Proceedings of the 11th International Symposium on
Computer Applications in Biotechnology (CAB 2010),
Leuven, Belgium, July 7-9, 2010
Julio R. Banga, Philippe Bogaerts, Jan Van Impe,
Denis Dochain, Ilse Smets (Eds.)

Possibilistic validation of a constraint-based model under data scarcity: application to *Pichia pastoris* cultures

M. Tortajada*[1]. F. Llaneras[1]**
J. Picó**

**Biopolis S.L., Dept. of Microbial Biotechnology*
Valencia, Spain (e-mail: marta.tortajada@biopolis.es).
***AI2, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022*
Valencia, Spain (frallaes@upvnet.upv.es; jpico@ai2.upv.es)
[1]These authors contributed equally to this work

Abstract: Constraint-based modelling allows building structured models of cells without accounting for intracellular kinetics. These models can be combined with experimental data to estimate the (pseudo-steady) state or phenotype exhibited by cells at given conditions, standing out as a useful analytical tool. In this work, a simplified, constraint-based model of *Pichia pastoris*, a widely recognized platform for recombinant protein expression, is derived from its metabolic network. Then, the model is validated against experimental data provided by different research groups: possibility theory is used to analyse the consistency between model and measurements. Afterwards, the biomass growth rate is estimated to illustrate the ability of the model to predict non-measured fluxes. The approach followed in this contribution is particularly useful in scenarios lacking data; it makes it possible to link the extracellular behaviour of *Pichia pastoris* during cultivation with its internal state, being a promising tool for optimization and monitoring industrial processes.

Keywords: modelling, constraints, possibility theory, *Pichia pastoris*, uncertainty, metabolic flux analysis

1. INTRODUCTION

Pichia pastoris is a methylotrophic yeast worldwide recognized as a workhorse for recombinant protein expression for its high-density cell growth, ability to produce post-translational modifications and good protein yield/cost ratio. Cloned under its strong, tightly regulated alcohol oxidase promoter, heterologous genes are expressed while *P. pastoris* grows on methanol as sole or combined carbon source. Recombinant protein expression improvement in *P. pastoris* has been usually addressed heuristically, although a few publications on modelling and optimisation can be found in the literature. Among these, a few explore more structured models representing intracellular behaviour (Ren, 2003; Solà, 2007).

Nowadays, microbial systems are being increasingly studied under a system-level approach in which the collection of biocatalytic reactions involved in metabolism is assembled in networks (Palsson, 2002). The mass balances around the nodes of these networks, the *m* internal metabolites, can be described by a matrix equation:

$$\frac{d\mathbf{c}}{dt} = \mathbf{N} \cdot \mathbf{v} \qquad (1)$$

where $\mathbf{c}$ is a vector of metabolite concentrations and $\mathbf{v}$ is the vector of reaction rates, or fluxes, representing the mass flow through each of the *n* reactions in the network.

Since reaction kinetics are still rarely known, internal metabolites are often assumed to be at steady state, so (1) results into a system of linear equations. Then, other constraints can be imposed; for instance, it is common to consider the irreversibility of certain reactions using inequalities. In this way, a constraint-based model is assembled (Llaneras, 2008; Schilling, 2000):

$$\mathcal{MOC} = \begin{cases} \mathbf{N}\cdot\mathbf{v} = \mathbf{0} \\ \mathbf{D}\cdot\mathbf{v} \geq \mathbf{0} \end{cases} \qquad (2)$$

Where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D_{ii}} = 1$ if the flux *i* is irreversible (otherwise 0). The constraint-based model defines a space of feasible states: only flux vectors fulfilling the constraints are states that cells can exhibit.

At this point, one can perform a metabolic flux analysis (MFA), which, generally speaking, is the exercise of estimating the fluxes at given conditions by combination of the model and a set of measured fluxes (Heijden, 1994). A typical difficulty to be tackled by MFA is the scarcity of measurements common in industry, and thus Possibilistic MFA, particularly well suited for these situations, will be used in this contribution (Llaneras, 2009).

In following sections a model of *P. pastoris* will be described and validated against experimental data. Afterwards, its ability to predict non-measured fluxes will be illustrated by estimating the biomass growth rate.

