Molecular Modeling for Systems Pharmacology:
Integrative modeling of biomolecular assemblies and networks

1/20/15
Andrej Sali (sali@salilab.org)

While it may be hard to live with generalization, it is inconceivable to live without it. Peter Gay, Schnitzler’s Century (2002).
1. Hierarchy of biological organization and systems pharmacology

2. Modeling

3. Integrative modeling of assembly structures

4. Integrative modeling of molecular networks
Contents

1. Hierarchy of biological organization and systems pharmacology

2. Modeling

3. Integrative modeling of assembly structures

4. Integrative modeling of molecular networks
Representation: molecular structure (particles in space), localization (contents of volume elements), networks of components, inter-compartmental flows, …
Systems Pharmacology to the rescue

What is Systems Pharmacology?

1. Systems pharmacology is the application of systems biology principles to the field of pharmacology. It seeks to understand how medicines work on various systems of the body. Instead of considering the effect of a drug to be the result of one specific drug-protein interaction, systems pharmacology considers the effect of a drug to be the outcome of the network of interactions a drug may have. (Wikipedia)

2. Systems Pharmacology brings a new combination of mathematical and experimental tools to bear on the discovery and analysis of therapeutic drugs. System-level understanding of pharmacological effects will help to identify new uses for existing drugs, identify those patients most likely to benefit from mono and combination therapies, and make drug discovery and development faster, cheaper, and more predictable. (http://isp.hms.harvard.edu)
Systems Pharmacology Bibliography


7. http://www.nature.com/psp
Contents

1. Hierarchy of biological organization and systems pharmacology

2. Modeling

3. Integrative modeling of assembly structures

4. Integrative modeling of molecular networks
Scientific progress

1. Iteration of *Hypothesis* and *Experiment*
Scientific progress

1. Iteration of *Hypothesis* and *Experiment*

Hypothesis vs Theory
Scientific progress

1. Iteration of *Hypothesis* and *Experiment*

Hypothesis vs Theory

Thomas Kuhn, 1962. *The Structure of Scientific Revolutions*
Scientific progress

1. Iteration of *Hypothesis* and *Experiment*

   ![Diagram of the Kuhn cycle]

   **Hypothesis vs Theory**

   Thomas Kuhn, 1962. *The Structure of Scientific Revolutions*

2. Exploratory research (eg, genome sequencing).
Modeling

Scientific modelling is a scientific activity the aim of which is to make a particular part or feature of the world easier to understand, define, quantify, visualize, or simulate. It requires selecting and identifying relevant aspects of a situation in the real world and then using different types of models for different aims, such as conceptual models to better understand, operational models to operationalize, mathematical models to quantify, and graphical models to visualize the subject. Modelling is an essential and inseparable part of scientific activity, and many scientific disciplines have their own ideas about specific types of modeling. (Wikipedia)
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

NATURE

April 25, 1953
Molecular Structure of Nucleic Acids

A Structure for Deoxyribose Nucleic Acid

X-ray diffraction

Composition
Stoichiometry
Chemical complementarity
To understand and modulate cellular processes, we need their models. These models are best generated by considering all available information.
Contents

1. Hierarchy of biological organization and systems pharmacology

2. Modeling

3. Integrative modeling of assembly structures

4. Integrative modeling of molecular networks
Structural biology:
Maximize accuracy, resolution, completeness, and efficiency of the structural coverage of macromolecular assemblies

Motivation: Models will allow us to understand how machines work, how they evolved, how they can be controlled, modified, and perhaps even designed.

There may be thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized, involved in a few hundred core biological processes.
Integrative Structural Biology
for maximizing accuracy, resolution, completeness, and efficiency of structure determination

Use structural information from any source: measurement, first principles, rules; resolution: low or high resolution to obtain the set of all models that are consistent with it.

A description of integrative structure determination


While it may be hard to live with generalization, it is inconceivable to live without it. Peter Gay, Schnitzler’s Century (2002).
Challenges in interpreting the data in terms of a structural model

- **Sparseness**
  due to incompleteness of measurements

- **Error**
  due to measurement and other imperfections

- **Ambiguity**
  *eg*, due to multiple copies of a protein in a system

- **Incoherence (mixture)**
  due to multiple states of a system in a heterogenous sample
Pushing the envelope of structural biology by integration of all available information

