Equation Discovery for Whole-Body Metabolism Modelling

Marvin Meeng[1], Arno Knobbe[1], Arne Koopman[1], Jan Bert van Klinken[2], and Sjoerd A. A. van den Berg[2]

[1] LIACS, Leiden University, the Netherlands, `meeng@liacs.nl`
[2] LUMC, Leiden University, the Netherlands

This paper is concerned with the modelling of whole-body metabolism. The analysis is based on data collected from a range of experiments involving mice in *metabolic cages*, whose food consumption, activity and respiration has been monitored around the clock. Our aim is to model the dependencies between these different variables by means of (differential) equations, as is customary in systems biology. However, compared to the common setting of modelling on the cellular level, where changes in concentrations are mostly instantaneous, in whole-body metabolism we need to take into account the relatively slow process of food digestion. As a result, the effects of eating will only be visible in the activity and body-heat variables with a certain delay. To further complicate the modelling, the digestive delay depends on the different rates of metabolism of carbohydrates and fat. We accommodate for these (varying) delays in digestion, by adding different time-shifted versions of the primary variables to the data, and applying different levels of smoothing. These newly constructed variables can be interpreted as representations of available blood sugars, with different hypothetic rates of digestions. The Lagramge tool [2,3,4] was used to induce ordinary and differential equations that model the enriched data. Lagramge is an equation discovery tool that finds equations of arbitrary (configurable) complexity, and subsequently performs the parameter fitting to the data.

1 The metabolic cage

Our data was gathered at the LUMC in a project concerning the Metabolic Syndrome. In that study, 16 genetically identical mice were divided into two equal-sized groups, one was put on a low (LFD), the other on a high fat diet (HFD). During a 3-day period, various variables were recorded every 7.5 minutes while the mice were in a metabolic cage. Such a cage creates a closed environment in which the amount of oxygen and carbon dioxide can be controlled.

For the experiments below, the following variables are used: VO_2 (oxygen consumption), VCO_2 (carbon dioxide production), RER (respiratory exchange ratio), HEAT (aka. energy expenditure), F (food consumed) and X (total X-activity). The activity was measured using a number of infrared beams in the cage. RER and HEAT are calculated as follows:

$$RER = VCO_2/VO_2$$
$$HEAT = 3.185 \cdot VO_2 + 1.232 \cdot VCO_2$$

Though just a simple ratio, RER is very useful, as it gives direct insight into the energy source an organism digested to fuel its activity. Digestion of pure carbohydrates would result in a RER of 1.00, pure fat in 0.707, and a 50/50 diet would result in a RER of 0.85 [1]. This allows to differentiate between the two diet groups.

2 Equation Discovery with Lagramge

The equation discovery tool Lagramge was used to generate equations that might capture the essential variables involved during the various stages of metabolism, along with their interplay. Lagramge is capable of discovering both ordinary (OE) and differential (DE) equations. To restrict the search space, the structure of equations can be defined through a context free grammar, which also allows domain specific knowledge to be included, in the form of formulas. Such formulas, then need no longer be discovered, but are available to be included into new candidate equations right from the start. Three different grammars were tested, a *Linear*, *Universal* and *Metabolic Cage* (*MC*) grammar. Because of space limitations, we only present results for the *MC* grammar (shown below), which is somewhat inspired by the *Universal* grammar [4], but includes the information of the RER and HEAT equations above.

$$
\begin{aligned}
&\text{E} \rightarrow \text{E} + \text{F} \mid \text{E} - \text{F} \mid \text{E} \cdot \text{F} \mid \text{E} / \text{F} \mid \text{const} \\
&\text{F} \rightarrow \text{RER} \mid \text{HEAT} \mid \text{V} \\
&\text{RER} \rightarrow \text{V} / \text{V} \\
&\text{HEAT} \rightarrow \text{const} \cdot \text{V} + \text{const} \cdot \text{V}
\end{aligned}
$$

3 Experiments

Three variables were chosen as targets for separate experiments: RER, HEAT and X. Both OEs and DEs were sought using an exhaustive search setting of depth 4, as this turned out to be a good trade-off between formula complexity and computation time. Note that depth refers to the number of refinements by means of one of the rules in the grammar.

