Research in Molecular Medicine

Original Article

2015, Vol: 3, Issue: 1, Pages: 18-23

DOI: 10.7508/rmm.2015.01.004

Identifying and prioritizing genes related to Familial hypercholesterolemia QTLs using gene ontology and protein interaction networks

Ali Kazemi-Pour [1], Bahram Goliaei [2*], Hamid Pezeshk [3], Behjat Kalantari khandani [4]

[1] PhD student of Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
[2] Professor of Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
[3] Professor of statistics, Science College, University of Tehran, Tehran, Iran.
[4] Assistant Professor of Internal Medicine, Kerman University of Medical Sciences, Kerman, Iran.

Received: 17 Aug 2014

Revised : 4 Sep 2014

Accepted: 10 Sep 2014

Corresponding Authors:
Bahram Goliaei
Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
Phone: +98-2161113356
E-mail: goliaei@ibb.ut.ac.ir

Abstract

Background: Gene identification represents the first step to a better understanding of the physiological role of the underlying protein and disease pathways, which in turn serves as a starting point for developing therapeutic interventions. Familial *hypercholesterolemia* is a hereditary *metabolic* disorder characterized by high low-density lipoprotein cholesterol levels. Hypercholesterolemia is a quantitative trait that is controlled by interactions among several quantitative trait loci**.** Many biological data is presented in the context of biological networks and evaluation of biological networks is considered as the essential key to understanding complex biological systems.

Materials and Methods: In this research, we used combination of information about quantitative trait loci of hypercholesterolemia with information of gene ontology and protein–protein interaction network for identification of genes associated with hypercholesterolemia.

Results: For hypercholesterolemia disease, we introduced 16 new genes which were in quantitative trait loci regions and were associated with the hypercholesterolemia disease in terms of gene ontology characteristics.

Conclusion: Combination of linkage information (QTLs) and genomics information (gene ontology and protein–protein interaction network data) is highly able to identify genes associated with diseases**.**

Keywords: Complex disease; Disease-gene prediction; Familial *hypercholesterolemia*; Protein interaction network; Quantitative trait loci

Please cite this article as: Kazemi-Pour A, Goliaei B, Pezeshk H, Kalantari khandani B. Identifying and prioritizing genes related to Familial hypercholesterolemia QTLs using gene ontology and protein interaction networks. Res Mol Med. 2015; 3 (1): 18-23.

Introduction

Complex disorders

Identification of genes associated with diseases is one of the most important research priorities in the field of health. Associating genes with diseases is a fundamental challenge in human health with applications to understanding disease mechanisms, diagnosis and therapy(1). The identification of genes involved in human hereditary diseases is often time-consuming and expensive.

Complex diseases or quantitative trait do not obey the standard Mendelian patterns of inheritance. The vast majority of diseases fall into this category, including several congenital defects and a number of adult-onset diseases. Some examples include Alzheimer's disease, scleroderma, asthma, Parkinson's disease, multiple sclerosis, osteoporosis, connective tissue diseases, kidney diseases, autoimmune diseases, and many more (2).

Familial hypercholesterolemia is a condition characterized by very high levels of cholesterol in the blood which is known to increase the risk of several adverse health effects including atherosclerosis, heart attack, and stroke (3). Hypercholesterolemia is a quantitative trait that is controlled by interactions among several quantitative trait loci (QTLs) combined with environmental influences.

Protein network and Gene Ontology

Today, bioinformatics and genomics data have given the researchers very useful tools in study of genes associated with diseases. Genes and proteins do not work independently, but are organized into co-regulated units that perform a common biological function. Complex disorders are determined by the combined effects of many loci and are affected by gene networks or biological pathways. Systems biology approaches are of great importance in the identification of candidate genes associated with complex diseases or traits at the system level (4).

Systems biology approach the analysis of the relationship between the genes and proteins as a whole, to understand the disease Phenotype. Within a cell, Proteins interact with each other, and those interactions represented by a network(5).

The most important goals of Systems biology are protein function prediction, interaction prediction identification of disease candidate genes and drug and identification of candidate genes (6-8). The protein network is one of the most frequently used type of evidence for disease gene prediction (9).

