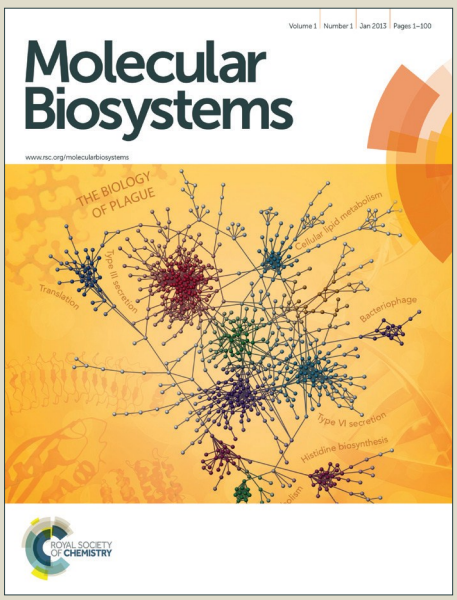
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Principal elementary modes analysis (PEMA)

Abel Folch-Fortuny[1,*], Rodolfo Marques[2], Inês A. Isidro[2], Rui Oliveira[2] and Alberto Ferrer[1]

[1]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, 46022 València, Spain.

[2]REQUIMTE/CQFB, Chemistry Department, FCT/ Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Abstract

Principal component analysis (PCA) has been widely applied in fluxomics to compress data into a few latent structures in order to simplify the identification of metabolic patterns. These latent structures lack a direct biological interpretation due to the intrinsic constraints associated to a PCA model. Here we introduce a new method that significantly improves the interpretability of the principal components with a direct link to metabolic pathways. This method, called Principal elementary modes analysis (PEMA), establishes a bridge between a PCA-like model, aimed at explaining the maximum variance in flux data, and the set of elementary modes (EMs) of a metabolic network. It provides an easy way to identify metabolic patterns in large fluxomics data sets in terms of the simplest pathways of the organism metabolism. The results using a real metabolic model of *Escherichia coli* show the ability of PEMA to identify the EMs that generated the different simulated flux distributions. Actual flux data of *E. coli* and *Pichia pastoris* cultures confirm the results observed in the simulated study, providing a biologically meaningful model to explain flux data of both organisms in terms of the EMs activation. The PEMA toolbox is freely available for non-commercial purposes on http://mseg.webs.upv.es.

* *E-mail address: abfolfor@upv.es*

1 Introduction

Principal component analysis (PCA) is one of the most applied statistical methods in Systems Biology. Its ability to compress large amounts of data, combining different kinds of variables, allows distinguishing between biologically relevant information and noise. This information is contained in a set of new variables built by PCA, the so-called principal components (PCs). In the context of fluxomics, PCA has been widely applied[1-3] with two main goals: (i) identify which parts of the metabolism retain the main variability in flux data and (ii) relate them to the behaviour of the organism, *e.g.* substrates consumption and protein production. This way, the PCs identify subsets of reactions based on the correlation structure of the flux data.

However, in the context of fluxomics PCA has some limitations. It is difficult to drive the PCs into a biologically meaningful solution, since PCA is a hard modelling method. For example, the main active pathways in a metabolic network could not be orthogonal, so PCA would be unable to describe them accurately in their PCs. To overcome these problems Multivariate Curve Resolution - Alternating Least Squares algorithm[4] (MCR-ALS) has been proposed to improve the biological interpretation of the components[5]. This method permits the incorporation of constraints, such as non-negativity and selectivity, when building the components. Finally, as with PCA, different sets of reactions or pathways emerge as the driving forces guiding the fluxes in the metabolic network.

Here we propose a new method to improve the interpretability of the components extracted by PCA and MCR-ALS, using the topology of the network to obtain the biologically relevant pathways in the model. This method is called Principal elementary modes analysis (PEMA). Its main advantage, over the previous methods, is that instead of building artificial components based solely on the correlation structure of the data (and some *a priori* knowledge in the case of MCR-ALS), the components are selected from the complete set of elementary modes (EMs) of the metabolic network. The EMs are the simplest representations of pathways across a metabolic network. Basically, each EM connects substrates with end-products concatenating reactions in a thermodynamically feasible way. The EMs analysis of a metabolic network allows extracting meaningful

information of a fluxome data set, since a given metabolic state can be represented as a linear combination of a specific subset of EMs. The PEMA algorithm is designed to identify the most relevant set of active EMs in flux data, using a strategy akin to PCA in dimensionality reduction.

Some methods have been proposed in the literature to select a set of representative or active EMs. One such attempt is the concept of the α−spectrum[6], which involves a linear optimization to determine how the extreme pathways (a systemically independent subset of EMs) contribute to a given steady-state flux distribution. This algorithm allows the determination of maximum and minimum possible weightings for each extreme pathway. A different approach involves the quadratic decomposition of a single steady-state flux into a set of EMs[7]. In this algorithm, a particular set of EMs is chosen, based on the minimization of the weighting vector length. A reinterpretation of this methodology was also proposed by projecting the flux space into the yield space[8], thus restricting the search for active EMs in a bounded convex space. The PEMA algorithm is quite different from the previous approaches. On the one hand, since PEMA is considering the whole set of EMs, instead of only the extreme pathways, the flux data can be interpreted with fewer pathways. On the other hand, PEMA finds the common set of active EMs in several flux distributions, reducing substantially the number of pathways needed to explain a complete flux data set.

2 Methods

2.1 Principal Component Analysis (PCA)

PCA is a multivariate projection method aimed at finding the underlying patterns of data that represent their main features[9]. The projection is achieved defining new variables, the so-called principal components (PCs), which are built as linear combinations of the original variables, exploiting the correlations among them. The PCA model equation is:

$$\mathbf{X}=\mathbf{T}\cdot\mathbf{P}^{T}+\mathbf{F} \quad (1)$$

where **X** is the original data set, **T** is the score matrix, containing the new uncorrelated

variables (PCs), **P** is the loading matrix, which contains the coefficients for the linear combinations of the original variables, and **F** is the error matrix. The number of components extracted by PCA is usually assessed taking into account the eigenvalues of the decomposition and the cumulative explained variance of the components[10].

2.2 Elementary modes

The concept of elementary mode (EM) is key for the analysis of metabolic networks. The set of EMs arises from the stoichiometric matrix, and each EM is defined as a minimal set of cellular reactions able to operate at steady-state, with each reaction weighted by the relative flux they need to carry for the mode to function[11]. The EMs are usually organized in a data matrix, **EM**, having the EMs by columns, the reactions in the metabolic network by rows, and the relative fluxes in its entries.

The set of EMs is obtained from convex analysis[12] and it is unique for a given metabolic network. Since this set represents a convex basis, any particular steady-state flux distribution can be obtained as a non-negative linear combination of EMs. Current algorithms for the computation of EMs face a common problem when dealing with highly interconnected metabolic networks[13]. In such cases, the combinatorial explosion of the number of EMs renders the analysis of large networks difficult. Very recently, two new methods[14,15] have been proposed to compute the EMs in large networks in a fast and efficient way.

2.3 Principal elementary modes analysis (PEMA)

PEMA is proposed with the aim of improving the interpretability of the PCA results. This way PEMA uses the set of EMs as the candidates for the PCs. Let **X** be a flux data set with N observations or experiments and K fluxes. The PEMA model is as follows:

$$\mathbf{X}=\mathbf{\Lambda}\cdot\mathbf{PEM}^{T}+\mathbf{F} \qquad (2)$$

where **PEM** is the $K\times E$ principal elementary modes matrix, formed by a subset of E EMs from the entire **EM** matrix; $\mathbf{\Lambda}$ is the $N\times E$ weightings matrix; and **F** is the $N\times K$ residual matrix. It is worth noting that the values in $\mathbf{\Lambda}$ are forced to be positive, since from a

network-based point of view, each possible steady-state flux distribution can be expressed as a non-negative combination of EMs[16].

