Lecture 12: Epistasis and the General Problem

R. Ranganathan
Green Center for Systems Biology, ND11.120E

So…the full complexity of interactions within systems.

Sewall Wright
1889 - 1988

Ronald Fisher
1890 - 1962

Jeffries Wyman
1901 - 1995

Alan Fersht
1943 -
So, large-scale **non-linear dynamical systems**. First, we need a general theory of interactions…

<table>
<thead>
<tr>
<th>Linear</th>
<th>n = 1</th>
<th>n = 2 or 3</th>
<th>n &gt;&gt; 1</th>
<th>continuum</th>
</tr>
</thead>
<tbody>
<tr>
<td>exponential growth and decay</td>
<td>second order reaction kinetics</td>
<td>electrical circuits</td>
<td>Diffusion</td>
<td></td>
</tr>
<tr>
<td>single step conformational change</td>
<td>linear harmonic oscillators</td>
<td>molecular dynamics</td>
<td>Wave propagation</td>
<td></td>
</tr>
<tr>
<td>fluorescence emission</td>
<td>simple feedback control</td>
<td>systems of coupled harmonic oscillators</td>
<td>quantum mechanics</td>
<td></td>
</tr>
<tr>
<td>pseudo first order kinetics</td>
<td>sequences of conformational change</td>
<td>equilibrium thermodynamics</td>
<td>viscoelastic systems</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonlinear</th>
<th>Fixed points</th>
<th>Anharmonic oscillators</th>
<th>Systems of non-linear oscillators</th>
<th>Nonlinear wave propagation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bifurcations, multi stability</td>
<td>relaxation oscillations</td>
<td>non-equilibrium thermodynamics</td>
<td>Reaction-diffusion in dissipative systems</td>
</tr>
<tr>
<td></td>
<td>irreversible hysteresis</td>
<td>predator-prey models</td>
<td>protein structure/function</td>
<td>Turbulent/chaotic flows</td>
</tr>
<tr>
<td></td>
<td>overdamped oscillators</td>
<td>van der Pol systems</td>
<td>neural networks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chaotic systems</td>
<td></td>
<td>the cell</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ecosystems</td>
<td></td>
</tr>
</tbody>
</table>
A portion of core metabolism...
the essential characteristics of complex systems…

heterogeneity and non-linearity…
the essential characteristics of **complex systems**…

**heterogeneity** and **non-linearity**…

Some parts are much more important than others…

Parts don’t act independently….the whole displays behaviors that are much more than the summed action of the parts
complex systems show…

**heterogeneity** and **non-linearity**…

so, for biological systems, what are the **relevant parts**? And what is the **extent of non-linearity** in the interactions between parts?
complex systems show…

**heterogeneity** and **non-linearity**…

so, for biological systems, what are the **relevant parts**? And what is the **extent of non-linearity** in the interactions between parts?

How can we design a **process** to systematically address these questions? For next time…
What is the **extent of non-linearity** in the interactions between parts? We need a general formalism for “interactions”…
A general theory of **epistasis**

...in the context of proteins for clarity, but the ideas are completely general.
Epistasis is defined as the **non-independence** of the parts that make up a system.

~100 amino acids
What does **non-independence** mean, exactly?
What does **non-independence** mean, exactly?

Protein $X$ is in thermal equilibrium with its surroundings at some temperature absolute zero. Then due to thermal agitation, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, \ldots, x_N\}$$
What does non-independence mean, exactly?

Protein $X$ is in thermal equilibrium with its surroundings at some temperature absolute zero.

Then due to thermal agitation, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, \ldots, x_N\}$$

Take $x_0$ as the least energy state ... most populated at equilibrium.

We know that:

$$\frac{P(x_n)}{P(x_0)} = e^{-\frac{A6_n}{k_BT}}$$
What does non-independence mean, exactly?

Protein $X$ is in thermal equilibrium with its surroundings at some temperature absolute zero.

