REVIEW

Protein structure and function : A review with application

Hyejeong Cheon

Correspondence:
hyejeonc@stud.ntnu.no
Department of physics, NTNU, Trondheim , Norway
Full list of author information is available at the end of the article

Abstract

Protein is a main biomolecule involved in a metabolism of organisms, and understanding its function is important for biomedical applications. The structure of a protein is based on different levels of organization, such as the tertiary and quaternary structures with specific functions and domains. On the other hand, data of the structure and function of a protein have various uses in bioinformatics, such as predicting its function through its primary structure. In this paper, the structure of proteins and a bioinformatical application of it are reviewed.

Keywords: Protein; Structural biology; Hidden Markov Model; Secondary structure; Bioinformatics

Introduction

The various metabolic processes needed in living organisms are programmed by the genome, and they are mostly conducted by kinds of proteins which interact with each other [1, 2]. Therefore, the knowledge of the protein functions and how these are regulated is important for understanding biological mechanisms and also for medical purposes [3, 4]. Proteins are large molecules composed of one or more chains of amino acids arranged in a specific order [5]. Amino acids are linked together in a linear configuration by peptide bonds, and such peptide sequence defines its primary structure. Secondary structure is defined as locally assembled structures that form within a polypeptide as a result of interactions between the peptide backbones. The 3D structure of a polypeptide is called its tertiary structure, and it is mainly due to interactions between side chains of the amino acids making up the polypeptide. Quaternary structure refers to functional protein complexes that involve more than one polypeptide chain, for example, hemoglobin and ion channels [6, 7].

The structural research of proteins has developed based on the fact that the structure of polypeptides directly affect the function of proteins [1]. The improvement of crystallographic and spectroscopic methods, including XRD (X-ray diffraction), AFM (Atomic force microscopy), NMR (Nuclear magnetic resonance) and cryo-EM (Cryogenic electron microscopy) allows the structural research progress [8, 9, 10, 11] and the data on protein structure and function have been accumulated in the Protein Data Bank (PDB) [12, 13]. Based on the PDB, analyzing the structure of proteins and studying their functions from the point of view of informatics also have been derived; therefore, the structure of proteins has become an important research field for understanding and predicting biological processes.

In the first part of this paper, I review the structure of proteins from the basic units to high-dimensional structures with examples of various proteins and their function.

The second part gives a brief overview of a bioinformatic algorithm related to the protein structure.

Protein structure and function

Amino acids

Amino acids are the basic building blocks of proteins that make up the proteome of living organisms. Amino acids refer to molecules which contain both a carboxyl (-COOH) and an amine (-NH_2) functional group. That is, amino acids are present as zwitterions in aqueous solution on neutral pH condition [1, 14].

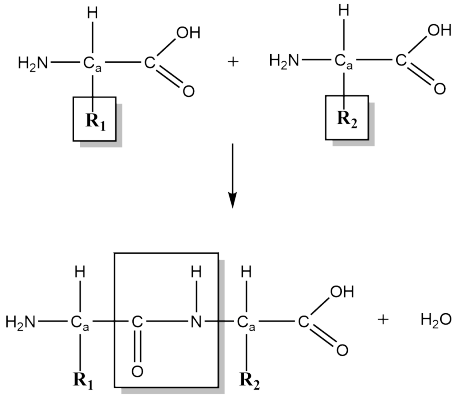

Figure 1 Structure of amino acids and formation of a peptide bond by two amino acids.

Table 1 Table of 20 amino acids and their properties [15].

