Journal of
Data Mining in Genomics & Proteomics

Kondo and Miyazaki J Data Mining Genomics Proteomics 2015, 6:2
DOI: 10.4172/2153-0602.1000175

Research Article Open Access

Protein Functional Site Prediction Using a Conservative Grade and a Proximate Grade

Yosuke Kondo and Satoru Miyazaki*

Department of Medicinal and Life Science, Faculty of Pharmaceutical Sciences, Tokyo University of Science, Japan

Abstract

So far, in order to predict important sites of a protein, many computational methods have been developed. In the era of big-data, it is required for improvements and sophistication of existing methods by integrating sequence data in the structural data. In this paper, we aim at two things: improving sequence-based methods and developing a new method using both sequence and structural data. Therefore, we developed an originally modified evolutionary trace method, in which we defined conservative grades calculated from a given multiple sequence alignment and a proximate grade in order to evaluate predicted active sites from a viewpoint of protein-ion, protein-ligand, protein-nucleic acid, protein-protein interaction by use of three-dimensional structures. In other words, the proximate grade also can evaluate an amino acid residue. When we applied our method to translation elongation factor Tu/1A proteins, it showed that the conservative grades are evaluated accurately by the proximate grade. Consequently, our idea indicated two advantages. One is that we can take into account various cocrystal structures for evaluation. Another one is that, by calculating the fitness between the given conservative grade and the proximate grade, we can select the best conservative grade.

Keywords: Evolutionary trace; Three-dimensional structure; Elongation factor

Introduction

When a protein works, a specific site to bind an ion or a molecule may exist. Identification of binding sites is important to investigate how the protein works and binds ions or molecules. In order to identify such an important site, it is necessary to prepare a mutant type of the protein, whose amino acid residue is mutated into another one, and then a difference of binding affinity between the mutant type and the wild type is investigated. However, mutating amino acid residues one by one takes an amount of time and costs. Therefore, it is effective for developing a method to narrow down the amino acid residues.

For electing the candidate sites, there are many computational methods, which are based on (i) sequence, (ii) structure and (iii) sequence and structure [1-6]. Sequence-based methods usually assume that such an important site is conservative against mutation and therefore important sites and others should have been mutated in different patterns. In order to detect such patterns, various methods have been developed [7]. One of the sequence-based methods is a method based on Shannon entropy (SE) [8,9]. However, the SE-based method may have three problems. The first one is that the SE-based method, in which twenty standard amino acids are regarded as characters, does not consider properties of amino acids. Therefore, a method based on SE of residue properties [10] or a sum of pairs [11] was proposed. The second one is that the SE-based method does not consider a background distribution of amino acids. Therefore, other information-theoretical method such as relative entropy [12] or Jensen Shannon divergence [13] was proposed. The third one is that the SE-based method, in which a rate of an amino acid is calculated, cannot take into account which amino acid is included in a sequence. Therefore, some methods based on windowing [13], weighting [14] or phylogenetic analysis was proposed. One of the methods based on a phylogenetic tree is an evolutionary trace (ET) method [15], which has been extended as weighted ET (WET) [16], integer-valued ET (iv-ET) and real-valued ET (rv-ET) methods [17]. Additionally, other methods based on phylogenetic trees are ConSurf [18] and Rate4Site [19,20] algorithms.

Although a variety of sequence-based methods have been already compared each other [13,21], what difference makes a difference is difficult to understand because such methods do not be explained by an idea. Therefore, we consider a map, a mathematical formula, on a multiple sequence alignment (MSA) and aim at constructing an exhaustive method. As part of this effort, we propose a method currently including some existing methods such as the method based on SE or SE of residue properties, the method based on a sum of pairs with/without weighting and the iv-ET or the rv-ET method.

Even if a variety of methods are executable, how are the methods evaluable? There may exist two approaches: confirmation by site-directed mutagenesis and visualization onto a three-dimensional structure. The former is more consistent with identification of binding sites because the latter is verifiable that a site is proximate from ions or molecules. In spite of that, the latter has been still used because of indefinability of protein functional sites. Therefore, on the basis of benchmark sets such as catalytic sites, ligand-binding sites or protein-protein interfaces [13], the predictive ability has been evaluated. However, the latter is immature because of usually conducting only a structure [15,22]. This mainly causes two problems. The first one is that the latter neglects a protein which binds various ions or molecules because an entry in the Protein Data Bank (PDB) [23] does not always include all states of the protein structure. The second one is that the latter cannot take account of proteins which are derived from an ancestor. Therefore, protein structures derived from different organisms are incomparable with each other. To solve these problems, we consider another map, which measures proximity of amino acid residues and ions or molecules, and then two maps are integrated.

***Corresponding author:** Satoru Miyazaki, Department of Medicinal and Life Science, Faculty of Pharmaceutical Sciences, Tokyo University of Science, Yamazaki 2641, Noda-shi, Chiba 278-8510, Japan, Tel: +81-4-7121-3630; E-mail: smiyazak@rs.noda.tus.ac.jp

Received June 21, 2015; **Accepted** July 09, 2015; **Published** July 16, 2015

Citation: Kondo Y, Miyazaki S (2015) Protein Functional Site Prediction Using a Conservative Grade and a Proximate Grade. J Data Mining Genomics Proteomics 6: 175. doi:10.4172/2153-0602.1000175

Citation: Kondo Y, Miyazaki S (2015) Protein Functional Site Prediction Using a Conservative Grade and a Proximate Grade. J Data Mining Genomics Proteomics 6: 175. doi:10.4172/2153-0602.1000175

Mathematical Formulation of Mappings of an MSA

Notation of fundamental elements

Let M=(*m*ij) denote a given MSA and here *m*ij denote an amino acid symbol of site *j* on sequence *i* in the MSA. Let ${}_kM=\left[m_{1k},m_{2k},....,m_{nk}\right]^t$ be column *k* on the MSA and we consider a mapping

$$f_x:M\to\left[0,\infty\right].$$

Mapping by a character type

In this section, we define mathematical formulation of a mapping by similarity of the amino acid symbols on ${}_iM$. Let ${}_i^tM\in{}_iM$ denote ${}_iM$ at time point t = 1, 2. . . N + 1, where *N* is a number of internal nodes on a phylogenetic tree reconstructed from the given MSA (Figure 1A), and be represented by a field of sets. For example, ${}_4^tM$;

$${}_4^1M=\left\{\left\{R,R,L,R,R,R\right\}\right\}$$
$${}_4^2M=\left\{\left\{R,R,L,R,R\right\},\left\{R\right\}\right\}$$
$${}_4^3M=\left\{\left\{R,R,L,R\right\},\left\{R\right\},\left\{R\right\}\right\}$$
$${}_4^4M=\left\{\left\{R,R,L\right\},\left\{R\right\},\left\{R\right\},\left\{R\right\}\right\}$$
$${}_4^5M=\left\{\left\{R,R\right\},\left\{L\right\},\left\{R\right\},\left\{R\right\},\left\{R\right\}\right\}$$

Let there be $g_x:{}_iM\to\left[0,1\right]$, which here maps to 1 if there exists ${}_i^tM_u\in{}_i^tM$ which comprises two or more types of characters and, in other cases, maps to 0. By $g_x\left({}_i^tM\right)$ only, ${}_1M-{}_5M$ are indistinguishable. If $g_x\left({}_i^1M\right)$ and $g_x\left({}_i^2M\right)$ are summed, ${}_1M$and the others are distinguishable. If $g_x\left({}_i^1M\right)$, $g_x\left({}_i^2M\right)$and $g_x\left({}_i^3M\right)$are summed, ${}_1M$ and ${}_2M$ and the others are distinguishable. Therefore, let

$$f_x\left({}_iM\right):=\sum_{t=1}^{T}g_x\left({}_i^tM\right)\tag{1}$$

Where T=1,2,…N.

