Briefings in Bioinformatics Advance Access published March 21, 2010

A guide to web tools to prioritize candidate genes

Léon-Charles Tranchevent, Francisco Bonachela Capdevila*, Daniela Nitsch*, Bart De Moor, Patrick De Causmaecker and Yves Moreau*

Submitted: 8th January 2010; Received (in revised form): 8th February 2010

Abstract

Finding the most promising genes among large lists of candidate genes has been defined as the gene prioritization problem. It is a recurrent problem in genetics in which genetic conditions are reported to be associated with chromosomal regions. In the last decade, several different computational approaches have been developed to tackle this challenging task. In this study, we review 19 computational solutions for human gene prioritization that are freely accessible as web tools and illustrate their differences. We summarize the various biological problems to which they have been successfully applied. Ultimately, we describe several research directions that could increase the quality and applicability of the tools. In addition we developed a website (http://www.esat.kuleuven.be/gpp) containing detailed information about these and other tools, which is regularly updated. This review and the associated website constitute together a guide to help users select a gene prioritization strategy that suits best their needs.

Keywords: *gene prioritization; candidate gene; disease gene;* in silico *prediction; review*

BACKGROUND

One of the major challenges in human genetics is to find the genetic variants underlying genetic disorders for effective diagnostic testing and for unraveling the molecular basis of these diseases. In the past decades, the use of high-throughput technologies (such as linkage analysis and association studies) has permitted major discoveries in that field [1, 2]. These technologies can usually associate a chromosomal region with a genetic condition. Similarly, one can also use expression arrays to 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 size of the chromosomal regions returned, typically several megabases [3]. The working hypothesis is often that only one or a few genes are really of primary interest (i.e. causal). Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task. Typically, a biologist would have to go manually through the list of candidates, check what is currently known about

Corresponding author. Yves Moreau, Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium. Tel: +32 (0)16 32 8645; Fax: +32 (0)16 32 1970; E-mail: yves.moreau@esat.kuleuven.be

*These authors contributed equally to this work.

Léon-Charles Tranchevent is a PhD student at the Katholieke Universiteit Leuven. His main research topic is the development of computational solutions for the identification of disease causing genes through the fusion of multiple genomic data.

Francisco B. Capdevila, is a PhD student at the Katholieke Universiteit Leuven. His main research interest is the application of machine learning techniques, specially clustering, in gene prioritization.

Daniela Nitsch is a PhD student at the Katholieke Universiteit Leuven. Her research focus on the identification of disease causing genes through the exploration of gene and protein network based techniques.

Bart De Moor is a full Professor at the Department of Electrical Engineering of the Katholieke Universiteit Leuven. His research interests are in numerical linear algebra and optimization, system theory and system identification, quantum information theory, control theory, data-mining, information retrieval and bioinformatics.

Patrick De Causmaecker is an Associate Professor at the Department of Computer Science at the Katholieke Universiteit Leuven, Head of the CODeS Research Group on Combinatorial Optimisation and Decision Support.

Yves Moreau is a Professor at the Department of Electrical Engineering and a Principal Investigator of the *SymBioSys* Center for Computational Systems Biology of the Katholieke Universiteit Leuven. His two main research themes are the development of (i) statistical and information processing methods for the clinical diagnosis of constitutional genetic and (ii) data mining strategies for the identification of disease causing genes from multiple omics data.

each gene, and assess whether it is a promising candidate or not. The bioinformatics community has therefore introduced the concept of gene prioritization to take advantage of both the progress made in computational biology and the large amount of genomic data publicly available. It was first introduced in 2002 by Perez-Iratxeta *et al.* [4] who already described the first approach to tackle this problem. Since then, many different strategies have been developed [5–34], among which some have been implemented into web applications and eventually experimentally validated. A similarity between all strategies is their use of the 'guilt-by-association' concept: 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 [35–37]. For example, when studying type 2 diabetes (T2D), KCNJ5 appears as a good candidate through its potassium channel activity [38], an important pathway for diabetes [39], and because it is known to interact with ADRB2 [40], a key player in diabetes and obesity. This notion of similarity is not restricted to pathway or interaction data but rather can be extended to any kind of genomic data. Recently, initial efforts have been made to experimentally validate these approaches. For instance, in 2006, two independent studies used multiple tools in conjunction to propose new meaningful candidates for T2D and obesity [41, 42]. More recently, Aerts *et al.* [43] have developed a computationally supported genetic screen whose computational part is based on gene prioritization (Figure 1).

