Lipid Metabolism and Comparative Genomics

Problem presented by

Janette Jones, Brendan O'Malley, Patrick Warren,
Laura Pickersgill and John Melrose

Unilever Corporate Research

Problem statement

The Study Group participants were asked to focus on two questions relating to metabolism. The first of these concerned modelling lipoprotein metabolism such that differences in the biology of the healthy and obese states can be encompassed as well as changes in the size and composition of lipoprotein particles. The use of ordinary differential equation models to infer rate constants from experimental data was discussed and a partial differential equation model was developed to describe the dynamics of lipoprotein particle formation.

The second problem focused on understanding how certain components of well understood biological networks can help in determining the functionality of other networks, in which certain components are not so well determined. The study group participants focused on issues regarding comparative genomics to discuss this question. A number of issues were addressed including the derivation of parameter values from experimental data via a stoichiometric matrix and the importance of randomly sampling and comparing sections of known networks.

Study Group contributors

David Broomhead (University of Manchester)
Jens Gravesen (Technical University of Denmark)
Poul Hjorth (Technical University of Denmark)
James Ing (University of Aberdeen)
John King (University of Nottingham)
Bill Lionheart (University of Manchester)
Eirik Mo (Norwegian University of Science and Technology)

James Parrott (University of Bristol)
Jonathan Rougier (University of Durham)
Marcus Tindall (University of Oxford)
Eddie Wilson (University of Bristol)

Report prepared by

Marcus Tindall (University of Oxford)
Jonathan Rougier (University of Durham)
Laura Pickersgill (Unilever Corporate Research)
John Melrose (Unilever Corporate Research)
Brendan O'Malley (Unilever Corporate Research)
Janette Jones (Unilever Corporate Research)

1 Detailed Problem Statement

The Study Group participants were asked to focus on two problems related to the modelling of metabolic networks. The first concerned dysregulated lipid metabolism which is a feature of many diseases including metabolic syndrome, obesity and coronary heart disease. As one of the world's largest foods companies Unilever has a responsibility to produce health promoting foods and thus requires better understanding of lipoprotein metabolism and the key drivers of its dysregulation in order to develop products that further improve heart health. The Study Group was asked to develop a simple model of the kinetics of lipoprotein metabolism between healthy and obese states incorporating the activities of key enzymes.

A second question concerned the use of comparative genomics in understanding and comparing metabolic networks in bacterium. In modelling metabolic networks a key step is understanding the interactions between network components where some parts are experimentally well mapped and other regions are data poor. We take, as an example system for this problem, bacterial metabolism where some species, e.g. *E. coli*, are well mapped. From this we wish to infer parts of a comparative metabolic network, which have not been measured, in another bacterium. Metabolic maps are often constructed using comparative genomics methods, in which one compares the genome of new organism with that of a previously characterised organism, to infer the presence of enzymes and metabolic pathways in the new organism. The first interesting mathematical problem is how one can quantify, in a statistical sense, such a metabolic map. In particular where there are different levels of confidence for presence of different parts of the map. The next and most important question is how one can design a measurement strategy to maximise the confidence in the accuracy of the metabolic map.

2 Lipid Metabolism

Lipoproteins are a family of macromolecular complexes, which function to transport dietary and endogenously produced lipids through the aqueous environment of the plasma. Lipoprotein metabolism describes the way in which lipids are transported throughout the various physiological compartments of the body. Chylomicrons (CM), the largest lipoproteins, carry exogenous triglyceride (triglyceride) and cholesterol (C) from the intestine via the thoracic duct to the venous system. CM are heterogeneous particles consisting of a triglyceride and cholesterol ester (CE) core surrounded by a phospholipid (PL) layer which contains free cholesterol (C) and various apolipoproteins. In the capillaries of peripheral tissues such as adipose tissue and skeletal muscle, 90CM-triglyceride is removed through the action of lipoprotein lipase (LPL). Fatty acids and glycerol, derived from hydrolysis of CM, enter the adipocytes and muscle cells and are used as fuel, stored or metabolised. As a result of repeated lipase action, progressively smaller, triglyceride-poorer CM remnants (CMR) are formed. CMR are removed from the circulation by the liver in a receptor-mediated process involving interaction with apolipoproteins on CMR.

The liver releases endogenous triglyceride and CE in very low density lipoproteins (VLDL). VLDL can be separated into two subclasses, $VLDL_1$ and $VLDL_2$, which differ in their triglyceride-content and hence size. Under conditions of increased hepatic triglyceride content, the rate of production of triglyceride-rich $VLDL_1$ released is increased. The same lipases that act on CM quickly degrade endogenous triglyceride in VLDL, giving rise to intermediate density lipoproteins (IDL) that have had much of their triglyceride and surface apolipoproteins removed. IDL is further degraded by removal of more triglyceride, producing low density lipoprotein (LDL). LDL has a plasma residence time of 2 to 3 days. Approximately 70% of LDL is removed by the liver via LDL receptors (LDLR) present on the surface of hepatocytes and other cells which bind to apolipoprotein B100. In addition, a small but significant amount of LDL appears to be removed from the circulation by non-receptor mediated pathways, including uptake by scavenger receptors on macrophages that may migrate into arterial walls and form foam cells which lead to atherosclerotic plaque formation.

Dysregulated lipoprotein metabolism is a feature of many disease states including obesity. In the obese state there are many alterations in behaviour of the lipoprotein metabolism system. Many obese subjects exhibit a reduced LPL activity compared to lean controls. This results in an increased resonance time of CM in the plasma and increased delivery of triglyceride-rich CM to the liver. The hepatic triglyceride content is further increased in the obese state due to an increased flux of fatty acids entering the portal vein from adipose tissue. Consequently, there is an increase in the amount of large triglyceride-rich $VLDL_1$ produced, which due to the reduced LPL activity and competition from uncleared CM, is converted to LDL at a reduced rate. In addition cholesterol ester transfer protein (CETP) is upregulated in the obese state. This enzyme catalyzes the exchange of triglyceride in VLDL and IDL for CE in LDL and the lipoprotein responsible for reverse cholesterol transport, high density lipoprotein (HDL) as demonstrated in Figure 1. As a result of this enzyme activity, the lipoprotein contents are remodelled resulting in less C being returned to the liver via HDL and increased triglyceride content of LDL and hence the generation of an altered LDL known as small dense LDL (sdLDL). Sd LDL have an increased residence time in the plasma, hypothesised to be a result of a change in conformation of apoB100 which prevents binding to the LDLR. In addition, LDLR expression is down-regulated in obesity, further contributing to the increase in plasma LDL-C. Consequently these atherogenic LDL particles are less likely to be returned to the liver and more likely to contribute to the development of atherosclerotic plaques.

The main focus of the work undertaken here is to obtain kinetic rates for the production of the VLDL, IDL and LDL lipoprotein species from experimental data from obese individuals. The ultimate aim is to combine this kinetic data with other biological measurements and to compare lipoprotein metabolism in healthy and obese subjects to gain insight into what governs the formation of sdLDL.

