Inference of Co-Evolving Site Pairs
An Excellent Predictor of Contact Residue Pairs in Protein 3D structures

Sanzo Miyazawa

JSBi2013, 10/29-10/31/2013

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.

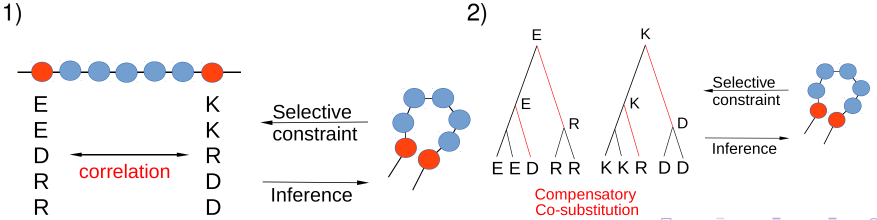
Background

- 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.



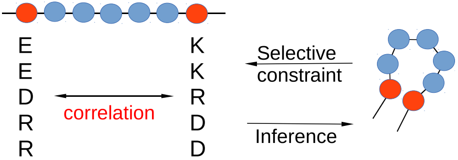
Two approaches to infer co-evolving site pairs

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

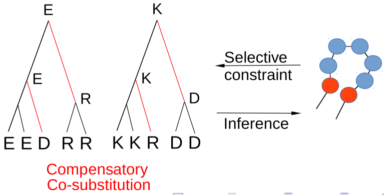
   Recently remarkable prediction accuracy of contact residue pairs was achieved by extracting essential correlations of amino acid types between residue positions by Bayesian graphical models and with a direct information (DI) score.

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

1)



2)



Two approaches to infer co-evolving site pairs

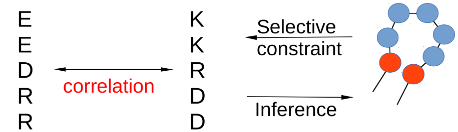
1. From the equilibrium distribution of amino acid sequences;
   ex. Direct Information (DI) score based on an inverse Potts problem.

   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.

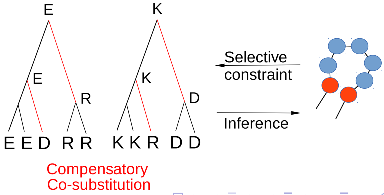
2. From the dynamic process of amino acid substitutions: The present approach.

   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)



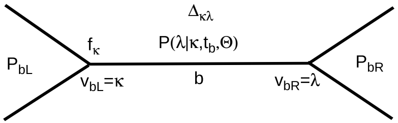
Methods: Mean of characteristic changes ($\Delta_{\kappa\lambda}$) at site i in branch b

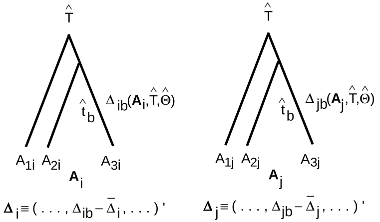
Likelihood of an alignment $\mathcal{A}$ in a tree T under a codon substitution model Θ : $P(\mathcal{A}|T,\Theta)$

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.





$$P(\mathcal{A}_i, v_{bL}=\kappa, v_{bR}=\lambda|T,\Theta) \equiv P_{bL}(\mathcal{A}_i|v_{bL}=\kappa, T,\Theta) f_\kappa P(\lambda|\kappa, t_b,\Theta) P_{bR}(\mathcal{A}_i|v_{bR}=\lambda, T,\Theta) \quad (1)$$

$$P(\mathcal{A}_i|T,\Theta) = \sum_\kappa \sum_\lambda P(\mathcal{A}_i, v_{bL}=\kappa, v_{bR}=\lambda|T,\Theta) \quad (2)$$

$$(\hat{T},\hat{\Theta}) = \arg\max_{T,\Theta} \prod_i P(\mathcal{A}_i|T,\Theta) \quad (3)$$

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*:

$$\Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta}) = \sum_{\kappa,\lambda} \frac{\Delta_{\kappa,\lambda} P(\mathcal{A}_i, v_{bL} = \kappa, v_{bR} = \lambda | \hat{T}, \hat{\Theta})}{P(\mathcal{A}_i | \hat{T}, \hat{\Theta}, \theta_\alpha)} \quad (4)$$

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

$$\boldsymbol{\Delta}_i \equiv (\ldots , \Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta}) - \frac{\sum_b \Delta_{ib}(\mathcal{A}_i, \hat{T}, \hat{\Theta})}{\sum_b 1} , \ldots)' \quad (5)$$

Correlation coefficient matrix of the feature vectors between sites:

$$(C)_{ij} \equiv r_{\Delta_i \Delta_j} = \frac{(\boldsymbol{\Delta}_i, \boldsymbol{\Delta}_j)}{\|\boldsymbol{\Delta}_i\| \|\boldsymbol{\Delta}_j\|} \quad (6)$$

Partial correlation coefficient matrix of the feature vectors between sites:

$$\mathcal{C}_{ij} \equiv r_{\Pi_{\perp\{\Delta_{k\neq i,j}\}}\Delta_i \Pi_{\perp\{\Delta_{k\neq i,j}\}}\Delta_j} \equiv \frac{(\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_i \ , \ \Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_j)}{\|\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_i\| \ \|\Pi_{\perp\{\Delta_{k\neq i,j}\}}\boldsymbol{\Delta}_j\|} = \frac{-(C^{-1})_{ij}}{((C^{-1})_{ii}(C^{-1})_{jj})^{1/2}} \quad (7)$$

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: $\Delta^s_{ib} \sim 1 - \exp(-\mu_i \hat{t}_b) \propto \mu_i \overline{\Delta^s_{\bullet b}} \implies C_{ij} \gg 0$

   Most of the phylogenetic bias can be removed from C_{ij} by a linear regression on $\mathbf{\Delta}^s_k$, $(k \neq i, j)$, 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: $\Delta^s_{ib} \sim 1 - \exp(-\mu_i \hat{t}_b) \propto \mu_i \overline{\Delta^s_{\bullet b}} \implies C_{ij} \gg 0$

   Most of the phylogenetic bias can be removed from C_{ij} by a linear regression on $\mathbf{\Delta}^s_k$, $(k \neq i, j)$, 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:

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

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

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.
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. Change of hydrophobicity: $\Delta^{h}_{\kappa,\lambda} \equiv e_{a_\lambda r} - e_{a_\kappa r}$,
   where $e_{a_\kappa r}$ is the mean contact energy of amino acid a_κ.
6. Changes of β propensity: $\Delta^{\beta}_{\kappa,\lambda} \equiv \beta\text{\_sheet\_propensity}_{a_\lambda} - \beta\text{\_sheet\_propensity}_{a_\kappa}$
7. Changes of turn propensity: $\Delta^{t}_{\kappa,\lambda} \equiv \text{turn\_propensity}_{a_\lambda} - \text{turn\_propensity}_{a_\kappa}$
8. Change of aromatic interactions: $\Delta^{ar}_{\kappa,\lambda} \equiv \delta_{\text{aromatic\_side\_chains}, a_\lambda} - \delta_{\text{aromatic\_side\_chains}, a_\kappa}$
9. Change of branched side-chains:
   $\Delta^{br}_{\kappa,\lambda} \equiv \delta_{\text{aliphatic\_branched\_side\_chains}, a_\lambda} - \delta_{\text{aliphatic\_branched\_side\_chains}, a_\kappa}$
10. Change of cross-link capability: $\Delta^{cl}_{\kappa,\lambda} \equiv \delta_{\text{cross\_link}, a_\lambda} - \delta_{\text{cross\_link}, a_\kappa}$
11. Change of inonic side-chains: $\Delta^{ion}_{\kappa,\lambda} \equiv \delta_{\text{inonic\_side\_chains}, a_\lambda} - \delta_{\text{inonic\_side\_chains}, a_\kappa}$