With this review, we aim at describing the current options for a biologist who needs to select the most promising genes from large candidate gene lists. We have selected strategies for which a web application was available, and we describe how they differ from each other and, when applicable, how they were successfully applied to real biological questions. In addition, since it is likely that novel methods will be proposed in the near future, we have also developed a website termed 'Gene Prioritization Portal' (available at: http://www.esat.kuleuven.be/gpp/) that represents an updatable electronic review of this field.

SELECTING THE GENE PRIORITIZATION TOOLS

In this study, we review 19 gene prioritization tools that fulfill 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 gene prioritization with reasonable performance [5]. 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 molecular pathways and functions [44]. 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.* [45] and Punta *et al.* [46]. Our second criterion is that a functional web application should be available for the proposed strategy. Since the end users of these tools are not expert in computer science, approaches only providing a set of scripts, or some code to download have been discarded. Furthermore, we focus our analysis on the noncommercial solutions and thus require the web tools to be freely accessible for academia. Using these criteria, we were able to retain a total of 19 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 2) that supports users in their decision process.

In the following section, we summarize the gene prioritization tools that we have retained. The corresponding references and the URL of their web applications are presented in Table 1. Several approaches combine different data sources. SUSPECT ranks candidate genes by matching sequence features, gene expression data, Interpro domains, and GO terms [6]. CANDID uses several heterogeneous data sources, some of them chosen to overcome bias [7]. Endeavour is, however, using 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 [8].

Among the tools using different data sources, ToppGene, SNPs3D, GeneDistiller and Posmed include mouse data within their algorithms, but in a different manner. ToppGene combines mouse phenotype data with human gene annotations and literature [9]. SNPs3D identifies genes that are candidates for being involved in a specified disease based on literature [10]. GeneDistiller uses mouse

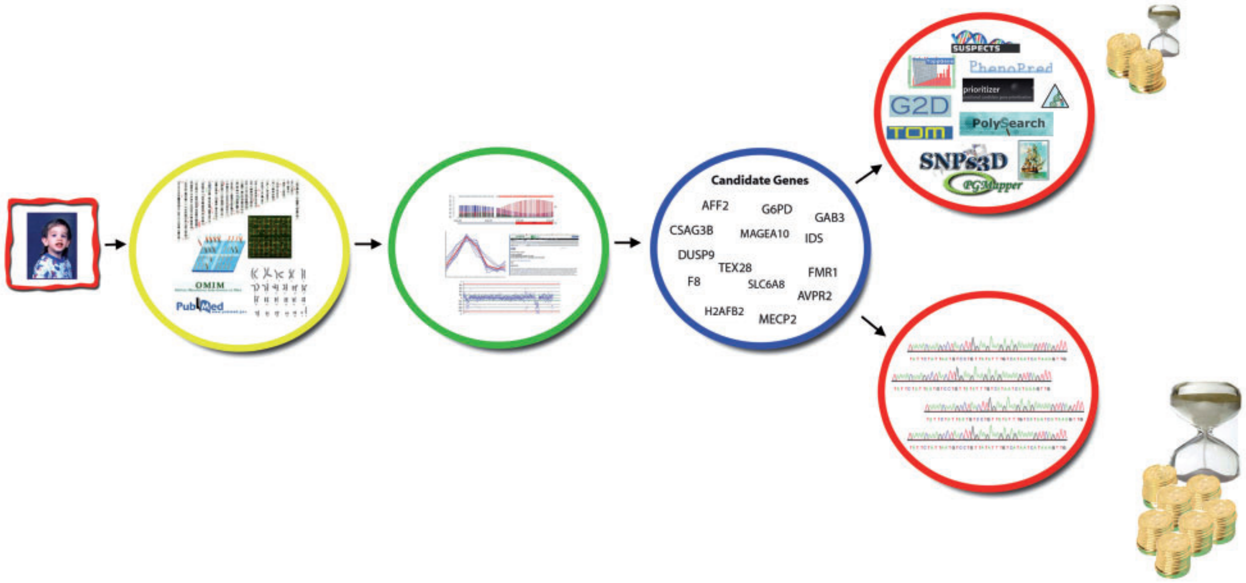

Figure 1: A major challenge in human genetics is to unravel the genetic variants and the molecular basis that underlay genetic disorders. In the past decades, geneticists have mainly used high-throughput technologies (such as linkage analysis and association studies). These technologies usually associate a chromosomal region, possibly encompassing dozens of genes, with a genetic condition. Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task. The use of computational solutions, such as the ones reviewed in that paper, could reduce the time and the money spent for such analysis without reducing the effectiveness of the whole approach.

