Text Mining for Genes Involved in the Regression of Atherosclerosis

Adam Cankaya, Graduate Student
Dr. Ravi Shankar, Professor
Department of Computer Engineering and Computer Science, Florida Atlantic University
February 1, 2019

Abstract

There is evidence for the regression of atherosclerosis associated with improving plasma lipoprotein profiles induced by lowering concentrations of atherogenic apolipoprotein B (ApoB). The reduction of ApoB to levels associated with regression can be induced by multiple factors such as environment, nutrition, and pharmaceuticals. The genetic profile of a patient will help determine how effective these treatments can be in lowering ApoB concentrations and regressing atherosclerosis.

This paper will focus on the regression of atherosclerosis - returning a patient to the lipid concentration levels they maintained before the presence of heart disease. We will not consider the preventive or predictive factors that go into the development of atherosclerosis nor will we consider the reversal process of healing damage done to the arteries.

Instead we focus on factors that influence the serum levels of ApoB. A set of genes and gene products potentially involved in the regression process is curated using the MEDIE natural language search engine.

I. Background

The initial phase of atherosclerosis is “the subendothelial accumulation of apolipoprotein B-containing lipoproteins (ApoB).” The particles are modified by oxidation and lead to activation of endothelial cells to secrete chemokines. The chemokines act as chemoattractants that interact with receptors on monocytes to “recruit” them into the atherogenic process. The monocytes interact with the endothelial cells via interaction with a variety of selectins eventually causing them to differentiate into macrophages that “contribute to formation of necrotic core and fibrous cap thinning that characterizes the vulnerable plaque.” [1]

Regression of atherosclerosis is dependent upon lowering the concentration of ApoB levels to the point that the endothelial cells are no longer activated and do not trigger an atherogenic response. Multiple factors can play a role in determining the ApoB concentration including patient lifestyle (diet, exercise, body shape, smoking) and external factors (environment, genetics).

II. Methods

First, a manual reading of two literature reviews [1, 2] of previous studies on atherosclerosis regression attempts was completed. From these papers a list of genes associated with the atherosclerosis regression process was manually curated. Only genes specifically associated with processes related to ApoB concentrations were considered. The genes were then manually searched in OMIM to gather information on their known gene products. All of this is curated into Table 1 below.

Next, a manual reading of a review of text mining tools for the biomedical domain [3] led to using the MEDIE [4] natural language search engine. MEDIE was used in an attempt to gather additional genes not mentioned in the literature reviews. MEDIE allows users to search for abstracts using a natural language query that includes a subject, a verb, and an object. The subject and object can be either a gene, gene product, or a disease. These types can be specified under the "advanced search" option. When a subject or object is a disease it can be associated with a UMLS identification. By default genes are highlighted in red on the search results.

For example, a MEDIE search for "[what] regresses atherosclerosis" gives 25 abstracts. The subject is left blank in the search terms (instead of actually typing in "what") as to not limit results. By leaving the subject blank MEDIE returns its own subject for each result. Some of the subjects returned are "Vitamin E", "Niacin", "results of major trials of drugs", "they", "we", and multiple pharmaceutical names.

Another MEDIE search was done using just an object of "ApoB (UMLS:C1862606)". By just searching for an object, MEDIE returns its own values for subjects and verbs for each search result. This can give us some insight into terms to use to narrow the results. For example, the first 100 results from the search with just an object of "ApoB" includes verbs such as "performed", "discovered", "associated", "measured", "assessed", "reduced", "correlated" and "investigated".

By picking some of these most interesting verbs additional filtered results related to regressing ApoB levels can be found. For example, searching for "[what] reduced ApoB" also gives over 100 results with dozens of genes and gene products listed. Some interesting subjects include "this diet", "a supervised exercise program" and "Omega-3 fatty acids". By searching for "[what] regulates ApoB" more subjects come up such as "insulin" and "a range of factors including developmental, nutritional, environmental, and metabolic stimuli".