Amino acid	3-Letter name	Residue (R_n) property
Hydrophobic, nonpolar amino acids		
Alanine	Ala (A)	Hydrocarbon chain
Cysteine	Cys (C)	Sulfhydryl (-SH) group
Glycine	Gly (G)	One hydrogen atom (small side chain)
Isoleucine	Ile (I)	Hydrocarbon chain
Leucine	Leu (L)	Hydrocarbon chain
Methionine	Met (M)	Hydrocarbon chain containing sulfur
Phenylalanine	Phe (F)	Aromatic group
Proline	Pro (P)	Cyclic ring
Tryptophan	Trp (W)	Large, aromatic side chain
Valine	Val (V)	Hydrocarbon chain
Hydrophilic, polar amino acids		
Glutamine	Gln (Q)	-
Asparagine	Asn (N)	-
Serine	Ser (S)	Hydroxyl group (-OH) which forms a hydrogen bond
Threonine	Thr (T)	Hydroxyl group (-OH) which forms a hydrogen bond
Tyrosine	Tyr (Y)	Phenolic hydroxyl group forming hydrogen bond
Hydrophilic, poitively charged amino acids		
Arginine	Arg (R)	-
Histidine	His (H)	Imidazole side chain
Lysine	Lys (K)	-
Hydrophilic, negatively charged amino acids		
Aspartic acid	Asp (D)	-
Glutamic acid	Glu (E)	-

The chemical structure of an amino acid with side chain R_1 and R_2 is described in Figure 1, as $H_2NC_\alpha HR_nCOOH$. A central carbon atom (C_α) in the amino acids is combined with four different groups: a carboxyl group, an amine group, a hydrogen

atom and a side chain group (Residue, R_n). In this structure, Rn determines a nature of each amino acid. The property of an amino acid is classified as hydrophilic, hydrophobic, polar, nonpolar, charged (acidic/basic) or uncharged as indicated in Table 1.

Hydrophilic amino acids possess an ionized or polarized side chain. Among hydrophilic amino acids, Arginine, Histidine and Lysine are positively charged, and Aspartic acid and Glutamic acid are negatively charged. Asparagine, Glutamine, Cysteine, Serine and Threonine have uncharged but polar side chains.

Hydrophobic amino acids have side chains which have low water solubility. Phenylalanine, Tyrosine and Tryptophan are examples of amino acids with side chains containing an aromatic group. Alanine, Valine, Isoleucine, Leucine and Methionine have a hydrocarbon side chain.

Cystein, Glycine, and Proline have unique properties due to their side chains. Cystein has a side chain containing a sulfhydryl group (-SH). When the two sulfhydryl groups are oxidized, they form a disulfide bond (-S-), which allows a polypeptide to stabilize its conformation folded [16]. Glycine is the smallest amino acid and it can fill a small space which no other amino acids can fit in. That is, a polypeptide with a high proportion of Glycine is very flexible [1]. If Glycine is mutated, the structure of the protein is likely to change because it is difficult that other amino acids fill this space instead of Glycine. Strictly speaking, Proline is an imino acid, but it is called an amino acid generally. Proline has a cyclic side chain connected to the β carbon. This distinctive side chain gives a peptide conformational rigidity [17].

Two or more amino acids can be linked by a peptide bond. A short amino acid chain (2-50 amino acids) is referred to as a peptide, and a long chain of amino acids (>50 amino acids) are called a polypeptide [18]. The carboxyl group of an amino acid forms a peptide bond to the amine group of the adjacent amino acid through condensation reaction (See Figure 1). A peptide has a directionality with one end of the amine group as N-terminus and the other end of the carboxyl group as C-terminus. The linear structure of a polypeptide based on the sequence of amino acids from the N-terminus to the C-terminus is called the primary structure.

Secondary structure

Proteins have higher order structure based on the primary structure, which is called the secondary structure. The main components of secondary structure are the α-helix and the β-sheet. The α-helix is the most identifiable structure which is suggested by Pauling and co-workers in 1951 [19]. Generally, one turn of the α-helix contains 3.6 amino acids in average and the nth oxygen of the C=O group forms a hydrogen bond with the hydrogen of the $(n+4)$th NH group. Therefore, except the first NH (N-terminus) and the last C=O (C-terminus), all NH and C=O form hydrogen bonds in the α-helix structure.

The β-sheet is another secondary structure that is observed frequently. The β-sheet consists of several strands formed by 5-10 amino acids. Amino acids in one strand form hydrogen bonds with amino acids in the neighbor strand. When two neighbor strands are in the same direction, they are called a parallel β-sheet, and an antiparallel β-sheet denominates strands in the opposite direction. Furthermore, there is a complex structure called the Rosman fold which refers to a mixed structure of parallel and antiparallel sheets [20].