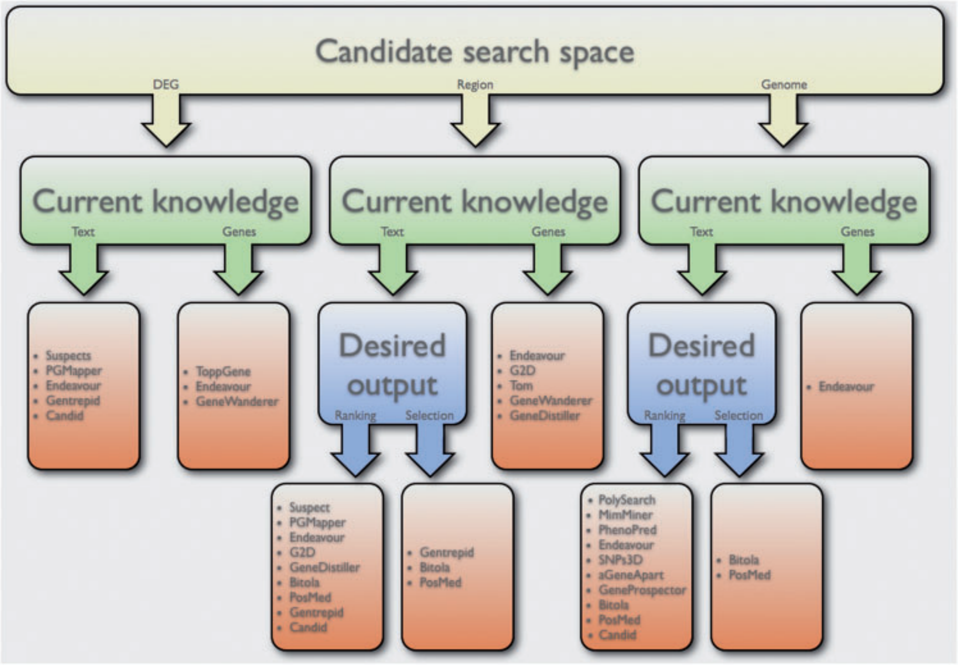

Figure 2: Decision tree that categorizes the 19 gene prioritization tools according to the outputs they use and the outputs they produce. This tree is designed to support the end users in their decision so that they can choose the tools that suit best their needs. By starting from the first question on the top and by going down, the user can determine a list of tools that can be used; in addition, the Figure 3 that describes the data sources used by the tool can also be used to support the decision.

Table 1: Overview of the 19 tools reviewed in the current study with their corresponding publications and website

Tool	References	Website
SUSPECT	[6]	http://www.genetics.med.ed.ac.uk/suspects/
ToppGene	[9]	http://toppgene.cchmc.org/
PolySearch	[15]	http://wishart.biology.ualberta.ca/polysearch/index.htm
MimMiner	[16]	http://www.cmbi.ru.nl/MimMiner/cgi-bin/main.pl
PhenoPred	[23]	http://www.phenopred.org
PGMapper	[21]	http://www.genediscovery.org/pgmapper/index.jsp
Endeavour	[8, 32]	http://www.esat.kuleuven.be/endeavour
G2D	[33, 34]	http://www.ogic.ca/projects/g2d.2/
TOM	[13, 14]	http://www-micrel.deis.unibo.it/~tom/
SNPs3D	[10]	http://www.SNPs3D.org
GenTrepid	[20]	http://www.gentrepid.org/
GeneWanderer	[22]	http://compbio.charite.de/genewanderer
Bitola	[17]	http://www.mf.uni-lj.si/bitola/
CANDID	[7]	https://dsgweb.wustl.edu/hutz/candid.html
PosMed	[12]	http://omicspace.riken.jp
GeneDistiller	[11]	http://www.genedistiller.org/
aGeneApart	[18]	http://www.esat.kuleuven.be/ageneapart
GeneProspector	[19]	http://www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do

phenotype to filter genes [11] and Posmed utilizes among other data sources orthologous connections from mouse to rank candidates [12].

G2D 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 [4]. TOM efficiently employs functional and mapping data and selects relevant candidate genes from a defined chromosomal region [13, 14].

