On the optimal contact potential and sequence conservation modes of proteins

Akira R. KINJO
Institute for Protein Research, Osaka University

The 11th Japan-Korea-China Bioinformatics Training Course & Symposium
2013-06-18

Disclaimer

- Old stories only,
- nothing new,
- nothing useful...

Why Protein Structure Prediction is Important?

- Biology is all about genotype-phenotype mapping*.
- Genotype is nothing but DNA/protein sequence.
- Phonotype is nothing but form (structure) and motion (behavior).
- Protein is *both* genotype *and* phonotype.
- Protein folding is the physical process of genotype-phenotype mapping.
- Structure prediction is about understanding principles of this mapping.

∗**Caution**: a highly biased opinion.

Introduction

- Sequence: Eigenvalue decomposition of amino acid substitution matrices,
  A. R. Kinjo and K. Nishikawa (2004).

- Structure: The optimal contact potential for structure prediction,
  A. R. Kinjo and S. Miyazawa (2008).

- Sequence, again: Singular value decomposition of position-specific substitution matrices,
  A. R. Kinjo and H. Nakamura (2008).

Conservation Modes of Amino Acid Sequence

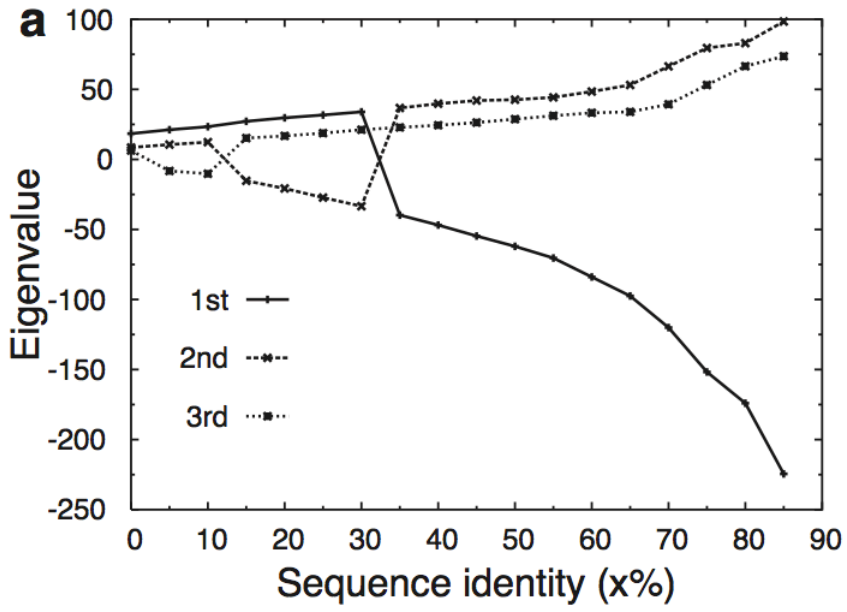
A. R. Kinjo and K. Nishikawa, *Bioinformatics* **20**:2504-2508 (2004)

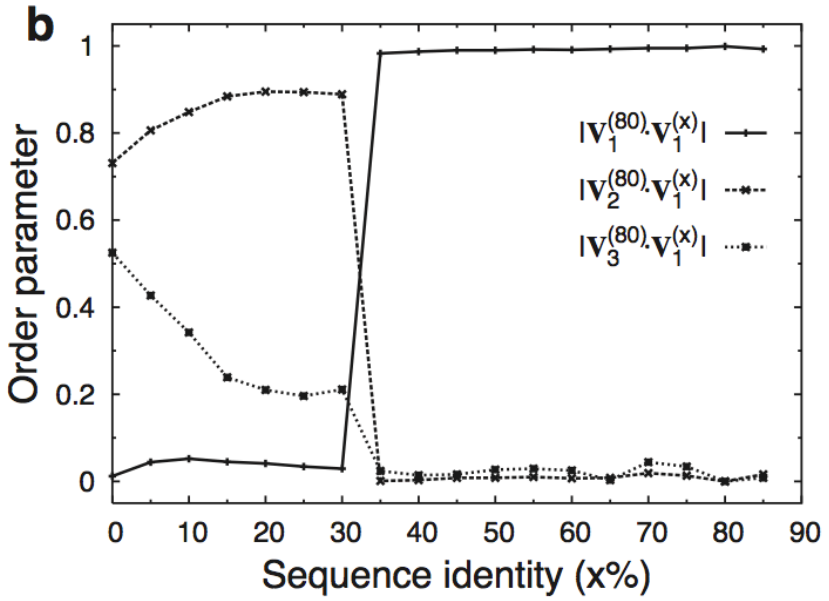
Eigenvalue decomposition of AASM

$$M^{(x)} = \sum_{\alpha=1}^{20} \lambda_{\alpha}^{(x)} \mathbf{v}_{\alpha}^{(x)} \mathbf{v}_{\alpha}^{(x)'}. \tag{1}$$

