Research History

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

at KMUTT D-Lab on February 9, 2018

A full publication list is available in google scholar (https://scholar.google.com/citations?user="Sanzo Miyazawa").

1. Education

Education

- April 1966 - March 1970
  Tokyo Metropolitan University, Department of Physics
- April 1970 - May 1973
  Nagoya University, Graduate Division, Department of Physics
- June 1973 - March 1978
  Kyushu University, Graduate division, Department of Biology

- March 1978
  Doctor of Science from Nagoya University
  - Relationship between the type of amino acid substitutions in protein evolution and site position in protein strucure.

2-1. At NIH, 1/3

Research/teaching experience

- April 1978 - March 1979
  Postdoctoral Trainee, Department of Biology, Kyushu University
- May 1979 - Nov. 1985
  Visiting Fellow/Associate,
  Laboratory of Experimental and Computational Biology,
  CCR, NCI, National Institutes of Health, USA

Protein folding process:

- Miyazawa, Sanzo, and Jernigan, Robert L.:
  Equilibrium folding and unfolding pathways for a model protein.
  Biopolymers, Vol. 21, pp. 1333-1363, (1982).
- Miyazawa, Sanzo, and Jernigan, Robert L.:
  Most probable intermediates in protein folding-unfolding with a non-interacting globule-coil model.
  Biochemistry, Vol. 21, pp. 5203-5213, (1982).

2-1. At NIH, 2/3

- **Protein folding process:**
  - Miyazawa, Sanzo, and Jernigan, Robert L.:
    Equilibrium folding pathways for model proteins.
    Journal of Statistical Physics, Vol. 30, pp. 549-559, (1983).
- **Knowledge-based interaction potential between residues from known protein 3D structures**
  - Miyazawa, Sanzo, and Jernigan, Robert L.:
    Estimation of effective interresidue contact energies in protein crystal structures: quasi-chemical approximation.
    Macromolecules, Vol. 18, pp. 534-552, (1985).
    Top 10th most cited until 2006 in Macromolecule.

2-1. At NIH, 3/3

- **Others:**
  - Miyazawa, Sanzo:
    Cooperative ligand binding on multi-dimensional lattices: Bethe approximation.
    Biopolymers, Vol. 22, pp. 2253-2271, (1983).
  - Miyazawa, Sanzo:
    Statistical mechanics of supercoiling-induced B-to-Z transitions in a closed circular DNA: one-dimensional model system with a double quadratic displacement potential and long range interactions.
    Journal of Chemical Physics, Vol. 83, pp. 859-883, (1985).

2-2. At NIG

- Dec. 1985 - March 1991
  Associate Professor, Center for Genetic Information Research
  National Institute of Genetics in Japan
- Oct. 1988 - March 1991
  Associate Professor, Department of Genetics,
  The Graduate University for Advanced Studies

Task: to establish DNA Sequence Data Bank of Japan (DDBJ)
subtask: to establish a IP network with GenBank and EMBL

- Miyazawa, Sanzo:
  DNA Data Bank of Japan: present status and future plans.
  In: Computers and DNA, Santa Fe Institute Studies in the Sciences of Complexity. Vol. VII, Eds. G. I. Bell and T. G. Marr, Reading MA: Addison-Wesley, ISBN 0-201-51505-9, pp. 47-61, (1990).

2-3. At Gunma Univ., 1

- April 1991 - March 2007
  Associate Professor, Department of Basic Science,
  Faculty of Engineering, Gunma University
- April 2007 - March 2013
  Associate Professor, Department of Information Science,
  Graduate School of Engineering, Gunma University

  Teaching:

  - For undergraduates: C language, Data structure, Sort and Merge
  - For graduates: Bioinformatics; protein structure, stochastic Hopfield network, sequence alignment, hidden Markov model, sequence analysis, genome science, phylogeny, Bayesian network, SVM

  Design and Management of Campus LAN
- August - September 2003, August - September 2004, September 2005
  Visiting Professor, L. H. Baker Center for Bioinformatics,
  Iowa State University, USA

2-3-1.Probabilistic pairwise alignment method, 1/6

Probabilistic alignment method

- Miyazawa, Sanzo:
  A reliable sequence alignment method based on probabilities of residue correspondences.
  Protein Engineering, Vol. 8, pp. 999-1009, (1995).

