Measuring the Similarity of Proteomes using Grammar-based Compression via Domain Combinations

Morihiro Hayashida[1], Hitoshi Koyano[2] and Jose C. Nacher[3]
[1]*Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, Matsue, Shimane, Japan*
[2]*School of Life Science and Technology, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan*
[3]*Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, Japan*

Keywords: Grammar-based Compression, Kolmogorov Complexity, Protein Domain Combination.

Abstract: Revealing evolution of organisms is one of important biological research topics, and is also useful for understanding the origin of organisms. Hence, genomic sequences have been compared and aligned for finding conserved and functional regions. A protein can contain several domains, which are known as structural and functional units. In the previous work, a proteome, whole kinds of proteins in an organism, was regarded as a set of sequences of protein domains, and a grammar-based compression algorithm was developed for a proteome, where production rules in the grammar represented evolutionary processes, mutation and duplication. In this paper, we propose a similarity measure based on the grammar-based compression, and apply it to hierarchical clustering of seven organisms, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Escherichia coli*. The results suggest that our similarity measure could classify the organisms very well.

1 INTRODUCTION

To understand evolutionary processes of organisms is important, and many researchers are interested in the origin of organisms. DNA sequences are changed by mutation, recombination, gene duplication, and so on, which are responsible for genetic variation as well as evolution of new genes and species. For classifying the evolutionary lineage of an organism, 16S ribosomal RNA is often used because the gene is included in every organism and the evolution rate is slow (Woese and Fox, 1977).

Comparing DNA and protein sequences is a fundamental task in molecular biology and bioinformatics field, and many sequence search tools such as FASTA (Lipman and Pearson, 1985) and BLAST (Altschul et al., 1990) have been developed. Recent high-throughput sequencing technologies produce very long, full-length reads and very large datasets, and need more efficient alignment methods. Minimap2 was three times as fast as existing methods at comparable accuracy to map DNA or long mRNA sequences against a large reference database (Li, 2018). Compression techniques are often useful for efficient analyses of DNA and protein sequences, and for saving storage space. GReEn was developed for compressing genome resequencing data using a reference genome sequence, and outperformed several existing methods in storage space requirements and running times (Pinho et al., 2012). LFQC is a lossless non-reference compression method for FASTQ files, and achieved better compression ratios on several datasets (Nicolae et al., 2015). For our purpose, however, it is difficult to extract evolutionary construction of DNA and protein sequences using compression in a simple manner.

Protein domains are part of a protein, often form globular structures, and are known as functional units (Doolittle, 1995). It is observed that the same kind of domain can be contained in distinct proteins. Several computational methods that make use of domain combinations have been developed for prediction of interacting proteins (Hayashida et al., 2011), identification of small protein complexes (Ruan et al., 2013), analysis of the scale-free behavior of protein-protein interaction networks (Nacher et al., 2009) among others. As an evolutionary model of domain combinations in a proteome, a model with mutation and duplication of domains was proposed (Nacher et al., 2006). They defined a specific network, called protein domain net-

Hayashida, M., Koyano, H. and Nacher, J.
Measuring the Similarity of Proteomes using Grammar-based Compression via Domain Combinations.
DOI: 10.5220/0008913101170122
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS*, pages 117-122
ISBN: 978-989-758-398-8; ISSN: 2184-4305

work, constructed from domain combinations of proteins in an organism. In the network, a node represents a protein, and an edge is added if two proteins corresponding to the two nodes of the edge have the same domain. Then, it was reported that the degree distribution of the protein domain network generated by their evolutionary model had the same tendency as those by actual proteomes such as *H. sapiens* and *M. musculus* obtained from Pfam and InterPro domain databases (El-Gebali et al., 2018; Mitchell et al., 2018).

From the viewpoint of information theory, a similarity measure between data structures can be derived. The Kolmogorov complexity of an object was defined as the length of a shortest program outputs the object (Li and Vitanyi, 1997), which cannot be computed realistically, and has been approximated by the compressed size. The conditional Kolmogorov complexity of an object relative to another object was similarly defined, which means the length of a shortest program outputs an object using another object. If two objects are very similar, the conditional Kolmogorov complexity becomes small. The normalized compression distance (NCD) was proposed using the conditional Kolmogorov complexity (Li et al., 2004). They compressed several whole mitochondrial DNA sequences using string compressors such as bzip2 and GenCompress (Chen et al., 2001), and constructed a hierarchical tree based on the distance. In addition to NCD, the universal compression distance (UCD) and compression distance (CD) were used for classifying biological sequences, structures, and networks (Ferragina et al., 2007; Hayashida and Akutsu, 2010).

