A new bioinformatics method for predicting active sites of multifunctional proteins

Yosuke Kondo

A dissertation
submitted to Department of Pharmaceutical Sciences,
Graduate School of Pharmaceutical Sciences,
Tokyo University of Science
in candidacy for the degree of
doctor of philosophy

Abstract

In order to develop new drugs more efficiently, it is important to understand how a drug is bound to a target molecule. Proteins often become a target for drugs but it is hard to identify which amino acid dominates the total binding energy. For identifying majorly contributory amino acids, we need site-directed mutagenesis, which requires mutant proteins. However, mutating amino acids one by one takes a large amount of costs. Therefore, we need to develop a method for narrowing down the candidate amino acids.

This thesis aims at getting rid of such a time-consuming task for biological or biochemical experiments by improving prediction methods for protein functional sites. So far, lots of such prediction methods have been developed and the prediction performance has been examined by using various proteins. However, existing methods have some problems in regard with how to explain or evaluate the methods. In order to solve the problems, this thesis suggests two things. The first one is an idea to integrate some existing methods. As described above, there exist many protein functional site prediction methods. However, each method is explained by each explanation and is not generalized. It is difficult to investigate how much the performance is changed if how or what we change the methods. We regard how to treat amino acid residues as parameters and then compare the methods. In addition, alignment gaps are also treated as a parameter. This means that how to treat gaps is different in each existing method and therefore this study investigates how much the performance is changed by changing how to treat gaps. Meanwhile, in addition to prediction of important sites for a whole protein family, we consider a method to predict important sites of a particular subfamily. We thereby also discuss conservation of amino acids from a different point of view. The second one is improving evaluation methods. By using a three-dimensional complex structure of protein-ion, protein-small compound, protein-RNA or protein-protein, we can investigate whether the binding sites are predicted or not. However, existing evaluation methods use one of the complex structures. This implies that we cannot always evaluate a protein binding with various molecules and the evaluation methods therefore are not appropriate for a multifunctional protein. This study proposes a method to utilize multiple structures for evaluation.

For evaluation of prediction performance of existing or proposed methods, translation elongation factor 1A (EF1A) is used. EF1A was discovered as a factor promoting protein biosynthesis in the 1960s.

In the 1990s, EF1A was also found as an actin binding protein and further analyses found that EF1A involves many functions such as nuclear export, protein degradation, apoptosis and virus functions. In order to investigate how each function works and how the functions relate to each other, identifications of the functional sites are required. EF1A is therefore a multifunctional protein and we cannot deny a possibility that there are undiscovered functional sites. In this study, such unknown functional sites are also predicted by the proposed method.

This thesis consists of six chapters. Chapter 1 describes introduction of this study. Chapter 2 defines basic words and symbols. Chapter 3, 4 or 5 introduces a new method and applies the method to EF1A. Chapter 6 describes conclusions of this study.

Chapter 1 described backgrounds and significances of this study and explained features of proteins and the reason why we need informatics methods.

Chapter 2 defined terminologies and notations such as amino acid symbols, multiple sequence alignment and existing or proposed conservation analysis. The proposed method provides two measures. The first one is a value to measure how much the site is conserved and investigates two types of conservations by considering a site which is conserved in the whole family or conserved in a particular subfamily but not in the whole family. The second one is a value to measure how much the site is proximate from ions or molecules on multiple three-dimensional structures. Finally, how to estimate correlations between the measures was described in order to evaluate that the conservation can distinguish whether or not the site is proximate from ions or molecules.

In Chapter 3, the measures defined in Chapter 2 were used for analysis of EF1A. One of the conservations showed that when the conservation is visualized onto three-dimensional structures, the conservation can detect binding sites of aminoacyl tRNA, release factor subunit 1, pelota homologue, EF-Ts, EF-1B, GTP, GDP, pulvomycin, thiocillin, kirromycin and magnesium ion. However, this evaluation showed appearances visualized onto the three-dimensional structures and therefore has not achieved numerical evaluation. By using the proximity defined in Chapter 2, the proposed methods and existing methods are evaluated by comparing correlations between the conservation and the proximity. This showed that the proposed method has a higher performance than existing methods. However, we did not know what made improving the performances. By using the idea to integrate existing methods, we investigate why the performances are increased. The results showed that there are two important things. The first one is how to treat alignment gaps. Although gaps have been treated as a same thing in most of the existing methods, the performances were increased when the gaps are treated as different things. The second one is utilization of a phylogenetic tree. The performances were increased when the conservation was largely changed whether a site is diverged in near to the root or near to the leaf node. Finally, a threshold of the conservation was determined from the evaluation results and then determined predicted functional sites.

In Chapter 4, EF1A was analyzed by using a specific conservation. This means that although in Chapter 3 we predicted a site conserved in a whole protein family, in Chapter 4 we consider a site conserved in a particular subfamily but not in the whole family. As a result, the specific conservation could detect known amino acid residues involving functional divergences. This implies that the specific conservation is useful for comprehensive functional site prediction of EF-Tu/EF1A.

In Chapter 5, the binding site of FNIII14 was investigated by using three-dimensional structural information. FNIII14 involves anoikis, which is apoptosis triggered by cell detachment from the extracellular matrix, by binding membrane-associated EF1A. This implies that because EF1A involves not only protein biosynthesis but also cell death, it is very important for investigating the crosstalk between these cellular functions. The complex structure was predicted by using a docking tool. This showed that there are complement sites for FNIII14 in the structure of EF1A. Then, we used a protein simulation method and investigate stability of the complex and stable binding amino acid residues. Furthermore, by combining the results of Chapter 3, we narrowed down important amino acid residues for binding with FNIII14.

Chapter 6 described conclusions and future works. This study could develop an original method integrating some existing methods for predicting protein functional sites and introduce a measure which can evaluate such methods more accurately by using multiple structures. By this evaluation, the proposed method showed a higher performance than the existing methods. Furthermore, this study introduced a specific conservation and investigated functional divergences of EF-Tu/EF1A. This showed that the specific conservation could detect amino acid residues involving functional divergences of EF-Tu/EF1A. In addition, we predicted complex structures of EF1A and FNII14 using structural analyses and narrowed down candidate amino acid residues by combining the results of above conservation analysis. At present, based on the prediction results, biochemical experiments are to be conducted. FNIII14 is useful for increasing effects of anticancer drugs by the detachment activity. It is expected that our results expedite development of a more effective FNIII14-like substance or a small compound which has a similar effect of FNIII14. Additionally, such drugs are applied to cancers which have resistance of anticancer drugs and solve the clinical problem. Meanwhile, the proposed prediction methods are useful for any proteins. Especially, the conservation-based method is useful for various situations. In other words, if we have only one protein sequence, the method is executable. The results of this study showed that our method is effective for EF1A but is only applied for EF1A. Therefore, it becomes a next issue that our method is applied for various proteins and investigated how much the prediction performance is.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The basis of pharmacological activity is interaction of drugs with biological molecules such as deoxyribonucleic acids (DNA), ribonucleic acids (RNA), proteins, lipids and sugars. Although such molecules can become a target for drugs, most of the drug targets are proteins. Therefore, understanding how proteins bind drugs enables us to develop more valid drugs.

Proteins are a macromolecule whose unit is made from an amino acid, which simultaneously has two chemical functional groups, an amino group (-NH_2) and a carboxyl group (-COOH). Two amino acids are dimerized by constructing an amide group (-NHCO-), which is obtained by dehydration of the amino group and the carboxyl group, and connections of many amino acids by amide bonds make a protein. Meanwhile, an amino acid has leftovers of the amino group and carboxyl group. Basically, because there are twenty types of the leftovers, there exist twenty standard amino acids, which are often represented as twenty alphabets (see Chapter 2.1) and a line of the alphabets is regarded as a protein sequence or a primary structure.

Although we can measure a distance between two amino acids on the sequence, all the distances are not always linear with a distance on the three-dimensional space because some proteins are spontaneously folded and construct a tertiary structure. Folding is one of the basic features of proteins and some biological activities are triggered by interaction of a folded protein and ions or molecules. The interaction is a physical contact between both and therefore there exist binding sites of ions or molecules in proteins. Identification of binding sites is important to investigate how proteins bind ions or molecules and how drugs are bound to proteins.

In order to identify such binding sites, one of the effective analyses is site-directed mutagenesis. In this analysis, a mutant protein, whose amino acid residue is mutated into another one, is prepared and then differences between the mutant and the wild type are investigated by measuring binding affinities of the protein and binding molecules. However, this analysis requires a large amount of time and costs because of necessity of mutating amino acid residues one by one. In order to reduce the possible amino

acids, we need a method to narrow down the amino acid residues.

For developing such methods, there exists two types of useful data. The first one is sequence information. A protein sequence is derived from a gene, which is a DNA sequence, and therefore the amino acid mutation is regulated by the gene mutation, which is a point mutation occurred in the DNA sequence. Considering how a gene has been mutated in the evolutionary process enables us to hypothesize the relationships between the protein evolution and function as a mutation rate of an amino acid is changed by whether the amino acid involves the protein function or not, respectively. Based on this hypothesis, binding sites should be predicted by investigating differences of the mutation rates on sequence information. The second one is three-dimensional structural information. One of the effective utilizations of protein structures is energy estimation from interatomic forces. This enables us to calculate stabilities of complex structures such as protein-ion, protein-DNA, protein-RNA, protein-protein and protein-small compound. We can thereby obtain stable amino acids largely accounting for the total binding energy in the complex structure.

Many sequence or structural data are freely available for anyone because many biological databases are open to the public. However, because of existence of a huge amount of biological data, we should consider a method to extract useful information. In this study, such objectives are achieved by developing a new informatics method and improving the prediction performance.

The contents of this thesis are as follows. Chapter 1 describes backgrounds and significances of this study. Chapter 2 describes notations of fundamental elements. In Chapter 3, conservation analyses are conducted and then existing methods and proposed methods are compared. In Chapter 4, from a different point of view of Chapter 3, conservation analyses are conducted by using a specific conservation. In Chapter 5, the results of Chapter 3 are combined with structural analysis. Chapter 6 describes conclusions and future works.

Chapter 2

Notations of fundamental elements

2.1 Amino acid symbols

Standard amino acids are often represented as twenty alphabets: A (alanine), C (cysteine), D (aspartic acid), E (glutamic acid), F (phenylalanine), G (glycine), H (histidine), I (isoleucine), K (lysine), M (methionine), N (asparagine), P (proline), Q (glutamine), R (arginine), S (serine), T (threonine), V (valine), W (tryptophan) and Y (tyrosine). In this study, these alphabets are considered to as elements of a field of set, $\mathcal{A}$, which is, for example, definable as

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

or

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

2.2 Multiple sequence alignment

A line of the symbols can be regarded as a protein sequence. A protein sequence comes from a gene, which consists of DNAs (each of which is linked by phosphodiester bond) and sometimes whose DNA has been mutated, inserted or deleted in the evolutionary process. Such gene dynamics can generate plenty of protein sequences. In order to compare the sequences, we often put the symbols in a row and create a sequence alignment. If we aligned two sequences, the alignment is called as a pairwise sequence alignment. If we aligned three or more sequences, the alignment is called as a multiple sequence alignment (MSA). Let $\mathcal{M} = (m_{jk})$ denote a given MSA and here m_{jk} denote an amino acid symbol of site k on sequence j in the MSA. Let ${}_i\mathcal{M} = [m_{1i}, m_{2i}, \ldots, m_{ni}]^t$ be column i on the MSA (see Figure 2.1A) and refer to as an alignment site

2.3 Quantification of alignment sites

In order to extract useful information or objectively compare each alignment site, we often convert ${}_i\mathcal{M}$ into a value (see Figure 2.2A). Therefore, let us consider a mapping $f_x : \mathcal{M} \to [\, 0, \infty\,)$, in other words, we associate ${}_i\mathcal{M}$ with a non-negative real number.

2.4 Existing conservation analysis

The mapping f_x enables us to describe some existing methods. For example, the Shannon entropy (SE) based method [99, 101] is defined as

$$f_{se}\left({}_i\mathcal{M}\right) = -\sum_{x\in\mathcal{A}} p_i\left(x\right)\ln p_i\left(x\right), \tag{2.1}$$

where $p_i\left(x\right)$ is a probability of amino acids in ${}_i\mathcal{M}$. The sum of pairs (SP) based method [116] is defined as

$$f_{sp}\left({}_i\mathcal{M}\right) = \frac{1}{N}\sum_{l=1}^{N}\sum_{m=1}^{N} w\left(l\right) S\left(x_l, x_m\right), \tag{2.2}$$

where N is the number of sequences, $w\left(l\right)$ is a weight of sequence l and

$$S\left(x,y\right) = \frac{m\left(x,x\right) - m\left(x,y\right)}{m\left(x,x\right)}, \tag{2.3}$$

where $m\left(x,y\right)$ is a substitution score of amino acids x and y. The integer-valued evolutionary trace (iv-ET) method [81] is defined as

$$f_{iet}\left({}_i\mathcal{M}\right) = 1 + \sum_{l=1}^{N-1}\begin{cases} 0 & \text{(if site } i \text{ conserved within each group } g\text{)} \\ 1 & \text{(otherwise)}\,, \end{cases} \tag{2.4}$$

where g is determined from a phylogenetic tree which is reconstructed under the hypothesis that the evolutionary rate is constant. The real-valued ET (rv-ET) method [81] is defined as

$$f_{ret}\left({}_i\mathcal{M}\right) = 1 + \sum_{l=1}^{N-1}\frac{1}{l}\sum_{g=1}^{l}\left[-\sum_{x\in\mathcal{A}} p_{ig}\left(x\right)\ln p_{ig}\left(x\right)\right], \tag{2.5}$$

where $p_{ig}\left(x\right)$ is a probability of amino acids in group g of ${}_i\mathcal{M}$.

2.5 Proposed conservation analysis

This section defines notations of the proposed method. Let ${}_i^t\mathcal{M} \in {}_i\mathcal{M}$ denote ${}_i\mathcal{M}$ at time point $t = 1, 2, \ldots, N+1$, where N is the number of internal nodes on a phylogenetic tree reconstructed from the

given MSA (see Figure 2.1A or 2.2B), and be represented by a field of sets. For example, ${}^{t}_{4}\mathcal{M}$;

- ${}^{1}_{4}\mathcal{M} = \{\{R, R, L, R, R, R\}\}$
- ${}^{2}_{4}\mathcal{M} = \{\{R, R, L, R, R\}, \{R\}\}$
- ${}^{3}_{4}\mathcal{M} = \{\{R, R, L, R\}, \{R\}, \{R\}\}$
- ${}^{4}_{4}\mathcal{M} = \{\{R, R, L\}, \{R\}, \{R\}, \{R\}\}$
- ${}^{5}_{4}\mathcal{M} = \{\{R, R\}, \{L\}, \{R\}, \{R\}, \{R\}\}$.

Let there be $g_x : {}_{i}\mathcal{M} \rightarrow [0, 1]$, which here associates ${}^{t}_{i}\mathcal{M}$ with 1 if there exists ${}^{t}_{i}\mathcal{M}_u \in {}^{t}_{i}\mathcal{M}$ which comprises two or more types of characters and, in the other cases, associates ${}^{t}_{i}\mathcal{M}$ with 0. By $g_x\left({}^{1}_{i}\mathcal{M}\right)$ only, ${}_{1}\mathcal{M}$ – ${}_{5}\mathcal{M}$ are indistinguishable. If $g_x\left({}^{1}_{i}\mathcal{M}\right)$ and $g_x\left({}^{2}_{i}\mathcal{M}\right)$ are summed, ${}_{1}\mathcal{M}$ and the others are distinguishable. If $g_x\left({}^{1}_{i}\mathcal{M}\right)$, $g_x\left({}^{2}_{i}\mathcal{M}\right)$ and $g_x\left({}^{3}_{i}\mathcal{M}\right)$ are summed, ${}_{1}\mathcal{M}$, ${}_{2}\mathcal{M}$ and the others are distinguishable. Therefore, let

$$f_1\left({}_{i}\mathcal{M}\right) := \sum_{t=1}^{T} g_x\left(h_x\left({}^{t}_{i}\mathcal{M}, \mathcal{A}, {}_{i}\mathcal{G}\right)\right), \tag{2.6}$$

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

As shown in Figure 2.1B, let $h_x : {}^{t}_{i}\mathcal{M} \rightarrow [0, 1]$ be included in g_x and $g_x\left({}^{t}_{i}\mathcal{M}\right)$ be represented as following three definitions:

$$g_1\left({}^{t}_{i}\mathcal{M}\right) := \begin{cases} 0 & \left(\forall {}^{t}_{i}\mathcal{M}_u \in {}^{t}_{i}\mathcal{M}, h_x\left({}^{t}_{i}\mathcal{M}_u\right) \leq \tau\right) \\ 1 & \left(\exists {}^{t}_{i}\mathcal{M}_u \in {}^{t}_{i}\mathcal{M}, h_x\left({}^{t}_{i}\mathcal{M}_u\right) > \tau\right), \end{cases} \tag{2.7}$$

where τ is a threshold of $h_x\left({}^{t}_{i}\mathcal{M}_u\right)$,

$$g_2\left({}^{t}_{i}\mathcal{M}\right) := \frac{1}{\mid {}^{t}_{i}\mathcal{M} \mid} \sum_{{}^{t}_{i}\mathcal{M}_u \in {}^{t}_{i}\mathcal{M}} h_x\left({}^{t}_{i}\mathcal{M}_u\right), \tag{2.8}$$

where $\mid {}^{t}_{i}\mathcal{M} \mid$ is the number of multisets in ${}^{t}_{i}\mathcal{M}$ and

$$g_3\left({}^{t}_{i}\mathcal{M}\right) := h_x\left({}^{t}_{i}\mathcal{M}_{*u}\right), \tag{2.9}$$

where ${}^{t}_{i}\mathcal{M}_{*u}$ is a multiset which is separated at time point $t + 1$. For example, ${}^{t}_{4}\mathcal{M}_{*u}$;

- ${}^{1}_{4}\mathcal{M}_{*u} = \{R, R, L, R, R, R\}$
- ${}^{2}_{4}\mathcal{M}_{*u} = \{R, R, L, R, R\}$
- ${}^{3}_{4}\mathcal{M}_{*u} = \{R, R, L, R\}$
- ${}^{4}_{4}\mathcal{M}_{*u} = \{R, R, L\}$

- ${}^{5}_{4}\mathcal{M}_{*u} = \{R, R\}$.

In order to consider alignment gaps, which involve in insertion or deletion of amino acids, let ${}_{i}\mathcal{G} \subset {}_{i}\mathcal{M}$ denote a field of sets of gaps in ${}_{i}\mathcal{M}$. ${}_{i}\mathcal{G}$ enables as to treat alignment gaps as a parameter of f_1. For example, ${}_{i}\mathcal{G}$ is definable as ${}^{1}_{i}\mathcal{G} = \left\{\left\{{}^{1}_{i}\gamma, {}^{2}_{i}\gamma, \ldots, {}^{G}_{i}\gamma\right\}\right\}$ or ${}^{G}_{i}\mathcal{G} = \left\{\left\{{}^{1}_{i}\gamma\right\}, \left\{{}^{2}_{i}\gamma\right\}, \ldots, \left\{{}^{G}_{i}\gamma\right\}\right\}$, where G is the number of gaps. By using above $\mathcal{A}$ and ${}_{i}\mathcal{G}$, let us define $h_x\left({}^{t}_{i}\mathcal{M}_u\right)$ as following three definitions:

$$h_1\left({}^{t}_{i}\mathcal{M}_u\right) := \begin{cases} 0 & (\forall l \in {}^{t}_{i}\mathcal{M}_u, \exists \mathcal{X} \in \mathcal{A} \cup {}_{i}\mathcal{G}; l \in \mathcal{X}) \\ 1 & (otherwise)\,, \end{cases} \tag{2.10}$$

$$h_2\left({}^{t}_{i}\mathcal{M}_u\right) := -\sum_{\mathcal{X} \in \mathcal{A} \cup {}_{i}\mathcal{G}} p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) \log_{|\mathcal{A} \cup {}_{i}\mathcal{G}|} p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right), \tag{2.11}$$

where $|\ \mathcal{A} \cup {}_{i}\mathcal{G}\ |$ is the number of sets in $\mathcal{A} \cup {}_{i}\mathcal{G}$ and

$$p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) = \frac{1}{|\ {}^{t}_{i}\mathcal{M}_u\ |} \sum_{l \in {}^{t}_{i}\mathcal{M}_u} \begin{cases} 0 & (l \notin \mathcal{X}) \\ 1 & (l \in \mathcal{X})\,, \end{cases} \tag{2.12}$$

where $|\ {}^{t}_{i}\mathcal{M}_u\ |$ is the number of characters in ${}^{t}_{i}\mathcal{M}_u$ and if $p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) = 0$, $p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) \log_{|\mathcal{A} \cup {}_{i}\mathcal{G}|} p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right)$ is regarded as 0 and

$$h_3\left({}^{t}_{i}\mathcal{M}_u\right) := \frac{1}{|\ {}^{t}_{i}\mathcal{M}_u\ |} \sum_{l \in {}^{t}_{i}\mathcal{M}_u} \sum_{m \in {}^{t}_{i}\mathcal{M}_u} w\left(l\right) s\left(l, m\right), \tag{2.13}$$

where $w\left(l\right)$ is a weight of sequence l and

$$s\left(l, m\right) = \begin{cases} 0 & (l \in \mathcal{X} \in {}_{i}\mathcal{G} \wedge m \in \mathcal{Y} \in {}_{i}\mathcal{G} \wedge \mathcal{X} = \mathcal{Y}) \\ \frac{S_{max} - S_{min}}{S_{max}} & (l \in \mathcal{X} \in {}_{i}\mathcal{G} \wedge m \in \mathcal{Y} \in {}_{i}\mathcal{G} \wedge \mathcal{X} \neq \mathcal{Y}) \\ \frac{S_{max} - S_{min}}{S_{max}} & (l \in \mathcal{X} \in {}_{i}\mathcal{G} \wedge m \in \mathcal{Y} \in \mathcal{A}) \\ \frac{S(l,l) - S_{min}}{S(l,l)} & (l \in \mathcal{X} \in \mathcal{A} \wedge m \in \mathcal{Y} \in {}_{i}\mathcal{G}) \\ \frac{S(l,l) - S(l,m)}{S(l,l)} & (l \in \mathcal{X} \in \mathcal{A} \wedge m \in \mathcal{Y} \in \mathcal{A})\,, \end{cases} \tag{2.14}$$

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

Table 2.1 shows that f_1 can be computed by determining the seven parameters. Table 2.2 shows that existing methods can be defined by the parameters of f_1.

2.6 Specific conservation

In this study, amino acid conservation are considered from a different point of view. This means that we consider amino acids are conserved in a particular subfamily but not conserved in the whole family

Table 2.1: Seven parameters for computation of f_1

Parameter	Value	Description
T	$1, 2, \ldots, N$	a time point
g_x	g_1, g_2, g_3	$g_x : {}_i\mathcal{M} \rightarrow [0, 1]$
τ	[0,1]	a threshold of g_1
h_x	h_1, h_2, h_3	$h_x : {}_i^t\mathcal{M} \rightarrow [0, 1]$
$\mathcal{A}$	${}^{20}\mathcal{A}, {}^{9}\mathcal{A}$	a field of sets of amino acids
${}_i\mathcal{G}$	${}_i^1\mathcal{G}, {}_i^G\mathcal{G}$	a field of sets of gaps in site i
$w\,(l)$	$(0, \infty)$	a weight of sequence l

Table 2.2: Relationships of the parameters and prediction methods

Mapping	T	g_x	τ	h_x	$\mathcal{A}$	${}_i\mathcal{G}$	$w\,(l)$	Reference
f_{se}	1	g_3	–	h_2	${}^{20}\mathcal{A}$	${}_i^1\mathcal{G}$	–	[99]
f_{se}	1	g_3	–	h_2	${}^{9}\mathcal{A}$	${}_i^1\mathcal{G}$	–	[101]
f_{sp}	1	g_3	–	h_3	$\mathcal{A}$	${}_i^1\mathcal{G}$	unused	[116]
f_{sp}	1	g_3	–	h_3	$\mathcal{A}$	${}_i^1\mathcal{G}$	used	[116]
f_{iet}	N	g_1	$\tau \in (0, 1)$	h_1	${}^{20}\mathcal{A}$	${}_i^1\mathcal{G}$	–	[81]
f_{ret}	N	g_2	–	h_2	${}^{20}\mathcal{A}$	${}_i^1\mathcal{G}$	–	[81]

and a specific conservation of base group a and target group b is defined as

$$C\left({}_i^a\mathcal{M}, {}_i^b\mathcal{M}\right) = \frac{g_3\left({}_i^a\mathcal{M}\right) + 1}{g_3\left({}_i^b\mathcal{M}\right) + 1} - 1, \tag{2.15}$$

where $a < b$.

2.7 A proximity on three-dimensional structures

In order to evaluate above conservations, we next consider another quantification (see Figure 2.2A). In this section, ${}_i\mathcal{M}$ is not regarded as character type of data but coordinate type. Let $\mathbb{R}$ denote a set of real numbers and there be $e : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow [\,0, \infty\,)$. Let $\mathcal{R} \subset \mathbb{R}^3$, $\mathcal{Q} \subset \mathbb{R}^3$ and

$$e\left(\mathcal{R}, \mathcal{Q}\right) := \min_{(\boldsymbol{r}, \boldsymbol{q}) \in \mathcal{R} \times \mathcal{Q}} \left(\parallel \boldsymbol{r} - \boldsymbol{q} \parallel_2\right), \tag{2.16}$$

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

Let us consider structure k, which contains a protein and ions or molecules (see Figure 2.2C). Let ${}_i^k\mathcal{R} \subset \mathbb{R}^3$ denote atomic coordinates of amino acid residue i in structure k and ${}^k\mathcal{Q} \subset \mathbb{R}^3$ denote atomic coordinates of ions or molecules in structure k. Let K denote the number of structures and the sequences be aligned. Let $\left\{{}_i^1\mathcal{R}, {}_i^2\mathcal{R}, \ldots, {}_i^{K-G}\mathcal{R}\right\} \subseteq {}_i\mathcal{M}$ denote a set of residues in ${}_i\mathcal{M}$ and $\left\{{}_i^1\gamma, {}_i^2\gamma, \ldots, {}_i^G\gamma\right\} = {}_i\mathcal{G} \subset {}_i\mathcal{M}$ denote a set of gaps in ${}_i\mathcal{M}$. Let

$$f_2\left({}_i\mathcal{M}\right) := \min_{{}_i^k\mathcal{R} \in {}_i\mathcal{M} \setminus {}_i\mathcal{G}} \left[e\left({}_i^k\mathcal{R}, {}^k\mathcal{Q}\right)\right]. \tag{2.17}$$

2.8 Correlations between f_1 and f_2

In order to evaluate f_1 by f_2, we consider two types of correlations between f_1 and f_2 based on Spearman's ρ [106] and an ROC curve.

2.8.1 Spearman's ρ

Let $\mathcal{F}_x \ni f_x\left({}_i\mathcal{M}\right)$ denote a multiset of $f_x\left({}_i\mathcal{M}\right)$ and be represented as $\mathcal{F}_x \ni {}_i^1V_x \leq {}_i^2V_x \leq \cdots \leq {}_i^IV_x$, where I is the number of ${}_i\mathcal{M}$. Let r denote a rank function and

$$r\left({}_i^{j+k-1}V_x\right) = j - 1 + \frac{t_n + 1}{2}, \tag{2.18}$$

where $j = 1, 2, \ldots, I$, $k = 1, 2, \ldots, t_n$ and t_n is a size of the tied rank. Here, a Spearman's ρ [106] is defined as

$$\rho = \frac{T_1 + T_2 - \sum_{i=1}^{I}\left[r\left({}_i^lV_1\right) - r\left({}_i^mV_2\right)\right]^2}{2\sqrt{T_1T_2}}, \tag{2.19}$$

where $l = 1, 2, \ldots, I$, $m = 1, 2, \ldots, I$,

$$T_1 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_1}\left(t_n^3 - t_n\right)}{12} \tag{2.20}$$

and

$$T_2 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_2}\left(t_n^3 - t_n\right)}{12}, \tag{2.21}$$

where N_1 and N_2 are the numbers of tied ranks in $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively.

2.8.2 ROC curve

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

$$t_j \begin{cases} < v_1 & (j = 0) \\ = \dfrac{v_j + v_{j+1}}{2} & (j = 1, 2, \ldots, J-1) \\ > v_J & (j = J)\,. \end{cases} \tag{2.22}$$

Let c_2 denote a cutoff of $f_2\left({}_i\mathcal{M}\right)$. Let I_f denote the number of ${}_i\mathcal{M}$ which satisfies $f_2\left({}_i\mathcal{M}\right) > c_2$ and I_t denote the number of ${}_i\mathcal{M}$ which satisfies $f_2\left({}_i\mathcal{M}\right) \leq c_2$. Let $I_{fp}\left(t_j\right)$ denote the number of ${}_i\mathcal{M}$ which satisfies $f_2\left({}_i\mathcal{M}\right) > c_2$ and $f_1\left({}_i\mathcal{M}\right) \leq t_j$ and $I_{tp}\left(t_j\right)$ denote the number of ${}_i\mathcal{M}$ which satisfies

$f_2\left({}_i\mathcal{M}\right) \leq c_2$ and $f_1\left({}_i\mathcal{M}\right) \leq t_j$. Let a false positive rate

$$p\left(t_j\right) = \frac{I_{fp}\left(t_j\right)}{I_f}, \tag{2.23}$$

a true positive rate

$$q\left(t_j\right) = \frac{I_{tp}\left(t_j\right)}{I_t}, \tag{2.24}$$

and an area under the curve

$$AUC = \frac{1}{2}\sum_{j=0}^{J-1}\left[p\left(t_{j+1}\right) - p\left(t_j\right)\right] \cdot \left[q\left(t_{j+1}\right) + q\left(t_j\right)\right]. \tag{2.25}$$

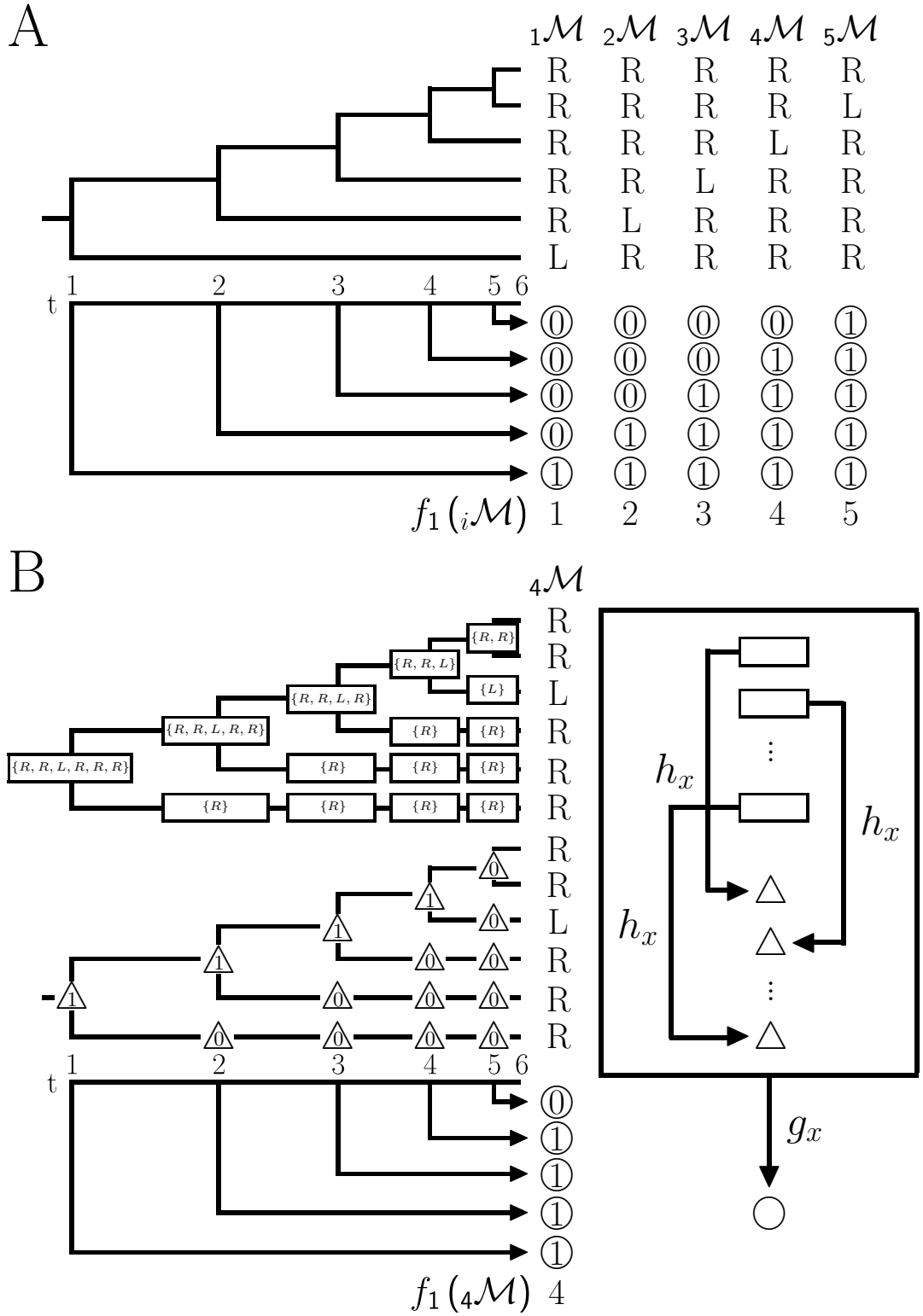


Figure 2.1: **Idea of a mapping by character type.** (A) An idea of f_1. $_1\mathcal{M}$ – $_5\mathcal{M}$ are comprised of 5 R and 1 L and each character attaches a leaf node of a rooted phylogenetic tree under a hypothesis that the evolutionary rate is constant. Numbers in ascending order are assigned from the root to leaf nodes as time point t. In f_1, after a value in a circle is assigned to $_i\mathcal{M}$, values in circles are summed. (B) Ideas of g_x and h_x. In g_x, after h_x associates characters in a square to a value in a triangle, values in triangles are associated to a value in a circle.

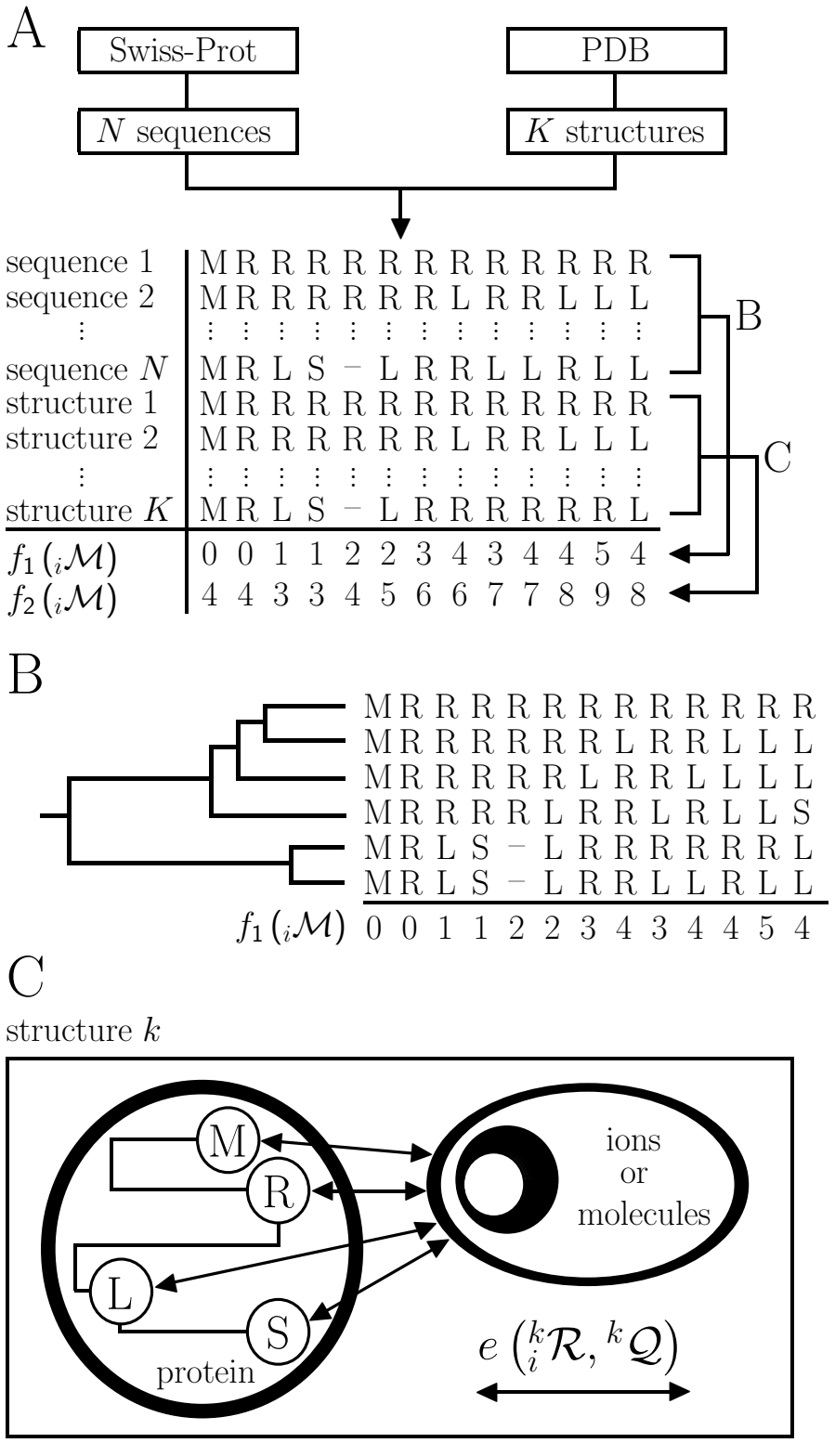


Figure 2.2: **Computations of f_1 and f_2.** (A) N sequences and K structures are extracted from the Swiss-Prot and the PDB, respectively. After all the sequences are aligned, $f_1\,({}_i\mathcal{M})$ and $f_2\,({}_i\mathcal{M})$ are computed by (B) and (C), respectively. (B) After a phylogenetic tree is written from sequences, $f_1\,({}_i\mathcal{M})$ is computed. (C) In structure k, ${}^k_i\mathcal{R}$ and ${}^k\mathcal{Q}$ denote coordinates of an amino acid residue and coordinates of ions or molecules, respectively. After proximity of ${}^k_i\mathcal{R}$ and ${}^k\mathcal{Q}$ is measured as $e\left({}^k_i\mathcal{R}, {}^k\mathcal{Q}\right)$ and computed on K structures, $f_2\,({}_i\mathcal{M})$ is computed.

