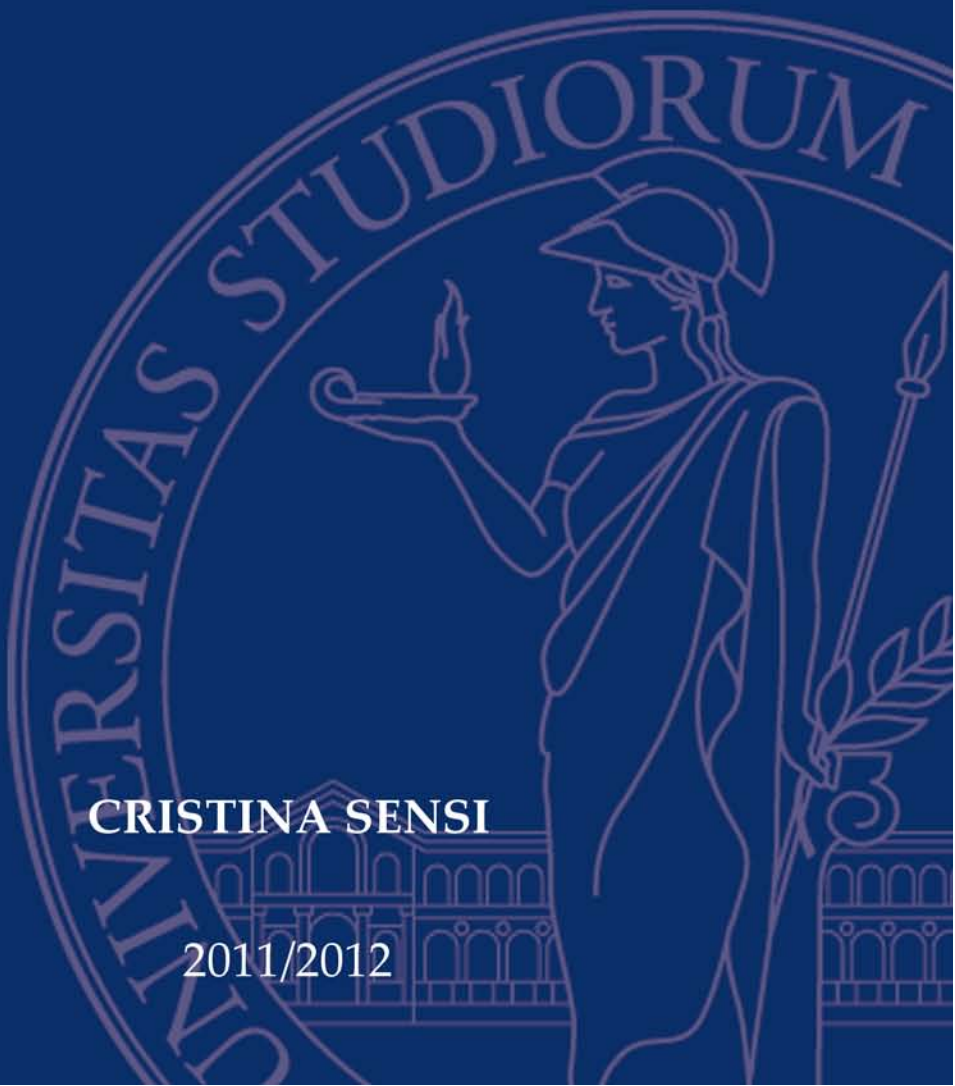
Graduate School in Biochemical, Nutritional and Metabolic Sciences

UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE AGRARIE E ALIMENTARI

LCAT MOLECULAR MODELING: APPLICATIONS TO STRUCTURE-FUNCTION RELATIONSHIPS AND TO INHIBITOR DISCOVERY



CRISTINA SENSI

2011/2012

UNIVERSITÀ DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO
Scienze Biochimiche, Nutrizionali e Metaboliche

FACOLTA' DI SCIENZE AGRARIE E ALIMENTARI

CORSO DI DOTTORATO
Nutrizione Sperimentale e Clinica
XXV Ciclo

LCAT MOLECULAR MODELING: APPLICATIONS TO STRUCTURE-FUNCTION RELATIONSHIPS AND TO INHIBITOR DISCOVERY

Settore scientifico-disciplinare BIO/10

Dottorando
CRISTINA SENSI

Tutor
ELISABETTA GIANAZZA

Co-tutor
IVANO EBERINI
ANITA FERRARETTO

Coordinatore del dottorato
AMBROGINA PAGANI

A.A. 2011/2012

"E os marinheiros, perguntou ela, [...], Disseram-me que já não há ilhas desconhecidas, e que, mesmo que as houvesse, não iriam eles tirar-se do sossego dos seus lares e da boa vida dos barcos de carreira para se meterem em aventuras oceânicas, à procura de um impossível, como se ainda estivéssemos no tempo do mar tenebroso, E tu, que lhes respondeste, Que o mar é sempre tenebroso"

(J. Saramago, O conto da Ilha Desconhecida)

PREFACE

A series of large population studies revealed the existence of a strong inverse correlation between plasma levels of high-density lipoprotein cholesterol (HDL·C) and cardiovascular risk. This inverse relation is observed at the level of low-density lipoprotein cholesterol, including very low concentrations achieved in statin-treated patients [Barter, 2007a]. The central role of HDL in reverse cholesterol transport (RCT), the process by which cholesterol is removed from arterial wall macrophages and routed to the liver for disposal, is widely considered the main mechanism responsible for HDL-mediated atheroprotection [Cuchel 2006]. The efflux of unesterified cholesterol from peripheral cells to extracellular acceptors is the first and rate-limiting step of RCT. This movement can occur simply by aqueous diffusion, or through three pathways involving distinct cell membrane proteins [Jessup 2006]:
(i) an ATP-dependent pathway mediated by the ATP binding cassette A1 (ABCA1) transporter;
(ii) an ATP-dependent pathway mediated by the ABCG1 transporter;
(iii) an ATP-independent pathway involving the scavenger receptor type BI (SR-BI).
ApoA-1 interacts with ABCA1, acquiring cellular phospholipids and cholesterol, and thereby forming preβ-HDL [Yokoyama 2006], which can remove further cell cholesterol via ABCA1 [Favari 2004]. In the circulation, preβ-HDL are acted on by lecithin:cholesterol acyltransferase (LCAT), which esterifies cholesterol and generates α-HDL particles, the preferential cell cholesteroil acceptors through the ABCG1- and SR-BI-mediated pathways [Jessup 2006].
The LCAT-synthesized cholesteryl esters are transported to the liver through two major routes: (i) transfer by the cholesteryl ester transfer protein (CETP) to triglyceride-rich lipoproteins, which are metabolized into remnants and then removed by the liver, and (ii) direct uptake by the liver through SR-BI. In addition to the two routes for HDL-cholesteryl esters transport, there is also an SR-BI-dependent pathway for the direct uptake of HDL unesterified cholesterol in the liver [Cuchel 2006]. Beside their major function in RCT, HDL exert a series of non-lipid related atheroprotective activities, such as maintaining of endothelial cell homeostasis [Calabresi 2003].
HDL have become a major target for the development of novel therapies for the treatment of atherosclerotic cardiovascular disease [Linsel-Nitschke 2005]. However, doubts on the clinical benefit achievable with such interventions have been raised by the failure of torcetrapib (Barter 2007b), a small lipophilic molecule that binds CETP. The failure of torcetrapib, however, does not mean that the concept of targeting HDL to further reduce cardiovascular risk is definitely dead. Indeed, other CETP inhibitors lacking aldosterone and blood pressure effects are at advances stages of clinical development [Rennings 2008; Bloomfield 2009].

LCAT (lecithin:cholesterol acyltransferase) catalyzes the transacylation of the sn-2 fatty acid of lecithin to the free 3-OH group of cholesterol, generating cholesteryl ester (CE) and lysolecithin [Jonas 2000]. By this way, the LCAT reaction accounts for the synthesis of most of the plasma CE. The preferred lipoprotein substrate for LCAT is a newly assembled small, discoidal HDL migrating in preβ position on agarose gel, thus called preβ-HDL [Nakamura 2004]. The formation and accumulation of CE within these particles leads to conversion into spherical, α-migrating HDL, which comprise most of the plasma HDL. preβ-HDL have a short plasma half-life, being rapidly cleared through the kidney [Rye 2004], white mature α-HDL

have a much slower turnover. LCAT thus plays a central role in intravascular HDL metabolism and in the determination of plasma HDL levels.
In addition to its function in HDL metabolism, LCAT has also long been believed to play a critical role in macrophage RCT. However, the most recent findings have questioned the critical role of LCAT in RCT. Human LCAT overexpression in mice remarkably increases plasma HDL levels but does not enhance macrophage RCT in vivo. Conversely, LCAT-deficient mice display an almost preserved macrophage RCT in vivo despite a severe plasma HDL reduction [Tanigawa 2009]. These mouse findings are consistent with early data in humans, showing that sera form carriers of genetic LCAT deficiency can efficiently remove cholesterol from macrophages [Calabresi 2009b] and that HDL are capable of directly delivering large amounts of unesterified cholesterol to the liver via SR-BI, without LCAT-mediated conversion into CE [Schwartz 1982; Schwartz 2004].
The effect of LCAT on atherogenesis is still controversial. Studies in animals overexpressing or lacking LCAT have provided inconsistent results. LCAT overexpression in mice or rabbits remarkably increases plasma HDL levels, and is associated with protection against diet-induced atherosclerosis in rabbits but not in mice. Mice lacking the LCAT gene show the expected reductions in plasma HDL levels and have less or more atherosclerosis than their counterparts with active LCAT, mostly depending on the plasma levels of apoB-containing lipoproteins [Rousset 2009]. Early cross-sectional studies in humans have reported either decreased [Solajić-Bozicević1991; Solajić-Bozicević 1994]; or increased [Wells 1986] LCAT activity in patients with angiographically proven coronary artery disease.
A recent study evaluated carotid atherosclerosis in Italian carriers of genetic LCAT deficiency, showing that the inheritance of a mutated LCAT genotype has a remarkable gene-dose-dependent effect in reducing carotid IMT [Calabresi 2009a]. These surprising results suggest that the presence of defective LCAT may not preclude cholesterol removal from the arterial wall, and efficient reverse cholesterol transport. Indeed, sera from the same carriers of LCAT gene mutations displayed an enhanced capacity for ABCA1-mediated cell cholesterol efflux than sera from non affected family members, due to the greater content of preβ-HDL particles [Calabresi 2009b]. Recently, the first prospective study relating LCAT concentration to future cardiovascular events showed that low LCAT plasma levels are not associated with increased atherosclerosis in the general population. Interestingly, a gender-specific analysis showed that higher LCAT levels in women are associated with an increased cardiovascular risk [Holleboom 2010]. More recently, Baldassarre and colleagues measured plasma LCAT concentration in European individuals associated with a high cardiovascular risk [Baldassarre 2010], to look at the relation between LCAT and atherosclerosis, and their results were similar to those of Calabresi and colleagues [Calabresi 2011].
These recent data, together with findings about genetic LCAT deficiency, challenge the idea that LCAT is atheroprotective and suggest that, despite positive effects on plasma HDL concentration, elevating LCAT expression and/or activity is not a promising therapeutic strategy to reduce cardiovascular risk. On the contrary, decreasing LCAT activity has certainly negative effects on plasma HDL concentration but possibly positive effects on HDL structure/function, and may represent a new therapeutic strategy to reduce cardiovascular risk. This option has never been tested before, and it represents the basis of this thesis.

References

Baldassarre D. et al., 2010 Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study. Eur Heart J. 2010 Mar;31(5):614-22.

Barter P.J. et al., 2007a HDL cholesterol, very low levels of LDL cholesterol, and cardiovascular events. Treating to New Targets Investigators. N Engl J Med. 357(13):1301-10.

Barter P.J. et al., 2007b Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med. 357(21):2109-22.

Bloomfield D. et al. 2009 Efficacy and safety of the cholesteryl ester transfer protein inhibitor anacetrapib as monotherapy and coadministered with atorvastatin in dyslipidemic patients. Am Heart J. 157(2):352-360.e2.

Calabresi L. et al., 2003 Endothelial protection by high-density lipoproteins: from bench to bedside. Arterioscler Thromb Vasc Biol. 23(10):1724-31.

Calabresi L. et al., 2009a Functional lecithin: cholesterol acyltransferase is not required for efficient atheroprotection in humans.Circulation. 120(7):628-35.

Calabresi L. et al., 2009b A novel homozygous mutation in CETP gene as a cause of CETP deficiency in a Caucasian kindred. Atherosclerosis. 205(2):506-11.

Calabresi L. et al., 2011 Plasma lecithin:cholesterol acyltransferase and carotid intima-media thickness in European individuals at high cardiovascular risk. J Lipid Res. 52(8):1569-74.

Cuchel M., Rader D.J. 2006 Macrophage reverse cholesterol transport: key to the regression of atherosclerosis? Circulation. 113(21):2548-55.

Favari E. et al., 2004 Depletion of pre-beta-high density lipoprotein by human chymase impairs ATP-binding cassette transporter A1- but not scavenger receptor class B type I-mediated lipid efflux to high density lipoprotein. J Biol Chem. 279(11):9930-6.

Holleboom A.G. et al.,2010 Plasma levels of lecithin:cholesterol acyltransferase and risk of future coronary artery disease in apparently healthy men and women: a prospective case-control analysis nested in the EPIC-Norfolk population study. J Lipid Res. 51(2):416-21.

Jessup W. Et al., 2006 Roles of ATP binding cassette transporters A1 and G1, scavenger receptor BI and membrane lipid domains in cholesterol export from macrophages. Curr Opin Lipidol. 17(3):247-57.

Jonas A. 2000 Lecithin cholesterol acyltransferase. Biochim Biophys Acta. 1529(1-3):245-56.

Linsel-Nitschke P., Tall A.R. 2005 HDL as a target in the treatment of atherosclerotic cardiovascular disease. Nat Rev Drug Discov. 4(3):193-205.

Nakamura Y. et al., 2004 Molecular mechanism of reverse cholesterol transport: reaction of pre-beta-migrating high-density lipoprotein with plasma lecithin/cholesterol acyltransferase. Biochemistry. 43(46):14811-20.

Rennings A.J., Stalenhoef A. 2008 JTT-705: is there still future for a CETP inhibitor after torcetrapib? Expert Opin Investig Drugs. 17(10):1589-97.

Rousset X. et al., 2009 Lecithin: cholesterol acyltransferase--from biochemistry to role in cardiovascular disease. Curr Opin Endocrinol Diabetes Obes. 16(2):163-71.

Rye K.A., Barter P.J 2004 Formation and metabolism of prebeta-migrating, lipid-poor apolipoprotein A-I.Arterioscler Thromb Vasc Biol. 24(3):421-8.

Schwartz C.C. et al., 1982 Central role of high density lipoprotein in plasma free cholesterol metabolism. J Clin Invest. 70(1):105-16.

Schwartz C.C. et al., 2004 Lipoprotein cholesteryl ester production, transfer, and output in vivo *in humans. J Lipid Res. 45(9):1594-607.*

Solajić-Bozicević N, et al., 1991 Lecithin:cholesterol acyltransferase activity in patients with acute myocardial infarction and coronary heart disease. Artery. 18(6):326-40.

Solajić-Bozicević N, et al., 1994 Lecithin-cholesterol acryltransferase activity in patients with coronary artery disease examined by coronary angiography. Clin Investig. 72(12):951-6.

Tanigawa H. et al., 2009 Lecithin: cholesterol acyltransferase expression has minimal effects on macrophage reverse cholesterol transport in vivo*. Circulation 120(2):160-9.*

Wells I.C. et al., 1986. Lecithin: cholesterol acyltransferase and lysolecithin in coronary atherosclerosis.Exp Mol Pathol. 45(3):303-10.

Yokoyama S. 2006 ABCA1 and biogenesis of HDL. J Atheroscler Thromb. 13(1):1-15.

Index

1	***STATE OF THE ART***	***11***
1.1	*BIOLOGICAL INTRODUCTION*	*13*
	Atherosclerosis	*13*
	Cholesterol metabolism	*13*
	LCAT Biochemistry	*15*
	LCAT in HDL metabolism	*18*
	LCAT in animal models	*19*
	Human genetic disorders of LCAT	*19*
	LCAT and atherosclerosis	*21*
	Therapeutic regulation of LCAT	*22*
1.2	*INFORMATIC INTRODUCTION*	*24*
	Bioinformatics	*24*
	Potential energy surface	*25*
	Forcefield	*26*
	Energy minimization	*30*
	Classical molecular dynamics simulations	*33*
	Low mode molecular dynamics	*36*
	Protein structure modeling	*37*
	Mutagenesis	*39*
	Molecular docking	*39*
	References	41
2	***AIMS OF THE STUDY***	***47***
	References	50
3	***RESULTS AND DISCUSSION***	***51***
3.1	*LCAT STRUCTURE*	*53*
3.2	*MATERIALS AND METHODS (Theme 1)*	*54*
3.3	*RESULTS AND DISCUSSION (Theme 1)*	*56*
	LCAT homology modeling	*56*
	Definition of the active site	*61*
	Mutations mapping	62
3.4	*CONCLUSIONS (Theme 1)*	*71*
	References	72
3.5	*LCAT MODULATORS*	*74*
3.6	*MATERIALS AND METHODS (Theme 2)*	*75*
3.7	*RESULTS AND DISCUSSION (Theme 2)*	*77*
	High throughput screening results	*77*
	De novo design results	*81*
	In vitro and in vivo *results*	*82*
3.8	*CONCLUSIONS (Theme 2)*	*86*
	References	87
APPENDIX		89
ACKNOWLEDGMENT		91

Chapter 1
State Of The Art

1.1 BIOLOGICAL INTRODUCTION

Atherosclerosis

The term arteriosclerosis denotes a thickening of artery walls. Atherosclerosis, a particular kind of arteriosclerosis, concerns exclusively the large vessels [Ross 1986]. The disease has its own evolution, but it is also the cause of such acute cardiovascular events as myocardial infarction, stroke and peripheral vascular disease [Ross 1993]. Atherosclerosis is induced by an imbalance between noxious stimuli, such as dyslipoproteinemia or hypercholesterolemia, and the healing responses of the artery wall [DePalma 1978].

Complications of atherosclerosis still account for half of the deaths in the industrialized societies. In addition, the quality of life of millions of people is adversely affected by angina and heart failure caused by coronary artery disease (CAD), by intermittent claudication secondary to peripheral vascular disease, and by transient ischemic episodes secondary to cerebrovascular disease. It is for these reasons that so much attention is directed toward understanding the etiology of hyperlipidemia and developing effective therapeutic strategies.

Our current knowledge is based on a large number of animal studies and on many large, randomized, double-blind studies in human beings that prove beyond doubt the cause-and-effect relationship between hypercholesterolemia and morbidity/mortality from CAD, particularly for individuals with multiple risk factors and/or with existing CAD [Brown 1993; Superko 1994; Tyroler 1987]. A control of hyperlipidemia for such individuals is highly effective in reducing risk as demonstrated dramatically in the Scandinavia 4S study, which showed that cholesterol reduction not only decreased CAD mortality by 42% over a 6-years period, but also led to a 30% reduction in overall mortality. There was also a 30% reduction in cerebrovascular events [Scandinavian Simvastatin Survival Study 1994].

Cholesterol metabolism

A primary risk factor predisposing to heart disease is an abnormally elevated level of cholesterol in the blood; cholesterol accumulation in the blood is correlated with development of atherosclerotic plaques.

Cholesterol is only slightly soluble in water, but it is insoluble in blood, and for this reason it is transported in the circulatory system as component of large macromolecular complexes, termed lipoproteins. These lipid::protein complexes are particles with an exterior composed of amphiphilic proteins and lipids whose outward-facing surfaces are water-soluble and inward-facing surfaces are lipid-soluble, allowing movement of apolar lipids through aqueous environments. A total of nine major apolipoproteins are found in human lipoproteins [Mackness 1992]. The standard lipoproteins are classified according to their increasing density: chylomicrons, very-low-density lipoprotein (VLDL), intermediate-density lipoprotein (IDL), low-density lipoprotein (LDL), and high-density lipoprotein (HDL). Because lipids are of much lower density than proteins, the lipid content of a lipoprotein class is inversely related to its density.

In addition to transporting cholesterol through the blood, lipoproteins have cell-targeting signals that direct the lipids they carry to specific tissues.

Despite their differences in lipid and protein composition, all lipoproteins share common structural features, notably a spherical shape that can be detected by electron microscopy (Figure 1.1).

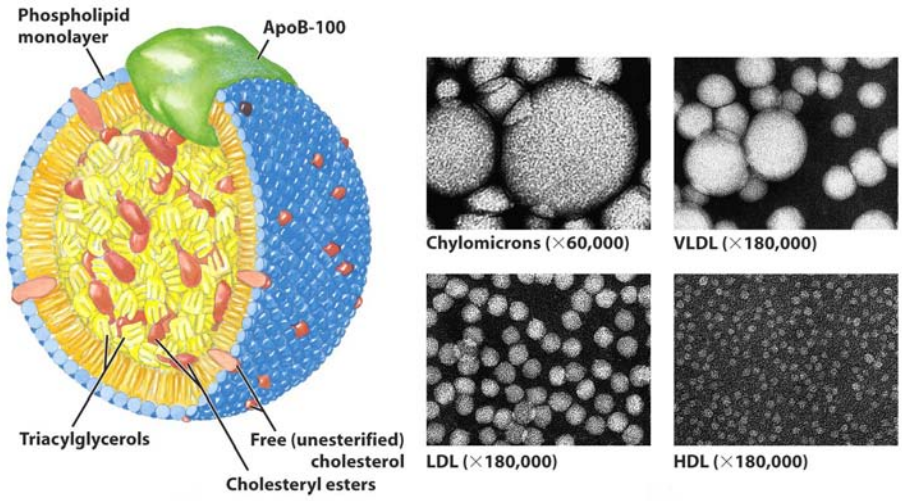


Figure 1.1. Lipoprotein scheme and electron microscopic analysis of lipoprotein fractions [from http://www.utpa.edu/faculty/ahmad/Adv-Biochem/Lipid%20biosynthesis-3_files/frame.htm]

Chylomicrons represent the form through which dietary fat is transported from intestine to peripheral tissue, notably heart, muscle, and adipose tissue. VLDL molecules play a comparable role for triacylglycerols synthesized in liver. The triacylglycerols in both lipoproteins are hydrolysed to glycerol and fatty acids at the inner surfaces of capillaries in the peripheral tissues. This hydrolysis involves activation of extracellular enzyme lipoprotein lipase by apoliprotein C-II.

VLDL are produced by the liver and contain apolipoprotein B100 and apolipoprotein E in their shells. During transport in the bloodstream, the blood vessels cleave and absorb more triacylglycerol from IDL molecules, which contain an even higher percentage of cholesterol. The IDL molecules have two possible fates: half enter metabolism by hepatic lipases, taken up by the LDL receptor on the liver cell surfaces, and the other half continue to lose triacylglycerols in the bloodstream until they form LDL molecules, which contain the highest percentage of cholesterol.

LDL is the principal form through which cholesterol is transported to tissues; each molecule contains approximately 1500 molecules of cholesterol ester.

Each LDL particle contains one molecule of apolipoprotein B100, which is recognized by the LDL receptor in peripheral tissues. Upon binding of apolipoprotein B100, many LDL receptors become localized in clathrin-coated pits. Both the LDL and its receptor are internalized by endocytosis to form a vesicle within the cell. The vesicle then fuses with a lysosome, which has an enzyme called lysosomal acid lipase that hydrolyses the cholesterol esters. Now within the cell, the cholesterol can be used for membrane biosynthesis or esterified and stored within the cell, so as to not interfere with cell membranes. Synthesis of the LDL receptor is regulated by SREBP, the same regulatory protein as is used to control synthesis of cholesterol *de novo* in response to cholesterol presence in the cell. When the cell has abundant cholesterol, LDL receptor synthesis is blocked so new cholesterol in the form of LDL molecules cannot be taken up. On the converse, more LDL receptors are made when the cell is deficient in cholesterol. When this system is deregulated, many LDL molecules appear in the blood without receptors on the peripheral tissues. These LDL molecules are oxidized and taken up by macrophages, which become engorged and form foam cells. These cells often become trapped in the walls of blood vessels and contribute to atherosclerotic plaque formation. Differences in cholesterol

homeostasis affect the development of early atherosclerosis (carotid intima-media thickness). These plaques are the main causes of heart attacks, strokes, and other serious medical problems, leading to the association of so-called LDL cholesterol (actually a lipoprotein) with "bad" cholesterol.
HDL play the primary role in returning excess cholesterol from tissues to the liver for metabolism or excretion, in a process known as reverse cholesterol transport (RCT). Having large numbers of large HDL particles correlates with better health outcomes. In contrast, having small numbers of large HDL particles is independently associated with atheromatous disease progression within the arteries: for this reason, cholesterol in HDL is called “good” [Montgomery 1990; Franceschini 1991].
Cholesterol in plasma lipoproteins exists both as free sterol and as cholesterol esters. Cholesterol esters are considerably more hydrophobic than cholesterol itself. Cholesterol esterification has been described for the first time in 1935 by Sperry [Sperry 1935], but he attributed esters production to plasma heating to 55-60°C. In 1962, Glomset [Glomset 1962] identified an enzyme responsible for cholesterol ester synthesis in plasma: Lecithin:cholesterol acyltransferase (LCAT) (EC2.3.1.43), also known as Phosphatidylcholine-sterol acyltransferase. During the following years, Glomset described the key role of LCAT in the RCT pathway, the process in which excess cholesterol is removed from macrophages in arterial walls by HDL and delivered to the liver for biliary excretion.
After discovery of the LCAT role in cholesterol metabolism, and in the formation and maturation of HDL, interest in the enzyme increased. Currently, numerous animal models lacking or overexpressing LCAT, including mice, hamsters, rabbits and monkeys, have been generated to investigate the role of LCAT in lipoprotein metabolism, in RCT and in atherosclerosis, and over 60 different mutations in *LCAT* gene have been described.