In PEMA algorithm, the PEMs are chosen from the complete set of EMs in a step-wise fashion. The weightings associated to the PEMs are obtained by solving Equation 2:

$$\mathbf{\Lambda}=\mathbf{X}\cdot\mathbf{PEM}\cdot(\mathbf{PEM}^{T}\cdot\mathbf{PEM})^{-1} \quad (3)$$

Unlike the loadings in PCA, the PEMs are not orthonormal, so Equation 3 usually requires the computation of the pseudo-inverse of $\mathbf{PEM}^{T}\cdot\mathbf{PEM}$.

The first step of PEMA consists of calculating the weightings for each EM. So, initially, **PEM** and $\mathbf{\Lambda}$ are column vectors. Then the explained variance by each EM is obtained as follows[10]:

$$EV = 100\% \cdot (\|\mathbf{X}\|^2 - \|\mathbf{F}\|^2)/\|\mathbf{X}\|^2 \quad (4)$$

The EMs are sorted by *EV*, and the EM explaining most variance becomes the first PEM, with its associated $\mathbf{\Lambda}$ values. Afterwards, the variance explained jointly by the first PEM and each of the rest of EMs is calculated, and the pairs of EMs are sorted again by *EV*. The EM explaining more variance (jointly with the first PEM) becomes the second PEM, with their corresponding new $\mathbf{\Lambda}$ values. This procedure is iterated until reaching the maximum number of EMs. Since the weightings are recalculated for the 1st-*i*th PEMs when the (*i*+1)th PEM is computed, the amount of variance explained by the current set of PEMs is maximum.

When the PEMs are extracted step-wise, selecting the EMs explaining most variance at each step, the greedy solution is obtained. This is the usual procedure in PCA. The loadings are built in such a way that they explain as much variance in data as possible, and additionally, the resulting loadings are orthonormal. However, with PEMA, the EMs are not orthonormal (neither orthogonal). Therefore, the greedy solution may not be the best subset of EMs for explaining the data, since the choice of the first PEM influences the variance in data that the following PEMs could explain.

Two tuning parameters are introduced in the algorithm to cope with the previous problem. The greedy selection of the EMs is improved using a relaxation parameter R. This parameter makes the algorithm considers the best R EMs for the current PEM, and based on the variance explained extracting more PEMs, the best EM from the set of R is selected. This relaxation step can be done for several consecutive selections of PEMs. The branch point number, B, marks up to which PEM the relaxed selection is performed. Figure 1 shows an example of how the tuning parameters affect the selection of EMs. For instance, with R=3 and B=2, if one PEM is selected in the PEMA model it will be EM_1, since it is the EM explaining most variance; if two PEMs are selected it is possible that EM_1 and any of its 2nd PEM candidates (EM_6, EM_{11}, or EM_{19}) explain less variance that, for example, EM_4 and EM_8, so these last two will be the EMs selected in the PEMA model with two PEMs, and so on. The greedy approach accumulates the selected PEMs, but with R>1 the EMs may change completely from one PEM to the next one, in order to explain more variance with a fixed number of PEMs.

The number of PEMs evaluations, *i.e.* the number of times that the algorithm solves Equation 4 for all EMs, can be calculated using R and B. Let M be the maximum number of PEMs to be extracted by PEMA. Then, the number of evaluations, A, has the following expression:

$$A = \sum_{i=1}^{M} R^i + (M - B) \cdot R^B \tag{5}$$

where A grows exponentially with the number of branch points B. This way, the computation time required for each possible pair (R, B) can be estimated using Equation 5 and the computation time of the greedy approach (R=B=1 and A_{greedy}=M).

PEMA is an heuristic approach to solve the problem *which EMs do reconstruct the flux data?* The mathematical formulation of this problem consists of minimizing the 2-norm of $\mathbf{X}\text{-}\mathbf{\Lambda}\cdot\mathbf{PEM}^T$ subject to $\mathbf{PEM} \subseteq \mathbf{EM}$. The problem with this formulation is that it represents a mixed integer nonlinear programming problem, and since the number of fluxes and EMs may be extremely high, it is justified the application of an heuristic algorithm to find a

suboptimal solution to this problem. The proposed problem could be solved using genetic algorithms, however, different models have to be fit in order to get solutions with different number of PEMs. As well, the solution may change drastically depending on the initial points and the genetic operator chosen. This kind of algorithms improve an objective function, which can be the explained variance as in PEMA, but at some steps of the algorithm the search within the feasible space is performed in a random fashion, while PEMA focuses at each step in selecting the EMs explaining most variance. In this way, a single run of PEMA presents several solutions with a different number of PEMs.

2.4 Data preprocessing

PCA aims at explaining the main variability in data using a few PCs. If the original variables have strongly different means and/or variances, the PCs may focus on explaining only the variables with the highest values and/or variances, disregarding the small variance associated to the rest of variables.

PEMA has the same problem as PCA, so the flux data has to be preprocessed. While in PCA it is relatively easy to scale and mean center the original data, in PEMA, since the EMs are fixed, this is a subtle issue. To maintain the biological meaning of the EMs, if **X** is scaled column-wise by their standard deviations, the EM matrix has to be modified scaling row-wise all the EMs by the same values. The scaling of the **X** and **EM** matrices gives, initially, equal importance to all fluxes in the data, since their variances are equal to 1. This preprocessing is always recommended, since the variance of external fluxes can be exponentially greater than internal fluxes.

The mean centering of the PEMA model must not be done. When the data matrix **X** is mean centered, irreversible reactions would take negative fluxes thus the directionality of the fluxes is lost. In this way, if **X** is mean centered the PEMs are no longer able to fit the flux data. One way to overcome the mean centering problem is fitting additional PEMA models excluding the variables with the highest means. Once computed, the global and the local models can be compared in terms of EMs activation and reaction usage, to assess whether the global model is accounting for the fluxes with small values.

184 *2.5 PEMA algorithm*

185 The PEMA algorithm consists of the following steps:

186 1. Scale column-wise the original flux data **X** by their standard deviations.

187 2. Scale row-wise the elementary modes matrix, **EM**, using the standard deviations of the
188 original data set.

189 3. Choose the number of relaxations (R) and branch points (B).

190 4. Obtain the different PEMA models with 1 PEM, 2 PEMs, ..., M PEMs, solving Equation
191 2.

192 5. Select the number of EMs based on the aim of the study.

193 6. Recalculate the weightings **Λ** and the explained variance with the original flux data
194 (without scaling).

195 Practitioners should start with the greedy approach (R=B=1) and then, using the prediction
196 of the computation time, select different configurations to compare the models. To span the
197 different solutions that PEMA produces when changing the parameters, users are
198 encouraged to follow the configurations presented in section 3.1 (see also Table 1). For
199 large datasets, *e.g.* genome-scale networks with millions of EMs, the computation of the
200 greedy solution may take several hours. To avoid this long computation time, users can pre-
201 select a subset of relevant EMs prior to applying PEMA.

202 Also, the number of PEMs selected in each model, as in PCA, depends on the aim of the
203 study[17,18]. In this way, the scree plot (see next section) may help to select the EMs
204 explaining most variance in the flux data.

205

206 # 3 Results

207 *3.1 Escherichia coli simulated study*

A simulated study is proposed here to validate the performance of PEMA. The study consists of simulating different flux data sets, using several subsets of elementary modes (EMs), in order to assess if PEMA algorithm is capable of identifying them. The metabolic model of *Escherichia coli*, presented in reference [19], is used for this purpose (see Figure 2). The set of reactions can be found *online* in the Supplementary Materials section. The set of 255 EMs from the metabolic network of *E. coli* are obtained using EFMTOOL[20].

