University of Hertfordshire UH

Modelling SMA Disease GWAS Data Using Probabilistic Graphical Models

Presented by Manal Helal
To Computational Biology Journal Club

8/10/2021

Aim

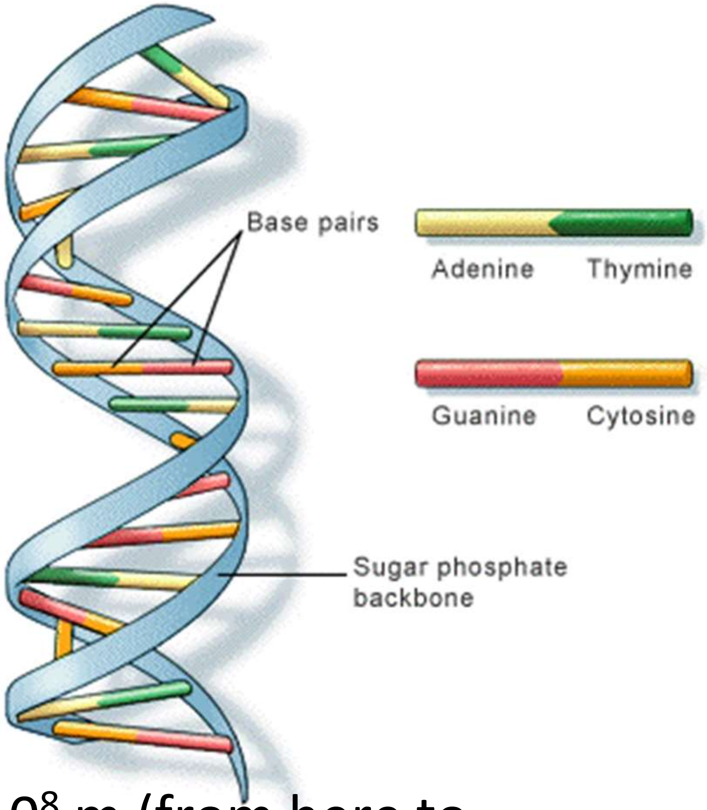
This paper models Spinal Muscular Atrophy (SMA) disease as a case study applying Probabilistic Graphical Models and Bayesian Inference. We automated the network construction using Gene Wide Association Study (GWAS) dataset augmented with gene interactions from GeneMania Database.

We aim to deliver a case study of applying data science on the public molecular databases to model diseases and report analysis metrics and graphical representations.

Objectives

1. Research SMA disease molecular causes and interactions reported in the literature.
2. Research Probabilistic Graphical Models (PGM) and their various applications in molecular interactions and disease modelling.
3. Build a PGM model for SMA disease using GWAS data and analyse its performance.

Literature Review

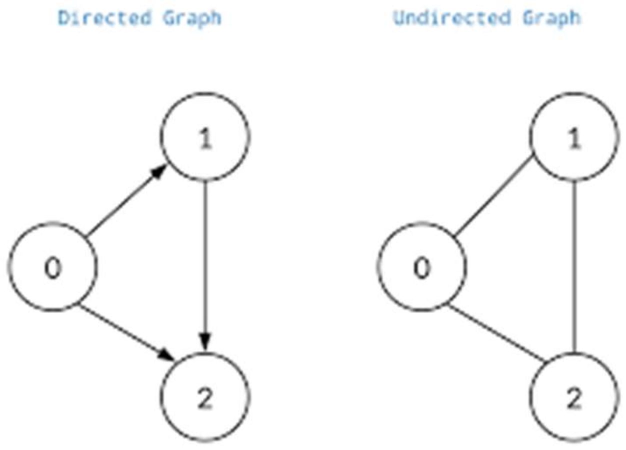


- There are around 40 trillion cells in the human body of 220 different types to do the different functions either solely or in groups to form tissues that form organs.
- Each cell contains the full DNA (Deoxyribonucleic acid) that encodes genetic code in the 23 chromosomes inherited from each parent.
- This long double helix shaped string measures 3,85 x 10^8 m (from here to the moon), contains a sequence of genes encoding all molecular activities performed by the body cells.
- **Transcription:** Each gene starts with a start codon (3 nucleobases basis) read by an RNA polymerase enzyme to start transcribing shorter single strand messenger RNA (mRNA) [2].

Literature Review – Cont’d

- **Translation**: These mRNA travels outside the nucleus and starts translating its code codon by codon to form the amino acids encoding the produced protein.
- This process of transcription and translation is ongoing with time delays or is activated by cell signals. This changes the cells’ mRNA and protein contents all the time.
- Proteins are performing most cellular activities and various functions (independently or in groups) such as: growth and DNA repair, DNA replication, metabolism (catalysing metabolic reactions), responding to stimuli (signalling and immune responses), providing structure to cells and organisms (maintaining shape by scaffolding), and transporting molecules from one location to another.
- DNA nucleobases are affected by environmental factors that causes mutations (changes, insertions or deletions) of some bases. These SNP (single-nucleotide polymorphism) sometimes lead to various diseases and phenotype [2] .

Analysis Method



- Biological networks are dynamic networks that model variable cell contents over variable time delays. The time delays are subject to the different environmental interactions, causing cell contents of mRNA and proteins to vary at the different time steps.

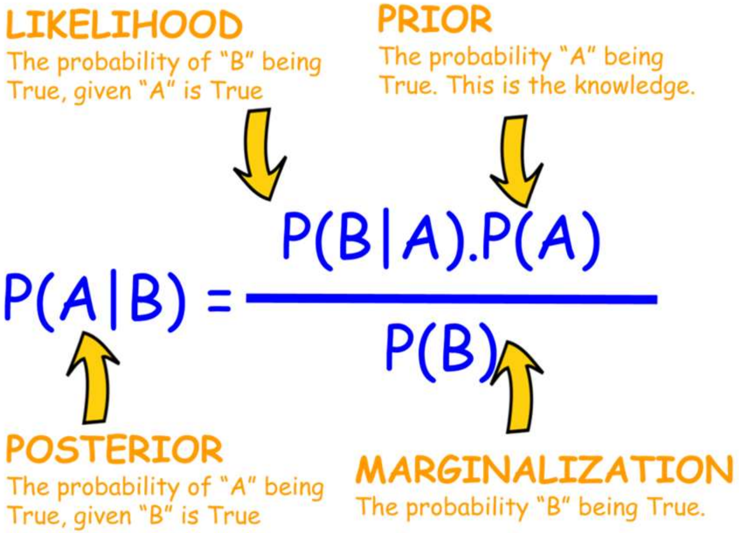
- Using this type of network, we can apply dynamic network models to infer causality and generate predictions.

- **Graphical models** are usually applied to provide graphical representation of uncertain Data represented as graphs of random variables as nodes with edges modelling interactions or correlations. These edges can be Directed and undirected.

- The generated network allows applying inference algorithms, whether exact or approximate for Decision making, and to learn the network parameters and structure, with and without complete data [3].

Bayes Theorem

- Bayesian inference derives the posterior probability according to Bayes' theorem as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data.