Tools that are mainly based on literature and text mining are PolySearch, MimMiner, BITOLA, aGeneApart and GenePropector. PolySearch extracts and analyses relationships between diseases, genes, mutations, drugs, pathways, tissues, organs and metabolites in human by using multiple biomedical text databases [15]. MimMiner analyses the human phenome by text mining to rank phenotypes by their similarity to a given disease phenotype [16] and BITOLA mines MEDLINE database to discover new relations between biomedical concepts [17]. aGeneApart 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 [18]. GeneProspector 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 [19].

Finding associations between genes and phenotypes is the focus of Gentrepid and PGMapper.

Whereas Gentrepid predicts candidate disease genes based on their association to known disease genes of a related phenotype [20], PGMapper 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 [21].

Tools, such as GeneWanderer, Prioritizer, Posmed and PhenoPred, make use of genomewide networks. GeneWanderer is based on protein–protein interaction and uses a global network distance measure to define similarity in protein–protein interaction networks [22]. PhenoPred 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 [23]. Instead of using a human protein–protein interaction network, Posmed 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 [12].

Although we have limited our analysis to the tools freely accessible via a web interface, we are aware of other gene prioritization methods that were excluded of the present analysis but that still represent important contributions to the field. First,

Box 1: 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 a 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 (such as Reactome [58] and Kegg [59]) define a 'bio-molecular view' of the genes, while PPI networks (such as HPRD [60] and MINT [61]) define an 'interactome view'. 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 [8, 29]. Supplementary Table 1 contains the list of the 12 data sources we have defined.

Inputs

Two distinct 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 genomic condition defines a quantitative trait locus (QTL), the candidates are therefore the genes lying in that region. Another possibility is a list of genes differentially expressed in a tissue of interest that are not necessary from the same chromosomal location. Alternatively, the whole human genome can be used. An overview of the inputs required by the applications can be found in Table 2.

Outputs

For the 19 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 only the most promising 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 ranking only these. Several tools benefit from an additional output, a statistical measure, often a *P*-value, which 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. 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 2.

Text mining

It is the automatic extraction of information about genes, proteins and their functional relationships from text documents [62].

several gene prioritization methods, such as CAESAR [24], GeneRank [25] and CGI [26] propose interesting alternatives (e.g. natural language processing based disease model [24]), 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 [27], POCUS [28], Prioritizer [29]). 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 [30, 47–53]. For instance, Lombard *et al.* [47] have prioritized 10 000 candidates for the fetal alcohol syndrome (FAS) using a complex set of 29 filters. Their analysis reveals interesting therapeutic targets like TGF-β, MAPK and members of the Hedgehog signaling pathways. Another example is the network-based classification of breast cancer metastasis developed by Chuang *et al.* [48]. These approaches are, however, case specific and cannot be easily ported to another disease. Last, alternative techniques to circumvent recurrent problems in gene prioritization are currently under development. As an illustration, Nitsch *et al.* [31] 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.

DESCRIPTION OF THE GENE PRIORITIZATION METHODS

The genomic data are at the core

We have defined a data source as a type of data that represents a particular view of the genes (see Box 1—'Gene view') and thus can correspond to several related databases. Data sources are at the core of the gene prioritization problem since both high coverage and high quality data sources are needed to make accurate predictions. In total, 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, which describes for each tool which data sources it uses (Supplementary Table 1). We can observe from the data source landscape map that text mining is by far the most widely used data source since 14 out of the 19 tools are using co-occurrence or functional text mining. Most of the approaches make use of a wide range of data sources covering distinct views of the genes, but four tools rely exclusively on text mining (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 and only by two tools at most although they describe unique features that might not always be captured by the other data sources. However, the rule should not be to include as many data sources as possible but rather to reach a critical mass of data beyond which accurate predictions can be made.

Inputs and outputs of the methods

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. We furthermore consider two possibilities for the prior knowledge as it can be defined by a set of genes or by a set of keywords. The retrieval of a training set requires the knowledge of, at least, one disease causing gene, but preferably more than one. In addition, the set needs to be homogeneous, meaning that it usually contains between 5 and 25 genes that, together, describe a specific biological process. When no disease gene can be found, members of the pathways disturbed by the diseases are also an option (Thienpont *et al.*, manuscript in preparation). 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. In the second case, 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). Regarding the candidate search space, we have distinguished between a locus, a differentially expressed genes (DEG) list, and the whole genome. A locus is a set of neighboring genes (e.g. all genes from the cytogenetic band 22q11.23) while the genes in a DEG list are not necessarily located at the same locus. Although these two options are similar, the distinction we made is important since several tools allow the definition of a locus but not of DEG list and vice versa. Alternatively, nine tools allow the exploration of the full genome, in case no candidate gene set can be defined beforehand.