2.1 Experimental Data and ODE models

The lipoprotein literature includes a number of experimental articles describing the rates of transition from VLDL to IDL to LDL. Such measurements are generally obtained from

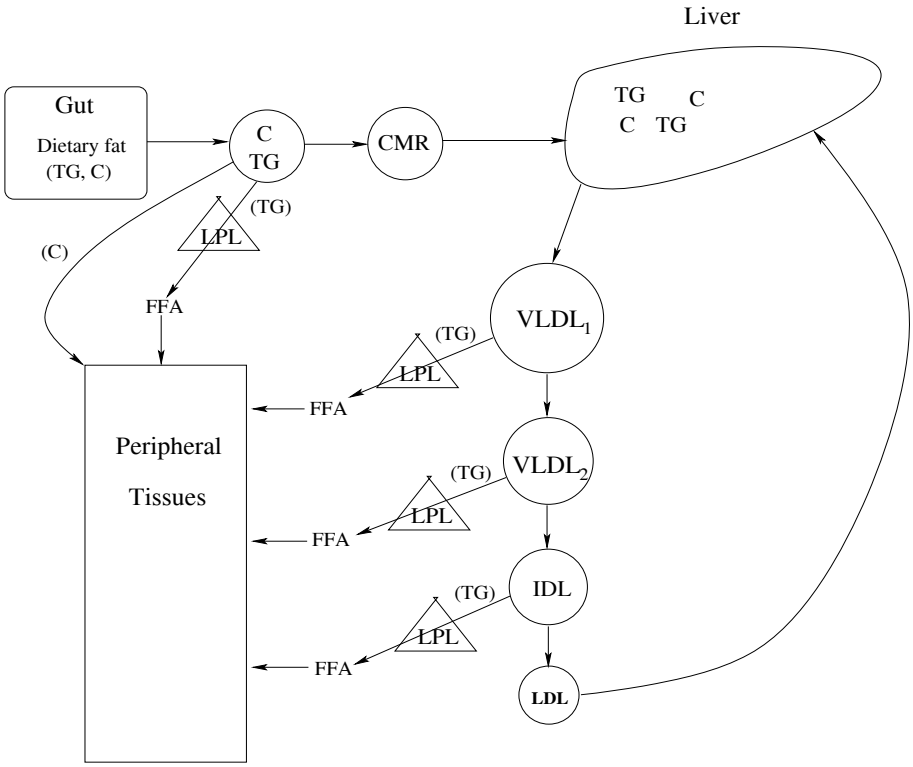


Figure 1: A schematic representation of the main biochemical steps in lipoprotein metabolism between the human gut, liver and peripheral tissue for a healthy person. The total triglyceride (TG) and cholesterol (C) content of the lipoprotein molecules varies as they move through the metabolic process.

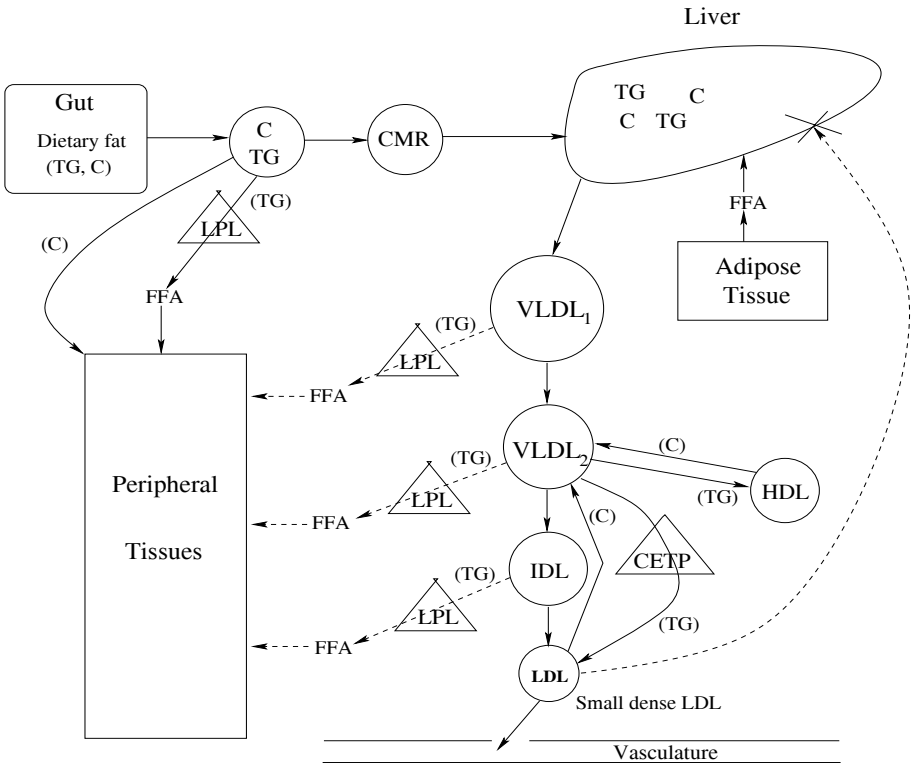


Figure 2: A schematic representation of the main biochemical steps in lipoprotein metabolism between the gut, liver and peripheral tissue for an obese person. Note the down-regulation of LPL activity, the up-regulation of HDL and CETP leading to an increase in cholesterol and triglyceride transfer and a reduction in liver-LDL receptor activity. The dotted lines indicate a down regulation in the function of that pathway in comparison to a healthy individual.

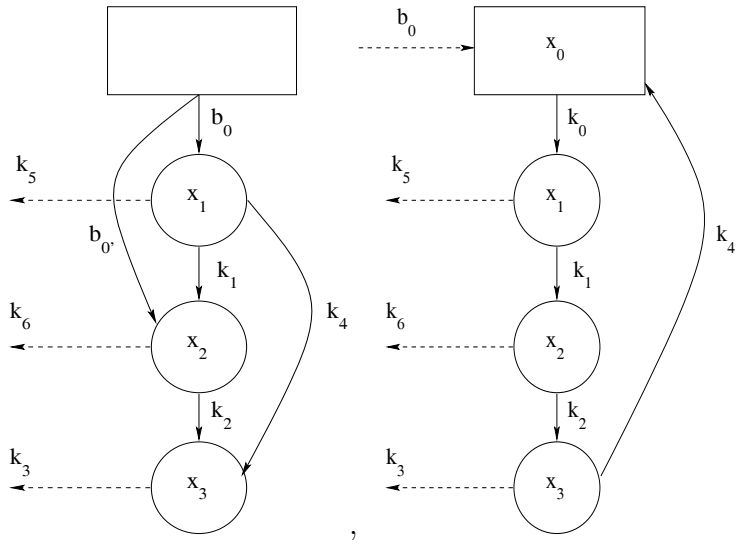


Figure 3: Current ODE models of lipid metabolism which rely on knowing *in vivo* rates of lipoprotein production by the liver.

trials where the participants' lipoprotein content are measured. Compartmental models have been employed to estimate rate constants for the production and catabolism of these lipoprotein species using measurements of changes in lipoprotein concentrations over time. The Study Group participants sought to obtain parameter estimates from an example of this work, Pont et al. (2002), but this proved difficult without access to the raw data set and a detailed understanding of how and why the data had been normalised.

This work, however, highlighted the value of constructing dynamical compartmental or ordinary differential equation (ODE) models in order to possibly obtain dynamic rate parameters from future experimental data. In doing so it became clear that many models, such as that shown in Figure 3, rely on being able to determine the rate of production of the VLDL lipoproteins by the liver from human trial data. The Study Group participants postulated that it may be more appropriate, if possible, to be able to conduct experiments which did not necessarily rely on knowing the rate of production of VLDL_1, thus removing the need for human participant data. The remaining production rates of VLDL_2 could possibly then be determined using *in vitro* experiments. This would then result in the revision of the compartmental models as shown in Figure 4.