The simulated study is as follows: 100 different data sets are generated using from 1 to 10 EMs selected at random from the EM matrix. Ten different configurations of PEMA are applied on the present data, varying the values of the relaxations and branches *R*-*B*: 1-1, 5-1, 10-1, 20-1, 2-2, 5-2, 10-2, 3-3, 5-3, 4-4. The configurations are sorted in increasing computation time.

The identifiability of each PEMA configuration can be assessed computing how many times the complete set of EMs that generated the simulated flux data is identified. This information is presented in Table 1. As expected, for a fixed value of *B*, the higher is *R* the better tends to be the solution. Also, the more branch points are considered the more sets of EMs tend to be completely identified.

Even though not all the EMs are identified when the number of generating ones increases, all PEMA configurations are able to detect a subset of them. The precision and recall of the EMs identifications are shown in Figure 3. The high precision implies that most of the EMs identified are true ones, and also the high recall implies that the method identified most of the original EMs. With the exception of the greedy approach, all PEMA configurations are able to identify 80-100% of the original 3-4 EMs. The most complex configurations, *i.e.* when *B*=3 or *B*=4, maintain this level of accuracy with 5-6 generating EMs.

It is also interesting to check the mean number of PEMs identified by the different configurations and the percentage of explained variance. Since there exists a high degree of redundancy in any EM matrix, different linear combinations of EMs can represent a given flux distribution. This is clearly seen in Figure 4. Up to 5-6 generating EMs, the most complex PEMA configurations identify the same number of PEMs, matching the original ones (see Figure 4a). From 7 generating EMs onwards, the average number of PEMs grows

slower, identifying between 7 and 8 PEMs on average with 10 generating EMs. However, the percentage of explained variance by these PEMs remains very high, more than 99% having 7-10 generating EMs (see Figure 4b). The reduction in the number of EMs might also be due to the fact that some of the randomly selected EMs, with a random weighting on the model, have a small contribution to the variance in comparison to the EMs with greater coefficients. A table with the minimum, mean and maximum values for Figure 4a and the standard deviations for Figure 4b) can be found in Supplementary Materials.

3.2 E. coli real data

The flux data of *E. coli* presented in reference [19] is used in this section to check the performance of PEMA with real data. Each observation in this dataset describes a flux distribution in *E. coli*, after a specifically targeted gene knock-out. The metabolic network and EMs set considered here are the same as in the simulated study (see Figure 2). The flux data matrix, **X** considered in this paper has 21 observations (rows) and 42 fluxes (columns). In these 21 observations, a subset of the original 32 observations, the same set of reactions is considered. The flux data set can be found *online.*

Based on the results of the simulated study, the tuning parameters R and B are both set to 4, to obtain more accurate results. The computation time of PEMA in this case is 2 minutes (2.9 GHz Intel Core I7, 8GB RAM 1600 MHz), while the computation time of the greedy approach is less than a second. Figure 5a shows the cumulative scree plot of the PEMs. This kind of plot is usually employed in PCA to assess the appropriate number of principal components. Here, 8 PEMs are selected: EM_{125}, EM_{167}, EM_{254}, EM_{27}, EM_{235}, EM_{16}, EM_{143} and EM_{145}, explaining 97.8% of variance with the scaled data, and 99.4% of the real variance. As opposed to PCA, in PEMA the PEMs are usually explaining common sources of variability. This can be seen in Figure 5b, where the direct sum of all variances explained by the PEMs is 150%. For instance, EM_{125} explains more than 80% of variance in data, but this variance is shared with other PEMs. Nevertheless, the PEMs explaining most variance can be considered the most relevant in the model.

The degree of orthogonality of the PEMs can be obtained by dividing the variance explained by the model (99.4%) by the sum of the explained variances of each PEM. Here,

the degree of orthogonality is 66.3%, which implies that the solution obtained by the PEMA is strongly non-orthogonal and, therefore, quite different from the PCA one.

To assess if some observation is not well modelled the percentage of explained variance per observation can be computed (see Figure 6a). Also the observed versus predicted plot can be used to visualise the differences at a data point level (see Figure 6b). In the present case, the percentage of explained variance is 97-99% for all observations, and the predicted values lay close to the true ones.

The PEMA model can be easily interpreted using an adaptation of the classical PCA loadings and scores plot. This way, Figures 7-8 shows the principal elementary modes plot and the weightings plot, respectively. The PEMs plot shows which reactions are active for a specific EM, while the weightings plot represents the contribution weight of each PEM on each observation (*i.e.* knock-out). A first look at the selected PEMs shows that the whole set captures the formation of all metabolic requirements for cell synthesis, that is, reactions 31-41 (see Figure 7). EM_{125} is the PEM explaining most variance in data (see Figure 5b). This pathway depicts the glucose flux into glycolysis and TCA, without any exchange fluxes for cell synthesis metabolites. This leads to a high rate of NADH production, which generally is used to synthesize ATP. For this, EM125 can be interpreted as the cell's catabolic pathway, while the remainder PEMs capture the fluxes for cell synthesis metabolites, thus representing anabolic pathways leading to synthesis of biomass.

Since EM_{125} is related to the catabolism, it has a strong weight in each knock-out (see Figure 8). Nevertheless, some observations seem to have a greater impact in this PEM than others, in particular the knock-outs 2, 3, 10, 14, 15 and 16, representing the genes *glk*, *pgm*, *gpmB*, *rpiB*, *tktB* and *talB*. The *pgm* gene codifies the phosphoglucomutase that converts G6P into G1P and its deletion would likely direct the carbon flux to glycolysis or the pentose phosphate pathway. The genes *rpiB*, *tktB* and *talB*, also scoring a high weight, are related to pentose phosphate reactions.

The EMs related to anabolic metabolism represent all the remaining exchange fluxes that produce the cell synthesis metabolites. EMs 16, 27, 143, 145 and 167 connect glucose directly to the pentose phosphate pathway, which is fundamental in the metabolism, since it

generates NADPH, a reduced equivalent important in biosynthetic processes[21]. Moreover, EM_{16} and EM_{167} are responsible for balancing the metabolic fluxes towards E4P and R5P, being the sole PEMs that predict the fluxes of these metabolites to cell synthesis. With a few exceptions, the knock-out experiments have similar weight values inside each anabolic PEM. These exceptions are the observations 1, 5, 8, 12 and 14, representing the knockouts *galM*, *pfkB*, *gapC*, *pykF* and *rpiB*. This group of genes has low weightings in EM_{254} and EM_{235}, meaning that these flux modes have a minor impact in the metabolism of these mutants, that is, a lower flux in the synthesis of Pyr, 3-PG, 2-KG and OAA for biomass synthesis. Conversely higher weightings from these mutants are observed for EM_{27} and EM_{16}, that is, in the production of E4P, PEP and G6P. Another curious aspect of EM_{16} and EM_{27} is the activation of the glyoxylate bypass. This pathway is known to be active in low glucose concentrations[22], but repressed when glucose becomes available in higher concentrations[23,24]. The observations 18 to 21 reflect *E. coli* wild-type cultured at a dilution rate of 0.2 h^{-1}, used as control experiments. In these observations, positive fluxes for the gyoxylate pathway were registered, possibly due to a low glucose feed to the culture.

Finally, all the PEMs have a zero coefficient for fermentative pathways (reactions 28-30), therefore these fluxes are not being represented by the model. However, looking at the original data, all the observations have zero values for fluxes 28 and 29. Regarding flux 30, few observations (4 out of 21) have a non-zero value for it. For the latter case, since PEMA, as PCA, aims at explaining the covariance between the original variables using the PEMs, if most of the values in a variable are 0 it is difficult for PEMA to identify the EM generating these slight differences. The extraction of more PEMs may correct that, however, the risk of overfitting is higher and the model would become less parsimonious.