Regarding the outputs, two types were considered, a ranking and a selection of the candidate genes. In a ranking scenario, all the candidates are ranked so that the most promising candidate can be found at the top, while for a selection, a subset of the original candidate set, containing only the most promising candidates, is returned. From the 19 tools, four perform a selection of the candidates and three of these four perform a selection followed by a ranking. In addition, we record which tools further measure the significance of their results via any statistical method. Of interest, a selection can then be obtained from a ranking by using a threshold on this statistical measure. Table 2 shows an overview of the input data required by the tools as well as the output they produce. Also, a clustering of the tools regarding to their inputs and outputs is presented in Figure 3. 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 the tools they want to use. By answering these questions, users define first, which genes are candidate; second, how the current knowledge is represented; and third (when necessary), what is the desired output type.

The importance of biological validation

Since the methods we are interested in are predictive, an important criterion for selection is the performance. The tools reviewed here were all originally published together with the results of a benchmark analysis as a proof of concept. It is however difficult to

Table 2: Description of the inputs needed by the tools and the outputs produced by the tools

Tool	Inputs					Output		
	Training data		Candidate genes			Ranking	Selection of candidates	Test statistic
	KnownGenes	Keywords	Region	DEG	Genome			
SUSPECT		×	×	×		×		
ToppGene	×			×		×		×
PolySearch		×			×	×		×
MimMiner		×			×	×		
PhenoPred		×			×	×		
PGMapper		×	×	×		×		
Endeavour	×	×	×	×	×	×		×
G2D	×	×	×			×		×
TOM	×		×				×	
SNPs3D		×			×	×		
GenTrepid		×	×	×		×	×	
GeneWanderer	×		×	×		×		×
Bitola		×	×		×	×	×	
CANDID		×				×		×
aGeneApart		×			×	×		×
GeneProspector		×			×	×		
PosMed		×	×		×	×	×	×
GeneDistiller	×	×	×			×		

We distinct two types of inputs: the prior knowledge about the genetic disorder of interest and the candidate search space. The prior knowledge 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. The candidate search space is either a locus linked to a genomic condition or a list of genes differentially expressed in a tissue of interest (DEG) or the whole human genome. The output is either a ranking of the candidate genes or a selection of the most promising candidates. In addition, a statistical measure that estimates how likely it is to obtain that result by chance alone. More details about the inputs and outputs can be found in the Box 1.