Alignment A_l

$$A_l \equiv \begin{bmatrix} \dots & a_2 & a_3 & - & - & a_4 & \dots \\ \dots & - & b_3 & b_4 & b_5 & b_6 & \dots \end{bmatrix} \tag{1}$$

Total alignment score $S(A_l)$ of A_l

Assuming that there is no correlation between site correspondences,

$$S(A_l) \equiv \sum_{\{(i,j)\in A_l\}} s(a_i, b_j) \; - \text{(penalty for gaps)} \tag{2}$$

where

$s(a_i, b_j)$ *a similarity score for a pair of amino acids a_i and b_j*

Probabilistic pairwise alignment method, 2/6

The maximum similarity alignment

$$\text{The maximum similarity alignment } \equiv A \text{ such that } S(A) = \max_{l} S(A_l) \tag{3}$$

The probabilistic alignment

We assume that each alignment A_l is probable with the probability $P(A_l)$.

$$P(A_l) \propto \exp(\frac{S(A_l)}{T}) \tag{4}$$

where T is a scaling parameter.

$$\begin{aligned} \text{The most probable alignment} \quad &\equiv \quad A \text{ such that } P(A) = \max_{l} P(A_l) \\ &= \quad A \text{ such that } S(A) = \max_{l} S(A_l) \\ &= \quad \text{The maximum score alignment} \end{aligned} \tag{5}$$

Probabilistic pairwise alignment method, 3/6

Probabilities of residue-residue correspondences

The probability $p(a_i, b_j)$ that two sites a_i and b_j correspond to each other in all feasible alignments can be represented by

$$p(a_i, b_j) = \frac{1}{Z} \quad Z_{i-1,j-1} \exp(\frac{s(a_i, b_j)}{T}) \quad Z'_{i+1,j+1} \tag{6}$$

$$\begin{aligned} p(a_i, -) &= 1 - \sum_{j=1}^{n} p(a_i, b_j) \\ p(-, b_j) &= 1 - \sum_{i=1}^{m} p(a_i, b_j) \end{aligned} \tag{7}$$

where $Z'_{i+1,j+1}$ is the partition function for partial sequences of a consisting of a_{i+1} to a_m and b from b_{j+1} to b_n. m and n are the sequence length of a and b.

This method is well known as the **transfer matrix** method in statistical physics and the **Viterbi** algorithm with the forward and backward algorithm for HMM in the field of information science.

Probabilistic pairwise alignment method, 4/6

Probability alignment

An probability alignment, which consists of the most probable correspondences, can be made by iteratively choosing a site pair with the maximum probability as follows.

1. Set $i_1 = 1$, $i_2 = m$, $j_1 = 1$, and $j_2 = n$.
2. Calculate a site pair (a_i, b_j) such that $p(a_i, b_j) = \max_{i_1 \leq k \leq i_2, j_1 \leq l \leq j_2} p(a_k, b_l)$, $p(a_i, b_j) \geq p(a_i, -)$, and $p(a_i, b_j) \geq p(-, b_j)$.
3. If there is no such a site pair, align $-$ to all sites of $i_1 \leq i \leq i_2$ and of $j_1 \leq j \leq j_2$.
4. If (a_i, b_j) is such a site pair, choose it as one of residue-residue correspondences in the alignment. Then, repeat the steps of 2 to 4 to align the remaining segments until all the sites are aligned.

This alignment may include residue correspondences that do not correspond to the most probable one for either one, and whose probabilities are not significantly high.

Probabilistic pairwise alignment method, 5/6

A threshold of the probability for reliable residue correspondences

- A site pair (a_i, b_j) with $p(a_i, b_j) > 0.5$ is the most probable correspondence for a given site a_i and for b_j.
- All correspondences with $p(a_i, b_j) > 0.5$, $p(a_i, -) > 0.5$, and $p(-, b_j) > 0.5$ can constitute an alignment.

Therefore residue correspondences with $p > 0.5$ are highly probable correspondences in the probability alignment.

