International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Candidate Gene Predictions: Bioinformatics Significance in Linkage Analysis

Naureen Aslam Khattak[1]

[1]PMAS- Arid Agriculture University, Rawalpindi, Pakistan

Abstract: ***The rationale behind this review article is based on the identification of candidate genes in linkage analysis by using in-silico strategies. In recent development, it is noted that computational approaches are widely used to find out disease-causing genes. This is achieved by applying various greedy algorithms, which are constructed either on mathematical modeling, or computational search and alignment methods, or both. The advantage of computer-assisted techniques is that, it reduces the search domain of disease-causing genes which may comprise of 100 to 1000's candidate genes mapped on a single locus. In this context, the common criterion for in-silico candidate-gene identification relies on gene ontology, protein-protein interaction, sequence based features, functional annotation, data mining, and microarray-expression data. In addition to improve the disease-causing genes identification, development of disease-predicting software by incorporating existing data (wet-lab) will be a straightforward and effective step.***

Keywords: Bioinformatics, Linkage analysis, Candidate gene, Computational Method, Insilico gene identification, Gene prediction tools, Genetic disorders

1. Introduction

Identification of disease-causing genes and their position in genomic regions is an intricate process [1] which required time and high laboratory expenditure's. The frequent strategy used to elucidate such target genes highly depend on a technique known as" linkage analysis [2]. In this method genome scan is performed, associated with genetic marker to address the position of susceptible genes [3]. It is also used to map mendalian traits (monogenic) [2] and testing for the co-segregation of alleles for genetic markers [4].

Traditional approaches to find susceptible loci (position of a gene) and disease- causing genes in these regions are expensive, laborious and time consuming process. The presence of hundreds to thousands candidate genes in a single locus and the possibility of each gene as a potential target is the main obstacle to find out the culprit gene [3]. In this connection, it is painstaking to consider each candidate gene for experimental validations, without the probability of its involvement in disease. Refinement of locus could possibly reduce the boundary of mapped regions , however the end results may comprised of high number of candidate genes [3,5]

Moreover, already identified disease- causing genes are small in number as compare to discovered mapped loci for several genetic disorders [6]. One possible reason is the high cost associated with refining of mapped regions [3]. In this scenario, if the biological information, that correlate susceptible genes to different phenotype of interest may exclude from consideration, then the ratio of known disease causing genes over identified mapped loci will further decrease [6]. Geneticists are now utilizing computational strategies to perform efficient and quick analysis. These *insilico* approaches prioritized the candidate genes based on different criteria such as gene ontology, protein-protein interaction, sequence features, literature analysis or by high throughput data mining techniques retrieved from diverse public databases [7] .Computational methods show promising results in some cases, in contrast to blind approach used in the linkage analysis

Therefore, computer-aided techniques in combination with experimental validations may serve as a powerful mechanism to ranked the target genes [8] Genes that has direct or indirect biological association with in a particular phenotype of interest is known as candidate genes [9]

Human genome sequencing provides a gateway for the development of sophisticated computational approaches [10], which are mainly focused on data annotation and integration [11]. The fundamental concept of computerized candidate gene predictions state that similar phenotypes are related to the genes with same or associated functions [12]. Complex statistical methods are applied to rule out the potential candidate genes. This step rule out hundreds of candidate genes with maximum probability of involvement in disease phenotype and then short listed the biologically associated candidate genes within respective genetic disorder [13].

The most commonly used data sources for this purpose are semi-automated system, genotype-phenotype mapping methods, bipartite distribution of disease with non disease approach and hybrid techniques [14]. In addition to phenotypic data, protein–protein interactions and ontology data are also incorporated in some tools [15].

Currently three main approaches used computational method. Ontology , Computational and Integrated approach [16]. Ontology method is principally stands on the bioinformatics analyses for positional candidate gene prioritization [17]. Computation strategies comes under multiple web-resources methods to ranked the target gene involved in disease phenotype and utilized statistical algorithms include data-mining analysis [18] ,hidden Markov analysis [19] , cluster analysis (similarity-based method) [20], kernel-based data fusion analysis [21], machine learning [22] and K-nearest neighbor (KNN) classification algorithm [23].

A hybrid technique consists of multiple digital resources or integration of most applicable data is employed to prioritize the diseased genes. This candidate identification approach cover experimental validations and database resources which may comprised of literature and gene ontology [24]. In addition it uses gene structure variation, DNA-protein interactions, protein-protein interactions (interactome), and signaling pathways [25]. Genes involved in pathway and gene ontology are identified by a combined analysis of integrated approach [26] , text- and data-mining integrated method [27], and genetic maps and QTL combined analysis [28].