LCAT Biochemistry
The human *Lcat* gene is a 4.5 kb gene localized at the q21-22 region of chromosome 16; it is composed of 6 exons separated by 5 introns, containing 1.5 kb of coding sequence [Jonas 2000]. It was sequenced and cloned for the first time in 1986 by McLean et al. [McLean 1986].
Lcat gene is mainly expressed in the liver but the enzyme is also produced in testes and by astrocytes in brain, where it catalyzes the cholesterol esterification in nascent apoE-containing lipoproteins secreted from glia.
The synthesized protein is secreted into plasma and in cerebral spinal fluid. The plasma concentration of LCAT is about 6 µg/ml and varies in adults with age, gender, alimentary habits and smoking. The half-life of human LCAT in plasma is estimated in 4-5 days.
The mature LCAT contains 416 amino acids, with a molecular mass of approx. 63 kDa, which is more than 20% greater than the predicted protein mass. The extra mass is due to different post-translational modifications: two O-glycosylations on Thr407 and Ser409 consisting of sialylated galactose β, and four N-glycosylations (Asn20, Asn84, Asn272 and Asn384) containing sialylated triantennary and/or biantennary complexes. The carbohydrate content is about 25% of its total mass, with the majority being N-linked. Removal of the carbohydrate moieties from isolated human LCAT by neuraminidase is associated with a 60% increase in the enzymatic activity [Doi 1983], but inhibition of glycosylation in CHO cells reduces the enzymatic activity without affecting LCAT protein secretion [Collet 1991]. Different site-directed mutagenesis studies have been performed to investigate the post-translational modifications. Qu and colleagues [Qu 1993] in transfected COS-6 cells mutated Asn into Thr; they showed that Asn272 is essential for secretion of active LCAT, while Asn84 is critical for its full activity, but not for intracellular processing. In another study in COS-1 cells, Karmin and colleagues [Karmin 1993] substituted Asn with Gln, showing that glycosylation at all four sites is required to generate the full-size mature LCAT protein, but deletion of only one of the

N-linked glycosylation sites does not affect intracellular processing and secretion. N-linked glycosylations are connected with the catalytic activity of the enzyme: substitution of Asn84 or Asn272 with Gln leads to a 82% and 62% decrease in activity, while replacement of Asn384 leads to a substantially increased activity. The pattern of N-linked glycosylation, however, has profound effects on the catalytic activity of the enzyme. Substitutions of Asn84 and Asn272 with Gln result in a loss of activity respectively of 82% and 62%, while replacement of Asn384 leads to a substantially increased activity. Furthermore, N-glycosylation seems to be important for determining substrate specificity towards native HDL and LDL.
The biological significance of the two O-glycosylation sites has not been clarified.

LCAT is responsible for the conversion of free cholesterol into cholesteryl esters. The catalysis is greatly stimulated by the major HDL apolipoprotein, apoA-I, and to a lesser extent by apoC-I, apoE, and apoA-IV.
The LCAT reaction consists in a transesterification, in which a fatty acid at the sn-2 position of phosphatidylcholine, or lecithin, is transferred to the free hydroxyl group of cholesterol, and in the meantime phosphatidylcholine is converted into lysophosphatidylcholines. This occurs in two steps. After binding to a lipoprotein, LCAT cleaves the fatty acid in sn-2 position of phosphatidylcholine and transfers it onto Ser181. This step is mediated by the phospholipase activity of LCAT and depends on the phosphatidylcholine composition. Next, the fatty acid is transesterified to the 3-β-hydroxyl group on the A-ring of cholesterol to form cholesteryl ester (Figure 1.2). Finally, the cholesteryl esters produced, more hydrophobic than free cholesterol, migrate into the hydrophobic core of HDL. However, at an atomic level, the mechanism is not yet well described.
However, at an atomic level, the mechanism is not yet well described.
This reaction would proceed by a chemical mechanism similar to that proposed by Kinnunen et al. for lipoprotein lipase [Kinnunen 1982] and shared by serine-histidine esterases [Brady 1990; Nemukhin 2006].

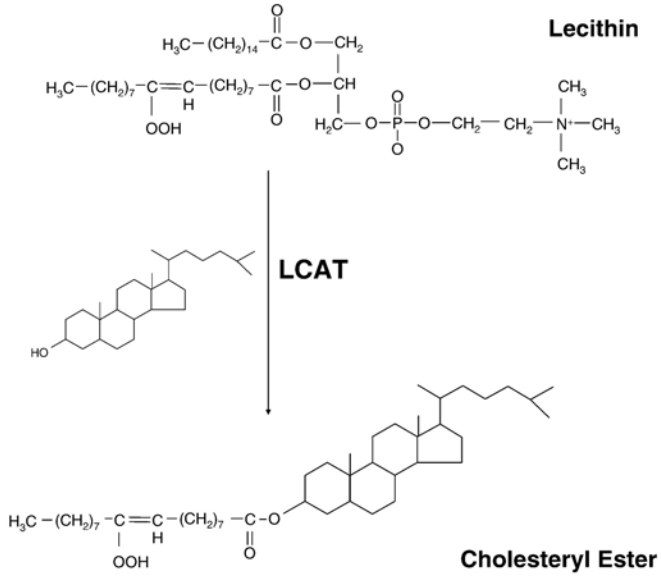


Figure 1.2. Summary of LCAT role: theoretical enzymatic mechanism [from Peter 2007].

In nature, the aptitude to convert esters, amides and peptides has evolved independently on several distinct occasions. The most common structural theme is enzymes with a serine-histidine-aspartate/glutamate triad. So far, similar triads were observed in many proteins, comprising proteases, esterases and lipases and, in some cases, Asp was replaced with Glu. LCAT shares with these enzymes the catalytic triad, consisting of Ser 181, His 244 and Asp 275.
The triad catalytic mechanism involves nucleophilic attack on the carboxyl carbon by Ser Oγ. For transferring the Ser hydroxyl proton to the leaving group, nucleophilicity is enhanced through acid/base catalysis by His, forming an acylenzyme. In the deacylation reaction, His assists the nucleophilic attack on the acylenzyme ester by an acyl acceptor (*e.g.* water or an amine and, in LCAT, cholesterol) by transferring its proton to Ser Oγ. In the reaction, a tetrahedral intermediate is expected. The activation energy barrier for the reaction is lowered by stabilizing interactions between the transition state and the enzyme, notably between the transition state and features of the enzyme forming an "oxyanion hole".
Ser has a crucial role in the catalytic mechanism and is the core residue of the active site. Its Oγ has to be accessible from both the inside and the outside of the protein. This requires the absence of bulky obstructing side chains. For this reason, in its immediate vicinity in the sequence, glycine residues are often observed [Rawlings 1994], such as in α/β hydrolase family members; for instance, in LCAT, we can recognize a GHSGL pattern.
His role is as crucial as that of Ser. Its sequence position is variable, but in three-dimensional space it is almost invariably close to Ser with its Nε2, while forming a hydrogen bond to the carboxylate with its Nδ1 (2.6-3.0 Å in trypsin crystal structures, 2.9 Å in inhibited elastase).
The role of Asp, which is invariably strongly hydrogen-bonded to Nδ1 of His, is to polarize the system and make Ser unusually reactive. In detail, it is important for the stabilization of the positively charged protonated His, stabilizing its correct tautomer and conformation.

Human LCAT preferentially acts on phospholipids containing 18:1 or 18:2 fatty acids, whereas rat and mouse LCAT prefer phospholipids containing 20:4 fatty acids [Grove 1991]. Other phospholipids, such as phosphatidylethanolamine, can also participate in the LCAT reaction, whereas other lipids, such as sphingomyelin, can inhibit LCAT [Bolin 1996; Sparks 1989].
Approximately 75% of plasma LCAT activity is associated with cholesterol on HDL, the referred natural substrate, but LCAT is also able to bind and produce cholesteryl esters on LDL and other apoB-containing lipoproteins. The LCAT activity is thus distinct in α-activity, if LCAT acts on cholesterol bound to apoA-I-containing lipoproteins (HDL particles), and β-activity, if LCAT works on cholesterol bound to apoB-containing lipoproteins (VLDL and LDL particles).
Among other defined structural elements in LCAT are two sets of disulfide bridges, Cys50–Cys74 and Cys313–Cys356. The Cys50–Cys74 disulfide was reported to be essential for the binding of LCAT to lipoprotein surfaces [Baldassarre 2010] and appears to be located in the putative 'lid' region that may cover the active site [Calabresi 2011]. Peelman [Peelman 1998] reported that deletion of amino acids at positions 56 to 68 gave a mutant LCAT that was not active on any substrate.

In different studies [Kosek 1999], some kinetic parameters of LCAT α-activity were collected. The equilibrium dissociation constants (Kd) for the interaction of pure human LCAT with LDL, HDL2, HDL3, and reconstituted discoidal HDL (rHDL) are as follows: rHDL = HDL3 </= HDL2 < LDL, with Km=0.10 mM for HDL3, Km=0.4 mM for HDL2 and Km=0.97 mM for LDL and with relative reactivities (app Vmax/app Km) of 100, 16, 1.3 and 6.5%, respectively.

LCAT in HDL metabolism

LCAT plays a central role in intravascular HDL metabolism and in the determination of plasma HDL levels (Figure 1.3). The LCAT preferred lipoprotein substrate is a newly assembled small, discoidal-shaped HDL generated through the interaction of lipid-free or lipid-poor apolipoprotein A-I with the adenosine triphosphate-binding cassette transporter A1 (ABCA1) located on the plasma membrane, in liver or intestine. These nascent discoidal HDL are called preβ HDL.

Upon its association with the preβ HDL particle, cholesterol is esterified by LCAT and migrates into the hydrophobic core of the lipoprotein, converting the particles into mature, spherical-shaped, and α-migrating HDL (α-HDL) [Nakamura 2004].

Importantly, upon esterification of cholesterol in HDL, LCAT maintains the gradient of free cholesterol between the cellular membrane and the surface of the HDL particle, which is thought to generate a continuous flow of cholesterol from the cell to lipoproteins and prevent the transfer of cholesterol back to the cell.

α-HDL can be converted back to preβ-HDL through the concerted action of the cholesteryl ester transfer protein (CETP) and of a variety of lipases. Preβ HDL have a short plasma half-life, being rapidly cleared through the kidney, whereas mature α-HDL have a much slower turnover [Rye 2004].

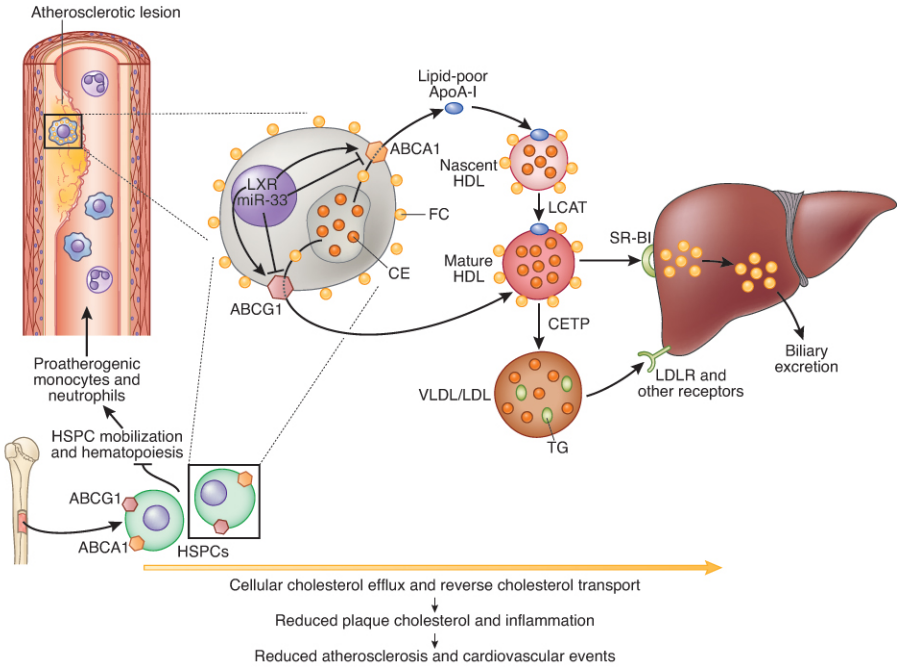


Figure 1.3. Pathways influencing HDL cholesterol metabolism and flux and potential relationship with atherosclerosis [from Rader 2012].

In addition to its function in HDL metabolism, LCAT has also long been believed to play a crucial and critical role in macrophage reverse cholesterol transport (RCT). Through this process, HDL remove excess cholesterol from macrophages within the arterial wall and subsequently deliver it to the liver for biliary excretion - a function that may well explain HDL-mediated atheroprotection. In their paper, Glomset and colleagues [Glomset 1962; John 1970]

described how LCAT activity could contribute to cell cholesterol egress, assuming that "LCAT normally reduces this excess lipid through a combination of direct and indirect effects".
The effect of LCAT on the flux of cholesterol may depend both on the type and metabolic status of the cells, and on the environment of HDL in the extracellular medium.

LCAT in animal models

For several reasons, LCAT role in RCT, and particularly as an atheroprotective factor, is under debate. To better investigate this feature, several experimental systems were developed in various animal models, but they provided inconsistent results.
Human LCAT overexpression in mice remarkably increases plasma HDL levels, but does not enhance macrophage RCT *in vivo* and LCAT-deficient mice display a largely preserved macrophage RCT despite a severe plasma HDL reduction, mostly due to high plasma content of preβ-HDL and enhanced macrophage cholesterol removal via ABCA1 [Tanigawa 2009]. These mouse findings are consistent with early data in humans, showing that HDL are capable of directly delivering large amounts of free cholesterol to the liver, without LCAT-mediated conversion into cholesteryl esters [Calabresi 2009b]. LCAT has also been overexpressed in transgenic rabbits [Hoeg 1996], which - differently from mice - express CETP [Rousset 2009]. As observed in mice, overexpression of LCAT in rabbits also increases HDL but, differently from mice, it decreases LDL and reduces atherosclerosis with respect to control animals. Notably, the double transgenic mice expressing both LCAT and CETP develop less atherosclerosis than LCAT transgenic mice, although more than wild-type mice.
Transient expression of human LCAT in squirrel monkeys with adenovirus also raises HDL and decreases apoB-lipoproteins, due to increased catabolism [Amar 2009].
Studies of mice lacking *LCAT* gene (LCAT KO mice) have also advanced our knowledge of the role of LCAT on HDL metabolism. LCAT KO mice have markedly reduced plasma total cholesterol, cholesteryl esters, HDL, apoA-I, and an increase in plasma triglycerides. The amount of α-HDL is strikingly decreased and the residual HDL is mostly preβ-HDL [Rousset 2009].
Unexpectedly, LCAT deficiency in mice significantly reduces diet-induced atherosclerosis when on a high cholesterol diet, despite causing a marked decrease of HDL [Lambert 2001]. In this case, LCAT deficiency is associated with a significant decrease of apoB-containing lipoproteins.
LCAT also transesterifies and hydrolyzes platelet activating factor and oxidized phospholipids with long chains in the sn-2 position. Thus, LCAT is expected to contribute to the antioxidant or anti-inflammatory properties of HDL.
LCAT deficiency in mice is associated with enhanced insulin sensitivity. Furthermore, it was recently reported that LCAT-deficient mice, especially females, are protected against high-fat, high-sucrose (HFHS) diet-induced obesity. These protective metabolic phenotypes are associated with protection against diet-induced hepatic and adipocyte ER stress, but the mechanistic link with the enzymatic action of LCAT needs further investigation [Li 2011].
Therefore, in addition to its role in RCT, LCAT directly or indirectly interferes with several other physiological processes that might affect the development of atherosclerosis. Overall, the results from the various animal models indicate that there is a complex interaction between LCAT and atherosclerosis, which depends on the diet. It appears, however, that the anti-atherogenic effect of LCAT more closely correlates with its ability to lower plasma levels of apoB-containing lipoproteins than on its ability to raise HDL.

Human genetic disorders of LCAT

Over 60 different mutations in the human *LCAT* gene have been described [Kuivenhoven 1997] (mutations are collected in http://www.lcat.it/database.html), which cause two rare

LCAT deficiency syndromes, namely familial LCAT deficiency [Gjone 1968], FLD (MIM n. 245900) and fish-eye disease [Carlson 1979], FED (MIM n. 136120).
LCAT deficiency is a rare disorder. In general, the prevalence of LCAT mutations in subjects with low HDL cholesterol has been estimated at 2–16% in Finnish and Caucasian Canadian patients, respectively. In the recently published results of the Copenhagen City Heart Study in which the regulatory and coding regions of LCAT were re-sequenced in individuals with the 2% lowest (n=180) plasma HDL cholesterol levels, no carriers of loss-of-function mutations in LCAT were identified, indicating that these are extremely rare in the general population. In the Netherlands, however, functional mutations in *LCAT* were found in almost a third (29%) of patients with low HDL cholesterol, thus constituting a common cause of low HDL cholesterol in referred patients in this country. Interestingly, a recent genome-wide association study (GWAS) in more than 100,000 individuals of European ancestry identified a single-nucleotide polymorphism (SNP) in *LCAT* as the strongest marker of isolated variation in HDL cholesterol levels [Haase 2010].
FLD was reported for the first time in 1967 in a Norwegian family [Norum 1967]. In this family, three adult sisters showed extremely low plasma LCAT activity, reduced levels of cholesteryl esters, and reduced plasma LCAT concentration. Years later it was established that homozygosis for a single nucleotide substitution in codon 252 of exon 6 in the gene, leading to the exchange of a methionine (ATG) for a lysine residue (AAG), was responsible for the LCAT deficiency in this family.
In FLD mutants, LCAT production is abolished or the synthetized enzyme lacks both α-LCAT and β -LCAT activity; instead, in FED mutants, α -LCAT activity is missing, but β -LCAT activity is preserved. Some LCAT mutations have been shown to selectively affect LCAT activity on HDL, but not all mutations can be neatly categorized as affecting only the esterification of cholesterol on HDL or LDL, suggesting that some patients with FED may differ from FLD by having more residual LCAT activity on both HDL and LDL [Klein 1995].
Both conditions are characterized by low HDL and corneal opacities, but FLD subjects have a more severe deficiency of LCAT and can develop other signs and symptoms. In detail, homozygous FLD cases have very little cholesteryl esters in plasma and unesterified cholesterol accumulates in all plasma lipoprotein fractions; in addition to the corneal opacity, they reportedly present anemia, and many of them unpredictably develop renal disease. The FED cases have subnormal plasma cholesteryl ester/unesterified cholesterol ratios and a milder clinical phenotype. Heterozygous carriers have an intermediate lipid/lipoprotein phenotype, with a significant LCAT gene-dose-dependent effect on plasma cholesterol esterification, lipid, and apolipoprotein levels [Carlson 1985].
However, the differential diagnosis cannot be defined *a priori*; to discriminate FLD and FED pathology in carriers of two mutant LCAT alleles, it is mandatory to measure the ability of plasma to esterify cholesterol incorporated into endogenous lipoproteins (α-LCAT and β -LCAT activity) and into a standard HDL substrate (α-LCAT).
In 1997, Kuivenhoven et al. proposed a new classification system for natural LCAT mutations based on the different biochemical and clinical phenotypes observed in LCAT-deficient patients as well as on the biochemical characteristics of mammalian cell cultures expressing the mutants of the *LCAT* gene [Kuivenhoven 1997]. Importantly, this new system also allows the classification of heterozygous mutations. Intermediate phenotypes, *i.e.* due to mutations not causing a total loss of LCAT activity (FLD), but causing more than "just" partial loss of LCAT activity against HDL (FED) are also taken into account. For the *in vivo* classification, five criteria were proposed: 1) LCAT activity towards HDL analogues; 2) cholesterol esterification rate (endogenous lipoproteins); 3) ratio of plasma FC to EC; 4) plasma LCAT concentration; 5) clinical symptoms of the disease. For the *in vitro* experiments, two criteria were proposed: 1) specific activity of the mutant protein towards HDL analogues and native LDL, and 2) activity

of the mutant protein towards heat-inactivated plasma. The classification system describes five classes. The first class contains null mutations of the *LCAT* gene. This means that patients in this class display a total loss of catalytic activity of LCAT and that they have the clinical phenotype of FLD. The second class contains missense mutations that cause complete or nearly complete loss of catalytic activity of the *LCAT* gene. The third class contains both missense mutations and minor deletions in the *LCAT* gene that are responsible for an intermediate phenotype, meaning that there is either partial loss of activity against LDL or combined partial loss of activity against both HDL and LDL. This class thus also includes patients that show FED symptoms, but do not develop all the symptoms of the FLD syndrome. The fourth class contains the mutations responsible for the typical symptoms of the FED syndrome. This class thus includes missense mutations that result in specific loss of activity against HDL analogues, but activity against LDL or other apoB-containing lipoproteins is preserved. The fifth and last class contains three mutations that the group of Kuivenhoven et al. was unable to categorize; this category was therefore named "unclassified mutations".

LCAT and atherosclerosis

As LCAT was considered the main driving force in the reverse cholesterol transport pathway, it was soon thought that the enzymatic activity of LCAT could be involved in the protection against atherosclerotic lesion formation [Rousset 2009; Calabresi 2010]. In 1973, Hovig and Gjone demonstrated lipid deposition in renal arteries and veins of patients with FLD [Hovig 1973]. Furthermore, the spleen contains numerous lipid-laden cells, which are assumed to be partly responsible for the splenomegaly found in FLD patients. Notably, only 35% of the total cholesterol is esterified in arterial lipid depositions of FLD patients, as compared to 75 % in normal atheromas. In 1982, Carlson showed that FED patients did not suffer from premature atherosclerosis in spite of their extremely low HDL cholesterol levels [Carlson 1982]. This was surprising, as FED patients have very low levels of HDL and it thus was expected that these patients would have an increased risk for atherosclerosis.

Ever since then, the effect of LCAT on human atherogenesis has been controversial. Early cross-sectional studies have reported either decreased or increased LCAT activity in patients with angiographically proven CAD; moreover, an enhanced cholesterol esterification rate was shown to be a strong predictor of the presence of angiographically proven atherosclerotic lesions.

Recently, a relatively large study on carriers of LCAT defects has reported not only reduced HDL but also a marked increase in C-reactive protein and in intima/media thickness (IMT) of the artery [Calabresi 2009a]. No significant change in IMT was observed in homozygotes, but an increased incidence of CHD was reported when heterozygotes were compared with controls. Similar findings for heterozygous subjects were observed in a 25 years follow-up study of a large Canadian LCAT-deficient family and in 13 unrelated Italian families with FLD and FED. These results suggest that, while heterozygosis for LCAT deficiency is associated with increased IMT and CHD, this may not be true for homozygous subjects; however, this puzzling result could potentially be explained by the low number of homozygous subjects studied. An alternative explanation is that homozygous FLD and FED patients may be partially protected from their low HDL, because they often also have reduced levels of LDL compared to heterozygotes and controls.

LCAT is not an extensively polymorphic protein and only a few studies examining genetic variants of the *LCAT* gene in the general population have been carried out.

Recently, a study reported greater IMT and elevated LCAT activity in subjects with metabolic syndrome, suggesting that higher LCAT activity may not be beneficial [Dullaart 2008]. A similar positive association of LCAT was also observed in patients with angiographically proven CHD [Wells 1986]. These results are in contrast, however, with several earlier papers

describing either a negative correlation or no association between LCAT activity and CHD. These seemingly contradictory results may potentially be explained by the fact that most of these studies are relatively small and do not examine the other proteins and enzymes in the RCT pathway, which can potentially alter the effect of LCAT on atherosclerosis. For example, low LCAT activity when also linked with elevated levels of pre-beta HDL was associated CHD [Miida 1996; Savel 2012]. Another possible explanation for these contradictory results is that there may be other biochemical markers, such as the fractional esterification rate of apoB-depleted plasma (FER_{HDL}), that are better than the *in vitro* LCAT activity assay for assessing the HDL maturation process [Frohlich 2003]. Finally, it is important to note that it is impossible to determine from epidemiologic studies whether LCAT is playing a causal role in promoting or decreasing atherosclerosis or instead may be up or down regulated by some sort of compensatory response.

Therapeutic regulation of LCAT

Therapeutic up-regulation of LCAT function has gained interest in the recent years, not only as enzyme replacement therapy for LCAT deficiency syndromes, but also as a potential new therapeutic strategy for reducing atherosclerosis. Strategies for therapeutically raising LCAT activity include recombinant LCAT protein administration, viral expression of LCAT, and small molecule activators of LCAT.

Intravenous infusion of recombinant LCAT in LCAT knockout mice with or without expression of human apoA-I rapidly raises HDL cholesterol and relieves other lipid abnormalities. Moreover, a preliminary report indicates that subcutaneous injection of recombinant LCAT stimulates reverse cholesterol transport and attenuates atherosclerosis progression in New Zealand White rabbits [Zhou 2009].

Kuroda and colleagues have focused on developing a long-lasting LCAT replacement therapy via transplantation of human *LCAT* gene-transduced autologous adipocytes [Kuroda 2012]. LCAT from the transduced adipocytes improves the abnormal HDL particles from a FED patient *in vitro*. Furthermore, LCAT can be steadily detected in adipocyte-transplanted mice four weeks after transplantation. Lastly, the therapeutic potential of a small molecule activator of LCAT, compound A, is being explored for the treatment of atherosclerosis. Compound A increases LCAT activity at a micromolar concentration, by interacting with the free sulfhydryl group in Cys31 near the catalytic site of LCAT. Cys 31 is a conserved residue in multiple species and, on purely theoretical grounds, compound A is able to activate LCAT from multiple species, including mouse, hamster, rhesus monkey, and man. Intraperitoneal administration of 20 mg/kg of compound A increases HDL cholesterol acutely in C57Bl/6 mice and in high-fat diet fed Syrian golden hamsters, while non-HDL cholesterol and triglycerides are reduced [Chen 2012]. Also chronic daily administration of 20 mg/kg and 60 mg/kg via oral gavage into high-fat diet fed Syrian golden hamsters leads to a dose-dependent increase in HDL cholesterol. VLDL cholesterol is decreased at the dose of 20 mg/kg, but no further decrease is seen after administration of 60 mg/kg. Gall bladder bile acids at termination are increased 2-fold, indicative of enhanced RCT upon chronic treatment with the LCAT activating compound A. The effects of these studies with respect to the generation of a more anti-atherogenic lipoprotein profile look promising.

Another interesting strategy for atherosclerosis therapy has been focused on CETP modulation: many studies suggested that CETP inhibition could exert a cardio-protective effect, specifically stimulating cell efflux [Barter 2007a; Calabresi 2009b; Niesor 2010]. Zhen and colleagues recently applied the adeno-associated viral vector serotype 8 (AAV8) for liver-directed delivery of human LCAT in heterozygous LDL receptor knockout mice expressing CETP. AAV8-hLCAT administration resulted in a hLCAT concentration of 300 μg/mL, which

declined slightly over the course of the experiment to 220 μg/mL, a level which is estimated to be 20-fold higher than the physiological concentration of LCAT. The mice showed a markedly raised HDL cholesterol and increased particle size, while LDL cholesterol, plasma triglycerides, and plasma apoB were reduced.

