Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(3):1126-1133

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Data mining strategies for identification of HNF4A MODY gene using gene prioritize tool

Soniyapriyadharishni A. K. and P. B. Ramesh Babu

Department of Bioinformatics, Bharath University, Chennai, Tamil Nadu, India

---

ABSTRACT

One of the major challenges in human genetics is to find the genetic variants underlying genetic disorders for effective diagnostic testing and for unravelling the molecular basis of these diseases. Finding the most promising genes among large lists of candidate genes has been defined as the gene prioritization problem. In the past decades, the use of high-throughput technologies (such as mining technologies, linkage analysis and association studies) has permitted major discoveries in that field. These technologies can usually associate a chromosomal region with a genetic condition. It is a recurrent problem in genetics in which genetic conditions are reported to be associated with chromosomal regions. To overcome this, several different computational approaches have been developed to tackle this challenging task. In this study, we used 19 computational solutions for human MODY gene prioritization that are accessible as web tools and illustrate their differences. Where various biological problems to which they have been successfully applied. Ultimately, we described several research directions that could increase the quality and applicability of the tools. With the help of developed website (http://www.esat.kuleuven.be/gpp) containing detailed information about these genes and other tools, this 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: Bio creative, Curator, HNFA, MODY, Portal, Prioritization.

INTRODUCTION

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 each gene, and assess whether it is a promising candidate or not. Thus 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*. who already described the first approach to tackle this problem. Since then, many different strategies have been developed (Zhang P *et al*.2006), 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 (Smith, Jimenaz *et al*.2008). For example, when studying T2D, KCNJ5 appears as a good candidate through its potassium channel activity (Iizuka M, Kubo Y *et al*.1995), an important pathway for diabetes (Wasda *et al*.2002), and because it is known to interact with ADRB2 (Huang Q *et al*.2008), 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 (Tiffin N, Elbers *et al*.2007). More recently, Aerts *et al*.2009 have developed a computationally supported genetic screen whose computational part is based on gene prioritization **(Figure 1).**

With this assessment, the dissertation work cross-purpose 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 used a developed website termed **'Gene Prioritization Portal' (available/http://www.esat.kuleuven.be/gpp/)** that represents an updatable electronic review of this field.

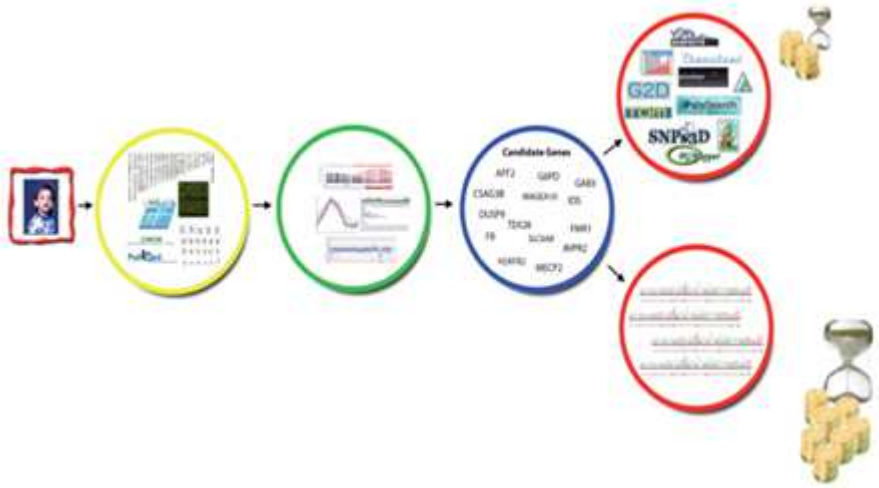

Figure 1: A major challenge in human genetics is to unravel the genetic variants and the molecular basis that underlay genetic disorders

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 could reduce the time and the money spent for such analysis without reducing the effectiveness of the whole approach