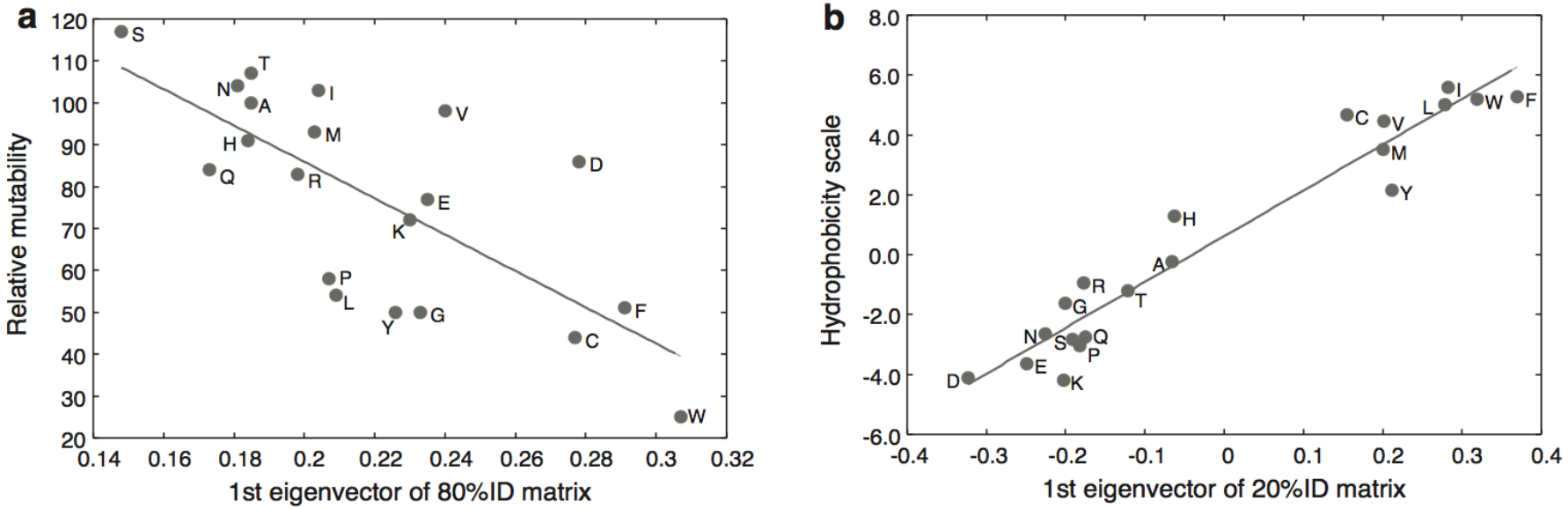
- AASM for each x%ID range.
- Sorted in decreasing order of $|\lambda_{\alpha}|$.

A transition at the “twilight zone.”





Meaning of eigenvectors



Matching eigenvectors with the AAindex database.

- All elements of $\mathbf{v}_1^{(80)}$ are of the same sign.
- Elements of $\mathbf{v}_1^{(20)}$ contain both signs.

Summary (1)

- EVD of AASM shows a transition of conservation modes around the twilight zone.
- In high %ID ranges, "mutability" dominates, and its contribution is negative.
- In low %ID ranges, hydrophobicity dominates, and its contribution can be positive or negative.

Sequence alignment is accurate as long as mutability is dominant?

What is the Optimal Contact Potential for Structure Prediction?

A. R. Kinjo and S. Miyazawa, *Chemical Physics Letters* **451**:132-135 (2008)

The contact potential

Given a sequence S and a conformation C,

- A generalized sequence-dependent contact potential: $\mathcal{E}(S) = (\mathcal{E}_{ij})$.
- Contact matrix: $\Delta(C) = (\Delta_{ij})$ (1 or 0).

$$E(C,S) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{E}_{ij}(S)\Delta_{ij}(C) = \frac{1}{2}\left[\mathcal{E}(S),\Delta(C)\right] \tag{2}$$

The lower bound of contact potential

The Cauchy-Schwarz inequality says:

$$[\mathcal{E}, \Delta] \geq -\|\mathcal{E}\|\|\Delta\| \tag{3}$$

where the equality holds if and only if

$$\mathcal{E} = \epsilon\Delta \tag{4}$$

for some $\epsilon(< 0)$.

Remark:

$$\|\Delta\|^2 = 2N_c \tag{5}$$

where $N_c = (1/2)\sum_{i,j}\Delta_{ij}$ is the total number of contacts.