The governing equations for the revised model result in a linear system of ordinary differential equations (ODE). Applying a linear systems approach to the problem these can be written as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{u} \tag{1}$$

$$\mathbf{y} = \mathbf{B}\mathbf{x} \tag{2}$$

where $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, $\mathbf{A}$ is the coefficient matrix and $\mathbf{u}$ represents the source terms of each equation.

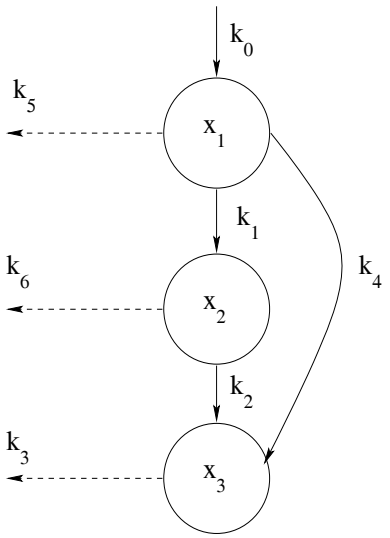


Figure 4: A revised form of the ODE compartmental model shown in Figure 3. This compartmentalisation removes the need for *in vivo* data, and thus the rate constants can be inferred from *in vitro* data.

Using such an approach experimental data will then allow us to estimate $\mathbf{A}$, the distributions for the values of the rate constants k_{ij}.

2.2 Experimental Insight

Lipoprotein subclasses have been traditionally classified in terms of physicochemical properties (e.g. electrophoretic mobility and density after separation by ultracentrifugation). As a lipoprotein is converted from one subclass to another the main changes that occur are in: (1) triglyceride and cholesterol content; and (2) apolipoprotein complement. It is this first difference which leads us to consider the formation of each lipoprotein subclass within a predefined triglyceride-cholesterol (triglyceride-C) particle number density space as demonstrated in Figure 5. We assume for the time being that the changes in triglyceride-C will be large enough and sufficient to discriminate between lipoprotein subclasses without the need for consideration of changes in apolipoprotein content.

The total volume V of a lipoprotein molecule can thus be defined as

$$V = N_T V_T + N_C V_C, \tag{3}$$

where N_T is the number of triglyceride particles per lipoprotein molecule, N_C the number of cholesterol particles and V_T and V_C represent the respective volume of each triglyceride and cholesterol particle.

Essentially it is the rate of change of lipoprotein molecule density which experimentalists measure and from this data they wish to obtain an estimate of change in triglyceride

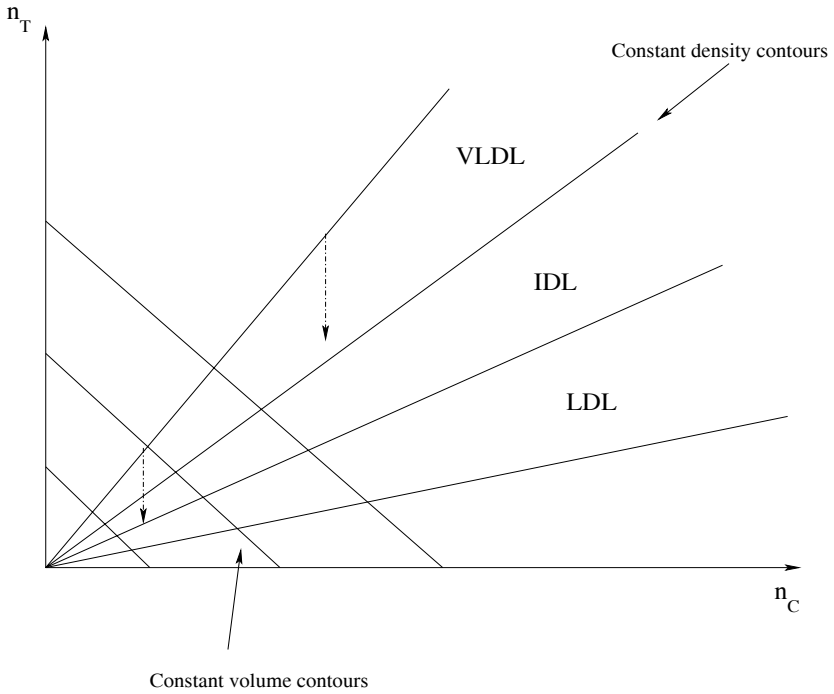


Figure 5: The formation of each lipoprotein molecule within triglyceride-cholesterol space.

particle activity. We can now formulate relationships between the state variables and experimentally measurable parameters in order to obtain relationships between these two quantities.

By considering Figure 5 we can define the ratio of N_T and N_C per lipoprotein to be given by

$$\tan\theta = \frac{N_T}{N_C}, \tag{4}$$

in essence defining a 'pseudo-polar' co-ordinate relationship.

The changes in lipoprotein density and activity can be found by differentiating equations (3) and (4) with respect to time to obtain

$$\begin{pmatrix} \frac{dV}{dt} \\ \frac{d\theta}{dt} \end{pmatrix} = \begin{pmatrix} V_T & V_C \\ \cos^2\theta & \frac{-\sin^2\theta}{n_T} \end{pmatrix} \begin{pmatrix} \frac{dN_T}{dt} \\ \frac{dN_C}{dt} \end{pmatrix}. \tag{5}$$

Equations (3) and (4) also allow us to define the following relationship for the density of a lipoprotein molecule in terms of the mass and volume of each triglyceride and cholesterol particle

$$\rho = \frac{m}{V} = \frac{m_T + m_C}{N_T V_T + N_C V_C} = \frac{m_T \tan\theta + m_C}{V_T \tan\theta + V_C}. \tag{6}$$

Differentiating this expression with respect to time reveals how the density of the lipoprotein molecule varies

$$\frac{d\rho}{dt} = \frac{(V_C m_T - V_T m_C)}{(V_C + \cos\theta V_T)^2} \frac{dN_T}{dt}. \tag{7}$$

Thus the difference in change in lipoprotein density versus triglyceride content differs by a pre-factor, dependent upon the mass and volume of the cholesterol and triglyceride particles.

We note that any model formulated in (N_T, N_C) space will allow us to discern between both the change in density of the lipoprotein molecule as well as the activity of the addition and removal of triglyceride and cholesterol particles. Importantly molecules which move along positive contours in (N_T, N_C) space will maintain a constant density, whilst particles moving along negative contours will maintain a constant volume as shown in Figure 5.

2.3 A PDE model of Lipid Metabolism

The previous section leads us to consider a description of lipoproteins which incorporates both the dynamic details of their transition between each of their subclasses, and the effect that changes in triglyceride and cholesterol content has on this transition.

We develop a model formulation which can account for lipoprotein particles containing arbitary amounts of triglyceride N_T and cholesterol N_C, these being treated as continuous variables. Thus we introduce a density function $\rho(N_T, N_C, t)$ to describe at each time the number density of particles having composition (N_T, N_C); the model could be generalised to higher-dimensional composition spaces, thought this would compound the computational difficulties. The total volume of particles is governed by equation (3). The rate of change in triglyceride and cholesterol density is governed by

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J} + f(N_T, N_C), \tag{8}$$

where $\mathbf{J}$ represents the flux of cholesterol or triglyceride molecules onto or off the lipoproteins, $f(N_T, N_C)$ is a source-sink term and

$$\nabla = \frac{\partial}{\partial N_T}\mathbf{i} + \frac{\partial}{\partial N_C}\mathbf{j}. \tag{9}$$

The flux of triglyceride and cholesterol onto the lipoprotein particles is governed by

$$\mathbf{J} = \mathbf{u}\rho(N_T, N_C), \tag{10}$$

where $\mathbf{u}$ represents the rate at which triglyceride or cholesterol is added or removed. In the case of a healthy individual this process is purely governed by the activity of LPL, such that $\mathbf{u} = k_L L(\mathbf{x}, t)$, where $L(\mathbf{x}, t)$ represents the density of LPL, k_L the rate of triglyceride transfer and $\mathbf{x} = (N_T, N_C)$. In the work which follows we assume $L(\mathbf{x}, t) = L$ is constant.

