Physical Principles of Molecular Information Systems

Physics Colloquium

Weizmann

(G. Brodsky Mol Cell 2010)
Noisy molecular channels: Rate-Distortion theory

• Molecular spaces: \( S, M \).
• Mapping: \( \Phi: S \leftrightarrow M \).
• Functionals (map): \( \mathcal{F}(\Phi) \rightarrow \text{“fitness”} \)
• Optimize(fitness): \( \Phi^* = \arg\max \mathcal{F}(\Phi) \)
  \[ \rightarrow \text{Design principles, Coding transition...} \]
• Topological aspects.
Examples: four fundamental living systems

Homologous Recombination

Molecular codes (genetic code)

Photosynthesis

Chromosome organization
Examples: four essential living systems

Homologous Recombination

Photosynthesis

Molecular codes (genetic code)

Chromosome organization
Recombination machinery recognizes homologous DNA

- Exchange between two homologous DNAs.
- Essential for:
  - *Genome integrity* (repair machinery).
  - *Genetic diversity* via crossover and sex (horizontal transfer).
  - affects speciation.
- Task: Detect correct, homologous DNA target among many incorrect lookalikes.

(Savir &TT, Plos 1 2007, Mol Cell 2010)
Why DNA is extended in homologous recombination?

- Mediated by RecA.
- Extension by 50%.
- Costs $3-4 k_BT/bp$.
- Mechanism to detect Right DNA in large pool of similar targets.
- General mechanism: Conformational proofreading.
Homologous recombination maps sequences to decisions

- Recombination advances in 3-base steps.
- Each step Mapping from DNA target space to decisions:
  - Problem: Optimization to withstand noise in sequence recognition.
  - \( S \) → \( M \)
    - **Proceed** recombination
    - **Abort** recombination

\( S \) → Sequence recognition

\( M \)
Molecular recognition is a decision problem

- Each 3-base step is **decision-making process**: 2x2 channel.

  - Each event = (input, output) bears cost/benefit, $C(\text{event})$.

  - **Fitness**: $\mathcal{F} = \sum_{\text{event}} C(\text{event}) \times P(\text{event})$.

  - Fitness depends on structural parameters and can be optimized.

(Savir & TT, Plos 1 2007, Mol Cell 2010)
Extension maximizes recognition fitness

- Fitness depends on energetics via binding probabilities:
  \[ \mathcal{F}(C, P_b(E_b)) \approx P_{\text{comp}} - P_{\text{non-comp}} \]
- One effective d.o.f. = \( E_{\text{extension}} \)
  \[ E_b = E_{\text{pairing}} - E_{\text{extension}} \]
- Extension shifts binding to optimal fitness.
- Experimental extension \( \sim \) optimal.
- General design principle of molecular recognition systems?

\[ P_B = \frac{1}{1 + e^{-E_b/k_B T}} \]
Optimal recognition: When off-target is right on

- Structural \textit{mismatch or energy barrier}
- Reduce Right, \textit{but} reduces Wrong even more.

- Result: Enhancement of fitness $\mathcal{F} \approx P_{\text{right}} - P_{\text{wrong}}$.

- \textbf{Conformational Proofreading:}
  Optimal fitness at non-zero mismatch or extension.

  - Optimal recognizer is \textit{off-target}
  - Not lock-and-key (\textit{induced fit})

Examples: four essential living systems

Homologous Recombination

Molecular codes (genetic code)

Chromosome organization

Photosynthesis
Rubisco captures atmospheric $\text{CO}_2$ to make sugar

- Photosynthesis fixates carbon into organic forms.
- Rubisco captures $\text{CO}_2$.
- Complex of 8 large + 8 small subunits (540 kDa).

(Savir, Noor, Milo & TT, PNAS 2010)
Impact of Rubisco’s (in)efficiency on the biosphere

- Most abundant protein on Earth.
- Catalyzes most carbon fixation.
- **Very slow** catalysis rate (~ 3-10 CO₂/sec).
- **Low specificity:** confuses O=C=O and O=O.

- Can Rubisco be improved? – little success so far.
- Already optimized by evolution?
- Inefficiency due to biochemical constraints?

