Projection to latent pathways (PLP): a constrained projection to latent variables (PLS) method for elementary flux modes discrimination

Ana R Ferreira1,2†, João ML Dias†1, Ana P Teixeira2,3†, Nuno Carinhas2,3†, Rui MC Portela†1, Inês A Isidro†1, Moritz von Stosch4† and Rui Oliveira1,2*†

Abstract

Background: Elementary flux modes (EFM) are unique and non-decomposable sets of metabolic reactions able to operate coherently in steady-state. A metabolic network has in general a very high number of EFM reflecting the typical functional redundancy of biological systems. However, most of these EFM are either thermodynamically unfeasible or inactive at pre-set environmental conditions.

Results: Here we present a new algorithm that discriminates the "active" set of EFM on the basis of dynamic envirome data. The algorithm merges together two well-known methods: projection to latent structures (PLS) and EFM analysis, and is therefore termed projection to latent pathways (PLP). PLP has two concomitant goals: (1) maximisation of correlation between EFM weighting factors and measured envirome data and (2) minimisation of redundancy by eliminating EFM with low correlation with the envirome.

Conclusions: Overall, our results demonstrate that PLP slightly outperforms PLS in terms of predictive power. But more importantly, PLP is able to discriminate the subset of EFM with highest correlation with the envirome, thus providing in-depth knowledge of how the environment controls core cellular functions. This offers a significant advantage over PLS since its abstract structure cannot be associated with the underlying biological structure.

Background

An elementary flux mode (EFM) can be defined as a minimal set of enzymes able to operate at steady state, with the enzymes weighted by the relative flux they need to carry for the mode to function [1]. The universe of EFM of a given metabolic network define the full set of non-decomposable steady-state flux distributions that the network can support. Any particular steady-state flux distribution can be expressed as a non-negative linear combination of EFM. Motivated by these unique properties, EFM analysis has become a widespread technique for systems level metabolic pathways analysis [1-8].

The number of EFM of a metabolic network is in general very high, denoting the innate adaptability and robustness of biological systems. As illustrative example, the central carbon metabolism of a genome-scale reconstructed Escherichia coli metabolic network has approximately 26 million EFM [9]. However, not all of these pathways are thermodynamically feasible or even physiologically reachable [10]. Over the last decade several methods were proposed to reduce the number of EFM founded on different principles (Table 1).

Some of the proposed methods reduce EFM based solely on structural information of the metabolic network. de Figueiredo et al. [11] presented a method to enumerate the EFM in increasing order of number of reactions. This approach enabled to identify the K-shortest EFM in Escherichia coli and Corynebacterium glutamicum metabolic networks, which are in principle energetically more efficient. Song and Ramkrishna [12] proposed a reduction algorithm based on the effect of EFM on the convex hull volume. This allowed the a priori reduction, without any experimental data, from the initial 369 to 35 EFM for a yeast metabolic network fermenting both glucose and xylose.
EFM can also be discriminated on the basis of reaction thermodynamics. Wlaschin et al. [13] demonstrated with experimentally determined intracellular fluxes that EFM weights are inversely correlated with the entropy generated by the involved metabolic reactions. This suggests that evolution induced cellular regulatory patterns to favor efficient pathways with low entropy generation. Zhao et al. [14] demonstrated with the computational algorithm to reduce EFM based on the correlation of EFM weighting factors with measured environment factors, which we call projection to latent pathways (PLP). The underlying principles are: (i) only a moderate number of EFM are active at given environmental conditions, (ii) the environment plays a critical role in their regulation, and (iii) active EFM deliver a characteristic environmental footprint that can be used for their identification. In what follows we present all mathematical details underlying PLP and compare it with PLS in relation to a case study.

### Results

#### Projection to Latent Pathways (PLP) Algorithm

**Problem statement**

By applying steady-state material balance equations to a metabolic network with $m$ metabolites and $q$ metabolic reactions, the following system of linear algebraic equations is obtained:

\[ \mathbf{N} \cdot \mathbf{r} = 0 \]  
\[ \mathbf{r}_k > 0 \]

with $\mathbf{r}$ a vector of $q$ metabolic fluxes, $\mathbf{r}_k$ the subset of fluxes associated to irreversible reactions and $\mathbf{N}$ a $m \times q$ stoichiometric matrix. It is a well-known property of system (1) that its null space solution takes the form of a polyhedral cone [21]. Furthermore, the convex basis of

