Biological Design Automation for Optimal Cell Factories

Giuseppe Nicosia
Computational Synthetic & Systems Biology Lab
Dept. of Mathematics & Computer Science
University of Catania, Italy
www.dmi.unict.it/~nicosia/

BioCAD tools and algorithms

1. Optimisation
   1. Global/Local Optimisation
   2. Single/Multi-objective Optimisation (discrete for the genes and continuous for fluxes)
   3. $\epsilon$-dominance Analysis

2. Sensitivity Analysis (SA)
   1. Reaction-oriented SA (RoSA)
   2. Species-oriented SA (SoSA)
   3. Pathway-oriented SA, discrete (Gene Sets) or continuous (Fluxes) (PoSA)

3. Robustness Analysis (RA)
   1. Global RA
   2. Local RA
   3. Glocal RA
   4. Pathway-oriented RA

4. Identifiability Analysis (Genotype-Phenotype relationships)

Requests in Electronic Design Automation VS. Biological Design Automation

- Which are the most important parameters/parts/modules of the given Device/Circuit/System*?
- How many Feasible Devices/Circuits/Systems?
- How many Optimal and/or Suboptimal Devices/Circuits/Systems?
- How many Pareto Optimal Devices/Circuits/Systems?
- How many Robust Pareto Optimal Devices/Circuits/Systems?
- Which is the set of Robust Pareto Optimal Devices/Circuits/Systems that can be successfully manufactured?

* Device = gene/protein/enzyme
Circuit = pathway/organelle
Board/Chip = cell
System (system-of-systems) = tissue/organ/organism

10^7 to 10^8 candidate solutions/strains/pathways
Mixed Integer-Discrete-Continuous Constrained Multi-Objective Optimization Problem

Find

\[ X = \{x_1, x_2, \ldots, x_n\} = \left[ X^{(i)}, X^{(d)}, X^{(c)} \right]^T \]

To Minimize/maximize

\[ f_m(x), \quad m = 1, 2, \ldots, M; \]

Subject to

\[ g_j(x) \geq 0, \quad j = 1, 2, \ldots, J; \]

\[ h_k(x) = 0, \quad k = 1, 2, \ldots, K; \]

\[ x_i^{(L)} \leq x_i \leq x_i^{(U)} \quad i = 1, \ldots, N. \]

Where \( X^{(i)}, X^{(d)}, X^{(c)} \) denotes feasible subsets of integer, discrete and continuous variables respectively. While both integer and discrete variables have a discrete nature, only discrete variables can assume floating point values (they are often unevenly spaced): \([L_i, U_i, S_i]\)

Integer and discrete variables required different handling.

If a solution \( x \) satisfies all of the \((J+K)\) constraints and all of the \(2N\) variable bounds, it is known as a feasible solution.

**Design variables:** fluxes and/or gene sets, or Down- and Up- Regulation of Enzymes

**Objective functions:** Biomass vs. ATP, Biomass vs. Succinate (Ethanol, 1,4-Butanediol)

**Constraints:** \( O_2 = 0, \; GLC \geq 10, \; 5 \leq Ca \leq 10, \; Ph \) value

**Acetate** = + 66% with respect to Wild Type

$y_{34} = 1$
$y_{784} = 1$
$y_{432} = 1$

$y_{63} = 1$
$y_{222} = 1$
$y_{562} = 1$
$y_{24} = 1$

Knockout = 3

Knockout = 4

Acetate = + 130% with respect to Wild Type

**Genetic Design via MOO**
Algorithm 1 optBioCAD Pseudo-code.

1: optBioCAD \((model, d, dup, \tau_B, \rho, \beta, s_a)\)
2: \(t \leftarrow 0\);
3: \(BC_{arch} \leftarrow Create\_Archive(s_a)\);
4: \(P^{(t)} \leftarrow Initialise(d)\);
5: Evaluate\((P^{(t)}, model)\);
6: EvaluateConstraints\((P^{(t)}, model)\);
7: \textbf{while} \neg \text{Stop\_Condition}(t) \textbf{do}
8: \quad P_{cop} \leftarrow Copying\((P^{(t)}, dup)\);
9: \quad P_{LS} \leftarrow Local\_Search\_Operator\((P_{cop}, \rho)\);
10: \quad P_{GS} \leftarrow Global\_Search\_Operator\((P_{hyp}, \beta)\);
11: \quad Evaluate\((P_{GS}, model)\);
12: \quad EvaluateConstraints\((P_{GS}, model)\);
13: \quad Diversity\_Enforcing\((P^{(t)}, P_{GS}, \tau_B)\);
14: \quad BC_{arch} \leftarrow (BC_{arch} \cup P^{(t)} \cup P_{GS});
15: \quad P^{(t+1)} \leftarrow Selection\((P^{(t)}, P_{GS}, BC_{arch})\);
16: \quad t \leftarrow t + 1;
17: \textbf{end while}
18: \textbf{return} \((P^{(t)})\); /* output the best \(d\) candidate solutions */
Design Flow – 1/2