Conditions for the lower bound

- If $\mathcal{E} = \epsilon\Delta(C_n)$ holds for the native conformation C_n, then $\mathcal{E}$ is a Gō potential.
- For the native conformation to be the unique GMEC*, it should be maximally compact (maximal $\|\Delta\|^2 = 2N_c$).

*GMEC: Global Minimum Energy Conformation.

Spectral relations

Let's examine more generous lower bounds using SVD:

$$\Delta = \sum_{\alpha=1}^{N} \sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T, \tag{6}$$

$$\mathcal{E} = \sum_{\alpha=1}^{N} \tau_\alpha \mathbf{x}_\alpha \mathbf{y}_\alpha^T. \tag{7}$$

The von Neumann trace formula says

$$[\mathcal{E}, \Delta] \geq -\sum_{\alpha=1}^{N} \sigma_\alpha \tau_\alpha \tag{8}$$

where the equality holds if and only if

$$(\mathbf{u}_\alpha^T \mathbf{x}_\beta)(\mathbf{v}_\alpha^T \mathbf{y}_\beta) = -\delta_{\alpha,\beta}. \tag{9}$$

More on spectral relations

In terms of EVD, we have

$$\Delta = \sum_{\alpha=1}^{N} \lambda_\alpha \mathbf{u}_\alpha \mathbf{u}_\alpha^T, \tag{10}$$

$$\mathcal{E} = \sum_{\alpha=1}^{N} \epsilon_\alpha \mathbf{x}_\alpha \mathbf{x}_\alpha^T \tag{11}$$

where $|\lambda_\alpha| = \sigma_\alpha$, $|\epsilon_\alpha| = \tau_\alpha$. Thus, the I.b. restated:

$$[\mathcal{E}, \Delta] \geq -\sum_{\alpha=1}^{N} \sigma_\alpha \tau_\alpha = \sum_{\alpha=1}^{N} \lambda_\alpha \epsilon_\alpha \tag{12}$$

with $\lambda_\alpha \epsilon_\alpha \leq 0$ for all $\alpha = 1, \cdots, N$.

Yet more on spectral relations

- Assume that the l.b. is met and that $\mathrm{rank}(\mathcal{E}) = \mathrm{rank}(\Delta)$.

Then, by Sylvester's law of inertia, there exists a non-singular matrix S such that

$$\mathcal{E} = -S\Delta S^T, \tag{13}$$

i.e., $\mathcal{E}$ is *congruent to Δ. "Structure prediction" is a matter of matrix inversion:

$$\Delta = -S^{-1}\mathcal{E}S^{-T}. \tag{14}$$

But non-native structures may have lower energies... But we have

$$[\mathcal{E}, \Delta] \geq -\sum_{\alpha=1}^{N} \sigma_\alpha \tau_\alpha \geq -\sqrt{\sum_\alpha \sigma_\alpha^2}\sqrt{\sum_\alpha \tau_\alpha^2} = -\|\mathcal{E}\|\|\Delta\|. \tag{15}$$

1D approximation (1)

Just pick the first eigencomponent of $\mathcal{E}$:

$$\mathcal{E} \approx \epsilon_1 \mathbf{x}_1 \mathbf{x}_1^T \tag{16}$$

(although it is NOT a very good approximation). The l.b. ($= \epsilon_1 \lambda_1$) is obtained if

$$\mathbf{x}_1 = \pm \mathbf{u}_1. \tag{17}$$

Empirically known facts:

1. if $\mathbf{x}_1$ is set to some kind of hydrophobicity scale, it is highly correlated to the native $\mathbf{u}_1$.

2. Actual $\mathbf{u}_1$ is well correlated with contact numbers.

1D approximation (2)

Average over columns. Let

$$\langle \mathcal{E}_{i\bullet} \rangle = \frac{1}{N} \sum_{j=1}^{N} \mathcal{E}_{ij} \tag{18}$$

$$n_i = \sum_{i=1}^{N} \Delta_{ij} \tag{19}$$

and $\mathbf{e} = (\langle \mathcal{E}_{1\bullet} \rangle, \cdots, \langle E_{N\bullet} \rangle)^T$ and $\mathbf{n} = (n_1, \cdots, n_N)$.

$$E(C, S) \approx \frac{1}{2} \mathbf{e}^T \mathbf{n} \geq -\frac{1}{2} \|\mathbf{e}\| \|\mathbf{n}\| \tag{20}$$

where the equality holds iff $\mathbf{e} = \epsilon \mathbf{n}$.

Summary (2)

- The optimal contact potential is Gō-like.
- Hence, it must be sequence position-specific.
- 1D approximations are not perfect.

How can we construct sequence position-dependent contact potentials?

A logical conclusion: We cannot have the optimal contact potential for structure prediction.