Probabilistic pairwise alignment method, 6/6

However, the condition of Equation 13 is not sufficient to say that site pairs with $p(a_i, b_j) > 0.5$, $p(a_i, \phi) > 0.5$, and $p(\phi, b_j) > 0.5$ can constitute an alignment. In addition, the sequence order among residue correspondences must be compatible with an alignment, that is, the following condition must be satisfied for a set of site pairs to be able to constitute an alignment.

> *Lemma* *Let $p(a_i, b_j) > 0.5$ and $p(a_k, b_l) > 0.5$. If $i < k$, then $j < l$.*
> *Proof* *Any alignment with the match/mismatch pair of a_i and b_j cannot have any match/mismatch pair of a_k and b_l with $i < k$ and $j \geq l$. Thus, if $p(a_i, b_j) > 0.5$, then $\sum_{l=1}^{j} p(a_k, b_l) < 0.5$ for $i < k$. Therefore, when $p(a_i, b_j) > 0.5$ and $p(a_k, b_l) > 0.5$, if $i < k$, then $j < l$.*

Thus, all correspondences with $p(a_i, b_j) > 0.5$, $p(a_i, \phi) > 0.5$, and $p(\phi, b_j) > 0.5$ can constitute an alignment, and therefore are highly probable correspondences in the probability alignment that is constructed by the procedure already described in this section.

2-3. At Gunma Univ., 3

- **Knowledge-based residue potential to estimate the folding energy of protein, 1:**
    - Miyazawa, Sanzo, and Jernigan, Robert L.:
      Protein stability for single substitutions mutants and the extent of local compactness in the denatured state,
      Protein Engineering, Vol. 7, pp. 1209-1220, (1994).
    - Miyazawa, Sanzo, and Jernigan, Robert L.:
      Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading.
      Journal of Molecular Biology, Vol. 256, pp. 623-644, (1996).
    - Miyazawa, Sanzo, and Jernigan, Robert L.:
      Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues.
      PROTEINS: Structures, Function, and Genetics, Vol. 34, pp. 49-68, (1999).

2-3. At Gunma Univ., 4

- **Knowledge-based residue potential to estimate the folding energy of protein, 2:**
  - Miyazawa, Sanzo, and Jernigan, Robert L.:
    Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition.
    PROTEINS: Structures, Function, and Genetics, Vol. 36, pp. 347-356, (1999).
  - Miyazawa, Sanzo, and Jernigan, Robert L.:
    An empirical energy potential with a reference state for protein fold and sequence recognition.
    PROTEINS: Structures, Function, and Genetics, Vol. 36, pp. 357-369, (1999).
  - Miyazawa, Sanzo and Jernigan, Robert L.:
    Long- and short-range interactions in native protein structures are consistent/minimally-frustrated in sequence space.
    PROTEINS: Structures, Function, and Genetics, Vol. 50, pp. 35-43, (2003).

2-3. At Gunma Univ., 5

- **Protein sequence-structure probabilistic alignment by evaluating residue-residue interactions with the knowledge-based potential in the mean field approximation; in other words, threading sequence into structure**
  - Miyazawa, Sanzo, and Jernigan, Robert L.:
    A new substitution matrix for protein sequence searches based on contact frequencies in protein structures.
    Protein Engineering, Vol. 6, pp. 267-278, (1993).
  - Miyazawa, Sanzo and Jernigan, Robert L.:
    Identifying sequence-structure pairs undetected by sequence alignments.
    Protein Engineering, Vol. 13, pp. 459-475, (2000).
  - Miyazawa, Sanzo:
    Protein Sequence-Structure Alignment Based on Site-Alignment Probabilities.
    Genome Informatics, Vol. 11, pp. 141-150, (2000).

2-3. At Gunma Univ., 6

- **Knowledge-based residue potential to estimate the folding energy of protein, 3:**
  - Miyazawa, Sanzo and Jernigan, Robert L.:
    How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins?
    Journal of Chemical Physics, Vol. 122, 024901(pp. 1-18), (2005).
  - Miyazawa, Sanzo and Kinjo, Akira R.:
    Properties of contact matrices induced by pairwise interactions in proteins.
    Physical Review E, Vol. 77, 051910/1-10, (2008).

