Topic Course

# Gene and protein evolution

Winter 2016
Department of Molecular Genetics
University of Toronto

**Lecture 5**

Hue Sun Chan

## Synergy between the studies of protein biophysics and protein evolution

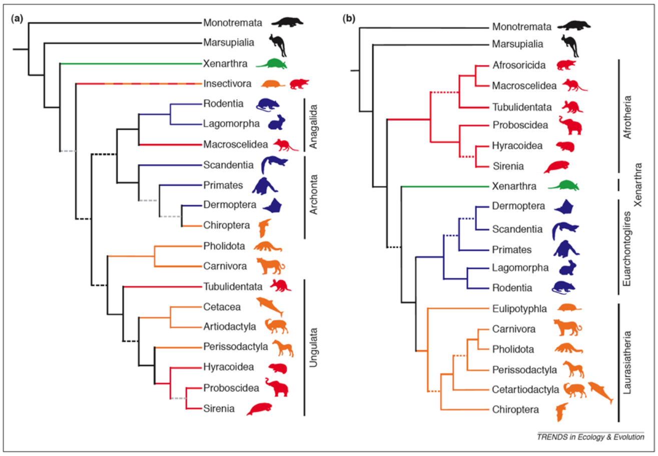
- Beyond pure sequence analysis:
  Constraints imposed by protein structure on molecular evolution
- Utilizing evolutionary information in the study of protein energetics
- Structural, thermodynamic, and dynamic consequences of mutations on folded proteins; and their evolutionary implications
- Biophysical and evolutionary constraints on protein thermodynamic and kinetic stability, "marginal" stability of natural proteins

**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).

## Substitution matrices are used to construct phylogenetic trees

A phylogeny, or evolutionary tree, is a branching diagram depicting an inferred evolutionary relationships among a set of organisms or groups of organisms (taxon; *plural*: taxa) based on their observed genetic similarities.

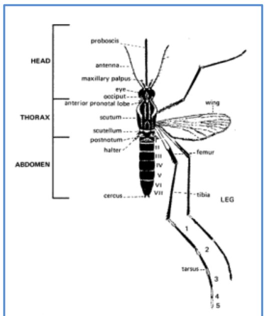
**An example:**



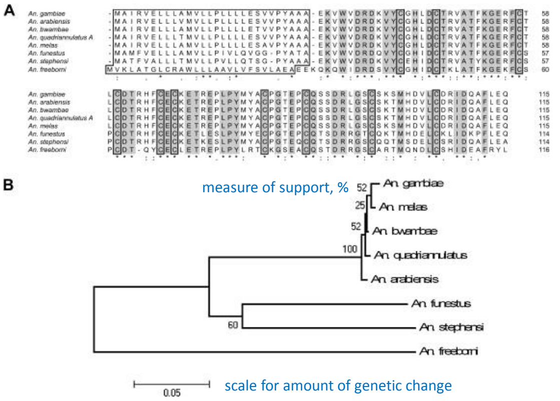
The left tree (a) is based on evidence from morphology, while the right tree (b) is based on evidence from genetics. The two trees are similar, but they are not identical. [Figure from: Springer, Stanhope, Madsen & de Jong (2004) Molecules consolidate the placental mammal tree. *Trends in Ecology and Evolution* **19:**430-438.]

**Substitution matrices are used to score putative mutations**

**Another example:** gSG6-P1 peptide is a potential specific salivary biomarker of exposure to *Anopheles* mosquitoes bites.



"There are approximately 3,500 species of mosquitoes grouped into 41 genera. **Human malaria is transmitted only by females of the genus *Anopheles*.** Of the approximately 430 *Anopheles* species, only 30-40 transmit malaria (i.e., are "vectors") in nature." (From CDC website)



```
An. gambiae            -MAIRVELLLAMVLLPLLLLESVVPYAAA-EKVWVDRDKVYCGHLDCTRVATFKGERFCT  58
An. arabiensis         -MAIRVELLLVMVLLPLLLLESVVPYAAA-EKVWVDRDKVYCGHLDCTRVATFKGERFCT  58
An. bwambae            -MAIRVELLLAMVLLPLLLLETVVPYAAA-EKVWVDRDKVYCGHLDCTRVATFKGERFCT  58
An. quadriannulatus A  -MAIRVELLLTMVLLPLLLLESVVPYAAA-EKVWVDRDKVYCGHLDCTRVATFKGERFCT  58
An. melas              -MAIRVELLLAMVLLPLLLLESAVPYAAA-EKVWVDRDKVYCGHLDCTRVATFKGERFCT  58
An. funestus           -MAMFVELLLAMVLLPIVLQVGG-PYATA-EKVWVDRDQVYCGHIDCTRVATFKGERFCS  57
An. stephensi          -MATFVALLLTMVLLPVLLQTSG-PYAAA-EKVWVDRDKVYCEHIDCTRVATYKGERFCS  57
An. freeborni          MVKLATGLCRAWLLLAAVLVFSVLAEAEEKQKQWIDRDSVYCGHIDCTKLATFKGEKFCS  60

An. gambiae            LCDTRHFCECKETREPLPYMYACPGTEPCQSSDRLGSCSKSMHDVLCDRIDQAFLEQ  115
An. arabiensis         LCDTRHFCECKETREPLPYMYACPGTEPCQSSDRLGSCSKTMHDVLCDRIDQAFLEQ  115
An. bwambae            LCDTRHFCECKETREPLPYMYACPGTEPCQSSDRLGSCSKTMHDVLCDRIDQAFLEQ  115
An. quadriannulatus A  LCDTRHFCECKETREPLPYMYACPGTEPCQSSDRLGSCSKTMHDVLCDRIDQAFLEQ  115
An. melas              LCDTRHFCECKETREPLPYMYACPGTEPCQSSDRLGSCSKSMHDVLCDRIDQAFLEQ  115
An. funestus           PCDTRHFCECKETLESLPYMYECPGTEPCQTSDRRGTCQKTMHDELCKLIDKPFLEQ  114
An. stephensi          PCDTRHFCECKETKESLPYMYACPGTEQCQTSDRRGSCQQTMSDELCSRIDQAFLEA  114
An. freeborni          PCDT-QYCECLETREPLPYLRTCKGSEACQSTDRRGSCARTMQNDLCSHIDEAFRYL  116
```

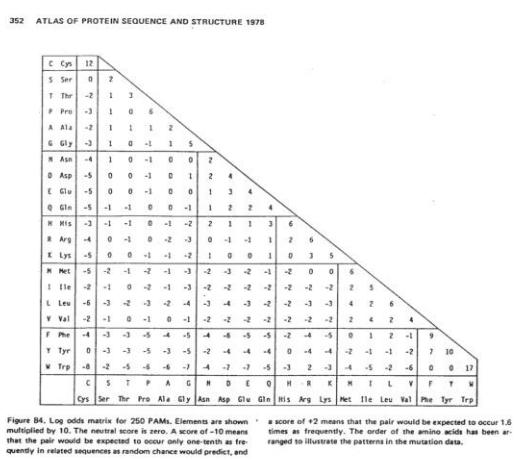
measure of support, %

scale for amount of genetic change

Sequences of the *Anopheline* gSG6 proteins . (**A**) Clustal alignment of *Anopheline* gSG6 proteins. Signal peptides and conserved Cysteines are boxed. Conserved sites are shaded. (**B**) Phylogenetic tree constructed from the alignment of the nucleotide sequence encoding the mature gSG6 polypeptides. [From: Drame et al. (2013) In: *Anopheles Mosquitoes - New Insights into Malaria Vectors*. S. Manguin ed., Chapter 23. DOI: 10.5772/55613.]

## Traditional substitution matrices (e.g. PAM and BLOSUM) were derived with sequence information without structural input. As a consequence, they may not encapsulate some of the important biophysical implications of mutations.

**PAM** (Point Accepted Mutation) matrices are a series of 20×20 substitution matrices each representing a % level of sequence divergence. E.g., PAM1 is calibrated for 1 mutation per 100 amino acids; PAM250 = $(\text{PAM1})^{250}$. The PAM matrices are based on data from closely related protein sequences (alignments of 71 sequence families each sharing at least 85% identity) and were derived from the frequencies of various mutations in presumed phylogenetic trees constructed using maximum parsimony for these families. The resulting log-odds (**Lod**) matrices are the logarithmic ratios of the observed mutation frequency divided by the probability of substitution expected if the mutations were random (target *minus* background; > 1 means the mutation is more likely than random). The log-odds matrices can then be used to score sequence alignments in general.



352 ATLAS OF PROTEIN SEQUENCE AND STRUCTURE 1978

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C Cys | 12 | | | | | | | | | | | | | | | | | | | |
| S Ser | 0 | 2 | | | | | | | | | | | | | | | | | | |
| T Thr | -2 | 1 | 3 | | | | | | | | | | | | | | | | | |
| P Pro | -3 | 1 | 0 | 6 | | | | | | | | | | | | | | | | |
| A Ala | -2 | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | |
| G Gly | -3 | 1 | 0 | -1 | 1 | 5 | | | | | | | | | | | | | | |
| N Asn | -4 | 1 | 0 | -1 | 0 | 0 | 2 | | | | | | | | | | | | | |
| D Asp | -5 | 0 | 0 | -1 | 0 | 1 | 2 | 4 | | | | | | | | | | | | |
| E Glu | -5 | 0 | 0 | -1 | 0 | 0 | 1 | 3 | 4 | | | | | | | | | | | |
| Q Gln | -5 | -1 | -1 | 0 | 0 | -1 | 1 | 2 | 2 | 4 | | | | | | | | | | |
| H His | -3 | -1 | -1 | 0 | -1 | -2 | 2 | 1 | 1 | 3 | 6 | | | | | | | | | |
| R Arg | -4 | 0 | -1 | 0 | -2 | -3 | 0 | -1 | -1 | 1 | 2 | 6 | | | | | | | | |
| K Lys | -5 | 0 | 0 | -1 | -1 | -2 | 1 | 0 | 0 | 1 | 0 | 3 | 5 | | | | | | | |
| M Met | -5 | -2 | -1 | -2 | -1 | -3 | -2 | -3 | -2 | -1 | -2 | 0 | 0 | 6 | | | | | | |
| I Ile | -2 | -1 | 0 | -2 | -1 | -3 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 2 | 5 | | | | | |
| L Leu | -6 | -3 | -2 | -3 | -2 | -4 | -3 | -4 | -3 | -2 | -2 | -3 | -3 | 4 | 2 | 6 | | | | |
| V Val | -2 | -1 | 0 | -1 | 0 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 2 | 4 | 2 | 4 | | | |
| F Phe | -4 | -3 | -3 | -5 | -4 | -5 | -4 | -6 | -5 | -5 | -2 | -4 | -5 | 0 | 1 | 2 | -1 | 9 | | |
| Y Tyr | 0 | -3 | -3 | -5 | -3 | -5 | -2 | -4 | -4 | -4 | 0 | -4 | -4 | -2 | -1 | -1 | -2 | 7 | 10 | |
| W Trp | -8 | -2 | -5 | -6 | -6 | -7 | -4 | -7 | -7 | -5 | -3 | 2 | -3 | -4 | -5 | -2 | -6 | 0 | 0 | 17 |
| | C<br>Cys | S<br>Ser | T<br>Thr | P<br>Pro | A<br>Ala | G<br>Gly | N<br>Asn | D<br>Asp | E<br>Glu | Q<br>Gln | H<br>His | R<br>Arg | K<br>Lys | M<br>Met | I<br>Ile | L<br>Leu | V<br>Val | F<br>Phe | Y<br>Tyr | W<br>Trp |