### Table 1 Classification of methods for EFM reduction

| Principle                        | Method                                      | Data required                      | References |
|----------------------------------|---------------------------------------------|------------------------------------|------------|
| Network connectivity and stoichiometry | K-shortest EFM: Enumerates the EFM in increasing order of number of reactions. | Parameter free | [11]       |
|                                  | Yield Analysis: Excludes EFM with negligible contribution to convex hull in yield space. |                      | [12]       |
| Thermodynamics                   | Fractional contributions of EFM: Estimates the EFM Coefficients based on calculated EFM thermodynamic properties. | Thermodynamic data | [13]       |
|                                  | Maximum Entropy Principle: Calculates the EFM Coefficient by maximizing Shannon's entropy, which is an indirect measure of system complexity. |                      | [14]       |
| (Non)linear programming          | $\alpha$-spectrum: Uses linear optimization to maximize and minimize the weightings of each metabolic pathway that produces steady state flux distributions. | 'omics' data can be used to shrink the $\alpha$-spectrum. | [15,16,38]|
|                                  | Flux regulation coefficients: Estimates the EFM coefficients that optimize a given performance function (e.g. minimum error in flux or yield prediction). | Fluxomics and possibly other omic datasets | [17]       |
|                                  | Quadratic program: Calculates the weights for a large set of EFM by using quadratic program to reconstruct flux distributions from subsets of EFM. |                      |            |
| Enzyme kinetics                  | Quantitative elementary mode analysis of metabolic pathways: Combines structural and kinetic modelling to assess the effect of changes in enzyme kinetics on the usage of EFM. | Enzyme kinetic parameters | [19]       |
system (1) is formed by a large number of base vectors, which are the EFM studied in this paper:

\[
\mathbf{r} = \sum_{i=1}^{n_{\text{em}}} \lambda_i \cdot \mathbf{e}_{mi}
\]

with \( \mathbf{e}_{mi} \) a \( q \times 1 \) vector of reaction weighting factors that defines EFM \( i \) and \( \lambda_i \) a scalar variable defining the partial contribution of \( \mathbf{e}_{mi} \) to the overall flux phenotype, \( \mathbf{r} \), and \( n_{\text{em}} \) the number of EFM.

In this paper we study the reduction of EFM on the basis of dynamical envirome data sets. The basic premise is that measured fluxome vectors can be systematically deconvoluted into genetic dependent factors (the structure of EFM, \( \mathbf{e}_{mi} \)) and envirome dependent factors (the partial contribution of each EFM to flux phenotype, \( \lambda_i \)). To implement this method, we developed a discrimination algorithm that works according to the following criteria:

1. Maximisation of explained variance of flux data sets, \( \mathbf{r} = \{ \mathbf{r}(t) \} \)
2. Maximisation of correlation of \( \lambda_i \) against envirome data, \( \mathbf{X} = \{ \mathbf{x}(t) \} \)
3. Minimisation of the number of active EFM

with \( \mathbf{X} = \{ \mathbf{x}(t) \} \) a \( np \times nx \) matrix of \( np \) independent observations of envirome vectors \( \mathbf{x}(t) \) (dim(\( \mathbf{x} \)) = \( nx \)), \( \mathbf{R} = \{ \mathbf{r}(t) \} \) a \( np \times nr \) matrix of \( np \) independent observations of reaction rates, \( \mathbf{r}(t) \) (dim(\( \mathbf{r} \)) = \( q \)). These criteria are equivalent to a covariance maximisation problem (covariance maximisation implies maximisation of correlation and minimisation of redundancy) between envirome data, \( \mathbf{X} \), and observed flux data, \( \mathbf{R} \), under the constraint of a plausible set of EFM:

\[
\begin{align*}
\text{Maximize} & \quad \text{cov} (\mathbf{X}, \mathbf{R}) \\
\text{s.t.} & \quad \mathbf{R} = \mathbf{A} \times \mathbf{E} \mathbf{M}^T \\
& \quad \mathbf{A} = \mathbf{X} \times \mathbf{C}^T
\end{align*}
\]

with \( \mathbf{E} \mathbf{M} = \{ \mathbf{e}_{mi} \} \) a \( nr \times nem \) matrix of nem EFM, \( \mathbf{e}_{mi} \) (dim(\( \mathbf{e}_{mi} \)) = \( q \)), \( \mathbf{A} = \{ \mathbf{a}(t) \} \) a \( np \times nem \) matrix of weight vectors \( \mathbf{a}(t) \) of EFM (dim(\( \mathbf{a} \)) = \( nem \)) and \( \mathbf{C} \) a \( nem \times nx \) matrix of regression coefficients.