LIKELIHOOD
The probability of "B" being True, given "A" is True
PRIOR
The probability "A" being True. This is the knowledge.
P(A|B) = P(B|A).P(A) / P(B)
POSTERIOR
The probability of "A" being True, given "B" is True
MARGINALIZATION
The probability "B" being True.

A: SMA case
B: SNP 1 is present

P(A) % of SMA Children
P(B) % of the population having SNP1

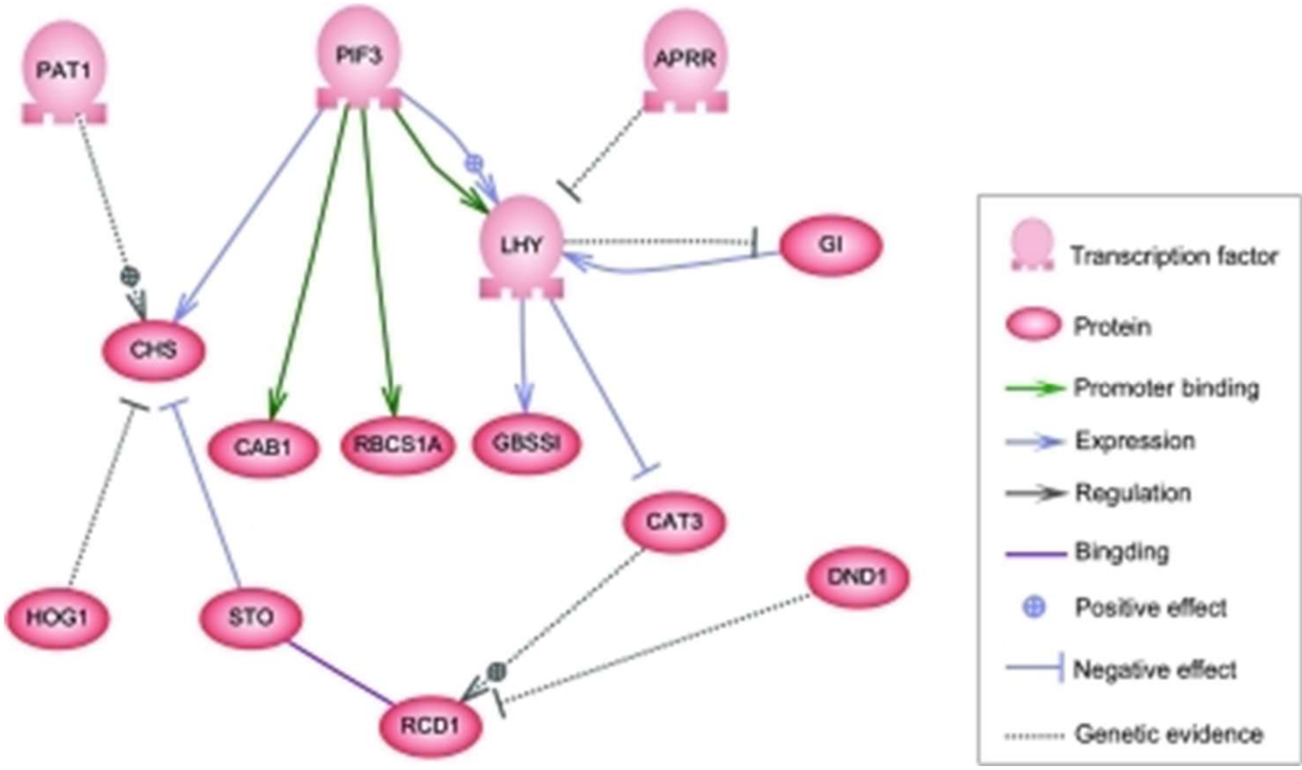
P(B|A) → Probability of SNP 1 being present given an SMA Child -- Prior
P(A|B) → Probability of having SMA if SNP 1 is present -- Posterior

Bayesian Probabilistic Graphical Models (PGM)

Bayesian network models (directed graphs) are used to capture the behavior of the interactome being modeled. This enables the predictions corresponding to experimental observations

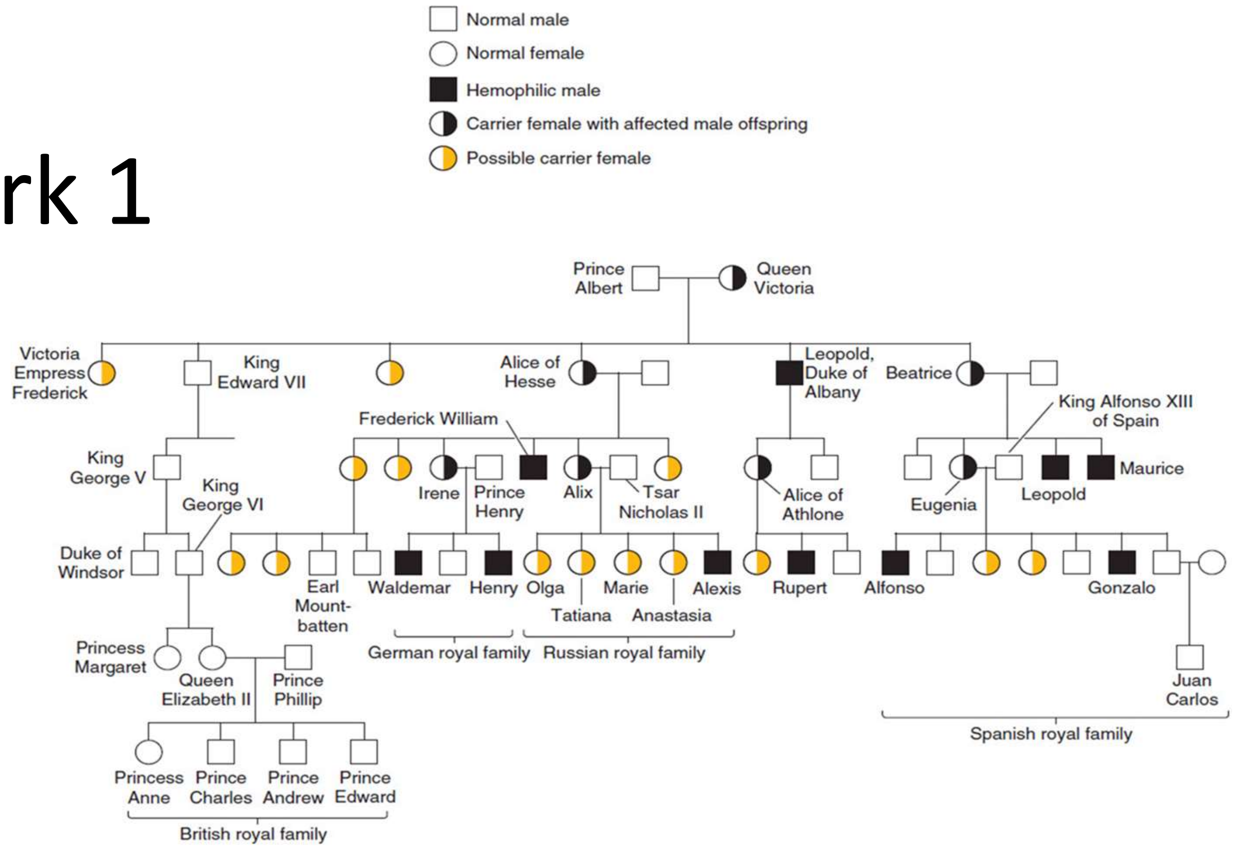
of future developments,

or inference of causality.



