Topic Course

Gene and protein evolution

Winter 2016
Department of Molecular Genetics
University of Toronto

Lecture 6

Hue Sun Chan

Theory of protein sequence space organization and the dynamics of molecular evolution

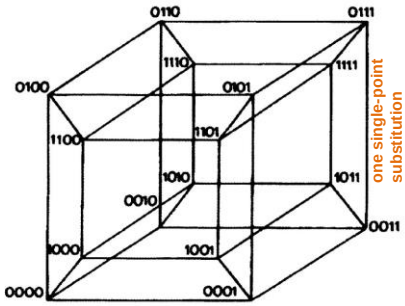
- Sequence-space topology, neutral nets, and their biophysical basis
- Simple exact models of the mapping between protein sequences and structures; the hydrophobic-polar model
- Superfunnels and mutational robustness, correlation between thermodynamic and mutational stabilities
- Recombination, local conformational preference, and autonomous folding units
- Promiscuous functions and conformational switches
- Escape from adaptive conflicts: mutational robustness and subfunctionalization

Reference: Sikosek & Chan (2014) Biophysics of protein evolution and evolutionary protein biophysics. *J Royal Society Interface* **11:**20140419 (http://rsif.royalsocietypublishing.org/content/11/100/20140419.full).

Models of structural evolution of proteins are based on a presumed mapping between protein sequences and structures and a presumed relationship between structure and function

Evolution as adaptive walk in sequence space

- Sequence space may be pictured as a multi-dimensional "rugged landscape"



Example of a model sequence space: A *four-dimensional* Boolean *hypercube*, in which each of the 16 vertices represents one of the possible strings of four 0 or 1 values. Here each such string is interpreted as a specific tetrapeptide with two possible types of amino acid at each position.

The *NK* model

The spin-glasses inspired ***NK*** model defines a combinatorial space consisting of every string ***s*** of length ***N***, each site of the string is chosen from an alphabet of ***A*** alternative states – e.g., the 20 amino acids. A scalar value (which may be identified with "fitness" ***F***) is assigned to each string, where

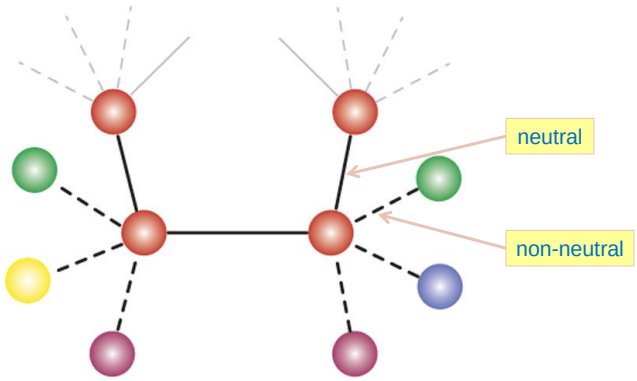
$$F(s) = \frac{1}{N}\sum_{i=1}^{N} F^{(K)}{}_i(s_i; s_{i1}, \ldots, s_{iK})$$

If a distance metric is defined between strings (e.g. Hamming distance, i.e., number of amino acid substitutions), the resulting construct is a fitness landscape.

***Figure from the following references for the NK model*:**

Kauffman & Levin (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128:**11-45.
Kauffman & MacReady (1995) Search strategies for applied molecular evolution. *J Theor Biol* **173:**427-440.

Example of genotype-phenotype map in the modeling of evolution

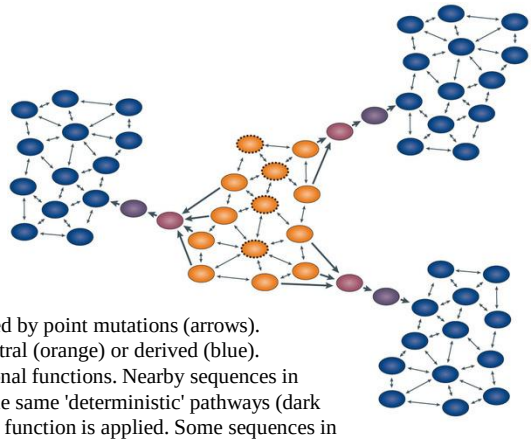


Schematic representation of the genotype–phenotype map used in the analysis of Draghi et al.* Each circle corresponds to a genotype; colours denote phenotypes. A proportion *q* of mutations are neutral (solid lines) and the remaining mutations are non-neutral (dashed lines). A non-neutral mutation changes an individual's phenotype to one of the *K* accessible alternatives that form the individual's phenotypic neighbourhood. When *K* is smaller than the total number of alternative phenotypes in the landscape, *P*, individuals may have different phenotypic neighbourhoods. The central pair of adjacent genotypes shown here express the same phenotype, but they have different phenotypic neighbourhoods.

***Figure from*:** Draghi, Parsons, Wagner & Plotkin (2010)
Mutational robustness can facilitate adaptation. *Nature* **463:**353-355.

Concepts of genotype-phenotype (sequence-structure) map and neutral net have long been used in theories of protein evolution

Neutral net or network: sequences encoding for the same structure/function and are interconnected by single-point mutations



Protein sequences (ovals) are connected by point mutations (arrows). The colours represent functions: ancestral (orange) or derived (blue). Transitional colours represent transitional functions. Nearby sequences in the ancestral neutral network follow the same 'deterministic' pathways (dark arrows) when selection for the derived function is applied. Some sequences in the neutral network cannot achieve the derived function without first taking a permissive functionally neutral step through the network (ovals with dashed outlines).

***Figure from*:** Harms & Thornton (2013) Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nature Review Genetics* **14:**559-571.

Although theories based on *ad hoc* genotype-phenotype maps or maps that were not derived from explicit-chain protein models (e.g. the *NK* model) have led to much advance, it has become increasingly clear that sequence-structure maps based on explicit-chain biophysics are essential for gaining insights into the evolution of protein folds.

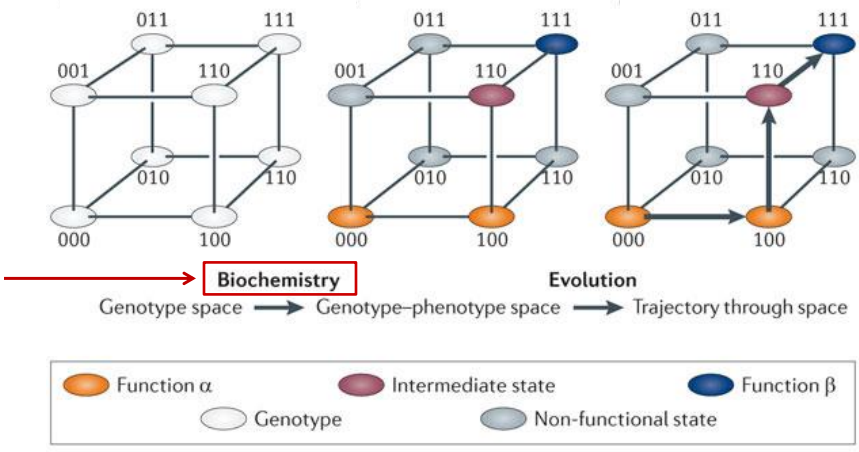


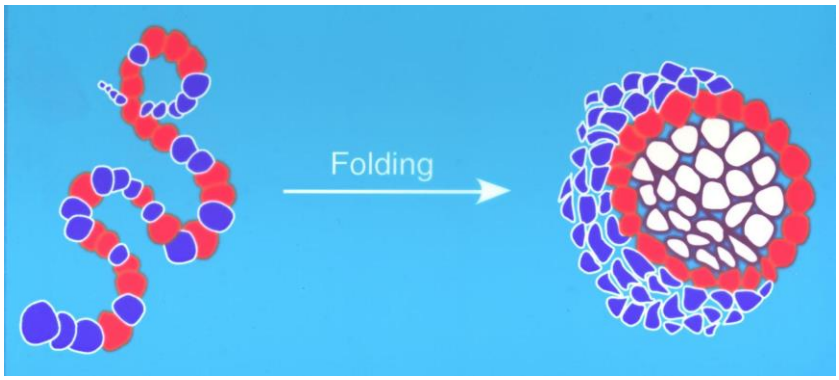
***Figure from*:** Harms & Thornton (2013) Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nature Review Genetics* **14:**559-571.

