A Scalable Algorithm to Explore the Gibbs Energy Landscape of Genome-scale Metabolic Networks

Enzo Marinari

(*Sapienza*, Rome, Italy)

Innsbruck, September 21, 2012

This is work published in: A Scalable Algorithm to Explore the Gibbs energy Landscape of Genome-scale Metabolic Networks, D. De Martino, M. Figliuzzi, A. De Martino and E.M., PLoS Comp. Bio 8(6): e1002562 (2012).

First paper about MinOver for metabolic networks: Identifying essential genes in Escherichia coli from a metabolic optimization principle, C. Martelli, A. De Martino, E.M., M. Marsili and I. Pérez Castillo, PNAS **106** (2009) 2607;

An interesting application to a very different system: Cell to cell lactate shuttle in the brain is independent of glutamatergic activity: a constraint based modeling perspective, F. A. Massucci, M. Di Nuzzo, F. Giove, B. Maraviglia, I. Perez Castillo, E.M. and A. De Martino, December 2011;

Summary

Metabolic networks and Gibbs Energy Landscape

FBA (Flux Balance Analysis) and VN (Von Neumann) modeling

Fluxes and free energies

Feasibility and Loop Removal

A new algorithm

An analysis of the human red blood cell metabolic network

Loop removal on E. coli network

Cellular metabolism at genome scale: constraint based models.

Minimal assumptions about the steady state.

Stoichiometric constraints:

- mass balance
- thermodynamic feasibility

Viable configuration of reaction fluxes + test for thermodynamic feasibility.

This can be computationally very demanding.

Fast and scalable algorithm (a relaxation procedure) to reconstruct the Gibbs energy + remove unfeasible reaction cycles.

(red blood cell + Escherichia coli)

Predict chemical activity and response to perturbations of cells - without relying on kinetic details (many parameters, imprecise data available).

Specify minimal physico-chemical constraints to describe the reaction network.

1. flux vectors need to satisfy mass balance conditions.
2. the direction of each reaction should guarantee decrease in Gibbs energy.

(1) (i.e. mass balanced flux vectors) does not imply (2): unfeasible cycle could plague the network.

This is true even when using at best estimates of chemical potentials in physiological conditions.

Requesting thermodynamic consistency of the model also for estimate of metabolite concentration.

Information on feasible Gibbs energy ranges: exploit the patterns of reactions interconnections encoded in the stoichiometry to narrow the experimental bounds.

Here: landscape of Gibbs free energies compatible with a given vector of reaction directions using all stoichiometric information via heuristics inspired by perceptron learning.

Metabolic reaction network.

Matrix Ξ encoding the stoichiometric coefficients ξ_i^μ for metabolite $\mu = 1, \ldots M$ in reaction $i = 1, \ldots N$.

A positive ξ_i^μ characterizes the product of the forward reaction. A negative ξ_i^μ characterizes the substrate of the forward reaction.

Concentration c^μ of metabolite μ:

$$\dot{c}^\mu = \sum_{i=1}^{N} \xi_i^\mu \nu_i \left(\mathbf{c}, \mathbf{k}, \ldots\right)$$

ν_i: net flux of reaction i

$\mathbf{k}$: vector of reaction constants

The equation can also depend on many other factors, for example enzyme availability and kinetics, transport processes etc.

Usual approach: assume stationarity $\Rightarrow \nu \mid \Xi\nu = 0$. M linear mass balance equations (MBE) + bounds $\nu_i^{\min} \leq \nu_i \leq \nu_i^{\max}$.

Usually $N > M$, i.e. system is under-determined.

$$S \equiv N - \text{ rank } (\Xi)$$

dimension of solution space.

Find optimized solutions (flux configurations) optimal with respect to specific biological functionalities: for example maximize biomass or ATP production, or minimize glucose consumption (effective nutrient usage) or the total flux of intracellular reactions (maximal enzymatic efficiency). This is FBA.

For example describe optimal bacterial growth in wild type and knock-out conditions (Palsson et al., Segrè et al.),

Our approach. Apply Von Neumann modeling (originally for economics, producers and consumers).

No cost function: solution space is left unconstrained. Extracellular medium is fixed.

A metabolite is producible if $\exists \nu |$

$$\sum_i \xi_i^\mu \nu_i > 0$$

Feasible production profiles are solution of

$$\Xi \nu \geq 0$$

$y \equiv \Xi \nu$ tells us if a metabolite is producible ($y^\mu > 0$) or not ($y^\mu = 0$) in a given feasible flux state.

This is VNC (Von Neumann flux stability constraint): for each chemical species at stationarity overall consumption cannot exceed the supply.

Producible metabolites: cellular outtakes or other processes (bug in the network...).

