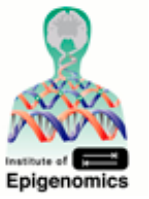
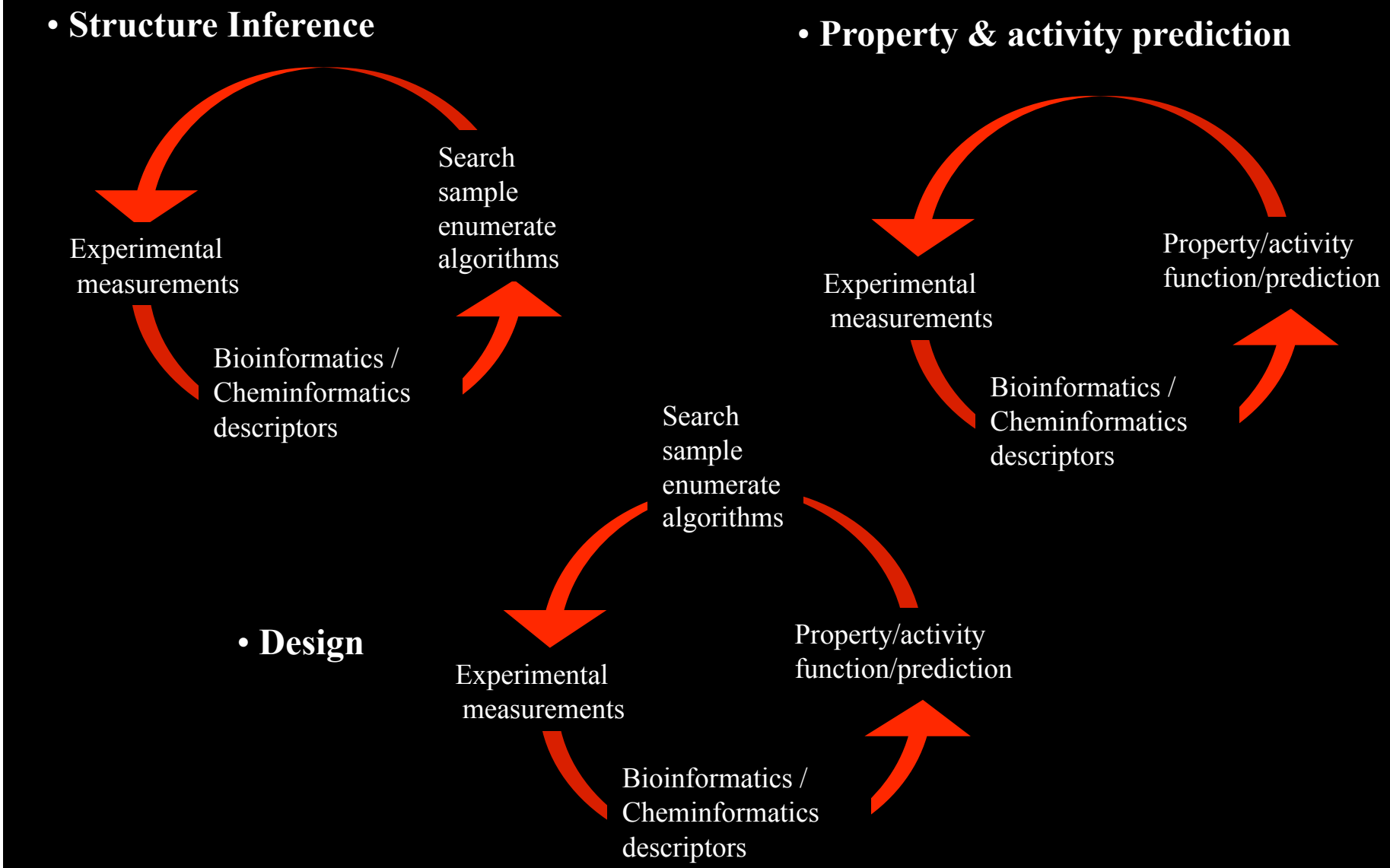
A systems chemical biology approach to infer and design metabolic networks

Jean-Loup Faulon,
Evry University, UniverSud Paris
Epigenomics Program, Genopole®, CNRS
523 Terrasses de l'Agora
91034 ÉVRY cedex
jean-loup.faulon@epigenomique.genopole.fr
http://jfaulon.wikispaces.com



Inference and design of chemicals and biological sequences and networks



Prediction and inference using residue and atom neighborhood

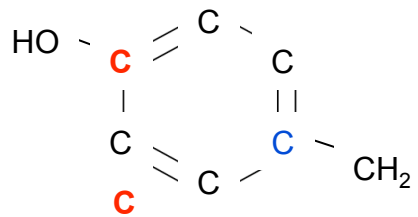
Residue neighborhood (strings, k-mers)

…EKKAIPQEKK…

k=1 (h=0) P
k=3 (h=1) IPQ
k=5 (h=2) AIPQE

Leslie *et al.* PSB 2002

Atom neighborhood (atom signature)



h=0 C
h=1 C(C =C O)
h=2 C(C(=C) =C(C) O(H))
h=3 C(C(=C(C))=C(C(=C))O(H))

Faulon *et al.* J. Chem. Info. 1994

canonization:
Faulon *et al.* J. Chem. Info. 2004

Protein string spectrum

$^{k=3}\sigma$(EKKAIPQEKK) =

EKK KKA KAI AIP IPQ PQE QEK
(2 1 1 1 1 1 1)

Sequence assembly
Function prediction

Molecular signature

$^{h=1}\sigma$(Tyrosine) =

C (C =C O) C (C =C) C (C C =C) O(C H) C (C H H)
(1 4 1 1 1)

Structure elucidation
Property prediction

QSAR & kernel methods to detect homology and similarity

Chemical structure-activity relationships: examples with atom neighborhood

- Atom neighborhood/signature (1994)

- Tripos' Hologram HQSAR (1996)

-Filimonov & Poroikov multilevel atom-neighborhood PASS descriptors (2001)

- Solov'ev, Varnek , & Wipff, sequences & augmented atoms (2001)

- Glen & Bender atom environment (2003)

HIV-1 Protease Inhibitors IC_{50}, 130 compounds in training set								
Height	0	1	2	3	4	5	6	7
q^2 (MLR)	0.75	0.80	0.77	0.81	0.77	0.69	0.65	0.66
Log *P*, 1,000 compounds in training set								
q^2 (MLR)	0.61	0.83	0.78	0.58	0.31	0.23	0.16	0.11
Glass Transition T_g 262 polymers in training set								
q^2 (MLR)	0.65	0.78	0.75	0.81	0.65	0.60	0.44	0.32
q^2 (Linear SVR)	0.64	0.80	0.81	0.81	0.73	0.22	0.29	0

Faulon et al. JMGM 2002, JCICS 2003-a&b, JCAMD 2005, Ind. Eng. Chem. Res 2005, JCIM 2006, JMGM 2009

Strings spectra are prevalent in many computational biology applications

Protein structure prediction

- SCOP family prediction *Leslie et al., PSB2002*.
- Beta sheet ordering: *Brown, Martin, Strauss, Faulon, J. Mol. Model., 2006*

Protein functional annotation

- Subcellular localization prediction: *Hua & Sun, Bioinformatics 2001*
- TIGR Protein function prediction: *Martin, Brown, Faulon, Genome-To-Life review 2005*
- Catalytic activity prediction (EC number second level): *Kunik et al. CSB2005*
- Enzyme (EC) number prediction: *Faulon et al. Bioinformatics 2008*
- Phosphorylation site prediction, *Gray et al. Annals Operation Research 2008*

Network inference: Protein-protein interactions (PPI), Enzyme-metabolite interactions, Traget-ligand interactions, DNA binding site predictions