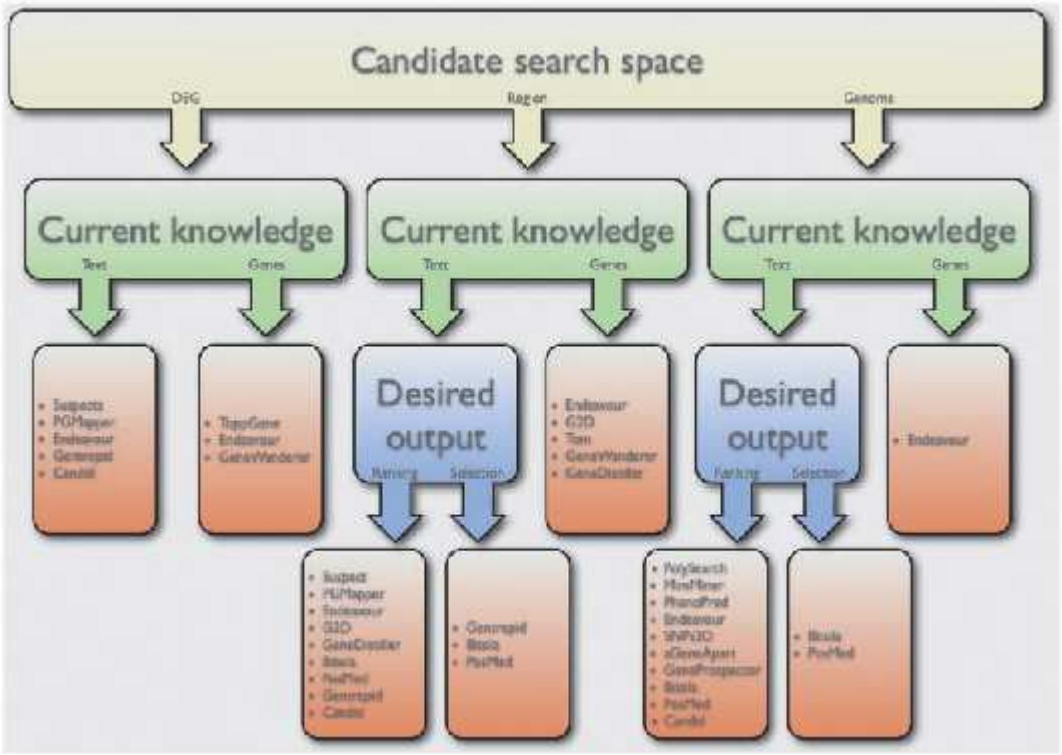

Figure 2: Decision tree that categorizes the19 gene prioritization tools according to the outputs they use and the outputs they produce. This tree is designed to support the end users in their decision so that they can choose the tools that suit best their needs

2. Selecting the gene prioritization tools

In this study, we operated 19 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 gene prioritization with reasonable performance (Sheng H, Zhang P *et al*. 2006). 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 (Myers CL, Hibbs *et al*. 2006).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*. and Punta *et al*. 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 then on commercial 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.

3. Identification of HNF4A MODY gene using Gene prioritize tool

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 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 (**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. Thus, with the help of biological Curator and some data mining strategies HNF4A MODY genes are identified using Gene prioritize tool.

Table 1: Tools and their Data descriptions

Tools	Description and Use
SUSPECT	Ranks genes from a given chromosomal region of interest regarding to a specific disease or a set of candidate genes by matching sequence features
ToppGene	Prioritize or rank candidate genes based on functional similarity to training gene list
PolySearch	Used in biomedical text mining is to handle "comprehensive" or "associative" queries
MimMiner	Used to retrieve the related phenotypes of a specific phenotype.
PhenoPred	Prioritize genes based on their likelihood to be associated with a given disease or to prioritize diseases for a given query gene.
PGMapper	Obtains the information of all genes of a disease locus
Endeavour	A software used for the computational prioritization of candidates genes, based on a set of training genes
G2D	Used for prioritizing candidates genes for inherited diseases
TOM	Used for the efficient extraction of candidate genes for hereditary diseases
SNPs3D	Used to assigns molecular functional effects of non-synonymous SNPs based on structure and sequence analysis.
GenTrepid	Assumes that common phenotypes are associated with dysfunction in proteins that participate in the same complex or pathway.
GeneWanderer	Solve the probability to be involved in particular disease or phenotype.
Bitola	Used to decrease the number of candidate relations
CANDID	Prioritizes candidate genes for genetic analysis
aGeneApart	It shows text based gene profiling for human dysmorphy
GeneProspector	It display information about genes in relation to human diseases, risk factors and other phenotypes
PosMed	It ranks biomedical resources such as genes, metabolites, diseases and drugs, based on the statistical significance of associations between a user-specified phenotypic keyword
GeneDistiller	Used to find an interval begins with a genetic marker and end with a physical position