Nicosia et al, PLOS ONE '15.
The Overall Design Flow – part 2/2

Nicosia et al, PLOS ONE ’15.
E. coli Designing
We test the computational framework in the genome-scale metabolic network of *E. coli* iAF1260 (Feist et al., 2007)

- 2382 reactions (299 exchange fluxes)
- 1039 metabolites
- 913 gene sets (1040 in *E. coli* 2011)
- 1260 genes
- 36 pathways
- 3 compartments

**Aim:** maximise a metabolite of interest (e.g., acetate/succinate), and simultaneously ensure the *biomass formation*, with the *minimum knockout cost*.
A) Acetate vs. Biomass maximisation in different environmental conditions
B) Succinate vs. Biomass maximisation in different environmental conditions
Pareto Front obtained by the 2-objective optimisation to maximise the acetate production and biomass formation in E. coli (parameters of GDMO: pop=1000, gen=1500).

Anaerobic condition, GLC = 10 mmolh⁻¹gdW⁻¹, values in brackets represent the variation with respect to the wild type configuration.

<table>
<thead>
<tr>
<th>Acetate</th>
<th>Biomass</th>
<th>Knockout</th>
<th>Genes</th>
<th>Pathways</th>
<th>Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3014</td>
<td>0.23106</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.1962</td>
<td>0.096328</td>
<td>1</td>
<td></td>
<td>Pyruvate Metabolism, Cofactor and Prosthetic Group Biosynthesis</td>
<td>acetaldehyde dehydrogenase (acylating)</td>
</tr>
<tr>
<td>(71.0095%)</td>
<td>(-58.2994%)</td>
<td>1</td>
<td>(((b0351) OR (b1241)))</td>
<td>D-alanine transaminase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(((b0870) OR (b2551)))</td>
<td>alanine transaminase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b3744)</td>
<td>Glucose-6-phosphate isomerase</td>
<td>L-allo-Threonine Aldolase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b2500)</td>
<td>Phosphoribosylglycinaminitransferase</td>
<td>asparagine synthetase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b4025)</td>
<td>Inorganic Ion Transport and Metabolism</td>
<td>Threonine aldolase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(((b2987) OR (b3493)))</td>
<td>Glycine and Serine metabolism</td>
<td>asparagine synthetase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b4388)</td>
<td>Tyrosine, Tryptophan, and Phenylalanine Metabolism</td>
<td>acetaldehyde dehydrogenase (acylating)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b3708)</td>
<td></td>
<td>D-alanine transaminase</td>
</tr>
<tr>
<td>13.7911</td>
<td>0.13035</td>
<td>3</td>
<td>(((b0351) OR (b1241)))</td>
<td>Threonine and Lysine Metabolism, Glycine ar L-allo-threonine dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>(68.13%)</td>
<td>(-43.5725%)</td>
<td>3</td>
<td>(((b0351) OR (b1241)))</td>
<td>D-serine dehydrogenase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(b1539)</td>
<td></td>
<td>L-serine dehydrogenase</td>
</tr>
<tr>
<td>18.4549</td>
<td>0.096233</td>
<td>11</td>
<td>(((b0351) OR (b1241)))</td>
<td>Pyruvate Metabolism, Transport Inner Membrane</td>
<td>D-lactate transport via proton symport (periplasm)</td>
</tr>
<tr>
<td>(122.3111%)</td>
<td>(-58.3406%)</td>
<td>11</td>
<td>(((b0351) OR (b1241)))</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(((b2975) OR (b3603)))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GDMO (us) vs. GDLS (G. Church)