- **Molecular phylogeny:**
  - Miyazawa, Sanzo:
    Selective constraints on amino acids estimated by a mechanistic codon substitution model with multiple nucleotide changes.
    PLoS One, Vol. 6, e17244/1-22, (2011).
  - Miyazawa, Sanzo:
    Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences.
    PLoS One, Vol. 6, e28892/1-20, (2011).
  - Miyazawa, Sanzo:
    Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic an alysis.
    BMC Evol. Biol., 13, 257, 2013.

2-3. At Gunma Univ., 8

- **Characteristics of protein evolution:**
  - Miyazawa, Sanzo:
    Selection maintaining protein stability at equilibrium.
    J. Theor. Biol., 391, 21-34, 2016.
  - Miyazawa, Sanzo:
    Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness.
    J. Theor. Biol., 433, 21-38, 2017.

Prediction of contact residue pairs in protein structure, 1

Prediction of contact residue pairs based on co-substitution between sites in protein structures

- Miyazawa, Sanzo:
  Prediction of contact residue pairs based on co-substitution between sites in protein structures.
  PLoS One, Vol. 8, e54252/1-20, (2013).
- Miyazawa, Sanzo:
  Prediction of structures and interactions from genome information.
  https://arxiv.org/abs/1709.08021 , (2017).
  This manuscript is supposed to be printed as a chapter of a book, "Integrative Structural Biology with Hybrid Methods" as one of the book series: Advances in Experimental Medicine and Biology from Springer.

Background

- Residue-residue interactions, which fold a protein into a unique 3D structure and make it play a specific function, impose structural and functional constraints on each amino acid.
- Structural and functional constraints are recorded
  - in amino acid orders in homologous protein sequences and also
  - in the evolutionary trace of amino acid substitutions.

Background

- Residue-residue interactions, which fold a protein into a unique 3D structure and make it play a specific function, impose structural and functional constraints on each amino acid.
- Structural and functional constraints are recorded
    - in amino acid orders in homologous protein sequences and also
    - in the evolutionary trace of amino acid substitutions.
- Structural and functional constraints arise from interactions between sites mostly in close spatial proximity.
- As a result, the types of amin acids and amino acid substitutions must be correlated between sites particularly in close spatial proximity.

Background

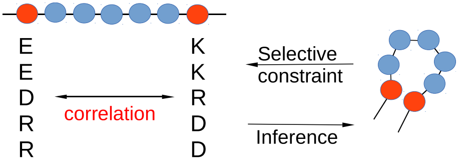
- Residue-residue interactions, which fold a protein into a unique 3D structure and make it play a specific function, impose structural and functional constraints on each amino acid.
- Structural and functional constraints are recorded
  - in amino acid orders in homologous protein sequences and also
  - in the evolutionary trace of amino acid substitutions.
- Structural and functional constraints arise from interactions between sites mostly in close spatial proximity.
- As a result, the types of amin acids and amino acid substitutions must be correlated between sites particularly in close spatial proximity.
- A present challenge is to extract only direct dependences between sites by excluding indirect correlations through other sites and phylogenetic bias.

Two approaches to infer co-evolving site pairs

1. From the equilibrium distribution of amino acid sequences;
   ex. Direct Information (DI) score based on an inverse Potts problem.

2. From the dynamic process of amino acid substitutions: The present approach.

1)

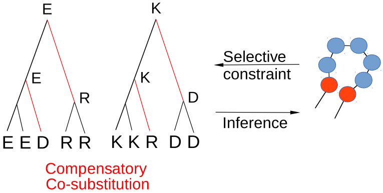


correlation

Selective constraint

Inference

2)



Compensatory Co-substitution

Selective constraint

Inference

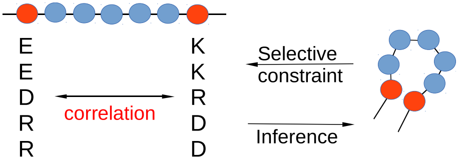
Two approaches to infer co-evolving site pairs

1. From the equilibrium distribution of amino acid sequences;
   ex. Direct Information (DI) score based on an inverse Potts problem.

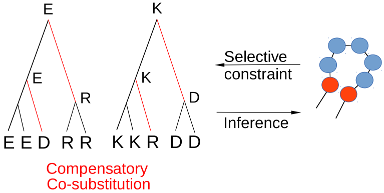
   Recently remarkable prediction accuracy of contact residue pairs was achieved by extracting essential correlations of amino acid types between residue positions by Bayesian graphical models and with a direct information (DI) score.