Data was preprocessed in two ways. First, for all variables but X, the data was modestly smoothed using the standard Gaussian kernel, $\mathcal{G}(\mu, \sigma)$, with $\mu = 0$, and $\sigma = 1$. As time points are relatively far apart, some smoothing was deemed necessary to compensate for boundary effects. X was left out of this procedure as, compared to eg. food digestion, this is the most abrupt process, and any spread in dependencies between X and other variables is already achieved by their smoothing. Furthermore, for the F variable, data was additionally smoothed using $\mathcal{G}(0, 0.5)$ and $\mathcal{G}(0, 1.5)$ to account for any gradual effects the consumption of food may have.

More than any of the other variables, the energetic effects of food consumption depend on time. As a second step, we therefore added four versions of each F variable, to accommodate potential different rates of metabolism. The time
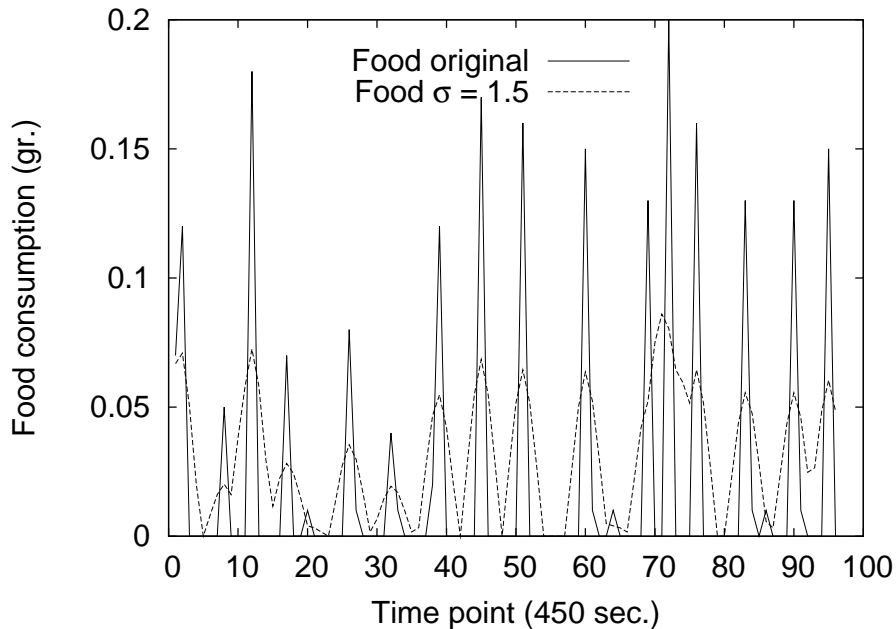
Fig. 1. F Smoothed using $\mathcal{G}(0, 1.5)$

delays were 15, 30, 60 and 90 minutes, which resulted in 15 different F variables in total, five for each kernel version. These were all simultaneously present in the data file, the rationale being that Lagramge might include the important variables from concurrent time scales all in a single equation. Here, the concurrent time scales are related to the various rates and stages of carbohydrate and fat metabolism. Figure 1 shows a smoothing of the F variable for a HFD mouse (first 12 hours are depicted only).

3.1 Results

Table 1 shows some of the best equations found using the MC grammar in an exhaustive search of depth 4. Here best is defined as smallest error. These errors, and the correlation between the measured values and the predicted ones, are shown in table 2, which also shows the number of trees Lagramge evaluated to reach the results. Lagramge has two heuristic functions which it can use to minimize the error in its search for the best equations. All results presented here use the SSE (sum of squared error) function. Note that the errors for X seem very large compared to those of RER and HEAT, but the input domain of RER and HEAT is roughly between 0 and 1, while that of X ranges from 0 to 1,000. For X this yields an SSE range of 0 - 1,000,000, so the largest error 73,088 is about 7%. For OEs and DEs the target is denoted like T_d and T'_d respectively, where d indicates the diet group. For the food variable (F), the

σ subscript denotes the sigma used for the smoothing kernel and M denotes the shift in minutes. A superficial scan of these results shows that a variety of equation syntaxes are used, with linear equations dominating the RER results. Furthermore, most equations involve at least one delayed F variable, with only a single equation being based on the (slightly smoothed) direct F variable. This clearly illustrates the effect that absorbed nutrients have on the mechanisms by which fuel selection is regulated. Also, table 1 shows that the activity X is a major determinant of energy expenditure, as would be expected.