As shown in Figure 1B, let $h_x:{}_i^tM\to\left[0,1\right]$be included in g_x and $g_x\left({}_i^tM\right)$ be represented as following three definitions:

$$g_1\left({}_i^tM\right):=\begin{cases}0\left(\forall{}_i^tM_u\in{}_i^tM,h_x\left({}_i^tM_u\right)\le\tau\right)\\1\left(\exists{}_i^tM_u\in{}_i^tM,h_x\left({}_i^tM_u\right)>\tau\right)\end{cases}\tag{2}$$

Where τ is a threshold of $h_x\left({}_i^tM\right)$,

$$g_2\left({}_i^tM\right):=\frac{1}{\left|{}_i^tM\right|}\sum_{{}_i^tM_u\in{}_i^tM}h_x\left({}_i^tM_u\right)\tag{3}$$

Where $\left|{}_i^tM\right|$ is a number of multisets in ${}_i^tM$ and

$$g_3\left({}_i^tM\right):=h_x\left({}_i^tM_{*u}\right)\tag{4}$$

Where ${}_i^tM_{*u}$ is a multiset which is separated at time point $t+1$. For example, ${}_i^tM_{*u}$;

$${}_4^1M_{*u}=\left\{R,R,L,R,R,R\right\}$$
$${}_4^2M_{*u}=\left\{R,R,L,R,R\right\}$$
$${}_4^3M_{*u}=\left\{R,R,L,R\right\}$$
$${}_4^4M_{*u}=\left\{R,R,L\right\}$$
$${}_4^5M_{*u}=\left\{R,R\right\}$$

Let A denote a field of sets of amino acid symbols and ${}_iG\subset{}_iM$ denote a field of sets of gaps in ${}_1M$. For example, A is definable as

$${}^{20}A=\{\{A\},\{C\},\{D\},\{E\},\{F\},\{G\},\{H\},\{I\},\{K\},\{L\},\{M\},\{N\},\{P\},\{Q\},\{R\},\{S\},\{T\},\{V\},\{W\}\}$$

Or ${}^{9}A=\{\{M,L,V,I\},\{H,R,K\},\{S,T\},\{A,G\},\{D,E\},\{Q,N\},\{F,W,Y\},\{P\},\{C\}\}$

and ${}_tG$ is definable as ${}_i^1G=\{\{{}_i^1\gamma,{}_i^2\gamma,...,{}_i^G\gamma\}\}$ or ${}_i^GG=\{\{{}_i^1\gamma,\},\{{}_i^2\gamma,\},...,\{{}_i^G\gamma\}\}$ where G is a number of gaps.

Let $h_x\left({}_i^tM_u\right)$ be represented as following four definitions:

$$h_1\left({}_i^tM_u\right):=\begin{cases}0(\forall l\in{}_i^tM_u,\exists X\in A\cup{}_tG;l\in X)\\1(otherwise)\end{cases}\tag{5}$$

where $\left|A\cup{}_iG\right|$ is a number of sets in $A\cup{}_iG$ and

$$p\left({}_i^tM_u,X\right)=\frac{1}{\left|{}_i^tM_u\right|}\sum_{l\in{}_i^tM_u}\begin{cases}0(l\notin X)\\1(l\in X)\end{cases}\tag{7}$$

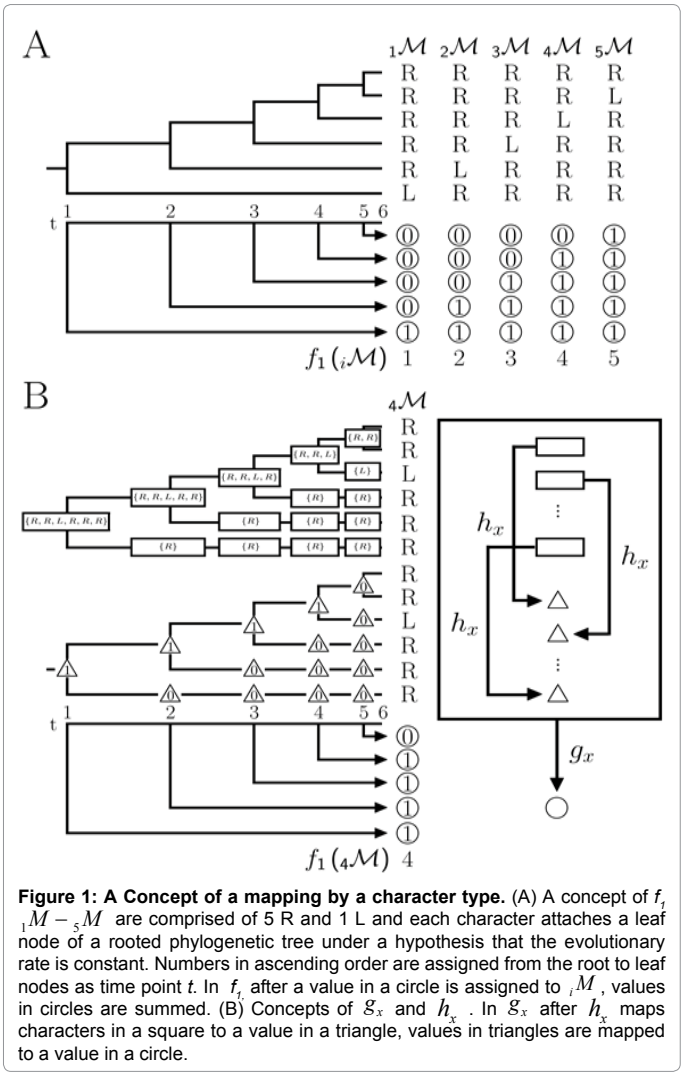

Figure 1: A Concept of a mapping by a character type. (A) A concept of f_t ${}_1M-{}_5M$ are comprised of 5 R and 1 L and each character attaches a leaf node of a rooted phylogenetic tree under a hypothesis that the evolutionary rate is constant. Numbers in ascending order are assigned from the root to leaf nodes as time point *t*. In f_t after a value in a circle is assigned to ${}_iM$, values in circles are summed. (B) Concepts of g_x and h_x . In g_x after h_x maps characters in a square to a value in a triangle, values in triangles are mapped to a value in a circle.