Unlike the two mentioned structures, the loop (turn) structure has various lengths and unspecified shapes. A protein, from a macro perspective, is regarded as a combination of secondary structures. The loop is a structure that connects secondary structures. In other words, unlike the α-helix or β-sheet, amino acids in the loop form hydrogen bonds with surrounding molecules including water, not the amino acids in the peptide itself.

The secondary structure of a protein is apparently related to its primary structure. Some sequences of amino acids tend to form a specific secondary structure. For example, Proline is not available to form a hydrogen bond with others because of its cyclic side chain, therefore, it is rarely observed within the α-helix except the starting and ending point. It also makes a kink on a polypeptide chain, so it is often observed at the starting point of the α-helix. As discussed, Glycine is the common amino acids providing high flexibility on a chain, so it is observed in the loop region [1, 21].

Supersecondary structure (motif)

A combination of described secondary structures can form a local sub-structure which is called a motif. Some motifs perform functions but there are motifs that provide stability on a polypeptide chain without special functions. An example of a particular function is DNA binding, and the DNA-binding motif is frequently found in transcription factors, that is, proteins with gene regulation function [22].

One of the simplest motifs, containing two α-helices and one loop is the so-called Helix-Turn-Helix (HTH). This motif binds to DNA and Ca^{2+} ion, and the motif which binds to DNA is sometimes recognized as Helix-Loop-Helix (HLH) motif. Ribosomal protein R7 is one of examples which contain the HTH motif [23].

Other examples are the β-hairpin motif, composed of two antiparallel β-strands, and the Greek key motif with four antiparallel β-strands. Both of these maintain very stable structure. For example, the β-hairpin motif can be found in toxins such as Erabutoxin, a snake venom. This explains the high stability of the toxins when released in a different environment than the one where it was produced [24].

Some motifs are formed with specific amino acids. The zinc finger (ZnF) is a small motif for forming coordination complex of Zn^{2+} ion(s) with Cysteine and Histidine. The classic ZnF motif (Cys_2His_2) is composed of two Cysteines and two Histidines with a β-hairpin and a α-helix. There are some other cases of ZnF forming a coordinate bond with four or more Cysteines. The ZnF has DNA/RNA/protein-binding properties as shown in one example of transcription factor IIIA (TFIIIA) of Arabidopsis thaliana (Mouse-ear cress) [25, 26, 27, 28, 29].

Two α-helices with Leucine resulting in the formation of a dimer, is called Leucine zipper. Since Leucine is highly hydrophobic, two Leucines bind by strong hydrophobic interaction. The two same helices create a homodimeric Leucine zipper (see Figure 2). The part that does not participate in the formation of dimer binds to the gene such as DNA like a tong [31, 32].

Tertiary structure

The three-dimensional structure that a polypeptide chain adopts is called the tertiary structure. The tertiary structure may be formed by folded secondary structures

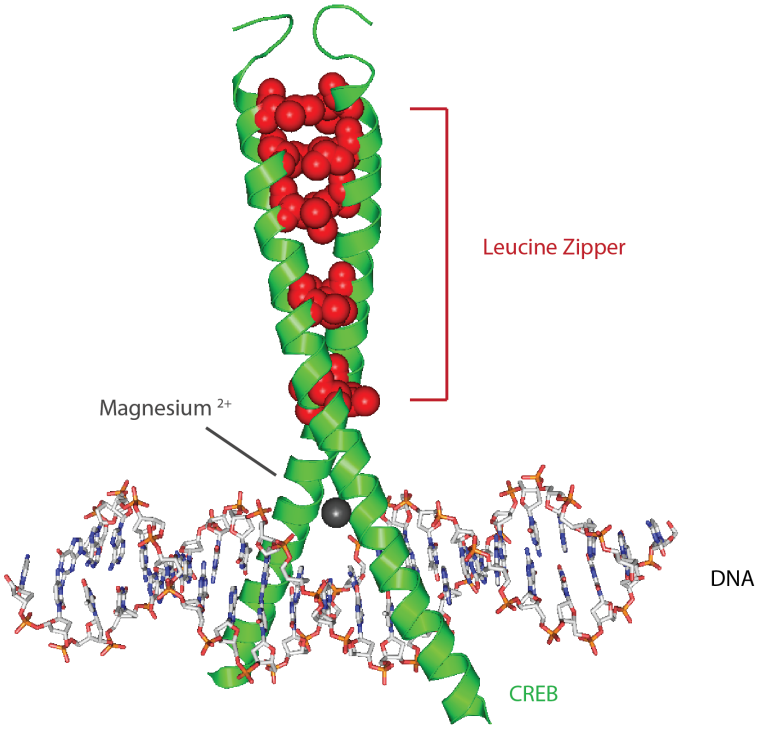

Figure 2 A scheme of Leucine zipper binding to DNA. Red compounds are Leucines. The image is reused with permission [30].