- **Is Rubisco optimal, constrained or both?**
Rubisco maps available substrates to products

- Four d.o.f.: 2 affinities to CO$_2$ and O$_2$: \( \begin{pmatrix} K_C & K_O \\ v_C & v_O \end{pmatrix} \)
  - 2 catalysis rates

\[
\begin{align*}
\text{CO}_2 & \rightarrow +1 \cdot \text{CO}_2 \rightarrow \text{sugar} \\
\text{O}_2 & \rightarrow -\frac{1}{2} \cdot \text{CO}_2 \text{ (effective loss)} \\
\text{empty} & \rightarrow \emptyset
\end{align*}
\]

- Overall fitness = net photosynthesis rate

\[
F = \sum_{\text{binding}} \frac{d}{dt} (\text{CO}_2)_b \times P_b = v_C P_{\text{CO}_2} - \frac{1}{2} v_O P_{\text{O}_2} + 0 \cdot P_{\text{empty}}
\]

- Optimal Rubisco?

\[
\begin{align*}
P_{\text{CO}_2} &= \frac{[\text{CO}_2]}{1 + \frac{[\text{CO}_2]}{K_C} + \frac{[\text{O}_2]}{K_O}} \\

P_{\text{O}_2} &= \frac{[\text{O}_2]}{1 + \frac{[\text{CO}_2]}{K_C} + \frac{[\text{O}_2]}{K_O}}
\end{align*}
\]
Cross-species analysis exhibits strong correlation

- Analysis of data from 28 eukaryotes and prokaryotes:

<table>
<thead>
<tr>
<th>Species</th>
<th>$K_C$ [μM]</th>
<th>$v_C$ [1/sec]</th>
<th>$S$</th>
<th>$K_O$ [μM]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cyanobacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synechococcus 6301</td>
<td>340</td>
<td>11.6</td>
<td>43</td>
<td>972</td>
</tr>
<tr>
<td>Synechococcus 7002</td>
<td>246</td>
<td>13.4</td>
<td>52</td>
<td>1300</td>
</tr>
<tr>
<td>Cocoloris peniocystis</td>
<td>121</td>
<td>7.3</td>
<td>47</td>
<td>1220</td>
</tr>
<tr>
<td>Aphanizomenon flos aquae</td>
<td>105</td>
<td>9</td>
<td>48</td>
<td>990</td>
</tr>
<tr>
<td>Rhodospirillum rubrum</td>
<td>80</td>
<td>6.7</td>
<td>41</td>
<td>220</td>
</tr>
<tr>
<td><strong>Photosynthetic Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodopseudomonos sphaeroides II</td>
<td>80</td>
<td>12.3</td>
<td>62</td>
<td>290</td>
</tr>
<tr>
<td>Chromatium vinosum</td>
<td>37</td>
<td>9</td>
<td>63</td>
<td>840</td>
</tr>
<tr>
<td>Rhodopseudomonos sphaeroides I</td>
<td>36</td>
<td>11</td>
<td>61</td>
<td>660</td>
</tr>
<tr>
<td>Scenedesmus obliquus</td>
<td>38</td>
<td>11</td>
<td>61</td>
<td>480</td>
</tr>
<tr>
<td><strong>Green algae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>29</td>
<td>5.8</td>
<td>54</td>
<td>410</td>
</tr>
<tr>
<td>Euglena gracilis</td>
<td>25</td>
<td>5.8</td>
<td>54</td>
<td>810</td>
</tr>
<tr>
<td>Zea mays</td>
<td>34</td>
<td>4.4</td>
<td>78</td>
<td>309</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>30</td>
<td>5.4</td>
<td>78</td>
<td>289</td>
</tr>
<tr>
<td><strong>C4 plants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flaveria australasica</td>
<td>22</td>
<td>3.84</td>
<td>77.2</td>
<td>640</td>
</tr>
<tr>
<td>Amaranthus edulis</td>
<td>18.2</td>
<td>4.14</td>
<td>77.5</td>
<td>640</td>
</tr>
<tr>
<td>Amaranthus hybridus</td>
<td>16</td>
<td>3.8</td>
<td>82</td>
<td>309</td>
</tr>
<tr>
<td>Potulaca oleracea</td>
<td>13.6</td>
<td>5.9</td>
<td>78</td>
<td>289</td>
</tr>
<tr>
<td><strong>C3 plants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lolium perenne</td>
<td>16</td>
<td>3.7</td>
<td>80</td>
<td>500</td>
</tr>
<tr>
<td>Spinacia oleracea</td>
<td>14</td>
<td>2.5</td>
<td>80</td>
<td>480</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>14</td>
<td>2.5</td>
<td>90</td>
<td>730</td>
</tr>
<tr>
<td><strong>Nongreen algae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phaeodactylum tricornutum</td>
<td>28</td>
<td>3.4</td>
<td>81</td>
<td>600</td>
</tr>
<tr>
<td>Griffithsia monilis</td>
<td>9.3</td>
<td>2.6</td>
<td>81</td>
<td>666</td>
</tr>
<tr>
<td>Galdieria sulfuraria</td>
<td>3.3</td>
<td>1.2</td>
<td>82</td>
<td>415</td>
</tr>
</tbody>
</table>