2. METHODS

2.1 Consistency analysis

The constraint-based model has been validated against experimental datasets taken from the literature. Basically, the consistency analysis is performed checking that the flux states shown by cells do not violate the constraints imposed by the model. However, this simple approach would be impractical because, as measurements are imprecise, they do not *exactly* satisfy the constraints. Such difficulty is overcome by taking into account uncertainty:

$$\mathbf{v'_m} = \mathbf{v_m} + \mathbf{e_m} \tag{3}$$

Where $\mathbf{e_m}$ represents the error or deviation between the actual fluxes $\mathbf{v_m}$ and the measured values $\mathbf{v'_m}$. Model and measurements will be consistent if there is a flux vector $\mathbf{v}$ fulfilling (2) and (3) for a reasonably small $\mathbf{e_m}$. Otherwise, we will conclude that model and measurements are inconsistent.

A simple way of analysing the consistency between model and measurements is to find the flux vector $\mathbf{v}$ fulfilling (2) and (3) that minimises the (variance-weighted) sum of errors in the measurements:

$$\min \quad \Phi = \mathbf{e_m^T} \cdot \mathbf{F^{-1}} \cdot \mathbf{e_m} \quad \text{s.t.} \quad \mathcal{MOC} \tag{4}$$

Where it is assumed that $\mathbf{e_m}$ are distributed normally with a mean value of zero and a variance-covariance matrix $\mathbf{F}$.

When only linear equalities are used in *MOC*, the residual ϕ is a stochastic variable following a χ^2-distribution (of order equal to the degree of redundancy of the system being solved), enabling the use of χ^2-test to analyze the consistency (Stephanopoulos, 1998). As our model contains inequalities, the χ^2-test cannot be used, though the residual ϕ still provides an indication of consistency.

2.2 Consistency analysis: Possibilistic MFA

The consistency analysis can also be formulated as a possibilistic, constraint satisfaction problem. A flux vector fulfilling the model constraints (2) and compatible with the measurements will be "possible", otherwise "impossible". This idea can be refined to cope with measurements errors by introducing the notion of "degree of possibility".

We introduce a set of constraints considering measurement imprecision, as in (3), but where $\mathbf{e_m}$ is substituted by some non-negative decision variables μ and ε:

$$\mathcal{MEC} = \begin{cases} \mathbf{v'_m} = \mathbf{v_m} + \varepsilon_1 - \mu_1 + \varepsilon_2 - \mu_2 \\ \varepsilon_1, \mu_1 \geq 0 \\ 0 \leq \varepsilon_2 \leq \varepsilon_2^{\max} \\ 0 \leq \mu_2 \leq \mu_2^{\max} \end{cases} \tag{5}$$

In this way, the assertion $\mathbf{v'_m}=\mathbf{v_m}$ is relaxed, conforming a possibility distribution in $(\mathbf{v'_m},\mathbf{v_m})$ associated to a cost index J (under a non-interactivity assumption):

$$J = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \tag{6}$$

where α and ß are row vectors of user-defined, sensor accuracy coefficients (if sensor error is symmetric, both vectors should be equal).

The cost index J reflects the log-possibility of a particular flux vector $\mathbf{v}$. The interpretation of (5) and (6) may be: "$\mathbf{v'_m}=\mathbf{v_m}$ is fully possible; the more $\mathbf{v'_m}$ differs from $\mathbf{v_m}$, the less possible such situation is".

At this point, the maximum possibility (minimum-cost) flux vector $\mathbf{v_{mp}}$ can be obtained solving a LP problem:

$$\begin{aligned} \min \quad & J = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \\ \text{s.t.} \quad & \mathcal{MOC} \cup \mathcal{MEC} \end{aligned} \tag{7}$$

being its degree of possibility $\pi(\mathbf{v_{mp}}) = \exp(J_{\min})$.

This degree of possibility provides an indication of the consistency between model (*MOC*) and measurements (*MEC*): a possibility equal to one must be interpreted as complete agreement between the model and the original measurements; lower values of possibility imply that certain degree of error in the measurements is needed to find a flux vector fulfilling the model constraints. See (Llaneras, 2009) for more details on the possibilistic framework.