Nicholls and colleagues [Nicholls 2011] tested the effect of CEPT inhibitors in patients with dyslipidemia and demonstrated that, compared with placebo or statin in monotherapy, these inhibitors could increase HDL-C and decrease LDL-C levels. Zanotti and colleagues obtained similar results using a different CETP inhibitor [Zanotti 2011]. However, this therapeutic strategy was found to be unsuccessful in early clinical trials, and it was discontinued due to off-target adverse effects.

Nevertheless, considering the complex interaction of LCAT with blood lipoproteins, extensive studies on the effects on atherosclerosis susceptibility should be performed in the future to be able to draw any sound conclusion on the therapeutic applicability of these new strategies.

1.2 INFORMATICS INTRODUCTION

Bioinformatics

Bioinformatics is a field of science in which biology, computer science, and information technology merge into a single discipline to analyse biological information using computers and statistical techniques. Major applications of bioinformatics include sequence alignment, genome assembly, protein structure prediction, protein-protein and protein-ligand docking, gene expression prediction, and molecular dynamics simulations.

Molecular modeling indicates the general process of describing and studying complex chemical systems in terms of a realistic atomic model, with the aim to understand, predict and simulate macroscopic properties based on detailed knowledge at an atomic scale. The knowledge of the three-dimensional structure of a protein of interest is useful to plan experiments aimed at elucidating its function or at developing compounds that specifically interact with it (drug design).

Macroscopic physical properties can be distinguished in:

- static equilibrium properties, such as the binding constant of an inhibitor to an enzyme, the average potential energy o f a system, or the radial distribution function in a liquid;
- dynamic or non-equilibrium properties, such as the viscosity of a liquid, diffusion processes in membranes, the dynamics of phase changes, reaction kinetics, or the dynamics of defects in crystals.

The choice of the specific technique to be used depends on the question asked and on the feasibility of the method to yield reliable results at the present state of the art. Ideally, the time-dependent Schrödinger equation describes the properties of molecular systems with high accuracy, but anything more complex than the equilibrium state of a few atoms cannot be handled at this *ab initio* level. Thus approximations are necessary; the higher the complexity of a system and the longer the time span of the processes of interest is, the more severe the required approximation are. At a certain point the *ab initio* approach must be augmented or replaced by empirical parameterization of the model used. Where simulations based on physical principles of atomic interactions still fail, due to the complexity of the systems, molecular modeling is based entirely on a similarity analysis of know structural and chemical data. The QSAR methods (Quantitative Structure-Activity Relations) and many homology-based protein structure predictions belong to the latter category.

Macroscopic properties are always ensemble averages over a representative statistical ensemble (either equilibrium or non-equilibrium) of molecular systems. For the generation of a representative equilibrium ensemble two methods are available: Monte Carlo simulations and molecular dynamics (MD) simulations. For the generation of non-equilibrium ensembles and for analysis of dynamic events, only the second method is appropriate. While Monte Carlo simulations are simpler than MD (they do not require the computations of forces), they do not yield significantly better statistics than MD in a given amount of computer time. Therefore MD is a more universal technique. If a starting configuration is very far from equilibrium, the forces may be excessively large and the MD simulation may fail. In those cases a robust energy minimization (EM) is required. Another reason to perform an energy minimization is the removal of kinetic energy from the systems: if several “snapshots” from a dynamical simulation must be compared, energy minimization reduces the thermal noise in the structures and potential energies, so that they can be compared better [Van der Spoel et al., 2006].

Potential energy surface

The fundamental computation at the core of a forcefield-based simulation is the calculation of the potential energy for a given configuration af atoms. The calculation of this energy, along with its first and second derivatives with respect to the atomic coordinates, yields the information necessary for minimization, harmonic vibrational analysis, and dynamics simulations. This calculation is actually performed by the simulation engine, or forcefield-based program. Simulation engines are the computational packages that handle the application of forcefields in minimization, dynamics, and other molecular mechanics simulations.

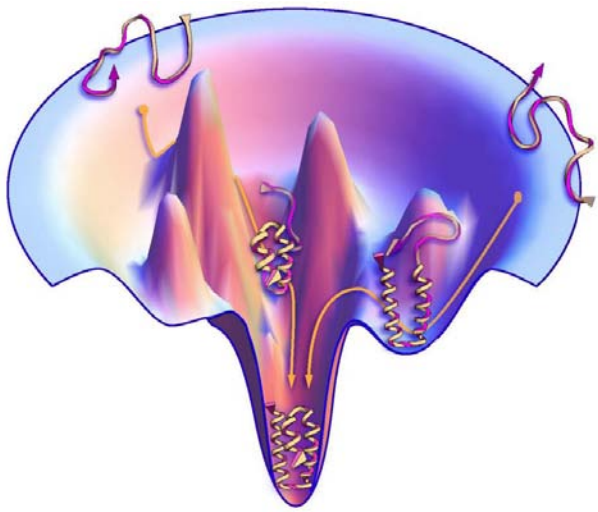


Figure 1.4. Schematic potential energy funnel for proteins folding (modified from Stony Brook University, 2012)

The functional form of the potential energy expression and the entire set of parameters needed to fit the potential energy surface constitute the forcefield [Ermer, 1976]. The energy expression is the specific equation set up for a particular model and including (or not) any optional terms. For example, a forcefield would contain bond-stretching parameters for all combinations of atoms for which it was parameterized, as well as defined, summed functional forms for the bond-stretching term. The corresponding energy expression would contain bond-stretching parameters for only those combinations of bonded atoms actually found in the model being studied, as well as the specific bond-stretching terms for the number and types of bonds in that model.

It is important to understand that the forcefield (both the functional form and the parameters themselves) represents the single largest approximation in molecular modeling. The quality of the forcefield, its applicability to the model at hand, and its ability to predict the particular properties measured in the simulation directly determine the validity of the results. Molecular modeling programs are the graphical user interfaces that can be used to prepare models, set up forcefield, and access the simulation engines.

The complete mathematical description of a molecule, including both quantum mechanical and relativistic effects, is a formidable problem, due to the small scale and large velocity. However these intricacies can be ignored, because molecular mechanics and dynamics are based on empirical data that implicitly incorporate all the relativistic and quantum effects. Since no

complete relativistic quantum mechanical theory is suitable for the description of molecular systems, the non-relativistic, time-dependent form of the Schrödinger description is used:

$$H\Psi(R,r) = E\Psi(R,r) \tag{1.1}$$

where H is the Hamiltonian for the systems, Ψ is the wavefunction, and E is the energy. In general, Ψ is a function of the coordinates of the nuclei (R) and the electrons (r). Although this equation is quite general, it is too complex for any practical use, so approximations are made. Noting that the electrons are several thousands of times lighter than nuclei and therefore move much faster, Born and Oppenheimer proposed what is know as Born-Oppenheimer [Born & Oppenheimer, 1927] approximation: the motion of the electrons can be decoupled from that of the nuclei, giving two separate equations. The first equation describes the electronic motion:

$$H\psi(R;r) = E\psi(r;R) \tag{1.2}$$

and depends only parametrically on the position of the nuclei. Note that this equation defines an energy $E(R)$, which is a function of only the coordinates of the nuclei. This energy is usually called the potential energy surface. The second equation the describes the motion of the nuclei on this potential energy $E(R)$:

$$H\phi(R) = E\phi(R) \tag{1.3}$$

The direct solution of Eq 1.2 is the solution of *ab initio* quantum chemical codes. Semiempirical codes also solve Eq. 1.2, but they approximate many of the integrals needed with empirically fit functions. The common feature of these programs, though, is that they solve for the electronic wavefunction and energy as a function of nuclear coordinates. In contrast, simulation engines provide an empirical fit to the potential energy surface [MSI, 1998].

Solving Eq 1.3 is important if one is interested in the structure or time evolution of a model. Eq 1.3 is the Schrödinger equation for the motion of the nuclei on the potential energy surface. In principle, Eq 1.2 could be solved for the potential energy E, and then Eq 1.3 could be solved. However, the effort required to solve Eq 1.2 is extremely large, so usually an empirical fit to potential energy surface, commonly called a forcefield (V) is used. Since the nuclei are relatively heavy objects, quantum mechanical effects are often insignificant, in which case Eq 1.3 can be replaced by Newton's equation of motion:

$$-\frac{dV}{dR} = m\frac{d^2R}{dt^2} \tag{1.4}$$

The solution of Eq 1.4 using an empirical fit to the potential energy surface E(R) is called molecular dynamics. Molecular mechanics ignores the time evolution of the system and instead focuses on finding particular geometries and their associated energies or other static properties. This includes finding equilibrium structures, transition states, relative energies, and harmonic vibrational frequencies [MSI, 1998].

Forcefield

The forcefield contains the necessary building blocks for the calculations of energy and force:

- a list of atom types;
- a list of atomic charges;
- atom-typing rules;

- functional forms for the components of the energy expression;
- parameters for the function terms;
- for some forcefield, rules for generating parameters that have not been explicitly defined;
- for some forcefield, a defined way of assigning functional forms and parameters.

The forcefield commonly used for describing molecules employs a combination of internal coordinates and terms (bond distances, bond angles, torsions, etc.), to describe part of the potential energy surface due to interactions between bonded atoms, and nonbond terms. The functional forms range from simple quadratic forms to Morse functions, Fourier expansions, Lennard-Jones potentials, etc.

The goal of the forcefields is to describe entire classes of molecules with reasonable accuracy. In a sense, the forcefield interpolates and extrapolates form the empirical data of the small set of models used to parameterize the forcefield to a larger set of related models. Some forcefields aim for high accuracy for a limited set of element types, thus enabling good prediction of many molecular properties. Other forcefields aim for the broadest possible coverage of the periodic table, with necessarily lower accuracy.

The physical significance of most of the types of interactions in a forcefield is easily understood, since describing a model's internal degrees of freedom in terms of bonds, angles, and torsions seems natural. The analogy of vibrating balls connected by springs to describe molecular motion is equally familiar. However, it must be remembered that such models have limitations. Consider for example the difference between such a mechanical model and quantum mechanical bonds.

Covalent bonds can, to a first approximation, be described by the harmonic oscillator in Figure 1.5.

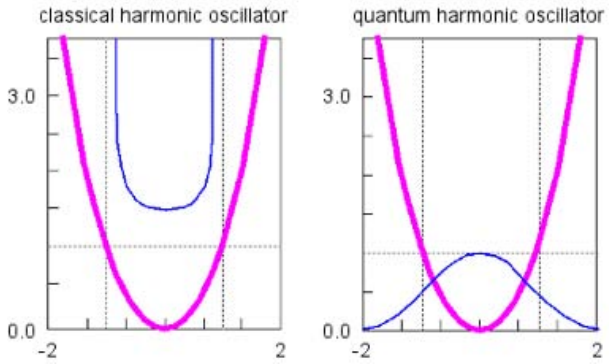


Figure 1.5. Energy and probability of a mechanical and quantum particle in a harmonic energy well. The energy is indicated by the heavy lines and probability by the thin lines. The total energy of the system is indicated by the pale horizontal line. The classical (mechanical) probability is highest when the particle reaches it maximum potential energy (zero velocity) and drops to zero between these points. The quantum mechanical probability is highest where the potential energy is lowest, and there is a finite probability that the particle can be found outside the classical limits (pale vertical lines) [from MSI, 1998].

A ball poised at the intersection of the light horizontal line with the parabolic energy surface would begin to roll down, converting its potential energy to kinetic energy until, exactly at the same height as it had started, it would pause momentarily before rolling back. The interchange of kinetic and potential energy in such a mechanical system is intuitive.

The probability of finding the ball at any point along its trajectory is inversely proportional to its velocity at that point (which is opposite to the probability for a real atom). The probability is plotted above the parabolic curve. The probability is greatest near the high-energy limits of its

trajectory (where it is moving quickly), because the total energy cannot exceed the initial potential energy defined by the intersection of the total energy with the parabola.
Describing a quantum mechanical "trajectory" is impossible [Heisemberg, 1930], because the uncertainty principle prevents an exact, simultaneous specification of both position and momentum. However, the probability that the quantum mechanical ball will be at a given point on the parabola can be quantified. The quantum mechanical probability function plotted in the right panel of Figure 1.5 is very different from the mechanical systems. First, the highest probability is at the energy minimum, which is the opposite of the mechanical case. Second, the quantum mechanical ball can actually be found beyond the classical limits imposed by total energy of the system (tunneling). Both these properties can be attributed to the uncertainty principle.
With such a different qualitative picture of fundamental physical principles, it is reasonable to use a mechanical approach for obviously quantum mechanical entities like bonds? In practice, many experimental properties such as vibrational frequencies, sublimation energies, and crystal structures can be reproduced with a forcefield, not because the systems behave mechanically, but because the forcefield is fit to reproduce relevant observables and therefore includes most of the quantum effects empirically. Nevertheless it is important to appreciate the fundamental limitations of a mechanical approach.
Applications beyond the capability of most forcefield methods include:

- electronic transitions (photon absorption);
- electron transport phenomena;
- proton transfer (acid/base reactions).

The true power of the atomistic description of a model embodied in the energy expression lies in the three major areas:

- the first is that forcefield-based simulations can handle large systems, since these simulations are several orders of magnitude faster (and cheaper) than quantum-based calculations. Forcefield-based simulations can be used for studying condensed-phase molecules, macro-molecules, crystal morphology, inorganic and organic interphases, etc., where the properties of interest are not sensitive to quantum effects (e.g., phase behaviour, equations of state, bond energies, etc);
- the second is the analysis of the energy contributions at the level of individual interactions, or classes thereof. For instance, it is possible to decompose the energy into bond energies, angle energies, nonbond energies, etc. or even to the level of a specific hydrogen bond or van del Waals contact, in order to understand a physical observable or to make a prediction;
- the third area, which is described as application of constraints and restraints, lies in the modification of the energy expression to bias the calculation. One can impose constraints (absolute conditions), such as fixing an atom in space and not allowing it to move. The user can also add extra terms to the energy expression to restraint or force the system in certain ways. For instance, by adding an extra torsion potential to a particular bond, it is possible to force the torsion angle toward a desired value. (One can apply constraints also for quantum-based energy calculations) [MSI, 1998].

The energy expression
The actual coordinates of a model combined with the forcefield data create the energy expression (or target function) for the model. This energy expression is the equation that describes the potential energy surface of a particular model as a function of its atomic coordinates. The potential energy of a system can be expressed as a sum of valence (or bond), cross-term, and non-bond interactions.

$$E_{total} = E_{valence} + E_{crossterm} + E_{nonbond} \tag{1.5}$$

The energy of valence interactions is generally accounted for by diagonal terms, namely, bond stretching (E_{bond}), valence angle vending (E_{angle}), dihedral angle torsions ($E_{torsion}$) and inversion ($E_{inversion}$ o E_{oop}) terms, which are part of nearly all forcefields for covalent systems. A Urey-Bradley term (E_{UB}) may be used to account for interactions between atom pairs involved in 1-3 configurations (*i.e.,* atoms bound to a common atom):

$$E_{valence} = E_{bond} + E_{angle} + E_{torsion} + E_{oop} + E_{UB} \tag{1.6}$$

Modern (second-generation) forcefields generally achieve higher accuracy by including cross-terms to account for such factors as bond or angle distortions caused by nearby atoms. Cross-terms can include the following terms: stretch-stretch, stretch-bend-stretch, bend-bend, torsion-stretch, torsion-bend-bend, bend-torsion-bend, stretch-torsion-stretch.
The energy of interaction between non-bonded atoms is accounted for by van der Waals (E_{vdW}), electrostatic ($E_{Coulomb}$), and (on some older forcefields) hydrogen bond (E_{hbond}) terms:

$$E_{nonbond} = E_{vdW} + E_{Coulomb} + E_{hbond} \tag{1.7}$$

Restraints that can be added to an energy expression include distance, angle, torsion, and inversion restraints. Restraints are useful if one is, for example, interested in the structure of only part of a model. As a simple example of a complete energy expression, consider the following equation, which might be used to describe the potential energy surface of a water model:

$$\mathrm{V(R)} = K_{oh}(b - b_{oh}^0)^2 + K_{oh}(b' - b_{oh}^0)^2 + K_{hoh}(\Theta - \Theta_{hoh}^0)^2 \tag{1.8}$$

Where K_{oh}, b_{oh}^0, K_{hoh} and Θ_{hoh}^0 are parameters of the forcefield, b is the current bond length of one O-H bond, b' id the length of the other O-H bond, and O is the H-O-H angle. In this example, the forcefield defines:

- the coordinates to be used (bond lengths and angles)
- the functional form (a simple quadratic in both types of coordinates)
- the parameters (the force constants K_{oh} and K_{hoh}, as well as the reference values b_{oh}^0 and Θ_{hoh}^0)

The reference O-H bond length and reference H-O-H angle are the values for an ideal O-H bond and H-O-H angle at zero energy, which is not necessarily the same as their equilibrium values in a real water molecule.
Eq 1.8 is an example of an energy expression as set up for a simple molecule. Eq 1.9 is an example of the corresponding general, summed forcefield function:

$$\mathrm{V(R)}= \sum_b D^b \left[1\text{-}e^{-a(b\text{-}b_0)}\right]^2 + \sum_\Theta H_\Theta\,(\Theta\text{-}\Theta_0)^2 + \sum_\Theta H_\Theta\,[1 + s\cos(n\Phi)] +$$
$$+ \sum_\chi H_\chi \chi^2 + \sum_b \sum_{b'} F_{bb'}\,(b\text{-}b_0)\left(b'\text{-}b'_0\right) + \sum_\Theta \sum_{\Theta'} F_{\Theta\Theta'}\,(\Theta\text{-}\Theta_0)\left(\Theta'\text{-}\Theta'_0\right) +$$
$$+ \sum_b \sum_\Theta F_{b\Theta}\,(b\text{-}b_0)(\Theta\text{-}\Theta_0) + + \sum_\Theta \sum_{\Theta'} F_{\Theta\Theta'\Phi}\,(\Theta\text{-}\Theta_0)\left(\Theta'\text{-}\Theta'_0\right)\cos\Phi +$$

$$+\sum_{\chi}\sum_{\chi'}F_{\chi\chi'}\chi\chi'+\sum_{i}\sum_{j>i}\left[\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}+\frac{q_iq_j}{r_{ij}}\right] \tag{1.9}$$

The first four terms in this equation are sums that reflect the energy needed to stretch bond (b), bend angle (Θ) away from their reference values, rotate torsion angle (Φ) by twisting atoms about the bond axis that determines the torsion angle, and distort planar atoms out of the plane formed by the atoms they are bonded to (χ). The next five terms are cross-terms that account for interactions between the four types of internal coordinates. The final term represents the non-bond interactions as a sum of repulsive and attractive Lennard-Jones terms as well as Coulombic terms, all of which are a function of the distance r_{ij}between atom pairs. The forcefield defines the functional form of each term in this equation as well as the parameters such as D_b, α, and b_0. The forcefield also defines internal coordinates such as b, Θ, and χ as a function of the Cartesian atomic coordinates, although this is not explicit in Eq 1.9. We should note that the energy expression in Eq 1.9 is cast in a general form. The true energy expression for a specific model includes information about the coordinates that are included in each sum. For example, it is common to exclude interactions between bonded and 1-3 atoms in the summation representing the non-bond interactions. Thus a true energy expression might actually use a list of allowed interactions rather than the full summation implied in Eq 1.9 [MSI, 1998].

Many types of forcefield

Forcefields can be divided into the following categories:

- classical, first-generation forcefield applicable mainly to biochemistry;
- second-generation forcefields capable of predicting many properties;
- rule-based forcefields applicable to a broad range of elements in the periodic table;
- specific-purpose forcefields that are narrowly applicable to particular applications or types of models.

Energy minimization

The potential energy function of a (macro)molecular system is a very complex landscape in a large number of dimensions. For example, the energy of a conformation of ethane is a function of the 18 internal coordinates or 24 Cartesian coordinates that are required to completely specify the structure. The way in which the energy varies with the coordinates is referred to as the potential energy surface. For a system with N atoms the energy is thus a function of $3N-6$ internal or $3N$ Cartesian coordinates. It is therefore impossible to visualize the entire potential energy surface except for some simple cases where the potential energy is a function of just one or two coordinates. For example, the van der Waals energy of two argon atoms depends upon just one coordinate: the interatomic distance. Sometimes the user may wish to visualize just a part of the potential energy surface. For example, suppose one takes an extended conformation of pentane and rotate the two central carbon-carbon bonds so that the torsion angles vary from 0° to 360°, calculating the potential energy of each structure generated. The potential energy in this case is a function of just two variables and can be plotted as a contour diagram or as a isometric plot, as shown in Figure 1.6.

The potential energy surface has one deepest point, the global energy minimum, associated with the lowest energy, and a very large number of local minima. Minimum energy arrangements of the atoms correspond to stable states of the systems; any movement away from a minimum gives a configuration with a higher energy. At the minima all derivatives of the potential energy function with respect to the coordinates are zero and all second derivatives

are non-negative. The matrix of second derivatives, which is called Hessian matrix, has non-negative eigenvalue; only the collective coordinates that correspond to translation and rotation (for an isolated molecule) have zero eigenvalue. In between the local minima there are saddle points, where the Hessian matrix has only one negative eigenvalue. These points are the mountain passes through which the systems can migrate from one local minimum to another.

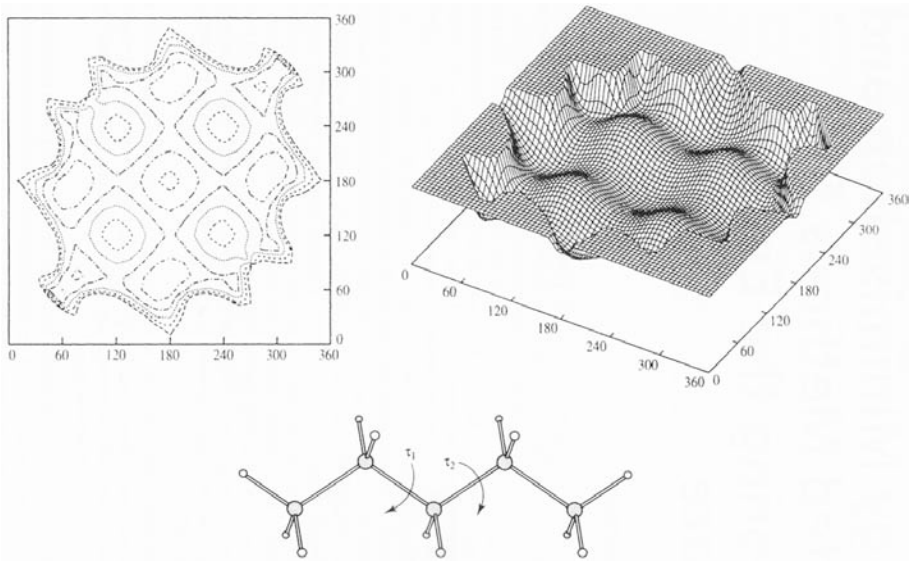


Figure 1.6. Variation in the energy of pentane with the two torsion angles indicated and represented as a contour diagram and isometric plot. Only the lowest-energy regions are shown. (from Leach, 2001)

Knowledge of all local minima, including the global one, and of all saddle points would enable us to describe the relevant structures and conformations and their free energies, as well as the dynamics of structures transitions. A minimization algorithm is used to identify those geometries of the system. Unfortunately, the dimensionality of the configurational space and the number of the local minima is so high that it is impossible to sample the space at a sufficient number of points to obtain a complete survey. In particular, no minimization method exists that guarantees the determination of the global minimum in any practical amount of time. However, given a starting configuration, it is possible to find the nearest local minimum. Nearest in this contest does not always imply nearest in a geometrical sense (*i.e.* the least sum of square coordinate differences), but means the minimum that can be reached by systematically moving down the steepest local gradient [Van der Spoel et al., 2006].
The minimization problem can be formally stated as follows: given a function f which depends on one or more independent variables $x_1, x_2, \ldots, x_i$, find the value of those variables where f has a minimum value. At the minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are positive:

$$\frac{\partial f}{\partial x_i} = 0 \qquad\qquad \frac{\partial^2 f}{\partial x_i^2} > 0 \tag{1.10}$$

The functions of most interest to us will be the quantum mechanics or molecular mechanics energy with the variables x_i being the Cartesian or the internal coordinates of the atoms. Molecular mechanics minimizations are nearly always performed in Cartesian coordinates,

with which the energy is a function of $3N$ variables; it is more common to use internal coordinates with quantum mechanics. For analytical functions, the minimum of a function can be found using standard calculus methods. However, this is not generally possible for molecular systems due to the complicated way in which the energy varies with the coordinates. Rather, minima are located using numerical methods, which gradually change the coordinates to produce configurations with lower and lower energies until the minimum is reached.
Minimization algorithms can be classified into two groups: those, which use derivatives of the potential energy with respect to the coordinates and those, which do not. Derivatives can be useful because they provide information about the shape of the potential energy surface, and, if used properly, they can significantly enhance the efficiency with which the minimum is located. There are many factors that must be taken into account when choosing the most appropriate algorithm (or combination of algorithms) for a given problem; the ideal minimization algorithm is one that provides the answer as quickly as possible, using the least amount of memory. No single minimization method has yet proved to be the best for all molecular modeling problems. In particular, a method that works well with quantum mechanics may not be the most suitable for use with molecular mechanics. This is party because quantum mechanics is usually used to model systems with fewer atoms than molecular mechanics; some operations that are integral to certain minimization procedures (such as matrix inversion) are trivial for small systems but formidable for systems containing thousands of atoms. Quantum mechanics and molecular mechanics also require different amounts of computational effort to calculate the energies and the derivatives of the various configurations. Thus an algorithm that takes many steps may be appropriate for molecular mechanics but inappropriate for quantum mechanics.
Most minimization algorithms can only go downhill on the potential energy surface and so they can only locate the minimum that is nearest (in a downhill sense) to the starting point. To locate more than one minimum or to locate the global potential energy minimum usually requires a means of generating different starting points, each of which is then minimized. Some specialized minimization methods can make uphill moves to seek minima lower in energy that the nearest one, but no algorithm has yet proved capable of locating the global energy minimum from an arbitrary starting position.
The input to a minimization program consists of a set of initial coordinates for the systems. The initial coordinates may come from a variety of sources. They may be obtained from an experimental technique, such as X-ray crystallography or NMR. In other cases a theoretical method is employed, such as a conformational search algorithm. A combination of experimental and theoretical approaches may also be used. For example, to study the behaviour of a protein in water one may take an X-ray structure of the protein and immerse it in a solvent bath, where the coordinates of the solvent molecules have been previously obtained from a Monte Carlo or molecular dynamics simulation [Leach, 2001].