Protein network applications in medical field include identifying new disease genes, the study of their network properties, identifying disease-related subnetworks and network-based disease classification (10). Networks have been exploited to find novel candidate genes, based on the assumption that neighbors of a disease-causing gene in a network are more likely to cause either the same or a similar disease. This indicates that network neighbors of known disease genes form an important class of candidates for identifying novel genes for the same disease (11).Nowadays, the cellular biology researches, have shifted from the molecular to the modular researches and the single gene or protein study is replaced by the study of the function of protein and gene complexes. Study proteins in network to detect and prioritize disease genes are better than traditional approaches that used only protein-phenotype associations (12).

Many specific examples have shown that individual genes that cause a given diseases phenotype tend to be linked at the biological levels as components of a multi-protein complex(12).

Gene Ontology and Functional annotations including biological processes, and molecular functions are another rich sources of evidence that are frequently used for disease-gene prediction (13). Complex disease genes associated with same disease more often tend to share a protein-protein interaction (PPI) and GO biological process compared to the genes associated with different diseases (11). Genes that interact directly or indirectly may have the same or similar functions in the biological processes in which they are involved and together contribute to the related disease phenotypes (14).

Quantitative trait locus

Diseases such as diabetes, cancers, hypercholesterolemia, Alzheimer etc. have polygenic inheritance and study of quantitative trait loci (QTLs) associated with these diseases is the main step for recognition of genes of these diseases. A quantitative trait locus (QTL) is a chromosomal region that contains a gene or genes that influence a quantitative trait. Quantitative trait locus analysis is a powerful method for localizing disease genes, but identifying the causal gene remains difficult. Thus, the major obstacle in identifying QTL genes is not detection of a QTL, but rather the expensive and time-consuming process of narrowing a QTL to a few candidate genes that can be rigorously tested. Using bioinformatics techniques with the experimental methods is a powerful way to narrow a QTL interval (15-16).

In this paper, we have studied the hypercholesterolemia, disease using quantitative trait locus integrated with the gene ontology and protein-protein interaction to predict the new candidate genes.

Table 1. Specifications *QTLs associated with* hypercholesterolemia *disease (*SCL_H: Serum cholesterol level QTL human)

***QTLs* Symbol**	**chromosome**	**Position**
SCL132_H	2	(60,139,274 - 86,139,274)
SCL85_H	2	(2,227,411 - 28,227,411)
SCL86_H	2	(8,086,103 - 34,086,103)
SCL87_H	3	(1 - 24,517,367)
SCL88_H	3	(1 - 24,517,367)
SCL98_H	3	(158,750,982 - 184,750,982)
SCL129_H	6	(147,967,069 - 171,055,059)
SCL128_H	7	(140,012,742 - 156,472,993)
SCL133_H	7	(70,789,500 - 96,789,500)
SCL89_H	9	(118,527,538 - 141,149,349)
SCL96_H	9	(72,784,197 - 98,784,197)
SCL130_H	12	(12,449,930 - 53,197,213)
SCL131_H	15	(46,713,930 - 72,713,930)
SCL134_H	16	(1 - 25,139,539)
SCL92_H	19	(1 - 25,712,901)
SCL93_H	19	(17,417,129 - 43,417,129)
SCL94_H	19	(17,417,129 - 43,417,129)
SCL97_H	19	(1 - 19,113,369)
SCL126_H	20	(559,258 - 49,561,921)
SCL95_H	22	(13,015,963 - 39,015,963)

Materials and Methods

Hypercholesterolemia (FHC) or familial hypercholesterolemia is a metabolic disease in which heredity is controlled as complex or multiple genes. Information of this disease has been recorded in OMIM database and with OMIM ID: 143890. So far, seven genes including EPHX2, ABCA1, APOA2, PPP1R17 (C7orf16), LDLR, ITIH4 and GHR which affect the occurrence of this disease has been specified (17). Study of molecular markers has led to identification of different QTLs related to FHC indicating the presence of unknown genes associated with this disease. Each QTL includes many genes and it is difficult to study all genes available in QTL for identification of gene associated with disease.

