MISSION-T2D

Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data

Work Package 3

Deliverable 3.3

Applying the approach to adjust for parameter discrepancy at the interfaces between levels

Document Information

Grant Agreement	Nº	600803	Acronym	MISSION-T2D
Full Title	Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data			
Project URL	http://www.mission-t2d.eu			
EU Project Officer	Name	Dr. Adina Ratoi		

Deliverable	No	3.3	Title	Applying the approach to adjust for parameter discrepancy at the interfaces between levels
Work package	No	3	Title	

Date of delivery	Contractual	30.08.2014	Actual	06.09.2014
Status	Version 1.2		Final	1.2

Nature	Prototype		Report	X	Dissemination		Other	

Dissemination level	Consortium+EU	
	Public	X

Target Group (If Public)		Society (in general)	
Specialized research communities	X	Health care enterprises	
Health care professionals		Citizens and Public Authorities	

Responsible Author	Name	Pietro Lio	Partner	UniCAM
	Email	pl219@cam.ac.uk		

Version Log

Issue Date	Version	Author (Name)	Partner
22.08.2014	1.0	Pietro Liò	UniCAM
28.08.2014	1.1	Filippo Castiglione, Paolo Tieri	CNR
01.09.2014	1.2	Pietro Liò	UniCAM

Executive Summary	In this deliverable we describe the work done in task 3.3. This deliverable shows important links with the other deliverables (3.2 and 3.4).
Keywords	Flux balance, optimisation, networks, disease complexity

Contents

1 Introduction

In this report we describe the work done under WP3 for the deliverable 3.3: Applying the approach to adjust for parameter discrepancy at the interfaces between levels based on our analysis of diabetes and inflammatory datasets. We have used a multi objective optimization procedure to investigate diseases complexity and extract parameters values. We have developed methods based on Pareto fronts and the other on multiple networks (multiplex). The methodologies described here are clearly linked with the parameters and data analysis presented in Deliverable 3.2 (Analysis of gene copy number, SNPs and other omics) and with the methodologies presented in Deliverable 3.4, (Partially observed Markov process models of inflammation and nutritional and lifestyle aspect that have impact on T2D and inflammation) in particular related to HMM. In the next sections we describe the methodology used for task 3.3.

2 General aspects of the Methods

Out methodology is exemplified in figure 2.1 and 2.2 (figure 2.1 is a detail of figure 2.2 at step 2). While figure 2.1 describe the general aspects of investigating parameters' values on the basis of an optimization procedure in a multiple network model, the figure 2.2 describes how this information is used.

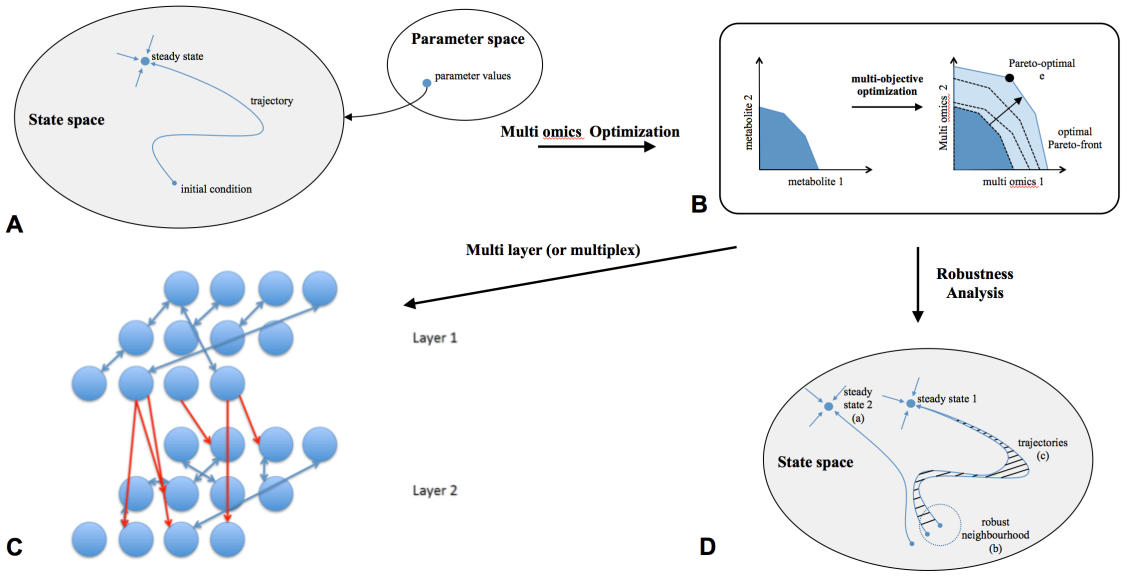


Figure 2.1 The common framework to analyze disease comorbidity. The evolution of conditions, of a disease could be seen as a trajectory in a state space (A). A multi optimization procedure of multi omic networks could be described using a pareto front (B). The information on the optimisation step could be used to build the multiplex (or multilayer) networks (C). Note that the step B provides estimation of its high-dimensional parameter space and evaluate the sensitivity and robustness (D).