The "objective function" can emerge here as a statistically robust production profile.

Here effective sampling is possible even for large networks.

Monte Carlo could be possible also on large networks: work in progress (compute volume of polytopes in many dimensions).

We study here:

1. Red Blood Cell Metabolism (a small network: 35 reactions, 39 metabolites, 12 uptake fluxes);
2. Escherichia Coli (we consider 1767 reactions among 1349 chemical species). According to standard reversibility assignments 292 reactions are reversible, 1475 are unidirectional.

- $T = 298K$ $(24.9C)$
- P=1 atm
- pH = 7.6

$$G = E - P\,V - T\,S$$

does not increase spontaneously in an open system at constant T, P.

Let $u_i = \text{sign}\{\nu_i\} = \pm 1$, ν_i flux of reaction i ($+1$ is for a forward reaction, -1 for a backward one). Gibbs energy change ΔG_i induced by the reaction must be such that $u_i \Delta G_i \leq 0$.

$S_{\alpha i}$: stoichiometric coefficient of metabolite α in reaction i: $S_{\alpha i} > 0$ product, $S_{\alpha i} < 0$ substrate.

$\mu = \{\mu_\alpha\}$. μ_α Gibbs energy per mole of species α:

$$\Delta G_i = \left(S^T\right)\ \mu_\alpha$$

$$X_i \equiv -u_i \sum_{\alpha=1}^{M} S_{\alpha i}\ \mu_\alpha \geq 0\ \ \forall i$$

We look for $\boldsymbol{\mu}$.

For fixed u_i solution space for $\boldsymbol{\mu}$ is convex.

Relaxation methods: iterations where variables are updated and violated inequalities get fixed.

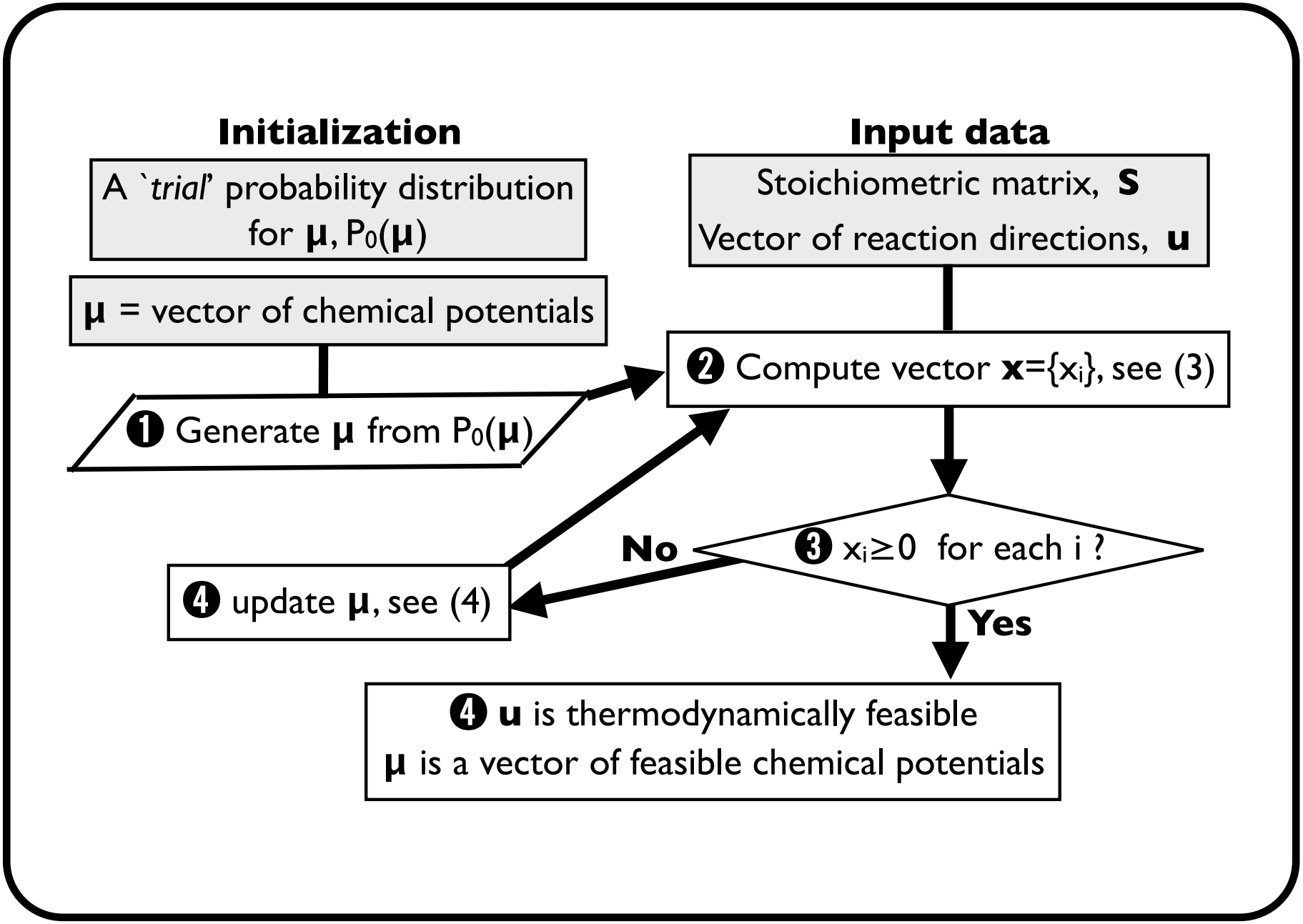
MinOver (Krauth Mézard 1987 for perceptron learning) is also used for computing fluxes in Von Neumann approach.

