MissiON-T2D

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

Work Package 3

Deliverable 3.4

Partially observed Markov process models of inflammation and nutritional and lifestyle aspect that have impact on T2D and inflammation

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	3.4	Title	Partially observed Markov process models of inflammation and nutritional and lifestyle aspect that have impact on T2D and inflammation
Work package	No	3	Title	

Date of delivery	Contractual	30.08.2014	Actual	04.09.2014
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
18.08.2014	1.1	Pietro Liò	UniCAM
28.08.2014	1.2	Pietro Liò	UniCAM

Executive Summary	In this deliverable we describe the work done in task 3.3. We have approaches the task of "Partially observed Markov process models of inflammation and nutritional and lifestyle aspect that have impact on T2D and inflammation" by implementing a Baum-Welch Hidden Markov models procedure for estimating diabetes and inflammatory diseases. Then we analysed diabetes in different age related datasets to identify age specific genes which are perturbed. We have developed a software that combines clinical, molecular, lifestyle data and gene ontology to make inference and visualise a morbidity profile.
Keywords	Baum Welch, HMM, Gene ontology, comorbidity profile

Contents

1 Introduction

Exploring associations among diabetes, obesity and inflammatory diseases at the molecular and clinical levels could greatly facilitate our understanding of pathogenesis, and eventually lead to better diagnosis and treatment. Combination of multiple types of omics, phenotype and ontology data identifies integrative biomarkers for the stratification of patients with clinical outcome. Beyond, behavioural and environmental aspects should also be considered in order to understand disease-disease associations. Recent research has increasingly demonstrated that many seemingly dissimilar diseases have common molecular mechanisms and strong associations. A comorbidity relationship exists between diseases whenever they affect the same individual substantially more than expected by chance. It represents the co-occurrence of diseases or presence of different illness or medical conditions simultaneously or one after another in the same patient. Comorbidity associations can be due to direct or indirect causal relationships and the shared risk factors among them. If two diseases have associated comorbidity, the occurrence of one of them in a patient may increase the likelihood of developing the other disease. Certain diseases, such as diabetes and obesity often co-occur in the same individual, sometimes one being considered a significant risk factor for the other. Comorbidity is an important factor for better risk stratification of patients and treatment planning. Diseases with similar genetic, environmental, and lifestyle risk factors may be co-morbid in patients or may be risk factors for additional conditions. Shared risk and environmental factors have similar consequences, prompting the co-occurrence of related diseases in the same patient. For an instance, many well-known and influential environmental factors such as smoking, diet, and alcohol intake are strongly associated with diabetes type 1 and type 2, and obesity. Also, many serious chronic diseases, such as cancer and diabetes, are complex diseases influenced by a combination of environment and epistasis between many genes. Therefore, a patient diagnosed for a combination of diseases and exposed to specific environmental, lifestyle and genetic risk factors may be at a considerable risk of developing several other genetically and environmentally related diseases. It is now well accepted that phenotypes are determined by genetic material under environmental influences. Recently, genome-wide association studies (gwas) have proved useful as a method for exploring phenotypic associations with diseases. Single-nucleotide polymorphisms (SNPs), a variation of a single nucleotide, are

assumed to play a major role in causing phenotypic differences between individuals. It has become possible to assess systematically the contribution of common SNPs to complex disease. In copy number variations (CNVs) longer stretches of DNA can get lost, duplicated, or rearranged in the genome of an individual that cause various phenotypic abnormalities. CNVs are significantly associated with the risk of complex human diseases including inflammatory autoimmune disorders, diabetes etc. The development of type 2 diabetes has also been known to be influenced by genetic and environmental factors. In this way, diseases may share many distinct types of relationships with varying levels of risk for disease comorbidity. Thus, a singular view of dependencies among diseases is not sufficient. As more and more ontology, phenotype, omics and environmental data sets become publicly available, it is beneficial to improve our understanding of human diseases and diseases comorbidities based on these new system-level biological data. The integration analysis of various 'omic' data has become increasingly widespread. Some studies have indicated that these limitations can be mitigated by integrating two or more omic datasets but a comprehensive understanding requires to inspect multiple sources of evidence. We developed a software for disease comorbidity risk assessment based on the gene-disease association, pathway disease association, DO (disease ontology) and clinical information that uses gene expression, miRNA-based relationships, shared environmental factors, ontology, SNPs, CNVs and phenotypic manifestations. Now we propose a computational framework that integrates more heterogeneous and important data including miRNA-target interactions, miRNA disease association, phenotype similarities of diseases, GO (gene ontology), SNPs, CNVs and known disease-environmental associations to capture the complex relationships between phenotypes, genotypes and clinical comordibidity.

In the next section we describe the main examples and the methodologies we have developed:

2 Methods

Diseases are connected when they share at least one significant dysregulated gene/miRNA/SNP/CNV/ GO/phenotype or environmental factor. Let a particular set of

human diseases D and a set of human genes G, gene-disease associations attempt to find whether gene g ε G is associated with disease d ε D.