2. From the dynamic process of amino acid substitutions: The present approach.

1)



2)



Two approaches to infer co-evolving site pairs

1. From the equilibrium distribution of amino acid sequences;

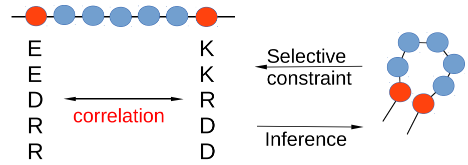
   ex. Direct Information (DI) score based on an inverse Potts problem.

   Recently remarkable prediction accuracy of contact residue pairs was achieved by extracting essential correlations of amino acid types between residue positions by Bayesian graphical models and with a direct information (DI) score.

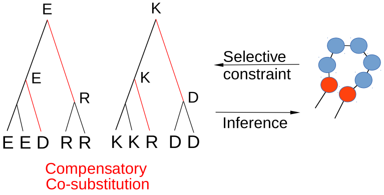
2. From the dynamic process of amino acid substitutions: The present approach.

   Here, we report an alternative approach of inferring co-evolving site pairs from concurrent and compensatory substitutions between sites in each branch of a phylogenetic tree.

1)



2)



Direct coupling analysis, 1/3

Maximum entropy model for the distribution of protein sequences

Let us consider a probability distribution $P(\boldsymbol{\sigma})$ of amino acid sequences, $\boldsymbol{\sigma} \equiv (\sigma_1, \ldots, \sigma_L)^T$ with $\sigma_i \in$ {amino acids, deletion}, single-site and two-site marginal probabilities of which are equal to a given frequency $P_i(a_k)$ of amino acid a_k at each site i and a given frequency $P_{ij}(a_k, a_l)$ of amino acid pair (a_k, a_l) for site pair (i, j), respectively.

$$P(\sigma_i = a_k) \equiv \sum_{\sigma} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k} = P_i(a_k) \qquad (8)$$

$$P(\sigma_i = a_k, \sigma_j = a_l) \equiv \sum_{\sigma} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} = P_{ij}(a_k, a_l) \qquad (9)$$

where $a_k \in$ {amino acids, deletion}, $k = 1, \ldots, q$, $q \equiv |\{\text{amino acids, deletion}\}| = 21$ $i, j = 1, \ldots, L$, and $\delta_{\sigma_i a_k}$ is the Kronecker delta. The distribution P_{ME} with the maximum entropy is

$$P_{\text{ME}}(\boldsymbol{\sigma}|h, J) = \frac{1}{Z} e^{-H_{\text{Potts}}(\boldsymbol{\sigma}|h,J)} \qquad (10)$$

where a Hamiltonian H_{Potts}, which is called that of the Potts model for $q > 2$ (or the Ising model for $q = 2$), and a partition function Z are defined as

$$-H_{\text{Potts}}(\boldsymbol{\sigma}|h, J) = \sum_{i} h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \quad , \quad Z = \sum_{\boldsymbol{\sigma}} e^{-H_{\text{Potts}}(\boldsymbol{\sigma}|h,J)} \qquad (11)$$

Direct coupling analysis, 2/3

Log-likelihood and log-posterior probability

Log-posterior-probability and log-likelihood for the Potts model are

$$\log P_{\text{post}}(h, J|\{\boldsymbol{\sigma}\}) \propto \ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J) + \log P_0(h, J) \tag{12}$$

$$\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J) = B \sum_{\boldsymbol{\sigma}} P_{\text{obs}}(\boldsymbol{\sigma}) \log P_{\text{ME}}(\boldsymbol{\sigma}|h, J) \tag{13}$$

where $P_{\text{obs}}(\equiv \sum_{\tau=1}^{B} \delta_{\boldsymbol{\sigma}\boldsymbol{\sigma}^\tau}/B)$ is the observed distribution of $\boldsymbol{\sigma}$ specified with $\{P_i(a_k)\}$ and $\{P_{ij}(a_k, a_l)\}$, and B is the number of instances; sequences $\boldsymbol{\sigma}^\tau$ are assumed here to be independently and identically distributed samples in sequence space. $P_0(h, J)$ is a prior probability of (h, J).

