Disease-Gene Association: Tools, Techniques and Trends

Shital Patil
Department of Computer Engineering and Information Technology
College of Engineering, Pune
Email:

Satish Kumbhar
Department of Computer Engineering and Information Technology
College of Engineering, Pune
Email:

ABSTRACT
Genetic make-up of an individual is responsible for expression of external characters. Genes express through creation of intermediate products such as amino acids and proteins in turn. Proteins structurally and functionally are responsible for causing phenotype change. Some properties like mutations in gene may cause abnormalities. This makes it necessary to relate particular gene with the diseases it caused. Essence of gene and gene product information has proven its role in many aspects of life and disease-gene association is closely concerned with it. It has got the paramount importance in many fields like genetic engineering, forensics and clinical diagnosis. The availability of essential experimental data from various genetic and protein related databases aided by state of the art computing technology has brought unforeseen revolution in the field of bioinformatics. Efforts on human genome project and GWAS (Genome Wide Association Studies) have brought radical development in sequencing and assembly. Post genomic era demands processing and usage of this valuable data in real practice to address real life scenarios. Various data mining and statistical approaches exist to address disease-gene association problem. We describe the analysis of such tools and techniques based on parameters like working principles, algorithms or methods used, speciality and limitations etc..

Keywords
Phenotype, genetic engineering, forensics, GWAS.

1. INTRODUCTION
Data deposited from various experiments being conducted though biological laboratories, research centers and academic institutions have contributed a lot in establishing relationship between related biological entities [1]. With evolution in biotechnology and computational biology the parallel techniques came into existence. These techniques applied statistical and computational methods to extract out the useful patterns in data. These patterns reveal certain useful inferences to be applied in real practice.

Recent standardizations and refinements have encouraged data mining methods in this area. Ample scope exists to mine and extract knowledge out of gene products like amino acid, protein and protein- protein interaction information. Few of the currently existing well known databases are Ensemble, Speedier, Unipart, Swiss- port, Genbank, PDB. Metabolic pathway databases like KEGG and Biocyc. On the other hand there is also huge deposition of clinical practice and disease description data. Examples of these databases are OMIM (Online Mendelian Inheritance in Man) [2] and LocusLink [3]. OMIM is database of disorders attributed to particular gene or group of genes associated to specific mendelian disorder. This database provides physicians and genetic researchers with comprehensive, updated and authoritative information about genetic disorders. Efforts being made to bridge the gap between genotype and phenotype data have got crucial importance.

Relationship between genes in genetic databases during expression, interaction between gene products like amino acids and proteins and finally the phenotype character they lead to is undoubtedly worth considering. The expressed character may be abnormal e.g. mutation in certain gene can lead to uncontrolled cell division and eventually become cause for cancer. In this case seemingly small mutation can disturb entire process of regulation and protein synthesis.

Now the questions arise- Can this mutation which occurred be mapped to corresponding disorder (cancer here)? After analyzing genetic history, is it possible to predict that the person is vulnerable to certain hereditary disease and what precautions he/she should take? Can appropriate drugs or therapies to control gene function and level of their expression be developed? Of course many biomedical laboratories have put toiling endeavors and come up with positive answers to these questions.

Here comes the importance of computational methods applied to scattered biological data. If one wants to do thorough research on particular disease and events that caused it, he/she has to consider meticulous details about each factor involved in gene expression to protein synthesis. Many biological experiments have already deposited data about each involved factor of the process. Thus there is a necessity to integrate information about these factors and computationally analyze them to infer useful patterns. This generates need for developing systems which utilize complementary available data from these databases and produce results which are of crucial use in clinical diagnosis, Medical healthcare and pharmaceutical drug development. This motivated the idea of finding and attributing genes which are responsible for causing certain disorder

Basic method to computationally determine responsible genes for causing disorder is called *candidate gene prioritization*. Gene description data used by some of these methods are annotation, protein-protein interactions and sequence features. Many of these systems use GO (gene ontology) terms [4]. Gene ontology offers the unified representation to gene and gene product related terms within all species. It creates and maintains controlled vocabulary of gene and gene product data. Here we cover the analysis of these tools and methods based on their working principles, resource data used and specialities of these systems.

2. PREVALENT TOOLS USING PUBLIC DATA RESOURCES

Contemporary tools that try to cater to the problem of disease gene association take into account various factors like linkage analysis, mutation analysis, similarity measures etc.