For the purpose of finding the genetic entropy that an individual organism contains, a compression algorithm based on evolutionary processes was developed (Hayashida et al., 2014). In their method, a proteome, that is, whole kinds of proteins in an organism was regarded as a family of sets of domains, and a grammar on sets was introduced based on evolutionary processes such as mutation, gene duplication, and gene fusion for compressing proteomes. In reality, domains, however, are lined in an adequate order in a protein. Hence, the modified compression algorithm was developed, where a protein was regarded as a sequence of domains (Hayashida et al., 2018). In this paper, we propose a similarity measure based on the modified grammar-based compression, and apply it to hierarchical clustering of seven organisms, *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, and *E. coli*. The results suggest that our similarity measure could classify the organisms very well.

2 METHODS

We briefly review the modified grammar-based compression of a proteome, and a similarity measure based on the Kolmogorov complexity, and explain our approach for measuring a similarity between proteomes using the grammar-based compression.

2.1 Grammar-based Compression

Let $\mathcal{D}$ and $\mathcal{P}$ be a set of domains and a set of proteins, respectively. We regard a protein $P_i (\in \mathcal{P})$ to be a sequence of domains $D_j (\in \mathcal{D})$. Then, for all $D_j \in \mathcal{D}$, P_i exists such that D_j is included in P_i. We consider the following problem: Given $\mathcal{P}$ and $\mathcal{D}$, find a minimum grammar $\mathcal{G}$ with two types R_m and R_d of production rules constructing all proteins $P_i \in \mathcal{P}$ from domains $D_j (\in \mathcal{D})$, where the size of the grammar $\mathcal{G}$ is defined by the sum of costs of all production rules of $\mathcal{G}$, R_m and R_d correspond to evolutionary processes, mutation and gene duplication, respectively.

In a production rule of R_m for a protein P_i, P_i is constructed from only domains $D_{i_j} (\in \mathcal{D})$ $(j = 1, \ldots, |P_i|)$, where $|P_i|$ denotes the length of sequence P_i. Then, the production rule is written as $P_i \leftarrow D_{i_1} \cdots D_{i_{|P_i|}}$. The cost for R_m is defined by

$$cost_{R_m}(P_i) = \lceil \log |\mathcal{D}| \rceil |P_i|, \qquad (1)$$

where $\lceil x \rceil$ denotes the ceiling function that returns the least integer greater than or equal to x, the base of the logarithm is two, and $\lceil \log |\mathcal{D}| \rceil$ means the amount in bits to specify one domain.

In a production rule of R_d for a protein P_i, P_i is constructed by duplicating another protein P_j. In the duplication, domains contained in P_j are duplicated, and several domains can be inserted and deleted. We calculate the Levenshtein distance (Levenshtein, 1965) for finding the minimum number of edit operations, insertion and deletion from P_j to P_i. Then, the cost for R_d is defined by

$$\begin{aligned} cost_{R_d}(P_i, P_j) = {} & \lceil \log |\mathcal{P}| \rceil + |P_i| \\ & + d_L(P_j, P_i)(\lceil \log |\mathcal{D}| \rceil + \lceil \log(|P_j| + 1) \rceil), \end{aligned} \qquad (2)$$

where $d_L(P_j, P_i)$ denotes the Levenshtein distance from P_j to P_i with insertion cost one, and $\lceil \log |\mathcal{P}| \rceil$ means the amount in bits to specify one protein to be duplicated. For example, Figure 1 shows an example of finding the Levenshtein distance from $P_1 = D_1 D_2 D_1$ to $P_2 = D_1 D_2 D_3 D_1$. D_3 is inserted to P_1. Then, $d_L(P_1, P_2) = 1$. P_1 has $|P_1| + 1 = 4$ candidate positions to insert domains, and D_3 is inserted in between the second domain and the third domain of P_1. Then, $cost_{R_d}(P_2, P_1) = \lceil \log |\mathcal{P}| \rceil + 4 + 1 \cdot (\lceil \log |\mathcal{D}| \rceil + \lceil \log 4 \rceil)$.

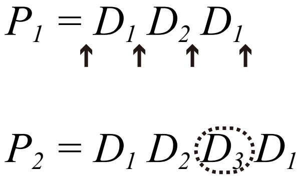

Figure 1: Example of finding the Levenshtein distance from $P_1 = D_1D_2D_1$ to $P_2 = D_1D_2D_3D_1$. In this example, $d_L(P_1, P_2) = 1$.