Conceptual Models of Protein Evolution

An atomistic sequence-structure map is currently not feasible to construct, because such a feat amounts to first solving the protein folding problem for all possible amino acid sequences. In view of this fundamental limitation in theory and computation, highly simplified *explicit*-chain models that aim to capture the essential biophysics of protein folding are being used to develop theories of protein evolution.

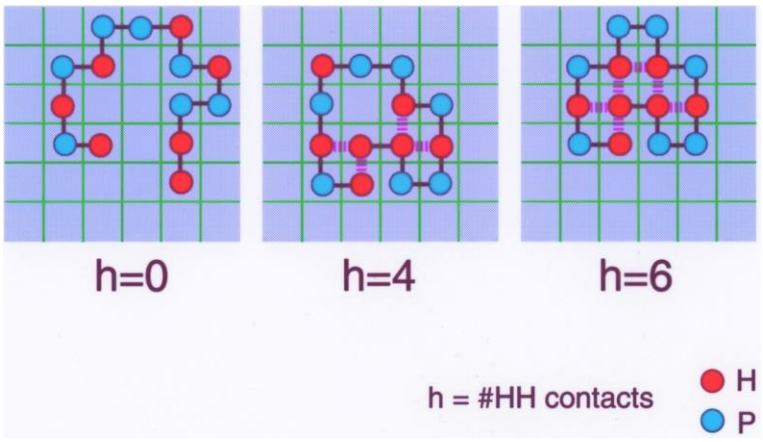
Hydrophobic Interaction is a Major Driving Force for Protein Folding

Kauzmann (1959), Dill (1990), etc.



Chan & Dill, *Physics Today* (1993)

A Simple Exact Biophysical Model: The Hydrophobic-Polar (HP) Model



Lau & Dill, *Macromolecules* (1989); Chan & Dill, *J Chem Phys* (1991) • ***Figure from***: Chan & Dill, *Physics Today* (1993)

The Hydrophobic-Polar (HP) Lattice Protein Model

- Hydrophobic-hydrophobic contacts are favorable
- Allows for exact (exhaustive) enumeration of sequences and conformations
- Provides an exact sequence-to-structure mapping for the study of evolution

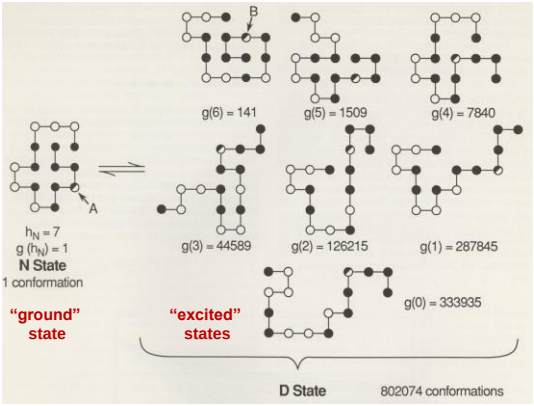
[Chan & Bornberg-Bauer, *Appl Bioinformatics* (2002) ; Xia & Levitt, *Curr Opin Struct Biol* **14**, 202 (2004)]

Short-chain 2D HP model capture several essential features of the sequence-to-structure mapping of real proteins [Dill et al & Chan, *Protein Sci* (1995)].

~ hydrophobic-polar statistics: Irbäck *et al*, *PNAS* **93,** 9533 (1996); Irbäck & Sandelin, *Biophys J* **79,** 2252 (2000)

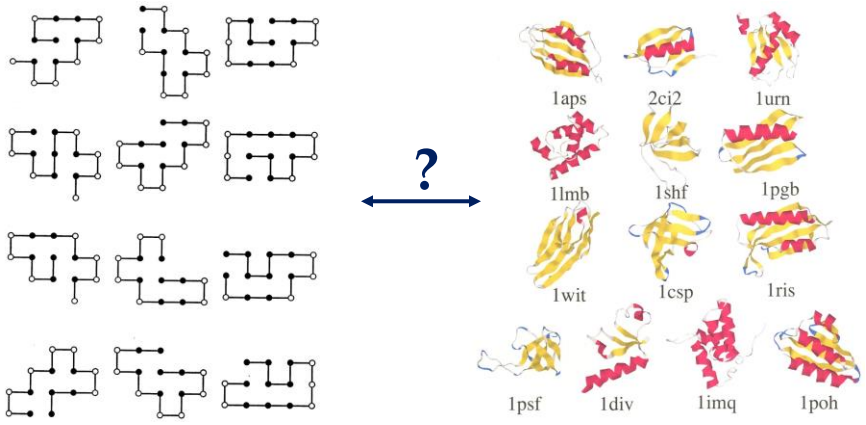
~ comparative ("homology") modeling: Moreno-Hernández & Levitt, *Proteins* **80**, 1683 (2012)

- only ~ 2% of sequences are encoding uniquely



***Drawing from*:** Shortle, Chan & Dill, *Protein Sci* (1992)

What properties of real proteins can simple HP model capture?



- Not every property; e.g., HP lattice proteins undergo much less cooperative folding than real proteins. But the HP model is apparently quite reasonable as a model for the mapping between protein sequences and their ground-state (lowest-free-energy) structures.

HP pattern is an important determinant of protein structure

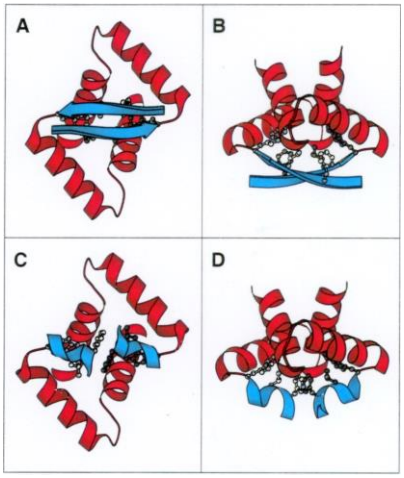


Figure from: Cordes, Walsh, McKnight & Sauer, *Science* **284**, 325 (1999)

wildtype arc repressor (8-46)

--QFNLRW--

↓↓

--QFLNRW--

Asn11↔Leu12 "switch" double mutant

("PH"→"HP")

The **N11L** "bridge" single-mutant folds into two nearly equally populated native structures

--QFNLRW--

↓

--QFLLRW--

("PH"→"HH")

Cordes *et al*. & Sauer, *Nature Struct Biol* (2000)

Sequence Pattern Statistics of HP Model Proteins are similar to that of Real Proteins

$\sigma_i = \pm 1$, where +1 for **H**; -1 for **P**

for real proteins, **H** = L, I, V, F, M, or W; **P** otherwise

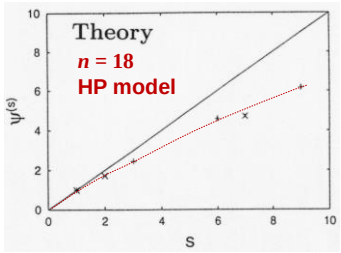
Consider mean-square block fluctuation

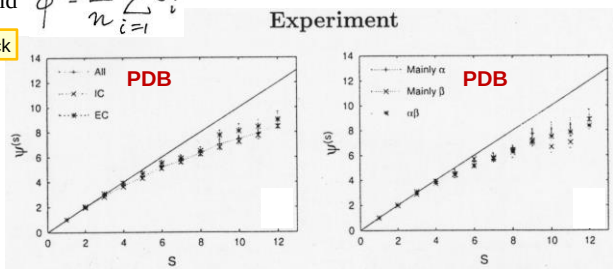
$$\psi^{(s)} = \frac{s}{n} \sum_{k=1}^{n/s} \psi_k^{(s)}$$

where

$$\psi_k^{(s)} \propto \left(\sigma_k^{(s)} - s\phi\right)^2 \quad \text{and} \quad \phi = \frac{1}{n} \sum_{i=1}^{n} \sigma_i$$

sum of σ_i over the *k*th *s*-block



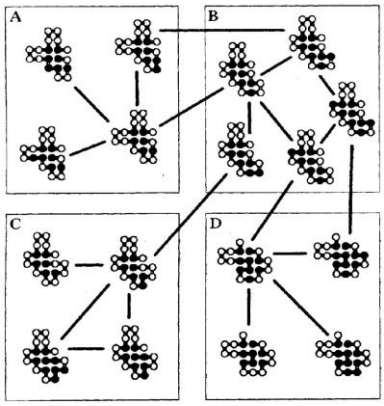


- **Similar nonrandomness is observed for uniquely encoding sequences in the HP model and for real proteins.**

Figures and analysis taken from: Irbäck & Sandelin, *Biophys J* **79**, 2252 (2000)

Early applications of the HP model to the study of evolution

Lipman & Wilbur* recognized that the two-dimensional (2D) HP model can provide a *biophysical* assessment of J. Maynard-Smith's idea (1970) that 'if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through non-functional intermediates' [*Nature* **225:**563-564]. The authors pointed out that Maynard-Smith's concept is related to S. Wright's more general notion of an adaptive landscape (1932), and noted further that in modeling adaptive landscapes, the 2D HP model sequence-structure map offers a useful, more physical alternative to ***NK*** approaches that treated this map as random. Lipman & Wilbur designate HP sequences that do not have a unique ground-state contact pattern as nonviable and thus mutations to these non-encoding sequences as "lethal". Within this modeling construct, Maynard-Smith's condition was found to be satisfied by the 2D HP model.