2.3 Possibilistic estimation of non-measured fluxes

Possibilistic MFA is also capable of estimating the metabolic fluxes based on the model and the available measurements. The simplest point-wise estimate is the minimum-cost flux vector resulting from (7), which contains the most possible value for each flux. However, a point-wise estimate is limited when multiple combinations might be reasonably possible. In this situation, intervals of flux values $[\,v_{i,\gamma}^{m}, v_{i,\gamma}^{M}\,]$ with a degree of *a posteriori* possibility higher than γ can be obtained solving two LP problems:

$$v_{i,\gamma}^{m} = \min v_i \quad \text{s.t.} \begin{cases} \mathcal{MOC} \cup \mathcal{MEC} \\ J - \log \pi(\mathbf{v_m}) < -\log \gamma \end{cases} \tag{8}$$

The upper bound is obtained replacing min by max.

Possibilistic intervals have a similar interpretation to "confidence intervals" ("credible intervals") in Bayesian statistics, and provide concise but rich flux estimation.

3. METABOLIC NETWORK OF *P. pastoris*

The metabolic network shown in Fig. 1 has been adapted from the stoichiometric model presented in (Dragosits,

2009). The objective of this representation is not to accurately reproduce the biochemistry of the yeast, but to produce a simplified model, capable to describe the main scenarios shown by *Pichia* cultures, in which to test useful analysis methodologies.

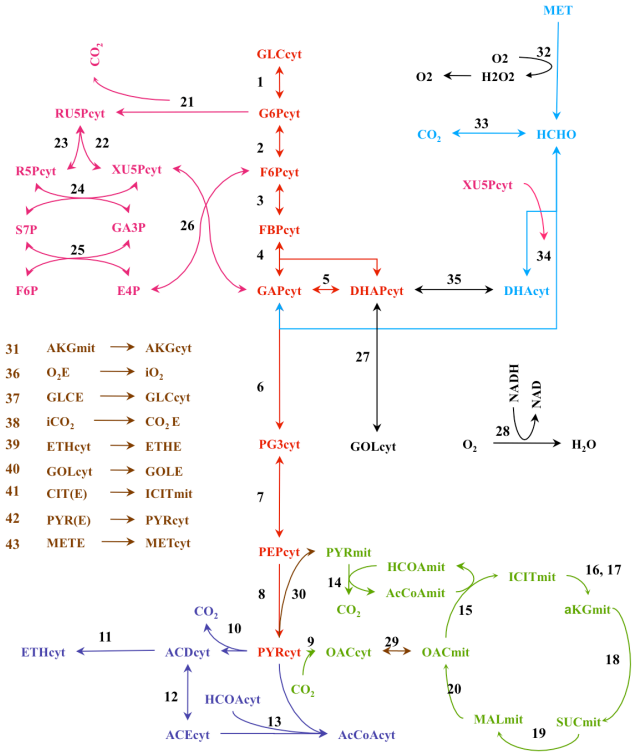

Fig 1. Metabolic network of *P. pastoris*

The network schematically represents the main catabolic pathways of the yeast *P. pastoris* for growth on glucose, glycerol and methanol, including glycolysis, citric acid cycle, pentose phosphate and fermentative pathways. Anabolic pathways are summarized by a biomass equation derived from yeast macromolecule composition, considering key precursors for each of its components and structural monomer formation reactions for aminoacids and nucleotides. The model considers compartimentation between mitochondrial and cytosolic pools for NAD^+, $NADP^+$, AcCoA, oxalacetate and pyruvate.

The resulting model contains 44 pathways reactions. A balanced, null production rate is assumed for 36 compounds resulting in 8 Degrees of freedom. The constraint-based model (2) used hereinafter is generated when irreversibilities are considered for all reactions except for {1-7; 12; 29; 33-35; 41}.

4. VALIDATION AGAINST EXPERIMENTAL DATA

A total of 11 different datasets were compiled from the literature and used to determine whether the simplified model described above is coherent with experimental data (table 2).

The datasets were used to check that the experimental measurements, which reflect the metabolic state of cells, are feasible states according to the model. Two different analysis of consistency were performed: one based on weighted least squares and another one based on possibility theory, both described in the methods section. The possibilistic approach is preferred in this case because the analysis of the residuals has some limitations due to the presence of inequality constraints on the model.