4. Curator's interactive candidate gene selection task

The proposed task involved in candidate gene selection using Gene Prioritization tool

Candidate gene selection using Gene prioritization tool can be processed in five detective steps each steps will elucidate the function of each region in the following sections.

Biocurator { *Gene portal* } Task 1:

Data source

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 and Kegg) define a 'bio-molecular view' of the genes, while PPI networks (such as HPRD and MINT) 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 **{Tabular Array1**} which contains the list of the 12 data sources that have identified for MODY candidates who participated in the MODY candidate selection scrutiny.

Biocurator { *Gene portal* } Task 2:

Input Evaluation – The User Interface

Two distinct types of inputs are distinguished: the prior knowledge about the genetic disorder of MODY and the candidate search space. On the one hand, the prior knowledge represents what is currently known about the MODY under study, it can be represented either as a set of genes i.e.HNF4A known to play a role in the MODY or as a set of key words such as D1 (MODY1) that describe the MODY disease. On the other hand, the candidate search space D1 candidates define HNF4A genes are positive candidates. For instance, a locus linked to a G2D defines a quantitative trait locus (QTL); the HNF4A 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 i.e. for HNF4A genes is displayed in {**Tabular Array 2}**.

Biocurator { *Gene portal* } Task 3:

Output interactive task evaluation results

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 D1 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). An overview of the outputs produced by the different applications of the concerned HNF4A genes are evaluated and displayed in {**Tabular Array 2}.**

Biocurator { *Gene portal* } Task 4:

Text-mining-based database curation proposed by biocurator

The flowchart for constructing the Text mining flow comprises of three stages: (i) dataset collection and pre-processing, (ii) candidate genes extraction and (iii) manual verification.

Stage 1: Dataset collection and pre-processing

In this stage, we collect MODY-related abstracts from PubMed and filter out those that are non-genetic by using a genetic research filter. The filter uses a list of key words that are frequently used in abstracts of genetic research such as 'polymorphism', 'alleles', 'variants' and regular expressions to determine whether the abstract is genetically related. Further these, information's will be used by the following stages for recognizing entities and extracting disease– gene pairs.

Stage 2: Candidate genes extraction

In Stage 2, we extracted D1 related candidate genes from the pre-processed dataset through the following steps.

First, a MODY gene named entity recognition system is employed to recognize disease terms in a sentence. Second, the GI system is used to recognize and link mentioned HNF4A genes to their corresponding Entrez Gene IDs. Based on the results of the previous steps, if a MODY term and HNF4A genes are present in the same sentence, they are extracted as a disease–gene (D-G) candidate pair. Finally, the D-G extraction system determines whether a relation indeed exists within this D-G pair.

Stage 3: Manual curation

Although the employed text-mining components have shown satisfactory scores **{Tabular Array 4}**, the text-mined candidate genes are examined by all Text mining MODY curators in Stage 3 to further ensure the quality of the curated content. In this stage, newly extracted D1 candidate genes and their corresponding evidence sentences and abstracts are presented to the MODY curators. MODY curators review each extracted candidate gene and remove the incorrect results.