Then due to thermal equilibration, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, ..., x_N\}$$

Now... we divide the protein into a bunch of parts... say for convenience into all its $M$ residues:

$$r \in \{r_1, r_2, r_3, ..., r_i, ..., r_M\}$$

How many observations of free energies of the parts $r$ do we need to guarantee knowledge of states $x$?
What does **non-independence** mean, exactly?

...the effect of a **single part** taken independently
What does **non-independence** mean, exactly?

...the effect of a **two parts** acting together
What does **non-independence** mean, exactly?

---

1. State $x_{ij}$ is not necessarily predictable from $x_i$ or $x_j$ if $r_i$ and $r_j$ are allowed to potentially be coupled.

2. If $r_i$ and $r_j$ are known to be independent, then $\Delta \Delta 6_{ij} = 0$ and things reduce to a more simple case.

3. Couplings cause non-linearities in system information (i.e., energetic variables).
What does *non-independence* mean, exactly?

If we are measuring an *equilibrium state function*, then independence means additivity in that variable.

And, non-independence (or epistasis) means *non-additivity*.
Thermodynamic additivity is **statistical independence**…

\[
\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta \Delta G_{ij}
\]

If \( x_i, x_j \) independent, \( \Delta \Delta G_{ij} = 0 \)

\[
\begin{align*}
\frac{P(x_i)}{P(x_0)} &= e^{-\frac{\Delta G_i}{kT}} \\
\frac{P(x_j)}{P(x_0)} &= e^{-\frac{\Delta G_j}{kT}} \\
\frac{P(x_{ij})}{P(x_0)} &= e^{-\frac{\Delta G_{ij}}{kT}}
\end{align*}
\]
Thermodynamic additivity is **statistical independence**...
What does non-independence mean, exactly?

….the effect of a two parts acting together
What does non-independence mean, exactly?
What does non-independence mean, exactly?
What does **non-independence** mean, exactly?

\[
\Delta G_n = \sum_{i=1}^{M} \Delta G + \sum_{i=2}^{M} \Delta^2 G + \sum_{i=3}^{M} \Delta^3 G + \ldots + \sum_{i=n}^{M} \Delta^n G + \ldots + \sum_{i=M}^{M} \Delta^M G
\]

\[
\binom{M}{n} = \frac{M!}{n!(M-n)!}
\]

...the case of **all parts** acting cooperatively
What does **non-independence** mean, exactly?

\[ \Delta G_n = \sum \Delta G + \sum \Delta^2 G + \sum \Delta^3 G + \ldots + \sum \Delta^n G + \ldots + \sum \Delta^M G \]

\[
\binom{m}{k} = \frac{m!}{k!(m-k)!}
\]

So... with no knowledge of independence: need

\[
\sum_{k=1}^{m} \frac{m!}{k!(m-k)!}
\]

Numbers to thermodynamically know all states \( x \).

...the case of **all parts** acting cooperatively
A warning....!!

**Note**….epistasis is the non-independence of the parts that make up a system. But, it is sometimes **not obvious** what this means for parameters we might measure in general. Is independence necessarily additivity of a thing?

Needs to be validated…
So…a formal theory of epistasis in systems. We will use point mutagenesis in proteins as a case study…
Epistasis is important for function...
...and for **evolution**

![Diagram showing fitness landscapes in sequence space for less epistatic and more epistatic scenarios.](image)
Epistasis can effectively block off mutational paths between functional genotypes...
...and for evolution

[3D diagrams showing fitness landscapes in sequence space, transitioning from less epistatic to more epistatic conditions]
So…a formal theory of epistasis in systems. We will use point mutagenesis in proteins as a case study…
Orders of epistasis...

**Fig. 1.** Abstract representation of variants involved in (A) single mutant, (B) double mutant, and (C) triple mutant experiments. Genotypes are denoted as $g = \{g_N, \ldots, g_1\}$ with $g_i \in \{0, 1\}$, where '0' or '1' indicate the state of the mutable entity (e.g., the gene or the amino acid in the protein). Each genotype $g$ has an associated phenotype $y_g$. The effect of a single, double, triple mutation is given by the red arrows. Two-way epistasis is defined as the differential effect of a mutation depending on the background in which it occurs, for example in (B) it is the phenotypic difference between mutating one position from '0' to '1' depending on the other position being either '0' or '1', i.e., $(y_{11} - y_{10}) - (y_{01} - y_{00})$. 
Orders of epistasis…

The **fourth order epistasis** is the degree to which a third order epistasis depends on a fourth mutation…and so on. In general, **$n^{th}$ order epistasis** is the difference of two $n-1$ dimensional epistases.
I. The biochemical view of epistasis...