However in the case of an obese person this activity is altered by the down regulation of LPL and the up-regulation of CETP and HDL. CETP works by transferring a cholesterol or triglyceride molecule onto each lipoprotein particle in exchange for an opposing type of particle, i.e. cholesterol transfer from CETP to the lipoprotein results in the transfer

of a triglyceride molecule from the lipoprotein to CETP and vice-versa. This activity can be expressed by

$$\mathbf{u} = -k_L L\mathbf{i} + \alpha P_C(\mathbf{i}-\mathbf{j}) + \beta P_T(\mathbf{i}-\mathbf{j}) - P_F(k_{FT}\mathbf{i} + k_{FC}\mathbf{j}), \tag{11}$$

where P_C, P_T and P_F respectively represent the concentration of CETP molecules with bound cholesterol, bound triglyceride and CETP molecules which are free of either molecule and α, β and k_{FL} are the respective transfer rates. We note that this formulation assumes equal transfer rates of triglyceride and cholesterol particles onto and off the respective CETP molecules bound with either particle type, but these rates can differ in the free CETP molecule case.

The changes in P_T, P_C and P_F concentrations are governed by

$$\frac{\partial P_T}{\partial t} = \int_0^{N_C^m}\int_0^{N_T^m} (\alpha P_C - \beta P_T + k_{FT}P_F)\rho dN_T dN_C \tag{12}$$

$$\frac{\partial P_C}{\partial t} = \int_0^{N_C^m}\int_0^{N_T^m} (-\alpha P_C + \beta P_T + k_{FC}P_F)\rho dN_T dN_C \tag{13}$$

$$\frac{\partial P_F}{\partial t} = \int_0^{N_C^m}\int_0^{N_T^m} -(k_{FT}+k_{FC})P_F\rho dN_T dN_C, \tag{14}$$

where N_C^m and N_T^m represent the maximum cholesterol and triglyceride content of the lipoprotein particles. This leads us to express the total CETP concentration P as

$$P = P_C + P_T + P_F. \tag{15}$$

For the source-sink term, we note that the liver both produces the initial lipoproteins and removes them, such that

$$f(N_T, N_C) = \tilde{R}(N_T, N_C) - R(N_T, N_C)\rho, \tag{16}$$

where the rate of removal is assumed to depend upon the particle density.

We now have a full system of equations to describe a distribution of lipoprotein molecules, accounting for the change in total cholesterol and triglyceride content, for both a healthy and obese individual. Substituting equations (11) and (16) into equation (8) gives the equation governing the change in lipoprotein density for an obese individual

$$\frac{\partial \rho}{\partial t} - \frac{\partial(k_L L\rho)}{\partial N_T} + P_C\left(\frac{\partial(\alpha\rho)}{\partial N_T} - \frac{\partial(\alpha\rho)}{\partial N_C}\right) + P_T\left(\frac{\partial(\beta\rho)}{\partial N_T} - \frac{\partial(\beta\rho)}{\partial N_C}\right) - P_F\left(\frac{\partial(k_{FT}\rho)}{\partial N_T} + \frac{\partial(k_{FC}\rho)}{\partial N_C}\right) = \tilde{R}(N_T, N_C) - R(N_T, N_C)\rho, \tag{17}$$

where the LPL activity is assumed to be constant and P_C, P_T and P_F are governed by equations (12)-(14). In the case of a healthy individual $\alpha = 0 = \beta = k_{FL}$.

In each case we assume the initial lipoprotein density is given by

$$\rho(0, t) = 0 \tag{18}$$

and the initial concentration of each CETP type molecule is

$$P_C(0) = P_{C0}, \qquad P_T(0) = P_{T0} \quad \text{and} \quad P_F(0) = P_{F0}. \tag{19}$$

2.3.1 Parameter Values

Our formulated model allows us to discern between the dynamic evolution of lipoprotein molecules through their various states, i.e. $VLDL_1$, $VLDL_2$, IDL and LDL, and the effect that an increase in triglyceride and/or cholesterol activity has on this transition. However, data for our model, as mentioned in Section 2.1, was not available at the time of the Study Group meeting. Hence it was necessary to make a number of assumptions regarding parameter values based upon current understanding of the metabolic network.

We note that there is an elevated rate of hepatic VLDL production and that hepatic LDL uptake is decreased in obese compared to healthy subjects. In the case of VLDL production there is a 60% increase in production in the obese case. For simplicity we take $\tilde{R} = 1$ for the healthy state and thus $\tilde{R} = 1.6$ for the obese case. In the obese case LDL uptake is decreased by 64% and we assume in the case of a healthy person that $R = 20$ and thus $R = 11.2$ for the obese case. We do not know the details of lipolysis kinetics, but in principle k_L could be a function of N_T and N_C; here we assume k_L is a constant and take $k_L = 0.5$ for a healthy person and $k_L = 0.4$ for an obese person.

We note that the healthy person choice of $k_L = 0.5$ gives us the required model behaviour, i.e. the lipoprotein molecule is re-absorbed by the liver. We then assume that LPL function is only slightly reduced for an obese person. For simplicity we take $L = 1.0$.

In the case of CETP activity we assume that the production rate of CETP molecules with bound triglyceride and cholesterol particles is equivalent ($\alpha = \beta = 1.0$). In addition we assume that the rate at which free CETP molecules are bound with either a triglyceride or a cholesterol particle are the same ($k_{FT} = k_{FC} = 1.0$). Finally we assume that initial concentration of unbound CETP molecules is zero and that of CETP molecules bound with either triglyceride or cholesterol is the same ($P_{FO} = 0.8$, $P_{T0} = 0.1 = P_{C0}$).

2.3.2 Solution Method

Our governing equations constitute a hyperbolic equation (equation (17)) and three integral equations (equations (12)-(14)). Equation (17) is solved using the weighted average flux method (WAF) as detailed in Toro (1999). We assume that the liver activity is governed by

$$\tilde{R}(N_T, N_C) = \begin{cases} 1, & N_T^3 \leq N_T \leq N_T^m, N_C^3 \leq N_C \leq N_C^m \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

and

$$R(N_T, N_C) = \begin{cases} 1, & N_T^1 \leq N_T \leq N_T^2, N_C^1 \leq N_C \leq N_C^2 \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

as demonstrated in Figure 6 for the 1-D case.