where $\left|{}_i^t M_u\right|$ is a number of characters in ${}_i^t M_u$ and if $p\left({}_i^t M_u, X\right)=0$, $p\left({}_i^t M_u, X\right)\log_{|A\cup {}_iG|} p\left({}_i^t M_u, X\right)$ is regarded as 0,

$$h_3\left({}_i^t M_u\right) := \frac{1}{\left|{}_i^t M_u\right|^2}\sum_{l\in {}_i^t M_u}\sum_{m\in {}_i^t M_u} s(l,m) \tag{8}$$

Where

$$s(l,m) = \begin{cases} 0\left(l\in X\in {}_iG \wedge m\in Y\in {}_iG \wedge X=Y\right) \\ \frac{S_{\max}-S_{\min}}{S_{\max}}\left(l\in X\in {}_iG \wedge m\in Y\in {}_iG \wedge X\neq Y\right) \\ \frac{S_{\max}-S_{\min}}{S_{\max}}\left(l\in X\in {}_iG \wedge m\in Y\in A\right) \\ \frac{S(l,l)-S_{\min}}{S(l,l)}\left(l\in X\in A \wedge m\in Y\in {}_iG\right) \\ \frac{S(l,l)-S(l,m)}{S(l,l)}\left(l\in X\in A \wedge m\in Y\in A\right) \end{cases} \tag{9}$$

Where ${}_{\max}$, $S_{\min}$, $S(l,l)$ and $S(l,m)$ are the maximum, the minimum, a diagonal element and an off-diagonal element in an amino acid substitution matrix, respectively, and

$$h_4\left({}_i^t M_u\right) := \frac{1}{\left|{}_i^t M_u\right|^2}\sum_{l\in {}_i^t M_u}\sum_{m\in {}_i^t M_u} s(l,m)w(l) \tag{10}$$

where l is a weight of sequence l.

Mapping by a coordinate type

Let $\mathbb{R}$ denote a set of real numbers and there be $e:\mathbb{R}^3\times\mathbb{R}^3\to[0,\infty)$. Let $R\subset\mathbb{R}^3$, $Q\subset\mathbb{R}^3$ and

$$e(R,Q) := \min_{(r,q)\in R\times Q}\left(\|r-q\|_2\right) \tag{11}$$

where $\|\cdot\|_2$ is an Euclidean norm.

Let us consider structure *k*, which contains a protein and ions or molecules. Let ${}_i^k R\subset\mathbb{R}^3$ denote atomic coordinates of amino acid residue *i* in structure *k* and ${}^kQ\subset\mathbb{R}^3$ denote atomic coordinates of ions or molecules in structure *k*. Let K denote a number of structures and the sequences are aligned. Let $\left\{{}_i^1 R, {}_i^2 R, \ldots, {}_i^{K-G} R\right\}\subseteq {}_iM$ denote a set of residues in ${}_iM$ denote a set of residues in ${}_iM$ and $\{{}_i^1\gamma, {}_i^2\gamma, \ldots, {}_i^G\gamma\} = {}_iG\subset {}_iM$ denote a set of gaps in ${}_iM$ .Let

$$f_2\left({}_iM\right) := \min_{{}_i^k R\in {}_iM\setminus {}_iG}\left[e\left({}_i^k R, {}^kQ\right)\right] \tag{12}$$

Materials and Methods

Data collection

In UniProtKB/Swiss-Prot release 2015_01 [24], entries which are annotated as 'Classic translation factor GTPase family. EF-Tu/EF-1A subfamily', do not include 'X' in the sequence and are not a fragment were 984 entries. In the PDB, entries which are referenced from above 984 entries and are determined by X-ray crystallography were 68 entries. 14 entries were excluded because of binding an immunoprotein [25] and forming a chimeric protein [26-29]. Consequently, as shown in Table 1, 54 entries including 103 chains were retained.

Computations of f_1 and f_2

As N=984 and K=103 in Figure 2, the sequences were aligned by the MAFFT 7 program [30]. 477 ${}_iM$ were extracted because of including residues which have coordinate data.

A difference between two sequences was computed by the maximum likelihood method [31] using the Jones-Taylor-Thornton model [32] as a substitution matrix and the Dayhoff method [33] for computing equilibrium frequencies. From all combinations of the differences, a phylogenetic tree was written by the unweighted pair group method with arithmetic mean [34]. $f_1\left({}_iM\right)$ was computed by changing, T, g_x, h_x, τ, A and ${}_iG$ For h_3 or h_4, the Gonnet matrix [35] was used. For h_4, a weight was computed by the Sibbald and Algos algorithm [36] and the iteration number was 100,000.

By separating each asymmetric unit, $f_1\left({}_iM\right)$ was computed and, in each entry, representative ions or molecules were shown in Table 1. However, because of uncertain functions, we excluded the following ions or molecules; sodium ion, acetate ion, sulfate ion, ammonium ion, sugar (sucrose), di(hydroxyethyl)ether, glyoxylic acid, 5-bromofuran-2-carboxylic acid, β-mercaptoethanol and water [37-43].

Correlations between f_1 and f_2

Let $[0,\infty)\supset F\ni f_1\left({}_iM\right)$ denote a subset of non-negative real numbers and a set of $f_1\left({}_iM\right)$ and be represented as $F\ni v_1<v_2<\ldots<v_J$. Let t_j denote a threshold and satisfy

$$t_j \begin{cases} <v_1 (j=0) \\ =\frac{v_j+v_{j+1}}{2}(j=1,2,\ldots,J-1) \\ >v_J (j=J) \end{cases} \tag{13}$$

Let denote a cutoff of $f_2\left({}_iM\right)$ and, in this study, c_2=3 Å. Let I_f denote a number of ${}_iM$ which satisfies $f_2\left({}_iM\right)>c_2$ and I_t denote a number of ${}_iM$ which satisfies $f_2\left({}_iM\right)\le c_2$. Let $I_{fp}\left(t_j\right)$ denote a number of ${}_iM$ which satisfies $f_2\left({}_iM\right)>c_2$ and $f_1\left({}_iM\right)\le t_j$ and $I_{tp}\left(t_j\right)$ denote a number of ${}_iM$ which satisfies $f_2\left({}_iM\right)\le c_2$ and $f_1\left({}_iM\right)\le t_j$. Let a false positive rate

$$p(t_j)=\frac{I_{fp}(t_j)}{I_f} \tag{14}$$

a true positive rate

$$q(t_j)=\frac{I_{tp}(t_j)}{I_t} \tag{15}$$

and an area under the curve

$$AUC=\frac{1}{2}\sum_{j=0}^{J-1}\left[p\left(t_{j+1}\right)-p(t_j)\right]\cdot\left[q\left(t_{j+1}\right)+q(t_j)\right] \tag{16}$$

