A solvable model of the genesis of amino-acid sequences via coupled dynamics of folding and slow genetic variation

ACC Coolen, S Rabello, CJ Pérez-Vicente and F Fraternali

Motivation: proteins have non-random disorder ...
Dynamics of folding and sequence selection
Finite-\( n \) replica analysis, replicated transfer matrices
The limit \( n \to \infty \), deterministic sequence selection
Numerical results, simulations
Summary and outlook
1. **MOTIVATION**

Proteins are disordered systems, but with non-random disorder ...
Primary structure: monomer sequence (the disorder), DHJKAFACGD ...
Secondary structure: local conformation of \( \alpha \)-helices, \( \beta \)-sheets, etc
Tertiary structure: 3D arrangement of secondary structure elements

‘Knowledge of a protein’s tertiary structure is a prerequisite for the proper understanding and engineering of its function.’

Problem for statistical mechanics
To use disordered systems techniques a la Parisi, we need a formula for the disorder statistics ...

- Random amino-acid sequences do not fold into unique conformations, amino-acid sequences of proteins have been selected during evolution
- Our options for ensembles of sequences:
  (i) find a formula for a nontrivial ensemble of random amino-acid sequences?
  (ii) empirically: download all sequences from biomedical data base?
2. MODEL DEFINITIONS

slow process:  

- genetic selection of sequences \( \lambda \)  
- Hamiltonian \( H_{\text{eff}}(\lambda) \)

fast process:  

- folding of residue orientations \( \phi \)  
- Hamiltonian \( H_f(\phi|\lambda) \)

- No defn of sequence statistics: define genetic dynamics of sequences  
- Simple Hamiltonians, focus on secondary structure  
- Solve coupled dynamics for disparate timescales using finite \( n \) replica method  
- Exploit 1D nature of proteins: replicated transfer matrices

\[ \lambda_i : \text{the local amino-acid type} \]
\[ \phi_i : \text{residue angle relative to ‘backbone’} \]

primary structure:  
\( (\lambda_1, \ldots, \lambda_N) \)

secondary structure:  
\( (\phi_1, \ldots, \phi_N) \)
The fast process: folding

Variables: angles \( \phi = (\phi_1, \ldots, \phi_N) \)
\[ \phi_i = \{0, 2\pi/q, \ldots, (q-1)2\pi/q\}, \]
\( q = 2, 3, \ldots \)

\[
H_f(\phi|\lambda) = -\frac{J_p}{N} \sum_{ij} \xi(\lambda_i) \xi(\lambda_j) \delta_{\phi_i,\phi_j} - J_s \sum_i \cos[(\phi_{i+1} - \phi_i) - (\phi_i - \phi_{i-1}) - a(\lambda_i)]
\]

- polarity energy:
  proxy for energy gain by folding in 3D
  \( \xi(\lambda) \): polarity of residue \( \lambda \),
  \( \xi > 0 \): hydrophobic, \( \xi < 0 \): hydrophilic

- steric energy: mechanical constraints,
  residues ‘stick out’, distort homogeneous winding
  \( a(\lambda) \): winding shift induced by residue \( \lambda \)
The slow process: genetic selection of sequences

sequence fitness:

(i) sequence must give protein with reproducible conformation
(ii) structure is useful, e.g. can act as catalyst of some reaction

translate into minimization of

\[ H_{\text{eff}}(\lambda) = U(\lambda) + V(\lambda) + F_f(\lambda) \]

- \(U(\lambda)\): biological utility as catalyst
- \(F_f(\lambda)\): free energy of folding process
  (low free energy = proxy for reproducible conformation)
- \(V(\lambda)\): energetic cost of not having strictly hydrophilic ‘surface residues’ and strictly hydrophobic ‘core residues’
stochastic minimization of Glauber type, noise level $\tilde{T}$:
genetic selection evolves to equilibrium state,

$$P_\infty(\lambda) \propto \exp[-\tilde{\beta}H_{\text{eff}}(\lambda)]$$

combined model solved in equilibrium
by calculating effective free energy

$$f_N = -\frac{1}{\beta N} \log \sum_\lambda e^{-\tilde{\beta}H_{\text{eff}}(\lambda)} = -\frac{1}{n\beta N} \log \sum_\lambda e^{-n\beta[U(\lambda)+V(\lambda)]} [Z_f(\lambda)]^n$$