- Protein-DNA: *Olman et al. J. Bioinform Comput Biol* 2003
- PPI: *Martin, Roe, Faulon, Bioinformatics 2005*
- Target-ligand interaction*: Oprea, Tropsha , Faulon, Nature Chemical Biology 2007*
- Enzyme-metabolite interactions, *Bioinformatics* 2008

Predicting protein-ligand interactions

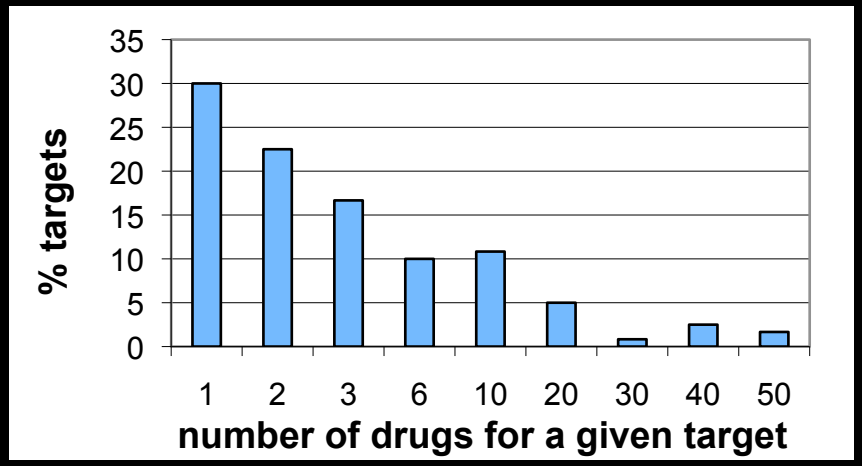
Why:

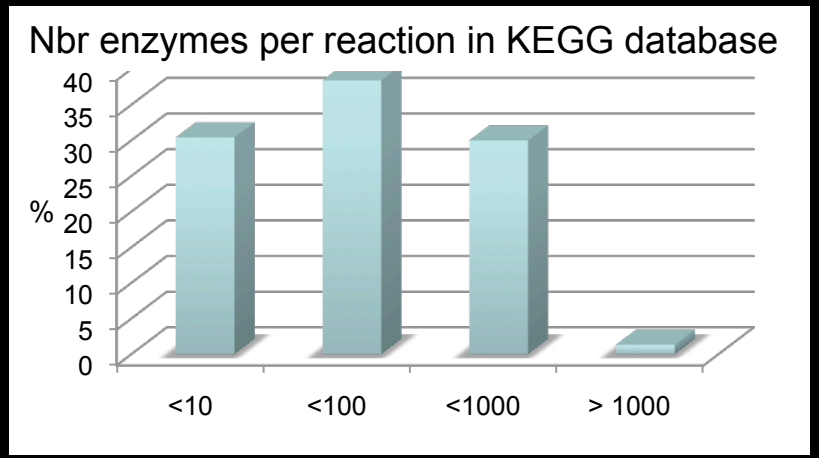
- **Signaling network inference (protein-protein, protein-ligand interactions)**
- **Drug discovery**
  - **Drug specificity, polypharmacology, promiscuous inhibitors**
- **Metabolic network inference, genome annotation**
  - **Enzyme promiscuity, searching enzymes for novel reactions**

How:

- **Chemical informatics**
  - **Classical QSAR predicts compounds binding to a given target**
- **Bioinformatics**
  - **Sequence homology predicts enzyme catalytic activity**

Training sets are required for each target or each metabolic reaction





Predicting protein-ligand interactions

Why:

- **Signaling network inference (protein-protein, protein-ligand interactions)**
- **Drug discovery**
  - **Drug specificity, polypharmacology, promiscuous inhibitors**
- **Metabolic network inference, genome annotation**
  - **Enzyme promiscuity, searching enzymes for novel reactions**

How:

- **Chemical informatics**
  - **Classical QSAR predicts compounds binding to a given target**
- **Bioinformatics**
  - **Sequence homology predicts enzyme catalytic activity**

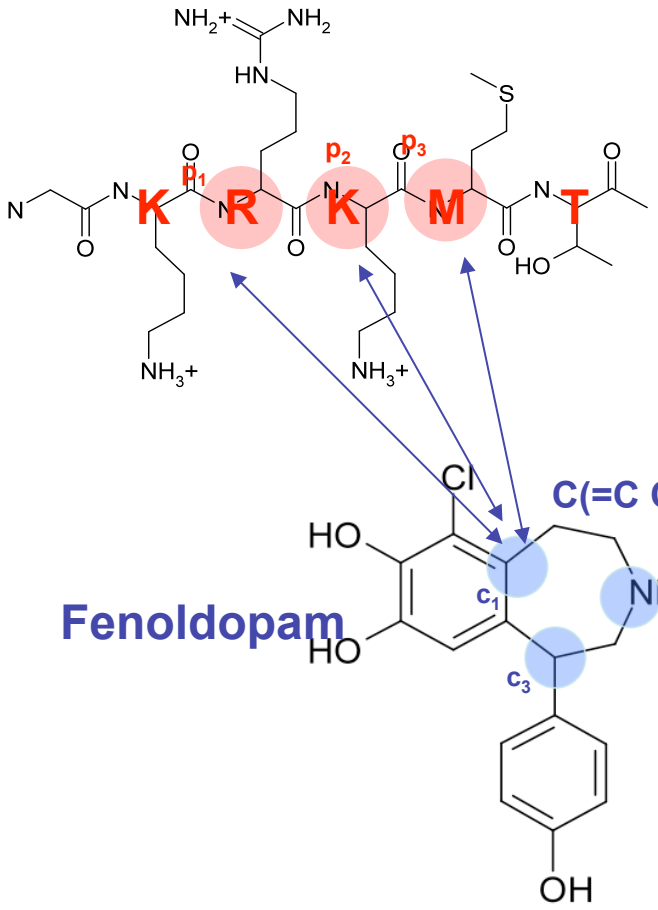
Training sets are required for each target or each metabolic reaction

Our work:

A kernel method where learning is performed on the pair of interacting partners

Can use heterogeneous datasets found in Kegg, MetaCyc, BRENDA, MDDR, PubChem, DrugBank,..

Representing protein-ligand interaction: the tensor product



DOPAMINE RECEPTOR DRD5

MLPPGSNGTAYPGQFALYQQLAQGNAVGGSAGAPPLGPSQVVTACLLTLLIIWTLLGNVL
VCAAIVRSRHLRANMTNVFIVSLAVSDLFVALLVMPWKAVAEVAGYWPFGAFCDVWVAFD
IMCSTASILNLCVISVDRYWAISRPFRYKRKMTQRMALVMVGLAWTLSILISFIPVQLNW
HRDQAASWGGLDLPNNLANWTPWEEDFWEPDVNAENCDSSLNRTYAISSSLISFYIPVAI
MIVTYTRIYRIAQVQIRRISSLERAAEHAQSCRSSAACAPDTSLRASIKKETKVLKTLSV
IMGVFVCCWLPFFILNCMVPFCSGHPEGPPAGFPCVSETTFDVFVWFGWANSSLNPVIYA
FNADFQKVFAQLLGCSHFCSRTPVETVNISNELISYNQDIVFHKEIAAAYIHMMPNAVTP
GNREVDNDEEEGPFDRMFQIYQTSPDGDPVAESVWELDCEGEISLDKITPFTPNGFH

Protein: KRK,…,RKM,…,KMT
p_1 p_2 p_3

Protein x Ligand :

KRK C(=C C C),…, KRK N(C C H),…,KRK C(C C C H)
RKM C(=C C C),…, RKM N(C C H),…, RKM C(C C C H)
KMT C(=C C C),…, KMT N(C C H),…, KMT C(C C C H)