Let $F_x\ni f_x\left({}_iM\right)$ denote a multiset of $f_x\left({}_iM\right)$ and represented as $F_x\ni {}_i^1V_x\le {}_i^2V_x\le\ldots\le {}_i^IV_x$, where *I* is a number of ${}_iM$. Let *r* denote a rank function and

$$r\left({}_i^{j+k-1}V_x\right)=j-1+\frac{t_n+1}{2} \tag{17}$$

where $j=1,2,\ldots,I, m=1,2,\ldots,I,\ k=1,2,\ldots,t_n$ and t_n is a size of the tied rank. Here, a Spearman's ρ [44] is defined as

$$\rho=\frac{1}{2\sqrt{T_1T_2}}\left\{T_1+T-\sum_{i=1}^{I}\left[r\left({}_i^lV_1\right)-r\left({}_i^mV_1\right)\right]^2\right\} \tag{18}$$

Where *l*=1, 2... *I*, *m*=1, 2... *I*,

Subfamily	Organism	PDB ID	Resolution	Ions or molecules
EF-Tu	*Bos taurus*, mitochondrial	1D2E 1XB2	1.94 2.20	GDP, Mg2+ Elongation factor Ts mitochondrial
	Escherichia coli	1EFC 2HCJ 3U6B 2BVN 4G5G 1D8T 3U6K 1DG1 1EFU 1EFM 3U2Q 2HDN 1ETU 4Q7J 1OB2 2FX3	2.05 2.12 2.12 2.30 2.30 2.35 2.45 2.50 2.50 2.70 2.70 2.80 2.90 2.90 3.35 3.40	GDP, Mg2+ GDP, TAC, Mg2+ GDP, Mg2+ ENX, GNP, Mg2+ Thiomuracin A derivative, GDP, Mg2+ Thiocillin GE2270, GDP, Mg2+ Thiocillin GE2270 analogue NVP-LDK733, GDP, Mg2+ GDP, Mg2+ Elongation factor Ts GD P Thicillin GE2270 analogue NVP-LFF571, GDP, Mg2+ GDP, TAC, Mg2+ GDP, Mg2+ Elongation factor Ts, Q β replicase Phe-tRNA, GNP, KIR, Mg2+ GDP, Mg2+
	Pseudomonas putida KT2440	4J0Q 4IW3	2.29 2.70	GDP, MES, MPD, Mg2+ Putative uncharacterized protein, GDP, Mg2+
	Thermus aquaticus	1EFT 1B23 1TTT 1TUI 1OB5	2.50 2.60 2.70 2.70 3.10	GNP, Mg2+ Cys-tRNA, GNP, Mg2+ Phe-tRNA, GNP, Mg2+ GDP, Mg2+ Phe-tRNA, ENX, GNP, Mg2+
	Thermus thermophilus	2C78 2C77 1EXM 4LBW 4H9G 1HA3 4LBV 4LBZ 4LC0 4LBY 1AIP 4V5L 4V5P 4V5Q 4V5R 4V5S 4V8Q 4V5G	1.40 1.60 1.70 1.74 1.93 2.00 2.03 2.22 2.22 2.69 3.00 3.10 3.10 3.10 3.10 3.10 3.10 3.10 3.60	GNP, PUL, Mg2+ Thiocillin GE2270, GNP, Mg2+ GNP, Mg2+ GNP, Mg2+ GNP, Mg2+ GDP, MAU, Mg2+ GNP, Mg2+ GNP, Mg2+ GNP, Mg2+ GNP, Mg2+ Elongation factor Ts 16S rRNA, 23S rRNA, Trp-tRNA, GCP, Mg2+ 16S rRNA, 23S rRNA, Trp-tRNA 16S rRNA, 30S rpS12, Trp-tRNA, GDP, KIR 16S rRNA, Trp-tRNA, GDP, KIR 16S rRNA, Trp-tRNA, GDP, KIR 16S rRNA, 23S rRNA, Small protein B SMPB, tmRNA δ, GDP, KIR, Mg2+ 16S rRNA, 23S rRNA, 30S rpS12, Thr-tRNA, GDP, KIR, Mg2+
aEF1A	*Aeropyrum pernix*	3VMF 3WXM	2.30 2.30	Peptide chain release factor subunit 1, GTP, Mg2+ Protein pelota homologue, GTP, Mg2+
	Sulfolobus solfataricus	1JNY 1SKQ		GD P GDP, Mg2+
eEF1A	*Oryctolagus cuniculus*	4C0S		GDP, Mg2+
	Saccharomyces cerevisiae	1F60 2B7C 1G7C 1IJE 2B7B 1IJF		Elongation factor 1Bα Elongation factor-1 β Elongation factor 1- β, 5GP Elongation factor 1- β, GD P Elongation factor-1 β, GD P Elongation factor 1- β, GD P

TAC; Tetracycline, ENX; Enacyloxin IIa, GNP; Phosphoaminophosphonic acid-guanylate ester, KIR; Kirromycin, MES; 2-(N-morpholino)-ethanesulfonic acid, MPD; (4S)-2-methyl-2,4-pentanediol, PUL; Pulvomycin, MAU; N-methyl kirromycin, GCP; Phosphomethylphosphonic acid guanylate ester, 5GP; Guanosine-5′ -monophosphate.

Table 1: 54 PDB entries of EF-Tu/EF-1A proteins.

$$T_1 = \frac{1}{12}\left[I^3 - I - \sum_{n=1}^{N_2} \left(t_n^3 - t_n \right) \right] \quad (19)$$

And

$$T_2 = \frac{1}{12}\left[I^3 - I - \sum_{n=1}^{N_2} \left(t_n^3 - t_n \right) \right] \quad (20)$$

where N_1 and N_2 are numbers of tied ranks in F_1 and F_2, respectively.

Visualization

$f_1({}_iM)$, $f_2({}_iM)$, AUC and Spearman's ρ were visualized by the matplotlib Python package [45]. A three-dimensional structure was visualized by the VMD program [46].

Results

Fitness between f_1 and f_2

If g_x, h_x, τ and A are same but ${}_iG$ is different, Table 2 shows that when ${}_iG = {}_i^1G$, the AUC or the Spearman's ρ is smaller than ${}_iG = {}_i^GG$. In the latter case, Figure 3 shows that when the time point increases, the AUC or the Spearman's ρ tends to increase.

Evaluation of predicted functional amino acid residues by f_2

Figure 4A shows that ${}_iM \in M$ is classifiable in 4 by $f_1({}_iM)$ and $f_2({}_iM)$ using a receiver operating characteristic (ROC) curve [47] in Figure 4B. Figures 4C and 4D show that the left sides tend to have small $f_1({}_iM)$ and small $f_2({}_iM)$ but the right sides tend to have large $f_1({}_iM)$ and large $f_2({}_iM)$.