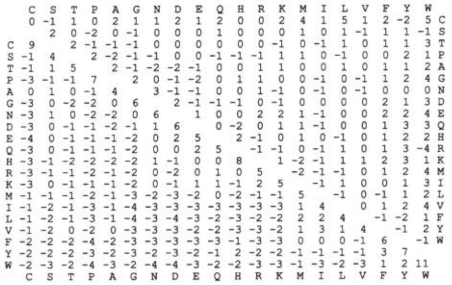
Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of −10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

**From:** Dayhoff et al. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, Maryland.

**BLOSUM** (**BLO**cks **SU**bstitution **M**atrix) matrices are based on the amino acid statistics observed in blocks of *ungapped* aligned regions of protein families (i.e., a blocks data base). Unlike the PAM matrices, the entries of the BLOSUM matrices were derived from the observed pairings of amino acids in the blocks *without regard to any presumed evolutionary process*. To calibrate the matrices for application to align sequences with different expected degrees of diversity, sequences with a given % identity or higher within a block are clustered to adjust (reduce) their contributions to the pairing statistics. E.g., the BLOSUM62 matrix clusters all sequences with 62% identity or higher together such that their total contribution to the matrix is equivalent to one sequence that has less than 62% identity with any other sequence in the block.

● All BLOSUM matrices are based on actual alignments. Unlike the PAMs, they are not extrapolated from restricted data derived from closely related proteins.

● The log-odds BLOSUMs and PAMs are similar if they have similar relative entropies (measure how much the entries are different from background; see example). But there are consistent differences: “BLOSUM 62 is less tolerant to substitutions involving hydrophilic amino acids, while it is more tolerant to substitutions involving hydrophobic amino acids.”



BLOSUM62

BLOSUM62 – PAM160

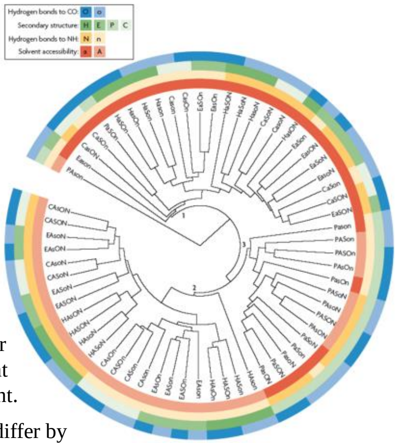
| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -1 | 1 | 0 | 2 | 1 | 1 | 2 | 1 | 2 | 0 | 0 | 2 | 4 | 1 | 5 | 1 | 2 | -2 | 5 | C |
| | | 2 | 0 | -2 | 0 | -1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | -1 | 1 | 1 | -1 | S |
| C | 9 | | 2 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | -1 | 1 | 0 | 1 | 1 | 3 | T |
| S | -1 | 4 | | 2 | -2 | -1 | -1 | 0 | 0 | -1 | -1 | -1 | 1 | 1 | 0 | -1 | 0 | 0 | 2 | 1 | P |
| T | -1 | 1 | 5 | | 2 | -1 | -2 | -2 | -1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | A |
| P | -3 | -1 | -1 | 7 | | 2 | 0 | -1 | -2 | 0 | 1 | 1 | 0 | 0 | -1 | 0 | -1 | 1 | 2 | 4 | G |
| A | 0 | 1 | 0 | -1 | 4 | | 3 | -1 | -1 | 0 | 0 | 1 | -1 | 0 | -1 | 0 | -1 | 0 | 0 | 0 | N |
| G | -3 | 0 | -2 | -2 | 0 | 6 | | 2 | -1 | -1 | -1 | 0 | -1 | 0 | 0 | 0 | 0 | 2 | 1 | 3 | D |
| N | -3 | 1 | 0 | -2 | -2 | 0 | 6 | | 1 | 0 | 0 | 2 | 2 | 1 | -1 | 0 | 0 | 2 | 2 | 4 | E |
| D | -3 | 0 | -1 | -1 | -2 | -1 | 1 | 6 | | 0 | -2 | 0 | 1 | 1 | -1 | 0 | 0 | 1 | 3 | 3 | Q |
| E | -4 | 0 | -1 | -1 | -1 | -2 | 0 | 2 | 5 | | 2 | -1 | 0 | 1 | 0 | -1 | 0 | 1 | 2 | 2 | H |
| Q | -3 | 0 | -1 | -1 | -1 | -2 | 0 | 0 | 2 | 5 | | -1 | -1 | 0 | -1 | 1 | 0 | 1 | 3 | -4 | R |
| H | -3 | -1 | -2 | -2 | -2 | -2 | 1 | -1 | 0 | 0 | 8 | | 1 | -2 | -1 | 1 | 1 | 2 | 3 | 1 | K |
| R | -3 | -1 | -1 | -2 | -1 | -2 | 0 | -2 | 0 | 1 | 0 | 5 | | -2 | -1 | -1 | 0 | 1 | 2 | 4 | M |
| K | -3 | 0 | -1 | -1 | -1 | -2 | 0 | -1 | 1 | 1 | -1 | 2 | 5 | | -1 | 1 | 0 | 0 | 1 | 3 | I |
| M | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -3 | -2 | 0 | -2 | -1 | -1 | 5 | | -1 | 0 | -1 | 1 | 2 | L |
| I | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | 1 | 4 | | 0 | 1 | 2 | 4 | V |
| L | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -4 | -3 | -2 | -3 | -2 | -2 | 2 | 2 | 4 | | -1 | -2 | 1 | F |
| V | -1 | -2 | 0 | -2 | 0 | -3 | -3 | -3 | -2 | -2 | -3 | -3 | -2 | 1 | 3 | 1 | 4 | | -1 | 2 | Y |
| F | -2 | -2 | -2 | -4 | -2 | -3 | -3 | -3 | -3 | -3 | -1 | -3 | -3 | 0 | 0 | 0 | -1 | 6 | | -1 | W |
| Y | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -2 | -1 | 2 | -2 | -2 | -1 | -1 | -1 | -1 | 3 | 7 | | |
| W | -2 | -3 | -2 | -4 | -3 | -2 | -4 | -4 | -3 | -2 | -2 | -3 | -3 | -1 | -3 | -2 | -3 | 1 | 2 | 11 | |
| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |

FIG. 2. BLOSUM 62 substitution matrix (*Lower*) and difference matrix (*Upper*) obtained by subtracting the PAM 160 matrix position by position. These matrices have identical relative entropies (0.70); the expected value of BLOSUM 62 is −0.52; that for PAM 160 is −0.57.

**From:** Henikoff & Henikoff (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89:**10915-10919.

## Mutations are constrained by biophysical factors, especially for essentially neutral mutations that leave a protein fold unchanged

“… the rate of amino acid substitution reflects both Darwinian selection for functionally advantageous mutations and selectively neutral evolution operating within the constraints of structure and function. During neutral evolution, whereby mutations accumulate by random drift, amino acid substitutions are constrained by factors such as the formation of intramolecular and intermolecular interactions and the accessibility to water or lipids surrounding the protein. These constraints arise from the need to conserve a specific architecture and to retain interactions that mediate functions in protein families and superfamilies.”



● Mutation rates depend on location in the protein structure → Different amino acid substitution probabilities for different local environments.

● Hierarchical clustering the 64 substitution matrices for 64 different local environments show clearly that solvent accessibility of the mutation site is the primary constraint.

● Three distinct clusters 1, 2, & 3 are observed: 1 & 2 differ by solvent accessibility, whereas cluster 3 is distinguished by a positive mainchain φ (can only be accommodated by Gly without energetic penalty). The next level of physical constraint is the presence of a sidechain hydrogen bond to the mainchain NH. …

***Results, quotes and figure taken from*:** Worth, Gong & Blundell (2009) Structural and functional constraints in the evolution of protein families. *Nature Reviews* **10:**709-720.

## Evolution rate *d*N/*d*S is sensitive to structural context of the mutated site and well correlated with relative solvent accessibility (RSA)

- Solvent accessible surface area (SASA) were normalized to the 99th percentile within each residue type to produce relative solvent accessibility (RSA).

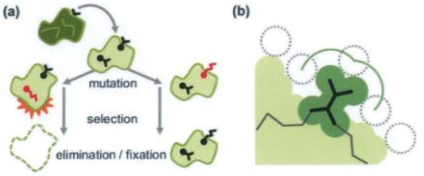


FIG. 1.—Evolutionary implications of the residue microenvironment. (*a*) A cartoon diagram of a protein is shown in cross section, highlighting two residues: one completely buried in the core and another partially exposed to solvent. Mutations occur at both sites, but whether or not they go to fixation depends on the properties of the residue microenvironments and their effects on the overall stability and function of the protein. (*b*) One quantitative property of the residue microenvironment is shown in detail. Here, a solvent molecule (dotted circle) traces the solvent accessible surface of a particular residue, shown in heavy wireframe.