In biochemistry the idea is to start with one genotype (the "wild-type") and use this as a reference for all effects of mutations....

wild-type binding free energy \((\text{in kcal/mol})\)

\(y_0\)
The biochemical view of epistasis...

In biochemistry the idea is to start with one genotype (the “wild-type”) and use this as a reference for all effects of mutations....

wild-type binding free energy (in kcal/mol)

\[ y_0 \]

the zeroth order term
The biochemical view of epistasis...

The single mutation experiment....
The biochemical view of epistasis...

The effect of a single mutant....

\[ \varepsilon_1 = y_1 - y_0 \]

a first-order term
The biochemical view of epistasis...

A double mutation experiment....
The biochemical view of epistasis…

The second order epistasis of two mutations is the degree to which the effect of one mutation depends on the background of a second…
The biochemical view of epistasis...

\[ \varepsilon_{111} = (y_{111} - y_{101} - y_{110} + y_{100}) - (y_{011} - y_{001} - y_{010} + y_{000}) \]

The **third order epistasis** is the degree to which a second order epistasis depends on a third mutation…
The biochemical view of epistasis...

The **fourth order epistasis** is the degree to which a third order epistasis depends on a fourth mutation...
The biochemical view of epistasis...

\[
\Delta^n G_{1,...,n} = \Delta G_{1,...,n}^0 + (-1)^1 \sum_{i_1 < i_2 < \ldots < i_{n-1}}^{n} \Delta G_{i_1,i_2,\ldots,i_{n-1}}^0
+ (-1)^2 \sum_{i_1 < i_2 < \ldots < i_{n-2}}^{n} \Delta G_{i_1,i_2,\ldots,i_{n-2}}^0 + \ldots + (-1)^n \Delta G_{i_1,i_2,\ldots,i_n}^0
\]

In general, there is a hierarchical expansion of terms....
The biochemical view of epistasis...

In matrix form...

\[ \bar{e} = G \bar{y} \]
The biochemical view of epistasis...

\[ \bar{\varepsilon} = G \bar{y} \]

\begin{itemize}
\item the phenotypes
\item the epistases
\end{itemize}
The biochemical view of epistasis...

\[ \bar{e} = G \bar{y} \]

...this is a **transform**, just like Fourier or Laplace. A one-to-one mapping between a space of phenotypes and epistases.

You've already been introduced to many transforms....
The concept of transforms

The phenotypes

\[ \bar{\epsilon} = G \bar{y} \]

the epistases

In general, a transform is a mapping of variables from one space to another. Why do this?
The concept of transforms

In general, a **transform** is a mapping of variables from one space to another. Why do this? Well... problems that are initially defined in one space - or one parametrization - might in fact be a lot simpler in another...
The concept of transforms

An example….

(LXXIV) times (XXVIII) → A multiplication problem in Roman numerals…
The concept of transforms

(LXXIV) times (XXVIII)

Step 1. Make a transformation

(74) times (28) → A multiplication problem in Hindu-Arabic numerals…
The concept of transforms

Step 1. Make a transformation

(LXXIV) times (XXVIII)

Step 2. Solve it

(74) times (28)

2072

The answer
The concept of transforms

Step 1. Make a transformation

(LXXIV) times (XXVIII)

Step 2. Solve it

(74) times (28)

Step 3. Inverse transform

2072

MMLXXII → The answer in Roman numerals...
The concept of transforms

\[ \frac{dA}{dt} = -kA, \quad A(0) = A_0 \]
The concept of transforms

**Step 1.** Make a transformation

\[ \frac{dA}{dt} = -kA, \quad A(0) = A_0 \]

\[ \mathcal{L}\{f(t)\} = F(s) = \int_0^\infty f(t)e^{-st}dt \]