For all $P_i \in \mathcal{P}$, exactly one production rule is selected for each protein if the size of grammar $\mathcal{G}$ is the smallest. The problem of finding the minimum grammar $\mathcal{G}$ for a proteome $\mathcal{P}$ with domains $\mathcal{D}$ can be transformed into the minimum spanning tree problem for an edge-weighted directed graph $G(V, E, w)$ with a set V of vertices, a set E of edges, and edge weight $w(e)$ of $e(\in E)$ as follows. Suppose that v_0 is a special vertex representing a protein without any domain, and v_i is corresponding to a protein P_i. Then, $V = \{v_0\} \cup \{v_i | P_i \in \mathcal{P}\}$, $E = \{(v_0, v_i) | P_i \in \mathcal{P}\} \cup \{(v_i, v_j) | P_i, P_j \in \mathcal{P}\}$, and $w(v_0, v_i) = cost_{R_m}(P_i)$, $w(v_i, v_j) = cost_{R_d}(P_j, P_i)$. The minimum spanning tree problem can be solved in polynomial time, and at least one edge for each protein belongs to the minimum spanning tree. From the solution, the production rule for P_i in the minimum grammar, and the compressed size $C(\mathcal{P})$ can be obtained.

$$\begin{aligned} C(\mathcal{P}) &= \sum_{P_i \in \mathcal{P}} \{\delta_{P_i}^{R_m} cost_{R_m}(P_i) \\ &\quad + \sum_{P_j \in \mathcal{P}} \delta_{P_i, P_j}^{R_d} cost_{R_d}(P_i, P_j)\}, \end{aligned} \tag{3}$$

where $\delta_{P_i}^{R_m} = 1$ if a mutation-type production rule is selected to P_i in the optimal solution, otherwise 0, and $\delta_{P_i, P_j}^{R_d} = 1$ if a duplication-type production rule from P_j is selected, otherwise 0. The uncompressed size is calculated by $\delta_{P_i}^{R_m} = 1, \delta_{P_i, P_j}^{R_d} = 0$ for all P_i.

2.2 Similarity Measure

We can compress a proteome $\mathcal{P}$ by finding the minimum grammar as mentioned in the previous section. The compressed size is the size of the minimum grammar. In general, the conditional Kolmogorov complexity $K(o_i | o_j)$ of an object o_i given another object o_j is defined as the size of a minimum program that takes o_j and returns o_i (Li and Vitanyi, 1997). If o_i is similar to o_j, $K(o_i | o_j)$ becomes small. The normalized information distance is defined by

$$\frac{\max\{K(o_i | o_j), K(o_j | o_i)\}}{\max\{K(o_i), K(o_j)\}}, \tag{4}$$

where $K(o_i)$ is the Kolmogorov complexity defined as $K(o_i | \varepsilon)$ given no objects. Since $K(o_i)$ and $K(o_i | o_j)$ are not computable, the size $C(o_i)$ compressed by a compressor is used for $K(o_i)$, and $K(o_i | o_j)$ is approximated by $C(o_i \cdot o_j) - C(o_j)$, where $o_i \cdot o_j$ means a concatenation of o_i and o_j. Thus, the universal compression distance (UCD) is defined by

$$\begin{aligned} & UCD(o_i, o_j) \\ &= \frac{\max\{C(o_i \cdot o_j) - C(o_j), C(o_j \cdot o_i) - C(o_i)\}}{\max\{C(o_i), C(o_j)\}}. \end{aligned} \tag{5}$$

2.3 Our Compression Approach

For our purpose of measuring the similarity of proteomes, we introduce the sizes of the minimum grammars for two proteomes $\mathcal{P}_i$ and $\mathcal{P}_j$ as the compressed sizes $C(\mathcal{P}_i)$ and $C(\mathcal{P}_j)$, respectively. Since $\mathcal{P}_i \cup \mathcal{P}_j = \mathcal{P}_j \cup \mathcal{P}_i$, substituting $C(\mathcal{P}_i \cup \mathcal{P}_j)$ to Eq.(5), we have the distance between $\mathcal{P}_i$ and $\mathcal{P}_j$ as

$$d(\mathcal{P}_i, \mathcal{P}_j) = \frac{C(\mathcal{P}_i \cup \mathcal{P}_j) - \min\{C(\mathcal{P}_i), C(\mathcal{P}_j)\}}{\max\{C(\mathcal{P}_i), C(\mathcal{P}_j)\}}. \tag{6}$$