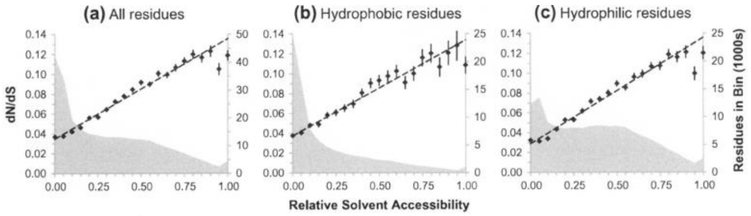


FIG. 2.—Correlating RSA and evolutionary rate. (*a*) Yeast codons are binned according to the RSA values of their associated residues; $d_N/d_S$ is then calculated. The distribution of residues across the bins is shown in the background (right vertical axis). (*b*) The trend and distribution from (*a*) restricted to hydrophobic residues. (*c*) The trend and distribution from (*a*) restricted to hydrophilic residues.

***Results and figures taken from***: Franzosa & Xia (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* **26:**2387-2395.

## RSA-dependent evolution model provides a significantly better fit to sequence data than does a traditional, RSA-independent model

- A “more accurate evolutionary null expectation”: A biophysics/structure-based neutral baseline beyond simply associating *d*N/*d*S > 1 with positive selection and *d*N/*d*S < 1 with negative selection.

- The method goes beyond previous approaches by finding sites of purifying selection that would have been missed and by including some sites of positive selection that would not have been identified.

- The method has been applied to influenza proteins hemagglutinin and neuraminidase. Hemagglutinin is an important immunity target. Thus the sialic acid-binding site is expected to be under positive selection, as it is a major target for host antibody binding.

- Relative to models that assume site-independent mutation rates, models that incorporate RSA dependence perform better according to the Akaike information criterion (AIC), which is a measure of the relative quality of statistical models. It rewards good fits and penalizes models with a large number of parameters.

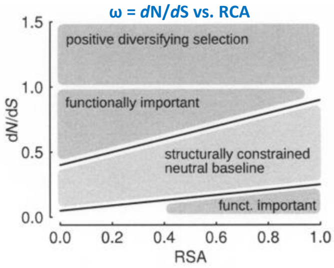

ω = *d*N/*d*S vs. RCA

FIG. 1. Regions of interest in $\omega$–RSA plot. Most sites in proteins fall into a trapezoidal region we consider the neutral baseline. Sites with $\omega > 1$ are generally considered to be under positive diversifying selection. In addition to such sites, our method can also identify sites with an $\omega < 1$ but either larger or smaller than expected given their RSA. These sites fall into the triangular regions below $\omega = 1$ that are either above or below the neutral baseline. Sites in these regions experience either an accelerated or a reduced rate of evolution relative to the baseline and are likely to be functionally important.

***Results and figure taken from***: Meyer & Wilke (2013) Integrating sequence variation and protein structure to identify sites under selection. *Mol Biol Evol* **30:**36-44.

(*cont'd*)

- Hemagglutinin forms a homotrimer on the surface of influenza virus, RSAs computed from this multimeric form provide a better evolutionary model than those computed using the monomeric form.

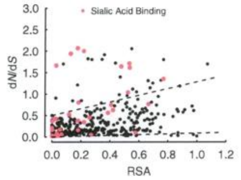


Fig. 4. Average $\omega$ versus RSA for hemagglutinin, obtained from the optimal model (three slopes and three intercepts). Dashed lines indicate the trapezoidally shaped neutral baseline (as ascertained by eye). Sites highlighted in red are within 8 Å of the sialic acid-binding region. Sites above the upper dashed line are significantly enriched in sites near the sialic acid-binding region (Fisher's exact test, OR = 6.6, $P = 6.1 \times 10^{-5}$).

- By incorporating RSA, highly solvent-exposed sites with high *d*N/*d*S that also have large RSA are inferred as not being positively selected. For hemagglutinin, the best-fitting model suggested that at least one site with elevated evolutionary rate should be considered part of the neutral baseline (fig. 4). In total, for hemagglutinin 33 sites were predicted to undergo accelerated evolution (above the upper dashed line in fig. 4) and 9 sites were predicted to be exceptionally conserved (below the lower dashed line in fig. 4).

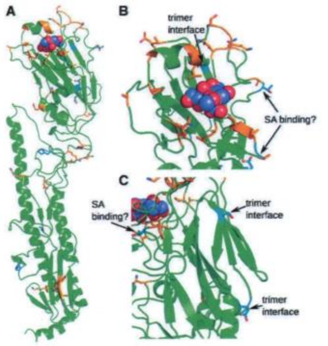
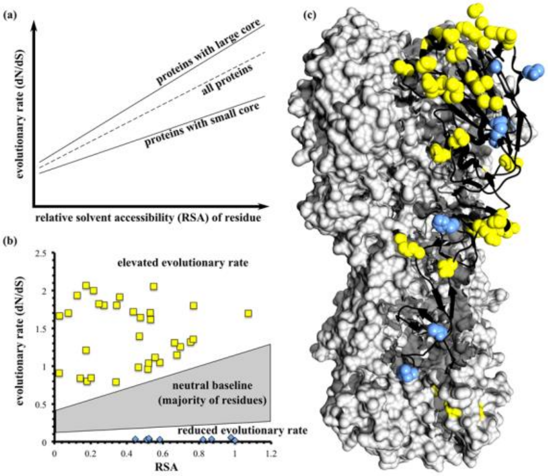


Fig. 5. Sites of interest identified for hemagglutinin. Sites that fall above the upper dashed line in fig. 4 are colored orange. Sites that fall below the lower dashed line in fig. 4 are colored light blue. The polypeptide backbone is colored green. Sialic acid is represented by the space-filling model near the top of the molecule. (A) View of the entire hemagglutinin monomer. (B) View of the sialic acid-binding region. Sites that are highlighted as "SA binding?" are unusually conserved and close to (though not within 8 Å of) the sialic acid. Sites that are highlighted as "trimer interface" are unusually conserved and seem to be important for trimerization. (C) View of the trimer-interface region. Labeling of sites is as in part (B).

***Results and figures taken from***: Meyer & Wilke (2013) Integrating sequence variation and protein structure to identify sites under selection. *Mol Biol Evol* **30:**36-44.

## Summary:

## Biophysical context-dependent interpretation of *d*N/*d*S enhances evolutionary models

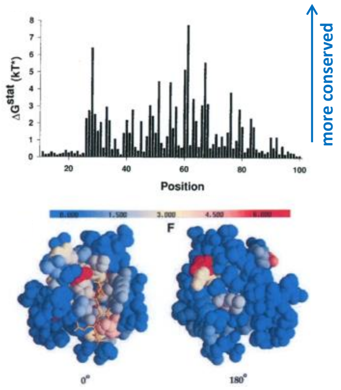


(a) RSA is strongly correlated with ω ≡ *d*N/*d*S (b) RSA improves the neutral baseline for the detection of positive and negative selection in influenza proteins. Yellow data points have elevated evolutionary rates but can be either below or above the conventional ω = 1 divide that typically distinguishes positive (ω > 1) and negative/purifying (ω < 1) selection. Blue data points show sites evolving at reduced evolutionary rates compared with the new neutral baseline. (c) The homo-trimeric haemagglutinin is an influenza surface glycoprotein. One of the three monomeric units is shown as black ribbons with the residues under positive or negative selection highlighted as spheres using the same colour code as that in (b). A majority of the positively selected sites are found around the region that is most frequently targeted by antibodies (top right) and are thus under strong selection pressure to diversify.

***Figure from***: Sikosek & Chan (2014) *J Royal Society Interface* **11:**20140419.

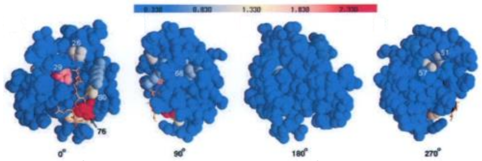
## Inferring residue-residue contacts in protein structures from evolutionary data

**Statistical analysis of the PDZ family**



The more conserved positions (colored white to red) in a structure-based multiple sequence alignment of 274 eukaryotic PDZ domains constitute a rough description of the protein interaction surface of the fold.

**Statistical coupling analysis (SCA)** is based on multiple-sequence alignment of a large number (an "evolutionarily well sampled" set) of sequences folding to essentially the same structure. Using such a dataset, variation in the relative abundance of amino acid types at a given position conditioned upon the amino acid type at another position may be ascertained. Coupling between residues that are spatially far apart and otherwise not obvious can be revealed by the *conservation weighted covariance matrix* for all pairs of positions in a protein constructed using SCA. Here is an example:

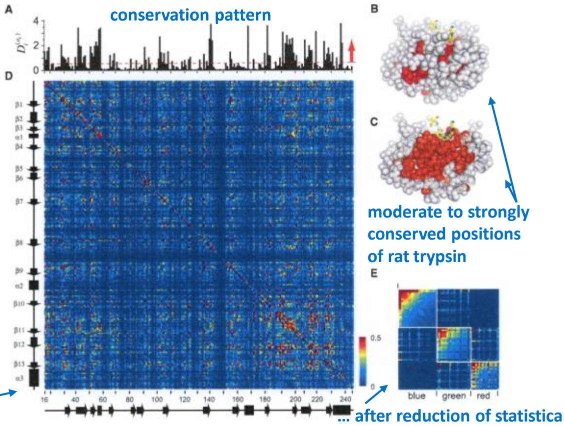


Mapping of the SCA data on the PDZ fold when position 76 is perturbed. Each view is successively rotated by 90° from the 1st. The coupled positions exhibit energetic interactions at sites both spatially close to and distant from the point of perturbation. A range of weak (blue) to strong (red) coupling is indicated by the color scale.