Chapter 3

Protein functional site prediction by conservation

3.1 Introduction

In order to predict protein functional sites, there are many computational methods, which are based on (i) sequence, (ii) structure and (iii) sequence and structure [41, 28, 34, 47, 70, 42]. 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 [115]. One of the sequence-based methods is an SE-based method [99, 101]. 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 (SEP) [120] or SP [116] was proposed. The second one is that the SE-based method does not consider a background distribution of amino acids. Therefore, other information-theoretic method such as relative entropy [119] or Jensen Shannon divergence [17] 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 [17], weighting [36] and phylogenetic analysis were proposed. One of the methods based on a phylogenetic tree is an evolutionary trace (ET) method [74], which has been extended as weighted ET (WET) [66], iv-ET and rv-ET methods [81]. Additionally, other methods based on phylogenetic trees are ConSurf [8] and Rate4Site [95, 80] algorithms.

Although a variety of sequence-based methods have been already compared each other [17, 49], what difference makes a difference is difficult to understand because such methods do not be explained by one idea. Therefore, we consider a mapping, a mathematical formula, on a multiple sequence alignment

(MSA) and aim to construct 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 SEP, the method based on SP 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, based on the benchmark sets such as catalytic sites, ligand-binding sites or protein-protein interfaces [17], the predictive ability has been evaluated. However, the latter is immature because of usually conducting only one structure [74, 9]. 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) [16] 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 mapping, which measures proximity of amino acid residues and ions or molecules, and then two mappings are integrated [63].

3.2 Materials and Methods

3.2.1 Data collection

In UniProtKB/Swiss-Prot release 2015_01[21], 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 the 984 entries and are determined by X-ray crystallography were 68 entries. 14 entries were excluded because of binding an immunoprotein [22] and forming a chimeric protein [57, 110, 111, 112]. Consequently, as shown in Table 3.1, 54 entries including 103 chains were retained.

3.2.2 Computations of f_1 and f_2

As $N = 984$ and $K = 103$ in Figure 2.2, the sequences were aligned by the MAFFT 7 program [55]. 477 ${}_i\mathcal{M}$ were extracted because of including residues which have coordinate data.

A dissimilarity between two sequences was computed by the maximum likelihood method [58] using the Jones-Taylor-Thornton model [50] as a substitution matrix and the Dayhoff method [24] for computing equilibrium frequencies. From all combinations of the dissimilarities, a phylogenetic tree was written by the unweighted pair group method with arithmetic mean (UPGMA) [103]. f_1 (${}_i\mathcal{M}$) was computed by changing T, g_x, h_x, τ, $\mathcal{A}$ and ${}_i\mathcal{G}$. For h_3, the Gonnet matrix [14] was used and a weight was computed

Table 3.1: **54 PDB entries of EF-Tu/EF-1A proteins.**

Subfamily	Organism	PDB ID	Resolution	Ions or molecules
EF-Tu	*Bos taurus*, mitochondrial	1D2E	1.94	GDP, Mg^{2+}
		1XB2	2.20	Elongation factor Ts mitochondrial
	Escherichia coli	1EFC	2.05	GDP, Mg^{2+}
		2HCJ	2.12	GDP, TAC, Mg^{2+}
		3U6B	2.12	GDP, Mg^{2+}
		2BVN	2.30	ENX, GNP, Mg^{2+}
		4G5G	2.30	Thiomuracin A derivative, GDP, Mg^{2+}
		1D8T	2.35	Thiocillin GE2270, GDP, Mg^{2+}
		3U6K	2.45	Thiocillin GE2270 analogue NVP-LDK733, GDP, Mg^{2+}
		1DG1	2.50	GDP, Mg^{2+}
		1EFU	2.50	Elongation factor Ts
		1EFM	2.70	GDP
		3U2Q	2.70	Thicillin GE2270 analogue NVP-LFF571, GDP, Mg^{2+}
		2HDN	2.80	GDP, TAC, Mg^{2+}
		1ETU	2.90	GDP, Mg^{2+}
		4Q7J	2.90	Elongation factor Ts, Qβ replicase
		1OB2	3.35	Phe-tRNA, GNP, KIR, Mg^{2+}
		2FX3	3.40	GDP, Mg^{2+}
	Pseudomonas putida KT2440	4J0Q	2.29	GDP, MES, MPD, Mg^{2+}
		4IW3	2.70	Putative uncharacterized protein, GDP, Mg^{2+}
	Thermus aquaticus	1EFT	2.50	GNP, Mg^{2+}
		1B23	2.60	Cys-tRNA, GNP, Mg^{2+}
		1TTT	2.70	Phe-tRNA, GNP, Mg^{2+}
		1TUI	2.70	GDP, Mg^{2+}
		1OB5	3.10	Phe-tRNA, ENX, GNP, Mg^{2+}
	Thermus thermophilus	2C78	1.40	GNP, PUL, Mg^{2+}
		2C77	1.60	Thiocillin GE2270, GNP, Mg^{2+}
		1EXM	1.70	GNP, Mg^{2+}
		4LBW	1.74	GNP, Mg^{2+}
		4H9G	1.93	GNP, Mg^{2+}
		1HA3	2.00	GDP, MAU, Mg^{2+}
		4LBV	2.03	GNP, Mg^{2+}
		4LBZ	2.22	GNP, Mg^{2+}
		4LC0	2.22	GNP, Mg^{2+}
		4LBY	2.69	GNP, Mg^{2+}
		1AIP	3.00	Elongation factor Ts
		4V5L	3.10	16S rRNA, 23S rRNA, Trp-tRNA, GCP, Mg^{2+}
		4V5P	3.10	16S rRNA, 23S rRNA, Trp-tRNA
		4V5Q	3.10	16S rRNA, 30S rpS12, Trp-tRNA, GDP, KIR
		4V5R	3.10	16S rRNA, Trp-tRNA, GDP, KIR
		4V5S	3.10	16S rRNA, Trp-tRNA, GDP, KIR
		4V8Q	3.10	16S rRNA, 23S rRNA, Small protein B SMPB, tmRNA δ, GDP, KIR, Mg^{2+}
		4V5G	3.60	16S rRNA, 23S rRNA, 30S rpS12, Thr-tRNA, GDP, KIR, Mg^{2+}
aEF1A	*Aeropyrum pernix*	3VMF	2.30	Peptide chain release factor subunit 1, GTP, Mg^{2+}
		3WXM	2.30	Protein pelota homologue, GTP, Mg^{2+}
	Sulfolobus solfataricus	1JNY	1.80	GDP
		1SKQ	1.80	GDP, Mg^{2+}
eEF1A	*Oryctolagus cuniculus*	4C0S	2.70	GDP, Mg^{2+}
	Saccharomyces cerevisiae	1F60	1.67	Elongation factor 1Bα
		2B7C	1.80	Elongation factor-1 β
		1G7C	2.05	Elongation factor 1-β, 5GP
		1IJE	2.40	Elongation factor 1-β, GDP
		2B7B	2.60	Elongation factor-1 β, GDP
		1IJF	3.00	Elongation factor 1-β, GDP

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.

by the Sibbald and Algos algorithm [102] and the number of iterations was 100,000.

By separating each asymmetric unit, $f_2\,({}_i\mathcal{M})$ was computed and, in each entry, representative ions or molecules were shown in Table 3.1. However, because of the 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 [35, 87, 65, 60, 31, 91, 117].

3.2.3 Visualization

$f_1\,({}_i\mathcal{M})$, $f_2\,({}_i\mathcal{M})$, AUC and Spearman's ρ were visualized by the matplotlib Python package [44]. A three-dimensional structure was visualized by the VMD program [43].

3.3 Results

3.3.1 Prediction by f_1

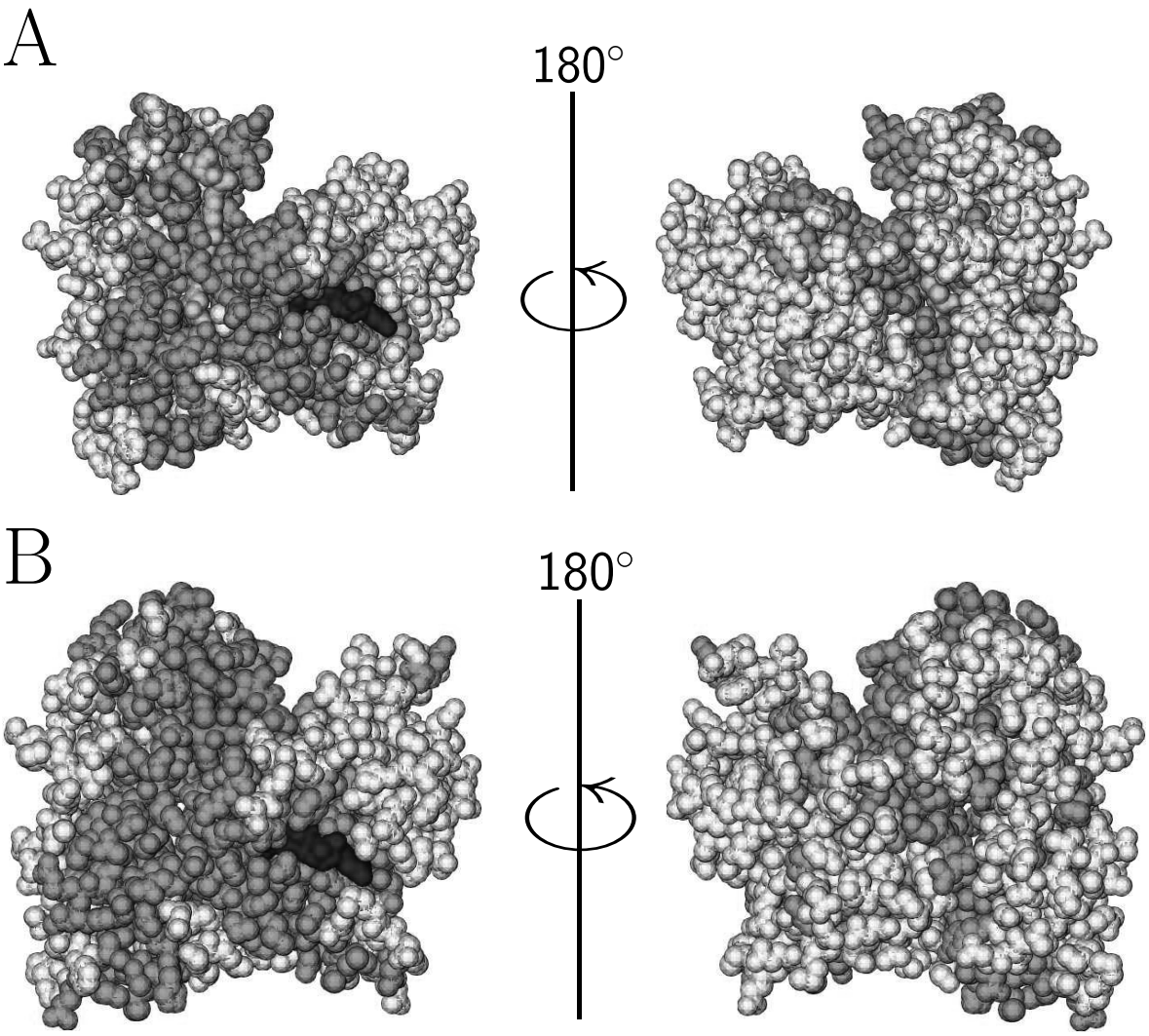


Figure 3.1: **The three-dimensional structures of** *Termus thermophilus* **EF-Tu (PDB ID: 2C78, 2C77).** $f_1\,({}_i\mathcal{M})$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$. The amino acid residues were divided in half by $f_1\,({}_i\mathcal{M})$. If $f_1\,({}_i\mathcal{M})$ is small, the atoms were colored by pink. If $f_1\,({}_i\mathcal{M})$ is large, the atoms were colored by white. The blue spheres show atoms belonging to GTP. (A) The green spheres show atoms belonging to the pulvomycin (PDB ID: 2C78). (B) The cyan spheres show atoms belonging to the thiocilin (PDB ID: 2C77).

Based on the 984 sequences of EF-Tu/1A, $f_1\,({}_i\mathcal{M})$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$. Figures 3.1, 3.2, 3.3, 3.4 and 3.5 show that if the $f_1\,({}_i\mathcal{M})$ is small, the amino acid residue is more likely to be proximate from aminoacyl tRNA, release factor subunit 1, pelota homologue, EF-Ts, EF-1B, GTP, GDP, pulvomycin, thiocillin, kirromycin and magnesium ion.

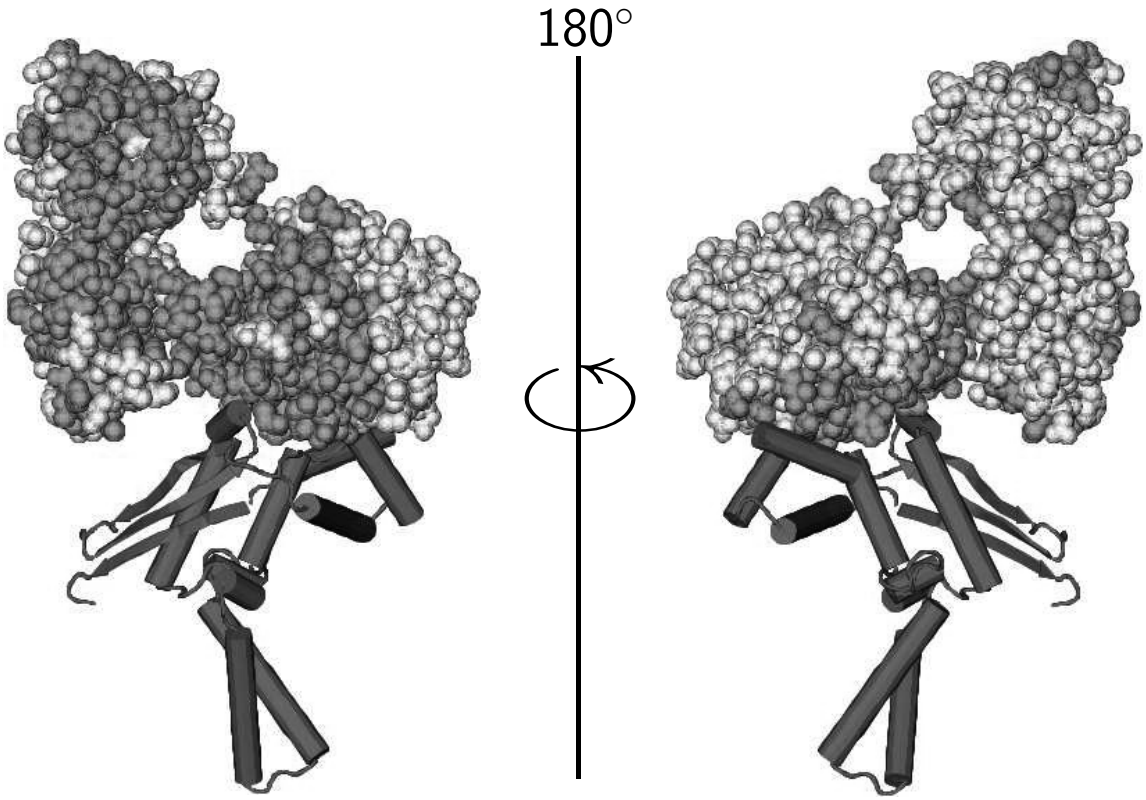


Figure 3.2: **The three-dimensional structure of** *Termus thermophilus* **EF-Tu (PDB ID: 1AIP).** $f_1\left({}_i\mathcal{M}\right)$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}_i^G\mathcal{G}$. The amino acid residues were divided in half by $f_1\left({}_i\mathcal{M}\right)$. If $f_1\left({}_i\mathcal{M}\right)$ is small, the atoms were colored by pink. If $f_1\left({}_i\mathcal{M}\right)$ is large, the atoms were colored by white. The tan colored chain shows atoms belonging to the EF-Tu.

3.3.2 Correlations between f_1 and f_2

$f_1\left({}_i\mathcal{M}\right)$ was calculated by using various combinations of the parameters. If g_x, h_x, τ and $\mathcal{A}$ are same but ${}_i\mathcal{G}$ is different, Table 3.2 shows that when ${}_i\mathcal{G} = {}_i^1\mathcal{G}$, the AUC or the Spearman's ρ is smaller than ${}_i\mathcal{G} = {}_i^G\mathcal{G}$. In the latter case, Figure 3.6 shows that when the time point increases, the AUC or the Spearman's ρ tends to increase.

3.3.3 Evaluation of predicted functional amino acid residues by f_2

Figure 3.7A shows that ${}_i\mathcal{M} \in \mathcal{M}$ is classifiable in 4 by $f_1\left({}_i\mathcal{M}\right)$ and $f_2\left({}_i\mathcal{M}\right)$ using a receiver operating characteristic (ROC) curve [67] in Figure 3.7B. Figures 3.7C and 3.7D show that the left sides are more likely to have small $f_1\left({}_i\mathcal{M}\right)$ and small $f_2\left({}_i\mathcal{M}\right)$ but the right sides are more likely to have large $f_1\left({}_i\mathcal{M}\right)$ and large $f_2\left({}_i\mathcal{M}\right)$.

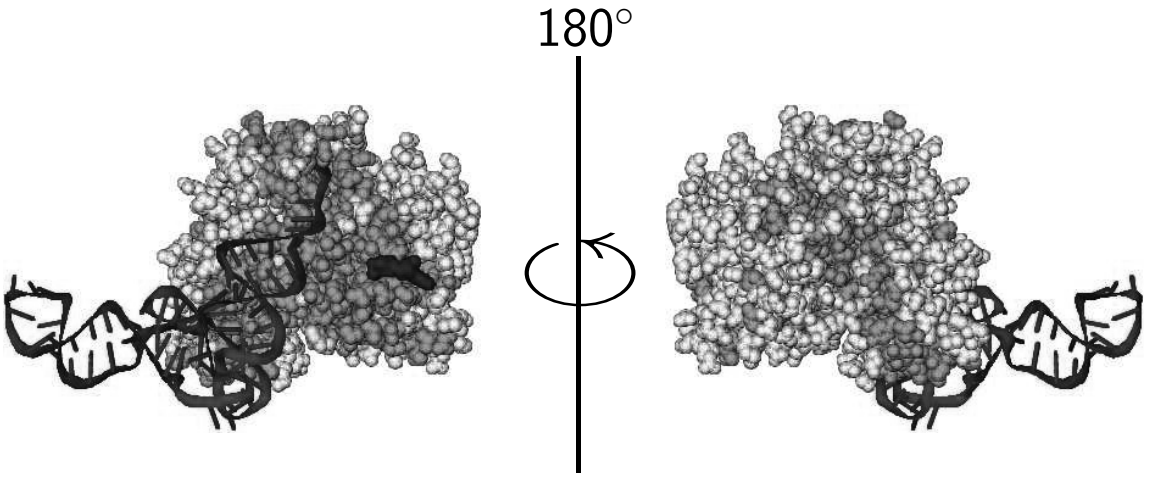


Figure 3.3: **The three-dimensional structure of** *Termus thermophilus* **EF-Tu (PDB ID: 4V5Q).** $f_1\left({}_i\mathcal{M}\right)$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$. The amino acid residues were divided in half by $f_1\left({}_i\mathcal{M}\right)$. If $f_1\left({}_i\mathcal{M}\right)$ is small, the atoms were colored by pink. If $f_1\left({}_i\mathcal{M}\right)$ is large, the atoms were colored by white. The purple chain shows tRNA. The blue spheres show atoms belonging to the GDP. The yellow spheres show atoms belonging to the kirromycin.

3.4 Discussion

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

Figures 3.1, 3.2, 3.3, 3.4 and 3.5 show that the $f_1\left({}_i\mathcal{M}\right)$ can predict binding sites of aminoacyl tRNA, release factor subunit 1, pelota homologue, EF-Ts, EF-1B, GTP, GDP, pulvomycin, thiocillin, kirromycin and magnesium ion but only showed appearances visualized onto the three-dimensional structures. Therefore, we conducted numerical evaluation by using the ROC curve and Spearman's ρ.

If $T = 1$, $g_x = g_3$, $h_x = h_2$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{1}_{i}\mathcal{G}$, the method is the method based on SE [99]. If $T = 1$, $g_x = g_3$, $h_x = h_2$, $\mathcal{A} = {}^{9}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{1}_{i}\mathcal{G}$, the method is the method based on SEP [120]. If $T = 1$ is changed to $T = N$ in the former and the latter, Figure 3.6 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 $T = 1$, $g_x = g_3$, $h_x = h_3$, ${}_i\mathcal{G} = {}^1_i\mathcal{G}$ and $w(l) = unused$ the method is the method based on SP [116]. If $T = 1$, $g_x = g_3$, $h_x = h_3$, ${}_i\mathcal{G} = {}^1_i\mathcal{G}$ and $w(l) = used$ the method is the method based on SP with weighting (SPW) [116]. If $T = 1$ is changed to $T = N$ in the former and the latter, Figure 3.6 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 ρ. However, in the above case, if ${}_i\mathcal{G} = {}^1_i\mathcal{G}$ is changed to ${}_i\mathcal{G} = {}^G_i\mathcal{G}$ in the former and the latter, Figure 3.6 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's ρ.

If $T = N$, $g_x = g_1$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^1_i\mathcal{G}$, the method is the iv-ET method [81]. If $T = N$, $g_x = g_2$, $h_x = h_2$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^1_i\mathcal{G}$, the method is equivalent to the rv-ET method [81]. If ${}^1_i\mathcal{G}$ is changed to ${}^G_i\mathcal{G}$ in the former and the latter, Table 3.2 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 latter, considering that each gap is different is effective for improving the AUC and the Spearman's ρ. Thus, $f_1({}_i\mathcal{M})$ is evaluable by $f_2({}_i\mathcal{M})$ and our methods improved some existing methods.

EF-Tu/EF-1A proteins are responsible for protein biosynthesis [91, 6] and we selected cocrystal structures involving the function. Therefore, if $f_2({}_i\mathcal{M})$ is small, an amino acid residue in ${}_i\mathcal{M}$ is proximate from a region involving protein biosynthesis. If $f_2({}_i\mathcal{M})$ is large, the amino acid residues in ${}_i\mathcal{M}$ are not proximate from the region. Figures 3.7A, 3.7C and 3.7D show the proximate region and the non-proximate region and Figure 3.7B shows that, on the ROC curve of $f_1({}_i\mathcal{M})$, the AUC is 0.739, which indicates that the proximate region tends to become small $f_1({}_i\mathcal{M})$ but the non-proximate region tends to become large $f_1({}_i\mathcal{M})$. Figure 3.7B shows the threshold of ${}_i\mathcal{M}$ is 134. This means that we can determine the threshold of functional sites from the ROC curve. In addition, Table 3.2 shows that the Spearman's ρ is 0.5801, which indicates that $f_1({}_i\mathcal{M})$ tends to be small if $f_2({}_i\mathcal{M})$ is small and $f_1({}_i\mathcal{M})$ tends to be large if $f_2({}_i\mathcal{M})$ is large. However, a complete linear correlation between $f_1({}_i\mathcal{M})$ and $f_2({}_i\mathcal{M})$ was not obtainable and therefore not all of $f_1({}_i\mathcal{M})$ can explain $f_2({}_i\mathcal{M})$. This may indicate that $f_1({}_i\mathcal{M})$ and $f_2({}_i\mathcal{M})$ can measure a similar thing each other but cannot always measure a same thing and, by $f_1({}_i\mathcal{M})$ or $f_2({}_i\mathcal{M})$, 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, $f_1\,({}_i\mathcal{M})$ and $f_2\,({}_i\mathcal{M})$ can quantify an amino acid residue.

3.5 Conclusions

Methods to mapping an MSA, which is represented as a character type and a coordinate type, were used for EF1A and we propose two usages. The first one is to evaluate f_1 by f_2. The second one is to determine predicted functional amino acid residues by the use of f_2. Our methods showed a better performance and reliability for functional site prediction of EF-Tu/EF1A.

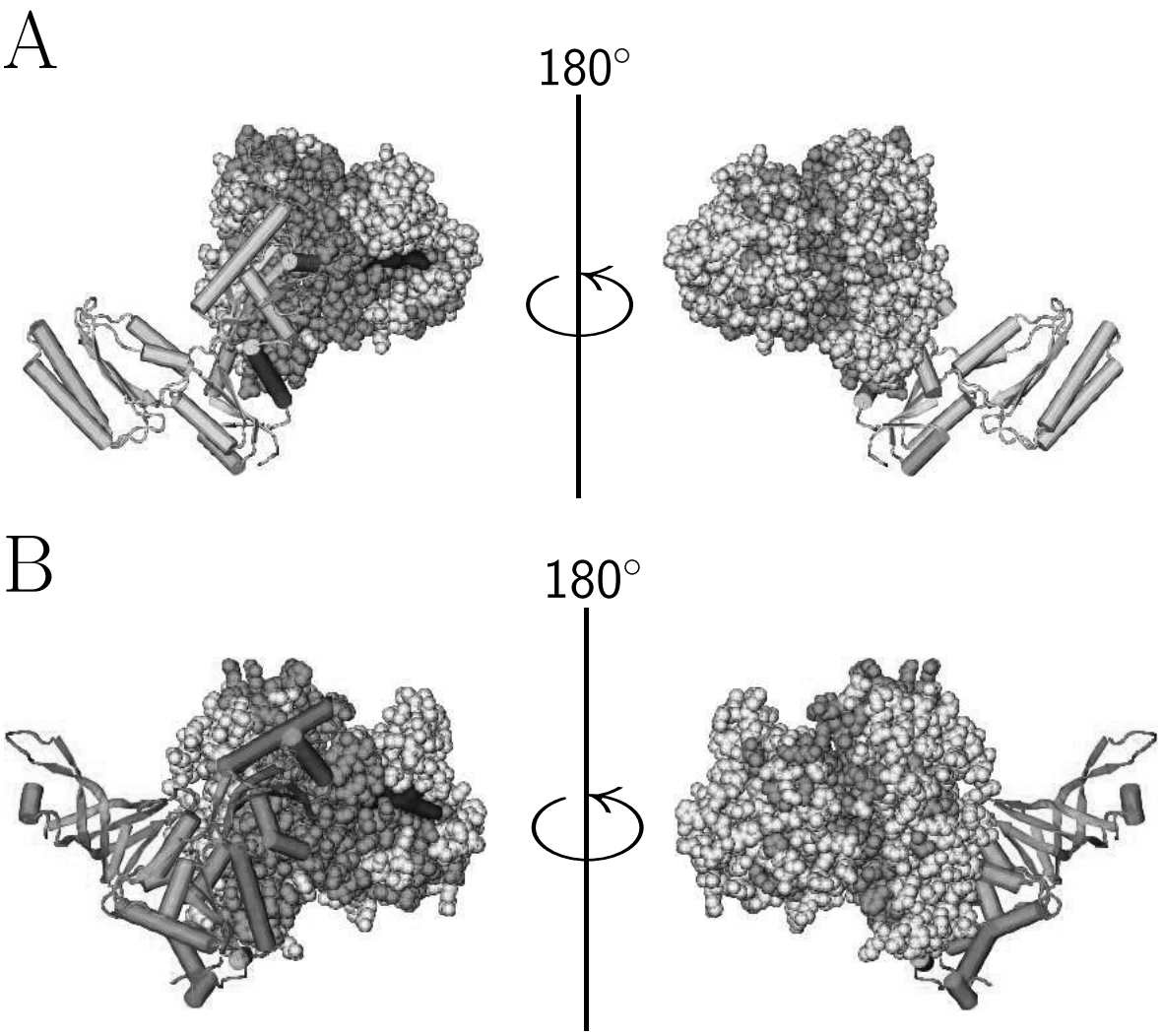


Figure 3.4: **The three-dimensional structures of** *Aeropyrum pernix* **EF1A (PDB ID: 3VMF, 3WXM).** $f_1\,({}_i\mathcal{M})$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$. The amino acid residues were divided in half by $f_1\,({}_i\mathcal{M})$. If $f_1\,({}_i\mathcal{M})$ is small, the atoms were colored by pink. If $f_1\,({}_i\mathcal{M})$ is large, the atoms were colored by white. The blue spheres show atoms belonging to GTP. (A) The yellow chain shows the release factor subunit 1 (PDB ID: 3VMF). (B) The green chain shows the pelota homologue (PDB ID: 3WXM).

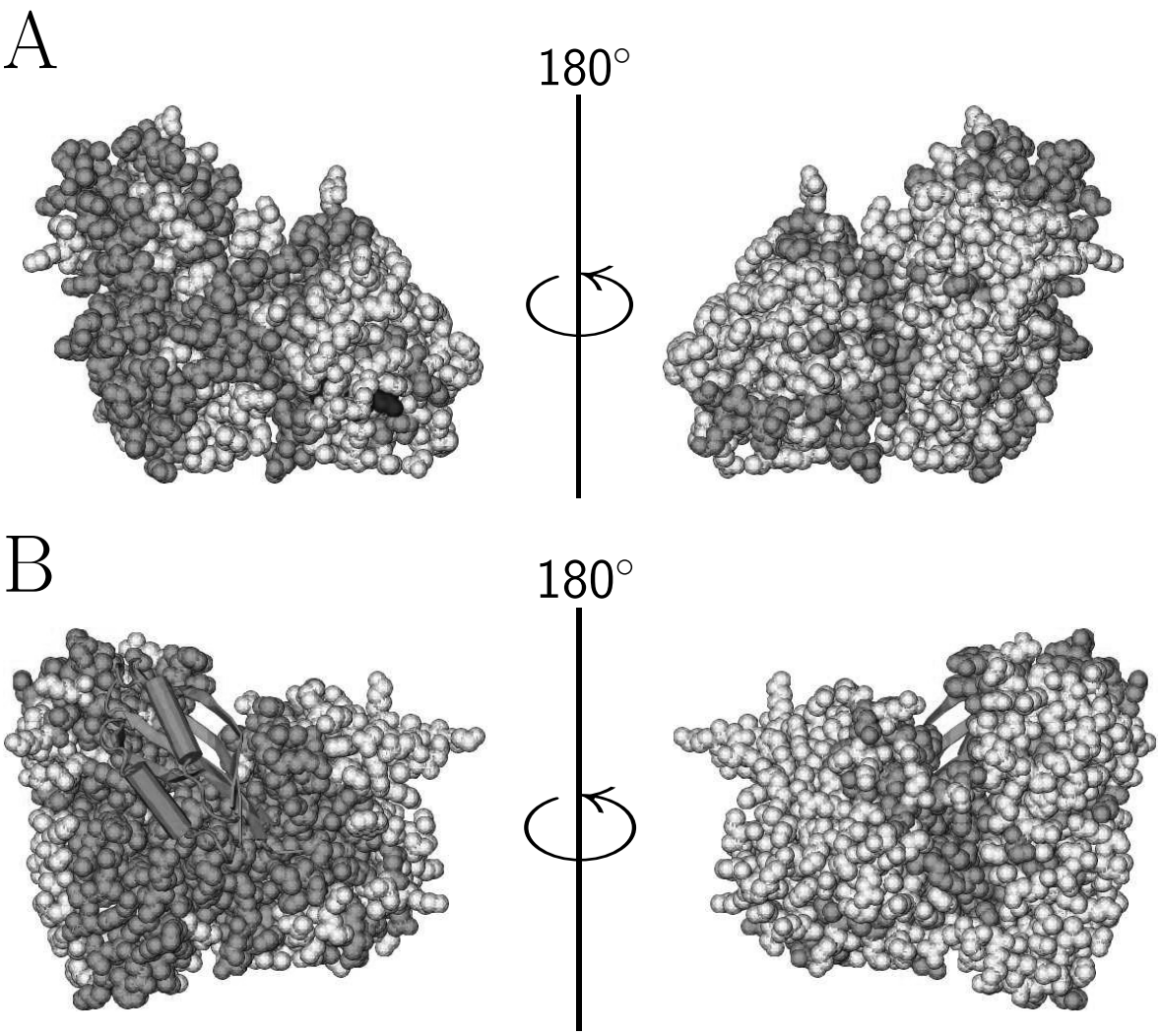


Figure 3.5: **The three-dimensional structures of** *Orytolagus cuniculus* **EF1A and** *Saccharomyces cerevisiae* **EF1A (PDB ID: 4C0S, 1F60).** $f_1\,({}_i\mathcal{M})$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}_i^G\mathcal{G}$. The amino acid residues were divided in half by $f_1\,({}_i\mathcal{M})$. If $f_1\,({}_i\mathcal{M})$ is small, the atoms were colored by pink. If $f_1\,({}_i\mathcal{M})$ is large, the atoms were colored by white. (A) The blue spheres show atoms belonging to the GDP and the yellow sphere shows a magnesium ion (PDB ID: 4C0S). (B) The cyan chain shows the EF-1Bα (PDB ID: 1F60).