Table 2. Specifications 16 genes were identified as genes associated with hypercholesterolemia disease

	Gen Symbol	Gene Position	GO term	Term
SCL85_H	APOB	21,224,301 - 21,266,945	GO:0010886 GO:0006642	positive regulation of cholesterol storage, triglyceride mobilization
SCL86_H	SNX17	27,593,363 - 27,600,400	GO:0006707	cholesterol catabolic process
SCL132_H	UGP2	63,840,950-63,891,562	GO:0006011	carbohydrate metabolic process & co-annotated with GO:0009103 lipopolysaccharide biosynthetic process
SCL87_H	SCL88_H	14,145,147-14,178,672	GO:0000075	cell cycle checkpoint & co-annotated with GO:0043550 regulation of lipid kinase activity
XPC	SCL133_H	80,602,188-80,679,277	GO:0006629	Lipid metabolic process, & co-annotated with GO: 0008203, cholesterol metabolic process, GO: 0017127, cholesterol transporter activity ….
CD36	SCL89_H	125,234,848-125,241,387	GO:0042149	cellular response to glucose starvation, & co-annotated with GO:0050995, negative regulation of lipid catabolic process, GO:0060621, negative regulation of cholesterol import, …
HSPA5	SCL96_H	92,031,134-92,115,474	GO:0030148	sphingolipid biosynthetic process, & co-annotated with GO:0009245, lipid A biosynthetic process, GO:0019216, regulation of lipid metabolic process, GO:2000189, positive regulation of cholesterol homeostasis
SPTLC1	SCL130_H	56,360,571..56,362,799	GO:0008203	cholesterol metabolic process, & co-annotated with GO:0006629 , lipid metabolic process, GO:0006707, cholesterol catabolic process, GO:0017127, cholesterol transporter activity
SCL131_H	COPS2	49,125,274-49,155,657	GO:0007165	signal transduction & co-annotated with GO:0016042, lipid catabolic process, GO:0006629, lipid metabolic process
SCL134_H	CREBBP	3,725,054-3,880,120	GO:0007165	Signal transduction. Reactome: REACT_22279, An association has been curated linking CREBBP and cellular lipid metabolic process in Homo sapiens.
SCL134_H	STUB1	680,115-682,768	GO:0051604	protein maturation, & co-annotated with GO:0090181, regulation of cholesterol metabolic process, GO:0019915, lipid storage
SCL134_H	SOCS1	11,254,417-11,256,182	GO:0046627 GO:0045444	negative regulation of insulin receptor signaling pathway, fat cell differentiation, , & co-annotated with GO:0010887, negative regulation of cholesterol storage, GO:0010888, negative regulation of lipid storage, GO:0019216, regulation of lipid metabolic process, GO:0071397, cellular response to cholesterol GO:0019915 , lipid storage …
SCL92_H SCL97_H	PRKACA	14,091,688-14,117,747	GO:0046827	positive regulation of protein export from nucleus, & co-annotated with GO:0046889, GO:0045833, positive and negative regulation of lipid biosynthetic process
SCL92_H SCL97_H	AP1M2	10,572,671-10,587,315	GO:0061024	Membrane organization, An association has been curated linking AP1M2 and membrane organization in Homo sapiens, Original References(s): Reactome: REACT_11123, & co-annotated with, GO: 0006497, protein lipidation, GO: 0017127, cholesterol transporter activity, GO: 0033344, cholesterol efflux…
SCL126_H	PLTP	45,898,620-45,912,364	GO:0006629 GO:0006869	lipid metabolic process, lipid transport, & co-annotated with GO:0008203, cholesterol metabolic process, GO:0033344, cholesterol efflux, GO:0006707, cholesterol catabolic process, …
SCL126_H	PLCG1	41,137,519-	GO:0009395	phospholipid catabolic process
		41,175,719	GO:0016042	lipid catabolic process GO:0008203, cholesterol metabolic process, GO:0045540, regulation of cholesterol biosynthetic process, GO:0033344, cholesterol efflux, GO:0030301, cholesterol transport, GO:0010873, GO:0060621, positive and negative regulation of cholesterol esterification, …

We used combination of protein–protein interactions network (PPI) and gene ontology (GO) to identify the genes associated with FHC disease, among different genes in its QTLs. The numbers of seven known genes associated with the disease were used as "seed genes" for identification of the genes neighborhood in Location of 119 neighborhood genes was compared with chromosome location of QTLs of the disease and the genes which were out of the QTLs zone were excluded.