Unconstrained maximisation of covariance can be performed by the widely used method projection to latent structures (PLS), also known as partial least squares. Figure 1 shows the structural differences between PLS and PLP. Since PLP is derived from PLS, in the lines below we first review PLS decomposition and then show how it can be extended to PLP.

**Projection to Latent structures (PLS)**

PLS is a multivariate linear regression technique between an input (predictor) matrix, \( \mathbf{X} \), and an output response matrix, \( \mathbf{Y} \). It differs from traditional multivariate linear regression in that it decomposes both the predictor and the response matrices into reduced sets of uncorrelated latent variables, which are then linearly regressed against each other.

The most widely used PLS algorithm is the NIPALS (non-iterative partial least squares) algorithm [22], which provides the basis for PLP derivation. NIPALS proceeds according to the following steps:

1. Set the initial \( ny \times 1 \) \( \mathbf{Y} \)-loading vector, \( \mathbf{q} \), equal to an arbitrarily chosen nonzero row of \( \mathbf{Y} \), \( \mathbf{y}_t \)

\[
\mathbf{q} = \frac{\mathbf{Y}_t^T}{\|\mathbf{Y}_t\|}
\]

in case of univariate PLS, \( ny = 1 \) and \( q = 1 \)

2. Compute the \( np \times 1 \) \( \mathbf{Y} \)-score vector, \( \mathbf{u} \)

\[
\mathbf{u} = \mathbf{Y} \cdot \mathbf{q}
\]

3. Compute the \( nx \times 1 \) weight vector, \( \mathbf{w} \)

\[
\mathbf{w} = \frac{\mathbf{X}^T \cdot \mathbf{u}}{\|\mathbf{X}^T \cdot \mathbf{u}\|}
\]

4. Compute the \( np \times 1 \) \( \mathbf{X} \)-score vector, \( \mathbf{t} \)

\[
\mathbf{t} = \mathbf{X} \cdot \mathbf{w}
\]

5. Recalculate the \( \mathbf{Y} \)-loading vector, \( \mathbf{q} \)

\[
\mathbf{q} = \frac{\mathbf{Y}_t^T \cdot \mathbf{t}}{\|\mathbf{Y}_t \cdot \mathbf{t}\|}
\]

6. Repeat steps 1-5 until the convergence criterion \( \|\mathbf{t} - \mathbf{t}_{\text{old}}\| < \epsilon \) is obeyed with, for instance, \( \epsilon = 1 \times 10^{-6} \). In case of univariate PLS, Eq. 8 yields \( q = 1 \) hence no iterations are performed.

7. Compute the \( \mathbf{X} \) data block loadings, \( \mathbf{p} \), and rescale accordingly:

\[
\mathbf{p} = \frac{\mathbf{X}^T \cdot \mathbf{t}}{\|\mathbf{t} \|}
\]

\[
\mathbf{p}_{\text{new}} = \frac{\mathbf{p}}{\|\mathbf{p}\|}
\]

\[
\mathbf{t} = \mathbf{t} \cdot \|\mathbf{p}\|
\]

\[
\mathbf{w} = \mathbf{w} \cdot \|\mathbf{p}\|
\]

8. Compute the regression coefficient of the inner linear model
\[ b = \frac{u^T \cdot t}{t^T \cdot t} \]  

9. Compute the X and Y residuals

\[ E_X = X - t \cdot p^T \]  
\[ E_Y = Y - b \cdot t \cdot p^T \]  

10. Then go back to step 1 and repeat the procedure for the next latent variable after making

\[ X = E_X \]  
\[ Y = E_Y \]  

Steps 1-10 are repeated for \( k = 1, \ldots, \) Fac latent variables resulting into the following overall decomposition:

\[ X = T \cdot W^T + E_X \]  
\[ Y = U \cdot Q^T + E_Y \]  
\[ U = T \cdot B^T + E_U \]  

with \( E_i \) residuals matrices. Finally, the prediction of \( Y \) from \( X \) is given by

\[ \hat{Y} = X \cdot RC^T \]  

with \( RC \) the \( ny \times nx \) regression coefficients matrix given by

\[ RC = Q \cdot B \cdot W^T \]  

For more details about PLS and NIPALS see Geladi and Kowalski [23].