• Size

• Static systems in single and multiple states

• Dynamic systems

• Bulk and single molecule views

• Impure samples

• Overlapping with other domains such as systems biology
Configuration of 456 proteins in the Nuclear Pore Complex

with M. Rout & B. Chait

Quantitative Immunoblotting
30 relative abundances

Affinity Purification
Overlay Assay
75 composites 7 contacts

Protein-protein Proximities

Protein Stoichiometry

Protein Localization

Electron Microscopy
electron microscopy map

Immuno-Electron Microscopy
10,615 gold particles

Ultracentrifugation
30 S-values 1 S-value

Bioinformatics and Membrane Fractionation
30 protein sequences

Configuration of 456 proteins in the Nuclear Pore Complex

with M. Rout & B. Chait

Protein Shape

Protein Stoichiometry

Protein-protein Proximities

Symmetry

Protein Localization

Quantitative Immunoblotting

30 relative abundances

Affinity Purification Overlay Assay

75 composites 7 contacts

Bioinformatics and Membrane Fractionation

30 protein sequences

Ultracentrifugation

30 S-values 1 S-value

Immunoelectron Microscopy

10,615 gold particles

Electron Microscopy

electron microscopy map

**Integrative Modeling Platform (IMP)**

http://integrativemodeling.org

R. Pellarin, M. Bonomi, B. Raveh, S. Calhoun, C. Greenberg, G.Dong.

Open source, versions, documentation, wiki, examples, mailing lists, unit testing, bug tracking, ...

---

**Representation:**

- Atomic
- Rigid bodies
- Coarse-grained
- Multi-scale
- Symmetry / periodicity
- Multi-state systems

**Scoring:**

- Density maps
- EM images
- Proteomics
- FRET
- Chemical and Cys cross-linking
- Homology-derived restraints
- SAXS
- Native mass spectrometry
- Statistical potentials
- Molecular mechanics forcefields
- Bayesian scoring
- Library of functional forms (ambiguity, ...)

**Sampling:**

- Simplex
- Conjugate Gradients
- Monte Carlo
- Brownian Dynamics
- Molecular Dynamics
- Replica Exchange
- Divide-and-conquer enumeration

**Analysis:**

- Clustering
- Chimera
- Pymol
- PDB files
- Density maps
Modeling aspects

Find simplest possible models that reproduce the measured data within their uncertainty.

1. Representation
2. Scoring function
3. Sampling algorithm
4. Assessment
Representation
Hierarchical model representation facilitates using imprecise information

Ignorance (residues per bead)
Scoring function: Least-squares modeling of a structure

\[ S(X) = \sum_i w_i [d_i - f_i(X)]^2 \]

When \( w_i = 1/\sigma_i^2 \), we get a “maximum likelihood” approach, for which \( S(X) = \chi^2 \).
**Scoring function: Bayesian modeling of structure(s)**

\[
p(M | D, I) \propto p(D | M, I) \cdot p(M | I)
\]

- **Posterior** is the probability density of model \( M \), given data \( D \) and information \( I \).

- Model \( M \) can include coordinates of one or more structures \( X \) as well as additional parameters (noise levels, weights, calibration parameters, ...).

- **Likelihood** is the probability density of observing data \( D \), given model \( M \) and prior information \( I \) (by relying on a model of noise and a forward model, which computes data \( D \) given model \( M \)).

- **Prior** is the probability density of model \( M \), given prior information \( I \).

\[
p(AB) = p(BA) = p(A) \cdot p(B/A) = p(B) \cdot p(A/B)
\]

Sampling, optimization, enumeration
Sampling, optimization, enumeration
Assessment
(in the absence of Bayesian inference)
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.
2. Precision of the ensemble of good-scoring models.
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.

2. Precision of the ensemble of good-scoring models.

3. Check model against unused data (cross-validation).
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.
2. Precision of the ensemble of good-scoring models.
3. Check model against unused data (cross-validation).
4. Known precision / accuracy for “similar” cases.
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.
2. Precision of the ensemble of good-scoring models.
3. Check model against unused data (cross-validation).
4. Known precision / accuracy for “similar” cases.
5. Non-random patterns in the model.
Assessment
(in the absence of Bayesian inference)

1. Existence of a good-scoring model.
2. Precision of the ensemble of good-scoring models.
3. Check model against unused data (cross-validation).
4. Known precision / accuracy for “similar” cases.
5. Non-random patterns in the model.

Modeling facilitates assessing the data as well as models in terms of precision and accuracy.
Steps in Comparative Protein Structure Modeling

START

Template Search

Target – Template Alignment

Model Building

Model Evaluation

Yes

END

No

OK?