Ligand: C(=C C C),…, N(C C H),…,C(C C C H)
c_1 c_2 c_3

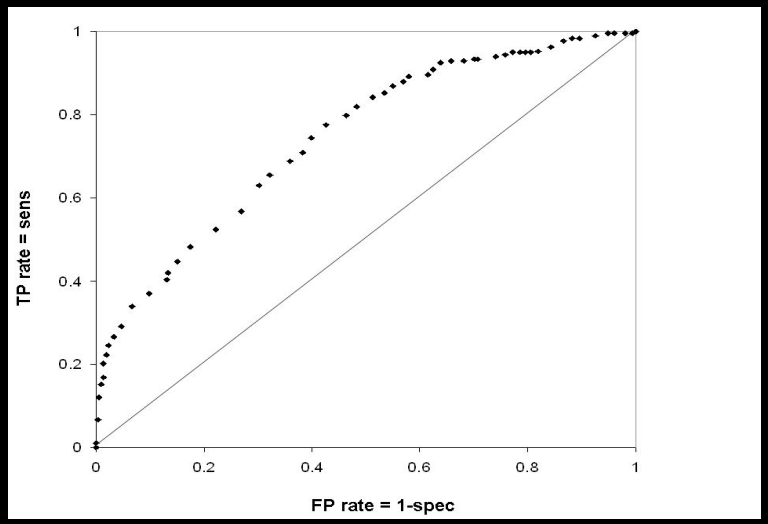
$$\sigma(P) = (p_1, p_2, \ldots, p_n)$$
$$\sigma(C) = (c_1, c_2, \ldots, c_n)$$
$$\sigma(P \otimes C) = (p_1c_1, \ldots, p_1c_n, p_2c_1, \ldots, p_2c_n, \ldots, p_nc_1, \ldots, p_nc_n)$$

Martin, Roe, Faulon, Bioinformatics 2005
Faulon et *al.*, Bioinformatics 2008
Martin, Brown, Faulon, in *Advances in Biochemical Engineering/Biotechnology,, Springer-Verlag.* 2008

Finding new interactions?

KEGG database (Nov 2007) comprises 873 drug- target pairs between 121 human targets and 551 drugs. Tensor product gives accuracy > 90% with 5-fold cross validation

Independent test set composed of 298 interactions between FDA approved drugs and targets (extracted from DrugBank database). Only 32 interactions are in training set (KEGG)



Accuracy = 67%
Area Under the ROC Curve = 0.74

class	Nbr such examples	Drug-target example
Target & drug not in training	94	Pinozide – Opioid receptor OPRD1
Target not in training	16	Fenoldopam – Dopamine receptor DRD5
Drug not in training	145	Perphenazine -Dopamine receptor DRD2

These predictions cannot be made with classical QSAR or sequence homology

Predicting protein-ligand interactions

Why:

- **Signaling network inference (protein-protein, protein-ligand interactions)**
- **Drug discovery**
  - **Drug specificity, polypharmacology, promiscuous inhibitors**
- **Metabolic network inference, genome annotation**
  - **Enzyme promiscuity, searching enzymes for novel reactions**

How:

- **Chemical informatics**
  - **Classical QSAR predicts compounds binding to a given target**
- **Bioinformatics**
  - **Sequence homology predicts enzyme catalytic activity**

- **Training sets are required for each target or each metabolic reaction**
- **Virtual screening requires 3D structures**

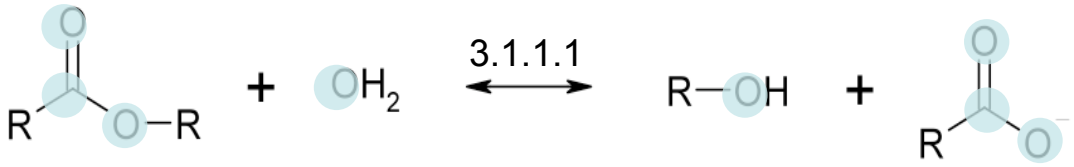
Our work:

- **A a kernel method where learning is performed on the pair of interacting partners**

- **Can use heterogeneous datasets found in Kegg, MDDR, PubChem, DrugBank,..**

Representing enzyme-reaction interaction

Faulon JCICS 1994 : σ(reaction) = σ(product) - σ(reactant)



σ(3.1.1.1)= σ(Alcohol)+σ(Carboxylate)-σ(Carboxylic ester)-σ(H_2O)

esterase YpfH (*E. coli* UTI89)

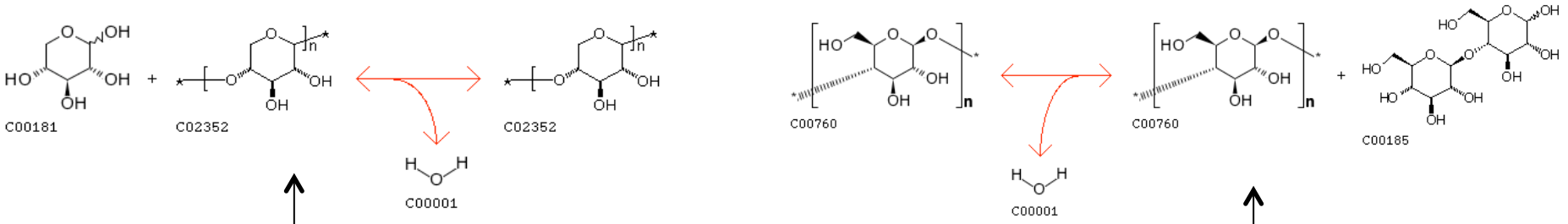
MKHDHFVVQSPDKPAQQLLLLFHGVGDNPVAMG
EIGSWFAPLFPDALVVSVGGAEPSGNP
AGRQWFSVQGITEDNRQARVNAIMPTFIETVRYW
QKQSGVGANATALIGFSQGAIMALES
IKAEPGLASRVIAFNGRYASLPETASTATTIHLIHGG
EDPVIDLAHAVAAQEALISAGGD
VTLDIVEDLGHAIDNRSMQLALDHLRYTIPKHYFD
EALSGGKPGDDDVIEMM

```
  0.0   [O=]([C=])
  1.0   [O-]([C])
 -1.0   [O]([R][C])
  1.0   [O]([R][H])
 -1.0   [O]([H][H])
  1.0   [C]([=O][O-][R])
 -1.0   [C]([=O][O][R])
```

```
1 MKH
1 KHD
1 HDF
...
```