More positively: Once we have the optimal contact potential, we are done.

How important is structural information?

(an analysis in hindsight)

A. R. Kinjo and H. Nakamura, *PLoS One* **3**:e1963 (2008)

Singular value decomposition (SVD) of PSSM

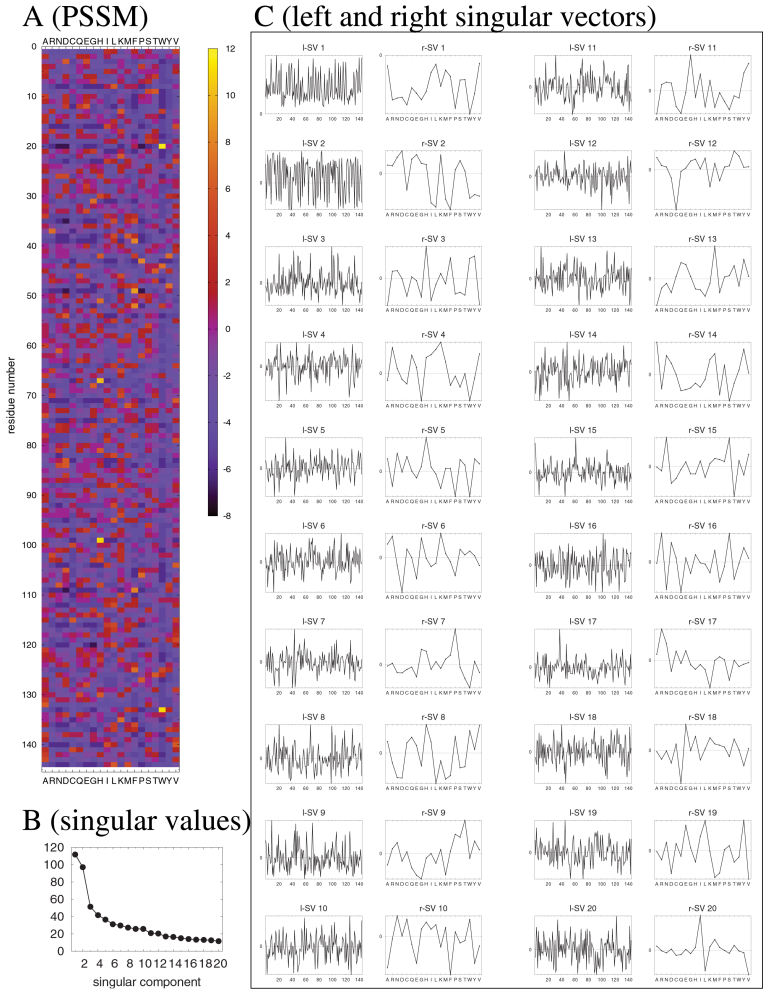
- PSSM (position-specific scoring matrix) is an $N \times 20$ matrix.
- Any matrices can be SVDed.

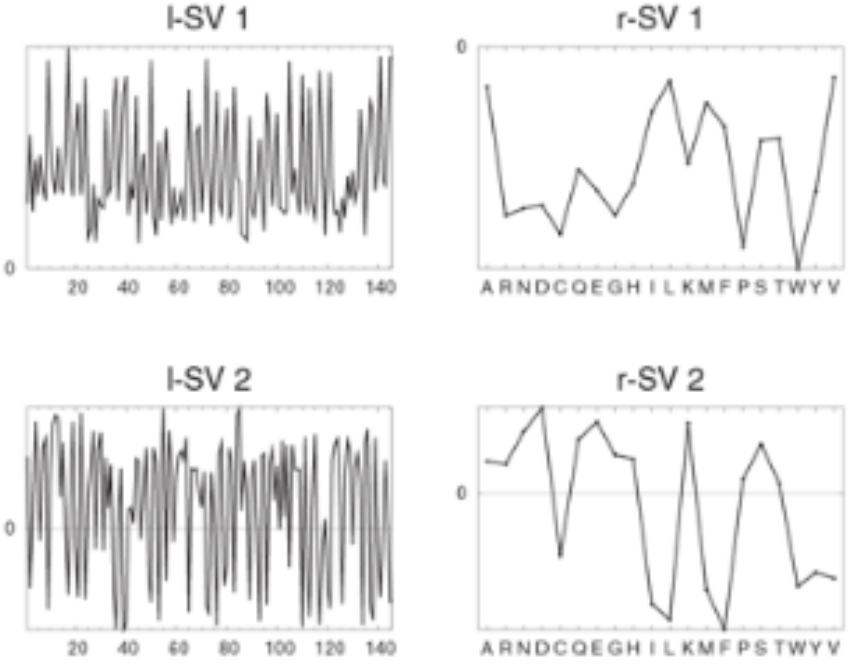
$$M = U\Sigma V^T = \sum_{\alpha=1}^{20} \sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T \tag{21}$$

where

- $\sigma_\alpha (\geq 0)$ is a *singular value*,
- $\mathbf{u}_\alpha$ is a *left singular vector* of N dimension,
- $\mathbf{v}_\alpha$ is a *right singular vector* of 20 dimension,