If G_i and G_j are the sets of significant up and down dysregulated genes associated with diseases i and j respectively then the number of shared dysregulated genes (n^g_{ij}) associated with both diseases i and j is as follows:

$$n^g_{ij} = N(G_i \cap G_j)$$

We calculated the similarity between a pair of diseases, indicating how many entities (gene, SNP, CNV, miRNA, HPO or environmental factor) are shared. For example, for generating gene-sharing, we generated a list of genes known to be associated with each disease, and the disease similarity (correlation) was calculated based on how many genes are shared between a pair of diseases. The similarity is defined as

$$Sim(i,j) = \frac{N(G_i \cap G_j)}{\sqrt{N(G_i)} * \sqrt{N(G_j)}}$$

where N(Gi) and N(Gj) are the number of genes linked to disease i and j respectively, and N(Gi \Gj) is the number of genes associated to both disease i and disease j. SNP-sharing, CNV-sharing, miRNA sharing, HPO-sharing and environmental factor were also generated with the same approach used for gene-sharing. Hypergeometric test is implemented for enrichment analysis. It is used to assess whether the number of selected genes or ontology associated with disease is larger than expected. To determine whether any disease annotate a specified list of genes at frequency greater than that would be expected by chance, we calculate a p-value using the hypergeometric distribution. Significance of the enrichment analysis is assessed by the hypergeometric test and the p value is adjusted by false discovery rate (FDR). Then the p-value is calculated using the following formula:

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N is the total number of reference genes, M is the number of genes that are associated to the disease of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are associated to the disease. In case of GO term the p-value reports the likelihood of finding n genes annotated with a particular GO term in the set of interest by chance alone, given the number of genes annotated with that GO terms in the reference set. A biological process, molecular function or cellular location (represented by a GO term) is called enriched if the p-value is less than 0.05. The co-occurrence refers to the number of shared miRNAs/genes/ontology/SNPs/CNVs between two diseases. Each common neighbour is calculated based on the Jaccard Index method to measure the strength of co-occurrence, where association score for a node pair is as:

$$Ass_{i,j} = \frac{N(G_i \cap G_j)}{N(G_i \cup G_j)}$$

A third variation which usually improves performance significantly is the Adamic and Adar measure [Adamic and Adar, 2003], which weights the impact of neighbour disease nodes inversely with respect to their total number of connections as follows:

$$AssScore(i,j) = \sum_{n \in N(G_i \cap G_j)} \frac{1}{log(degree(n))}$$

This inverse frequency approach is based on the principle that rare relationships are more specific and have more impact on the disease similarity.

Finally we calculate the disease-disease interaction score. The score indicates the strength of the interaction between the diseases based on the protein interaction. The interaction score (ϕ_{ij}) is assigned for each disease pair i and j as follows:

$$\phi_{ij} = log(n_{ij}^{g} * N + Z) - \log(NG_i * NG_j + Z)$$

Here, NG_i and NG_j are the total number of genes for the disease, i and j, respectively; n^g ij is the total number of common genes between the two diseases. N is the size of entire proteins involved in the disease protein network. Z is a constant (Z = 1) introduced to avoid out-of bound errors, if $NG_i = NG_j = n^g_{ij} = 0$. The expected result of Φ_{ij} is positive, when the disease pair is over-represented and negative, when the disease pair is under-represented. Co-occurrence refers to the number of shared patients. This weighting scheme is used to avoid bias based on disease prevalence. The mutual information weight $W(d_i, d_j)$ between two diseases d_i and d_j is defined as

$$W(d_i, d_j) = log\left(\frac{p(d_i, d_j)}{p(d_i) * p(d_j)}\right)$$

where the numerator is the observed co-occurrence (joint probability) and the denominator is the random expectation of co-occurrence (product of marginal probabilities). The use of semantic similarity between biological processes to estimate disease similarity could enhance the identification and characterization of disease similarity besides identifying novel biological processes involved in the diseases. Graph-based methods using the topology of GO graph structure are used to compute semantic similarity. Semantic values of GO term are calculated based on the DAG of corresponding diseases. Semantic similarity for any pair of GO term is calculated based on disease semantic value.

Formally, a GO term a can be represented as a graph DAGa = (a; Ta;Ea), where Ta is the set of all GO terms in DAGa, including term a itself and all of its ancestor terms in the GO graph, and Ea is the set of corresponding edges that connect the GO terms in DAGa. To encode the semantic of a GO term in a measurable format to enable a quantitative comparison, Wang firstly defined the semantic value of term a as the aggregate contribution of all terms in DAGa to the semantics of term a [Wang et al, 2010]. Terms closer to term a in DAGa contribute more to its semantics. Thus, the contribution of a GO term t in DAGa is defined to the semantics of GO term a as the S value of the term t related to term a, Sa(t), which can be calculated as:

$$S_a(t) = \begin{cases} S_a(a) = 1 \ \ if \ t = a \\ S_a(t) = max\{w_e * D_a(t') | t' \in \ \ children \ of \ (t)\} \ \ if \ t \neq a \end{cases}$$

where we is the semantic contribution factor for edge e (e 2 Ea) linking term t with its child term t 0. It is assigned between 0 and 1 according to the types of associations. Term a contributes to its own is defined as one. Then the semantic value of GO term a, SV (a) and the semantic value of GO term b, SV (b) are calculated as:

$$SV(a) = \sum_{t \in T_a} S_a(t), \quad SV(b) = \sum_{t \in T_b} S_b(t)$$

Thus for the given two GO terms a and b, the semantic similarity between these two terms is defined as:

$$S_{sim}(a,b) = \sum_{t \in T_a \cap T_b} \frac{S_a(t) + S_b(t)}{SV(a) + SV(b)}$$

To gain more insight into the shared molecular mechanism of associated human genetic diseases, mapping was implemented from disease phenotype to gene based on the disease-gene association. With the accumulation of large amounts and multiple types of experimental data, prediction of gene-phenotype associations has emerged as a very productive subfield with great importance for the understanding of human disease. Given a particular set of human phenotypes (typically diseases) D, a set of human genes G and evidence E, these methods attempt to find whether gene g ε G is associated with phenotype d ε D. Note that evidence E can be gene-disease associations obtained through genetic studies. To quantitatively describe the phenotypic similarity between different phenotype record Pi and Pj , according to [Zhang et al, 2010] we defined the similarity measure as cosine of the angle between their corresponding phenotype feature vectors using the following formula:

$$Sim(P_i, P_j) = \frac{\sum_{k=1}^{N} w_{k,i} * w_{k,j}}{\sqrt{\sum_{k=1}^{N} (w_{k,i})^2} * \sqrt{\sum_{k=1}^{N} (w_{k,j})^2}}$$

where N is the total mapping concepts, $w_{k,i}$ and $w_{k,j}$ were the k-th term, weight in phenotype record Pi and Pj , respectively.