**Projection to latent pathways (PLP)**

PLP can be viewed as a constrained version of PLS that maximises the covariance between \( X \) and \( R \) under the constraint of known EFM. PLP performs essentially the
same decomposition described by Eq. 18-22. The main difference lies in the computation of the output loadings, \( Q \). Since EFM are unique and non-decomposable flux solutions, any observed flux distribution can be expressed as a non-negative weighted sum of EFM (Eq. 2). Thus, EFM \( em \), can be interpreted as latent variables (or principle components of a metabolic network) while the weights \( \lambda_s \) can be interpreted as score values of such latent variables. According to this analogy, PLS was modified as follows:

1. For each EFM \( k \), set the loadings equal to \( em \) and compute the respective score vector, \( \lambda_k \):

\[
q_k = em_k
\]
\[
\lambda_k = R \cdot q_k (\equiv u_k)
\]

2. Perform a univariate PLS (with \( q = 1 \)) with input \( X \) and target \( Y = \lambda_k \) for \( Fac \) latent variables as described in the previous section and compute the predicted \( \lambda_k \):

\[
\hat{\lambda}_k: \text{predicted } \lambda_k \text{ from univariate PLS}
\]

3. Compute the predicted \( R \) by the \( k \) EFM and the respective explained variance:

\[
\hat{R}_k = \hat{\lambda}_k \cdot q_k^T
\]
\[
\text{var}_k(\%) = 100 \cdot \left( 1 - \frac{\sum_i (R - \hat{R}_k)^T \cdot (R - \hat{R}_k)}{\sum_i R^T \cdot R} \right)
\]

4. Repeat steps 1-3 for every EFM \( k = 1, ..., nem \) and choose the best, \( kopt \), as the one that exhibits the highest variance value given by Eq. 27.

\[
kopt: \text{EFM with highest } \text{var}_k \text{ value}
\]

5. Remove \( kopt \) from the list of EFM and make

\[
R = R - \hat{R}_{kopt}
\]

6. Go back to step 1 and repeat the procedure for a maximum number of EFM or until the explained variance of \( R \) does not increase any further.

With this procedure the output loadings, \( Q \), hold a subset of EFM from matrix \( EM \) while the output scores, \( U \), are equivalent to the EFM weights matrix, \( \Lambda \):

\[
R = \Lambda \times EM^T + E_R
\]

As such, while PLS and PLP are structurally equivalent, the loadings and scores in PLS are abstract variables while in PLP they have a physical interpretation:

1. The number of latent variables in PLS is analogous to the number of active EFM in PLP. Thus the subset of EFM that explain most of the variance of \( R \) are interpreted as the set of metabolic pathways activated by environmental factors.

2. The regression coefficients vector, \( RC_{kopt} \), of the inner univariate PLS, being directly associated with EFM, show the contribution of each environmental factor to the up- or down-regulation of EFM.

The PLS and PLP algorithms were coded in Matlab™ (Mathworks, Inc). The code is freely available for academic use under a free academic license and can be downloaded at http://www.dq.fct.unl.pt/sbegroup.

In what follows we compare both algorithms in relation to a case study.

**Case study: recombinant BHK cell line**

Data of a recombinant baby hamster kidney (BHK) cell line expressing a fusion glycoprotein IgG1-IL2 was used to compare PLS and PLP. The data set comprises 134 observations acquired from 7 independent bioreactor experiments operated in batch and fed-batch modes. The predictor matrix, \( X \) (dim(\( X \)) = 134 × 26), includes measured data of 26 environmental factors (pH, osmolarity and concentrations of viable cells, glucose, lactate, ammonia, IgG1-IL2 and 19 amino acids) while the target matrix, \( R \) (dim(\( R \)) = 134 × 24), comprises 24 production or consumption fluxes of extracellular compounds. Further details about the data can be found elsewhere ([20]).

A relatively small BHK metabolic network comprising 35 metabolites and 57 metabolic reactions was constructed. Its EFM were computed using Metatool 5.0 [24] resulting in 251 EFM. Details can be found as Additional Files 1 and 2. These 251 EFM were used as constraints to PLP decomposition.

**Comparing PLP and PLS decomposition results**

The full data set was divided into two partitions of randomly selected points with equal size for calibration and validation (with 67 points each). The results of a single run of PLS and PLP decomposition for the calibration data set are shown in Tables 2 and 3 respectively. PLS decomposition stops at latent variable 18, when the \( X \) variance reaches 100%. The final explained \( R \) variance is 90.1%. As for PLP, decomposition progresses up to the 17th EFM, explaining 82.5% of \( R \) variance, thus 7.5% less than PLS. PLP decomposition stops when the threshold degree of correlation between \( \lambda_s \) and \( X \) can no longer be satisfied (\( r^2 > 0.75 \) and p-value < 0.05, see Table 3). This procedure ensures that the identified EFM are the ones...
Assessment of EFM reduction consistency