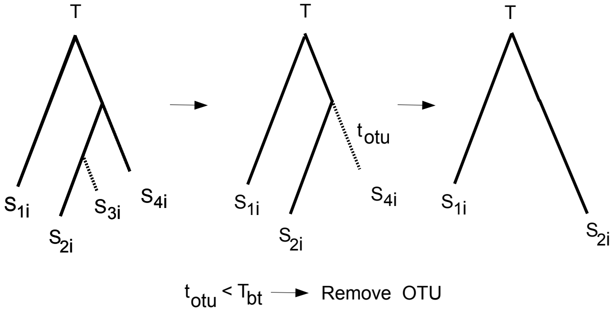
Protein families used

Pfam ID*	Seed**		Full§		Target protein domain		Fold	#sites
	#seqs	Length	#seqs	Length	Uniprot ID§§	PDB ID†	type	/Length††
Trans_reg_C	362	114	35180	269	OMPR_ECOLI/156-232	1ODD-A:156-232	α	76/77
CH	202	249	5756	650	SPTB2_HUMAN/176-278	1BKR-A:5-107	α	101/103
7tm_1	64	434	26656	2354	OPSD_BOVIN/54-306	1GZM-A:54-306	α (tm)‡	248/253
SH3_1	61	56	8993	210	YES_HUMAN/97-144	2HDA-A:97-144	β	48/48
Cadherin	57	129	18808	494	CADH1_HUMAN/267-366	2O72-A:113-212	β	91/100
Trypsin	71	348	14720	1356	TRY2_RAT/24-239	3TGI-E:16-238	β	212/216
Kunitz_BPTI	151	81	3090	209	BPT1_BOVIN/39-91	5PTI-A:4-56	$\alpha + \beta$	53/53
KH_1	399	104	11484	280	PCBP1_HUMAN/281-343	1WVN-A:7-69	$\alpha + \beta$	57/63
RRM_1	79	79	31837	580	ELAV4_HUMAN/48-118	1G2E-A:41-111	$\alpha + \beta$	70/71
FKBP_C	174	247	11034	845	O45418_CAEEL/26-118	1R9H-A:26-118	$\alpha + \beta$	92/93
Lectin_C	44	136	6530	801	CD209_HUMAN/273-379	1SL5-A:273-379	$\alpha + \beta$	103/107
Thioredoxin	50	123	16281	609	THIO_ALIAC/1-103	1RQM-A:1-103	α/β	99/103
Response_reg	57	157	103232	804	CHEY_ECOLI/8-121	1E6K-A:8-121	α/β	110/114
RNase_H	65	246	13801	574	RNH_ECOLI/2-142	1F21-A:3-142	α/β	128/140
Ras	61	229	13525	1461	RASH_HUMAN/5-165	5P21-A:5-165	α/β	159/161

* Pfam release 26.0 (November 2011) was used.
** The number of sequences and the length of alignment included in the Pfam seed alignment.
§ The number of sequences and the length of alignment included in the Pfam full alignment.
§§ Target protein member in the Pfam family.
† A protein structure corresponding to the target protein domain.
‡ Transmembrane α.

OTUs with short branches in Pfam full alignments are removed:

- Including closely-related sequences requires computationally intensive calculation, although it is not much informative.
- The subsets of a full alignment and their NJ trees are made by removing OTUs that are connected to the parent nodes with branches shorter than a certain threshold (T_{bt}), although seed sequences and a target protein are not removed.



Only ungapped positions in the target proteins are extracted from the alignments and used.

Results : Coevolution score ρ_{ij} for site pair (i, j)

Direct correlation between sites for concurrent substitutions must be positive:

$$\rho^{s}_{ij} \equiv \max\left(\mathcal{C}^{s}_{ij},\ 0 \right) \tag{8}$$

Results : Coevolution score ρ_{ij} for site pair (i, j)

Direct correlation between sites for concurrent substitutions must be positive:

$$\rho^{s}_{ij} \quad \equiv \quad \max\,(\,\mathcal{C}^{s}_{ij},\ 0\,) \tag{8}$$

For other characteristic variables the condition of concurrent substitutions between sites is a premise:

$$\rho^{x}_{ij} \quad \equiv \quad \mathrm{sgn}\,\mathcal{C}^{x}_{ij}\,(|\rho^{s}_{ij}\mathcal{C}^{x}_{ij}|)^{1/2} \qquad \text{for} \quad x \in \{v, c, hb, h, \ldots\} \tag{9}$$

Results : Coevolution score ρ_{ij} for site pair (i, j)

Direct correlation between sites for concurrent substitutions must be positive:

$$\rho_{ij}^{s} \equiv \max\left(\mathcal{C}_{ij}^{s},\ 0 \right) \tag{8}$$

For other characteristic variables the condition of concurrent substitutions between sites is a premise:

$$\rho_{ij}^{x} \equiv \operatorname{sgn} \mathcal{C}_{ij}^{x} \left(|\rho_{ij}^{s}\mathcal{C}_{ij}^{x}|\right)^{1/2} \quad \text{for} \quad x \in \{v, c, hb, h, \ldots\} \tag{9}$$

Direct correlations of volume, charge, and H-B capability changes for compensatory substitutions must be negative:

$$\max\left(-\rho_{ij}^{v},\ 0 \right),\ \max\left(-\rho_{ij}^{c},\ 0 \right),\ \max\left(-\rho_{ij}^{hb},\ 0 \right)$$

Coevolution score based on each characteristic change

Characteristic variable	$\rho^x_{ij} \geq \rho^s_{ij} \geq r_t$ * TP§	FP§	PPV†	$\rho^x_{ij} \leq -\rho^s_{ij} \leq -r_t$ * TP§	FP§	PPV†
	over all protein families					
Substitution	687	642	0.52			
Volume	18	20	0.47	73	10	**0.88**‡
Charge	6	8	0.43	134	54	**0.71**‡
Hydrogen bond	4	11	0.27	125	51	**0.71**‡
Hydrophobicity	23	13	**0.64**‡	23	16	**0.59**‡
α propensity	14	20	0.41	9	10	0.47
β propensity	24	17	**0.59**‡	30	14	**0.68**‡
Turn propensity	21	18	**0.54**‡	17	15	**0.53**‡
Aromatic interaction	30	10	**0.75**‡	16	14	**0.53**‡
Branched side-chain	26	16	**0.62**‡	20	8	**0.71**‡
Cross link	23	12	**0.66**‡	5	9	0.36
Ionic side-chain	27	15	**0.64**‡	14	18	0.44

* The E-value $E_t = 0.001$ (the P-value $P_t = E_t / n_{\text{pairs}}$).