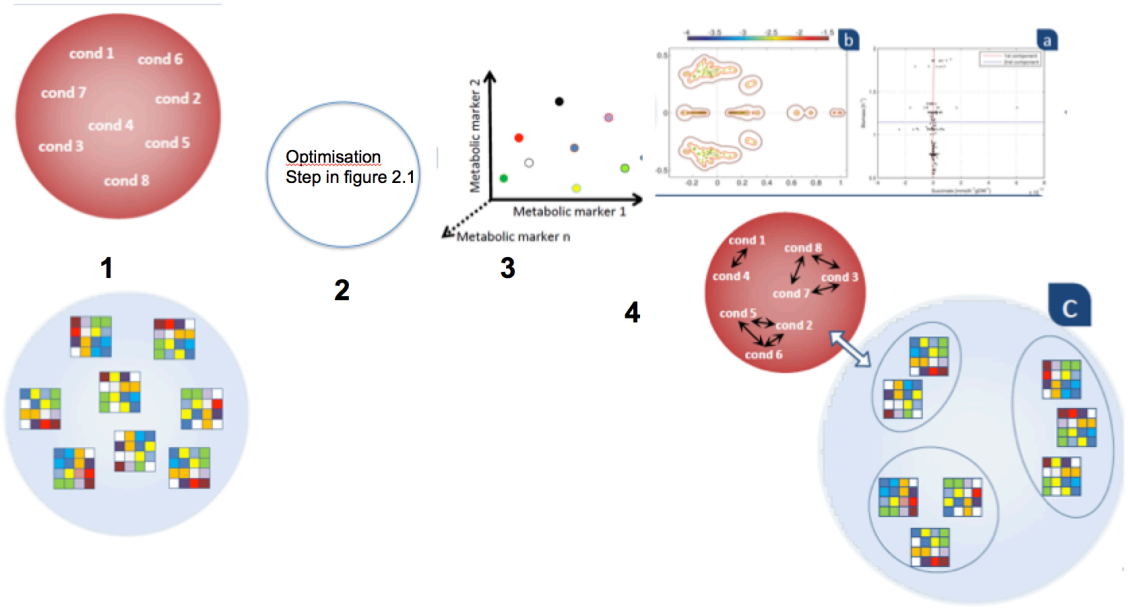


Figure 2.2 The response to different comorbid states (diabetes and inflammatory) and lifestyle and environmental conditions (1 top); is measured through microarray expression profiling (but could be through an integration of omics)(1 bottom) and other omics (not shown). We use multi objective optimization algorithms (2) to evaluate the environmental conditions and detect their community structure (3). Machine learning methods provide means to discover associations between states and conditions (4).

In the next sections we describe the steps represented in the figures 2.1 and 2.2. The optimization task is conducted with respect to a single objective function or a set of competing, conflicting, and non-commensurate objectives having nonlinear interdependence. It is necessary, hence, to use proper heuristics and algorithms to optimize the objective functions while satisfying several constraints. Recently, in multi-

objective optimization has been found important applications in a growing number of translational medicine fields. It has shown to have significant benefits compared to single-objective optimisation. Using a multi-objective optimization algorithm, we have discovered Pareto frontiers between two competing and conflicting objectives. The sensitivity analysis (SA) concerns the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input. In particular, SA tries to identify the most influential parameters of a given model; understanding which are the most important parameters of a model could be extremely difficult since it is common to deal with non-linear, highly noise and computational expensive models. The intent in robust optimization is to search for optimal design solutions which are as good as possible, and at the same time any variation in their performance (i.e. objective and/ or constraint functions), due to uncertainty, is within an acceptable range. It is important to remark the differences between Robustness (RA) and SA; RA aims to evaluate which is the probability of a system to remain in a reference state under perturbations, while, SA perturbs a system in order to find which is the aspect that mainly affects its behavior and to detect the dependencies among input parameters and between input and output. The robustness is a dimensionless metric that assesses the yield of a given system, it is the property of the system itself to undergo mutations remaining in a reference state and continuing to perform its tasks in a reliable way. Certainly In biomedical systems, robustness is generally regarded as a desiderable feature. Two kind of robustness analysis will be performed; the global robustness analysis applies a stochastic noise to each parameter; while, the local robustness analysis applies the noise one parameter at time

We detail our framework to decipher metabolic heterogeneity in different inflammatory diseases using flux balance analysis (FBA) and multi-objective optimization. Our approach aims to use gene expression data to constrain the metabolic model and to identify interesting metabolic changes that could be experimentally validated. In our investigation, we have used Recon2, and gene expression profile from GEO Omnibus repository. Recon2 is arguably the most comprehensive and complete human metabolic network to date and has been assembled using a combination of genetic, biochemical and phenotypic data. The model, available in the systems biology markup language (SBML) format, can be analysed using constraint-based reconstruction and analysis. In FBEPA, we link the gene expression profiles with the FBA fluxes of the associated reactions.