2. Bioinformatics Databases and Gene Identification

Comprehensive bioinformatics databases e.g. UCSC Genome browser [29], Ensembl Genome browser [30], National center for Biotechnology information (NCBI) [31], Gene Cards [32], Human protein reference database (HPRD) [33] , BioGPS [34] and Online Mendelian inheritance in Man (OMIM) [35] provides useful on mapped loci and diseases causing genes.

2.1 Map Viewer: Retrieval of Genes on Mapped locus

Map viewer is an application of NCBI database and helpful to identify the genes present in mapped region of interest. List of cloned identified genes, predicted transcripts, pseudo-genes and hypothetical proteins present can be easily extracted from Map Viewer (Homo sapiens Genome Build 37.1) [31].

2.2 Gene Cards: The Human Gene Database

GeneCards is an integrated database that provides information on all annotated and predicted human genes with domain, expression, functions, orthologs , paralogs, localization and pathways information [32].

2.3 Human Protein Reference Database (HPRD)

Human Protein Reference Database consist of human proteome information including protein domain, post-translation modification, protein- protein interactions and disease association for query protein [33]

2.4 BioGene Portal (BioGPS)

BioGPS is free gene annotation portal which gives complete information of molecular function, biological process and cellular location of query gene with related protein [34]

2.5 Online Mendelian Inheritance in Man: (OMIM)

Online inheritance in man database consists of a detail collection of known genes present in disease etiology, or catalogues all the known diseases with a genetic component. Updating of database is done on a regular basis. OMIM can be used to search the reported gene in the region of interest to develop a genotype-phenotype relation for concerned disease phenotype and then select the reported gene which has (i) role in disease (ii) clinical presentation related to the disease [35]

3. Diseased Gene Prediction Servers

3.1 PROSPECTR

PRiOrization by Sequence & Phylogenetic Extent of CandidaTe Regions (PROSPECTR) is free online available server that target sequence based approach on parameters like gene and protein length for sequence homology to discriminate the genes that may involve in disease with the genes with least chance of involvement on mapped loci. The input consist of the marker ID, base pair location, and cytogenetic band to extract the list of genes present at the mapped locus, which in turn gives a prioritized list of candidate genes. The results comprised of (i) classification of gene being involved in disease phenotype or not (ii) Scoring criteria and (iii) contribution of the factor in scoring diseases. The example of candidate genes successfully identified by PROSPECTRE includes ABCA2, (Alzheimer's), COL4A1 and CO4A2 (osteoporosis), SLITs (Schizophrenia) [Source: (http://www.genetics.med.ed.ac.uk/prospectr/) [22].

3.2 Suspects

Suspects (http://www.genetics.med.ed.ac.uk/suspects/) deals with the genes appeared in particular trait and most probably share the same pathways , therefore a strong hint that these genes may likely share same domains, annotation and patterns of expression. Prioritized mapped genes can be retrieve with parameters like (i) sequence features (include BLAST to search protein coding regions) and (ii) gene ontology (iii) Interpro domains shared and (iv) gene expression profile. Online inheritance in man (OMIM), Human genome mutation database (HGMD) and Genetic Association database (GAD) are utilized to get information about disease gene and finally scored according to correlation of matching profiles. A list display which shows all the candidate genes.. Examples of candidate genes analyzed by suspect includes TLR3 on chromosome 4, a candidate gene in premature degenerative osteoarthropathy and COL11AB on chromosome 11 in deafness [36].

3.3 Endeavour

The Endeavour candidate gene identification web server (http://www.esat.kuleuven.be/endeavour) relies on three stages of training, scoring and fusion. Firstly, all reported genes are retrieve for target diseases ,and a set of genes present at linked locus. After that, scoring of each gene is perform by utilizing different parameters like gene annotations, protein interactions, sequence data, literature to build model set. Model sets generate and scored each candidate genes. Finally, global ranking represent the target genes [37] Selection of YPEL1 for wet lab experiments eventually confirmed the phenotypic association in zebra fish. [38] Five susceptible loci have been prioritized to disclose molecular link between disorders obesity and Type

II diabetes [39].

3.4 Gene Wanderer

The Gene Wanderer is a protein – protein interaction (PPI) server.(http://compbio.charite.de/genewanderer/GeneWanderer). The prioritization of target genes are based on their presence in particular disease or phenotype. Gene Wanderer combines two sources of information for candidate gene predictions. First, a list of known disease causing genes are required. The second information is chromosomal location (linkage interval) . Then gene hunting begins with any listed parameters such as (i) random walk, (ii) shortest path, (iii) diffusion kernel and (iv) interaction. For example, the protein interaction network used direct interaction and shortest path methods to identify the genes PSMB8 and PSMB9 as potential disease genes associated with bare lymphocyte syndrome type 1 [40].