SVD of PSSM: example





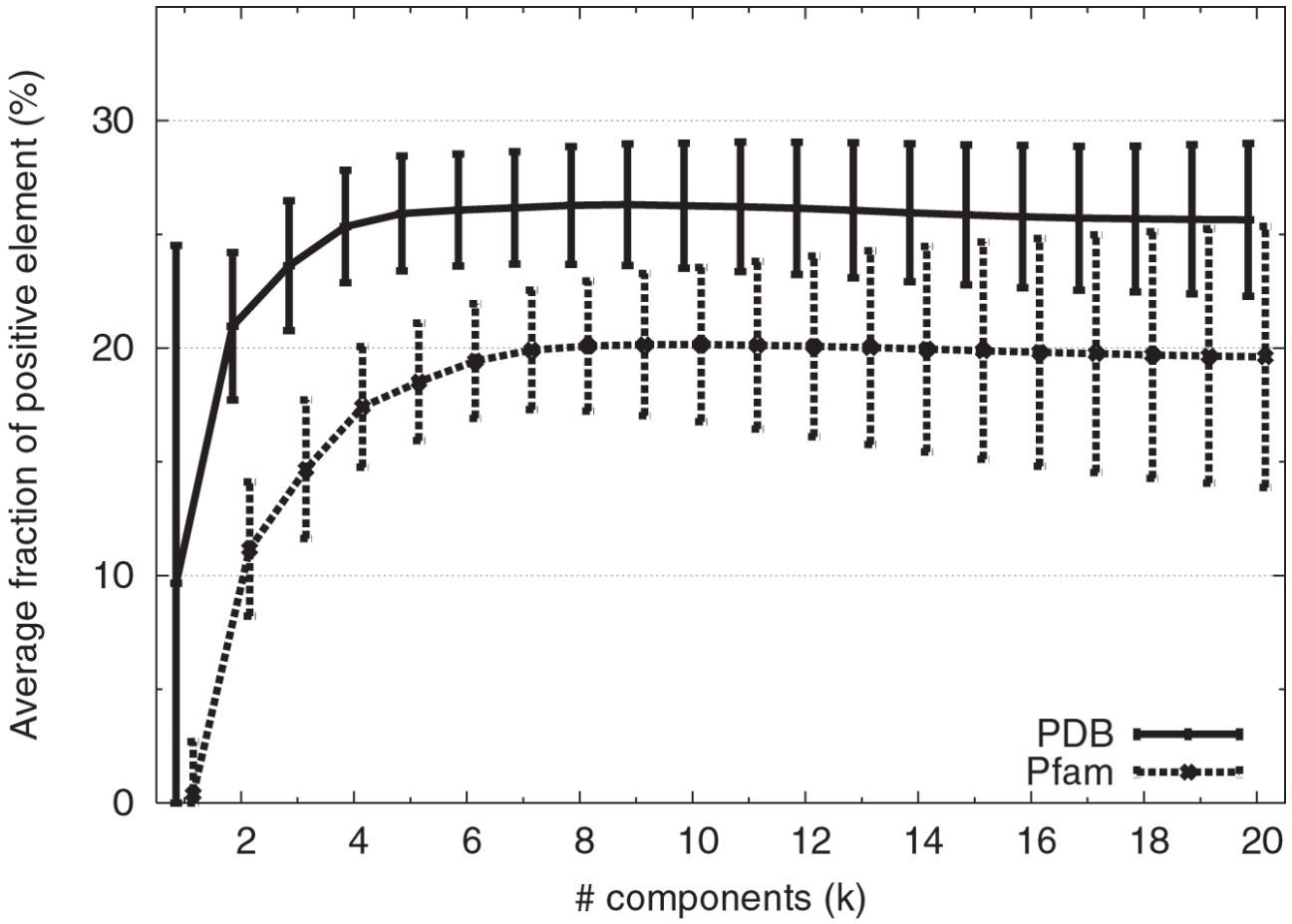
Interpretation of singular vectors

$$M = \sum_{\alpha=1}^{20} \sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T \tag{22}$$

$\mathbf{u}_\alpha$		$\mathbf{v}_\alpha$
left singular vector	dual	right singular vector
N	$\leftrightarrow$	20
1D structure		amino acid index