2.1 GeneSeeker

This application utilizes information from various databases containing chromosomal locations, gene expression and phenotypes data [5].

Principle: It assumes the possibility that responsible gene will show better expression in tissues suffering a particular disorder. Notable thing with this system is the use of related data from other species like mouse which helps identify location and function of similar human gene considering analogy [6].

Resources: This tool utilizes well known datasets by categorizing them in three distinctive groups. First group involves localization databases in human such as MIMMAP and GDB etc. MIMMAP is restructured version of OMIM.OXFORD grid is used to translate human to mouse mapping of locations on chromosome. The SRS (Sequence Retrieval Systems) are used which perform range based search on chromosomes. The second group involves mouse genome related databases such as MGD. Group three involves expression and phenotype databases e.g. PubMed nature library of medicine, OMIM and UniPort etc. UniPort helps to solve the problem of naming inconsistencies by providing synonym information for different terminologies used by different databases to refer to same entities to a certain extent.

Algorithm: Working flow of GeneSeeker is shown in Figure 1.

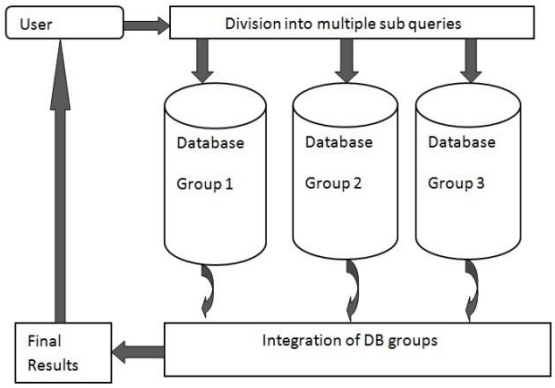

Figure 1: Working flow of GeneSeeker [5].

- Cytogenic localization is processed through first database group.
- Oxford-grid is used by second database group to find the similar region on mouse chromosome.
- Tissues of interest and phenotypes of disease can be specified by user which is split and processed by third database group.
- Finally gene lists obtained from three database groups are put together based on boolean logic specified in user query and results are displayed.

Speciality: GeneSeeker is efficient in identifying disease genes especially related to the syndromes where affected tissues show changed expression patterns.

Limitations: Operation of GeneSeeker depends on gene nomenclature which is different across different databases this problem needs to be handled effectively.

2.2 TOM

TOM is a web based integrator. It avoids time consuming process of linkage analysis of families, detection of gene transmission path and confirmation of location of certain disease causing gene.

Principle: TOM considers the knowledge of another disease causing gene in same chromosomal area or genetic intervals of genes related to the disease in query [7]. The algorithm in one locus case takes query genes as input and checks the combination of functionally related genes in the specified linkage region of interest. The next two loci option is used when there is not enough of pre-existing information about the genes related to relevant disease in consideration.

Resources: This system utilizes data related to gene characteristics such as genetic mapping, expression profiling and function related annotations from GO (Gene Ontology) [8].

The combination of gene characteristics helps collect genes with same description and exclude those that differ in functionality. This uses probabilistic approaches for selection which specify numeric space and rejection criteria for genes. Functional roles are distinctively identified using hypergeometric distribution.

Algorithm: The three step filtering algorithm involves following steps.

- Using genome sequence information, list of genes mapped to particular chromosomal region is obtained
- Transcriptome data from public repositories is used. Only those genes in list are retained which have related expression variation in datasets. It may be among genes themselves (Two Loci) or with seeds (One Locus). This is done by defining expression neighborhood. Selection criteria involve definition of genome regions and transcriptional profile distances. The protein involved in cellular processes is checked against the seed genes and functionally unrelated candidates are discarded out of selection process.
- Third and final step is optional which filters candidate genes based on their functional roles. GO helps in understanding of molecular functions of biological process related to genes by providing enriched vocabulary.

Speciality: Power of this tool resides in constantly utilized updated gene expression data from microarray experiments. ***Limitations:*** The tool lacks support from pathway databases in case of poor description of disease related data.

2.3 DGP

There is a classical tool named DGP which is based on the observed results and patterns of protein structure [9].

