1. **List of models available**

Currently in the GenePattern tool 10 models are available while it should be 8.

The models to remove are :

MT\_recon\_2\_2.mat

MT\_recon\_2.mat

MT\_recon\_1.mat

MT\_quek14.mat

MT\_inesMouseModel.mat

Furthermore, the name of the model when selecting the options should be different to avoid confusion of the user. See below the list of models and the name they should be associated with:

MT\_recon\_2\_2\_entrez.mat 🡺 Human\_recon\_2\_2

MT\_iHsa.mat 🡺 Human\_iHsa

MT\_iCHOv1\_final.mat 🡺 CHO\_iCHOv1

MT\_iRno.mat 🡺 Rat\_iRno

MT\_iMM1415.mat 🡺 Mouse\_iMM1415

1. **TO Modify – Description of the files to submit a job**

Currently, this is the description :

|  |  |
| --- | --- |
|  | “*A .mat or .csv expression matrix with rows corresponding to genes and columns corresponding to samples. Should contain a header row starting with "genes" and containing all of the sample names.”*  Introduce the following instead  “*Gene expression matrix can be provide in .mat or .csv.*  *For .mat files, you need to provide a structure variable (named e.g. data) containing a cell field named “genes” containing the NCBI Entrez gene ID (one gene by cell) and a double field named “value” containing a matrix with gene expression value for each of the sample you want to evaluate (i.e. with rows corresponding to genes and columns corresponding to samples). See available example dataTest.mat. (*attach a link for downloading the following document “*dataTest.mat*”)  *For csv. files, you need to provide an expression matrix with rows corresponding to genes and columns corresponding to samples. The first column is the list of genes NCBI Entrez gene ID and the first header row should start with “genes” and be followed by the name of each sample present in your dataset. See available example dataTest.csv.(*attach a link for downloading the following document “*dataTest.csv*”)  *The list of genes and associated NCBI Entrez ID present in each reference model can be downloaded here (*attach a link for downloading the following document “NCBI\_Entrez\_gene\_ID\_models.xls”)*”* |

1. **To add at the level of the “Home Page”**

* Ideally the logo plus a short description of the tool
* text to add

*CellFie : Cellular Functions InferencE*

*An alternative approach to capture the breadth of cellular functions by performing a functional analysis of existing biological networks.*

*We curated hundreds of metabolic functions from literature and defined them with regard to their metabolic pathways usage in existing genome-scale models of mammalian metabolism.*

*This platform can be used to comprehensively quantify the propensity of a cell line or tissue to be responsible for a metabolic function.*

*We welcome any comments, bug reports, and feature requests. Please send all feedback to* [*arichelleres@gmail.com*](mailto:arichelleres@gmail.com)

1. **To add at the level of the “results page” when a job is finished**

We need to provide explanations on the outputs generated by Cellfie. The gene pattern tool provides 5 outputs. Add the following description:

Stdout.txt - Text file containing the log information of the job submitted

taskInfo.csv - File containing the descriptive information about the 195 tasks assessed

score.csv - File containing the matrix of relative quantifications of the activity of the 195 metabolic tasks (rows) for each sample (columns) of the input data set. Note that if the results present metabolic tasks associated with a negative scores, it means that they cannot be calculated either due to the lack of gene expression data or due to the lack of gene information in the reference model used (see <https://doi.org/10.1371/journal.pcbi.1006867.s006> for the list of supported metabolic task for each reference model).

score\_binary.csv – File containing the binary version of the metabolic task score matrix.

Cellfieout.mat - Matlab files containing the information related to the tasks (i.e., taskInfos), and the two score matrices (score and score\_binary).

1. **Add a documentation page link (see text below) at the level of the homePage**

**Computation framework of the metabolic tasks**

We define a metabolic task as the capacity of producing a defined list of output products when only a defined list of input substrates is available. For each available reference genome-scale model, we computed the list of tasks the model successfully passed (i.e. list of functions the model is able to describe). If a task successfully passes in a model, we determined the set of reactions associated to this task. Therefore, based on the GPR rules, we can enumerate the genes that may contribute to the acquisition of this metabolic function.

We developed a computation framework for attributing a score to each metabolic task based on the availability of transcriptomic data. To this end, we first attribute a gene activity level (GAL) for each gene present in the model and for which expression data have been provided.

GAL=5\*log(1+ gene expression value / threshold)

We propose multiple options to define the threshold (see section *Threshold Definition* for more details).