TARGET

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPERASFQWMNDK

TEMPLATE

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIEPE

MSVIPKRLGNCETSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE


http://salilab.org/

05/27/2006
Contents

1. Hierarchy of biological organization and systems pharmacology

2. Modeling

3. Integrative modeling of assembly structures

4. Integrative modeling of molecular networks
Integrating **structural biology and systems biology**

CJ Ryan, P Cimermancic, ZA Szpiech, A Sali, RD Hernandez, NJ Krogan

Annotation of enzyme function

1. Prediction of the function in the context of the pathway

2. Prediction of the pathway based on all available information

3. Automated (\textit{i.e.}, thorough, objective, ...) method
What is the pathway, given a set of potential enzymes and metabolites as well as at least one enzyme and/or metabolite in it?
A tiny subset of structure-, imaging- and systems-based information that (in principle) informs network prediction

Structural biology
- Virtual screening (VS)
- *In silico* transformations and chemical similarity calculations
- High-throughput screening

Systems biology and imaging
- Coexpression of genes
- Genome context
- Genetic interactions
- Metabolite levels
- Protein functional links
- Orthology
- Tomography
- Super-resolution optical microscopy
- FRET spectroscopy

![Diagram with various scientific concepts and terms related to proteomics, chemogenomics, virtual screening, cheminformatics, genome neighborhood, protein interaction predictions, and more.

- Protein interaction predictions
- Virtual screening
- Cheminformatics
- Gene coexpression
- High-throughput screening
- Genome neighborhood
- Proteomics
- Chemogenomics
- Metabolomics
- Regulons

*AAATTCGTA CAAATGGTA CAAATGGTA ATATGGTT*
Mindset: From integrative structure modeling towards integrative pathway mapping

Alber et al. Annual Reviews in Biochemistry, 2008
Mindset: From integrative structure modeling towards integrative pathway mapping

Alber et al. Annual Reviews in Biochemistry, 2008
Mindset: From integrative structure modeling towards integrative pathway mapping

- Components
- Mindset: From integrative structure modeling towards integrative pathway mapping
- Sampling and scoring
- Biological network
- RESTRAINS
  - X-ray crystallography
  - NMR
  - Electron microscopy
  - Cheminformatics
  - Cross-linking
- Operon/genome context
- Docking
- Co-purification
- Bioinformatics, physics
- Cheminformatics
- Operon/genome context
- Docking
- Co-purification
- IMP
- Biological network
- Potential Components
- Model
Mindset: From integrative structure modeling towards integrative pathway mapping

Restraints
- X-ray crystallography
- NMR
- Electron microscopy
- Cheminformatics
- Cross-linking
- Operon/genome context
- Docking
- Co-purification
- Bioinformatics, physics

Potential Components

IMP

Sampling and scoring

Metabolic Pathway

Model
Enzymes and ligands are individual nodes
Edges reflect a relationship between an enzyme and its substrate/product
Scoring

• Translation of information into restraints that quantify the degree of consistency between a pathway model and the corresponding information
• Scoring function is a sum of individual restraints to rank alternative models

\[ Z_{comp} = Z_{sea} + Z_{rxns} + Z_{dock} + Z_{screen} \]
Sampling
Monte Carlo with simulated annealing

1) Starting with a model, select a node to alter identity

2) Select a new node identity from all possible node identities & make change in model

3) If change improves score, accept the new model. Otherwise, accept the new model with a probability determined by the difference in scores.

4) Repeat until “convergence”

$$\Delta S = \text{Score(model}_{i+1}) - \text{Score(model}_i)$$

Prob. of acceptance = $\exp(-\Delta S/T)$
Example I: reconstructing glycolysis

Here, we are treating glycolysis as a linear pathway.
Four-stage overview
Glycolysis

Gathering information

- 2,742 potential substrates & products
- 10 glycolytic enzymes
- Docking scores for ligand-enzyme pairs
- SEA e-values between all enzymes
- Chemical transformation scores for all substrate-enzyme-product trios

Designing model representation and evaluation

- Fixed-length graph: 10 protein nodes and 11 ligand nodes
- Monte Carlo with simulated annealing 5 M iterations per run & 400 runs

Sampling good models

Analyzing models and information

- Assessing top models over 3.5 standard deviations above mean

SEA calculations by Magdalena Korczynska, Henry Lin, and Brian Shoichet
**Assessment: reconstructing enzyme order**