• temperature ratio $n = \tilde{\beta}/\beta$
• folding partition function $Z_f(\lambda) = \sum_\phi \exp[-\beta H_f(\phi|\lambda)]$
• solvable with replica method (finite $n$ version)
• $n \to 0 (\tilde{T} \to \infty)$: free energy of system with quenched random sequences
• effective free energy as generator of observables:

$$H_f(\phi|\lambda) \to H_f(\phi|\lambda) + \chi NG(\phi, \lambda) : \langle \langle G(\phi, \lambda) \rangle \rangle_{\text{fast}} \Rightarrow_{\text{slow}} \lim_{\chi \to 0} \frac{\partial}{\partial \chi} f_N$$
Connections with earlier studies

  random sequences (no genetic dynamics), but included hydrogen bonds

  genetic dynamics, but only (long-range) polarity forces, $J_s = J_g = 0$

Simple choices for remaining parameters

- Sequence utility potential: $U(\lambda) = \sum_i u(\lambda_i)$, $u(\lambda) = \mu \xi(\lambda) + \nu \cos[a(\lambda)]$
- Energetic cost of polarity imbalance: $V(\lambda) = J_g N v(\frac{1}{N} \sum_i \xi(\lambda_i) - k^*)$
- periodic boundary conditions, $N$ even
- chemical characteristics of amino-acids statistically indep:

$$w(\xi, \eta) = \frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} \delta[\xi - \xi(\lambda)] \delta[\eta - \cos[a(\lambda)]] = w(\xi)w(\eta)$$

$\Lambda$ : nr of amino acid species, i.e. 20
Assumed amino-acid properties

\(a(\lambda)\): winding shift of residue \(\lambda\)

\(\xi(\lambda)\): polarity of residue \(\lambda\)

- independence of polarity and steric properties?

- preferred overall polarity

\[k^* = N^{-1} \sum_i \xi(\lambda_i)\]

(Eisenberg scale)
3. REPLICA ANALYSIS OF THE MODEL

write $Z^n_i(\lambda)$ in terms of $n$ replicas of the system, sum over sequences $\lambda$ before sum over conformations, $f = \lim_{N \to \infty} f_N = \text{extr}_Z \varphi_n(z)$

\[
n \varphi_n(z) = J_p \sum_{\alpha \phi} z_{\alpha \phi}^2 + n J_g[v\left(\frac{1}{n} \sum_{\alpha \phi} z_{\alpha \phi} - k^*\right) - \left(\frac{1}{n} \sum_{\alpha \phi} z_{\alpha \phi}\right) v'\left(\frac{1}{n} \sum_{\alpha \phi} z_{\alpha \phi} - k^*\right)] - \frac{1}{\beta} \log \Lambda
\]

\[
- \lim_{N \to \infty} \frac{1}{\beta N} \log \sum_{\phi_1 \ldots \phi_n} \prod_{i} M[\phi_{i-1}, \phi_i, \phi_{i+1}|z]
\]

\[
M[\phi_{i-1}, \phi_i, \phi_{i+1}|z] = \frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} e^{\beta \xi(\lambda)} \sum_{\alpha} [2J_p z_{\alpha \phi} - J_g v'\left(\frac{1}{n} \sum_{\alpha \phi} z_{\alpha \phi} - k^*\right)] + \beta J_s \sum_{\alpha} \cos[\phi_{i+1} - \phi_i - 2\phi - a(\lambda)] - n\beta u(\lambda)
\]

structure: replicated transfer matrix product
embedded within a mean-field calculation
in principle solvable!
only order pars with one replica index, so RS ok
**Simplest case** $q = 2$: $\phi_i \in \{-\pi/2, \pi/2\}$

$\phi_i = \sigma_i \pi/2$, with $\sigma_i = \pm 1$

$$H_f(\sigma|\lambda) = -\frac{J_p}{2N} \sum_{ij} \xi(\lambda_i)\xi(\lambda_j)[1 + \sigma_i\sigma_j] - J_s \sum_i \cos[a(\lambda_i)]\sigma_{i+1}\sigma_{i-1}$$