***Results and Figures from*:** Lockless & Ranganathan (1999) *Science* **286:**295-299.

## Existence of Independent Evolutionary Units (Sectors) in a Protein Structure is Suggested by Statistical Coupling Analysis (SCA)

Statistical analysis of correlated evolution between amino acid residues in serine proteases is suggestive of a structural partition of the protein into three quasi-independent groups of evolutionarily correlated amino acids term ''protein sectors.''



conservation pattern

moderate to strongly conserved positions of rat trypsin

SCA matrix for a sequence alignment of 1470 members of the protease family

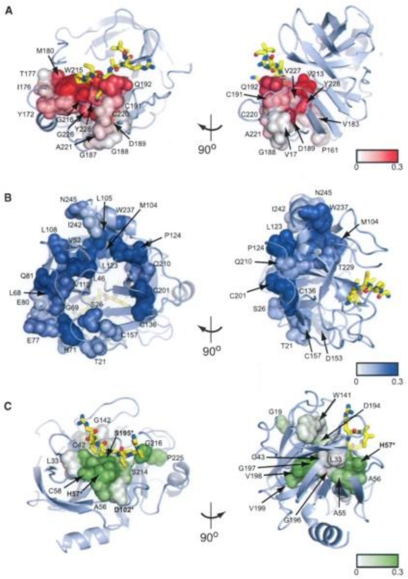
… after reduction of statistical noise and of global coherence

**Figure 1. Position-Specific and Correlated Conservation in the S1A Protease Family**
(A) The conservation of each position *i* in a multiple sequence alignment of 1470 members of the S1A family, computed by the relative entropy $D_i^{(a_i)}$ (position numbering according to bovine chymotrypsin, and graph is aligned with the matrix below).
(B and C) Mapping of the moderate to strongly conserved positions in a surface view (B) and a slice through the core (C) of rat trypsin shows a simple and intuitive arrangement. Residues with $D_i^{(a_i)} > 0.5$ (in orange) occupy the protein core and regions contacting substrate, while less conserved positions are mostly located on the surface. The cutoff is chosen to color ~50% of residues to illustrate the pattern of conservation in the protein structure.
(D) SCA matrix $\tilde{C}_{ij}$ for a sequence alignment of 1470 members of the protease family, showing a pattern of correlated conservation that is distributed throughout the primary structure and across secondary structure elements.
(E) SCA matrix after reduction of statistical noise and of global coherent correlations (see the Supplemental Experimental Procedures and Supplemental Discussion). The 65 positions that remain fall into three groups of positions (red, blue, and green, termed "sectors"), each displaying strong intragroup correlations and weak intergroup correlations. In each sector, positions are ordered by descending magnitude of contribution (Figure S3), showing that sector positions comprise a hierarchy of correlation strengths.

***Results and Figure from*:** Halabi, Rivoire, Leibler & Ranganathan (2009) *Cell* **138:**774-786.

## Protein Sectors

**Three sectors are identified in the S1A family of serine proteases**



### Characteristics of the sectors:

- Statistical independence

  demonstrated by SCA and further correlation entropies analysis.

- Structural connectivity

  "the physical connectivity of each sector is striking, given that no information about tertiary structure was used in their identification."

- Biochemical independence

  the main effects of alanine substitutions in the red and blue sectors are in catalytic activity (red) and thermal stability (blue).

- Independent sequence divergence

  suggesting that "no single measure of the divergence of protein sequences can correctly represent their differences in functional properties"

***Results and Figure from*****:** Halabi, Rivoire, Leibler & Ranganathan (2009) *Cell* **138:**774-786.

## Direct-Coupling Analysis (DCA) of Co-evolution of Amino Acid Residues

**Goal/Motivation:** Ascertain the correlation between amino acid occupancy at residue positions as an evolutionary predictor of spatial proximity of residues in folded proteins.

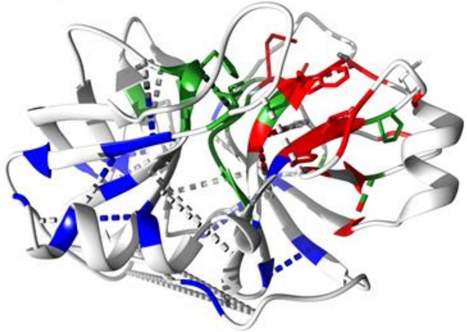
**Basic Hypothesis:** "If two residues of a protein or a pair of interacting proteins form a contact, a destabilizing amino acid substitution at one position is expected to be compensated by a substitution of the other position over the evolutionary timescale, in order for the residue pair to maintain attractive interaction."

**Relationship with SCA:** Covariance analysis identifies direct contacts *as well as secondary correlations* between non-interacting residues induced by the substitution patterns of directly interacting residues. Direct-coupling analysis (DCA) was formulated to disentangle direct from indirect correlations.

**The Main Mathematical Idea behind DCA Approaches:** Infer/construct a global statistical model of *direct interactions* for entire protein sequences under the requirement that the model generates the empirically observed amino acid frequency counts (which include both direct and indirect coupling). A "best" model is chosen among many possibilities by applying additional constraints such as maximum informational entropy.

***References*****:** Weigt, White, Szurmant, Loch & Hwa (2009) *PNAS* **106:**67-72; Morcos, Pagnani, Lunt, Bertolino, Marks, Sander, Zecchina, Onuchic, Hwa & Weigt (2011) *PNAS* **108:**E1293-E1301.

## Protein sectors determined by statistical analyses of the correlation between mutational changes at different positions in a protein: SCA and DCA



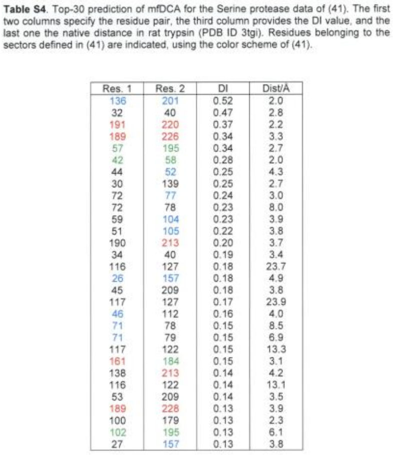


**Table S4**. Top-30 prediction of mfDCA for the Serine protease data of (41). The first two columns specify the residue pair, the third column provides the DI value, and the last one the native distance in rat trypsin (PDB ID 3tgi). Residues belonging to the sectors defined in (41) are indicated, using the color scheme of (41).

| Res. 1 | Res. 2 | DI | Dist/Å |
|---|---|---|---|
| 136 | 201 | 0.52 | 2.0 |
| 32 | 40 | 0.47 | 2.8 |
| 191 | 220 | 0.37 | 2.2 |
| 189 | 226 | 0.34 | 3.3 |
| 57 | 195 | 0.34 | 2.7 |
| 42 | 58 | 0.28 | 2.0 |
| 44 | 52 | 0.25 | 4.3 |
| 30 | 139 | 0.25 | 2.7 |
| 72 | 77 | 0.24 | 3.0 |
| 72 | 78 | 0.23 | 8.0 |
| 59 | 104 | 0.23 | 3.9 |
| 51 | 105 | 0.22 | 3.8 |
| 190 | 213 | 0.20 | 3.7 |
| 34 | 40 | 0.19 | 3.4 |
| 116 | 127 | 0.18 | 23.7 |
| 26 | 157 | 0.18 | 4.9 |
| 45 | 209 | 0.18 | 3.8 |
| 117 | 127 | 0.17 | 23.9 |
| 46 | 112 | 0.16 | 4.0 |
| 71 | 78 | 0.15 | 8.5 |
| 71 | 79 | 0.15 | 6.9 |
| 117 | 122 | 0.15 | 13.3 |
| 161 | 184 | 0.15 | 3.1 |
| 138 | 213 | 0.14 | 4.2 |
| 116 | 122 | 0.14 | 13.1 |
| 53 | 209 | 0.14 | 3.5 |
| 189 | 228 | 0.13 | 3.9 |
| 100 | 179 | 0.13 | 2.3 |
| 102 | 195 | 0.13 | 6.1 |
| 27 | 157 | 0.13 | 3.8 |

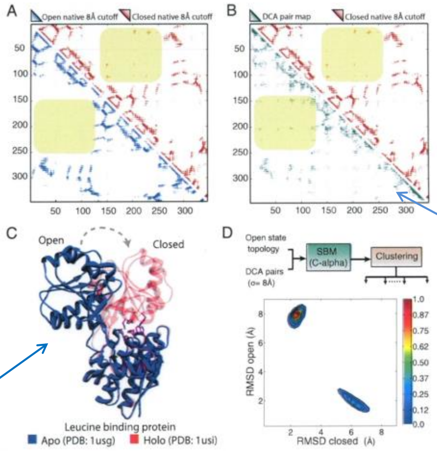
Co-evolving residues in rat trypsin (3TGI), a serine protease. Protein sectors are networks of co-evolving residues with independent functions. Here, the three sectors of this class of serine proteases determined by Halabi et al. (*Cell* 2009) are shown in red (substrate specificity), blue (thermal stability) and green (catalysis). Known functional residues are shown as sticks. These sectors were identified using the statistical coupling analysis (SCA) approach, whereas a different approach, direct coupling analysis (DCA), yielded a partially different set of co-evolving residues (dashed lines, see Table on the right from Morcos et al. *PNAS* 2011).

***Figure from*****:** Sikosek & Chan (2014) *J Royal Society Interface* **11:**20140419. ***Table from*****:** Morcos et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS* **108:**E1293-E1301.

## Utilizing Evolutionary Information in the Biophysical Study of Protein Structure and Dynamics

Aligned sequences contain biophysical information. Combination of contact preference from direct coupling analysis (DCA) and structure-based (native-centric, or Gō-like) models (SBMs) can predict intermediates or hidden states that are functionally important.

**Example:** Leucine binding protein. (SBM for the open state + DCA info) is sufficient to predict the closed state.



DCA prediction

**Fig. 1.** A hybrid SBM+DCA model of the L-leucine binding protein is able to uncover its two-state (apo/holo) conformational landscape. *A* compares the native open and closed contact maps and *B* compares a DCA contact map with the native closed state. In *A*, comparing the native contact map of the open conformation (PDB ID 1usg; lower triangular map) and the closed conformation (PDB ID 1usi; upper triangular map) shows a clear set of contacts (shaded box) that are exclusive to the closed state. In *B* a predicted contact map using highly ranked DCA residue pairs (lower triangular map) shows a very accurate reconstruction of the complete map that includes the extra contacts in the closed conformation (upper triangular map). (*C*) Structural comparison between the apo and the holo states of the L-leucine binding protein, showing domain closure. (*D*) Integrating a SBM of the open-state topology with DCA contacts produces a distinct bimodal landscape, as opposed to the single-basin distribution observed when we use the same number of extra contacts but randomly distributed.