3 Flux balance methodology

Flux balance analysis, also known as, constraint-based modeling (CBM) is a mathematical approach for analyzing the flux of metabolites through a genome-scale reconstructions of metabolic networks, and is summarized in Figure 3.1 taken from [Thiele and Paulsson 2010]. FBA requires the genome-scale metabolic network reconstruction to be represented as a stoichiometric matrix S where the rows correspond to the metabolites and the column coincide with reactions in the metabolic network. Under the steady state assumption, there is no net change of mass in the system and the mass is conserved. The column vector v contains the flux through the system. Under the steady state assumption, the matrix multiplication of the stoichiometric matrix S and column vector **v** provides the linear equations to be solved through linear programming and the product of the matrix multiplication must equal zero ($\mathbf{S} \bullet \mathbf{v} = 0$).

We define the objective function $\mathbf{Z} = \mathbf{c} \bullet \mathbf{v}$, and in the objective function, we define the reaction which we want to minimize or maximize. 1 is assigned to the position of the reaction of interest in the column vector c. The objective function is constrained by the steady state assumption $\mathbf{S} \bullet \mathbf{v} = 0$ and the lower and upper bounds of the metabolic flux $lb \leq v \leq ub$. The lower and upper are vectors represents the lowest and highest reaction rate possible for each reaction. These constraints reduce the possible solution space, representing the possible flux distribution of v and FBA finds the flux distribution that optimizes the objective function.

We used Recon2, a genome-scale human metabolic network, for our flux balance analysis. We defined the biomass reaction as the objective function [Feist and Palsson, 2010], and constrained the model with gene expression values from different diabetes and inflammatory datasets. The constraints were calculated and mapped according to the Recon2 gene transcript protein reaction associations. If the gene required all the transcripts for its function, we assigned the minimum gene expression value of the transcript as its expression value. If the gene required only one transcript, we assigned the maximum gene expression value of the transcript as its expression level. If gene expression for the unique enzyme or transporter is not found, we assumed the gene was normally expressed. In our investigation, the gene expression value from 0 to 1 indicates down-regulation, 1 signals normal expression, and above 1 signifies up regulation. We calculated both logarithmic and linear constraints by multiplying the default lower and upper bounds with log2 of the gene expression value and with raw

gene expression value, respectively. Thereafter, we identified the metabolic flux that optimise for biomass reaction using linear programming software Gurobi and glpk. We performed flux balance analysis with varying amounts of glucose, glutamine, malate and lactate and under aerobic and anaerobic conditions.

It is noteworthy that COBRA, in addition to flux balance analysis, allows the user to perform single and double gene deletion studies. Given the constraints we wanted to identify the gene(s) that is most vulnerable to perturbation.