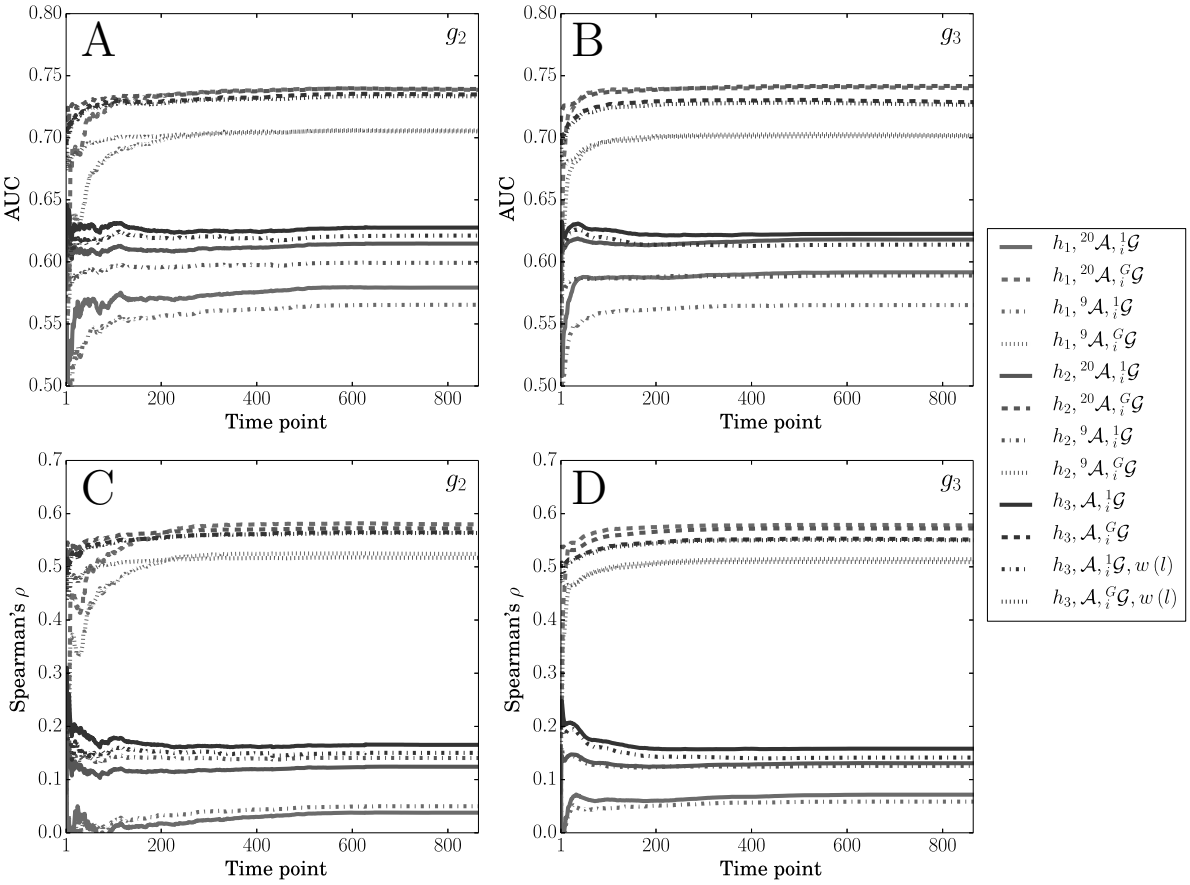


Figure 3.6: **Dependence on time points.** The time point is T in Eq. (2.6). (A) (B) AUC and (C) (D) Spearman's ρ were computed using g_x, h_x, $\mathcal{A}$ and ${}_i\mathcal{G}$ as shown in the figures. For calculation of h_3, we considered without or with sequence weights as $w\,(l)$.

Table 3.2: **Correlations between f_1 and f_2.**

g_x	h_x	$w\,(l)$	$\mathcal{A}$	${}_i\mathcal{G}$	τ	AUC	Spearman's ρ
g_1	h_1	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	$0 < \tau < 1$	0.5896	0.1221
				${}^{G}_{i}\mathcal{G}$	$0 < \tau < 1$	0.6242	0.3650
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	$0 < \tau < 1$	0.5700	0.1015
				${}^{G}_{i}\mathcal{G}$	$0 < \tau < 1$	0.6184	0.3509
	h_2	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	0.1	0.6036	0.1436
					0.2	0.6376	0.1977
					0.3	0.6420	0.2015
					0.4	0.6120	0.1585
				${}^{G}_{i}\mathcal{G}$	0.1	0.7207	0.5566
					0.2	0.7412	0.5773
					0.3	0.7374	0.5613
					0.4	0.7037	0.5028
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	0.1	0.5757	0.1171
					0.2	0.6023	0.1659
					0.3	0.6151	0.1715
					0.4	0.5885	0.1160
				${}^{G}_{i}\mathcal{G}$	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	unused	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	0.1	0.5901	0.1288
					0.2	0.6052	0.1467
					0.3	0.6366	0.2315
					0.4	0.6391	0.2763
				${}^{G}_{i}\mathcal{G}$	0.1	0.6758	0.4474
					0.2	0.6975	0.4734
					0.3	0.6905	0.4823
					0.4	0.6865	0.4922
		used	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	0.1	0.5872	0.1200
					0.2	0.5939	0.1387
					0.3	0.5916	0.1460
					0.4	0.5995	0.1733
				${}^{G}_{i}\mathcal{G}$	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	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5916	0.0718
				${}^{G}_{i}\mathcal{G}$	–	0.7399	0.5780
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5652	0.0587
				${}^{G}_{i}\mathcal{G}$	–	0.7020	0.5145
	h_2	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6180	0.1308
				${}^{G}_{i}\mathcal{G}$	–	0.7417	0.5722
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5890	0.1257
				${}^{G}_{i}\mathcal{G}$	–	0.7012	0.5091
	h_3	unused	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6225	0.1579
				${}^{G}_{i}\mathcal{G}$	–	0.7287	0.5517
		used	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6138	0.1412
				${}^{G}_{i}\mathcal{G}$	–	0.7265	0.5501
g_3	h_1	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5792	0.0378
				${}^{G}_{i}\mathcal{G}$	–	0.7386	0.5801
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5655	0.0501
				${}^{G}_{i}\mathcal{G}$	–	0.7052	0.5244
	h_2	–	${}^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6147	0.1241
				${}^{G}_{i}\mathcal{G}$	–	0.7393	0.5723
			${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.5992	0.1405
				${}^{G}_{i}\mathcal{G}$	–	0.7059	0.5170
	h_3	unused	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6276	0.1653
				${}^{G}_{i}\mathcal{G}$	–	0.7349	0.5650
		used	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–	0.6211	0.1502
				${}^{G}_{i}\mathcal{G}$	–	0.7335	0.5637

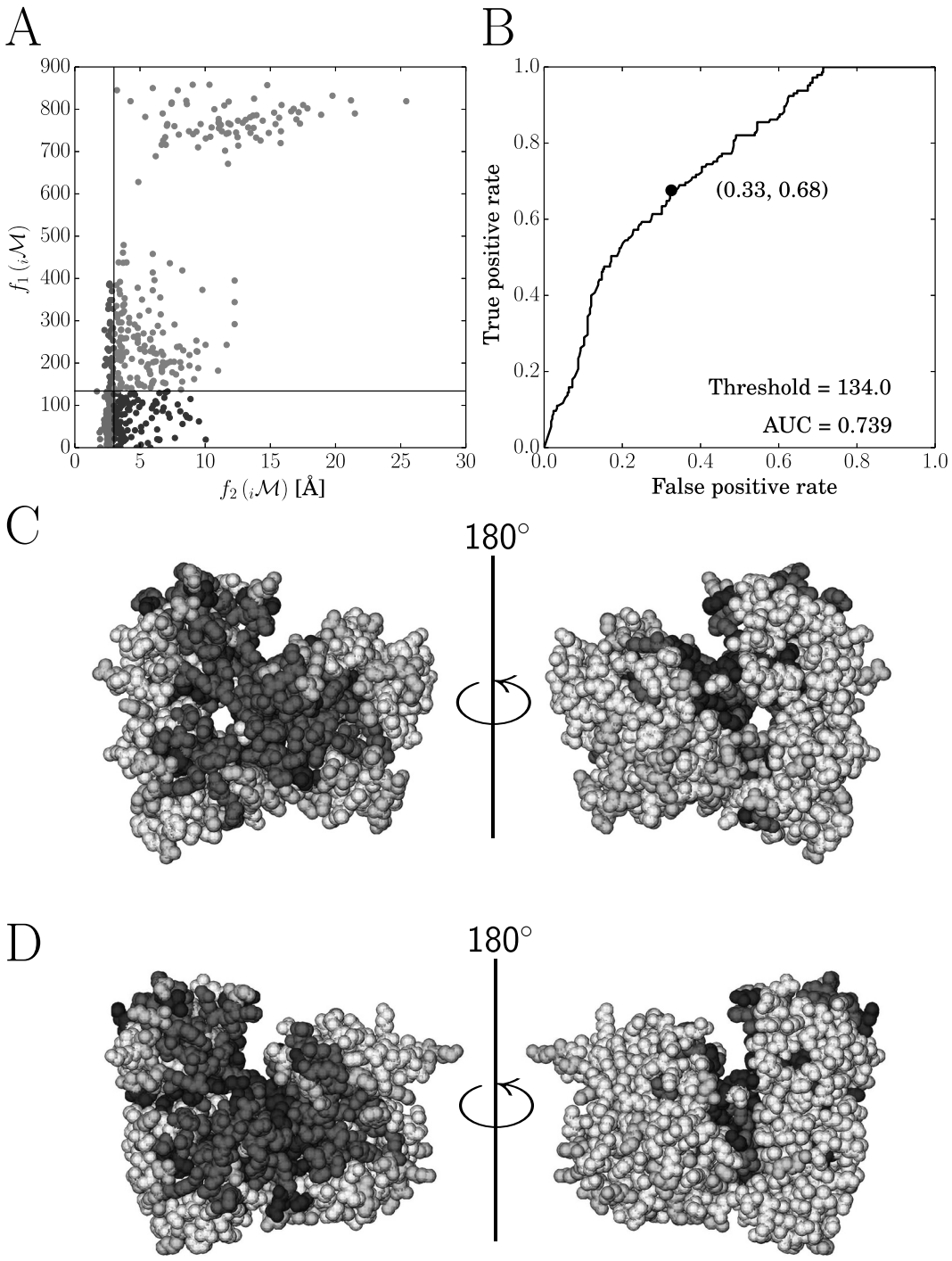


Figure 3.7: **Visualization of f_1 and f_2.** $f_1\,({}_i\mathcal{M})$ was computed using g_2, h_2, ${}^{20}\mathcal{A}$ and ${}^{G}_{i}\mathcal{G}$. (A) A scatter plot of $f_1\,({}_i\mathcal{M})$ and $f_2\,({}_i\mathcal{M})$. However, one point whose $f_1\,({}_{468}\mathcal{M}) = 822$ and $f_2\,({}_{468}\mathcal{M}) = 45.56$ was not shown. ${}_i\mathcal{M} \in \mathcal{M}$ was classified into 2 by whether $f_2\,({}_i\mathcal{M})$ is equal to or smaller than 3 Å or larger than 3 Å because if 3 Å was changed to 2 Å, the number of the smaller sites was 4. (B) By regarding the former as true and the latter as false, the ROC curve was written using $f_1\,({}_i\mathcal{M})$. The threshold was determined so that (true positive rate + 1 - false positive rate) is maximum and, eventually, ${}_i\mathcal{M} \in \mathcal{M}$ was classified into 4, which were visualized onto three-dimensional structures of (C) *Thermus thermophilus* EF-Tu [91] and (D) *Saccharomyces cerevisiae* EF1A [6].

Chapter 4

Protein functional site prediction by specific conservation

4.1 Introduction

Chapter 3 used a conservation which investigates important sites for all subfamilies of the protein family. This means that the conservation is more likely to be small if the site has been conserved in all subfamilies but the conservation is more likely to be large if the site has not been conserved in a subfamily. This chapter changes the point of view of conservation and uses a conservation which investigates important sites for a particular subfamily. This means that the conservation is more likely to be large if the site has been conserved in a particular subfamily but has not been conserved in the whole family. In addition, the conservation is more likely to be small if the site has not been conserved in the subfamily or has conserved in the whole family. Therefore, we define a specific conservation which can detect sites that have been conserved in the target subfamily but have not been conserved in the whole family. This chapter investigates whether or not the specific conservation can detect important sites for functional divergences of *Mycoplasma pneumoniae* EF-Tu (MpEF-Tu) and *Mycoplasma genitalium* EF-Tu (MgEF-Tu) and eEF1A1 and eEF1A2 [61].

4.2 Materials and Methods

4.2.1 Data collection

From UniProtKB/Swiss-Prot release 2013_11 (541,762 entries) [7], 1,024 entries were extracted by searching 'GTP-binding elongation factor family. EF-Tu/EF-1A subfamily' in the annotation of protein similarities. 41 entries were excluded because 39 entries were annotated as a fragment and 2 entries included

'X' in their sequences. Consequently, 983 entries were retained as sequences of the EF-Tu/EF1A proteins.

4.2.2 Computation of a specific conservation

The sequences were aligned by the MAFFT 7 program [55]. A dissimilarity between two sequences was computed by the PHYLIP protdist program [29] using the Jones-Taylor-Thornton model [50] as an amino acid substitution model. A phylogenetic tree was reconstructed by the UPGMA in the PHYLIP neighbor program and the sequences were divided into groups at each node of the phylogenetic tree. For computation of SP-based specific conservations, the Gonnet substitution matrix [14] was used and a weight of sequences was computed by the Sibbald and Argos method [102] and the number of iterations was 100,000.

4.2.3 Visualization

The AUC was calculated by using the pROC R package [96], A heat map of the AUCs was created by the matplotlib Python package [44]. The sequence alignment and the phylogenetic tree were visualized by the Discovery Studio 3.1 software package [2] and the MEGA 5.2 software package [113], respectively.

4.3 Results

4.3.1 Analysis of EF-Tu in Mycoplasma species

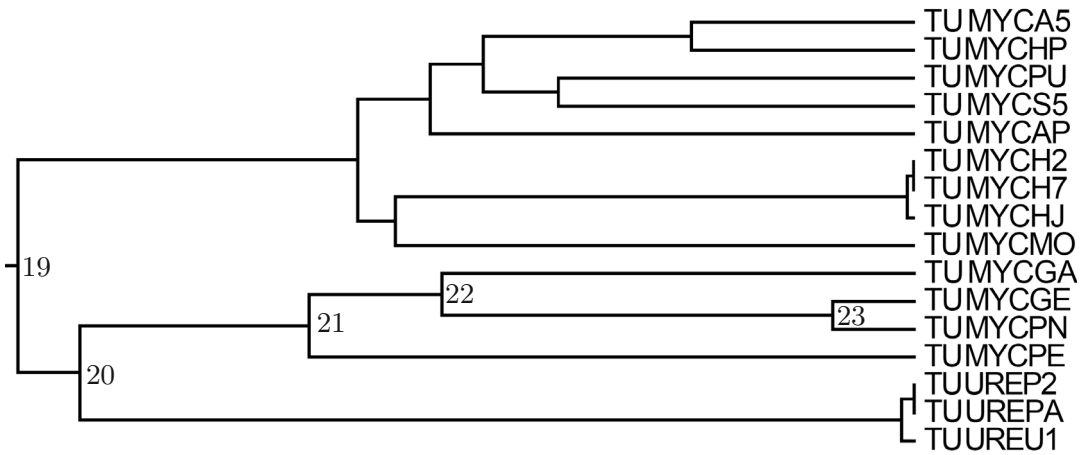


Figure 4.1: **Evolutionary branches of EF-Tu in Mycoplasma species.** The evolutionary branches include the MpEF-Tu and MgEF-Tu nodes (TUMYCPN and TUMYCGE). The numbers are assigned to each node by counting from the root to the MpEF-Tu node.

Based on the 983 sequences of EF-Tu/1A, we constructed the MSA and the phylogenetic tree. Figure 4.1 shows that the nodes of MpEF-Tu and MgEF-Tu are divided into the 23rd node when numbers 1

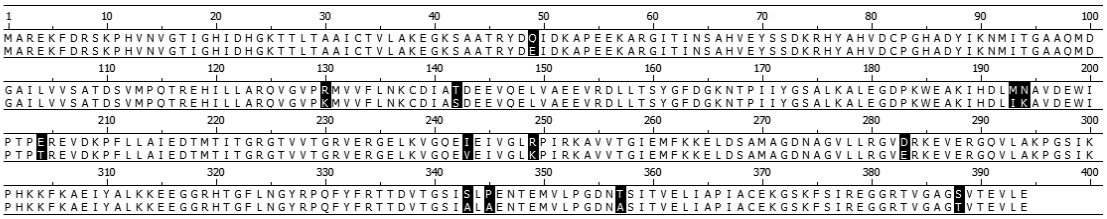


Figure 4.2: **Pairwise sequence alignment of EF-Tu in Mycoplasma species.** The 2 sequences in Mycoplasma species were extracted from the MSA of the EF-Tu/EF1A family. Alignment sites which do not have the residues of MpEF-Tu and MgEF-Tu were excluded. A white background site consists of unique amino acids and black background site consists of two types of amino acids.

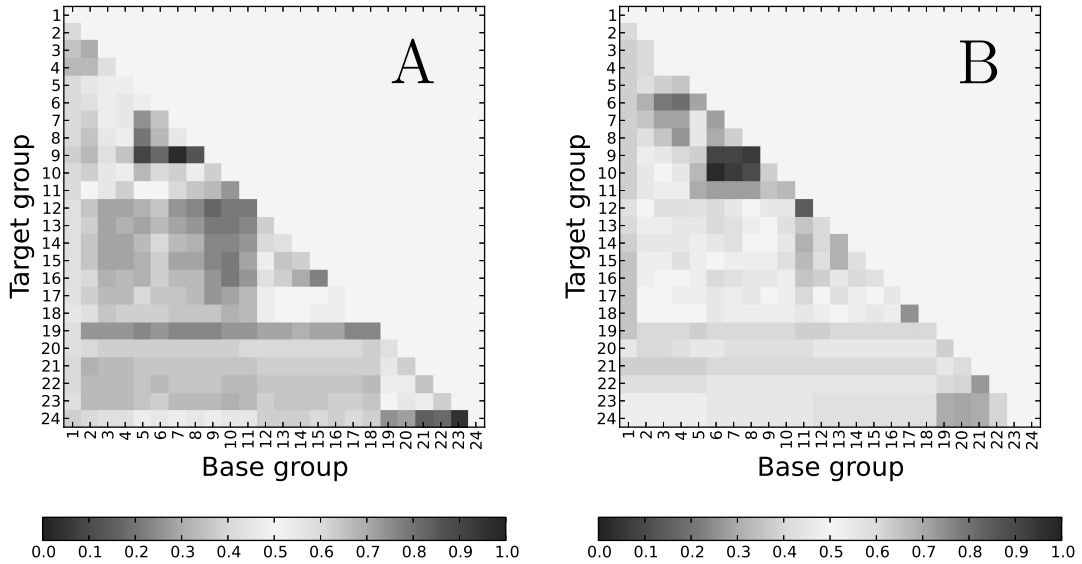


Figure 4.3: **Performance evaluation for predicting fibronectin binding sites in EF-Tu.** From specific conservations based on (A) SPW or (B) SE, the AUC was calculated by using the M193, N194, E204, S343, P345 and T357 residues, which involve fibronectin binding of MpEF-Tu [11]. The positive direction for the ROC curve was defined as ascending order of the specific conservations.

to 24 are assigned to each node from the root to the leaf node of MpEF-Tu. Figure 4.2 shows that there are 13 different sites between the sequences of MpEF-Tu and MgEF-Tu. Figure 4.3A shows that $C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$ is the second highest AUC and $C\left({}^{7}_{i}\mathcal{M}, {}^{9}_{i}\mathcal{M}\right)$ is the highest AUC. Figure 4.3B shows that the AUC of $C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$ is 0.5 and $C\left({}^{6}_{i}\mathcal{M}, {}^{10}_{i}\mathcal{M}\right)$ is the highest AUC. Table 4.1 shows that each $C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$ in the SP- or SE-based method is higher than each $C\left({}^{7}_{i}\mathcal{M}, {}^{9}_{i}\mathcal{M}\right)$ or $C\left({}^{6}_{i}\mathcal{M}, {}^{10}_{i}\mathcal{M}\right)$, respectively.

4.3.2 Analysis of eEF1A1 and eEF1A2

Figure 4.4 shows that the nodes of human eEF1A1 and eEF1A2 are divided into the 21st node when numbers 1 to 28 are assigned to each node from the root to the leaf node of human eEF1A1. Figure 4.5 shows that there are 36 different sites between the sequences of human eEF1A1 and eEF1A2. Figure 4.6A shows that $C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$ is the second highest AUC and $C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$ is the highest AUC. Figure

Table 4.1: **Specific conservations with high performances for predicting fibronectin binding sites in EF-Tu** The 'X' letters denote the MpEF-Tu residue involving fibronectin binding [11]. The threshold shows the arithmetic mean between two conservations whose sum of the true positive rate and 1 - false positive rate becomes maximum when the threshold is used as a cutoff value of the conservations. $f_1\left({}_i\mathcal{M}\right)$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$.

$h_x, \mathcal{A}, {}_i\mathcal{G}, w\,(l)$	$h_3, {}^{20}\mathcal{A}, {}^{1}_{i}\mathcal{G}, w\,(l)$		$h_2, {}^{20}\mathcal{A}, {}^{1}_{i}\mathcal{G}$			
Residues	$C\left({}^{7}_{i}\mathcal{M}, {}^{9}_{i}\mathcal{M}\right)$	$C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$	$C\left({}^{6}_{i}\mathcal{M}, {}^{10}_{i}\mathcal{M}\right)$	$C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$	Fibronectin binding	$f_1\left({}_i\mathcal{M}\right)$
Q49E	-0.00082	0.14000	0.00043	0.22767		291
R130K	-0.00006	0.11945	0.00081	0.22767		266
T142S	0.00079	0.08863	0.00038	0.22767		229
M193I	0.00278	0.15784	0.00323	0.22767	X	176
N194K	0.00166	0.23084	0.00064	0.22767	X	437
E204T	0.00253	0.25597	0.00271	0.22767	X	378
I243V	-0.00197	0.06348	-0.00074	0.22767		265
R249K	-0.00161	0.11945	-0.00013	0.22767		458
D283E	-0.00022	0.13905	0.00058	0.22767		220
S343A	0.00052	0.12061	0.00134	0.22767	X	388
P345A	0.00190	0.29930	0.00307	0.22767	X	191
T357A	0.00280	0.17341	0.00179	0.22767	X	277
S388T	-0.00322	0.08863	0.00059	0.22767		204
AUC	0.97619	0.95238	0.97619	0.50000		
Threshold	0.00023	0.14892	0.00061			
Sensitivity	0.85714	1.00000	0.85714			
Specificity	1.00000	0.83333	1.00000			

4.6B shows that $C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$ is the highest AUC and $C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$ is the second highest AUC. Table 4.2 shows that each $C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$ is higher than each $C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$ in the SP- or SE-based method.

4.4 Discussion

4.4.1 Specific conservation

Landgraf et al. [66] have proposed an SPW-based variability. The range of the variability is 0.0 to 1.0. The higher the value implies, the more the alignment site consists of hardly substituting amino acids. Mihalek et al. [81] have proposed an SE-based variability. The range of the variability is 0.0 to $\ln 20$. The higher the value implies, the more the alignment site consists of various letters. In our thesis, two changes have been added to this variability. One is that a gap site is regarded as an extra amino acid. The other is that the base of the logarithm is 21. The range of this variability is 0.0 to 1.0.

Landgraf et al. [66] have also proposed a specific conservation using their variability. The conservation is obtained from two variabilities. One is the variability computed from all sequences in the MSA. The other is the variability computed from a group which includes a target protein. The range of the score is -0.5 to 1.0. The higher the score implies, the more the variability of the group relatively decreases than the variability of all sequences. They have discussed influences of grouping by determining a threshold of the specific conservation from random sequence clusters. However, the conservation obtained from their grouping method strongly depends upon how to construct a set of protein sequences. Therefore, by comprehensive grouping of the evolutionary branches, we consider a method which can determine an evolutionary divergence highly correlating with the functional divergence.

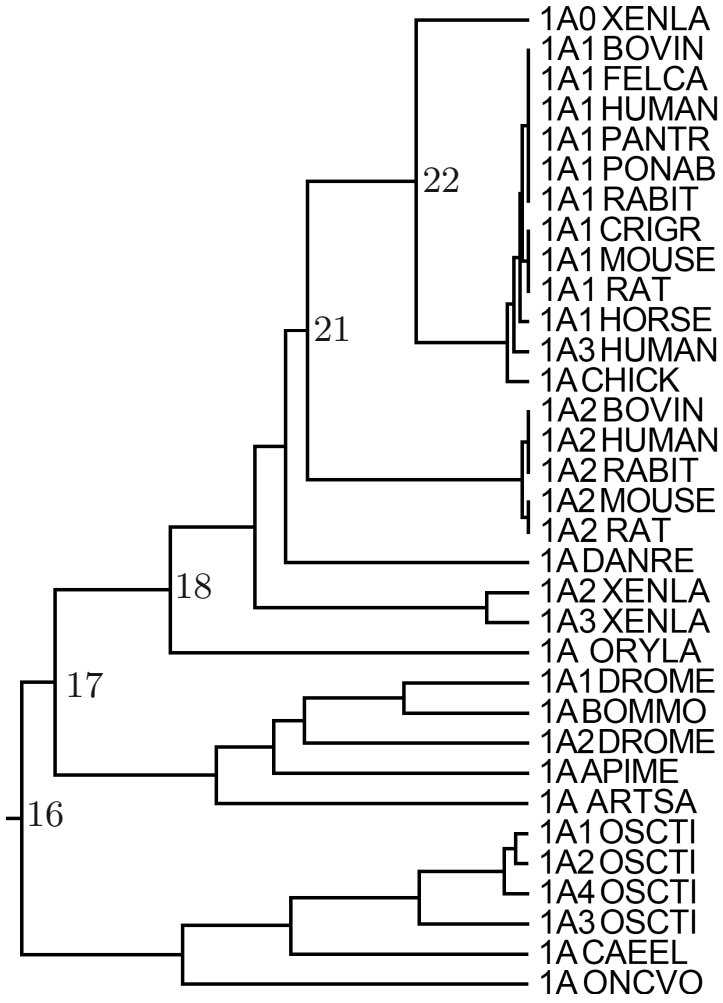


Figure 4.4: **Evolutionary branches of eEF1A1 and eEF1A2.** The numbers are assigned to each node by counting from the root to the human eEF1A1 node.

4.4.2 Functional divergences of EF-Tu/EF1A

Some bacteria use the membrane-localized EF-Tu to infect their host cell. Mycoplasma species can be found in many different hosts, whereas the individual species have a strict host, organ and tissue specificity [37]. *M. pneumoniae* is isolated from the human respiratory tract [118] and the urogenital tract [30]. *M. genitalium*, which triggers sexually transmitted disease [48], has also been associated with the human respiratory tract [12]. The tissue specificity may be able to be determined by genetic distinctions between them. Balasubramanian et al. [11] have shown that MpEF-Tu has higher fibronectin binding affinity than MgEF-Tu. They have identified fibronectin binding residues by site-directed mutagenesis studies of MpEF-Tu.

Meanwhile, some species have several eEF1A encoding genes and *Homo sapiens* has two variant human eEF1A forms, referred to as eEF1A1 and eEF1A2, shared with ∼96% similarity between their protein sequences [76]. Both eEF1A proteins seem to exhibit similar translation activities although they have different binding affinities of GTP/GDP and eEF1Bα [53, 77]. In addition, the expression patterns

of the two isoforms are exclusive in the human tissues [73, 59]. eEF1A2 is highly expressed in skeletal muscle, heart muscle and brain whereas eEF1A1 is lowly expressed in these tissues but highly expressed in other tissues like lung, liver, and placenta. Loss of expression of eEF1A2 in mice has an effect on motor neuron degradation [19, 86].

4.4.3 Fibronectin binding residues

We explored two EF-Tu/1A subfamilies strongly involving changes of 6 fibronectin binding residues between MpEF-Tu and MgEF-Tu. The conservations with the highest AUCs are $C\left({}^{7}_{i}\mathcal{M}, {}^{9}_{i}\mathcal{M}\right)$ and $C\left({}^{6}_{i}\mathcal{M}, {}^{10}_{i}\mathcal{M}\right)$ in the SPW- and SE-based method, respectively. Comparisons of $C\left({}^{23}_{i}\mathcal{M}, {}^{24}_{i}\mathcal{M}\right)$ show that the above conservations are considerable lower values. This implies that there may be no significance of the differences of the values. The conservations with the second highest AUC is the conservations between the node dividing the sequences of MpEF-Tu and MgEF-Tu in the SPW-based method. This implies that the functional divergence involving fibronectin binding between MpEF-Tu and MgEF-Tu highly correlates with the node dividing the sequences of MpEF-Tu and MgEF-Tu. The value of the conservation depends on the substitution scores of amino acids because conservations calculated from a leaf node is always 0. This shows that the property of amino acids is an essential factor for the functional difference of the fibronectin binding function.

Balasubramanian et al. [11] have created the structural model of the MpEF-Tu protein by homology modeling using the *Thermus thermophilus* EF-Tu as a template. The fibronectin binding residues, which are divided into binding regions 1 and 2, are surface accessible residues in the three-dimensional structure. Our result shows that these surface residues have very high correlation with the specific conservations because Table 4.1 shows such large AUCs.

4.4.4 Actin binding residues

We explored two EF-Tu/EF1A subfamilies strongly involving changes of 2 actin binding residues between the human eEF1A1 and eEF1A2. The SPW- or SE-based conservations whose AUCs are the first and second highest values also include $C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$. Comparisons of $C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$ and $C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$ show that the above conservations are considerable lower values. This implies that there may be no significance of the differences of the values. In Figure 4.4, SPW- or SE-based specific conservations whose AUCs are the first and second highest values include $C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$. This implies that the functional divergence involving actin binding between the human eEF1A1 and eEF1A2 highly correlates with the node dividing the sequences of human eEF1A1 and eEF1A2. Figure 4.5 shows that there are 29 sites consisting of unique letters in the sequences of the 22nd node of Figure 4.4. This implies that the value of the conservation depends upon the substitution score of amino acids. This shows that the property of amino acids is an essential factor for the functional difference of the actin binding function.

Table 4.2 shows that the first and second highest values of 36 SE-based conservations are 0.25547 and 0.19407, respectively. The residues of human eEF1A1 and eEF1A2 with these conservations include the 2 actin binding residues. The values of other 27 residues are also 0.19407. This shows that we cannot investigate differences of the 27 residues by using the SE-based conservations. Conversely, there are various values in the SPW-based conservations. We can suggest a possibility for further analyses of eEF1A1 and eEF1A2. Soares et al. [105] have created structural models of the human eEF1A1 and eEF1A2 proteins by homology modeling using the yeast eEF1A as a template. They have compared interaction surfaces and different residues between the human eEF1A1 and eEF1A2 proteins based upon the three-dimensional structures. Table 4.3 shows hypothetical functional residues described by Soares et al. [105]. In the hypothetical residues involving guanosine binding, actin binding and phosphorylation, N197H, A326C and F393S have the highest specific conservations, respectively. In addition, we can also compare the specific conservations with the sites involving post translational modifications [104] or origomerization [75]. This shows that the specific conservation is useful for determining key residues for further analyses of EF1A.

4.5 Conclusions

This chapter described specific conservations of subfamilies. This method showed that the functional divergences involving actin and fibronectin binding of the EF-Tu/1A molecules highly correlate with the evolutionary branches which divide the sequences. Such quantitative descriptions, which are based upon SP and SE, is effective for predicting important amino acid residues of EF-Tu/EF1A. These methods may prompt identification of key residues involving functional divergence of protein subfamilies or subtypes.

Meanwhile, Tables 4.1 and 4.2 show that $f_1\,({}_i\mathcal{M})$ is more likely to be larger than 134, which is a threshold determined in Figure 3.7B. This implies that $f_1\,({}_i\mathcal{M})$ cannot always detect all functional sites of EF-Tu/EF1A. Therefore, it is necessary for comprehensively predicting functional sites of multifunctional proteins that conservation of amino acid residues is considered from a different point of view.

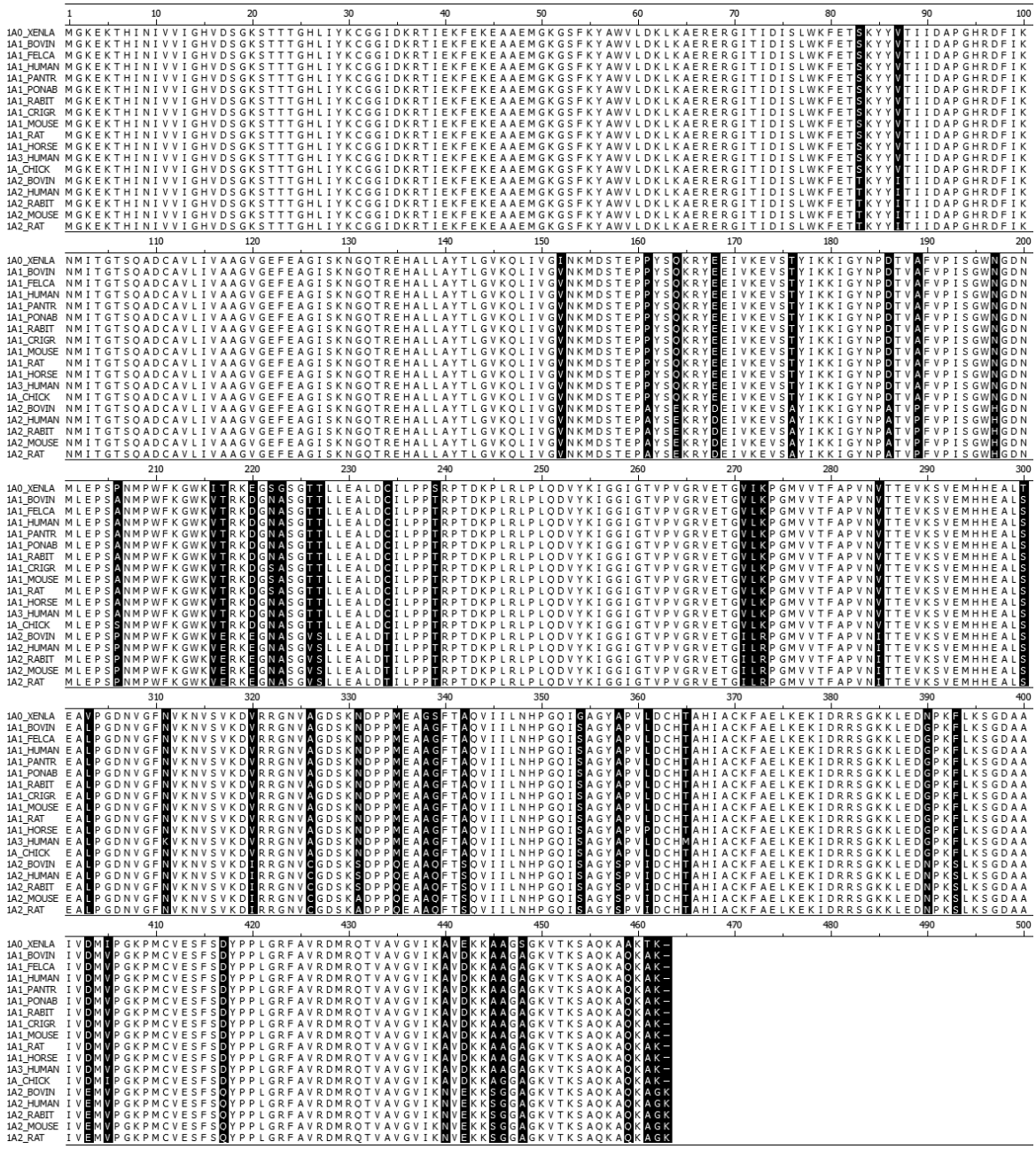


Figure 4.5: **MSA of eEF1A1 and eEF1A2.** The 18 sequences were extracted from the MSA of the EF-Tu/EF1A family. Alignment sites which do not have amino acid residues of the 18 sequences were excluded. A white background site consists of unique amino acids and a black background site consists of two or more types of amino acids.

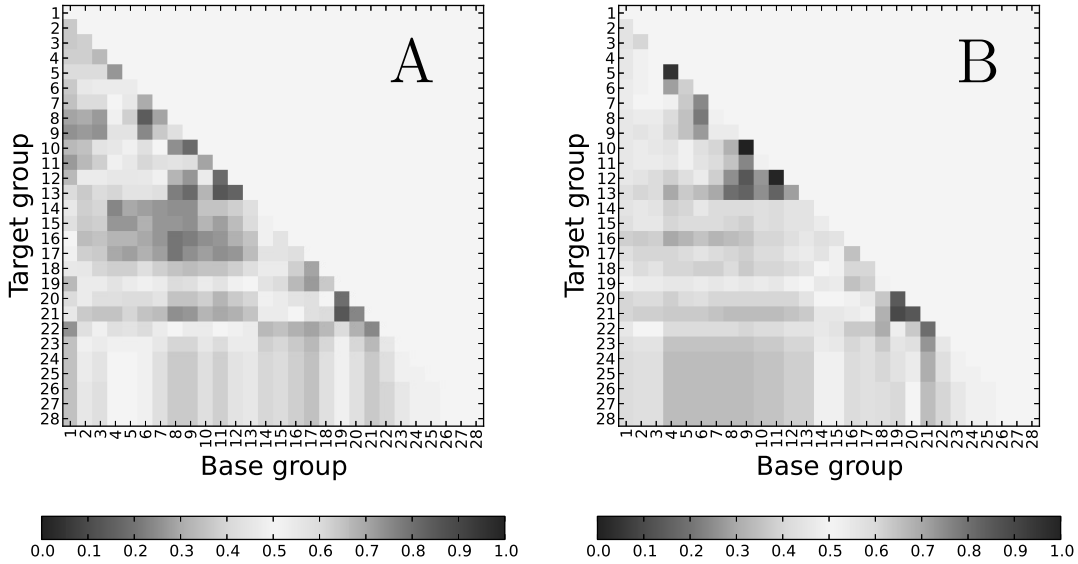


Figure 4.6: **Performance evaluation for predicting actin binding sites in EF1A.** From specific conservations based on (A) SPW or (B) SE, the AUC was calculated by using the N331S and M335Q residues, which are equivalent to the N329 and K333 residues involving actin binding of the yeast eEF1A molecule [32]. The positive direction for the ROC curve was defined as ascending order of the specific conservations.