as a result of interactions among side chains of amino acids. The resulting structure of a polypeptide can be globular or fibrous. Protein folding typically resulted from one or more intramolecular attractions which may be a hydrogen bond, electrostatic interaction, hydrophobic interaction, a disulfide bond due to Cysteines and Van der Waals interaction. The secondary structure appears from a hydrogen bond between peptide bonds (C=O - HN) but the tertiary structure appears due to the formation of hydrogen bonds among side chains (R_n-R_m).

A functional unit called the domain made by folding is the main characteristic of tertiary structure. A polypeptide can have more than one domain. It is difficult to distinguish clearly between a domain and a motif, but a domain is characterized with a specific function. Therefore, a domain is often named according to its function such as DNA-binding domain or membrane-binding domain. A domain is more stable than a motif and one or more motifs can be a part of a domain.

Domains are classified by the combination of secondary structures, such as the α domain, the β domain, the $\alpha+\beta$ domain, and the α/β domain. The α domain consists of only α-helices (*i.e.* λ suppressor) and the β domain consists of only β-sheets (*i.e.* γ-crystalline). The $\alpha+\beta$ domain refers to a domain where the α-helix is located on one side and the β-sheet is on the other (*i.e.* Src oncoprotein). The α/β domain refers to a domain where a β-sheet forms a core and α-helices surround it (*i.e.* Tyoredoxin) [33, 34, 35].

Quaternary structure

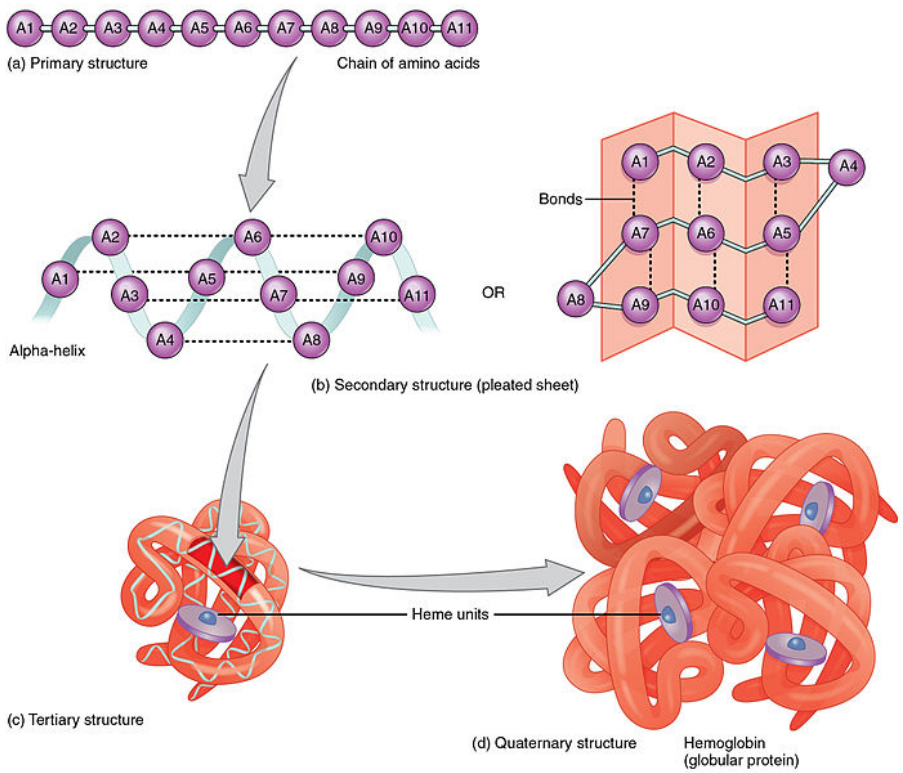

Figure 3 The primary (a), secondary (b), tertiary (c) and quaternary (d) structure of hemoglobin. The image is reused with permission [36].