**Table 1.** Best equations found for each setting and diet type.

Target	Equation
RER_{LFD}	$0.826 + F_{\sigma1.5,M30} + 1.646 \cdot F_{\sigma1.5,M15} - 2.094 \cdot F_{\sigma1.5,M60}$
RER_{HFD}	$0.791 + F_{\sigma1.5,M30} + 0.762 \cdot F_{\sigma1.5,M15} - 0.318 \cdot F_{\sigma1.5,M60}$
RER'_{LFD}	$-0.004 \cdot F_{\sigma1.5,M90} + 0.006 \cdot F_{\sigma0.5,M0} + 0.182 \cdot F_{\sigma0.5,M90}$
RER'_{HFD}	$-0.008 \cdot F_{\sigma1.5,M30} + 0.011 \cdot F_{\sigma0.5,M15} + 0.627 \cdot F_{\sigma0.5,M30}$
$HEAT_{LFD}$	$0.441 + 1.219 \cdot 10^{-4} \cdot RER - 0.406 \cdot X$
$HEAT_{HFD}$	$0.376 + F_{\sigma1.5,M30} + 2.146 \cdot 10^{-4} \cdot F_{\sigma1.5,M60} + 0.348 \cdot X$
$HEAT'_{LFD}$	$(-0.024 + F_{\sigma0.5,M15}) \cdot (0.015 \cdot F_{\sigma0.5,M15} - 6.907 \cdot F_{\sigma0.5,M90})$
$HEAT'_{HFD}$	$(0.873 - RER) \cdot (-0.007 \cdot F_{\sigma0.5,M30} + 77.452 \cdot F_{\sigma1.0,M30})$
X_{LFD}	$(-0.221 + VO_2) \cdot (17217.2 \cdot F_{\sigma0.5,M0} - 16822.8 \cdot VO_2)$
X_{HFD}	$(-0.206 + VO_2) \cdot (6075.65 \cdot HEAT - 784.577 \cdot VO_2)$
X'_{LFD}	$-0.011 - F_{\sigma0.5,M15} + F_{\sigma0.5,M0}/HEAT$
X'_{HFD}	$-0.706 \cdot HEAT + 1.400 \cdot F_{\sigma0.5,M90} - 9.481 \cdot VCO_2$

Table 2. Error and correlation of best equations and number of trees evaluated.

Target	Error	Correlation	Trees
RER_{LFD}	$3.833 \cdot 10^{-3}$	0.20	117369
RER_{HFD}	$1.600 \cdot 10^{-3}$	0.52	
RER'_{LFD}	$12.156 \cdot 10^{-3}$	-0.06	139369
RER'_{HFD}	$4.712 \cdot 10^{-3}$	-0.06	
$HEAT_{LFD}$	$2.441 \cdot 10^{-3}$	-0.57	117369
$HEAT_{HFD}$	$2.589 \cdot 10^{-3}$	0.78	
$HEAT'_{LFD}$	$2.683 \cdot 10^{-3}$	0.05	139195
$HEAT'_{HFD}$	$5.693 \cdot 10^{-3}$	-0.19	
X_{LFD}	44657.4	-0.62	163572
X_{HFD}	24848.0	0.84	
X'_{LFD}	72839.1	0.26	190641
X'_{HFD}	73088.9	-0.24	

For all three targets, figure 2 shows the number of occurrences of each version of the $F_{\sigma 0.5}$ variable in the top 1,000 DEs for each diet group. Here we can clearly see a different pattern for the two groups. Compared to the HFL group there are many more occurrences of the non-shifted variable $F_{\sigma 0.5, M0}$ in the LFD group for $HEAT'$ and X', while $F_{\sigma 0.5, M15}$ is much more frequent for the HFD group for RER' and $HEAT'$. This indicates that the energy from high-carb nutrition is available in the blood stream quicker than for the high-fat diet.