Table 4.2: **Specific conservations with high performances for predicting actin binding sites in EF1A** The 'X' letters denote the human eEF1A1 and eEF1A2 residues which are equivalent to the N329 and K333 residues involving actin binding of the yeast eEF1A molecule [105]. The threshold shows the arithmetic mean between two conservations whose sum of the true positive rate and 1 - false positive rate becomes maximum when the threshold is used as a cutoff value of the conservations. $f_1\left({}_i\mathcal{M}\right)$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$.

$h_x, \mathcal{A}, {}_i\mathcal{G}, w\left(l\right)$	$h_3, {}^{20}\mathcal{A}, {}^{1}_{i}\mathcal{G}, w\left(l\right)$		$h_2, {}^{20}\mathcal{A}, {}^{1}_{i}\mathcal{G}$			
Residues	$C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$	$C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$	$C\left({}^{6}_{i}\mathcal{M}, {}^{8}_{i}\mathcal{M}\right)$	$C\left({}^{21}_{i}\mathcal{M}, {}^{22}_{i}\mathcal{M}\right)$	Actin binding	$f_1\left({}_i\mathcal{M}\right)$
S83T	0.02403	0.08153	0.01511	0.19407		438
V87I	-0.00468	0.06226	0.00442	0.19407		43
P161A	0.02264	0.24512	0.00125	0.19407		820
Q164E	0.03766	0.13165	0.01125	0.19407		307
E168D	0.00917	0.13241	0.00984	0.19407		207
T176A	0.00567	0.15339	0.00740	0.19407		204
D186A	-0.00321	0.24408	0.00252	0.19407		188
A189P	-0.00725	0.28690	-0.00221	0.19407		109
N197H	-0.00107	0.25931	0.00426	0.19407		278
A206P	0.00098	0.05119	0.01027	0.08149		768
T217E	0.02192	0.23347	0.01228	0.19407		819
D220E	0.01581	0.03949	-0.00070	0.11018		628
T226V	-0.00024	0.22298	0.00951	0.19407		384
T227S	-0.00233	0.07602	0.00579	0.19407		328
C234T	0.00033	0.29774	0.00084	0.19407		479
V271I	-0.00281	0.06226	-0.00071	0.19407		395
K273R	-0.00389	0.11880	-0.00113	0.19407		292
V285I	0.03537	0.06226	0.01110	0.19407		372
V320I	0.00474	0.06226	-0.00023	0.19407		206
A326C	0.00771	0.31076	0.00082	0.19407		188
N331S	0.00691	0.20197	0.00510	0.25547	X	288
M335Q	0.03814	0.27765	0.01750	0.19407	X	111
G339Q	-0.02090	0.16000	-0.00380	0.15548		350
A342S	-0.00375	0.10550	-0.00168	0.19407		219
A358S	0.03212	0.10550	0.01011	0.19407		122
L361I	0.02974	0.02915	0.01903	0.15548		87
G390N	-0.00206	0.08175	-0.00219	0.11018		225
F393S	0.01559	0.34059	0.02170	0.19407		99
D403E	0.00208	0.11476	0.01418	0.19407		419
D417Q	-0.00424	0.19312	-0.00304	0.19407		790
A440N	-0.01732	0.24580	-0.00275	0.19407		312
D442E	0.00573	0.03949	0.00231	0.11018		438
A445S	-0.00473	0.10550	0.00362	0.19407		801
A446G	-0.00585	0.14825	0.00157	0.11018		795
K462G	-0.02254	0.33389	-0.00195	0.19407		
-463K	-0.00848	0.23160	-0.00183	0.19407		
AUC	0.83824	0.75000	0.77941	0.80147		
Threshold	0.00632	0.19754	0.00476	0.22477		
Sensitivity	0.67647	0.64709	0.61765	1.00000		
Specificity	1.00000	1.00000	1.00000	0.50000		

Table 4.3: **Estimated functional residues involving divergences between eEF1A1 and eEF1A2 [105].**

Residues	Hypothetical function
Q164E	Adjacent to the guanosine binding pocket
E168D	Adjacent to the guanosine binding pocket
N197H	Adjacent to the guanosine binding pocket
A206P	Adjacent to the guanosine binding pocket
T217E	Phosphorylatable
T227S	Phosphorylatable
V320I	Adjacent to aminoacyl-tRNA and eEF1Bα binding sites
A326C	Adjacent to an actin binding residue
N331S	Actin binding
M335Q	Actin binding
A358S	Adjacent to an actin binding residue
F393S	Phosphorylatable
A445S	Phosphorylatable

Chapter 5

Protein functional site prediction by sequence and structure

5.1 Introduction

The well known function of the translation elongation factor Tu (EF-Tu) and eukaryotic translation elongation factor 1A (eEF1A) is a carrier of aminoacyl tRNAs to the A site of the ribosome in bacteria and eukaryotes, respectively [5]. The eEF1A·GTP·tRNA complex elongates a new polypeptide chain upon the ribosome and this elongation event triggers GTP hydrolysis. The inactive GDP form of the eEF1A is recycled to the active GTP form by guanine nucleotide exchange factors including eukaryotic translation elongation factor 1Bα (eEF1Bα) [6]. These functions are inferred as a primary function of the EF-Tu/EF1A family.

Moonlighting functions have been discovered in the eEF1A proteins [27, 78, 1]. In around 1990, *Dictyostelium discoideum* eEF1A was identified as an actin-binding protein [121]. Gross et al. [32] have identified actin binding residues by site-directed mutagenesis of yeast eEF1A. Meanwhile, *Saccharomyces cerevisiae* has two eEF1A encoding genes [84]. Overexpression of these genes resulted in defective budding and enlarged cells [83] and eEF1A2 induces filopodia production in rodent and human cell lines [4]. This indicates that eEF1A2 regulates actin remodeling and cell motility. Additionally, eEF1A2 is likely to be an important human oncogene [114]. Overexpression of the eEF1A2 causes ovarian cancer probably triggered by inhibition of apoptosis [68]. The protection mechanism exerted by eEF1A2 may correlate with regulation of caspase-3 activation whereas the increase in eEF1A1 protein levels may facilitate rapid death of cells [97]. However, it is still unknown how each function works and the functions relate to each other.

In around 2012, it was discovered that EF1A binds a peptide called FNIII14, which involves in

anoikis that is apoptosis triggered by detachment from the extra cellular matrix [46]. This function is important to investigate how the cellular functions between protein biosynthesis and cell death are related. For further analysis of this crosstalk, it is needed to identify the binding site of FNIII14. However, the human EF1A1 consists of 462 amino acid residues. Therefore, identifications of important amino acids are time- and money-consuming. This chapter narrows down the candidates for the binding site of FNIII14 by using structure-based methods such as docking tools and molecular dynamics (MD) simulation. Finally, we combine the results of the structure-based method with the sequence-based method in Chapter 3.

5.2 Materials and Methods

5.2.1 Modeling of the complex structure

Human eEF1A1 was modeled by homology modeling using *Sacharomyces cerevisiae* EF1A (PDB ID: 1F60) as a template by the Discovery Studio 3.1 software package [2]. Human eEF1A1 was aligned to *Sacharomyces cerevisiae* EF1A by P160 and P161 as gaps. The model of human eEF1A1 was constructed by the Build Homology Modeling protocol.

FNIII14 was constructed from the crystal structure of fibronectin (PDB ID: 1FNH) by extracting T236-K257. The C-terminal lysine was converted into cysteine by the Discovery Studio 3.1 software package [2].

We used the ZDOCK program [20] for docking of eEF1A1 and FNIII14 by the number of grids, width and rotation angle as 128×128×128, 1.2 Å and 6°. We defined domains of eEF1A1 as domain 1: M1-T261, domain 2: V262-G327 and domain 3: D328-K462 and domain 3 of eEF1A1 was blocked because experimental results showed that domain 3 has the highest binding affinity (data not shown). Then, we conducted two dockings by changing blocking regions. In Docking 1, we blocked the domains 1 and 2 of eEF1A1. In Docking 2, we blocked the domains 1 and 2 of eEF1A1 except for the regions within 12 Å from domain 3.

5.2.2 MD simulation and its analysis

We used the AMBER software [18] for MD simulations and specified the AMBER2003 force field [26]. Then, we constructed a rectangular box whose size is ≈ 80×80×90 $Å^3$ and put ≈ 100,000 TIP-3P water molecules [51] and 2 chloride ions into the box. We conducted energy minimizations of 200 steps by the steepest descent method and 6,800 steps by the conjugate gradient method. 10 ps MD simulations were conducted by increasing the reference temperature to 300 K by ntb = 1. The time step was set to 2 fs by using the SHAKE algorithm for hydrogen atoms [98]. The cutoff of the Van der Waals (VdW) energy was set to 9 Å. The Coulomb energy was computed by the particle mesh Ewald (PME) algorithm [23].

Next, 10 ns MD simulations were conducted by ntb = 2 and the reference pressure was set to 1.0 bar. For analysis of the MD simulations, we sampled the MD trajectories every 10 ps. Based on the MD trajectories, a root mean square deviation (RMSD) was calculated. Let p_i and q_i denote a vector and

$$RMSD = \min_{R,v} \sqrt{\frac{1}{n}\sum_{i=1}^{n} |q_i - R\,(p_i + v)|^2}, \tag{5.1}$$

where R is a rotational matrix, v is a translational vector and n is the number of particles [52].

The binding energy is defined as

$$\Delta G_{bind} = G_{com} - (G_{rec} + G_{lig})\,, \tag{5.2}$$

where G_{com}, G_{rec} and G_{lig} are free energies of complex, receptor and ligand, respectively. The binding energy was calculated by molecular mechanics Poisson-Boltzmann surface area method [40, 109, 108] and is approximated as

$$\Delta G_{bind} = \Delta H - T\Delta S \approx \Delta \overline{E}_{MM} + \Delta \overline{G}_{solvation} - T\Delta S, \tag{5.3}$$

where ΔH is a difference of enthalpy, T is a temperature, ΔS is a difference of entropy, $\overline{E}_{MM}$ is a mean of differences of molecular mechanics energies and calculated as

$$\Delta \overline{E}_{MM} = \Delta \overline{E}_{internal} + \Delta \overline{E}_{elec} + \Delta \overline{E}_{vdw}, \tag{5.4}$$

where $\Delta \overline{E}_{internal}$ is a mean of differences of internal energies, $\Delta \overline{E}_{elec}$ is a mean of differences of electrostatic energies and $\Delta \overline{E}_{vdw}$ is a mean of differences of Van der Waals energies and $\Delta \overline{G}_{solvation}$ is a mean of differences of solvation energies and calculated as

$$\Delta \overline{G}_{solvation} = \Delta \overline{G}_{hydrophobic} + \Delta \overline{G}_{hydrophilic}, \tag{5.5}$$

where $\Delta \overline{G}_{hydrophobic}$ is a hydrophic contribution and a mean of differences of surface area energies and is calculated as

$$\Delta \overline{G}_{hydrophobic} = \gamma \Delta \overline{A} + b, \tag{5.6}$$

where γ is a coefficient, which was set to 0.00542, $\Delta \overline{A}$ is a mean of differences of surface areas, b is an intersection, which was set to 0.92 and $\Delta \overline{G}_{hydrophilic}$ is a hydrophilic contribution and a mean of differences of Poisson-Boltzmann energies, which was calculated by setting inter and exterior dielectric constants as 1 and 80 and Delphi II [39] of grid spacing as 0.5 Å.

For calculation of interacting residues from MD simulations, we defined that two residues are inter-

acted if the length between atoms belonging to EF1A1 and YTIYVIAL, which is the important sequence of FNIII14, is 3 Å or less and the rate of the interactions was computed from the MD trajectories.

5.2.3 Visualization

RMSD was computed by the ProDy Python package and visualized by matplotlib Python package [44]. Three-dimensional structures were visualized by the VMD program [43].

5.3 Results

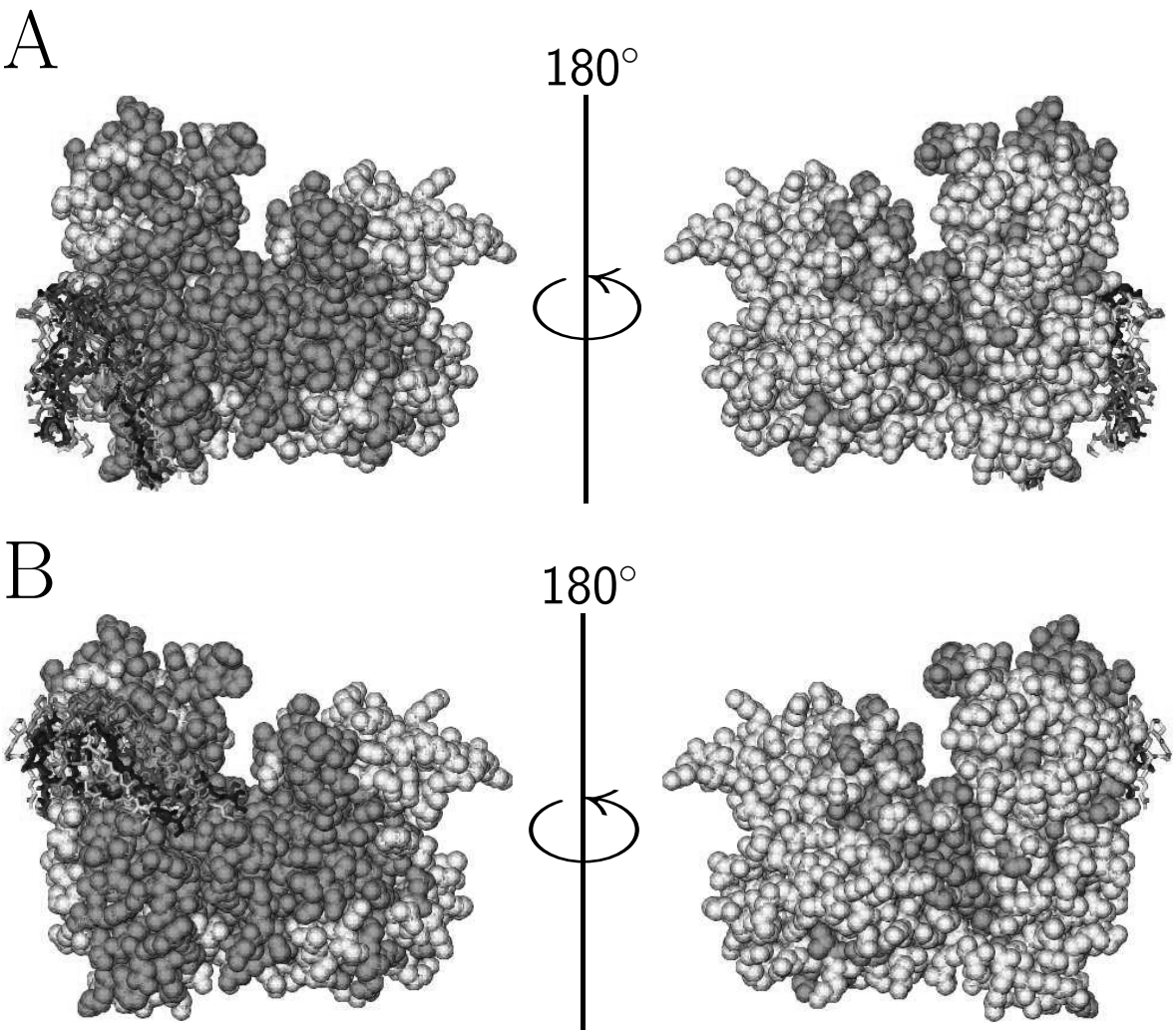


Figure 5.1: **Predicted complex structures of FNIII14 and eEF1A1.** The complex structures were obtained by the ZDOCK program by blocking (A) domains 1 and 2 and (B) domains 1 and 2 except for the border region of domain 3, respectively. (A) The rankings of ZDOCK scoring were 1, 2, 3, 6, 7 and 10. (B) The rankings of ZDOCK scoring were 1, 2, 3, 4 and 7. $f_1\left({}_i\mathcal{M}\right)$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}^{G}_{i}\mathcal{G}$. The amino acid residues were divided in harf by $f_1\left({}_i\mathcal{M}\right)$. If $f_1\left({}_i\mathcal{M}\right)$ is small, the atoms were colored by pink. If $f_1\left({}_i\mathcal{M}\right)$ is large, the atoms were colored by white.

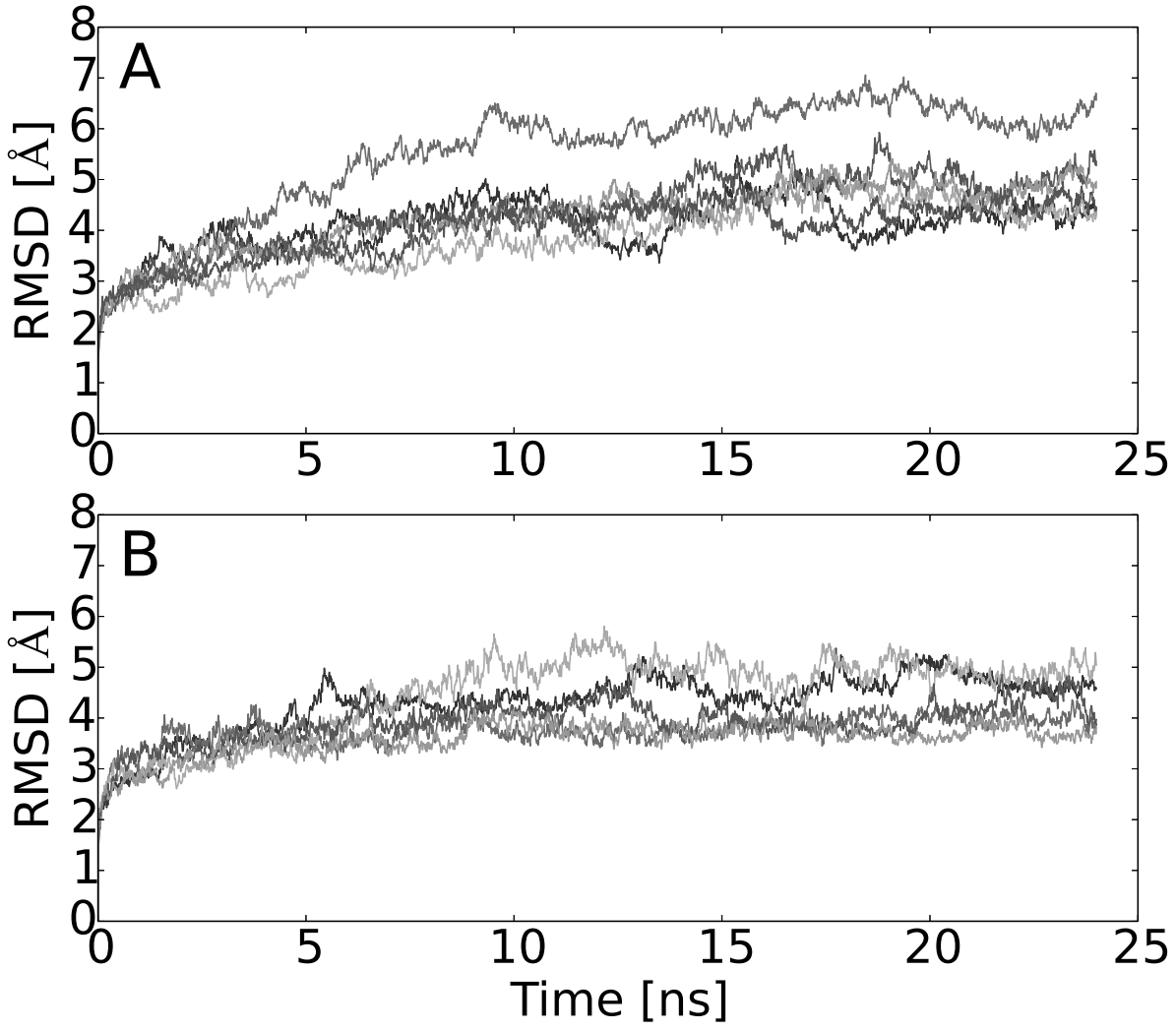


Figure 5.2: **RMSD of MD simulations.** Abscissas show time of the last 24 ns MD simulation. Ordinates show RMSD [Å] of (A) 6 independent MD trajectories from initial structures of Docking 1 and (B) 5 independent MD trajectories from initial structures of Docking 2.

Complex structures between human eEF1A1 and FNIII14 were predicted by the ZDOCK software. Figures 5.1A and 5.1B show that the predicted complex structures which have high ZDOCK scores are similar in Dockings 1 and 2, respectively, and the FNIII14s are bound to the regions whose amino acid residues have small $f_1\,({}_i\mathcal{M})$. In order to take account of dynamics of the complex structure, we conducted MD simulations using the complex structures as initial structures. Figure 5.2 shows that some trajectories do not have large changes in the RMSDs in the last 15 ns of the MD simulations and therefore the trajectories were used for computations of the binding energy and the rate of interaction. Table 5.1 shows that the mean of the binding energies of dockings 1 and 2 are almost same. Table 5.2 shows that some amino acid residues have small $f_1\,({}_i\mathcal{M})$ and a large rate of interaction.

5.4 Discussion

Figures 5.1A and 5.1B show that the complex structures whose ZDOCK scores, which consists of shape complementarity, electrostatics and desolvation, have high values are each similar. This implies that the binding regions are complement for the C-terminal region of FNIII14. However, Table 5.1 shows that the binding energies are almost same and therefore we could not determine which complex structure is stable. Meanwhile, Figure 5.1 shows that the binding regions of FNIII14s are the proximate regions predicted by the sequence analysis in Chapter 3. Therefore, the sequence analysis and the structural analysis showed similar predicted regions of eEF1A1. Additionally, as shown in Table 5.2, we computed the rate of interaction. The rate means that if the rate is small, the residue is not constantly proximate from FNIII14 but if the rate is large, the residue is constantly proximate from FNIII14. Therefore, amino acid residues which have small $f_1\left({}_i\mathcal{M}\right)$ and a large rate of interaction can be considered as important residues for binding with FNIII14.

5.5 Conclusions

The sequence and the structure analyses showed similar predicted regions of eEF1A1 and candidate functional amino acid residues can be selected from Table 5.2. In the future works, the predicted functional residues should be verified by experimental approaches such as site-directed mutagenesis. We hope that the detachment activity of FNIII14 is used for cancers which have resistance of anticancer drugs [54] and applied for clinical problems by developing a more useful FNIII14-like substance or a small compound which has a similar effect of FNIII14.

Table 5.1: **Binding energies of complex structures between FNIII14 and eEF1A1.**

Docking 1					
Complex	$\Delta\overline{E}_{elec}$	$\Delta\overline{E}_{vdw}$	$\Delta\overline{G}_{hydrophobic}$	$\Delta\overline{G}_{hydrophilic}$	$\Delta\overline{G}_{bind}$
1	-377.149	-75.393	-9.503	404.901	-57.140
2	-388.549	-57.471	-8.257	409.733	-44.542
3	-318.291	-70.949	-8.434	352.641	-45.033
4	-340.263	-51.481	-7.330	364.477	-34.599
5	-278.515	-61.470	-7.888	302.059	-45.815
6	-331.041	-58.237	-8.236	360.948	-36.566
Mean	-338.968	-62.500	-8.275	365.793	-43.949
Docking 2					
Complex	$\Delta\overline{E}_{elec}$	$\Delta\overline{E}_{vdw}$	$\Delta\overline{G}_{hydrophobic}$	$\Delta\overline{G}_{hydrophilic}$	$\Delta\overline{G}_{bind}$
1	-393.201	-93.633	-12.108	439.697	-59.242
2	-335.289	-81.710	-10.487	376.478	-51.009
3	-415.172	-61.606	-9.245	449.527	-36.497
4	-335.931	-53.589	-8.283	364.951	-32.852
5	-351.509	-63.574	-9.265	383.235	-41.113
Mean	-366.220	-70.822	-9.878	402.778	-44.143

(kcal/mol)

Table 5.2: **Sequence and structural analyses of eEF1A1.** $f_1\,({}_i\mathcal{M})$ was calculated by $T = N$, $g_x = g_3$, $h_x = h_1$, $\mathcal{A} = {}^{20}\mathcal{A}$ and ${}_i\mathcal{G} = {}_i^G\mathcal{G}$. The rate of interaction is calculated from the MD simulations. The residues whose rates of interactions are 0.1 or more were shown.

Docking 1			Docking 2		
Residue	$f_1\,({}_i\boldsymbol{M})$	Rate of interaction	Residue	$f_1\,({}_i\boldsymbol{M})$	Rate of interaction
H349	46	0.300	K100	2	0.716
P350	66	0.375	N101	22	0.691
G351	88	0.401	T104	50	0.335
Q352	88	0.559	G105	0	0.542
I353	79	0.292	T106	11	0.649
S354	140	0.457	S107	16	0.407
A355	283	0.243	P249	61	0.252
G356	112	0.485	L250	227	0.263
Y357	19	0.834	Q251	48	0.825
A358	122	0.970	D252	57	0.247
P359	35	0.712	V253	96	0.233
V360	42	0.744	R266	69	0.178
H367	31	0.148	P282	23	0.190
I368	119	0.260	V317	101	0.283
A369	41	0.664	K318	269	0.267
C370	50	0.250	V320	206	0.188
K371	261	0.514	R321	192	0.962
K392	130	0.178	R322	57	0.990
F393	99	0.198	G323	19	0.583
G407	163	0.301	H364	43	0.258
K408	278	0.200	T365	43	0.683
R427	34	0.323	A366	116	0.946
D428	36	0.658	H367	31	0.893
M429	28	0.696	I368	119	0.440
R430	54	0.252	K408	278	0.586
V433	110	0.335	P409	151	0.248
			R427	34	0.335
			R430	54	0.188

Chapter 6

Conclusions

The results in Chapter 3 showed that the conservation analysis can detect regions which involve protein biosynthesis. However, there were two problems. The first one is that the predicted residues are not all the known functional residues of EF-Tu/EF1A. In other words, the results in Chapter 4 showed that the conservation analysis in Chapter 3 cannot always detect functional sites because Tables 4.1 and 4.2 show that actin- or fibronectin-binding residues were not predictable. However, the functional residues were detected by the specific conservations in Chapter 4. This implies that an appropriate conservation should be selected based upon a type of the protein functions. In addition, the results also showed that the existing methods were difficult to predict protein functional sites comprehensively but a protein that has many functional sites was applicable for the proposed method. The second one is that there were many predicted residues as shown in Figure 3.7. However, the predicted residues were further narrowed down by the structural analysis in Chapter 5. This showed that combination of sequence and structural analyses is useful for determining candidate amino acid residues for site-directed mutagenesis.

Meanwhile, a feature of sequence-based methods is widely applicable for protein functional site prediction. This means that the method can determine candidate functional amino acid residues by following procedures: (1) Selection of one protein sequence (2) Collection of sequences by similarity search (3) Construction of an MSA for the collected sequences (4) Reconstruction of a phylogenetic tree from the MSA (5) Calculation of conservations from the MSA and the phylogenetic tree (6) Determination of candidate residues based on the conservations. The candidate amino acid residues are useful for site-directed mutagenesis and we can reduce time and costs for the experiments.

There are future works from biological and informatics aspects, respectively. As for the former, the results are useful for investigating other functions such as nuclear export, protein degradation, cytoskeltal regulation and viral functions and may expedite identification of functional sites involving unknown functions of EF-Tu/EF1A. As for the latter, there are two future plans. The first one is integration of other existing methods. Chapter 2 integrated some existing methods such as SE- and

SP-based methods and ET methods but not integrated some methods such as the relative entropy based method, the Jensen-Shannon divergence based method and ConSurf and Rate4Site algorithms. Therefore, our idea is not inclusive all the existing methods. In the future works, an integrated method of other existing methods should evaluate how much the performance is changed by how to treat amino acids or their background distribution, how to reconstruct a phylogenetic tree or how to take account of its branch lengths. The second one is application for other proteins because this study only investigated EF-Tu/EF1A but not other proteins. Analyzing whether other proteins are applicable or not is necessary to evaluate the prediction performance.

Abbreviations

A	alanine
ACh	acethylcholine
aEF	archaeal elongation factor
ANM	anisotropic network model
AUC	area under the curve
BoNT	Botulinum neurotoxin
C	cysteine
D	aspartic acid
DNA	deoxyribonucleic acid
E	glutamic acid
eEF	eukaryotic elongation factor
EF	elongation factor
ET	evolutionary trace
F	phenylalanine
G	glycine
GDP	guanosine diphosphate
GTP	guanosine triphosphate
H	histidine
HA	hemagglutinin
I	isoleucine

iv-ET integer-valued evolutionary trace

K lysine

L leucine

L-PTC large progenitor toxin complex

M methionine

MD molecular dynamics

Mg *Mycoplasma genitalium*

Mp *Mycoplasma pneumoniae*

MSA Multiple sequence alignment

M-PTC minimally functional progenitor toxin complex

N asparagine

NMA normal mode analysis

NTNHA non-toxic non-hemagglutinin

P proline

PCA principal component analysis

PDB Protein Data Bank

PME particle mesh Ewald

Q glutamine

R arginine

RMSD root mean square deviation

RNA ribonucleic acid

ROC receiver operating characteristic

rv-ET real-valued evolutionary trace

S serine

SE Shannon entropy

SEP Shannon entropy of residue properties

SNAP-25	synaptosomal-associated protein of 25 kDa
SP	sum of pairs
SPW	sum of pairs with weighting
T	threonine
tRNA	transfer ribonucleic acid
V	valine
VAMP	vesicle-associated membrane protein
VdW	Van der Waals
W	tryptophan
WET	weighted evolutionary trace
Y	tyrosine

Acknowledgements

At this point, I would like to thank everyone who helped finishing this thesis.

Special thanks go to my mentor, Prof. Satoru Miyazaki, who has been always there to listen and provided invaluable help throughout my undergraduate and graduate careers. His constant support has been essential to my development as a researcher. Without his patience and steadfast guidance, this work would not have been possible.

I owe great gratitude to my supervisor, Prof. Toshiyuki Kaji, for providing me an opportunity to access to the laboratory and for the time devoted to putting me through in the course of undertaking this project. My deepest gratitude is also to Prof. Fumio Fukai for providing useful information in my research. My sincere thanks also go to Prof. Takao Aoyama and Prof. Masataka Mochizuki for taking the time to serve on my committee.

I would also like to thank all the current and former members at the laboratory for their support and for some much needed humor and entertainment. In particular, I am grateful to Ph.D. Yeondae Kwon for interesting discussion and providing insightful comments and valuable feedback on my work over the past several years.

Last but not the least, I am thankful to my family for supporting me spiritually throughout in my life. I want to acknowledge the financial and moral contributions of my parents, Mr. and Mrs. Kondo, and my sibling to finishing this program. My family's unwavering love and support are the foundation for all I have accomplished. Special thanks go to my family for their support.

Bibliography

[1] W. Abbas, A. Kumar, and G. Herbein. The eEF1A proteins: at the crossroads of oncogenesis, apoptosis, and viral infections. *Front Oncol*, 5:75, 2015.

[2] Accerlys Inc. *Discovery Studio, Version 3.1.* San Diego, CA, 2011.

[3] A. Amadei, A. Linssen, and H. Berendsen. Essential dynamics of proteins. *Proteins-structure function and genetics*, 17(4):412–425, 1993.

[4] A. Amiri, F. Noei, S. Jeganathan, G. Kulkarni, D. E. Pinke, and J. M. Lee. eEF1A2 activates Akt and stimulates Akt-dependent actin remodeling, invasion and migration. *Oncogene*, 26(21):3027–3040, 2007.

[5] G. Andersen, P. Nissen, and J. Nyborg. Elongation factors in protein biosynthesis. *Trends Biochem Sci*, 28(8):434–441, 2003.

[6] G. Andersen, L. Pedersen, L. Valente, I. Chatterjee, T. Kinzy, M. Kjeldgaard, and J. Nyborg. Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A : eEF1Bα. *Mol Cell*, 6(5):1261–1266, 2000.

[7] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, E. Alpi, R. Antunes, J. Arganiska, E. B. Casanova, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, W. M. Chan, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, E. Dimmer, F. Fazzini, P. Gane, A. Fedotov, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, J. Jacobsen, R. Jones, D. Legge, W. Liu, J. Luo, A. MacDougall, P. Mutowo, A. Nightingale, S. Orchard, S. Patient, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, T. Sawford, H. Sehra, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, M. Corbett, M. Donnelly, P. van Rensburg, M. Goujon, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, P.-A. Binz, M.-C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, E. de Castro, L. Cerutti, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James,

F. Jungo, G. Keller, V. Lara, P. Lemercier, J. Lew, D. Lieberherr, X. Martin, P. Masson, A. Morgat, T. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.-L. Veuthey, M. Zerara, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, H. Huang, A. Kukreja, K. Laiho, P. McGarvey, D. A. Natale, T. G. Natarajan, N. V. Roberts, B. E. Suzek, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, M. S. Yerramalla, J. Zhang, and U. Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res*, 41(D1):D43–D47, 2013.

[8] A. Armon, D. Graur, and N. Ben-Tal. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307(1):447–463, 2001.

[9] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*, 38(2):W529–W533, 2010.

[10] A. Bakan, L. M. Meireles, and I. Bahar. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

[11] S. Balasubramanian, T. R. Kannan, P. J. Hart, and J. B. Baseman. Amino acid changes in elongation factor Tu of *Mycoplasma pneumoniae* and *Mycoplasma genitalium* influence fibronectin binding. *Infect Immun*, 77(9):3533–3541, 2009.

[12] J. Baseman, S. Dallo, J. Tully, and D. Rose. Isolation and characterization of *Mycoplasma genitalium* strains from the human respiratory tract. *J Clin Microbiol*, 26(11):2266–2269, 1988.

[13] D. A. Benefield, S. K. Dessain, N. Shine, M. D. Ohi, and D. B. Lacy. Molecular assembly of botulinum neurotoxin progenitor complexes. *Proc Natl Acad Sci USA*, 110(14):5630–5635, 2013.

[14] S. Benner, M. Cohen, and G. Gonnet. Amino-acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*, 7(11):1323–1332, 1994.

[15] H. Berendsen, J. Postma, W. Vangunsteren, A. Dinola, and J. Haak. Molecular-dynamics with coupling to an external bath. *J Chem Phys*, 81(8):3684–3690, 1984.

[16] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, JAN 1 2000.

[17] J. A. Capra and M. Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[18] D. A. Case, T. A. Darden, Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Schafmeister, W. S. Ross, and P. A. Kollman. *AMBER 9*. University of California, San Francisco, 2006.

[19] D. Chambers, J. Peters, and C. Abbott. The lethal mutation of the mouse wasted (*wst*) is a deletion that abolishes expression of a tissue-specific isoform of translation elongation factor 1α, encoded by the *Eef1a2* gene. *Proc Natl Acad Sci U S A*, 95(8):4463–4468, 1998.

[20] R. Chen, L. Li, and Z. Weng. ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, 52(1):80–87, 2003.

[21] T. U. Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res*, 43(D1):D204–D212, 2015.

[22] S. Dai, F. Crawford, P. Marrack, and J. W. Kappler. The structure of HLA-DR52c: comparison to other HLA-DRB3 alleles. *Proc Natl Acad Sci U S A*, 105(33):11893–11897, 2008.

[23] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J Chem Phys*, 98(12):10089–10092, 1993.

[24] M. O. Dayhoff and R. M. Schwartz. Chapter 22: a model of evolutionary change in proteins. In *in Atlas of Protein Sequence and Structure*, 1978.

[25] S. Dineen, M. Bradshaw, and E. Johnson. Neurotoxin gene clusters in *Clostridium botulinum* type A strains: sequence comparison and evolutionary implications. *Curr Microbiol*, 46(5):345–352, 2003.

[26] Y. Duan, C. Wu, S. Chowdhury, M. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem*, 24(16):1999–2012, 2003.

[27] S. Ejiri. Moonlighting functions of polypeptide elongation factor 1: from actin bundling to zinc finger protein R1-associated nuclear localization. *Biosci Biotechnol Biochem*, 66(1):1–21, 2002.

[28] C. Fang, T. Noguchi, and H. Yamana. Simplified sequence-based method for ATP-binding prediction using contextual local evolutionary conservation. *Algorithms Mol Biol*, 9, 2014.