Similar Previous Work 1

- PGM have been applied in [4] to model pedigrees (individuals such as genotype in a particular phenotype group) to model the heredity in the meiosis process.



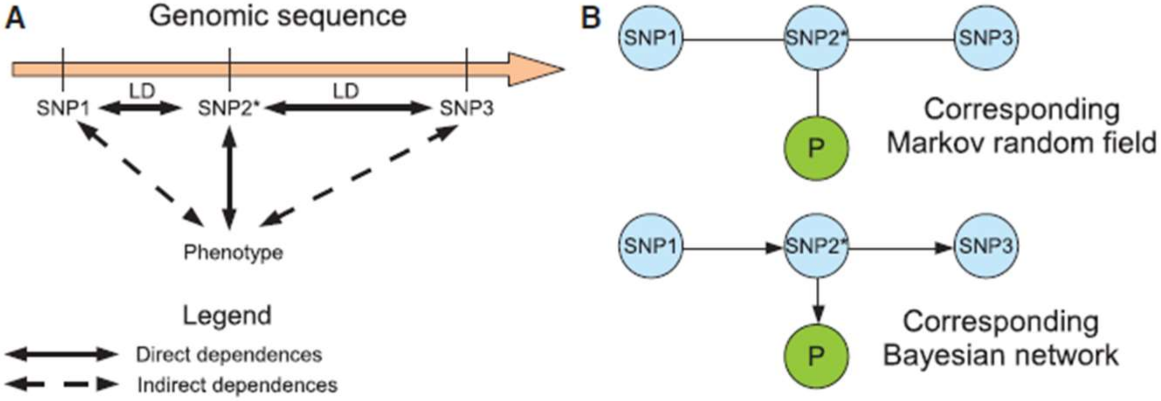
Well-publicized pedigree of haemophilia in the royal families of Europe.

G. B. Schaefer and J. N. Thompson, ‘Chapter 9: Family History and Pedigree Analysis’, in *Medical genetics: an integrated approach*, New York: McGraw Hill, 2014.

Similar Previous Work 2

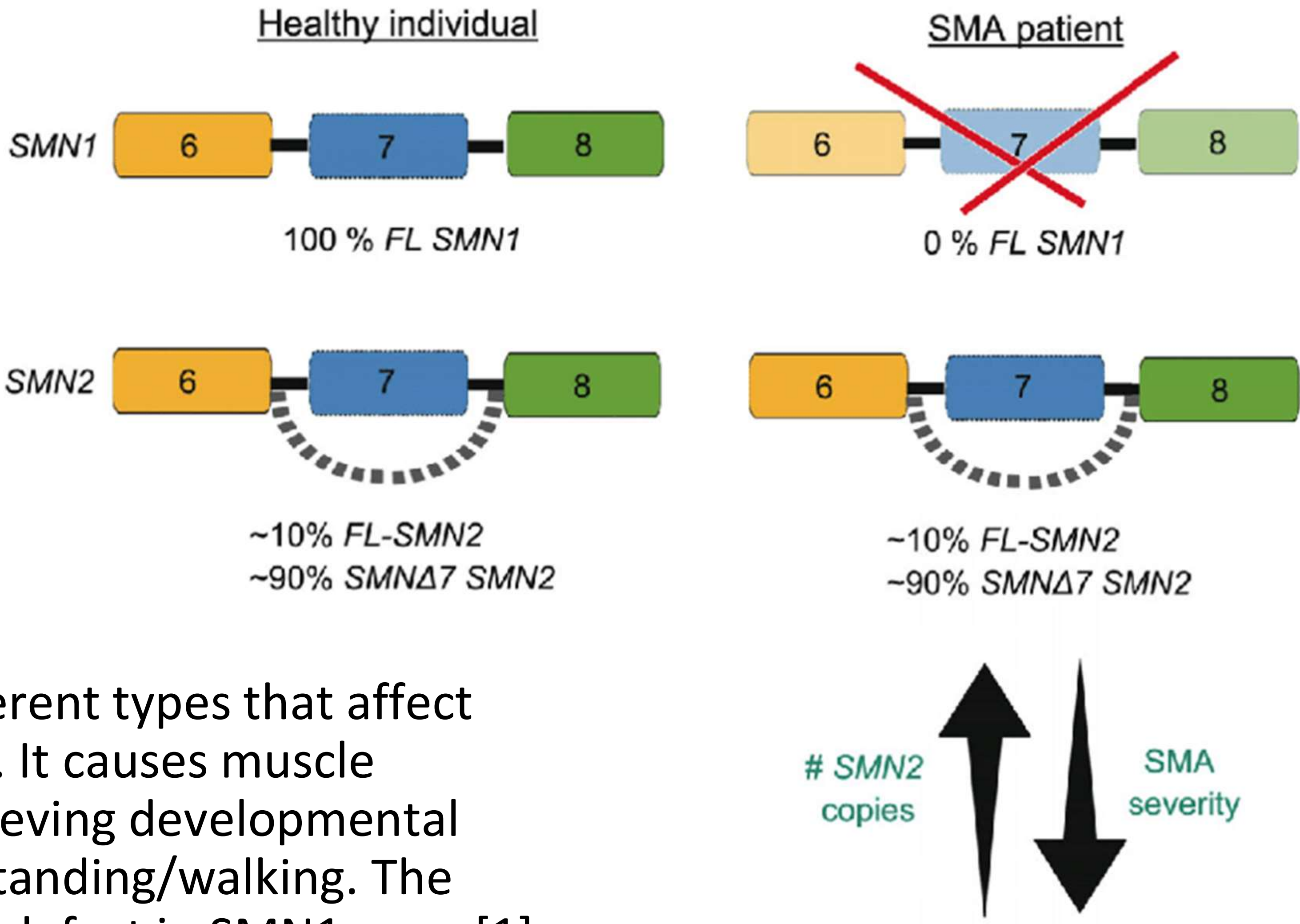
- PGMs are also modelled using Single Nucleotide Polymorphism (SNP) as the random variables capturing the dependencies between SNPs. Various models are discussed in [5] to query the SNP-phenotype association.

- Other similar work added more details, such as adding gene-gene interactions, apply data integration of genetics, gene expressions and proteomics to capture complete biological processes



(A) Distinction between direct, i.e. causal, and indirect, i.e. due to linkage disequilibrium (LD), SNP-phenotype dependences. SNP2* is the causal SNP (or the closest SNP to the unobserved causal mutation). SNP1 and SNP3 are in LD with SNP2*. (B) An MRF and a Bayesian network modeling the situation is presented in Figure 9A.The node P represents the phenotype.

What is SMA?



SMA is genetic disease of different types that affect newborn babies up to 2 years. It causes muscle weakness and inability to achieving developmental milestones, difficulty sitting/standing/walking. The disease is caused by a genetic defect in SMN1 gene [1].

