Deciphering Genomic Heterogeneity and the Internal Composition of Tumour Activities through a Hierarchical Factorisation Model

José Carbonell-Caballero, Antonio López-Quílez, David Conesa, Joaquín Dopazo

November 8, 2021

1 Integration of Somatic Mutations with Gene Expression Values

The set of somatic mutations contained in the genome of a given patient could have a direct effect on the activity of its genes. The prediction of this type of alterations is usually performed by a somatic variant caller(**?**), whose output consists of a list of observed somatic mutations supported by specific contextual information, such as the genomic position or the observed nucleotide change. In our protocol, this information is directed to the *SnpEff* (**?**) tool that evaluates the effect of individual mutations over carrier genes. In particular, *SnpEff* summarises the effect of a given mutation according to four general categories (*HIGH*, *MODERATE*, *LOW*, and *MODIFIER*) that describe different degrees of mutations effect (Table **??**).

Table S1: Categories provided by the *SnpEff* tool for characterising the effect of a given mutation in a particular gene.

Effect	Description	Weight (ω)
HIGH	Significant structural changes, such as the presence of a premature stop codon, or a change in the reading frame	0.99
MODERATE	Mutations that do not result in an overall change in protein structure, but result in change or loss of specific amino acids	0.75
LOW	Mutations in the coding sequence that do not result in an amino acid change in the protein	0.1
MODIFIER	Mutations that do not affect the coding region of the gene	0.05

Then, we proceed to aggregate this information at the gene level, creating the matrix X_g^v that depicts the combined effect of somatic mutations in the same genes. In particular, for a gene i, and an individual j, the mutations are aggregated as follows:

$$X^v_{g,i,j} = \left[1 - \sum_{v}^{V^-_{i,j}} \omega(\psi(v))\right] \left[1 + \sum_{v}^{V^+_{i,j}} \omega(\psi(v))\right], \tag{1}$$

where $V^-_{i,j}$ y $V^+_{i,j}$ correspond to the set of mutations that produce a loss and gain of function, respectively, ψ to the most severe effect produced by the v mutation in any of the carrier gene isoforms, and ω to its predefined weight (Table **??**). To determine if a particular mutation provides a gain of function, it will be necessary to assess whether the number of affected individuals is higher than expected. To do this, a hypothesis testing is performed by taking as a reference distribution the number of affected individuals across the whole set of somatic mutations in the same gene. In this case, a gain of function is accepted when the p-value obtained by the contrast is below 0.05. In particular:

$$p = 1 - \phi^-1(\eta), \tag{2}$$

where η corresponds to the number of individuals affected by the evaluated mutation, and ϕ^{-1} to the cumulative density function of the reference distribution.

Finally, once the mutations effects have been aggregated in X_g^v we obtained the final gene activity matrix as:

$$X_g = X_g^e \odot X_g^v, \tag{3}$$

where $\odot$ corresponds to the matrix dot product, and X_g^e to the gene expression matrix.

2 Altered Biological Processes

The set of cellular functions that take place in cells are driven by the combined action of different molecular pathways. On this basis, quantifying the molecular pathways allows us, in turn, to estimate the activity of those cellular functions that may be altered in the disease.

For a complete collection of cellular processes we used the Gene Ontology (G0)(**?**) database. GO is an ontology of terms that describe cellular activities at different levels of granularity, specifying the proteins involved in each specific biological process. To estimate the activity of a given biological process, we need to determine which signalling pathways are involved in its regulation. To do this, the annotations provided by GO at the gene level were projected into signalling cascades. More precisely, each signalling cascade inherited the GO annotations that are specific to its effector protein, since the aim of the cascade is to activate this protein and thus the processes it regulates(**?**).

3 Signalling Pathway Quantification: Building the *Hipathia* Function ($\hbar$)

Hipathia(**?**) models a set of signalling pathways by using directed graphs, where the nodes represent individual proteins or complexes, and the edges their interactions. As a preliminary step, *Hipathia* decomposes each signalling pathway into one or more subgraphs representing individual signalling cascades. Each subgraph depicts how cells respond to a particular stimulus from the extracellular space, in order to activate a final effector protein responsible for orchestrating a medium term cellular response.