§ Contacts are defined as residue pairs within 5 Å and separated by more than 5 residues.

† PPV $=$ TP$/($TP $+$ FP$)$; TP and FP are the numbers of true and false positives.

Results : Coevolution score ρ_{ij} for site pair (i, j)

Direct correlation between sites for concurrent substitutions must be positive:

$$\rho_{ij}^{s} \quad \equiv \quad \max\left(\ \mathcal{C}_{ij}^{s},\ 0\ \right) \tag{8}$$

For other characteristic variables the condition of concurrent substitutions between sites are a premise:

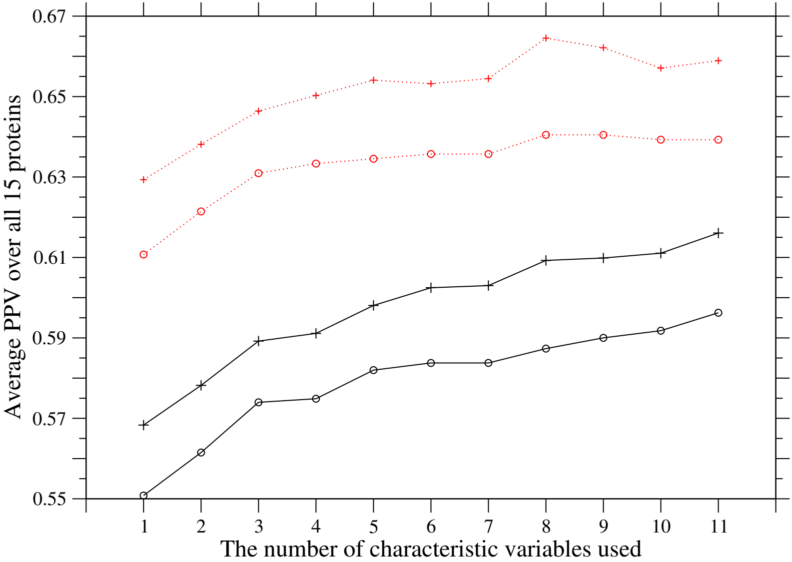
$$\rho_{ij}^{x} \quad \equiv \quad \operatorname{sgn} \mathcal{C}_{ij}^{x}\ (|\rho_{ij}^{s}\mathcal{C}_{ij}^{x}|)^{1/2} \qquad \text{for} \quad x \in \{v, c, hb, h, \ldots\} \tag{9}$$

Overall coevolution score ρ_{ij} for site pair (i, j) is defined as:

$$\begin{aligned} \rho_{ij} \quad \equiv \quad & \max[\rho_{ij}^{s}, \max(-\rho_{ij}^{v}, 0), \max(-\rho_{ij}^{c}, 0), \max(-\rho_{ij}^{hb}, 0), \\ & |\rho_{ij}^{h}|, |\rho_{ij}^{\beta}|, |\rho_{ij}^{t}|, |\rho_{ij}^{ar}|, |\rho_{ij}^{br}|, \max(\rho_{ij}^{cl}, 0), \max(\rho_{ij}^{ion}, 0)] \end{aligned} \tag{10}$$

Basically, site pairs are selected for contacts in the decreasing order of the overall coevolution score ρ_{ij}.

Dependences of PPV on the number of characteristic variables used



PPV = TP / (TP + FP).

Contacts are defined as residue pairs within 5 Å and separated by more than 5 residues.

Accuracy of contact prediction based on the overall coevolution score I

Pfam ID (PDB ID)	N_{otu} * (t_{bt})	#contacts /#sites**	TP+FP †	PPV†† C_{ij}^{s} ‡		PPV†† $\mathcal{C}_{ij}^{s}$ ‡‡	PPV†† ρ_{ij} §	DI§§
Trans_reg_C	7720	111/76	27	0.222	$\ll$	0.630	**0.667**	0.556
(1ODD-A:156-232)	(0.12)	1.5	37	0.189	$\ll$	0.541	**0.622**	0.432
CH	2960	172/101	43	0.047	$\ll$	0.395	0.465	**0.488**
(1BKR-A:5-107)	(0.01)	1.7	57	0.053	$\ll$	0.439	**0.491**	0.439
7tm_1	6302	372/248	93	0.011	$\ll$	0.333	**0.344**	0.194
(1GZM-A:54-306)	(0.10)	1.5	124	0.008	$\ll$	0.290	**0.306**	0.169
SH3_1	4160	89/48	22	0.227	$\ll$	0.727	**0.682**	0.636
(2HDA-A:97-144)	(0.01)	1.9	29	0.241	$\ll$	0.621	**0.655**	0.552
Cadherin	7617	220/91	55	0.291	$\ll$	0.764	**0.836**	0.818
(2O72-A:113-212)	(0.06)	2.4	73	0.274	$\ll$	0.726	**0.767**	0.753
Trypsin	6688	636/212	159	0.396	$\ll$	0.642	**0.673**	0.591
(3TGI-E:16-238)	(0.10)	3.0	212	0.344	$\ll$	0.575	**0.613**	0.533

** Contacts are defined as residue pairs within 5 Åand separated by more than 5 residues.

‡ The prediction with the correlation coefficient of substitution number vector.

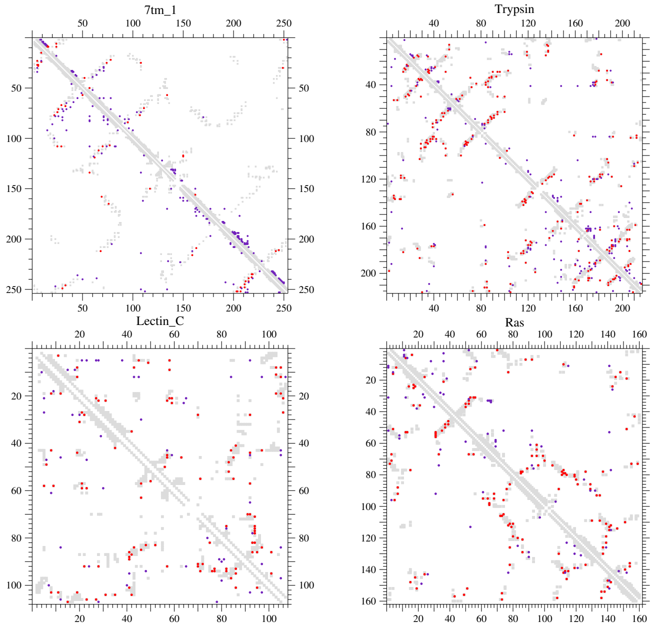
‡‡ The prediction with the partial correlation coefficient of substitution number vector.

§§ The prediction with the Direct Information (DI); a conservation filter is used (Marks et al., *PLoS One*, **6**, e28766, 1911).