Minimization algorithms
Three possible energy minimization methods are:

- those that require only function evaluations. Examples are the simplex method and its variants. A step is made on the basis of the results of previous evaluations. If derivative information is available, such methods are inferior to those that use this information;
- those that use derivative information. Since the partial derivatives of the potential energy with respect to all coordinates are known in MD programs (these are equal to minus forces) this class of methods is very suitable as modification of MD programs.
- those that use second derivative information as well. These methods are superior in their convergence properties near the minimum: a quadratic potential function is

minimized in one step. The problem is that for *N* particles a $3N$ x $3N$ matrix must be completed, stored and inverted. Apart from the extra programming to obtain second derivatives, for most systems of interest this is beyond the available capacity. There are intermediate methods building up the Hessian matrix on the fly, but they also suffer from excessive storage requirements.

The steepest descent method is of the second class. It simply takes a step in the direction of the negative gradient (hence in the direction of the force), without any consideration of the history built up in previous steps. The step size is adjusted such that the search is fast but the motion is always downhill. This is a simple and sturdy method: its convergence can be quite slow, especially in the vicinity of the local minimum. Although steepest descent is certainly not the most efficient algorithm for searching, it is robust and easy to implement. The vector r is defined as the vector of all *3N* coordinates. Initially a maximum displacement h_0 (*e.g.* 0.01nm) must be given. First the force *F* and potential energy are calculated. New positions are calculated by:

$$r_{n+1} = r_n + \frac{F_n}{\max(|F_n|)} h_n \tag{1.11}$$

where h_n is the maximum displacement and F_n is the force, or the negative gradient of the potential *V*. The notation $\max(|F_n|)$ means the largest of the absolute values of the force components. The forces and energy are again computed for the new positions. If $(V_{n+1} < V_n)$, the new positions are accepted and $h_{n+1} = 1.2h_n$. If $(V_{n+1} \geq V_n)$, the new position are rejected and $h_n = 0.2h_n$. The algorithm stops when either a user-specified number of force evaluations has been performed (*e.g.* 100), or when the maximum of the absolute values of the force (gradient) components is smaller than a specified value ϵ. Since force truncation produces some noise in the energy evaluation, the stopping criterion should not be made too tight to avoid endless iterations. A reasonable value for ϵ can be estimated from the root mean square force *f* a harmonic oscillator would exhibit at a temperature *T*. This value is:

$$f = 2\pi\upsilon\sqrt{2mkT} \tag{1.12}$$

where υ is the oscillator frequency, *m* the (reduced) mass, and *k* Boltzmann's constant. For a weak oscillator with a wave number of 100 cm^{-1} and a mass of 10 atomic units, at a temperature of 1 K, $f = 7.7$ $kJmol^{1}nm^{1}$. A value for ϵ between 1 and 10 is acceptable.

Conjugate gradient is slower than steepest descent in the early stages of the minimization, but becomes more efficient closer to the energy minimum, The parameters and stop criterion are the same as for steepest descent. The conjugate gradient method [Zimmerman, 1991] uses gradient information from previous steps. In general, steepest descents will bring close to the nearest local minimum, but performs worse far away from the minimum [Lindahl et al., 2001].

Classical molecular dynamics simulations

While minimization computes the forces on the atoms and changes their positions to minimize the interaction energies, dynamics computes forces and moves atoms in response to the forces. MD simulations solve Newton's equations of motion for a system of *N* interacting atoms:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i, i = 1 \ldots N \tag{1.13}$$

The forces are the negative derivatives of a potential function $V(r_1, r_2, \ldots, r_n)$:

$$F_i = \frac{\partial V}{\partial r_i} \tag{1.14}$$

The equations are solved simultaneously in small timesteps. The systems is followed for some time, taking care that the temperature and pressure remain at the required values, and the coordinates are written to an output file at regular intervals. The coordinates as a function of time represent a trajectory of the systems. After initial changes, the system will usually reach an equilibrium state. By averaging over an equilibrium trajectory many macroscopic properties can be extracted from the output file.

Dynamics simulations are useful in studies of the time evolution of a variety of systems at non-zero temperatures, for example, biological molecules, polymers, or catalytic materials, in a variety of states, for examples, crystals, aqueous solutions, or in the gas phase.

The major applications of molecular dynamics are:

- performing conformational searches. During dynamics, a system undergoes conformational and momentum changes so that different parts of the phase space accessible to the model can be explored. The conformational search capability of dynamics is one of its most important uses;
- generating statistical ensembles. By providing several mechanisms for controlling the temperature and pressure of simulated systems, molecular dynamics allows the generation of statistical ensembles from which various energetic, thermodynamic, structural and dynamics properties can be calculated. For such studies, it is important that the calculation visits various conformational states with the correct statistical frequency;
- studying the motions of molecules. Although modern crystallography has provided a window into the static structure of both small and large molecules, the thought of intermolecular collisions and conformational variation is always present. After all, building of substrates by proteins, folding of proteins and peptides into unique shapes, the dynamic behavior of polymers, and chemical reactions themselves would be inconceivable without the concept of molecular motion. Studies of model motions can be used to derive properties such as diffusion coefficients.

Other approaches to simulating molecular motion and generating conformational searches exist. For example, a dynamics trajectory can be constructed from a set of normal modes to represent the vibrations of a model. While this is a fast method, it is restricted to harmonic motion about a single energy minimum. An approach to doing conformational searches is the Monte Carlo method. While this method can sample conformational space so as to produce meaningful statistical ensembles, it does not provide dynamic information about the model, since particles of the model system are simply moved randomly according to some statistical rules [MSI, 1998].

It is useful at this point to consider the limitations of MD simulations:

- the simulations are classical:
  using Newton's equation of motion automatically implies the use of classical mechanics to describe the motion of atoms. This is all right for most atoms at normal temperatures, but there are exceptions. Hydrogen atoms are quite light and the motion of protons is sometimes of essential quantum mechanical character. For example, a proton may tunnel through a potential barrier in the course of a transfer over a hydrogen bond. Such processes cannot be properly treated by classical dynamics. Liquid helium at low temperature is another example in which classical mechanics breaks down. The statistical mechanics of a classical harmonic oscillator differs appreciably from that of a real quantum oscillator when the resonance frequency ν approximates or exceeds $k_b T/h$. Now at room temperature the wavenumber

$\sigma = 1/\lambda = \nu/c$ at which $h\nu = k_b T$ is approximately 200 cm^{-1}. Thus all frequencies higher than 100 cm^{-1} may misbehave in classical simulations. This means that practically all bond and bond-angle vibrations are suspect, and even hydrogen-bonded motions as translation or librational H-bond vibrations are beyond the classical limit. To circumvent these limitations, apart from real quantum-dynamical simulations, one can do one of two things:

(a) If performing MD simulations using harmonic oscillators for bonds, one should make corrections to the total internal energy $U = E_{kin} + E_{pot}$ and specific heat C_V (and to entropy S and free energy A or G if those are calculated). The corrections to the energy and specific heat of a one-dimensional oscillator with frequency ν are [McQuarrie, 1976]:

$$U^{QM} = U^{cl} + kT\left(\frac{1}{2}x - 1 + \frac{x}{e^x - 1}\right) \tag{1.15}$$

$$C_V^{QM} = C_V^{cl} + k\left(\frac{x^2 e^2}{(e^x - 1)^2} - 1\right) \tag{1.16}$$

where $x = h\nu/kT$. The classical oscillator absorbs too much energy (kT), while the high frequency quantum oscillator is in its ground state at the zero-point energy level of $\frac{1}{2}h\nu$.

(b) One can treat the bonds (and bond angles) as constraints in the equation of motion. The rationale behind this is that a quantum oscillator in its ground state resembles a constrained bond more closely than a classical oscillator. A good practical reason for this choice is that the algorithm can use larger timesteps when the highest frequencies are removed. In practice the timestep can be made four times as large when bonds are constrained than when they are oscillators [Van Gunsteren & Berendsen, 1977]. The flexibility of the latter is rather essential to allow for the realistic motion and coverage of configurational space [Van Gunsteren & Berendsen, 1977].

- electrons are in the ground state.
  In MD its common to use a conservative forcefield that is a function of the positions of atoms only. This means that the electronic motions are not considered: the electrons are supposed to adjust their dynamics instantly when the atomic positions change (the Born-Oppenheimer approximation), and remain in their ground state. This is almost always correct but, of course, electron transfer process and electronically excited states can not be treated. Neither can chemical reactions be treated properly.
- forcefield are approximate.
- Forcefield provide the forces. They are not really a part of the simulation method and their parameters can be user-modified as the need arises or knowledge improves. But the form of the forces that can be used in a particular program is subject to limitations. The majority of the forcefields is pair-additive (apart from long-range Coulomb forces), cannot incorporate polarizabilities, and does not contain fine-tuning of bonded interactions - but is quite useful and fairly reliable for bio-macromolecules in aqueous solution.
- the forcefield is pair-additive.
  This means that all non-bonded force result from the sum of non-bondend pair interactions. Non pair-additive interactions, the most important example of which is interaction through atomic polarizability, are represented by effective pair potentials.

Only average non pair-additive contributions are incorporated. This also means that the pair interactions are not pure, *i.e.,* they are not valid for isolated pairs or for situations that differ appreciably from the test systems on which the models were parameterized. In fact, the effective pair potentials are generally adequate. But the omission of polarizability also means that electrons in atoms do not provide a dielectric constant as they should. For example, real liquid alkanes have a dielectric constant of slightly more than 2, which reduces the long-range electrostatic interaction between (partial) charges. Thus the simulations will exaggerate the long-range Coulomb terms.

- long-range interactions are cutoff.
  MD programs always use a cutoff radius for the Lennard-Jones interactions and sometimes for the Coulomb interactions as well. Due to the minimum-image convention (only one image of each particle in the periodic boundary conditions is considered for a pair interaction), the cutoff range cannot exceed half the box size. That it still sufficient for large systems, and trouble is only expected for systems containing charged particles. But there can be accumulation of charges at the cutoff boundary or very wrong energies. For such systems one should consider using one of the implemented long-range electrostatic algorithms, such as particle-mesh Ewald [Essmann et al., 1995; Darden et al., 1993].
- boundary conditions are unnatural.
  Since system size is small (even 10 000 particles is small), a cluster of particles will have a lot of unwanted boundary with its environment (vacuum). When simulating bulk systems, periodic boundary conditions are used to avoid real phase boundaries. But liquids are not crystals, so something unnatural remains. For large systems the errors are small, but for small systems with a lot of internal correlation, the periodic boundaries may enhance internal correlation. This is especially important when using lattice sums for long-range electrostatics, since these are known to sometimes introduce extra ordering [Van der Spoel et al., 2006].

Molecular dynamics (MD), Monte Carlo (MC) simulations, or hybrids thereof are general methods to generate a physically meaningful ensemble of conformations. These methods rely on a suitably well-parameterized force field that describes the forces to which the atoms in the system are subject. A trajectory is generated, typically at constant temperature, and the resulting conformations can be subjected to clustering and further analysis. These methods are general and can work with large, complex, multicomponent systems. However, one serious drawback is that they explore the conformation space very slowly, since most of the simulation time is spent around the equilibrium energy state(s) and low frequency conformation transitions are observed relatively rarely. Special techniques can been used, *e.g.,* for polypeptide chains, to greatly improve the sampling properties of Monte Carlo; however, the long simulation times make MD and MC simulations more suitable for the thermodynamic analysis of molecular systems rather than for conformational analysis.

Low Mode molecular dynamics

The problem of protein folding is a long-time studied and fascinating topic. The idea that the three-dimensional protein structure is determined by its primary structure, *i.e.* by the physicochemical properties of the amino acids composing the biopolymer, was already hypothesized in the 30's. Anfinsen, with his experiment, demonstrated that a denatured enzyme is able to refold and to completely recover its biological activity [Anfinsen 1973]. Since the three-dimensional structure of a protein is closely related to its sequence, it seems straightforward to compute its structure by a purely biophysical-based 'brute-force' approach.

However, protein folding does not seem to take place in this exhaustively explorative fashion since, according to Levinthal's evaluations, a very small protein made of 100 residues should spend approximately 1027 years to sample all its conformational space [Levinthal 1969]. The fact that protein folding time is in the order of magnitude of μs to ms suggests that other factors can drive this process. Different hypotheses were put forward about the molecular mechanisms of protein folding. The most common ones make reference to kinetics or to thermodynamics, such as classical molecular dynamics. However, the exhaustive search of the conformational space is a computationally demanding procedure, especially for big proteins, and cannot actually be afforded through either classical MD or random walk approaches. For this reason, scientists developed alternative methods aimed at empowering the exploration of the protein conformational space, including approaches based on the guided search of the conformational space. In 1996, Kolossvàry and Guida reported on Low Mode Search (LowMode), which is a conformational search method based on low-frequency modes [Kolossvary 1996].
In this method, for each iteration the molecular system is perturbed along low-frequency vibrational mode directions in a systematic way. The perturbed conformation is then subjected to coordinate optimization. The fundamental principle of low mode search is very appealing: use low-frequency mode transitions to "hop" from one low-energy state to another. An assumption is made that such transitions will lead to a full exploration of the low energy conformation ensemble; however, this assumption is physically motivated and appears reasonable. LowModeMD is not intended to be a high-throughput method, and emphasis is placed on the efficient production of low-strain-energy local minima of a potential energy function. The method is stochastic, in that there is no attempt to systematically follow vibrational mode directions.
In comparison with classical MD, it features a very efficient way for searching for minima troughs on the potential energy surface. LowModeMD can be used for studying the flexibility of structural protein regions, such as loops, or for the modeling of ligands in the binding pocket.

Protein structure modeling

Protein structures are primarily obtained using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, but these methods are time consuming, expensive, and not feasible for all proteins.
All the data about the already known proteins are organized in public databases classified in sequence and structure databases. Among these, we just recall the most relevant ones:

Sequence:	UniProt [http://www.uniprot.org/]
	NCBI [http://www.ncbi.nlm.nih.gov/guide/proteins/]
Structure:	RCSB PDB [http://www.rcsb.org/pdb/home/home.do]
	CATH [http://www.cathdb.info/]
	SCOP [http://scop.mrc-lmb.cam.ac.uk/scop/]

The Universal Protein Resource (UniProt) provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequences and functional information. The National Center for Biotechnology Information (NCBI) also maintains a non-redundant database of protein primary structures from several species. RCSB (Research Collaboratory for Structural Bioinformatics) PDB (Protein Data Bank) contains three-dimensional structures of biopolymers solved mainly through X-ray crystallography and NMR. At the start of 2012, PDB contains more than 80 000 structures. The CATH database is a manually-managed classification (supported by automated techniques) of protein domain structures. The SCOP database provides a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As reported in its web site, it provides a broad survey of all known protein folds, detailed

information about the close relatives of any particular protein, and a framework for future research and classification. As seen from the above, only a small number of protein structures have been solved, but the knowledge of protein structure is a critical step toward understanding their functions.
The structural properties of a protein are univocally determined by its primary structure, *i.e*. its amino acid sequence. For several years, scientists have been studying the possibility to predict the three-dimensional structure of a protein just starting from its amino acid sequence. This is not a plain issue but, with time, useful methods for predicting the structural properties, local or global, of a protein from its sequence have been significantly improved.

The protein structure modeling moves from the idea that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins and then that some tertiary structures are shared between different proteins.
The comparative protein modelling indeed provides the use of existing structures solved as a starting point for the construction of new structural models. The approach to be used for building a three-dimensional protein model depends upon the data that can be retrieved from literature. The most used and most accurate procedure is homology or comparative modeling; it is based on the idea that proteins with similar sequence share the same folding. The necessary conditions for producing a good three-dimensional model through the homology modeling approach is that the similarity between the target sequence and the template structure(s) is detectable and that a reliable alignment between them can be obtained.
If you are unable to identify evolutionarily related templates to the protein of interest by sequence alignment, a good alternative strategy can be the protein threading approach. This method scans the amino acid sequence of an unknown structure against a database of solved structures and tries to adapt the primary structure on all the tertiary structure available, assigning a folding score according to a scoring function, which evaluates the compatibility of the sequence to the structures.
Nevertheless, in some cases, both comparative modeling strategies can fail, principally when the evolutionary distance between query and template is too high. In these cases, possible solutions are *ab initio* or *de novo* protein modelling methods. These methods build three-dimensional protein models "from scratch", *i.e.,* based on physical principles rather than on previously solved structures. However, these procedures appear to have a high computational cost and to be time-consuming. To predict protein structure *de novo* for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers or distributed computing.
Once modelled the protein backbone, it is necessary an accurate packing of the amino acid side chains. Different methods exist to predict the correct side-chain geometry, building a set of different side chain conformations known as "rotamers", and then energetically evaluate them.
A very interesting method for protein model construction is Robetta [Kim 2004]. The Robetta server is an automated protein structure prediction service offered by the Baker laboratory for non-commercial *ab initio* and comparative modelling [Raman 2009].
First, Robetta searches for structural homologs using different algorithms, such as BLAST or PSI-BLAST, then parses the target sequence into individual domains, or independently folding units of proteins, by matching the sequence to structural families in the Pfam database; if domains present structural homology with known proteins, Robetta follows a "template-based model" protocol. Domains without structural homologs are modelled by Rosetta *de novo* method. From all these fragments, Robetta produces a decoy (possible structure) to be then energetically evaluated. The final structure prediction is selected by taking the lowest energy model as determined by a low-resolution Rosetta energy function. Finally, side-chain contributions are modelled using a protocol for Monte Carlo conformational search.

Mutagenesis

Mutagenesis studies on proteins are carried out with different aims. They are useful in deciphering the structure/function relationships, in predicting the effect of a mutation on the development of a specific pathology, in the evaluation of the physicochemical effects of specific mutations on protein folding and stability, etc. The effect of single point mutations can be estimated keeping into account the distance from active residues, the change in the local hydrophobic/hydrophilic balance, the change in steric hindrance, the disruption of salt bridges, etc.

Several approaches for studying the effects of mutations have been set up, tested, and described in literature. Some of them are based on the knowledge of the protein structure; others are just based on comparative considerations. A widely used method is PolyPhen-2 [Adzhubei 2010]. PolyPhen-2 is an interesting tool that can predict the impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. It is based on a pipeline combining multiple sequence alignment with a probabilistic classifier based on machine-learning methods.

When the three-dimensional structures of the target proteins are available, more accurate evaluations can be performed. For single point mutations, local structure and energetic evaluations can be carried out after side chain repacking. This procedure is more complex if the insertion or deletion is larger. Some very efficient sampling methods can be useful in order to evaluate the possible conformations allowed after the primary structure modification, for instance, the LowMode MD simulations method, applied on the mutated residue and on its environment.

Molecular docking

In computational biology, molecular docking is a method to predict the structure (or structures) of the intermolecular complex formed between two or more molecules, or rather, the reciprocal preferred orientation of two (or more) interacting molecules. Docking is widely used to predict the strength of association (binding affinity) between two molecules.

Docking is used to investigate on the associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids, that play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (*e.g.,* agonism *vs* antagonism). For instance, this method is frequently used to predict the binding orientation of small putative drug compounds to their protein targets in order to evaluate the affinity (and the activity, even if not directly) of these small molecules, and hence docking plays a relevant role in the rational design of drugs.

Considerable efforts have been directed towards improving the methods used to predict docking, because of the high biological and pharmaceutical significance of molecular docking

The “docking problem” thus deals with the generation and evaluation of plausible structures of intermolecular complexes. The docking problem involves many degrees of freedom. There are six degrees of translational and rotational freedom per molecule with respect to another one, as well as the conformational degrees of freedom of each molecule.

This problem can be manually managed, but this “hands-on” approach can be very complex and time consuming. Automatic docking algorithms can be less biased than human operators and usually consider many more complex possibilities. Various algorithms have been developed to tackle the docking problem. These can be characterised according to the number of degrees of freedom they ignore. Thus, the simplest algorithms treat the two molecules as rigid bodies and explore only the six degrees of translational and rotational freedom. This method is named rigid molecular docking, and molecules are treated as rigid objects that

cannot change their spatial shape during the docking process. The earliest algorithms for docking small molecule ligands into the binding sites of proteins and DNA used this approximation; they were developed by Kuntz et al. in 1982 [Kuntz 1982].

References

Adzhubei I.A. et al., 2010 A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-9.

Amar M.J. et al., 2009 Adenoviral expression of human lecithin-cholesterol acyltransferase in nonhuman primates leads to an antiatherogenic lipoprotein phenotype by increasing high-density lipoprotein and lowering low-density lipoprotein.Metabolism. 58(4):568-75.

Anfinsen C.B. 1973 Principles that govern the folding of protein chains. Science 181 (4096): 223–230.

Baldassarre D. et al., 2010 Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study. Eur Heart J. 31(5):614-22.

Barter P.J. et al., 2007 HDL cholesterol, very low levels of LDL cholesterol, and cardiovascular events. Treating to New Targets Investigators. N Engl J Med. 357(13):1301-10.

Bolin D.J., Jonas A. 1996 Sphingomyelin inhibits the lecithin-cholesterol acyltransferase reaction with reconstituted high density lipoproteins by decreasing enzyme binding. J Biol Chem. 271(32):19152-8.

Born M., Oppenheimer J.R. 1927 On the quantum theory of molecules. Ann. Phys. 84:457-84.

Brady L. et al., 1990 A serine protease triad forms the catalytic center of a triacylglycerol lipase Nature 343, 767 – 770.

Brown B.G. et al., 1993 Lipid lowering and plaque regression. New insights into prevention of plaque disruption and clinical events in coronary disease. Circulation 87(6):1781-91.

Calabresi L. et al., 2009a Functional lecithin: cholesterol acyltransferase is not required for efficient atheroprotection in humans.Circulation. 120(7):628-35.

Calabresi L. et al., 2009b A novel homozygous mutation in CETP gene as a cause of CETP deficiency in a Caucasian kindred. Atherosclerosis. 205(2):506-11.

Calabresi L., Franceschini G. 2010 Lecithin:cholesterol acyltransferase, high-density lipoproteins, and atheroprotection in humans.Trends Cardiovasc Med. 20(2):50-3.

Calabresi L. et al., 2011 Plasma lecithin:cholesterol acyltransferase and carotid intima-media thickness in European individuals at high cardiovascular risk. J Lipid Res. 52(8):1569-74.

Carlson L.A., Philipson B. 1979 Fish-eye disease. A new familial condition with massive corneal opacities and dyslipoproteinaemia. Lancet. 2(8149):922-4.

Carlson L.A., Holmquist L. 1985 Paradoxical esterification of plasma cholesterol in fish eye disease. Acta Med Scand. 217(5):491-9.

Chen Z. et al., 2012. Small molecule activation of lecithin cholesterol acyltransferase lipoprotein metabolism in mice and hamsters. Metabolism. 61: 470-481.

Collet X., Fielding C.J. 1991. Effects of inhibitors of N-linked oligosaccharide processing on the secretion, stability, and activity of lecithin:cholesterol acyltransferase. Biochemistry. 30: 3228-3234.

Darden T. et al., 1993 An N·log(N) method for Ewald sums in large systems J. Chem. Phys. 98:10089-92.

De Palma R.G., Clowes A.W. 1978 Aug Interventions in atherosclerosis: a review for surgeons. Surgery. 84(2):175-89.

Doi Y., Nishida T. 1983 Microheterogeneity and physical properties of human lecithin-cholesterol acyltransferase. J. Biol. Chem. 258: 5840-5846.

Dullaart R.P. et al., 2008 A Plasma lecithin: cholesterol acyltransferase activity is elevated in metabolic syndrome and is an independent marker of increased carotid artery intima media thickness. J Clin Endocrinol Metab. 93(12):4860-6.

Ermer O. 1976 Calculation of molecular properties using force fields. Applications in organic chemistry, Structure and Bonding, 27:161-211.

Essman U. et al., 1995 A smooth particle mesh Ewald potential J. Chem. Phys. 103:8577-93.

Franceschini G. et al., 1991 Reverse cholesterol transport: physiology and pharmacology. Atherosclerosis. 88(2-3):99-107.

Frohlich J., Dobiásová M. 2003 Fractional esterification rate of cholesterol and ratio of triglycerides to HDL-cholesterol are powerful predictors of positive findings on coronary angiography. Clin Chem. 49(11):1873-80.

Gjone E., Norum K.R. 1968 Familial serum cholesterol ester deficiency. Clinical study of a patient with a new syndrome. Acta Med Scand. 183(1-2):107-12.

Glomset J.A. 1962. The mechanism of the plasma cholesterol esterification reaction: plasma fatty acid transferase. Bioclrim Biopfvys Acta 65:128-135.

Glomset J.A., et al., 1970 Plasma lipoproteins in familial lecithin:cholesterol acyltransferase deficiency: lipid composition and reactivity *in vitro* John A. J Clin Invest. 49(10): 1827–1837.

Grove D., Pownall H.J. 1991 Comparative specificity of plasma lecithin:cholesterol acyltransferase from ten animal species. Lipids. 26:416–20.

Haase, C.L., et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 466: 707-713.

Heisenberg W., Physikalische Prinzipien der Quantentheorie, Leipzig: Hirzel (1930).

Hoeg J.M. et al., 1996 Overexpression of lecithin:cholesterol acyltransferase in transgenic rabbits prevents diet-induced atherosclerosis. Proc Natl Acad Sci U S A. 93(21):11448-53.

Hovig T., Gjone E. 1973 Familial plasma lecithin: cholesterol acyltransferase (LCAT) deficiency. Ultrastructural aspects of a new syndrome with particular reference to lesions in the kidneys and the spleen.Acta Pathol Microbiol Scand A. 81(5):681-97.

Jonas A. 2000 Lecithin cholesterol acyltransferase. Biochim Biophys Acta. 1529(1-3):245-56.

Karmin O., et al., 1993. Lecithin:cholesterolacyltransferase: role of N-linked glycosylation in enzyme function. Biochem. J. 294: 879-884.

Kim D.E. et al., 2004 Protein structure prediction and analysis using the Robetta server. Nuclei c Acids Res. 32 Suppl 2:W526-31.

Kinnunen P.K., et al., 1982 Dynamics of C-apolipoproteins in the enzymatic interconversions of plasma lipoproteins. Ric Clin Lab. 12(1):41-9.

Klein H.G. et al., 1995 *In vitro* expression of structural defects in the lecithin-cholesterol acyltransferase gene. J Biol Chem. 270(16):9443-7.

Kolossvary I., Guida W.C., 1996 Low mode search— an efficient, automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides. J. Am.Chem. Soc. 118, 5011-5019.

Kosek A.B. et al., 1999 Binding affinity and reactivity of lecithin cholesterol acyltransferase with native lipoproteins. Biochem. Biophys. Res. Commun. 258:548-551.

Kuivenhoven J.A. et al., 1997 The molecular pathology of lecithin:cholesterol acyltransferase (LCAT) deficiency syndromes.J Lipid Res. 38(2):191-205.

Kuntz I.D. et al. 1982 A geometric approach to macromolecule-ligand interactions. Journal of Molecular Biology 161(2): 269–88.