The LaPlace transform….
The concept of transforms

Step 1. Make a transformation

\[
\frac{dA}{dt} = -kA, \quad A(0) = A_0
\]

\[
L\{\frac{dA}{dt}\} = sA(s) - A(0)
\]

\[
A(s) = \frac{A_0}{s+k}
\]

Step 2. Solve it

The LaPlace solution…
The concept of transforms

Step 1. Make a transformation

\[ \frac{dA}{dt} = -kA, \quad A(0) = A_0 \]

Step 2. Solve it

\[ L\left\{ \frac{dA}{dt} \right\} = L\{-kA\} \]

\[ A(s) = \frac{A_0}{s + k} \]

Step 3. Inverse transform

\[ L^{-1}(A(s)) = L^{-1}\left[ \frac{A_0}{s + k} \right] \]

So...

\[ f(t) \quad F(s) \]

\[ e^{-kt} \quad \frac{1}{s + k} \]
The concept of transforms

**Step 1.** Make a transformation

\[
\frac{dk}{dt} = -kA, \quad A(0) = A_0
\]

**Step 2.** Solve it

\[
L\left\{\frac{dk}{dt}\right\} = k^2 \left\{-kA\right\}
\]

\[
A(s) = \frac{A_0}{s + k}
\]

**Step 3.** Inverse transform

\[
A(t) = A_0 e^{-kt}
\]

The time-domain solution…
The biochemical view of epistasis...

...this is a **transform**, just like Fourier or Laplace. A one-to-one mapping between a space of phenotypes and epistases. What kind of transform? wait...
The biochemical view of epistasis...

\[ \bar{e} = G \bar{y} \]

the epistasis operator
The biochemical view of epistasis...

\[ \bar{\varepsilon} = G \bar{y} \]

for 3 mutations...third order

\[
\begin{pmatrix}
\varepsilon_{000} \\
\varepsilon_{001} \\
\varepsilon_{010} \\
\varepsilon_{011} \\
\varepsilon_{100} \\
\varepsilon_{101} \\
\varepsilon_{110} \\
\varepsilon_{111}
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
y_{000} \\
y_{001} \\
y_{010} \\
y_{011} \\
y_{100} \\
y_{101} \\
y_{110} \\
y_{111}
\end{pmatrix}
\]
The biochemical view of epistasis…

\[ \bar{\varepsilon} = G \bar{y} \]

for 3 mutations…third order

\[
\begin{pmatrix}
\varepsilon_{000} \\
\varepsilon_{001} \\
\varepsilon_{010} \\
\varepsilon_{011} \\
\varepsilon_{100} \\
\varepsilon_{101} \\
\varepsilon_{110} \\
\varepsilon_{111}
\end{pmatrix} =
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 & 1
\end{pmatrix} \times
\begin{pmatrix}
y_{000} \\
y_{001} \\
y_{010} \\
y_{011} \\
y_{100} \\
y_{101} \\
y_{110} \\
y_{111}
\end{pmatrix}
\]

\[ \varepsilon_{011} = (y_{011} - y_{001}) - (y_{010} - y_{000}) \]
The biochemical view of epistasis...

\[ \bar{\varepsilon} = G \bar{y} \]

for 3 mutations...third order

\[
\begin{pmatrix}
\varepsilon_{000} \\
\varepsilon_{001} \\
\varepsilon_{010} \\
\varepsilon_{011} \\
\varepsilon_{100} \\
\varepsilon_{101} \\
\varepsilon_{110} \\
\varepsilon_{111}
\end{pmatrix} = 
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1
\end{pmatrix}
\] * 

\[
\begin{pmatrix}
\varepsilon_{111} = [(y_{111} - y_{101}) - (y_{110} - y_{100})] - [(y_{011} - y_{001}) - (y_{010} - y_{000})]
\end{pmatrix}
\]
The biochemical view of epistasis...

\[ \bar{e} = G \bar{y} \]