In practice, *Hipathia* estimates the pathway activity in a set of samples from their gene expression values. We represent *Hipathia* as a function described as

$$\hbar : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times n} \tag{4}$$

$$X_p = \hbar(X_g), \tag{5}$$

with m being the number of genes, p the number of signalling cascades, and n the number of individuals.

To quantify the level of activity of a given signalling cascade, *Hipathia* simulates the propagation of a virtual signal that diffuses through the network from the initial nodes to the final node. Here, the virtual signal attempts to reproduce the flow of information that occurs when a ligand from the extracellular space is captured by a membrane receptor protein, triggering a cascade of interactions that culminates in the activation of the effector protein. To account for gene activity, *Hipathia* modulates the flow through each node in the network according to its level of activity. Furthermore, due to each protein can be activated or inhibited by several proteins or complexes, the output flow from a particular node j is defined by using the following equation:

$$s_j^{(t)} = v_j \left[1 - \prod_{a \in A} (1 - s_a^{(t-1)}) \right] \left[\prod_{i \in I} (1 - s_i^{(t-1)}) \right], \tag{6}$$

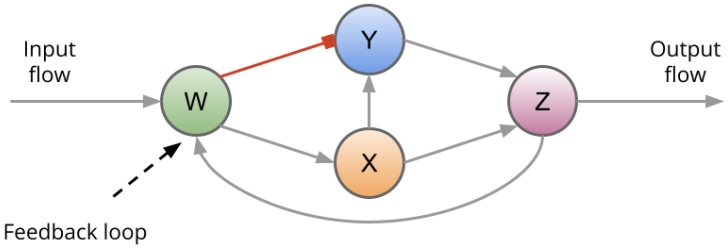
where $s_j^{(t)}$ corresponds to the output flow of node j at the iteration t, and v_j its activity value, with $s_a^{(t-1)}$ and $s_i^{(t-1)}$ being the output flow value calculated in the previous iteration at each of its A activators and I inhibitors, respectively. This equation is applied to each node as the virtual signal propagates through the subgraph, with the amount of flow at the final node being the value that directly describes the cascade activity.

Due to the presence of feedback loops inside molecular pathways, *Hipathia* employs an iterative process that is maintained until the output flow value remains stable. This approach allows *Hipathia* to efficiently model any kind of signalling network regardless of its complexity, but in contrast, makes its integration into more general optimisation processes, such as MF, more complicated. In particular, since the flow value is not obtained from an analytic solution, we lack a set of equations that could be used to calculate the corresponding partial derivatives in a classic optimisation process, such as those based on a gradient descend.

In order to overcome this limitation, a customised version of *Hipathia* was prepared (Figure **??**), providing a specific equation for each signalling cascade, equivalent to the iterative process of *Hipathia* for a predefined number of cycles (typically 5). As a result, the *Hipathia* function is redefined as $\hbar = [J_1, J_2, ..., J_p]$, where J_i corresponds to the equation associated with the i-th signalling cascade included in the tool.

To do this, a controlled execution of *Hipathia* was performed, taking as input a randomly generated activity matrix. During the iterative process, the output flow equation of *Hipathia* (eq. **??**) was emitted after visiting each node. This equation describes the output flow of a given node at a specific iteration, defined according to the incoming flow of its particular inhibitors and activators in the previous iteration. This process generated a set of equations defining the output flow value of each node across all performed iterations. Then, the terms defined at time t were recursively substituted by the equivalent terms including their regulators at time $t-1$,

until $t = 0$ was reached. As a result, at the end of the recursive process, we obtained an expression that describes the output flow value of the final node as a function of all previous regulators at time $t = 0$, corresponding to the input matrix values. Finally, the set of obtained equations was validated against the values provided by the original tool version.