Principle: Working principle is based on observed results and patterns of protein structure. Long proteins that have highly conserved amino acid sequences are more prone to mutations which cause abnormalities while short proteins with many paralogues are less vulnerable to disease causing mutations

Resources: It uses OMIM, Locus Link, Swiss Port, Tremble

Long proteins that have highly conserved amino acid sequences are more prone to mutations which cause abnormalities while proteins with paralogues are less vulnerable to disease causing mutations. From the results shown in Figure 2 it is clear that as the protein length increases, disease prone proteins outnumber healthy human proteins.

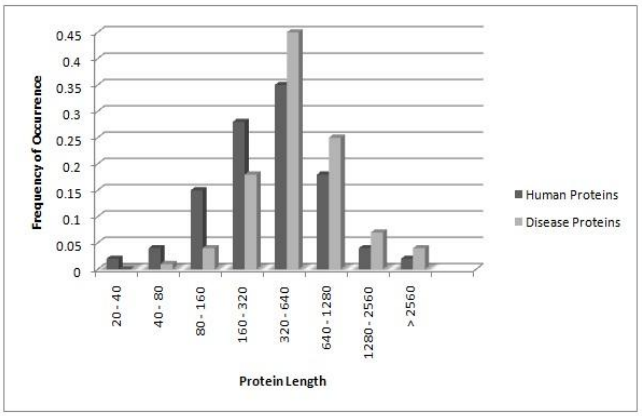

Figure 2: Protein length as number of amino acids with distribution of human proteins and disease proteins [9].

3. TOOLS UTILIZING SIMILARITY MEASURES

Many annotation based systems do work well provided the annotations are correct and complete. If annotations do not show these features or are inconsistent, these systems may produce irrelevant results. In this case use of sequence feature based methods will be more effective and reliable.

PROSPECTR is one of the efficient tools which utilize sequence based features for selecting disease genes. [10].

3.1 PROSPRECTR

It is a simple and easy way to know about the genes involved in mendelian and oligogenic diseases.

Principle: It is experimentally proven that the genes responsible for hereditary diseases have some distinctive characters such as larger gene size [11].The approach is based on targeting such features and select those genes which are more likely to be involved in disease and eliminating others using decision tree model.

Resources: Data resources used involve OMIM, Ensembl etc. Preprocessing: For creating training set and building feature definition comparative study of genes from two sets was done. First set involved few known genes from Ensemble not known in priori to have relation with any disease. Second set involved genes from Ensembl involved in disease and listed in OMIM. Using Mann-Whitney test certain distinctive characteristics were discovered between genes from both of these sets.

- There was significant difference in protein and cDNA sizes of two sets.
- Genes present in OMIM list were larger in size and encoded longer proteins.
- Genes listed in OMIM had better conserved reciprocal hit homologues with other species like mice.
- Percentage of gene product secreted was higher in genes listed in OMIM.
- Number of exon was also a crucial parameter in examination.
- Genes listed in OMIM were found to be expressed more in specific tissues.
- The significant differences were found in distance to neighboring genes.

Algorithm: The implementation involves use of alternating decision tree algorithm. This algorithm is efficient because it offers high degree of accuracy with small number of rules specification [12]. Weka was used as the platform for machine learning experiments. Trivial example of decision tree is shown in Figure 3.

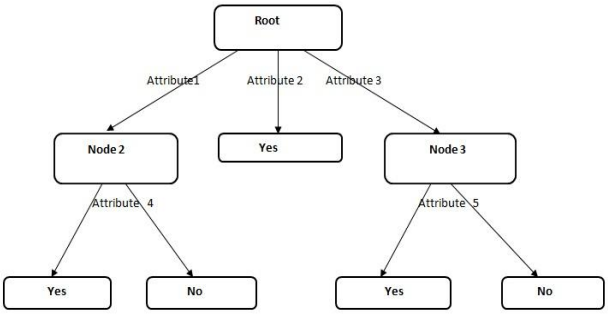

Figure 3: Example of decision tree.

PROSPECTR algorithm is as follows.

- The rules of specification are derived from set of control and disease genes i.e. genes from first and second group in preprocessing step.
- The rules are iteratively added to the decision tree with descending order of their predictive power. This means, the rules with strong predictive power are added first and then those with less predictive power are included. Rule is added as entirely new node or as child of any previously existing node.
- The gene is classified starting right from the root and following particular rule the branch to next node is followed.
- As an instance, consider gene length as rule based parameter, if it exceeds certain value or not is tested. If yes one branch is followed otherwise another branch is followed. If value concerning any parameter is unknown then no branch is followed.
- The classification is based on classification scores. Negative classification score associated with some gene makes it more likely to be involved in disease as opposed to positively scored gene which is less likely to be responsible for causing disease.