3.3 Pichia pastoris real data

A second real case study is analysed here: a fluxome for the growth of recombinant *P. pastoris*. This data set was based on a statistical design of experiments to test the effects of culture media factors in the flux data. The media composition was prepared according to the Invitrogen's guidelines for *P. pastoris* fermentation, and consists mainly on mineral salts. 26 shake flask experiments were performed with variations on 11 media factors

selected for statistical design. Glycerol was used as carbon source in every experiment.

The metabolic network for the central carbon metabolism of *P. pastoris* used here is largely based on the network proposed in reference [25], with adaptations from other central carbon[26] and genome-scale networks[27]. The network consists of 43 metabolic reactions, 34 internal metabolites and 10 exchange reactions (see Figure 9). The main catabolic reactions are represented in this network, namely glycolysis and gluconeogenesis pathways, the tricarboxylic acid cycle (TCA), the pentose-phosphate pathway, anaplerotic, fermentative and phosphorylative oxidation pathways. A biomass formation reaction is also included in the model, from selected internal metabolites based on *P. pastoris* cells macromolecular compositon[28]. There exist 158 EMs in the metabolic model. The flux data set and the elementary modes matrix can be found *online* in the Supplementary Material section.

The results of PEMA with this data set are the same using either the greedy approach and the most complex approach presented here (R=B=4), which takes 35 seconds. This indicates that the results are stable against the different PEMA configurations. 99.5% of the scaled data is explained using 3 PEMs, with a degree of orthogonality of 70% (*i.e.* the variance explained by the 3 PEMs sums 141%). As in the previous real case study, this implies that PCA cannot obtain these results using orthogonal components. The cumulative scree plot and the variance explained by each PEM are shown in Figure 10.

All scenarios are being represented by the selected EMs, as can be seen in the explained variance per observation plot (see Figure 11a); and the observed versus predicted plot (see Figure 11b) shows an even better fitting than with *E. coli*, which could be due to different levels of noise in the flux data sets.

Figure 12 shows the PEMs and weightings plots. The PEMs identified are EM_{147}, EM_{10} and EM_{149}. The binary version of the PEM plot appears in the Supplementary Material section. The binary version of the weightings plot is not included, since all observations use all PEMs.

The first PEM consumes glycerol (reactions 35 and 29) and crosses half of the glycolytic pathway (reactions 4-7) to activate the TCA cycle (reactions 15, 17-20), clearly

representing the cell's catabolism. EM10 uses also reactions 35, 29 and 4-7 to activate the TCA cycle, but in this case reaction 16 is used instead of 17. It also activates the pentose phosphate pathway (reactions 8-13), leading to the synthesis of redox equivalents (NADPH), but also precursor metabolites for the synthesis of biomass. For this reason, this PEM groups the reactions for the cell's anabolism. At the end, this is the PEM responsible of the biomass production in all observations. The last PEM assimilates glycerol in the same way as EM_{147} and afterwards focuses on the production of ethanol (reactions 25 and 39). The occurrence of ethanol synthesis during aerobic respiration in yeast is a common feature (Crabtree effect). Nonetheless, unlike most yeasts, *P. pastoris* does not typically exhibit a significant ethanol production, favouring the aerobic metabolism. This fact is well captured by the relative lower explained variance of EM_{149} in comparison to EM_{147} (see Figures 10 and 12b).

Finally, as expected, no EM related to methanol assimilation (reactions 30-32 and 26) and final products such as pyruvate or citrate (reactions 41 and 42, respectively) is selected, since all fluxes are 0 for these reactions.

4 Discussion

In this paper a new method called principal elementary modes analysis (PEMA) is presented with the aim of improving the interpretability of a traditional PCA modelling in fluxomics. PEMA builds a PCA-like model using the complete set of elementary modes (EMs) in order to identify which ones, the PEMs, are the driving forces generating the flux distributions.

The simulated study on *E. coli* shows the high identifiability of PEMA. The most complex PEMA configurations are able to detect completely 1-4 generating EMs and, a high percentage of them, up to 6-7 EMs. Even though not all the EMs are identified by PEMA, the method provides always a parsimonious solution explaining more than 99% of variance. The analysis of actual flux data of the same organism confirms the tendency shown with the simulated fluxes. 8 PEMs are identified explaining 99.4% of variance in the flux data.

This way, most of the PEMs identified are describing the glucose consumption, the glycolytic pathway and the TCA cycle, but afterwards, each of them has a different function in the cell synthesis. The results obtained with *P. pastoris* are coherent with *E. coli*'s. In this case 3 PEMs are selected describing accurately the metabolic pathways being activated when glycerol is used as main carbon source in aerobic conditions.

A significant number of graphical tools, all of them integrated in the PEMA toolbox, are provided in this paper. The cumulative scree plot, the observed versus predicted plot, and the variance explained per observation plot can be used to decide the number of PEMs to extract. The plot showing the variance explained by each PEM and the PEMs and weightings plots are useful to exploit the PEMA model in terms of relevance and biological interpretation of the PEMs, and their activation among the observations.

Additionally, the theoretical estimation of the runs of PEMA algorithm when the tuning parameters change permits to establish a relatively accurate upper bound of the computation time, based on the greedy approach solution. This allows designing wisely a set of trials to compare the results of the different configurations of PEMA.

5 Conclusion

In this work, PEMA is developed to explain the inherent variability on a fluxomics dataset, while preserving biological meaning. This can be regarded as an exploratory technique that allows researchers to interpret a data set by uncovering the most representative pathways operating in a cell.

There is a potential use of this methodology in bioprocess engineering applications, such as the development of structured metabolic models in cell culture fermentations. PEMA can be useful in the identification of a specific set of EMs that explains variations in cellular metabolic rates under certain operational conditions, such as temperature and pH. This would allow the improvement of the process kinetics' modelling by the incorporation of biological knowledge from the cellular system.

Author's contributions

AF-F, RM and IAI performed the analyses. AF-F and RM wrote the manuscript. AF-F wrote the code of the PEMA toolbox. RO and AF conceived the study and reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Research in this study was partially supported by the Spanish Ministry of Economy and Competitiveness and FEDER funds from the European Union through grants DPI2011-28112-C04-02 and DPI2014-55276-C5-1R. We would also acknowledge Fundacão para a Ciência e Tecnologia for PhD fellowships with references SFRH/BD/67033/2009, SFRH/BD/70768/2010 and PTDC/BBB-BSS/2800/2012.

References

1. B. Sariyar, S. Perk, U. Akman and A. Hortasu, *Journal of Theoretical Biology*, 2006, **242**(2), 389– 400.

2. C. Barrett, M. Herrgard, and B.O. Palsson, *BMC Systems Biology*, 2009, **3**, 30.

3. J.M. González-Martínez, A. Folch-Fortuny, F. Llaneras, M. Tortajada, J. Picó and A. Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2014, **134**, 89–99.

4. J. Jaumot, R. A. de Juan, and R. Tauler, *Chemometrics and Intelligent Laboratory Systems*, 2015, **140**, 1-12.

5. A. Folch-Fortuny, M. Tortajada, J.M. Prats-Montalbán, F. Llaneras, J. Picó and A. Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2015, **146**, 293–303.

6. S.J. Wiback, R. Mahadevan and B.O. Palsson, *Journal of Theoretical Biology*, 2003, **3**, 313-324.

7. J.-M. Schwartz and M. Kanehisa, *Bioinformatics*, 2005, **21**, 204-205.

8. H.-S. Song and D. Ramkrishna, *Biotechnology and Bioengineering*, 2009, **2**, 554-568.

9. J.E. Jackson, *A User's Guide to Principal Components*. Wiley Series in Probability and Statistics, Wiley Online Library, 2004.