For each of the phenotype clusters, mapping was implemented from disease phenotypes to their associated disease genes based on the disease-gene association list in the OMIM database. Thus, we can get the corresponding gene subsets mapped to different phenotype clusters. Each OMIM phenotype was mapped to the hierarchy of HPO to retrieve the matched HPO terms. Then, a new HPO similarity is calculated for each pair of phenotypes by Jaccard similarity coefficient

$$Sim_{HPO} = \frac{|P1 \cap P2|}{|P1 \cup P2|}$$

where P1 and P2 are the set of the matched HPO terms of the two phenotypes, respectively. The way to assign terms to objects is to add annotations. In our case, the objects represent genes and terms corresponding to phenotypes (HPO terms) or biological processes (GO terms). The specificity of the terms associated with genes allows us to calculate the most significant relationships between them, which use to be related to its proximity to the root.

Individual diseases are usually annotated to multiple phenotypic features. In order to calculate the similarity between two diseases, d1 and d2, we adapt a method previously developed for estimating protein similarity with GO [Pesquita et al, 2008], whereby each feature of d1 is matched with the most similar feature of d2 and the average is taken over all such pairs of features:

$$sim(d1 \rightarrow d2) = avg\left[\sum_{s \in d1} \max_{t \in d2} sim(s,t)\right]$$

Equation above is not symmetric with respect to d1 and d2, the final similarity metric is defined as the mean of this Equation taken in both orientations:

$$sim(d1,d2) = \frac{1}{2} * sim(d1 \rightarrow d2) + \frac{1}{2} * sim(d2 \rightarrow d1)$$

This metric is used to define the similarity between two diseases.

3 Partially observed Markov process approach

We considered two layers of information: disease 1 (say diabetes) and disease 2 (say inflammatory disease). They can be both modelled as hidden Markov models. We generally assume to have data of several patients to determine the transition probabilities. Each (HMM) layer represents one disease. We consider one disease driving the other, at least partially. Each node, represents one state of the system under consideration. Statistics from different omics is used. This is only sensible if the different parts under consideration are distinct for each driver/driven sequence of chains. Thought on driver/driven sequence: given a network of "model diabetes inflammation", find augmenting paths. They give nice sequences of driver/driven nodes. The notation used is the following:

X_t: The driving Markov Chain (with $t = 1, \dots T$)

Y_t: The driven Markov Chain (with $t = 1, \dots T$)

$O_t^{(1)}$: Observations of the driving chain

$O_t^{(2)}$: Observations of the driven chain

O_t: The pair of observations

$A_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$ (adj. matrix of driving chain)

$B_{ij} = \mathbb{P}(Y_{t+1} = j | Y_t = i)$ (adj. matrix of driven chain)

$B_{ij}^{(k)} = \mathbb{P}(Y_{t+1} = j | Y_t = i, X_{t+1} = k)$ (adj. matrix of driven chain)

$P_i(k) = \mathbb{P}(O_t^{(1)} = k | X_t = i)$ (emission prob. of driving chain)

$Q_i(k) = \mathbb{P}(O_t^{(2)} = k | Y_t = i)$ (emission prob. of driven chain)

$I(i) = \mathbb{P}(X_1 = i)$ (initial distr. of driving chain)

$I(i) = \mathbb{P}(Y_1 = i)$ (initial distr. of driven chain)

e.g. $\mathbb{P}(O_{1:t})$: Probability of observing the first t observations

$\mathbb{1}(\text{statement})$: $= 1$ if statement is true, else $= 0$.

For the driving Chain this is the usual Baum-Welch algorithm, as the driving chain is just a traditional HMM. Its derivation is given for completeness and comparison to the adopted version.

$$\gamma_i(t) = \mathbb{P}(X_t = i|O^{(1)}) = \frac{\mathbb{P}(X_t = i, O^{(1)})}{\mathbb{P}(O^{(1)})} = \frac{\overbrace{\mathbb{P}(O^{(1)}_{t+1:t}|X_t = i)}^{\beta_i^{(1)}(t)}\overbrace{\mathbb{P}(X_t = i, O^{(1)}_{1:t})}^{\alpha_i^{(1)}(t)}}{\sum_{j=1}^n \mathbb{P}(O^{(1)}, X_t = j)}$$
$$= \frac{\alpha_i^{(1)}(t)\beta_i^{(1)}(t)}{\sum_{j=1}^n \alpha_j^{(1)}(t)\beta_j^{(1)}(t)}$$

$$\xi_{ij}(t) = \mathbb{P}(X_t = i, X_{t+1} = j|O^{(1)}) = \frac{\mathbb{P}(X_t = i, X_{t+1} = j, O^{(1)})}{\mathbb{P}(O^{(1)})}$$
$$= \frac{\mathbb{P}(O^{(1)}_{t+2:T}|X_{t+1} = j)\mathbb{P}(O^{(1)}_{t+1}|X_{t+1} = j)\mathbb{P}(X_{t+1} = j|X_t = i)\mathbb{P}(X_t = i, O^{(1)}_{1:t})}{\sum_{j=1}^n \alpha_j^{(1)}(t)\beta_j^{(1)}(t)}$$
$$= \frac{\beta_j^{(1)}(t+1)P_j(O^{(1)}_{t+1})A_{ij}\alpha_i^{(1)}(t)}{\sum_{j=1}^n \alpha_j^{(1)}(t)\beta_j^{(1)}(t)}$$

