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 cocystal 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], weighing [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 RateSite [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

Copyright: © 2015 Kondo Y. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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 $\mathcal{M} = [m_1, m_2, \ldots, m_N]$ be column $k$ on the MSA and we consider a mapping $f_k : \mathcal{M} \to [0, \infty]$.

Mapping by a character type

In this section, we define mathematical formulation of a mapping by similarity of the amino acid symbols on $\mathcal{M}$. Let $\mathcal{M} \in \mathcal{M}$ denote $\mathcal{M}$ at time point $t = 1, 2, \ldots, N + 1$, 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, $\mathcal{M}$:

- $\mathcal{M} = \{[R, R, L, R, R, R]\}
- \mathcal{M} = \{[R, R, L, R, R]\}
- \mathcal{M} = \{[R, R, L, R, R]\}
- \mathcal{M} = \{[R, R, R, L, R]\}
- \mathcal{M} = \{[R, R, R, L]\}

Let there be $g_\tau : \mathcal{M} \to [0, 1]$, which here maps to 1 if there exists $\mathcal{M}_x \in \mathcal{M}$ which comprises two or more types of characters and, in other cases, maps to 0. By $g_\tau(\mathcal{M})$, only $\mathcal{M} \in \mathcal{M}$ and $\mathcal{M} \in \mathcal{M}$ are indistinguishable. If $g_\tau(\mathcal{M})$ and $g_\tau(\mathcal{M})$ are summed, $\mathcal{M}$ and $\mathcal{M}$ are summed. Therefore, let

$$f_\tau(\mathcal{M}) = \sum_{\mathcal{M}_x \in \mathcal{M}} g_\tau(\mathcal{M}_x)$$

where $\tau = 1, 2, \ldots, N$.

As shown in Figure 1B, let $h_\tau : \mathcal{M} \to [0, 1]$ be included in $g_\tau$ and $g_\tau(\mathcal{M})$ be represented as following three definitions:

$$g_\tau(\mathcal{M}) = \begin{cases} 0 & \forall \mathcal{M}_x \in \mathcal{M}, h_\tau(\mathcal{M}_x) \leq \tau \\ \{2\} & \exists \mathcal{M}_x \in \mathcal{M}, h_\tau(\mathcal{M}_x) > \tau \end{cases}$$

where $\tau$ is a threshold of $h_\tau(\mathcal{M})$.

$$g_\tau(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{\mathcal{M}_x \in \mathcal{M}} h_\tau(\mathcal{M}_x)$$

where $\mathcal{M}_x$ is a number of multisets in $\mathcal{M}$ and $g_\tau(\mathcal{M}) = h_\tau(\mathcal{M}_x)$.

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

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

$G$ is definable as $G = \{[i], [j], \ldots, [g]\}$ or $G = \{[i], [j], \ldots, [g]\}$ where $G$ is a number of gaps.

Let $h_\tau(\mathcal{M})$ be represented as following four definitions:

$$h_\tau(\mathcal{M}) = \begin{cases} 0 & \forall \mathcal{M}_x \in \mathcal{M}, \exists \mathcal{X} \in A \cup \mathcal{G}, [l] \in \mathcal{X} \\ \text{otherwise} \end{cases}$$

where $A \cup \mathcal{G}$ is a number of sets in $A \cup \mathcal{G}$ and $p(\mathcal{M}_x, x) = \frac{1}{|\mathcal{M}_x|} \sum_{l \in \mathcal{X}} 0(\mathcal{M}_x, x)$.

