all the positives they found were ranked in the top 15%. This shows that gene prioritization tools, when integrated into such workflows, can increase their efficiency for a decreased cost.

In this study, we review the methods that were developed for human disease candidate gene prioritization and that are directly accessible via a web application. We are however aware of other gene prioritization techniques that were excluded of the present analysis but that still represent an important contribution to the field. First, several gene prioritization methods, such as CAESAR (Gaulton et al.), GeneRank (Morrison et al.), and CGI (Ma et al.) propose interesting alternatives (e.g., natural language processing based disease model [ref Caesar]), however, they only provide a standalone application to install and run locally. We believe that a web application is essential since it does not require an extensive IT knowledge to be installed and used. Second, there are methods that were once pioneers in that field and for which web applications were provided in the past, but are not accessible any more (e.g., TrAPSS (Braun et al.), POCUS (Turner et al.), Prioritizer (Franke et al.)). Prioritizer recently moved from a living web application to a program to download and was therefore excluded prior to publication. Third, several studies also present case specific approaches tailored to answer a specific problem (refs). For instance, Lombard et al. have prioritized 10,000 candidates for the fetal alcohol syndrome (FAS) using a complex set of 29 filters (ref). Their analysis reveals interesting therapeutic targets like TGF-β, MAPK and members of the Hedgehog signalling pathways. Another example is the network-based classification of breast cancer metastasis developed by Chuang et al. (ref). These approaches are however case specific and can not be easily ported to another disease. Last, alternative techniques to circumvent recurrent problems in gene prioritization are currently developed. As an illustration, Nitsch et al. have proposed a data-driven method in which knowledge about the disease under study comes from an expression data set instead of a training set or a keyword set (ref). Altogether, these methods represent significant advances indicating that this field is still an emerging field. It is therefore most likely that novel methods will be developed in the future and that the existing ones will be improved. To overcome the limitations due to the static nature of this review, we have developed a website that is its online side, its aim is to represent an up-to-date electronic version of the present review. This website, termed 'XX', contains, for every tool, a detailed sheet that summarises the necessary information such as the inputs needed and the data sources used. It also builds tables that describe the general data source usage and the general input/output usage that are equivalent

to the tables 2 and 3 of the current publication. We think that this website should serve as a reference to guide users through the labyrinth of the gene prioritization tools

With the use of advanced high-throughput technologies, the amount of genomic data is growing exponentially and the quality of the gene prioritization methods is also increasing accordingly. However, several avenues need to be explored in the coming years to increase even further the potential of these tools. We already mentioned the interface, that is sometimes overlooked in the software development process. More at the data level, some efforts have already been made to use the huge amount of data available for species close to human (Chen et al., ref). Already, several tools described in the current review are including rodent data (e.g., ToppGene, PhenoPred, Posmed). However, the development of gene prioritization approaches combining in parallel many data sources from different organisms is still to come. Another important development is the inclusion of clinical and patient related data. Decipher (ref) already represents a first step in that direction since it includes aCGH data from patients and allow the text based prioritization of the genomic alterations, detected in the aCGH data, with respect to the phenotype of the patient. Efforts should also be made to include data sources that have been, so far, rarely included such as chemical components. Another important track is to explore different computational approaches to improve once more the algorithms that are at the core of the gene prioritization methods. Preliminary results show that the use of SVM based algorithm on kernel data can reduce the error in prediction (ref).

Conclusion

This review describes 18 human candidate gene prioritization methods that can offer services freely through a web interface. We have organised the tools according to their characteristics helping therefore users to decide which tools are best suited for their analysis. We have also developed a website, the electronic version of the review, aiming at guiding the biologists lost in the labyrinth of gene prioritization tool.

GLOSSARY

Gene prioritization:

The gene prioritization problem has been defined as the identification of the most promising candidate genes from a large list of candidates with respect to one biological process of interest.

Data sources:

Data sources are at the core of the gene prioritization problem since the quality of the predictions directly correlates with the quality of the data used to make these predictions. The different genomic data sources can be defined as different views on the same object, a gene. For instance, pathway databases like Reactome (ref) and Kegg (ref) define a 'bio-molecular view' of the genes while PPIs networks such as HPRD (ref) and MINT (ref) define their 'interactome views'. In

engineering drawing, a single view only contains limited information about the object while the combination of several views can define a much more precise picture. This reveals to be also true in genetics, a single data type might not be powerful enough to predict the disease causing genes accurately while the use of several complementary data sources allow much more accurate predictions (ref). The table 2 contains the list of the 12 data sources we have defined.

Inputs:

Two types of inputs can be distinguished: the prior knowledge about the genetic disorder of interest and the candidate search space. On the one hand, the prior knowledge represents what is currently known about the disease under study, it can be represented either as a set of genes known to play a role in the disease or as a set of keywords that describe the disease. On the other hand, the candidate search space defines which genes are candidates. For instance, a locus linked to a trait defines a quantitative trait locus (QTL), the candidates are therefore the genes lying in that region. Another possibility is an expression QTL (eQTL), it is a list of genes differentially expressed in a tissue of interest that are not necessary from the same chromosomal location. Alternatively, the full human genome can be used. An overview of the inputs needed by the applications can be found in table 3.

Outputs:

For the 18 selected applications, the output is either a ranking of the candidate genes, the most promising genes being ranked at the top, or a selection of the most promising candidates, meaning that not all genes are returned. Several tools are performing both at the same time (Gentrepid, Bitola, PosMed), that is first selecting the most promising candidates and then rank only these. Several tool benefit from an additional output, a statistical measure, often a p-value, that estimates how likely it is to obtain that ranking by chance alone. The statistical measure is often of crucial importance since there will always be a gene ranked in first position even if none of the candidate genes is really interesting. Notice then that a selection can be obtained from a ranking by using the statistical measure, e.g., simply by choosing a threshold above which all the genes are considered as promising. You can find an overview of the outputs produced by the different applications in table 3.

Text mining:

Text mining is the process of deriving formatted information, such as gene co-occurence, from raw text, such as publication abstracts.