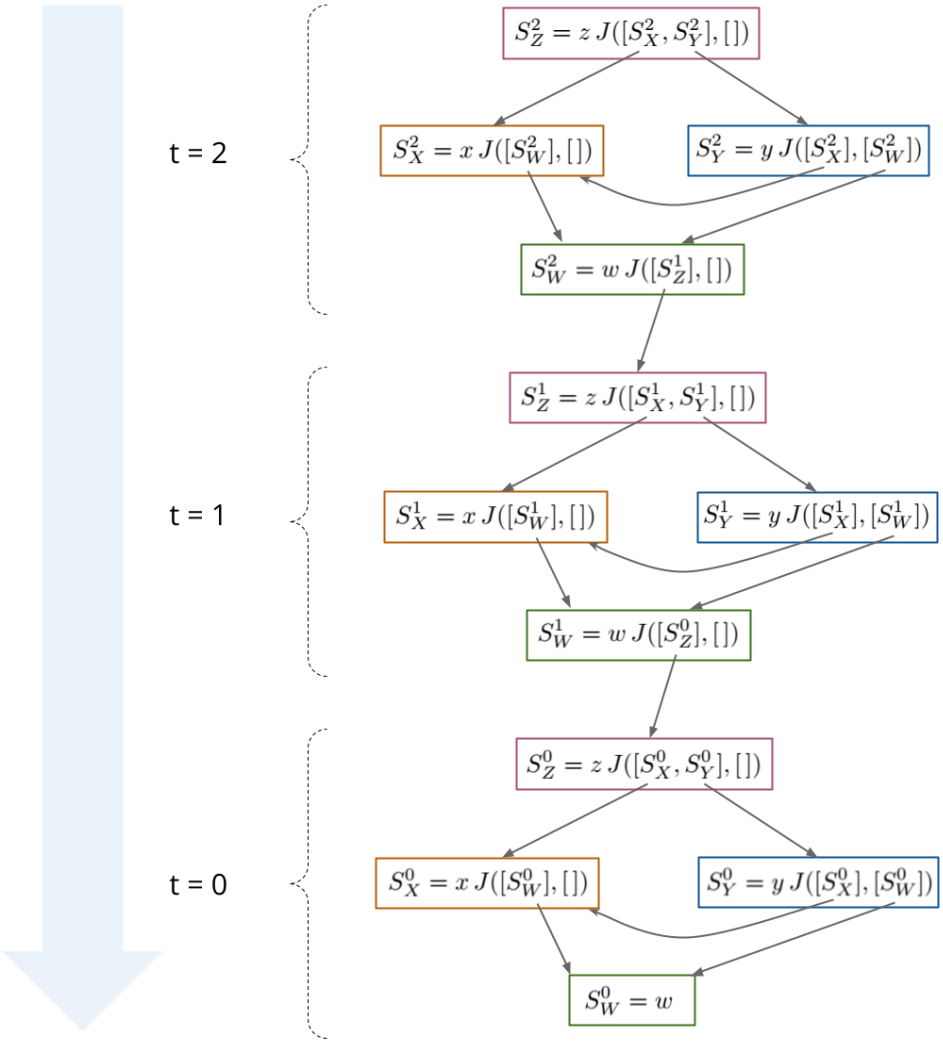


Figure S1: Illustrative example to represent the strategy followed to obtain an equation equivalent to the execution of *Hipathia* during 3 cycles.

4 Simulations

Simulating biological functions allows us to evaluate in detail the properties of the proposed hierarchical factorisation model. In this case, the design of the simulations will attempt to mimic the hierarchical structure that genes and pathways show in cells functioning, thus providing the ideal framework for assessing the benefits of a hierarchical model as opposed to an independent factorisation in both levels.

The first step in the simulation is focused on defining the latent components at the pathway level. To this end, it is necessary to choose the set of m_p cascades that will regulate the biological function to be simulated, extracted from the whole set of signalling cascades available in *Hipathia*. Then, the m_g genes included in those selected cascades are collected, thus obtaining the final set of genes and cascades included in the simulation.