2.3.3 1-D Model Solutions

To begin the analysis of our model, we consider the case of a distribution of lipoprotein molecules with constant cholesterol particle density. Solutions to the model in N_T space

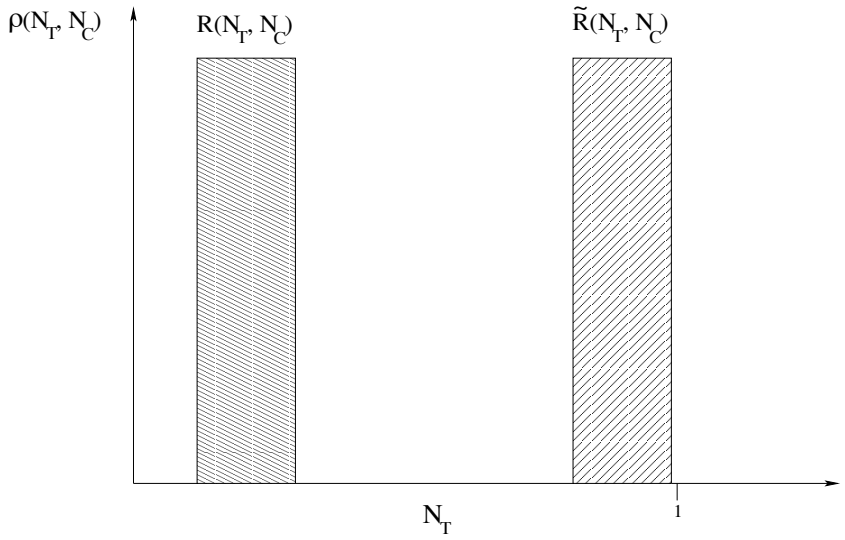


Figure 6: The activity of the liver as a function of triglyceride activity.

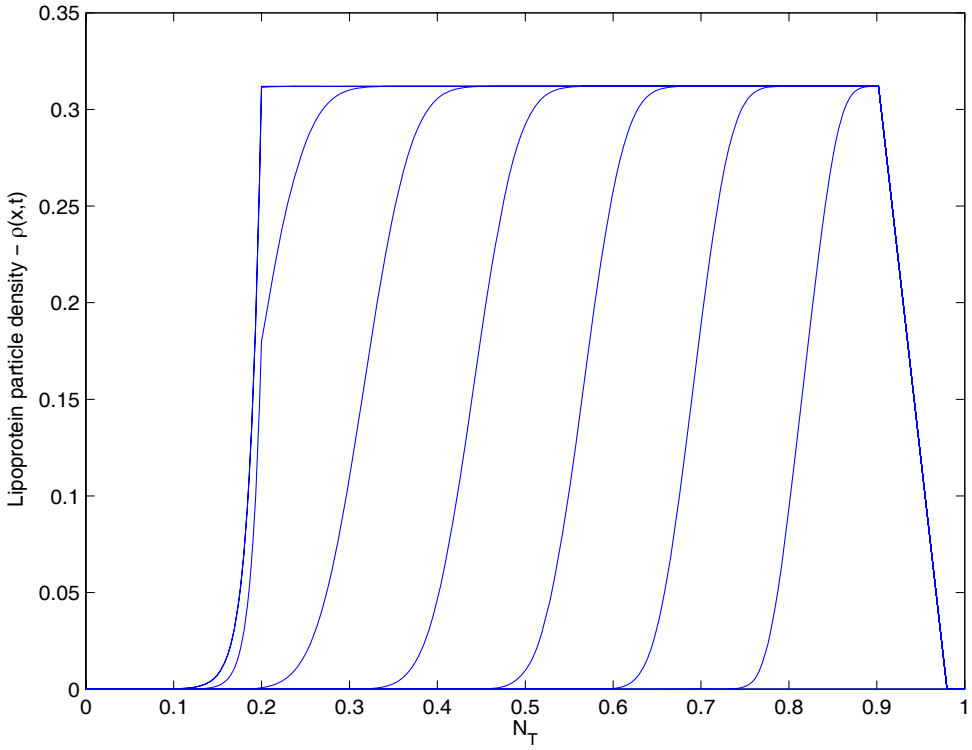


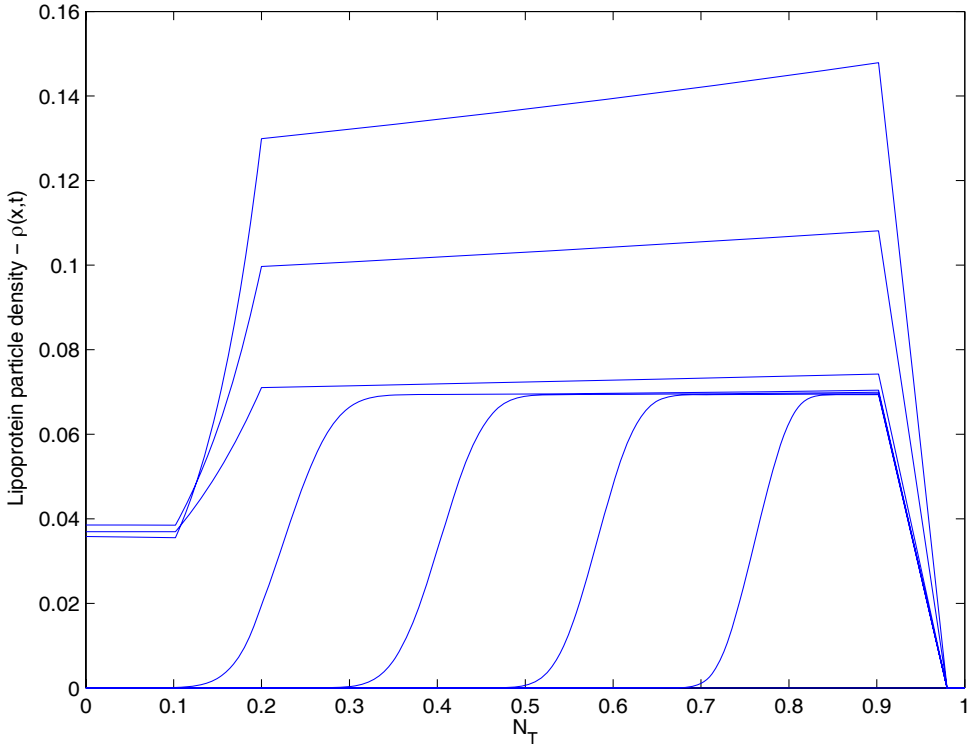
Figure 7: The change in triglyceride density of a lipoprotein molecule for a healthy person. In this case $\alpha = \beta = k_{FT} = k_{FC} = 0$, $\tilde{R} = 1.0$ and $R = 20.0$ and $k_L = 0.5$.

will then allow us to see what effect each mechanism, in particular LPL and CETP, has on the increase in triglyceride content for healthy and obese individuals.