Let us define cross entropy[**?**] as the negative log-posterior-probability per instance.

$$\begin{aligned} S_0(h, J|\{P_i\}, \{P_{ij}\}) &\propto -(\log P_{\text{post}}(h, J|\{\boldsymbol{\sigma}\}))/B \\ &\equiv S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\}) + R(h, J) \end{aligned} \tag{14}$$

where the cross entropy S_{Potts}, which is the negative log-likelihood per instance for the Potts model, and the negative log-prior per instance R are defined as follows.

$$S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\}) \equiv -\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J)/B \tag{15}$$

$$= \log Z(h, J) - \sum_i \sum_k h_i(a_k) P_i(a_k) - \sum_i \sum_k \sum_{j>i} \sum_l J_{ij}(a_k, a_l) P_{ij}(a_k, a_l) \tag{16}$$

Direct coupling analysis, 3/3

Given marginal probabilities, the estimates of fields and couplings are those minimizing the cross entropy.

$$(h, J) = \arg\min_{(h,J)} S_0(h, J|\{P_i\}, \{P_{ij}\}) \ , \ S_0(\{P_i\}, \{P_{ij}\}) \equiv \min_{(h,J)} S_0(h, J|\{P_i\}, \{P_{ij}\}) \tag{18}$$

Since $S_0(\{P_i\}, \{P_{ij}\})$ is the Legendre transform of $(\log Z(h, j) + R(h, J))$ from (h, J) to $(\{P_i\}, \{P_{ij}\})$, these optimum h and J can also be calculated from

$$h_i(a_k) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_i(a_k)} \quad , \quad J_{ij}(a_k, a_l) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_{ij}(a_k, a_l)} \tag{19}$$

In most methods for contact prediction, residue pairs are predicted as contacts in the decreasing order of score $(\mathcal{S}_{ij})$ calculated from fields $\{J_{ij}(a_k, a_l)|1 \leq k, l < q\}$; see Eqs. **??** and **??**.

Direct coupling analysis, 4/3

Various methods to estimate (h, J)

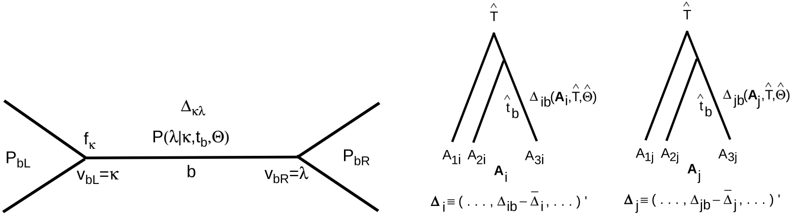
- Boltzmann machine
- Message passing algorithm to estimate marginal probabilities
- Mean field approximation for the inverse Potts model
- Continuous multivariate Gaussian approximation for $P(\sigma)$ with ℓ_1 or ℓ_2 regularization term for precision matrix
- Gaussian approximation for $P(\sigma)$ with a normal-inverse-Wishart prior
- Pseudo-likelihood approximation
- Adaptive cluster expansion of cross-entropy for sparse Markov random field
- ...

Methods: Mean of characteristic changes ($\Delta_{\kappa\lambda}$) at site i in branch b

Likelihood of an alignment $\mathcal{A}$ in a tree T under a codon substitution model Θ : $P(\mathcal{A}|T,\Theta)$

Codon substitutions from κ to λ occur with $P(\lambda|\kappa, t_b, \Theta, \theta_\alpha)$ for branch length t_b.

- Substitutions are assumed to occur independently at each site; $P(\mathcal{A}|T,\Theta) = \prod_i P(\mathcal{A}_i|T,\Theta)$
- Protein evolution is assumed to be in the stationary state in a time-homogeneous and -reversible Markov process.
  $\longrightarrow$ Any node can be regarded as a root node; let us regard the left node v_{bL} of branch b as a root.