In order to define the $W_p \in \mathbb{R}^{m_p \times k_p}$ matrix, containing the k_p latent components at pathway level, we start initialising a random matrix at gene level ($W_g^0 \in \mathbf{R}^{m_g X k_p}$) obtained as $W_g^0 \sim Beta(10, 10)$. Then, the *Hipathia* function is applied to obtain de W_p matrix:

$$W_p = \hbar(W_g^0). \tag{7}$$

We then proceed to define the S matrix, which describes the hierarchical structure of both sets of components. To do this, for each gene level component we randomly assign the pathway level component to which is associated. After this, we proceed to optimise the gene level components, in order to provide the same *Hipathia* values as their associated pathway level components. To this end, we initialise the gene level components matrix with a Beta distribution ($W_g \sim Beta(10, 10)$), and apply the following objective function:

$$\min \left\| \hbar(Wg) - W_p S^T \right\|_f^2. \tag{8}$$

From this expression, we derive the following update rule for W_g:

$$W_g = W_g - \eta \left[\frac{\partial \hbar(W_g)}{\partial W_g} \hbar(W_g) - \frac{\partial \hbar(W_g)}{\partial W_g} W_p S^T \right], \tag{9}$$

where $\eta = 0.0001$ corresponds to the learning factor and $\frac{\partial \hbar(W_g)}{\partial W_g}$ to the partial derivative between genes and pathways defined in the hierarchical model description.

Once the component matrices (W_g and W_p) and their hierarchical relationships (S) have been defined, we proceed to optimise the corresponding mixing matrices (H_p and H_g). To do this, we define the following objective function:

$$\min \left\| \hbar(W_g H_g) - W_p H_p \right\|_f^2, \tag{10}$$

from which we derive the update rules:

$$H_g = H_g - \eta \left[W_g^T \frac{\partial \hbar(W_g H_g)}{\partial W_g H_g} \hbar(W_g H_g) - W_g^T \frac{\partial \hbar(W_g H_g)}{\partial W_g H_g} W_p H_p \right] \tag{11}$$

$$H_p = H_p - \eta \left[W_p^T W_p H_p - W_p^T \hbar(W_g H_g) \right]. \tag{12}$$

Finally, the obtained model matrices (W_g, W_p, H_g, H_p, and S) are stored to be used during the evaluation of the hierarchical model.

5 Characterising Relevant Genes

Although at the biological level the entire set of genes and cascades included in a given factorisation regulate the biological function under analysis, it does not necessarily imply that all elements were altered in the disease. Therefore, it is important to determine which genes are really relevant and which of them do not show a high degree of variability between the subtypes included in the cohort. In addition, we need to consider in which components the most relevant genes are acting, providing, in turn, a way to characterise each component based on the profile of its most relevant genes.

To determine if a given gene is relevant in a particular component, we have to identify whether the gene shows an extreme value compared to other components. To this end, we consider a gene as relevant in those components where it shows a value greater than 1.5 times the interquartile range of the distribution formed by all the component values.

In addition, we need to assess whether a small variation in a given gene could potentially provide strong differences in pathway space. This notion is based on the topological structure of molecular pathways, where some proteins with a central role are not allowed to show a high range of variability. To evaluate whether a given gene is relevant in the pathway space, we apply the function *Hipathia* k_g times to each component, setting the value of all included genes except the evaluated gene, which sequentially takes the value obtained through all components. In this case, the gene will be considered relevant if any signalling cascade in the original component shows an extreme value across the repeats.

6 Hierarchical model plot legend

The following figure graphically depicts the obtained internal composition (Ω) of a group of individuals in the context of a given biological process.