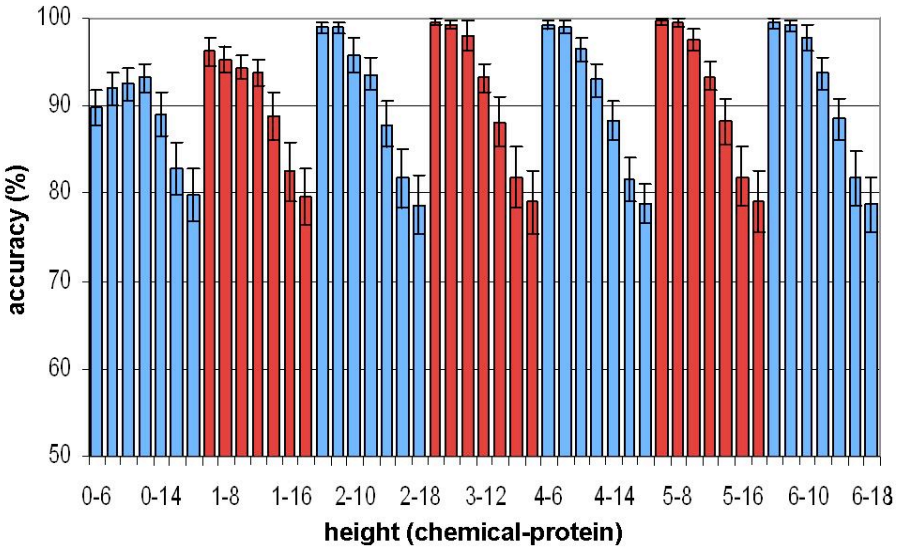
```
 1.0  [O-]([[C])MKH        1.0  [O-]([C])KHD        -1.0  [O-]([C])HDF ...
-1.0  [O]([R][C])MKH      -1.0  [O]([R][C])KHD      -1.0  [O]([R][C])HDF ...
 1.0  [O]([R][H])MKH       1.0  [O]([R][H])KHD       1.0  [O]([R][H])HDF ...
-1.0  [O]([H][H])MKH      -1.0  [O]([H][H])KHD      -1.0  [O]([H][H])HDF ...
 1.0  [C]([=O][O-][R])MKH 1.0  [C]([=O][O-][R])KHD 1.0  [C]([=O][O-][R])HDF ...
-1.0  [C]([=O][O][R])MKH -1.0  [C]([=O][O][R])KHD -1.0  [C]([=O][O][R])HDF ...
```

Probing enzyme promiscuity



Xylan hydrolysis
(3.2.1.8)
52 sequences in
KEGG

Training set



Cellulose hydrolysis
(3.2.1.4)
175 sequences in
KEGG

Test set

- **Cellulases from *T. fusca*, *S. degradans*, *X. campestris* found to catalyse xylans**
  ***T. fusca* cellulases bind to Chitin (*Li, Wilson, Biotech. & Bioengineering, 2008*)**

Finding enzyme catalyzing novel reactions

Test set composed of enzymes and reactions accepted by NC-IUBMB in Sept. 2006 (none were in training set)

Training set KEGG database

855,722 interactions
3,905 reactions
255,304 enzymes

EC class	# Positive Pairs	Acc.	Prec.	Sens.	Spec.
EC 1.1.1.290 4-phosphoerythronate dehydogenase	59	88.7	82.6	98.3	79.1
EC 1.13.11.52 indoleamine 2,3-dioxygenase	13	76.9	76.9	76.9	76.9
EC 1.13.11.53 acireductone dioxygenase (Ni2+-requiring)	11	86.4	83.3	90.9	81.8
EC 1.2.1.71 succinylglutamate-semialdehyde dehydrogenase	55	87.5	82.3	95.5	79.5
EC 1.2.1.72 erythrose-4-phosphate dehydrogenase	46	88.0	80.8	100.0	76.1
EC 1.8.4.11 peptide-methionine (S)-S-oxide reductase	390	79.5	99.8	59.1	99.9
EC 2.6.1.81 succinylornithine transaminase	21	81.0	72.7	100.0	61.9
EC 3.1.3.77 acireductone synthase	160	89.4	82.5	100.0	78.8
EC 3.3.2.9 microsomal epoxide hydrolase	17	84.3	82.3	88.2	80.4
EC 3.5.1.96 succinylglutamate desuccinylase	49	88.6	81.6	100.0	77.1
EC 3.5.3.23 N-succinylarginine dihydrolase	51	90.2	83.7	100.0	80.4
EC 4.2.1.109 methylthioribulose 1-phosphate dehydratase	12	87.5	80.0	100.0	75.0
Average		85.7±4.3	82.4±6.3	92.4±12.6	78.9±8.4

Faulon et al. Bioinformatics 2008

Sequence design

Designing sequence using inverse QSAR (i-QSAR)

QSAR:

- activity = *f* (descriptors)

Inverse-QSAR:

- descriptors = f^{-1}(activity)

- assemble sequences matching descriptors

i-QSAR: Design ICAM-1 binding sequences (1)

Peptides	IC_{50}	Peptides	IC_{50}
1. CLLRMRSAC	480	9. CLLRMRSVC	700
2. CILRMRSAC	190	10. CLLRMRSIC	580
3. CVLRMRSAC	>1000	11. CALRMRSIC	>1000
4. CLIRMRSAC	720	12. CLARMRSIC	>1000
5. CLVRMRSAC	>1000	13. CLLRARSIC	>1000
6. CLLKMRSAC	105	14. CLLRMASIC	>1000
7. CLLRMKSAC	90	15. CLLRMRAIC	710
8. CLLRMRSLC	>1000	16. CILKMKSAC	40

1. QSAR/QSPR analysis

47 descriptors in the training set

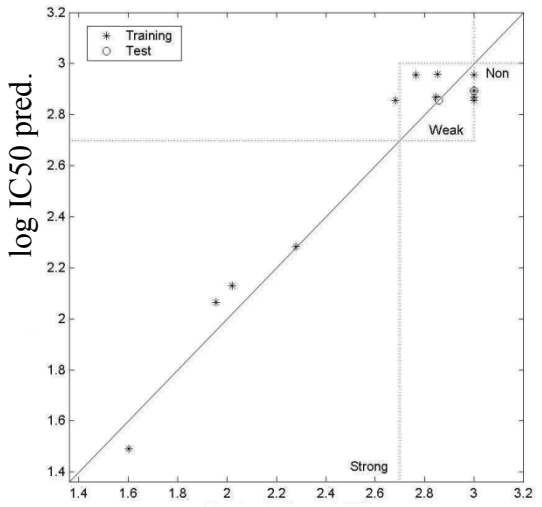
x1	C(CL)	x6	L(LR)	x11	R(MS)	x16	S(AK)	x21	I(LR)
x2	C(CV)	x7	L(IR)	x12	R(IM)	x17	S(AR)	x22	A(CS)
x3	C(CI)	x8	L(CI)	x13	M(RR)	x18	S(RV)	x23	K(MS)
x4	C(AC)	x9	L(KL)	x14	M(KR)	x19	I(CS)	x24	K(LM)
x5	L(CL)	x10	R(LM)	x15	S(IR)	x20	I(CL)	x25	V(CS)

.....

MRL, SVR,PLS,…

QSAR equation:

$$Log_{10}(IC_{50}) = 2.81 - 0.739\,x2 - 0.574\,x8 + 0.662\,x13 + 0.728\,x31 + 0.727\,x41 - 0.644\,x37$$



i-QSAR: Design ICAM-1 binding sequences (2)

1. QSAR/QSPR analysis

2. Integer equation builder

Descriptors in the training set

x1	C(CL)	x6	L(LR)	x11	R(MS)	x16	S(AK)	x21	I(LR)
x2	C(CV)	x7	L(IR)	x12	R(IM)	x17	S(AR)	x22	A(CS)
x3	C(CI)	x8	L(CI)	x13	M(RR)	x18	S(RV)	x23	K(MS)
x4	C(AC)	x9	L(KL)	x14	M(KR)	x19	I(CS)	x24	K(LM)
x5	L(CL)	x10	R(LM)	x15	S(IR)	x20	I(CL)	x25	V(CS)

.....