Fraction of positive PSSM elements

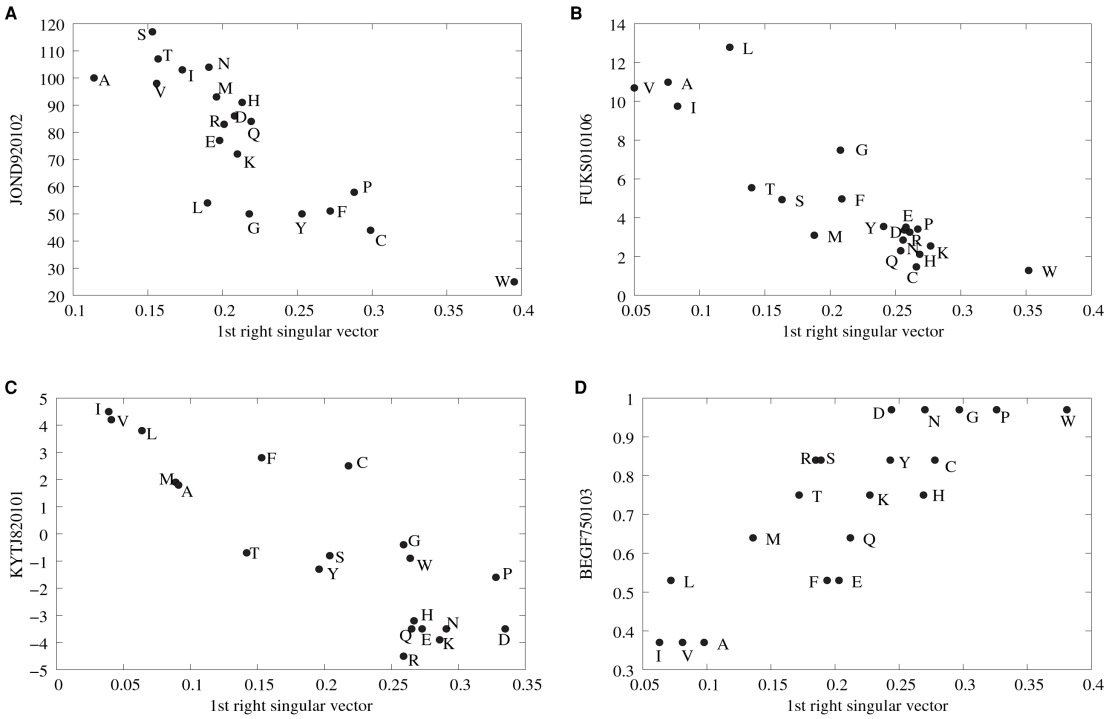
Partial PSSM: $M_k = \sum_{\alpha=1}^{k} \sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T$



The first *right* singular vector *vs.* AAindex

rank	PDB	Pfam
1	JOND920102 (10)	SNEP660101 (9)
2	FUKS010106 (7)	DESM900101 (7)
3	MCMT640101 (6)	KYTJ820101 (6)
4	MEEJ810101 (6)	WOLS870102 (6)
5	BEGF750103 (5)	JOND920102 (6)
6	KYTJ820101 (4)	FUKS010106 (5)
7	ROBB790101 (4)	BEGF750103 (3)
8	KIDA850101 (3)	CORJ870108 (3)
9	ROBB760108 (3)	LEVM780106 (2)
10	MIYS990101 (3)	AURR980120 (2)

The description of each AAindex ID can be found at http://www.genome.jp/dbget-bin/www_bfind? aaindex.
doi:10.1371/journal.pone.0001963.t002



Mutability, interior AA composition, Kyte-Doolittle, β-turn, ...*Note the sign!*

1st component contributes negatively!