These abstracts contain genes/products previously identified in Table 1 such as LXRα and ABCA1, but also identifies dozens of other genes and gene products. Some of the most applicable additional results are summarized in Table 2 below. Once again the gene and product terms are manually searched on OMIM to confirm relevance and gene/product relationships.

The MEDIE results also mention phrases such as “oxidized LDL” and “elevation of HDL cholesterol” that give us immediate insight into the atherosclerosis process. They mention potential causes of atherosclerosis including “infection-induced atherosclerosis” and “hypothyroidism”. The results even immediately point to modes of regression such as “azithromycin treatment”, “statins”, “replacement of a high cholesterol-saturated fat diet by another cholesterol free-unsaturated fat diet”, and “supplementation with HDL”.

III. Tabulated Results

Table 1: Manually curated selection of genes and their potential contribution to processes related to regression of atherosclerosis by the lowering of ApoB levels.

Gene	**Gene product**	**Potential role in atherosclerosis regression**	**OMIM #**
APOB, APOE	Apolipoprotein A, Apolipoprotein E	Forms a lipoprotein with LDL that allows the LDL to be transported in blood	107730, 107741
MSR1	type A scavenger receptor	Regulates uptake of oxidized LDL by macrophages	153622
SCARB1	CD36, member of type B scavenger receptor family	Regulates uptake of oxidized LDL by macrophages	601040
CCR7	Chemokine (C-C motif) receptor 7	Dendritic cell emigration from aortic lesions, promotes regression via emigration of CD68+ cells	600242
NR1H3	LXRα (Nuclear oxysterol liver X receptor alpha)	Increased levels found in foam cells during regression	602423
ABCA1	ABCA-1 (ATP-binding cassette 1)	Anti-atherogenic target of LXRα	600046
MIR33A	MiR-33 (Micro-RNA 33A)	Intronic mRNA that inhibits hepatic expression of ABCA-1, reduces HDL-C concentration, resulting in decreased cholesterol efflux	612156
CETP	CETP (Cholesteryl ester transfer protein)	Mediates exchange of lipids between lipoproteins.	118470

Table 2: Genes and gene products associated with atherosclerosis regression curated from searches using the MEDIE natural language search engine.

Gene	Gene product	Potential role in atherosclerosis regression	OMIM #
PPARA, PPARG	Peroxisome Proliferator-Activated Receptor-Alpha and Receptor-Gamma	Serve as lipid sensors, induces peroxisomes which contribute to the oxidation of fatty acids	601487
LDLR, LDLRAP1	Low Density Lipoprotein Receptor	Promotes cholesterol uptake in cells	605747
PCSK9	Proprotein convertase subtilisin/kexin type 9	Reduces LDLR levels	607786
ABHD5	Abhydrolase domain-containing 5	Major intermediate in membrane and storage lipid biosynthesis, coactivator of adipocyte triglyceride lipase	604780
ATF4	Activating Transcription Factor 4	Influences plasma lipoprotein levels	604064
MTTP	Microsomal triglyceride transfer protein	Catalyzes the transport of cholesterol between phospholipid surfaces	157147
LCAT	Lecithin cholesterol acyltransferase	Converts cholesterol to cholesteryl esters on surface of HDL	606967
PTEN	Phosphatase and tensin homolog	Regulates secretion of ApoB containing lipoproteins	601728
HNF4A	Hepatocyte nuclear factor-4alpha	Transcription factor that regulates genes involved in lipid metabolism	600281
PTPN11	Protein-tyrosine phosphatase non-receptor type 11	Influences serum ApoB and LDL levels	176876
mTOR	Mechanistic target of rapamycin	Mediation of inhibition of ApoB by insulin	601231
MAP2K4	Mitogen-activated protein kinase 4	Potential role in ApoB particle assembly	601335
APOBEC1	Apolipoprotein B mRNA-editing enzyme	Metabolic regulation of ApoB mRNA editing, converts ApoB-48 (intestine)	600130

	catalytic polypeptide 1	from ApoB-100 (liver)	
FGA, FGB, FGG	Fibrinogen polypeptide alpha, beta, gamma	Associated with increased secretion of ApoB	134820

IV. Enrichment Analysis

The genes and OMIM numbers from tables 1 and 2 are first converted to HGNC. The HGCN values listed are simply taking from the "HGNC Approved Gene Symbol" field on the OMIM entry's page.