**solution:**

$$f = \text{extr}_{m,k} \left\{ \frac{1}{2}J_p(m^2+k^2) + J_g[v(k-k^*)-kv'(k-k^*)] - \frac{\log \Lambda}{\beta n} - \frac{1}{\beta n} \log \lambda(m, k) \right\}$$

$\lambda(m, k)$: largest eigenvalue of $2^n \times 2^n$ transfer matrix

$$M_{\sigma \sigma'}(m, k) = \langle e^{\beta[\sum_\sigma J_p + \sum_\sigma J_g v'(k-k^*)]} \rangle_{\sigma} \langle e^{\beta[\sum_{\sigma} \xi(\lambda_i)\xi(\lambda_j)[1 + \sigma_i\sigma_j] - \mu - J_g v'(k-k^*)]} \rangle_{\xi}$$

$$\langle g(\xi) \rangle = \int d\xi \, w(\xi) g(\xi),$$

$$\langle g(\eta) \rangle = \int d\eta \, w(\eta) g(\eta)$$
physical meaning of \( \{m, k\} \):

\[
    m = \lim_{N \to \infty} \frac{1}{N} \sum_i \langle \xi(\lambda_i) \langle \sigma_i \rangle \rangle_{\text{slow}} \\
    k = \lim_{N \to \infty} \frac{1}{N} \sum_i \langle \xi(\lambda_i) \rangle_{\text{slow}}
\]

saddle-point equations:

\[
    m = \frac{\sum_{\sigma \sigma'} u^{\text{L}}_{\sigma} \sigma_1 Y_{\sigma \sigma'} u^{\text{R}}_{\sigma'}}{\lambda(m, k) \sum_{\sigma} u^{\text{L}}_{\sigma} u^{\text{R}}_{\sigma'}} \\
    k = \frac{\sum_{\sigma \sigma'} u^{\text{L}}_{\sigma} Y_{\sigma \sigma'} u^{\text{R}}_{\sigma'}}{\lambda(m, k) \sum_{\sigma} u^{\text{L}}_{\sigma} u^{\text{R}}_{\sigma'}}
\]

where

\[
    Y_{\sigma \sigma'} = \langle e^{\beta \eta [J_s \sigma \cdot \sigma' - n \nu]} \rangle_{\eta} \langle \xi e^{\eta \beta \xi [J_p (k + \frac{n}{\eta} \sum_{\alpha} \sigma_{\alpha}) - \mu - J_g v' (k - k^\star)]} \rangle_{\xi}
\]

\[
    \sum_{\sigma'} M_{\sigma \sigma'} u^{\text{R}}_{\sigma'} = \lambda(m, k) u^{\text{R}}_{\sigma}, \quad \sum_{\sigma'} u^{\text{L}}_{\sigma'} M_{\sigma' \sigma} = \lambda(m, k) u^{\text{L}}_{\sigma}
\]
Solution of replicated eigenvalue problem

\[ u^R_\sigma = \int dx \, \Phi(x)e^{\beta x \sum \sigma_\sigma}, \quad u^L_\sigma = \int dy \, \Psi(y)e^{\beta y \sum \sigma_\sigma} \]

from replicated spins to effective fields:

\[ \lambda \Phi(x) = \int dx' \Lambda_\Phi(x, x') \Phi(x') \quad \lambda \Psi(x) = \int dx' \Lambda_\Psi(x, x') \Psi(x') \]

\[ \Lambda_\Phi(x, x') = \langle \langle \delta[x - \xi J_p m - A(x', \eta J_s)]e^{n\beta[B(x', \eta J_s) + \xi(J_p k - \mu - J_g v(k - k^*) - \nu \eta)\rangle\rangle}_{\xi, \eta} \]

\[ \Lambda_\Psi(x, x') = \langle \langle \delta[x - A(x' + \xi J_p m, \eta J_s)]e^{n\beta[B(x' + \xi J_p m, \eta J_s) + \xi(J_p k - \mu - J_g v(k - k^*) - \nu \eta)\rangle\rangle}_{\xi, \eta} \]

with

\[ A(x, y) = \beta^{-1} \tanh^{-1}[\tanh(\beta x) \tanh(\beta y)] \]

\[ B(x, y) = \frac{1}{2} \beta^{-1} \log[4 \cosh[\beta(x + y)] \cosh[\beta(x - y)]] \]

everything follows from \( \Phi, \Psi \) ...
simplify, play around ...