Start from trial probability distribution

$$P_0(\boldsymbol{\mu}) = \prod_{\alpha=0}^{M} P(\mu_\alpha)$$

P_0^α uniform over a given μ range.

P_0^α contains prior biochemical information (i.e. centered around experimental value with width connected to experimental errors and such to span a few orders of magnitude).


Initialization
A `trial' probability distribution for μ, P0(μ)
μ = vector of chemical potentials
❶ Generate μ from P0(μ)
Input data
Stoichiometric matrix, S
Vector of reaction directions, u
❷ Compute vector x={xi}, see (3)
❸ xi≥0 for each i ?
No
❹ update μ, see (4)
Yes
❹ u is thermodynamically feasible
μ is a vector of feasible chemical potentials

The algorithm is based on the following steps:

1. Generate a chemical potential vector $\boldsymbol{\mu} = \{\mu_\alpha\}$ from $P_0(\boldsymbol{\mu})$.
2. Compute $\mathbf{x} = \{x_i\}$ and $i_0 = \arg\ \min_i x_i$ (i.e., i_0 is the index of the least satisfied constraint).
3. If $x_{i_0} \geq 0$ then $\boldsymbol{\mu}$ is a thermodynamically consistent chemical potential vector for $\mathbf{u}$; exit (or go to 1 to obtain a different solution).
4. If $x_{i_0} < 0$, update $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} \ \rightarrow \ \boldsymbol{\mu} - \lambda u_{i_0} \mathbf{S}_{i_0}$$

(where $\lambda > 0$ is a constant and $\mathbf{S}_j$ is the j-th column of matrix $\boldsymbol{S}$), go to 2 and iterate.

The $-\lambda u_{i_0} \boldsymbol{S}_{i_0}$ drives the adjustment of chemical potentials: at every iteration the least satisfied constraint gets improved.

If a solution exists, convergence to a solution is guaranteed $\forall \lambda > 0$ (and the time of convergence depends on λ).

We can prove it: suppose that a solution $\boldsymbol{\mu}^*$ exists

$$-u_i(\boldsymbol{S}_i \cdot \boldsymbol{\mu}^\star) \geq c \qquad \forall i \ ,$$

with $c > 0$ constant. We have that after ℓ steps

$$\begin{aligned} \boldsymbol{\mu}(\ell) \cdot \boldsymbol{\mu}^\star &= \boldsymbol{\mu}(\ell-1) \cdot \boldsymbol{\mu}^\star - \lambda s_{i_0(\ell-1)} (\boldsymbol{S}_{i_0(\ell-1)} \cdot \boldsymbol{\mu}^\star) \\ &\geq \boldsymbol{\mu}(\ell-1) \cdot \boldsymbol{\mu}^\star + \lambda c \\ &\geq \boldsymbol{\mu}(0) \cdot \boldsymbol{\mu}^\star + \ell \lambda c \ , \end{aligned}$$

and

$$\boldsymbol{\mu}(\ell) \cdot \boldsymbol{\mu}(\ell) \leq \boldsymbol{\mu}(0) \cdot \boldsymbol{\mu}(0) + \ell \lambda^2 A \ ,$$