Dataset

- Genome Wide Association Studies (GWAS) produces huge data representing gene expression counts by producing fragments that are probabilistically assigned to specific genes.
- An SMA GWAS study in a fruit fly model (Drosophila) was conducted in [6] collecting gene expressions variations between control (normal) and case (smn mutants) in the second and third instar larval stages.
- This GWAS SMA study showed that 3158 genes’ expression in the 1.5 fold change between 2 different time intervals and 2 different cell types (Brain and Muscle).

Pre-processing First Step

- The data was presented as different tables for the different cells and stages for upregulated and downregulated separately.
- The first pre-processing step was collecting the data per gene, per stage and per tissue type, and consider downregulation as -ve values of the change fold.
- This step produced an n × p data matrix X, comprising n observations which are cell types and time step or stage in the experiment for p molecular variables, which are genes.

Pre-processing Second Step

- The second step was to find the Entrez Ref Sequence ID from python Bio package code, which searches for the gene name in the nuccore database.
- Preference was given to IDs starting with “NM_” then “XM_” prefix. These prefixes represent protein coding transcribed mRNA. The same gene name is used for many types of molecular sequences in the Entrez database.
- This collection of IDs are generated from a sequence of searches from the most preferred RefSeq and mRNA and animal organisms, to the most unrestricted search, then compared to select the most relevant as possible.
- The difficulty in this step, is the availability of many variants, and many different types of the same organism. This reduced the genes to 3149 genes with RefSeq ID.

Pre-processing Third Step

- The third pre-processing step is to do Pathway Enrichments to reduce the dimensionality. The genes are enriched using the Reactome knowledgebase [13] to identify the pathways they are involved in. Only the genes regulating some pathways are used in the analysis. This step is processed using KNIME platform pipeline for gene expression enrichments.
- The workflow starts by identifying the Fold Change of the counts of genes in several samples using edgeR BioConductor tool to identify the most significantly differentially expressed genes.
- The significantly expressed genes as illustrated in Figure 1 are then clustered hierarchically based on their expression pattern. This step produces a heatmap as illustrated in Figure 2 and dendrogram as illustrated in Figure 3 containing 23 significant genes that has common expression patterns. A pathway enrichment analysis step identified the pathways invoked by the significantly expressed genes as illustrated in Figure 4.

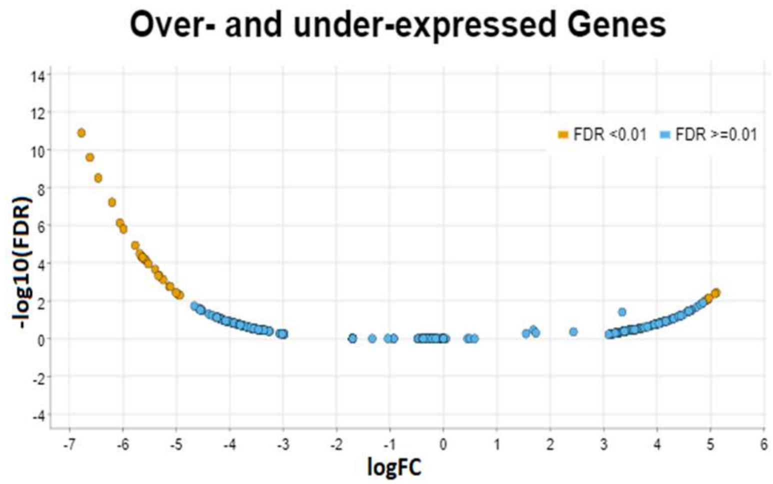


Figure 1: Over- and under-expressed Genes

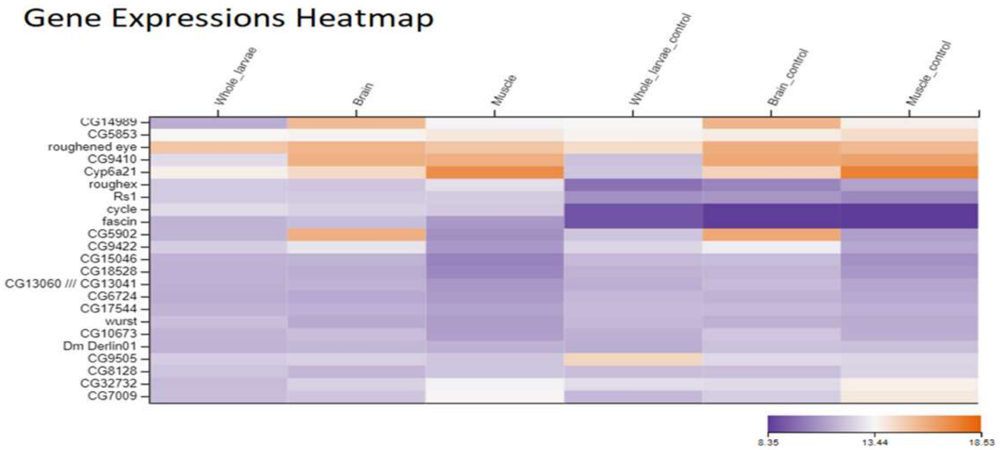


Figure 2: Heatmap for significantly expressed genes across samples from Whole Larves, Brain, and Muslce in stage 3 as compared to the Control in stage 2

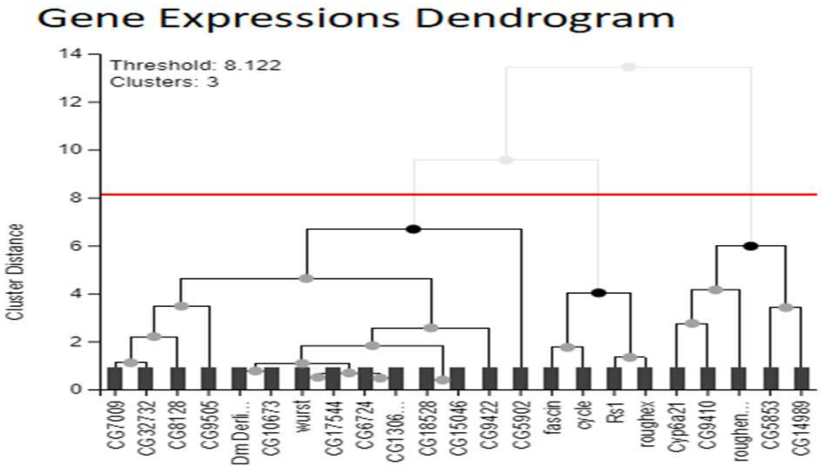


Figure 3: Dendrogram for significantly expressed genes based on similar expression patterns

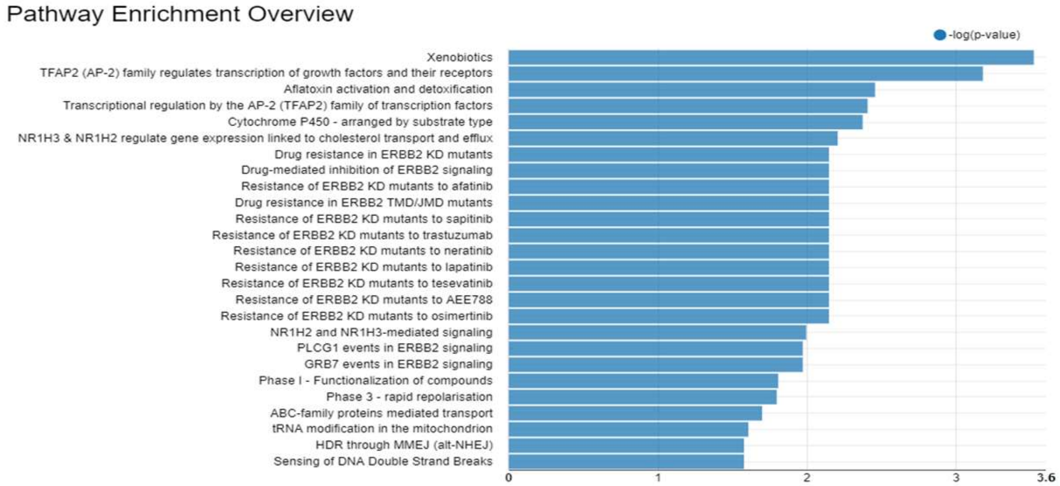


Figure 4: Identified Pathways related to significantly expressed genes.