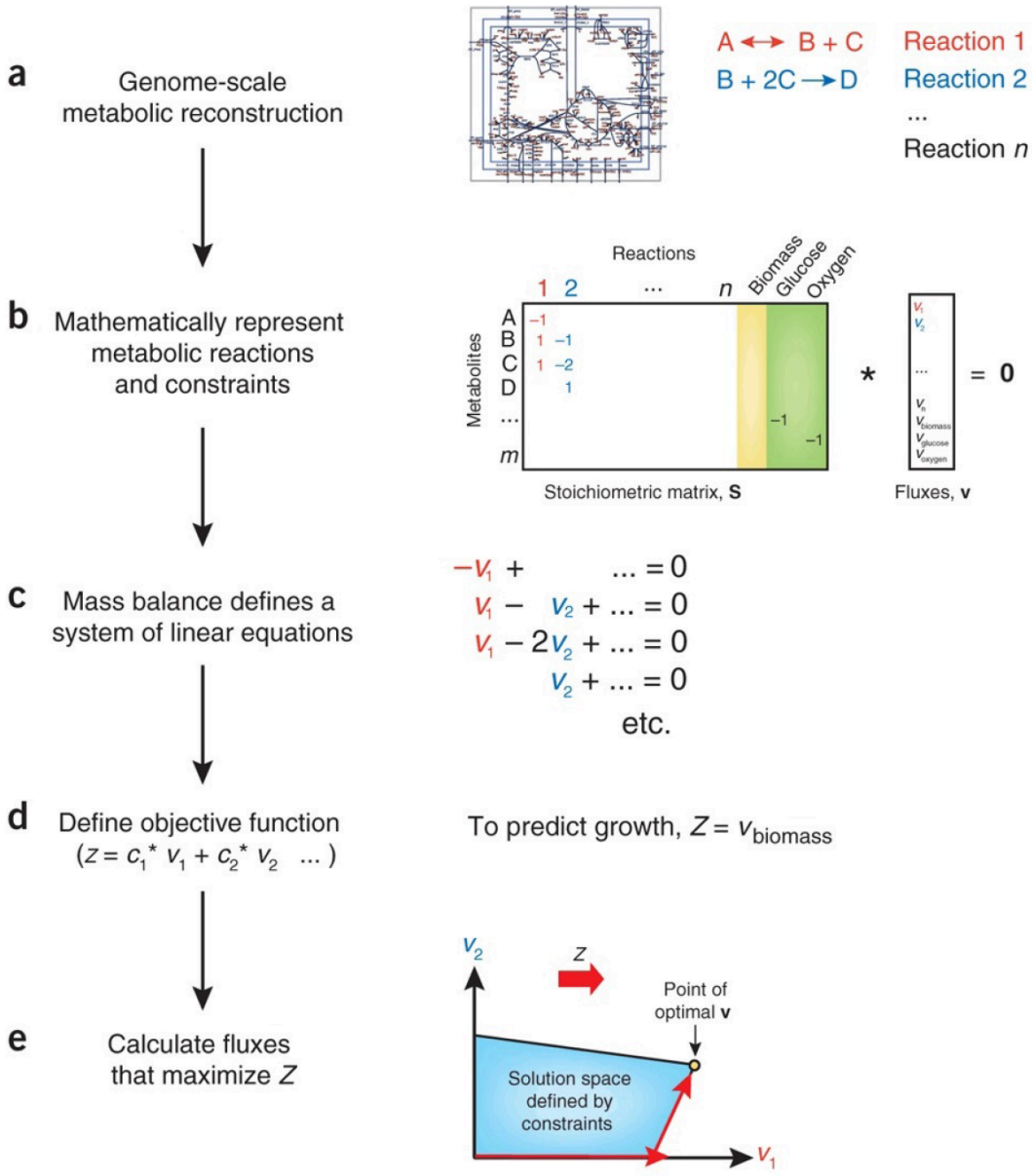


Figure 3.1 (from B. Paulsson) The figure summarizes flux balance analysis performed on Recon2 [Thiele and Palsson, 2010]. (a & b) Genome-scale metabolic network reconstruction is represented as a stoichiometric matrix **S**. (c) We assume that the

system is in a steady state. Therefore no new molecules are produced from the reaction, and the system can be represented as solvable linear equations. (d) To perform flux balance analysis, we define the objective **Z** and (e) linear programming calculates the metabolic flux that optimizes for the objective. We have chosen the biomass reaction as the reaction of interest in our investigation to predict the maximum growth rate possible given the constraints imposed by our gene expression. FBA approach finds the metabolic state in order to optimize a particular objective function as the maximization of biomass or ATP production. So, the problem can be formulated as a linear programming problem:

$$\begin{aligned} &\text{maximize (or minimize)} && f'v \\ &\text{such that} && Sv = 0 \\ & && v_j^L \leq v_j \leq v_j^U, j = 1, \ldots, n \end{aligned}$$

We would like to add more details on this approach. Therefore In the next section we describe the details of the of the multi objective optimization method.

4 Multi objective optimisation methods

Through the multiobjective optimization, we obtain the Pareto front, which represents how the organism maximizes and achieves the defined objectives. The Pareto front consists of a collection of non-dominated points that has been selected through nondominated sorting generic algorithm II (NSGA-II), and the point is described to be non-dominated, if there are no other points in the solution space that can better achieve and maximize the selected objectives. The Pareto front represents how the organism balances the defined objectives. Through the multiobjective optimization, we wanted to investigate how different tumor subtypes arbitrate the needs of biosynthesis, and we hoped to observe differences in how normal and tumors cells mediate the needs of cellular homeostasis and biosynthesis. To perform multiobjective optimization, we first apply the constraints calculated using gene expression values for metabolic enzymes and transporters from different diabetes and inflammatory diseases.

Metabolic network reconstructions available for analysis using flux balance analysis only have one objective and therefore, we introduced another objective to have two objectives for optimization. We refer to the two objectives as natural and synthetic

objective. We selected various rate-limiting reactions of biomass reactions as our natural and synthetic objectives. In the multi objective optimization process, the user first defines the number of population and generation for the algorithm.
A population is a set of individuals and each individual is assigned a randomly generated gene expression value for metabolic enzymes and transporters that influences how the model achieves the two objectives. In each generation, the individual that presents the best solution for the objectives is selected from a population through a evolutionary process. In the evolutionary process, the individuals are selected at random and their fitness is evaluated as a function of how they maximize for the objective. The individuals that are fitter are selected as parents and are used to beget the next generation.

Note that in addition to the robustness and sensitivity analysis one could also do the identifiability analysis. A non-identifiable component is a part of the system for which no unique solution exists. There are two different kinds of non-identifiability: (i) the structural non-identifiability occurs when some components are functionally related and therefore they cannot be determined unambiguously; (ii) the practical non-identifiability occurs when it is not possible to estimate precisely the component, due to low amount or quality of data available.
The area under the pareto is known as the hypervolume.

$$I_H(X) = \int_{\mathbb{R}^2} \mathbf{1}_{H(X,O)}(z)\,\mathrm{d}z,$$

this is an important estimator of the feasibility of modifying the optimization towards one objective or another . The area tell us about number of metabolic pathways which could be considered as a sort of betweenness to reach any point in the network. Finally we complete the description with the step B of figure 2.1: the multiplex representation of a group of patients.

5 Multiplex networks

Traditionally, complex networks have been used to model systems. In a complex network each unit of the system is represented as a network node (or vertex), and the interactions between elements are represented by a connection or edge. Weights are assigned to edges, quantifying the strength of the connection. This method of modeling

may be an oversimplification of systems, where we lose some information such as temporal or context-related properties of the interactions between nodes.