Forward part:

$$\alpha_i^{(1)}(t) = \mathbb{P}(O^{(1)}_{1:t}, X_t = i) = \sum_{j=1}^n \mathbb{P}(O^{(1)}_{1:t}, X_t = i, X_{t-1} = j)$$
$$= \sum_{j=1}^n \mathbb{P}(O^{(1)}_t|X_t = i)\mathbb{P}(X_t = i|X_{t-1} = j)\mathbb{P}(O^{(1)}_{1:t-1}, X_{t-1} = j) = \sum_{j=1}^n P_i(O_t)A_{ji}\alpha_j^{(1)}(t-1)$$

$$\alpha_i^{(1)}(1) = P_i(O_1)I(i)$$

Backward part:

$$\beta_i^{(1)}(t) = \mathbb{P}(O^{(1)}_{t+1:t}|X_t = i) = \sum_{j=1}^n \mathbb{P}(O^{(1)}_{t+1:T}, X_{t+1} = j|X_t = i)$$
$$= \sum_{j=1}^n \mathbb{P}(O^{(1)}_{t+2:T}|X_{t+1} = j)\mathbb{P}(O^{(1)}_{t+1}|X_{t+1} = j)\mathbb{P}(X_{t+1} = j|X_t = i) = \sum_{j=1}^n \beta_j^{(1)}(t+1)P_j(O^{(1)}_{t+1})A_{ij}$$

$$\beta_i(T) = 1$$

For the driven Chain the main difference know is, that for to find $P(Y_{t+1} = i|Y_t = j)$ one needs to condition on X_{t+1}. This why, for the steps from now on, we need to condition now on all observations O!

$$\delta_i(t) = \mathbb{P}(Y_t = i|O) = \frac{\mathbb{P}(O^{(2)}, Y_t = i|O^{(1)})}{\mathbb{P}(O^{(2)})} = \frac{\overbrace{\mathbb{P}(O^{(2)}_{t+1:T}|Y_t = i, O^{(1)})}^{\beta_i^{(2)}(t)}\overbrace{\mathbb{P}(O^{(2)}_{1:t}, Y_t = i, O^{(1)})}^{\alpha_i^{(2)}(t)}}{\sum_{j=1}^n \alpha_j^{(2)}(t)\beta_j^{(2)}(t)}$$

$$\eta_{ij}(t) = \mathbb{P}(Y_t = i, Y_{t+1} = j|O) = \frac{\mathbb{P}(Y_t = i, Y_{t+1} = j, O^{(2)}|O^{(1)})}{\sum_{j=1}^n \alpha_j^{(2)}(t)\beta_j^{(2)}(t)}$$

$$\begin{aligned}
&\mathbb{P}(Y_t = i, Y_{t+1} = j, O^{(2)}|O^{(1)}) \\
&\quad = \mathbb{P}(O^{(2)}_{t+2:T}|Y_{t+1} = j)\mathbb{P}(O^{(2)}_{t+1}|Y_{t+1} = j)\mathbb{P}(Y_{t+1} = j|Y_t = i, O^{(1)})\mathbb{P}(Y_t = i, O^{(2)}_{1:t}) \\
&\quad = \beta_j^{(2)}(t+1)Q_j(O^{(2)}_{t+1})\alpha_i^{(2)}(t)\sum_{k=1}^n \mathbb{P}(Y_{t+1} = j|Y_t = i, X_t = k)\mathbb{P}(X_t = k|O^{(1)}) \\
&\quad = \beta_j^{(2)}(t+1)Q_j(O^{(2)}_{t+1})\alpha_i^{(2)}(t)\sum_{k=1}^n B_{ij}^{(k)}\gamma_k(t)
\end{aligned}$$

Thus

$$\eta_{ij}(t) = \beta_j^{(2)}(t+1)Q_j(O_{t+1}^{(2)})\alpha_i^{(2)}(t)\frac{\sum_{k=1}^{n} B_{ij}^{(k)}\gamma_k(t)}{\sum_{j=1}^{n}\alpha_j^{(2)}(t)\beta_j^{(2)}(t)}$$

Forward part:

$$\alpha_i^{(2)}(t) = \mathbb{P}(O_{1:t}^{(2)}, Y_t = i|O^{(1)}) = \sum_{j=1}^{n}\mathbb{P}(O_{1:t}^{(2)}, Y_t = i, Y_{t-1} = j|O^{(1)})$$

$$= \sum_{j=1}^{n}\mathbb{P}(Y_t = i|Y_{t-1} = j, O^{(1)})\mathbb{P}(O_t^{(2)}|Y_t = i)\alpha_j^{(2)}(t-1) = \sum_{j=1}^{n}\alpha_j^{(2)}(t-1)Q_i(O_t^{(2)})\sum_{k=1}^{n}B_{ij}^{(k)}\gamma_k(t)$$

$$\alpha_i^{(2)}(1) = Q_i(O_1^{(2)})J(i)$$

Backward part:

$$\beta_i^{(2)}(t) = \mathbb{P}(O_{t+1:T}^{(2)}|Y_t = i, O^{(1)}) = \sum_{j=1}^{n}\mathbb{P}(O_{t+1:T}^{(2)}Y_{t+1} = j|Y_t = i, O^{(1)})$$

$$= \sum_{j=1}^{n}\mathbb{P}(O_{t+2:T}^{(2)}|Y_{t+1} = j, O^{(1)})\mathbb{P}(O_{t+1}^{(2)}|Y_{t+1} = j)\mathbb{P}(Y_{t+1} = j|Y_t = i, O^{(1)})$$

$$= \sum_{j=1}^{n}\beta_j^{(2)}(t+1)Q_j(O_{t+1}^{(2)})\sum_{k=1}^{n}B_{ij}^{(k)}\gamma_k(t)$$

$$\delta_i^{(2)}(T) = 1$$

This formulas above are coded and implemented together with the gene ontology and the integration of clinical and molecular data (see examples at pages 21-23).