10. R. Bro and A.K. Smilde, *Analytical Methods*, 2014, **6**(9), 2812–2831.

11. S. Schuster, D.A. Fell and T. Dandekar, *Nature Biotechnology*, 2000, **18**(3), 326–332.

12. S. Schuster, C. Hilgetag, J.H. Woods and D.A. Fell, *Journal of Mathematical Biology*, 2002, **45**, 153-181.

13. S. Klamt, and J. Stelling, *Molecular Biology Reports*, 2002, **29**(1-2), 233–236.

14. L.-E. Quek and L.K. Nielsen, *BMC Systems Biology*, 2014, **8**(1), 94.

15. M.B. Badsha, R. Tsuboi and H. Kurata, *Biochemical Engineering Journal*, 2014, **90**, 121–130.

16. F. Llaneras and J. Picó, *Journal of Bioscience and Bioengineering*, 2008, **105**(1), 1–11.

17. J. Camacho and A. Ferrer, *Journal of Chemometrics*, 2012, **26**(7), 361–373.

18. J. Camacho and A. Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2014, **131**, 37–50.

19. N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, T. Nakayama, T. Nishioka, K. Shimizu, H. Mori and M. Tomita, *Science*, 2007, **316**(5824), 593–597.

20. M. Terzer and J. Stelling, *Bioinformatics*, 2009, **19**, 2229-2235

21. M. Madigan, J. Martinko and J. Parker, *Brock Biology of Microorganisms*, Pearson Education, Inc., New Jersey, 2003.

22. A. Nanchen, A. Schicker, O. Revelles and U. Sauer, *Journal of Bacteriology*, 2008, **190**(7), 2323–2330.

458 23. A. Nanchen, A. Schicker and U. Sauer, *Applied and Environmental Microbiology*,
459 2006, **72**(2), 1164–1172.

460 24. R. Carlson and F. Srienc, *Biotechnology and Bioengineering*, 2004, **85**(1), 1–19.

461 25. M. Tortajada, F. Llaneras and J. Pico, *BMC Systems Biology*, 2010, **4**, 115.

462 26. K. Baumann, M. Carnicer, M. Dragosits, A.B. Graf, J. Stadlmann, P. Jouhten, H.
463 Maaheimo, B. Gasser, J. Albiol, D. Mattanovich and P. Ferrer, *BMC Systems Biology*,
464 2010, **4**(1), 141.

465 27. B.K. Chung, S. Selvarasu, A. Camattari, J. Ryu, H. Lee, J. Ahn, H. Lee and D.-Y. Lee,
466 *Microbial Cell Factories*, 2010, **9**(1), 50.

467 28. M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, M. Sauer, F.
468 Altmann, P. Ferrer and D. Mattanovich, *Journal of Proteome Research*, 2009, **8**(3), 1380–
469 1392.

470

Tables

Table 1. Complete identifications of the generating elementary modes.

Configuration	Number of generating elementary modes						
R-B	1	2	3	4	5	6	7-10
1-1	10/10	7/10	2/10	2/10	0/10	0/10	0/10
5-1	10/10	10/10	5/10	3/10	1/10	1/10	0/10
10-1	10/10	10/10	5/10	4/10	1/10	0/10	0/10
20-1	10/10	10/10	5/10	5/10	1/10	0/10	0/10
2-2	10/10	9/10	5/10	4/10	1/10	0/10	0/10
5-2	10/10	10/10	5/10	2/10	1/10	0/10	0/10
10-2	10/10	10/10	7/10	7/10	2/10	1/10	0/10
3-3	10/10	9/10	7/10	6/10	4/10	1/10	0/10
5-3	10/10	10/10	7/10	8/10	5/10	1/10	0/10
4-4	10/10	10/10	7/10	8/10	6/10	3/10	0/10

List of Figure captions

Figure 1. Relaxation (*R*) and branch point (*B*) parameters. When *B*=*R*=1 the EM explaining most variance is chosen and fixed at each step. If these parameters change, different subsets are considered for each PEM identification.

Figure 2. Metabolic network of *E. coli* considered for the present study.

Figure 3. Precision and recall of the different configurations. Precision is calculated by dividing the sum of the true identified EMs by the sum of the true identified plus the false identified ones. The recall is calculated by dividing the true identified EMs divided by the true ones plus the true non-identified ones.

Figure 4. a) Mean number of identified EMs. b) Mean percentage of explained variance.

Figure 5. a) PEMA Cumulative scree plot and b) Percentage of variance explained by each PEM in *E. coli* study: 8 PEMs are selected explaining 97.4% of variance in the scaled data.

Figure 6. a) Explained variance per observation and b) Observed versus predicted plot in *E. coli* study.

Figure 7. Principal elementary modes plot in *E. coli* study. The PEMs are represented by columns and the corresponding reactions by rows. Blue squares represent positive values, and dashed red squares the negative ones. The darker the colour, the more highly positive/negative is the value.

Figure 8. Weightings plot in *E. coli* study. The weightings of the PEMs are represented by columns and the observations by rows. The darker the colour, the more important is the PEM for the corresponding observation.

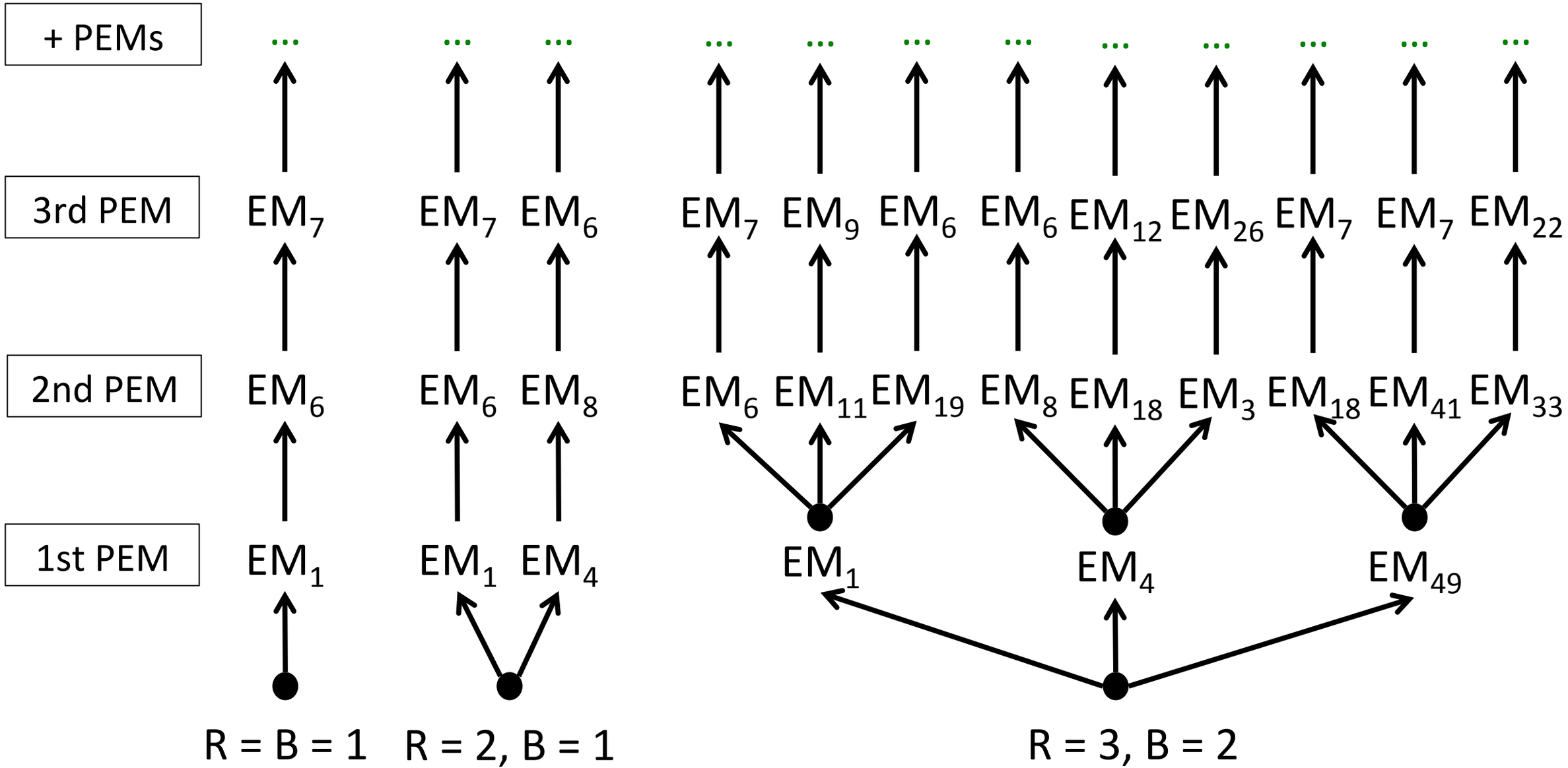
Figure 9. Metabolic network of *P. pastoris* considered for the real case study.