Kuroda M. et al., 2012 Ceiling culture-derived proliferative adipocytes are a possible delivery vehicle for enzyme replacement therapy in lecithin:cholesterol acyltransferase deficiency. Gene. Ther. Mol. Biol., in press.

Lambert G. et al., 2001 Analysis of glomerulosclerosis and atherosclerosis in lecithin cholesterol acyltransferase-deficient mice. J Biol Chem. 276(18):15090-8.

Leach A.R., Molecular Modeling: Principles and Applications, Prentice Hall (2001).

Levinthal, C. 1969 How to Fold Graciously Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois: 22–24.

Li L. et al., 2011 Lecithin cholesterol acyltransferase null mice are protected from diet-induced obesity and insulin resistance in a gender-specific manner through multiple pathways. J Biol Chem. 286(20):17809-20.

Lindahl E. et al., 2001 GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. 7:306-17.

Mackness B., Durrington P.N. Lipoprotein separation and analysis for clinical studies. Oxford University Press, Oxford, UK 1992.

MacLean J. et al. 1986 Cloning and expression of human lecithin-cholesterol acyltransferase cDNA. Proc. Natl. Acad. Sci. U.S.A. 83 (8): 2335–9.

Nicholls S.J. et al., 2011 Effects of the CETP inhibitor evacetrapib administered as monotherapy or in combination with statins on HDL and LDL cholesterol: a randomized controlled trial. JAMA. 306(19):2099-109.

Niesor E.J. et al. 2010 Modulating cholesteryl ester transfer protein activity maintains efficient pre-β-HDL formation and increases reverse cholesterol transport. J Lipid Res. 51(12):3443-54.

MacPherson P.A.C. et al., 2007 A dual role for lecithin:cholesterol acyltransferase (EC 2.3.1.43) in lipoprotein oxidation Free Radical Biology and Medicine Volume 43, Issue 11, Pages 1484–1493.

MacQuarrie D.A., Statistical Mechanics, Harper and Row, New York (1976).

Miida T. et al., 1996 Pre beta 1-high-density lipoprotein increases in coronary artery disease. Clin Chem. 42(12):1992-5

MSI - Forcefield-Based Simulations. Sacraton road, San Diego, CA (1998).

Montgomery R. et al., Lipoproteins. In: Biochemistry. A caseoriented approach, the C.V. Mosby Co., St. Louis (1990).

Nakamura Y. et al., 2004 Molecular mechanism of reverse cholesterol transport: reaction of pre-beta-migrating high-density lipoprotein with plasma lecithin/cholesterol acyltransferase. Biochemistry. 43(46):14811-20.

Nemukhin A.V. et al., 2006 Serine hydrolase catalytic sites: Geometry invariants and modeling catalytic activity Mendeleev Communications. 16, 290.

Norum, K.R., Gjone E. 1967. Familial plasma lecithin:cholesterol acyltransferase deficiency. A new inborn error of metabolism. Biochim. Biophys. Acta. 20: 231-243.

Peelman F. et al. 1998 A proposed architecture for lecithin cholesterol acyl transferase (LCAT): identification of the catalytic triad and molecular modeling. Protein Sci. 7(3):587-99.

Portman O.W., Sugano M. 1964 Factors Influencing the Level and Fatty Acid Specificity of the Cholesterol Esterification Activity in Human Plasma. Arch Biochem Biophys. 105:532–40.

Qu, S.J. et al., 1993. Effects of site-directed mutagenesis on the N-glycosylation sites of human lecithin:cholesterol acyltransferase. Biochemistry. 32: 8732-8736.

Rader D.J., Tall A.R. 2012 The not-so-simple HDL story: Is it time to revise the HDL cholesterol hypothesis? Nature Medicine 18, 1344–1346.

Raman S. et al.2009 Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 77 Suppl 9:89-99

Rawlings N.D., Barrett A.J. 1994 Families of serine peptidases. Methods Enzymol. 244:19-61.

Ross R. 1986 The pathogenesis of atherosclerosis--an update. N Engl J Med. 314(8):488-500.

Ross R. 1993 The pathogenesis of atherosclerosis: a perspective for the 1990s. Nature. 362(6423):801-9.

Rousset X. et al., 2009 Lecithin: cholesterol acyltransferase--from biochemistry to role in cardiovascular disease. Curr Opin Endocrinol Diabetes Obes. 16(2):163-71.

Rye K.A., Barter P.J. 2004 Formation and metabolism of prebeta-migrating, lipid-poor apolipoprotein A-I.Arterioscler Thromb Vasc Biol. 24(3):421-8.

Savel J. et al. 2012 Very low levels of HDL cholesterol and atherosclerosis, a variable relationship--a review of LCAT deficiency.Vasc Health Risk Manag. 8:357-61.

Scandinavian Simvastatin Survival Study 1994 Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S) Lancet. 344(8934):1383-9.

Sparks D.L., Pritchard P.H. 1989 The neutral lipid composition and size of recombinant high density lipoproteins regulates lecithin:cholesterol acyltransferase activity. Biochem Cell Biol. 67(7):358-64.

Sperry W.M. 1935. Cholesterol esterase in blood. J Biol Chem 111: 467.

Stony Brook University (2012) Stony Brook Researchers Look Back on Scientific Advances Made as a Result of a 50-Year Old Puzzle [http://commcgi.cc.stonybrook.edu/am2/publish/General_University_News_2/Stony_Brook_Researchers_Look_Back_on_Scientific_Advances_Made_as_a_Result_of_a_50-Year_Old_Puzzle.shtml].

Superko H.R., Krauss R.M. 1994 Coronary artery disease regression. Convincing evidence for the benefit of aggressive lipoprotein management. Circulation. 90(2):1056-69.

Tanigawa H. et al., 2009 Lecithin: cholesterol acyltransferase expression has minimal effects on macrophage reverse cholesterol transport *in vivo*. Circulation. 120(2):160-9.

Tyroler H.A. 1987 Review of lipid-lowering clinical trials in relation to observational epidemiologic studies. Circulation. 76(3):515–522.

Van der Spoel D. et al. Gromacs User Manual Version 3.3 Nijenborgh 4, 9747 AG Groningen, The Netherlands (2006).

Van Gunsteren W.F., Berendsen H.J.C. 1977 Algorithms for macromolecular dynamics and constraint dynamics, Mol. Phys. 34:1311-27.

Wells I.C. et al., 1986 Lecithin: cholesterol acyltransferase and lysolecithin in coronary atherosclerosis. Exp Mol Pathol. 45(3):303-10.

Zanotti I. et al., 2011 The thienotriazolodiazepine Ro 11-1464 increases plasma apoA-I and promotes reverse cholesterol transport in human apoA-I transgenic mice. Br J Pharmacol. 164(6):1642-51.

Zhou M. et al. 2009 Lecithin cholesterol acyltransferase promotes reverse cholesterol transport and attenuates atherosclerosis progression in New Zealand White Rabbits Circulation. 120: S1175b.

Zimmerman K. 1991 All purpose molecular mechanics simulator and energy minimize, J. Comp. Chem. 12:310–19

Chapter 2
Aims Of The Study

A series of large population studies have revealed the existence of a strong inverse correlation between plasma levels of high-density lipoprotein cholesterol (HDL-C) and cardiovascular risk. Accordingly, HDL has become a major target for the development of novel therapies for the treatment of atherosclerotic cardiovascular disease.
LCAT is the enzyme responsible for the catalysis of the transacylation of the sn-2 fatty acid of lecithin to the free 3-0H group of cholesterol, generating cholesteryl ester (CE) and lysolecithin. LCAT plays a central role in HDL structure and metabolism and promotes reverse cholesterol transport.[Jonas 2000]
The effect of LCAT on atherogenesis has been controversial, but recent studies [Calabresi 2009; Holleboom 2010] suggest that the presence of defective LCAT may not preclude cholesterol removal from the arterial wall and efficient reverse cholesterol transport. These recent data, together with findings in genetic LCAT deficiency, challenge the notion that LCAT is atheroprotective and suggest that, despite positive effects on plasma HDL-C concentration, elevating LCAT expression and/or activity is not a promising therapeutic strategy to reduce cardiovascular risk. On the contrary, decreasing LCAT activity has certainly negative effects on plasma HDL-C concentration but possibly positive effects on HDL structural function, and may represent a new therapeutic strategy to reduce cardiovascular risk. In this perspective, the availability of new potent and selective LCAT inhibitors is mandatory.
The objective of my studies is the identification of molecules able to modulate the LCAT enzyme, with the aim of identifying new drugs for the prevention and treatment of atherosclerotic cardiovascular diseases.
This aim was been realized using a combined approach of computational biology and biochemistry tools, based on the followed steps:

- *In silico* building of 3D LCAT structure by a combination of comparative and *ab initio* modeling and *in silico* evaluation of the known mutations impact on the LCAT structure/functions.
- *In silico* high-throughput screening of putative LCAT modulators and evaluation of the effects of LCAT modulators on cholesterol esterification *in vitro* and *in vivo* on atherosclerosis models.

Globally, this project will produce innovative pharmacological entities to be further developed for a completely new therapeutic approach to the treatment of atherosclerotic cardiovascular diseases.

References

Calabresi L. et al., 2009 Functional lecithin: cholesterol acyltransferase is not required for efficient atheroprotection in humans. Circulation120(7):628-35

Holleboom A.G. et al., 2010 Plasma levels of lecithin:cholesterol acyltransferase and risk of future coronary artery disease in apparently healthy men and women: a prospective case-control analysis nested in the EPIC-Norfolk population study. J Lipid Res. 2010 Feb;51(2):416-21.

Jonas A. 2000 Lecithin cholesterol acyltransferase. Biochim Biophys Acta. 1529(1-3):245-56

McLean J. et al. 1986 Cloning and expression of human lecithin-cholesterol acyltransferase cDNA. Proc. Natl. Acad. Sci. U.S.A. 83 (8): 2335–9.

Chapter 3
Results And Discussion

3.1 *LCAT STRUCTURE*

The identification of amino acids relevant in controlling LCAT structure and function is mandatory to design a therapeutic strategy to reduce cardiovascular risk that aims at modulating the activity of this enzyme. Knowledge of the topology of LCAT or, better, of its atomic structure is expected to be very helpful to understand its catalytic mechanism and its role in the metabolism of HDL. A molecular tool to perfect the study of the structure/function relationship of an enzyme is then the evaluation of the impact of known mutations its the activity. Several mutations in the *LCAT* gene, which result in familial LCAT deficiency (FLD) or fish-eye disease (FED), have been described.

To date, the structure of LCAT has not been resolved experimentally [Jauhiainen 1986; Yang 1987; Peelman 1999]. The major problems are connected with the enzyme purification from human plasma, where LCAT is present in extremely low concentrations and strongly bound to lipoproteins. An alternative approach may be protein modeling through *in silico* procedures.

LCAT belongs to the α/β hydrolase fold family and shares the Ser/Asp−Glu/His triad with lipases, esterases and proteases, but, as already discussed by Peelman et al. in 1998 [Peelman 1998], modeling LCAT structure faces a number of problems. LCAT lacks appropriate templates for a straightforward homology modeling: the protein has a very low sequence identity, if a high secondary structure similarity, with respect to all available templates. The LCAT N−terminus (residues 73−210) was already modeled on human pancreatic lipase by Peelman et al. [Peelman 1998]. The completion of the active site (aa 333−399), which extends beyond the modeled N-terminus, was carried out using the coordinates of the corresponding residues of *C. antarctica* lipase; the remaining parts of LCAT were not modeled. Peelman et al never made available to the scientific community the spatial coordinates of their LCAT model.

A new and complete model of LCAT will be appropriate to map and analyze the known mutations leading to FLD and FED and to envisage their structural implications. Interesting demonstrations on how molecular modelling can be successfully used to predict the functional importance of specific amino acids and their mutations are reported in [Campbell 2007; Peelman 2001; Razzaghi 2001; Kobayashi 2002].

In order to build a complete LCAT model, it is important to find better templates among the newly crystallized α /β hydrolases and to follow the most up-to-date strategies, such as a combined approach that includes secondary structure prediction, folding recognition, and 'chimeric' homology modeling [Weigelt 2010]. In parallel, tools, such as Polyphen and Panther [Sunyaev 2001; Thomas 2004; Brunham 2005], that rely on straightforward physical principles and perform comparative evaluations, may be used to predict the possible impact of various amino acid substitutions on the structure and function of LCAT.

3.2 MATERIALS AND METHODS (Theme 1)

Comparative modeling

The human LCAT sequence was downloaded from the UniProt-Protein Knowledgebase database [entry UniProt ID: P04180]. Starting from its sequence, a model was built based on multiple templates.

We identified two parts of the protein, a N-terminal 'domain', consisting of residues 1-210, and a C-terminal 'domain', composed of residues 200-416, and we separately submitted both parts to the Fold Recognition PSIPRED default procedure. We found 2VTV (PhaZ7depolymerase from *Paucimonas lemoignei*, UniProt ID: Q939Q9. Identity: 19.5%) as a suitable template for LCAT N-terminal 'domain' and 2VEO (Lipase A from *Candida antarctica*, UniProt ID: D4PHA8. Identity: 14.5%) as a template for LCAT C-terminal 'domain'. All the modeling procedures were carried out with modules of the suite Molecular Operating Environment 2008.10 (MOE).

The alignment of the sequences of target and template proteins was produced with the Align program of MOE using default parameters and was manually adjusted making reference to BLAST outputs. This alignment was set as reference for all the homology modeling procedures.

Comparative model building was carried out with the MOE Homology Model program. 2VTV was set as template for LCAT residues 1-210 and 2VEO for residues 200-416. Ten independent models were built and refined, and then the highest-scoring intermediate model was submitted to a further round of energy minimization (EM). Both for the intermediate and the final structures the refinement procedures consisted in EM runs based on the AMBER99 forcefield, with the reaction field solvation model.

The two disulfide bonds of the protein, between cysteines 50 and 74, and between cysteines 313 and 356, were created through the MOE Builder module.

The quality of the final model was carefully checked with the MOE Protein Geometry module to make sure that the stereochemical quality of the proposed structure was acceptable.

In parallel, we submitted the primary structure of LCAT to the Robetta Web Server, using default parameters. From the output, we selected the model with correct general topology, correct geometry of the catalytic triad, and most favorable setting to form cysteine disulfide bridges.

From the above, we built a chimeric model, setting the previous distant homology modeling and the Robetta *ab initio* model as primary template for residues 1-91; the option 'Use Selected Residues to Override Template(s)' was checked in order to override the primary template with the more appropriate ones only for the selected residues.

All models were minimized and geometrically and energetically evaluated as described above. Both disulfide bonds were set as described above.

Binding site analysis

The LCAT binding site was identified through the MOE Site Finder module, which uses a geometric approach to calculate possible binding sites in a receptor starting from its three-dimensional atomic coordinates. This method is based not on energy models but on alpha spheres, which are a generalization of convex hulls [Eldesbrunner 1995].

Mutations

We built LCAT mutant structures, introducing single mutations through the MOE Mutate module. For each mutation, we searched the best side chain orientation using MOE Rotamer Explorer module and we performed EM runs.

To evaluate the impact of mutations on LCAT, we submit the know mutations to two different prediction tools available online: Polyphen (polymorphism phenotyping; http://genetics.bwh.harvard.edu/pph2/) [Adzhubei 2010] and PANTHER (Protein ANalysis THrough Evolutionary Relationships; http://www.pantherdb.org/) [Thomas 2006], using default parameters.

Low-mode molecular dynamics

To evaluate the impact of all the collected mutations on the protein structure around the mutated residue, we applied the low-mode molecular dynamics (LowModeMD) approach, focusing a MD trajectory along the low-mode vibrations and searching with high efficiency for minima troughs on the potential energy surface. We run these computations with the Conformational Search program of MOE Conformations module, which uses an efficient implicit method for estimating the low-frequency modes, based on the attenuation of high-range velocities, as described in detail in [Labute 2010].

For both wild-type and mutated amino acid, we selected the nearest residues (closer than 4.5 Å); these were left free to move during the LowModeMD, whereas the residues more than 4.5 Å away were fixed (not free to move, but used for the potential calculations); all other residues were defined as inert (fixed and not used for calculations). The simulation was carried out with default parameters, except for strain cut-off, which was set at 100 kcal/mol. One hundred conformations were generated and ranked according to the value of the potential energy of the conformation.

For all the structure produced, potential energy and distance from the reference structure were calculated, building a RMSD matrix for each mutation.

3.3 RESULTS AND DISCUSSION (Theme 1)

LCAT homology modeling

For the modeling procedure we used the entire sequence of human LCAT; without its signal peptide, it consists of 416 amino acids [UniProt ID: P04180]. Submitting the entire sequence of LCAT to the PDB Search module of the MOE Suite, using default parameters, no suitable template was identified, as reported in Figure 3.1.

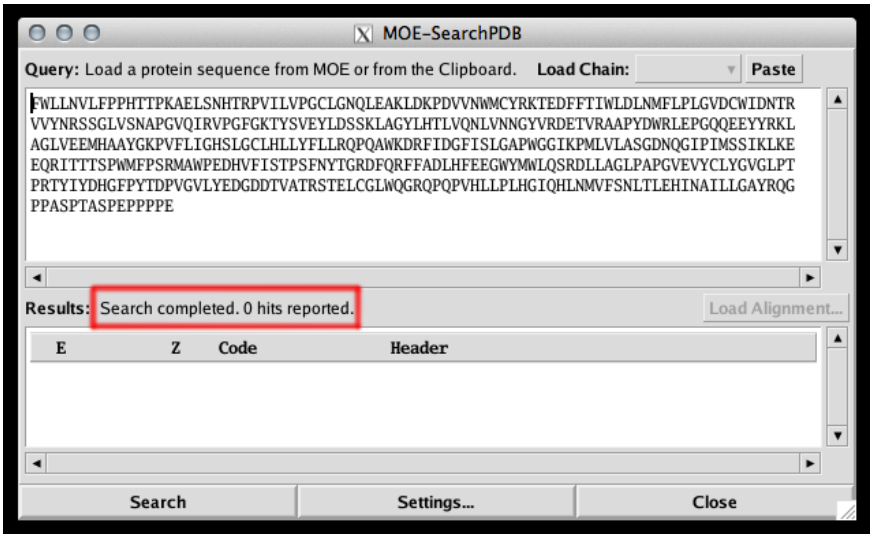


Figure 3.1. Screenshot of MOE-SearchPDB Output after submitting the LCAT entire sequence.

Submitting the same sequence to the Fold Recognition module (GenTHREADER) of the PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) did not identify any suitable template, and the best matching protein with a solved structure was a hydrolase from *Lactobacillus plantarum* [PDB ID 3LP5, UniProt ID: F9UMW5]. However, the latter shares with LCAT only 16% identity - a level of similarity insufficient to produce any acceptably accurate model. All the identified templates belong to the superfamily fold of α/β-hydrolases, confirming previous hypotheses that LCAT belongs to this superfamily.

The low sequence similarity inside the superfamily is due to the fact that the α/β-hydrolase fold is common to a number of hydrolytic enzymes of widely different phylogenetic origin and catalytic function. The enzymes are believed to have diverged from a common ancestor, preserving only the arrangement of the catalytic residues. All of them present a catalytic triad, which is the most conserved structural feature in the fold. This very high sequential and topological variability prevented us from recognizing an appropriate template for the homology modeling procedures. We had better success after splitting the LCAT primary structure approx. in two halves, which allowed the identification of two templates relevant to LCAT distant homology modeling.

One of the most innovative strategies to overcome the lack of an appropriate template is building homology models of distinct parts of a protein, and then merging them in a single model that can be defined 'chimeric'. This procedure has already been applied successfully by

our research group for modeling GPR17, a GPCR involved in ischemic and neurodegenerative diseases [Eberini 2011].
We thus split LCAT sequence into two parts of comparable length: a N-terminal 'domain', consisting of residues 1-210, and a C-terminal 'domain', composed of residues 200-416.
Submitting each part to the Fold Recognition PSIPRED procedure, we obtained putative templates with better identity than with *Lactobacillus plantarum* hydrolase, but still under 30%.
In Figure 3.2, we report the alignment between LCAT N-terminal 'domain' and 2VTV (PhaZ7depolymerase from *Paucimonas lemoignei*, UniProt ID: Q939Q9. Identity: 19.5%) and LCAT C-terminal 'domain' and 2VEO (lipase A from *Candida antarctica*, UniProt ID: D4PHA8. Identity: 14.5%).

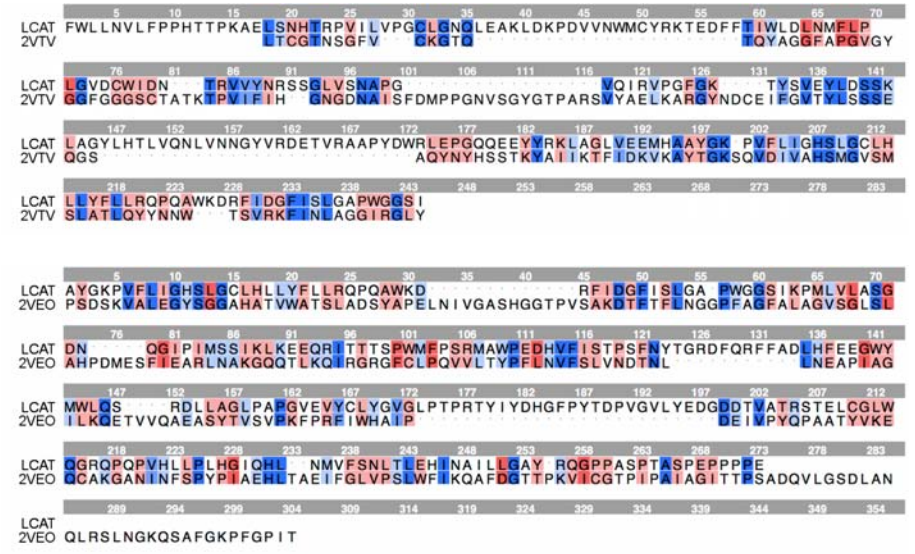


Figure 3.2. Alignment between LCAT N-terminal 'domain' and 2VTV and LCAT C-terminal 'domain' and 2VEO, color coded by similarity (BLOSUM62).

We independently modeled the two parts of LCAT: after loading the PSIPRED alignments into the suite MOE, we manually modified them to arrive at the best alignments, to optimize the overlapping residues of the two templates, and to minimize the gaps (alignment reported in Figure 3.3).
The parameters for the homology modeling procedures are described in detail under "Materials and Methods". One hundred different models of the whole LCAT were built, energetically and geometrically scored, and visually inspected.

We carefully checked the overall topology, the geometry (distances and orientation) of the catalytic triad, and the distance between the cysteine pairs forming the two disulfide bridges in LCAT. The two cystines have key structural roles: the first, Cys 50–74, stabilizes the N-terminal region, which is the putative sequence for lipoprotein recognition [Peelman 1999]; the second, Cys 313–356, is involved in the active site stabilization.
All the produced models had a correct topology, presenting seven strands in the central β-sheet with correct relative orientation. However, of all models produced, only nine structures had the cysteines of each pair sufficiently close to build of a disulfide bond. The best model had a

distance of 5.5 Å for the pair Cys 50 – Cys 74 and a distance of 5.18 Å for the pair Cys 313 – Cys 356.

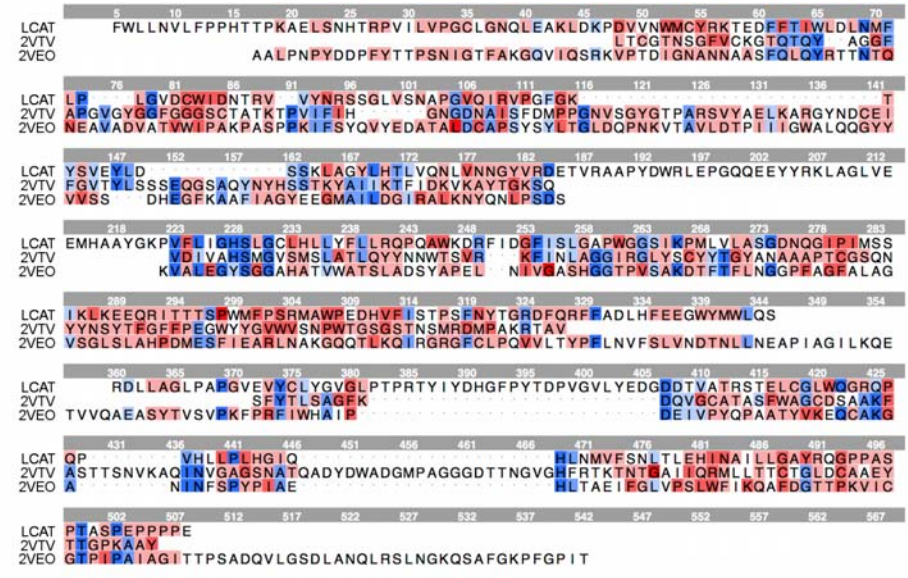


Figure 3.3. Global alignment used during the modeling procedures, color-coded by similarity (BLOSUM62).

However, none of these models showed a good geometry for the catalytic triad. This problem can be ascribed to the peculiar modeling procedure, since it is not easy to manage at the same time the best global alignment with the local geometric constraints imposed by the distances of the residues of the catalytic triad. In order to collect more knowledge about LCAT structure and to manage these structural issues, we decided to produce, in addition, an *ab initio* model of LCAT; to this purpose, we submitted the primary structure of LCAT, without its signal peptide, to the Robetta Web Server, using default parameters.

Only five of all the generated models had the correct general topology, and two of them lacked the correct geometry of the catalytic triad, and so were discarded without further investigations. In the three remaining models, the possibility to form cysteine disulfide bridges was evaluated: the C-terminal cystine was correctly predicted in all of them, whereas no acceptable predictions were obtained for the N-terminal cystine. Looking into the catalytic triad, one of the three models presented a very favorable interaction network among the three residues. Furthermore, also the localization of the oxyanion hole was correctly predicted. We selected this one as the best model obtained from the *ab initio* Robetta strategy.