Uncertainty in measurements was described as follows. In all weighted least squares problems a standard deviation of 10% was assigned to each measurement. To perform the possibilistic analysis of consistency, measurements uncertainty has to be represented in possibilistic terms. It was considered that values near the measured ones (less than 5% deviation) are fully possible, to account for *systemic* errors. A decreasing possibility was assigned to larger deviations: values with a deviation of 20% have a possibility of 0.5 and those with a deviation of 30% a possibility of 0.15. Notice that possibility was defined by conjunction, so that if two measurements are deviated, for instance with possibilities 0.8 and 0.5 respectively, their joint possibility will be 0.4. Hence, a maximum possibility of 0.36 means that there is an error between 10% and 20% in one measurement, or an error between 5% and 10% in two measurements, etc.

The results for each dataset are shown in table 2, where the minimized, sum of squared residuals (ϕ) and the possibility of the most possible flux vector (π) are given. The last column shows the measurements uncertainty needed to find a flux vector in full agreement with the model constraints (π=1).

In general, the consistency between model and experimental data is quite good. The dataset D1, which corresponds to *Pichia* growing on glucose, shows very good agreement. The measured data has full possibility (π=1), meaning that there is a flux vector fully compatible with model and measurements. In fact, as shown in the last column, a band of 1% around the measured values is sufficient to enclose this flux vector. Notice also that the residual is very low. Datasets A1 and A2 also show a good agreement. The discrepancy between measurements and model is bigger for A3, which possibility is 0.25, but still a band of 10% of deviation around the measurements is enough to enclose a flux vector compatible with the model. The larger discrepancy in A4, which corresponds to a scenario with high protein productivity, and similar results obtained for datasets B1-B3, reveal the existence of non-modelled phenomena. The agreement is quite good for C1-C3, but the increase of model and measurements discrepancy along with higher protein expression is also noticeable.

In summary, the constraint-based model shows acceptable agreement with the experimental data reported by different groups for *P. pastoris* cultures.

Table 2. Experimental data and model consistency

Ref*	μ	Q_{Glu}	Q_{Gly}	Q_{Met}	Q_{et}	OUR	CPR	Q_P	Consistency**		
	Cmmol $g^{-1}h^{-1}$	mmol $g^{-1}h^{-1}$	"	"	"	"	"	mg $g^{-1}h$	ϕ	π	To π=1
D1	3.86	0.97	0.00	0.00	0.00	2.02	2.07	0.002	0.03	1.00	2%
A1	1.88	0.00	1.09	0.00	0.00	2.16	1.56	0.002	0.28	1.00	7%
A2	2.07	0.00	0.95	0.63	0.00	2.70	1.70	0.001	1.20	0.73	12%
A3	1.72	0.00	0.74	1.48	0.00	3.90	2.10	0.004	2.81	0.25	20%
A4	2.02	0.00	0.57	2.33	0.00	4.85	2.21	0.006	5.36	0.09	29%
B1	6.17	0.00	2.75	0.00	0.00	3.62	2.35	0.000	0.07	1.00	4%
B2	6.18	0.00	2.22	1.87	0.00	7.19	4.18	0.000	5.24	0.04	23%
B3	6.24	0.00	2.23	2.73	0.00	7.20	3.60	0.009	2.34	0.32	19%
C1	2.32	0.00	0.74	2.22	0.00	3.58	2.05	0.012	0.06	1.00	3%
C2	2.32	0.00	0.37	3.33	0.00	4.44	2.55	0.021	0.79	1.00	10%
C3	2.32	0.00	0.00	4.44	0.00	5.29	2.82	0.022	1.63	0.49	15%

Ethanol, citrate and piruvate are not produced nor consumed. Q_p indicates heterologous protein specific production, OUR and CPR oxygen and carbon dioxide uptake and production rates, respectively.
*D: (Dragosits, 2009); A&B: (Solà, 2007); C: (Jungo, 2007)
**Minimized sum of squared residuals (ϕ), possibility of the most possible flux vector (π) and degree of measurements uncertainty to π=1.

6. USING THE MODEL TO PREDICT GROWTH

At this point, the biomass growth rate for each dataset was estimated by applying Possibilistic MFA based on the constraint-based model and the available measurements (except, of course, the measured growth rate). Details can be found in the methods section.