\[ m = \int d\xi dh \ W(h, \xi) \ \xi \tanh(\beta h) \quad k = \int d\xi dh \ W(h, \xi) \ \xi \]

\[ W(h, \xi) = \frac{p(\xi) \cosh^n[\beta h] \int dx \ \Psi(x) \Psi(h - x - J_p m \xi)}{\int d\xi dh' p(\xi') \cosh^n[\beta h'] \int dx \ \Psi(x) \Psi(h' - x - J_p m \xi')} \]

in which

\[ p(\xi) = \frac{w(\xi) e^{n\beta \xi(J_p k - \mu - J_g v(k-k^*))}}{\int d\xi' w(\xi') e^{n\beta \xi'(J_p k - \mu - J_g v(k-k^*))}} \]

\[ \Psi(x) = \frac{\int dx' \Phi(x') \int d\eta w(\eta) \delta[x - A(x', \eta J_s)] e^{\beta[B(x', \eta J_s) - \nu \eta]}}{\int dx' \Phi(x') \int d\eta w(\eta) e^{\beta[B(x', \eta J_s) - \nu \eta]}} \]

\[ \Phi(x) = \int d\xi \ p(\xi) \Psi(x - J_p m \xi) \]

formulas for \( f \) and for

\[ \pi(\xi, \eta) = \lim_{N \to \infty} \frac{1}{N} \sum_i \langle \langle \delta[\xi - \xi_i] \delta[\eta - \cos[a(\lambda_i)]] \rangle \rangle_1 \]

e.g. \( \pi(\xi, \eta) = \pi(\xi) \pi(\eta), \ \pi(\xi) = \int dh \ W(h, \xi) \)
Simple solutions and special cases

- state without secondary structure (always a soln): \( m = 0 \)

\[
\Psi(x) = \Phi(x) = \delta(x), \quad W(h, \xi) = p(\xi)\delta(h), \quad k = \frac{\int d\xi \xi w(\xi)e^{\beta [Jp k - \mu - gv(k-k^*)]}}{\int d\xi w(\xi)e^{\beta [Jp k - \mu - gv(k-k^*)]}}
\]

- infinite temperature: \( \beta = 0 \)

\[
\Psi(x) = \delta(x), \quad W(\xi, h) = w(\xi)\delta(h), \quad m = 0, \quad k = \int d\xi \xi w(\xi)
\]

\[
\lim_{\beta \to 0} \beta f = -n^{-1} \log \Lambda - \log 2
\]

- Random sequences: \( n \to 0 \)

\[
\Psi(x) = \int dy \, \Psi(y) \langle \delta[x - A(y + J_p m \xi, \eta J_s)] \rangle_{\xi, \eta}, \quad \Phi(x) = \langle \Psi(x - J_p m \xi) \rangle_{\xi}
\]

\[
m = \int dx dx' \, \Phi(x') \Psi(x) \langle \langle \xi \tanh[\beta(x + \xi J_p m + A(x', \eta J_s))] \rangle \rangle_{\xi, \eta}
\]

recovers Skantzos et al 2001

(random bond chain methods, ratios of constrained partition functions)
4. DETERMINISTIC SEQUENCE SELECTION

choose \( v(u) = \frac{1}{2} u^2 \),
define natural polarity balance \( k_0 = \frac{k^* - \mu / J_g}{1 - J_p / J_g} \)

take \( n \to \infty \) in system below:

\[
\Psi(x) = \frac{\int dx' \int d\xi' p(\xi') \Psi(x') \int d\eta' w(\eta') \delta[x - A(x' + J_pm\xi, \eta J_s)] e^{n\beta[B(x' + J_pm\xi, \eta J_s) - \nu \eta]}}{\int dx' \int d\xi p(\xi) \Psi(x') \int d\eta w(\eta) e^{n\beta[B(x + J_pm\xi, \eta J_s) - \nu \eta]}}
\]