Since the previous homology modeling procedure produced a better structured N-terminal 'domain', we decided to merge all the structural information obtained from both, the homology and the *ab initio* procedures, in a new, final LCAT model, by replacing the first part of the protein (residues 1-91) in the *ab initio* model with the same residues from the homology modeling procedure, as reported in Figure 3.4

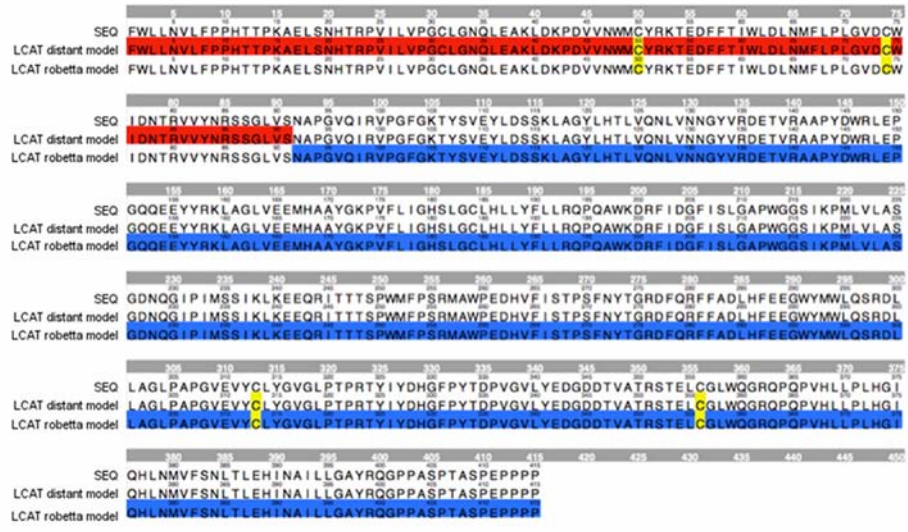


Figure 3.4. Alignment of different LCAT models: the distant homology model and the *ab initio* model. In red, the portion of LCAT distant homology model takes into account and in blue, the *ab initio* model used.

This final model structure of LCAT showed the correct topology, the correct mutual orientation of the catalytic triad residues, and the correct distance between all the cysteines involved in the formation of disulfide bridges. Table 3.1 reports geometric and energetic parameters of the final LCAT model.

Figure 3.5 shows the LCAT structure, after the formation of the two cystine bonds, and after a refinement step through energy minimization cycles. Figure 3.6 compares the triad in our LCAT model with the catalytic triad of a serine protease. subtilisin, from *Bacillus lentus*: reciprocal distance and residue orientation appear correctly predicted [Kuhn 1998].

Table 3.1. Scoring, according to different functions, of the final LCAT model [Kcal/mol]

	Contact Energy	**Packing**	**GB/VI**	**U**	**E sol**	**E ele**	**E vdw**	**E bond**
LCAT model	-186.76	13.22	-18273.37	2005.04	8100.11	-6947.34	4400.61	4954.83

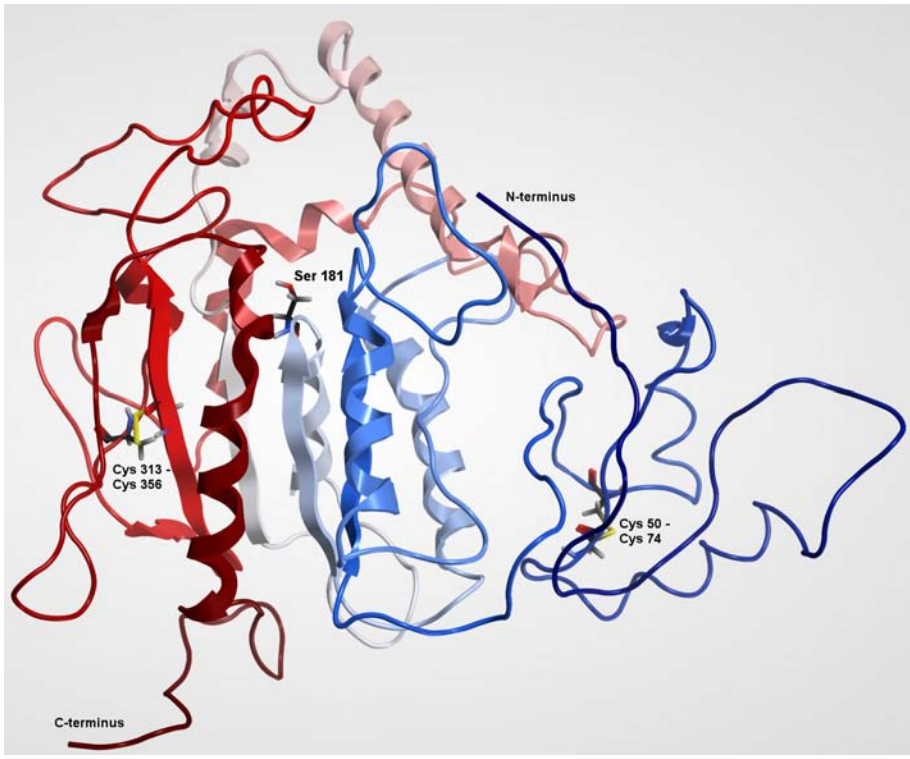


Figure 3.5. Final LCAT structure color-coded by terminus. In evidence the catalytic Serine 181 and the two cystines.

Then, we compared the distances between the catalytic residues with those reported by Peelman et al [*Peelman Protein Science 1998*]: distance between Oγ in Ser 181 and N2 in His 377 was 5.41 Å *versus* an expected distance of 2.5 Å, the distance between Oδ in Asp 345 and N1 in His 377 is 4.63 Å *versus* an expected distance of 2.9 Å. The distance between the oxyanion hole (Phe 103) and the catalytic triad was 6.75 Å *versus* an expected distance of 5 Å.

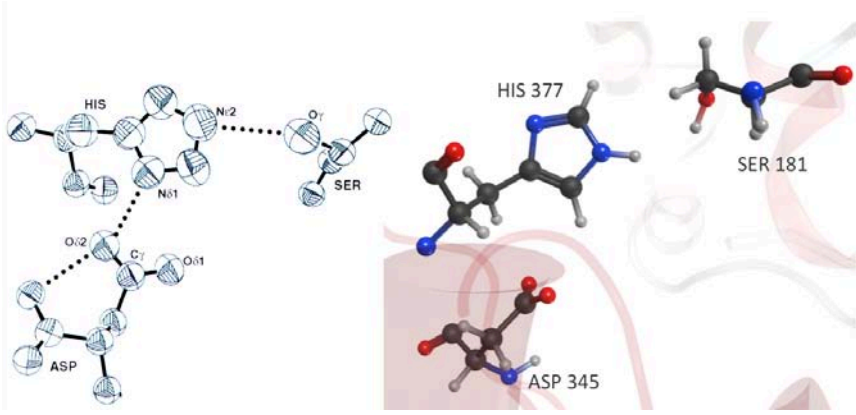


Figure 3.6. Comparison between Bacillus lentus subtilisin (in grey) and LCAT (colored) catalytic triads [modified from Kuhn 1998]

Definition of the active site

So far, the low identity of LCAT with other serine lipases and esterases has hindered the acquisition of more details about its structure.

The role of catalytic serine and histidine has been known since 1986, but the role of other putative catalytic amino acids, the presence of the oxyanion hole, and the role of the disulfide bridge have been debated for a long time. The catalytic triad has since been characterized as consisting of Ser181, Asp 347 and His 377, and Phe103, which is close to the triad, has been identified as the oxyanion hole. However, nothing is known at an atomic level of either the LCAT binding pocket or the lipoprotein interaction region.

The catalytic activity of LCAT, which is both phospholipasic and acyltransferasic, requires the possibility for both a lecithin and a cholesterol molecule to alternatively enter the catalytic site. For this reason we expected to find in the LCAT structure a hydrophobic pocket large enough to accomodate the substrates and the product, with an accessible catalytic triad placed at the basis of the pocket (active site), and not completely solvent-exposed.

We used the MOE's Site Finder tool to identify the binding site of LCAT natural ligands. MOE's Site Finder employs an alpha shape construction whereby points (atoms) on the protein's surface are identified, and the centers of spheres defined by combinations of four such points are marked. Where these clusters localize, there are potential binding sites on the protein's surface, excluding potential water sites. This approach falls into the category of geometric methods, because it does not provide an estimate of the interaction energy or van der Waals forces, methods with different levels of difficulty and computational costs [Miranker, 1991]. Furthermore, MOE's Site Finder does not use grids [Hendlich 1997] and its method is more efficient in terms of both memory use and CPU use. Finally, the MOE approach estimates the relative positions and accessibility of the receptor atoms very effectively and uses hydrophobic/hydrophilic classification to separate water sites from the more likely hydrophobic sites [Labute 2001].

The analysis of LCAT through the MOE Site Finder module revealed 27 putative binding sites; the top-scoring contained 346 contact atoms, among which 73 were hydrophobic and 256 involved sidechain atoms.

As can be seen in Figure 3.7, the identified binding pocket has all the expected features. The binding pocket is quite large, able to bind at the same time both the reaction substrates; it has hydrophobic features, and the catalytic triad faces the binding site, but is not fully solvent-exposed. In detail, from the figure, we can see that the binding pocket is divided into two hydrophobic regions (in yellow), separated by a thin, more exposed region (in blue).

In most lipases, it has been reported that a mobile lid covers the substrate binding site, and enzyme activation occurs upon binding to a hydrophobic substrate interface. In detail, in water, the access channel to the active site is closed by lid, and it opens only upon binding to a hydrophobic interface. Usually, the lid consists of a single short helix, and it is closed onto the active site. In contrast to the core of lipases, whose architecture is highly conserved, lids are less conserved elements, with significant variations in their length and different relative positions in the various lipases. In LCAT, several experiments have pointed out a highly amphipathic region: an α-helix in the N-terminal domain, closed, as in other lipases, by a disulphide bridge (Cys 50–Cys 74) [Peelman 1999]. This region is identified as an interfacial recognition domain for (apo)lipoproteins, and it could not only serve as a binding site for the hydrophobic substrate but further include a 'tilted' peptide, which is thought to destabilize the lipid substrate and facilitate the diffusion of a monomeric phospholipid or triglyceride into the active site cavity of the enzyme [Peelman 1997].

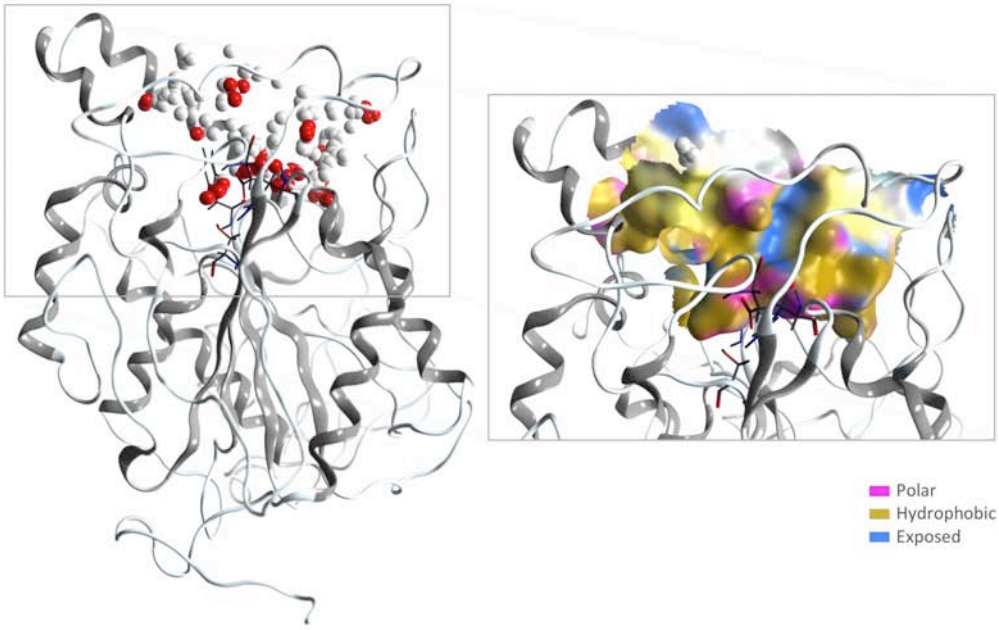


Figure 3.7. LCAT binding site identified by MOE's Site Finder. The alpha spheres are depicted in the whole LCAT structure, whereas in the right panel the molecular surface of LCAT active site is shown (hydrophobic regions in yellow, and solvent-exposed regions in blue).

This region is very distant from the enzyme active site, which is in the core of the protein – a region that is more conserved and easier to model. In our model we cannot properly define the N-terminal domain, because it is not sufficiently conserved, and because this common/shared lid is completely folded only in lipid environments.

Mutations mapping

Mutations in the *LCAT* gene cause both FLD and FED syndromes, both characterized by very low levels of HDL.

The new enzyme structure is a key to classify LCAT mutations on the basis of their effects on enzyme structure and function, and to define the genotype-phenotype relationship in carriers of LCAT mutations. Noteworthy demonstrations of how molecular modeling can be successfully used to predict the functional importance of specific amino acids and their mutations are reported in [Campbell 2007; Peelman 2001; Razzaghi 2001; Kobayashi 2002].

Mapping mutations on the new LCAT model may be useful to understand at a molecular level FED/FLD differences, and can allow predicting the impact of newly identified mutations on LCAT structure and function.

Among the different mutations described in the literature and collected at http://www.lcat.it/database.html, we focused our attention on 12 missense mutations, listed in Table 3.2. We discarded the mutation Thr (-13) Met, since it occurs in the signal peptide and does not directly affect LCAT three-dimensional structure.

All the listed missense mutations involve residues that are highly conserved in the LCAT sequence of vertebrate species for which LCAT primary structure is known.

It is important to note that our dataset includes two mutations affecting the same amino acid, but determining the two alternative pathologies.

Table 3.2. List of LCAT mutations and linked pathologies.

Mutant	***Mutant phenotype***
T-13M	FED
V46E	FED
A141T	FED
R244H	FED
T274A	FED
S91P	FLD
R140C	FLD
R147W	FLD
S181N	FLD
K218N	FLD
T274I	FLD
V309M	FLD
L372R	FLD

As reported in Figure 3.8, mutations don't cluster by disease but are spread over all the LCAT structure; it turns then impossible to topologically define a region responsible for FED and another for FLD.

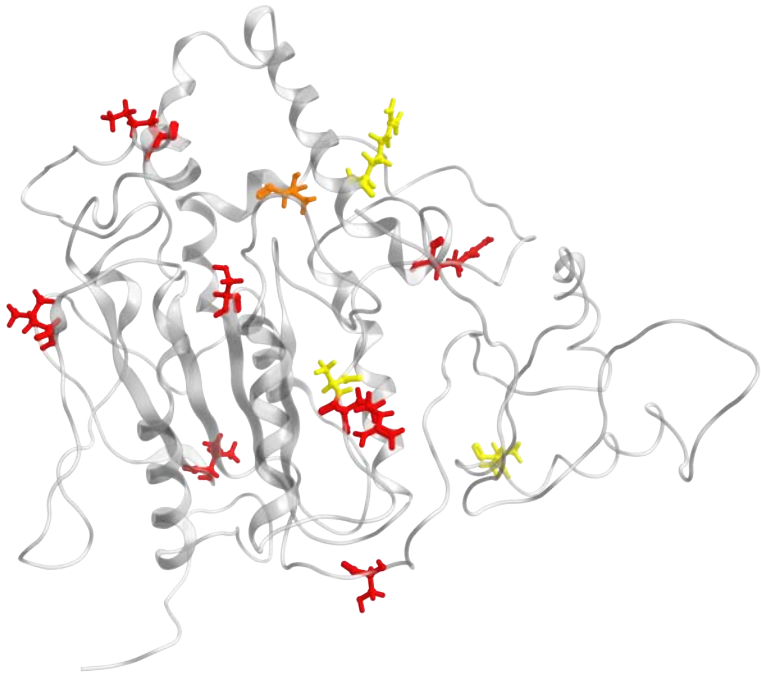


Figure 3.8. Mapping of LCAT mutations. In yellow residues involved in FED, in red those involved in FLD, in orange T274.

As a first approach to evaluate the impact of mutations on LCAT, we used two prediction tools available online: Polyphen (polymorphism phenotyping) [Adzhubei 2010] and PANTHER (Protein ANalysis THrough Evolutionary Relationships) [Thomas 2006]. Both are based on homology sequence analysis and on the evaluation of the physico-chemical properties of the amino acids, and use a machine-learning approach.
Polyphen predictions rely on physical and comparative considerations that estimate the impact of an amino acid replacement on the three-dimensional structure and on the function of the protein. Instead, PANTHER estimates the probability of a deleterious functional effect due to a given coding variant (Pdeleterious scores) by measuring substitution position-specific evolutionary conservation among evolutionarily-related proteins, without offering an explicit provision whether or not a mutation is possibly disease-causing but only yielding a probability evaluation.
As reported in Table 3.3, either tool produces a different output. Polyphen predicts all the test mutations as potentially damaging, except S91P and L372R, which are indicated as possibly damaging; this sorting is not useful to correctly classify the mutations that determine either FED or FLD diseases.
Conversely, PANTHER appears to be able to discriminate among the mutations those with a more severe impact on the carriers, with an average of 0.668 [pdeleterious] for FED mutations and 0.874 [pdeleterious] for FLD mutations.

Then, we analyzed in detail the effects of FED and of FLD mutations on LCAT three-dimensional structure. LowMode simulations of the wild-type and of the mutated protein were run to evaluate the impact of each mutation on the structure of the protein around the mutated residue.
For each structure, potential energy and distance from the reference structure were evaluated, constructing a RMSD matrix for each mutation. We analyzed each RMSD matrix to understand the impact of the mutation on the structure, checking if the mutation was able to destabilize the environment (increase in RMSD), with or without impact on the overall structure of the enzyme. The structures, scored for potential energy, are plotted in the RMSD matrices, in which structures with low RMSD (from 0.0 to 2.0 Å) are color-coded in blue, while structures with high RMSD (from 2.5 to 4Å) are color-coded in red.

All the results produced are summarized in Table 3.3.

Val 46 Glu
Val 46 has a pivotal role in a hydrogen-bond network involving also Asp 44, Trp 48 and Trp 75, and is located in a region pointed out as responsible for lipoprotein interaction. Mutation of Val 46 into Glu sharply reduces the environment hydrophobicity. However, mutation does not seem to destabilize the hydrogen bond network. The mutation modifies the local structural environment, but does not seem to affect the global structure of the protein and its activity. LowMode simulation doesn't show any significant modification in the stability of the environment around the mutation.

Ala 141 Thr
Ala 141 is located in the central β-sheet, a critical region for LCAT structure; its mutation into Thr changes significantly the environment, making it more hydrophilic. This mutation introduces a bigger residue, which bumps on the adjacent amino acids. Minimizing the mutated structure, the energy becomes less favorable, going from -10329 Kcal/mol to -6574 kcal/mol, and the β-strand starts to unfold. In contrast with these observations, during the LowMode simulation no significant differences in behavior are detected *versus* the wild-type protein.

Arg 244 His

Arg 244 has a pivotal role in a hydrogen bond network involving also Glu 149, Lys 240 and Thr 248. Mutation doesn't break the hydrogen bonds network, but LowMode simulation shows a significant decrease in stability of the mutation environment.

Thr 274 Ala

Thr 274 is close to Phe 103, which contributes to the definition of the oxyanion hole (2.5 Å), and its mutation into Ala increases the environment hydrophobicity. We can assume that the local conformational rearrangement due to the mutation causes a change in the Phe 103 side chain orientation. Also in this case, LowMode simulation shows a significant decrease in stability in the mutated amino acid environment.

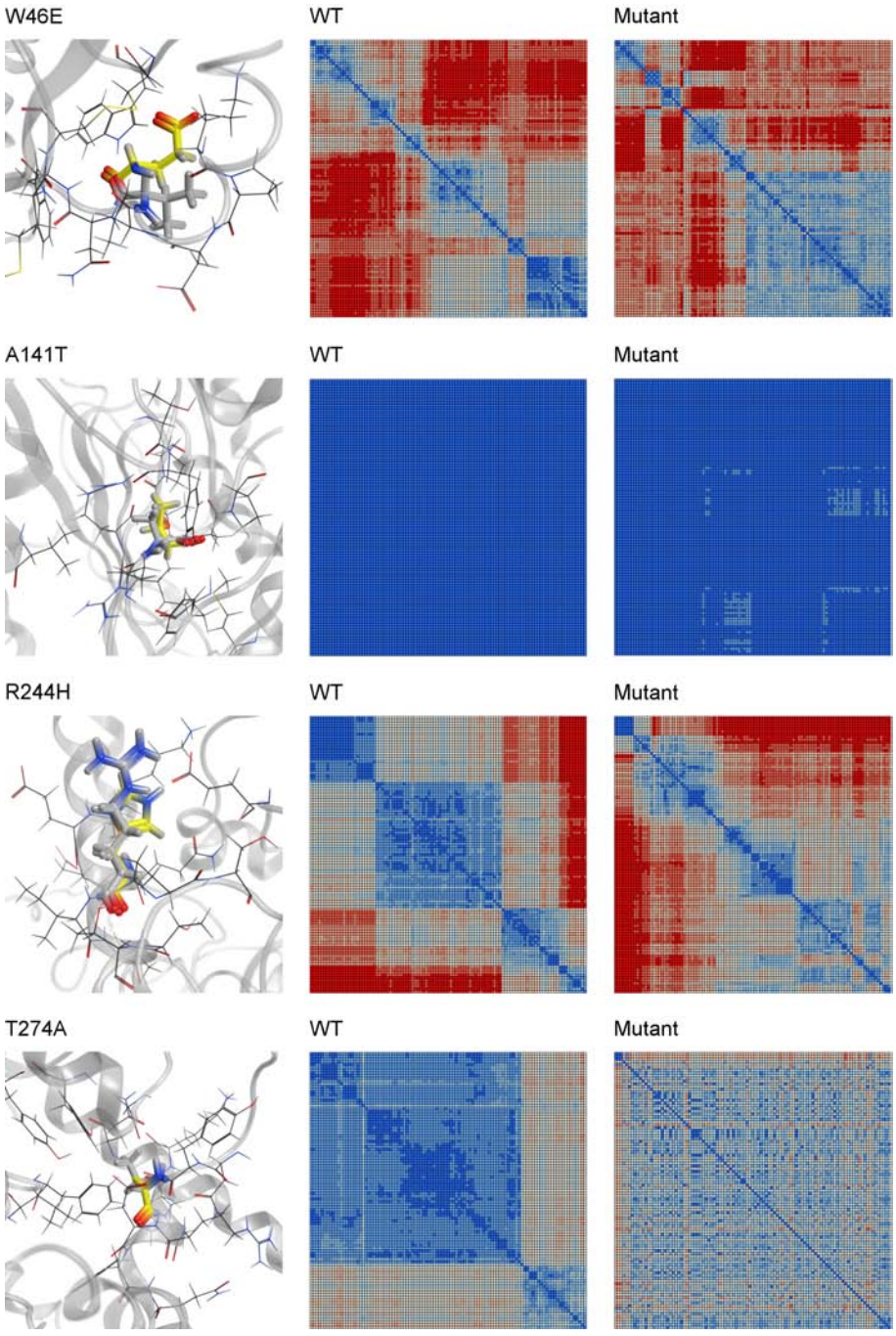


Figure 3.9. LowMode MD output. In figure are reported the FED mutations and the RMSD matrix for each mutation; in blue structure with low RMSD (from 0.0 to 2.0 Å), and in red structures with high RMSD (from 2.5 to 4Å).

FLD mutations determine a greater impact on LCAT structure:

Ser 91 Pro
Ser 91 is involved in a hydrogen-bond network also including Asp 136 and Lys 173. Mutation of Ser 91 increases the environment hydrophilicity and breaks the hydrogen-bond with Lys 173, impairing the α-helix structure around this residue.

Arg 140 Cys
Arg 140 has a pivotal role in a hydrogen bond network involving also Tyr 83, Thr 138, and Val 139. Mutation of Arg 140 changes this environment, increasing the hydrophilicity and breaking the hydrogen bond network. This rearrangement produces the loss of an important network of structural interactions.

Arg 147 Trp
Arg 147 is involved in a hydrogen bond network with Ser 255, Trp 259, Pro 260 and Asp 262. With the change from a charged/basic hydrophilic to an aromatic hydrophobic amino acid, this mutation alters the environment and impairs the hydrogen bond network. In addition, Trp bumps on Lys 159, determining a wide distortion of the local structure.

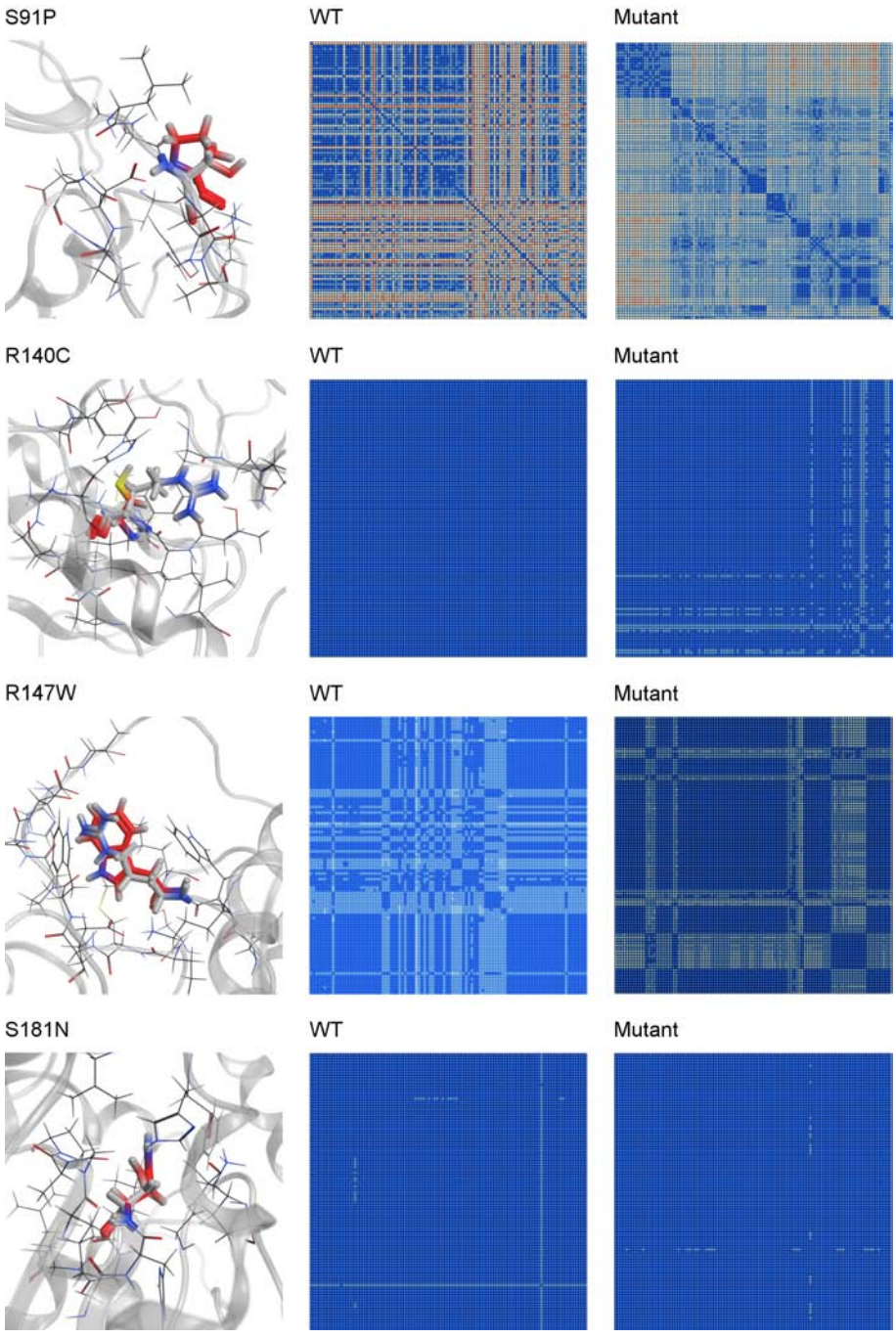
Ser 181 Asn
Ser 181 plays a direct role in the catalytic activity of LCAT. Any mutation of this amino acid completely blocks LCAT enzymatic activity.