**Black**: Pareto front obtained by GDMO.

**Red**: optimal results obtained by GDLS [G. Church, Lun et al., Molecular Systems Biology, 2009] [G. Church et al, PLOS Comp. Biol. 2013]
OptFlux *(Rocha et al, BMC Bioinf ’08)*  
OptGene *(Patil et al, BMC Bioinf ‘05)*  
GDLS *(Lun, G. Church et al, MSB ‘09)*  
OptKnock *(Bugard et al, Biotech & Bioeng’03)*  
GDBB *(Lun et al, Bioinformatics’12)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetate</td>
<td>8.30</td>
<td>15.129</td>
<td>15.138</td>
<td>15.914</td>
<td>n.a.</td>
<td>12.565</td>
<td>13.797</td>
<td>19.150</td>
<td>n.a.</td>
<td>10.610</td>
</tr>
<tr>
<td>Succinate</td>
<td>0.077</td>
<td>10.007</td>
<td>9.874</td>
<td>n.a.</td>
<td>9.727</td>
<td>9.069</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Biomass [1/h]</td>
<td>0.23</td>
<td>n.a.</td>
<td>n.a.</td>
<td>0.0500</td>
<td>0.0500</td>
<td>0.1181</td>
<td>0.1165</td>
<td>0.1296</td>
<td>0.053</td>
<td>0.087</td>
</tr>
<tr>
<td>K cost</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>14</td>
<td>26</td>
<td>54</td>
<td>53</td>
<td>3</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>


**FBA-GPR Models of *E. Coli* (2007 & 2011)**
Robustness Analysis - E. coli iAF1260

- **GR**: Global Robustness [1]
- **LR**: Local Robustness [1]
- **R**: Glocal Robustness [2]
- **PoRA**: Pathway-oriented Robustness [3]

<table>
<thead>
<tr>
<th>Wild Type</th>
<th>OptFlux</th>
<th>OptGene</th>
<th>GDLS</th>
<th>GDMO</th>
<th>OptKnock</th>
<th>GDMO</th>
<th>GDMO</th>
<th>GDMO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Succinate</td>
<td>0.077</td>
<td>10.007</td>
<td>9.874</td>
<td>n.a.</td>
<td>9.727</td>
<td>9.069</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Biomass [1/h]</td>
<td>0.23</td>
<td>n.a.</td>
<td>n.a.</td>
<td>0.0500</td>
<td>0.0500</td>
<td>0.1181</td>
<td>0.1165</td>
<td>0.1296</td>
</tr>
<tr>
<td>K cost</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>14</td>
<td>26</td>
<td>54</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>GR (%)</td>
<td>54.76/53.68</td>
<td>n.a.</td>
<td>n.a.</td>
<td>13.76</td>
<td>16.6</td>
<td>43.24</td>
<td>43.08</td>
<td>45.32</td>
</tr>
<tr>
<td>LR(%)</td>
<td>54.0/54.67</td>
<td>n.a.</td>
<td>n.a.</td>
<td>16.0</td>
<td>21.33</td>
<td>40.0</td>
<td>40.60</td>
<td>39.33</td>
</tr>
<tr>
<td>R</td>
<td>1.30/1.34</td>
<td>n.a.</td>
<td>n.a.</td>
<td>1.45</td>
<td>1.45</td>
<td>1.18</td>
<td>1.02</td>
<td>0.78</td>
</tr>
<tr>
<td>PoRA (%)</td>
<td>100.0/99.33</td>
<td>n.a.</td>
<td>n.a.</td>
<td>19.33</td>
<td>28.67</td>
<td>87.33</td>
<td>76.67</td>
<td>81.33</td>
</tr>
</tbody>
</table>

The Robustness estimates how robust is a strain when it undergoes small perturbations.

Flux Design in E. coli iAF1260
Power law & specific operational regions

Overproduction of BDO in *E. coli*

optBioCAD/GDMO vs. Genomatica Inc. (and BASF)

- 2 Genome-scale metabolic networks of *E. coli* (iJR904 and iJO1366) 
  - 931/2251 reactions
  - 625/1136 metabolites
  - 904/1366 genes
  - 729/1041 enzymes

- Synthetic pathway of **1,4 butanediol** (BDO)

- Genetic and Flux Design to overproduce BDO in *E. coli*

- BioCAD software

BDO is an inorganic compound; it is used industrially as a solvent and in the manufacture of some types of plastics, elastic fibers and polyurethanes.

BDO currently is manufactured entirely from petroleum-based feedstocks.