[29] J. Felsenstein. *PHYLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

[30] M. Goulet, R. Dular, J. Tully, G. Billowes, and S. Kasatiya. Isolation of *Mycoplasma pneumoniae* from the human urogenital tract. *J Clin Microbiol*, 33(11):2823–2825, 1995.

[31] M. K. Groftehauge, M. O. Therkelsen, R. Taaning, T. Skrydstrup, J. P. Morth, and P. Nissen. Identifying ligand-binding hot spots in proteins using brominated fragments. *Acta Crystallogr F Struct Biol Commun*, 69(9):1060–1065, 2013.

[32] S. R. Gross and T. G. Kinzy. Improper organization of the actin cytoskeleton affects protein synthesis at initiation. *Mol Cell Biol*, 27(5):1974–1989, 2007.

[33] S. Gu, S. Rumpel, J. Zhou, J. Strotmeier, H. Bigalke, K. Perry, C. B. Shoemaker, A. Rummel, and R. Jin. Botulinum neurotoxin is shielded by NTNHA in an interlocked complex. *Science*, 335(6071):977–981, 2012.

[34] M. Gültas, G. Düzgün, S. Herzog, S. J. Jäger, C. Meckbach, E. Wingender, and S. Waack. Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. *BMC Bioinformatics*, 15, 2014.

[35] S. E. Heffron, S. Mui, A. Aorora, K. Abel, E. Bergmann, and F. Jurnak. Molecular complementarity between tetracycline and the GTPase active site of elongation factor Tu. *Acta Crystallogr D Biol Crystallogr*, 62:1392–1400, 2006.

[36] S. Henikoff and J. Henikoff. Position-based sequence weights. *J Mol Biol*, 243(4):574–578, 1994.

[37] R. Herrmann and B. Reiner. *Mycoplasma pneumoniae* and *Mycoplasma genitalium*: a comparison of two closely related bacterial species. *Curr Opin Microbiol*, 1(5):572–579, 1998.

[38] B. Hess, H. Bekker, H. Berendsen, and J. Fraaije. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*, 18(12):1463–1472, SEP 1997.

[39] B. HONIG and A. NICHOLLS. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[40] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model*, 51(1):69–82, 2011.

[41] Y.-F. Huang and G. B. Golding. Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Comput Biol*, 10(1), 2014.

[42] Y.-F. Huang and G. B. Golding. FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. *Bioinformatics*, 31(4):523–531, 2015.

[43] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J Mol Graphics Modell*, 14(1):33–38, 1996.

[44] J. D. Hunter. Matplotlib: A 2D graphics environment. *Comput Sci Eng*, 9(3):90–95, MAY-JUN 2007.

[45] K. Inoue, Y. Fujinaga, T. Watanabe, T. Ohyama, K. Takeshi, K. Moriishi, H. Nakajima, K. Inoue, and K. Oguma. Molecular composition of *Clostridium botulinum* type A progenitor toxins. *Infect Immun*, 64(5):1589–1594, MAY 1996.

[46] K. Itagaki, T. Naito, R. Iwakiri, M. Haga, S. Miura, Y. Saito, T. Owaki, S. Kamiya, T. Iyoda, H. Yajima, S. Iwashita, S.-I. Ejiri, and F. Fukai. Eukaryotic translation elongation factor 1A induces anoikis by triggering cell detachment. *J Biol Chem*, 287(19):16037–16046, 2012.

[47] J.-O. Janda, A. Popal, J. Bauer, M. Busch, M. Klocke, W. Spitzer, J. Keller, and R. Merkl. H2rs: deducing evolutionary and functionally important residue positions by means of an entropy and similarity based analysis of multiple sequence alignments. *BMC Bioinformatics*, 15, 2014.

[48] J. Jensen. *Mycoplasma genitalium*: the aetiological agent of urethritis and other sexually transmitted diseases. *J Eur Acad Dermatol Venereol*, 18(1):1–11, 2004.

[49] F. Johansson and H. Toh. A comparative study of conservation and variation scores. *BMC Bioinformatics*, 11, 2010.

[50] D. Jones, W. Taylor, and J. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–282, 1992.

[51] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Kkein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79(2):926–935, 1983.

[52] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 1976.

[53] S. Kahns, A. Lund, P. Kristensen, C. Knudsen, B. Clark, J. Cavallius, and W. Merrick. The elongation factor 1 A-2 isoform from rabbit: cloning of the cDNA and characterization of the protein. *Nucleic Acids Res*, 26(8):1884–1890, 1998.

[54] R. Kato, T. Ishikawa, S. Kamiya, F. Oguma, M. Ueki, S. Goto, H. Nakamura, T. Katayama, and F. Fukai. A new type of antimetastatic peptide derived from fibronectin. *Clin Cancer Res*, 8(7):2455–2462, 2002.

[55] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[56] T. Kawabata. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys J*, 95(10):4643–4658, 2008.

[57] R. T. Kidmose, N. N. Vasiliev, A. B. Chetverin, G. R. Andersen, and C. R. Knudsen. Structure of the Qβ replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins. *Proc Natl Acad Sci U S A*, 107(24):10884–10889, 2010.

[58] H. Kishino, T. Miyata, and M. Hasegawa. Maximum-likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol*, 31(2):151–160, 1990.

[59] S. Knudsen, J. Frydenberg, B. Clark, and H. Leffers. Tissue-dependent variation in the expression of elongation factor-1α isoforms: isolation and characterization of a cDNA encoding a novel variant of human elongation-factor 1α. *Eur J Biochem*, 215(3):549–554, 1993.

[60] K. Kobayashi, K. Saito, R. Ishitani, K. Ito, and O. Nureki. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1α complex. *Nucleic Acids Res*, 40(18):9319–9328, 2012.

[61] Y. Kondo, Y. Kwon, and S. Miyazaki. Detection of key residues involving functional divergence into the translation elongation factor Tu/1A family using quantitative measurements for specific conservation of protein subfamilies. *J Comput Sci Syst Biol*, 7:054–061, 2014.

[62] Y. Kondo and S. Miyazaki. Dynamics of the large progenitor toxin complex of *Clostridium botulinum*. *J Comput Sci Syst Biol*, 9(1):006–009, 2015.

[63] Y. Kondo and S. Miyazaki. Protein functional site prediction using a conservative grade and a proximate grade. *J Data Mining Genomics Proteomics*, 6:175, 2015.

[64] T. I. Lam, L. H. Stanker, K. Lee, R. Jin, and L. W. Cheng. Translocation of botulinum neurotoxin serotype A and associated proteins across the intestinal epithelia. *Cell Microbiol*, 17(8):1133–1143, 2015.

[65] M. J. LaMarche, J. A. Leeds, K. Amaral, J. T. Brewer, S. M. Bushell, J. M. Dewhurst, J. Dzink-Fox, E. Gangl, J. Goldovitz, A. Jain, S. Mullin, G. Neckermann, C. Osborne, D. Palestrant, M. A. Patane, E. M. Rann, M. Sachdeva, J. Shao, S. Tiamfbok, L. Whitehead, and D. Yu. Antibacterial optimization of 4-aminothiazolyl analogues of the natural product GE2270 A: identification of the cycloalkylcarboxylic acids. *J Med Chem*, 54(23):8099–8109, 2011.

[66] R. Landgraf, D. Fischer, and D. Eisenberg. Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng*, 12(11):943–951, 1999.

[67] T. Lasko, J. Bhagwat, K. Zou, and L. Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*, 38(5):404–415, 2005.

[68] J. Lee. The role of protein elongation factor eEF1A2 in ovarian cancer. *Reprod biol Endocrinol*, 1:69, 2003.

[69] K. Lee, S. Gu, L. Jin, T. T. N. Le, L. W. Cheng, J. Strotmeier, A. M. Kruel, G. Yao, K. Perry, A. Rummel, and R. Jin. Structure of a bimodular botulinum neurotoxin complex provides insights into its oral toxicity. *PLoS Pathogens*, 9(10), OCT 2013.

[70] K. Lee, K.-H. Lam, A.-M. Kruel, S. Mahrhold, K. Perry, L. W. Cheng, A. Rummel, and R. Jin. Inhibiting oral intoxication of botulinum neurotoxin a complex by carbohydrate receptor mimics. *Toxicon*, 107, Part A:43 – 49, 2015. Highlights of the TOXINS 2015 Meeting.

[71] K. Lee, W.-H. Lam, A. M. Kruel, K. Perry, A. Rummel, and R. Jin. High-resolution crystal structure of HA33 of botulinum neurotoxin type B progenitor toxin complex. *Biochem Biophys Res Commun*, 446(2):568–573, 2014.

[72] K. Lee, X. Zhong, S. Gu, A. M. Kruel, M. B. Dorner, K. Perry, A. Rummel, M. Dong, and R. Jin. Molecular basis for disruption of E-cadherin adhesion by botulinum neurotoxin A complex. *Science*, 344(6190):1405–1410, 2014.

[73] S. Lee, A. Francoeur, S. Liu, and E. Wang. Tissue-specific expression in mammalian brain, heart, and muscle of S1, a member of the elongation factor-1α gene family. *J Biol Chem*, 267(33):24064–24068, 1992.

[74] O. Lichtarge, H. Bourne, and F. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–358, 1996.

[75] T. Liu, Y. Yang, D. Wang, Y. Xiao, G. Du, L. Wu, M. Ding, L. Li, and C. Wu. Human eukaryotic elongation factor 1A forms oligomers through specific cysteine residues. *Acta Biochim Biophys Sin*, 47(12):1011–1017, 2015.

[76] A. Lund, S. Knudsen, H. Vissing, B. Clark, and N. Tommerup. Assignment of human elongation factor 1α genes: *EEF1A* maps to chromosome 6q14 and *EEF1A2* to 20q13.3. *Genomics*, 36(2):359–361, 1996.

[77] F. Mansilla, I. Friis, M. Jadidi, K. Nielsen, B. Clark, and C. Knudsen. Mapping the human translation elongation factor eEF1H complex using the yeast two-hybrid system. *Biochem J*, 365(3):669–676, 2002.

[78] M. K. Mateyak and T. G. Kinzy. eEF1A: thinking outside the ribosome. *J Biol Chem*, 285(28):21209–21213, 2010.

[79] T. Matsumura, Y. Sugawara, M. Yutani, S. Amatsu, H. Yagita, T. Kohda, S.-I. Fukuoka, Y. Nakamura, S. Fukuda, K. Hase, H. Ohno, and Y. Fujinaga. Botulinum toxin A complex exploits intestinal M cells to enter the host and exert neurotoxicity. *Nat Commun*, 6, 2015.

[80] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791, 2004.

[81] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, 336(5):1265–1282, 2004.

[82] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI coarse-grained force field: extension to proteins. *J Chem Theory Comput*, 4(5):819–834, 2008.

[83] R. Munshi, K. Kandl, A. Carr-Schmid, J. Whitacre, A. Adams, and T. Kinzy. Overexpression of translation elongation factor 1A affects the organization and function of the actin cytoskeleton in yeast. *Genetics*, 157(4):1425–1436, 2001.

[84] K. Nagashima, M. Kasai, S. Nagata, and Y. Kaziro. Structure of the two genes coding for polypeptide chain elongation factor 1α (EF-1α) from *Saccharomyces cerevisial*. *Gene*, 45(3):265–273, 1986.

[85] T. Nakamura, T. Tonozuka, A. Ide, T. Yuzawa, K. Oguma, and A. Nishikawa. Sugar-binding sites of the HA1 subcomponent of *Clostridium botulinum* type C progenitor toxin. *J Mol Biol*, 376(3):854–867, 2008.

[86] H. Newbery, T. Gillingwater, P. Dharmasaroja, J. Peters, S. Wharton, D. Thomson, R. Ribchester, and C. Abbott. Progressive loss of motor neuron function in wasted mice: effects of a spontaneous null mutation in the gene for the eEF1A2 translation factor. *J Neuropathol Exp Neurol*, 64(4):295–303, 2005.

[87] P. Nissen, S. Thirup, M. Kjeldgaard, and J. Nyborg. The crystal structure of Cys-tRNA$^{\mathrm{Cys}}$-EF-Tu-GDPNP reveals general and specific features in the ternary complex and in tRNA. *Structure*, 7(2):143–156, 1999.

[88] I. Ohishi and G. Sakaguchi. Oral toxicities of *Clostridium botulinum* type C and type D toxins of different molecular sizes. *Infect Immun*, 28(2):303–309, 1980.

[89] I. Ohishi, S. Sugii, and G. Sakaguchi. Oral toxicitees of *Clostridium botulinum* toxins in response to molecular size. *Infect Immun*, 16(1):107–109, 1977.

[90] C. Oostenbrink, A. Villa, A. Mark, and W. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem*, 25(13):1656–1676, 2004.

[91] A. Parmeggiani, I. Krab, T. Watanabe, R. Nielsen, C. Dahlberg, J. Nyborg, and P. Nissen. Enacyloxin IIa pinpoints a binding pocket of elongation factor Tu for development of novel antibiotics. *J Biol Chem*, 281(5):2893–2900, 2006.

[92] M. Parrinello and A. Rahman. Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J Appl Phys*, 52(12):7182–7190, 1981.

[93] X. Periole, M. Cavalli, S.-J. Marrink, and M. A. Ceruso. Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *J Chem Theory Comput*, 5(9):2531–2543, 2009.

[94] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[95] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(Suppl 1):S71–S77, 2002.

[96] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Mueller. pROC: an open-source package for R and S plus to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 2011.

[97] L. Ruest, R. Marcotte, and E. Wang. Peptide elongation factor eEF1A-2/S1 expression in cultured differentiated myotubes and its protective effect against caspase-3-mediated apoptosis. *J Biol Chem*, 277(7):5418–5425, 2002.

[98] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*, 23(3):327 – 341, 1977.

[99] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991.

[100] G. Schiavo, M. Matteoli, and C. Montecucco. Neurotoxins affecting neuroexocytosis. *Physiol Rev*, 80(2):717–766, 2000.

[101] P. Shenkin, B. Erman, and L. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins*, 11(4):297–313, 1991.

[102] P. Sibbald and P. Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol*, 216(4):813–818, 1990.

[103] P. H. A. Sneath and R. R. Sokal. *Numerical taxonomy: the principles and practice of numerical classification*. W. H. Freeman, San Francisco, 1973.

[104] D. C. Soares and C. M. Abbott. Highly homologous eEF1A1 and eEF1A2 exhibit differential post-translational modification with significant enrichment around localised sites of sequence variation. *Biol Direct*, 8, 2013.

[105] D. C. Soares, P. N. Barlow, H. J. Newbery, D. J. Porteous, and C. M. Abbott. Structural models of human eEF1A1 and eEF1A2 reveal two distinct surface clusters of sequence variation and potential differences in phosphorylation. *PLoS ONE*, 4(7), 2009.

[106] C. Spearman. The proof and measurement of association between two things. *Int J Epidemiol*, 39(5):1137–1150, 2010.

[107] Y. Sugawara, M. Yutani, S. Amatsu, T. Matsumura, and Y. Fujinaga. Functional dissection of the *Clostridium botulinum* Type B hemagglutinin complex: identification of the carbohydrate and E-Cadherin binding sites. *PLoS ONE*, 9(10), 2014.

[108] H. Sun, Y. Li, M. Shen, S. Tian, L. Xu, P. Pan, Y. Guan, and T. Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys Chem Chem Phys*, 16(40):22035–22045, 2014.

[109] H. Sun, Y. Li, S. Tian, L. Xu, and T. Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys Chem Chem Phys*, 16(31):16719–16729, 2014.

[110] D. Takeshita and K. Tomita. Assembly of Qβ viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. *Proc Natl Acad Sci U S A*, 107(36):15733–15738, 2010.

[111] D. Takeshita and K. Tomita. Molecular basis for RNA polymerization by Qβ replicase. *Nat Struct Mol Biol*, 19(2):229–237, 2012.

[112] D. Takeshita, S. Yamashita, and K. Tomita. Mechanism for template-independent terminal adenylation activity of Qβ replicase. *Structure*, 20(10):1661–1669, 2012.

[113] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 28(10):2731–2739, 2011.

[114] S. Thornton, N. Anand, D. Purcell, and J. Lee. Not just for housekeeping: protein initiation and elongation factors in cell growth and tumorigenesis. *J Mol Med*, 81(9):536–548, 2003.

[115] W. Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, 2002.

[116] W. Valdar and J. Thornton. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–124, 2001.

[117] L. Vogeley, G. Palm, J. Mesters, and R. Hilgenfeld. 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(20):17149–17155, 2001.

[118] K. Waites and D. Talkington. *Mycoplasma pneumoniae* and its role as a human pathogen. *Clin Microbiol Rev*, 17(4):697–728, 2004.

[119] K. Wang and R. Samudrala. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7, 2006.

[120] R. M. Williamson. Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J Theor Biol*, 174(2):179–188, 1995.

[121] F. Yang, M. Demma, V. Warren, S. Dharmawardhane, and J. Condeelis. Identification of an actin-binding protein from *Dictyostelium* as elongation factor-1A. *Nature*, 347(6292):494–496, 1990.

[122] G. Yao, K. Lee, S. Gu, K.-H. Lam, and R. Jin. Botulinum neurotoxin A complex recognizes host carbohydrates through its hemagglutinin component. *Toxins*, 6(2):624–635, FEB 2014.

Appendices

Appendix A

Accession numbers

In the UniprotKB/SwissProt, all accession numbers used in this study are below: Q5R4R8, P50256, A3CTG3, Q8U152, O59949, A5CCA0, Q53871, A1AIF3, A9KWA0, Q40450, Q9ZEU3, P33165, B2S0H9, A0K3L0, P50371, Q97EH5, Q83ES6, B8DLL9, B0TX03, Q5FKR8, Q72NF9, Q2RFP5, Q1BDD3, B6JET1, B4S5M9, B1JDW6, A8GVB2, A4XBP8, P64028, Q8P1W4, Q2JUX4, Q9TMM9, P0DH99, Q05639, P54959, A8ABM5, P27592, P35021, Q3A9R3, A7NR65, A7FNJ0, A5D5I8, Q8DCQ7, Q1KVS9, A7Z0N5, C5C0J3, A9M5Q2, B9MQH1, P42473, Q877L9, Q3Z7S9, A4W5A0, Q39Y08, P56003, A5VJ92, A0L5V8, Q4A7K0, Q9TKZ5, B4SBU5, Q46IW4, Q981F7, A8GT71, Q83JC4, Q3K1U4, A3CP09, B7IHU4, Q8D240, Q27139, P86939, P0CN31, Q12WT3, Q00080, B6YVG2, B1IPW0, A3Q968, A6TWJ8, Q2K9L8, A1JS52, A8MLC4, B7IT17, Q7VRP0, O31297, A7H4R3, Q3KM40, Q8FS84, B1YGU8, P50064, P42480, A5IHR6, Q6F0J5, Q6MU81, Q5YPG4, B3QZH5, P09591, Q2IXR2, Q21M86, C1CLI6, B9K884, Q877P8, P86934, Q9YAV0, Q18EY5, A4YCR6, Q8LPC4, Q7N9B1, P60338, Q1R5U4, Q0HNT9, A9NEN4, A8EW02, A8F982, A1QZR2, A2S7F9, B1IGF6, B1WQY4, A8LLG2, A7NEC7, P43926, Q1GAQ0, Q03YI2, B1MGH7, P9WNN1, Q0P3M7, A2C4U5, Q1Q8P2, Q8KT97, A6WHR4, A8YZP5, Q1J7N4, A5GIP0, Q83NT9, P62629, Q2HJN8, A8MAJ1, A6UV43, Q41011, P19486, A7ZSL4, A8G8E0, Q1CN86, Q3IJV1, Q5GRY3, Q8UE16, A0R8H8, B3DT29, B0CH34, A7GZK6, Q9Z9A7, P18905, Q1IX70, C4Z2R9, C5D3R5, P42477, A1TYJ5, B8ZSC1, Q2YAZ9, Q18CE4, Q48D34, Q7UMZ0, Q160Y4, Q01SX2, O33594, Q5M5I8, Q47LJ1, Q3BWY6, Q59QD6, O29325, P07810, Q09069, A4VHL6, Q73IX6, Q1H4N9, P46280, A3M1F6, P13552, A6KYK9, C0QVZ4, Q3JMP6, A5N4N1, B7K834, Q2GFN6, Q0BKB8, Q7VJ74, A8YUS2, B8DAY7, C1AL18, Q5F5Q8, A7HWP7, A2BYN4, Q8XGZ0, Q8KTA6, A4YBY5, Q8CQ81, Q48UK5, Q3AW53, B5ZC31, P05303, Q9Y713, P51554, Q74MI6, A3DMQ1, A1USC1, Q21SF0, B0RU84, A5CCL4, Q0AUG3, A3N246, A8HTW6, A9ISD9, C0ZIH6, Q2SU25, Q255F3, Q8XFP8, P50373, Q8R603, B5Z8K3, A2RMT1, B1HMZ0, P13927, P64027, Q6CZW6, Q7V500, A9WSW5, Q8KTA3, A8G1F0, B5XKI1, P42476, A5F3K0, P0CT31, P62632, Q90835, A7I656, Q6L202, C5A5P4, A1AGM6, A9KW88, Q66FQ9, A4VHM8, B0RU96, Q0VSL7, B7JKB7, B7GU46, A7I3U7, B0B7N8, C3PKP2, P33168, P02991,

A5GAW4, B0UV21, Q1MPT8, A6W394, B2HSL3, A6Q1L5, A0T0K6, A6UZH4, Q0SFF4, Q5LMR5, Q1GP97, B4U3U1, B9DRL9, A6LLL1, Q5GWR8, Q5R1X2, Q2HJN9, P14963, Q8PUR8, A3MV69, Q03033, Q0AIJ7, P40174, A8A779, A3DBA0, P29544, B7GJ65, Q5L890, Q661E5, P33167, P42474, B0RB36, A9NAK7, Q727D5, Q0RRS3, P19457, C5CGR6, P09953, A1UBL1, Q04FQ4, P51287, B0KK53, A8GPF2, P0A1H6, A6QEK0, C1C881, A0LIH6, P68158, P06805, P17508, Q8SS29, A9A9U3, O93729, Q0W8G2, Q1H4Q1, Q43467, Q6AP73, A1AX82, Q664R7, Q2GJ61, Q814C4, B7J241, Q05FI3, Q7M7F1, Q250N4, A6GYU7, B2GIL2, Q055E6, B0JSE0, Q4A597, A6X0A2, A4SUU7, A5VXN3, A4WVL0, A9MT05, Q2YSB3, B5E653, Q2LQA3, P42481, P68104, P14865, P41203, A2STF0, P32186, P43643, A1WVC4, Q0SZX8, Q3A9P8, A7NS01, Q1CCT9, A7HBL7, A7GK18, Q79G84, Q057A2, B9KFF9, P56292, C4LL63, B8J1A0, P42475, Q0BUQ2, A6W5T5, B0SAF6, A2SLF9, Q98QG1, Q2G8Y2, A1VIP8, C3K2X8, A3PGI1, Q57H76, Q2G0N0, Q04N79, A6Q6H4, Q72GW4, Q66RN5, P17507, P0CN30, Q2FRI3, Q01520, A1RXW9, Q9ZT91, P0CE47, Q8EK81, A1USL2, Q21RV6, A7FNN8, B4RYQ8, C1ET37, Q492B2, O31298, O69303, B0BBV3, A4QBH0, C6C171, C4KZP9, A9H3R7, Q5QWA3, Q5ZYP5, Q9MUP0, Q73SD1, A1SNN5, A5DN78, Q3ZJ24, Q07KJ2, A4FPM7, A7WYX6, P72483, C1CSB0, A5IM81, P02992, P08736, P0CT54, Q96WZ1, Q464Z4, Q9HDF6, Q5JFZ4, Q0TCC0, Q089R8, A1WCN6, A5WH42, Q8PC51, Q0ABH7, Q73F98, Q06J54, B8D9U9, A8FKQ5, B3QY22, Q6NJD5, P14634, P49411, Q5X873, Q6KI66, Q3SSW8, Q02T82, C1AYS3, Q1GDV0, A5V604, A8AWA0, C1CF71, P13537, Q9P9Q9, A2Q0Z0, Q0WL56, P90519, Q57770, Q01765, C6A4R7, Q1R5Y2, Q0HNV1, Q6FZL2, Q134R0, Q1C1T4, B3ETZ7, B7HJ46, Q2L2G6, B8D851, A1VYI6, Q8KAH0, P42439, Q30X13, O50340, Q5FTY1, A6T3K6, Q5WZL4, Q605B0, Q8EX18, B2J5B1, Q25820, Q1IFW8, Q6N4Q4, A8M531, A6TZ25, P33170, Q30TQ5, Q4MYA4, Q5NQ65, P34823, P02993, A6UQ14, P25698, A1W2Q5, A5WGK9, Q5GSU2, Q1D776, Q826Z7, A0LRL8, A0JZ88, A9VP75, A5ELM9, A3NEI1, Q3APH1, A0PXT1, B7JUP5, Q5FFE6, Q5NID9, B0TC54, Q74JU6, Q8Y422, Q2SSW8, P48864, Q6MDN0, Q318N5, P85834, P48865, P33169, Q4L3K9, P64030, Q3AMT6, A1WHC3, P13549, Q40034, P41745, P34825, Q01372, P17196, Q8R7V2, Q2RQV8, A1JIH3, Q3A6P9, Q9KUZ6, A6VKH7, A1KB29, Q6MJ00, Q2YM08, Q13TF5, Q0SQC8, Q47JA5, Q4G342, Q5L3Z9, B2UUW8, Q88VE0, Q2W2H3, Q600B6, Q2GD83, Q3B6G3, A2BT83, Q1MIE3, B0BUR2, Q32B27, Q06SH3, A4VZZ3, A0T100, Q877T5, Q8W4H7, O27132, O42820, Q3ILP4, Q8DD27, A1WVD6, Q0SY20, Q6FF97, A1R8U9, P33166, Q89J82, A1V8A5, Q5L5H6, B1KSM7, Q85FT7, P0A6N2, B2SFC9, C4K4F8, B2GBC2, C1KZK6, P0A559, A1T4L6, A6LE88, A2CC87, Q1XDK1, A8F2E9, B0TM14, B9DKV8, B1ICR4, A5GW14, B1AJG3, P0CT53, Q9HM89, A5ULM5, P26751, A5D5K0, Q5SHN6, P0CE48, Q8EK70, B0CCD0, P17745, Q65PA9, B5RPI0, A3MRT8, Q11HA6, A5I7K8, Q46WC7, A5EX84, Q14JU2, A5UHC1, B3WE38, B1MY04, B3PMU1, P9WNN0, B1ZPC5, A3PEZ7, Q4FQG6, Q92GW4, A1S204, Q6GBT9, Q1JHV6, P18668, Q118Z2, P62630, P28295, B0R8C3, B8GIQ3, A1RRJ3, Q3A6R2, Q0AUH8, B1IVA7, A3Q980, Q5FCW3, O50293, Q6HPR0, Q7TT91, Q39KI2, Q1ACI3, A6LPP6, Q0K5Z9, Q54HB2, A8LC58, A5U9R1, Q039K9, A6YG72, A0QL35, A5U071,

Q6YQV8, Q4ZMP2, A8EZL8, Q9Y700, Q6GJC0, Q1JMR3, B1XI63, O83217, P68105, P50257, P16018, Q6LXI1, O59153, P02994, P57939, P29542, Q0TA85, Q089Q6, Q43364, O66429, Q9Z9L6, Q79GC6, Q0BJ48, Q8M9W7, A7FZ71, A7MKI5, A1VAK4, Q2JFH8, Q7TTF9, Q03QN5, Q9XD38, Q8BFR5, Q20EU5, Q6A6L7, Q889X3, Q1RHL9, P0A1H5, P99152, Q1JCT6, Q2JMX7, Q73PN3, P34824, Q5VTE0, Q92005, Q8TYP6, Q979T1, Q4QMW6, Q0I0B9, Q8R7T8, Q2RQU6, Q8ZAN8, Q2II78, B7HQU2, Q0SN31, P59506, Q5HVZ7, P0CD71, Q4JT41, C0Q9Y7, A7HM54, Q28UW7, Q6ACZ0, B3E156, P23568, Q8YP63, P72231, Q4K519, Q211E6, A9MHG0, Q2FJ92, P69952, A8Z5T8, B1LBP2, Q27140, Q00251, A2BN41, Q3IUD8, P25166, Q6FZC0, Q134S7, Q8PC59, P57966, P60339, B0BQZ3, B6YQ04, Q1LSY4, P42471, A4JAM5, B3EH93, A9KRZ4, Q11Q98, Q33451, C1A6Q3, Q9ZK19, Q02WY9, B9E8Q0, A4T1R2, A1KRF9, Q03F25, Q7UZY7, A6MW28, P0A3A9, Q31VV0, P42479, P64031, C5BQ44, A7MXE4, P68103, P62631, P29520, Q41803, Q9YIC0, Q976B1, B0SUQ7, A5USJ1, Q1C2U1, Q7MYE8, Q7MGR1, A0KQ95, C3P9Q3, B8DTV7, P64024, A4XI37, B3EP63, A3DJ00, A5FQQ5, Q839G8, O50306, Q2EEV7, B2G6R2, Q65QG6, Q4A9G1, Q81ZS3, A1ALS6, A4SCQ7, Q925Y6, P0A3B0, A6U842, Q8E645, A4VTQ7, Q01698, P42482, P29521, Q71V39, A0RUM4, Q8TRC4, A5DPE3, P17197, A8A5E6, A3DA74, A4TGY7, Q88QN7, Q4URC5, B2UQY9, Q81VT2, Q8G5B7, P64025, A0RQJ3, P17746, Q47UU9, Q9R342, C4ZB99, A4IJI7, Q0I1U9, Q38WR7, Q0ANN1, P30768, Q3J8Q0, A9BHA7, Q15NP2, C0ZVT7, Q1AU14, A9ETD1, P95724, Q03LX0, A8F4Q9, Q2NZX1, Q2HJN4, Q8GTY0, P31018, A6VGV6, A4WKK8, O24534, Q1D7V1, Q82DQ0, A7ZUJ2, A8GKK1, P40175, Q3MDM5, Q63H92, B5RM34, Q1BRT3, Q99QM0, A5CUB6, A9KD33, A4J0Z5, P26184, Q6B8Y0, Q1IHG6, B0SSH9, C5CC66, A3PV96, P49462, B2RL52, Q3K5X4, C3PPA9, Q2S1P8, Q5HIC7, Q5XD49, P33171, Q3SLQ1, P17506, P40911, Q2NEL1, Q6LVC0, Q9KV37, Q4QMT5, Q0I0A7, A1TJ05, Q5P334, Q5WLR4, P49410, Q62GK3, A9WFP3, A7GJ76, Q9TLV8, P0A6N3, Q2A1M0, Q2S8Z8, Q04B37, Q927I6, A5IYA9, A0PM42, A8G708, A1T056, Q8KT99, Q12SW1, P64029, A2RFQ4, Q7U4D1, Q83GW1, P10126, Q2HJN6, A4FWE9, Q9V0V7, P86933, Q0C1F4, Q3YWT3, B0T2B5, A5UYI1, A4TS36, Q5PBH1, B9IZJ2, P50062, Q1R0H7, B1VET1, B8G1W4, A5FIJ9, A0M3Z6, A6TEX7, Q04PT6, B8ELG5, A0QS98, Q8ETY4, Q123F6, A4XZ92, Q3J5S4, Q5PIW4, A5IQA2, P0DA82, Q67JU1, Q31IY4, P0CT32, P41752, A6TWI4, Q2K9N2, Q4URD7, Q0AF46, P29543, A5FZW7, Q2NJ20, Q8KHX9, A4YSJ0, Q63PZ6, Q822I4, Q0TMN0, P17245, Q5HAS0, A4IW92, B6JN44, Q9CEI0, A0ALY8, P18906, P64026, Q7VA05, O21245, C4K2I2, A0KRL0, Q49V58, P0DA83, P74227, A5CW32, Q32PH8, P19039, A0B7D6, P17786, Q88QP8, Q7MH43, Q0BYB2, Q3YV04, B7H1K5, B8HD11, Q8A463, P52854, A3P0B5, A2CI56, B8I5N8, B8HVR7, Q3YRK7, A0Q874, Q17VM8, Q042T5, Q71WB9, A1KGG5, B9L7I8, A1B002, A9BCK0, Q1LI13, Q8KTA1, A8GYW2, Q5HRK4, B8ZL95, Q0ID59, P50068, P0CY35, P14864, P53013, O49169, O64937, Q04634, Q6AP86, A1AVJ8, Q8ZJB2, Q6LLV5, Q73H85, A4SHU2, C3LJ80, A1A0T1, A5VR08, A7ZCN0, A1BJ36, P50372, Q3ZXX3, Q2N9A8, Q748X8, A4G9U0, Q1WU83, P22679, Q1QN32, Q4FLK5, Q9TJQ8, P75022, Q8KT95, Q2NQL7, Q8E0H1, Q5M101, Q8DI42 and Q8NL22.

Appendix B

Source code

B.1 Main function

This section describes a source code to compute $f_1({}_i\mathcal{M})$, $f_2({}_i\mathcal{M})$ and their correlations. Because describing all the source codes needs a very large space, only the main function written in C++ was shown below.

Listing B.1: A main function of the source code

```
1  int main()
2  {
3  /*
4   1. To create a protein family database from the UniprotKB/SwissProt database
5  - UniprotXML is an input file of UniProt data in XML format.
6  - FamilyDB is an output database file.
7  */
8    const char* UniprotXML = "~/uniprot/uniprot_sprot_2015_01.xml";
9    const char* FamilyDB = "ProteinFamily.db";
10   if(! CreateProteinFamilyDB(UniprotXML, FamilyDB)) return EXIT_FAILURE;
11
12 /*
13  2. To insert PDB data into the family database
14 - FamilyID is an identifier for protein family (20: EF-Tu/EF-1A family).
15 - PDBDirPath is a file path of the PDB data in XML format.
16 */
17   const unsigned int FamilyID = 20;
18   const char* PDBDirPath = "~/pdb";
19   if(! InsertPDBXML(FamilyDB, FamilyID, PDBDirPath)) return EXIT_FAILURE;
20
21 /*
22  3. To write a fasta file which contains all analysing sequences
```

```
23 - If the sequence length is smaller than MinSeqLen or larger than MaxSeqLen, the
       sequence is excluded.
24 - SeqFile is an output file which contains sequences derived from the SwissProt and PDB.
25 */
26   const unsigned int MinSeqLen = 20;
27   const unsigned int MaxSeqLen = 600;
28   const char* SeqFile = "Sequence.fasta";
29   if(! WriteSeqFile(FamilyDB, FamilyID, MinSeqLen, MaxSeqLen, SeqFile)) return
         EXIT_FAILURE;
30
31 /*
32  4. To execute the MAFFT program
33 - MafftOption is parameters of the MAFFT program.
34 - MafftFile is an output file of the MAFFT program.
35 */
36   const char* MafftOption = "--localpair --maxiterate 1000";
37   const char* MafftFile = "Mafft.fasta";
38   if(! ExecuteMAFFT(SeqFile, MafftOption, MafftFile)) return EXIT_FAILURE;
39
40 /*
41  5. To select alignment sites which have structural data
42 - SelectedSitesFile is an output file in fasta format.
43 */
44   const char* SelectedSitesFile = "SelectedSites.fasta";
45   if(! SelectAlignmentSites(MafftFile, SelectedSitesFile)) return EXIT_FAILURE;
46
47 /*
48  6. To insert infomation of a phylogenetic tree into the conservation database
49 - SubstitutionModel is a substitution model of amino acids (Dayhoff or JTT model).
50 - EquilibriumFrequencyModel is how to calculate equilibrium frequencies of amino acids.
       If it is OWN, Dayhoff or JTT, the frequencies are calculated from the input
       alignment or based on the Dayhoff or JTT model, respectively.
51 - When EquilibriumFrequencyModel is OWN, sequence weights are considered if
       SequencePairWeight is ture.
52 - TimeConstant is a time constant for computing sequence dissimilarities by the maximum
       likeliwood method.
53 - ConservationDB is an output database file.
54 */
55   const char* SubstitutionModel = "JTT"; //Dayhoff or JTT
56   const char* EquilibriumFrequencyModel = "OWN"; //OWN, Dayhoff, or JTT
57   const bool SequencePairWeight = true;
58   const double TimeConstant = 0.01;
59   const char* ConservationDB = "Conservation.db";
60   if(! InsertPhylogeneticTreeIntoConservationDB(SelectedSitesFile, SubstitutionModel,
```

```
        EquilibriumFrequencyModel, SequencePairWeight, TimeConstant, ConservationDB))
        return EXIT_FAILURE;
61
62  /*
63   7. To insert conservations into the conservation database
64  - Iteration is the number of iterations for computing sequence weights.
65  */
66    const unsigned int Iteration = 100000;
67    if(! InsertConservationsIntoConservationDB(ConservationDB, Iteration)) return
          EXIT_FAILURE;
68
69  /*
70   8. To insert correlations into the correlation database
71  - Cutoff is a cutoff [angstrom] which determines whether an amino acid residue is
        proximate from ions or molecules or not.
72  - If HomoInteraction is true, proximities from same proteins are taken account.
73  - ExcludingMolecule contains the molecules which are excluded when computating
        proximities.
74  - CorrelationDB is an output database file.
75  */
76    const double Cutoff = 3.0;
77    const bool HomoInteraction = false;
78    const char* ExcludingMolecule = "ACETATE ION,AMMONIUM ION,CHLORIDE ION,SULFATE ION,
          SODIUM ION,5-BROMOFURAN-2-CARBOXYLIC ACID,BETA-MERCAPTOETHANOL,DI(HYDROXYETHYL)
          ETHER,GLYOXYLIC ACID,SUGAR (SUCROSE)";
79    const char* CorrelationDB = "Correlation.db";
80    if(! InsertCorrelationstIntoCorrelationDB(FamilyDB, MafftFile, ConservationDB, Cutoff,
          HomoInteraction, ExcludingMolecule, CorrelationDB)) return EXIT_FAILURE;
81  }
```

B.2 Output

This section describes three output databases, which contain calculation results of above program. The protein family database contains data extracted from the Swiss-Prot and PDB (see Figure B.1). The conservation database contains data of an MSA, phylogenetic tree and some conservations based on h_x (see Figure B.2). The correlation database contains $f_1\,({}_i\mathcal{M})$, $f_2\,({}_i\mathcal{M})$ and their correlations (see Figure B.3).