A recursive **generative function** for \( n^{th} \)-order epistasis....

\[ G_{n+1} = \begin{pmatrix} G_n & 0 \\ -G_n & G_n \end{pmatrix} \quad \text{with} \quad G_0 = 1 \]
The biochemical view of epistasis...

\[ G_{n+1} = \begin{pmatrix} G_n & 0 \\ -G_n & G_n \end{pmatrix} \text{ with } G_0 = 1 \]

for \( n = 0 \)…zeroth order (wild-type)
The biochemical view of epistasis...

\[
G_{n+1} = \begin{pmatrix}
G_n & 0 \\
-G_n & G_n
\end{pmatrix} \text{ with } \ G_0 = 1
\]

for \( n = 1 \)…first order (a single mutation experiment)

\[
\begin{pmatrix}
\varepsilon_{000} \\
\varepsilon_{001} \\
\varepsilon_{010} \\
\varepsilon_{011} \\
\varepsilon_{100} \\
\varepsilon_{101} \\
\varepsilon_{110} \\
\varepsilon_{111}
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 & 1
\end{pmatrix} \begin{pmatrix}
y_{000} \\
y_{001} \\
y_{010} \\
y_{011} \\
y_{100} \\
y_{101} \\
y_{110} \\
y_{111}
\end{pmatrix}
\]
The biochemical view of epistasis...

\[ G_{n+1} = \begin{pmatrix} G_n & 0 \\ -G_n & G_n \end{pmatrix} \quad \text{with} \quad G_0 = 1 \]

for \( n = 2 \)...second order (a double mutation experiment)
The biochemical view of epistasis...

\[ G_{n+1} = \begin{pmatrix} G_n & 0 \\ -G_n & G_n \end{pmatrix} \text{ with } G_0 = 1 \]

for \( n = 3 \)…third order (a triple mutation experiment)
In the biochemical view of epistasis...we take one particular genotype (the “wild-type”) as our reference...

\[ \bar{e} = G \bar{y} \]
In this view...

(1) Large third order effect….all three mutations represent a collectively cooperative group.
In this view...

(1) Large third order effect….all three mutations represent a collectively cooperative group.

(2) Large first order effects of the H372A and T-2F single mutations
In this view...

(1) Large third order effect…all three mutations represent a collectively cooperative group.

(2) Large first order effects of the H372A and T-2F single mutations

(3) Large second order epistatic coupling between G330T and T-2F
Another view of epistasis...background averaging!

In this view....there is no reference sequence. Values are averaged over all genotypes. For example...
Another view of epistasis...background averaging!

In biochemistry, the second order epistasis would be:

\[ \varepsilon_{11} = (y_{11} - y_{01}) - (y_{10} - y_{00}) \]
Another view of epistasis...background averaging!

In biochemistry, the second order epistasis would be:

\[ \varepsilon_{11} = (y_{11} - y_{01}) - (y_{10} - y_{00}) \]

In this view, it is an average in the background of two genotypes... \( y_{0**} \) and \( y_{1**} \)

\[ \varepsilon_{*11} = \frac{[(y_{111} - y_{101}) - (y_{110} - y_{100})] + [(y_{011} - y_{001}) - (y_{010} - y_{000})]}{2} \]
Background averaged epistasis…

Has a matrix form too….

\[ \bar{e} = H \bar{y} \]

…the background-averaged epistasis operator

…also a **transform**!
Background averaged epistasis…

Has a matrix form too….