Reverse Engineering the Gene Regulatory Network (GRN)

- A feasible PGM model to create from the observed genes' expressions is a one that aims to reverse engineer the Gene Regulatory Network (GRN) regulating the genes expressed in the experiment.
- We need to embed existing biological knowledge to create the network that identifies a parent gene to a child gene (parent gene regulates the expression of a child gene).
- We searched GeneMania for gene interactions using the most significant gene identified in the previous step. Only 19 genes were identified by GeneMania and their network was extracted as illustrated in Figure 5.

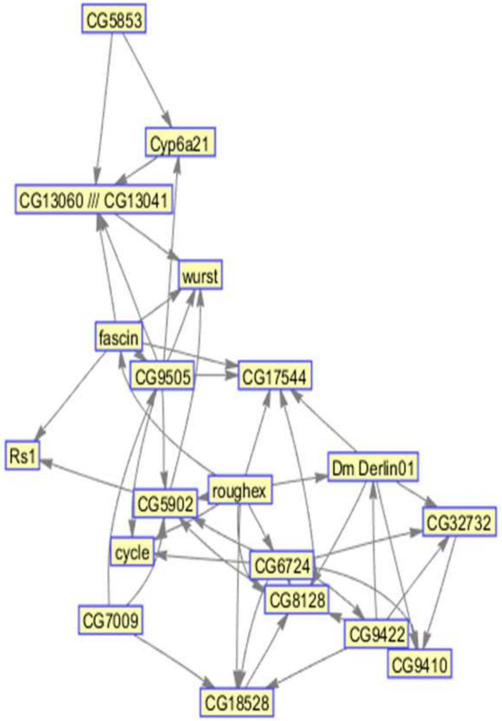
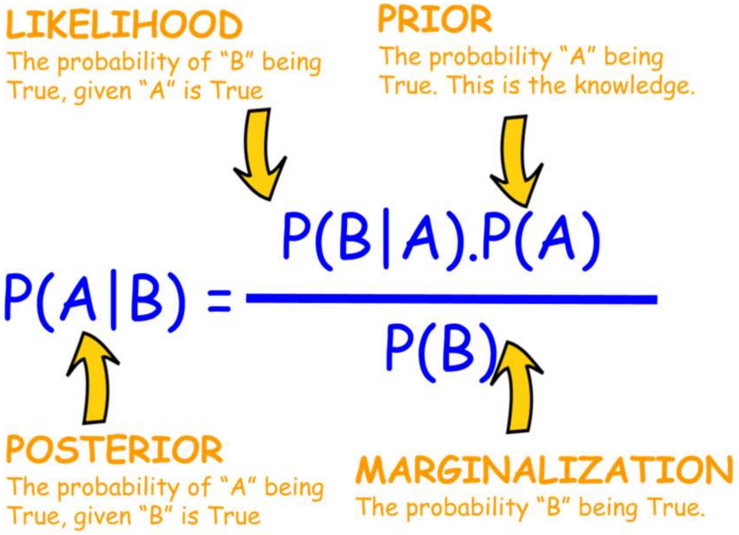


Figure 5: Genes Interactions Network extracted from GeneMania

Bayesian Inference Methods



LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR

The probability of "A" being True, given "B" is True

MARGINALIZATION

The probability "B" being True.

- We applied a Bayesian variational inference (VI) method, which is loopy-belief propagation (LBP) to analytically estimate the network parameters and latent variables from the observed variables using Expectation Maximisation (EM) algorithm. The approach estimates the parameters that fit the given data by alternating between estimating the parameters values, and maximizing the posterior estimation (MAP) until convergence.
- We also applied an MCMC method using the Hamiltonian Monte Carlo algorithm to infer the parameters estimates of the posterior distribution of each gene and enable using them for prediction.

The VI experiment

- The pre-processed 19 genes and their known co-expression were used to build a probabilistic factor graph network (FGN) to infer the marginal posterior distribution of the latent variables (biological function maintained by the messages communicated by the gene expressions).
- The loopy-belief propagation (LBP) is applied to estimate the marginal posterior distributions on all gene logical variables and compared its predicted marginals with the genes observed states from the input network [7].

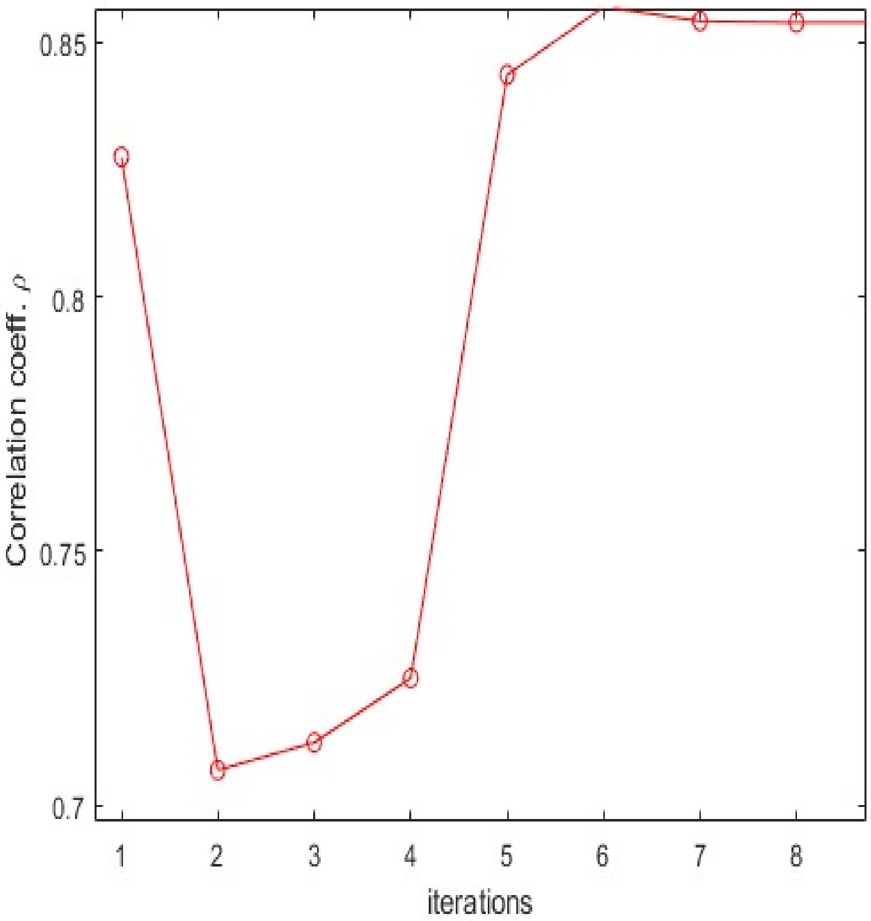


Figure 6: Pearson correlation plots of LBP message-passing convergence with increasing iteration for 2-states discretization levels in SMA response network