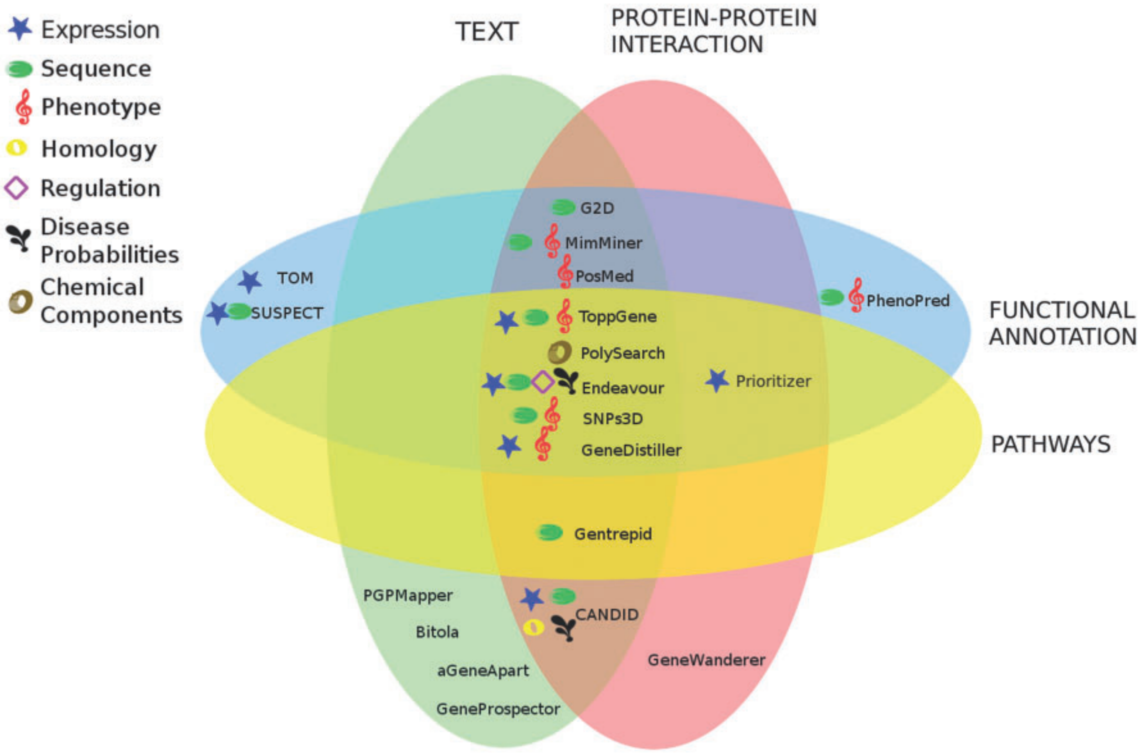

Figure 3: Repartition of the 19 tools according to the data sources they use. The four data sources most commonly used are Text (functional and interactions mining), protein–protein interactions, functional annotations and pathways and are therefore represented as large ellipses. The additional seven data sources are represented with symbols.

compare the performance of these benchmarks directly since their setups are different (different diseases, different genes). Although a rigorous comparison is still missing, various studies that compare several gene prioritization tools by analyzing their performance on a particular disease have been performed (e.g. on T2D [41, 42, 54]). An overview is presented in Supplementary Table 2. Although it is of primary importance, the performance obtained through a benchmark analysis represents more a proof of concept than a critical performance assessment. Therefore, it is only an estimation of the real performance (e.g. for a real biological application) and it is also most likely benchmark specific. That is the reason why we believe that the definition of the desired inputs/outputs and data sources, and the knowledge of real biological applications are also crucial.

Beside these benchmarks, several biological applications have been described in the literature. Supplementary Table 3 gives an overview of these applications. Interestingly, three of them analyzed T2D associated loci and are using several gene prioritization tools in conjunction [41, 42, 54]. Elbers *et al.* [42] analyzed five loci previously reported to be linked with both T2D and obesity that encompass more than 600 genes in total. 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, five genes were involved in immunity and defense (e.g. TLR2, FGB) and it is 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.* [41] prioritized T2D and obesity associated loci and proposed another set of 164 promising candidates. 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 set of 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.* [54] compared the finding from recent genome-wide association studies (GWAS) to the predictions made by eight gene prioritization methods. 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.* [43] in fruit fly. The aim of a genetic screen is to discover *in vivo* associations between genotypes and phenotypes. A forward 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 and then only the genes ranked in the top 30% of every locus are assayed in a secondary screen. Additionally, it was shown that 30% is a conservative threshold since all the positives 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.

Intuitive interfaces

Beside the data, the inputs/outputs and the performance, what is critical for a tool to be used is its interface. Ideally, it has to be an intuitive interface that accepts simple inputs and provides detailed outputs. A past success and reference in bioinformatics is basic local alignment search tool (Blast) for which only a single sequence needs to be provided [55]. In return, Blast provides the complete detailed alignments together with cross-links to sequence databases 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 probably corresponding to the simple input—detailed output paradigm described above. Besides, the presence of an advanced mode that allows users to fine tune the analysis is also clearly an advantage (e.g. defining a threshold for the Blast *e*-value).

Several gene prioritization tools such as MimMiner, PhenoPred, aGeneApart and GeneProspector can already be fed with a single disease name that represents the simplest training input possible. However, an advanced mode to fine tune the analysis is missing for these applications. The outputs generated by the tools are very detailed and almost always contain cross-references to external databases (e.g. Hugo, EnsEMBL, RefSeq). However, only few tools present detailed information about the data underlying the ranking of the candidate genes. This data is crucial for the user who needs to determine which candidates should be investigated further. This is probably the weakest point of most of the current tools although several tools like Suspects and G2D already propose preliminary solutions. In addition, most of the tools benefit from a user manual and a dedicated help section that help users to understand how they should interact with the interface.

FUTURE DIRECTIONS

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 have already mentioned the interface, which 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 [9–12]. Already, several tools described in the current review include rodent data (e.g. SNPs3D, ToppGene, GeneDistiller, 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 [56] already represents a first step in that direction since it includes aCGH data from patients and allow text mining prioritization (using the core engine of aGeneApart [18]) 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 and miRNA data. Another important research track is to explore different computational approaches to improve once more the algorithms that are running the gene prioritization methods. Preliminary results have shown, for example, that kernel methods are more efficient than simpler statistical methods such as Pearson correlation or binomial based over-representation [57]. The last challenge of this field is its necessary adaptation to the shift observed in genetics towards the study of more complex disorders that is though to be more difficult than the study of the Mendelian diseases.