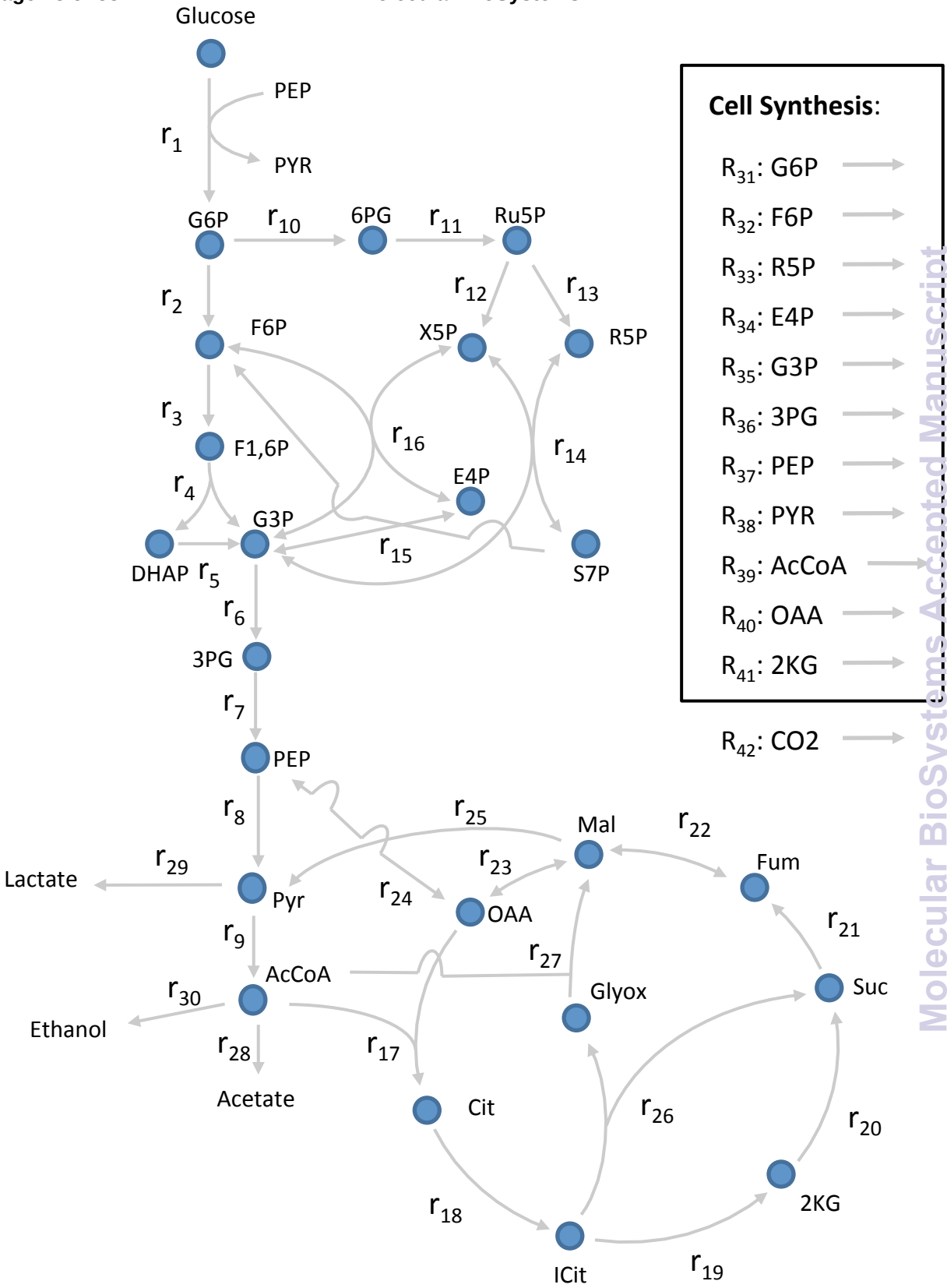
Figure 10. a) PEMA Cumulative scree plot and b) Percentage of variance explained by each PEM in *P. pastoris* study: 3 PEMs are selected explaining 99.5% of variance in the

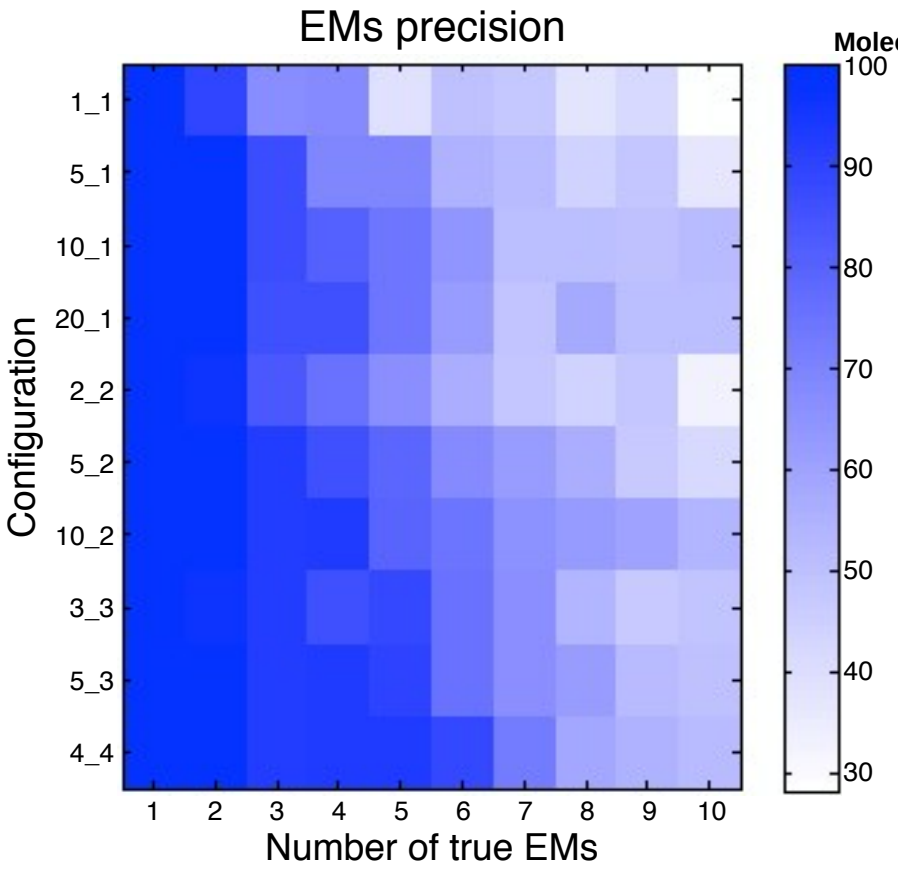
scaled data.