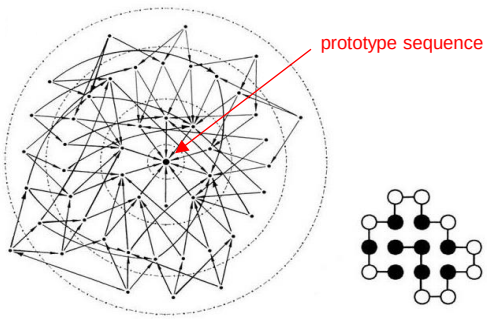


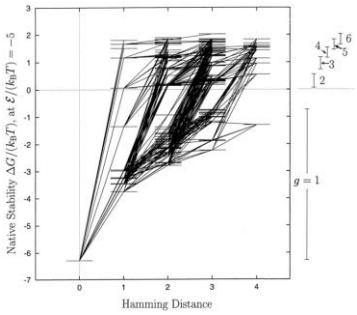
● Importance of neutral mutations: Any path from phenotype A to C requires at least one neutral mutation in phenotype B (figure from Lipman & Wilbur, 1991*).

Lau & Dill (1990) Theory for protein mutability and biogenesis. *PNAS* **87:**638-642.
Chan & Dill (1991) Sequence space soup of proteins and copolymers. *J Chem Phys* **95:**3775-3787;
* Lipman & Wilbur (1991) Modeling neutral and selective evolution of protein folding. *Proc R Soc B-Biol Sci* **245:**7-11.

Superfunnels:

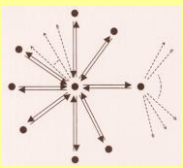
Correlation between Thermodynamic and Mutational Stabilities





Neutral net of sequences encoding for a given structure tends to organized around a "prototype sequence" with maximum thermodynamic stability and mutational stability (robustness).

● Sequences with higher mutational robustness tend to have higher steady-state populations under evolutionary dynamics



cf. van Nimwegen *et al*, *PNAS* **96,** 9716 (1999)

Bornberg-Bauer & Chan (1999) *PNAS* **96:**10689-10694;
Wroe, Bornberg-Bauer & Chan (2005) *Biophys J* **88:**118-131.

Master Equations may be used to model evolutionary population dynamics in the limit of an infinite population. Finite-population effects may be modeled by Monte Carlo simulations.

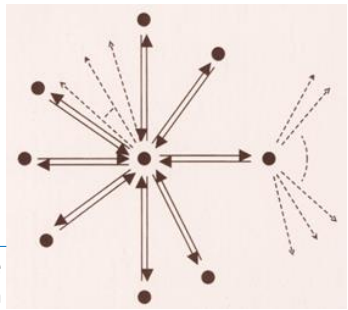
$$\frac{dP_i}{dt} = -\mu n P_i + \mu \sum_{j=1}^{A_i} P_{\nu_i(j)}$$

population of sequence *i*

# of viable neighbors of *i*

mutation rate

chain length



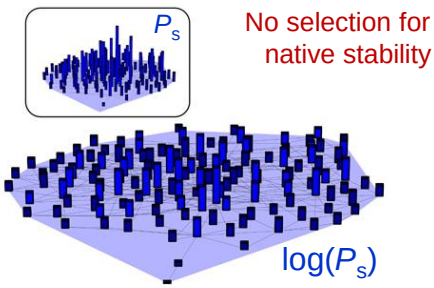
Mutational robustness tends to increase evolutionary population

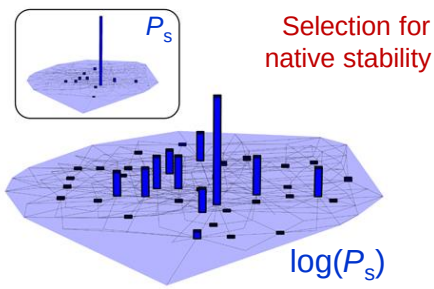
The steady-state population of a sequence is governed by the balance between gains from mutational influx from adjacent viable sequences (——►) and losses from viable mutations (——►) and lethal mutations (----------►) of the given sequence. Hence a sequence with more viable neighbors tend to accumulate a higher steady-state evolutionary population.

Bornberg-Bauer & Chan (1999) *PNAS* **96:**10689-10694;
Chan & Bornberg-Bauer (2002) *Applied Bioinformatics* **1:**121-144.

Interplay of network topology and fitness effects on evolutionary population

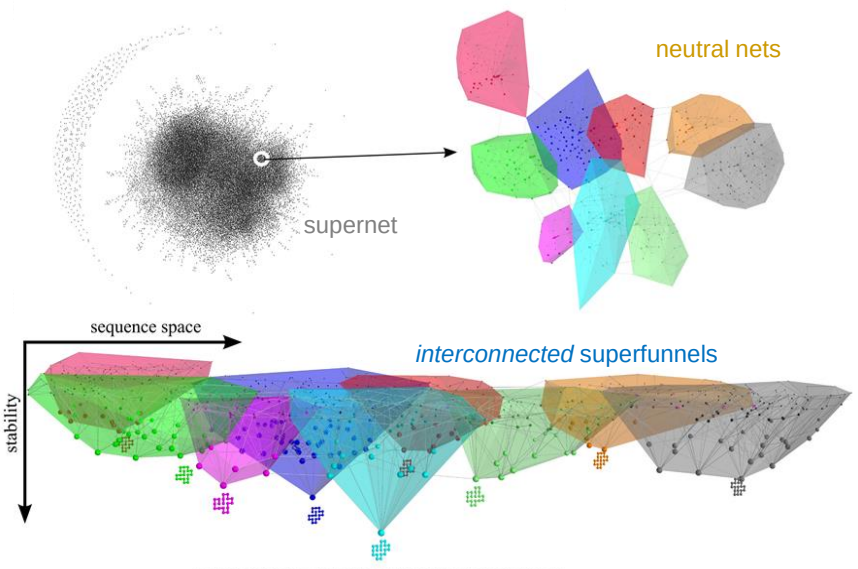
Mutational robustness alone may not be sufficient to account for the experimentally observed concentration of evolutionary populations at prototype-like sequences. Selection for native stability and kinetic stability is evident.





Sikosek & Chan, *J R Soc Interface* (2014)

Neutral Nets and Supernets in Sequence Space

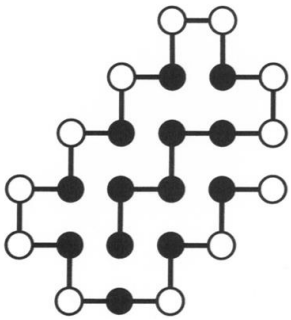


Sikosek & Chan, *J R Soc Interface* (2014)

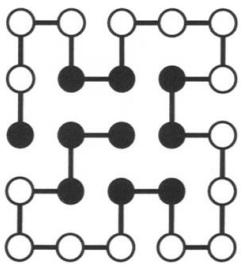
Other 2D lattice Protein Models

Models that restrict conformational variation to maximally compact conformations are computationally more tractable but are less physical

Example:

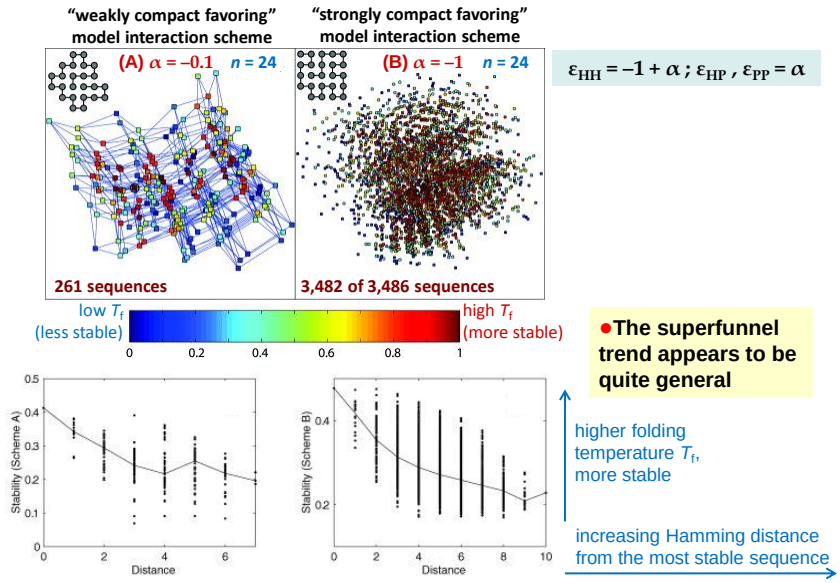


The most frequently encoded structure in the standard 2D HP model of chain length $n = 25$, shown by one of the 326 sequences that encode uniquely for it [Irbäck & Troein (2002) *J Biol Phys* **28:**1-15]. All 5,768,299,665 square-lattice $n = 25$ conformations are considered to arrive at this ground-state structure.