Observation about algorithm: The observation revealed that there should be moderate number of nodes present in decision tree. Less number of nodes leads to lose the predictive power

because they have weak discriminatory power and can not reach confident level of classification. On the other hand decision tree containing large number of nodes show decreased decision strength due to addition of less predictive capacity nodes (rules) to subsequent levels.

Speciality: It works well even in cases where input data is new and hard to classify.

Limitations: Decisions are based on sequence related properties but it should offer good support for annotation based analysis as well.

3.2 G2D

Principle: Basic assumption made by this technique is that, if there is a disease D1 with associated gene G1. There is now another disease D2 in query having same phenotypic characteristics as that of D1 with undiscovered associated gene G2. In this case some functions of G1 and G2 are co-related and relevant to those phenotypes [13].

Resources: The databases like MEDLINE contain information about the references about medical literature annotated using MeSH ontology. It also uses OMIM, GO, RefSeq database.

Algorithm: Working of G2D is as shown in Figure 4.

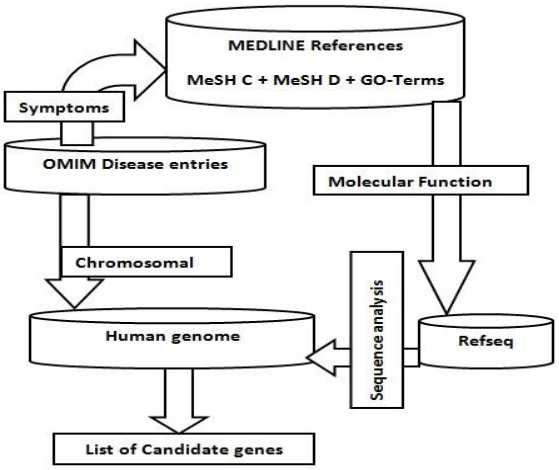

Figure 4: G2D algorithm [13].

- The algorithm takes MeSH C (disease terms) from OMIM as input. These terms are then searched for corresponding MeSH D terms (Chemicals and Drugs).
- Then the reference terms from it are extracted. The terms from OMIM which have more MEDLINE links get more weight values than those having lesser references.
- The functional annotations of proteins are taken from RefSeq database [3] considering related GO terms. Using RefSeq database, BLAST homology search is performed with genomic region where disease is mapped to.
- All hits in that genomic region are recorded, stored and sorted. If recorded hits show values less than or equal to specified E value, the genes found are then declared to be candidate genes.

Speciality: As this approach is based on sequence features as well as functional annotation information it is noteworthy to conclude that results for queries for certain disease genes get refined over time with increased accuracy of human genome as well as frequent efforts to annotate human sequences and their homologues in different organisms like mouse. Method has ability to point pseudogenes for human examination. The system is updated to include three algorithms as per specific requirements [14]. These techniques involve phenotype method, known genes method and protein- protein interaction method as extension.

4. GFINDer

Collecting, integrating and analyzing information from regularly updating web based data resources helps refine existing results and explore new knowledge. This functionality is offered by systems like GFINDer (Genome Function Integrated Discover) [15].

Principle: The system dynamically aggregates functional and phenotype related annotations and allows analysis and mining of user selected gene list.

Resources: GO, Pam, normalized OMIM and Interpro etc.

The system has three tier architecture namely data tier, processing tier and user tier which perform their obvious roles efficiently. It is shown in Figure 5.

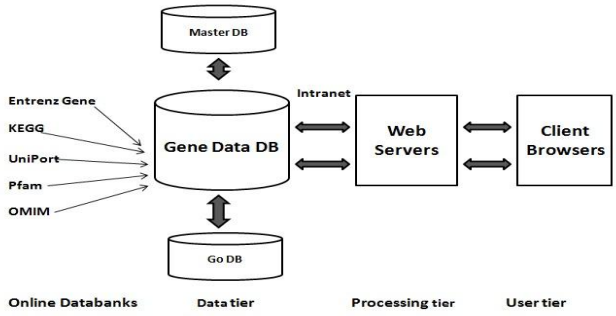

Figure 5: GFINDer System..