Figure 1: A Concept of a mapping by a character type. (A) A concept of $f_\tau$ and $f_\tau$ 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 $t$. After a value in a circle is assigned to $f_\tau$, values in circles are summed. (B) Concepts of $h_\tau$ and $h_\tau$ and $h_\tau$ map characters in a square to a value in a triangle, values in triangles are mapped to a value in a circle.
where $|M|$ is a number of characters in $\mathcal{M}$, and if $P(\mathcal{M}, X) = 0$, $P(\mathcal{M}, X) \log_{10} P(\mathcal{M}, X)$ is regarded as 0,

$$h_i(\mathcal{M}) := \frac{1}{|M|} \sum_{i \in \mathcal{M}} \sum_{m \in \mathcal{I}} s(l, m)$$  \hspace{1cm} (8)

Where

$$s(l, m) = \begin{cases} 
0 & (i \in X \in \mathcal{G}, m \in X \in \mathcal{G}, X = Y) \\
S_{\mathcal{M}} - S_{\mathcal{M}} & (i \in X \in \mathcal{G}, m \in X \in \mathcal{G}, X \neq Y) \\
S_{\mathcal{M}} - S_{\mathcal{M}} & (i \in X \in \mathcal{A}, m \in Y \in \mathcal{G}) \\
S_{\mathcal{M}} & (i \in X \in \mathcal{A}, m \in Y \in \mathcal{A}) \\
S_{\mathcal{M}} - S_{\mathcal{M}} & (i \in X \in \mathcal{A}, m \in Y \in \mathcal{A}) \\
S_{\mathcal{M}} & (i \in X \in \mathcal{A}, m \in Y \in \mathcal{A}) 
\end{cases}$$  \hspace{1cm} (9)

Where $S_{\mathcal{M}}$ and $S_{\mathcal{A}}$ are the maximum, a diagonal element and an off-diagonal element in an amino acid substitution matrix, respectively, and

$$h_i(|M|) := \frac{1}{|M|} \sum_{i \in \mathcal{M}} \sum_{m \in \mathcal{I}} s(l, m)$$  \hspace{1cm} (10)

where $I$ is a weight of sequence $I$.

**Mapping by a coordinate type**

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

$$e_R(\mathcal{Q}) := \min_{v \cdot q = 0} \|v\|_2$$  \hspace{1cm} (11)

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

Let us consider structure $k$, which contains a protein and ions or molecules. Let $\mathcal{R} \subset \mathbb{R}^3$ denote atomic coordinates of amino acid residue $i$ in structure $k$ and $\mathcal{Q} \subset \mathbb{R}^3$ denote atomic coordinates of ions or molecules in structure $k$. Let $K$ denote a number of structures and the sequences are aligned. Let $\{\mathcal{M}, \mathcal{R} \subset \mathbb{R}^3 \} \subset \mathcal{M}$ denote a set of residues in $\mathcal{M}$ and $\{\mathcal{M}, \mathcal{Q} \subset \mathbb{R}^3 \} = \mathcal{G}, \mathcal{C} \subset \mathcal{M}$ denote a set of gaps in $\mathcal{M}$. Let

$$f_i(\mathcal{M}) := \min_{x \in \mathcal{M}, \mathcal{Q} \subset \mathcal{Q}} \left[ e_C(\mathcal{R}, \mathcal{Q}) \right]$$  \hspace{1cm} (12)

**Materials and Methods**

**Data collection**

In UniProtKB/Swiss-Prot release 2015_01 [24], entries which are annotated as ‘Classic translation factor GT-Pase 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 $f_i$ and $f_2$**

As $N=984$ and $K=103$ in Figure 2, the sequences were aligned by the

MAFFT 7 program [30], 477 $\mathcal{M}$ 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]. $f_i(\mathcal{M})$ was computed by changing, $T, S, R, h_1, \tau, A$ and $G$. For $h_i$ or $h_2$, the Gonnet matrix [35] was used. For $h_i$, a weight was computed by the Sibbald and Algos algorithm [36] and the iteration number was 100,000.

By separating each asymmetric unit, $f_i(\mathcal{M})$ 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, $\beta$-mercaptoethanol and water [37-43].