Table 3: The genes and corresponding HGCN terms used for enrichment analysis

Item #	Gene	HGCN		Item #	Gene	HGCN
1	APOB	APOB		15	ABHD5	ABHD5
2	APOE	APOE		16	ATF4	ATF4
3	MSR1	MSR1		17	MTTP	MTTP
4	SCARB1	SCARB1		18	LCAT	LCAT
5	CCR7	CCR7		19	PTEN	PTEN
6	NR1H3	NR1H3		20	HNF4A	HNF4A
7	ABCA1	ABCA1		21	PTPN11	PTPN11
8	MIR33A	MIR33A		22	mTOR	MTOR
9	CETP	CETP		23	MAP2K4	MAP2K4
10	PPARA	PPARA		24	APOBEC1	APOBEC1
11	PPARG	PPARG		25	FGA	FGA
12	LDLR	LDLR		26	FGB	FGB
13	LDLRAP1	LDLRAP1		27	FGG	FGG
14	PCSK9	PCSK9				

Next, the list of HGCN terms are put into the GO Enrichment Analysis tool on http://www.GeneOntology.org. Using the search values of "biological process" and "Homo sapiens" and the GO Ontology database released 2019-01-01 gives results with 26 out of 26

mapped ID's. There is one unmapped ID for MIR33A. The Bonferroni count is 8749. The results are exported and then filtered for entries with a p-value of less than 10^{-3}. This gives a total of 145 results.

The filtered GO term and P-value results are then fed into ReviGO at http://ReviGO.irb.hr. Doing so results in four of the terms being not found in the 2016 version of Gene Ontology used by RevGO - GO:1905952, GO:1905953, GO:1905954, and GO:62012.

Of the remaining 141 GO terms, 18 of them have a frequency greater than 10% and 40 of them have a frequency greater than 1%:

Table 4: The GO terms associated with the list of genes and with frequencies greater than 1%.

GO Term ID	Description	Frequency
GO:0008152	metabolic process	75.39%
GO:0071704	organic substance metabolic process	58.36%
GO:0044238	primary metabolic process	53.74%
GO:0044237	cellular metabolic process	53.06%
GO:0043170	macromolecule metabolic process	39.49%
GO:0006807	nitrogen compound metabolic process	38.74%
GO:0009058	biosynthetic process	31.61%
GO:1901576	organic substance biosynthetic process	30.37%
GO:1901360	organic cyclic compound metabolic process	30.32%
GO:0044249	cellular biosynthetic process	30.05%
GO:0051179	localization	18.50%
GO:0019538	protein metabolic process	18.49%
GO:1901564	organonitrogen compound metabolic process	17.89%
GO:0051234	establishment of localization	17.76%
GO:0006810	transport	17.62%
GO:0044281	small molecule metabolic process	15.14%
GO:0019222	regulation of metabolic process	11.94%

GO:0080090	regulation of primary metabolic process	11.68%
GO:0051716	cellular response to stimulus	9.56%
GO:0071702	organic substance transport	4.98%
GO:0006950	response to stress	4.58%
GO:0006629	lipid metabolic process	3.52%
GO:0065008	regulation of biological quality	3.40%
GO:0042221	response to chemical	3.07%
GO:0033036	macromolecule localization	3.03%
GO:0044255	cellular lipid metabolic process	2.70%
GO:0032501	multicellular organismal process	2.37%
GO:0048519	negative regulation of biological process	1.98%
GO:0048523	negative regulation of cellular process	1.83%
GO:0048518	positive regulation of biological process	1.74%
GO:0042592	homeostatic process	1.66%
GO:0051128	regulation of cellular component organization	1.59%
GO:0048522	positive regulation of cellular process	1.59%
GO:0009605	response to external stimulus	1.37%
GO:0050793	regulation of developmental process	1.21%
GO:0015711	organic anion transport	1.19%
GO:0048583	regulation of response to stimulus	1.12%
GO:0009893	positive regulation of metabolic process	1.05%
GO:0070887	cellular response to chemical stimulus	1.01%
GO:0031325	positive regulation of cellular metabolic process	1.00%