It is noted that $\lceil \log |\mathcal{P}_i \cup \mathcal{P}_j| \rceil$ can be different from $\lceil \log |\mathcal{P}_i| \rceil$ or $\lceil \log |\mathcal{P}_j| \rceil$ in Eqs (1) and (2). Hence, we calculate $C(\mathcal{P}_i)$, $C(\mathcal{P}_j)$ and $C(\mathcal{P}_i \cup \mathcal{P}_j)$ using $\lceil \log |\mathcal{P}_i \cup \mathcal{P}_j| \rceil$ and $\lceil \log |\mathcal{D}_i \cup \mathcal{D}_j| \rceil$ instead of $\lceil \log |\mathcal{P}_i| \rceil$, $\lceil \log |\mathcal{D}_i| \rceil$ and so on, where $\mathcal{D}_i$ and $\mathcal{D}_j$ denote sets of domains included in $\mathcal{P}_i$ and $\mathcal{P}_j$, respectively.

3 COMPUTATIONAL EXPERIMENTS

As protein domains, we used ProRule entries (Sigrist et al., 2005) included in UniProt database (release 2019_03) (The UniProt Consortium, 2019), which is a set of manually created rules concerning domains identified by PROSITE motifs, and contains the position of structurally and functionally critical amino acids. The PROSITE database uses two kinds of descriptors, patterns and profiles, to detect conserved regions. For biologically significant, highly conserved regions such as enzyme catalytic sites and regions involved in binding a metal ion, patterns or regular expressions are used. For other motifs, profiles that are represented by tables of position-specific amino acid weights and gap costs are used.

For seven organisms, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Escherichia coli*, we got ProRule identifiers and positions for each protein, and calculated the compressed sizes, $C(\mathcal{P}_i)$ and $C(\mathcal{P}_i \cup \mathcal{P}_j)$, for each proteome and pairs of proteomes. Then, the minimum spanning tree problem for the generated graph $G(V,E,w)$ was solved using Edmonds's optimum branching algorithm (Tarjan, 1977), which runs in time $O(E\log V)$ for sparse graphs and $O(V^2)$ for dense graphs. We performed hierarchical clustering using hclust function in R statistics software (https://www.r-project.org). In the clustering, according to the distance $D(X,Y)$ between clusters X and Y, two clusters with the smallest distance are merged into one cluster. For single linkage clustering, the distance is defined by

$$D(X,Y) = \min_{i\in X, j\in Y} d(\mathcal{P}_i, \mathcal{P}_j). \tag{7}$$

For complete linkage clustering, the distance is defined by

$$D(X,Y) = \max_{i\in X, j\in Y} d(\mathcal{P}_i, \mathcal{P}_j). \tag{8}$$

For the unweighted pair group method with arithmetic mean (UPGMA) that is used in phylogenetic analyses, the distance is defined by

$$D(X,Y) = \frac{\sum_{i\in X, j\in Y} d(\mathcal{P}_i, \mathcal{P}_j)}{|X||Y|}. \tag{9}$$

4 RESULTS

Table 1 shows the results on the number $|\mathcal{P}|$ of proteins, the number $|\mathcal{D}|$ of domains, the uncompressed, compressed sizes, and the compression ratio for single proteomes of seven organisms, *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, and *E. coli*, where the uncompressed size means the sum of $cost_{R_m}(P_i)$ for all $P_i \in \mathcal{P}$, that is, every protein is represented by a domain sequence. It is confirmed that the ratio of the compressed size to the uncompressed size for higher organisms was smaller than that for others. It means that higher organisms use gene duplication more frequently.

Table 2 shows the results on the compressed size $C(\mathcal{P}_i \cup \mathcal{P}_j)$ between proteomes of the seven organisms. From this table, the distances $d(\mathcal{P}_i, \mathcal{P}_j)$ between proteomes were calculated. Figures 2, 3, and 4 show the results on the dendrogram using the single linkage, complete linkage, and UPGMA clustering, respectively, for proteomes of the seven organisms. The structure of the hierarchical tree by the single linkage clustering was the same as that by the UPGMA and generally known phylogenetic trees, and was slightly different from that by the complete linkage clustering.