**Correlations between $f_i$ and $f_2$**

Let $[0, \infty) \subset \mathcal{F} \subset f_i(\mathcal{M})$ denote a subset of non-negative real numbers and a set of $f_i(\mathcal{M})$ be represented as $\mathcal{F} \ni v_1 < v_2 < \ldots < v_j$ . Let $l_f$ denote a threshold and satisfy

$$< v_i(j = 0) \leq v_j \leq v_i(j > 0) \leq v_j = v_j(j = 0) \leq v_j$$  \hspace{1cm} (13)

Let denote a cutoff of $f_i(\mathcal{M})$ and, in this study, $c_3 = 3$. Let $l_f$ denote a number of $\mathcal{M}$ which satisfies $f_i(\mathcal{M}) > c_1$ and $l_i$ denote a number of $\mathcal{M}$ which satisfies $f_i(\mathcal{M}) \leq c_1$ . Let $l_{g_\mathcal{F}}(s_i)$ denote a number of $\mathcal{M}$ which satisfies $f_i(\mathcal{M}) > c_i$ and $l_{g_\mathcal{F}}(s_i)$ denote a number of $\mathcal{M}$ which satisfies $f_i(\mathcal{M}) \leq c_i$ and $f_i(\mathcal{M}) \leq s_i$. Let a false positive rate

$$p(t) = \frac{l_{g_\mathcal{F}}(s_i)}{l_f}$$  \hspace{1cm} (14)

a true positive rate

$$q(t) = \frac{l_{g_\mathcal{F}}(s_i)}{l_i}$$  \hspace{1cm} (15)

and an area under the curve

$$AUC = \frac{1}{2} \sum_{i=1}^{l_f} [p(t_i) - p(t_{i-1})][q(t_i) + q(t_{i-1})]$$  \hspace{1cm} (16)

Let $F_i \ni f_i(\mathcal{M})$ denote a multiset of $f_i(\mathcal{M})$ and represented as $F_i \ni v'_1 < v'_2 < \ldots < v'_r$, where $I$ is a number of $\mathcal{M}$ . Let $r$ denote a rank function and

$$r(\mathcal{M}) = j = \frac{1 + \frac{l_i}{2}}{2}$$  \hspace{1cm} (17)

where $j = 1, 2, \ldots, I, m = 1, 2, \ldots, l, k = 1, 2, \ldots, l$ and $l_{\mathcal{F}}$ is a size of a tied rank. Here, a Spearman’s $\rho$ [44] is defined as

$$\rho = \frac{1}{2} \frac{1}{\sqrt{H_i + T - \sum_{i=1}^{l_f} |r(y'_i) - r(y'_j)|}}$$  \hspace{1cm} (18)

Where $l=1, 2, \ldots, I, m=1, 2, \ldots, I$. 

**References**

1. [Reference 1]
2. [Reference 2]
3. [Reference 3]
4. [Reference 4]
5. [Reference 5]
6. [Reference 6]
7. [Reference 7]
8. [Reference 8]
9. [Reference 9]
10. [Reference 10]
11. [Reference 11]
12. [Reference 12]
13. [Reference 13]
14. [Reference 14]
15. [Reference 15]
16. [Reference 16]
17. [Reference 17]
18. [Reference 18]
19. [Reference 19]
20. [Reference 20]
21. [Reference 21]
22. [Reference 22]
23. [Reference 23]
24. [Reference 24]
25. [Reference 25]
26. [Reference 26]
27. [Reference 27]
28. [Reference 28]
29. [Reference 29]
30. [Reference 30]
31. [Reference 31]
32. [Reference 32]
33. [Reference 33]
34. [Reference 34]
35. [Reference 35]
36. [Reference 36]
37. [Reference 37]
38. [Reference 38]
39. [Reference 39]
40. [Reference 40]
41. [Reference 41]
42. [Reference 42]
43. [Reference 43]
44. [Reference 44]