Statistical method used: Considering total of N genes out of which M unknown genes belong to annotation category A. That means rest N- M are outliers. If randomly subset K of these N genes is selected and assigned to the class at least x of these selected genes will relate to category A. Probability of this is calculated by hypergeometric distribution with parameters (N,M,K) [16].Alternative statistical test uses chi-square distribution to expedite the process of computation .***Speciality:*** Performs simultaneous gene comparisons and multilevel GO based evaluations. System is well tolerant to increasing overload over multiple queries. ***Limitation:*** Biochemical pathway and gene expression data is worth adding.

5. INTERACTION NETWORK BASED METHOD

Recently few methods have emerged which utilize network based approaches in graph theory to perform candidate gene prioritization. The recent work on these techniques provides insight into this approach..

5.1 Phenome-Interactome Network

Resources: This algorithm uses data from OMIM and Human Protein Reference Database (HPRD) [17].

Algorithm:

- First the phenotype similarity profile is represented as matrix and disease phenotypes are obtained from literature.
- Small similarity create noise, so in order to eliminate these components threshold is set and values below this threshold are ignored. Thus weighted phenotype similarity network is obtained. In this network vertices are disease phenotypes and edges incident on these vertices represent relationship between them. This network is termed as phenome [18].
- Protein- protein interaction network is obtained from HPRD. This network is referred as interactome. Biomart [19] tool is used to obtain associated genes in interactome and related disease phenotypes in phenome.
- With this data the heterogeneous network is created which is called phenome-Interactome network with already known interactions and associations. Figure 6 shows one such network. Each edge in this network is assigned numeric capacity value which represents confidence on the relationship between two vertices connected by that edge. Each edge has positive capacity value c (u, v).
- To associate gene with disease in query from candidate gene list, the maximum information method is used.

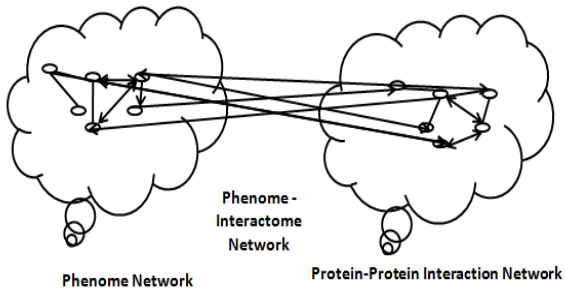

Figure 6: Phenome-Interactome network [18]..

Speciality: Better utilization of priorly known associations between genotype and phenotype to predict new candidate genes which may be unknown to have related to disease previously.

6. SUMMARY OF THE SYSTEMS

Summary of depicted systems is presented in nutshell in Table 1.This primarily includes working principle and data resources used.

Table 1. Feature Summary of Systems

System name	Principle/Feature	Datasets/Resources used
GeneSeeker	Efficient in identifying disease genes related to the syndromes where affected tissues show changed expression patterns. Use of mouse genome utilizing human genome analogy.	OXFORD, MIMMAP, GDB, MIMMAP, OMIM SRS(Sequence Retrieval Systems), PubMed and UniPort
TOM	Avoids long tedious process of linkage analysis. Refers another disease existing in same genomic region.	Gene characteristics and mapping, expression profiling and functional annotations. GO terms.
DGP	Based on observed results and patterns of proteins. Long proteins are more vulnerable to disease as compared to small.	OMIM, LocusLink, SwissPort TrEmbl
PROSPECTR	This utilizes sequence based features for selecting disease genes. Decision tree based model is used to specify rule based selection criteria.	Ensembl, OMIM
G2D	Considers assumption-disease D1 priorly associated with gene G1 helps to associate disease D2 with same phenotypes as D1 to unknown gene G2.	MEDLINE references (MeSH C, MeSH D.), GO, RefSeq.
GFINDer	Web based comprehensive integrator	GO, Pfam, normalized OMIM, InterPro
Phenome-interactome network	Network based method constructs disease phenotype relation network and protein-protein interaction network. Uses maximum flow method to find relations in this heterogeneous network	OMIM, Human Protein Reference Database (HPRD), Biomart tool.

7. CONCLUSION

Observing all the systems depicted, it is vivid to see that each system has some unique working principle. Data resources and algorithms used depend on the needs suitable for chosen method. Despite having these differences, these techniques contribute their major to the area of disease gene association. The results from these systems provide valuable aid in bioinformatics research, genomic medicine, disease diagnosis and clinical healthcare in case of hereditary or gene disorder related abnormalities.