***Results and Figure from*****:** Morcos, Jana, Hwa & Onuchic (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *PNAS* **110:**20533-20538.

## Biophysical consequences of protein mutations

**Possible effects of mutations on protein folding and interaction.** The folding landscape of a globular protein shows the correctly folded structure as the global free energy minimum, whereas a shallower minimum is a misfolded structure. Interactions of the original protein are indicated by black arrows; those of the mutant are indicated in red. Mutations can lead to misfolding and/or aggregation and/or misinteractions. Mutations can also lead to no apparent changes (neutral mutation). Some non-neutral mutations can lead to new functional interactions that would then be subject to evolutionary selection. Note that the depiction of interactions between folded proteins as a 'lock and key' fit between specific shapes is adopted here merely to simplify the schematic representation. The perspective conveyed by the present figure does not preclude more dynamic binding mechanisms such as induced fit and conformational selection.

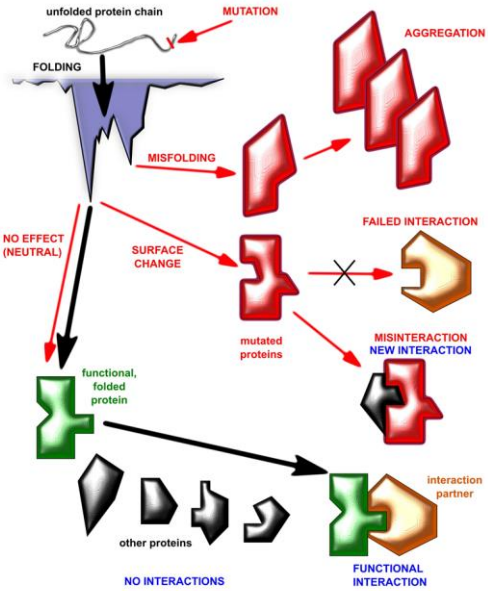


***Figure from*:** Sikosek & Chan (2014) *J Royal Society Interface* **11:**20140419.

## Thermodynamic and Kinetic Stability

In protein evolution studies, stability is often used as a proxy for the fidelity of a protein function, because a sufficient stability of the native state is often required for function. Although a protein's function is not equivalent to its stability, experimental support exists for a positive correlation between protein functionality and native stability.

### Most mutations on extant proteins are thermodynamically destabilizing

Among 290 single-residue substitutions of staphylococcal nuclease created artificially by Shortle and co-workers, 257 are destabilizing, five lead to stabilities essentially the same as that of the wild-type (approx. 5.5 kcal mol$^{-1}$), only 28 are stabilizing. Moreover, each destabilizing artificial mutation destabilizes by more than 2.08 kcal mol$^{-1}$ on average (maximum = 7.5 kcal mol$^{-1}$), whereas each stabilizing artificial mutation stabilizes by only 0.36 kcal mol$^{-1}$ on average (maximum = 1.0 kcal mol$^{-1}$).

***References*:** Shortle et al. (1990) *Biochemistry* **29:**8033-8041; Green et al. (1992) *Biochemistry* **31:**5717-5728; Meeker et al. (1996) *Biochemistry* **35:**6443-6449.

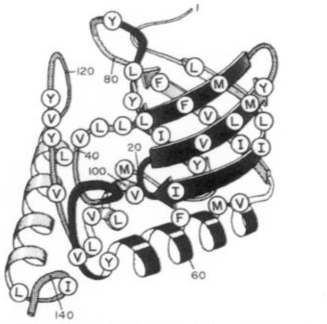


FIGURE 1: Ribbon diagram showing the positions of the large hydrophobic residues altered in mutant forms of staphylococcal nuclease. The approximate position of the $C_\alpha$ carbon is labeled by using the one-letter code. Numbers designate every 20th residue. This figure was drawn and copyrighted by Jane Richardson (1981) and is used with her permission.

(Figure from Shortle et al. 1990)

## Most mutations on extant proteins are thermodynamically destabilizing

… A similar trend is exhibited by the 98 artificial mutants of chymotrypsin inhibitor 2 studied by Fersht and co-workers (77 with a single substitution, 17 with two substitutions, and four with three substitutions): 90 artificial mutants are less stable than the wild-type (7.6 kcal mol$^{-1}$), only eight artificial mutants are more stable than the wild-type. On average, a destabilizing mutation destabilizes by 1.67 kcal mol$^{-1}$ (maximum = 4.93 kcal mol$^{-1}$ among single-substitution mutants), whereas a stabilizing mutation stabilizes by only 0.18 kcal mol$^{-1}$ (maximum = 0.42 kcal mol$^{-1}$). *These data suggest that the stabilities of natural proteins are close to, albeit not exactly at, the maximum achievable by sequences in the immediate sequence-space neighbourhood of the wild-type sequence.*

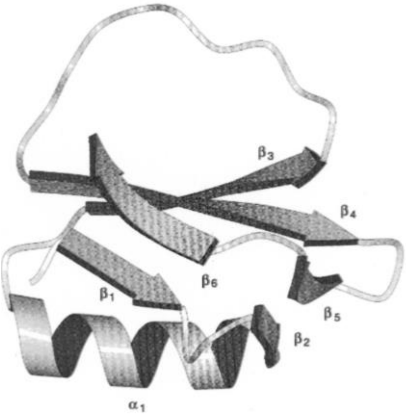


**Figure 2.** Schematic representation of the structure of CI2. The diagram was produced using the program MolScript (Kraulis, 1991).

***Figure and Results from*****:** Itzhaki et al. (1995) *J Mol Biol* **254:**260-288.

## Most mutations on extant proteins are thermodynamically destabilizing

**However, when larger numbers of amino acid substitutions are applied to a wild-type, an increase in thermodynamic and/or kinetic stability of 3–4 kcal mol$^{-1}$ has been observed in several proteins.**

**Barnase & Binase** are two members of the microbial RNase family. The two protein sequences have 85% identity (17 substitutions and 1 deletion) and almost identical native structures. A multiple mutant with all six mutations that individually stabilizes the wildtype (by ≤ 1.1 kcal/mol$^{-1}$ each) achieves an overall increase in the stability of barnase by 3.3 kcal mol$^{-1}$ relative to wildtype barnase and preserves the same activity.

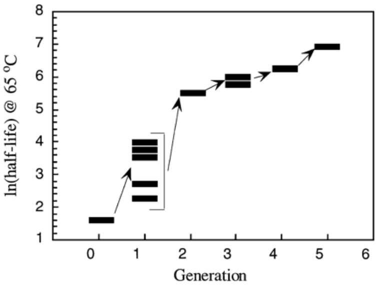
***Excerpts from*****:** Serrano et al. (1993) *J Mol Biol* **233:**305-312.

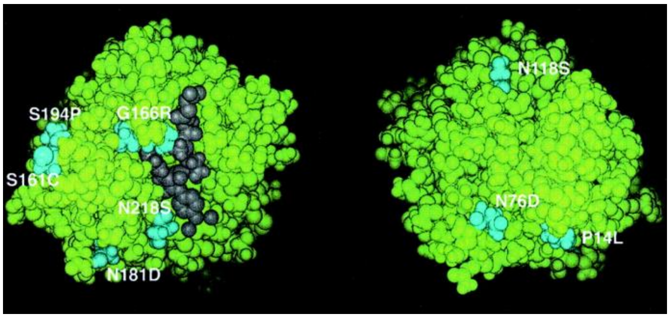
Directed evolution was used to convert *Bacillus subtilis* **subtilisin E** into an enzyme functionally equivalent to its thermophilic homolog thermitase from *Thermoactinomyces vulgaris*. Five generations of random mutagenesis, recombination and screening created **subtilisin E 5-3H5**, whose half-life at 83°C (3.5 min) and temperature optimum for activity (76°C) are identical with those of thermitase. The optimum temperature of the evolved enzyme is 17°C higher and its half-life at 65°C is >200 times that of wild-type subtilisin E. Thermitase differs from subtilisin E at 157 amino acid positions. However, only eight amino acid substitutions were sufficient to convert subtilisin E into an enzyme equally thermostable.

***Excerpts from*****:** Zhao & Arnold (1999) *Protein Eng* **12:**47-53.

## Creating a highly stable Subtilisin E mutant (8 substitutions from wildtype) by Directed Evolution

Random mutagenesis and screening yielded five thermostable first-generation variants, which were recombined to create the second-generation library, from which the single most thermostable variant was selected. A second round of mutagenesis and screening yielded three thermostable third-generation variants, etc.



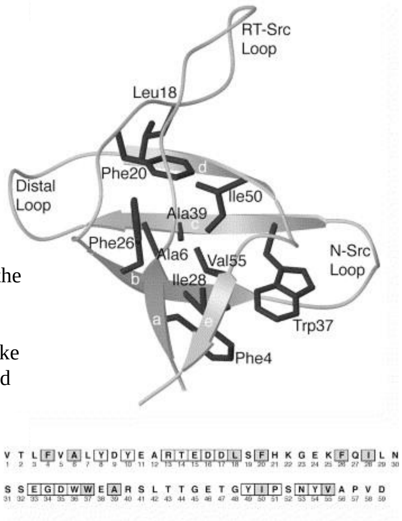


The eight stabilizing amino acid substitutions are all distributed over the surface of the enzyme and most of these positions are far from the active site.

***Results and Figures from*:** Zhao & Arnold (1999) Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Eng* **12:**47-53.

## Were natural proteins selected for thermodynamic stability? Were they selected for fast folding? Were they selected for slow unfolding?

An investigation of 48 natural mutants with single-site substitutions in the hydrophobic core of the SH3 domain indicated that conservation correlates well with unfolding rates but not the folding rates of the mutants.