PLS belongs to a class of multivariate regression techniques that can be used to model high dimensional data sets with low number of sampling points [25]. However, when the number of samples is too low, the partitioning into calibration and validation sets may have a high impact on the final model structure. Since stemming from PLS, the same problem does in principle apply to PLP. In order to assess EFM discrimination variability due to data partitioning, a bootstrapping technique was implemented, in which PLP and PLS were repeated 200 times with randomly selected calibration and validation partitions with 67 points each. Figure 3 shows the frequency of selection of EFM resulting from the bootstrapping analysis. The complete set of results is provided as Additional File 3. These results evidence a subset of frequently selected EFM, which include EFM1, EFM2, EFM4, EFM6, EFM11, EFM179 and EFM210 with frequency of selection higher than 75% and EFM69, EFM72, EFM173 and EFM206 with frequency of selection higher than 50%. Less frequently selected EFMs are very sensitive to the data partitioning and to experimental noise and thus less reliable to interpret.

Metabolic interpretability

As mentioned previously, while in PLS the output latent variables have no physical meaning, in PLP they are EFM. To illustrate this difference we plot in Figure 4 the output loadings of the first two PLS latent variable (Table 2) against the reaction weighting factors of the first two selected EFM (179 and 1) (Table 3).

It can be seen that the first PLS loadings vector $q_1$ calculated by Eq. 8 does show a residual correlation with the first selected EFM 179 structure ($r^2 = 0.26$). However, the second loadings vector $q_2$ shows no correlation at all with second select EFM 1. Despite the fact that both the calculation of the output loadings $q$ and the selection of EFM obey to the same criterion of maximization of the correlation between $X$ and $Y$, it is clear that the data structure identified by PLS cannot be easily associated with the underlying biological structure.

It is beyond the scope of this paper to present a detailed metabolic interpretation of the discriminated EFM by PLP (for a detailed analysis see [20]). Here we just highlight a few illustrative examples, the most frequently selected EFM for biomass synthesis is EFM179 followed by EFM173. The product formation EFM (EFM 1) is also frequently selected. The anaerobic conversion of glucose into lactate was also frequently selected (EFM 11). Serine transamination into glycine (EFM 6) was also among the most frequently selected EFM. EFM 4 corresponds to the glutaminolysis pathway, well known as a major carbon source for energy production in mammalian cells. In general, these are important pathways known to be active in mammalian cells.
Regression coefficients

While PLS regression coefficients are associated with latent variables lacking physical meaning, PLP regression coefficients are directly associated to the discriminated EFM (see Figure 5). Thus they provide information of how the envirome up- or down-regulates each EFM. This interpretation should however be done with care as regression coefficients cannot disclose between a cause and an effect. An EFM is per definition a non-decomposable sub-network. Most of them start and end in extra-cellular compounds. Each EFM produces a characteristic dynamic footprint in the environment in terms of consumed or produced metabolites, which is more an effect rather than a cause. Moreover, it is an important feature of PLS and per inheritance of PLP that the \(X\)-loadings are computed in a way to maximise predictive power of \(Y\) in detriment of interpretability of the individual contribution of \(X\) variables. Although many papers have attempted to develop interpretation of PLS regression coefficients (e.g. [26,27]), other techniques are in principle better suited for this purpose. Even so, main causal-effects can be extracted from the analysis of regression coefficients. For this analysis it is however important to calculate the confidence intervals of the regression coefficients, which can be obtained from the previously described bootstrapping technique [28]. From the \(z = 200\) PLP runs with randomly selected calibration and validation data sets, \(z = 200\) vectors of regression coefficients are calculated (see Additional File 3). The respective mean and standard deviation can be estimated as follows:

\[
\bar{B} = \frac{\sum_{i=1}^{z} B_i}{z} \quad (31)
\]

\[
S = \sqrt{\frac{1}{z-1} \sum_{i=1}^{z} (B_i - \bar{B})^2} \quad (32)
\]

The 95% confidence intervals can then be calculated from the \(t\)-student distribution with 0.975 half interval and \(z\)-Fac degrees of freedom

\[
B = \bar{B} \pm S \times t_{0.975,z-\text{Fac}} 
\]

As illustrative example, Figure 6 plots the confidence interval against the mean of the regression coefficients for the product formation EFM (EFM 1). It can be observed that only a subset of regression coefficients lay below the one half threshold line. These include the regression coefficients associated with pH, osmolality, glutamine, lactate, IgG, valine and lysine. These regression coefficients are the most statistically significant and thus more reliable interpretations can be withdrawn from them. As example, it is a rational result that the weighting factor of the product EFM 1 is highly correlated with

![Figure 2 Correlation between EFM weighting factors and envirome variables](http://www.biomedcentral.com/1752-0509/5/181)
the product concentration since the product results from EFM 1. All other identified environmental parameters are potential targets for manipulation in order to improve product synthesis. This analysis can be systematically extended to the full set of envirome components and full set of EFMs to support the concept of cell functional enviromics as defended in [20].