\[
m = \frac{\int d\xi p(\xi) \int dx \int dy \Psi(x) \Psi(y) \tanh[\beta(J_pm\xi + x + y)] \cosh^n[\beta(J_pm\xi + x + y)]}{\int d\xi p(\xi) \int dx \int dy \Psi(x) \Psi(y) \cosh^n[\beta(J_pm\xi + x + y)]}
\]

\[
k = \frac{\int d\xi p(\xi) \int dx \int dy \Psi(x) \Psi(y) \cosh^n[\beta(J_pm\xi + x + y)]}{\int d\xi p(\xi) \int dx \int dy \Psi(x) \Psi(y) \cosh^n[\beta(J_pm\xi + x + y)]}
\]

\[
p(\xi) = \frac{w(\xi) e^{n\beta(J_p - J_g)(k - k_0)}}{\int d\xi' w(\xi') e^{n\beta(J_p - J_g)(k - k_0)}}
\]
Form of $\Psi(x)$ for $n \to \infty$

- $\exists \Omega \subseteq [-J_s, J_s]: \Psi(x) = 0$ for $x \notin \Omega$, $\Psi(x) = e^{n\beta\psi(x)}$ for $x \in \Omega$
- $\max_{x \in \Omega} \psi(x) = 0$
- need to find $\Omega$ and $\lim_{n \to \infty} \psi(x)$

several pages later ...

$J_g > J_p$: $k = k_0$, heteropolar, $m = 0$ or $F_{\beta J_p}(m) = -\tanh(\beta J_s)$

$J_g < J_p$: $k = \pm 1$, homopolar, $m = 0$ or $F_{\beta J_p}(m) = \text{sgn}(\nu) \tanh(\beta J_s)$

with

$$F_x(m) = \frac{\tanh \left[ \frac{1}{2}xm - \frac{1}{2}\tanh^{-1}(m) \right]}{\tanh \left[ \frac{1}{2}xm + \frac{1}{2}\tanh^{-1}(m) \right]}$$

![Graph showing $F_x(m)$ with $x = 5$ and $x = 0.5$]
Phase diagrams

inhom polarity, swollen (IS): $\pi(\xi)$ continuous, $m = 0$
inhom polarity, collapsed (IC): $\pi(\xi) = \frac{1}{2}(1+k_0)\delta(\xi-1)+\frac{1}{2}(1-k_0)\delta(\xi+1), \ m \neq 0$
hom polarity, swollen (HS): $\pi(\xi) = \delta(\xi \pm 1), \ m = 0$
hom polarity, collapsed (HC): $\pi(\xi) = \delta(\xi \pm 1), \ m \neq 0$
hom polarity, mixed (HM): $\pi(\xi) = \delta(\xi \pm 1)$, coexistence of $m = 0$ and $m \neq 0$

$\nu > 0$: favours helices, ↑↓↑↓↑↓↑...  
$\nu < 0$: favours $\beta$-sheets, ↑↑↑↑↑↑↑...
5. NON-DETERMINISTIC SEQUENCE SELECTION

Transitions for finite $n$, increased genetic noise

Generally hard ...
except continuous transitions away from $m = 0$

$$m \rightarrow \Delta m, \quad k \rightarrow k + \Delta k, \quad \Psi(x) \rightarrow \delta(x) + \Delta \Psi(x)$$

gives

$$\Delta \Psi(x) = \frac{\int dx'[\Delta \Psi(x') - J_p k \Delta m \delta'(x')]}{\int d\eta w(\eta) \delta[x - A(x', \eta J_s)] e^{\eta [B(x', \eta J_s) - \nu \eta]}}$$

$$\Delta m = 2k \int dh \; \tanh(\beta h) \cosh^n(\beta h) \Delta \Psi(h) + \beta J_p \Delta m \int d\xi \; p(\xi) \xi^2 + O(\Delta^2)$$

soln:

$$\Delta \Psi_A(x) = \frac{\lambda J_p k}{\lambda - 1} \delta'(x) \Delta m \quad \lambda = \frac{\int d\eta w(\eta) \tanh(\beta \eta J_s) e^{\eta [B(0, \eta J_s) - \nu \eta]}}{\int d\eta w(\eta) e^{\eta [B(0, \eta J_s) - \nu \eta]}}$$
continuous \( m \neq 0 \) bifurcations:

\[
\Delta m \neq 0 : \quad 1 = \beta J_p \left[ \int d\xi \: \xi^2 p(\xi) - \frac{2\lambda k^2}{\lambda - 1} \right]
\]

\[
p(\xi) = \frac{w(\xi) e^{n\beta (J_p - J_g)(k - k_0)}}{\int d\xi' w(\xi') e^{n\beta (J_p - J_g)(k - k_0)}}
\]

\[
\lambda = \frac{\int d\eta \: w(\eta) \tanh(\beta \eta J_s) e^{n\beta [B(0,\eta J_s) - \nu \eta]}}{\int d\eta \: w(\eta) e^{n\beta [B(0,\eta J_s) - \nu \eta]}}
\]

\[e.g. \ J_g \leq \frac{1}{2} J_p, \ \nu = \frac{1}{2}\]

onset of discont transition at \( n = 2! \)

(as in other coupled dynamics models)
Numerical solution via population dynamics

\( n \) appears in exponents, which limits numerical analysis to \( n \leq 400 \)

\[(J_s, J_p, J_g) = (0.1, 1, 2), \ k_0 = 0.7, \ \mu = 0.2, \ \nu = 0.5\]

\( n \to \infty \): continuous IS\( \to \)IC transition at \( T_c = 1.183 \)

large but finite \( n \): discontinuous
prediction: \( \lim_{n \to \infty} \Psi(x) = \frac{1}{2} \delta(x + x^*) + \frac{1}{2} \delta(x - x^*) \)
finite \( n \) corrections: \( \mathcal{O}(n^{-1/2}) \) for \( \nu > 0 \), \( \mathcal{O}(n^{-1}) \) for \( \nu < 0 \)

\( (J_s, J_p, J_g) = (0.1, 1, 2) \), \( n = 200 \), \( k_0 = 0.2 \), \( \mu = 0.7 \), and \( T = 1.07 \)
\( \nu > 0 \): favours helices, ↑↓↑↓↑↓↑…
\( \nu < 0 \): favours \( \beta \)-sheets, ↑↑↑↑↑↑↑…
requires two nested equilibrations of disordered systems, inner ‘loop’ of the code: disordered Ising chain ... 

\[ N \text{ too small: no transitions, } N \text{ too large: no equilibration} \]

\[ N = 1000, \text{ at } T = 0.3 \text{ and } n = 200 \]
\[ \nu = J_g = \frac{1}{2}, J_p = 1, \text{ and and } k^* = 0. \]

For \( n \to \infty \): coexistence & remanence
7. SUMMARY AND OUTLOOK

nice:
• solvable models describing protein structure formation
circumvent the obstacle of non-random amino-acid sequences
• nested equilibration of slow/fast processes: finite $n$ replica method
short-range frozen random forces: diagonalization of replicated transfer matrix
• exact results for phase transitions,
especially for deterministic sequence selection, $n \rightarrow \infty$

not so nice:
• many simplifications:
one angle per residue (should be two), simple phenomenological Hamiltonian
no hydrogen bonds, only primary & secondary structure
• potential for evolution to homo-polar polymers,
artifact of Hamiltonian? probably ...
• statements on ensemble of hetero-polymers,
not solution of protein folding problem (not even approximate)
Future directions

*If driven by passion for theory ...*

- introduce contact maps to replace present long-range forces, structure similar to ‘small-world’ topologies,
  more sophisticated order parameters of finitely connected graphs, RSB, etc
- real-valued residue orientations, i.e. \( q \to \infty \)
  diagonalization of replicated transfer kernels

*If driven by passion for biology ...*

- increase level of biological detail:
  two residue angles, with real rather than discrete values,
  more realistic Hamiltonians:
    work our steric effects for real amino-acids
    include hydrogen bonds
  more realistic modeling of tertiary structure influence, via contact maps

there is overlap!