A multilayer network can be defined by, M= (G, C) where G is a family of graphs that can be directed, undirected, weighted or unweighted and defines the layers of M; there is a set of interlayer connections between nodes of different layers. A special type of multilayer network is the multiplex network in which the same nodes are present in all layers and where nodes can only have interlayer connections to their counterpart nodes. Multiplex networks are useful in modeling social systems, with nodes as individuals, layers representing the different types of social interaction or settings. Let us consider a multiplex network formed by M layers designated $G_1, \ldots, G_M$ with their respective adjacency matrices given by $A_1, \ldots, A_M$.

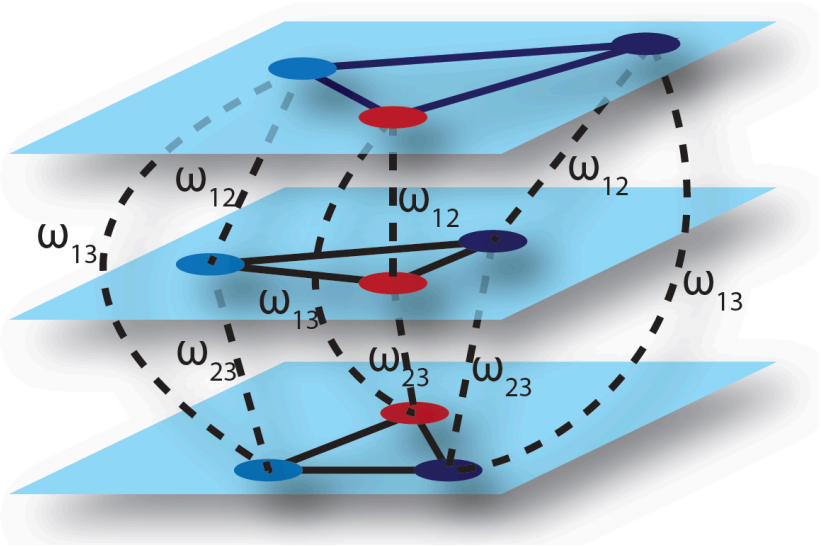


Figure 5.1 Multiplex formed by three layers, each representing a data type, and three nodes, each representing a patient (in other investigations we have carried out the layers represent Diabetes and inflammatory disease comorbidities).

Taking into account the previous assumptions and descriptions the matrix we have used could be written as :

$$M = \begin{pmatrix} A_1 & \omega_{12} I & \cdots & \omega_{1h} I \\ \omega_{21} I & A_2 & \cdots & \omega_{2h} I \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{h1} I & \omega_{h2} I & \cdots & A_h \end{pmatrix}$$

where multiplex network formed by M layers are designated as $G_1, \ldots, G_M$ with their respective adjacency matrices given by $A_1, \ldots, A_M$. The omegas represent the interlayer dependencies. In a single layered network with unweighted edges, a useful property is that the number of walks of length k between the nodes p and q is given by the p, q-entry of the kth power of the adjacency matrix of the network. In a multiplex network formed of unweighted graphs, it follows that the walks of length k in the multiplex are given by entries of Mk. Let's now discuss how to model patients with multiplex networks.

Let's imagine that we have a set of patients with the same disease (in our case diabetes). Each patient has different types of data describing them in some way. In each data type, patients have some level of similarity to each other and each data type has a level of dependency or interaction. In this case we can model the patients in a multiplex network. Each layer in the multiplex represents a particular type of data with a each node representing a patient in each layer of the multiplex. The edges between nodes in each layer represent a measure of similarity between patients in corresponding to the level of similarity between patients for the particular data type which the layer represents. The strength of interaction between each data type can be modeled by a weight connecting each layer in the multiplex. Figure 5.1 shows an example with three layers (data types) and three patients. Notice that all the edges in the layers are undirected as similarity is symmetric between patients. The same applies to the interlayer connections since data types will be equally interrelated.
Another example might be that we may have mRNA expression, DNA copy number, DNA methylation and clinical data describing diabetic patients. In this example, our multiplex network would comprise four layers, each with 1000 nodes. In each layer, each node has a weighted undirected edge connecting it to every other node in the same layer. In addition, each patient is connected to itself in every other layer by the strength of interaction between the data types. In this case, we consider that the

strength of interaction is undirected and symmetric, i.e. $\omega_{ij} = \omega_{ji}$. Now that we have modeled the similarities between patients using multiple sets of data, how can we interpret it? Our motivation is to use multiple types of data to understand or predict patient response. Response could mean anything from a response to a drug, time to relapse, etc. How can we do this using the properties the multiplex network? (diagram of multiplex and response layer). One way to do this is to compute an overall disease similarity between patients given all sets of data. We can find the disease similarity by aggregating the descriptive layers in some way, taking in to account the properties of the multiplex. If we imagine that the response can be representing by another network with similarities between patients, then we can compare the aggregate to the response network to gain some understanding of the relationship between the two. Let's consider that the edge weights between nodes provide a normalised measure of similarity between zero and one. We can define the weight of a path between two nodes in the multiplex to be the product of the edges between each node in each step of the path. Since the weight between nodes is a measure of similarity or information shared between the nodes, it follows that the weight of the path provides a measure of information flowing through the path.