Constraint Equations

$$2.81 - 0.739\,x2 - 0.574\,x8 + 0.662\,x13 + 0.728\,x31 + 0.727\,x41 - 0.644\,x37 - \text{target-IC}_{50} = 0$$

-x18 + x25 = 0
-x2 + x25 = 0
-x11 + x15 + x17 + x18 = 0
x16 – x23 = 0
x15 – x19 = 0
x16 + x17 – x22 = 0
x10 + x11 + x12 – 2x13 – x14 = 0
-x6 – x7 + x10 = 0
x12 – x21 = 0

x9 – x24 = 0
x7 + x8 – x20 – x21 = 0
-x1 + x5 + x8 = 0
-x3 + x19 + x20 = 0
(x5 + x6 + x9) % 2 = 0
x4 – x22 = 0
(x1 + x2 + x3 + x4) % 2 = 0
x14 – x23 – x24 = 0
Σ xi = 9

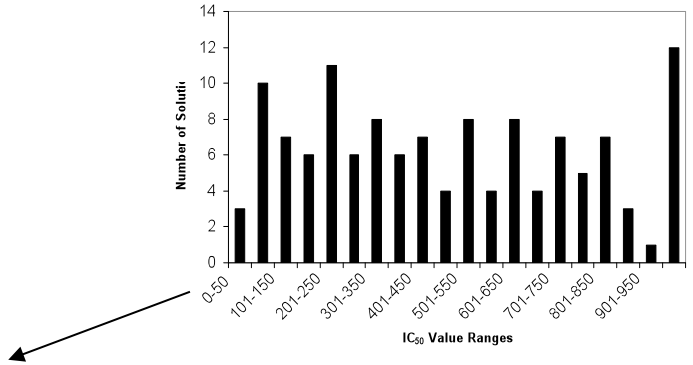
i-QSAR: Design ICAM-1 binding sequences (3)

1. **QSAR/QSPR analysis**

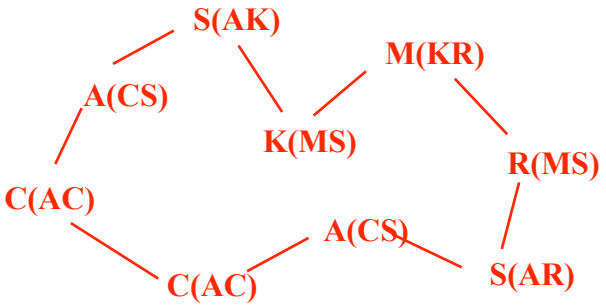
2. **Integer equation builder**

3. **Integer (diophantine) equation solver**

4. **Structure generator (sequence assembly)**



$$2\,C(AC) + 2\,A(CS) + S(AK) + K(MS) + M(KR) + R(MS) + S(AR) \rightarrow IC_{50} = 24.8$$



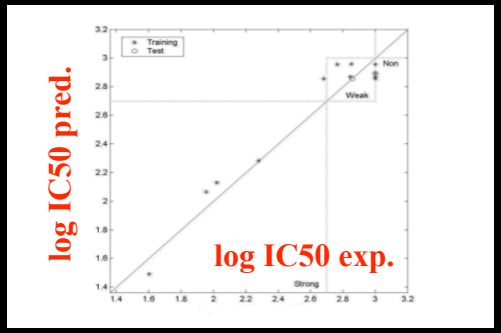
CASKMRSAC

Hamiltonian path in overlapping graph

Euler path in dual graph (Pevzner)

i-QSAR: Design ICAM-1 binding sequences

QSAR:
- activity = *f* (descriptors)



Inverse-QSAR:

- descriptors = f^{-1}(activity)

2 C(AC) + 2 A(CS) + S(AK) + K(MS) + M(KR) + R(MS) + S(AR) = f^{-1}(<25)

- assemble sequences matching descriptors

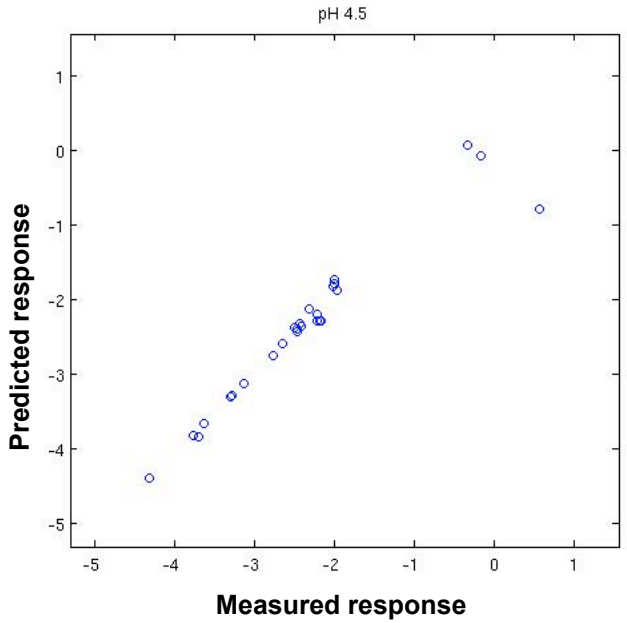
CASKMRSAC

• Larson's lab (UNM) synthesized and tested in vitro (ELISA) and in vivo (cellular aggregation blocking assay)
CASKMRSAC (IC_{50} = 23, *i*-QSAR predicts 24.8)
CVSKMRSVC (IC_{50} = 28, *i*-QSAR predicts 37.3)
• Up to date most potent inhibitors known for ICAM-1 (JMGM 2004)

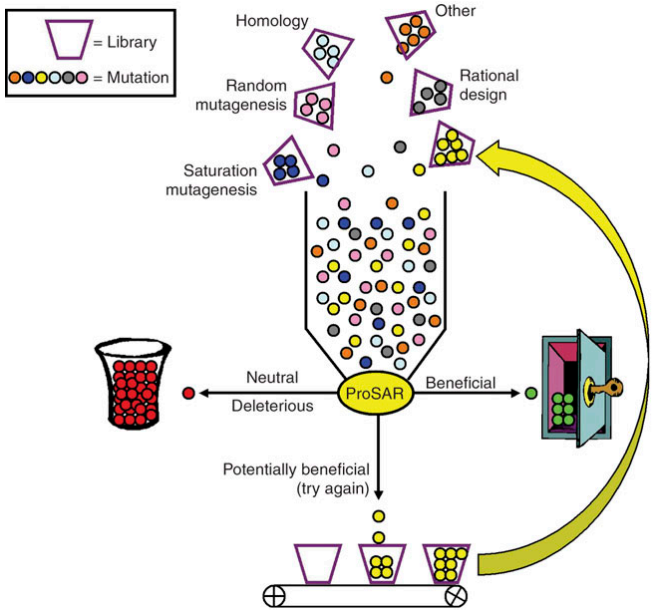
- Filed US Patent 2006

- Other *i*-QSAR examples : JMGM 2002, JCICS 2003, JCAMD 2005, Ind. Eng. Chem. Res 2005, JCIM 2006

Designing proteins using QSAR and i-QSAR



Activity measured using DNS assay for 25 cellulase mutants *(CelA from Alicyclobacillus acidocaldarius substrate carboxyl cellulose, pH 4.5)*