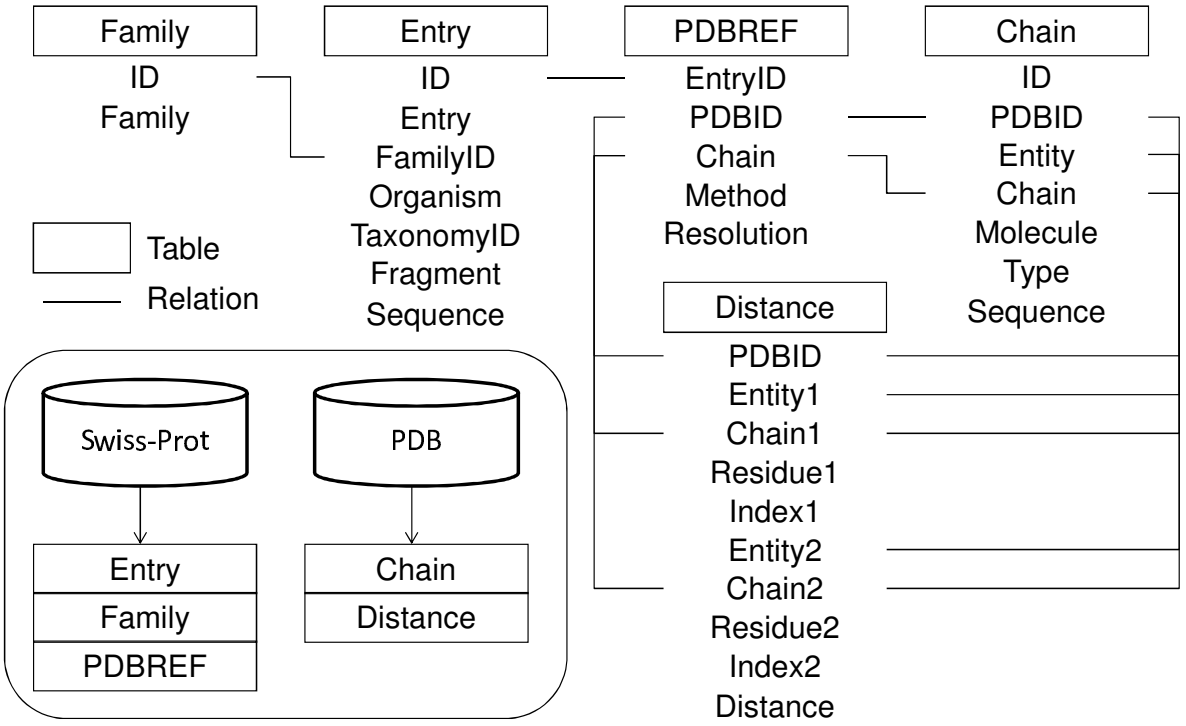


Figure B.1: **A schema of the protein family database.** From the Swiss-Prot database, tables of Entry, Family and PDBREF are created. From the PDB, tables of Chain and Distance are created.

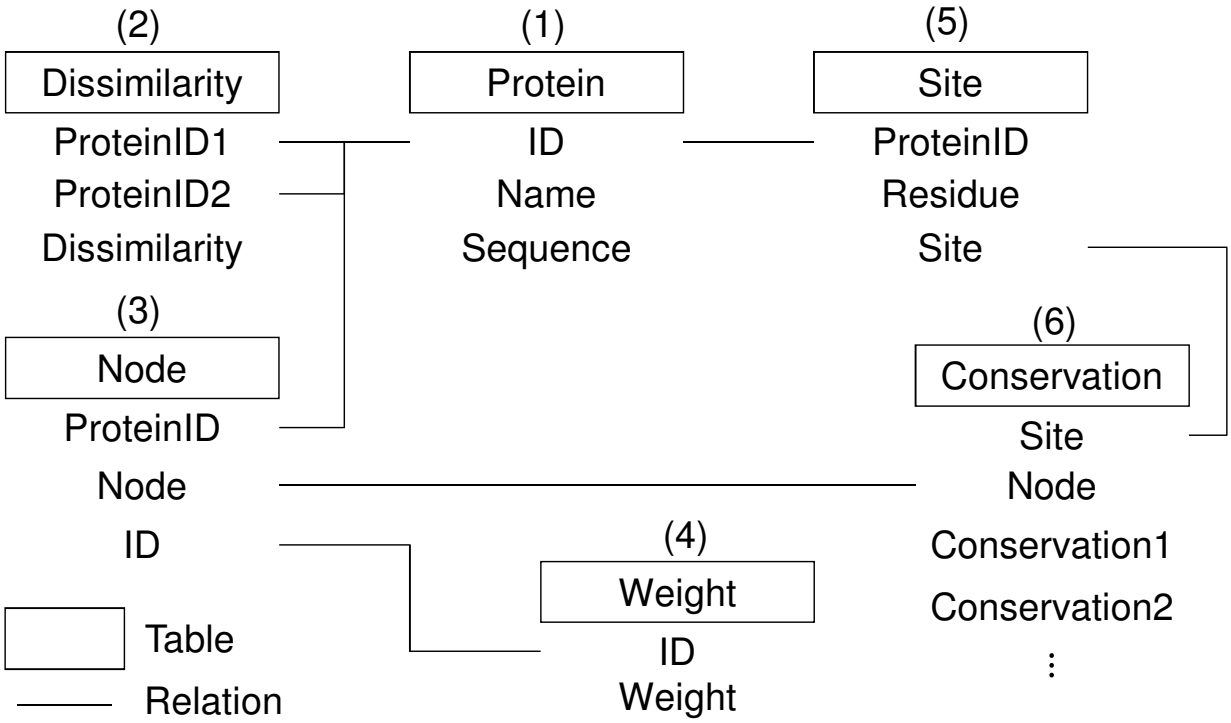


Figure B.2: **A schema of the conservation database.** From the MSA, tables of Protein, Dissimilarity, Node, Weight, Site and Conservation are created.

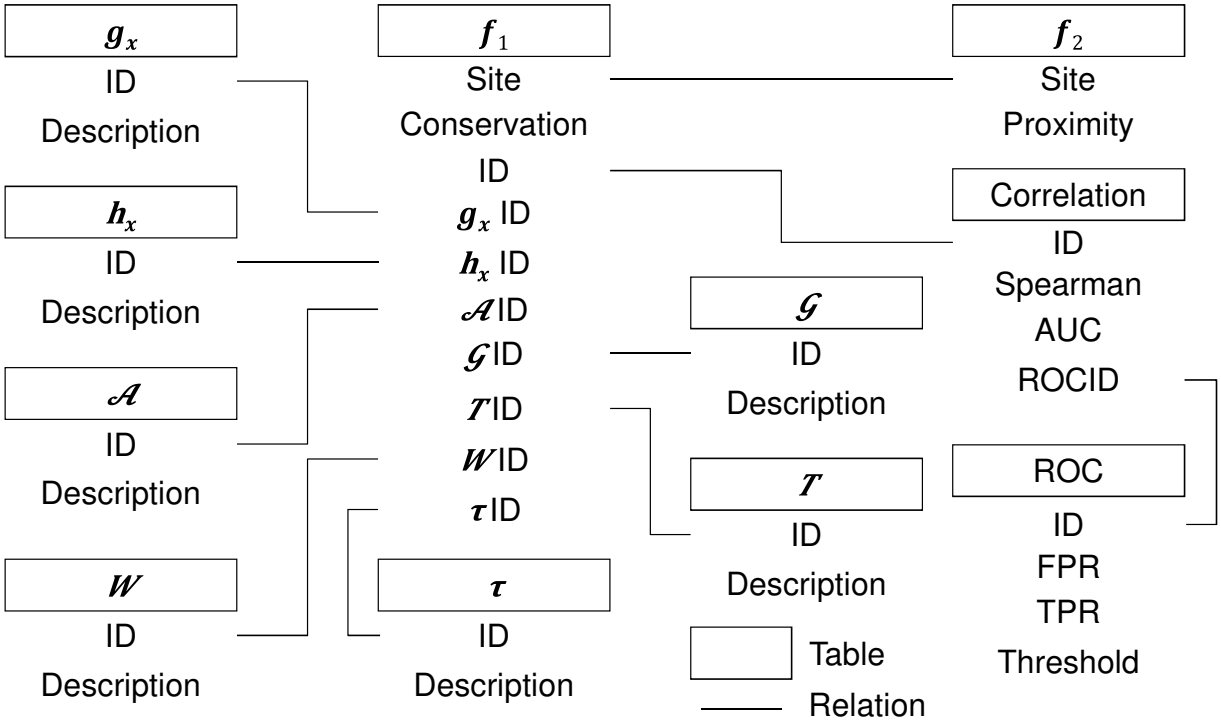


Figure B.3: **A schema of the correlation database.** From the conservation database, $f_1\,({}_i\mathcal{M})$ is calculated by changing the seven parameters. Meanwhile, $f_2\,({}_i\mathcal{M})$ is calculated from the protein family and conservation databases. Finally, correlations between $f_1\,({}_i\mathcal{M})$ and $f_2\,({}_i\mathcal{M})$ are calculated based on the Spearman's ρ and ROC curve.

Appendix C

Slides

This section shows slides used for explaining this thesis.

A new bioinformatics method for predicting active sites of multifunctional proteins

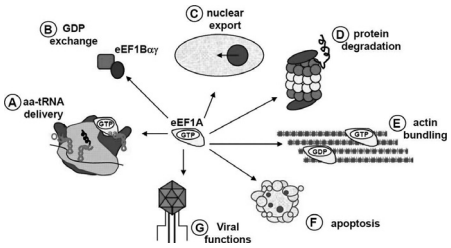
Yosuke Kondo

Department of Pharmaceutical Sciences,
Graduate School of Pharmaceutical Sciences, Tokyo University of Science

February 18, 2016

Introduction

Eukaryotic translation elongation factor 1A (eEF1A) has many functions



Maria K. Mateyak, and Terri Goss Kinzy J. Biol. Chem. 2010;285:21209-21213

Introduction

Problems and objectives

- Translation elongation factor 1A (EF1A) is a multifunctional protein
- However, it is still unknown how each function works and the functions relate to each other
- In order to investigate the unknown functions, identification of the functional sites is very helpful
- However, the EF1A consists of 462 amino acid residues
- In order to narrow down candidate functional sites, we used computational methods

Introduction

Sequence-based computational methods

- Shannon entropy (Sander et al, 1991)
- Shannon entropy of residue properties (Williamsom, 1995)
- Evolutionary trace (Lichtarge et al, 1996)
- ConSurf (Armon et al, 1998)
- Sum of pairs (Valdar and Thornton, 2001)
- Variance (Pei and Grishin, 2001)
- Rate4Site (Mayrose et al, 2004)
- Relative entropy (Wang and Samudrala, 2006)
- Jensen Shannon divergence (Capra and Singh, 2007)
- ⋮

Introduction

Problems and objectives

- There exist many sequence-based computational methods
- For elongation factor 1A, which method is the most appropriate?
- Meanwhile, while reading some papers, we noticed that a new method can be developed
- Therefore, we propose an original method and compare the method with existing methods

Amino acids

Representation

A: alanine	M: methionine
C: cysteine	N: asparagine
D: aspartic acid	P: proline
E: glutamic acid	Q: glutamine
F: phenylalanine	R: arginine
G: glycine	S: serine
H: histidine	T: threonine
I: isoleucine	V: valine
K: lysine	W: tryptophan
L: leucine	Y: tyrosine

Multiple sequence alignment

M R R R R R R R R R R R
M R R R R R R L R R L L L
M R R R R R L R R L L L L
M R R R R L R R L R L L S
M R L S – L R R R R R R L
M R L S – L R R L L R L L

Notation

- $\mathcal{M}$, multiple sequence alignment
- ${}_i\mathcal{M} \in \mathcal{M}$, site i of $\mathcal{M}$
- $f_x : \mathcal{M} \to [\,0, \infty\,)$, a mapping from $\mathcal{M}$ to $[\,0, \infty\,)$

Existing conservation analysis

Shannon entropy based methods

$$f_{se}\left({}_i\mathcal{M}\right) = -\sum_{x\in\mathcal{A}} p_i\left(x\right)\ln p_i\left(x\right) \tag{1}$$

Two methods

- Shannon entropy (SE)
  - $\mathcal{A} =$ $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, gap\}$
- Shannon entropy of residue properties (SEP)
  - $\mathcal{A} = \{MLVI, HRK, ST, AG, DE, QN, FWY, P, C, gap\}$

Existing conservation analysis

Sum of pairs based methods

$$f_{sp}\left({}_i\mathcal{M}\right) = \frac{1}{N}\sum_{l=1}^{N}\sum_{m=1}^{N} w\left(l\right) D\left(x_l, x_m\right) \tag{2}$$

$$D\left(x,y\right) = \frac{s\left(x,x\right) - s\left(x,y\right)}{s\left(x,x\right)} \tag{3}$$

Two methods

- Sum of pairs (SP): all sequence weights were 1
- Sum of pairs with sequence weights (SPW)
  - Sequence weights were calculated by the Sibbald and Argos algorithm (PR, Sibbald and P, Argos, 1990)

Existing conservation analysis

Integer-valued evolutionary trace (iv-ET) method

$$f_{iet}\left({}_i\mathcal{M}\right) = 1 + \sum_{l=1}^{N-1}\begin{cases} 0 & \text{(if site } i \text{ conserved within each group } g\text{)} \\ 1 & \text{(otherwise)} \end{cases} \tag{4}$$

Real-valued evolutionary trace (rv-ET) method

$$f_{ret}\left({}_i\mathcal{M}\right) = 1 + \sum_{l=1}^{N-1}\frac{1}{l}\sum_{g=1}^{l}\left[-\sum_{x\in\mathcal{A}} p_{ig}\left(x\right)\ln p_{ig}\left(x\right)\right] \tag{5}$$

Proposed method

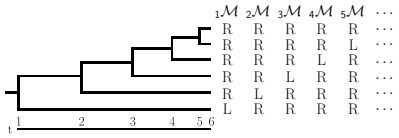
Notation

- $\mathcal{M}$, multiple sequence alignment
- ${}_i\mathcal{M} \in \mathcal{M}$, site i of $\mathcal{M}$
- $f_x : \mathcal{M} \to [\,0, \infty\,)$, a mapping from $\mathcal{M}$ to $[\,0, \infty\,)$

Integration of mappings

- $f_{se}, f_{sp}, f_{iet}, f_{ret} \longrightarrow f_1$

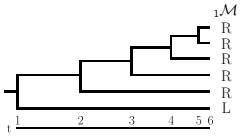
Representation of ${}_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

-
-
-
-
-

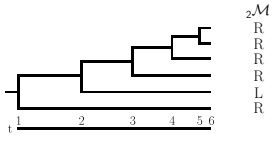
Representation of ${}_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

- ${}^1_1\mathcal{M} = \{\{R, R, R, R, R, L\}\}$
- ${}^2_1\mathcal{M} = \{\{R, R, R, R, R\}, \{L\}\}$
- ${}^3_1\mathcal{M} = \{\{R, R, R, R\}, \{R\}, \{L\}\}$
- ${}^4_1\mathcal{M} = \{\{R, R, R\}, \{R\}, \{R\}, \{L\}\}$
- ${}^5_1\mathcal{M} = \{\{R, R\}, \{R\}, \{R\}, \{R\}, \{L\}\}$

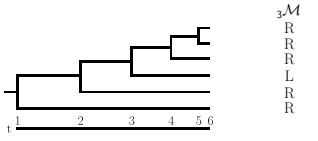
Representation of ${}_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

- ${}^1_2\mathcal{M} = \{\{R, R, R, R, L, R\}\}$
- ${}^2_2\mathcal{M} = \{\{R, R, R, R, L\}, \{R\}\}$
- ${}^3_2\mathcal{M} = \{\{R, R, R, R\}, \{L\}, \{R\}\}$
- ${}^4_2\mathcal{M} = \{\{R, R, R\}, \{R\}, \{L\}, \{R\}\}$
- ${}^5_2\mathcal{M} = \{\{R, R\}, \{R\}, \{R\}, \{L\}, \{R\}\}$

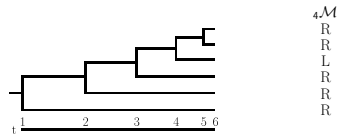
Representation of ${}_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

- ${}^1_3\mathcal{M} = \{\{R, R, R, L, R, R\}\}$
- ${}^2_3\mathcal{M} = \{\{R, R, R, L, R\}, \{R\}\}$
- ${}^3_3\mathcal{M} = \{\{R, R, R, L\}, \{R\}, \{R\}\}$
- ${}^4_3\mathcal{M} = \{\{R, R, R\}, \{L\}, \{R\}, \{R\}\}$
- ${}^5_3\mathcal{M} = \{\{R, R\}, \{R\}, \{L\}, \{R\}, \{R\}\}$

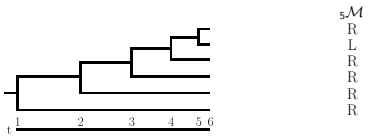
Representation of $_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

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

Representation of $_i\mathcal{M}$



${}^t_i\mathcal{M}$, a field of sets of characters

- ${}^1_5\mathcal{M} = \{\{R, L, R, R, R, R\}\}$
- ${}^2_5\mathcal{M} = \{\{R, L, R, R, R\}, \{R\}\}$
- ${}^3_5\mathcal{M} = \{\{R, L, R, R\}, \{R\}, \{R\}\}$
- ${}^4_5\mathcal{M} = \{\{R, L, R\}, \{R\}, \{R\}, \{R\}\}$
- ${}^5_5\mathcal{M} = \{\{R, L\}, \{R\}, \{R\}, \{R\}, \{R\}\}$

Definition of f_1

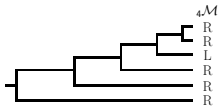
Notation

- T, a time point
- $g_x : {}_i\mathcal{M} \rightarrow [0, 1]$, a mapping from ${}_i\mathcal{M}$ to $[0, 1]$
- ${}^t_i\mathcal{M}$, a field of sets of characters

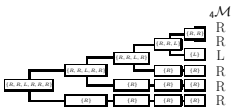
Definition

$$f_1\left({}_i\mathcal{M}\right) := \sum_{t=1}^{T} g_x\left(h_x\left({}^t_i\mathcal{M}, \mathcal{A}, {}_i\mathcal{G}\right)\right) \qquad (6)$$

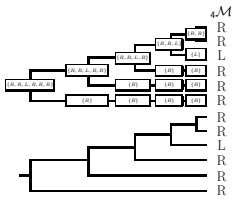
Explanation of g_x



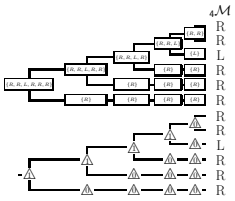
Explanation of g_x



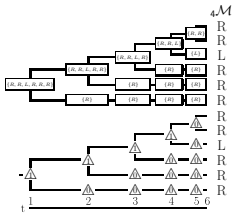
Explanation of g_x



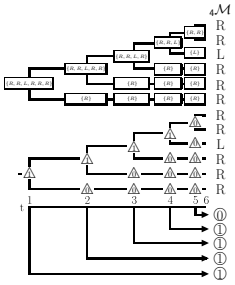
Explanation of g_x



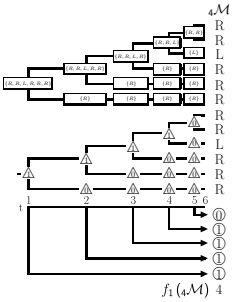
Explanation of g_x



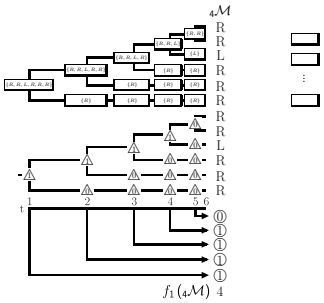
Explanation of g_x



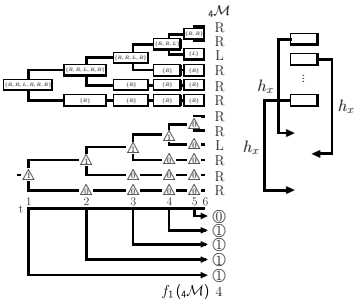
Explanation of g_x



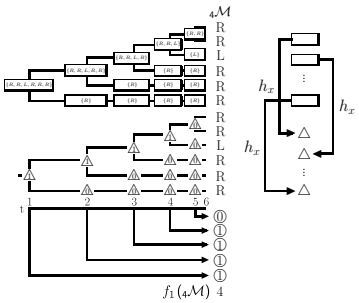
Explanation of g_x



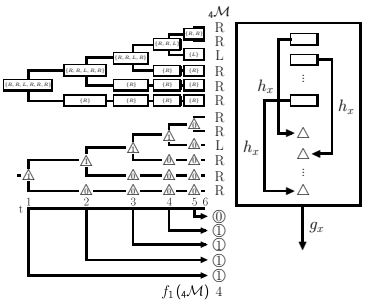
Explanation of g_x



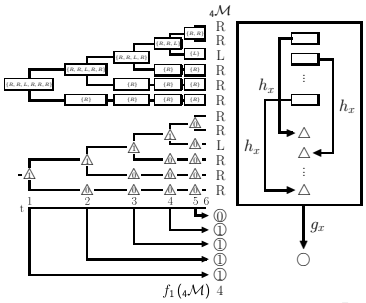
Explanation of g_x



Explanation of g_x



Explanation of g_x



Definition of g_1

Notation

- ${}^t_i\mathcal{M}$, a field of sets of characters
- $h_x : {}^t_i\mathcal{M} \to [0,1]$, a mapping from ${}^t_i\mathcal{M}$ to $[0,1]$
- τ, a threshold

Definition

$$g_1\left({}^t_i\mathcal{M}\right) := \begin{cases} 0 & \left(\forall {}^t_i\mathcal{M}_u \in {}^t_i\mathcal{M}, h_x\left({}^t_i\mathcal{M}_u\right) \le \tau\right) \\ 1 & \left(\exists {}^t_i\mathcal{M}_u \in {}^t_i\mathcal{M}, h_x\left({}^t_i\mathcal{M}_u\right) > \tau\right) \end{cases} \tag{7}$$

Definition of g_2

Notation

- ${}^t_i\mathcal{M}$, a field of sets of characters
- $|\cdot|$, a number of elements
- $h_x : {}^t_i\mathcal{M} \to [0,1]$, a mapping from ${}^t_i\mathcal{M}$ to $[0,1]$

Definition

$$g_2\left({}^t_i\mathcal{M}\right) := \frac{1}{\left|{}^t_i\mathcal{M}\right|} \sum_{{}^t_i\mathcal{M}_u \in {}^t_i\mathcal{M}} h_x\left({}^t_i\mathcal{M}_u\right) \tag{8}$$

Definition of g_3

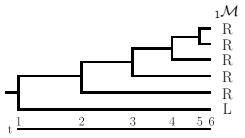
Notation

- ${}^t_i\mathcal{M}$, a field of sets of characters
- ${}^t_i\mathcal{M}_{*u}$, a multiset of characters
- $h_x : {}^t_i\mathcal{M} \to [0,1]$, a mapping from ${}^t_i\mathcal{M}$ to $[0,1]$

Definition

$$g_3\left({}^t_i\mathcal{M}\right) := h_x\left({}^t_i\mathcal{M}_{*u}\right) \tag{9}$$

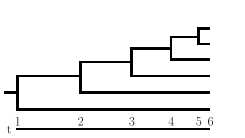
Representation of ${}^t_i\mathcal{M}_{*u}$



${}^t_i\mathcal{M}_{*u}$, a multiset of characters

- ${}^1_1\mathcal{M}_{*u} = \{R, R, R, R, R, L\}$
- ${}^2_1\mathcal{M}_{*u} = \{R, R, R, R, R\}$
- ${}^3_1\mathcal{M}_{*u} = \{R, R, R, R\}$
- ${}^4_1\mathcal{M}_{*u} = \{R, R, R\}$
- ${}^5_1\mathcal{M}_{*u} = \{R, R\}$

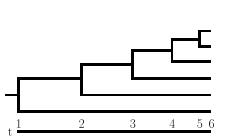
Representation of ${}^t_i\mathcal{M}_{*u}$



${}^t_i\mathcal{M}_{*u}$, a multiset of characters

- ${}^1_2\mathcal{M}_{*u} = \{R, R, R, R, L, R\}$
- ${}^2_2\mathcal{M}_{*u} = \{R, R, R, R, L\}$
- ${}^3_2\mathcal{M}_{*u} = \{R, R, R, R\}$
- ${}^4_2\mathcal{M}_{*u} = \{R, R, R\}$
- ${}^5_2\mathcal{M}_{*u} = \{R, R\}$

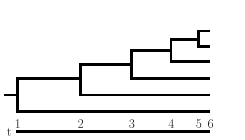
Representation of ${}^t_i\mathcal{M}_{*u}$



${}^t_i\mathcal{M}_{*u}$, a multiset of characters

- ${}^1_3\mathcal{M}_{*u} = \{R, R, R, L, R, R\}$
- ${}^2_3\mathcal{M}_{*u} = \{R, R, R, L, R\}$
- ${}^3_3\mathcal{M}_{*u} = \{R, R, R, L\}$
- ${}^4_3\mathcal{M}_{*u} = \{R, R, R\}$
- ${}^5_3\mathcal{M}_{*u} = \{R, R\}$

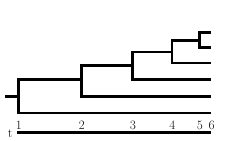
Representation of ${}^t_i\mathcal{M}_{*u}$



${}^t_i\mathcal{M}_{*u}$, a multiset of characters

- ${}^1_4\mathcal{M}_{*u} = \{R, R, L, R, R, R\}$
- ${}^2_4\mathcal{M}_{*u} = \{R, R, L, R, R\}$
- ${}^3_4\mathcal{M}_{*u} = \{R, R, L, R\}$
- ${}^4_4\mathcal{M}_{*u} = \{R, R, L\}$
- ${}^5_4\mathcal{M}_{*u} = \{R, R\}$

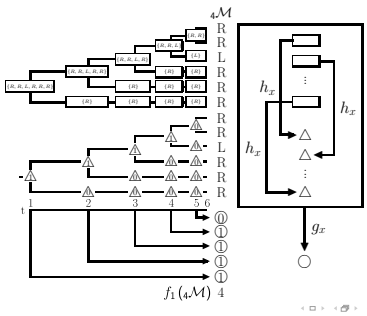
Representation of ${}^t_i\mathcal{M}_{*u}$



${}^t_i\mathcal{M}_{*u}$, a multiset of characters

- ${}^1_5\mathcal{M}_{*u} = \{R, L, R, R, R, R\}$
- ${}^2_5\mathcal{M}_{*u} = \{R, L, R, R, R\}$
- ${}^3_5\mathcal{M}_{*u} = \{R, L, R, R\}$
- ${}^4_5\mathcal{M}_{*u} = \{R, L, R\}$
- ${}^5_5\mathcal{M}_{*u} = \{R, L\}$

Explanation of h_x



Definition of $\mathcal{A}$

Notation

- $\mathcal{A}$, a field of sets of amino acids

Example

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

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

Definition of ${}_i\mathcal{G}$

Notation

- ${}_i\mathcal{G}$, a field of sets of gaps in site i

Example

- ${}^{1}_{i}\mathcal{G} = \left\{\left\{{}^{1}_{i}\gamma, {}^{2}_{i}\gamma, \ldots, {}^{G}_{i}\gamma\right\}\right\}$
- ${}^{G}_{i}\mathcal{G} = \left\{\left\{{}^{1}_{i}\gamma\right\}, \left\{{}^{2}_{i}\gamma\right\}, \ldots, \left\{{}^{G}_{i}\gamma\right\}\right\}$

Definition of h_1

Notation

- ${}^{t}_{i}\mathcal{M}_u$, a multiset of characters
- $\mathcal{A}$, a set of amino acid symbols
- ${}_i\mathcal{G}$, a set of gaps in site i

Definition

$$h_1\left({}^{t}_{i}\mathcal{M}_u\right) := \begin{cases} 0 & (\forall l \in {}^{t}_{i}\mathcal{M}_u, \exists \mathcal{X} \in \mathcal{A} \cup {}_i\mathcal{G}; l \in \mathcal{X}) \\ 1 & (otherwise) \end{cases} \quad (10)$$

Definition of h_2

Notation

- ${}^{t}_{i}\mathcal{M}_u$, a multiset of characters
- $|\cdot|$, a number of elements

Definition

$$h_2\left({}^{t}_{i}\mathcal{M}_u\right) := -\sum_{\mathcal{X} \in \mathcal{A} \cup {}_i\mathcal{G}} p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) \log_{|\mathcal{A} \cup {}_i\mathcal{G}|} p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) \quad (11)$$

$$p\left({}^{t}_{i}\mathcal{M}_u, \mathcal{X}\right) = \frac{1}{|\,{}^{t}_{i}\mathcal{M}_u\,|} \sum_{l \in {}^{t}_{i}\mathcal{M}_u} \begin{cases} 0 & (l \notin \mathcal{X}) \\ 1 & (l \in \mathcal{X}) \end{cases} \quad (12)$$

Definition of h_3

Notation

- ${}^{t}_{i}\mathcal{M}_u$, a multiset of characters
- $|\cdot|$, a number of elements
- $w(l)$, a weight of sequence l
- $s(l,m)$, a substitution score of amino acid l and m

Definition

$$h_3\left({}^{t}_{i}\mathcal{M}_u\right) := \frac{1}{|\,{}^{t}_{i}\mathcal{M}_u\,|} \sum_{l \in {}^{t}_{i}\mathcal{M}_u} \sum_{m \in {}^{t}_{i}\mathcal{M}_u} w(l)\, s(l,m) \quad (13)$$

Definition of $s(l.m)$

Notation

- S_{max}, S_{min}, the maximum or the minimum
- $S(l,l), S(l,m)$, a diagonal or an off-diagonal element

Definition

$$s(l,m) = \begin{cases} 0 & (l \in \mathcal{X} \in {}_i\mathcal{G} \wedge m \in \mathcal{Y} \in {}_i\mathcal{G} \wedge \mathcal{X} = \mathcal{Y}) \\ \frac{S_{max} - S_{min}}{S_{max}} & (l \in \mathcal{X} \in {}_i\mathcal{G} \wedge m \in \mathcal{Y} \in {}_i\mathcal{G} \wedge \mathcal{X} \neq \mathcal{Y}) \\ \frac{S_{max} - S_{min}}{S_{max}} & (l \in \mathcal{X} \in {}_i\mathcal{G} \wedge m \in \mathcal{Y} \in \mathcal{A}) \\ \frac{S(l,l) - S_{min}}{S(l,l)} & (l \in \mathcal{X} \in \mathcal{A} \wedge m \in \mathcal{Y} \in {}_i\mathcal{G}) \\ \frac{S(l,l) - S(l,m)}{S(l,l)} & (l \in \mathcal{X} \in \mathcal{A} \wedge m \in \mathcal{Y} \in \mathcal{A}) \end{cases} \quad (14)$$

Parameters of f_1

Table 1: Seven parameters for calculation of f_1

Parameter	Value	Description
T	$1, 2, \ldots, N$	a time point
g_x	g_1, g_2, g_3	$g_x : {}_i\mathcal{M} \to [0,1]$
τ	$[0,1]$	a threshold of g_1
h_x	h_1, h_2, h_3	$h_x : {}^{t}_{i}\mathcal{M} \to [0,1]$
$\mathcal{A}$	$^{20}\mathcal{A}, {}^{9}\mathcal{A}$	a field of sets of amino acids
${}_i\mathcal{G}$	${}^{1}_{i}\mathcal{G}, {}^{G}_{i}\mathcal{G}$	a field of sets of gaps in site i
$w(l)$	$(0, \infty)$	a weight of sequence l

Existing methods and a proposed method

Table 2: Relationships of the parameters and prediction methods

Method	T	g_x	τ	h_x	$\mathcal{A}$	${}_i\mathcal{G}$	$w(l)$
SE	1	g_3	–	h_2	$^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–
SEP	1	g_3	–	h_2	${}^{9}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–
SP	1	g_3	–	h_3	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	unused
SPW	1	g_3	–	h_3	$\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	used
iv-ET	N	g_1	$\tau \in (0,1)$	h_1	$^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–
rv-ET	N	g_2	–	h_2	$^{20}\mathcal{A}$	${}^{1}_{i}\mathcal{G}$	–
Proposed	N	g_3	–	h_1	$^{20}\mathcal{A}$	${}^{G}_{i}\mathcal{G}$	–

Definition of a specific conservation

Notation

- ${}_i^a\mathcal{M}$, a field of sets of characters at time point a
- ${}_i^b\mathcal{M}$, a field of sets of characters at time point b

Definition

$$C\left({}_i^a\mathcal{M}, {}_i^b\mathcal{M}\right) = \frac{g_3\left({}_i^a\mathcal{M}\right)+1}{g_3\left({}_i^b\mathcal{M}\right)+1} - 1 \qquad (15)$$

where $a < b$

Proposed method

Notation

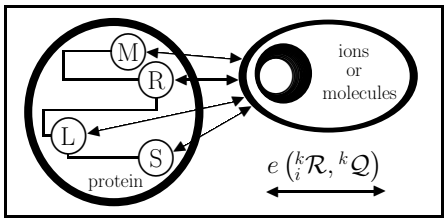
- $\mathcal{M}$, multiple sequence alignment
- ${}_i\mathcal{M} \in \mathcal{M}$, site i of $\mathcal{M}$
- $f_x : \mathcal{M} \to [\,0, \infty\,)$, a mapping from $\mathcal{M}$ to $[\,0, \infty\,)$

Two data types of ${}_i\mathcal{M}$

1. Character type $\longrightarrow f_1$
2. Coordinate type $\longrightarrow f_2$

Numerical evaluation of prediction methods

structure k



Numerical evaluation of prediction methods

Notation

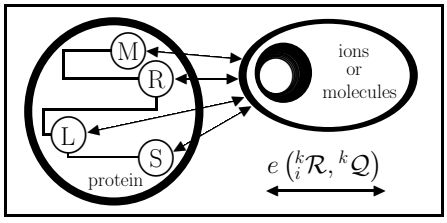
- $\mathbb{R}$, a set of real numbers
- $e : \mathbb{R}^3 \times \mathbb{R}^3 \to [\,0, \infty\,)$, a mapping from $\mathbb{R}^3 \times \mathbb{R}^3$ to $[\,0, \infty\,)$
- $\mathcal{R} \subset \mathbb{R}^3$, a subset of $\mathbb{R}^3$
- $\mathcal{Q} \subset \mathbb{R}^3$, a subset of $\mathbb{R}^3$
- $\| \cdot \|_2$, an Euclidean norm

Definition

$$e\left(\mathcal{R}, \mathcal{Q}\right) := \min_{(\boldsymbol{r},\boldsymbol{q})\in\mathcal{R}\times\mathcal{Q}} \left(\| \boldsymbol{r} - \boldsymbol{q} \|_2\right) \qquad (16)$$