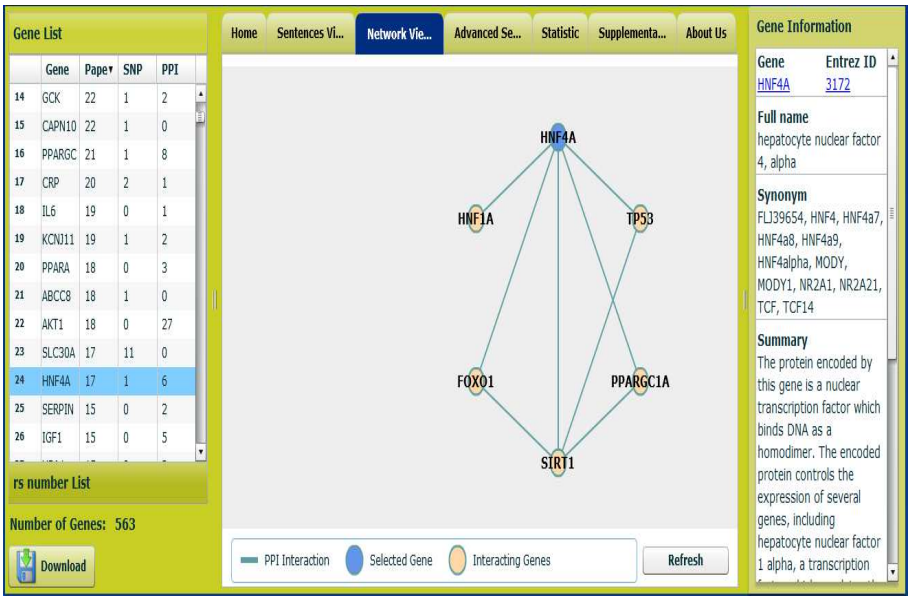

Figure3. The network visual illustration of HNF4A gene

Biocurator { *Gene portal* } Task 5:

The biological Validation of selected Candidate genes by user

Thus finally we curated, genes using gene prioritization tools in conjunction and reported 30 interesting MODY candidates. Some of them (D1) were already known to be involved in either diabetes or obesity (e.g. TCF1 and HNF4A, responsible for MODY) but some candidates were novel predictions. See, **the network visual illustration (Figure 3)**. Among them, five genes were involved in immunity and defence (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 glut-tony (e.g. AACS, PTGIS and the neuropeptide Y receptor family members). Using a similar strategy prioritized MODY associated loci and proposed another set of promising candidates. Of interest, 12 families of the 30 candidates with D1 were finally curated reported for further analysis. Although there is an overlap between the predictions of D1 and other candidate genes comparison table i.e. **{Tabular Array 3}**, some important discrepancies remain and can be explained by the fact that the two candidate genes (D1 and other genes) has to fulfil the BLAST (basic local alignment search tool) search for further biochemical and Data analysis process see BLAST evaluation result.

RESULTS

5.1 Biocreative Interactive Task Evaluation Results

Tabular array 1: Data sources used by the 19 tools. We have defined 12 distinct types, each type can correspond to several databases (e.g., Reactome and Kegg are two pathway databases). A black cross means that a data source is available for one tool. A red cross means that the user can add its own data source of that type

Tools	Text (coocurrance)	Text (functional mining)	PPIs	Functional Annotations	Pathways	Expression	Sequence	Phenotype	Conservation/ Homology	Regulation	Disease probabilities/scores	Chemical Components
SUSPECT				x		x	x					
ToppGene	x		x	x	x	x	x	x				
PolySearch		x	x	x	x							x
MimMiner		x	x	x			x	x				
PhenoPred			x	x			x	x				
PGMapper		x										
Endeavour		x	x	x	x	x x	x			x	x x	
G2D	x		x	x			x					
TOM				x		x						
SNPs3D		x	x	x	x		x	x				
GenTrepid		x	x		x		x					
GeneWanderer			x									
Bitola		x										
CANDID		x	x			x	x		x		x	
aGeneApart		x										
GeneProspector		x										
PosMed	x	x	x	x				x				x
GeneDistiller	x		x	x	x	x		x				

Tabular Array 2: Description of the input of HNF4A gene needed by the tools and the output data selection of D1 candidate produced by the tools