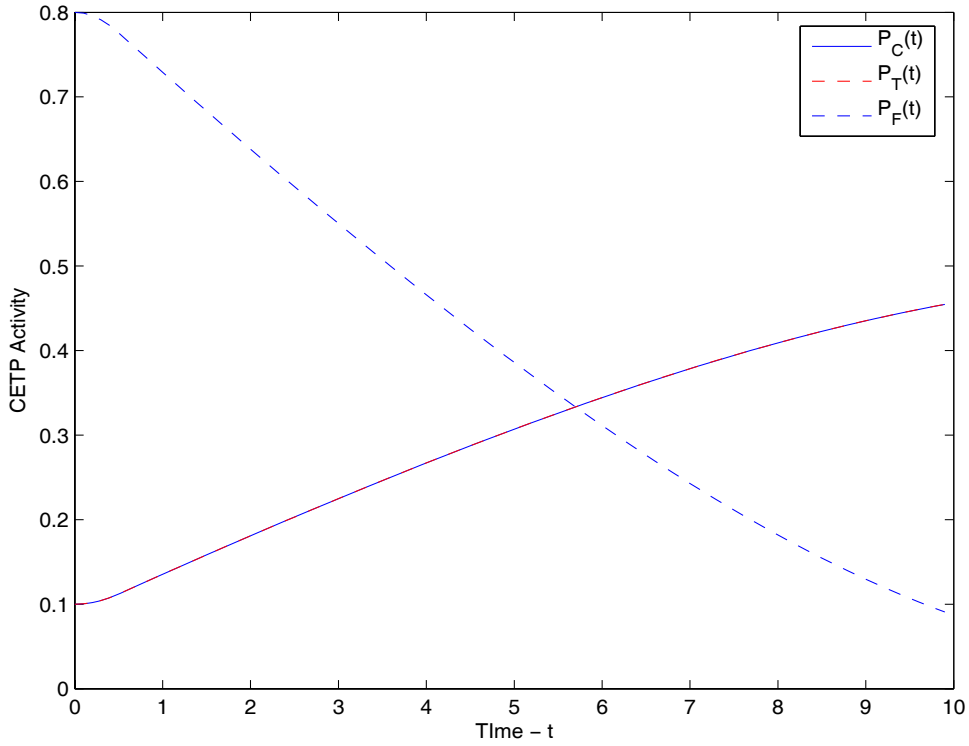
Figure 7 shows the time evolution of the density distribution for a distribution of lipoprotein molecules for a healthy individual. We note that the initial lipoproteins ($VLDL_1$) are produced by the liver and due to the activity of LPL, are slowly advected towards the origin, thereby reducing their triglyceride density. As the triglyceride density is reduced the resulting LDL molecules can be metabolised by the liver, thus removing the lipoprotein particles as shown.

In the case of an obese person, the up-regulation of CETP activity and down-regulation of LPL plays an important role in affecting the lipoprotein density as demonstrated in Figure 8(a). The lipoprotein particles are again produced by the liver, however, with decreased liver uptake and increased CETP activity as shown in Figure 8(b), not all of the resulting lipoproteins are absorbed by the liver thus leading to a non-zero lipoprotein density as $N_T \to 0$. Although LPL is down-regulated, the increased advection rate of the lipoproteins through N_T space as a result of increased CETP, leads to the observed increase in the triglyceride density of the lipoproteins. We note from Figure 8(b) that the activity of $P_C(t)$ and $P_T(t)$ is indistinguishable given $\alpha = \beta = 1$. The concentration of CETP molecules free of both triglyceride and cholesterol falls in time as expected and that of CETP molecules bound with either triglyceride or cholesterol rises.

The results of Figure 8(a) show that the model does not reach a steady-state in the specified time, indeed further simulations have shown that the time to steady-state is exceedingly long given the parameter value estimates. We note that the model solutions do show an increase in particle density which is unexpected and we believe this to be a result of only considering model solutions in 1-D. As a result further model analysis is required in 2-D before any further conclusions about the behaviour of our model can be drawn.



(a)



(b)

Figure 8: The change in triglyceride density of a lipoprotein molecule for an obese person (a) and the corresponding change in CETP particle activity (b). Here liver uptake of LDL ($R = 11.2$) and LPL activity ($k_L = 0.4$) are both reduced and CETP activity is up-regulated ($\alpha = 1.0 = \beta$, $k_{FC} = 1.0$ and $k_{FT} = 1.0$) with initial distributions of each CETP molecule given by $P_{C0} = 0.1$, $P_{T0} = 0.1$ and $P_{F0} = 0.8$.

3 Comparative Genomics

This section describes mathematical approaches that maybe useful in 'comparative genomics'. The notion of comparative genomics is described in section 3.2, along with some of the challenges. Prior to this, however, section 3.1 describes a representation of a metabolic network in terms of the network's *stoichiometric matrix*. This representation is helpful in providing a simple framework in which the application of mathematical methods can be rigorously described. While it may not be the most appropriate representation for certain types of problem, it should be straightforward to map the principles outlined here onto other approaches. Section 3.3 considers some of the issues associated with formulating a stochastic representation of a metabolic network, which is an important aspect of data-based inference when the network structure is uncertain. Section 3.4 considers the design question, concerning which aspects of a partially-specified network to measure.

3.1 The stoichiometric matrix

The following is a brief summary taken in part from Bernhard Palsson's notes, available at `http://gcrg.ucsd.edu/classes/4_slides_Smatrix.pdf`.

The stoichiometric matrix S is a compact way of representing a metabolic network in terms of the metabolites (rows of S) and the reactions and transport processes (columns of S). For the j^{th} reaction, described by the column $S_{(j)}$, a positive value indicates a metabolite that is formed, and a negative one a metabolite that is consumed; reversible reactions are entered twice, once for each direction. The actual numbers are the stoichiometric coefficients, representing the numbers of molecules involved. One enzyme can catalyse a number of reactions and, conversely, the same reaction can be catalysed by more than one enzyme. Thus there is not a 1-to-1 mapping between enzymes and columns. However, enzymes can be associated with columns so that, for example, the elimination of an enzyme (e.g. the removal of a gene from the genome) can be accounted for by the removal of one or more columns from S. A well-formed S will satisfy the property that the elements must balance during a chemical reaction, e.g. the same number of molecules of hydrogen must be present both before and after the reaction. It should also be balanced for charge and moiety.

The flux through the reactions in the network at time t is denoted by $v(t) \geq \mathbf{0}$; this inequality holds because we have entered the reversible reactions twice. The dynamic mass balance equation states that

$$\dot{x}(t) = S\, v(t) \tag{22}$$

where $x(t)$ is metabolite concentration at time t, and the dot denotes the time derivative d/dt. There are a number of ways in which S may be decomposed in order that we can better understand the evolution of $x(t)$ and $v(t)$. The simplest analysis concerns the steady state. At steady state we must have $\dot{x}(t) = \mathbf{0}$, so that $\bar{v}$ must be a point in the

null space of S, where the overbar denotes the steady state flux. The null-space of S can be described as a cone

$$\mathcal{C} \triangleq \left\{ \bar{v} : \bar{v} = \sum_{i=1}^{k} \alpha_i \, p_i, \quad \alpha_i, \ldots, \alpha_k \geq 0 \right\} \tag{23}$$

i.e. a non-negative linear combination of the k *extreme pathways* $p_1, \ldots, p_k$. A complex metabolic network can have a large number of extreme pathways, and a procedure for analysing these extreme pathways further is described in Price et al. (2003). Extreme pathways are very similar to *elementary flux modes*, as discussed in Schuster et al. (2000, p. 326): "... a minimal set of enzymes [mapping the components of p_i via the columns of S] that could operate at steady state". By 'minimal' is meant that where p_i is the only functioning pathway, inhibition of an enzyme involved in p_i would cause the steady state flux through p_i to be zero. Schuster et al. (2000, p. 330) note a number of uses for such an analysis, including drug target identification and the effect of gene deletion.

For our purposes below, we will assume that there exists a well-defined set of operations on a given stoichiometric matrix from which we can draw a series of conclusions, $h(S)$. Following the above analysis, this might be conclusion such as "if we inhibit enzyme x then the steady state flux through part A of the network is $\bar{v}_A$, but if we inhibit x' then the flux is $\bar{v}'_A$". In this case we might have $h(S) = (\bar{v}_A, \bar{v}'_A)$ or, if we are interested in the relative performance of the two enzymes, $h(S) = \bar{v}_A / \bar{v}'_A$.

3.2 Problems in comparative genomics

The problem posed by Unilever was how we should proceed when the stoichiometric matrix is not completely known. In particular, (i) How should we use data from flux measurements to 'fill in' the missing parts of the matrix? and (ii) What measurements should we make in order to do this 'filling in' most effectively? In statistics this type of question comes under the general heading of the design and analysis of *computer experiments*; Sachs et al. (1989), Koehler and Owen (1996) and Santner et al. (2003) provide reviews of this area, which is the subject of much on-going research.