There are a number of ways we can provide a new measure of similarity between two nodes given the properties of the multiplex network. One way would be to take a mean of the direct paths connecting each patient to and from another patient in each and every layer. We can define this mathematically as follows:

$$\bar{P}_{direct} = \frac{\sum_{i=1}^{h} \left(\boldsymbol{M}|_{p_i q_i} + \sum_{j=1, j \neq i}^{h} \boldsymbol{M}^2|_{p_i q_j} \right)}{h^2}$$

In many situations, a pair of nodes in a network does not communicate only through the shortest-path routes connecting both nodes, but also through all possible routes connecting both nodes. The number of these possible routes can be enormous. Moreover, the information can also go back and ford before connecting the pair of nodes. Network communicability, which was introduced by Estrada and Hatano in 2008, attempts to quantify such correlation effects in the communication between

nodes in complex networks. Estrada and Gomez-Gardenes defined communicability as a measure that “quantifies the number of possible routes that two nodes have to communicate with each other." In multiplex networks, the communicability, G, between two nodes p and q, is a weighted sum of all walks from p to q. This leads to the following relationships of the communicability between nodes p and q is given by:

$$G_{pq} = \boldsymbol{I} + \boldsymbol{M} + \frac{\boldsymbol{M}^2}{2!} + \cdots = \sum_{k=0}^{k} \frac{\boldsymbol{M}^k}{k!}\Big|_{pq}$$

$$G_{pq} = \left[e^{\boldsymbol{M}}\right]_{pq}$$

$$\boldsymbol{G} = e^{\boldsymbol{M}} = \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} & \cdots & \boldsymbol{G}_{1h} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} & \cdots & \boldsymbol{G}_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_{h1} & \boldsymbol{G}_{h2} & \cdots & \boldsymbol{G}_{hh} \end{pmatrix}$$

$$\hat{G}_{pq} = \frac{h}{\sum_{i=1}^{h} \frac{1}{[\boldsymbol{G}_i]_{pq}} + \sum_{j,k=1,j\neq k}^{h} \frac{1}{[\boldsymbol{G}_{jk}]_{pq}}}$$

with the conditions:

$$\text{minimise} \sum_{i,j=1}^{n} |a_{agg}(i,j) - a_{response}(i,j)|$$

$$\text{subject to } 0 \leq \omega_{ij} \leq 1, \forall_{i,j}$$

Finally our model could be visualised in the following way:

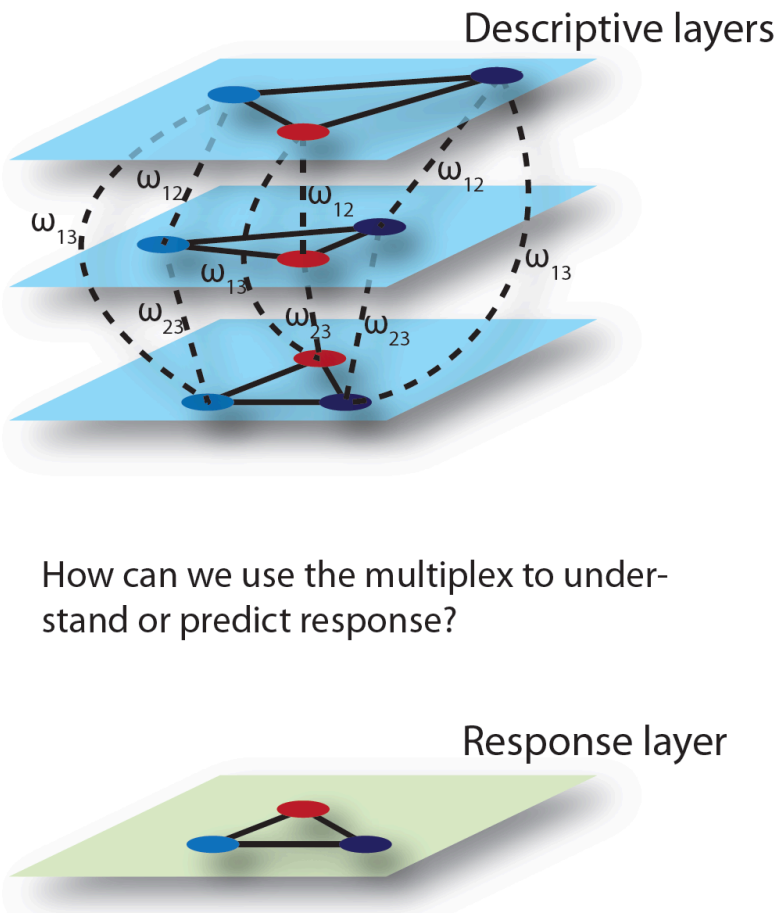


Figure 5. 1 using an aggregate layer or a response layer in a multiplex setting

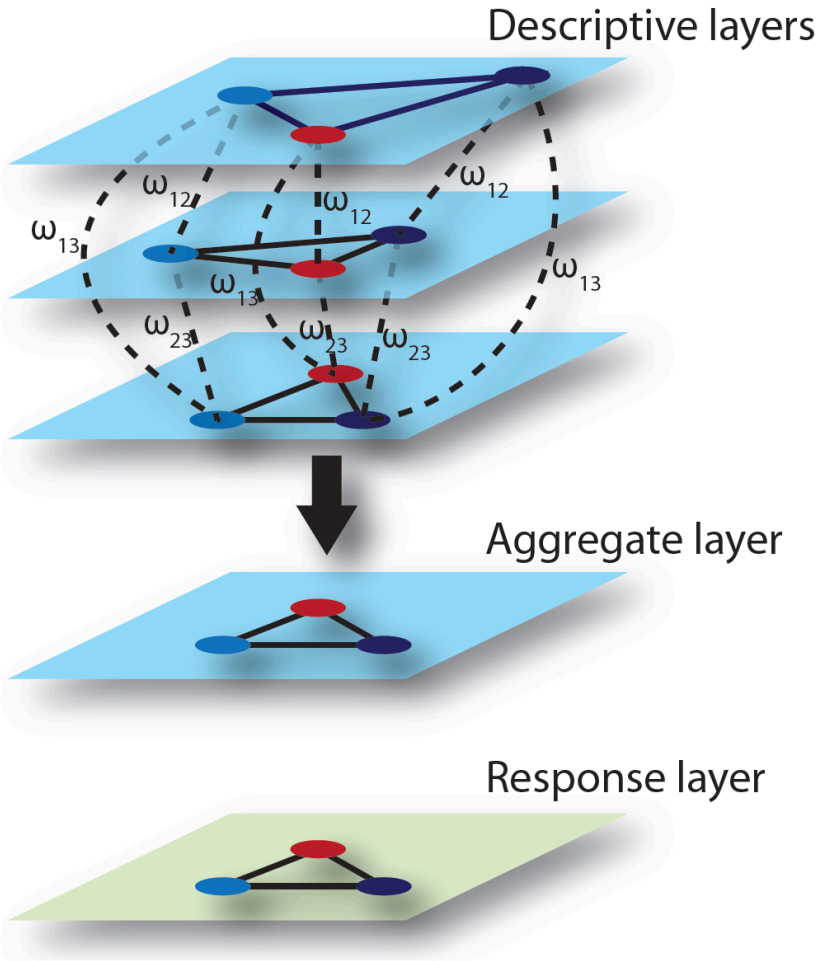


Figure 5.3 Varying ω using the method of aggregating on unfiltered layers

As previously discussed, the strength between layers in the multiplex, ω, represents a measure of dependency or strength of interaction between the layers. The edge weights between nodes represent a measure of similarity between nodes in the same layer, normalised between zero and one. Therefore, it is natural for the values of ω to represent a measure of dependence between zero and one, where zero and one indicate independence and total dependence between the layers respectively. The values of ω are not known a priori and therefore we can view them as parameters in our multiplex model. Clearly, if we vary the values of ω we should expect that the communicability between nodes in the multiplex will vary and hence the aggregate layer is a function of all the ω values. We wish to use the multiplex model to predict the response of a new patient, given knowledge of other patients. Therefore given the data

of a set of patients with known response, we want our aggregate network to match the response network as closely as possible, i.e the difference between the edge weights in the aggregate and response network should be minimised.

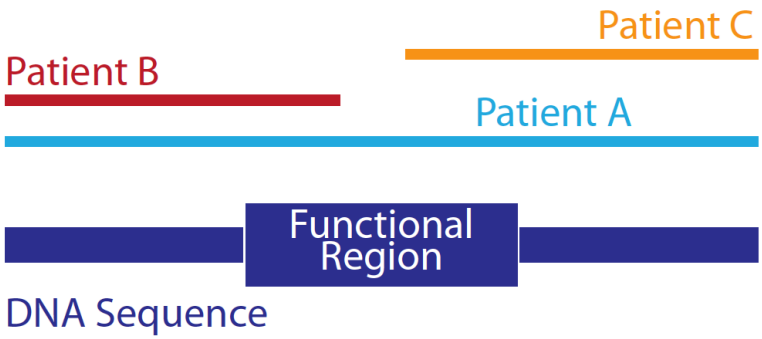
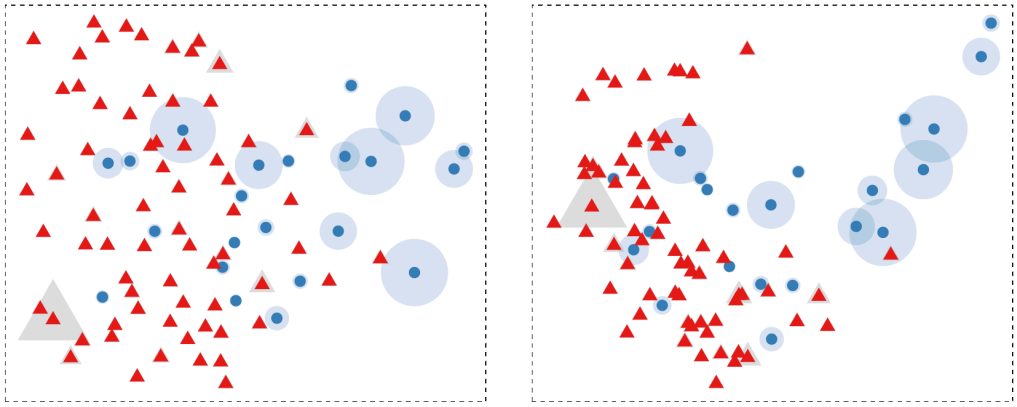