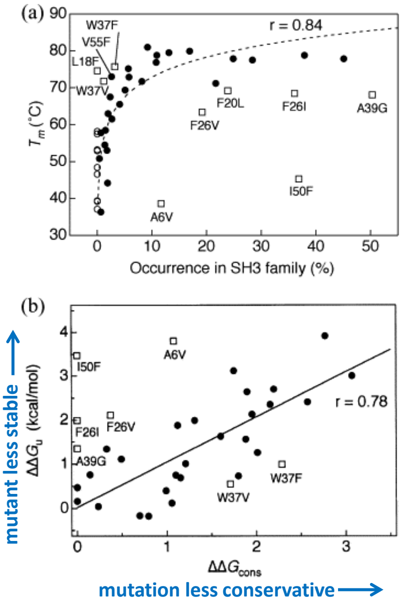


(Top) Fyn tyrosine kinase SH3 domain (1shf) showing the side-chains of the ten hydrophobic core positions. (Bottom) Sequence of the human Fyn SH3 with core residues highlighted by shaded boxes. Residues that make functional contacts in the binding surface are highlighted by unshaded boxes.

***Results and Figure from*:** Di Nardo, Larson & Davidson (2003) The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J Mol Biol* **333:**641-655.

## Conservation of hydrophobic residues in SH3 shows a reasonable positive correlation with thermodynamic stability of the folded state

Comparison of residue conservation in the SH3 family with the thermodynamic stability of the corresponding core mutants in Fyn. (a) Thermal stability of mutants as a function of the residue occurrence frequency in the family shows a logarithmic dependence. (b) The change in stability upon substitution ($\Delta\Delta G_u$; from kinetic analysis) compared to conservation energy ($\Delta\Delta G_{cons}$) shows a very good linear correlation. Some of the outliers (open squares) were rationalized by functional considerations or presumed energetic peculiarities of Fyn SH3.

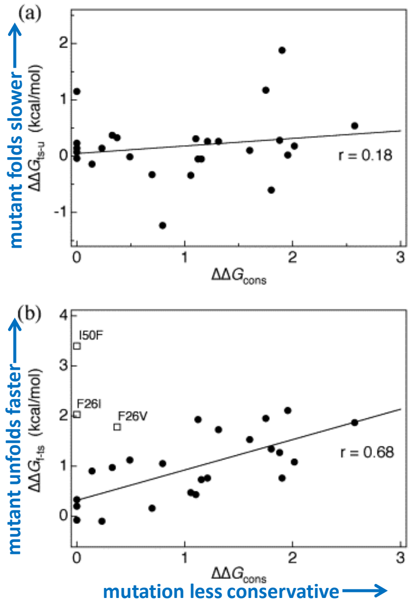


***Results and Figures from***: Di Nardo, Larson & Davidson (2003) *J Mol Biol* **333:**641-655.

## Conservation of hydrophobic residues in SH3 shows a reasonable negative correlation with unfolding rate but no correlation with folding rate

- Hence the positive correlation between conservation and thermodynamic stability arises almost exclusively from the negative correlation between conservation and unfolding rate.

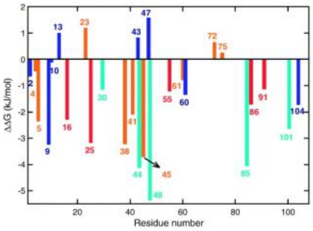
Comparison of kinetic data with residue conservation in the SH3 family. (a) There is little correlation ($r$=0.18) between folding rate and conservation. (b) There is a reasonably good correlation ($r$=0.68) between unfolding rate and conservation.



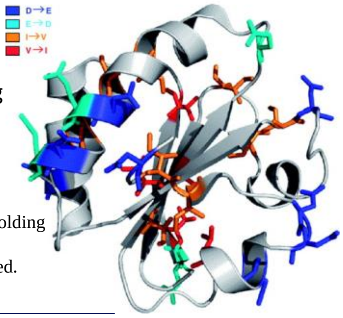
***Results and Figures from***: Di Nardo, Larson & Davidson (2003) *J Mol Biol* **333:**641-655.

## Protein kinetic stability is often more strongly selected by evolution than thermodynamic stability

Protein kinetic stability, i.e. a slow unfolding rate, is often more strongly selected by evolution than thermodynamic stability, most probably because kinetic instability (a faster unfolding rate) facilitates irreversible alteration processes such as amyloid formation and other forms of detrimental protein aggregation even if overall thermodynamic stability is maintained by a higher folding rate [Sanchez-Ruiz (2010) *Biophys Chem* **148:**1-15].



An investigation of 27 single-substitution variants of *E. coli* thioredoxin indicates that viable mutants can at most be 2 kcal mol$^{-1}$ less stable than the wild-type, but a significant correlation exists between slower unfolding rate and the occurrence frequency of a given residue in sequence alignments, again suggesting a significant natural selection for slower unfolding rates.



Experimental mutational effects on the thermodynamic stability (unfolding free energy) at pH 7. The different mutations and the location of the mutated residues in thioredoxin structure (pdb: 2H6X) are color coded.

***Results and Figure from*:** Godoy-Ruiz et al. (2006) *J Mol Biol* **362:**966-978.

## Evolutionary selection of protein kinetic stability

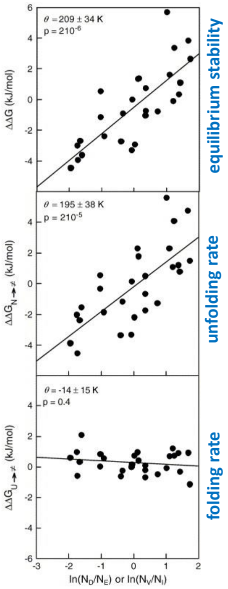
An empirical Boltzmann-like relationship between evolutionary population and thermodynamic stability of protein variants:

$$\frac{N_B}{N_A} \approx \exp\left(-\frac{\Delta\Delta G_{A\to B}}{R\theta}\right)$$

$N_B$, $N_A$ are the nos. of sequences, respectively, with residue type $B$ and $A$ at a given position of an alignment.

$R$ is the gas constant; the effective “temperature” θ may be viewed as a measure of evolutionary pressure on protein stability $\Delta G$.

- Significant correlation between the occurrence frequency of the involved residues with its mutational effect on equilibrium stability and activation free energy for unfolding but NOT activation free energy for folding.

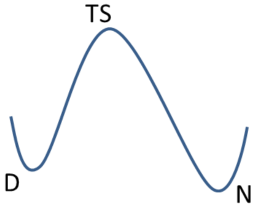


Correlations between effects of I → V and E → D mutations on equilibrium and activation free energy changes for *E. coli* thioredoxin folding-unfolding and occurrence frequencies of residues in sequence alignments. *p* is the statistical significance.

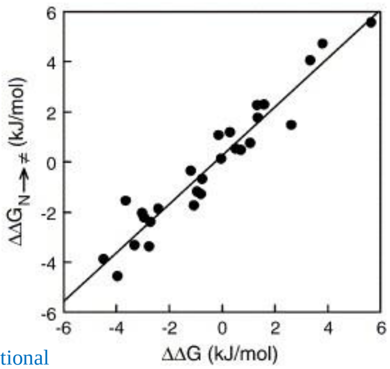
***Results and Figure from*:** Godoy-Ruiz et al. (2006) Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J Mol Biol* **362:**966-978.

## Folding Cooperativity might have been selected evolutionarily

Theoretical and experimental work suggest that a high free energy barrier separating the native and unfolded states is not a fundamental requirement of the physics of protein folding but, more likely, the outcome of natural selection.



**Cooperativity <=> high free energy barrier between D & N <=> very sparsely populated intermediate conformations**



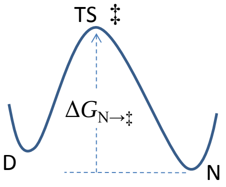
● Essentially all of the correlation between the mutational effects on the equilibrium stability of *E. coli* thioredoxin originates from the mutational effects on unfolding rate.

The slope of the scatter plot between the mutational effects on the activation free energy of unfolding and equilibrium stability is close to unity.

***Results and Figure from*****:** Godoy-Ruiz et al. (2006) *J Mol Biol* **362:**966-978.

## Folding Cooperativity might have been selected because kinetic stability is often a desirable trait for proteins

During thioredoxin evolution, it appears that natural selection eliminated variants that are only a few kJ/mol less stable thermodynamically than the *E. coli* wildtype (WT). This destabilization translates into only a few degrees in equilibrium denaturation temperature $T_m$. Since the WT $T_m$ is already quite high ≈ 90°C, it is unlikely that shifting of $T_m$ by a few degrees would be significantly detrimental.



● Therefore, it is reasonable to posit, at least in the case of thioredoxin, that natural selection does not directly act on thermodynamic stability but on some other linked factor. The evidence suggests strongly that this other factor is the kinetic stability of the protein *in vivo*.

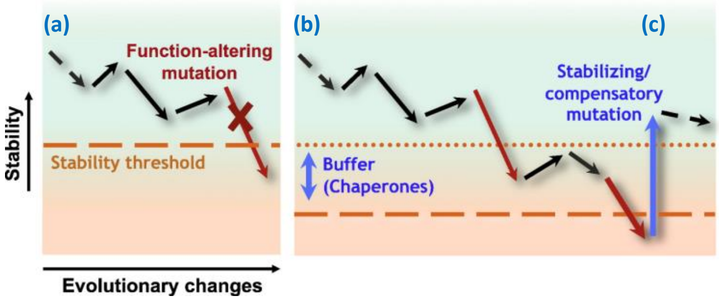
● Thermodynamic stability alone does not guarantee that the protein will remain in the functional native state for long *in vivo* because irreversible alterations such as aggregation -- even if they occur from lowly populated partially unfolded states -- may deplete the native state population progressively.

● The exponential dependence of the unfolding half-life time with $\Delta G_{N\rightarrow\ddagger}$ means that even moderate decreases in activation free energy could significantly compromise kinetic stability *in vivo*.

***Results and discussion from*****:** Godoy-Ruiz et al. (2006) *J Mol Biol* **362:**966-978.

**Manifestations of folding-related selection (e.g. against problematic folding kinetics):**

**Higher thermodynamic/kinetic stability ⇔ higher mutational tolerance**
**⇔ higher evolvability**
**Chaperonin overexpression promotes genetic variation and enzyme evolution**



Most mutations, particularly function-changing ones, are destabilizing. Stability loss is thus a major obstacle to protein evolution.