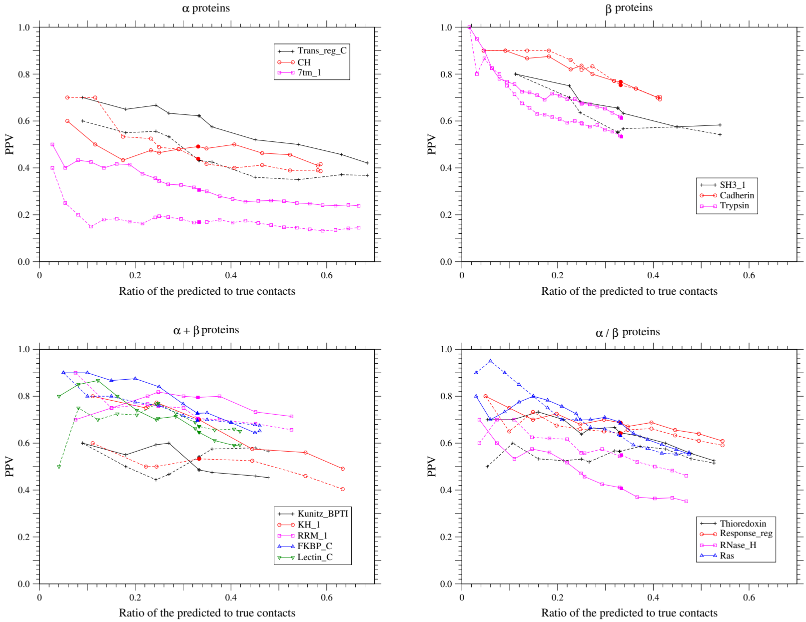
Accuracy of contact prediction based on the overall coevolution score II

Pfam ID (PDB ID)	N_{otu} * (t_{bt})	#contacts /#sites**	TP+FP †	PPV†† C^{s}_{ij} ‡		$\mathcal{C}^{s}_{ij}$ ‡‡	PPV†† ρ_{ij} §	DI §§
Kunitz_BPTI	2130	111/53	27	0.259	$\ll$	0.593	**0.593**	0.444
(5PTI-A:4-56)	(0.01)	2.1	37	0.216	$\ll$	0.514	0.486	**0.541**
KH_1	5114	90/57	22	0.455	$\ll$	0.682	**0.773**	0.500
(1WVN-A:7-69)	(0.01)	1.6	30	0.367	$\ll$	0.600	**0.700**	0.533
RRM_1	7684	133/70	33	0.273	$\ll$	0.758	**0.818**	0.758
(1G2E-A:41-111)	(0.15)	1.9	44	0.295	$\ll$	0.795	**0.795**	0.705
FKBP_C	5695	200/92	50	0.220	$\ll$	0.780	**0.840**	0.760
(1R9H-A:26-118)	(0.01)	2.2	66	0.197	$\ll$	0.667	**0.727**	0.697
Lectin_C	4479	246/103	61	0.197	$\ll$	0.656	0.705	**0.770**
(1SL5-A:273-379)	(0.01)	2.4	82	0.171	$\ll$	0.585	0.646	**0.671**
Thioredoxin	7483	188/99	47	0.213	$\ll$	0.660	**0.638**	0.532
(1RQM-A:1-103)	(0.06)	1.9	62	0.177	$\ll$	0.581	**0.645**	0.565
Response_reg	7613	202/110	50	0.000	$\ll$	0.680	**0.680**	0.660
(1E6K-A:8-121)	(0.46)	1.8	67	0.015	$\ll$	0.657	**0.687**	0.642
RNase_H	4782	273/128	68	0.162	$\ll$	0.456	0.471	**0.559**
(1F21-A:3-142)	(0.01)	2.1	91	0.132	$\ll$	0.407	0.407	**0.549**
Ras	6390	335/159	83	0.229	$\ll$	0.699	0.699	0.699
(5P21-A:5-165)	(0.02)	2.1	111	0.207	$\ll$	0.640	**0.685**	0.631

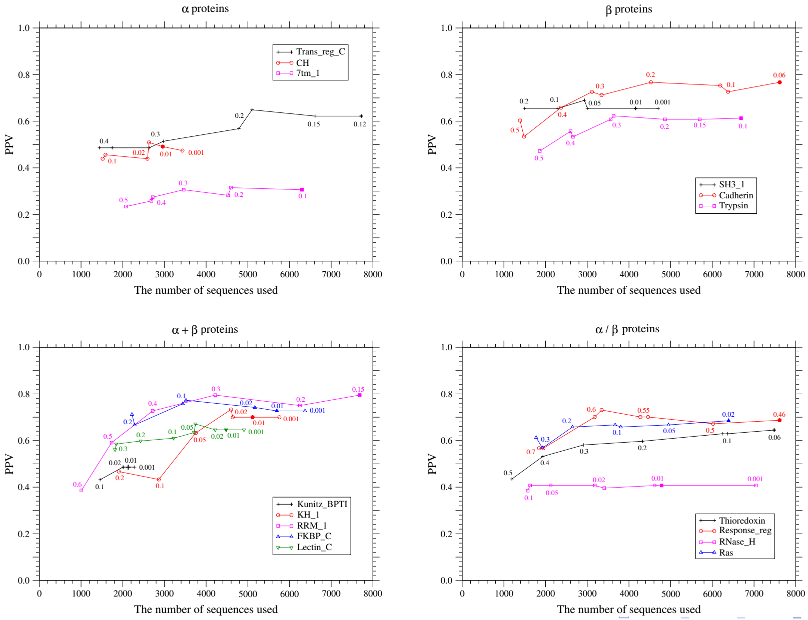
Coevolving (lower) versus DI (upper) residue pairs ($\leq$ 5 Å; TP, FP)



Dependences of PPV on the number of predicted contacts; solid: coevolution, dotted: DI



Dependences of PPV on the number of sequences used



Discussion I

- Prediction accuracy of residue contacts appears to be excellent enough for one to achieve reasonable 3D structure prediction.

  Besides, this excellent accuracy indicates that compensatory substitutions are significant in protein evolution.

  Limitations in prediction accuracy:

  - Statistical noise due to an insufficient number and insufficient diversities of sequences, incorrect matches in a multiple sequence alignment, and an incorrect phylogenetic tree. It is not practical and not cost-effective to optimize a phylogenetic tree, because of computationally intensive calculations and insignificant improvements.
  - Interactions between proteins, which are not taken into account here, in a protein complex.
  - Structural variance in homologous proteins.

Discussion II

- The present method based on co-substitution between sites could better detect non-specific interactions between closely-located residues but less detect residue-residue interactions maintaining secondary structures than the DI method based on the joint distributions of amino acid types between sites.
  Ex. Interactions within and between α helices in a membrane protein, 7tm_1.
- A present model can be regarded as a Gaussian graphical model, in which an undirected graph is assumed for site dependence.
  Because physical interactions between sites are not unidirectional, the Gaussian graphical model may be more appropriate for contact prediction than Bayesian graphical models, in which an acyclic directed graph is assumed.

Reference: PLoS One, 8, e54252/pp. 1-20, 2013.