The most frequently encoded structure in the modified HP model of Xia and Levitt (2002) [**99:**10382-10387] that restricts conformational variation to the 1,081 maximally compact conformations confined to a 5×5 square, shown by one of the 67,615 sequences encoding uniquely for this structure.

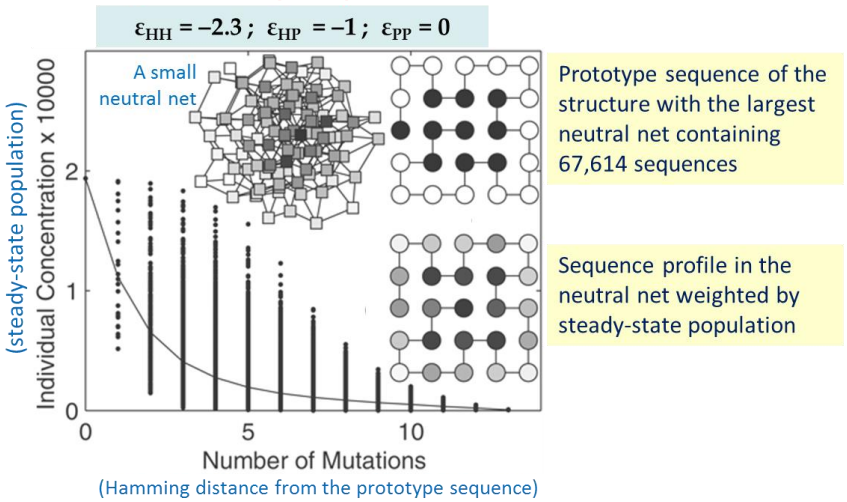
Size and Structure of Neutral Nets and Superfunnels Depend on Model Interaction Scheme



***Results and figures adapted from*:** Xia & Levitt (2004) Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* **55:**107-114.

Another 2D lattice model example of superfunnel

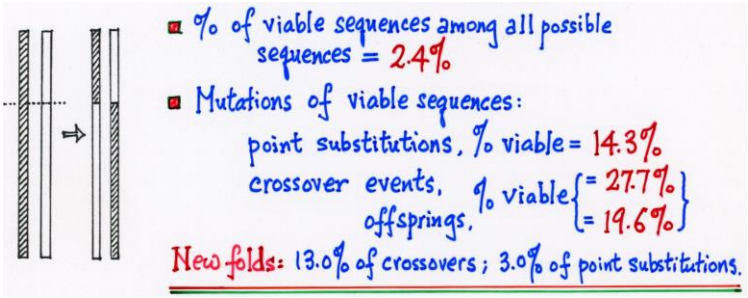
***n* = 25, ~ maximally compact 2D model**



***Results and figure from*:** Xia & Levitt (2002) Roles of mutation and recombination in the evolutionof protein thermodynamics. *PNAS* **99:**10382-10387.

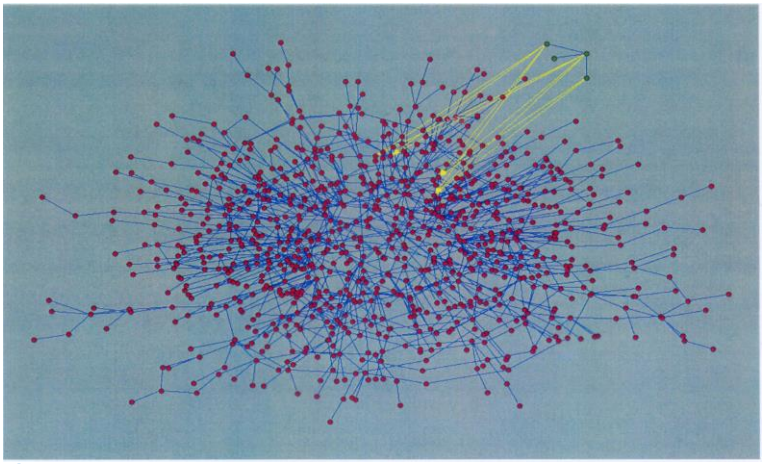
Recombinatoric exploration of folded structures

- more efficient than single-point mutations



***Results from*:** Cui, Wong, Bornberg-Bauer & Chan (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *PNAS* **99:**809-814.

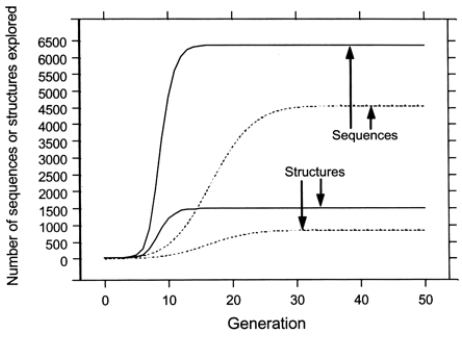
Sequences not reachable by single-point mutations are accessible by crossovers



● = neutral net; 897 neutral nets (n = 18; 4,553 sequences total) in the big net

Cui, Wong, Bornberg-Bauer & Chan, *PNAS* (2002)

Sequence and structure exploration facilitated by crossovers



The average number of distinct unique sequences visited or the average number of distinct structures these sequences encode, over 4,553 explorations, each using one of the model protein sequences in the largest net (supernet) as the starting point at generation 0. Exploration proceeds at any given generation to the next by including new unique sequences from (*i*) all possible single point mutations on all unique sequences visited thus far (dotted curves, point mutations only) or (*ii*) all possible crossovers between every pair of unique sequences visited thus far and those from (*i*) (solid curves, point mutations plus crossovers).

Discretized Master Equation of evolutionary population dynamics with point substitutions and crossovers

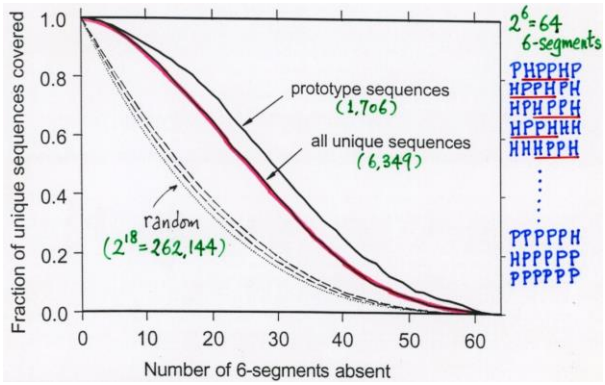
generation # — substitution & crossover rates — normalization — crossover connectivities — fitness of seq *i*

$$P_i(q+1) = \mathcal{N}(q)\left\{\left[-(\mu_m+\mu_c)P_i(q) + \frac{\mu_m}{n}\sum_{j=1}^{A_i} P_{\nu_i(j)}(q) + \frac{\mu_c}{n-3}\sum_{j<k}\sum_{s=1}^{2(n-3)} C(i|j,k;s)P_j(q)P_k(q)\right] + P_i(q)\right\} f_i,$$

***Results and figure from*:** Cui, Wong, Bornberg-Bauer & Chan (2002) *PNAS* **99:**809-814.

Local Signal for Sequence Uniqueness

- **Model proteins prefer certain local sequence patterns.**



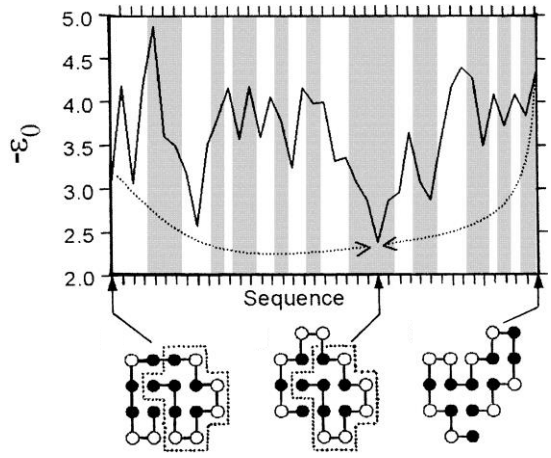
frequencies of 6-segments are sorted in different sets of n = 18 sequences

- Consistent with a subsequent experiment on β-lactamase showing that for a given number of amino acid substitutions, recombined variants are much more likely to retain function than variants generated by random point mutations [Drummond, Silberg, Meyer, Wilke & Arnold, *PNAS* (2005)].