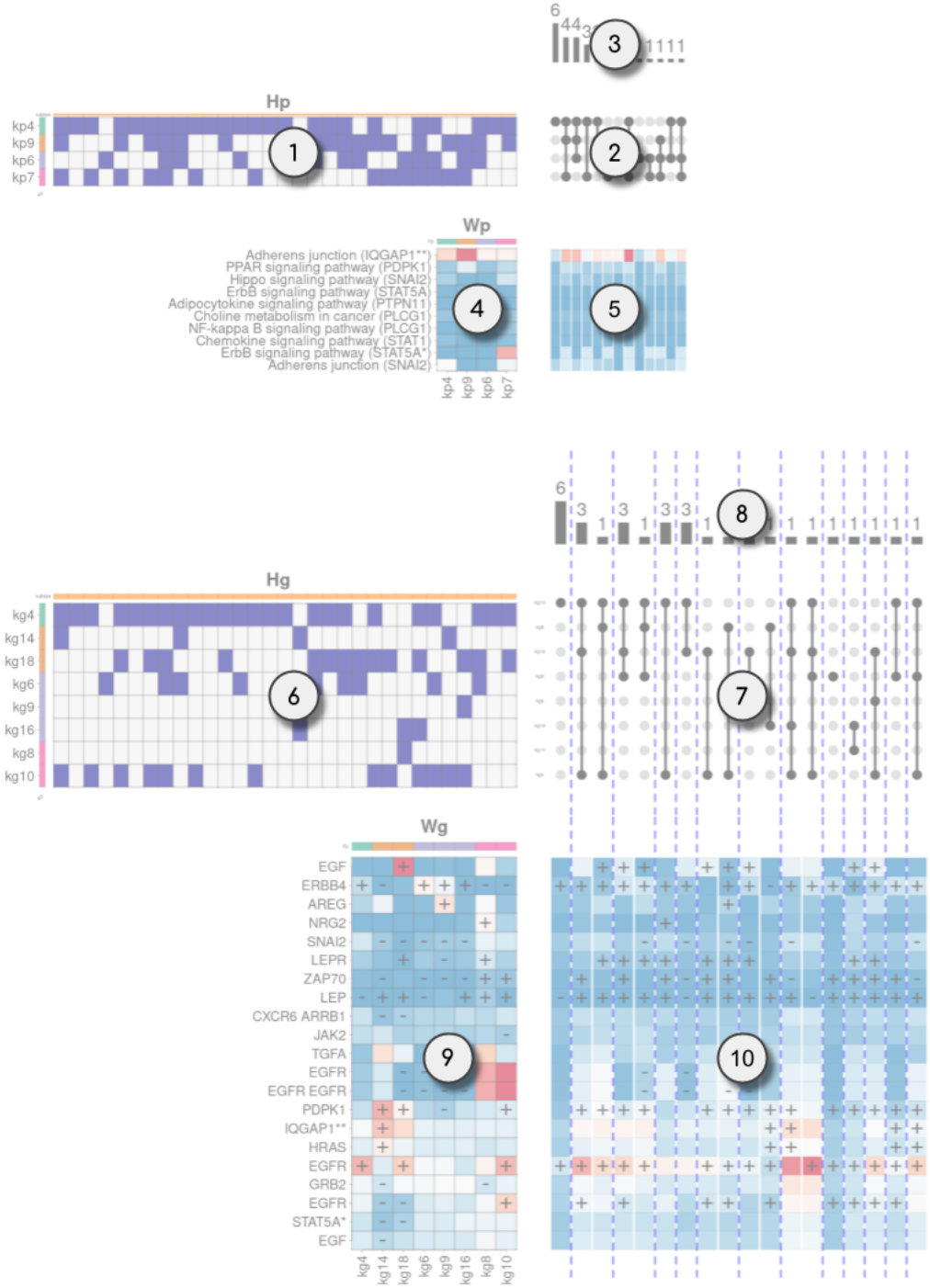


Figure S2: Graphical description of the internal composition (Ω) provided by the hierarchical model.

The figure is composed of the following panels:

1. Binarised pathway-level mixing matrix ($\overline{H_p}$);

2. Observed combinations of pathway-level components (C_p);
3. Sample frequencies of observed combinations of pathway-level components (φ_p);
4. Pathway-level component matrix (W_p);
5. Pathway-level meta-samples obtained from observed combinations (M_g);
6. Binarised gene-level mixing matrix ($\overline{H_g}$);
7. Observed combinations of gene-level components (C_g);
8. Sample frequencies of observed combinations of gene-level components (φ_g);
9. Gene-level component matrix (W_p);
10. Gene-level meta-samples obtained from observed combinations (M_p).

It is important to note that the W_g matrix only includes those genes that have been considered relevant in at least 1 component. Complementarily, the heatmap cells will show the label "+"/"-" to mark those genes that were considered relevant in a particular component.