The figure below shows the Flow diagram of the pipeline. A: we take as input preliminary diagnosis data of a patient and check the validation of the input. B: It performs the annotation and enrichment analysis. C: It preprocesses and updates required databases, performs statistical computation (hypergeometric and semantic similarity tests), and calculates relative risk between diseases. D: Comorbidity scores and disease network are provided as a result to the user. E: Visualisation of the comorbidity map and survival probability of patient considering comorbidity. Symbols D, E are used to indicate disease and environment respectively.

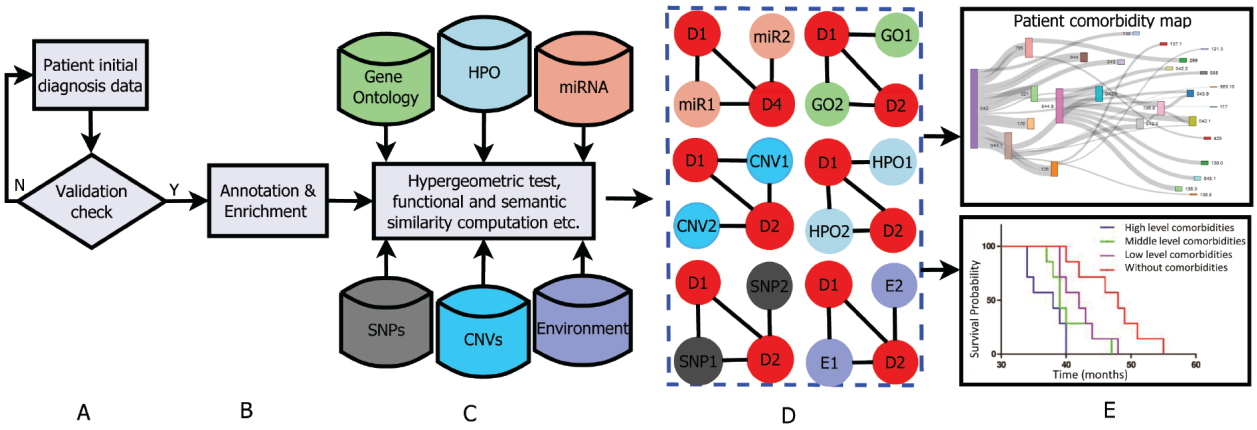

Figure 3.1 Schema of the pipeline.

4 Understanding the impact of lifestyle and behavior

The Gene Ontology (GO) is extremely useful for the exploratory analysis of microarray and other forms of high-throughput data [Ashburner et al, 2000]. GO is developed by the Gene Ontology Consortium (http://www.geneontology.org/) to describe gene products using controlled and structured vocabulary and is divided into three categories: biological process, molecular function and cellular component. The GO represents concepts, attributes, and relationships in the form of a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain. It gives the function of genes and the location of corresponding proteins. The functional similarity of GO terms is the basis for evaluating the possibility for a gene to be the causative gene of a disease. So, GO enable us to analyse disease association by adopting semantic similarity measures to expand our understanding of the relationships between different diseases. We have developed a function comorbidityGO for the computation of GO based disease comorbidity in an ontology sense. It is a GO-based enrichment analysis function to measure association among diseases and to explore their functional associations from gene sets. Statistical analysis is conducted to identify significant GO terms and to construct the functional profile of the disease gene set. Graph-based methods based on the [Wang et al, 2007] are implemented in the pipeline. We implemented a semantic similarity measurement to quantify the association between gene ontology and their associated diseases. The semantics of GO terms are encoded into a numeric format and the different semantic

contributions of the distinct relations are considered. Moreover, hypergeometric test is applied to a gene set to calculate the significance of a GO term, and the significant GO term sets are selected according to their p-values. Gene set enrichment analysis are used for predicting the significance of gene{disease and disease{disease associations. comorbidityGO function operates by using either of the following input: GO id, disease OMIM id, a list of gene symbols, Entrez gene ids or ICD-9 code of the patient disease. This function provides disease comorbidity associations and network based on the GO. comorbidityGO requires two parameters: id list and id type.

Type II Diabetes dataset called Expression data from human pancreatic islets, accession number GSE38642. This dataset was produced using GeneChip R Human Gene 1.0 ST arrays and contains information about 63 individuals, 54 of which are controls (non- diabetic). The dataset is relatively gender equalised (23 control females and 31 control males) with age range from 26 to 75 years. The fifty-four samples were divided into following groups:

Group Number	*Group Sex*	*Group Age Range*	*Number of samples*
1	Males	26 - 45 Years	3
2	Females	26 - 45 Years	3
3	Males	46 - 60 Years	15
4	Females	46 - 60 Years	12
5	Males	61 - 75 Years	13
6	Females	61 - 75 Years	8

Table 4.1

All the sample group combinations were analysed using Limma, leading to the following results (using the same Type I error threshold of = 0:05):

Compared Groups	*No. Differentially Expressed Genes* ($\alpha = 0.05$)	*No. Differentially Expressed Genes* ($\alpha = 0.05$) *with correction*
1 vs. 3	811	0
3 vs. 5	549	0
1 vs. 5	465	0
2 vs. 4	695	0
4 vs. 6	2,572	0
2 vs. 6	210	0
1 vs. 2	201	**8**
3 vs. 4	2693	**36**
5 vs. 6	459	**12**

Table 4.2

Similarly to the previous dataset, the differential gene expression analysis has detected no differentially expressed genes in relation to the age groups, while controlling for gender. Nevertheless, 56 genes were found to have differential expression in relation to the gender of the individual, while controlling for age group.