Inclusion of BDO synthetic pathway in *E. coli* model – **BDO production**

In silico analysis to find the key actors in **BDO overproduction**

## BioCAD results

<table>
<thead>
<tr>
<th>Deletions</th>
<th>EX_14btd(e)</th>
<th>EX_14btd(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHeR LDH_D</td>
<td>0,378</td>
<td>5,72</td>
</tr>
<tr>
<td>ADHeR LDH_D MDH</td>
<td>0,296</td>
<td>7,72</td>
</tr>
<tr>
<td>ADHeR LDH_D CO2t</td>
<td>0,363</td>
<td>6,88</td>
</tr>
<tr>
<td>ADHeR LDH_D PTAr</td>
<td>0,219</td>
<td>6,60</td>
</tr>
<tr>
<td>ADHeR LDH_D ACKr</td>
<td>0,219</td>
<td>6,60</td>
</tr>
<tr>
<td>ADHeR LDH_D THD2</td>
<td>0,367</td>
<td>6,56</td>
</tr>
<tr>
<td>ADHeR LDH_D PGI</td>
<td>0,195</td>
<td>6,21</td>
</tr>
<tr>
<td>ADHeR LDH_D TPI</td>
<td>0,199</td>
<td>6,13</td>
</tr>
<tr>
<td>ADHeR LDH_D FUM</td>
<td>0,250</td>
<td>6,04</td>
</tr>
<tr>
<td>ADHeR LDH_D C140SN</td>
<td>0,307</td>
<td>5,94</td>
</tr>
<tr>
<td>ADHeR LDH_D TKT2</td>
<td>0,375</td>
<td>5,86</td>
</tr>
<tr>
<td>ADHeR LDH_D GLCpts</td>
<td>0,333</td>
<td>5,83</td>
</tr>
<tr>
<td>ADHeR LDH_D GLUDy</td>
<td>0,352</td>
<td>5,78</td>
</tr>
<tr>
<td>ADHeR LDH_D RPE</td>
<td>0,376</td>
<td>5,78</td>
</tr>
<tr>
<td>ADHeR LDH_D PFK</td>
<td>0,360</td>
<td>5,76</td>
</tr>
<tr>
<td>ADHeR LDH_D FBA</td>
<td>0,360</td>
<td>5,76</td>
</tr>
<tr>
<td>ADHeR LDH_D FRD3</td>
<td>0,368</td>
<td>5,74</td>
</tr>
<tr>
<td>ADHeR LDH_D NADH8</td>
<td>0,368</td>
<td>5,74</td>
</tr>
<tr>
<td>ADHeR LDH_D CBMK2</td>
<td>0,374</td>
<td>5,73</td>
</tr>
<tr>
<td>ADHeR LDH_D MDH FORt</td>
<td>0,140</td>
<td>15,17</td>
</tr>
<tr>
<td>ADHeR LDH_D PFLi MDH</td>
<td>0,140</td>
<td>15,17</td>
</tr>
<tr>
<td>ADHeR LDH_D MDH ATPS4r</td>
<td>0,203</td>
<td>12,17</td>
</tr>
<tr>
<td>ADHeR LDH_D PGDHY PGI</td>
<td>0,131</td>
<td>11,86</td>
</tr>
<tr>
<td>ADHeR LDH_D EDA PGI</td>
<td>0,131</td>
<td>11,86</td>
</tr>
<tr>
<td>ADHeR LDH_D FUM ACKr</td>
<td>0,127</td>
<td>10,92</td>
</tr>
</tbody>
</table>

**203 solutions/strains**

- **ADHeR** : $\text{accoa} + (2) \text{h} + (2) \text{nadh} \Leftrightarrow \text{coa} + \text{etoh} + (2) \text{nad}$
- **LDH_D** : $\text{lac-D} + \text{nadh} \Leftrightarrow \text{h} + \text{nadh} + \text{pyr}$
- **MDH** : $\text{mal-L} + \text{nadh} \Leftrightarrow \text{h} + \text{nadh} + \text{oaa}$
- **PFLi** : $\text{coa} + \text{pyr} \rightarrow \text{accoa} + \text{for}$

**Knockout cost = 7**


BioCAD results

- MOO to maximise BDO production and biomass formation in synthetic E. coli model iJR904\textsuperscript{[1]}

Pareto fronts obtained by GDMO algorithm\textsuperscript{[2]}

- C=10 (12836 Pareto strains)
- C=50 (49876 Pareto strains)

versus 203 solutions

C is the maximum knockout number allowable

\textsuperscript{[1]} Reed et al. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR) Gen. Biol. 2003
\textsuperscript{[2]} Nicosia et al. (2012) Robust Design of Microbial Strains. Bioinformatics