| Subfamily            | Organism               | PDB ID  | Resolution | Ions or molecules                                                                 |
|---------------------|------------------------|---------|------------|-----------------------------------------------------------------------------------|
| **Bos taurus, mitochondrial** |                        |         |            |                                                                                   |
|                     |                        | 1E2E    | 1.94       | GDP, Mg2+                                                                          |
|                     |                        | 1XB2    | 2.20       | Elongation factor Ts mitochondrial                                                |
|                     | **Escherichia coli**   |         |            |                                                                                   |
|                     |                        | 1EFC    | 2.05       | GDP, Mg2+                                                                          |
|                     |                        | 2HCJ    | 2.12       | GDP, TAC, Mg2+                                                                     |
|                     |                        | 3U6B    | 2.12       | GDP, Mg2+                                                                          |
|                     |                        | 2BN    | 2.30       | ENX, GNP, Mg2+                                                                     |
|                     |                        | 4GSG    | 2.30       | Thiomuracin A derivative, GDP, Mg2+                                                |
|                     |                        | 1DST    | 2.35       | Thiolin GE2270, GDP, Mg2+                                                          |
|                     |                        | 3U6K    | 2.45       | Thiolin GE2270 analogue NVP-LDK733, GDP, Mg2+                                      |
|                     |                        | 1DQ1    | 2.50       | GDP, Mg2+                                                                          |
|                     |                        | 1EFU    | 2.50       | Elongation factor Ts                                                               |
|                     |                        | 1EFM    | 2.50       | GD P                                                                               |
|                     |                        | 3U2Q    | 2.70       | Thiolin GE2270 analogue NVP-LFF571, GDP, Mg2+                                      |
|                     |                        | 2HDN    | 2.70       | GDP, TAC, Mg2+                                                                     |
|                     |                        | 1ETU    | 2.80       | GDP, Mg2+                                                                          |
|                     |                        | 3U7J    | 2.90       | Elongation factor Ts, Qβ replicase                                                 |
|                     |                        | 1OB2    | 3.35       | Phe-tRNA, GNP, KIR, Mg2+                                                           |
|                     |                        | 2FX3    | 3.40       | GDP, Mg2+                                                                          |
| **Pseudomonas putida KT2440** |                   | 4J0Q    | 2.29       | GDP, MES, MPD, Mg2+                                                                |
|                     |                        | 4IW3    | 2.70       | Putative uncharacterized protein, GDP, Mg2+                                         |
| **EF-Tu**           | **Thermus aquaticus**  |         |            |                                                                                   |
|                     |                        | 1EFT    | 2.50       | GNP, Mg2+                                                                          |
|                     |                        | 1B23    | 2.60       | Cys-tRNA, GNP, Mg2+                                                                |
|                     |                        | 1TTT    | 2.60       | Phe-tRNA, GNP, Mg2+                                                                |
|                     |                        | 1TU    | 2.70       | GDP, Mg2+                                                                          |
|                     |                        | 1OB5    | 2.70       | Phe-tRNA, ENX, GNP, Mg2+                                                           |
| **Thermus thermophilus** |                      | 2C78    | 1.40       | GNP, PUL, Mg2+                                                                     |
|                     |                        | 2C77    | 1.60       | Thiolin GE2270, GNP, Mg2+                                                          |
|                     |                        | 1EXM    | 1.70       | GNP, Mg2+                                                                          |
|                     |                        | 4LBW    | 1.74       | GNP, Mg2+                                                                          |
|                     |                        | 4HG    | 1.93       | GNP, Mg2+                                                                          |
|                     |                        | 1HA3    | 2.00       | GNP, Mg2+                                                                          |
|                     |                        | 4LV    | 2.03       | GDP, MAU, Mg2+                                                                     |
|                     |                        | 4LBZ    | 2.22       | GNP, Mg2+                                                                          |
|                     |                        | 4LC0    | 2.22       | GNP, Mg2+                                                                          |
|                     |                        | 4LBY    | 2.69       | GNP, Mg2+                                                                          |
|                     |                        | 1AIP    | 3.00       | Elongation factor Ts                                                               |
|                     |                        | 4V5I    | 3.10       | 16S rRNA, 23S rRNA, Trp-tRNA, GCP, Mg2+                                            |
|                     |                        | 4V5P    | 3.10       | 16S rRNA, 23S rRNA, Trp-tRNA                                                      |
|                     |                        | 4V5Q    | 3.10       | 16S rRNA, 3OS rp512, Trp-tRNA, GDP, KIR                                            |
|                     |                        | 4VSR    | 3.10       | 16S rRNA, Trp-tRNA, GDP, KIR                                                      |
|                     |                        | 4VSS    | 3.10       | 16S rRNA, GDP, KIR, Mg2+                                                           |
|                     |                        | 4V6Q    | 3.10       | 16S rRNA, 23S rRNA, Small protein B SMPB, tmRNA, GDP, KIR                         |
|                     |                        | 4V5G    | 3.60       | 16S rRNA, 23S rRNA, 3OS rp512, Thr-tRNA, GDP, KIR, Mg2+                           |
| **aEF1A**           | **Aeropyrum pernix**   |         |            |                                                                                   |
|                     |                        | 3VMF    | 2.30       | Peptide chain release factor subunit 1, GTP, Mg2+                                  |
|                     |                        | 3WXM    | 2.30       | Protein pelota homologue, GTP, Mg2+                                                |
| **Sulfolobus solfataricus** |         | 1JNY    |            | GD P                                                                               |
| **Oryctolagus cuniculus** |                   | 1SKQ    |            | GDP, Mg2+                                                                          |
| **eEF1A**           | **Saccharomyces cerevisiae** |       |            |                                                                                   |
|                     |                        | 1F60    |            | Elongation factor 18s                                                               |
|                     |                        | 2BTC    |            | Elongation factor-1 β                                                               |
|                     |                        | 1G7C    |            | Elongation factor-1 β, 5GP                                                          |
|                     |                        | 1UJE    |            | Elongation factor-1 β, GD P                                                         |
|                     |                        | 2BTC    |            | Elongation factor-1 β, GD P                                                         |
|                     |                        | 1UJF    |            | Elongation factor-1 β, GD P                                                         |