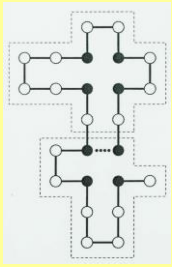
Cui, Wong, Bornberg-Bauer & Chan, *PNAS* (2002)

Tunneling and Autonomous Folding Units

Recombinatoric crossovers can help "tunnel" underneath rugged evolutionary landscape



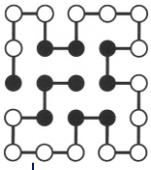
Autonomous folding units in the HP model



***Sequence from*:** Irbäck & Troein, *J Biol Phys* **28,** 1 (2002); ***Analysis and drawing from*:** Chan & Bornberg-Bauer, *Appl Bioinformatics* (2002)

Cui, Wong, Bornberg-Bauer & Chan, *PNAS* (2002)

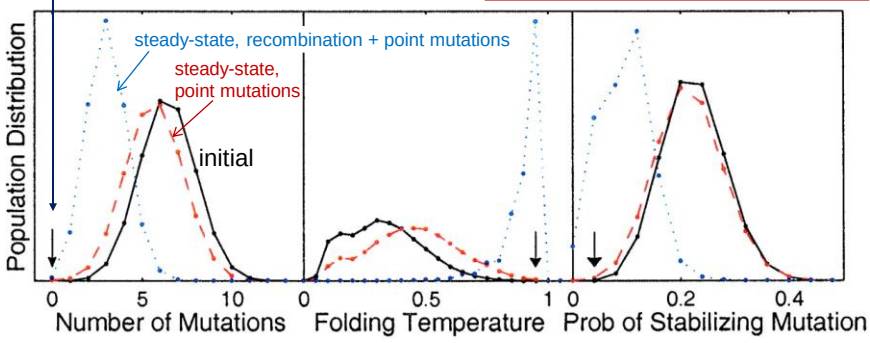
Recombinations are more effective than point substitutions in pushing evolutionary sequence population towards the prototype sequence



$\varepsilon_{HH} = -2.3$; $\varepsilon_{HP} = -1$, $\varepsilon_{PP} = 0$
~maximally compact model

prototype sequence

All sequences are equally populated initially. Population size N = 100,000 was sampled. Results in blue are for an simulated evolutionary process with recombination events dominating 1,000:1 over single-point mutation events.

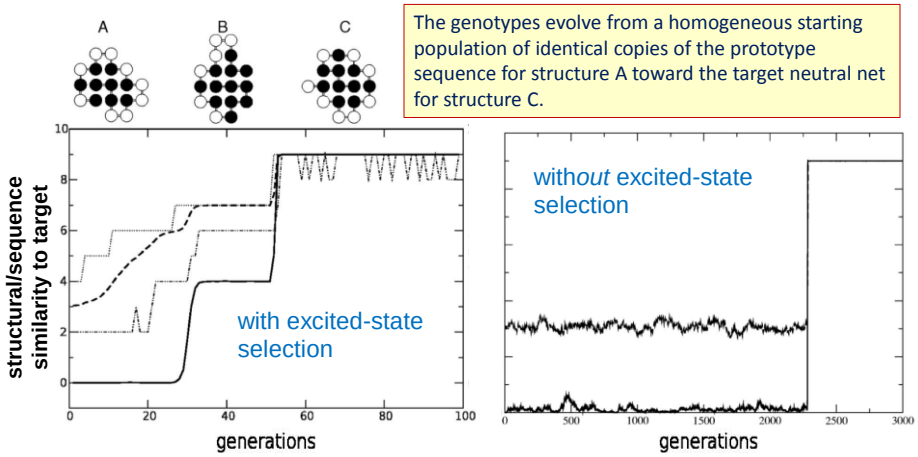


***Results and figure from*:** Xia & Levitt (2002) Roles of mutation and recombination in the evolutionof protein thermodynamics. *PNAS* **99:**10382-10387.

Latent Evolutionary Potentials:

Selection of excited-state (promiscuous) functions can speed up evolution dramatically

The genotypes evolve from a homogeneous starting population of identical copies of the prototype sequence for structure A toward the target neutral net for structure C.



***Results and figure from*:** Wroe, Chan & Bornberg-Bauer (2007) A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J* **1:**79-87.

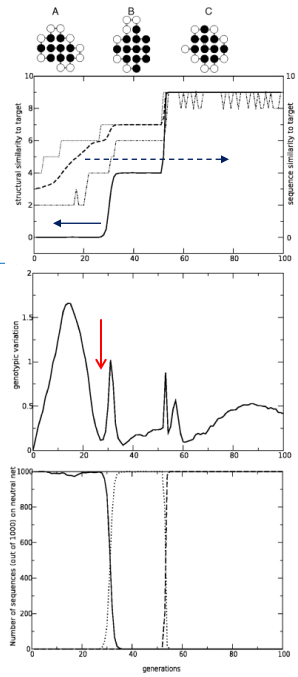
Model population dynamics with excited-state selections

- While the population drifts along network **A**, the changing sequence similarity indicates that it is getting closer to the target **C** but structurally the vast majority of the sequences still stay on the neutral network for structure **A**.

The upper and lower dotted curves give the maximum and minimum similarity of the sequences in the population to the target sequence.

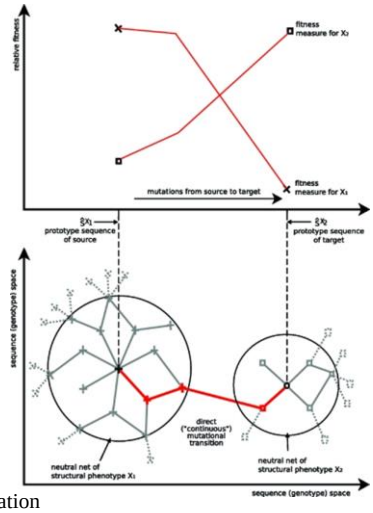
- Evolution of the average pairwise Hamming distance between all pairs of sequences in the population shows a reduction in genotypic variation immediately before the steps of adaptation, while at the same time structural distance from the originating structure increases.

Number of sequences (out of 1000), which code uniquely for structure **A** (solid curve), **B** (dotted), and **C** (dashed) as functions of generation.



***Results and figure from*:** Wroe, Chan & Bornberg-Bauer (2007) *HFSP J* **1:**79-87.

The simple 2D HP model provides a biophysical rationalization for the intriguing, and otherwise puzzling experimental observation that adaptation to new requirements can proceed while the "old," phenotypically dominant function is maintained along a series of seemingly neutral mutations.

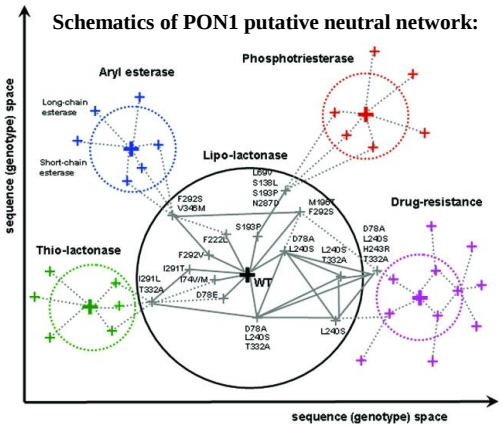


Evolutionary transitions between neutral nets. **Top:** possible development of the relative fitness during adaptation of a population which was initially optimal for one function (×, *left*) and then selected for another function (target: □, *right*). **Bottom:** an interpretation in terms of structural phenotypes according to Maynard-Smith's idea of a continuous phenotype space (there is one direct transition from one representative of one structural phenotype to the other) and the superfunnel paradigm (genotypes depicted in the center have more neutral neighbors and code for thermodynamically most stable structures).