\[ \bar{e} = H \bar{y} \]

\[
\begin{pmatrix}
-1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\
\end{pmatrix}
\begin{pmatrix}
-1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\
\end{pmatrix}
\begin{pmatrix}
y_{000} \\
y_{001} \\
y_{010} \\
y_{011} \\
y_{100} \\
y_{101} \\
y_{110} \\
y_{111} \\
\end{pmatrix}
\]
Background averaged epistasis…

$$
\begin{pmatrix}
  e_{+++} \\
  e_{++1} \\
  e_{+1*} \\
  e_{+11} \\
  e_{1**} \\
  e_{1*1} \\
  e_{11*} \\
  e_{111}
\end{pmatrix}
= 
\begin{pmatrix}
  1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
  1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\
  1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\
  1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\
  1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\
  1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \\
  1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
  1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\
\end{pmatrix}
\times
\begin{pmatrix}
  y_{000} \\
  y_{001} \\
  y_{010} \\
  y_{011} \\
  y_{100} \\
  y_{101} \\
  y_{110} \\
  y_{111}
\end{pmatrix}
$$

$$
\varepsilon_{*11} = \frac{[(y_{111} - y_{101})-(y_{110} - y_{100})] + [(y_{011} - y_{001})-(y_{010} - y_{000})]}{2}
$$

Makes sense?
Background averaged epistasis…

\[
\begin{pmatrix}
    c_{***} \\
    c_{**1} \\
    c_{*1*} \\
    c_{*11} \\
    c_{1**} \\
    c_{1*1} \\
    c_{11*} \\
    c_{111}
\end{pmatrix}
= 
\begin{pmatrix}
    1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
    1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\
    1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 \\
    1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\
    1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\
    1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
    1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \\
    1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1
\end{pmatrix}
\times
\begin{pmatrix}
    y_{000} \\
    y_{001} \\
    y_{010} \\
    y_{011} \\
    y_{100} \\
    y_{101} \\
    y_{110} \\
    y_{111}
\end{pmatrix}
\]

\[
\varepsilon_{*11} = 2 \left\{ (y_{111} - y_{101}) - (y_{110} - y_{100}) \right\} + \left\{ (y_{011} - y_{001}) - (y_{010} - y_{000}) \right\}
\]

There is a weighting for number of terms over which the average is taken…
Background averaged epistasis...

There is a weighting for number of terms over which the average is taken...
Background averaged epistasis...

\[
\begin{pmatrix}
E_{+++} \\
E_{++1} \\
E_{+1*} \\
E_{1**} \\
E_{1*1} \\
E_{11*} \\
E_{111}
\end{pmatrix} = 
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{pmatrix} \ast 
\begin{pmatrix}
Y_{000} \\
Y_{001} \\
Y_{010} \\
Y_{011} \\
Y_{100} \\
Y_{101} \\
Y_{110} \\
Y_{111}
\end{pmatrix}
\]

\[\varepsilon_{+++} = \frac{y_{111} + y_{101} + y_{110} + y_{100} + y_{011} + y_{001} + y_{010} + y_{000}}{8}\]

\[\bar{\varepsilon} = VH\bar{y} . \quad \text{...the background averaged epistasis}\]
Background averaged epistasis...

\[
\bar{\epsilon} = VH\bar{y}.
\]

the recursive **generative function** for \(n^{\text{th}}\)-order background averaged epistasis...

\[
V_{n+1} = \begin{pmatrix} \frac{1}{2}V_n & 0 \\ 0 & -V_n \end{pmatrix} \quad \text{with} \quad V_0 = 1
\]

\[
H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix} \quad \text{with} \quad H_0 = 1
\]
So…

(1) Large third order effect….all three mutations represent a collectively cooperative group.
(1) Large third order effect….all three mutations represent a collectively cooperative group.

(2) First order effects of all mutations essentially negligible.

this is because of higher order epistasis…H372A has a 2.05 kcal/mol effect in the wild-type background and a -1.71 kcal/mol effect in the T-2F background!

It’s not a general determinant of affinity!
So...

(1) Large third order effect... all three mutations represent a collectively cooperative group.

(2) First order effects of all mutations essentially negligible.

(3) Coupling of G330T and T-2F is lower

this is because of higher order epistasis... H372A has a 2.05 kcal/mol effect in the wild-type background and a -1.71 kcal/mol effect in the T-2F background!

It's not a general determinant of affinity!
Another case....