Predictive power

To test the predictive power, PLS and PLP models were calibrated with the calibration data set composed by 50% of data points and then simulated on the validation data set composed by the remaining 50% measured points. The PLS model with 18 latent variables explained 90.1% of $R^2$ variance in the calibration dataset but only 76.8% of the validation dataset. The quality of the results can be visually inspected in Figure 7. The degradation of accuracy in the validation dataset is rational given that the model is requested to predict data points, which may lay outside of the domain of experience defined by the calibration data set. As for PLP it is a very interesting result to verify that despite explaining a lower variance in the calibration data set (83.2% against 90.1% for PLP and PLS respectively), the accuracy of the validation data set was higher than that of PLS (81.9% against 76.8% for PLP and PLS respectively). Moreover, the variance of the validation data set is almost equal to that of the calibration data set, denoting a more consistent model, with higher predictive power than the PLS one (Figure 8).

In order to screen out the possibility of a casual better performance of PLP in relation to PLS due to the particular data partitioning employed, the same variance analysis was performed for the $z = 200$ PLP and PLS runs performed with randomly selected calibration and validation data points according to the bootstrapping technique previously described. The results show that the explained variance of the validation data set varied between 78.8-85.6%
for PLP and 50.4-82.7% for PLS (see Additional File 3). In 194 out of 200 runs the PLP outperformed the PLS, thus confirming that while PLS is consistently more accurate in describing the calibration data than PLP, the latter is consistently more accurate at predicting the validation data than PLS.

**Discussion**

The key PLS feature is identifying independent $X$ and $Y$-scores so that the relationship between successive pairs of scores is as strong as possible. PLS may be thus viewed as a robust form of redundancy analysis, seeking directions in the factor space that are associated with high variation in the response $Y$ but biasing them toward directions that are more accurately predicted. Due to its advantages in handling highly redundant data sets, PLS has become a widely used regression analysis technique in systems biology. It has been applied as an inference tool for predicting metabolic fluxes using isotopomer flux data [29], analysing genomic and proteomic data [25], identifying signalling networks by inducing cellular response to different stimuli [30-32] and network structure using metabolomic data [33]. Moreover, PLS has also been applied for the identification of active cellular pathways as a function of the environment using metabolic and gene expression profiles [34], detection of gene-gene interactions from microarrays data [35,36] and culture media optimization using nutritional profiling data [26,27].

The main disadvantage of PLS lies in its empirical data-driven nature with limited added-value in terms of mechanistic knowledge generation. Although carrying some internal structure, this structure is not inspired by any *a priori* mechanistic knowledge of the system. PLP may be viewed as a constrained version of PLS, attuned...
to the structure of the biological system under study. While in PLS the loadings and score are abstract variables, in PLP loadings and scores refer to well defined metabolic structures. Specifically, PLP explores EFM as “principle components” of a metabolic network. Indeed, EFM obey to the principle of non-decomposability, meaning that any particular flux distribution can be expressed as a nonnegative weighted sum of EFM. Thus the ranking obtained in PLP refers to active pathways as inferred by their level of correlation with the environmental state. In terms of data requirement, PLS belongs to the class of multivariate regression techniques particularly suitable to handle highly dimensional data sets even if the number of observations is limited [25]. PLS is typically used to model spectral data such as near infrared or 2D-fluorescence maps [37]. A basic requirement is that the number of latent variables must be lower than the number of observations in the calibration data set. This means that reliable linear models can be identified from a moderate number of observations of highly dimensional datasets. The same properties apply to PLP. A basic constraint is that the number of discriminated EFM cannot be higher than the number of observations in the calibration data set. However the method offers no restriction in terms of the dimensionality of the input data set.

Finally it should be commented on the computational power requirements, which scales linearly with the number of EFM. In the present study with 251 EFM, computation requirements are in the order of seconds in a common PC. For a genome scale network with several million of EFM, computation power might easily rise to the scale of days in a common PC.