Additional enrichment analysis was attempted by utilizing ReviGO to search for “molecular functions” and “cellular components” for homo sapiens. The molecular functions option does not return any GO terms with frequencies above 1%, while the cellular component

option only returns very generic GO terms such as “organelle” and “macromolecular complex”. Hence we do not consider any of the results for molecular functions or cellular components.

V. Discussion

There is still much more work to do in gathering genes and gene products before a genotype ontology can be linked to a phenotype ontology. The genes listed in this paper represent the high level of research into potential atherosclerosis regression. We mainly only considered serum levels of ApoB, but there are other apolipoproteins (such as ApoE), that potentially have an equally important part in the process. Also each individual type of apolipoprotein has multiple subtypes that can be converted between each other via mRNA editing. The relationship between ApoB concentrations versus the traditional HDL/LDL ratio as an indicator of atherosclerosis susceptibility must also be studied.

A set of GO terms with high frequencies was generated by ReviGO for biological processes, however the terms identified are rather generic - for example several "metabolic process" and "biosynthetic process" entries. It is speculated that the list of genes being used for enrichment analysis needs to be more specialized to focus on gene products exclusively or at least primarily only involved in the atherosclerosis process.

Or perhaps GO terms with smaller frequencies should be considered. For example, some of the frequencies below 1% appear very relevant to arthestoscolersis - “regulation of lipid storage” (0.010%), “negative regulation of cholesterol storage” (0.003%), and “cholesterol transport” (0.020%). When the “show/hide dispensable GO terms” option is used on ReviGO, a lot of these smaller frequency terms are still shown, indicating that ReviGO believes them to be relevant, despite smaller frequency.

VI. Conclusions

We have gathered over twenty gene and gene product pairs involved in the body’s regulation of blood serum ApoB concentration levels. Given that ApoB is the molecular component that allows cholesterol to be transported throughout the body, it is believed that lowering the ApoB concentration in a patient with atherosclerosis to pre-disease levels will result in the regression of atherosclerosis and halting of any further damage to the patient’s arteries. It is further hoped that by studying the influence of these genes we will better be able to make pharmaceutical treatment recommendations combined with changes to patient lifestyle and environmental factors .

VII. References

1. Feig, J. E. (2014). Regression of Atherosclerosis: Insights from Animal and Clinical Studies. *Annals of Global Health*, *80*(1), 13. doi:10.1016/j.aogh.2013.12.001

2. Chistiakov, D. A., Myasoedova, V. A., Revin, V. V., Orekhov, A. N., & Bobryshev, Y. V. (2017). The phenomenon of atherosclerosis reversal and regression: Lessons from animal models. *Experimental and Molecular Pathology, 102*(1), 138-145. doi:10.1016/j.yexmp.2017.01.013