**BioCAD results**

12836 Pareto optimal strains (C=10) vs BDO=15,1660 & Biomass=0,1398

<table>
<thead>
<tr>
<th>BDO</th>
<th>Biomass</th>
<th>kcost</th>
<th>Gene knockout</th>
<th>Pathway</th>
<th>Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.1683</td>
<td>0.13977 (-66.6334%)</td>
<td>6</td>
<td>b1241</td>
<td>Alternate Carbon Metabolism, Pyruvate Metabolism</td>
<td>LCADi, ADHEr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b3236</td>
<td>Citrate Cycle (TCA)</td>
<td>MDH</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b2975, b3603</td>
<td>Transport (Extracellular)</td>
<td>D-LACt2, GLYCLTt2r, L-LACt2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b2492, b0904</td>
<td>Transport (Extracellular)</td>
<td>FORt</td>
</tr>
</tbody>
</table>

BioCAD results

49876 Pareto optimal strains (C=50) vs BDO=15,1660 & Biomass=0,1398

<table>
<thead>
<tr>
<th>Genes</th>
<th>Pathway</th>
<th>Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHEr b1241</td>
<td>Alternate Carbon Metabolism, Pyruvate Metabolism</td>
<td>LCADI, ADHEr</td>
</tr>
<tr>
<td>LDH_D b2133 OR b1380</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDH b3236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDH b3236</td>
<td>(b0902 AND b0903 AND b2579) OR (b3114) OR</td>
<td></td>
</tr>
<tr>
<td>PFLi b3951ANDb3952</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BioCAD results

- **GDMO** considers the gene-protein-reaction (GPR) mapping and each bit of y is related to a gene set

- A gene set can be formed by
  - a single gene
  - more genes linked by a Boolean relationship

- The **method in Yim et al. 2011 does not consider the GPR map, and turns off/on the flux of the reactions**

- In order to compare our method with Yim et al. 2011 results, **we also perform knockout research in the reaction space.**
BioCAD results

Down- and Up- Regulation of Enzymes
Myristoyl-CoA Optimization for the iAF1260 E. coli: Fatty acids production

<table>
<thead>
<tr>
<th></th>
<th>BioCAD[**]</th>
<th>Redirector [*]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomass of the best strain</td>
<td>0.17 (+<strong>21.43%</strong>)</td>
<td>0.14</td>
</tr>
<tr>
<td>Myristoyl-CoA of the best strain</td>
<td>1.62 (+<strong>5.19%</strong>)</td>
<td>1.54</td>
</tr>
<tr>
<td>CPU time [s]</td>
<td><strong>2400s</strong></td>
<td>15200s</td>
</tr>
</tbody>
</table>

Anaerobic condition. Glucose uptake rate 8mmol-1 gDW-1.

Notable strain we obtain dominates the one obtained by G. Church et al [*].


Designing BioPlastic: Polylactic acid from Lactate acid production in S. cerevisiae
Lactate Production in S. cerevisiae by Redesigning the Metabolic Networks

- Organism: Saccharomyces cerevisiae S288c
- Model: iMM904
- Genome: PRJNA128
- Metabolites: 1226
- Reactions: 1577
- Genes: 905
- Database: http://bigg.ucsd.edu/models/iMM904/
- Publication PMID: 19321003
Lactate Production and Biomass Optimization in S. cerevisiae
A complete Computational Flow for Biological Design Automation

**SA:** SoSA, RoSA, PoSA

**IA:** Genotype-Phenotype relationships

**RA:** LR, GR, GlocalR

**Opt:** SOO, MOO

**Models:** ODEs, DAEs, FBA, FBA-GPR.

**Systems:** Pathways, Organelles & Organisms

**Results**

1. **1,4 Butanediol Production in E. coli**
2. **ATP maximization in the Mitochondrion**
3. Down- and Up- regulation of Enzymes for **Fatty acids production**
4. BioPlastic production by Engineered Yeast
Acknowledgments

Jole Costanza, Center for Genomic Science@IIT, Milano, Italy

Giovanni Carapezza, Nerviano Medical Sciences, Catania, Italy

Piero Conca, CNR – Catania, Italy

Giorgio Jansen, University of Catania, Italy

Andrea Patanè, University of Catania, Italy

Andrea Santoro, University of Catania, Italy
4th International Synthetic & Systems Biology Summer School – SSBSS 2017

July 17-21, 2017, Robinson College
University of Cambridge, UK

http://www.taosciences.it/ssbss-2017/

FB: https://www.facebook.com/ssbss.school/