P_{bL}
f_κ
$v_{bL}=\kappa$
$\Delta_{\kappa\lambda}$
$P(\lambda|\kappa,t_b,\Theta)$
b
$v_{bR}=\lambda$
P_{bR}
$\hat{T}$
$\hat{t}_b$
$\Delta_{ib}(\mathbf{A}_i,\hat{T},\hat{\Theta})$
A_{1i} A_{2i} A_{3i}
$\mathbf{A}_i$
$\mathbf{\Delta}_i \equiv (\ldots, \Delta_{ib} - \bar{\Delta}_i, \ldots)'$
$\hat{T}$
$\hat{t}_b$
$\Delta_{jb}(\mathbf{A}_j,\hat{T},\hat{\Theta})$
A_{1j} A_{2j} A_{3j}
$\mathbf{A}_j$
$\mathbf{\Delta}_j \equiv (\ldots, \Delta_{jb} - \bar{\Delta}_j, \ldots)'$

$$P(\mathcal{A}_i, v_{bL} = \kappa, v_{bR} = \lambda|T,\Theta) \equiv P_{bL}(\mathcal{A}_i|v_{bL} = \kappa, T, \Theta) f_\kappa P(\lambda|\kappa, t_b, \Theta) P_{bR}(\mathcal{A}_i|v_{bR} = \lambda, T, \Theta) \quad (20)$$

$$P(\mathcal{A}_i|T,\Theta) = \sum_\kappa \sum_\lambda P(\mathcal{A}_i, v_{bL} = \kappa, v_{bR} = \lambda|T,\Theta) \quad (21)$$

$$(\hat{T},\hat{\Theta}) = \arg\max_{T,\Theta} \prod_i P(\mathcal{A}_i|T,\Theta) \quad (22)$$

Phylogenetic tree:

Topology: Pfam reference tree

Branch lengths: by maximizing likelihood in a mechanistic codon substitution model

Mean of characteristic changes ($\Delta_{\kappa\lambda}$) by substitutions at site *i* in branch *b*:

$$\Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta}) \quad = \quad \sum_{\kappa,\lambda} \frac{\Delta_{\kappa,\lambda} P(\mathcal{A}_i, v_{bL} = \kappa, v_{bR} = \lambda | \hat{T}, \hat{\Theta})}{P(\mathcal{A}_i | \hat{T}, \hat{\Theta}, \theta_\alpha)} \tag{23}$$

Vector of the mean characteristic changes by substitutions at each site:

$$\boldsymbol{\Delta}_i \quad \equiv \quad (\ldots\ ,\ \Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta}) - \frac{\sum_b \Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta})}{\sum_b 1}\ ,\ \ldots)' \tag{24}$$

Correlation coefficient matrix of the feature vectors between sites:

$$(C)_{ij} \quad \equiv \quad r_{\Delta_i \Delta_j} = \frac{(\boldsymbol{\Delta}_i, \boldsymbol{\Delta}_j)}{\|\boldsymbol{\Delta}_i\| \|\boldsymbol{\Delta}_j\|} \tag{25}$$

Partial correlation coefficient matrix of the feature vectors between sites:

$$\mathcal{C}_{ij} \equiv r_{\Pi_{\perp\{\Delta_{k\neq i,j}\}}\Delta_i \Pi_{\perp\{\Delta_{k\neq i,j}\}}\Delta_j} \equiv \frac{(\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_i\ ,\ \Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_j)}{\|\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_i\|\ \|\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_j\|} = \frac{-\,(C^{-1})_{ij}}{((C^{-1})_{ii}(C^{-1})_{jj})^{1/2}} \tag{26}$$

Characteristic changes accompanied by substitutions whose correlation indicates coevolution between sites

1. Occurrence of amino acid substitutions: $\Delta^{s}_{\kappa,\lambda} \equiv 1 - \delta_{a_\kappa,a_\lambda}$ where a_κ is the type of amino acid corresponding to codon κ.

Characteristic changes accompanied by substitutions whose correlation indicates coevolution between sites

1. Occurrence of amino acid substitutions: $\Delta^{s}_{\kappa,\lambda} \equiv 1 - \delta_{a_\kappa,a_\lambda}$ where a_κ is the type of amino acid corresponding to codon κ.

   Phylogenetic bias: $\Delta^{s}_{ib} \sim 1 - \exp(-\mu_i \hat{t}_b) \propto \mu_i \overline{\Delta^{s}_{\bullet b}} \quad \Longrightarrow \quad C_{ij} \gg 0$

   Most of the phylogenetic bias can be removed from C_{ij} by a linear regression on $\boldsymbol{\Delta}^{s}_{k}$, $(k \neq i, j)$, and is not included in $\mathcal{C}_{ij}$.