- The elements of partial PSSM

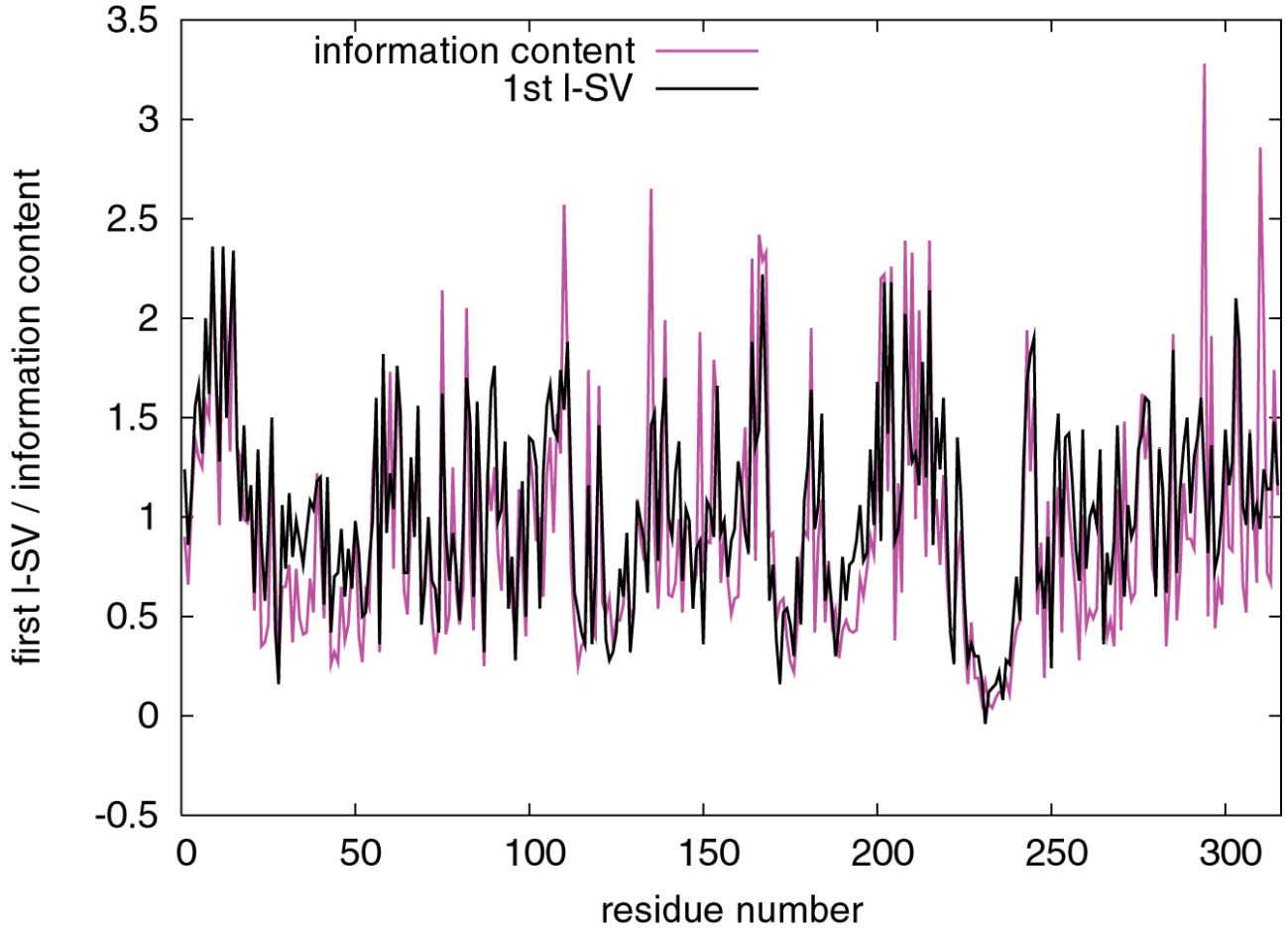
$$M_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \tag{23}$$

  are almost always negative!

- It disfavors any substitutions!

- Functional constraints?

The first *left* singular vector *vs.* ...?

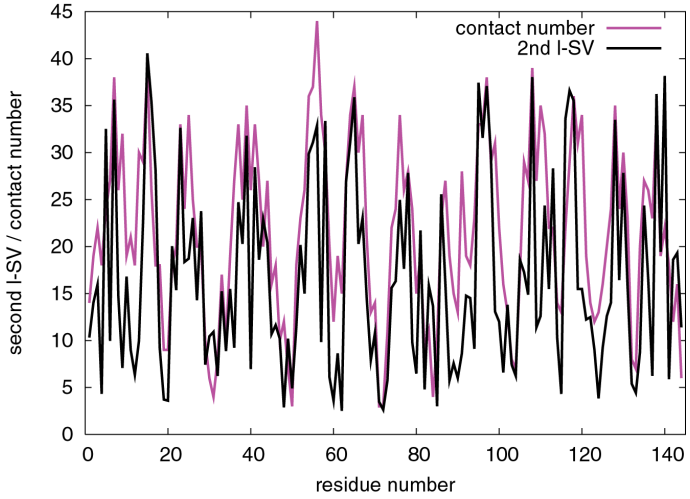


correlation coefficient: 0.543 (mean) / 0.601 (median).

The second singular vectors

left

contact number



right

hydrophobicity

rank	PDB	Pfam
1	MIYS990101 (33)	MIYS990101 (54)
2	BASU050101 (12)	GRAR740102 (12)
3	BASU050103 (11)	BASU050103 (10)
4	GRAR740102 (8)	MIYS990102 (7)
5	SWER830101 (7)	BASU050101 (6)
6	MIYS990102 (6)	SWER830101 (1)
7	ZHOH040103 (2)	ZHOH040103 (1)
8	CORJ870102 (2)	MIYS990105 (1)
9	KYTJ820101 (2)	FAUJ830101 (1)
10	FAUJ830101 (2)	CORJ870102 (1)

doi:10.1371/journal.pone.0001963.t003

Right SV Almost exclusively correlated with some kind of hydrophobicity scales!