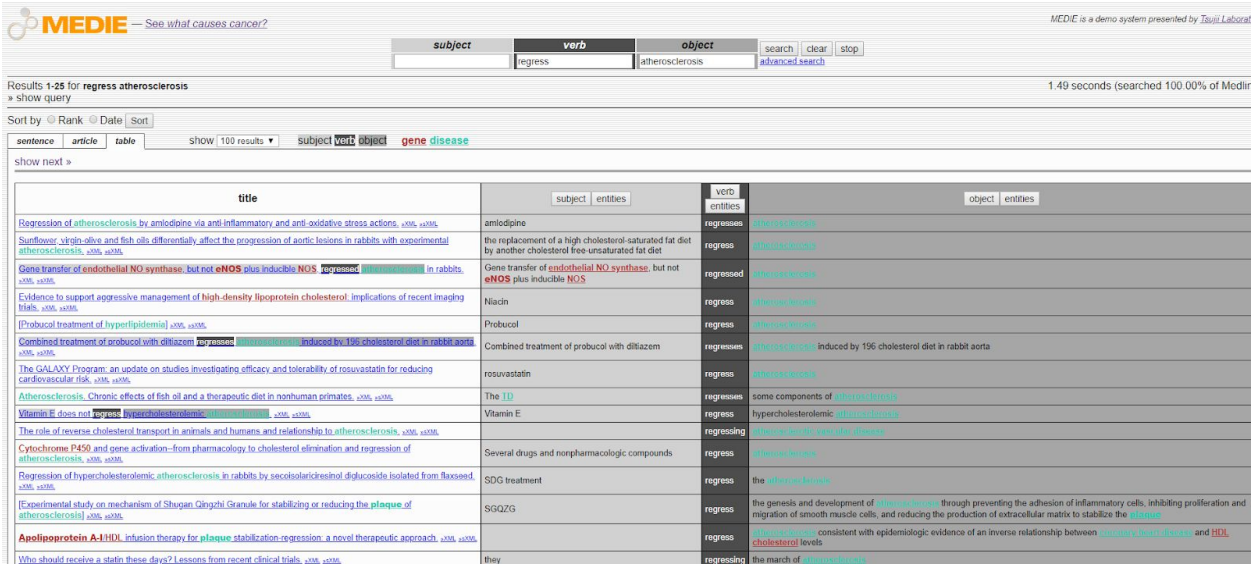
3. Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods, 74*, 97-106. doi:10.1016/j.ymeth.2015.01.015

4. MEDIE - Semantic retrieval engine for MEDLINE. Retrieved from http://www.nactem.ac.uk/medie/search.cgi. National Center for Text Mining.

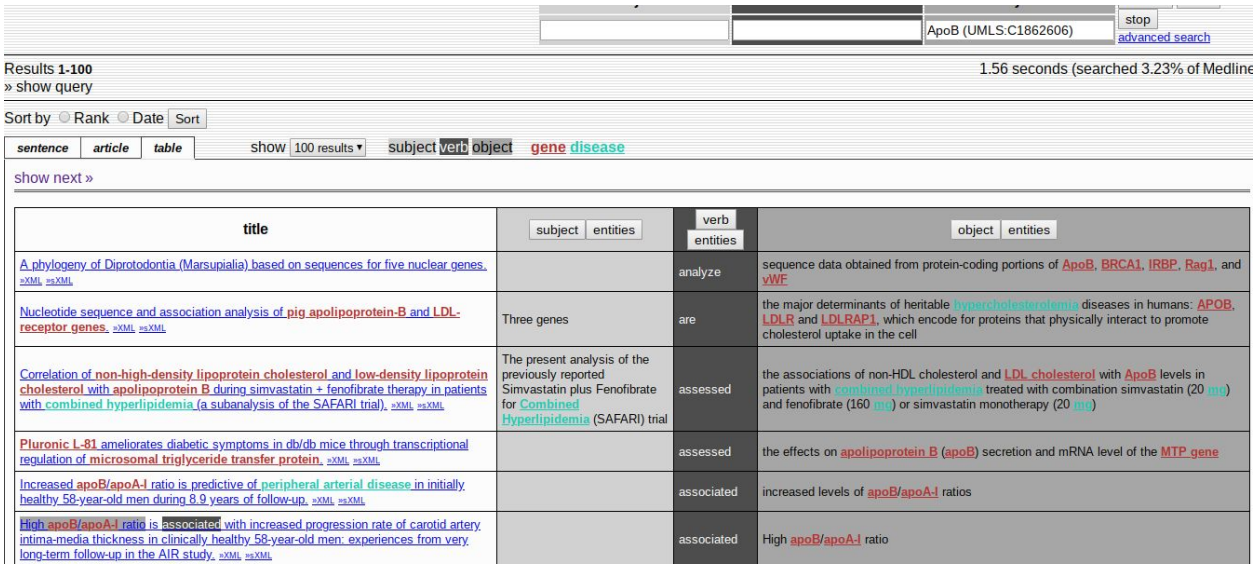
5. Supek F, Bošnjak M, Škunca N, Šmuc T. "*REVIGO summarizes and visualizes long lists of Gene Ontology terms*" PLoS ONE 2011. doi:10.1371/journal.pone.0021800