We further used the GPR rule associated to each reaction required for a task to decide which gene will be the main determinant of the enzyme abundance associated to this reaction and attribute the corresponding gene activity level. Therefore, each reaction involved in a task is associated to a reaction activity level (RAL) that corresponds to the preprocessed gene expression value of the gene selected as the main determinant for this reaction.

We also computed the significance of each gene selected with regard of its overall use in the observed condition. Actually, some genes will be mapped to multiples reactions (e.g. promiscuous enzyme). Therefore, we assume that it may exist some competition between the reactions using this gene. We define the significance of a gene (S) by its specificity for a reaction by the inverse of the number reactions in which this gene is used as main determinant.

Finally, the metabolic score can be compute as the mean of the product of the activity level of each reaction with the significance of its associated gene:

MT score= sum(RAL\*S)/number of reactions involved in the task

MT score provides a relative quantification of the activity of a metabolic task in a specific condition based on the availability of data for multiple conditions. Indeed, it has been shown that some important housekeeping genes always present very low expression value. Therefore, a metabolic function that will completely rely on this set of gene will actually always present very low MT score. Contrarily, some tasks can be associated with gene presenting very high expression levels. Therefore, MT scores cannot be compared across tasks but only across samples. To partly overcome this problem, we also propose this scoring approach in its binary version to determine whether a metabolic task is active or not based on a gene expression profile. To this end, the MT score no longer take into account the significance of gene determinant for each reaction but is just computed as the mean of the reaction activity levels. Doing so, a metabolic task will be considered as active if its MT score in its binary version has a value superior to 5log(2).

**Threshold definition**

We propose two approaches to describe the scheme of threshold imposition on expression value for gene and/or reaction to be considered as “active”.

1. *Global approach*, the threshold value is the same for all the genes. The global approach is mainly applied when only one sample is available (i.e. sample could be associated to a condition, a cell-type or a tissue) and/or no information is available in the literature to define expression threshold for a single gene. The user can define the global threshold using a value or a percentile
   * *value* : the user provides the gene expression value for which a gene will be considered as active;
   * *percentile*: the threshold value will be defined based on a percentile (defined by the user) of the distribution of expression value for all the genes and across all samples of the user dataset
2. *Local approach,* the threshold value is different for all the genes. The local approach is often applied when multiple samples are available as it allows having a relative assessment of the activity of a gene across samples. Multiple options are available to compute the local thresholds:
   * *Mean* - the threshold for a gene is defined as the mean expression value of this gene across all the samples, tissues, or conditions (option only available if more than 3 samples are provided by the user).
   * *MinMaxMean* - the threshold for a gene is determined by the mean of expression values observed for that gene among all the samples, tissues, or conditions BUT the threshold :(i) must be higher or equal to a lower bound and (ii) must be lower or equal to an upper bound. The lower an upper bound can be defined using a value introduced by the user or based on a percentile of the distribution of expression value for all the genes and across all samples of your dataset. The *MinMaxMean* option ensures that genes with very low expression values across all the samples will never be considered as active and genes with very high expression across samples are always considered as active (option only available if more than 3 samples are provided by the user).

**List of available genome-scale models**

Human models

* Human\_recon\_2\_2 - Swainston, N. et al. (2016) Recon 2.2: from reconstruction to model of human metabolism. Metabolomics, 12, 109.
* Human\_iHsa : Blais et al. (2017) Reconciled rat and human metabolic networks for comparative toxigenomics and biomarker predictions. Nature Communications, 8:14250

Rat model

* Rat\_iRno : Blais et al. (2017) Reconciled rat and human metabolic networks for comparative toxigenomics and biomarker predictions. Nature Communications, 8:14250

Mouse model

* Mouse\_iMM1415 : Sigurdsson et al. (2010) A detailed genome-wide reconstruction of mouse metabolism based on human Recon1. BMC Systems Biology, 4:140.

CHO model

* CHO\_iCHOv : Hefzi et al. (2016) A consensus Genome-scale reconstruction of Chinese Hamster Ovary cell metabolism. Cell Systems, 3(5):434-443.

**Information about the curation of metabolic tasks**

The curation has been done by first taking the union of previously published lists of metabolic task (Recon2 and iHsa). We removed duplicated tasks or lumped the ones that were relying on the description of similar metabolic functions. Each remaining task for which we didn’t found significant biological evidence has been removed. We also created 9 new tasks that were essential for the acquisition of already described metabolic function (i.e. intermediary biosynthetic steps for the acquisition of a task). Doing so, we obtained a collection of 195 tasks associated to 7 systems (energy, nucleotide, carbohydrates, amino acid, lipid, vitamin & cofactor and glycan metabolism).