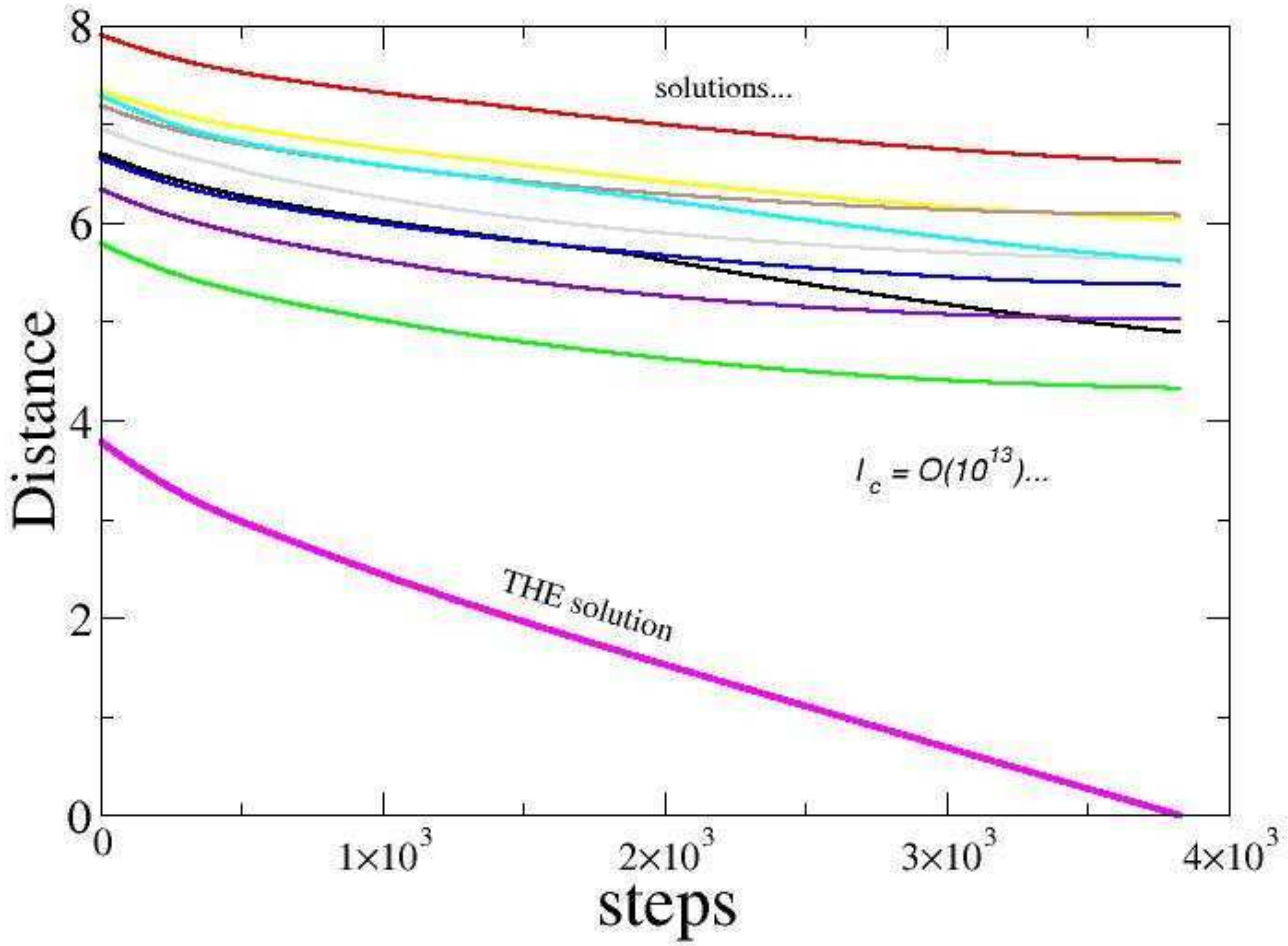
with

$$A = \max_i \sum_\alpha (S_{\alpha,i})^2 \ .$$

This implies

$$d(\ell) \equiv \frac{\boldsymbol{\mu}(\ell) \cdot \boldsymbol{\mu}^\star}{|\boldsymbol{\mu}(\ell)||\boldsymbol{\mu}^\star|} \geq \frac{\boldsymbol{\mu}(0) \cdot \boldsymbol{\mu}^\star + \ell \lambda c}{|\boldsymbol{\mu}^\star| \sqrt{|\boldsymbol{\mu}(0)|^2 + \ell \lambda^2 A}} \ .$$

By Cauchy-Schwarz $d(\ell) \leq 1$: imposing $d(\ell_c) = 1$ one finds an upper bound on the number of steps needed to converge.


8
6
4
2
0
Distance
solutions...
$l_c = O(10^{13})$...
THE solution
0
1×10^3
2×10^3
3×10^3
4×10^3
steps

Starting from a random vector if a solution exists we find it. Starting from different random vectors allows sampling the solutions (if there are many).

Many random starts: obtain a set of solutions of correlated $\{\mu_\alpha\}$.

The reinforcement steps build up the correlations.