The three sets of differentially expressed genes were submitted to the GOrilla for GO Term Enrichment analysis. GOrilla was set to use the Two unranked lists of genes option, where the target set submitted was the list of differentially expressed genes found, while the background set was a list of all the genes present in the arrays. The minimal p-value was left at default value 10^{-3}, revealing following results:

GO Term Enriched	**Enrichment *p-value***
histone demethylase activity	3.44×10^{-5}
demethylase activity	6.73×10^{-5}
oxidoreductase activity, acting on paired donors ...	2.02×10^{-4}
dioxygenase activity	6.65×10^{-4}

Table 4.3 Gender Differential Expression for Ages 26 to 45; The following GO terms were enriched

All four of these terms are part of the cellular catalytic activity and seem connected to the histone modification process. The likely explanation for seeing these GO Terms enriched in comparison between the two genders is that X chromosome inactivation in females is causing differential gene expression.

GO Term Enriched	Enrichment *p-value*
histone H3-K4 demethylation	5.23×10^{-5}
cellular macromolecule biosynthetic process	2.04×10^{-4}
histone lysine demethylation	4.14×10^{-4}
sex differentiation	4.14×10^{-4}
macromolecule biosynthetic process	4.34×10^{-4}
histone demethylation	5.26×10^{-4}
protein demethylation	7.20×10^{-4}
protein dealkylation	7.20×10^{-4}
translation	7.49×10^{-4}
translational initiation	9.42×10^{-4}

Table 4.4 Gender Differential Expression for Ages 46 to 60; The following genes were enriched in GO terms for cellular processes:

GO Term Enriched	Enrichment *p-value*
histone demethylase activity	3.21×10^{-8}
demethylase activity	1.31×10^{-7}
nucleic acid binding	5.02×10^{-6}
dioxygenase activity	1.38×10^{-5}
rRNA binding	1.39×10^{-5}
heterocyclic compound binding	2.45×10^{-5}
organic cyclic compound binding	3.09×10^{-5}
histone demethylase activity (H3-K4 specific)	5.23×10^{-5}
DNA binding	5.69×10^{-5}
oxidoreductase activity, acting on paired donors ...	7.03×10^{-5}

Table 4.5: Enriched GO Terms for Cellular Function in Ages 46 to 60

GO Term Enriched	Enrichment *p-value*
cytosolic small ribosomal subunit	2.20×10^{-5}
small ribosomal subunit	8.95×10^{-5}
polysome	9.42×10^{-4}

Table 4.6: Enriched GO Terms for Cellular Component in Ages 46 to 60

histone demethylase activity	6.74×10^{-5}
demethylase activity	1.32×10^{-4}
rRNA binding	1.32×10^{-4}
nucleic acid binding	3.85×10^{-4}
oxidoreductase activity, acting on paired donors ...	3.94×10^{-4}

Table 4.7: Enriched GO Terms for Cellular Component in Ages 61 to 75

The network of Cellular Function GO terms for the enriched GO terms is shown in the Tables. The GO terms enriched for this age group are almost identical to those GO terms enriched for the other age groups and therefore it seems logical to assume that the same explanation that was suggested for the previous age groups would be valid for this age group as well (specifically, that the reason for differential expression between the males and females is caused by sex-specific genes, such as X chromosome inactivation genes). Type II Diabetes dataset in the dataset there were no differentially expressed genes detected by the GEO2R tool in terms of age. For the differentially expressed genes in respect to the gender, the GO Term Enrichment analysis has shown that most of these genes are either sex-linked (i.e. located on Y chromosome), or related to histone modifications (especially de-methylation) - from this it was suggested that X chromosome inactivation in females may account for some of the differentially expressed genes observed. Similarly to the first dataset, also for this dataset it is concluded that no comorbidity was detected in the control samples which would relate to Type II diabetes.

Then we have used gene expression data for inflammatory comorbidities and for behavioural data. This below shows the type of comorbidity output

```
> comorbidityPatients("042", "ICD9")
      ICD.9.D1 ICD.9.D2 Prevalence.D1 Prevalence.D2 Co.occurrenceD1D2 RRij
 [1,] "011"    "018"    "16646"       "639"         "110"            "134.8425079963"
 [2,] "011"    "031"    "16646"       "3693"        "807"            "171.170619171347"
 [3,] "011"    "042"    "16646"       "1067"        "64"             "46.9840607226438"
 [4,] "011"    "112"    "16646"       "141325"      "752"            "4.16805883810342"
 [5,] "011"    "117"    "16646"       "9094"        "179"            "15.4181787263579"
 ......

      CI1                  CI2                  phi                      t
 [1,] "131.740584289535"  "138.017468654697"  "0.0334998290698798"  "12.600646934426"
 [2,] "170.62851113449"   "171.714449552972"  "0.1024054800024460"  "38.7007020715709"
 [3,] "45.1417917072126"  "48.9015140627741"  "0.0148728690404686"  "5.5917689302264"
 [4,] "4.15389463700166"  "4.18227133715455"  "0.0118565029777121"  "4.45752273294143"
 [5,] "15.1992449681935"  "15.6402660615956"  "0.0136184234763953"  "5.12004213530675"
 ......
```

Table 4.8 Comorbidity information extraction using ICD9 database.