Left SV Hence, some kind of core/surface properties such as contact numbers.

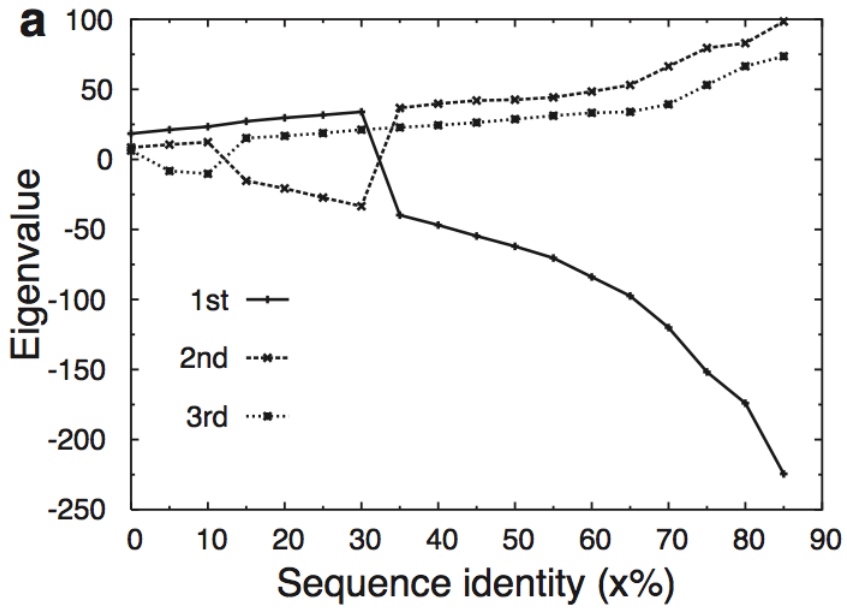
Summary (3)

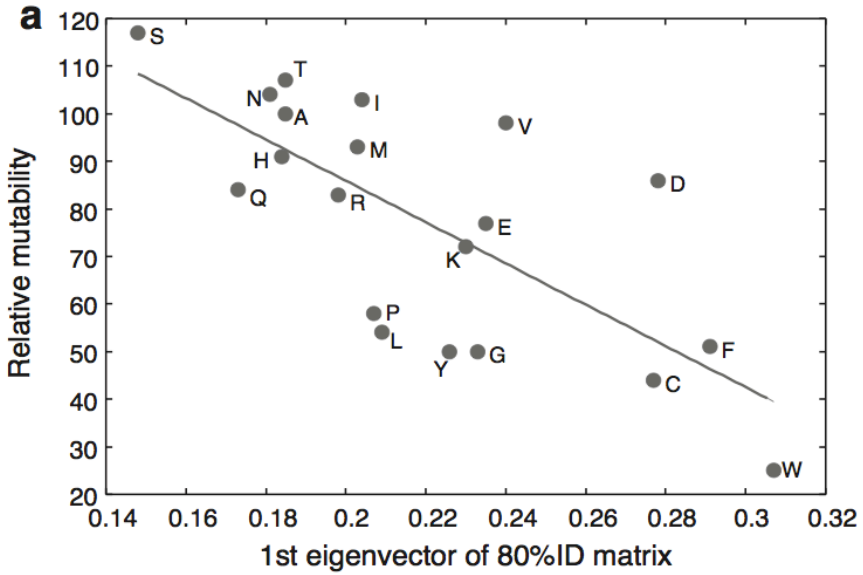
1. The 1st singular component contributes negatively.

2. The 1st left and right singular vectors are related to various conserved properties.

3. The 2nd singular component corresponds to hydrophobicity and structural stability.

Remember AASM?





- High %ID → mutability 1st (*negative contribution*), hydrophobicity 2nd → clear homology detection

- Low %ID → hydrophobicity 1st → vague homology detection

Reconsider PSSM

- Hydrophobicity (structural requirement) is secondary!
- Family-specific conservation mode is primary!

Function precedes structure!

Not *vice versa*.

Implications to optimal contact potential

- Extracting (purely) structural information may be difficult. (Since it's of secondary importance.)
- How can an optimal contact potential incorporate functional modes?
- Or, are we missing something important? e.g.,

$$\mathcal{E}(S) \rightarrow \mathcal{E}(S, C)$$

(sequence- *and* conformation-dependent potential)

By the way,...

Do we *REALLY* need

the optimal contact potential?

The End