Nature of solution space: unbounded cone passing through the origin. To obtain boundness one can use different approaches:

1. can clamp some μ_α keeping them fixed;
2. can assign a fixed range o variability to some or all μ_α :

$$\mu_\alpha^{\min} < \mu_\alpha < \mu_\alpha^{\max} ;$$

3. add global constraint (for example fix the potential for external metabolites in uptakes).

Important questions:

1. Minimal amount of a priori information needed to bound the solution space?
2. How to get uniform sampling? (Work in progress, G. De Concini, A. De Martino, D. De Martino, E.M.)

Here: we look at solutions close to prior biochemical information.

Our approach works "better" than relaxation and penalty methods even under noisy and inconsistent biochemical priors.

Identify and remove loops: an extension of the algorithm

A generic assignment of reaction directions can be unfeasible, i.e. such that $-u_i \sum_{\alpha=1}^{M} S_{\alpha i}\, \mu_\alpha \geq 0 \;\; \forall i$ does not have any non-trivial solution.

Farkas-Minkowski theorem: this happens if and only if there is at least one unfeasible loop, i.e. it $\exists$ a set $\mathcal{L}$ of reactions for which $\exists \{k_i > 0\}$ constant such that

$$\sum_{i \in \mathcal{L}} k_i u_i S_{\alpha,i} = 0 \qquad \forall \alpha \; .$$

In this case MinOver does not converge: the least satisfied constraint cycles along the loop. There is a problem in the network reconstruction (work in progress in analyzing a statistical theory that helps solving this problem: S. Colabrese, A. De Martino, D. De Martino, E.M.).

A simple way to correct an unfeasible set of reaction directions:

1. run MinOver for (large) number of iteration steps T. Keep track of the last K unsatisfied constraints, with K large;
2. select the reactions that appear more frequently in this set and search for a loop among these reactions;
3. if you find a loop, change the direction of one of its reactions.

All codes available from `http://chimera.roma1.infn.it/SYSBIO/`

The red blood cell metabolic network

Flux configurations from MC sampling of FBA solutions (Price, Schellenberger and Palsson 2004) and from Von Neumann (A. De Martino, Granata, EM, Martelli, Van Kerrebroeck 2010).

They are similar:

in FBA all reactions are bidirectional, all others are forward;

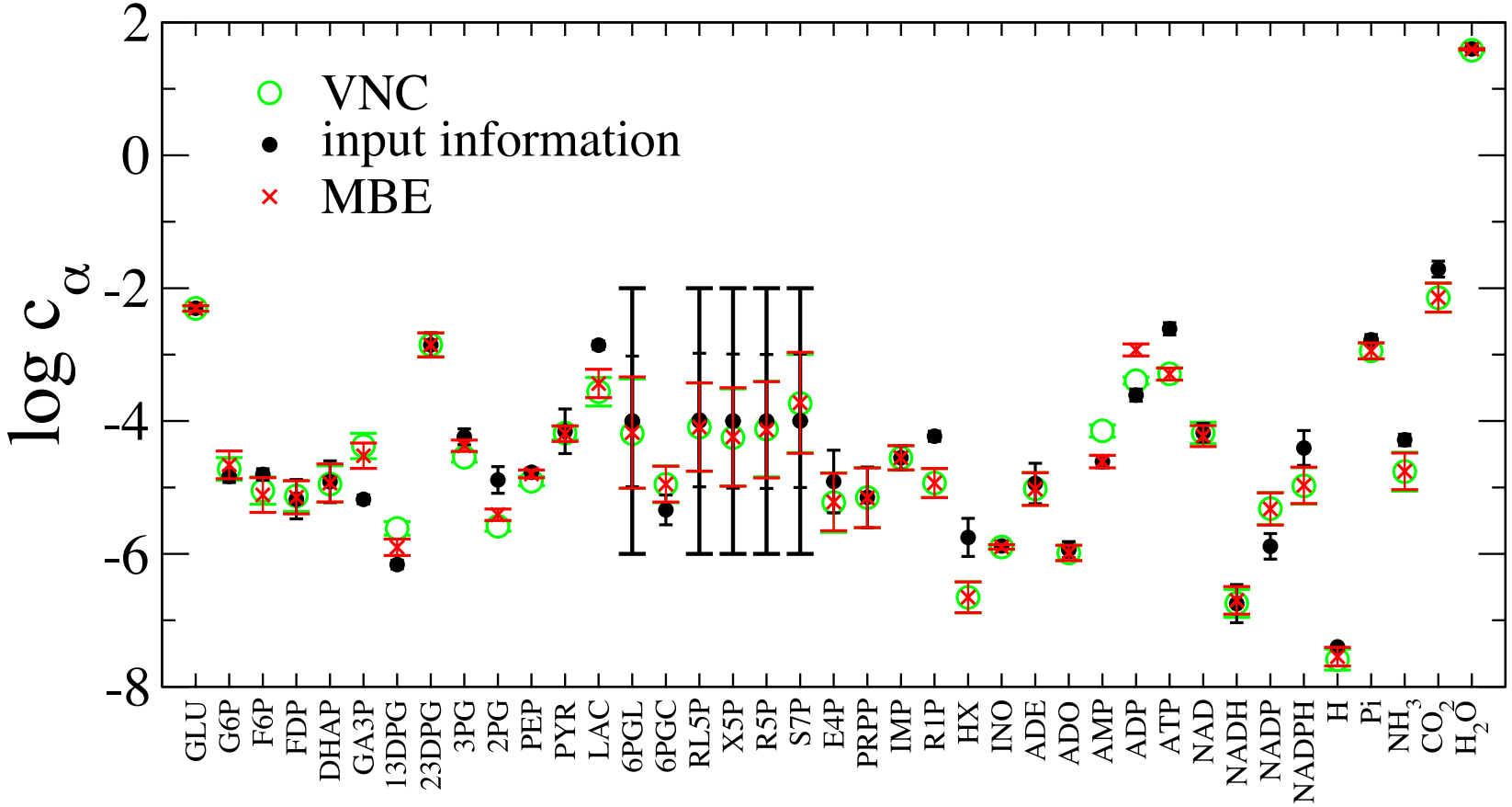
in VN one reaction is bidirectional (as in FBA) and two (that are bidirectional in FBA) are backward, all others are forward.