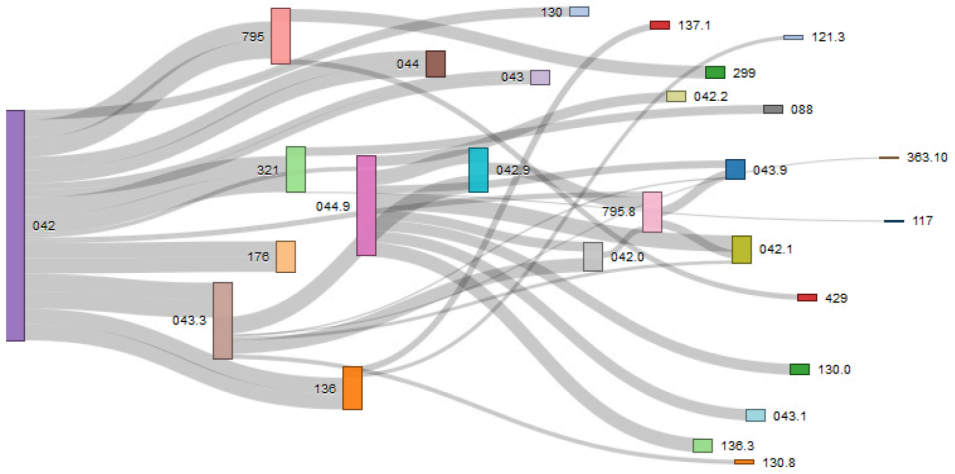

Figure 4.1 Example of the output figure comorbidityMap("042" , "ICD9") which then is used as input to the comorbidityPatients.

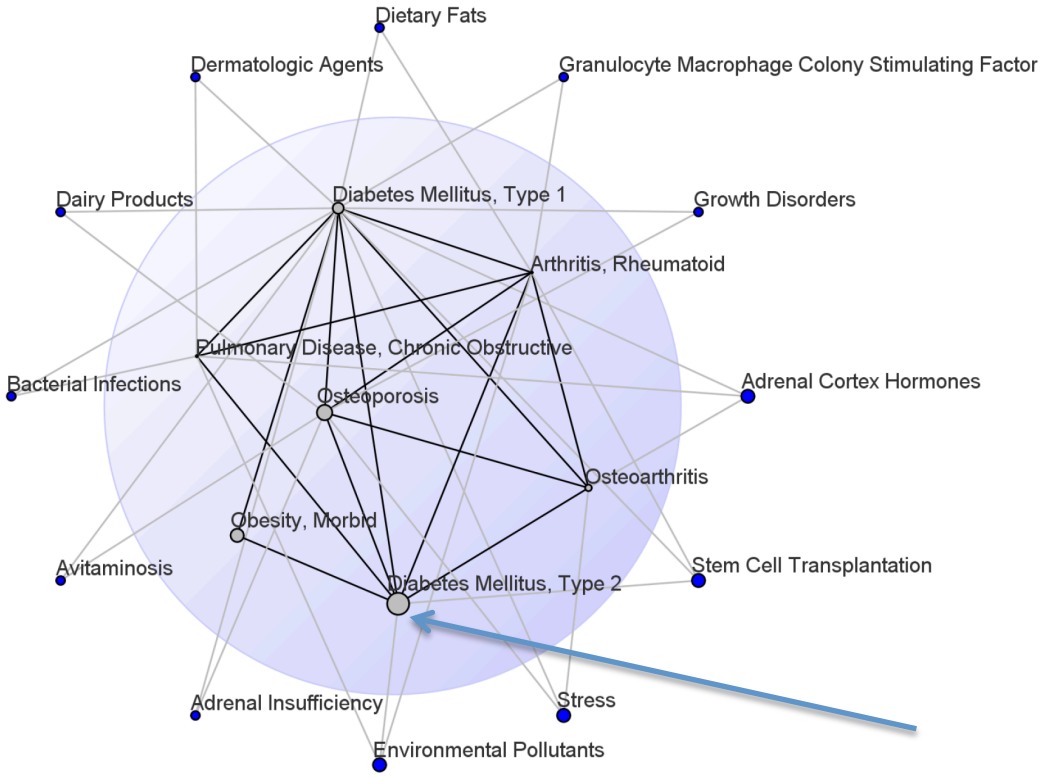

Figure 4.2 Here diseases are plotted in grey with their size proportional to the similarity to the differential expression. The arrow points to T2D patient group. Environmental

factors are plotted in blue with their size proportional to the number of diseases they are associated with. Edges between diseases indicate association of at least three identical genes while edges from environmental factors reflect direct association with the connected disease. Both types of edges are based on whether the genes or environmental factors have been mentioned together with the diseases in scientific journals.