3.5 Gene to Disease G2D

G2D, an online available web resource server (http://www.ogic.ca/projects/g2d_2/) for identification of causative genes . G2D take input of chromosomal region, DNA markers and base pair location. A data mining algorithm (fuzzy binary relations) is used to locate the disease gene. The candidate gene analysis strategy is the interaction of reported genes with the genes present in the chromosomal region on the basis of (i) phenotype, (ii) similarity to an already known gene and interaction with protein for inherited disorder. The scoring system relies on GO based on similarity (Relative scoring criteria) and R-Score. Reported examples based on G2D analysis is the ADAM9 gene present on chromosome 20 as a susceptible gene for asthma and the Amyloid beta A4 precursor protein binding as a candidate in Alzheimer's disease [18, 41].

3.6 Prioritizer (Positional Candidates Gene Prioritization)

A standalone software, Prioritizer stands on the principle of SNPs detection by using functional human gene network. More than 70,000 predicted protein-protein interaction data is accessible to evaluate the disease causing gene by assuming the fact that limiting number of molecular pathways that contribute to disease etiology. Based on the fact that most genes are functionally related to each other, Prioritizer identifies the most plausible candidate gene in diseases phenotype.(Sources:http://genenetwork.nl/prioritizer/download.php) [42].

4. Discussion

A hybrid approach of genetic mapping with computational / bioinformatics analysis is an important step to identify the most plausible candidate genes associated with genetic disorder [8, 5] .The large number of candidate genes present in loci and the traditional methods to evaluate each, one by one is a complicated task. To expedite the results, *in silico* techniques makes it easy to hunt the susceptible genes on mapped loci . Performance evaluation of (specificity and sensitivity) automated candidate gene prediction tools are complicate task because each individual group has its own set of control data [14]. Sequences and function based approaches are widely used to prioritize candidate genes. Many techniques employed both approaches. In first scenario (sequence based approach) , different parameters were used i-e molecular function, cellular location, biological process, comparison with identify disease and association of disease literature in genome sequences, genotype-phenotype relation [5]. The second scenario take in to account the structural parameters like protein size, conservation and function [43]. Although usage of computational approaches towards disease gene identification is quite fascinating but a major concern about these prioritization methods are (i) expenditure and attempt in describing novel disease gene, most likely when comparing with disease loci which are not under consideration. As a result, execution of mostly algorithms used known disease genes data set [40] (ii) Gene ontology data is under development, and required exploration of hidden genome information (iii) Structural information is more reliable as compared to annotation data but gene specific information is not fully complied [44] (iv) Commonly, sequences analysis techniques have border application with no constraint [9] whereas function based methods has limitations when applying on novel diseases genes without experimentally proved data [45]. Availability of more protein- protein interaction data can managed this drawback and may provide help in accurate prediction of disease causing genes based on both direct and in-direction interactions.

Multiple techniques based on different data information for candidate gene analysis can be more favorable as compared to used same input data adopted by several methods. By considering this fact prediction of candidate gene can more accurate for mapped loci [5].

5. Conclusion

Current review suggests the integration of biological methods with mathematical, statistical and computational techniques to uncover the biological mystery of disease gene identification. This approach may provides an efficient analysis of candidate genes , present on mapped loci and may serve as a gateway to treat inherited disorders effectively.

6. Acknowledgement

I am grateful to Ahmad Ali Ansari for his valuable comments to improve the review article.

References

[1] Huang QY, Li GHY, Cheung WMW, Song YQ, and Kung AWC (2008). Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. J Hum Genet 53:644–655
[2] Baron M (2001) .Millennium article The search for complex disease genes: fault by linkage or fault by association?.Mol phys 6: 143–149