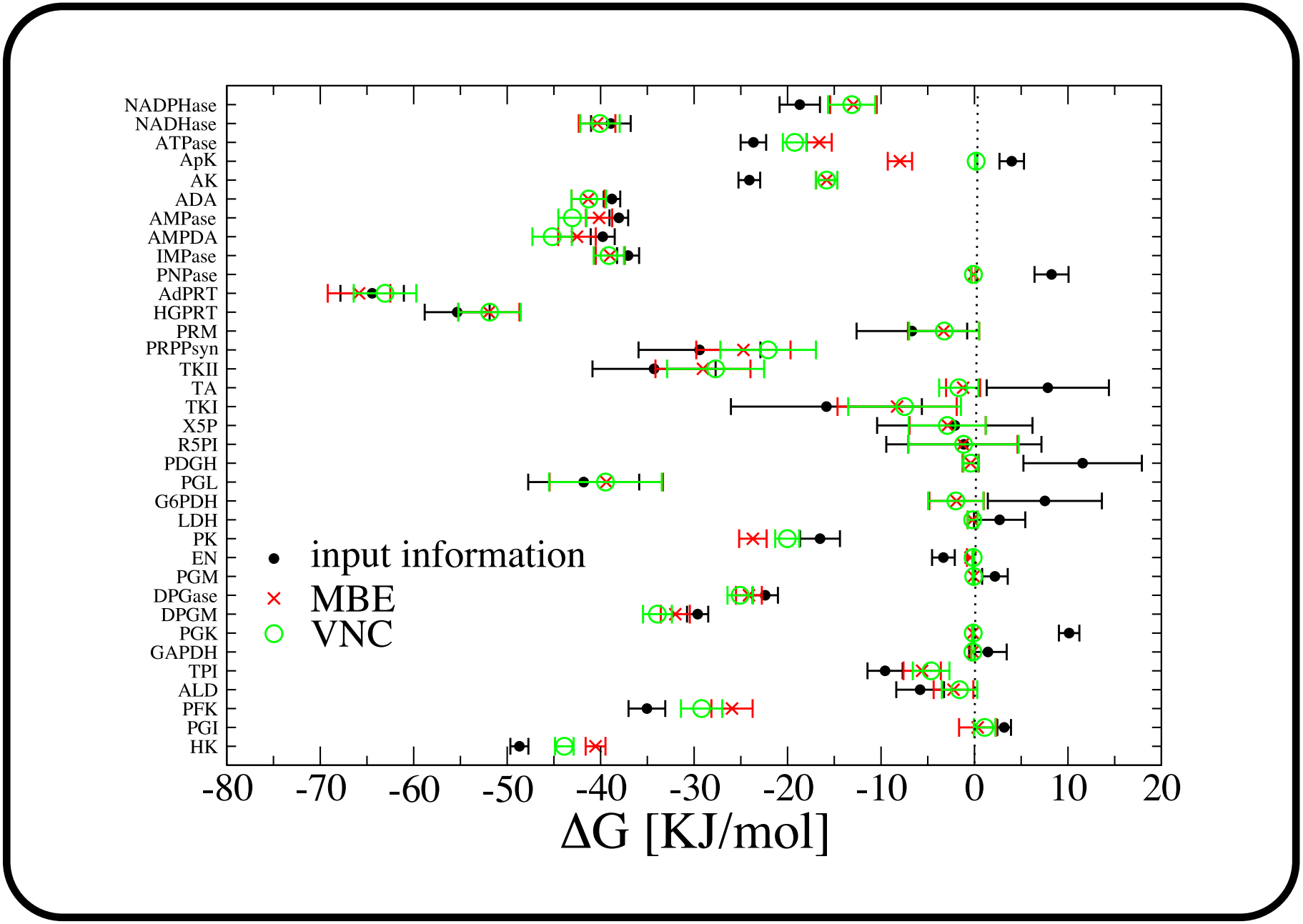
Both sets of solutions turn out to be thermodynamically feasible.