Finally, for target X, figure 3 (left) shows an example of one of the found equations (HFD group) compared to the original data, as a function of time. The right figure shows the fit between the actual measurement and its model, for the same equation. The linear correlation between these two functions is $r = 0.84$, table2 shows the correlation of the other equations.
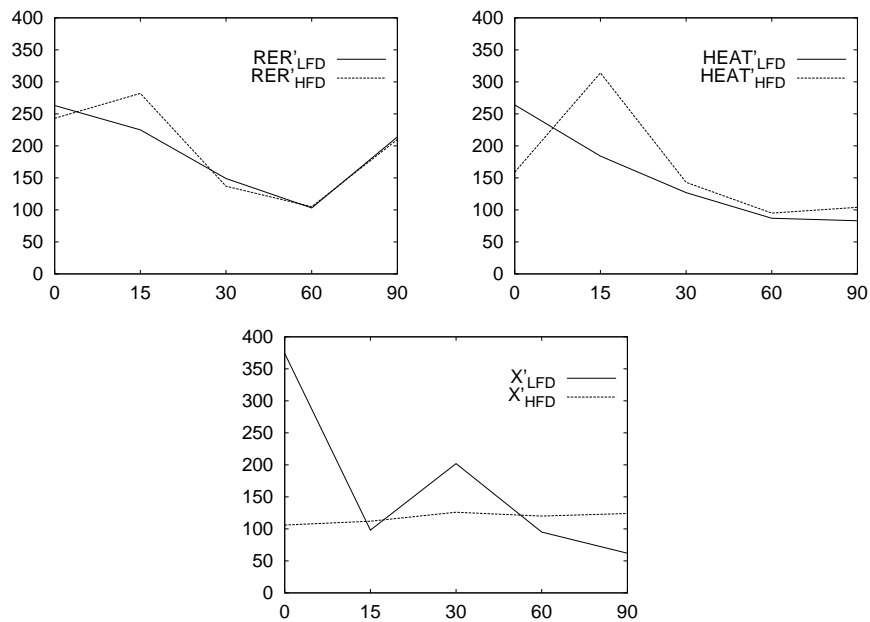
Fig. 2. Histograms of three target variables, each for two diets.

4 Conclusion

The experiments reported in this paper demonstrate that Lagramge can be an important tool for modelling in systems biology. It allows the induction of relatively elaborate algebraic and differential equations, including the fitting of parameters, without requiring excessive computation times. Especially where
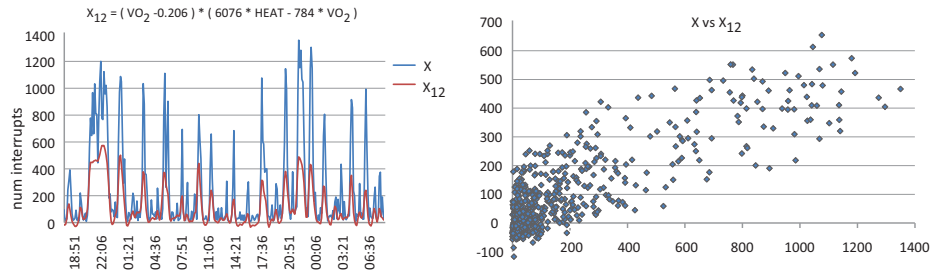
Fig. 3. Signal fit and scatter

modelling of whole-body metabolism is concerned, the use of various time-shifted variants of the primary data is essential, in order to account for different metabolic processes that have an inherent delay, the details of which may not directly be measurable in the system. The experiments show that the difference in metabolic rates of the two diets considered can be recognized from the difference in time shifts that occur in the respective equations.

5 Acknowledgements

References

1. McLean & Tobin, Animal and Human Calorimetry, Cambridge University Press 1987, ISBNO-521-30905-0.
2. Džeroski & Todorovski, Discovering Dynamics, ICML 1993.
3. Džeroski & Todorovski, Discovering Dynamics: From Inductive Logic Programming to Machine Discovery, JIIS 1995.
4. Todorovski & Džeroski, Declarative Bias in Equation Discovery, ICML 1997.