TAC: Tetracycline, ENX: Enacyloxin Ila, 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.


$$T_1 = \frac{1}{12} \left[ I^1 - I - \sum_{i=1}^{N_1} t_i^2 - t_i \right]$$ (19)

And

$$T_2 = \frac{1}{12} \left[ I^2 - I - \sum_{i=1}^{N_2} t_i^2 - t_i \right]$$ (20)

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

**Visualization**

$f_2(M), f_4(M)$, AUC and Spearman’s $\rho$ were visualized by the matplotlib Python package [45]. A three-dimensional structure was visualized by the VMD program [46].

**Results**

**Fitness between $f_i$ and $f_j$**

If $G_1, h_1, \tau$ and $A$ are same but $G$ is different, Table 2 shows that when $G = G'$, the AUC or the Spearman's $\rho$ is smaller than $G = \Gamma'G$. In the latter case, Figure 3 shows that when the time point increases, the AUC or the Spearman's $\rho$ tends to increase.

**Evaluation of predicted functional amino acid residues by $f_i$**

Figure 4A shows that $M \in M$ is classifiable in 4 by $f_1(M)$ and $f_3(M)$ using a receiver operating characteristic (ROC) curve [47] in Figure 4B. Figures 4C and 4D show that the left sides tend to have small $f_1(M)$ and small $f_3(M)$ but the right sides tend to have large $f_1(M)$ and large $f_3(M)$.

**Discussion**

Meanings of $f_1(M), f_2(M), f_3(M), f_4(M)$, AUC and Spearman’s $\rho$ are as follows. $f_1(M)$ becomes small when characters are only diverged in near to the root of the phylogenetic tree. $f_2(M)$ becomes large when characters are diverged in far from the root. $f_3(M)$ becomes small when at least one amino acid residue in $M$ is proximate from an ion or a molecule. $f_4(M)$ becomes large when amino acid residues in $M$ are not proximate from ions or molecules in all cocystal structures. If the AUC is 0.5, a correlation between $f_1(M)$ and being proximate and being non-proximate under a cutoff of $f_2(M)$ may not exist. If the AUC is close to 1, small $f_1(M)$ and large $f_2(M)$ correlate with being proximate and being non-proximate, respectively. If the AUC is close to 0, large $f_1(M)$ and small $f_2(M)$ correlate with being proximate and being non-proximate, respectively. If the Spearman's $\rho$ is 0, a linear correlation between $f_3(M)$ and $f_4(M)$ may not exist. If the Spearman's $\rho$ is close to 1 or -1, $f_3(M)$ and $f_4(M)$ have a positive or a negative linear correlation, respectively.