Concentrations are computed from (dilute solution approximation)

$$\log C_\alpha = \frac{\mu_\alpha - \mu_\alpha^{(0)}}{RT}$$

Water is clamped.


VNC
input information
MBE
log c α
2
0
-2
-4
-6
-8
GLU
G6P
F6P
FDP
DHAP
GA3P
13DPG
23DPG
3PG
2PG
PEP
PYR
LAC
6PGL
6PGC
RL5P
X5P
R5P
S7P
E4P
PRPP
IMP
R1P
HX
INO
ADE
ADO
AMP
ADP
ATP
NAD
NADH
NADP
NADPH
H
Pi
NH3
CO2
H2O


NADPHase
NADHase
ATPase
ApK
AK
ADA
AMPase
AMPDA
IMPase
PNPase
AdPRT
HGPRT
PRM
PRPPsyn
TKII
TA
TKI
X5P
R5PI
PDGH
PGL
G6PDH
LDH
PK
EN
PGM
DPGase
DPGM
PGK
GAPDH
TPI
ALD
PFK
PGI
HK
input information
MBE
VNC
-80
-70
-60
-50
-40
-30
-20
-10
0
10
20
ΔG [KJ/mol]

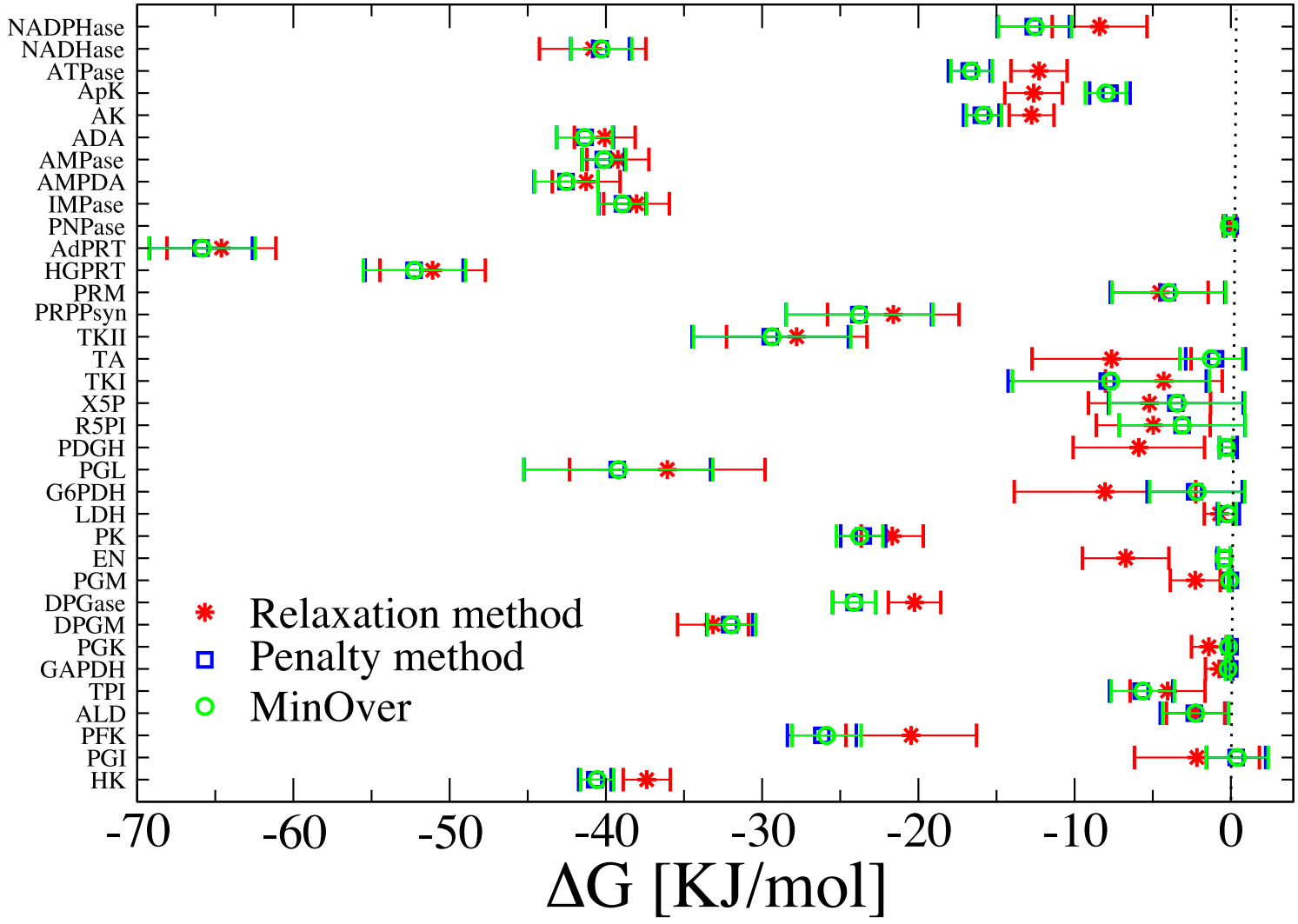
We compare to both penalty methods

$$\mu_\alpha \;\rightarrow\; \mu_\alpha - \lambda_1 \left[2(\mu_\alpha - \mu_\alpha(0)) + \lambda_2 \sum_{i \in \text{UNSAT}} S_{\alpha,i} \right] \qquad \forall \alpha \;,$$

and to a standard relaxation method defined by the update rule

$$\mu_\alpha \;\rightarrow\; \mu_\alpha - \frac{2x_{i_0}}{u_{i_0} \sum_\beta (S_{\beta,i_0})^2} \, S_{\alpha,i_0} \qquad \forall \alpha \;.$$

MinOver solutions are similar to the ones from penalty, but the method is close to 150 times faster.


NADPHase
NADHase
ATPase
ApK
AK
ADA
AMPase
AMPDA
IMPase
PNPase
AdPRT
HGPRT
PRM
PRPPsyn
TKII
TA
TKI
X5P
R5PI
PDGH
PGL
G6PDH
LDH
PK
EN
PGM
DPGase
DPGM
PGK
GAPDH
TPI
ALD
PFK
PGI
HK
Relaxation method
Penalty method
MinOver
-70
-60
-50
-40
-30
-20
-10
0
ΔG [KJ/mol]

Unfeasible loops in E. coli

1767 reactions, 1349 chemical species. Too large at today for full and reliable reconstruction of the Gibbs energy landscape (work in progress): we just cure it...