Numerical evaluation of prediction methods

structure k



Definition of f_2

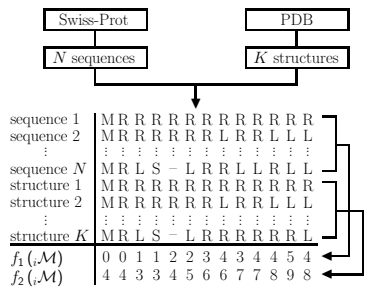
Notation

- K, a number of structures
- $\left\{{}_i^1\mathcal{R}, {}_i^2\mathcal{R}, \ldots, {}_i^{K-G}\mathcal{R}\right\} \subseteq {}_i\mathcal{M}$, a set of residues in ${}_i\mathcal{M}$
- $\left\{{}_i^1\gamma, {}_i^2\gamma, \ldots, {}_i^G\gamma\right\} = {}_i\mathcal{G} \subset {}_i\mathcal{M}$, a set of gaps in ${}_i\mathcal{M}$
- ${}_i^k\mathcal{R} \subset \mathbb{R}^3$, atomic coordinates of residue i in structure k
- ${}^k\mathcal{Q} \subset \mathbb{R}^3$, atomic coordinates of ions or molecules in structure k

Definition

$$f_2\left({}_i\mathcal{M}\right) := \min_{{}_i^k\mathcal{R}\in{}_i\mathcal{M}\setminus{}_i\mathcal{G}} \left[e\left({}_i^k\mathcal{R}, {}^k\mathcal{Q}\right)\right] \qquad (17)$$

Computations of f_1 and f_2



Correlations of f_1 and f_2

Two methods

1. Spearman's ρ
2. ROC curve

Rank function

Notation

- $\mathcal{F}_x \ni f_x\,({}_i\mathcal{M})$, a multiset of $f_x\,({}_i\mathcal{M})$
- $\mathcal{F}_x \ni {}^1_iV_x \leq {}^2_iV_x \leq \cdots \leq {}^I_iV_x$, elements of $\mathcal{F}_x$
- t_n, a number of tied ranks

Definition

$$r\left({}^{j+k-1}_{\ \ i}V_x\right) = j - 1 + \frac{t_n + 1}{2} \tag{18}$$

Spearman's ρ

Notation

- T_1, T_2, correction terms
- I, a number of sites
- r, a rank function
- $\mathcal{F}_x \ni {}^1_iV_x \leq {}^2_iV_x \leq \cdots \leq {}^I_iV_x$, elements of $\mathcal{F}_x$

Definition

$$\rho = \frac{T_1 + T_2 - \sum_{i=1}^{I}\left[r\left({}^l_iV_1\right) - r\left({}^m_iV_2\right)\right]^2}{2\sqrt{T_1T_2}} \tag{19}$$

Correction terms of Spearman's ρ

Notation

- N_1, N_2, a size of tied ranks

Definition

$$T_1 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_1}\left(t_n^3 - t_n\right)}{12} \tag{20}$$

$$T_2 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_2}\left(t_n^3 - t_n\right)}{12} \tag{21}$$

Threshold of an ROC curve

Notation

- $[\,0, \infty\,) \supset \mathcal{F} \ni f_1\,({}_i\mathcal{M})$, a subset of non-negative real numbers and a set of $f_1\,({}_i\mathcal{M})$
- $\mathcal{F} \ni v_1 < v_2 < \cdots < v_J$, elements of $\mathcal{F}$
- J, a number of elements in $\mathcal{F}$

Definition

$$t_j \begin{cases} < v_1 & (j = 0) \\ = \dfrac{v_j + v_{j+1}}{2} & (j = 1, 2, \ldots, J-1) \\ > v_J & (j = J) \end{cases} \tag{22}$$

False positive rate and true positive rate

Notation

- $I_{fp}\,(t_j)\,, I_{tp}\,(t_j)$, a number of sites of false positives and ture positives when the threshold is t_j
- I_f, I_t, a number of false or true sites

Definition

$$p\,(t_j) = \frac{I_{fp}\,(t_j)}{I_f} \tag{23}$$

$$q\,(t_j) = \frac{I_{tp}\,(t_j)}{I_t} \tag{24}$$

Area under the curve (AUC)

Notation

- J, a number of elements in $\mathcal{F}$
- $p\,(t_j)\,, q\,(t_j)$, a false positive rate and a true positive rate when the threshold is t_j
- $p\,(t_{j+1})\,, q\,(t_{j+1})$, a false positive rate and a true positive rate when the threshold is t_{j+1}

Definition

$$AUC = \frac{1}{2}\sum_{j=0}^{J-1}\left[p\,(t_{j+1}) - p\,(t_j)\right] \cdot \left[q\,(t_{j+1}) + q\,(t_j)\right] \tag{25}$$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	L	R	R	R	$\cdots$
R	L	L	R	R	$\cdots$
R	L	L	L	R	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	L	R	R	R	$\cdots$
R	L	L	R	R	$\cdots$
R	L	L	L	R	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$

$f_x({}_i\mathcal{M})$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	L	R	R	R	$\cdots$
R	L	L	R	R	$\cdots$
R	L	L	L	R	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$

$f_\alpha({}_i\mathcal{M})$ 0 0 1 1 1 $\cdots$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	L	R	R	R	$\cdots$
R	L	L	R	R	$\cdots$
R	L	L	L	R	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$
R	L	L	L	L	$\cdots$

$f_\beta({}_i\mathcal{M})$ 6 6 5 4 3 $\cdots$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	R	R	R	R	$\cdots$
R	R	R	R	L	$\cdots$
R	R	R	L	R	$\cdots$
R	R	L	R	R	$\cdots$
R	L	R	R	R	$\cdots$
L	R	R	R	R	$\cdots$

$f_x({}_i\mathcal{M})$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	R	R	R	R	$\cdots$
R	R	R	R	L	$\cdots$
R	R	R	L	R	$\cdots$
R	R	L	R	R	$\cdots$
R	L	R	R	R	$\cdots$
L	R	R	R	R	$\cdots$

$f_\beta({}_i\mathcal{M})$ 5 5 5 5 5 $\cdots$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	R	R	R	R	$\cdots$
R	R	R	R	L	$\cdots$
R	R	R	L	R	$\cdots$
R	R	L	R	R	$\cdots$
R	L	R	R	R	$\cdots$
L	R	R	R	R	$\cdots$

$f_\alpha({}_i\mathcal{M})$ 1 1 1 1 1 $\cdots$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	R	R	R	R	$\cdots$
R	R	R	R	L	$\cdots$
R	R	R	L	R	$\cdots$
R	R	L	R	R	$\cdots$
R	L	R	R	R	$\cdots$
L	R	R	R	R	$\cdots$

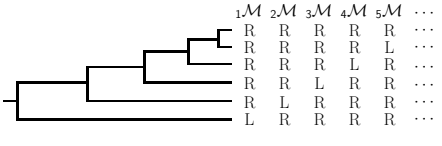
$f_x({}_i\mathcal{M})$

How do we define conservation?

${}_1\mathcal{M}$	${}_2\mathcal{M}$	${}_3\mathcal{M}$	${}_4\mathcal{M}$	${}_5\mathcal{M}$	$\cdots$
R	R	R	R	R	$\cdots$
R	R	R	R	L	$\cdots$
R	R	R	L	R	$\cdots$
R	R	L	R	R	$\cdots$
R	L	R	R	R	$\cdots$
L	R	R	R	R	$\cdots$

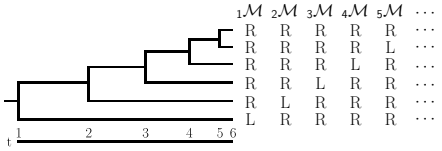
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



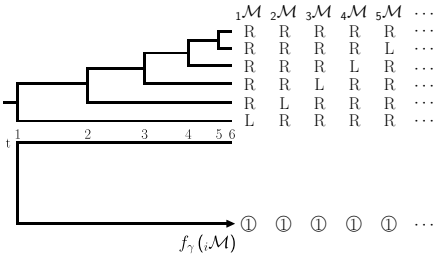
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



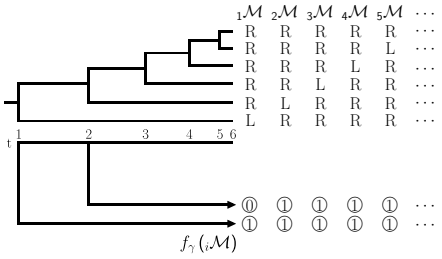
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



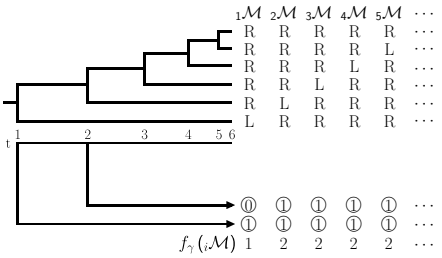
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



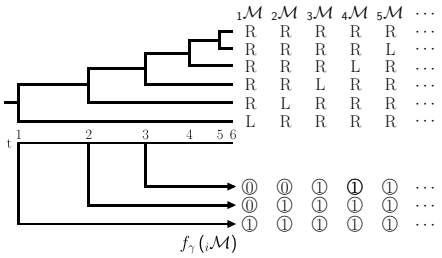
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



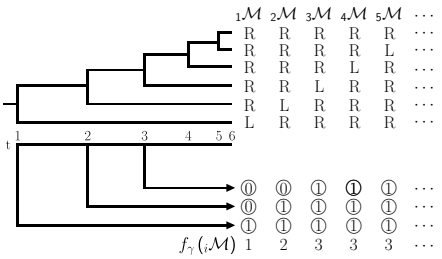
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



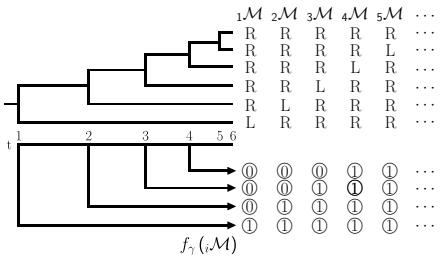
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?



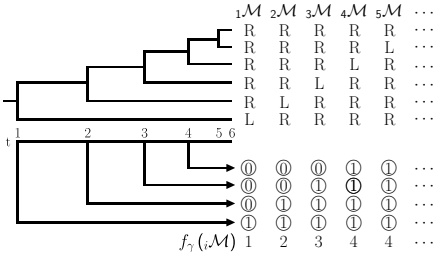
$f_\gamma({}_i\mathcal{M})$

How do we define conservation?

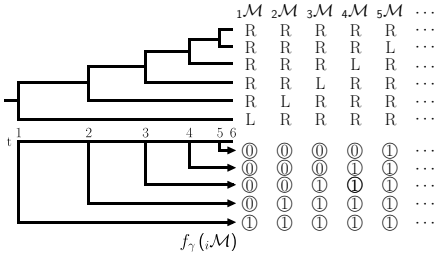


$f_\gamma({}_i\mathcal{M})$

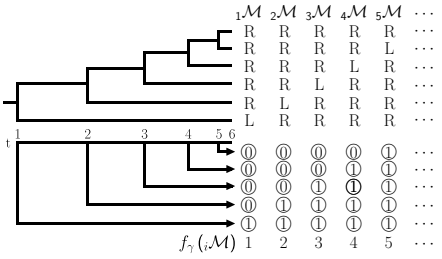
How do we define conservation?



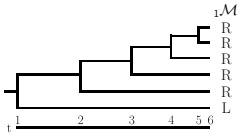
How do we define conservation?



How do we define conservation?



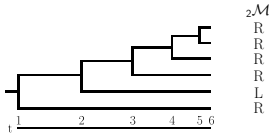
Representation of $_i\mathcal{M}$ in Chapter 3



$^t_i\mathcal{M}$, a multiset of characters

- $^1_1\mathcal{M} = \{R, R, R, R, R, L\}$
- $^2_1\mathcal{M} = \{R, R, R, R, R\}$
- $^3_1\mathcal{M} = \{R, R, R, R\}$
- $^4_1\mathcal{M} = \{R, R, R\}$
- $^5_1\mathcal{M} = \{R, R\}$

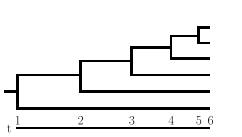
Representation of $_i\mathcal{M}$ in Chapter 3



$^t_i\mathcal{M}$, a multiset of characters

- $^1_2\mathcal{M} = \{R, R, R, R, L, R\}$
- $^2_2\mathcal{M} = \{R, R, R, R, L\}$
- $^3_2\mathcal{M} = \{R, R, R, R\}$
- $^4_2\mathcal{M} = \{R, R, R\}$
- $^5_2\mathcal{M} = \{R, R\}$

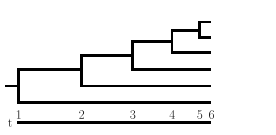
Representation of $_i\mathcal{M}$ in Chapter 3



$^t_i\mathcal{M}$, a multiset of characters

- $^1_3\mathcal{M} = \{R, R, R, L, R, R\}$
- $^2_3\mathcal{M} = \{R, R, R, L, R\}$
- $^3_3\mathcal{M} = \{R, R, R, L\}$
- $^4_3\mathcal{M} = \{R, R, R\}$
- $^5_3\mathcal{M} = \{R, R\}$

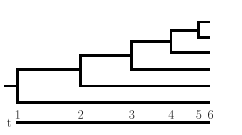
Representation of $_i\mathcal{M}$ in Chapter 3



$^t_i\mathcal{M}$, a multiset of characters

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

Representation of $_i\mathcal{M}$ in Chapter 3



$^t_i\mathcal{M}$, a multiset of characters

- $^1_5\mathcal{M} = \{R, L, R, R, R, R\}$
- $^2_5\mathcal{M} = \{R, L, R, R, R\}$
- $^3_5\mathcal{M} = \{R, L, R, R\}$
- $^4_5\mathcal{M} = \{R, L, R\}$
- $^5_5\mathcal{M} = \{R, L\}$

Definition of f_1

Notation

- ${}_i^t\mathcal{M}$, a multiset of characters
- N, a number of internal nodes
- $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} = \mathcal{A}$, a set of amino acid symbols
- $\left\{{}_i^1\gamma, {}_i^2\gamma, \ldots, {}_i^G\gamma\right\} = {}_i\mathcal{G}$, a set of gaps in site i

Definition

$$f_1\left({}_i\mathcal{M}\right) := \sum_{t=1}^{N} \begin{cases} 0 & \left(\forall l \in {}_i^t\mathcal{M}, \exists x \in \mathcal{A} \cup {}_i\mathcal{G}; l = x\right) \\ 1 & (otherwise) \end{cases} \tag{26}$$

Explanation of the proposed method

Features of the proposed method

- The proposed method requires that the phylogenetic tree is reconstructed by under the hypothesis that the evolutionary rate is constant
- Thanks to the phylogenetic tree, sites are distinguishable from each other even if amino acid contents in the site are same
- If $f_1\left({}_i\mathcal{M}\right)$ is small, site i is diverged in an early stage of the evolution but if $f_1\left({}_i\mathcal{M}\right)$ is large, site i is diverged more recently
- Although most of existing methods treat gaps as a same thing, the proposed method treats gaps as different things

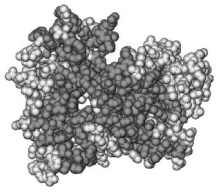
Computation of $f_1\left({}_i\mathcal{M}\right)$

Procedures

1. 984 sequences of EF-Tu/EF-1A proteins were collected from the Swiss-Prot database (Consortium TU, Nucleic Acids Res, 2015)
2. 984 sequences were aligned by the MAFFT program (Katoh K and Standley DM, Mol Biol Evol, 2013)
3. The phylogenetic tree was written from the multiple sequence alignment by the UPGMA method (Sneath PHA and Sokal RR, W. H. Freeman and Company, 1973)
4. $f_1\left({}_i\mathcal{M}\right)$ was calculated at each site

Visualization of predicted functional residues

PDB ID: 2C78



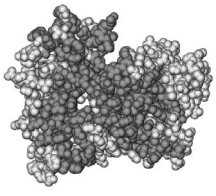
$f_1\left({}_i\mathcal{M}\right)$

small

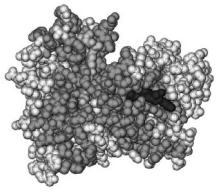
large

Visualization of predicted functional residues

PDB ID: 2C78

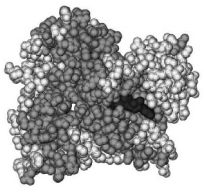


PDB ID: 2C78

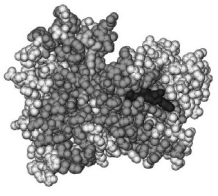


Visualization of predicted functional residues

PDB ID: 2C77



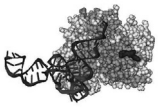
PDB ID: 2C78



Visualization of predicted functional residues

PDB ID: 1AIP

PDB ID: 4V5Q



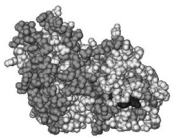
Visualization of predicted functional residues

PDB ID: 3VMF

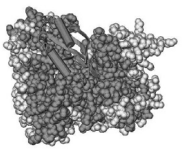
PDB ID: 3WXM

Visualization of predicted functional residues

PDB ID: 4C0S



PDB ID: 1F60



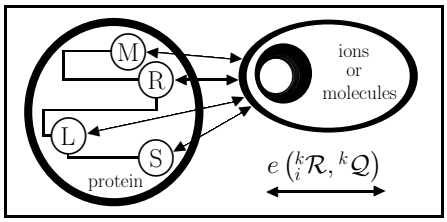
Visualization of predicted functional residues

Summary

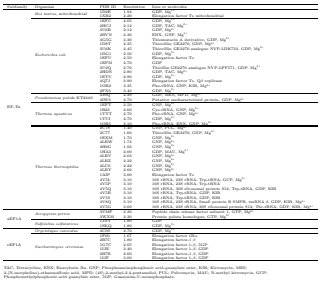
- If $f_1\,({}_i\mathcal{M})$ was small, site i was more likely to be proximate from aa-tRNA, EF-Ts, EF-1Bα, GTP, GDP, pulvomycin, thiocillin, kirromycin and Mg^{2+}
- Therefore, the proposed method could predict regions which are proximate from binding molecules
- However, the evaluation of the predicted functional sites only showed appearances visualized onto the three-dimensional structures
- Therefore, we tried numerical evaluation in order to compare the proposed method with existing methods

Numerical evaluation of the proposed method

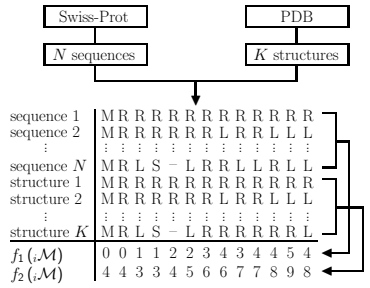
structure k



Structural data of elongation factor Tu/1A



Computations of f_1 and f_2

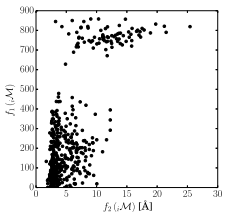


Correlations of f_1 and f_2

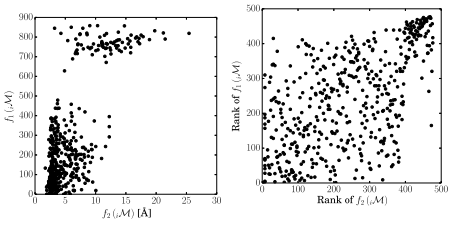
Two methods

1. Spearman's ρ
2. ROC curve

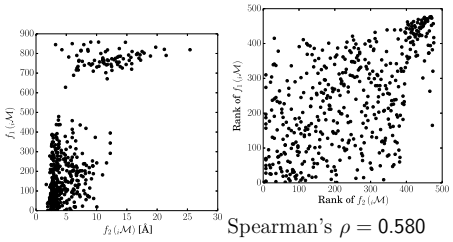
Correlations of f_1 and f_2



Correlations of f_1 and f_2

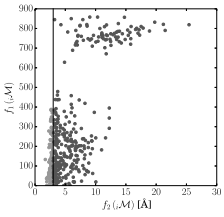


Correlations of f_1 and f_2

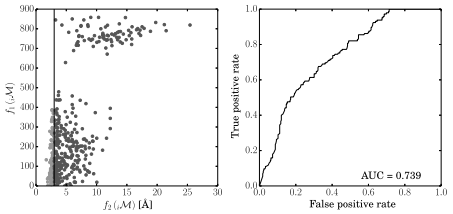


Spearman's $\rho = 0.580$

Correlations of f_1 and f_2



Correlations of f_1 and f_2



Numerical evaluation of the predicted method

Summary

- Existing methods were often evaluated by one of the three-dimensional structures
- However, elongation factor 1A binds various molecules
- Therefore, we used multiple structural data for evaluating the proposed method
- For this evaluation, Spearman's ρ and ROC curve may be useful

Comparisons of the prediction methods

Table 3: Spearman's ρ and AUC of the proposed method or existing methods

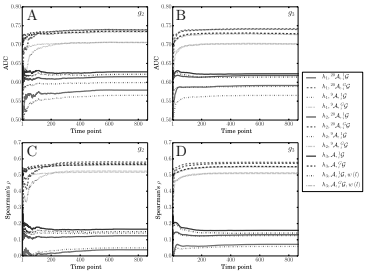
Method	Spearman's ρ	AUC
Proposed	0.580	0.739
SE	0.076	0.578
SEP	0.115	0.571
SP	0.198	0.608
SPW	0.226	0.609
iv-ET	0.122	0.590
rv-ET	0.131	0.618

Comparisons of the prediction methods

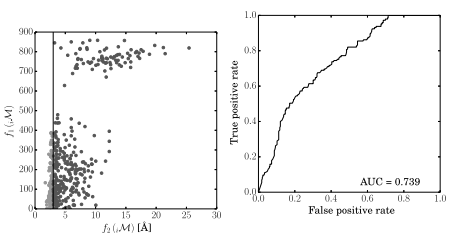
Summary

- The proposed method had higher Spearman's ρ and AUC than the SE-, SEP-, SP- or SPW-based or iv-ET or rv-ET method
- Therefore, the proposed method has a higher ability to distinguish whether a site is proximate from ions or molecules or not
- However, there are two problems
- The first one is which parameters of $f_1(_i\mathcal{M})$ are largely involved in the prediction performance
- The second one is that the threshold of $f_1(_i\mathcal{M})$ has not been determined

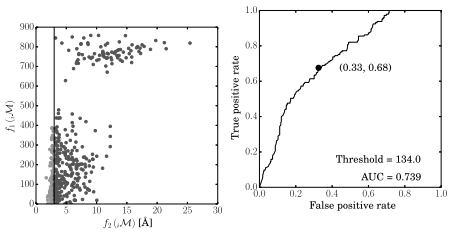
Parameters of f_1 and prediction performances



Determination of predicted functional residues

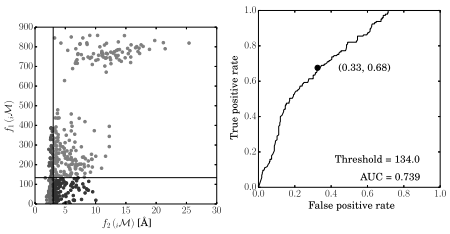


Determination of predicted functional residues



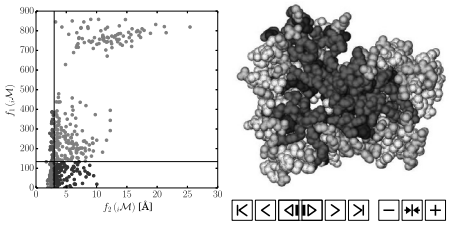
Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Determination of predicted functional residues



Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Determination of predicted functional residues



Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Conclusions in Chapter 3

Conclusions

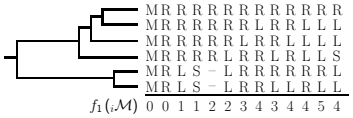
- We proposed a new method of sequence-based computational methods
- The proposed method could predict regions which are proximate from binding molecules
- Multiple structural data were used for evaluating the proposed method and existing methods
- The proposed method showed a higher performance than the existing methods
- The predicted functional residues may be useful for the unknown functions of elongation factor 1A

Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Introduction

In Chapter 3,



In Chapter 4,

- we consider a specific conservation, which can detect a site whose amino acids are conserved in a target subfamily but not conserved in the whole family

Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Definition of a specific conservation

Notation

- ${}^a_i\mathcal{M}$, a field of sets of characters at time point a
- ${}^b_i\mathcal{M}$, a field of sets of characters at time point b

Definition

$$C\left({}^a_i\mathcal{M}, {}^b_i\mathcal{M}\right) = \frac{g_3\left({}^a_i\mathcal{M}\right)+1}{g_3\left({}^b_i\mathcal{M}\right)+1} - 1 \qquad (27)$$

where $a < b$

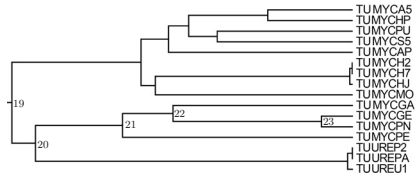
Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Functional divergences of EF-Tu/EF1A

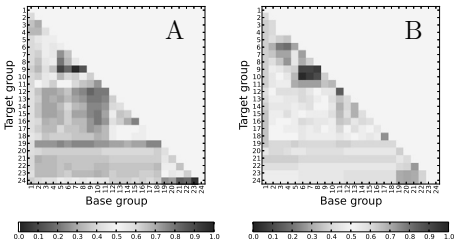
Two functional divergences between

1. *Mycoplasma pneumoniae* EF-Tu and *Mycoplasma genitalium* EF-Tu
   - Fibronectin binding affinity
   - Binding residues of EF-Tu (M193I, N194K, E204T, S343A, P345A and T357A)
2. eEF1A1 and eEF1A2
   - Actin binding affinity
   - Binding residues of eEF1A (N331S and M335Q)

Part of the phylogenetic tree of EF-Tu/EF1A



AUC of specific conservations in Mycoplasma species

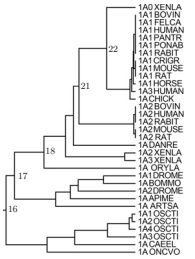


$C\left({}^{a}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right) = \frac{g_3\left({}^{a}_{i}\mathcal{M}\right)+1}{g_3\left({}^{b}_{i}\mathcal{M}\right)+1} - 1$, A: $x = sp$, B: $x = se$

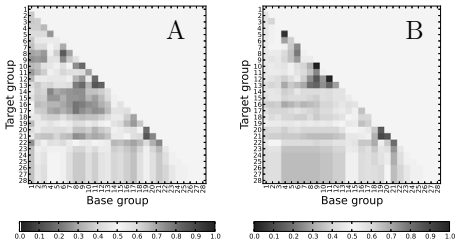
Specific conservations in Mycoplasma species

x	Sum of pairs		Shannon entropy			
Residues	$C\left({}^{a}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right)$	$C\left({}^{a}_{i}\mathcal{M}, {}^{a}_{i}\mathcal{M}\right)$	$C\left({}^{a}_{i}\mathcal{M}, {}^{a}_{i}\mathcal{M}\right)$	$C\left({}^{b}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right)$	Fibronectin binding	$f_3\left(_i\mathcal{M}\right)$
Q49E	-0.00082	0.14600	0.00043	0.22767		291
R130K	-0.00006	0.11945	0.00081	0.22767		266
T142S	0.00079	0.08863	0.00038	0.22767		229
M193I	0.00278	0.15784	0.00323	0.22767	X	176
N194K	0.00166	0.23084	0.00064	0.22767	X	437
E204T	0.00253	0.25597	0.00271	0.22767	X	378
I243V	-0.00197	0.06348	-0.00074	0.22767		265
R249K	-0.00161	0.11945	-0.00013	0.22767		458
D283E	-0.00022	0.13905	0.00058	0.22767		220
S343A	0.00052	0.12061	0.00134	0.22767	X	388
P345A	0.00190	0.29930	0.00307	0.22767	X	191
T357A	0.00280	0.17341	0.00179	0.22767	X	277
S388T	-0.00322	0.08863	0.00059	0.22767		204
AUC	0.97619	0.95238	0.97619	0.50000		
Threshold	0.00023	0.14892	0.00061			
Sensitivity	0.85714	1.00000	0.85714			
Specificity	1.00000	0.83333	1.00000			

Part of the phylogenetic tree of EF-Tu/EF1A



AUC of specific conservations in eEF1A1 and eEF1A2



$C\left({}^{a}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right) = \frac{g_3\left({}^{a}_{i}\mathcal{M}\right)+1}{g_3\left({}^{b}_{i}\mathcal{M}\right)+1} - 1$, A: $x = sp$, B: $x = se$

Specific conservations in eEF1A1 and eEF1A2

x	Sum of pairs		Shannon entropy			
Residues	$C\left({}^{a}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right)$	$C\left({}^{a}_{i}\mathcal{M}, {}^{a}_{i}\mathcal{M}\right)$	$C\left({}^{a}_{i}\mathcal{M}, {}^{a}_{i}\mathcal{M}\right)$	$C\left({}^{b}_{i}\mathcal{M}, {}^{b}_{i}\mathcal{M}\right)$	Actin binding	$f_3\left(_i\mathcal{M}\right)$
T63I	0.00403	0.09150	0.01551	0.19407		438
V87I	-0.00468	0.06226	0.00442	0.19407		43
F161A	0.02264	0.24512	0.00125	0.19407		820
Q164E	0.03766	0.13165	0.01125	0.19407		307
E166D	0.00917	0.12241	0.00984	0.19407		207
T176A	0.00567	0.15339	0.00740	0.19407		204
D186A	-0.00321	0.24408	0.00252	0.19407		188
A189P	-0.00725	0.28690	-0.00221	0.19407		109
N197H	-0.00107	0.25501	0.00426	0.19407		278
A206P	0.00098	0.05119	0.01027	0.08149		768
T217E	0.02192	0.23347	0.01228	0.19407		819
D220E	0.01581	0.03949	-0.00070	0.11018		628
T224S	-0.00233	0.07092	0.00579	0.19407		328
T227S	-0.00233	0.07092	0.00579	0.19407		328
C234T	0.00033	0.29774	0.00084	0.19407		479
V271I	0.00281	0.06226	-0.00071	0.19407		395
K273R	-0.00389	0.11880	-0.00113	0.19407		292
V285I	0.03537	0.06226	0.01110	0.19407		372
V326I	0.00474	0.06226	-0.00023	0.19407		206
A326C	0.00771	0.31076	0.00082	0.19407		188
N331S	0.00691	0.20197	0.00510	0.25547	X	288
M335Q	0.03814	0.27765	0.01750	0.19407	X	111
G339Q	-0.02090	0.16000	-0.00380	0.15548		350
A342S	-0.00375	0.10550	-0.00168	0.19407		219
A350S	0.03212	0.10550	0.01011	0.19407		122
L361I	0.02974	0.02915	0.01803	0.15548		87
G390N	-0.00256	0.08175	-0.00219	0.11018		225
D401E	0.00208	0.11476	0.00418	0.19407		439
D403E	0.00208	0.11476	0.00418	0.19407		439
D417Q	-0.00424	0.19312	-0.00304	0.19407		790
A440N	-0.01732	0.24580	-0.00275	0.19407		312
D442E	0.00573	0.03949	0.00231	0.11018		438
A445S	-0.00473	0.10550	0.00362	0.19407		801
A446G	-0.00585	0.14825	0.00157	0.11018		795
K462G	-0.02254	0.33389	-0.00195	0.19407		
-465K	-0.00849	0.23160	-0.00183	0.19407		
AUC	0.83824	0.76900	0.77941	0.80147		
Threshold	0.00632	0.19754	0.00476	0.22477		
Sensitivity	0.67647	0.64700	0.61765	1.00000		
Specificity	1.00000	1.00000	1.00000	0.50000		

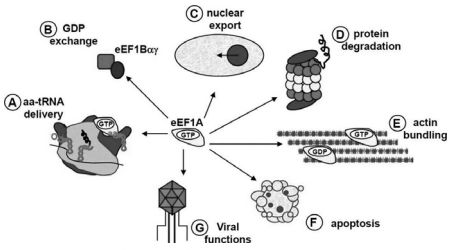
Conclusions in Chapter 4

Conclusions

- The specific conservation could detect fibronectin- or actin-binding residues
- Therefore, residues involving functional divergences of EF-Tu/EF1A were detected by the specific conservations
- In order to comprehensively predict protein functional sites, we should use a suitable conservation

Introduction

Eukaryotic translation elongation factor 1A (eEF1A) has many functions



Maria K. Mateyak, and Terri Goss Kinzy J. Biol. Chem. 2010;285:21209-21213

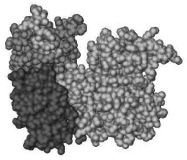
Introduction

Prediction of complex structures

- FNIII14 triggers anoikis by binding to membrane-localized eEF1A (Itagaki and Fukai et al. 2012)
- FNIII14 was extracted from the structure of fibronectin (PDB ID: 1FNH)
- Human eEF1A was modeled by homology modeling using yeast EF1A (PDB ID: 1F60) as a template
- The complex structures were predicted by rigid-body docking (ZDOCK: Chen et al. 2003)

Definition of domains

Human eEF1A1



domain 1
domain 2
domain 3

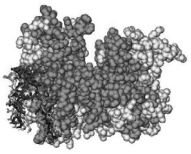
Docking of EF-Tu/EF1A

Two dockings blocked by

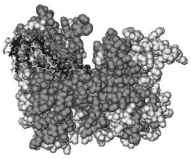
1. domains 1 and 2
2. domains 1 and 2 except for the border region of domain 3

Visualization of predicted functional residues

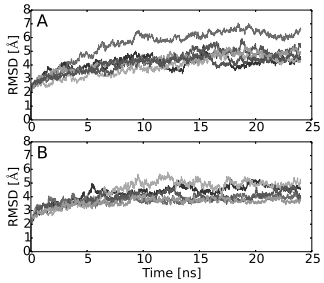
Docking 1



Docking 2



RMSD of MD trajectories



Binding energies of eEF1A1 and FNIII14

Docking 1					
Complex	ΔE_{VDW}	ΔE_{Coul}	$\Delta G_{Hydrophobic}$	$\Delta G_{Hydrophilic}$	ΔG_{bind}
1	-377.149	-75.393	-9.503	404.901	-57.140
2	-388.549	-57.471	-8.257	409.733	-44.542
3	-318.291	-70.949	-8.434	352.641	-45.033
4	-340.263	-51.481	-7.330	364.477	-34.599
5	-278.515	-61.470	-7.888	302.059	-45.815
6	-331.041	-58.237	-8.236	360.948	-36.566
Mean	-338.968	-62.500	-8.275	365.793	-43.949
Docking 2					
Complex	ΔE_{VDW}	ΔE_{Coul}	$\Delta G_{Hydrophobic}$	$\Delta G_{Hydrophilic}$	ΔG_{bind}
1	-393.201	-93.633	-12.108	439.697	-59.242
2	-335.289	-81.710	-10.487	376.478	-51.009
3	-415.172	-61.606	-9.245	449.527	-36.497
4	-335.931	-53.589	-8.283	364.951	-32.852
5	-351.509	-63.574	-9.265	383.235	-41.113
Mean	-366.220	-70.822	-9.878	402.778	-44.143

(kcal/mol)

Sequence and structural analysis of eEF1A1

Docking 1			Docking 2		
Residue	f_i [-Å²]	Rate of interaction	Residue	f_i [-Å² method 3]	Rate of interaction
H349	46	0.300	K100	2	0.716
P350	66	0.375	N101	22	0.691
G351	88	0.401	T104	50	0.335
Q352	88	0.559	G105	0	0.542
I353	79	0.292	T106	11	0.649
S354	140	0.457	S107	16	0.407
A355	263	0.243	P249	61	0.252
G356	112	0.485	L250	227	0.263
Y357	19	0.834	Q251	48	0.825
A358	122	0.970	D252	57	0.247
P359	35	0.712	V253	96	0.233
V360	42	0.744	R266	69	0.178
H367	31	0.148	P282	23	0.190
I368	119	0.260	V317	101	0.283
A369	41	0.664	K318	269	0.267
C370	50	0.256	V320	206	0.188
K371	261	0.514	R321	192	0.962
K392	130	0.178	R322	57	0.990
F393	99	0.198	G323	19	0.583
G407	163	0.301	H364	43	0.258
K408	278	0.200	T365	43	0.683
R427	34	0.323	A366	116	0.946
D428	36	0.658	H367	31	0.893
M429	28	0.696	I368	119	0.440
R430	54	0.252	K408	278	0.586
V433	110	0.335	P409	151	0.248
			R427	34	0.335
			R430	54	0.188

Conclusions in Chapter 5

Conclusions

- The structural analysis showed similar regions which were predicted by sequence analysis in Chapter 3
- Therefore, we could narrow down functional amino acid residues by combining sequence and structural analyses
- The predicted binding residues should be verified by experimental approaches such as site-directed mutagenesis

Conclusions and future works

Relationships between Chapters 3, 4 and 5

- The results in Chapter 3 showed that the proposed method can predict the regions involving protein biosynthesis
- However, there are two problems
- The first one is that all the functional sites cannot be predicted such as fibronectin- or actin binding residues
- However, in Chapter 4, the functional sites could be predicted by specific conservations
- The second one is that there are many predicted functional sites in the results of Chapter 3
- In Chapter 5, the functional residues were narrowed down by combining sequence and structural analyses

Conclusions and future works

Relationships between Chapters 3, 4 and 5

- Chapter 5 investigated unknown functional sites
- Chapters 3 and 4 investigated known functional sites
- The results in Chapters 3 and 4 showed that the proposed method can comprehensively predict functional sites of EF-Tu/1A
- The results also showed that the existing methods are difficult to predict protein functional sites comprehensively
- Therefore, thanks to the proposed method, a protein which has many functional sites could become applicable for such prediction methods

Conclusions and future works

Effectiveness of the prediction methods

- This study focused on functional site prediction from protein sequences
- This method can determine candidate functional amino acid residues by following procedures
  1. Selection of one protein sequence
  2. Collection of sequences by similarity search
  3. Construction of a multiple sequence alignment
  4. Reconstruction of a phylogenetic tree
  5. Calculation of conservations
  6. Determination of candidate residues
- Reduction of time and costs for experiments by using huge sequence data registered in biological databases

Conclusions and future works

Future works

1. Biological aspects
   - The results may be useful for further analysis of unknown functions of elongation factor Tu/1A
2. Informatics aspects
   - Other existing methods such as relative entropy or Jensen Shannon divergence based method and ConSurf or Rate4Site algorithm should be integrated into the proposed method because this study only integrated SE- or SP-based methods and iv-ET or rv-ET method
   - Our proposed method should be tried various proteins

Thank you very much for listening

Appendix D

Related research

This section describes a related research which used protein structural analyses [62]. Based on the parameter settings of the structural analyses, MD simulations in Chapter 5 were conducted. The subject of this research is structural analyses of a large progenitor toxin complex in *Clostridium botulinum*. Experimental results of the toxin complex suggested that the complex structure is flexible and the flexibility is important for binding to the intestinal surface. However, such flexibility has been investigated by experimental analyses but not by computer simulations. Therefore, this research conducted computational analysis and aimed to investigate what motions are included in the complex structure and how the complex structure is changed.