Abstract

Residue-residue interactions that fold a protein into a unique three-dimensional structure and make it play a specific function impose structural and functional constraints in varying degrees on each residue site. Selective constraints on residue sites are recorded in amino acid orders in homologous sequences and also in the evolutionary trace of amino acid substitutions. A challenge is to extract direct dependences between residue sites by removing phylogenetic correlations and indirect dependences through other residues within a protein or even through other molecules. Rapid growth of protein families with unknown folds requires an accurate *de novo* prediction method for protein structure. Recent attempts of disentangling direct from indirect dependences of amino acid types between residue positions in multiple sequence alignments have revealed that inferred residue-residue proximities can be sufficient information to predict a protein fold without the use of known three-dimensional structures. Here, we propose an alternative method of inferring coevolving site pairs from concurrent and compensatory substitutions between sites in each branch of a phylogenetic tree. First, branch lengths of the Pfam phylogenetic tree are optimized as well as other parameters by maximizing a likelihood of the tree in a mechanistic codon substitution model. Substitution probability and physico-chemical changes (volume, charge, hydrogen-bonding capability and others) accompanied by substitutions at each site in each branch of a phylogenetic tree are estimated with the likelihood of each substitution, and their direct correlations between sites are used to detect concurrent and compensatory substitutions. In order to extract direct dependences between sites, partial correlation coefficients of the characteristic changes along branches between sites, in which linear multiple dependences on feature vectors at other sites are removed, are calculated and used to rank coevolving site pairs. Accuracy of contact prediction based on the present coevolution score is comparable to that achieved by a maximum entropy model of protein sequences for 15 protein families taken from the Pfam release 26.0. Besides, this excellent accuracy indicates that compensatory substitutions are significant in protein evolution.

Reference: PLoS One, 8, e54252/pp. 1-20, 2013.

A mechanistic codon substitution model: PLoS One 6:e17244 (2011); PLoS One 6:e28892 (2011)

- Codon substitution model: $P(\lambda|\kappa, t_b, \Theta, \theta_\alpha) \equiv (\exp Rt)_{\kappa\lambda}$
- Substitution Rate: $R_{\mu\nu} = C_{\text{onst}}\, M_{\mu\nu} \frac{f_\nu}{f_\nu^{\text{mut}}} e^{w_{\mu\nu}}$ for $\mu \neq \nu$

where

$M_{\mu\nu}$	*is the mutation rate from codon μ to ν,*
f_ν^{mut}	*is the equilibrium frequency of codon ν in nucleotide mutations,*
f_ν	*is the equilibrium codon frequency,*
$\frac{f_\nu}{f_\nu^{mut}} e^{w_{\mu\nu}}$	*is the average rate of fixation, and*
$w_{\mu\nu}$	*is the selective constraints for mutations from μ to ν.*

- Codon mutation rates $M_{\mu\nu}$ are approximated by 9 parameters, assuming nucleotide mutations occur independently at each position:

$m_{tc\|ag}/m_{[tc][ag]}$, $m_{ag}/m_{tc\|ag}$, $m_{ta}/m_{[tc][ag]}$, $m_{tg}/m_{[tc][ag]}$, $m_{ca}/m_{[tc][ag]}$	*the ratios of nucleotide mutation rates*
m	*the relative ratio of multiple nucleotide changes*
f_a^{mut}, f_c^{mut}, *and* f_g^{mut}	*the equilibrium nucleotide frequencies in nucleotide mutat*

- Selective constraints $w_{\mu\nu}$: $w_{\mu\nu} = \beta w_{\mu\nu}^{\text{LG}} + w_0$, where β and w_0 are parameters and $w_{\mu\nu}^{\text{LG}}$ was one estimated from observed substitution data matrices (LG).
- The variation of selective constraints $w_{\mu\nu}$ is approximated by a discrete gamma distribution of shape parameter α with four categories.
- Codon frequencies f_ν are estimated from amino acid sequences with the assumption of equal codon usage.
- Other 12 parameters estimated for each set of Pfam seed sequences are used.
- Tree topologies inferred by the neighbor joining (NJ) method are assumed as true ones.

Correlation coefficients of concurrent substitutions between sites

Pfam ID	T_{bt}*	n_{otu}*	$C_{ij}^s \geq r_t$**		r_t** $> C_{ij}^s > 0$		$0 > C_{ij}^s > -r_t$**		$-r_t$** $\geq C_{ij}^s$	
			TP:FP §‡	PPV §§	TP:FP §‡	PPV §§	TP:FP §‡	PPV §§	TP:FP §‡	PPV §§
Trans_reg_C	0.12	7720	102:2282	0.04	1:30	0.03	0:0	–	0:0	–
CH	0.01	2960	167:4226	0.04	2:73	0.03	0:2	0.0	0:0	–
7tm_1	0.1	6302	358:28576	0.01	0:0	–	0:0	–	0:0	–
SH3_1	0.01	4160	74:674	0.10	7:60	0.10	0:5	0.0	0:0	–
Cadherin	0.06	7617	214:3333	0.06	1:46	0.02	0:7	0.0	0:0	–
Trypsin	0.1	6688	617:20312	0.03	0:0	–	0:0	–	0:0	–
Kunitz_BPTI	0.01	2130	86:799	0.10	11:48	0.19	0:2	0.0	0:0	–
KH_1	0.01	5114	78:1116	0.07	1:41	0.02	0:4	0.0	0:0	–
RRM_1	0.15	7684	119:1839	0.06	0:0	–	0:0	–	0:0	–
FKBP_C	0.01	5695	199:3445	0.05	0:10	0.0	0:1	0.0	0:0	–
Lectin_C	0.01	4479	234:4319	0.05	1:19	0.05	0:0	–	0:0	–
Thioredoxin	0.06	7483	188:4180	0.04	0:3	0.0	0:0	–	0:0	–
Response_reg	0.46	7613	202:5266	0.04	0:1	0.0	0:0	–	0:0	–
RNase_H	0.01	4782	271:7152	0.04	0:5	0.0	0:0	–	0:0	–
Ras	0.02	6390	329:11304	0.03	0:0	–	0:0	–	0:0	–

* OTUs connected to their parent nodes with branches shorter than this threshold value are removed from each Pfam full alignment.

** The E-value $E_t = 0.001$ (the P-value $P_t = E_t / n_{\text{pairs}}$) in the t-distribution of $\text{df} = (2n_{\text{otu}} - 3) - 2$.

‡ Contacts are defined as residue pairs within 5 Å and separated by more than 5 residues.

§§ $\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$; TP and FP are the numbers of true and false positives.