During the research throughout the century numbers of classically Mendelian disorders have been successfully related to the disease causing gene, but relatively small numbers of genes related to complex diseases have been characterized. With continuous increase and refinement in genotype as well as phenotype data high opportunities to achieve more precise results even in complex diseases will arise. Support of clinical

practice data, biochemical reactions impacts is crucial. Recent breakthrough created by microarray technology and its applications has encouraged computational knowledge extraction. There is a need for concentrating on processing of genotype and phenotype data as integrated component and generate more precise and reliable methods.

8. REFERENCES

[1] M.Y.Galperin, "The molecular biology database collection: 2006 update," Nucleic Acids Research, vol. 34, no. Database-Issue, 2006Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, andV. A. McKusick, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," Nucleic AcidsResearch, vol. 33, no. Database-Issue, pp. 514–517, 2005. <Online>.Available: http://dx.doi.org/10.1093/nar/gki033

[3] K. D. Pruitt and D. R. Maglott, "Refseq and locuslink: NCBI gene centered resources," Nucleic Acids Research, vol. 29, no. 1, pp. 137–140, 2001. <Online>. Available http://dx.doi.org/10.1093/nar/29.1.137

[4] M. A. Harris and et al (!), "The gene ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, pp. D258–D261, Jan.2004

[5] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, H. G. Brunner, and G. Vriend, "Geneseeker: extraction and integration of human disease-related information from web-based genetic databases," Nucleic Acids Research, vol. 33, no. Web-Server-Issue, pp. 758–761, 2005. <Online>. Available: http://dx.doi.org/10.1093/nar/gki435

[6] H. Wang, H. Zheng, D. Simpson, and F. Azuaje, "Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data," Mar. 08 2006. <Online>. Available: http://www.biomedcentral.com/1471-2105/7/116

[7] S. Rossi, D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, Benini, and S. Volinia, "TOM: a web-based integrated approach for identification of candidate disease genes," Nucleic Acids Research,vol. 34, no. Web-Server-Issue, pp. 285–292, 2006. <Online>. Available: http://dx.doi.org/10.1093/nar/gkl340

[8] M. A. Harris and et al (!), "The gene ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, pp. D258–D261, Jan. 2004.

[9] N. ria LoA pez Bigas and C. A, "Genome-wide identification of genes likely to be involved in human genetic disease."

[10] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization," BMC Bioinformatics, vol. 6, p. 55, 2005. <Online>.Available: http://dx.doi.org/10.1186/1471-2105-6-55

[11] E.-W. A. Smith NGC, ": Human disease genes: patterns and predictions."2003.

[12] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in Proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, 1999, pp. 124–133.

[13] C. Perez-Iratxeta, M. Wjst, P. Bork, and M. A. Andrade, "G2D: a tool for mining genes associated with disease." <Online>.Available:http://www.pubmedcentral.gov/articlerender.fcgi?artid=1208881

[14] C. Perez-Iratxeta, P. Bork, and M. A. Andrade-Navarro, "Update of the G2D tool for prioritization of gene candidates to inherited diseases, "Nucleic Acids Research, vol. 35, no. Web-Server-Issue, pp. 212–216,2007. <Online>. Available: http://dx.doi.org/10.1093/nar/gkm223

[15] M. Masseroli, "Management and analysis of genomic functional and phenotypic controlled annotations to support biomedical investigation and practice," IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 4, pp. 376–385, 2007. <Online>.Available:http://dx.doi.org/10.1109/TITB.2006.884367

[16] G. Casella and R. L. Berger, "Statistical inference, 2nd ed. belmont,." Belmont, CA:Duxbury Press, 2002.

[17] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B.Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N., H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey, "Development of human protein reference database as an initial platform for approaching systems biology in humans," Jun. 21 2004. <Online>.Available:http://www.pubmedcentral.gov/articlerender.fcgi?artid=403728

[18] Y. Chen, T. Jiang, and R. Jiang, "Uncover disease genes by maximizing information flow in the phenome-interactome network," Bioinformatics [ISMB/ECCB], vol. 27, no. 13, pp. 167–176, 2011. <Online>.Available:http://dx.doi.org/10.1093/bioinformatics/btr213

[19]] S. Damian, H. Syed, B. Benoit, H. Richard, L. Darin, T. Gudmundur, and K. Arek, "Biomart – biological queries made easy," 2009