Altogether, the methods described in this review 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 whose aim is to represent an updatable electronic version of the present review. This web site, termed ‘Gene Prioritization Portal’ (available at: http://www.esat.kuleuven.be/gpp), contains, for every tool, a detailed sheet that summarizes 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 Table 2 and Supplementary Table 1 of the current publication. We believe that this website represents a first step to guide users through their gene prioritization experiments.

CONCLUSION

This review tries to clarify the world of gene prioritization to the final user through an exhaustive guide of 19 human candidate gene prioritization methods that are freely accessible through a web interface. This taxonomy has been done according to different characteristics of the tools, including the type of input, data sources used during the process of prioritization and the desired output. We think that this review is a useful tool not only to help the wet lab researchers to dive into gene prioritization, but also to guide them to select the most convenient method for their analysis.

To keep up with the especially fast evolving world of bioinformatics in general and gene prioritization in particular, we have developed a website http://www.esat.kuleuven.be/gpp/ that contains updated information of all the tools described in this review. We expect our portal to become a reference point in gene prioritization where not only users but also developers will find up-to-date information necessary for their research.

Key Points

- Numerous computational methods have been developed to tackle the gene prioritization problem in human; we have collected the methods that offer such web services freely.
- We have described how these methods differ from each other by the inputs they need, the outputs they produce and the data sources they use.
- We have furthermore described some of the biological applications to which gene prioritization approaches were successfully applied.
- A website that contains information about the available gene prioritization methods has been developed and will be updated on a regular basis.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

Research Council KUL [GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA]; the Flemish Government [G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, ICCoS, ANMMM, MLDM, G.0733.09, G.082409, GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dsquare]; the Belgian Federal Science Policy Office [IUAP P6/25]; and the European Research Network on System Identification (ERNSI) [FP6-NoE, FP6-IP, FP6-MC-EST, FP6-STREP, FP7-HEALTH].

References

1. Redon R, Ishikawa S, Fitch KR, *et al.* Global variation in copy number in the human genome. *Nature* 2006;**444**:444–54.
2. Marazita ML, Murray JC, Lidral AC, *et al.* Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *Am J Hum Genet* 2004;**75**:161–73.
3. Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000;**10**:1435–44.
4. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet* 2002;**31**:316–9.
5. Zhang P, Zhang J, Sheng H, *et al.* Gene functional similarity search tool (GFSST). *BMC Bioinformatics* 2006;**7**:135.
6. Adie EA, Adams RR, Evans KL, *et al.* SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 2006;**22**:773–4.
7. Hutz JE, Kraja AT, McLeod HL, *et al.* CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet Epidemiol* 2008;**32**:779–90.
8. Aerts S, Lambrechts D, Maity S, *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**:537–44.
9. Chen J, Xu H, Aronow BJ, *et al.* Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 2007;**8**:392.
10. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006;**7**:166.
11. Seelow D, Schwarz JM, Schuelke M. GeneDistiller–distilling candidate genes from linkage intervals. *PLoS ONE* 2008;**3**:e3874.
12. Yoshida Y, Makita Y, Heida N, *et al.* PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res* 2009;**37**:W147–52.
13. Rossi S, Masotti D, Nardini C, *et al.* TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res* 2006;**34**:W285–92.
14. Masotti D, Nardini C, Rossi S, *et al.* TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics* 2008;**24**:428–9.
15. Cheng D, Knox C, Young N, *et al.* PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008;**36**:W399–405.
16. van Driel MA, Bruggeman J, Vriend G, *et al.* A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**:535–42.
17. Hristovski D, Peterlin B, Mitchell JA, *et al.* Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005;**74**:289–98.
18. Van Vooren S, Thienpont B, Menten B, *et al.* Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res* 2007;**35**:2533–43.
19. Yu W, Wulf A, Liu T, *et al.* Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics* 2008;**9**:528.
20. George RA, Liu JY, Feng LL, *et al.* Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 2006;**34**:e130.
21. Xiong Q, Qiu Y Gu W. PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics* 2008;**24**:1011–3.
22. Köhler S, Bauer S, Horn D, *et al.* Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
23. Radivojac P, Peng K, Clark WT, *et al.* An integrated approach to inferring gene-disease associations in humans. *Proteins* 2008;**72**:1030–7.
24. Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. *Bioinformatics* 2007;**23**:1132–40.
25. Morrison JL, Breitling R, Higham DJ, *et al.* GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005;**6**:233.