***Results and figure from*:** Wroe, Chan & Bornberg-Bauer (2007) *HFSP J* **1:**79-87.

Tawfik and coworkers characterized ~300 variants of the enzyme PON1 that are apparently neutral, or close to neutral, with respect to PON1's levels of expression and native lactonase activity. Their activities with promiscuous substrates and ligands indicated significant changes in adaptive potentials.

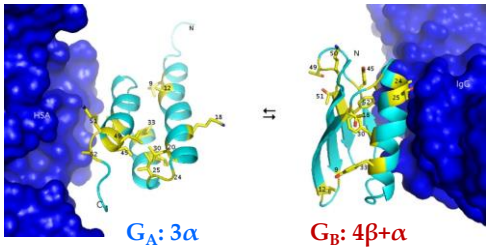
Schematics of PON1 putative neutral network:



***Results and figure from*:** Amitai, Gupta & Tawfik, *HFSP J* **1**, 67 (2007). See Tokuriki & Tawfik, *Science* **324**, 203 (2009) for a review.

"Inasmuch as the selection of nonnative functions and structures is operative, the attraction of any superfunnel towards its prototype may extend to proteins that are not yet part of its neutral network." – Sikosek et al, *PNAS* (2012)

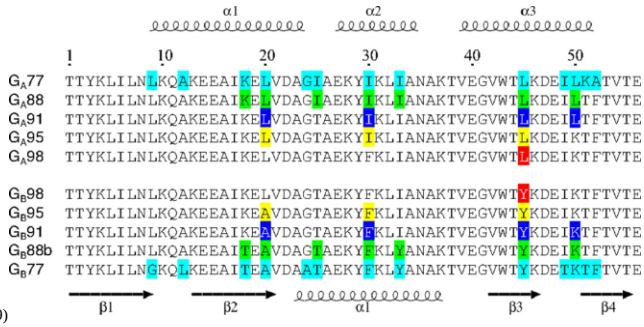
Conformational switches in real proteins



- Arc repressor homodimer: Double mutant L12↔N11: β → α; single mutant N11L ***bi-stable*** [Cordes *et al* & Sauer, *Science* **284**, 325 (1999); *Nature Struct Biol* **7**, 1129 (2000)].

- One-mutation switch between the human serum albumin-binding domain (G_A) and the IgG-binding domain (G_B) of *Streptococcus* protein G [Alexander, He, Chen, Orban & Bryan, *PNAS* **106**, 21149 (2009)]. NMR structures were determined for G_A95 & G_B95 (the 56aa sequences are 95% identical). Three substitutions in G_A95 or G_B95 shift the equilibrium from >99% in one structure to >99% in the other.

G_A, G_B sequence variants:



***Figures from*:** Alexander *et al* (2009)

The G_A-G_B conformational switch

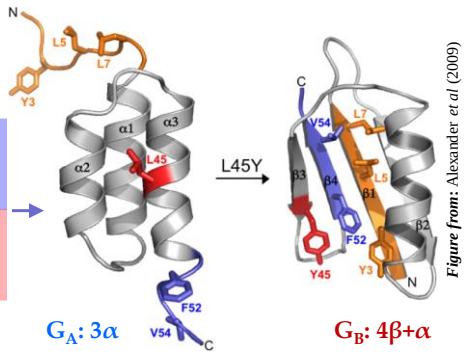
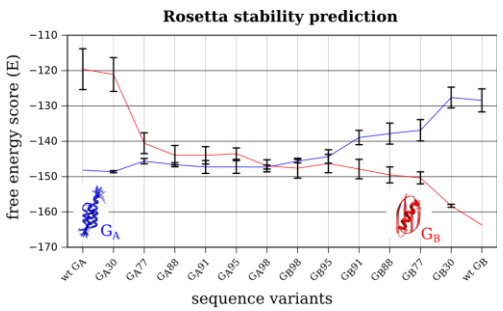


***Figure from*:** Alexander *et al* (2009)

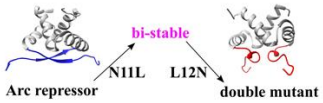
Model calculations we performed on the experimental G_A/G_B series of sequences indicate that the stability of the excited (latent) state gradually increases (i.e., with decreasing free energy) as the switch point is approached.

Sikosek, Bornberg-Bauer & Chan, *PLoS Comput Biol* (2012)

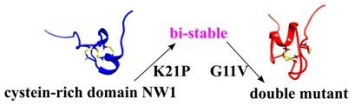


Experimentally designed bi-stable proteins and mutation-induced conformational switches

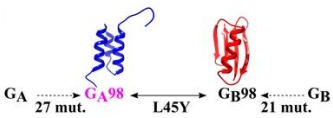
Bouvignies *et al.*, *Nature* (2011)



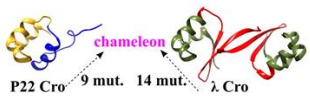
Cordes *et al.*, *Nature Struct Biol* (2000)



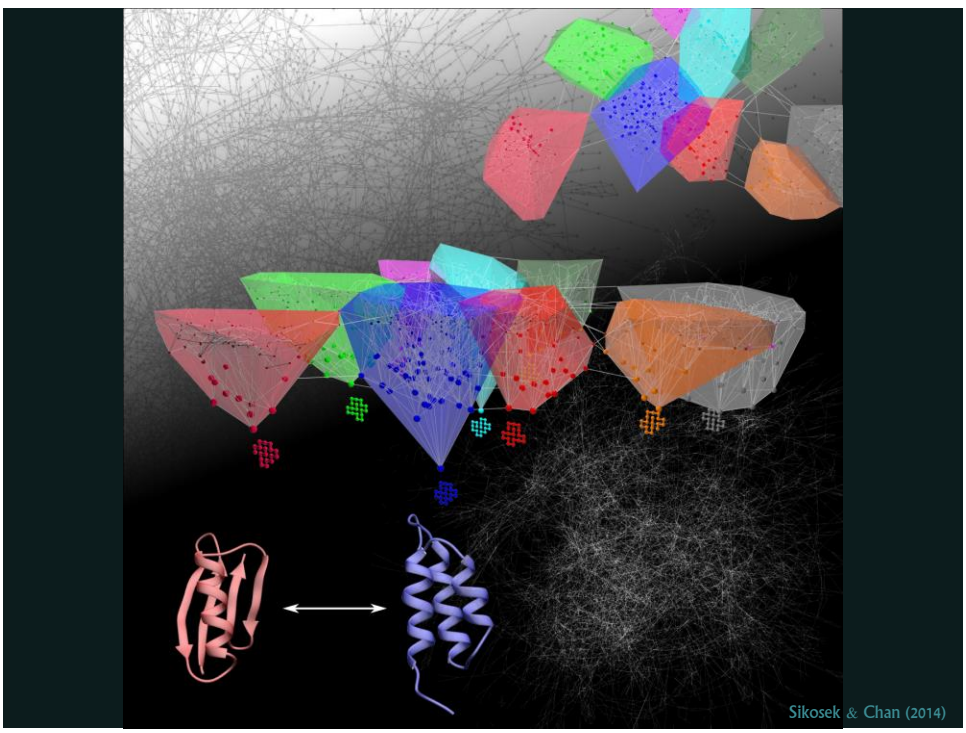
Meier *et al.*, *Curr Biol* (2007)



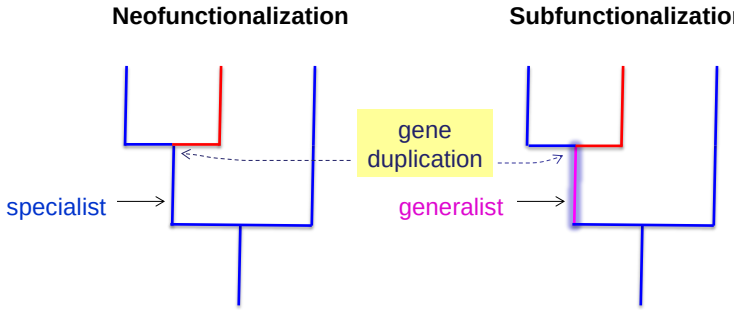
Alexander *et al.*, *PNAS* (2009)