Specificity, $S = \frac{\text{carboxylation} / [\text{CO}_2]}{\text{oxygenation} / [\text{O}_2]} = \frac{v_C}{K_C} / \frac{v_O}{K_O}$
Rubisco’s parameter space is effectively 1D

- Data resides in 4D space \((S, v_C, K_C, K_O)\).
- But constrained to 1D power law (linear in log scale).
- PCA analysis ( >90% of variability is 1D) → power law correlations:

\[
K_C \propto v_C^2 \\
K_O \propto v_C^{0.5 \pm 0.1} \\
S \propto v_C^{-0.5 \pm 0.1}
\]

(Savir, Noor, Milo & TT, PNAS 2010)
Rubisco are nearly optimal to their habitats

- Max(\(F\)) as function of \(v_C\) in given habitat (\(O_2\) and \(CO_2\)).

\[
F(v_C) = \frac{v_C - 0.003 \cdot \frac{[O_2]}{[CO_2]} \cdot v_C^{3/2}}{1 + 1.3 \cdot \frac{1}{[CO_2]} \cdot v_C^2 + 0.005 \cdot \frac{[O_2]}{[CO_2]} \cdot v_C^{3/2}}.
\]

- All organism classes are nearly optimal.
Interplay of evolution and constraints shapes Rubisco

- Hint: stronger fluctuations for parameters that affect $F$ weakly ($K_O$ vs. $K_C$).
- Possible test: point mutation survey.
Questions, future directions, generalization...

- **Structural Mechanism?**

- **Response to long term climate changes?** (CO₂, Temp....)

- **Constrained plasticity** in low dimensional landscapes:
  
  **Generic phenomenon in proteins?**
  
  – preliminary data from other strongly selected proteins.
Examples: four essential living systems

Homologous Recombination

Photosynthesis

Molecular codes (genetic code)

Chromosome organization
The genetic code is main info channel of life

- Genetic code: maps 3-letter words in 4-letter DNA language \((4^3 = 64\) codons) to protein language of 20 amino acids.

- Proteins are amino acid polymers.

- Diversity of amino-acids is essential to protein functionality.
The genetic code maps codons to amino-acids

- Molecular code = map relating two sets of molecules (spaces, “languages”) via molecular recognition.
- Spaces defined by similarity of molecules (size, polarity etc.)

<table>
<thead>
<tr>
<th>20 amino-acids</th>
<th>64 codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly, Ala, Val, Leu, Ile, Pro</td>
<td>GAG, GAC, GCG, GCC, GUG, GUC, GGA, GGU</td>
</tr>
<tr>
<td>Gly, Val, Leu, Ile, Pro</td>
<td>GGA, GGU</td>
</tr>
<tr>
<td>Cys, Ser, Thr, Trp, Phe, Met</td>
<td>UCA, UCU, UCC, UCG, UUA, UUG</td>
</tr>
<tr>
<td>Lys, Arg, His, Asn, Glu, Tyr</td>
<td>UUU, UUA, UUC, UUG, CUU, CUC</td>
</tr>
<tr>
<td>Asp, Glu, Trp, Ser, Thr, Phe, Met</td>
<td>GGU, GUG</td>
</tr>
</tbody>
</table>

Genetic Code

- tRNA
- amino acid
- codon
The genetic code is a smooth mapping

- Degenerate (20 out of 64).
- Compactness of amino-acid regions.
- Smooth (similar “color” of neighbors).

Generic properties of molecular codes?
Fitter codes have minimal distortion

\[ Q = \langle C_{\alpha \omega} \rangle = \sum_{\text{paths}} P_{\text{path}} C_{\alpha \omega} = \text{Tr}(E \cdot R \cdot D \cdot C) \]

- Distortion of noisy channel, \( Q \) = average distortion of AA.
- \( R \) defines topology of codon space.
- \( C \) defines topology of amino-acid space.