D.1 Abstract

Botulinum neurotoxins (BoNTs) are one of the potent toxins in nature but the toxicity is immediately eliminated by the harsh environment in the digestive tract because BoNTs are just a protein. However, BoNTs can get to synaptic junctions thanks to support of other proteins by aggregating with each other, combining with BoNTs and forming a large progenitor toxin complex (L-PTC). In order to explain how the complex formation enables the BoNTs to intrude into our body, we found that the three-dimensional structure of the L-PTC consists of an ovoid body with three legs and speculated important roles in the body and the legs. In the legs part, it is helpful for promoting absorption especially from the small intestine. Because experimental results showed that the legs are flexible and have specific binding sites of saccharides, the flexibilities may help the L-PTC to easily access the binding sites to the saccharides on the intestinal surface. However, such flexibilities have been only investigated by experimental methods. This means that we still have not objectively discussed what motions are generated from the shape of the L-PTC and how the structure is changed gradually. Therefore, we developed a new method integrating normal mode analysis and principal component analysis in order to measure the dynamics of the large-

sized protein consisting of several subunits. The results showed that the L-PTC had characteristic motions which have large movements of the three legs. In addition, the flexible motions were appeared regardless of the theoretical models. Our application to measure the dynamics of the L-PTC suggested the importance of the flexibility which enables the L-PTC to break the epithelial barrier. We hope that the activity of the L-PTC is applied for developing a new oral drug delivery system.

D.2 Introduction

At the neuromuscular junction between the presynaptic terminal of a motor neuron and the postsynaptic membrane of a skeletal muscle, acethylcholine (ACh), 2-acetoxy-N,N,N-trimethylethanaminium, works as a neurotransmitter. The reason why AChs synthesized in the presynaptic cell and taken into a vesicle are released from the presynaptic cell is a membrane fusion, which is triggered by binding synaptosomal-associated protein of 25 kDa (SNAP-25) and vesicle-associated membrane protein (VAMP) on the vesicle surface with syntaxin on the inner cellular membrane [100]. As a result, ACh can bind to an ACh receptor on the postsynaptic membrane. However, if AChs are not released from the presynaptic cell, we cannot transmit signals from the motor neuron to the skeletal muscle. Botulinum neurotoxin (BoNT) inhibits releasing AChs from the presynaptic cell by cleaving SNAP-25, VAMP or syntaxin. Before this cleavage, it should be required to eventually get BoNTs to reach the presynaptic terminal.

One of the intruding roots of BoNTs is oral. In the digestive tract, however, BoNTs are inactivated by the acidic condition or enzymes. Consequently, if BoNTs are taken orally, we can inhibit BoNTs to get to the presynaptic terminal. However, several other proteins called non-toxin components, which are simultaneously produced with BoNTs by *Clostridium botulinum*, help BoNTs to delivery into the blood-stream. The non-toxic components, which consist of a non-toxic non-hemagglutinin (NTNHA) and three hemagglutinins (HA70, HA17 and HA33) [25], can form various stoichiometric complexes with BoNTs. Such complexes are called as progenitor toxin complexes (PTCs) and include the minimally functional PTC (M-PTC) and the large PTC (L-PTC) [89, 88, 45]. Recent studies showed that the stoichimetries of the M-PTC and the L-PTC are BoNT:NTNHA = 1:1 [33] and BoNT:NTNHA:HA70:HA17:HA33 = 1:1:3:3:6 [69], respectively.

Some researchers suggested that the structures of the M-PTC and the L-PTC play important roles in the delivery. There are two effects. The NTNHA of the M-PTC is effective for protecting against inactivation of the BoNT by directly binding with the BoNT [33]. The HA parts of the L-PTC are effective for promoting absorption from the intestinal surface [69]. In the latter, for two reasons that the HA has specific binding sites of saccharides such as sialic acid, galactose, lactose, luctulose or isopropyl β-D-1-thiogalactopyranoside [85, 122, 71, 70] and the HA parts are flexible [13], the flexibilities may help the L-PTC to easily access the binding sites to the saccharides on the intestinal surface. However,

such flexibilities have not been investigated by computer analysis or simulation. This has caused two problems. The first one is that we cannot define what motions are generated from the structure of the L-PTC. The second one is that we cannot know how the structure of the L-PTC is changed gradually.

In this paper, we construct the three-dimensional structure of the L-PTC from crystal structures and then consider three theoretical models: anisotropic network model (ANM), coarse grained (CG) model and united atom model. Based on each model, the dynamics of the L-PTC are estimated by normal mode analysis (NMA) or MD simulation. Finally, the results of the NMA and MD simulations are compared.

D.3 Materials and Methods

D.3.1 Modeling of the L-PTC

From the Protein Data Bank (PDB) [16], we downloaded PDB ID : 3V0A (BoNT-NTNHA) [33], PDB ID : 4LO4 (a trimer of HA70s) [69] and PDB ID : 4LO7 (HA70-HA17-HA33) [69]. In order to model an HA complex, HA70:HA17:HA33 = 3:3:6, the C-terminal region of HA70 in 4LO7 was superimposed to each C-terminal region of HA70 in 4LO4 by the Align and Superimpose Proteins protocol of Discovery Studio 3.1 [2]. In order to model an L-PTC, we conducted two things: constructing N-loop of NTNHA in 3V0A by the Loop Refinement MODELER protocol and superimposing the HA complex and BoNT-NTNHA to L-PTC (EMDB ID : EMD2417) [69] by the gmconvert and the gmfit programs by setting 10 as the number of Gaussian distribution functions and 10,000 as the number of initial configurations by random generation [56] and then some structures of the L-PTC were modeled.

D.3.2 ANM analysis

We used the ProDy Python package for ANM analysis [10]. In one of the modeled structures, 6,535 Cα atoms were extracted and a cutoff and a γ parameter were set to 1.5 nm and 1.0, respectively.

D.3.3 MD simulations

We used the GROMACS software [94] for MD simulation. Although various force fields and parameters were tried, we eventually selected two computational conditions. In the first one, we specified the GROMOS53A6 force field [90]. Then, we constructed a rectangular box whose size is $\approx 28\times28\times24$ nm^3 and put $\approx$ 600,000 TIP-3P water molecules [51] and 107 sodium ions into the box. We conducted energy minimizations of 10,000 steps by the steepest descent method and 10,000 steps by the conjugate gradient method. 0.1 ns MD simulations were conducted by increasing the reference temperature to 310 K by the velocity rescale method. The time step was set to 2 fs by using the LINCS algorithm

for hydrogen atoms [38]. The switch method was used and pair list, long range and short range cutoffs were set to 1.2 nm, 1.1 nm and 1.0 nm, respectively. The Coulomb energy was computed by the PME algorithm [23]. Next, 10 ns MD simulations were conducted with the Parrinello-Rahman method [92] and the reference pressure was set to 1.0 bar. Independent MD simulations were conducted 10 times and the MD trajectories were analyzed by Cartesian principal component analysis (PCA) [3] by extracting 100 snapshots from each independent MD simulation.

In the second one, we specified the MARTINI21 force field [82] with an elastic network which is 500 kJ mol^{-1} nm^{-2} of the spring constant, 0.5 and 0.9 nm of the lower and the upper cutoffs, respectively [93]. Then, we constructed a rectangular box whose size is $\approx 29\times28\times25$ nm^3 and put $\approx$ 160,000 CG water molecules [82] into the box. We conducted energy minimizations of 10,000 steps by the steepest descent method and $\leq$ 10,000 steps by the conjugate gradient method. 1.0 ns MD simulations were conducted by the Berendsen method [15] and the reference temperature was set to 310 K. The time step was set to 20 fs by using the LINCS algorithm [38]. The shift method was used and the pair list cutoff, the long or short range cutoff of VdW energy and the short range cutoff of Coulomb energy were set to 1.4 nm, 1.2 nm, 0.9 nm and 0.0 nm, respectively. Next, 1,000 ns MD simulations were conducted by the Berendsen method and the reference pressure was set to 1.0 bar. Independent MD simulations were conducted 10 times and the MD trajectories were analyzed by Cartesian PCA [3] by extracting 1,000 snapshots from each independent MD simulation.

In order to estimate similarities between the ANM analysis and/or PCAs [11], two eigenvectors were represented as $\boldsymbol{A} = (a_1, a_2, \ldots, a_{3n})$ and $\boldsymbol{B} = (b_1, b_2, \ldots, b_{3n})$ and cosine similarity S was computed as

$$S = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{||\boldsymbol{A}||\,||\boldsymbol{B}||} = \frac{\sum_{i=1}^{3n} a_i b_i}{\sqrt{\sum_{i=1}^{3n} a_i^2}\sqrt{\sum_{i=1}^{3n} b_i^2}}, \tag{D.1}$$

where n is the number of Cα atoms or backbone atoms.

D.3.4 Visualization

Eigenvalues, projections and heat maps were visualized by the matplotlib Python package [44]. Three-dimensional structures were visualized by the VMD program [43].

D.4 Results

We conducted ANM analysis or PCAs of 10 or 1,000 ns MD simulations of the L-PTC. Figure 1A shows that 20 collective motions which involve to the whole structure of the L-PTC were obtained by the

ANM analysis. Figure 1B shows that, in all eigenvalues, rates of eigenvalues of PCA axes 1 and 2 are 27.8% and 16.4%, respectively. Figure 1C shows that, in all eigenvalues, rates of eigenvalues of PCA axes 1 and 2 are 36.7% and 23.1%, respectively. Figure 1D shows that the trajectories move a narrower space than the trajectories shown in Figure 1E. Figure 1E shows that some trajectories move well to the positive or the negative direction on PCA axis 1 or 2. For example, the red trajectories move to the negative direction on PCA axis 1, the green trajectories move to the positive direction on PCA axis 1 and the cyan trajectories move to the positive direction on PCA axis 2.

We computed cosine similarities from the ANM analysis and PCAs. Figure 2 shows that the highest value is 0.750, which is calculated from ANM mode 1 and PCA axis 2 by the CG model. The second highest value is 0.705, which is calculated from ANM mode 2 and PCA axis 1 by the CG model. The third highest value is 0.608, which is calculated from PCA axis 1 by the united atom model and PCA axis 1 by the CG model.

The followings are details of the movements of the L-PTC. Animation S1 shows that when the left leg moves up and the body moves left, the right leg moves down and the middle leg moves right and vise versa. Therefore, when the left leg moves down and the body moves right, the right leg moves up and the middle leg moves left. Animation S2 shows that the middle leg and the body move up and down and the left and the right legs move down and up, respectively. Animation S3 shows that the middle, the left and the right legs move up and down.

D.5 Discussion

Dynamics of the L-PTC were estimated by three computational models. The first one only considers Cα atoms and the forces between atoms are represented as a harmonic potential. The second one is the CG model whose amino acid residues are represented as 1-5 particles. The forces between particles are also represented as a harmonic potential. As other forces, bond, angle and dihedral angle and Coulomb and VdW interactions and water effects are considered. The third one is the united atom model whose carbon atom and hydrogen atoms connecting to the carbon atom are represented as 1 particle. As forces, bond, angle and dihedral angle and Coulomb and VdW interactions and water effects are also considered. On the basis of the computational models, the motion of the L-PTC was investigated. ANM analysis enables us to obtain sets of direction vectors of Cα atoms. The motion defined by the direction vectors is called as a collective motion and we can select the motion which involves all parts of the structure. MD simulation enables us to obtain trajectories of the atoms or the particles and PCA enables us to obtain direction vectors involving the large variance of the trajectories.

The ANM analysis and the PCAs of MD trajectories were compared by a square of the cosine similarity. The range of the square of the cosine similarity is from 0 to 1. The larger the value is,

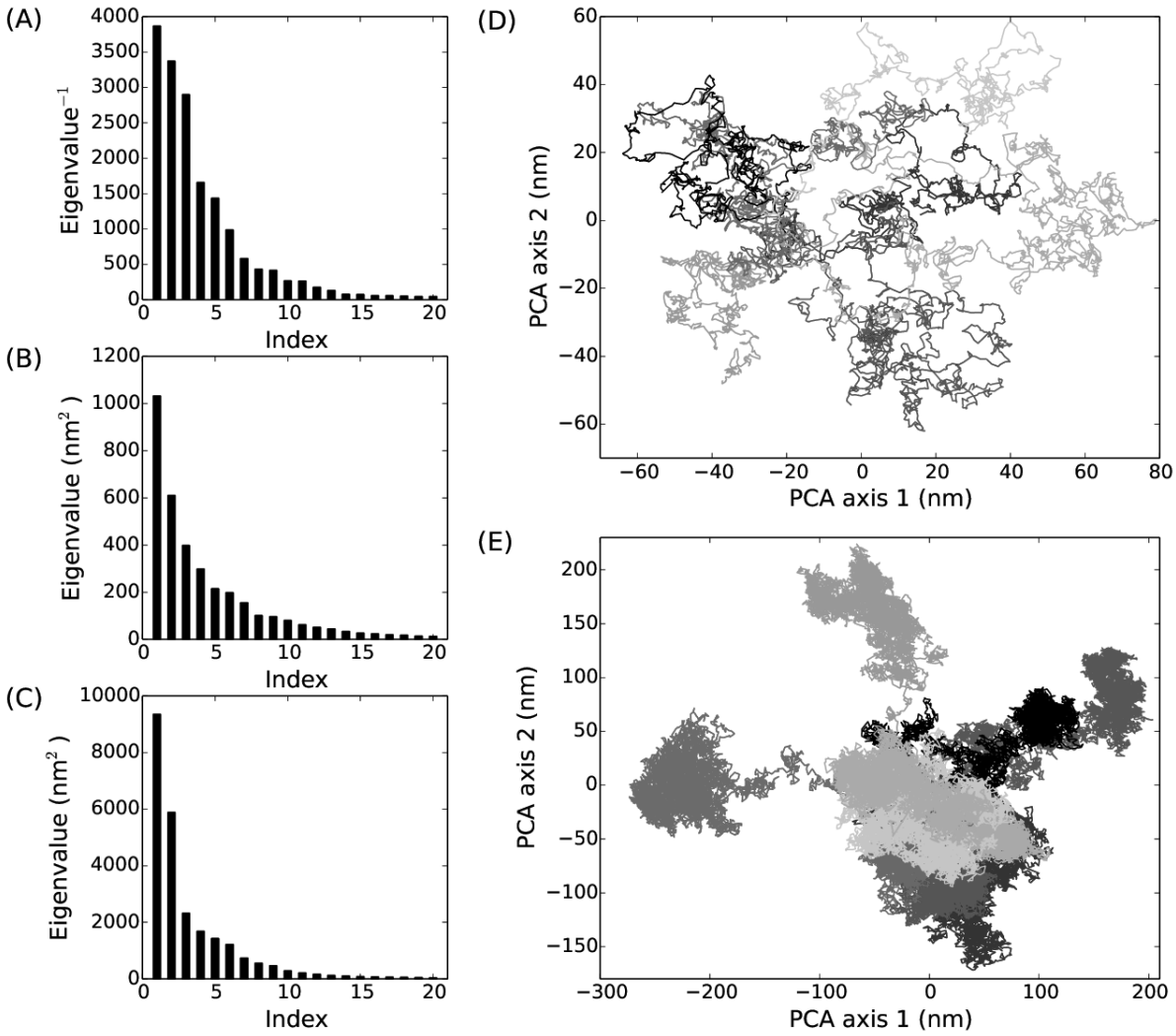


Figure D.1: **Eigenvalues and projections of MD trajectories.** (A) Abscissas show eigenvalue indices and the ordinate shows inverses from seventh to twenty-sixth lowest eigenvalues. The bar graphs are based on the results of (B) the united atom model and (C) the CG model and abscissas show eigenvalue indices and ordinates show the twenty highest eigenvalues of the PCAs of MD trajectories. The line graphs are based on the results of (D) the united atom model and (E) the CG model and show projections of MD trajectories. In each PCA, 10 independent simulations, each of which comprises 10 ns or 1,000 ns MD trajectories, were merged and projected onto the first and the second principal axes.

the more similar the vector direction is. In Figure 2, the highest value is 0.750 and the second highest value is 0.705. This means that the MD trajectories by the CG model include the motion resembling to ANM modes 1 and 2 as a large movement. As described above, the ANM analysis and the MD simulation by the CG model also consider harmonic potentials. Therefore, such a harmonic potential may be important for the motions which have flexible three legs. Additionally, the results also show that such motions are included even if the model considers waters or not. This implies that the shape of the L-PTC is important for the motions.

Meanwhile, the third highest value in Figure 2 is 0.608. This means that the MD trajectories by the CG model include the motion resembling to MD trajectories by the united atom model as a large

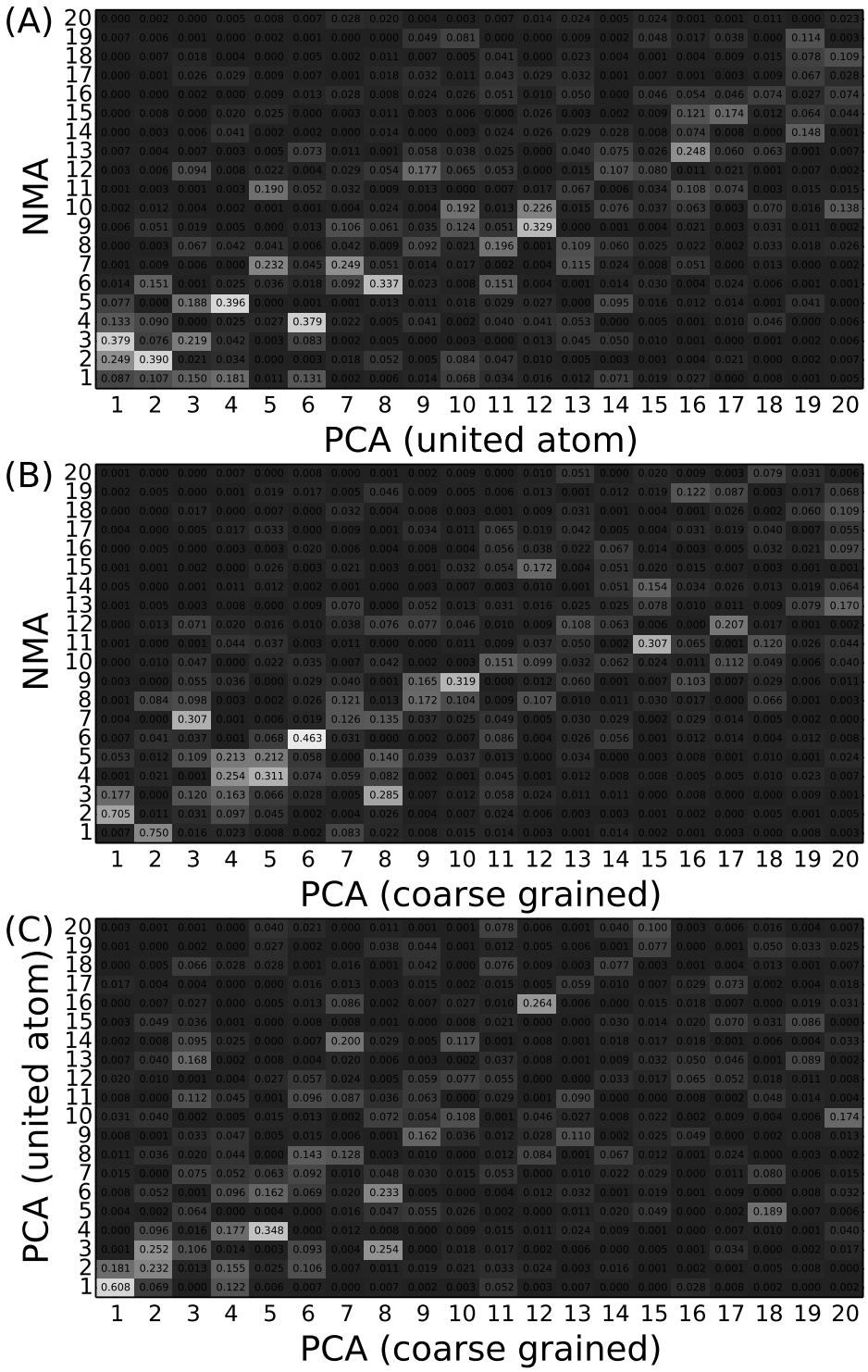


Figure D.2: **Similarities between the NMA and PCAs.** 20×20 combinations of squares of cosine similarities were computed from (A) ANM analysis and PCA of the united atom model, (B) ANM analysis and PCA of the CG model and (C) PCA of the united atom model and PCA of the CG model.

movement. Therefore, the motion like Animation S3 shown by the red trajectories in Figure 1E is included in MD trajectories by the united atom model because the red trajectories move well to the negative direction on PCA axis 1. This indicates that the movements such as Animation S3 could

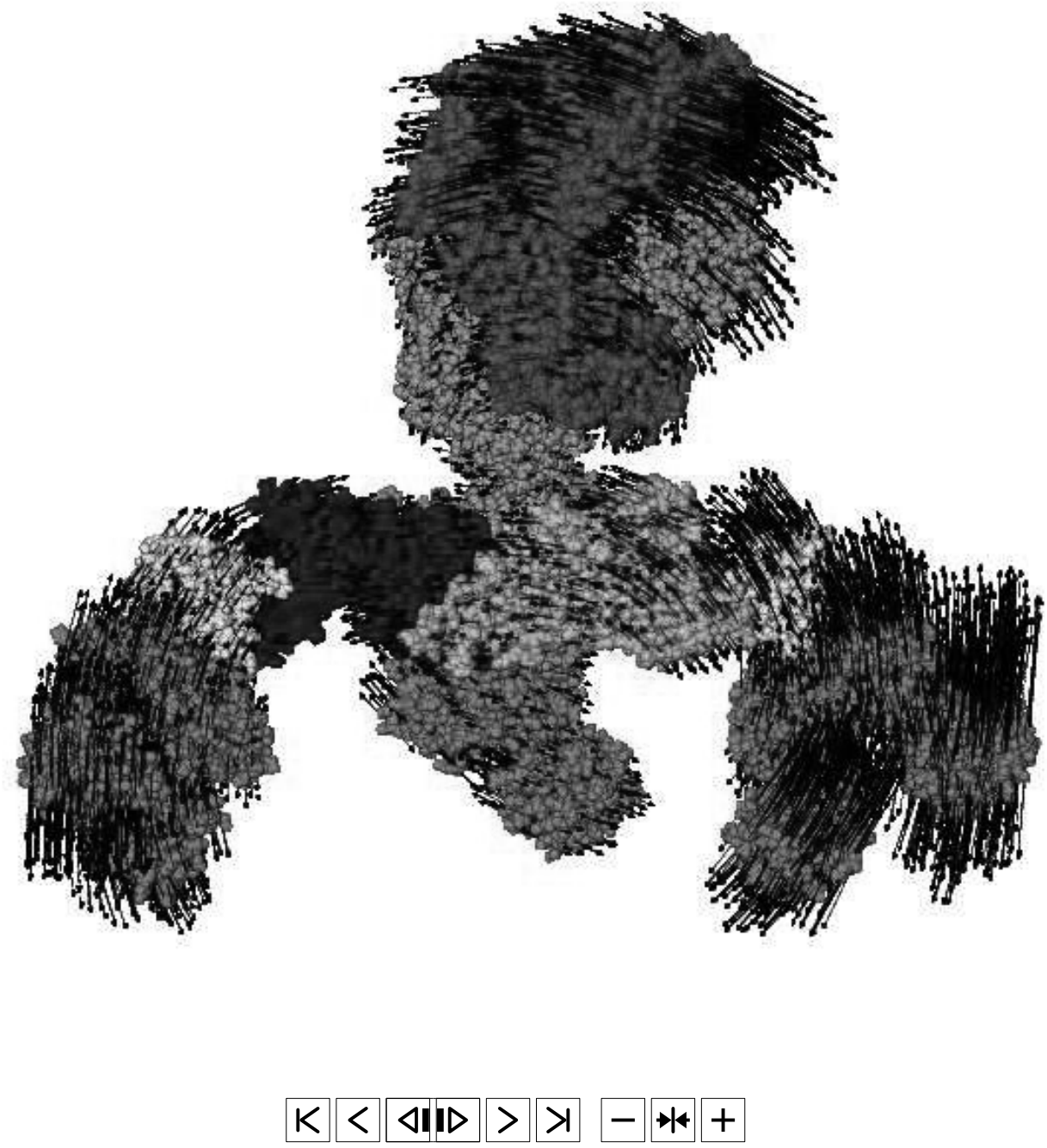


Figure D.3: **N**ormal mode 1. Black arrows show an eigenvector whose corresponding eigenvalue is the smallest in all eigenvalues except 6 zero eigenvalues. Molecular colors are same with Animation 3.

be also reproduced by the model which does not consider the harmonic potential. Therefore, if the simulation time of the L-PTC by the united atom model increased, we might be able to obtain the motion like Animation S3. For these reasons, such flexible motions could be obtained regardless of the three theoretical models.

Recent studies showed that the legs also interact with glycoprotein 2 [79] or E-cadherin [107, 72]. This implies that for binding with these molecules, the flexibilities of the three legs are also important. Therefore, we should take account of such flexibility for understanding how the L-PTC facilitates

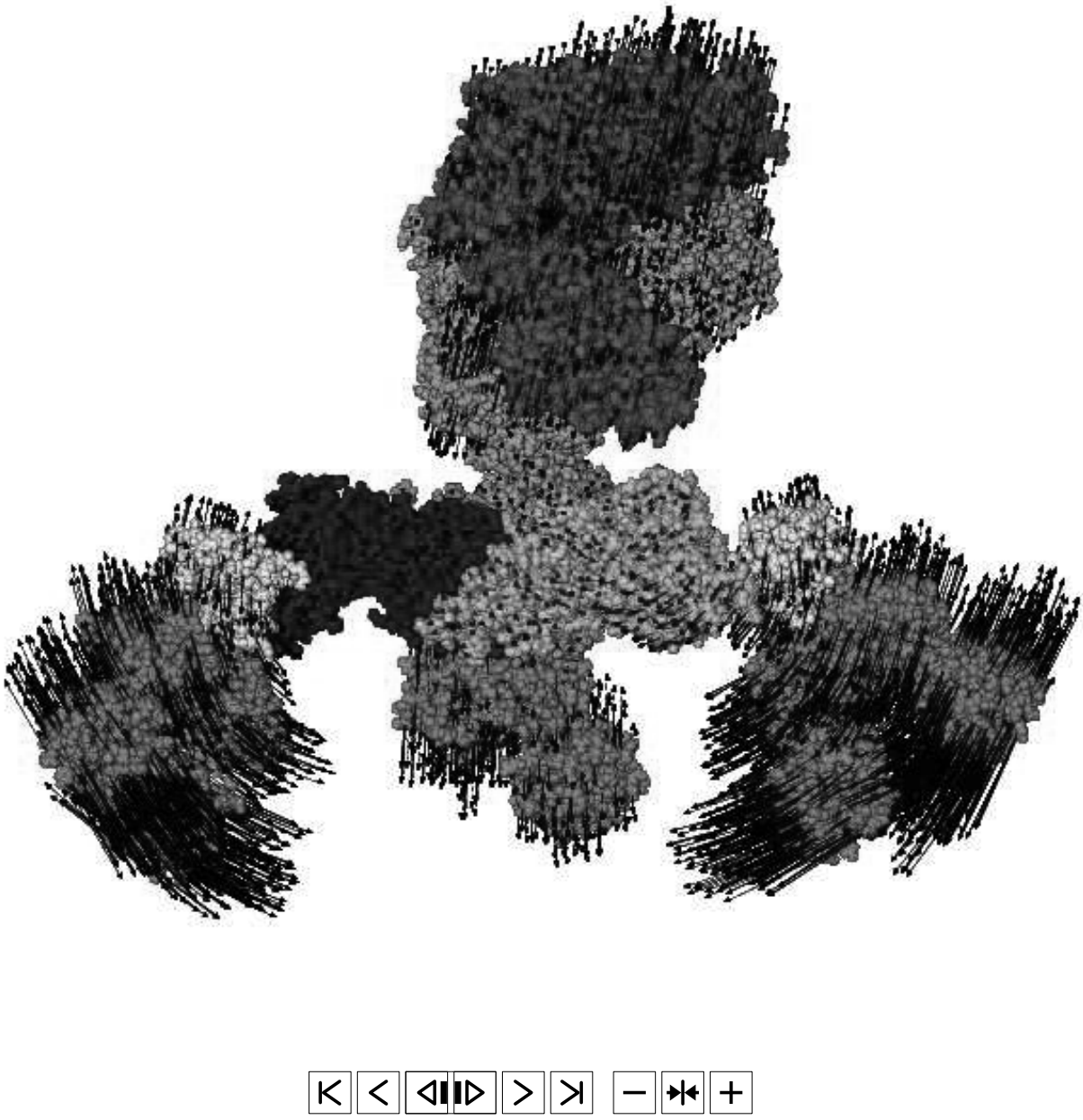


Figure D.4: **N**ormal mode 2. Black arrows show an eigenvector whose corresponding eigenvalue is the second smallest in all eigenvalues except 6 zero eigenvalues. Molecular colors are same with Animation 3.

translocation and disruption of the intestinal epithelial barrier [72, 64].

D.6 Conclusions

In this study, a complex structure, which consists of 14 subunits, was investigated by computer analysis and simulation. The results showed that the L-PTC had flexible movements of the three legs and the flexible motions were included in different types of the theoretical models. This indicates that the structure which is an ovoid body with three legs is important for large movements of the three legs. Based on the flexibility, we suggest a story that the flexible structure enables the L-PTC to easily access

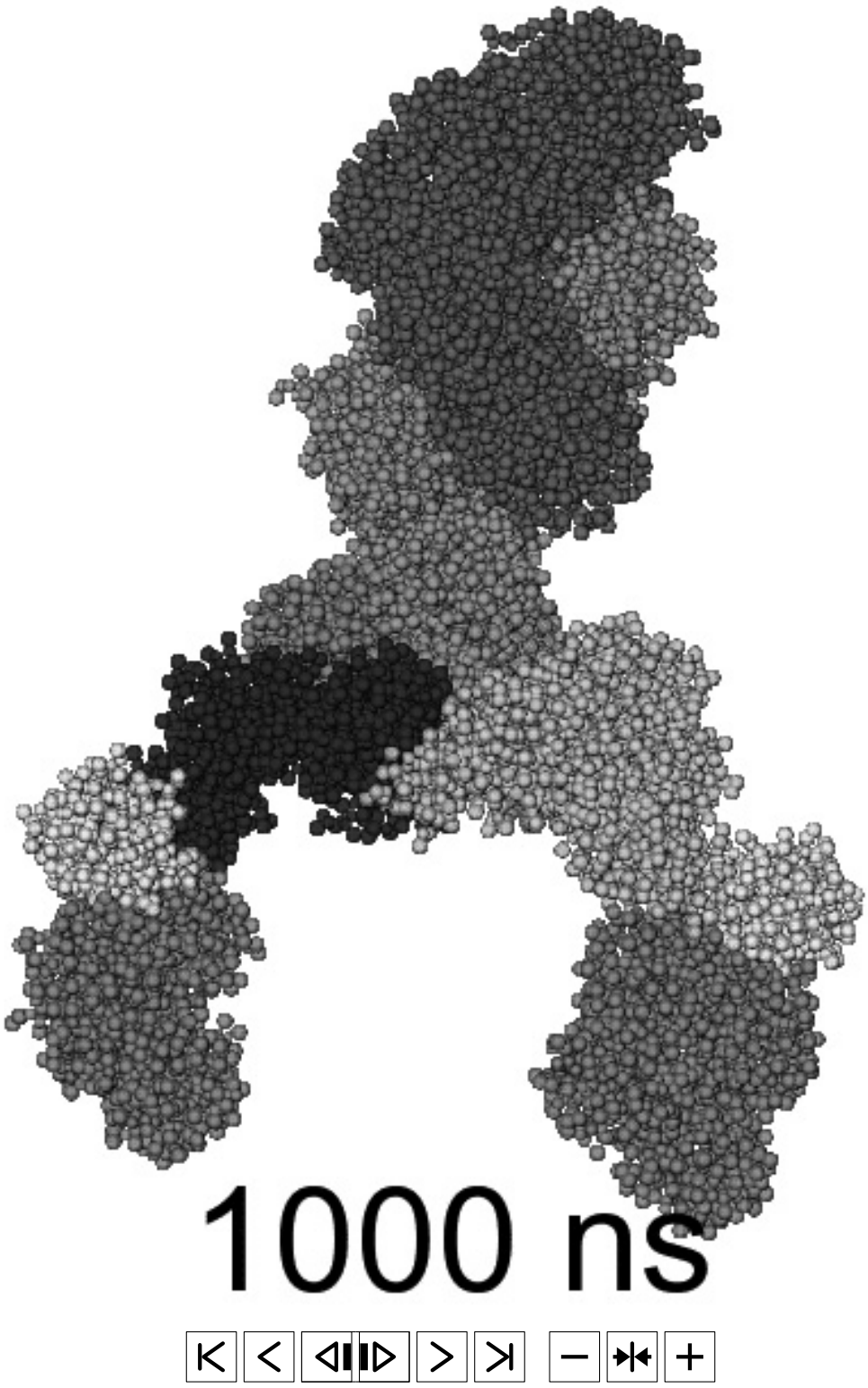


Figure D.5: **1**,000 ns MD trajectories. The trajectories show the red line in Figure 1E. The L-PTC was represented by the CG model. Magenta and pink spheres show atoms belonging to the BoNT and NTNHA, respectively. Green, blue and cyan spheres show atoms belonging to each HA70. Yellow and orange spheres show atoms belonging to HA17 and HA33, respectively.

to saccharides on the intestinal surface and then the L-PTC can stay in the surface for a long time and make an opportunity to be absorbed from the surface. Therefore, we supposed that the flexibility is

effective for making an initial step to translocate the epithelial cell and should consider relationships between such flexibility and interactions with saccharides, glycoprotein 2 and E-cadherin. In addition, such flexibility is generated by formation of the quaternary structure. This implies that the complex formation is important for obtaining an ability which does not appear in each subunit only.