3.2.1 Calibrated prediction

In a computer experiment we seek a probabilistic assessment of some quantity of interest subject to various sources of uncertainty. In the simplest case, our quantity of interest is $h^* \triangleq h(S^*)$ and our uncertainty concerns S^*, the 'correct' stoichiometric matrix. Note that the concept of a 'correct' value for S is not uncontroversial, and an important aspect of computer experiments is accounting for 'model inadequacy' (see, e.g., Kennedy and O'Hagan, 2001; Craig et al., 2001; Goldstein and Rougier, 2005a,b). For simplicity, however, we will ignore the issue of model inadequacy in what follows. Rougier (2004) provides further detail on the following calculations, at a fairly non-technical level, although in the context of climate prediction.

Our objective in this case is to compute, approximately, the distribution function

$$\begin{aligned} F_{h^*}(y) &\triangleq \Pr\big(h^* \leq y\big) \\ &= \int_S \mathbf{1}\big(g(S) \leq y\big)\ dF_{S^*}(S) \end{aligned} \tag{24}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function and F_{S^*} is the distribution function of S^*, i.e. describes our uncertainty about S^*. In other words, for any given value of y we add up all of the probability in F_{S^*} which corresponds to a value of $g(S)$ which is less than or equal to y. This type of calculation is often referred to as *uncertainty analysis* (Haylock and O'Hagan, 1996), and F_{h^*} is termed the *prior predictive distribution for* h^*.

In some situations we are able to perform an experiment to measure some of the fluxes in a metabolic network, which we can duplicate using some function of the stoichiometric matrix, $g(S)$, say. In this case our measured data has the form

$$z = g(S^*) + e \tag{25}$$

where e denotes (uncertain) measurement errors. For simplicity we assume that e has a gaussian distribution with zero mean vector and known variance matrix Σ^e. Our objective is now to compute the distribution function of h^* conditional upon the event that $z = \tilde{z}$, where $\tilde{z}$ is the observed value of z. Using Bayes's theorem,

$$\begin{aligned} F_{h^*|z}(y) &\triangleq \Pr\big(h^* \leq y \mid z = \tilde{z}\big) \\ &= c \int_S \mathbf{1}\big(g(S) \leq y\big)\ \mathrm{Lik}_{\tilde{z}}(S)\ dF_{S^*}(S) \end{aligned} \tag{26a}$$

where '|' denotes 'conditional upon', $c \triangleq \Pr(z = \tilde{z})^{-1}$ and $\mathrm{Lik}_{\tilde{z}}(\cdot)$ is the *likelihood function*,

$$\begin{aligned} \mathrm{Lik}_{\tilde{z}}(S) &\triangleq \Pr(z = \tilde{z} \mid S^* = S) \\ &= \phi\big(\tilde{z} - g(S)\ ;\ \mathbf{0}, \Sigma^e\big) \end{aligned} \tag{26b}$$

where $\phi(\cdot\ ;\ \cdot, \cdot)$ denotes a gaussian Probability Density Function (PDF) with given mean and variance; here the likelihood function is the PDF for the measurement error. This type of calculation is often referred to as *calibrated prediction* (Goldstein and Rougier, 2005a), and $F_{h^*|z}$ is referred to as the *posterior predictive distribution*.

Comparing (24) and (26), we can see that introducing the data causes us to weight the integrand according to how well each candidate value for S is able to replicate the data $\tilde{z}$, subject to acceptable amounts of measurement error. In the extreme case where Σ^e becomes large the likelihood becomes flat, and (26) becomes (24). This is the case where the data are measured so imprecisely as to be useless in terms of selecting the correct value for S. The value of Σ^e must be carefully specified, because it controls the curvature of the likelihood function, and thus controls the degree to which observations on z can tell us about S^*.

3.2.2 Numerical approximations

The simplest approach to computing (26) is to use Monte Carlo integration, for which

$$\hat{F}^{(n)}_{h^*|z}(y) \triangleq \sum_{i=1}^{n} w_i \, \mathbf{1}\big(h(S_i) \leq y\big) \tag{27}$$

where $w_i \propto \mathrm{Lik}_{\tilde{z}}(S_i)$ and $\sum_{i=1}^{n} w_i = 1$, and $S_1, \ldots, S_n$ are sampled independently from F_{S^*}. In other words, for any given value of y we sample n candidate values for S^* from F_{S^*}, and sum the (normalised) weights of those candidates for which $h(S_i) \leq y$. By the Strong Law of Large Numbers (SLLN), we have

$$\lim_{n\to\infty} \hat{F}^{(n)}_{h^*|z}(y) = F_{h^*|z}(y). \tag{28}$$

In the case of uncertainty analysis, where we do not have any data, the same approach is used, only all of the weights are set equal to n^{-1}.

In practice, we can improve on our estimate of $F_{h^*|z}$, because there are usually better alternatives to simple Monte Carlo integration. In particular, adopting *importance sampling* with *variance reduction techniques* can have a dramatic effect on the quality of our estimate (i.e. its rate of convergence to the correct value as a function of n). These issues are covered in Ripley (1987), and discussed in more detail in Evans and Swartz (2000, ch. 6).

The posterior predictive distribution calculation given in (26) and estimated in (27) performs what is sometimes known as *model averaging*. We do not know the correct value S^*, but we average $h(\cdot)$ over a number of candidate values, $S_1, \ldots, S_n$, according to the posterior PDF for S^*

$$\Pr(S^* = S_i \mid z = \tilde{z}) \propto \mathrm{Lik}_{\tilde{z}}(S_i) \; dF_{S^*}(S_i) \tag{29}$$

(informally). An alternative approach is to identify the single best candidate value for S^*, and treat this as though it was the correct value, which would give us a point prediction for h^* but no measure of uncertainty. This is obviously a good approach when the data z are highly informative about S^* and n is large, because in this case we would expect almost all of the probability to be concentrated onto one candidate value. There are several ways of identifying a single best value. Statisticians usually choose the posterior mean, for which the calculation involves an integration very much like (27). An alternative is to choose the Maximum A Posteriori (MAP) estimate, which can be found by optimisation, and may in some circumstances be much cheaper to compute. For the MAP estimate we need to find

$$S^*_{\mathrm{MAP}} \triangleq \mathrm{argmax}_S \, \Pr(S^* = S \mid z = \tilde{z}) \, . \tag{30}$$

To do this calculation effectively would require a discrete-state-space optimisation algorithm such as TABU search (see, e.g., Battiti, 1996).

There is a catch, however, with using a point estimate such as S^*_{MAP}. Metabolic networks are unlikely to give rise to a posterior PDF for S^* which is highly concentrated, because by

the fact that they have evolved, they display a high level of robustness. This is discussed in Barabási and Oltvai (2004). 'Robustness' here means that there are many ways for the network to function, and thus many ways in which the same data z might have arisen: effectively $g(\cdot)$ is a many-to-one function. This is very bad news for calibration, i.e. learning about S^*, but not necessarily bad news for prediction, i.e. learning about h^*, since it may be that all the possible candidates which might give rise to $\tilde{z}$ give roughly the same prediction for $h(\cdot)$. However, this is not something we can assert, and consequently it is prudent to proceed on the basis that we will not be able to make a reliable identification of S^*, and consequently when we think about h^* we should favour a model-averaging approach.