Tools	Input Data					Output Data		
	Training data		Candidate gene			Ranking	Selection of D1 candidates	Test statistic
	HNF4A gene	MODY1	Chrom Reg 20	DEG	Genome			
SUSPECT	x	x	x	x		x	x	
ToppGene		x		x		x	x	
PolySearch		x			x	x		
MimMiner		x			x	x		
PhenoPred		x				x		
PGMapper		x	x			x		
Endeavour	x	x		x	x	x	x	
G2D	x	x	x	x		x	x x	
TOM	x		x				x	
SNPs3D		x	x					
GenTrepid		x	x			x		
GeneWanderer	x			x		x	x	
Bitola		x				x		
CANDID	x	x	x	x		x	x	
aGeneApart		x		x	x	x	x	
GeneProspector		x			x	x		
PosMed		x	x		x	x	x x	
GeneDistiller	x	x	x			x		

Tabular Array 3: Comparison of D1 and D3 candidate genes in Text mining verification interface

	True positive	False positive	False negative	Precision	Recall	*F*-score	Number of documents
TCF	65	42	23	0.797	0.771	0.784	152
IPF	105	35	29	0.75	0.784	0.766	118
MODY1	30	14	16	0.714	0.69	0.702	93
MODY3	22	17	05	0.784	0.774	0.779	145
Overall	222	108	73	0.771	0.763	0.767	502

5.2 BLAST task search result

BLAST search for MODY1 gene of D1 Candidates

>gi|147883853|gb|EF591040.1| Homo sapiens hepatocyte nuclear factor 4, alpha (HNF4A) gene, complete cds, responsible for MODY1 (Maturity onset diabetes young type 1)

AAATTGAAGCCCTGAGAAGAGACGATGTAAATAAAGTTATCGGAATAGTGACTGAGCTAGTTCCTAAC
CCCAGGTCTCCTGACGTCAAATCTAGCCTCTCTTGTCCTCCACAACAGTTGCTCCCTCTCCCCTTTCCCT
ACAGGCAATAGCACTCCCCAGTCATCTCAGCTTCAGAGGTCAGATCAAGTGGACAGAATAAAGCTAA
GCAGGGCAGAGAAAGGGCTTCTAGCAGTGGAAAGAACATGAAGATGCCTCCTCATACCACTCGAACA
CACATGCCCTGACTCCCGATTTGCTCACTCATTAATCCACCCACACATTCACTCATTCACTCACTCATTA
ATTCACCCACCCATTCACTCATTAATTCACCCACCCATTTGCTCACTTACTCAGTAATTTACCCTCATTC
GTTCATTCACTCACTCATTCATGCATTCACTCAACTCACTCATTAATTCTCCCTGTCTCTAATTCACTCA
CTCCTAATTCACCCACTCATCAATTCCCCCACACACTCACTCATTAATTGGTGCTCCCACTAATTCGCCC
ACTCTTGCTCACAGGTGTCACCGAGGCCCTCCAACCTGCCCTTCCAGCCAGCAGTGGAGGACAGGCTC
CCAGGCCTCCCAAGTCTCAGGCTTTTCTTCCTCTGCCCTCACTCTCTGCCTTCTACATCAAGACTTTACT
TCCCCAGATTGATTGGCAGCCCTTGAAAATGTCTGCACAGAAGGCAATGAGGGCTGGAGGGAGTGAG
AAGAACAGAGTGCACCATGAGCTTCAGACAGCCAGGACCAGCTGTGTAATTTTCAGGGCCCCGTCCAA
AACAAAAATGCACAATCCCTTGTTCAAAAGTCATGATGAATTTTAGGACAACCACAGCAGAGCACTGA
ACCAGGAACCAGGACTTTCTAAGGGTTGGGTTGCCTGTGACTGCACTGGCGATACCCCCACAAAGCCC
ACTCTGAAGGTAGGAGACGGGTGGAGAGAAACAGGGGGATGGCAAGGGGGATACGAAACAGGGAGA
GGGAGGAGGGGGAAGAGGATGGACGTCTACCAGGCCCCACTTGGTGCTTGATTTATGCCATCTCATTT
CCTTCTCAAACCACCCTTTGAAGTTGATTGTACATTTTACAGAAAAGGAAACTGAGGCTCGGAGAGGA
GAATCATTTACCCAAGGTCCCAGTTAGTAGACGGTAG…….