Twenty two genes located in QTLs zone were selected as candidates for FHC and studied in terms of gene ontology characteristics.

To determine GO information of candidate genes, *QuickGO EBI* was used (19). For each gene, Biological Process term was studied.

The GO information of many genes is not completely known (20), therefore, co-annotated terms of the desired genes were specified and their GO characteristics were specified and studied.

The selected genes were studied and ranked with MedSim method in terms of association with FHC. MedSim is a novel approach for ranking candidate genes for a particular disease based on functional comparisons involving the Gene Ontology. It uses functional annotations of known disease genes to assess the similarity of diseases as well as the disease relevance of candidate genes (21).

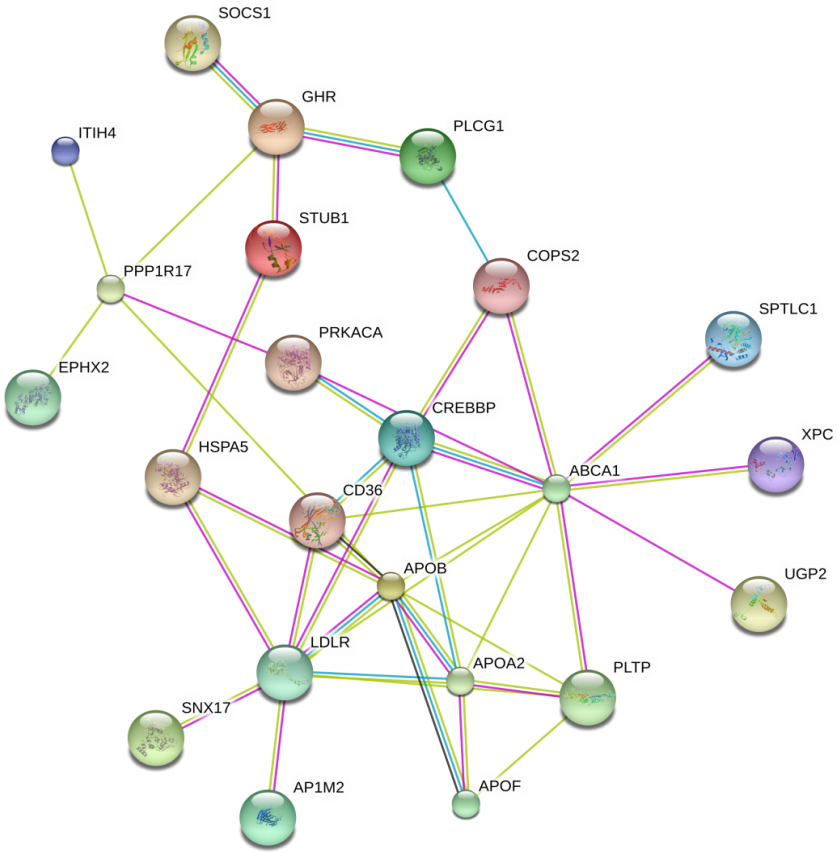

Figure 1. Interaction network of known genes and candidate genes for hypercholesterolemia

Results and Discussion

We developed a novel method for candidate disease gene identification. The method combines information about known genes and QTLs associated with disease with interaction network data, and their GO annotations to prioritize disease gene candidates.