3.3 Priors on models

One point that has not been addressed is how we formulate F_{S^*}, the distribution function on candidate stoichiometric matrices. The model-averaging and optimisation approaches described above make two different demands on how we formulate F_{S^*}. For model-averaging we need to be able to sample S from F_{S^*}, while for, say, the MAP estimate $\hat{h} \triangleq h(\hat{S}_{\mathrm{MAP}})$ we need to be able to compute the value $dF_{S^*}(S)$. For an optimisation approach like TABU search we also need to define some kind of neighbourhood structure over candidate values for S^*, so that we can imagine taking incremental steps through S-space.

This is an area where expert judgement is crucial. At a preliminary assessment, though, it seems as though sampling candidate values for S^* will be the easier task. This is because there seems to be a reasonable understanding of how complex metabolic networks arise from simpler ones, and this includes a stochastic description of processes that give rise to characteristic features of metabolic networks such as their scale-free and hierarchical topology (see, e.g., Barabási and Oltvai, 2004, Box 2, p. 105). Thus it ought to be possible to start with a known stoichiometric matrix S_1, which represents the bits of the network that are to be taken as known, and to propose a method for 'growing' it in a stochastic manner to come up with a candidate for the complete stoichiometric matrix S^*, which has a larger collection of metabolites and reactions.

3.4 Design issues

Finally, we consider briefly the design question: which fluxes ought to be measured? We imagine a number of candidate measurements, z', z'' and so on. What makes z' a better set of measurements than z''? The usefulness of any particular measurement should be directly related to the inference that is to be drawn from the experiment. In our case, the inference is summarised in the distribution function of h^*, and it is natural in this case to propose that our yardstick for valuing a measurement is in terms of how much that measurement allows us to reduce our uncertainty about h^*, relative to where we are at the moment, using z.

If we actually had the actual measurements $z' = \tilde{z}'$ and $z'' = \tilde{z}''$ then we could compare the two quantities

$$v(\tilde{z}') \quad \text{and} \quad v(\tilde{z}'') \tag{31a}$$

where

$$v(\tilde{z}') \triangleq \mathrm{Var}\big(h^* \mid (z = \tilde{z}, z' = \tilde{z}')\big) \tag{31b}$$

and likewise for $v(\tilde{z}'')$. That is, we could augment our data $\tilde{z}$ with the additional measurements one at a time, compute the predictive uncertainty in terms of the posterior predictive variance of h^*, and choose whichever of $\tilde{z}'$ and $\tilde{z}''$ gives rise to the largest drop in uncertainty, i.e. choose to measure z' rather than z'' if and only if $v(\tilde{z}') < v(\tilde{z}'')$.

To get around the problem of not actually having $\tilde{z}'$ and $\tilde{z}''$, we use *pseudo-data*. That is, we generate 'fake' data on the basis of a reasonable choice for S^*, using the function $g(\cdot)$. This type approach has been suggested in the context of computer experiments by Craig et al. (2001, section 8). To a limited extent the result depends on the values of the pseudo-data. But in fact it often turns out that the actual value $\tilde{z}'$ is less important that the fact of measuring z'. A good example where this is completely true is the multivariate gaussian distribution. If the collection $(h^*, z', z'' \mid z = \tilde{z})$ is gaussian, then $v(\tilde{z}')$ does not depend on the actual value $\tilde{z}'$ at all, but only on the joint variance structure of the marginal distribution $(h^*, z') \mid z = \tilde{z}$; likewise for $v(\tilde{z}'')$. Therefore an objective function for choosing between measuring z' and z'' that is based on the predictive variance can often be quite insensitive to the precise values of the pseudo-data.

4 Conclusions and Future Work

The focus of the work presented here has been two-fold. Firstly the issue of normal and dysregulated lipoprotein production and metabolism has been considered in a number of contexts. Initial compartmentalisation of the various lipoproteins particles (VLDL, IDL, and LDL) led us to consider the application of ODE type models to the problem. However, during the course of the Study Group it became apparent that the evolution of the lipoprotein particles through each of their states reflects a change in triglyceride and cholesterol density of the particle. This has led to a PDE continuum description of a distribution of lipoprotein particles in triglyceride and cholesterol space.

Initial 1-D results from the model have shown that it produces the qualitatively observed behaviour of particle production and uptake by the liver for a healthy person and difficulties in particle uptake in the obese case. However, further model analysis in 2-D is required to fully understand and capture the model behaviour.

The second problem related to comparative genomics has highlighted a number of statistical based methods for dealing with the issue of comparing unknown networks to those which are better understood. The importance of the stoichiometric matrix has been highlighted along with priors (in the Bayesian case) as well as the design of experiments for measuring parameter values.

In order to progress the modelling presented in this report the following is a suggested list of future work.

- Obtain more detailed parameter estimates in order to further populate the PDE model.
- Solve the PDE based model in 2-D in order to account for changes in both the triglyceride and cholesterol particle density for a given lipoprotein particle.
- Consider the affect of upregulated HDL activity.

References

Barabási, A.-L., Oltvai, Z. N., February 2004. Network biology: Understanding the cell's functional organisation. Nature Reviews 5, 101–114.

Battiti, R., 1996. Reactive search: Toward self-tuning heuristics. In: Rayward-Smith, V., Osman, I., Reeves, C., Smith, G. (Eds.), Modern Heuristic Search Methods. John Wiley and Sons Ltd., Ch. 4, pp. 61–83.

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001. Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association 96, 717–729.

Evans, M., Swartz, T., 2000. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford: Oxford University Press.

Goldstein, M., Rougier, J., 2005a. Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM Journal on Scientific Computing 26 (2), 467–487.

Goldstein, M., Rougier, J., 2005b. Reified Bayesian modelling and inference for physical systems, under review, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf`.

Haylock, R., O'Hagan, A., 1996. On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In: Bernado, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 5. Oxford, UK: Oxford University Press, pp. 629–637.

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B 63, 425–464, with discussion.

Koehler, J., Owen, A., 1996. Computer experiments. In: Ghosh, S., Rao, C. (Eds.), Handbook of Statistics, 13: Design and Analysis of Experiments. North-Holland: Amsterdam, pp. 261–308.

Pont, F., Duvillard, L., Florentin, E., Gambert, P., Vergés, B., 2002. Early kinetic abnormalities of apob-containing lipoproteins in insulin-resistant women with abdominal obesity. Arterioscler. Thromb. Vasc. Biol. 22 (10), 1726–32.

Price, N. D., Reed, J. L., Papin, J. A., Famili, I., Palsson, B. O., February 2003. Analysis of metabolic capabilities using singular value decomposition of extre pathway matrices. Biophysical Journal 84, 794–804.

Ripley, B., 1987. Stochastic Simulation. New York: John Wiley & Sons.

Rougier, J., 2004. Prediction of future climate using an ensemble of computer simulator evaluations, available at `http://www.maths.dur.ac.uk/stats/people/jcr/EnsemblesA4.pdf`.

Sachs, J., Welch, W., Mitchell, T., Wynn, H., 1989. Design and analysis of computer experiments. Statistical Science 4 (4), 409–423, with discussion, 423–435.

Santner, T., Williams, B., Notz, W., 2003. The Design and Analysis of Computer Experiments. New York: Springer.

Schuster, S., Fell, D. A., Dandekar, T., March 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology 18, 326–332.

Toro, E., 1999. Riemann Solvers and Numerical Methods for Fluid Dynamics, 2nd Edition. Springer Verlag.