Conclusions
In this work we have developed an algorithm for the discrimination of active EFM on the basis of dynamical envirome data called projection to latent pathways (PLP). The algorithm is designed to maximise the covariance between envirome data and observed flux data under the constraint of universe of genes translated into a plausible set of EFM. In general lines, the algorithm discriminates a minimal set of envirome correlated EFM that maximise the variance of measured flux data. Thus the algorithm may be viewed as a reverse, envirome-to-function metabolic reconstruction methodology as opposed to the generally accepted gene-to-function reconstruction approach. Although presented here as a method to analyse envirome data sets, PLP has broader scope. It is rather a general methodology for statistical elimination of redundant metabolic structures that, in a broader sense, has the potential to bring together all

Figure 5 PLP regression coefficients. Regression coefficients of selected EFM quantify the contribution of each environmental factor in X to the respective EFM weighting factor.
Figure 6 Regression coefficients confidence intervals for EFM 1. Confidence interval as function of regression coefficients obtained for the product formation EFM (EFM 1). Black full circles are envirome factors. The light and dark blue regions correspond to confidence intervals higher than 50% and 100% of the nominal value of the regression coefficient, respectively.

Figure 7 Predicted metabolic fluxes by PLS. Predicted against measured fluxes computed by the PLS model for the BHK data set. Blue circles and red triangles represent the calibration and validation data points, respectively.
layers of ‘omic’ information under a common computational framework.

Additional material

Additional File 1: BHK metabolic network Biochemical reactions/pathways, enzymes and biomass composition considered in the metabolic model of BHK cells.

Additional File 2: BHK elementary modes List of elementary modes obtained from the BHK metabolic network (Additional File 1). Elementary modes are represented in reduced form in terms of extracellular metabolites.

Additional File 3: Environme factors regression coefficients List of PLP regression coefficients and the respective confidence intervals resulting from the bootstrapping analysis.

Acknowledgements

Financial support for this work was provided by the Portuguese Fundação para a Ciência e Tecnologia through projects PTDC/EBB-EBI/103761/2008 and MIT/Pt7BS-BB/0082/2008, PhD grant SFRH/BD/36285/2007 (ARF), SFRH/BD/36676/2007 (NC), SFRH/BD/51577/2011 (IAI) and SFRH/BD/70768/2010 (IAI) and SFRH/BD/36990/2007 (MvS) and Post-Doc grant SFRH/BPD/46277/2008 (JD).

Author details

1REQUIMTE, Systems Biology & Engineering Group, DQ/FCT, Universidade Nova de Lisboa, Campus Caparica, Portugal.
2Instituto de Biologia Experimental e Tecnológica (IBET), Apartado 12, 2781-901 Oeiras, Portugal.
3Instituto de Tecnologia Química e Biológica - Universidade Nova de Lisboa (ITQB-UNL), Apartado 127, 2781-901 Oeiras, Portugal.
4LEPAE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal.

Authors’ contributions

The project was conceived by RO. Experiments were performed by AT and NC. PLP algorithm was developed and implemented by JD and RO. Algorithms (PLP and PLS) benchmarking and data analysis was performed by AR, MvS, IAI and RP. Manuscript writing was performed by JD and RO. All authors read and approved the final manuscript.

Received: 9 May 2011  Accepted: 1 November 2011
Published: 1 November 2011

References

1. Schuster S, Fell DA, Dandekar T: A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 2000, 18(3):326-332.
2. de Figueiredo LF, Schuster S, Kaleta C, Fell DA: Can sugars be produced from fatty acids? A test case for pathway analysis tools. Bioinformatics 2008, 24(22):2615-2621.
3. Klamt S, Stelling J: Two approaches for metabolic pathway analysis? Trends Biotechnol 2003, 21(2):64-69.
4. Palsson BO, Price ND, Papin JA: Development of network-based pathway definitions: the need to analyze real metabolic networks. Trends Biotechnol 2003, 21(5):195-198.
5. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO: Comparison of network-based pathway analysis methods. Trends Biotechnol 2004, 22(8):400-405.
6. Schuster S, Dandekar T, Fell DA: Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 1999, 17(2):53-60.
7. Trinh CT, Walschin A, Sienc F: Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl Microbiol Biotechnol 2009, 81(3):813-826.
8. Wagner C: Nullspace approach to determine the elementary modes of chemical reaction systems. J Phys Chem B 2004, 108(7):2425-2431.
9. Terzer M, Stelling J: Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics 2008, 24(19):2229-2235.
10. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 2000, 97(10):5528-5533.

11. de Figueiredo LF, Podhorodecki A, Rubio A, Kaleta C, Beasley JE, Schuster S, Planes F. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics 2009, 25(23):3158-3165.

12. Song HS, Ramkrishna D. Reduction of a Set of Elementary Modes Using Yield Analysis. Biotechnol Bioeng 2009, 102(2):554-568.