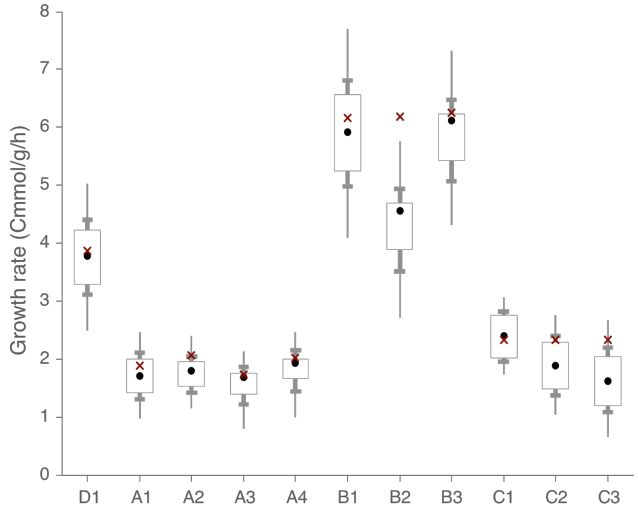

Fig. 2. Prediction of growth rate using Possibilistic MFA. Crosses denote the measured values, and circles the most possible estimate. The intervals of conditional possibilities of 0.8, 0.5 and 0.1 are also depicted.

As shown in fig. 2, for the majority of the analysed scenarios, in the 3 studied references, the estimated growth rate is found to be in very good agreement with the measured one. The most possible estimate is slightly accurate for two datasets —A2 and C3— and only one dataset, B2, is not enclosed by the interval estimate of possibility 0.1. This significant deviation observed in B2 could be indicative of non-modelled phenomena but also of larger measurement uncertainty within this particular dataset. This late option seems more likely considering the coherence found in the remaining datasets, both for this reference and others.

These results provide further validation of the model, pointing out that, even with limitations, the model has predictive capacity. This conclusion is strengthened by the fact that estimated variable, the growth rate, is highly connected along the whole network because the synthesis of biomass requires the participation of several precursors.

7. USING THE MODEL TO ESTIMATE THE WHOLE FLUX DISTRIBUTION

Once the model has been validated, the same approach used to estimate the growth rate could be used to estimate all the internal, non-measured fluxes.

For illustration purposes, only the whole flux distribution in the scenario A2 estimated with Possibilistic MFA is shown in Fig. 3. Notice that this estimation could not be done with standard MFA because the measurements were insufficient to produce a determined system (the network has 8 degrees of freedom and there are 7 measurements, one dependent, so the systems remain underdetermined with 2 degrees of freedom). The dependent measurement makes the system redundant, and thereby the estimation is more reliable.

Moreover, the results show that it is not necessary to completely offset the underdeterminacy of the system (nor invoking optimality criteria) to get narrow estimates for all the fluxes in the network. Possibilistic MFA estimates all the fluxes thanks to the irreversibility constraints.

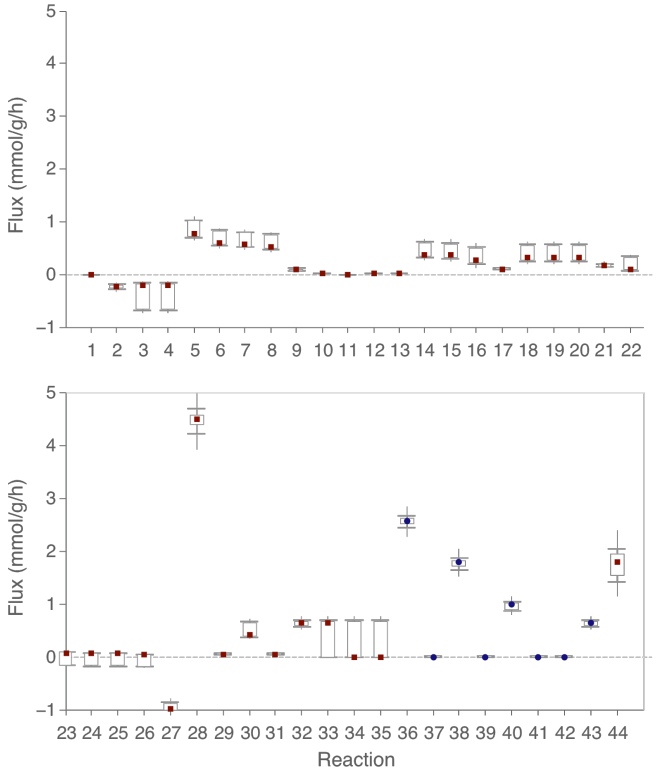

Fig. 3. Estimated fluxes in scenario A2. The most possible fluxes (circles and squares for measured and non measured fluxes, respectively) and the intervals of conditional possibilities 0.8, 0.5 and 0.1 are depicted.