Table 3 shows the results on the rate of the number of duplication-type production rules including two proteins from distinct organisms to that from one organism for each pair of proteomes of the seven organisms. The rates in between *H. sapiens* and *M. musculus*, and between *D. melanogaster* and *C. elegans* were over a half. It means that many orthologous proteins exist in the organisms. For example, the number of rules including two proteins from distinct organisms in between *H. sapiens* and *M. musculus* was 2,970. Among those, for generation of ACACA_HUMAN protein in UniProt identifier, the rule that duplicates ACACA_MOUSE protein was selected as an optimal solution, where both proteins contained ProRule identifiers of PRU00409, PRU01066, PRU01136, and PRU01137, and are known as acetyl-CoA carboxylases. Conversely, the rate in between *E. coli* and another organism was small. As a reason, it is also considered that the num-

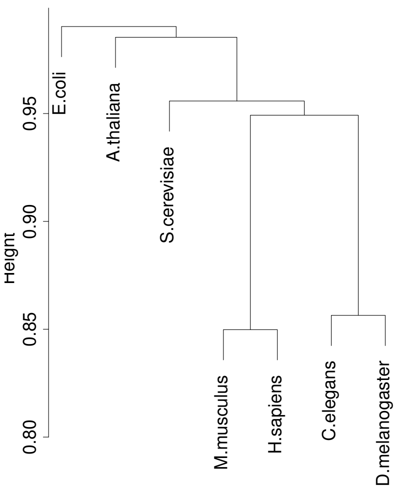

Figure 2: Result on the dendrogram using single linkage clustering for proteomes of the seven organisms.

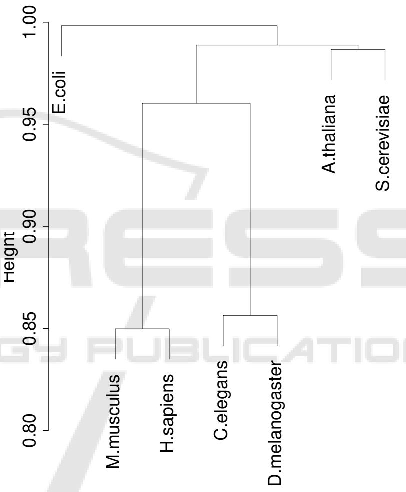

Figure 3: Result on the dendrogram using complete linkage clustering for proteomes of the seven organisms.

Table 1: Results on the number of proteins, the number of domains, the uncompressed, compressed sizes, and the compression ratio for single proteomes of the seven organisms.

organism	# proteins	# domains	uncompressed (A)	compressed (B)	B/A
H. sapiens	7292	666	164290	114564	0.697
M. musculus	6124	665	138330	97925	0.708
D. melanogaster	1105	415	21609	17560	0.813
C. elegans	1318	416	23985	20466	0.853
S. cerevisiae	1337	372	17217	15481	0.899
A. thaliana	5213	426	69444	59313	0.854
E. coli	896	309	11115	10045	0.904

Table 2: Results on the compressed size between proteomes of the seven organisms.

	H. sapiens	*M. musculus*	*D. melanogaster*	*C. elegans*	*S. cerevisiae*	*A. thaliana*
M. musculus	199674	–	–	–	–	–
D. melanogaster	132262	112699	–	–	–	–
C. elegans	135310	115765	35613	–	–	–
S. cerevisiae	133241	113703	35511	38812	–	–
A. thaliana	180512	163583	82244	85653	74467	–
E. coli	125742	109104	30061	33595	28153	74683

Table 3: Results on the rate of the number of duplication-type production rules including two proteins from distinct organisms to that from one organism for each pair of proteomes of the seven organisms.

	H. sapiens	*M. musculus*	*D. melanogaster*	*C. elegans*	*S. cerevisiae*	*A. thaliana*
M. musculus	0.57	–	–	–	–	–
D. melanogaster	0.26	0.22	–	–	–	–
C. elegans	0.25	0.27	0.53	–	–	–
S. cerevisiae	0.17	0.18	0.36	0.30	–	–
A. thaliana	0.18	0.18	0.15	0.19	0.27	–
E. coli	0.025	0.026	0.062	0.068	0.14	0.059