Smooth codes minimize distortion

- Noise confuses close codons.
- Smooth code: close codons = close amino-acids.
  → minimal distortion.

Optimal code must balance contradicting needs for smoothness and diversity.

Max smoothness
Min diversity

Min smoothness
Max diversity

1 20 64
# amino-acids

(TT, Bio Phys 2008)
Code’s cost is the rate of the channel

- Diversity requires high specificity = high $\varepsilon_{\text{binding}}$.
- Cost $\sim <\varepsilon_{\text{binding}}>$.
- $P_{\text{binding}} \sim \text{Boltzmann: } E \sim e^{\varepsilon b / T}$.

$$I \sim \sum_{\alpha,i} E_{\alpha i} \ln E_{\alpha i} + \sum_{\alpha,i} D_{j\omega} \ln D_{j\omega} \sim \langle \varepsilon_{\alpha i} \rangle_E + \langle \varepsilon_{j\omega} \rangle_D$$

- Cost $I = \text{Channel Rate}$ (bits/message)
Code fitness combines rate and distortion of map

\[ \mathcal{F}(\beta, E, D, R, C) = -\beta Q - I \]

- Fitness \( \mathcal{F} \) is “free energy” with inverse “temperature” \( \beta \).
- *Gain* \( \beta \) increases with organism complexity and environment richness.
- Evolution varies the gain \( \beta \).
- Population of self-replicators evolving according to code fitness \( \mathcal{F} \): mutation, selection, random drift.

(TT, PRL, Bio Phys 2008)
Code emerges at a critical coding transition

• Low gain $\beta$ : Cost too high
  $\rightarrow$ no specificity $\rightarrow$ **no code**.

• $\beta$ increases: **Code emerges**
  channel starts to convey information ($I \neq 0$).

• Continuous 2$^{nd}$ order phase transition.

• Emergent map is smooth, low mode of $R$.

*Rate-distortion theory* (Shannon 1956)
Errors define the topology of the genetic code

- Codon graph = codon vertices + 1-letter difference edges (mutations).
- Lowest excited modes of graph-Laplacian $R$.
- Single maximum for lowest excited modes (Courant).
- Every mode corresponds to amino-acid:
  
  \[ \# \text{low modes} = \# \text{amino-acids}. \]

\[ K_4 \times K_4 \times K_4 \]

\[ \text{\rightarrow single contiguous domain for each amino-acid.} \]

\[ \text{\rightarrow Smoothness.} \]
Coloring number limits number of amino-acids

- Coloring number = \( \text{Min( # colors that suffices to color a map) } \)
- Topological invariant (function of genus):

\[ chr(\gamma) = \left\lfloor \frac{1}{2} \left( 7 + \sqrt{1 + 48\gamma} \right) \right\rfloor. \]

(Ringel & Youngs 1968)

\[ \text{max(# amino-acids)} = chr(\gamma) \]

- From Courant’s theorem + “convexity” (tightness).
- Genetic code: \( \gamma = 25-41 \rightarrow \text{coloring number} = 20-25 \text{ amino-acids} \)

Examples: four essential living systems

Homologous Recombination

Molecular codes (genetic code)

Photosynthesis

Chromosome organization
The problem of chromosomes

- Genes are divided into chromosomes.
- What is the optimal (?) number?
- What is the optimal (?) organization?
- Relation to cell type?

<table>
<thead>
<tr>
<th>Organism</th>
<th># chr</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. pilosula (ant)</td>
<td>2</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>8</td>
</tr>
<tr>
<td>Arabisdopsis</td>
<td>10</td>
</tr>
<tr>
<td>C. elegans</td>
<td>12</td>
</tr>
<tr>
<td>Rye</td>
<td>14</td>
</tr>
<tr>
<td>Corn</td>
<td>20</td>
</tr>
<tr>
<td>Chinese hamster</td>
<td>22</td>
</tr>
<tr>
<td>Budding yeast</td>
<td>32</td>
</tr>
<tr>
<td>Earthworm</td>
<td>36</td>
</tr>
<tr>
<td>Cat</td>
<td>38</td>
</tr>
<tr>
<td>Syrian Hamster</td>
<td>44</td>
</tr>
<tr>
<td>Human</td>
<td>46</td>
</tr>
<tr>
<td>Tobacco</td>
<td>48</td>
</tr>
<tr>
<td>Silkworm</td>
<td>56</td>
</tr>
<tr>
<td>Horse</td>
<td>64</td>
</tr>
<tr>
<td>Dog</td>
<td>78</td>
</tr>
<tr>
<td>Goldfish</td>
<td>100</td>
</tr>
<tr>
<td>Adder’s tongue</td>
<td>1400</td>
</tr>
</tbody>
</table>