A high-order cooperativity in gating of the K+ channel pore....
The biochemical view of epistasis is a “local” one… taking a single genotype as an arbitrary reference.

\[ \bar{\varepsilon} = G \bar{y} \]
The biochemical view of epistasis is a “local” one… taking a single genotype as an arbitrary reference.

\[ \bar{\varepsilon} = G \bar{y} \]

…the biochemical view of epistasis is a “local” one… taking a single genotype as an arbitrary reference.

…its like doing a Taylor’s (local) expansion of the fitness landscape around a particular point (the reference genotype). A detailed analysis of a particular solution.
The bottom line…

The background averaged view of epistasis is a "global" one…taking averages over all possible genotypes.
The bottom line...

\[ \bar{\epsilon} = V H \bar{y}. \]

...its like doing a generalized Fourier expansion of the fitness landscape. A global analysis of all possible solutions given the design process.

the background averaged view of epistasis is a “global” one...taking averages over all possible genotypes.
in fact, this mapping has a name….it is a generalized form of a Fourier transform called the Hadamard transform.

\[ \tilde{\varepsilon} = V H \tilde{y}. \]
Fourier series…  Hadamard series…

a representation based on just **square** functions…
with components just +1 and -1.
The bottom line…

...in general, **background averaging** seems like the right thing to do if we want to learn the general rules for specifying systems.

\[ \bar{\varepsilon} = V H \bar{y}. \]
Ok...so **background averaged** epistasis is what we want?

\[ \bar{\epsilon} = V H \tilde{y}. \]

But...background averaging is a ridiculous concept in general. In examples, we analyzed single mutations at like 4 positions \((2^4)\). For 20 possible amino acids in a protein with 100 positions?
background averaged epistasis…

\[
\tilde{\epsilon} = V H \tilde{y}.
\]

So…the bottom line is that we need a general strategy to learn the epistasis vector from a practical set of experiment. Can we do this?
heterogeneity and non-linearity…

For most systems, we don't know either the relevant parts, or the non-linear interactions between them. Why?
**heterogeneity** and **non-linearity**…

For most systems, we don't know either the relevant parts, or the non-linear interactions between them. Why?

(1) We do experiments in **specific model systems** under specific, carefully controlled experimental conditions in the laboratory….what does this do?

(2) We do experiments typically one the scale of **one part at a time**…single mutations, single gene knockouts, single drug applications, single cell recordings…what does this do?

….what is **an optimal strategy** to address this problem?

\[ \tilde{c} = VH \tilde{y}. \]
So, large-scale **non-linear dynamical systems**. Next time….the problem of proteins

<table>
<thead>
<tr>
<th>Linear</th>
<th>n = 1</th>
<th>n = 2 or 3</th>
<th>n &gt;&gt; 1</th>
<th>continuum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exponential growth and decay</td>
<td>second order reaction kinetics</td>
<td>electrical circuits</td>
<td>Diffusion</td>
</tr>
<tr>
<td></td>
<td>single step conformational change</td>
<td>linear harmonic oscillators</td>
<td>molecular dynamics</td>
<td>Wave propagation</td>
</tr>
<tr>
<td></td>
<td>fluorescence emission</td>
<td>simple feedback control</td>
<td>systems of coupled harmonic oscillators</td>
<td>quantum mechanics</td>
</tr>
<tr>
<td></td>
<td>pseudo first order kinetics</td>
<td>sequences of conformational change</td>
<td>equilibrium thermodynamics</td>
<td>viscoelastic systems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonlinear</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed points</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bifurcations, multi stability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>irreversible hysteresis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overdamped oscillators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anharmonic oscillators</td>
<td></td>
<td></td>
<td></td>
<td>Nonlinear wave propagation</td>
</tr>
<tr>
<td>relaxation oscillations</td>
<td></td>
<td></td>
<td></td>
<td>Reaction-diffusion in dissipative systems</td>
</tr>
<tr>
<td>predator-prey models</td>
<td></td>
<td></td>
<td></td>
<td>Turbulent/chaotic flows</td>
</tr>
<tr>
<td>van der Pol systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chaotic systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>systems of non-linear oscillators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-equilibrium thermodynamics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protein structure/ function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neural networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the cell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ecosystems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

adapted from S. Strogatz