Discussion

Meanings of $f_1({}_iM)$, $f_2({}_iM)$, AUC and Spearman's ρ are as follows. $f_1({}_iM)$ becomes small when characters are only diverged in near to the root of the phylogenetic tree. $f_1({}_iM)$ becomes large when characters are diverged in far from the root. $f_1({}_iM)$ becomes small when at least one amino acid residue in ${}_iM$ is proximate from an ion or a molecule. $f_1({}_iM)$ becomes large when amino acid residues in ${}_iM$ are not proximate from ions or molecules in all cocrystal structures. If the AUC is 0.5, a correlation between $f_1({}_iM)$ and being proximate and being non-proximate under a cutoff of $f_2({}_iM)$ may not exist. If the AUC is close to 1, small $f_1({}_iM)$ and large $f_1({}_iM)$ correlate with being proximate and being non-proximate, respectively. If the AUC is close to 0, large $f_1({}_iM)$ and small $f_1({}_iM)$ correlate with being proximate and being non-proximate, respectively. If the Spearman's ρ is 0, a linear correlation between $f_1({}_iM)$ and $f_2({}_iM)$ may not exist. If the Spearman's ρ is close to 1 or -1, $f_1({}_iM)$ and $f_2({}_iM)$ have a positive or a negative linear correlation, respectively.

If T=1, $g_x = g_3$, $h_x = h_2$, $A = {}^{20}A$ and ${}_iG = {}_i^1G$, the method is the method based on SE [8]. If T=1, $g_x = g_3$, $h_x = h_2$, $A = {}^{9}A$ and ${}_iG = {}_i^1G$, the method is the method based on SE of residue properties [10]. If T=1 is changed to T=N in the former and the latter, Figure 3 shows that the AUC is from 0.5779 to 0.6147 and the Spearman's € ρ is from 0.0757 to 0.1241 and the AUC is from 0.5709 to 0.5992 and the Spearman's ρ is from 0.1152 to 0.1405, respectively. Therefore, in the former and the latter, distinguishing characters utilizing the phylogenetic tree is effective for improving the AUC and the Spearman's ρ.

If T=1, $g_x = g_3$, $h_x = h_3$ and ${}_iG = {}_i^1G$, the method is the method based on a sum of pairs [11]. If T=1, $g_x = g_3$, $h_x = h_4$ and ${}_iG = {}_i^1G$, the method is the method based on a sum of pairs with weighting [11]. If T=1 is changed to T=N in the former and the latter, Figure 3 shows that the AUC is from 0.6083 to 0.6276 and the Spearman's ρ is from 0.1982 to 0.1653 and the AUC is from 0.6093 to 0.6211 and the Spearman's ρ is from 0.2263 to 0.1502, respectively. Therefore, in the former and the latter, distinguishing characters utilizing the phylogenetic tree is effective for improving the AUC but not for the Spearman's . However, in the above case, if ${}_iG = {}_i^1G$ is changed to ${}_iG = {}_i^GG$ in the former and the latter, Figure 3 shows that the AUC is from 0.6941 to 0.7349 and the Spearman's ρ is from 0.4981 to 0.5650 and the AUC is from 0.6846 to 0.7335 and the Spearman's ρ is from 0.4749 to 0.5637, respectively. Therefore, in the former and the latter, distinguishing characters utilizing the phylogenetic tree and considering that each gap is different are effective for improving the AUC and the Spearman'sρ.

If T=N, $g_x = g_3$, $h_x = h_1$, $A = {}^{20}A$ and ${}_iG = {}_i^1G$, the method is the iv-ET method [17]. If T=N, $g_x = g_3$, $h_x = h_2$, $A = {}^{20}A$ and ${}^iG = {}_i^1G$, the method is equivalent to the rv-ET method [17]. If ${}_iG = {}_i^1G$ is changed to ${}_iG = {}_i^GG$ in the former and the latter, Table 2 shows that the AUC is from 0.5896 to 0.6242 and the Spearman's ρ is from 0.1221 to 0.3650 and the AUC is from 0.6180 to 0.7417 and the Spearman's ρ is from 0.1308 to 0.5722, respectively. Therefore, in the former and the

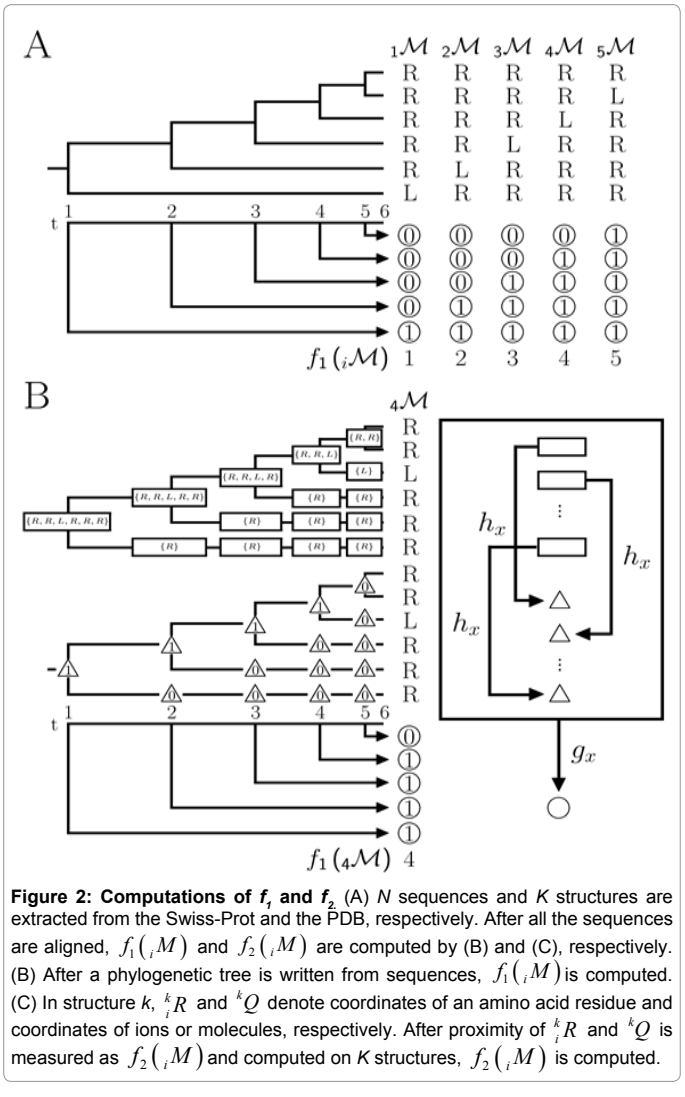

Figure 2: Computations of f_1 and f_2 (A) *N* sequences and *K* structures are extracted from the Swiss-Prot and the PDB, respectively. After all the sequences are aligned, $f_1({}_iM)$ and $f_2({}_iM)$ are computed by (B) and (C), respectively. (B) After a phylogenetic tree is written from sequences, $f_1({}_iM)$ is computed. (C) In structure *k*, ${}_i^kR$ and ${}^kQ$ denote coordinates of an amino acid residue and coordinates of ions or molecules, respectively. After proximity of ${}_i^kR$ and ${}^kQ$ is measured as $f_2({}_iM)$ and computed on *K* structures, $f_2({}_iM)$ is computed.