Nucleus of human fibroblast (Cremer et al., 2005)
The Game of Chromosome Organization

“Casino chip shuffling”

(A) Given stacks of multi-color Chips.
(B) Divide chessboard into “Blocks”.
(C) Cover board with stacks.
(D) For each color: Shuffle the Blocks such that all chips of this color will be close as possible to each other.
Mapping between 3D organization and expression

$S$
Real Space

Cell Types

$M$
Function Space

Expression

Spatial organization

Hypothesis: relation between real space and function space based on optimality
### Optimal chromosome organization?

**Hypothesis:** co-expressed genes or active genes with similar function tend to reside in the same or in close chromosomes.

<table>
<thead>
<tr>
<th>Optimal organization</th>
<th>Casino chip shuffling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell type/function</td>
<td>color</td>
</tr>
<tr>
<td>Gene</td>
<td>Chip stack</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Block</td>
</tr>
<tr>
<td>Nucleus</td>
<td>Chessboard</td>
</tr>
<tr>
<td>Chrom. reorganization</td>
<td>Block shuffling</td>
</tr>
<tr>
<td>Close active genes</td>
<td>Close Same color chips</td>
</tr>
</tbody>
</table>

**Diagram:**
- **Nucleus**
- **Expression**
- **Casino chip shuffling**
  - Color
  - Chip stack
  - Block
  - Chessboard
  - Block shuffling
  - Close Same color chips
Possible mechanisms of chromosomal interactions

Expression space
- Genetic networks, co-regulation, co-expression.

Physical space
- Physical proximity boosts efficiency of transcription factories?
- Small nuclear RNA (snRNA)?
  small nuclear ribonucleoproteins (snRNP)?
- ....

Smoothness

Transcription factories: Genes from same or from different chromosomes may associate with polymerases in the same factory. (Sutherland & Bickmore, Nat Rev Gen 2009)
Spatial organization and expression are correlated

- Expression distance: \[ \ln \left( \frac{\langle \text{activity/gene} \rangle_i}{\langle \text{activity/gene} \rangle_j} \right) \]
- Physical distance: \[ \langle r_i - r_j \rangle \]
- Average over 54 nuclei yields significant correlation.

[Diagram showing correlation between spatial organization and expression]
Is chromosome organization cell specific?

- HUVEC - human umbilical cord vein endothelial cell.
- Oocyte – female germ cells.
- Fibroblast – connective tissue.
- Lung – epithelial cells.
"Fitness" of chromosome organization

\[ A_{ij}^{fib} = e^{\frac{IPD_{ij}^{fib}}{\lambda}} \]

Transcription Factor Networks

\[ \mathcal{F} = \sum_{ij} A_{ij}^{fib} \sum_{f} (\Phi_{if}^{cell} - \Phi_{if}^{cell})^2 \]

Gene Expression

\[ \Phi_{if}^{cell} = \ln(I_{if}^{cell}) \]

Network Matrix

- \( \mathcal{F} \) is "Smoothness" measure
  \( (\sim \text{distortion } Q) \):
  How close are gene that perform the same function in given cell type?

\[ \text{Counts} \]

\[ \text{Lung, HUVEC, Oocyte, Fibroblast} \]

Normalized \( \mathcal{F} \)
Questions and directions

- Better optimality measures (transcription factories).
- Other cell types (preliminary evidence from T cells).
- Optimality transition as a function of chromosome #.
- Relations to the topology of the chromosome graph ($A_{ij}$).

Other molecular information channels:

- molecular recognition, transcription networks.
THANKS

WIS
Yoni Savir
Uri Alon
Elisha Moses
Shalev Itzkovitz
Guy Shinar
Roy Bar-Ziv
Tsevi Beatus
Elad Noor
Ron Milo

Rockefeller
Albert Libchaber
NUS, NCBS
G.V. Shivashankar
S. Maharana
V.K. Iyer
Geneva
Jean-Pierre Eckmann

more:  www.weizmann.ac.il\complex\tlusty