Figure 5.2 How can we aggregate the multiplex? We could use the amount of information shared indirectly through other patients. This information shared could be for example CNV or other omics discussed in Deliverable 3.2.

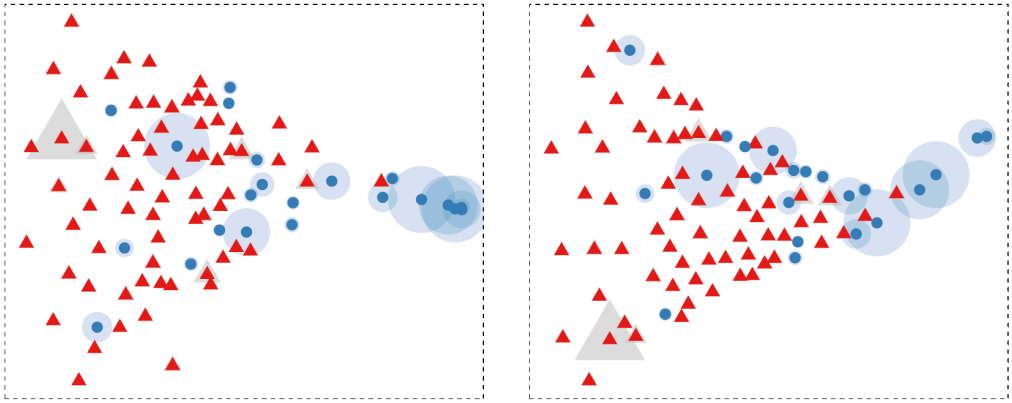
6 Results

As a summary, here we have built a multiplex model with the following characteristics: Patients have the same disease (diabetes) or comorbid diseases (diabetes and inflammation). We use Multiple types of data for all patients with Similar measure between patients for each data type. There is a level of dependency or interaction between data types; patients with the same disease make the network nodes while multiple types of data for all patients are the layers; similarity between patients for each data type are the weighted edges. The level of dependency or interaction between data types make the interlayer strengths. We build a graphical representation of distance between patients. The patients are clustered based on the various omics such as copy number aberration and gene expression data. Colour indicates cluster group. Patients that relapsed are represented by circles, the outer circles represent the relative time to relapse with larger circles indicating shorter relapse times. Patients that did not relapse are represented by triangles, the outer triangles represent the relative time that patients were monitored with larger triangles indicating a shorter study time.



: $\omega_{miRNA-mRNA} = 0.05$: $\omega_{miRNA-mRNA} = 1$

Figure 6.1 Patient clustering with omega set to 0.05 (left) and 1.



: $\omega_{miRNA-mRNA} = 0.05$: $\omega_{miRNA-mRNA} = 1$

Figure 6.2 Patient clustering with omega set to 0.05 (left) and 1.

The results reveal a striking and interesting aspects of the aggregation method. For example, we see that as ω miRNA−mRNA increases when using the direct strengths method, the patients that did not relapse increasingly cluster more strongly than the patients that relapsed did. In the indirect method, we see the opposite effect, as ω miRNA−mRNA increases the patients that relapsed tend to cluster together more strongly relative to how strongly the patients that did not relapse cluster together. This example can be seen in the graphical representations shown in figure 6.1 and 6.2

(more than one paper in progress).

7 Conclusion

We present here the patient-centered and comorbidity-centered.approaches for the aim of Applying the approach to adjust for parameter discrepancy at the interfaces between levels. The above multiplex models takes advantage of the data analysis and the data fusion methodologies presented in Deliverable 3.2 (Analysis of gene copy number, SNPs and other omics). The multiplex approach has clear connections with the Deliverable 3.4. Indeed in multiplex nodes could be patients but also disease features that are nodes of the HMM. Therefore the link between Deliverable 3.3 and the methodologies presented in Deliverable 3.4, (Partially observed Markov process models of inflammation and nutritional and lifestyle aspect that have impact on T2D and inflammation) represent a generalisation that we are exploring with the possibility of publishing different papers.

8 Bibliography

Adam M Feist and Bernhard O Palsson. (2010) The biomass objective function. Current Opinion in Microbiology, 13(3):344–349.

Ines Thiele Jeffrey Orth D and Palsson. (2010) What is flux balance analysis. Nature Biotechnology, 28(3):245–248

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., and Zanin, M. (2014). The structure and dynamics of multilayer networks. Physics Reports