Anderson *et al.*, *Protein Eng Des Sel* (2011)



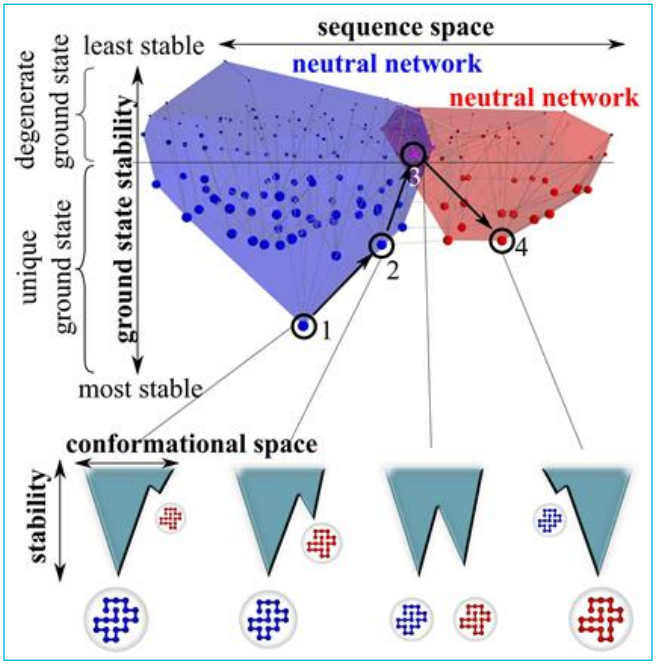
Role of excited-state selection in Escape from Adaptive Conflict



The "Escape from Adaptive Conflict" Perspective focuses on adaptation before gene duplication [Hittinger & Carrol, *Nature* **449**, 677 (2007); Des Marais & Rausher, *Nature* **454**, 762 (2008)]

Evolving a new folded structure:

traversing two superfunnels

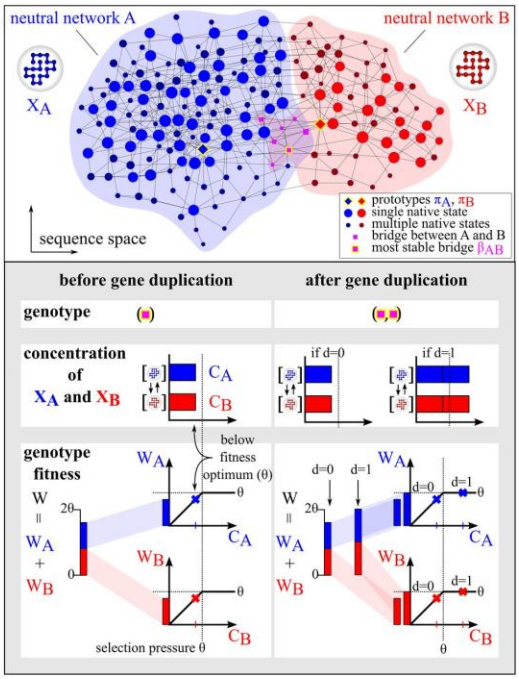


Sikosek *et al*., *PNAS* (2012); Sikosek *et al*., *PLoS Comput Biol* (2012); Sikosek & Chan, *J R Soc Interface* (2014)

Escape from Adaptive Conflict Follows from Weak Functional Trade-Offs and Mutational Robustness.

- Biophysics-based network connections.
- Evolutionary dynamics under mutations and gene duplications computed using both an analytical master equation and stochastic Monte Carlo simulations.
- Fitness is proportional to the stability (concentration) of the functional structures up to a certain optimum concentration above which fitness does not increase further with concentration.
- The optimum concentration corresponds to a measure of selection pressure.

Sikosek, Chan & Bornberg-Bauer (2012) *PNAS* **109:**14888-14893.



Master Equation of Evolutionary Population Dynamics with Single-Point Mutations and Gene Duplications

Now we need to consider two types of genotypes:
(i) those with a single gene and (ii) those with two genes.

(i) Genotypes with a single gene:

$$P_i(q+1) = \left[-(n\mu + \mu_d)P_i(q) + \mu \sum_{r=1}^{A_i} P_{\nu_i(r)}(q) + P_i(q)\right] \frac{\mathcal{N}(q) W_i}{\bar{W}(q)}$$

fitness of seq i

duplication rate

average fitness of generation q

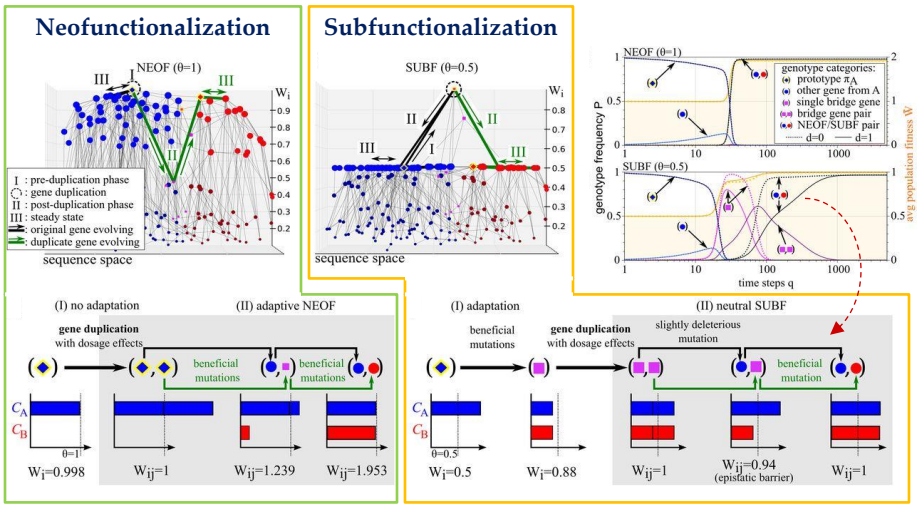
(ii) Genotypes with two genes:

$$P_{ij}(q+1) = \left[-2\mu n P_{ij}(q) + \mu \sum_{r=1}^{A_i} P_{\nu_i(r)j}(q) + \mu \sum_{s=1}^{A_j} P_{i\nu_j(s)}(q) + \mu^2 \sum_{r=1}^{A_i} \sum_{s=1}^{A_j} P_{\nu_i(r)\nu_j(s)}(q) + \delta_{ij}\mu_d P_i(q) + P_{ij}(q)\right] \frac{\mathcal{N}(q) W_{ij}}{\bar{W}(q)}$$

- **The formalism can also be applied to Monte Carlo simulations of finite evolutionary populations.**

Sikosek, Chan & Bornberg-Bauer (2012) Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *PNAS* **109:**14888-14893.

Neofunctionalization follows from strong selection pressures (strong trade-offs), whereas *Subfunctionalization* results from intermediate selection pressures (weak trade-offs)



Sikosek, Chan & Bornberg-Bauer (2012) *PNAS* **109:**14888-14893.

Escape from Adaptive Conflict (EAC) in the real world

Experiments showing that a reconstructed common ancestor of the fluorescent proteins in corals that emit either red or green light can emit light of both colors are indicative of EAC.

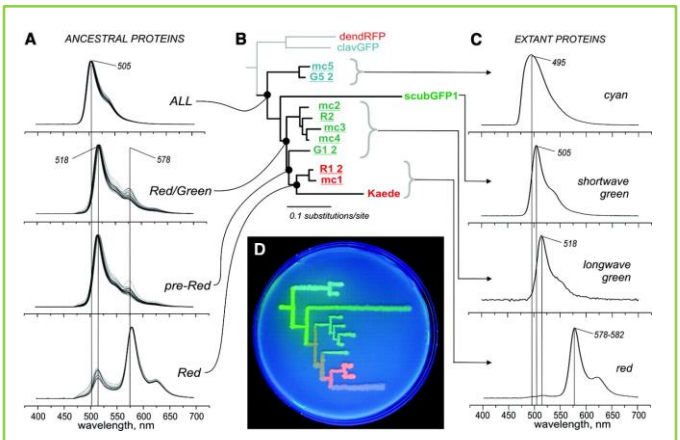
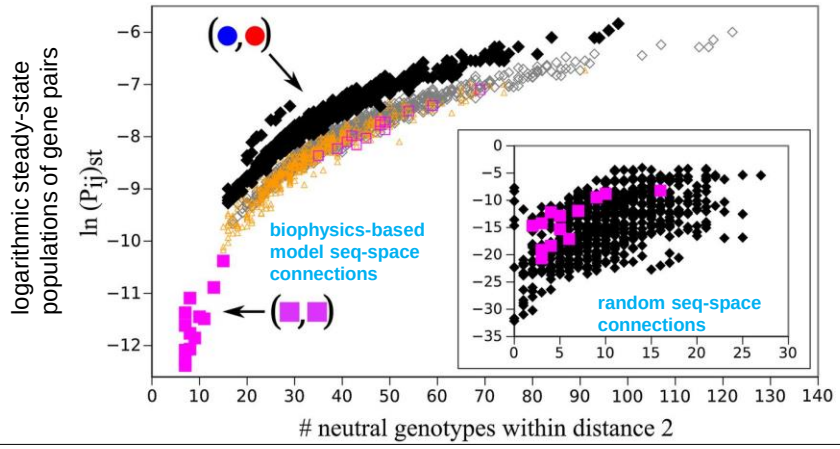