Figure 11. a) Explained variance per observation and b) Observed versus predicted plot in *P. pastoris* study.

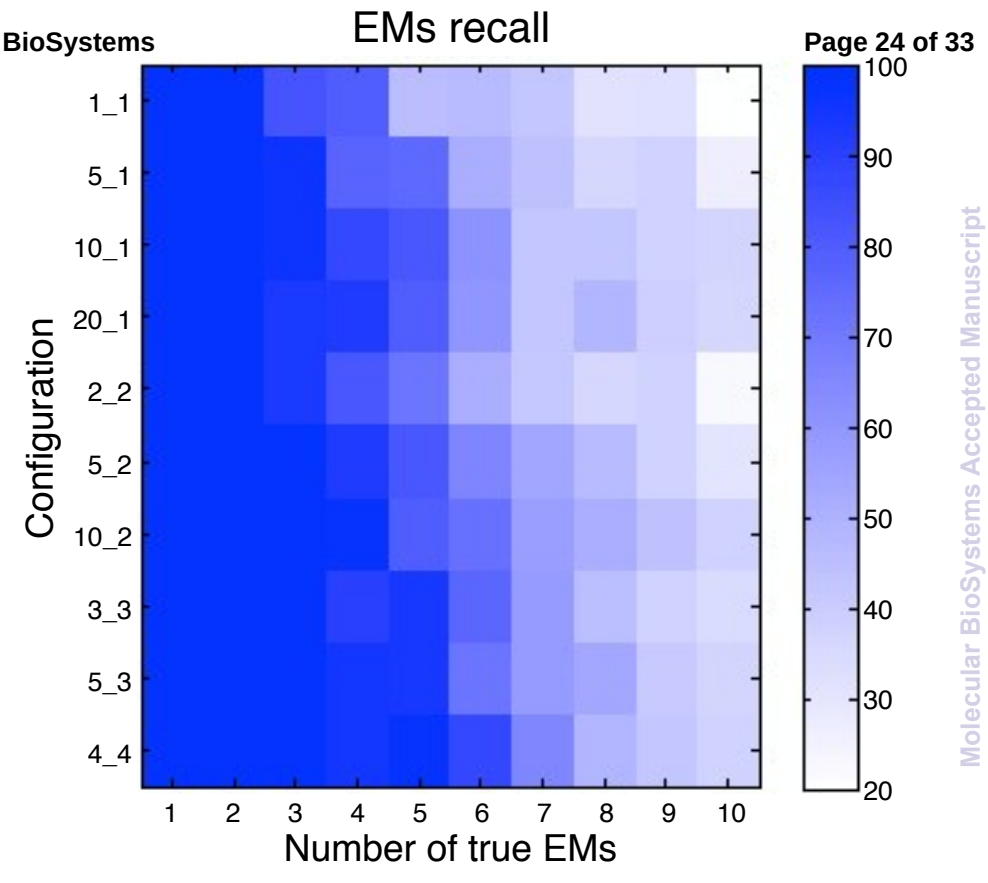
Figure 12. Principal elementary modes and weightings plots in *P. pastoris* study. The PEMs are represented by columns in both plots; reactions and observations appear row-wise in each plot, respectively. Blue squares represent positive values, and dashed red ones the negatives. The darker the colour, the more highly positive/negative is the value.



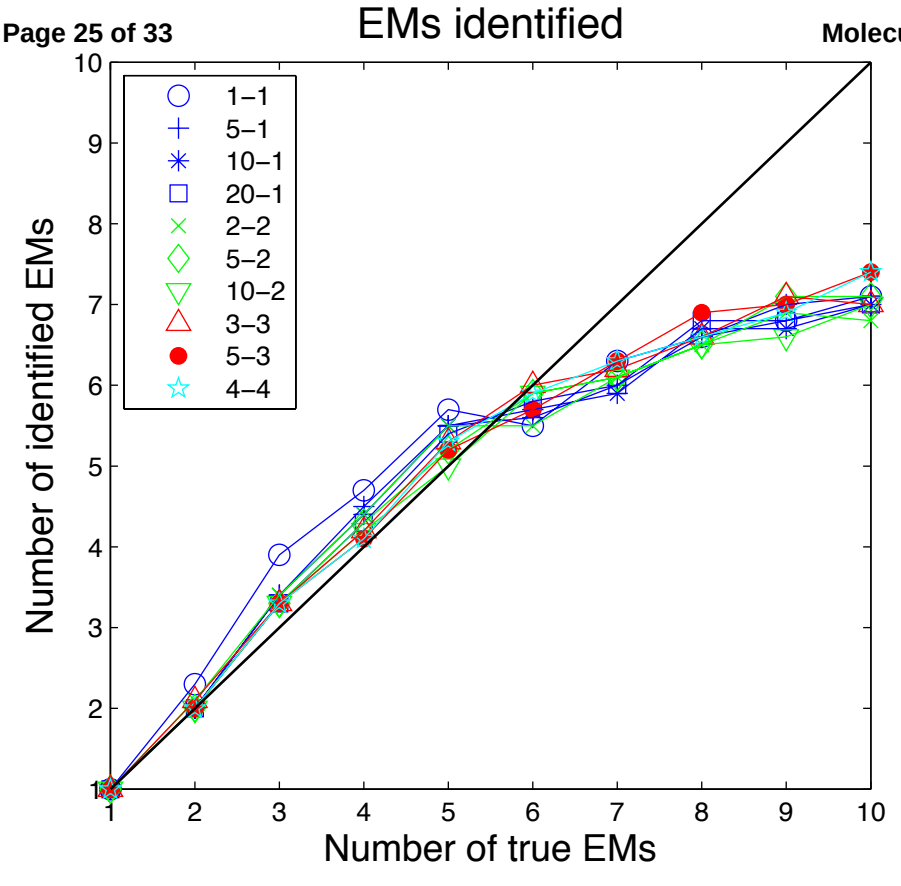




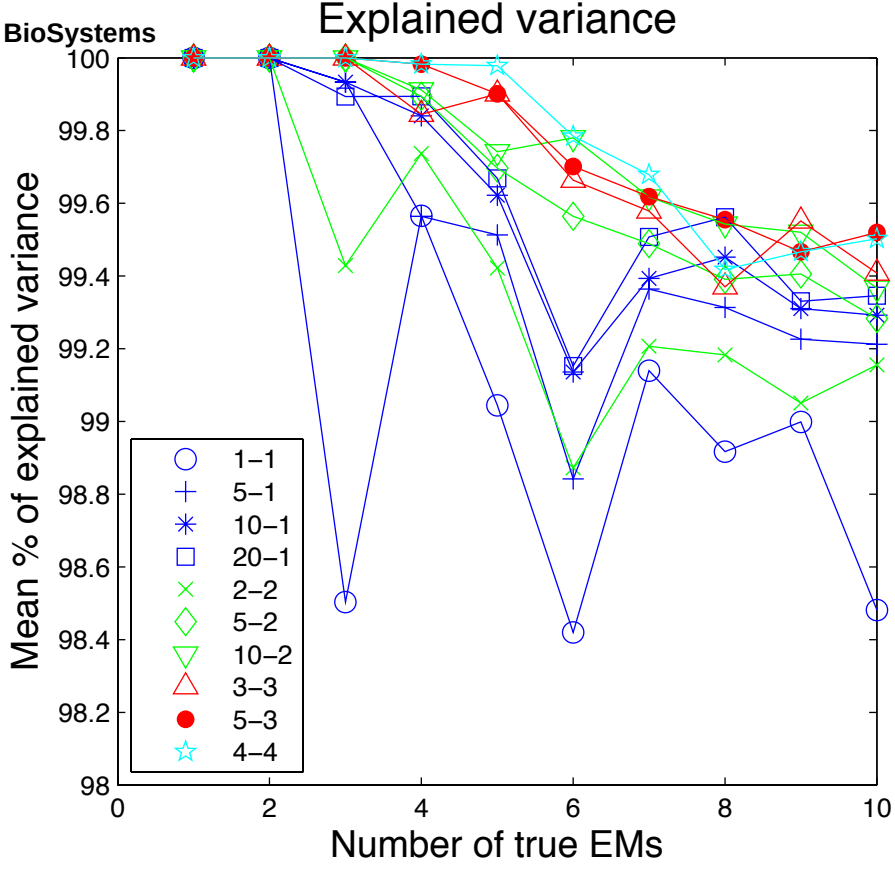
a)