**Evolutionary dynamics of protein stability.** (a) Mutations that lower the stability below a certain threshold will be purged out because they compromise *function expression*, even if the mutated protein by itself would have increased functionality. (b) Chaperone buffering lowers the stability threshold, allowing further functional mutations to be incorporated. (c) Subsequent stabilizing/compensatory mutations lift the protein above the stability threshold, thus allowing further destabilizing mutations.

***Reference*:** Tokuriki & Tawfik (2009) *Nature* **459:**668-673.
***Results and Figure from*:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates

### Laboratory (directed) enzyme evolution to convert *Pseudomonas aeruginosa* phosphotriesterase (PTE) to a proficient arylesterase

● Experiments suggested that GroEL/ES assists productive folding of PTE by binding individual nonnative conformational intermediates and thus shielding them from detrimental aggregation or degradation. In general, GroEL/ES affects folding kinetics and kinetic stability but it does not change thermodynamic stability. Buffering is transient and its capacity is limited by the number of GroEL/ES in the cell.

● Compensatory mutations (also known as suppressors) are those mutations that restore protein stability undermined by other mutations. In contrast to chaperonin buffering, the beneficial effect of compensatory mutations is that they change the intrinsic stability (thermodynamic and/or kinetic stability) of the protein above a certain selection threshold even in the absence of GroEL/ES action.

● The PTE directed evolution experiment shows that overexpression of GroEL/ES serves to buffer destabilizing mutations and smoothen the evolutionary trajectory. Subsequent compensatory mutations (which were encouraged by turning off overexpression of GroEL/ES) were observed to restore stability and thus enable more evolutionary exploration.

***Results from*:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates

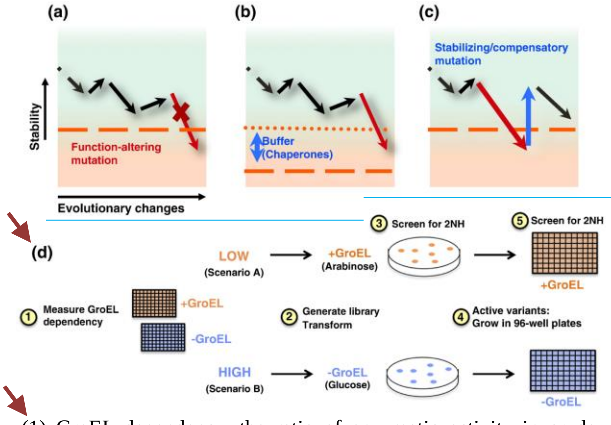
### Overview

A controllable GroEL/ES chaperone co-expression system enabled the authors “to vary the selection environment between buffering and compensatory, which smoothened the trajectory along the fitness landscape to achieve a > 104 increase in arylesterase activity. Biophysical characterization revealed that, in contrast to prevalent models of protein stability and evolution, the variants' soluble cellular expression did not correlate with *in vitro* stability, and compensatory mutations were linked to a stabilization of folding intermediates. Thus, folding kinetics in the cell are a key feature of protein evolvability.”

***Quoted from*****:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates

### Experimental Setup



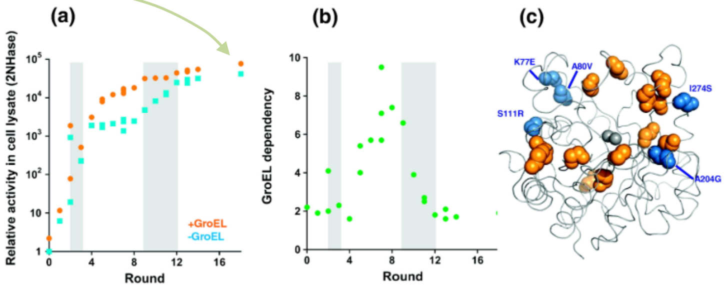
(1) GroEL dependency, the ratio of enzymatic activity in crude lysates with and without GroEL/ES, is measured. A library is generated, transformed into *E. coli* (2) and screened for 2NH activity (3). If GroEL dependency is low, chaperones and proteins are co-expressed (induction by arabinose) to accumulate functional mutations. If GroEL dependency is high, proteins are expressed in the absence of GroEL (suppression by glucose) to select for compensatory mutations. In both cases, positive variants are re-grown (4) and confirmed in a 96-well plate assay (5).

***Results and Figure from*****:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates

### Main Finding

5 orders of magnitude increase in the selected activity

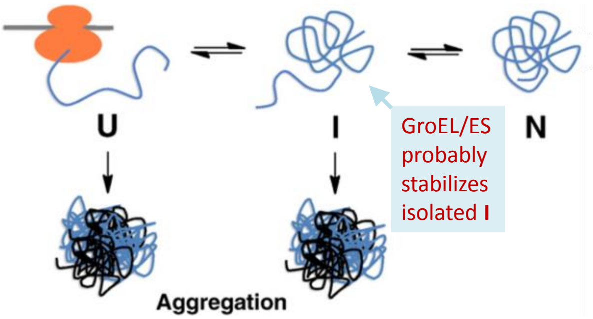


Changes of enzymatic activity and GroEL dependency along the laboratory evolutionary trajectory. (a) Hydrolysis of 2NH by the selected variants in cell lysate compared to wild type PTE in the presence (orange spheres) and absence (blue squares) of GroEL/ES. Shaded areas indicate where chaperone expression was switched off (Round 3 and Rounds 9–12). (b) Change of GroEL dependency (ratio of enzymatic activity in crude lysates following expression with and without GroEL/ES). (c) Crystal structure of wild-type PTE (PDB code 1DPM). The 13 functional mutations are shown in orange; the five compensatory mutations are shown in blue.

***Results and Figure from*:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## Evolvability is affected by protein folding kinetics in the cell

● PTE unfolding irreversible. ● Thermal stability does not correlate with expression level. ● The presence of partially folded intermediates as indicated by ANS binding is positively correlated with the soluble expression level and inversely correlated with GroEL/ES dependency.



GroEL/ES probably stabilizes isolated **I**

The chaperonins GroEL/ES likely stabilize the intermediate state and shift the flux towards **N**. Once PTE reaches the native state **N**, kinetic stability afforded by the high-energy barrier between **U** and **N** (or **U** and **I**) prevents the protein from unfolding and shifting to the aggregation state. In this situation, the amount of **N** in the cell is not governed by either the thermostability of **N**.

***Results and Figure from*:** Wyganowski, Kaltenbach & Tokuriki (2013) *J Mol Biol* **425:**3403-3414.

## Stability-mediated epistasis constrains evolution

The importance of the biophysical property stability (either thermodynamic stability by itself or as a proxy for kinetic stability or cellular expression level) in protein evolution is further demonstrated in a recent experiment indicating how the evolutionary trajectory of influenza nucleoprotein is probably constrained to avoid low-stability sequences.*

The phenomenon of epistasis, referring originally to non-additivity of genetic effects caused by gene interactions, can also manifest within a protein molecule. Mutational effects on stability or function at different sites of a protein can be non-additive when the sites are energetically coupled. A consequence of this biophysical property is that the overall evolutionary effect of multiple mutations can depend on the order in which the mutations are made. For the same given set of mutations, it may be that one temporal order of mutations is evolutionarily favoured because it entails a monotonic increase in fitness, whereas another order of mutations is disfavoured because it involves an intermediate step that decreases fitness. Several studies have demonstrated this type of epistatic behaviour in proteins and its constraints on evolutionary pathways.

***Results from*:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## Stability-mediated epistasis constrains the evolution of an influenza protein

Gong *et al*. examined whether such interactions have indeed constrained evolution of the influenza virus. Between 1968 and 2007, the nucleoprotein—which acts as a scaffold for the replication of genetic material—in the human H3N2 influenza virus underwent a series of 39 mutations. To test whether all of these mutations could have been tolerated by the 1968 virus, Gong et al. introduced each one individually into the 1968 nucleoprotein. They found that several mutations greatly reduced the fitness of the 1968 virus when introduced on their own, which strongly suggests that these ‘constrained mutations’ became part of the virus’s genetic makeup as a result of interactions with ‘enabling’ mutations.

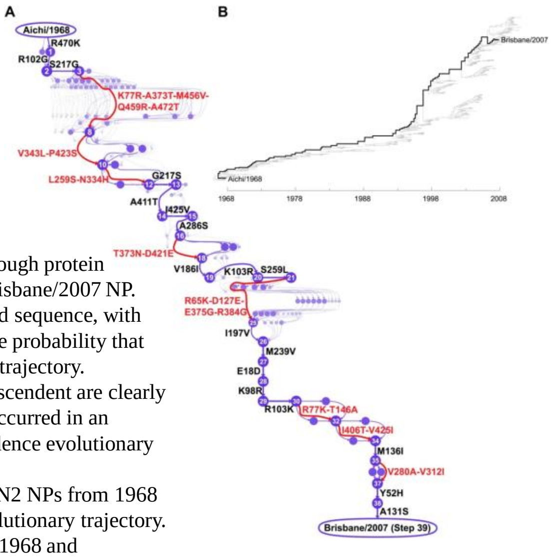
The constrained mutations decreased the stability of the nucleoprotein at high temperatures, while the enabling mutations counteracted this effect. It may, therefore, be possible to identify enabling mutations based on their effects on thermal stability. Intriguingly, the constrained mutations helped the virus overcome one form of human immunity to influenza, suggesting that interactions between mutations might limit the rate at which viruses evolve to evade the immune system.

Overall, these results show that interactions among mutations constrain the evolution of the influenza nucleoprotein in a fashion that can be largely understood in terms of protein stability. If the same is true for other proteins and viruses, this work could lead to a deeper understanding of the constraints that govern evolution at the molecular level.