If $T=1$, $g_1 = g_5, h_1 = h_5, A = \tau A$ and $G = \Gamma G$, 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 $\rho$ is from 0.1982 to 0.1653 and the AUC is from 0.6093 to 0.6211 and the Spearman’s $\rho$ 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 $\rho$. Therefore, in the above case, if $G = G'$ is changed to $G = \Gamma G$ in the former and the latter, Figure 3 shows that the AUC is from 0.6941 to 0.7349 and the Spearman’s $\rho$ is from 0.4981 to 0.5650 and the AUC is from 0.6846 to 0.7335 and the Spearman’s $\rho$ 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’s $\rho$.

If $T=N$, $g_2 = g_6, h_2 = h_6, A = \tau A$ and $G = \Gamma G$, the method is the iv-ET method [17]. If $T=N, g_3 = g_7, h_3 = h_7, A = \tau A$ and $G = \Gamma G$, the method is equivalent to the rv-ET method [17]. If $G = \Gamma G$ is changed to $G = \Gamma G$ in the former and the latter, Table 2 shows that the AUC is from 0.5896 to 0.6242 and the Spearman’s $\rho$ is from 0.1221 to 0.3650 and the AUC is from 0.6180 to 0.7417 and the Spearman’s $\rho$ is from 0.1308 to 0.5722, 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’s $\rho$.

**Figure 2: Computations of $f_i$ and $f_j$**

(A) $N$ sequences and $K$ structures are extracted from the Swiss-Prot and the PDB, respectively. After all the sequences are aligned, $f_1(M)$ and $f_3(M)$ are computed by (B) and (C), respectively.

(B) After a phylogenetic tree is written from sequences, $f_2(M)$ is computed.