As a result, we have identified 16 candidate genes, which may act as potential targets for hypercholesterolemia**.** Study of GO characteristics associated with 22 candidate genes for hypercholesterolemia which was in QTLs region of this disease indicated that 16 genes out of the candidate genes were associated with the disease considering GO information (Table 2). For example, APOB gene in QTL 85 (SCL85_H) plays role in positive regulation of cholesterol storage (GO: 0010886). Study of coexpression of APOB with the

known genes associated with FHC disease (seed gene) in COXPRESdb database (22), showed that APOB has coexpression with the known APOA2 and ITIH4 genes. Interference of APOB in hypercholesterolemia has been studied in different Among 22 candidate genes, for 7 genes including PDIA4, TGM2, LIMK1, SGTA, TYK2, SNTA1 and, NCOA6, there was not enough evidence for interfering in disease process in terms of GO characteristics.

Study of the neighborhood genes to seed genes demonstrated that APOF gene was a suitable candidate for FHC. But when the location of this gene was compared with QTLs associated with disease it was found that this gene is located near QTL: SCL130_H. Although APOF was not located in QTL region but the study of GO characteristics showed that this gene had strong relationship with FHC.

Table 3. Ranking the introduced genes for hypercholesterolemia based on similarity of GO (Biological Process term) to GO characteristics of the known genes of hypercholesterolemia with *Medsim* method. simRel, Lin, max simRel and max Lin are the methods for calculation of GO similarity score (21).

Rank	Gens name	BP simRel	BP LIN	BP max simRel	BP max Lin
1	APOF	0.97	1	1	1
2	PLCG1	0.8	0.81	0.98	1
3	SOCS1	0.79	0.79	1	1
4	PLTP	0.71	0.75	1	1
5	APOB	0.7	0.71	1	1
6	SNX17	0.68	0.7	1	1
7	AP1M2	0.64	0.66	0.91	1
8	CD36	0.62	0.63	1	1
9	HSPA5	0.55	0.56	0.84	0.84
10	PRKACA	0.54	0.56	0.98	1
11	CREBBP	0.52	0.55	0.98	1
12	SPTLC1	0.5	0.55	0.86	1
13	STUB1	0.47	0.49	0.84	0.84
14	UGP2	0.45	0.51	0.86	1
15	COPS2	0.36	0.39	0.55	0.57
16	XPC	0.36	0.38	0.65	0.66

As can be seen in Table 3, APOF has the first rank in similarity to the known genes of hypercholesterolemia based on Medsim method.

Results of APOF gene coexpression with the known genes (seed gene) showed that APOF gene has coexpression with EPHX, APOA2, ITIH4 and GHR which indicates the effect of APOF on FHC. Effect of this gene on transfer and esterification of cholesterol has been reported by Morton et al (24).

Sixteen introduced genes as the genes associated with FHC disease were studied and ranked -by MedSim method. In this method, similarity of GO characteristics of candidate genes was compared with GO characteristics of the known genes of disease. This study showed that the introduced genes have high similarity to the known FHC genes in terms of biological process (Table 3).

The simRel score is a functional similarity measure for comparing two GO terms with each other. It is based on Resnik's and Lin's similarity measures. The simRel score ranges from 0 for terms that have no similarity to 1 for terms with maximum similarity(25).

Conclusion

Each QTL includes many genes and identifying the target genes from a large number of candidates within these regions remains a challenge. Reducing QTL to a small number of testable candidate genes will be essential in quantitative trait analysis. This research showed that combination of linkage information (QTLs) and genomics information (GO, PPI) is highly capable of identifying genes associated with diseases. This combined method can be used for introduction of genes affecting diseases and also reduction of the number of candidate genes for quantitative trait in each QTL. Candidate genes can then be tested using a variety of experimental

methods, including RNA interference technology, deficiency complementation tests, knockouts, gene sequencing, pathway analysis, quantitative RT–PCR, Northern blots, Western blots, reporter gene assays, and various other protein assays**.**

Conflict of Interest

The authors declare no potential conflict of interest with respect to the authorship, and/or publication of this study.

Authors' Contributions

All the authors contribute in designing statistical tests, analyzing the results, introducing the hypothesis, and preparing the manuscript. All authors read and approved the final manuscript.