Partial correlation coefficients of concurrent substitutions between sites

Pfam ID	#contacts/#sites[‡]		$\mathcal{C}^s_{ij} \geq {r_t}^{**}$		${r_t}^{**} > \mathcal{C}^s_{ij} > 0$		$0 > \mathcal{C}^s_{ij} > -{r_t}^{**}$		$-{r_t}^{**} \geq \mathcal{C}^s_{ij}$	
			TP:FP[§‡]	PPV[§§]	TP:FP[§‡]	PPV[§§]	TP:FP[§‡]	PPV[§§]	TP:FP[§‡]	PPV[§§]
Trans_reg_C	103/75	1.4	32:57	0.36	59:1584	0.04	12:669	0.02	0:2	0.0
CH	169/100	1.7	16:17	0.48	125:2454	0.05	28:1828	0.02	0:2	0.0
7tm_1	366/247	1.5	36:84	0.30	263:15695	0.02	59:12787	0.005	0:10	0.0
SH3_1	81/46	1.8	24:17	0.59	46:516	0.08	11:206	0.05	0:0	–
Cadherin	215/90	2.4	40:8	0.83	132:1519	0.08	42:1857	0.02	1:2	0.33
Trypsin	617/210	2.9	115:75	0.61	383:11331	0.03	119:8899	0.01	0:7	0.0
Kunitz_BPTI	105/51	2.1	16:12	0.57	55:575	0.09	26:262	0.09	0:0	–
KH_1	79/55	1.4	19:15	0.56	50:707	0.07	10:438	0.02	0:1	0.0
RRM_1	119/68	1.8	45:36	0.56	63:1257	0.05	11:546	0.02	0:0	–
FKBP_C	199/91	2.2	66:51	0.56	103:2114	0.05	30:1288	0.02	0:3	0.0
Lectin_C	243/102	2.4	36:13	0.73	160:2401	0.06	39:1923	0.02	0:1	0.0
Thioredoxin	188/99	1.9	53:61	0.46	109:2677	0.04	26:1442	0.02	0:3	0.0
Response_reg	202/110	1.8	72:87	0.45	101:3182	0.03	28:1988	0.01	1:10	0.09
RNase_H	271/127	2.1	37:56	0.40	161:3700	0.04	72:3387	0.02	1:14	0.07
Ras	329/158	2.1	81:55	0.60	203:6472	0.03	44:4768	0.01	1:9	0.10

** The E-value $E_t = 0.001$ (the P-value $P_t = E_t / n_{\text{pairs}}$) in the t-distribution of df $= (2n_{\text{otu}} - 3) - 2$.

‡ Contacts are defined as residue pairs within 5 Å and separated by more than 5 residues. Both terminal sites are excluded from counting in this table.

§§ PPV $=$ TP$/($TP $+$ FP$)$; TP and FP are the numbers of true and false positives.

Effectiveness of partial correlation coefficients on contact prediction accuracy

Pfam ID*	#contacts /#sites**	TP + FP[§]	PPV($\equiv$ TP/(TP + FP)) C^s_{ij} [§§]	C^s_{ij} [†]	[††]	ρ_{ij} [‡]
Trans_reg_C	103/75	27	0.222	$\ll$ 0.630	$\simeq$ 0.630	$<$ 0.667
	1.4	37	0.189	$\ll$ 0.541	$<$ 0.595	$\simeq$ 0.595
CH	169/100	43	0.047	$\ll$ 0.395	$<$ 0.442	$<$ 0.535
	1.7	57	0.053	$\ll$ 0.439	$\simeq$ 0.439	$<$ 0.526
7tm_1	366/247	93	0.011	$\ll$ 0.333	0.290	$<$ 0.355
	1.5	124	0.008	$\ll$ 0.290	0.266	$<$ 0.315
SH3_1	81/46	22	0.227	$\ll$ 0.727	0.636	$<$ 0.682
	1.8	29	0.241	$\ll$ 0.621	0.586	$<$ 0.655
Cadherin	215/90	55	0.291	$\ll$ 0.764	0.691	$<$ 0.836
	2.4	73	0.274	$\ll$ 0.726	0.630	$<$ 0.767
Trypsin	617/210	159	0.396	$\ll$ 0.642	0.623	$<$ 0.673
	2.9	212	0.344	$\ll$ 0.575	0.571	$<$ 0.618
Kunitz_BPTI	105/51	27	0.259	$\ll$ 0.593	0.556	$<$ 0.630
	2.1	37	0.216	$\ll$ 0.514	0.459	$<$ 0.514
KH_1	79/55	22	0.455	$\ll$ 0.682	$<$ 0.773	0.727
	1.4	30	0.367	$\ll$ 0.600	$\simeq$ 0.600	$<$ 0.667
RRM_1	119/68	33	0.273	$\ll$ 0.758	$<$ 0.788	$<$ 0.818
	1.8	44	0.295	$\ll$ 0.795	0.750	$<$ 0.795
FKBP_C	199/91	50	0.220	$\ll$ 0.780	$<$ 0.880	0.840
	2.2	66	0.197	$\ll$ 0.667	$<$ 0.773	0.727
Lectin_C	243/102	61	0.197	$\ll$ 0.656	0.623	$<$ 0.705
	2.4	82	0.171	$\ll$ 0.585	0.537	$<$ 0.646
Thioredoxin	188/99	47	0.213	$\ll$ 0.660	$<$ 0.702	0.638
	1.9	62	0.177	$\ll$ 0.581	$<$ 0.661	0.645
Response_reg	202/110	50	0.000	$\ll$ 0.680	0.600	$<$ 0.680
	1.8	67	0.015	$\ll$ 0.657	0.522	$<$ 0.687
RNase_H	271/127	68	0.162	$\ll$ 0.456	$<$ 0.515	0.471
	2.1	91	0.132	$\ll$ 0.407	$<$ 0.440	0.407
Ras	329/158	83	0.229	$\ll$ 0.699	$\simeq$ 0.699	$<$ 0.735
	2.1	111	0.207	$\ll$ 0.640	$\simeq$ 0.640	$<$ 0.694

[††] In Eq. 10 for an overall coevolution score, $\rho^x_{ij} = \mathrm{sgn} C^x_{ij} (|\rho^s_{ij} C^x_{ij}|)^{1/2}$ with $x \neq s$ is supposed instead of Eq. 9; in other words, correlation coefficients are used instead of partial correlation coefficients for characteristic changes except co-substitution.

Contact prediction based on the overall coevolution score ρ_{ij}

Basically, sites pairs are selected for contacts in the decreasing order of the overall coevolution score ρ_{ij}.

Prediction rules in detail:

1. the coevolution scores of ρ_{ij}^{x} ($x \neq s$) are ignored for both terminal sites in multiple sequence alignments.
2. Also, if $\sum_j H(\rho_{ij} - r_t) > 15$, $\rho_{ij} \equiv \rho_{ij}^{s}$ will be used for site i, and
3. if $\sum_j H(\rho_{ij}^{s} - r_t) > 15$, $\rho_{ij} \equiv 0$ will be used and such a site will be excluded in contact prediction.

where r_t is the value corresponding to E-value $= 0.0001$ in the t-distribution.
Needless to say, the norm of any characteristic change vector is almost zero for invariant sites; $\|\boldsymbol{\Delta}_i\| \simeq 0$. Therefore, invariant sites are excluded in the present method for contact prediction.