g_x	h_x	A	${}_iG$	τ	AUC	Spearman's ρ
g_1	h_1	${}^{20}A$	${}^{1}_{i}G$	0<τ <1	0.5896	0.1221
			${}^{G}_{i}G$	0<τ <1	0.6242	0.3650
		${}^{9}A$	${}^{1}_{i}G$	0<τ <1	0.5700	0.1015
			${}^{G}_{i}G$	0<τ <1	0.6184	0.3509
	h_2	${}^{20}A$	${}^{1}_{i}G$	0.1	0.6036	0.1436
				0.2	0.6376	0.1977
				0.3	0.6420	0.2015
				0.4	0.6120	0.1585
			${}^{G}_{i}G$	0.1	0.7207	0.5566
				0.2	0.7412	0.5773
				0.3	0.7374	0.5613
				0.4	0.7037	0.5028
		${}^{9}A$	${}^{1}_{i}G$	0.2	0.5757	0.1171
				0.3	0.6023	0.1659
				0.4	0.6151	0.1715
				0.1	0.5885	0.1160
			${}^{G}_{i}G$	0.1	0.6997	0.5185
				0.2	0.6984	0.5097
				0.3	0.7084	0.5063
				0.4	0.6854	0.4797
	h_3	A	${}^{1}_{i}G$	0.1	0.5901	0.1288
				0.2	0.6052	0.1467
				0.3	0.6366	0.2315
				0.4	0.6391	0.2763
			${}^{G}_{i}G$	0.1	0.6758	0.4474
				0.2	0.6975	0.4734
				0.3	0.6905	0.4823
				0.4	0.6865	0.4922
	h_4	A	${}^{1}_{i}G$	0.1	0.5872	0.1200
				0.2	0.5939	0.1387
				0.3	0.5916	0.1460
				0.4	0.5995	0.1733
			${}^{G}_{i}G$	0.1	0.6602	0.4162
				0.2	0.6805	0.4470
				0.3	0.6888	0.4599
				0.4	0.6782	0.4675
g_2	h_1	${}^{20}A$	${}^{1}_{i}G$	-	0.5916	0.0718
			${}^{G}_{i}G$	-	0.7399	0.5780
		${}^{9}A$	${}^{1}_{i}G$	-	0.5652	0.0587
			${}^{G}_{i}G$	-	0.7020	0.5145
	h_2	${}^{20}A$	${}^{1}_{i}G$	-	0.6180	0.1308
			${}^{G}_{i}G$	-	0.7417	0.5722
		${}^{9}A$	${}^{1}_{i}G$	-	0.5890	0.1257
			${}^{G}_{i}G$	-	0.7012	0.5091
	h_3	A	${}^{1}_{i}G$	-	0.6225	0.1579
			${}^{G}_{i}G$	-	0.7287	0.5517
	h_4	A	${}^{1}_{i}G$	-	0.6138	0.1412
			${}^{G}_{i}G$	-	0.7265	0.5501

g_3	h_1	^{20}A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.5792 0.7386	0.0378 0.5801
		^{9}A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.5655 0.7052	0.0501 0.5244
	h_2	^{20}A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.6147 0.7393	0.1241 0.5723
		^{9}A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.5992 0.7059	0.1405 0.5170
	h_3	A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.6276 0.7349	0.1653 0.5650
	h_4	A	${}^{1}_{i}G$ ${}^{G}_{i}G$	- -	0.6211 0.7335	0.1502 0.5637

Table 2: Correlations between f_1 and f_2.

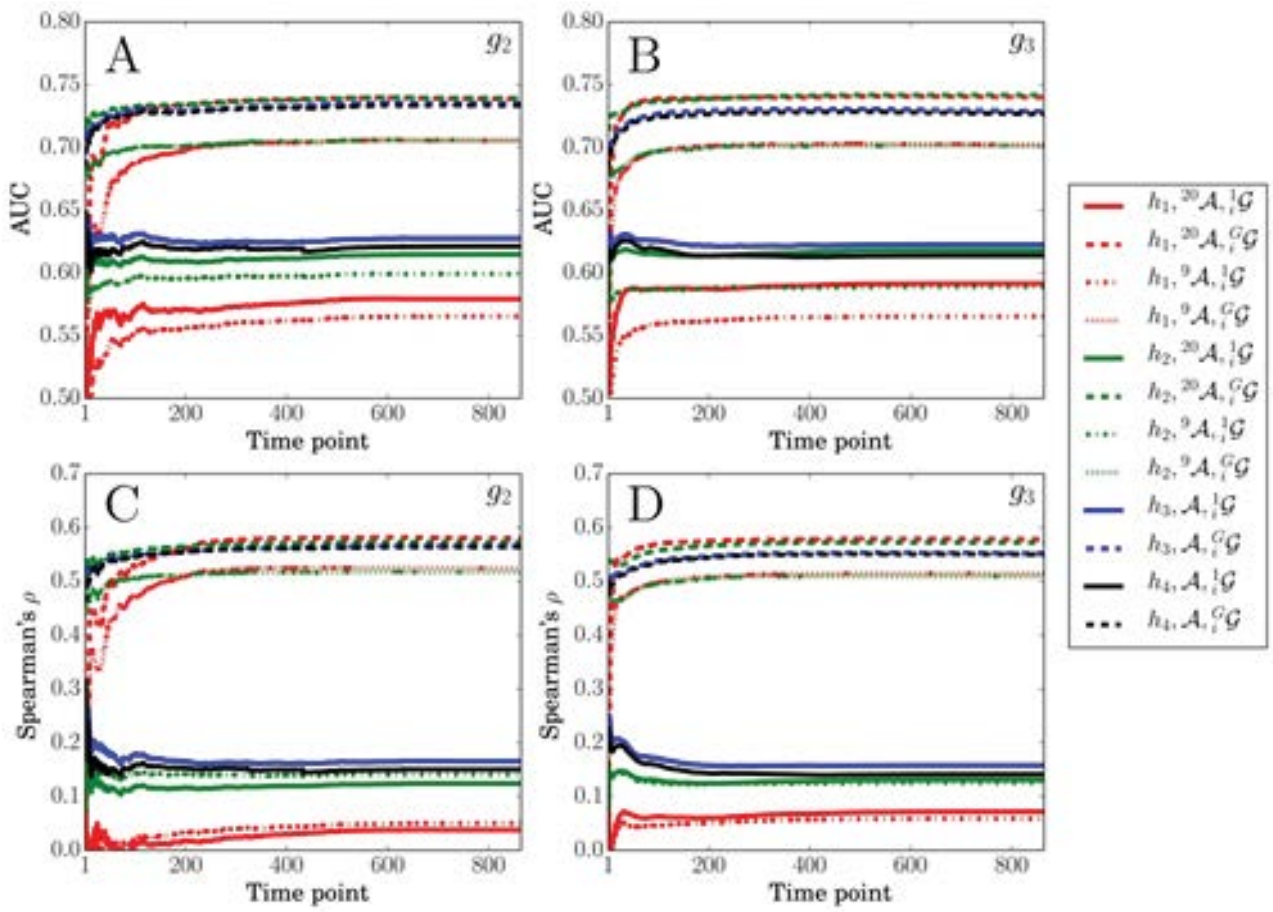

Figure 3: Dependence on time points. The time point is T in Eq. (1). (A) (B) AUC and (C) (D) Spearman's ρ were computed using g_x , h_x , A and ${}_{i}G$ as shown in the figures.