[3] Wayne ML and McIntyre LM (2002). Combining mapping and arraying: An approach to candidate gene identification. PNAS 99: 14903-6
[4] Lidral AC and Murray JC (2004). Genetic approaches to identify Disease gene for birth Defects with cleft lip Palate as a model. Clinic and mol teratology 7: 893-901
[5] Zhu M and Zhao S (2007). Review on Candidate gene identification Approach: Progress and challenges. Int J Biol Sci 7: 420-427
[6] Nguyen TH and Ho TB (2007). IEEE International Conference on Bioinformatics and Biomedicine A Semi-Supervised Learning Approach to Disease Gene: Nucleic Acid Res. 36:377-384
[7] Lou XY, Ma JZ, Yang MCK (2006). Improvement of mapping accuracy by unifying linkage and association analysis. Genetics 172: 647-661
[8] Mackay, T. F. C. (2001) Annu Rev Genet 35: 303–339
[9] Tabor HK, Risch NJ, Myers RM (2009). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3: 391-397
[10] Glazier AM, Nadeau JH, Aitman TJ (2002). Finding genes that underlie complex traits. Science 298: 2345-2349
[11] Zhang P, Zhang J, Sheng H (2006). Gene functional similarity search tool (GFSST). BMC Bioinformatics 7: 135
[12] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007). The human disease network. Proc Natl Acad Sci 104:8685–8690
[13] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA (2002). Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55
[14] Teber ET, Liu1 JY, Ballouz S, Fatkin D, Wouters MA (2009). Research on Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. BMC Bioinformatics 10: S1-S69
[15] Ma X, Lee H, Wang L, Sun F. CGI (2007). A new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23:215–221
[16] Bansal AK (2005). Review on Bioinformatics in microbial biotechnology – a mini review. Microb Cell Fact 28: 4-19
[17] Harhay GP, Keele JW (2003). Positional candidate gene selection from livestock EST databases using Gene Ontology. Bioinformatics 19: 249-255
[18] Perez-Iratxeta C, Bork P, Andrade MA (2002). Association of genes to genetically inherited diseases using data mining. Nature Genet 31: 316-319
[19] Pellegrini-Calace M, Tramontano A (2006). Identification of a novel putative mitogen-activated kinase cascade on human chromosome 21 by computational approaches. Bioinformatics 22:775-778
[20] Freudenberg J, Propping P (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18:110-115
[21] De Bie T, Tranchevent LC, van Oeffelen LM (2007). Kernel-based data fusion for gene prioritization. Bioinformatics 23:125-132
[22] Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2005). Research article speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6:55. doi:10.1186/1471-2105-6-55
[23] Xu J, Li Y (2006). Discovering disease-genes by topological features inhuman protein-protein interaction network. Bioinformatics 22:2800-2805
[24] Hristovski D, Peterlin B, Mitchell JA (2005). Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74: 289-298.
[25] Sugaya N, Ikeda K, Tashiro T (2007). An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacol 20 :7-10
[26] Feng Z, Davis DP, Sásik R (2007). Pathway and gene ontology based analysis of gene expression in a rat model of cerebral ischemic tolerance. Brain Res 1177: 103-23
[27] Tiffin N, Kelso JF, Powell AR (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33: 1544-1552
[28] Arcade A, Labourdette A, Falque M (2004). BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. Bioinformatics 20: 2324-2326
[29] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun ;12(6):996-1006
[30] Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva at al. Ensembl 2015 .Nucleic Acids Res, 43 Database issue: D662-D669
[31] Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L: Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004, 32:D35-D40
[32] Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M. and Lancet, D. PathCards: multi-source consolidation of human biological pathways, Database (2015); doi:10.1093/database/bav006
[33] Prasad, T. S. K. et al. (2009) Human Protein Reference Database - 2009 Update. Nucleic Acids Res. 37, D767-72
[34] Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol.10(11):R130.
[35] McKusick V.A. (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn. Johns Hopkins University Press, Baltimore, MD.
[36] Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006). SUSPECTS: enabling fast and effective prioritization of positional candidates. BMC Bioinformatics 22 (6): 773–774
[37] Tranchevent L., Barriot R., Yu S., Van Vooren S., Van Loo P., Coessens B., Aerts S., De Moor B., Moreau

Y.Nucleic Acids Research, Web Server issue, vol. 36, no. 1, Jun. 2008, pp. 377-384.
[38] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B (2006). Gene prioritization through genomic data fusion. Nat Biotechnol 24:537–544
[39] Elbers C, Onland-Moret C, Franke L, Niehoff A, van der Schouw Y, Wijmenga C (2007). A strategy to search for common obesity and type 2 diabetes genes. Trends Endocrinol Metab 18:19–26
[40] Köhler S, Bauer S, Horn D, Robinson PN (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. The Am J Hum Genet 82: 949–958
[41] Perez-Iratxeta C, Wjst M, Bork P and Andrade AM. G2D: a tool for mining genes associated with disease. 2005. BMC Genetics 2005, 6:45 doi:10.1186/1471-2156-6-45
[42] Prioritizer : http://genenetwork.nl/prioritizer/download.php
[43] López-Bigas N, Ouzounis CA (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 10:3108-14
[44] Rossi S, Masotti D, Nardini C (2006). TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34 : 285–292
[45] Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006). Reconstruction of a functional human gene network with an application for prioritizing positional candidate gene. Am J Hum Genet. 6:1011-25