Characteristic changes accompanied by substitutions whose correlation indicates coevolution between sites

1. Occurrence of amino acid substitutions: $\Delta^{s}_{\kappa,\lambda} \equiv 1 - \delta_{a_\kappa,a_\lambda}$ where a_κ is the type of amino acid corresponding to codon κ.

   Phylogenetic bias: $\Delta^{s}_{ib} \sim 1 - \exp(-\mu_i \hat{t}_b) \propto \mu_i \overline{\Delta^{s}_{\bullet b}} \quad \Longrightarrow \quad C_{ij} \gg 0$

   Most of the phylogenetic bias can be removed from C_{ij} by a linear regression on $\boldsymbol{\Delta}^{s}_{k}$, $(k \neq i, j)$, and is not included in $\mathcal{C}_{ij}$.

2. Change of side chain volume: $\Delta^{v}_{\kappa,\lambda} \equiv \text{side\_chain\_volume}_{a_\lambda} - \text{side\_chain\_volume}_{a_\kappa}$

3. Change of side chain charge: $\Delta^{c}_{\kappa,\lambda} \equiv \text{side\_chain\_charge}_{a_\lambda} - \text{side\_chain\_charge}_{a_\kappa}$

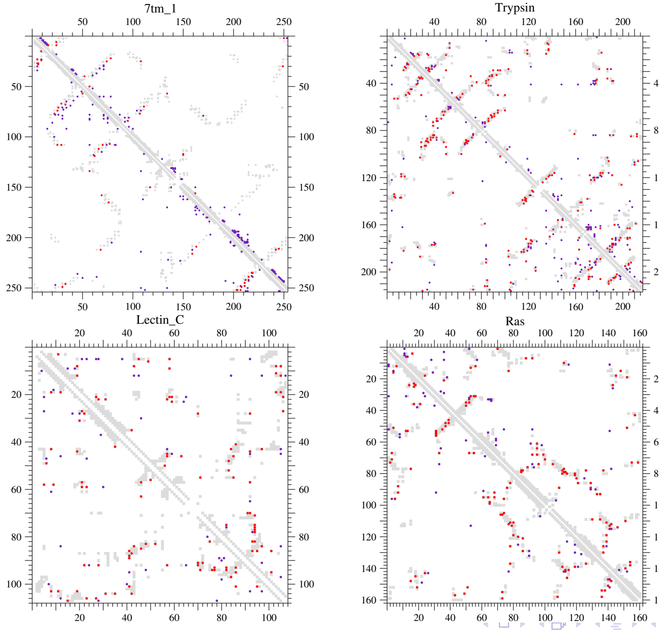
4. Change of hydrogen-bonding capability:

   $\Delta^{hb}_{\kappa,\lambda} \equiv$

   $\text{acceptor\_capability}_{a_\lambda} - \text{acceptor\_capability}_{a_\kappa} + \text{donor\_capability}_{a_\lambda} - \text{donor\_capability}_{a_\kappa}$

5. $\ldots$

Coevolving (lower) versus DI (upper) residue pairs (≤ 5 Å; TP, FP)



Deep learning method

A ultra-deep neural network has been also developed and applied for post-processing. (Wang and Sun et al., 2017)

- One-dimensional and two-dimensional deep neural networks (DNN)
- Each DNN is convolutional residual neural networks; 2D DNN performs 2D convolutional transformations, with respect to residue position, of pairwise features such as coevolutional information calculated by the direct coupling method.

This DNN performs very well, if the output from the direct coupling method is used, indicating that the connection of neurons must be improved to extract direct site couplings from one-dimensional sequence information only.

3. In conclusion

I have introduced my research history by focusing particularly on two topics, which some of you may be interested in.

Nowdays, many people who have the background of information science are studying in the filed of bioinformatics. However, there are subfields in which few information scientists study. They are molecular evolution and protein structure. Thus, teaching/discussing about

- statistical methods/algorithms for estimating phylogenetic tree
- various methods for predicting contacting residue pairs in protein structure

may be fruitful for us.