VIII. Appendices

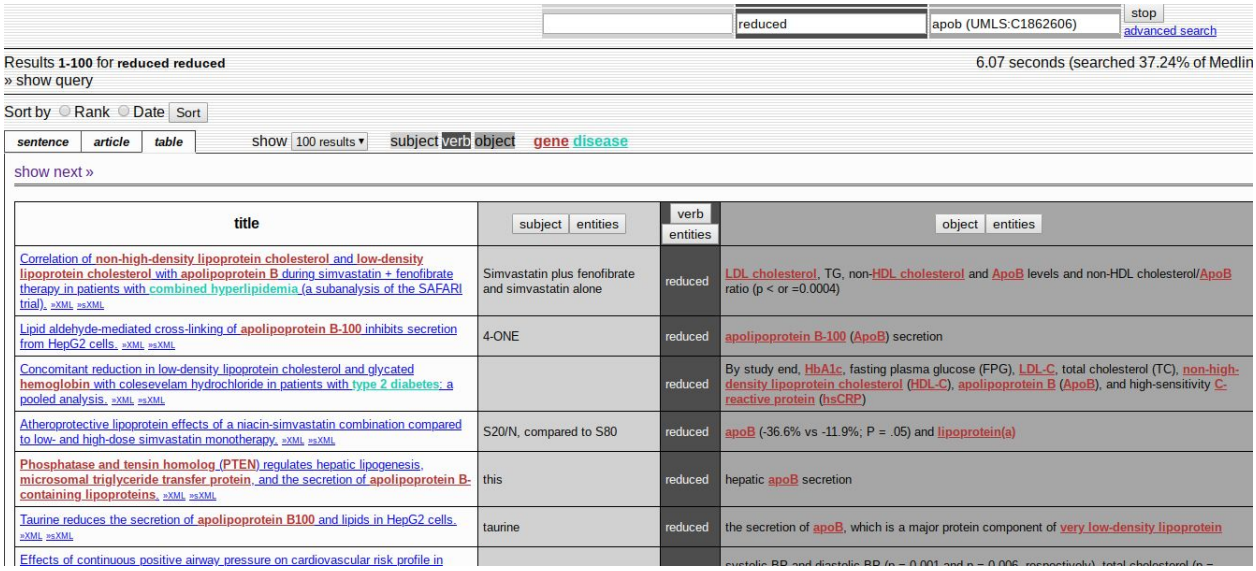
"[What] regresses atherosclerosis?"