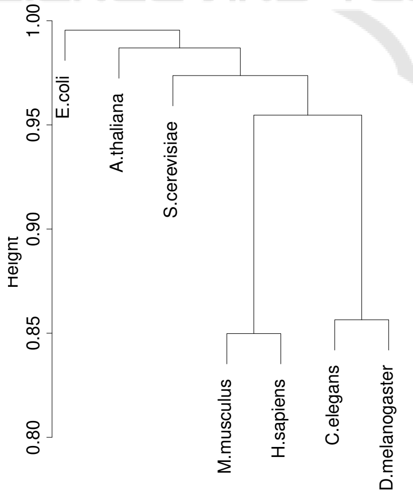

Figure 4: Result on the dendrogram using the unweighted pair group method with arithmetic mean (UPGMA) for proteomes of the seven organisms.

ber of proteins in *E. coli* was small. For example, the number of rules including two proteins from distinct organisms in between *E. coli* and *H. sapiens* was 67. Among those, for generation of ODP2_HUMAN protein, the rule that duplicates ODP2_ECOLI protein was selected, where ODP2_HUMAN contained two domains identified by PRU01066 and one domain by PRU01170, and ODP2_ECOLI contained three domains identified by PRU01066 and one domain by PRU01170. ODP2_HUMAN and ODP2_ECOLI are known as dihydrolipoyllysine-residue acetyltransferase components of pyruvate dehydrogenase complex. The production rule from ODP2_ECOLI to ODP2_HUMAN, rather than the rule from ODP2_HUMAN to ODP2_ECOLI, was selected because the cost of insertion of PRU01066 is larger than that of deletion of the domain.

5 CONCLUSIONS

We proposed a similarity measure based on the grammar-based compression for proteomes with sets of domain sequences, and applied it to hierarchical clustering of seven organisms, *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, and *E. coli*. The results suggest that our similarity measure could classify the organisms very well. As future work, we would like to analyze more organisms, to find the minimum grammar for generating proteomes of more organisms, and to investigate comprehensive evolutionary processes.

ACKNOWLEDGEMENTS

This work was partially supported by JSPS KAKENHI Grant Number JP19K12228.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Chen, X., Kwong, S., and Li, M. (2001). A compression algorithm for dna sequences. *IEEE Engineering in Medicine and Biology Magazine*, 20(4):61–66.

Doolittle, R. (1995). The multiplicity of domains in proteins. *Annual Review of Biochemistry*, 64:287–314.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., and Valiente, G. (2007). Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, 8:252.

Hayashida, M. and Akutsu, T. (2010). Comparing biological networks via graph compression. *BMC Systems Biology*, 4(Suppl. 2):S13.

Hayashida, M., Ishibashi, K., and Koyano, H. (2018). Analyzing order of domains in grammar-based compression of proteomes. In *The 24th International Conference on Parallel and Distributed Processing Techniques and Applications*, pages 278–281. CSREA Press.

Hayashida, M., Kamada, M., Song, J., and Akutsu, T. (2011). Conditional random field approach to prediction of protein-protein interactions using domain information. *BMC Systems Biology*, 5(Suppl. 1):S8.

Hayashida, M., Ruan, P., and Akutsu, T. (2014). Proteome compression via protein domain compositions. *Methods*, 67:380–385.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Adademii Nauk SSSR*, 163(4):845–848.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P. (2004). The similarity metric. *IEEE Transactions on Information Theory*, 50:3250–3264.

Li, M. and Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York.

Lipman, D. and Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C., Yong, S.-Y., and Finn, R. D. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360.

Nacher, J. C., Hayashida, M., and Akutsu, T. (2006). Protein domain networks: Scale-free mixing of positive and negative exponents. *Physica A*, 367:538–552.

Nacher, J. C., Hayashida, M., and Akutsu, T. (2009). Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs. *BioSystems*, 95:155–159.

Nicolae, M., Pathak, S., and Rajasekaran, S. (2015). LFWC: a lossless compression algorithm for FASTQ files. *Bioinformatics*, 31(20):3276–3281.

Pinho, A., Pratas, D., and Garcia, S. (2012). GReEn: a tool for efficient compression of genome resequencing data. *Nucleic Acids Research*, 40(4):e27.

Ruan, P., Hayashida, M., Maruyama, O., and Akutsu, T. (2013). Prediction of heterodimeric protein complexes from weighted protein-protein interaction networks using novel features and kernel functions. *PLoS ONE*, 8(6):e65265.

Sigrist, C. J. A., De Castro, E., Langendijk-Genevaux, P. S., Le Saux, V., Bairoch, A., and Hulo, N. (2005). ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics*, 21(21):4060–4066.

Tarjan, R. (1977). Finding optimum branchings. *Networks*, 7:25–35.

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47:D506–D515.

Woese, C. and Fox, G. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA*, 74(11):5088–5090.