We fix 1475 reactions that are putatively (from biochemical priori) irreversible, and we fix at random the 292 reactions that are deemed reversible (all reactions operate: worst case scenario).

Only $\sim 1.5\%$ of these configurations turned out to be thermodynamically feasible, i.e. free from loops.

With a simple algorithm we were able to cure all unfeasible instances.

All of inconsistencies: 23 loops in total.

Summing up...

Gibbs energy role:

- estimate concentrations;
- Gibbs energy landscape can contain important regulatory information.

Thermodynamic feasibility can be translated to topological constraints: an improvement in reversibility assignments can be crucial in network reconstruction.

Here: given a flux configuration, thermodynamic constraints can be written as simple stoichiometric inequalities for the chemical potentials.

An important feature of our algorithm: prior information is only used to initialize the algorithm and a large flexibility is allowed in deforming it until a viable solution is obtained.

Thanks to IIT (Italian Institute of Technology) for support through the Seed project Dream and through the Joint Sapienza-IIT nano-medicine Lab.

Our group, ChimeraBio, in Sapienza@Rome, working on bio-oriented problems:

Faculty:

EM (Sapienza), Andrea De Martino (IPCF-CNR@Sapienza), Daniele De Martino (IIT@Sapienza, NanoMedicine Joint Lab)

Post Doctoral Fellows:

Fabrizio Capuani (IIT), Benedetta Cerruti (IIT)

PhD Students:

Matteo Figliuzzi, Araks Martirosyan, Matteo Mori, Alessandro Seganti

Master Students:

Simona Colabrese, Guglielmo De Concini

Former coworkers

Silvia Grigolon (former master student, now PhD student in Paris Sud), Guido Uguzzoni (former master student, now PhD student in Parma), Francesca Di Patti (former IIT Postdoc, now Postdoc in Florence)