CONCLUSION

This curator study tries to clarify the world of gene prioritization to the final level through an exhaustive guide of 19 human MODY candidate gene prioritization methods that are freely accessible through a web interface and data mining strategies. 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. Thus this process 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 to select candidate genes for their biochemical analysis.

Finally, with the help of Gene prioritization tool we got DEG in contrast using Blast search the primers for D1 promising candidates can be designed easily and used to probate the mutational identification and sequence analysis of HNF4A genes cordially.

REFERENCES

[1] Adie EA, Adams RR, Evans KL, *et al*. *Bioinformatics* **2006**; 22:773–4.
[2] Aerts S, Lambrechts D, Maity S, *et al*. *Nat Biotechnology* **2006**; 24:537–44.
[3] Chen J, Xu H, Aronow BJ, *et al*. *BMC Bioinformatics* **2007**; 8:392.
[4] Cheng D, Knox C, Young N, *et al*. *NucleicAcidsRes* **2008**; 36:W399–405.
[5] George RA, Liu JY, Feng LL, *et al*. *NucleicAcidsRes* **2006**; 34:e130.
[6] Hristovski D, Peterlin B, Mitchell JA, *et al*. *IntJMed Inform* **2005**; 74:289–98.
[7] Hutz JE, Kraja AT, McLeod HL, *et al*. *GenetEpidemiol* **2008;** 32:779–90.
[8] Jorde LB. *GenomeRes* **2000**; 10:1435–44.
[9] Kohler S, Bauer S, Horn D, *et al*. *AmJHumGenet* **2008**; 82:949–58.
[10] Marazita ML, Murray JC, Lidral AC, etal. *AmJHumGenet* **2004**; 75: 161–73.
[11] Masotti D, Nardini C, Rossi S, *et al*. *Bioinformatics* **2008**; 24: 428–9.
[12] Perez-Iratxeta C, Bork P, Andrade MA. *NatGenet* **2002**; 31:316–9.
[13] Radivojac P, Peng K, Clark WT, *et al*. *Proteins* **2008**; 72:1030–7.

[14] Redon R, Ishikawa S, Fitch KR, *et al*. *Nature* **2006**; 444: 444–54.
[15] Rossi S, Masotti D, Nardini C, *et al*. *NucleicAcidsRes* **2006**; 34:W285–92.
[16] Seelow D, Schwarz JM, Schuelke M. *PLoS ONE* **2008**; 3:e3874.
[17] Soniyapriyadharishni.A.K, P.B.Rameshbabu : *Research journal of pharmaceutical and biological research*; **2014**; 4(2).
[18] Soniyapriyadharishni.A.K, P.B.Rameshbabu: *Research journal of pharmaceutical and biological research*; **2013**; 4(3):840-856.
[19] Van Driel MA, Bruggeman J, Vriend G, *et al*. *EurJHum Genet* **2006**; 14: 535–42.
[20] Van Vooren S, Thienpont B, Menten B, *et al*. *Nucleic Acids Res* **2007**; 35:2533–43.
[21] Xiong Q, Qiu Y Gu W. PGMapper: *Bioinformatics* **2008**; 24:1011–3.
[22] Yoshida Y, Makita Y, Heida N, *et al*. *NucleicAcidsRes* **2009**; 37:W147–52.
[23] Yu W, Wulf A, Liu T, *et al*. *BMC Bioinformatics* **2008**; 9:528.
[24] Yue P, Melamud E, Moult J. *BMC Bioinformatics* **2006**; 7:166.
[25] Zhang P, Zhang J, Sheng H, etal. *BMC Bioinformatics* **2006**; 7:135.