***Quoted from the Introduction of*:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## Evolution of the nucleoprotein of the human H3N2 influenza virus (1968 → 2007)

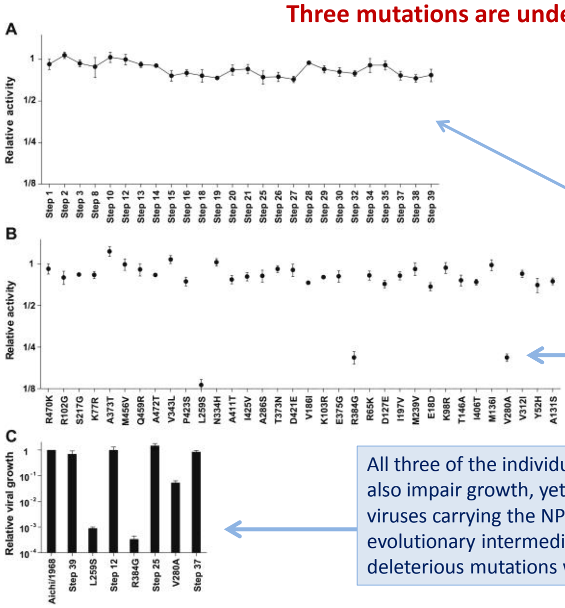
According to the authors’ probabilistic approach to estimate the posterior distribution of mutational events, the most probable trajectory consists of 39 mutational steps at 33 sites (5 mutations revert, e.g., I425V and V425I in steps 14-15 and 32-34; 1 site mutates to two identities – A373T, T373N).



**A.** Inferred evolutionary trajectory through protein sequence space from Aichi/1968 to Brisbane/2007 NP. Each circle represents a unique inferred sequence, with areas and intensities proportional to the probability that sequence was part of the evolutionary trajectory. Mutations for which the parent and descendent are clearly resolved are in black; mutations that occurred in an unknown order are in red. High-confidence evolutionary intermediates have numeric labels.
**B.** Phylogenetic tree of the human H3N2 NPs from 1968 to 2011 that were used to infer the evolutionary trajectory. The lines of descent connecting Aichi/1968 and Brisbane/2007 to their common ancestor are in black.

***Results and Figure from*:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## Three mutations are under strong epistatic constraint



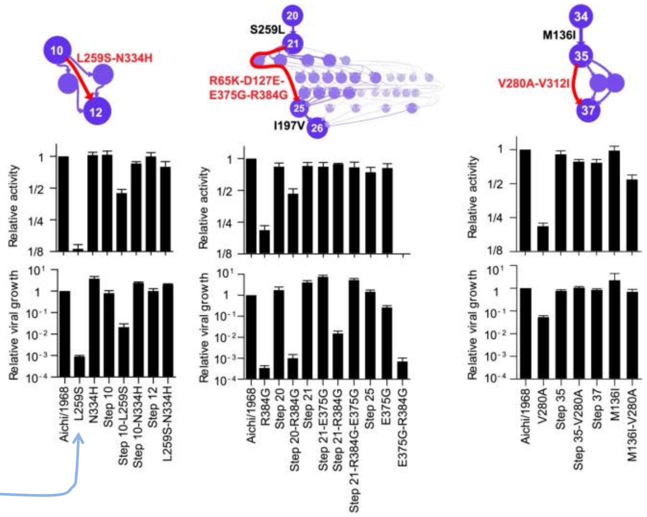
Three mutations are strongly deleterious when introduced individually into the parent Aichi/1968 NP, despite eventually becoming fixed along the evolutionary trajectory without apparent negative effect.

All three of the individual mutations that reduce activity also impair growth, yet there is no defect in the growth of viruses carrying the NPs of the first high-confidence evolutionary intermediates in which these individually deleterious mutations were actually fixed.

***Results and Figure from*:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

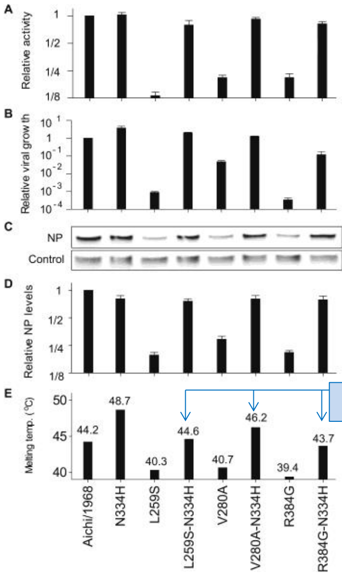
## Individually mutations that are deleterious for the initial Aichi/1968 NP sequence may not be deleterious in the evolutionary intermediates in which they occurred

For example, L259S impairs the transcriptional activity and viral growth of both the parent Aichi/1968 and the evolutionary intermediate Step 10, but is rescued by N334H in both backgrounds.



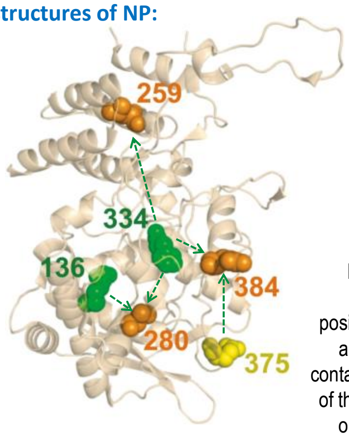
***Results and Figure from*****:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## The epistasis correlates with mutational effects on NP stability



rescue

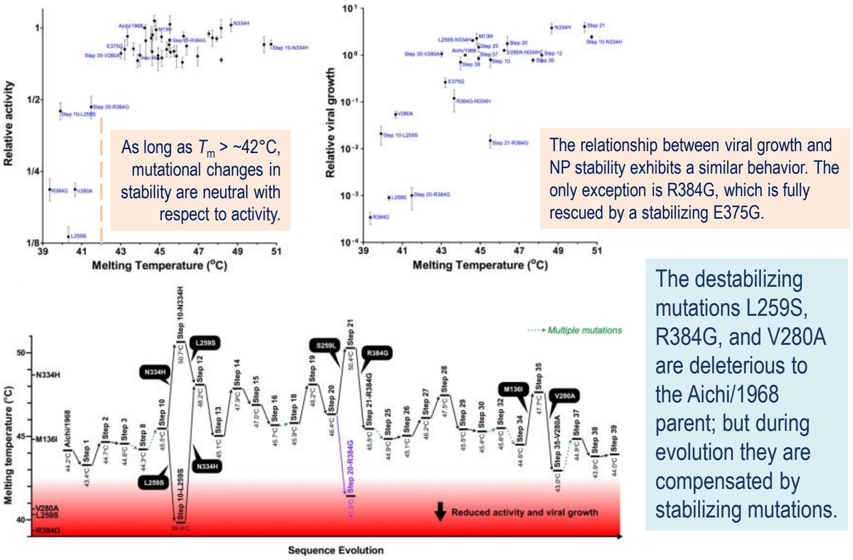
**… but the biophysical origin is not easy to discern: there is no obvious structural basis for the observed epistasis. The epistatically interacting mutations are not in contact in the solved crystal structures of NP:**



Note: the mutated positions are also not in contact in any of the known oligomeric structures

***Results and (unannotated) figures from*****:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## Most of the epistasis in the evolution of NP can be rationalized by counterbalancing stabilizing and destabilizing mutations



As long as $T_m$ > ~42°C, mutational changes in stability are neutral with respect to activity.

The relationship between viral growth and NP stability exhibits a similar behavior. The only exception is R384G, which is fully rescued by a stabilizing E375G.

The destabilizing mutations L259S, R384G, and V280A are deleterious to the Aichi/1968 parent; but during evolution they are compensated by stabilizing mutations.

***Results and Figure from*:** Gong, Suchard & Bloom (2013) *eLife* **2:**e00631.

## Natural Proteins are "Marginally" Stable

Native stability is required for globular proteins to function and to avoid misfolding and aggregation. It might seem that a higher native stability should always be desirable and therefore favoured by evolution. However, natural globular proteins are not extremely stable. Experimental data indicate an approximate native stability of 5–15 kcal mol$^{-1}$ for a natural globular protein with about 100 amino acids. This level of stability of natural globular proteins is often characterized as 'marginally stable'. 'Marginal' here points to the relatively small free energies of folding. Sometimes the term also refers to the fact that the net balance of 5–15 kcal mol$^{-1}$ for native stability is the result of a partial cancellation of two much larger free energies on the order of 100–200 kcal mol$^{-1}$ contributed by favourable intra-protein interactions on one hand and conformational entropy on the other.

***References*:**

Privalov & Gill (1988) Stability of protein structure and hydrophobic interaction. *Adv Protein Chem* **39:**191-234.
Dill (1990) Dominant forces in protein folding. *Biochemistry* **29:**7133-7155.

## Is Marginal Stability an Evolutionarily Selected Property?

If evolutionary selection for stability is expected, why are natural proteins only marginally stable? One possible reason is that native stability is not the only requirement on a functional globular protein. Conformational flexibility is crucial for certain functions. Therefore, adaptation towards increased conformational flexibility might have acted as a check against proteins evolving to become extremely stable, suggesting that marginal stability can be an adaptive trait.

However, is a strong selection pressure for marginal stability necessary to account for the observed marginal stability of natural proteins? Biophysical models have shown that marginal stability could be non-adaptive simply because the number of sequences encoding for a given structure generally decreases with native stability. Hence, even in the absence of any evolutionary selection, among sequences encoding for a given native structure, there are more sequences with low stabilities than with high stabilities. In this view, as long as a certain minimum stability requirement for folding and function is met, random mutational drift will lead an evolving population to a region of sequence space that encodes with marginal stabilities close to the minimum required stability.

***Reference*:** Taverna & Goldstein (2002) Why are proteins marginally stable? *Proteins* **46:**105-109;
Goldstein (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* **79:**1396-1407.
Reviewed in: Sikosek & Chan (2014) *J Royal Society Interface* **11:**20140419

## Summary

### Synergy between the studies of protein biophysics and protein evolution

● Evolutionary rate depends on the structural context of the mutated amino acid residue position in the protein. Biophysical information has been applied to improve the interpretation of *d*N/*d*S. Possible existence of protein sectors.

● Evolutionary/sequence alignment information may be exploited to infer biophysical properties of protein structure and dynamics by applying covariance and DCA analyses.

● Most single-point mutations are thermodynamically destabilizing – at least for a few proteins for which extensive data are available. But laboratory directed evolution shows that substantial stabilization is possible with multiple substitutions. There are likely both adaptive and non-adaptive (purely physical) origins of marginal stability of extant proteins.

● Kinetic stability is a likely evolutionary pressure more important than thermodynamic (equilibrium) stability. In this respect, equilibrium stability may be seen as largely a proxy for kinetic stability in protein evolution.

● The importance of kinetic/thermodynamic stability in protein evolution has been demonstrated in chaperonin overexpression direct evolution experiment and observation of stability-induced epistasis.