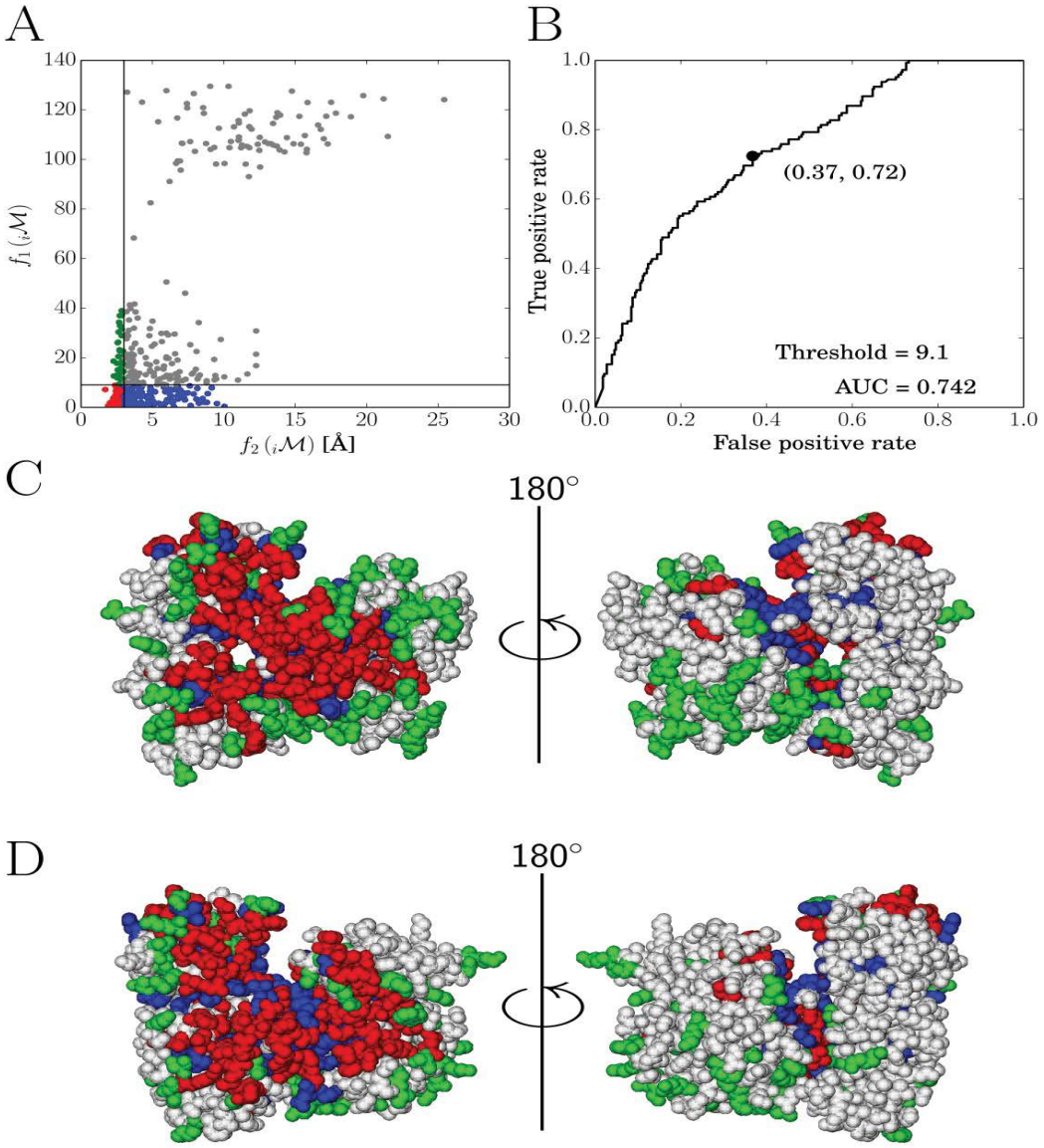

Figure 4: A scatter plot, an ROC curve and three-dimensional structures. $f_1({}_iM)$ was computed using g_2, h_2, ${}^{20}A$ and ${}^{G}_{i}G$ (A) A scatter plot of $f_1({}_iM)$ and $f_2({}_iM)$ However, one point whose $f_1({}_{468}M)$ 122.2 and $f_2({}_{468}M)$ 45.56 was not shown. ${}_iM \in M$ was classified into 2 by whether $f_2({}_iM)$ is equal to or smaller than 3 Å or larger than 3 Å. (B) By regarding the former as true and the latter as false, the ROC curve was written using $f_1({}_iM)$ The threshold was determined so that (true positive rate + 1 − false positive rate) is maximum and, eventually, $M \in M$ was classified into 4, which were visualized onto three-dimensional structures of (C) *Thermus thermophilus* EF-Tu [42] and (D) *Saccharomyces cerevisiae* EF1A [48].

latter, considering that each gap is different is effective for improving the AUC and the Spearman's ρ Thus, $f_1({}_iM)$ is evaluable by $f_2({}_iM)$ and our methods improved some existing methods.

EF-Tu/EF-1A proteins are responsible for protein biosynthesis [42,48] and we selected cocrystal structures involving the function. Therefore, if $f_2({}_iM)$ is small, an amino acid residue in ${}_iM$ is proximate from a region involving protein biosynthesis. If $f_2({}_iM)$ is large, the amino acid residues in ${}_iM$ are not proximate from the region. Figures 4A, 4C and 4D show the proximate region and the non-proximate region and Figure 4B shows that, on the ROC curve of $f_1({}_iM)$, the AUC is 0.742, which indicates that the proximate region tends to become small $f_1({}_iM)$ but the non-proximate region tends to become large $f_1({}_iM)$ In addition, Table 2 shows that the Spearman's ρ is 0.5722, which indicates that $f_1({}_iM)$ tends to be small if $f_2({}_iM)$ is small and $f_1({}_iM)$ tends to be large if $f_2({}_iM)c$ is large. However, a complete linear correlation between

and $f_2({}_iM)$ was not obtainable and therefore not all of $f_1({}_iM)$ can explain $f_2({}_iM)$. This may indicate that $f_1({}_iM)$ and $f_2({}_iM)$ can measure a similar thing each other but cannot always measure a same thing and, by $f_1({}_iM)$ and $f_2({}_iM)$, measurable things such as importance for binding ions or molecules or importance for maintaining the structure may be different. Thus, from a different point of view, $f_1({}_iM)$ and $f_2({}_iM)$ can evaluate an amino acid residue.