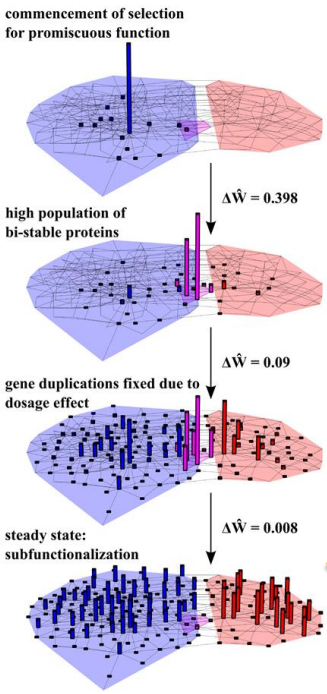
***Figure from*:** Ugalde, Chang & Matz, *Science* **305**, 1433 (2004)

Subfunctionalization (SUBF) as a Consequence of Neutral Network Topology

- A strong tendency toward mutationally robust genotypes (i.e., those with many sequence-space neighbors): a sequence-space "entropy" effect.
- Because of this "entropic" driving force, SUBF can be nonadaptive.

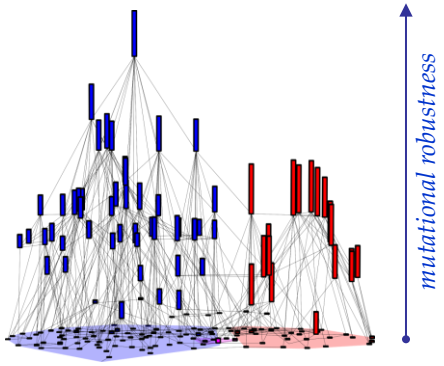


Sikosek, Chan & Bornberg-Bauer (2012) *PNAS* **109:**14888-14893.



Subfunctionalization can be driven solely by Mutational Robustness

generalists less robust than specialists



Sikosek & Chan, *J R Soc Interface* (2014)

Examples of 3D lattice protein models of evolution

- Maximally compact 27mers configured on the simple cubic lattice and restricted to a 3×3×3 cube, with a 20-letter alphabet

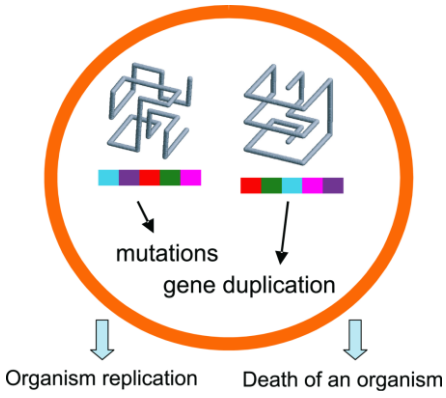


***Figure from*:** Zeldovich, Chen, Shakhnovich & Shakhnovich (2007) A first-principles model of early evolution: Emergence of gene families, species, and preferred protein folds. *PLoS Comput Biol* **3:**e139.

- **Useful and versatile construct, but non-maximally compact conformations – including open unfolded conformations – are *not* considered for the sake of computational tractability. This limitation has a significant impact on the physicality of the energetics of these 3D lattice protein models as for the 2D lattice protein models.**

Number of 27mer conformations in a 3×3×3 cube = 103,346

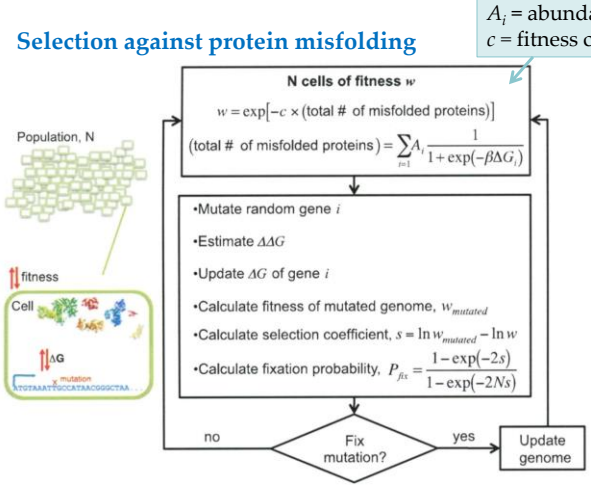
Total number of 27mer unrestricted conformations on the simple cubic lattice = 11,447,808,041,780,409

***Reference*:** Sikosek & Chan (2014) *J R Soc Interface* **11:**20140419

Possible biophysical basis for the anti-correlation between protein expression level and evolutionary rate

Selection against protein misfolding

A_i = abundance of gene i, $\beta \equiv 1/k_B T$, and c = fitness cost for each misfolded protein



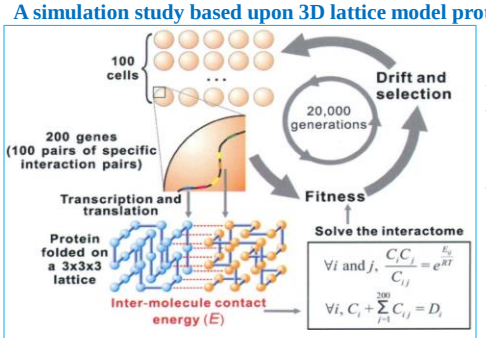
Assuming that proteins that tend to misfold more (those with less negative Δ*G*) are selected against, "multiscale" simulations suggest that the "purely biophysical" anti-correlation between pre-mutation Δ*G* and mutational ΔΔ*G* is a necessary requirement for the emergence of the experimental anti-correlation between protein expression level (abundance) and evolutionary rate.

Note: Lattice protein models not used in this simulation study.

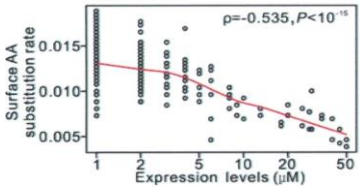
***Results and figure from*:** Serohijos, Rimas & Shakhnovich (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports* **2:**249-256.

Another probable biophysical constraint underpinning the anti-correlation between expression level and evolutionary rate is the need for a protein to avoid *misinteraction* with other proteins

A simulation study based upon 3D lattice model proteins. The authors* posit that the misfolding hypothesis does not fully explain the E-R (Expression-Rate) anti-correlation, especially for residues on protein surface. They propose that natural selection against protein–protein misinteraction, which wastes functional molecules and is potentially toxic, constrains the evolution of surface residues. Because highly expressed proteins are under stronger pressures to avoid misinteraction, surface residues are expected to show an E-R anti-correlation. Their findings indicate a pluralistic origin of the E-R anti-correlation.



- 100 expression levels were randomly generated and assigned to each pair of specific interacting partners.
- The native structure of a protein was fixed during the simulation of evolution.
- C_i is the concentration of free molecules of protein i, C_{ij} is the concentration of the protein complex composed of a protein i and a protein j, D_i is the total concentration of protein i in the cell.



***Results and figures from*:** *Yang, Liao, Zhuang & Zhang (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *PNAS* **109:**E831-E840.

Summary

- Biophysical sequence-to-structure mappings based upon simple explicit-chain protein models are versatile conceptual tools for addressing general principles of evolution.
- The superfunnel paradigm underscores a fundamental positive correlation between thermodynamic stability of the native structure and the mutational stability of a protein.
- Excited-state selection of promiscuous function can speed up evolution considerably.
- Subfunctionalization after duplication of a bi-stable gene with dual functions can be driven by sequence-space topology (i.e., mutational robustness) in an essentially nonadaptive manner.