(C) In structure $k$, $R$ and $Q$ denote coordinates of an amino acid residue and coordinates of ions or molecules, respectively. After proximity of $R$ and $Q$ is measured as $f_5(M)$ and computed on $K$ structures, $f_5(M)$ is computed.
| $g_i$ | $h_j$ | $A$ | $iG$ | $r$ | AUC | Spearman's $\rho$ |
|---|---|---|---|---|---|---|
| $20$ | $j$ | $iG$ | $0<\tau<1$ | $0.5896$ | $0.6242$ | $0.1221$ |
| | | $iG$ | | | | $0.3650$ |
| | $9$ | $iG$ | $0<\tau<1$ | $0.5700$ | $0.6184$ | $0.1015$ |
| | | | | | | $0.3509$ |
| | $20$ | $iG$ | $0.1$ | $0.6036$ | $0.6376$ | $0.1436$ |
| | | $iG$ | $0.2$ | $0.7207$ | $0.7412$ | $0.1977$ |
| | | | $0.3$ | | $0.7374$ | $0.2015$ |
| | | | $0.4$ | | $0.7037$ | $0.1585$ |
| | $9$ | $iG$ | $0.2$ | $0.5757$ | $0.6023$ | $0.1715$ |
| | | $iG$ | $0.3$ | $0.6151$ | $0.6391$ | $0.1715$ |
| | | | $0.4$ | $0.5885$ | | $0.1160$ |
| | | $iG$ | $0.1$ | | $0.6997$ | $0.5185$ |
| | | $iG$ | $0.2$ | | $0.6984$ | $0.5097$ |
| | | | $0.3$ | | $0.7084$ | $0.5083$ |
| | | | $0.4$ | | $0.6854$ | $0.4797$ |
| | $9$ | $iG$ | $0.1$ | $0.5901$ | $0.6052$ | $0.1288$ |
| | | $iG$ | $0.2$ | $0.6366$ | $0.6391$ | $0.2315$ |
| | | | $0.3$ | | $0.6865$ | $0.2763$ |
| | | $iG$ | $0.4$ | | $0.6758$ | $0.4474$ |
| | | $iG$ | $0.1$ | | $0.6975$ | $0.4734$ |
| | | | $0.2$ | | $0.6905$ | $0.4823$ |
| | | | $0.3$ | | $0.6888$ | $0.4922$ |
| | | $iG$ | $0.4$ | | $0.6602$ | $0.4470$ |
| | | $iG$ | $0.1$ | | $0.6805$ | $0.4599$ |
| | | | $0.2$ | | $0.6888$ | $0.4675$ |
| | $20$ | $iG$ | $0.1$ | | $0.5916$ | $0.5780$ |
| | | $iG$ | $0.2$ | | $0.7399$ | $0.5780$ |
| | | | $0.3$ | | | |
| | | | $0.4$ | | | |
| | | $iG$ | $0.5$ | | | |
| | $9$ | $iG$ | | | $0.5652$ | $0.0587$ |
| | | $iG$ | | | $0.7020$ | $0.5145$ |
| | | $iG$ | $0.1$ | | | |
| | | | $0.2$ | | | |
| | | | $0.3$ | | | |
| | | | $0.4$ | | | |
| | | $iG$ | $0.5$ | | | |
| | | $iG$ | $0.1$ | | $0.5890$ | $0.1257$ |
| | | | $0.2$ | | $0.7012$ | $0.5691$ |
| | | | $0.3$ | | $0.6225$ | $0.1579$ |
| | | | $0.4$ | | $0.7287$ | $0.5517$ |
| | $9$ | $iG$ | | | | $0.6138$ |
| | | $iG$ | | | | $0.7265$ | $0.1412$ |
| | | | | | | $0.5501$ |
Table 2: Correlations between $f_i$ and $f'_i$.

| $h_i$ | $g_i$ | $f_i$ | $f'_i$ | $\rho$ | $\rho$ |
|-------|-------|-------|-------|-------|-------|
| $h_1$ | $g_1$ | $f_1$ | $f'_1$ | 0.5792 | 0.3786 |
| $h_1$ | $g_2$ | $f_2$ | $f'_2$ | 0.6147 | 0.7393 |
| $h_2$ | $g_3$ | $f_3$ | $f'_3$ | 0.5992 | 0.7059 |
| $h_3$ | $g_4$ | $f_4$ | $f'_4$ | 0.6211 | 0.7335 |

Figure 3: Dependence on time points. The time point is $T$ in Eq. (1). (A) (B) AUC and (C) (D) Spearman's $\rho$ were computed using $g_i$, $h_i$, $f_i$, and $f'_i$ as shown in the figures.
latter, considering that each gap is different is effective for improving the AUC and the Spearman’s ρ. Thus, $f_1(M)$ is evaluable by $f_2(M)$ 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 $f_2(M)$ is small, an amino acid residue in $M$ is proximate from a region involving protein biosynthesis. If $f_3(M)$ is large, the amino acid residues in $M$ 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 $f_1(M)$, the AUC is 0.742, which indicates that the proximate region tends to become small $f_1(M)$ but the non-proximate region tends to become large $f_1(M)$. In addition, Table 2 shows that the Spearman’s ρ is 0.5722, which indicates that $f_1(M)$ tends to be small if $f_2(M)$ is small and $f_3(M)$ tends to be large if $f_2(M)$ is large. However, a complete linear correlation between $f_1(M)$ and $f_2(M)$ was not obtainable and therefore not all of $f_1(M)$ can explain $f_2(M)$. This may indicate that $f_1(M)$ and $f_2(M)$ can measure a similar thing each other but cannot always measure a same thing and, by $f_1(M)$ and $f_3(M)$, 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, $f_1(M)$ and $f_2(M)$ 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 meaning 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, Glaister 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.
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: 33-38.

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.