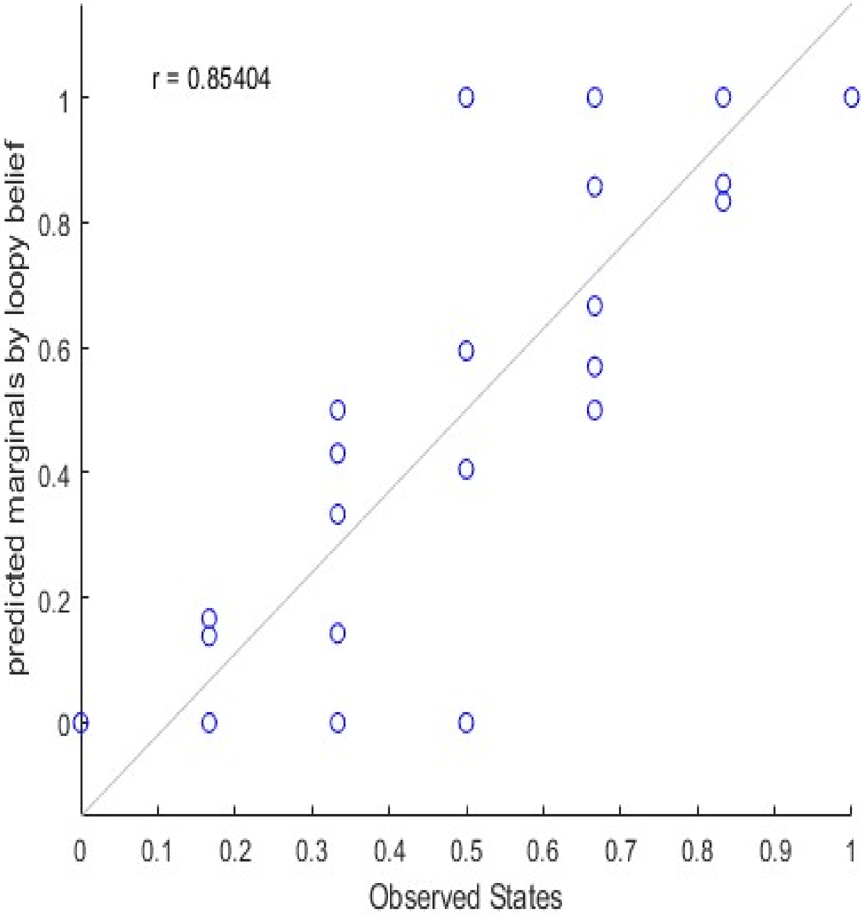


Figure 7: Pearson correlation plots between proportions of observed states and FGN inferred marginal posteriors for SMA response network: 2-states discretization, P-value 9.2836e-12 Correlation coefficient q is given in the plot

MCMC Bayesian Inference as modelled in [8]:

1. Identify the parameters to infer:

$$P(\theta_1, \theta_2, \dots, \theta_n) = \prod_{i-1}^{n} \begin{cases} P(\theta_i) & for\ a\ parentless\ paramter \\ P(\theta_i|\ parents(\theta_i)) & for\ a\ paramter\ with\ parent(s) \end{cases}$$

2. The conjugate uninformative Beta distribution for one gene (and Dirichlet Distribution for N genes) for the prior distribution since it is suitable to the random behaviour of percentages and proportions capturing the uncertainty in the expression levels of the various genes.

$$For\ a\ single\ gene\text{: } P(\theta_i) = Beta\ (\alpha, \beta)$$

$$For\ all\ the\ genes\ at\ one\ closed - form\text{: } P\left(\vec{\theta}\ \right) = Dirichlet\ (\ \vec{\alpha})$$

3. Identify the likelihood function based on the identified distributions, observing k_i fragments from a given gene i:

$$P(k_i|n, \theta_i) = Binomial\ (k_i|n, \theta_i) = \binom{k_i}{n} \theta_i^{k_i}\ (1 - \theta_i)^{n-k_i}$$

$$P\left(\vec{k}\middle|\vec{\theta}\right) = Multinomial\ \left(\vec{k}\middle|\vec{\theta}\right)$$

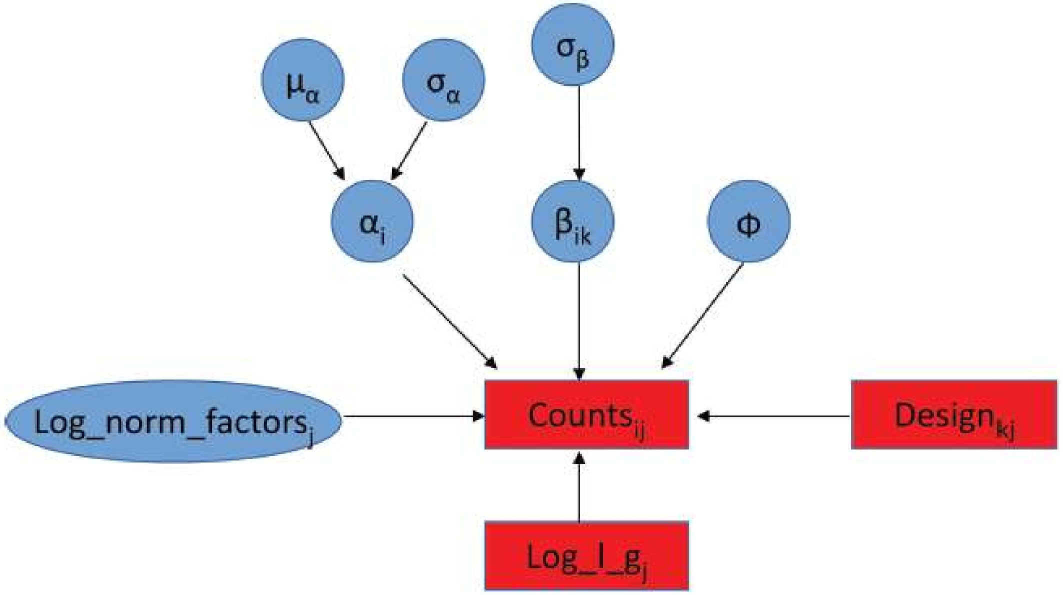
4. Identify the available dataset as follows: Gene expressions, number of genes in the model (n=19 the identified significant genes), number of samples (2 conditions – control vs SMN mutant), number of tissues (3: Whole larvae, Brain and Muscle Tissues)
5. Finally, the posterior is Dirichlet distribution as well:

$$P\left(\vec{\theta}\middle|\vec{k}\right) = \frac{P\left(\vec{k}\middle|\vec{\theta}\right) \times\ P\left(\vec{\theta}\right)}{P\left(\vec{k}\right)} = Dirichlet\ \left(\vec{\alpha} + \vec{k}\right)$$