Conclusions

Methods to map an MSA, which is represented as a character type and a coordinate type, were described and we propose two usages. The first one is to assess fitness between the first map and the second map. The second one is to evaluate predicted functional amino acid residues by use of the second map. Our methods show a better performance and reliability for functional site prediction of EF-Tu/EF-1A proteins.

References

1. Huang YF, Golding GB (2014) Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. PLoS Comput Biol 10.
2. Fang C, Noguchi T, Yamana H (2014) Simplified sequence-based method for ATP-binding prediction using contextual local evolutionary conservation. Algorithms Mol Biol 9.
3. Gultas M, Duzgun G, Herzog S, Juger SJ, Meckbach C, et al. (2014) Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. BMC Bioinformatics 15:96.
4. Janda JO, Popal A, Bauer J, Busch M, Klocke M, et al. (2014) H2rs: deducing evolutionary and functionally important residue positions by means of an entropy and similarity based analysis of multiple sequence alignments. BMC Bioinformatics 15.
5. Lee TW, Yang ASP, Brittain T, Birch NP (2015) An analysis approach to identify specific functional sites in orthologous proteins using sequence and structural information: application to neuroserpin reveals regions that differentially regulate inhibitory activity. Proteins 83: 135-152.
6. Huang YF, Golding GB (2015) FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. Bioinformatics 31: 523-531.
7. Valdar W (2002) Scoring residue conservation. Proteins 48: 227-241
8. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural mean ing of sequence alignment. Proteins 9: 56-68.
9. Shenkin P, Erman B, Mastrandrea L(1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11: 297-313.
10. Williamson RM (1995) Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J Theor Biol 174: 179-188.
11. Valdar W, Thornton J (2001) Protein-protein interfaces: Analysis of amino acid conservation in homodimers. Proteins 42: 108-124.
12. Wang K, Samudrala R (2006) Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics 7.
13. Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23: 1875-1882.
14. Henikof S, Henikof J (1994) Position-based sequence weights. J Mol Biol 243: 574-578.
15. Lichtarge O, Bourne H, Cohen F (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342-358.
16. Landgraf R, Fischer D, Eisenberg D (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng 12: 943-951.
17. Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336: 1265-1282.
18. Armon A, Graur D, Ben-Tal N (2001) ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307: 447-463.
19. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18: S71-S77.
20. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. Mol Biol Evol 21: 1781-1791.
21. Johansson F, Toh H (2010) A comparative study of conservation and variation scores. BMC Bioinformatics 11.
22. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38: W529- W533.
23. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.
24. Consortium TU (2015) Uniprot: a hub for protein information. Nucleic Acids Res 43: D204-D212, 2015.
25. Dai S, Crawford F, Marrack P, Kappler JW (2008) The structure of HLA-DR52c: comparison to other HLA-DRB3 alleles. Proc Natl Acad Sci U S A 105: 11893-11897.
26. Kidmose RT, Vasiliev NN, Chetverin AB, Andersen GR, Knudsen CR (2010) Structure of the Qβ replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins. Proc Natl Acad Sci U S A 107: 10884-10889.
27. Takeshita D, Tomita K (2010) Assembly of Qβ viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. Proc Natl Acad Sci U S A 107: 15733-15738.
28. Takeshita D, Tomita K(2012) Molecular basis for RNA polymerization by Qβ replicase. Nat Struct Mol Biol 19: 229-237.
29. Takeshita D, Yamashita S, Tomita K (2012) Mechanism for template-independent terminal adenylation activity of Qβ replicase. Structure 20: 1661-1669.
30. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.
31. Kishino H, Miyata T, Hasegawa M (1990) Maximum-likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31: 151-160.
32. Jones D, Taylor W, Thornton J (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282.
33. Dayhof MO, Schwartz RM (1978) Chapter 22: A model of evolutionary change in proteins. In: in Atlas of Protein Sequence and Structure.
34. Sneath PHA, Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman aqnd company, San Francisco.
35. Benner S, Cohen M, Gonnet G (1994) Amino-acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7: 1323-1332.
36. Sibbald P, Argos P (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 216: 813-818.
37. Hefron SE, Mui S, Aorora A, Abel K, Bergmann E, et al. (2006) Molecular complementarity between tetracycline and the GTPase active site of elongation factor Tu. Acta Crystallogr D Biol Crystallogr 62: 1 3 9 2 -1 4 0 0.
38. Nissen P, Thirup S, Kjeldgaard M, Nyborg J (1999) The crystal structure of Cys-tRNACys-EF-Tu-GDPNP reveals general and specific features in the ternary complex and in tRNA. Structure 7: 1 4 3 -1 5 6.
39. LaMarche MJ, Leeds JA, Amaral K, Brewer JT, Bushell SM, et al. (2011) Antibacterial optimization of 4-aminothiazolyl analogues of the natural product GE2270 A: identification of the cycloalkylcarboxylic acids. J Med Chem 54: 8099-8109.
40. Kobayashi K, Saito K, Ishitani R, Ito K, Nureki O (2012) Structural basis for translation termination by archaeal RF1 and GTP-bound EF1α complex. Nucleic Acids Res 40: 9319-9328.
41. Groftehauge MK, Therkelsen MO, Taaning R, Skrydstrup T, Morth JP, et al. (2013) Identifying ligand-binding hot spots in proteins using brominated fragments. Acta Crystallogr F Struct Biol Commun 69: 1 0 6 0 -1 0 6 5 .
42. Parmeggiani A, Krab I, Watanabe T, Nielsen R, Dahlberg C, et al. (2006) Enacyloxin IIa pinpoints a binding pocket of elongation factor Tu for development of novel antibiotics. J Biol Chem 281: 2 8 9 3 -2900 .
43. Vogeley L, Palm G, Mesters J, Hilgenfeld R (2001) Conformational change of elongation factor Tu (EF-Tu) induced by antibiotic binding - crystal structure of the complex between EF-Tu•GDP and aurodox. J Biol Chem 276: 17149-17155.
44. Spearman C (2010) The proof and measurement of association between two things. Int J Epidemiol 39: 1 1 59 -1 1 61 .

45. Hunter JD (2007) Matplotlib: a 2D graphics environment. Computing Sci Eng 9: 90-95.

46. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph Model 14: 3 3 -3 8 .

47. Lasko T, Bhagwat J, Zou K, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 38: 404-415.

48. Andersen G, Pedersen L, Valente L, Chatterjee I, Kinzy T, et al. (2000) Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A : eEF1Bα. Mol Cell 6: 1261-1266.