Lys 218 Asn
Lys 218 is involved in a hydrogen-bond network with Val 222, Thr 334, and Pro 336. This mutation modifies the environment, since the residue changes from a charged/basic to a non-charged amino acid. In addition, we observed that this mutation produces the disruption of the local hydrogen-bond pattern. LowMode MD simulation shows a significant decrease in stability around the mutation.

Thr 274 Ile
As described for the FED mutation, in which the same residue is involved, Thr 274 is very close to Phe 103 (the oxyanion hole), and its mutation can modify the side chain conformation of Phe 103, determining a local conformational rearrangement. Ile is associated with a bigger steric bulk than Ala, and this can cause a wider local rearrangement, moving the Phe 103 side chain almost completely into the binding pocket, and thus blocking LCAT enzymatic activity. Indeed, no room remains for the stabilization of the tetrahedral intermediate typical of this trans-esterification reaction. LowMode MD simulation shows in the mutated environment a reduction of local flexibility.

Val 309 Met
Val 309 mutation into Met causes Met to bump on Leu 191. This modifies the local protein structure in a region very close to the catalytic site.

Leu 372 Arg
Leu 372 is involved in a hydrogen bond network with Gly 275 and Ser 236. Its mutation modifies the environment, since a neutral hydrophobic amino acid is substituted with a charged hydrophilic residue, impairing the hydrogen bond network. Leu 372 is located in the binding pocket, in which changes in local hydrophobicity can seriously impair the molecular recognition of LCAT substrates.


S91P
WT
Mutant
R140C
WT
Mutant
R147W
WT
Mutant
S181N
WT
Mutant

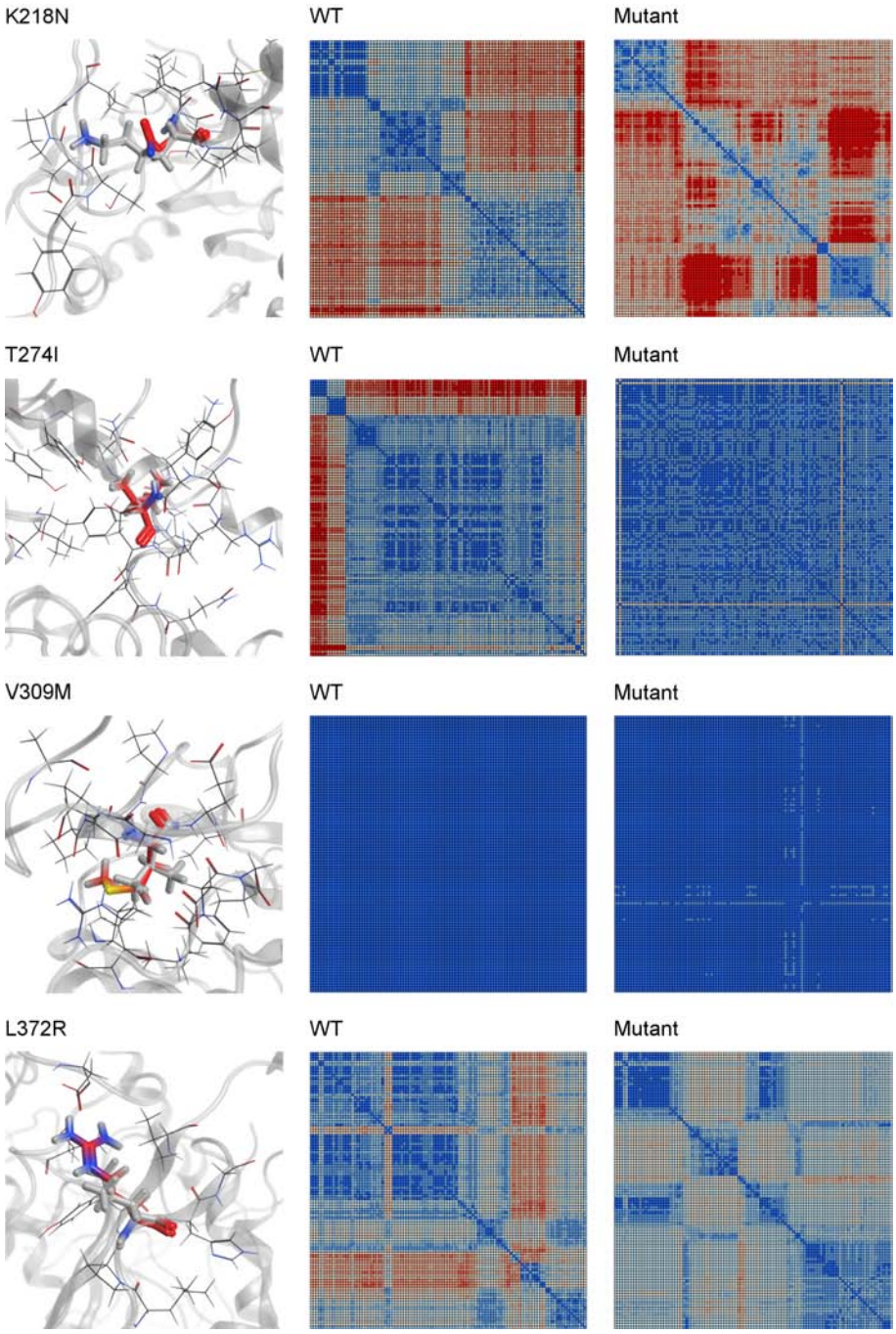


Figure 3.10. LowMode MD output. The FLD mutations are reported together with the RMSD matrix for each of them; structures with low RMSD (from 0.0 to 2.0 Å) are in blue, structures with high RMSD (from 2.5 to 4Å) are in red.

Table 3.3. Evolutive and energetic results about heterozygous LCAT mutations, obtained with different approaches: Panther, Polyphen and Lowmode MD.

Mutant	***Phenotype***	***Description***	***Panther results*** *[P deleterious]*	***Polyphen results***	***Lowmode results***
T-13M	FED	Mutation occurs in the signal peptide			
V46E	FED	Mutation makes environment more hydrophilic	0.474	Probably damaging	no significant differences
A141T	FED	Mutation make environment more hydrophilic	0.551	Probably damaging	no significant differences
R244H	FED	Mutation make environment less hydrophilic	0.794	Probably damaging	local effect (decrease in stability)
T274A	FED	Mutation causes local conformational rearrangement due to a change in the Phe103 side chain orientation	0.853	Probably damaging	local effect (decrease in stability)
S91P	FLD	Mutation causes impairing of α-helix structure	0.808	Possibly damaging	global effect (reduction of local flexibility)
R140C	FLD	Mutation produces the loss of an important network of structural interactions.	0.890	Probably damaging	no significant differences
R147W	FLD	Mutation changes the environment, passing from a charged hydrophilic to an aromatic hydrophobic amino acid, and impairs the hydrogen network	0.998	Probably damaging	no significant differences
S181N	FLD	Any mutation of this catalytic amino acid completely blocks the enzymatic activity.	0.989	Probably damaging	no significant differences
K218N	FLD	Mutation impairs the hydrogen-bond network	0.905	Probably damaging	local effect (decrease in stability)
T274I	FLD	Mutation causes wider local rearrangement than the T274A mutation	0.934	Probably damaging	local effect (reduction of local flexibility)
V309M	FLD	Mutation alters structure very close to the catalytic site	0.597	Probably damaging	no significant differences
L372R	FLD	Mutation alters local hydrophobicity and hydrogen-bond network in the binding pocket	0.871	Possibly damaging	no significant differences

3.4 CONCLUSIONS (Theme 1)

The solution of target structures by physicochemical methods is the best choice, when it can be pursued, but frequently severe experimental limitations make this way unsuitable both in the evaluation of protein structure-function relationships and in the deployment of a drug discovery project.
LCAT, as already explained in this thesis, is considered a very interesting but difficult target, and the very scanty number of published papers about its molecular structure demonstrates this. The use of the most up-to-date molecular modeling tools, however, allowed us to achieve interesting results in terms of structural knowledge. The combination of distant homology modeling, based on different templates, in order to model the protein core with a completely *de novo* three-dimensional structure prediction, and the merging of all the data coming from these procedures allowed us to build the first complete model for LCAT structure. This model, which satisfies all the published experimental data, was useful to localize and explain the effects of mutations identified in patients affected by LCAT genetic deficiencies with more reliability with respect to classical bioinformatics methods. Furthermore, the good agreement between the predicted and the clinical evidence helped us to validate the quality of the computed LCAT model. The usefulness of this model is multifaceted; the model can be used to further investigate the enzymatic activity of LCAT, to localize and evaluate the impact of newly identified mutations, and to develop new LCAT targeting strategies.

References

Adzhubei I.A. et al., 2010 A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-9.

Brunham L.R. et al., 2005 Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLoS Genet. 1(6):e83.

Calabresi L. et al. 2009a Functional lecithin: cholesterol acyltransferase is not required for efficient atheroprotection in humans. Circulation. 120(7):628-35.

Calabresi L. et al., 2009b A novel homozygous mutation in CETP gene as a cause of CETP deficiency in a Caucasian kindred.Atherosclerosis. 205(2):506-11.

Campbell G.R. et al., 2007 Tat mutations in an African cohort that do not prevent transactivation but change its immunogenic properties. Vaccine. 25(50):8441-7.

Eberini I. et al. 2011 *In silico* identification of new ligands for GPR17: a promising therapeutic target for neurodegenerative diseases.J Comput Aided Mol Des. 25(8):743-52.

Eldesbrunner H. et al. In: Proceedings of the 28th Hawaii international conference on systems science January 3-6, 1995, Kihei, Maui, Hawaii, USA. IEEE Computer Society (1995).

Hendlich M. et al., 1997 LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.J Mol Graph Model. 15(6):359-63, 389.

Jauhiainen M., Dolphin P.J. 1986 Human plasma lecithin-cholesterol acyltransferase. An elucidation of the catalytic mechanism. J Biol Chem. 261(15):7032-43.

Jonas A. 2000 Lecithin cholesterol acyltransferase. Biochim Biophys Acta. 1529(1-3):245-56.

Kobayashi Y. et al., 2002 Molecular modeling of the dimeric structure of human lipoprotein lipase and functional studies of the carboxyl-terminal domain. Eur J Biochem. 269(18):4701-10.

Kuhn P. et al., 1998 The 0.78 A structure of a serine protease: Bacillus lentus subtilisin. Biochemistry. 37(39):13446-52.

Labute P., Santavy M. 2001 Locating Binding Sites in Protein Structures. Chemical Computing Group Inc. http://www.chemcomp.com/journal/sitefind.htm.

Labute P. 2010 LowModeMD - implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. J Chem Inf Model 50:792–800.

McLean J. et al. 1986 Cloning and expression of human lecithin-cholesterol acyltransferase cDNA. Proc. Natl. Acad. Sci. U.S.A. 83 (8): 2335–9.

Miranker A., Karplus M. 1991 Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins, 11: 29–34].

Peelman F. et al. 1997 Structural and functional properties of the 154-171 wild-type and variant peptides of human lecithin-cholesterol acyltransferase. Eur J Biochem. 249(3):708-15.

Peelman F. et al. 1998 A proposed architecture for lecithin cholesterol acyl transferase (LCAT): identification of the catalytic triad and molecular modeling. Protein Sci. 7(3):587-99.

Peelman F. et al., 1999 Characterization of functional residues in the interfacial recognition domain of lecithin cholesterol acyltransferase (LCAT). Protein Eng. 12(1):71-8.

Peelman F. et al., 2001 Effect of mutations of N- and C-terminal charged residues on the activity of LCAT.Lipid Res. 42(4):471-9.

Razzaghi H. et al., 2001 Structure-function analysis of D9N and N291S mutations in human lipoprotein lipase using molecular modeling. J Mol Graph Model. 19(6):487-94, 587-90.

Sunyaev S, et al., 2001 Prediction of deleterious human alleles. Hum Mol Genet. 10(6):591-7.

Tanigawa H, et al., 2009 Lecithin: cholesterol acyltransferase expression has minimal effects on macrophage reverse cholesterol transport *in vivo*. Circulation. 120(2):160-9

Thomas P.D., Kejariwal A. 2004 Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A. 101(43):15398-403.

Thomas P.D. et al., 2006 Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools Nucl. Acids Res. 34 (suppl 2): W645-W650.

Weigelt J. 2010 Structural genomics-impact on biomedicine and drug discovery. Exp Cell Res. 316(8):1332-8.

Yang C.Y. et al. 1987 Lecithin:cholesterol acyltransferase. Functional regions and a structural model of the enzyme. J Biol Chem. 262(7):3086-91 -

3.5 LCAT MODULATORS

As recalled under "Introduction", a number of experimental data on mutated LCAT genotype suggest that the presence of defective LCAT might not preclude cholesterol removal from the arterial wall and efficient reverse cholesterol transport, despite low HDL levels [Degoma 2011]. Indeed, two recent studies have shown that low LCAT plasma levels are not associated with increased atherosclerosis risk in the general population and that higher LCAT levels are associated with an increased cardiovascular risk in women [Holleboom 2010; Baldassarre 2010].

However, till now, the idea of partially inhibiting LCAT activity with targeting molecules has never been put to a test. We are aware that LCAT inhibition can exert negative effects on plasma HDL concentration, but we believe that, on the other hand, it will have positive effects on HDL structure and function. In particular, the inhibition of LCAT activity will produce smaller HDL particles, which are more efficient both in cholesterol egress and in inhibiting the endothelial expression of cell adhesion molecules, with positive consequences for arterial protection.

To identify molecules able to specifically and selectively decrease LCAT activity, we followed two different strategies: an *in silico* high-throughput screening and an *in situ* drug design.

The first strategy is nowadays considered to be the best method for rapid identification of lead structures; it consists in a virtual and automatic evaluation of very large libraries of drug-like compounds [Ekins 2007]. The second strategy is the rational design of a molecule that mimics the *II* reaction intermediate, and that it is able to block LCAT, fully and selectively occupying its active site [Gomeni 2001; Noble 2000].

3.6 MATERIALS AND METHODS (Theme 2)

Molecular database preparation

The Asinex Platinum Collection (http://www.asinex.com/ download-zone.html) is a lead-like structural library containing approx. 130 000 in-house synthesized compounds. The SD file containing all the structures was downloaded and the MOE Conformation Import module was run on this file to produce a single low-energy conformation for each putative ligand contained in the Asinex SD file.

Molecular docking

The *in silico* screening was carried out with the Dock program contained in the MOE Simulation module. The full LCAT structure was set as Receptor. Before starting with the placement procedure, 1000 conformations were generated for each ligand by sampling their rotable bonds. The selected placement methodology was Triangle Matcher, in which the poses are generated by superposing triplets of ligand atoms and triplets of receptor site points. The protein site points are alpha spheres centers that represent locations of tight packing. Before scoring all the generated poses, duplicate complexes were removed. Poses are considered as duplicates if the same set of substrate-enzyme atom pairs are involved in hydrogen-bond interactions and the same set of ligand atom-protein residue pairs are involved in hydrophobic interactions. The accepted poses were scored according to the London dG scoring, which estimates the free energy of binding of the ligand from a given pose:

$$\Delta G = c + E_{flex} + \sum_{h-bonds} c_{HB} f_{HB} + \sum_{m-lig} c_M f_M + \sum_{atoms\_i} \Delta D_i \tag{3.1}$$

where c represents the average gain/loss of rotational and translational entropy; E_{flex} is the energy due to the loss of flexibility of the ligand (calculated from ligand topology only); f_{HB} measures geometric imperfections of hydrogen-bonds and takes a value in [0,1]; c_{HB} is the energy of an ideal hydrogen-bond; f_M measures geometric imperfections of metal ligations and takes a value in [0,1]; c_M is the energy of an ideal metal ligation; and D_i is the desolvation energy of atom i. The difference in desolvation energies is calculated according to:

$$\Delta D_i = c_i R_i^3 \left\{ \iiint_{u \notin A \cup B} |u|^{-6} du - \iiint_{u \notin B} |u|^{-6} du \right\} \tag{3.2}$$

where A and B are the protein and/or ligand volumes with atom i belonging to volume B; R_i is the solvation radius of atom i (taken as the OPLS-AA van der Waals sigma parameter plus 0.5 Å); and c_i is the desolvation coefficient of atom i. The coefficients $\{c, c_{HB}, c_M, c_i\}$ have been fitted from approx. 400 X-ray crystal structures of protein-ligand complexes with available experimental pK_i data. Atoms are categorized into about a dozen atom types for the assignment of the c_i coefficients. The triple integrals are approximated using Generalized Born integral formulas.

Only the top scoring solution was kept and submitted to a further refinement step, based on molecular mechanics (MM). In order to speed up the calculation, residues over a 6 Å cut-off distance away from the pre-refined pose were ignored, both during the refinement and in the

final energy evaluation. All receptor atoms were held fixed during the refinement. During the course of the refinement, solvation effects were calculated using the reaction field functional form for the electrostatic energy term. The final energy was evaluated using the MMFF94x forcefield with the Generalized Born solvation model (GBIV) [Wojciechowski 2004].
All the ligands contained in the Platinum library were screened according to the above procedure. The six top scoring compounds were resubmitted to the same docking procedure, keeping for each of them 300 poses. All of them were eventually bought from Asinex and tested in *in vitro* and *in vivo* experiments.
The same accurate docking procedure was applied to 'stirpex' (see in the following).
The estimated binding affinities and ligand efficiencies were calculated through the MOE LigX module. The pKi were computed through the binding free energies estimated with the London dG scoring function.

Protein ligand interaction fingerprints (PLIF)
PLIF is a method for summarizing the interactions between ligands and proteins using a fingerprint scheme. Hydrogen-bonds, ionic interactions, and surface contacts are classified according to the residue of origin, and built into a fingerprint scheme that is representative of a given database of protein::ligand complexes. All the 300 poses for the six selected compounds were submitted to the MOE PLIF program in its default configuration, except for the non-specific surface contacts which were disregarded; this generated a barcode plot summarizing the amino acid involvement in protein::ligand interactions.

***In vitro* and *in vivo* data**
CER - cholesterol esterification rate
The esterification of cholesterol within endogenous lipoproteins (cholesterol esterification rate [CER]) or incorporated into an exogenous standardized substrate (LCAT activity) was determined as previously described [Calabresi 2005]; [Murakami 1995]; [Franceschini 1990].
Plasma LCAT concentration was measured by an immuno-enzymatic assay [Murakami 1995].

Selectivity tests (trypsin/lipases/ACAT)
Trypsin inhibition was assessed by measuring the proteolytic activity of trypsin on a sensitive substrate (apoA-I) by SDS-PAGE [Ji & Jonas 1995].
The inhibition of lipase was evaluated with a colorimetric enzymatic assay, using the automatic analizer Cobas 8000 Hitachi (Roche Diagnostics).

Inhibition of acilCoA:cholesterol-acyltransferase (ACAT), an enzyme that catalyses esterification of cholesterol in cells and that has the same substrates as LCAT, was evaluated in peritoneal murine macrophages by measuring the esterification of oleic acid labelled with ^{14}C [Bernini 1997].

In vivo tests
C57BL mice have treated intraperitoneally with stirpex at a dose of 150 mg/kg. Blood samples were collected at different times after treatment [Zanotti 2011]. Plasma LCAT activity was measured by CER assay.

3.7 RESULTS AND DISCUSSION (Theme 2)

High-Throughput Screening results

We carried out the docking procedure in two steps: i) quick docking and ii) accurate docking (Figure 3.11). To evaluate the docking results in the refinement step of the procedures, we run molecular mechanics computations, and computed the final energy score with an empirical scoring function based on the MMFF94x forcefield with the generalized Born implicit solvation model (MM/GBIV).

We followed this procedure because the London dG function used after the ligand placement step does not take into account the van der Waals repulsive forces and, as such, only produced results with a favorable negative energy. Furthermore, since the molecular mechanics refinement step is based on the MMFF94x forcefield, it was more consistent to score the poses according to the same forcefield used during the calculations.

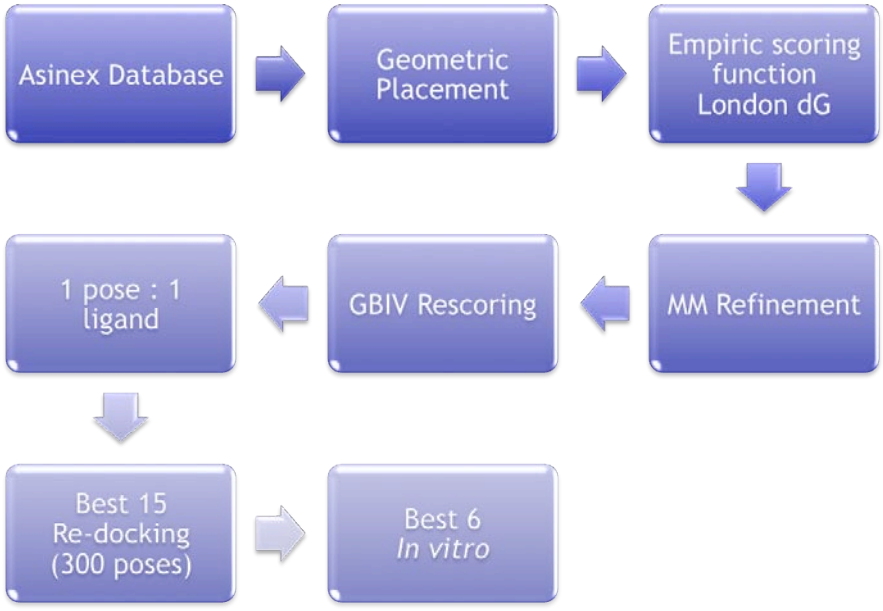


Figure 3.11. Complete docking pipeline.

All ligands in the tested Asinex database were evaluated using the quick docking procedure, refining and keeping only the best solution for each ligand. Figure 3.12 shows the MM/GBIV energy score plot for the 100 best docking solutions.

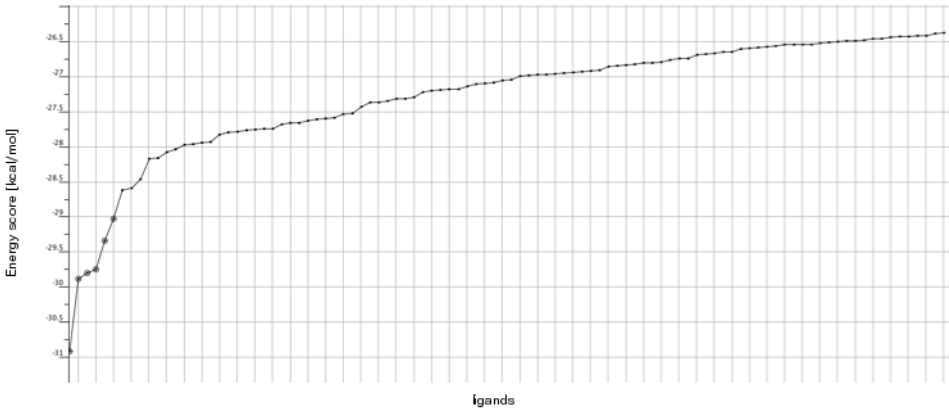


Figure 3.12. Plot of MM/GBIV energy score for the 100 best docking solutions.

The ligands corresponding to the six best poses show binding scores between -30.90 and -29.02 kcal/mol; these six compounds were submitted to the accurate docking procedure, generating 300 solutions (poses) for each ligand.
Figure 3.13 shows the plot of the binding energies for all the obtained solutions, sorted per molecule.

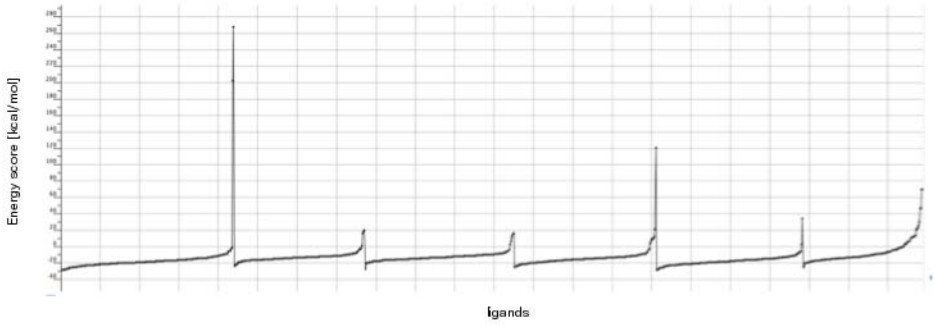


Figure 3.13. Plot of binding energies for the six best compounds, sorted per molecule.

Figure 3.14 reports the chemical structures of the six top-scoring compounds. They belong to different chemical classes, suggesting that the *in silico* screening allowed us to identify putative lead compounds with quite different chemical properties. In figure, the best poses of the selected compounds are reported. The active molecules completely block the access to the LCAT active site, even if they are located not very deeply in its binding pocket.