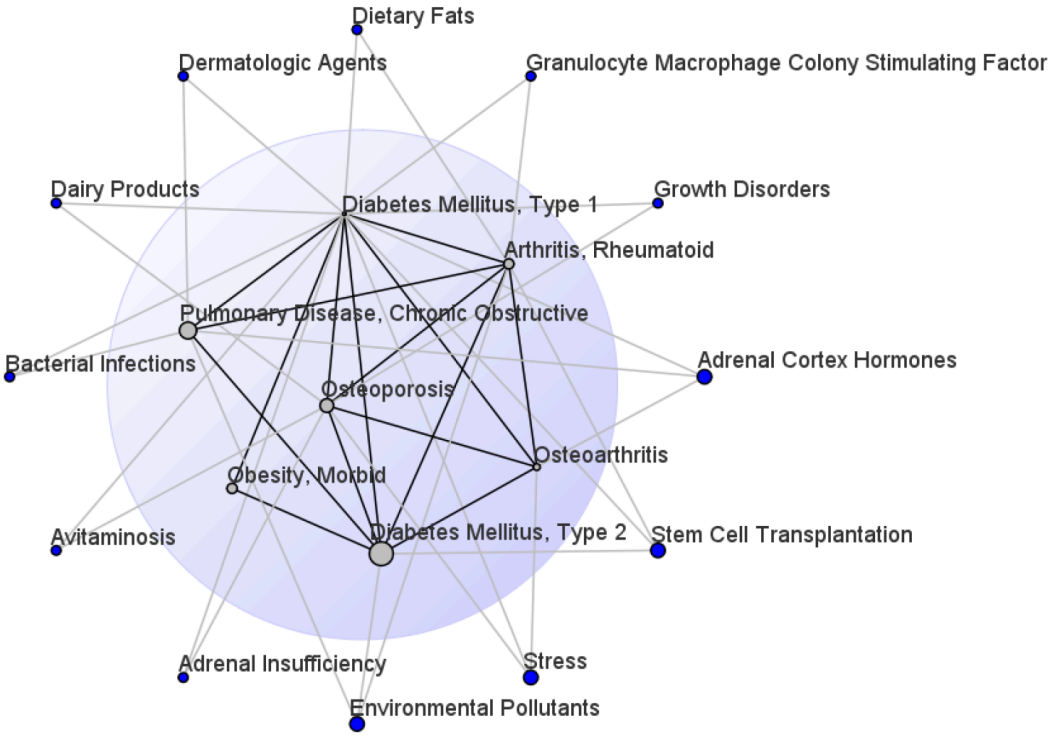

Figure 4.3 Nicotin abuse network. Color and size representation of the nodes and edges follow the same pattern as in the previous figure.

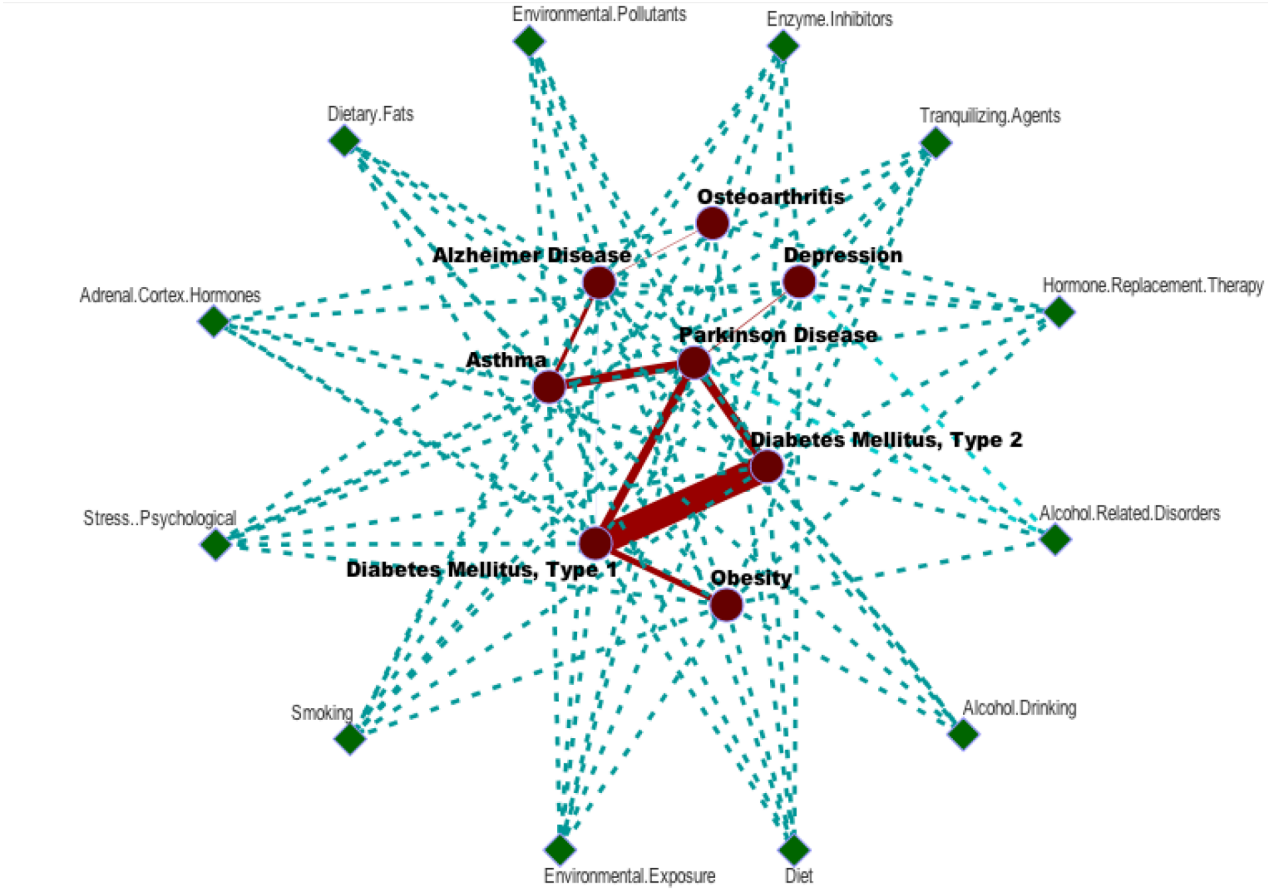

Figure 4.4 Comorbidity profile for diabetes and several inflammatory comorbidities and lifestyle conditions.

5 Conclusion

A pipeline for integrating multi omics, clinical and lifestyle information on diabetes and inflammatory disease is presented. This pipeline is based on different methodologies which use of hidden Markov models (Baum Welch) and gene ontology approaches.

6 Bibliography

Wang D, Wang J, Lu M, Song F, Cui Q (2010) Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. Bioinformatics 26: 1644-1650.

Zhang SH, Wu C, Li X, Chen X, Jiang W, et al. (2010) From phenotype to gene: detecting disease specific gene functional modules via a text-based human disease phenotyp network construction. FEBS letters 584: 3635{3643.

Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, et al. (2008) Metrics for go based protein semantic similarity: a systematic evaluation. BMC bioinformatics 9: S4.

Adamic LA, Adar E (2003) Friends and neighbors on the web. Social networks 25: 211-230.