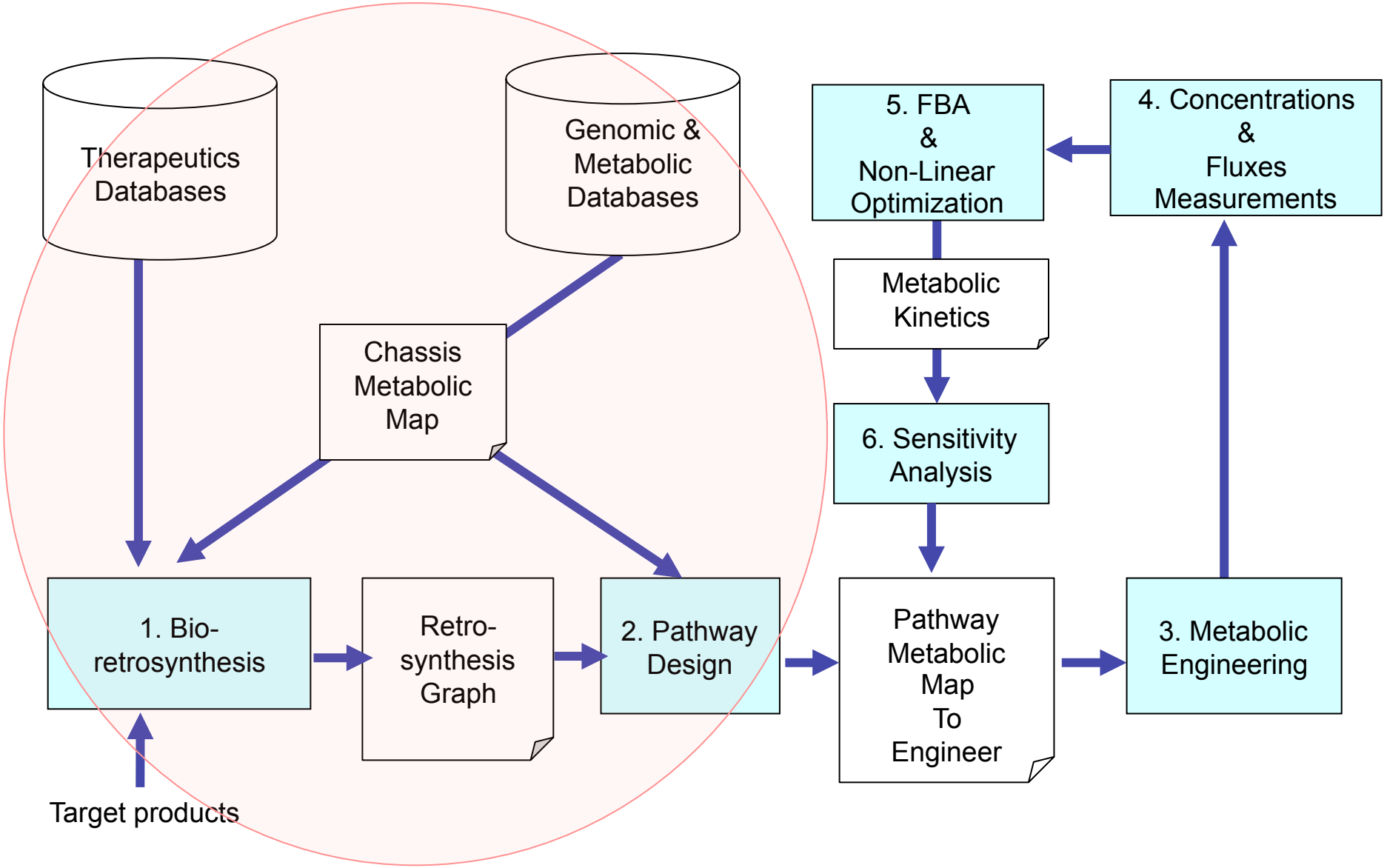


Fox et *al*. Improving catalytic function by ProSAR-driven enzyme evolution, Nature Biotech **25**, 338 - 344 (2007)

Engineered halohydrin dehalogenase incrases volumetric producing of Liptor 4,000 fold

Network design

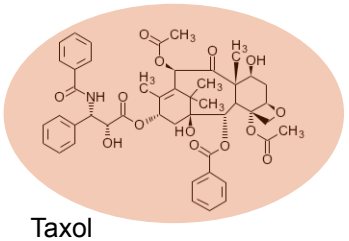
Metabolic engineering: a Bio-retrosynthesis approach



Metabolic engineering: Engineering E. coli to produce Taxol

Bio-retrosynthesis: Substrate and enzyme for Taxol?

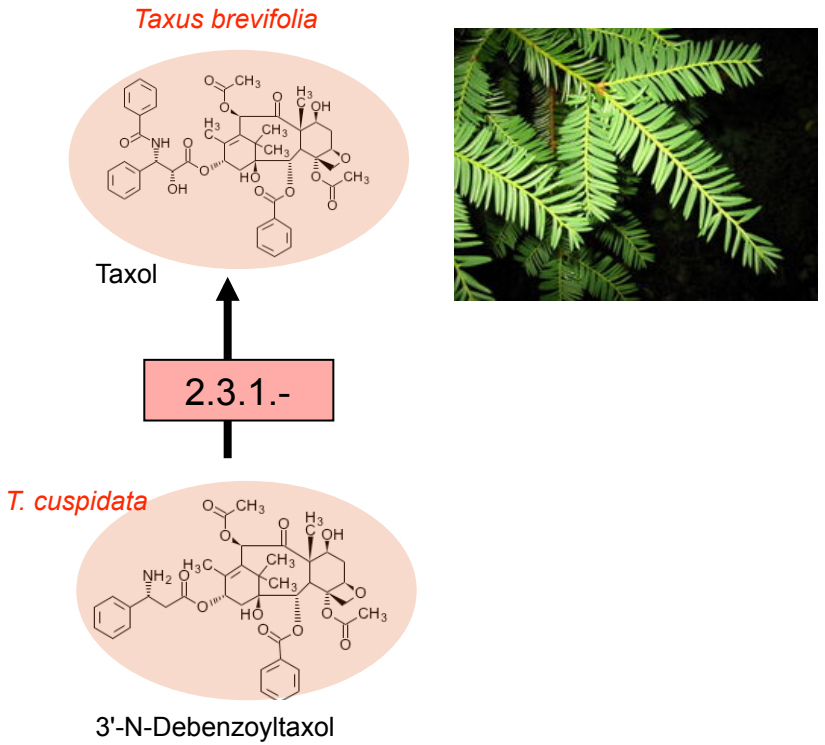
Taxus brevifolia



Taxol

Metabolic engineering: Engineering E. coli to produce Taxol

Bio-retrosynthesis: Substrate and enzyme for Taxol



Simple database search

Metabolic engineering: Engineering E. coli to produce Taxol

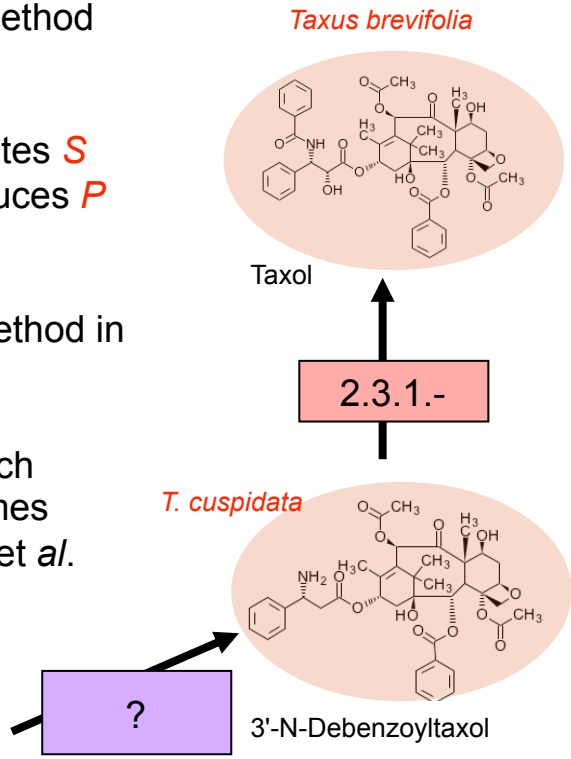
Bio-retrosynthesis: Substrate and enzyme for 3'N-Debenzoyltaxol?

- Code all known metabolic reactions *R* with neighborhood ($\sigma(R)$, method in Faulon, JCICS 1994)

- Search all metabolic substrates *S* such that *R* applied to *S* produces *P* = 3'N-Debenzoyltaxol

(i.e, $S \mid \sigma(S) = \sigma(P) - \sigma(R)$, method in Faulon, JCICS 2001)

- For each solution *S*, *R*, search genomes for candidate enzymes using tensor product (Faulon et *al*. Bioinformatics 2008)



Metabolic engineering: Engineering E. coli to produce Taxol

Bio-retrosynthesis: Substrate and enzyme for 3'N-Debenzoyltaxol?