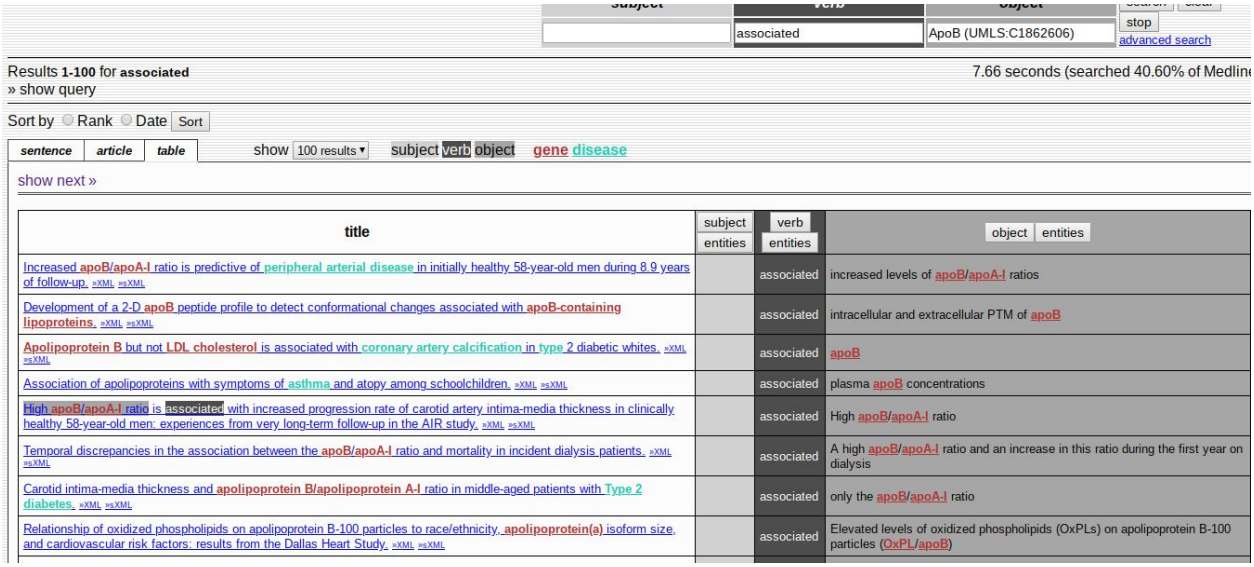
"[What] [Causes] ApoB?"



“[What] reduced ApoB?”



“[What] associated ApoB?”



	verb	object	
	associated	ApoB (UMLS:C1862606)	stop

advanced search

Results **1-100** for **associated** 7.66 seconds (searched 40.60% of Medlin

» show query

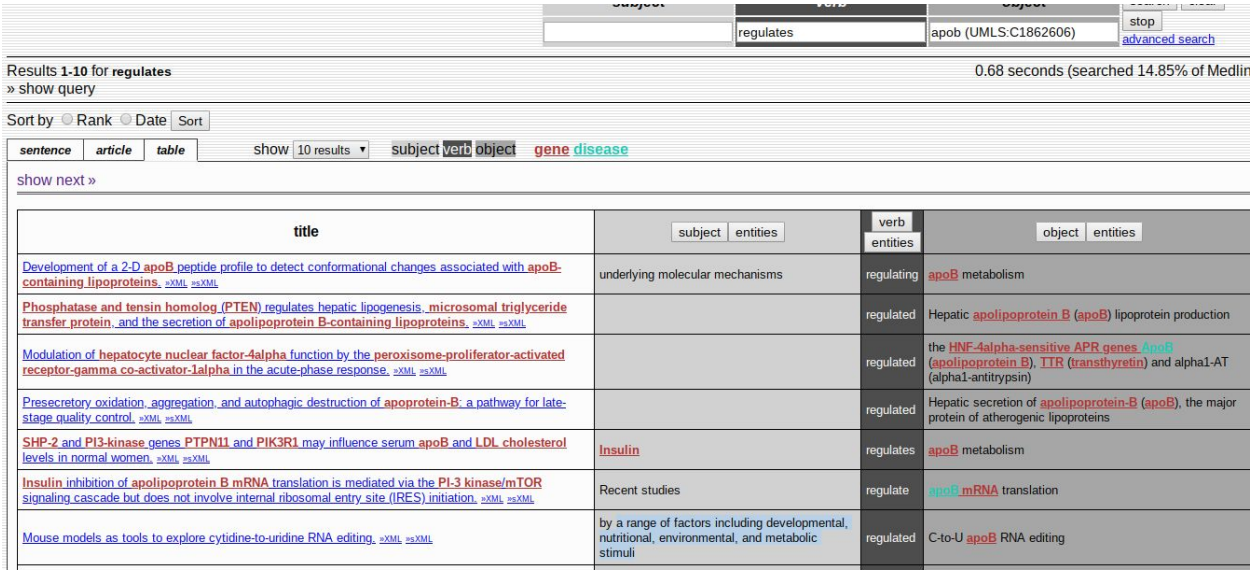
Sort by Rank Date Sort

sentence | article | table show 100 results subject verb object gene disease

show next »