26. Ma X, Lee H, Wang L, *et al*. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007;**23**: 215–21.
27. Braun TA, Shankar SP, Davis S, *et al*. Prioritizing regions of candidate genes for efficient mutation screening. *Hum Mutat* 2006;**27**:195–200.
28. Turner FS, Clutterbuck DR, Semple CAM. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003;**4**:R75.
29. Franke L, van Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**:1011–25.
30. Tiffin N, Okpechi I, Perez-Iratxeta C, *et al*. Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. *Physiol Genomics* 2008;**35**:55–64.
31. Nitsch D, Tranchevent L, Thienpont B, *et al*. Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE* 2009;**4**:e5526.
32. Tranchevent L, Barriot R, Yu S, *et al*. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 2008;**36**:W377–84.
33. Perez-Iratxeta C, Wjst M, Bork P, *et al*. G2D: a tool for mining genes associated with disease. *BMC Genet* 2005;**6**: 45.
34. Perez-Iratxeta C, Bork P, Andrade-Navarro MA. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res* 2007;**35**:W212–6.
35. Smith NGC, Eyre-Walker A. Human disease genes: patterns and predictions. *Gene* 2003;**318**:169–75.
36. Goh K, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
37. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;**409**:853–5.
38. Iizuka M, Kubo Y, Tsunenari I, *et al*. Functional characterization and localization of a cardiac-type inwardly rectifying K+ channel. *Recept Channels* 1995;**3**:299–315.
39. Wasada T. Adenosine triphosphate-sensitive potassium (K(ATP)) channel activity is coupled with insulin resistance in obesity and type 2 diabetes mellitus. *Intern Med* 2002;**41**: 84–90.
40. Lavine N, Ethier N, Oak JN, *et al*. G protein-coupled receptors form stable complexes with inwardly rectifying potassium channels and adenylyl cyclase. *J Biol Chem* 2002; **277**:46010–19.
41. Tiffin N, Adie E, Turner F, *et al*. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res* 2006; **34**:3067–81.
42. Elbers CC, Onland-Moret NC, Franke L, *et al*. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab* 2007;**18**:19–26.
43. Aerts S, Vilain S, Hu S, *et al*. Integrating computational biology and forward genetics in Drosophila. *PLoS Genet* 2009;**5**:e1000351.
44. Myers CL, Barrett DR, Hibbs MA, *et al*. Finding function: evaluation methods for functional genomic data. *BMC Genomics* 2006;**7**:187.
45. Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinformatics* 2005;**6**:34–43.
46. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol* 2008;**4**: e1000160.
47. Lombard Z, Tiffin N, Hofmann O, *et al*. Computational selection and prioritization of candidate genes for fetal alcohol syndrome. *BMC Genomics* 2007;**8**:389.
48. Chuang H, Lee E, Liu Y, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
49. Huang Q, Li GHY, Cheung WMW, *et al*. Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. *J Hum Genet* 2008;**53**:644–55.
50. Gajendran VK, Lin J Fyhrie DP. An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone* 2007;**40**:1378–88.
51. Alsaber R, Tabone CJ, Kandpal RP. Predicting candidate genes for human deafness disorders: a bioinformatics approach. *BMC Genomics* 2006;**7**:180.
52. Rasche A, Al-Hasani H, Herwig R. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics* 2008;**9**:310.
53. Furney SJ, Calvo B, Larrañaga P, *et al*. Prioritization of candidate cancer genes–an aid to oncogenomic studies. *Nucleic Acids Res* 2008;**36**:e115.
54. Teber ET, Liu JY, Ballouz S, *et al*. Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. *BMC Bioinformatics* 2009;**10**(Suppl. 1):S69.
55. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
56. Firth HV, Richards SM, Bevan AP, *et al*. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 2009; **84**:524–33.
57. De Bie T, Tranchevent L, van Oeffelen LMM, *et al*. Kernel-based data fusion for gene prioritization. *Bioinformatics* 2007; **23**:i125–32.
58. Vastrik I, D'Eustachio P, Schmidt E, *et al*. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007;**8**:R39.
59. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
60. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human Protein Reference Database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
61. Chatr-aryamontri A, Ceol A, Palazzi LM, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;**35**: D572–4.
62. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. *Genome Biol* 2005; **6**:224.