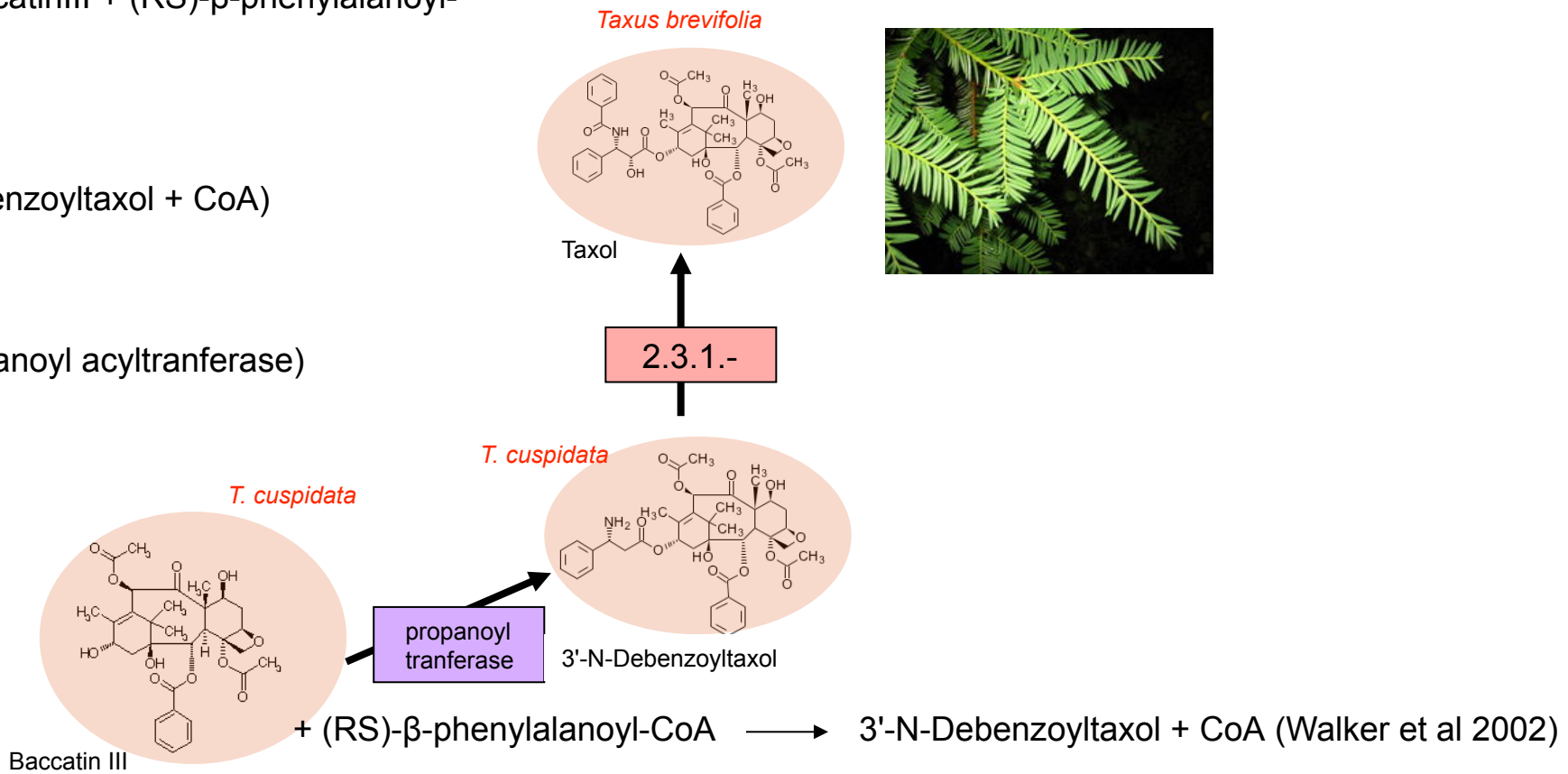
σ(BaccatinIII + (RS)-β-phenylalanoyl-CoA)

=

σ(Debenzoyltaxol + CoA)

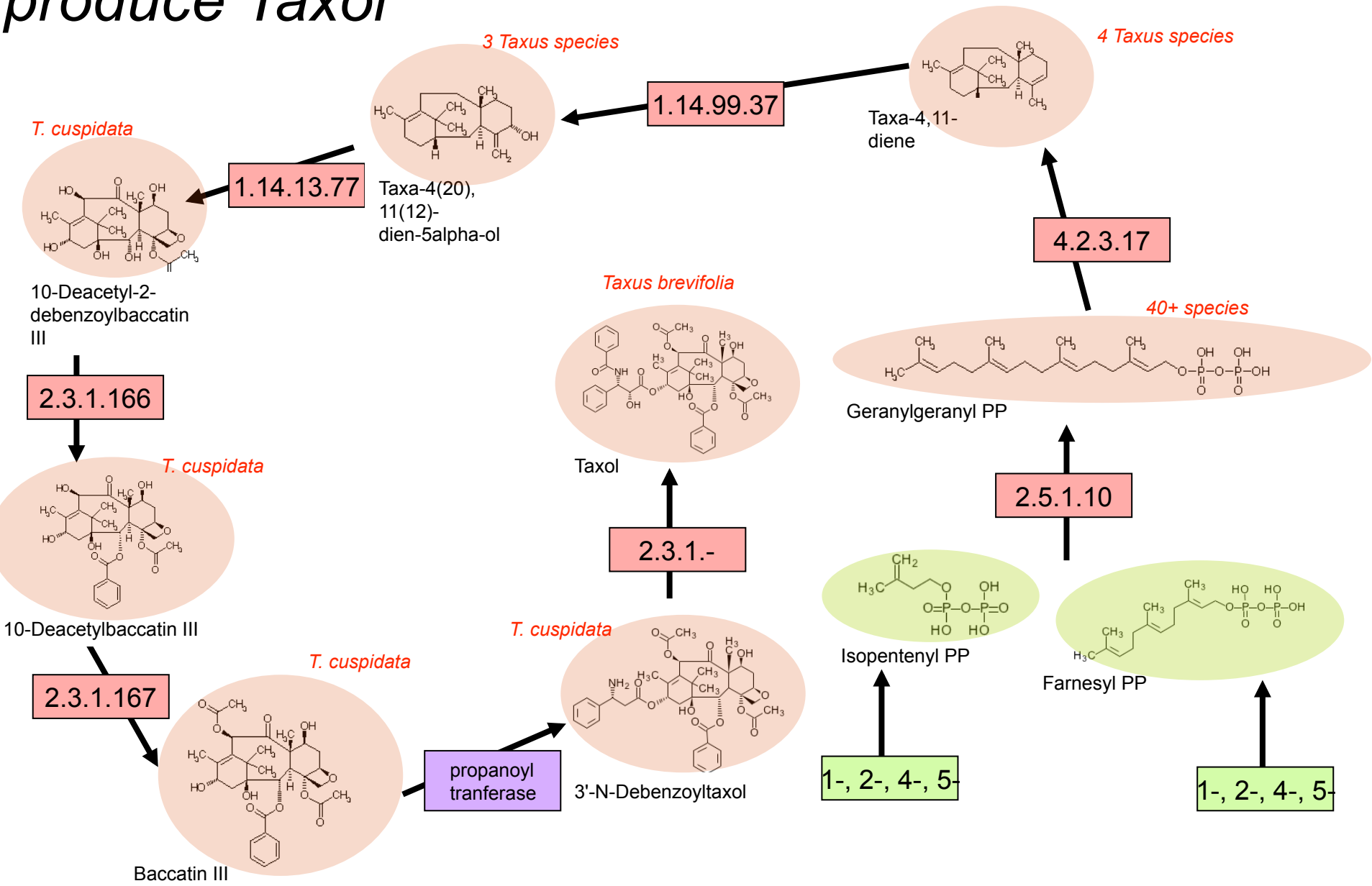
–

σ(propanoyl acyltranferase)