13. Zhao QY, Kurata H. Maximum entropy decomposition of flux distribution at steady state to elementary modes. J Biosci Bioeng 2009, 107(1):84-89.

14. Wilbach SJ, Mahadevan R, Palsson BO. Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. J Theor Biol 2003, 224(3):313-324.

15. Wilbach SJ, Mahadevan R, Palsson BO. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: The Escherichia coli spectrum. Biotechnol Bioeng 2004, 86(3):317-331.

16. Wang QZ, Yang YD, Ma HW, Zhao XM. Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. Bioinformatics 2003, 21(9):1049-1052.

17. Nookaew I, Meechai A, Thammarongtham C, Laoteng K, Ruanglek V, Wang QZ, Yang YD, Ma HW, Zhao XM. Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. Bioinformatics 2003, 21(9):1049-1052.

18. Schuster S, Hilgetag C. On Elementary Flux Modes in biochemical reaction systems at steady state. J Theor Biol 1994, 154:1-28.

19. Schwartz JM, Kanehisa M. Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. Bioinformatics 2003, 21(9):1049-1052.

20. Teixeira AP, Dias JML, Carinhas N, Sousa M, Clemente JJ, Cunha AE, von Stosch M, Alves PM, Carrondo MIT, Oliveira R. Cell functional enviroimics: Unravelling the function of environmental factors. BMC Syst Biol 2011, 5.

21. Song HS, Ramkrishna D. Reduction of a Set of Elementary Modes Using Yield Analysis. Biotechnol Bioeng 2009, 102(2):554-568.

22. Nookaew I, Meechai A, Thammarongtham C, Laoteng K, Ruanglek V, Cheevasathanak S, Nielsen J, Bhurimatana S. Identification of flux regulation coefficients from elementary flux modes: A systems biology tool for analysis of metabolic networks. Biotechnol Bioeng 2007, 97(6):1535-1549.

23. Schwartz JM, Kanehisa M. Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. BMC Bioinformatics 2006, 7:20.

24. Schuster S, Hilgetag C. On Elementary Flux Modes in biochemical reaction systems at steady state. J Theor Biol 1994, 154:1-28.

25. Wold H. Path models with latent variables: The NIPALS approach. In Quantitative sociology: International perspectives on mathematical and statistical modeling. Edited by: Bialock HM, Arganbargian A, Borodkin FM, Boudon R, Capechich V. New York: Academic; 1975:307-357.

26. Teixeira AP, Oliveira R, Alves PM, Carrondo MIT. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. Biotechnol Adv 2009, 27(6):762-732.

27. Llaneras F, Pico J. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. BMC Bioinformatics 2007, 8.

28. Miller-Jensen K, Janes KA, Brugge JS, Lauffenburger DA. Comprehensive modeling of cell signaling systems using a modified cell-based hybridoma modeling system. Biotechnol Bioeng 2002, 78(2):169-179.

29. Antoniewicz MR, Stephanopoulos G, Kelleher JK. Uncertainty estimation for multivariate regression coefficients. BMC Bioinformatics 2007, 8(1):76.

30. Ivakhno S, Armstrong JD. Non-linear dimensionality reduction of signaling network fluxes in scenarios where measurements are uncertain and/or insufficient. BMC Bioinformatics 2007, 8.

31. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 2000, 97(10):5528-5533.

32. Iwahashi T, Armstrong JD. Non-linear dimensionality reduction of signaling networks. BMC Syst Biol 2007, 1:7.

33. Iwahashi T, Armstrong JD. Non-linear dimensionality reduction of signaling networks. BMC Syst Biol 2007, 1:7.

34. Li Z, Chan C. Integrating gene expression and metabolic profiles. J Biol Chem 2004, 279(26):27124-27137.

35. Pihur V, Datta S. Reconstruction of genetic association networks from microarray data: a partial least squares approach. Bioinformatics 2008, 24(4):561-568.

36. Tenenhaus A, Guillemot V, Girard X, Frouin V. Gene Association Networks from Microarray Data Using a Regularized Estimation of Partial Correlation Based on PLS Regression. IEEE-ACM Trans Comput Biol Bioinform 2010, 7(2):251-262.

37. Teixeira AP, Oliveira R, Alves PM, Carrondo MIT. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. Biotechnol Adv 2009, 27(6):762-732.

38. Llaneras F, Pico J. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. BMC Bioinformatics 2007, 8.

Cite this article as: Ferreira et al.: Projection to latent pathways (PLP): a constrained projection to latent variables (PLS) method for elementary flux modes discrimination, BMC Systems Biology 2011, 5:181.