title	subject entities	verb entities	object entities
Increased apoB/apoA-I ratio is predictive of peripheral arterial disease in initially healthy 58-year-old men during 8.9 years of follow-up. »XML »sXML		associated	increased levels of apoB/apoA-I ratios
Development of a 2-D apoB peptide profile to detect conformational changes associated with apoB-containing lipoproteins. »XML »sXML		associated	intracellular and extracellular PTM of apoB
Apolipoprotein B but not LDL cholesterol is associated with coronary artery calcification in type 2 diabetic whites. »XML »sXML		associated	apoB
Association of apolipoproteins with symptoms of asthma and atopy among schoolchildren. »XML »sXML		associated	plasma apoB concentrations
High apoB/apoA-I ratio is associated with increased progression rate of carotid artery intima-media thickness in clinically healthy 58-year-old men: experiences from very long-term follow-up in the AIR study. »XML »sXML		associated	High apoB/apoA-I ratio
Temporal discrepancies in the association between the apoB/apoA-I ratio and mortality in incident dialysis patients. »XML »sXML		associated	A high apoB/apoA-I ratio and an increase in this ratio during the first year on dialysis
Carotid intima-media thickness and apolipoprotein B/apolipoprotein A-I ratio in middle-aged patients with Type 2 diabetes. »XML »sXML		associated	only the apoB/apoA-I ratio
Relationship of oxidized phospholipids on apolipoprotein B-100 particles to race/ethnicity, apolipoprotein(a) isoform size, and cardiovascular risk factors: results from the Dallas Heart Study. »XML »sXML		associated	Elevated levels of oxidized phospholipids (OxPLs) on apolipoprotein B-100 particles (OxPL/apoB)

"[What] regulates ApoB?"



	verb	object	
	regulates	apob (UMLS:C1862606)	stop

advanced search

Results **1-10** for **regulates** 0.68 seconds (searched 14.85% of Medlin

» show query

Sort by Rank Date Sort

sentence | article | table show 10 results subject verb object gene disease

show next »

title	subject entities	verb entities	object entities
Development of a 2-D apoB peptide profile to detect conformational changes associated with apoB-containing lipoproteins. »XML »sXML	underlying molecular mechanisms	regulating	apoB metabolism
Phosphatase and tensin homolog (PTEN) regulates hepatic lipogenesis, microsomal triglyceride transfer protein, and the secretion of apolipoprotein B-containing lipoproteins. »XML »sXML		regulated	Hepatic apolipoprotein B (apoB) lipoprotein production
Modulation of hepatocyte nuclear factor-4alpha function by the peroxisome-proliferator-activated receptor-gamma co-activator-1alpha in the acute-phase response. »XML »sXML		regulated	the HNF-4alpha-sensitive APR genes ApoB (apolipoprotein B), TTR (transthyretin) and alpha1-AT (alpha1-antitrypsin)
Presecretory oxidation, aggregation, and autophagic destruction of apoprotein-B: a pathway for late-stage quality control. »XML »sXML		regulated	Hepatic secretion of apolipoprotein-B (apoB), the major protein of atherogenic lipoproteins
SHP-2 and PI3-kinase genes PTPN11 and PIK3R1 may influence serum apoB and LDL cholesterol levels in normal women. »XML »sXML	Insulin	regulates	apoB metabolism
Insulin inhibition of apolipoprotein B mRNA translation is mediated via the PI-3 kinase/mTOR signaling cascade but does not involve internal ribosomal entry site (IRES) initiation. »XML »sXML	Recent studies	regulate	apoB mRNA translation
Mouse models as tools to explore cytidine-to-uridine RNA editing. »XML »sXML	by a range of factors including developmental, nutritional, environmental, and metabolic stimuli	regulated	C-to-U apoB RNA editing