Accuracy of contact prediction based on the overall coevolution score

Pfam ID*	#contacts /#sites**	*** TP + FP	PPV§§‡‡ DI §	PPV§§‡‡ ρ_{ij}	MDPNT†‡‡ DI §	MDPNT†‡‡ ρ_{ij}	MDTNP††‡‡ DI §	MDTNP††‡‡ ρ_{ij}
Trans_reg_C	111/76	27	0.556	**0.667**	1.30	**0.94**	4.20	**3.28**
	1.5	37	0.432	**0.622**	1.72	**1.16**	3.64	**2.82**
CH	172/101	43	**0.488**	0.465	**2.23**	2.55	4.59	**4.37**
	1.7	57	0.439	**0.491**	**2.12**	2.44	3.70	**3.30**
7tm_1	372/248	93	0.194	**0.344**	7.43	**5.31**	12.68	**7.71**
	1.5	124	0.169	**0.306**	7.30	**5.33**	12.18	**6.40**
SH3_1	89/48	22	0.636	**0.682**	0.83	**0.51**	**1.69**	2.34
	1.9	29	0.552	**0.655**	1.15	**0.62**	1.56	**1.51**
Cadherin	220/91	55	0.818	**0.836**	0.59	**0.25**	1.98	1.98
	2.4	73	0.753	**0.767**	0.64	**0.45**	1.60	1.60
Trypsin	636/212	159	0.591	**0.673**	1.75	**1.20**	3.26	**3.10**
	3.0	212	0.533	**0.613**	2.26	**1.65**	2.83	**1.94**
Kunitz_BPTI	111/53	27	0.444	**0.593**	1.40	**1.18**	2.31	**2.08**
	2.1	37	**0.541**	0.486	**1.13**	1.46	**1.86**	1.94
KH_1	90/57	22	0.500	**0.773**	0.99	**0.51**	**2.41**	3.29
	1.6	30	0.533	**0.700**	1.07	**0.56**	**2.16**	3.05
RRM_1	133/70	33	0.758	**0.818**	**0.52**	0.55	2.86	**2.36**
	1.9	44	0.705	**0.795**	0.83	**0.49**	2.49	**1.84**
FKBP_C	200/92	50	0.760	**0.840**	**0.53**	0.69	1.97	**1.85**
	2.2	66	0.697	**0.727**	0.94	**0.85**	1.66	**1.51**
Lectin_C	246/103	61	**0.770**	0.705	**0.80**	0.94	2.93	**2.67**
	2.4	82	**0.671**	0.646	1.19	**1.17**	2.54	**2.32**
Thioredoxin	188/99	47	0.532	**0.638**	0.98	**0.85**	3.43	**2.33**
	1.9	62	0.565	**0.645**	0.94	**0.91**	3.16	**1.86**
Response_reg	202/110	50	0.660	**0.680**	**0.86**	0.88	3.39	**3.06**
	1.8	67	0.642	**0.687**	1.01	**0.92**	2.54	**2.29**
RNase_H	273/128	68	**0.559**	0.471	**1.51**	1.53	**3.61**	5.44
	2.1	91	**0.549**	0.407	**1.55**	2.19	3.27	**3.07**
Ras	335/159	83	0.699	0.699	**0.94**	1.05	**2.98**	3.68
	2.1	111	0.631	**0.685**	**1.12**	1.45	**2.40**	2.51

** *Contacts are defined as residue pairs within 5 Å and separated by more than 5 residues.*

*** *Only predictions for* $TP + FP = \#contacts/4$ *and* $\#contacts/3$ *are listed.*

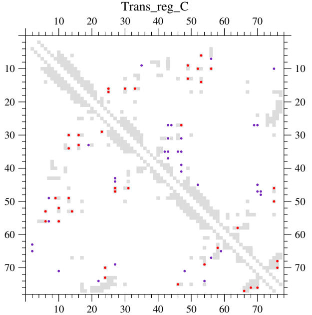
§§ *PPV stands for positive predictive value;* $PPV = TP/(TP + FP)$.

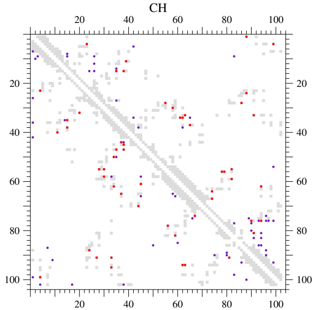
† *MDPNT stands for the mean Euclidean distance from predicted site pairs to the nearest true contact in the 2-dimensional sequence-position space.*

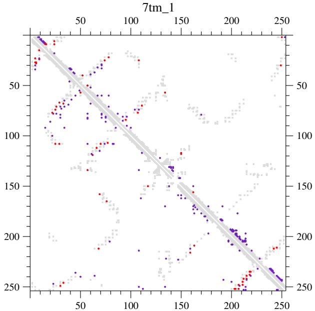
†† *MDTNP stands for the mean Euclidean distance from every true contact to the nearest predicted site pair in the 2-dimensional sequence-position space.*

‡‡ *DI means the prediction based on the direct information (DI) score to infer residue pair couplings from the joint distribution of amino acid types between sites in a multiple sequence alignment (Marks et al., 2011); only a conservation filter is used.*

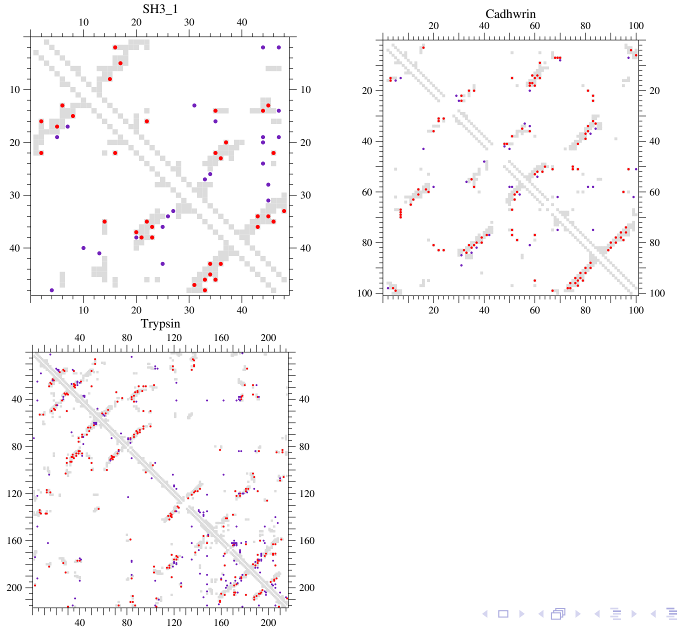
Coevolving (lower) versus DI (upper) residue pairs ($\leq$ 5 Å; TP, FP): α proteins