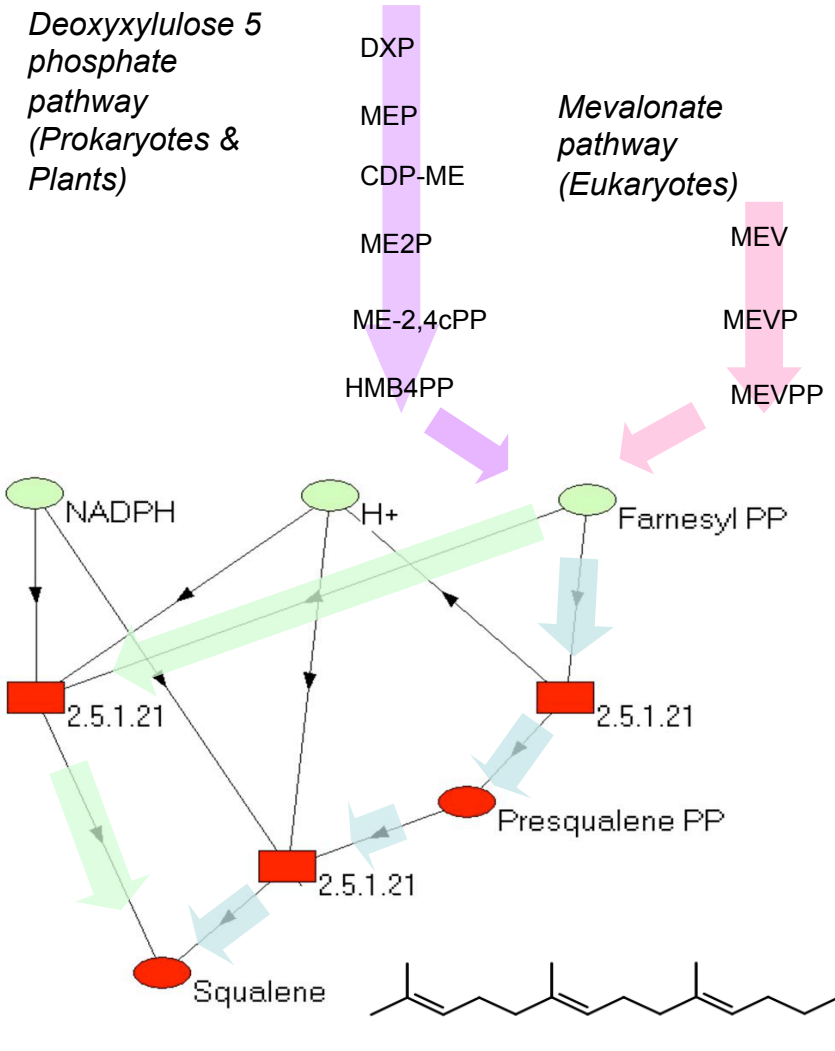


+ (RS)-β-phenylalanoyl-CoA ⟶ 3'-N-Debenzoyltaxol + CoA (Walker et al 2002)

Metabolic engineering: Engineering E. coli to produce Taxol



Bio-retrosynthesis: which pathway to engineer?



2x2 paths lead to squalene, do we need to engineer all? Which combination is best?

- Path enumeration
- Path ranking, a cost function, including

  - Number of exogenous enzymes and organisms
  - Number of predicted vs. documented enzymes is minimized.
  - Metabolites inhibition
  - Exogenous metabolites cytotoxicity
  - Difficulties of inserting exogenous enzymes into plasmids or chromosomes
  - Enzymes activity and pathway kinetics

Drug candidates for bio-retrosynthesis

Among 916 FDA approved small drugs taken from DRUGBANK

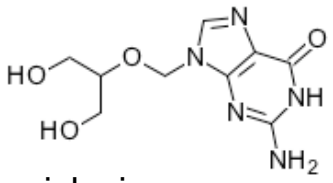
Which ones are also metabolites of *E. coli*?	- Metabolites searched in all *E coli* strains in KEGG	21
Which ones are also metabolites of *other organisms*?	- Metabolites searched in all other organisms stored in KEGG	42
Which ones are similar to some metabolites of *E. coli*?	- Similarity search using chemical fingerprints	36
Which ones are similar to some metabolites of *other organisms*?	- Similarity search using chemical fingerprints	32

Examples for bio-retrosynthesis

Which ones are also metabolites of *other organism*?

Taxol (diterpenoids biosynthesis)

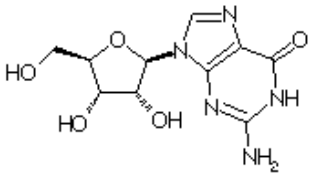
Which ones are similar to some metabolites of *E. coli*?

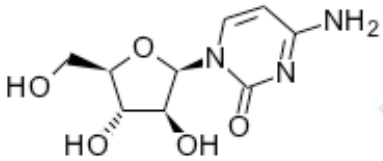


Ganciclovir

Tanimoto= 0.69

Guanosine
Purine metabolism

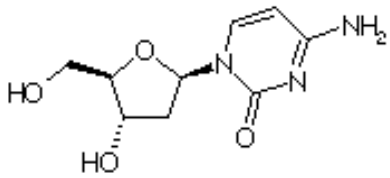




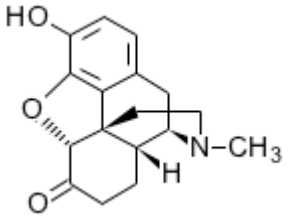
Cytarabine

Tanimoto= 0.80

Deoxycytidine
Pyrimidine metabolism



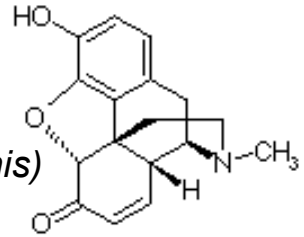
Which ones are similar to some metabolites of *other organism*?



Hydromorphone

Tanimoto= 0.82

Morphinone
Alkaloid biosynthesis I
(*Streptococcus sanguinis)*



Funding sources

- Genopole/ATIGE
  Bioretrosynthesis
- CNRS/Interface Physique Chimie Biologie
  Toxicity prediction

Source		Title
DOE/Bioenergy Researc Centers (www.jbei.org)	2007-08	Joint Bioenergy Institute
DOE/Sandia LDRD program	2005-08	Shotgun protein sequencing
DOE/Grand Challenge program	2005-08	Signaling networks Modeling immune system
DOE/Sandia LDRD program	2005-08	Molecular design
NIH/PubChem	2004-07	Molecular libraries screening centers
DOE/Sandia LDRD program	2002-05	Reverse engineering transciptional regulatory networks
DOE/Genome to Life Program	2002-05	Protein network inference

Acknowledgements

M. Misra, Sandia (QSAR)
Don Visco, R. Pophale, A. Kotu, D. Weiss, TTU (QSAR & iQSAR)
M. Brown, Sandia (QSAR & iQSAR)
C. Churchwell, Sandia (sequence design)

D. Roe, S. Martin, R. Carr, Sandia (SVMs and network inference)
K. Sale, Sandia (training sets & QSAR)
R. Sapra and B. Simmons, JBEI (enzyme assays)

T. Oprea, UNM (systems chemical biology)
R. Larson's lab, UNM (inhibitor assay)

http://jfaulon.wikispaces.com