References

1. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005;6(4):287-98. PMID: 15803198

2. Smilde TJ, van Wissen S, Wollersheim H, Kastelein JJ, Stalenhoef AF. Genetic and metabolic factors predicting risk of cardiovascular disease in familial hypercholesterolemia. Neth J Med. 2001;59(4):184-95. PMID: 11578794

3. Lim D, Kim NK, Park HS, Lee SH, Cho YM, Oh SJ, et al. Identification of candidate genes related to bovine marbling using protein-protein interaction networks. Int J Biol Sci. 2011; 7(7):992-1002. PMID: 11578794

4. Sanz-Pamplona R, Berenguer A, Sole X, Cordero D, Crous-Bou M, Serra-Musach J, et al. Tools for protein-protein interaction network analysis in cancer research. Clin Transl Oncol. 2012; 14(1):3-14. PMID: 22262713

5. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell. 2004; 15(6):853-65. PMID: 15383276

6. Sam L, Liu Y, Li J, Friedman C, Lussier YA. Discovery of protein interaction networks shared by diseases. Pac Symp Biocomput. 2007; 76-87. PMID: 17992746

7. Ruffner H, Bauer A, Bouwmeester T. Human protein-protein interaction networks and the value for drug discovery. Drug Discov Today. 2007;12(17-18):709-16. PMID: 17826683

8. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010; 26(8):1057-63. PMID: 20185403

9. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008; 18(4):644-52. PMID: 18381899

10. Barrenas F, Chavali S, Holme P, Mobini R, Benson M. Network properties of complex human disease genes identified through genome-wide association studies. PLoS One. 2009; 4(11):e8090. PMID: 19956617

11. Yang P, Li X, Wu M, Kwoh CK, Ng SK. Inferring gene-phenotype associations via global protein complex network propagation. PLoS One. 2011;6(7):e21502. PMID: 21799737

12. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A. 2007; 104(21):8685-90. PMID: 17502601

13. Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012; 279(5):678-96. PMID: 22221742

14. Liu H, Su J, Li J, Lv J, Li B, Qiao H, et al. Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network. BMC Syst Biol. 2011; 5:158. PMID: 21985575

15. DiPetrillo K, Wang X, Stylianou IM, Paigen B. Bioinformatics toolbox for narrowing rodent quantitative trait loci. Trends Genet. 2005; 21(12):683-92. PMID: 16226337

16. Burgess-Herbert SL, Cox A, Tsaih SW, Paigen B. Practical applications of the bioinformatics toolbox for narrowing quantitative trait loci. Genetics. 2008; 180(4):2227-35. PMID: 18845850

17. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002; 30(1):52-5. PMID: 11752252

18. Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, et al. The Rat Genome Database 2013--data, tools and users. Brief Bioinform. 2013; 14(4):520-6. PMID: 23434633

19. Huntley RP, Binns D, Dimmer E, Barrell D, O'Donovan C, Apweiler R. QuickGO: a user tutorial for the web-based Gene Ontology browser. Database (Oxford). 2009; 2009:bap010. PMID: 20157483

20. Wang J, Zhou X, Zhu J, Zhou C, Guo Z. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinformatics. 2010; 11:290. PMID: 20509916

21. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics. 2010; 26(18):i561-7. PMID: 20823322

22. Obayashi T, Okamura Y, Ito S, Tadaka S, Motoike IN, Kinoshita K. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Res. 2013; 41(Database issue):D1014-20. PMID: 23203868

23. Tellis CC, Moutzouri E, Elisaf M, Wolfert RL, Tselepis AD. The elevation of apoB in hypercholesterolemic patients is primarily attributed to the relative increase of apoB/Lp-PLA(2). J Lipid Res. 2013; 54(12):3394-402. PMID: 24092915

24. Morton RE, Gnizak HM, Greene DJ, Cho KH, Paromov VM. Lipid transfer inhibitor protein (apolipoprotein F) concentration in normolipidemic and hyperlipidemic subjects. J Lipid Res. 2008; 49(1):127-35. PMID: 17901467

25. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006; 7:302. PMID: 16776819