MCMC Inferred parameters:

Starting from counts in the Negative Binomial (NB) distribution (has been widely applied to model gene expression variability across different samples in popular tools such as edgeR or DESeq2) , an informed prior model “alpha” with a common distribution that is described by two other parameters, μ_α, σ_α (mu_alpha and sigma_alpha in the code). The following are the parameters to estimate simplifying them to normal distributions:

- $\vec{\mu}$ represents the gene expression across every gene in the transcriptome.
- $\vec{\beta}$ represents the differences in the expression level between stages 1 and 2.
- hyperparameter $\vec{\sigma}$, (sigma in the code) which will describe the expected variability of the observed changes in expression between both stages.
- hyperparameter σ_β will describe the expected variability of the observed changes in expression and is the expected standard deviation for $\vec{\beta}$



MODEL(DAG)

$\text{Counts}_{ij} \sim \text{NB}(\exp(\alpha_{ij} + \beta_{ik}\ \text{Design}_{kj} + \text{Log\_norm\_factors}_j + \text{log\_eff\_length}), \Phi)$

$\alpha_{ij} \sim \text{Normal}(\alpha_\mu, \alpha_\sigma)$

$\beta_{ik} \sim \text{N}(0, \sigma_\beta)$

PRIOR DISTRIBUTIONS

$\mu_\alpha \sim \text{Normal}(n/T, 1.17)$

$\sigma_\alpha \sim \text{Normal}(0, 2)$

$\sigma_\beta \sim \text{Normal}(0, 1)$

$\Phi \sim \text{Uniform}(0, 2000000000)$

$\text{Log\_norm\_factors}_j \sim \text{Normal}(0, 0.05)$

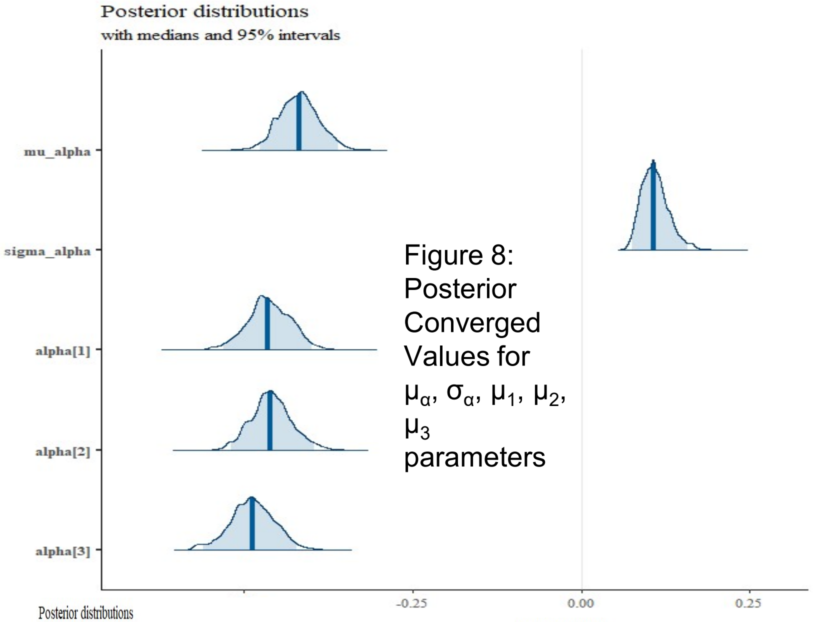


Figure 8: Posterior Converged Values for μ_α, σ_α, μ_1, μ_2, μ_3 parameters

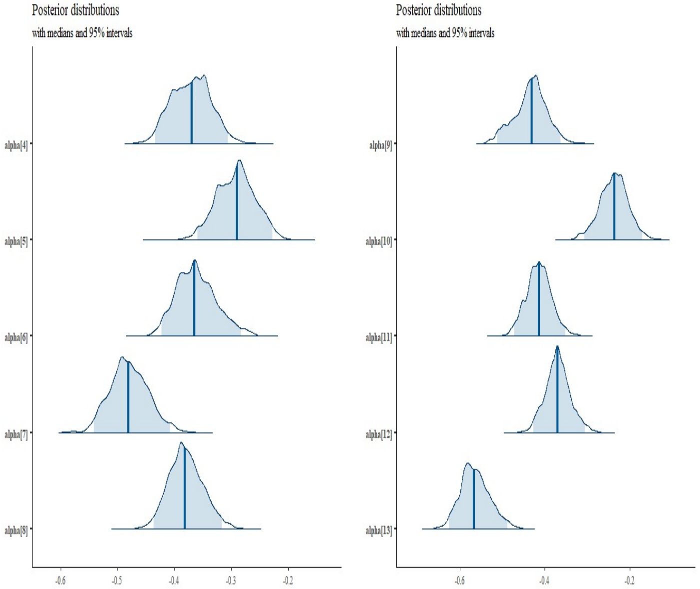


Figure 9: Posterior Converged Values for μ_4, μ_5, μ_6, μ_7, μ_8, μ_9, μ_{10}, μ_{11}, μ_{12}, μ_{13} parameters

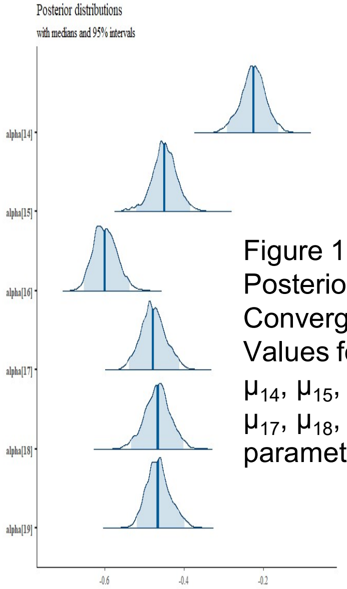


Figure 10: Posterior Converged Values for μ_{14}, μ_{15}, μ_{16}, μ_{17}, μ_{18}, μ_{19} parameters