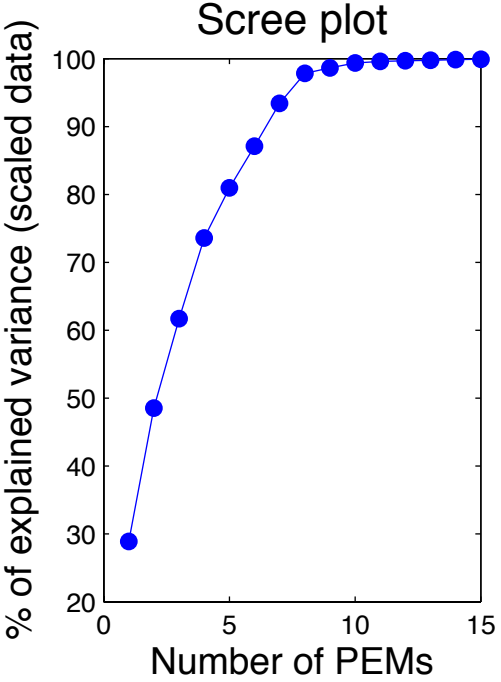
b)



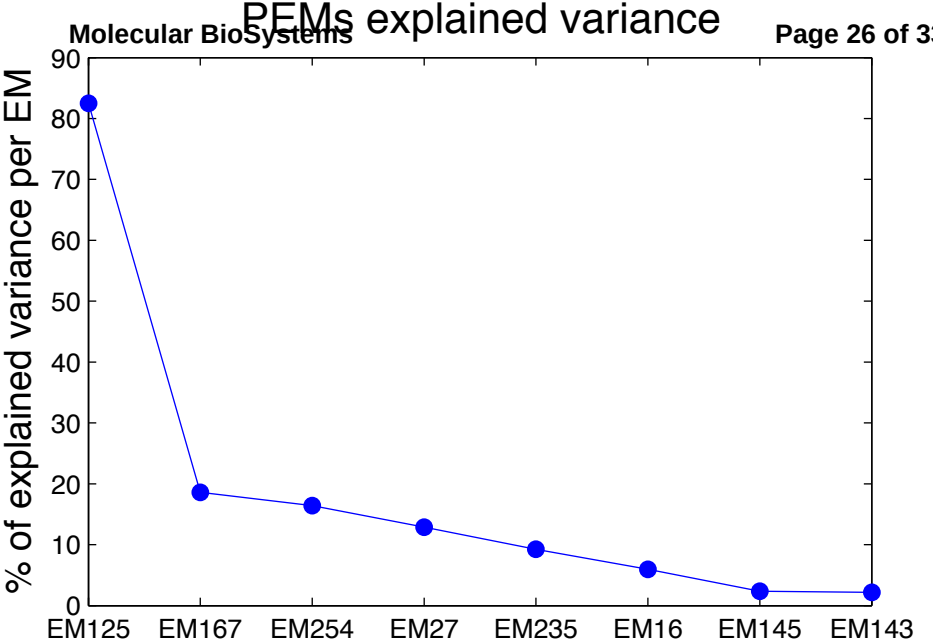
a)



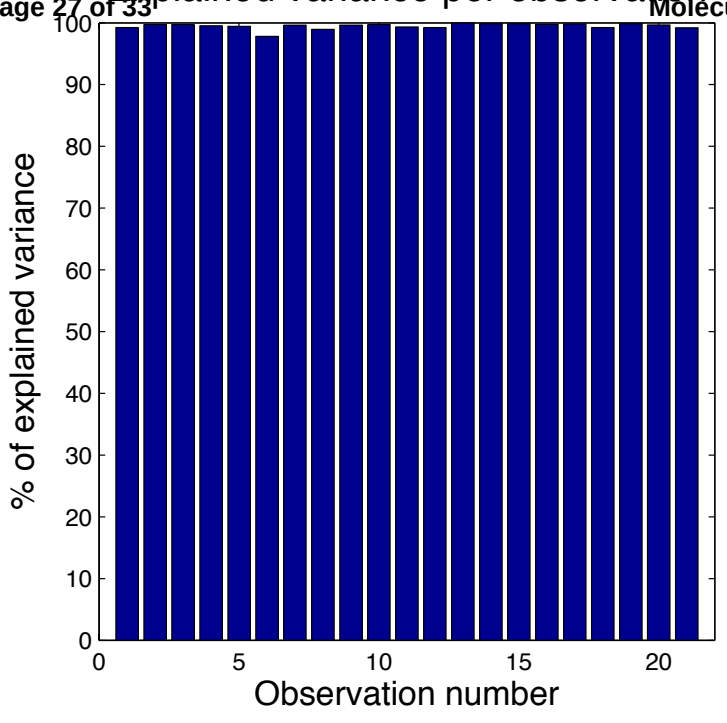
b)



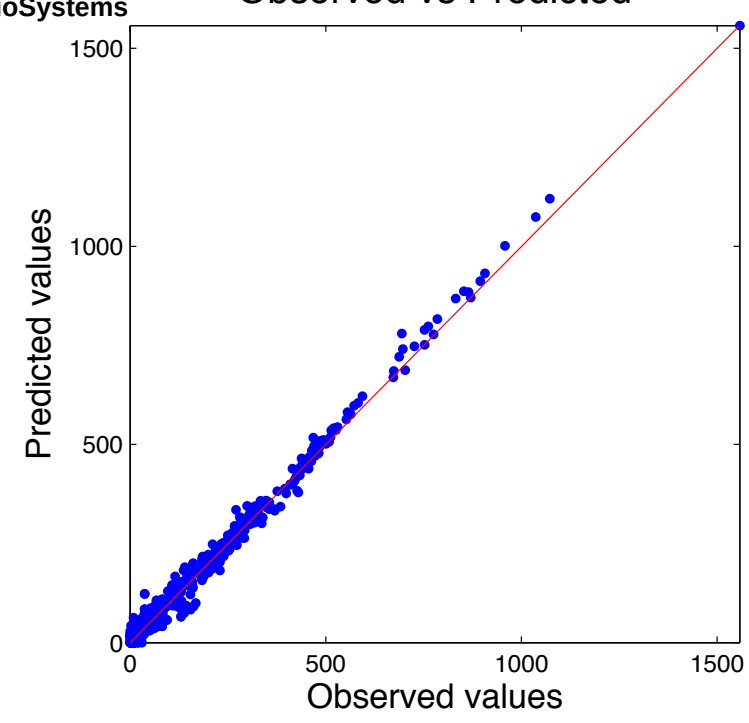
a)



b)



a)



b)

Principal elementary modes plot