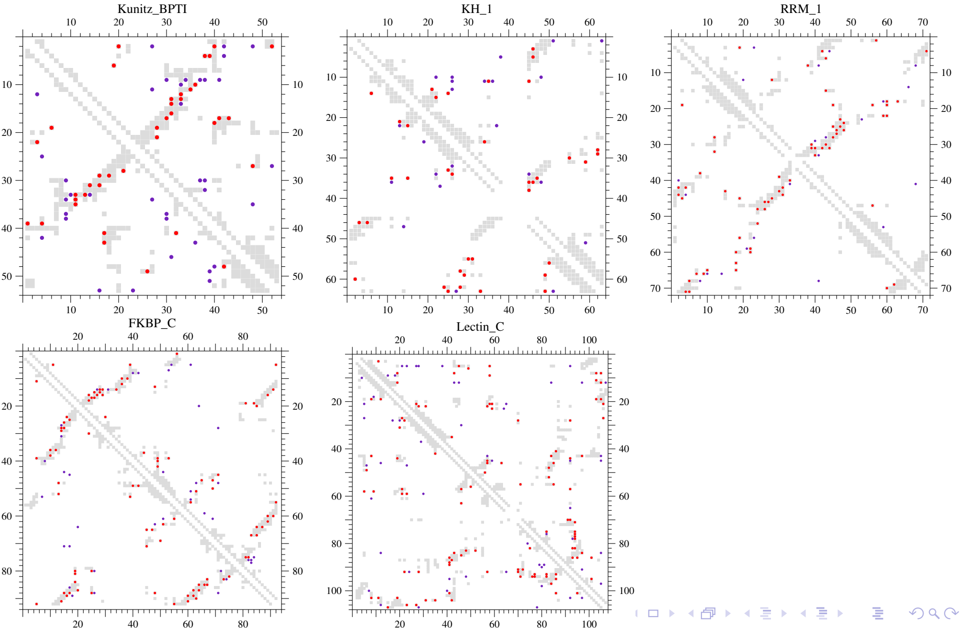




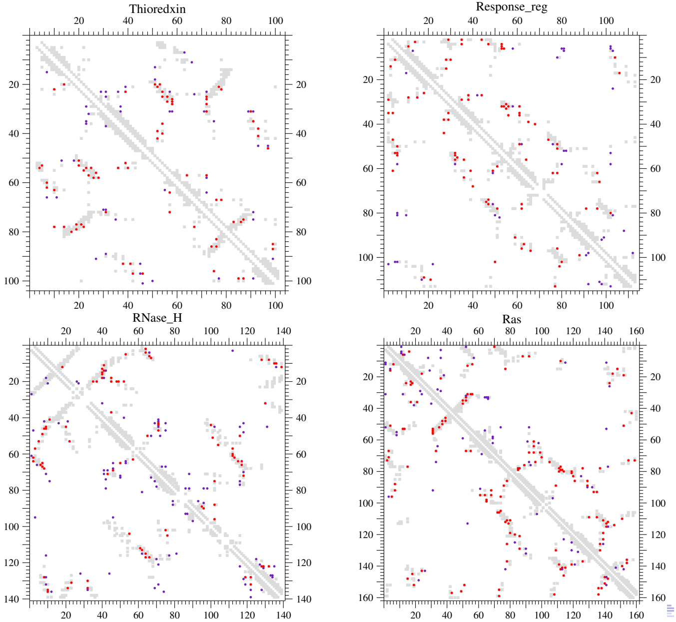
Coevolving (lower) versus DI (upper) residue pairs ($\leq$ 5 Å, TP, FP): β proteins



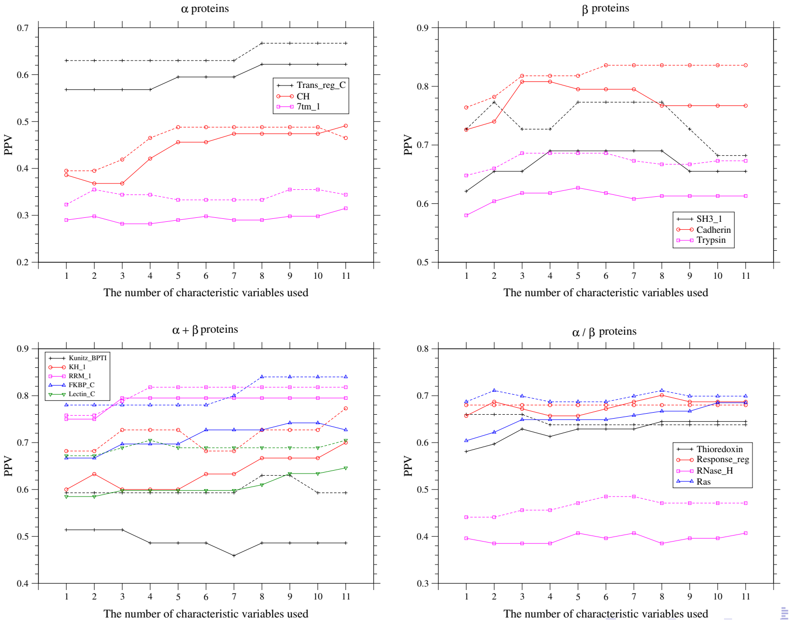
Coevolving (lower) vs. DI (upper) pairs ($\leq$ 5 Å, TP, FP): $\alpha + \beta$ proteins



Coevolving (lower) versus DI (upper) residue pairs ($\leq$ 5 Å, TP, FP): α/β proteins



Dependences of PPV on the number of characteristic variables used



Dependence of contact prediction accuracies on the topology of phylogenetic tree

Pfam ID*	#contacts /#sites**	TP + FP***	DI §§	PPV§ $\rho_{ij}^{§§§}$ Pfam tree †	FastTree2 ††	ExaML ††	Relative log-likelihood‡ Pfam tree †	FastTree2 ††	ExaML †††
Trans_reg_C	111/76	27	0.556	**0.667**	**0.667**		(−772541.8)	2768.9	
	1.5	37	0.432	**0.622**	0.595				
CH	172/101	43	**0.488**	0.465	0.419	0.395	(−246974.5)	1818.6	2988.1
	1.7	57	0.439	**0.491**	0.456	0.351			
7tm_1	372/248	93	0.194	0.344	**0.366**		(−1971205.1)	44545.9	
	1.5	124	0.169	**0.306**	**0.306**				
SH3_1	89/48	22	0.636	**0.682**	**0.682**	**0.682**	(−178181.5)	1214.8	2566.5
	1.9	29	0.552	0.655	0.586	**0.690**			
Cadherin	220/91	55	0.818	**0.836**	0.800		(−917754.4)	2891.1	
	2.4	73	0.753	**0.767**	0.740				
Trypsin	636/212	159	0.591	**0.673**	0.648		(−1843495.9)	5728.3	
	3.0	212	0.533	**0.613**	0.604				
Kunitz_BPTI	111/53	27	0.444	**0.593**	0.556	0.556	(−127989.5)	600.6	1731.1
	2.1	37	**0.541**	0.486	0.514	0.514			
KH_1	90/57	22	0.500	0.773	**0.818**		(−253902.4)	2428.0	
	1.6	30	0.533	**0.700**	**0.700**				
RRM_1	133/70	33	0.758	**0.818**	0.788		(−780196.4)	3056.8	
	1.9	44	0.705	**0.795**	0.773				
FKBP_C	200/92	50	0.760	**0.840**	0.800		(−455605.4)	3935.5	
	2.2	66	0.697	0.727	**0.773**				
Lectin_C	246/103	61	**0.770**	0.705	0.705		(−555599.9)	3073.6	
	2.4	82	**0.671**	0.646	0.610				
Thioredoxin	188/99	47	0.532	0.638	**0.660**		(−926791.5)	4137.4	
	1.9	62	0.565	**0.645**	**0.645**				
Response_reg	202/110	50	0.660	0.680	**0.700**		(−1654255.6)	2934.4	
	1.8	67	0.642	0.687	**0.716**				
RNase_H	273/128	68	**0.559**	0.471	0.456	0.485	(−364080.9)	4787.3	8280.3
	2.1	91	**0.549**	0.407	0.407	0.418			
Ras	335/159	83	0.699	0.699	**0.723**		(−932720.7)	9667.8	
	2.1	111	0.631	**0.685**	0.667				