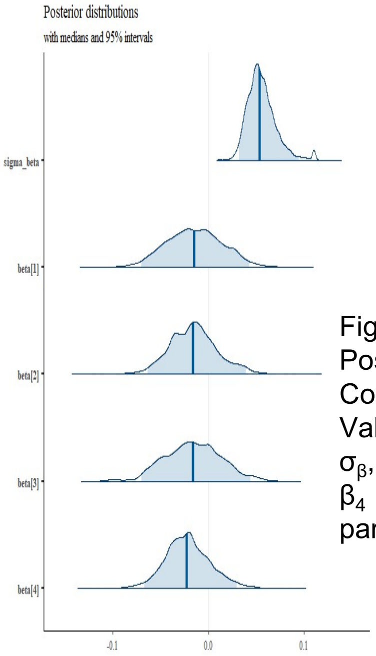


Figure 10: Posterior Converged Values for σ_β, β_1, β_2, β_3, β_4 parameters

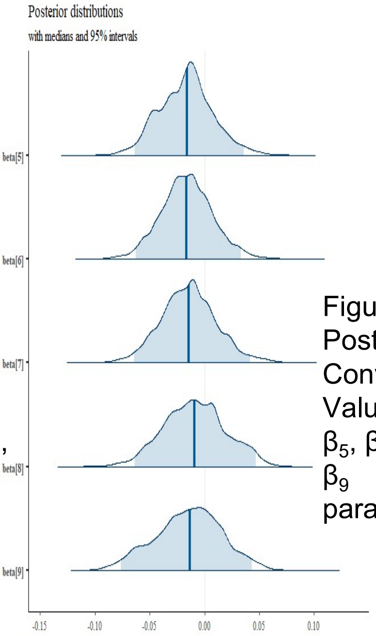


Figure 11: Posterior Converged Values for β_5, β_6, β_7, β_8, β_9 parameters

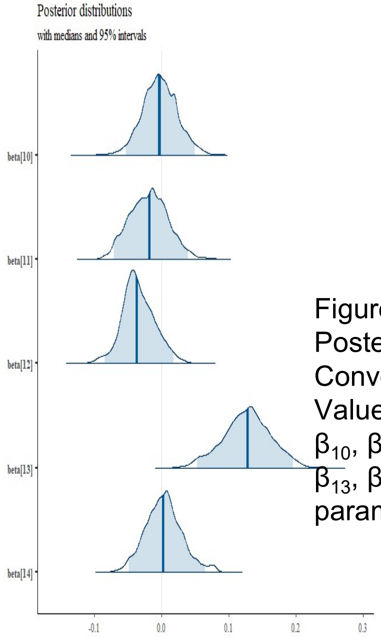
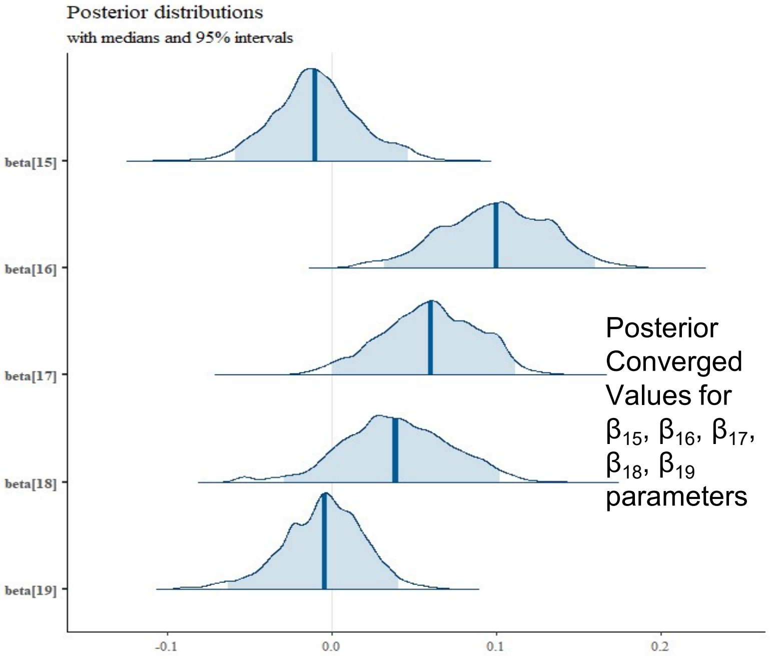


Figure 11: Posterior Converged Values for β_{10}, β_{11}, β_{12}, β_{13}, β_{14} parameters.



Posterior Converged Values for β_{15}, β_{16}, β_{17}, β_{18}, β_{19} parameters

Conclusion

- This experiment defined a pipeline to pre-processing of the GWAS generated gene expressions data to model a PGM to reverse engineer the GRN co-expression patterns in 2 different development stages of the SMA disease.
- The PGM was enriched with prior biological knowledge extracted from public databases that summarize biological experiments.
- Pathways linked to the most significantly expressed genes are identified.
- Comparing the VI and the MCMC approaches: MCMC are computationally expensive but have no bias and produce more accurate results than VI algorithms. VI approaches introduce a bias but performs a reasonable optimisation process suitable to very large-scale problems.

Limitations and Future work

- We used published datasets for SMA that was limited to 2 stages of the development of the disease.
- Both inference methods will benefit from bigger dataset across more stages of disease development or drug administration.
- Future work can aim to further connect with drugs databases such as ChEMBL for drugs that are known to target these genes expressions and pathways.
- Another suitable type of PGM is the Higher-Order Dynamic Bayesian Network (HO-DBN) that requires more than 2 stages of differential expressions.
- Various hierarchies can be added to the model by linking pathways, cis-regulatory modules (CRMs), Transcriptional Factors (TFs) or the cause/effect relationships between genes

Contribution

- To the best of our knowledge, this is the first SMA PGM model automated with data from a GWAS study and GeneMania.
- We applied Data Mining and Data Science techniques to come up with a pipeline of analytics to produce the posterior calculations using two approaches.
- These estimated posteriors imply some estimated latent variables for causality or biological functions that can be further investigated by connecting to other databases.
- Chemical interactions to inhibit a negative interaction, or promote a positive interactions is another possible future step.

References

[1] C. J. Sumner, S. Paushkin, and C.-P. Ko, Eds., *Spinal muscular atrophy: disease mechanisms and therapy*. Amsterdam: Elsevier/Academic Press, 2017.

[2] B. Alberts, *Molecular biology of the cell*, Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015.

[3] J. Wang, L. W.-K. Cheung, and J. Delabie, ‘Application of new probabilistic graphical models in the genetic regulatory networks studies’, p. 38.

[4] S. L. Lauritzen and N. A. Sheehan, ‘Graphical Models for Genetic Analyses’, Stat. Sci., vol. 18, no. 4, pp. 489–514, Nov. 2003, doi: 10.1214/ss/1081443232.

[5] R. Mourad, C. Sinoquet, and P. Leray, ‘Probabilistic graphical models for genetic association studies’, Brief. Bioinform., vol. 13, no. 1, pp. 20–33, Jan. 2012, doi: 10.1093/bib/bbr015.

[6] S. Lee, A. Sayin, S. Grice, H. Burdett, D. Baban, and M. van den Heuvel, ‘Genome-Wide Expression Analysis of a Spinal Muscular Atrophy Model: Towards Discovery of New Drug Targets’, PLoS ONE, vol. 3, no. 1, p. e1404, Jan. 2008, doi: 10.1371/journal.pone.0001404.

[7] S. Kotiang and A. Eslami, ‘A probabilistic graphical model for system-wide analysis of gene regulatory networks’, Bioinformatics, vol. 36, no. 10, pp. 3192–3199, May 2020, doi: 10.1093/bioinformatics/btaa122.

[8] V. Jiménez-Jiménez, C. Martí-Gómez, M. Ángel del Pozo, E. Lara-Pezzi, and F. Sánchez-Cabo, ‘Chapter 5: Bayesian Inference of Gene Expression’, in Bioinformatics, Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, Brazil and H. I. Nakaya, Eds. Exon Publications, 2021, pp. 65–87. doi: 10.36255/exonpublications.bioinformatics.2021.