8. CONCLUSIONS

The consistency of a constraint-based model for *Pichia pastoris* growth has been validated in several experimental scenarios resulting in good agreement between estimations and measurements. Besides, the predictive capacity of the model for cell growth rate, an attractive target for industrial fermentation monitoring and control, has been verified. Interestingly, the accuracy of predictions worsens for higher protein producing scenarios, indicating how the model, derived for a wild-type strain, is increasingly less applicable as wider resources are devoted to recombinant protein generation.

It must be highlighted that the model has been strictly constructed upon first-principles and sensible hypothesis, and can be now curated and its parameters tuned with further experimental data. Furthermore, after validation against intracellular data, the used approach will allow investigating the whole flux distribution and identifying patterns or alterations among intracellular fluxes as a result of changes in external, measurable fluxes.

ACKNOWLEDGEMENTS

This research has been partially supported by the Spanish Government (2st and 3rd authors are grateful to grants DPI2008-06880-C03-01 and A/016560/08). FLL is recipient of a fellowship from the Spanish Ministry of Science and Innovation (FPU AP2005-1442).

REFERENCES

Dragosits M., Stadlmann J., Albiol J., Baumann K., Maurer M., Gasser B., Sauer M., Altmann F., Ferrer P. and Mattanovich D. (2009). 'The effect of temperature on the proteome of recombinant *Pichia pastoris*.' J Proteome Res, 8 (3), 1380-92.

Llaneras F. and Picó J. (2008) 'Stoichiometric modelling of cell metabolism' *J Biosci Bioeng*, 105 (1), 1-11.

Llaneras F., Sala A. and Picó J. (2009). 'A possibilistic framework for constraint-based metabolic flux analysis.' *BMC Syst Biol,* 31 (3), 79.

Heijden R.T., Romein B., Heijnen J.J., Hellinga C., Luyben K.C. (1994) 'Linear Constraint Relations in Biochemical Reaction Systems: II. Diagnosis and Estimation of Gross' *Biotechnol Bioeng* 43 (1), 11-20.

Jungo C., Marison I., Stockar U. (2007) 'Mixed feeds of glycerol and methanol can improve the performance of *Pichia pastoris* cultures: A quantitative study based on concentration gradients in transient continuous cultures' *J Biotechnol*, vol. 128 (4) pp. 824-37

Palsson B. (2002) 'The challenges of *in silico* biology' *Nature Biotechnology*, 18 (11), 1147–1150.

Pfeiffer T., Sánchez-Valdenebro I., Nuño J.C., Montero F. and Schuster S. (1999) 'METATOOL: for studying metabolic networks' *Bioinformatics,* 15 (3) 251-7.

Ren H.T., Yuan J.Q. and Bellgardt K.H. (2003). 'Macrokinetic model for methylotrophic *Pichia pastoris* based on stoichiometric balance.' *J Biotechnol*, 5, 106 (1), 53-68.

Schilling C.H. and Palsson B.O. (2000). 'Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis'. *J Theor Biol*, 203 (3), 249-283.

Schuster S., Dandekar T. and Fell D.A. (1999). 'Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering'. *Trends Biotechnol*, 17(2), 53-60.

Schuster S., Hilgetag C., Woods J.H. and Fell D.A. (2002). 'Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism'. *J Math Biol*, 45 (2), 153-181.

Solà A., Jouhten P., Maaheimo H., Sánchez-Ferrando F., Szyperski T. and Ferrer P. (2007). 'Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates.' *Microbiology*, 153 (1), 281-90.

Stelling J., Klamt S., Bettenbrock K., Schuster S. and Gilles E.D. (2002). 'Metabolic network structure determines key aspects of functionality and regulation'. *Nature*, 420 (6912), 190-193.

Stephanopoulos GN, Aristidou A.A. (1998) 'Metabolic Engineering: Principles and Methodologies' San Diego: Academic Press, 725.