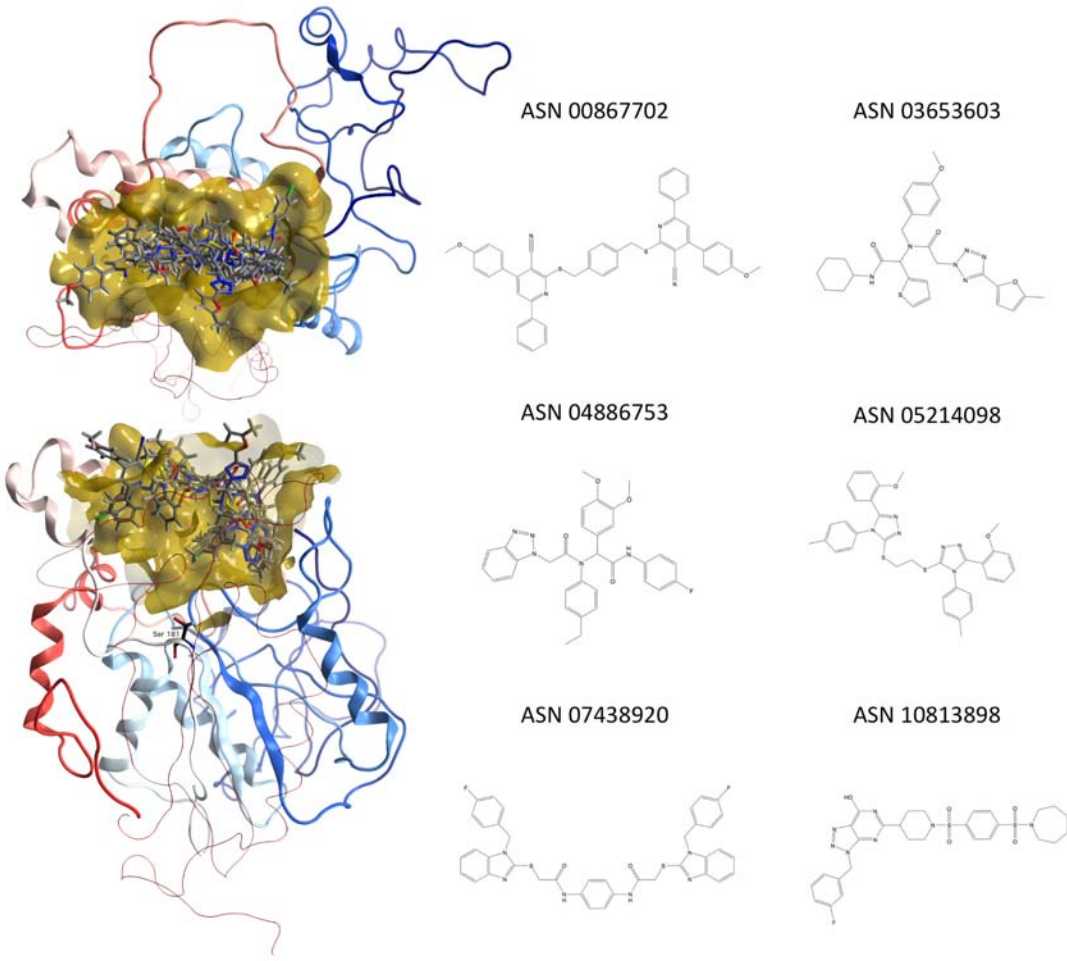


Figure 3.14, Chemical structures of the six top-scoring compounds and the corresponding lowest energy pose of these compounds interacting with LCAT binding site.

Table 3.4 reports some physico-chemical parameters for the selected ligands. The reported pKi values (-Log dissociation constant) are computed through the London dG scoring function and the Lig X MOE module, after a further refinement of the docked complexes. This and other empirical scoring functions are useful to rank the complexes according to their dissociation constant, as already discussed by Eberini et al [Eberini, 2011]. The docking score of the poses according to the MM/GBIV and the pKi values (affinities) computed through the London dG scoring function show a similar trend, suggesting that both these methods, based on different approaches, can be used to evaluate docking results and to compute approximate binding free energies for the system under investigation.

Table 3.4. Physico-chemical parameters for the six top-scoring ligands.

Ligand	MM/GBIV docking score (kcal/mol)	Affinity (pKi)
#1 ASN 00867702	-30.38	6.75
#2 ASN 03653603	-23.04	7.88
#3 ASN 04886753	-26.58	6.62
#4 ASN 05214098	-24.23	6.54
#5 ASN 07438920	-27.87	6.46
#6 ASN 10813898	-24.77	6.18

The PLIF barcode plot (Figure 3.15) analyzes all the poses from the docking refinement step for the six top-scoring compounds. Twenty-three amino acids are recognized as relevant in the interaction between LCAT and these compounds. Most of the residues positive to PLIF analysis belong to the set of amino acids previously identified by the MOE Site Finder module; exceptions are Leu 221, Glu 241, Arg 244, and Ile 326.
Gly 104, Glu 242, Ile 245, Tyr 327, and Phe 331 are the residues associated with the highest number of interactions with the tested ligands; all residues face the active site pocket.

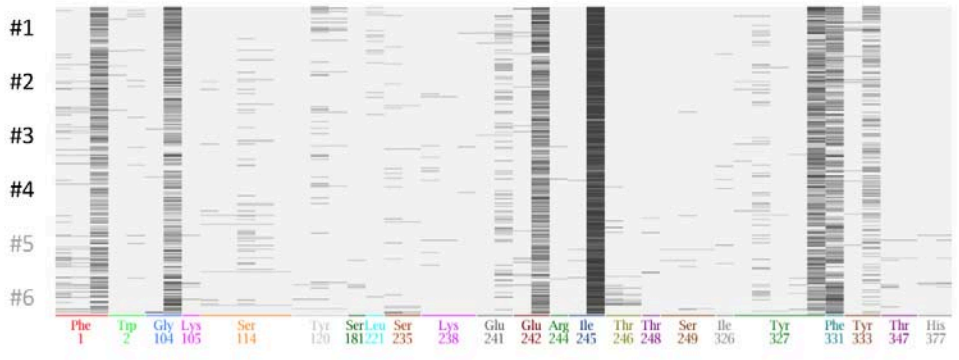


Figure 3.15. PLIF barcode plot. Each line corresponds to a ligand-LCAT interaction.

Observing in detail the identified molecules, these can be grouped into two distinct chemical families.
Family (I) contains compound 2 (ASN 03653603) and 3 (ASN 04886753), and has the general formula in Figure 3.16.

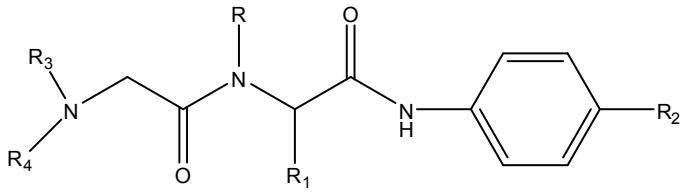


Figure 3.16. General formula for ASN 03653603 and ASN 04886753; where R is selected from the group comprising aryl, alkyl, benzyl optionally substituted; R1 is an optionally substituted aryl; R2 is selected from H, alkyl, F, Cl, Br, I; R3 and R4 are independently selected from the group comprising H, aryl, alkyl, benzyl optionally substituted, or R3 and R4 are closed to form a carbo- or hetero-cycle optionally substituted. In a preferred form, R is selected from aryl or benzyl optionally substituted in the para position; still more preferably, R is phenyl or methylphenyl substituted in the para position; R1 is an optionally substituted phenyl or a heteroaryl, still more preferably R1 is a thiophene; R2 is H or F; R3 and R4 are closed to form a hetero-cycle optionally substituted.

Family (II) collects compound 1 (ASN 00867702), 4 (ASN 05214098), 5 (ASN 07438920) and 6 (ASN 10813898), and has the general formula in Figure 3.17.

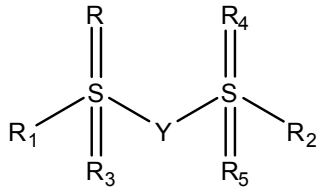


Figure 3.17. General formula for ASN 00867702, ASN 05214098, ASN 07438920 and ASN 10813898; where R, R3, R4, R5 can be independently present or absent, and when present, R, R3, R4, R5 are O, R1 and R2 are independently selected from alkyl or aryl optionally substituted, Y is an alkyl chain, aryl, alkyl -aryl optionally substituted and optionally containing heteroatoms or Y is a phenyl. One of these forms is the compound 6.
In a preferred form, R1 is alkyl and R2 is aryl, wherein said aryl is preferably in a chain to 8 carbon atoms. Even more preferably, said compound is selected from 1 long C8 or 5C8.
In a further embodiment, R, R3, R4, R5 are absent, R1 is p-methoxy benzyl-X, where X is a hetero aryl mono, bi or tri-substituted, and R2 is p-benzilmetossi Z, where Z is a hetero aryl mono, bi-or tri-substituted. Preferably, X and Z are equal to each other.

De novo design results

Once obtained the structural model of LCAT and clearly defined the molecular mechanism of the reaction catalysed, we designed a compound capable of acting as irreversible inhibitor of the enzyme, and useful as a laboratory pharmacological tool.
To design such an inhibitor while achieving the highest selectivity, we set to mimic the *II* intermediate of the reaction catalysed by the enzyme - a compound in which a cholesterol molecule and a fatty acid are bound together. Indeed, the *I* reaction intermediate could be similar for other enzymes that share the same catalytic triad, and this would certainly result in a loss of selectivity [Grochulski 1994].
For this reason, we designed a compound joining a molecule of cholesterol with a phosphonyl chloride group to a 17-carbon atom chain, an optimal length for the functionality of LCAT. This resulted in the synthesis of a heptadecylcholesteryl-(R,S)-phosphonyl chloridate, a compound that is able to fully occupy the active site of LCAT as identified in our *in silico* model (Figure 3.18) [Yang 1998].

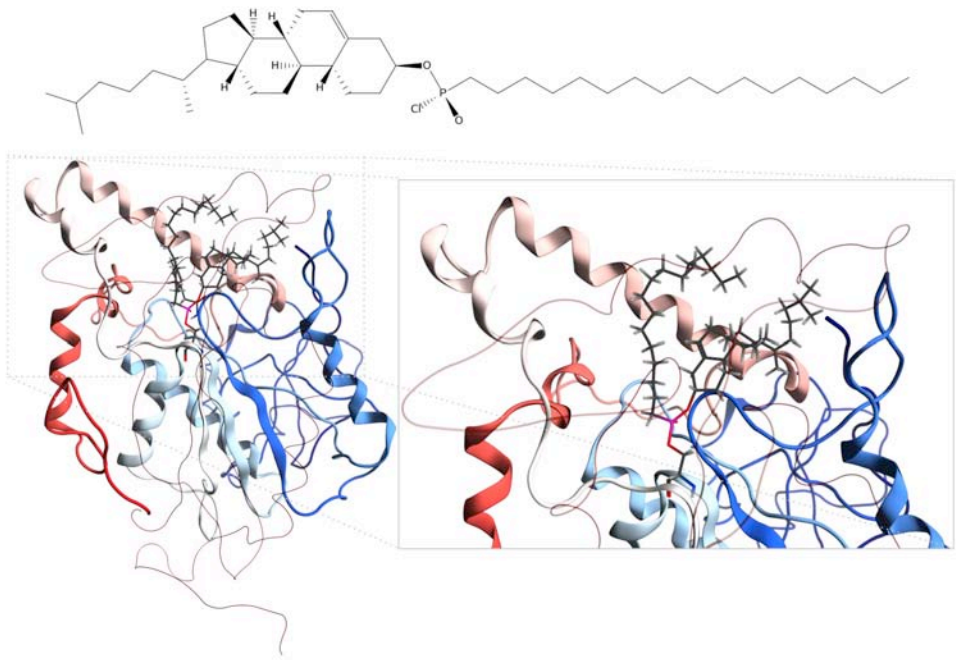


Figure 3.18. Chemical structures of the heptadecylcholesteryl (R, S) phosphonyl chloridate and corresponding lowest energy pose in LCAT binding site

Since this molecule has a chiral centre, both the enantiomers were tested *in silico*, and computational data suggested that the R-enantiomer is energetically favoured.
Henceforth, for convenience, we shall call "stirpex" the racemic mixture of heptadecylcholesteryl phosphonyl chloridate.
The binding energy of this molecule to LCAT (disregarding the covalent bond) was measured *in silico*, using the molecular docking protocol reported and the binding score is -29.40 kcal/mol.
Analyzing the docking results, we can recognize two distinct pocket portions able to bind: i) the cyclopentanoperhydrophenanthrene nucleus and ii) the long-chain fatty acid. As previously mentioned, both these regions are strongly hydrophobic.

***In vitro* and *in vivo* results**
The ability to affect the cholesterol esterification process of the identified LCAT inhibitors (both the HTS molecules and stirpex) was assessed *in vitro* by the measurement of the cholesterol esterification rate (CER) in control human plasma. In this experiment, modulators were added to control plasma and standardized substrates at different concentrations, and CER and LCAT activity were measured On the basis of the concentration-response curve obtained *in vitro*, the appropriate *in vivo* dose for stirpex was then established.
Figure 3.28 reports the dose-response curves of stirpex for the activity and selectivity assays *versus* ACAT, trypsin and lipase.

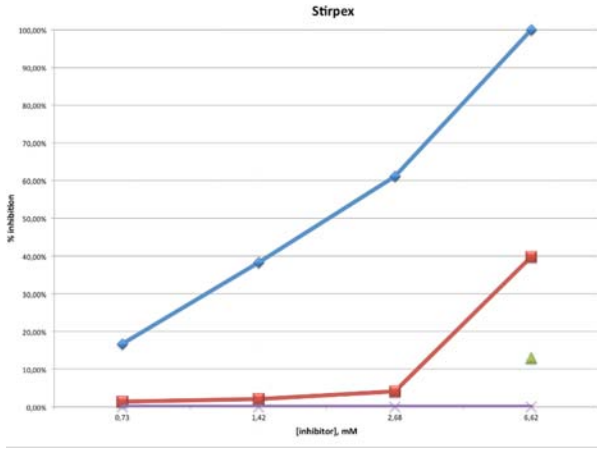


Figure 3.19. Dose-response curves of stirpex for the activity and selectivity assays *vs* LCAT (blue) ACAT (red), trypsin (green) and lipase (purple).

Figure 3.20 reports the dose-response curves of the Asinex compounds for the activity and selectivity assays *versus* ACAT and trypsin.

It is clear that all the molecules are able to completely inhibit the enzyme, although at different concentrations. All Asinex molecules completely inhibit LCAT at a concentration of 73 nmol, while stirpex only at 660 nmol. These results validate the computational model of LCAT three-dimensional structure: all the selected molecules are able to effectively interact with LCAT binding site.

We also performed selectivity tests for all the compounds *versus* other enzymes: one able to catalyze the formations of cholesterol esters, sterol O-acyltransferase 1 (ACAT), another able to hydrolyze proteins, trypsin. Stirpex was also tested *versus* a human lipase that shares with LCAT the same catalytic triad.
As reported in Figure 3.19, stirpex inhibits only LCAT, thus showing to be highly selective; this result indirectly confirms the quality of the active site model at an atomic level, since the compound was designed with high complementarity with respect to the active site modelled for LCAT.
As reported in Figure 3.20, all the molecules identified through HTS are unable to bind trypsin, and appear selective against other α/β hydrolases.
Compound 1 (ASN00867702) is slightly soluble and was not tested *in vitro*.

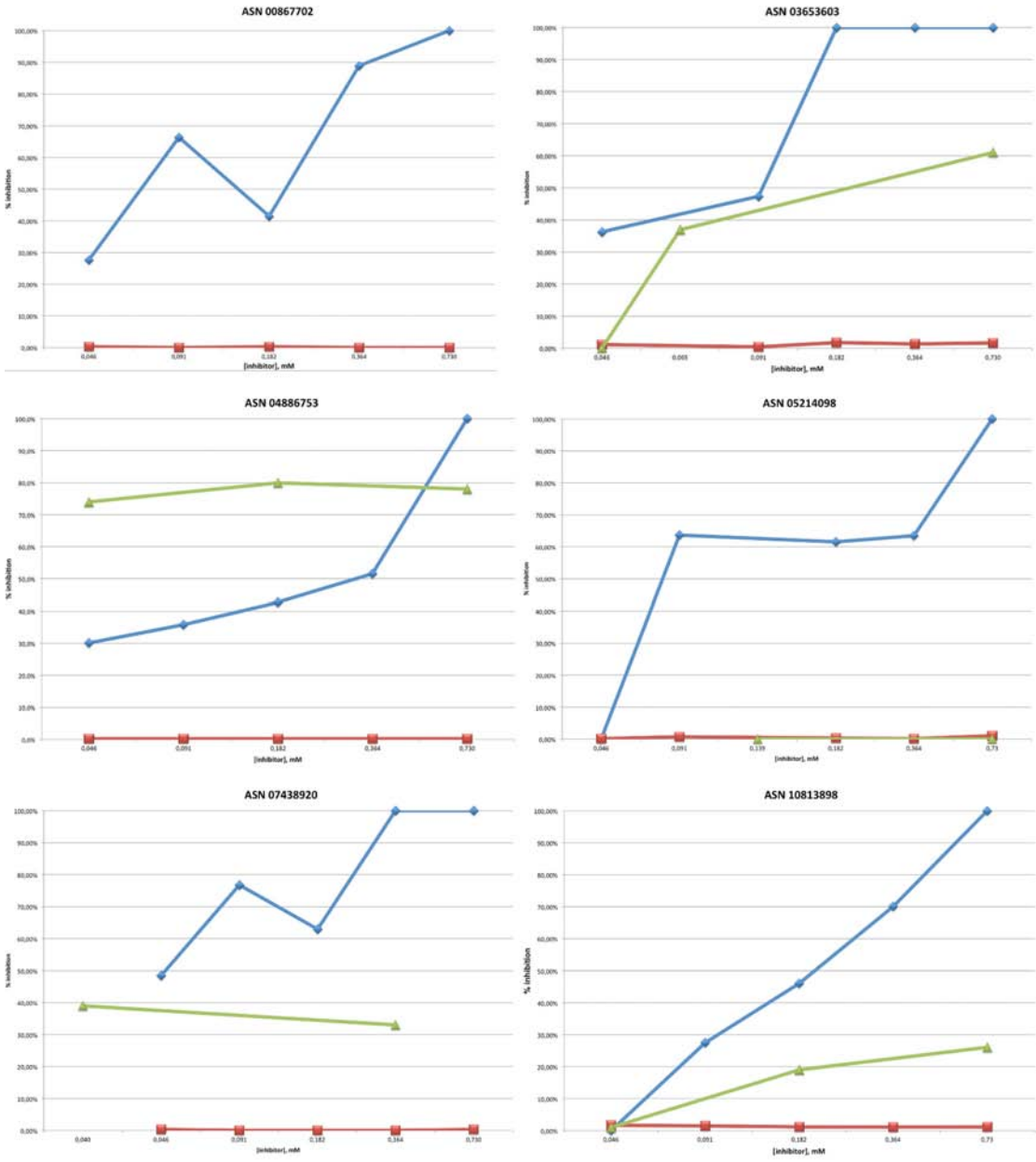


Figure 3.20. Dose-response curves of the six top-scoring compounds for the activity and selectivity assays *vs* LCAT (blue) ACAT (red), and trypsin (green).

However, some among the Asinex compounds, specifically compound 3 (ASN04886753) and 2 (ASN 03653603), can also inhibit ACAT. These compounds are thus incompletely selective, and were discarded without submitting them to the following analysis step. Compound 4 (ASN05214098) and 6(ASN10813898) showed limited inhibitory activity of ACAT, just at very high concentrations, and can therefore be defined as LCAT selective inhibitors.

In order to test the ability of our compounds to inhibit LCAT *in vivo*, an inhibition assay was carried out in the C57BL mouse model. Murine LCAT is highly similar to human (86% identity). Seven mice were treated intraperitoneally with stirpex at a dose of 150 mg/kg. Blood samples were collected at different times after treatment, and LCAT activity in plasma was evaluated. The time-course of the inhibition assay is summarized in Table 3.5.
For the time being, stirpex is the only LCAT inhibitor tested *in vivo*.

Table 3.5. Syrpex *in vivo* inhibition assay.

	LCAT activity (0h) *nmol CE/ml/h*	**LCAT activity** (6h) *nmol CE/ml/h*	**LCAT activity** (30h) *nmol CE/ml/h*	**LCAT activity** (45h) *nmol CE/ml/h*
Mouse 1		6.0	0.0	
Mouse 2		7.5	0.0	
Mouse 3		7.0	0.0	
Mouse 4	6.5		3.3	0.0
Mouse 5	2.3		2.5	0.0
Mouse 6	3.1		3.0	0.0
Mouse 7	6.3		0.0	0.0

The compound is absorbed, distributed and can efficiently bind to, and inhibit, LCAT. Complete inhibition occurs between 30 and 45 hours after administration.
These data further confirm the high specificity of stirpex, whereas the selectivity of some of the reversible compounds suggests that the LCAT binding site has been modeled in a sufficiently accurate fashion.

3.8 CONCLUSIONS (Theme 2)

The most up-to-date pharmacology strategies for drug discovery require the identification of specific targets that play a key role in the development of a specific pathology. The knowledge of the structure of the identified putative targets is a necessary step in the drug discovery pipeline. Differently, the possibility to modulate/operate the selected targets is assigned to a serendipitous process or to the deduction of molecules already in use for the modulation of other targets, frequently phylogenetically correlated. These two possibilities are not very effective: serendipity efficiency is completely unpredictable, whereas the use of molecules coming from the pharmacology of other targets generally fails in term of specificity and selectivity.

Recently, molecular modeling has made available a set of approaches, methods and tools useful to successfully manage this issue and to increase the structural knowledgebase, especially when working with difficult targets. The computational approach can be deployed in two different ways: i) the availability of large chemical databases and the computational power of the new computers allow a very efficient *in silico* high-throughput screening starting form the target structure and independently from any pharmacophoric hypothesis; ii) on the other hand, the knowledge of the active site and of the catalytic mechanism of a target can be exploited for *de novo* computer-aided drug design.

Our *de novo* design of an irreversible inhibitor is based on the knowledge of LCAT catalytic mechanism. A molecule able to mimic the *II* reaction intermediate, which is peculiar of LCAT and ACAT proteins, can be an interesting pharmacological tool to investigate the effects of LCAT inhibition on atherosclerotic disease. Up to now, stirpex is the first specific LCAT inhibitor, which does not demonstrate any inhibitory activity on other proteins sharing the same catalytic triad. For obvious reasons, stirpex cannot be proposed as a drug. Further modifications of this molecule could allow us to identify new active molecules belonging to the same chemical family. On the other hand, the high-throughput screening carried out on LCAT identified some very potent reversible inhibitors, which can be of some interest in the further drug development step. The demonstration that some of the identified molecules are selective for LCAT with respect to other Ser/Asp/His proteins, and especially with respect to ACAT (which is a very close paralogous protein), suggests that our distant comparative modeling strategy allowed us to predict with a good approximation the shape of the enzyme active site.

We deem that, at the state of the art, molecular modeling has become a necessary and efficient approach both to increase the knowledge of newly identified targets and to carry out in a rational way the first steps of the long pipeline towards drug discovery in pharmacology.

Refereces

Baldassarre D. et al. 2010 Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study. Eur Heart J. 31(5):614-22.

Bernini F. et al. 1997 Effect of lacidipine on cholesterol esterification: *In vivo* and *in vitro* studies. Br J Pharmacol 122:1209–1215.

Calabresi L. et al., 2005 The molecular basis of lecithin:cholesterol acyltransferase deficiency syndromes: a comprehensive study of molecular and biochemical findings in 13 unrelated Italian families. Arteriosclerosis, thrombosis, and vascular biology 25 (9): 1972–1978.

Degoma E.M., Rader D.J. 2011 Novel HDL-directed pharmacotherapeutic strategies. Nat Rev Cardiol. 8(5):266-77.

Ekins S. et al., 2007 *In silico* pharmacology for drug discovery: applications to targets and beyond.Br J Pharmacol. 152(1):21-37.

Franceschini G. et al., 1990 Apolipoprotein A-IMilano. Partial lecithin:cholesterol acyltransferase deficiency due to low levels of a functional enzyme. Biochim Biophys Acta. 1043: 1–6.

Gomeni R. et al., 2001 Computer-assisted drug development (CADD): an emerging technology for designing first-time-in-man and proof-of-concept studies from preclinical experiments. Eur J Pharm Sci. 13:261–270.

Grochulski P. et al., 1994 Analogs of reaction intermediates identify a unique substrate binding site in Candida rugosa lipase. Biochemistry. 33(12):3494-500.

Holleboom A.G. et al., 2010 Plasma levels of lecithin:cholesterol acyltransferase and risk of future coronary artery disease in apparently healthy men and women: a prospective case-control analysis nested in the EPIC-Norfolk population study. J Lipid Res. 51(2):416-21.

Ji Y. and Jonas A. 1995 Properties of an N-terminal proteolytic fragment of apolipoprotein AI in solution and in reconstituted high density lipoproteins. J Biol Chem. 270(19):11290-7.

Murakami T. et al., 1995 Triglycerides are major determinants of cholesterol esterification/transfer and HDL remodeling in human plasma. Arterioscler Thromb Vasc Biol. 15: 1819–1828.

Noble D., Colatsky T.J. 2000 A return to rational drug discovery: computer-based models of cells, organs and systems in drug target identification. Emerging therapeutic targets. 4:39–49.

Wojciechowski M., Lesyng B. 2004 Generalized Born Model, Analysis, Refinement and Applications to Proteins, J. Phys. Chem. B 108:18368–18376.

Yang G. et al., 1998 Tetrazole-Catalyzed Synthesis of Phosphonamidate Esters Tetrahedron Letters, Volume 39, Number 17, pp. 2449-2450(2)

Zanotti I. et al., 2011 The thienotriazolodiazepine Ro 11-1464 increases plasma apoA-I and promotes reverse cholesterol transport in human apoA-I transgenic mice. Br J Pharmacol. 164(6): 1642–1651.

Appendix

Domanda di Brevetto per INVENZIONE INDUSTRIALE
Numero domanda: MI2012A001980

Richiedenti
Laura Calabresi, Ivano Eberini, Guido Franceschini, Cristina Sensi

Titolo
"Inibitori dell'enzima lecitina:colesterolo aciltransferasi"

Descrizione
La presente invenzione è relativa a composti per uso medico dotati di attività inibitoria verso l'enzima lecitina:colesterolo aciltransferasi.

Acknowledgment

I would like to thank all the people who contributed to this work: Franco Bernini, Franco Bonomi, Laura Calabresi, Ivano Eberini, Anita Ferraretto, Guido Franceschini, Elisabetta Gianazza, Ambrogina Pagani, Chiara Parravicini, Sara Simonelli, Cesare R. Sirtori and Ilaria Zanotti, for the assistance they provided at all levels of research project.
In particular, I would like to thank Laura Calabresi and Sara Simonelli from the Center "E. Grossi Paoletti" and the analysis laboratory of Niguarda Hospital for the *in vitro* assays and Franco Bernini and Ilaria Zanotti for the *in vivo* tests.
So I would like to express my gratitude to Ivano Eberini and Elisabetta Gianazza, whose expertise, understanding and patience, added considerably to my Ph.D. experience.
I wish to express my love and gratitude to my beloved parents, for their understanding and endless love, through the duration of my studies.