**Glycolysis**

<table>
<thead>
<tr>
<th>Correct Pathway (Position in pathway →)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Reaction 1" /></td>
<td><img src="image2" alt="Reaction 2" /></td>
<td><img src="image3" alt="Reaction 3" /></td>
<td><img src="image4" alt="Reaction 4" /></td>
<td><img src="image5" alt="Reaction 5" /></td>
<td><img src="image6" alt="Reaction 6" /></td>
<td><img src="image7" alt="Reaction 7" /></td>
<td><img src="image8" alt="Reaction 8" /></td>
<td><img src="image9" alt="Reaction 9" /></td>
<td><img src="image10" alt="Reaction 10" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted Pathways</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Reaction 1" /></td>
<td><img src="image2" alt="Reaction 2" /></td>
<td><img src="image3" alt="Reaction 3" /></td>
<td><img src="image4" alt="Reaction 4" /></td>
<td><img src="image5" alt="Reaction 5" /></td>
<td><img src="image6" alt="Reaction 6" /></td>
<td><img src="image7" alt="Reaction 7" /></td>
<td><img src="image8" alt="Reaction 8" /></td>
<td><img src="image9" alt="Reaction 9" /></td>
<td><img src="image10" alt="Reaction 10" /></td>
<td></td>
</tr>
</tbody>
</table>

Ordered by frequency at each position.
# Assessment: synergy of different types of information

**Glycolysis**

<table>
<thead>
<tr>
<th>Step in pathway</th>
<th>Rank by VS score alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
</tr>
</tbody>
</table>
Assessment: synergy of different types of information

Glycolysis

<table>
<thead>
<tr>
<th>Step in pathway</th>
<th>Rank by VS score alone</th>
<th>Rank using VS scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>175</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
<td>369</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>163</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
<td>43</td>
</tr>
</tbody>
</table>
Assessment: synergy of different types of information

**Glycolysis**

<table>
<thead>
<tr>
<th>Step in pathway</th>
<th>Rank by VS score alone</th>
<th>Rank using VS scores</th>
<th>Rank using chemical transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>20</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>175</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
<td>369</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>163</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
<td>43</td>
<td>1</td>
</tr>
</tbody>
</table>
### Assessment: synergy of different types of information

#### Glycolysis

<table>
<thead>
<tr>
<th>Step in pathway</th>
<th>Rank by VS score alone</th>
<th>Rank using VS scores</th>
<th>Rank using chemical transformations</th>
<th>Rank using VS, chemical transformations, and SEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>20</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>54</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>10</td>
<td>93</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>175</td>
<td>83</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
<td>369</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>163</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
<td>43</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Prediction of pathway downstream of *H. influenzae* RdAW SBP TRAP transporter

Tripartite ATP-independent periplasmic transporters - family of transporters in bacteria and archaea for the uptake of organic acids

*Haemophilus influenzae*

“Bacterial flu” found in the upper respiratory system of host

Centers for Disease Control and Prevention
Virtual screening
Chemical transformations
High-throughput screening
Metabolic endpoints
Similarity ensemble approach

Gathering information

Genome neighborhood
Virtual screening
Chemical transformations
High-throughput screening
Metabolic endpoints
Similarity ensemble approach

Designing model representation and evaluation

6 proteins
6 protein-ligand pairs
5 enzymatic reactions
1 protein screened
1 endpoint
6 proteins

Sampling good models

Potential ligands
Potential proteins
Pathway model

Monte Carlo simulated annealing

Analyzing models and information

Assessing sampling
Assessing accuracy
Validated SPB TRAP pathway prediction

L-gulonate  $\rightarrow$ D-fructuronate  $\rightarrow$ D-mannonate

C9MHP2  $\rightarrow$ C9MHP1  $\rightarrow$ C9MHP6

$9.3 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$  $7.6 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$

D-glyceraldehyde-3P  $\rightarrow$ 2-keto-3-deoxy-6P-D-gluconate  $\rightarrow$ 2-keto-3-deoxy-D-gluconate

C9MHP7  $\rightarrow$ C9MHP5

$7.5 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$  $2.4 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$

$k_{\text{cat}}/K_m$

Crystal structure

in vitro assays by Daniel Wichlecki, Gerlt lab
Structure by Matthew Vetting et al., NYSGXRC
Metformin “network”


![Diagram of Metformin network](image-url)
In Conclusion

The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

This goal will be achieved by a **formal** integration of experiment, physics, and statistical inference, spanning all relevant size and time scales.