The primary, secondary and tertiary structures are considered within a polypeptide but the quaternary structure deals with multiple polypeptide chains. That is, a protein possesses a quaternary structure when its functional form consists of two or more polypeptide chains, associated via non-covalent bond including hydrogen bond, electrostatic interaction and hydrophobic interaction.

One polypeptide chain participating in a formation of the quaternary structure is called a subunit. The quaternary structure composed of identical subunits is called a homogeneous quaternary structure, and the quaternary structure composed of non-identical subunits is called a heterogeneous quaternary structure. As examples of quaternary structure, HIV-1 protease dimer consists of two identical subunits (homogeneous quaternary structure) and hemoglobin consists of two α subunits and two β subunits as $\alpha_2\beta_2$ [1, 37].

A protein with quaternary structure has the advantage of enabling allosteric regulation for enzymatic activities by conformational change. Hemoglobin, for example, allows one heme in each subunit to combine to one oxygen molecule. When an oxygen molecule binds to a heme, the oxygen-bound subunit deforms and enhances the oxygen affinity of the other heme-subunit complexes [38]. The primary, secondary, tertiary and quaternary structures of h emoglobin are schematically depicted in Figure 3.

Self-assembled structure

The process with which the proteins fold into their secondary, tertiary, and quaternary structures (if present) driven by mentioned interactions, is called self-assembly. Self-assembly can be divided into intramolecular and intermolecular self-assembly depending on whether it involves one or several polypeptides. Forming a structure within a polypeptide, such as folding mentioned in the tertiary structure, is regarded as the intramolecular self-assembly, and the quaternary structure with several different polypeptides is regarded as the intermolecular self-assembly [39]. The intermolecular self-assembly of proteins include both protein-protein interactions and metal-ligand interactions which is mediated by metal ions for forming a large assembly [40].

Self-assembly refers to a natural process of transforming components into an ordered system with reducing disorder in a system [41]. It is a general terminology for systems found in various areas, such as material science, chemistry and molecular/structural biology. The interesting point in self-assembly is that it does not cost energy of the system, in other words, the self-assembly refers a spontaneous process. This leads to a great interest in self-assembly, and it implies self-assembly plays a crucial role in the formation of different types of biomacromolecules including proteins and genetic materials.

Intrinsically disordered structure

The above discussion of protein structure has focused on ordered and folded structures. However, locally or globally unstructured proteins have been observed by SAXS (Small-angle X-ray scattering) and NMR, and some cases have been found to be related to their functions [42]. These proteins are called intrinsically disordered proteins (IDPs, also known as intrinsically unstructured proteins) and the regions that are disordered locally are called intrinsically disordered regions (IDRs). Most eukaryotic cell proteome has both IDRs and structured areas [43].

An IDP has a polypepide chain that is less complex. In particular, bulky hydrophobic amino acids do not tend to be observed, but charged or polar amino acids appear often in IDRs of IDPs. The IDR has the advantage that it is capable of rapid conformational changes. Also, IDP plays an important role in cellular signalling in that it can bind with different target molecules as needed [42].

Some IDPs tranform into structured proteins by folding to bind onto a specific target substrate. This process is called *coupled folding and binding mechanism*. The coupled folding and binding mechanisms include folding processes of the IDP after binding to the target molecules (Induced-fit model) and folding processes before association of IDP to the target (Conformational selection model) [44, 45].

Application of structural analysis

So far we discussed the secondary, tertiary and quaternary structures of proteins based on the primary structure, that is its sequence of amino acids. Study of relationship between a sequence of amino acids and the structure of a protein in this regard is actively undergoing, and prediction algorithms and programs are existent by using abundant database.

A study field of solving problems related to biology with methods of mathematics or information technology is called bioinformatics, in particular, the area of analyzing and predicting the structure of biomacromolecules such as proteins, RNAs and DNAs is called structural bioinformatics.

Various programs can be used as tools for structural bioinformatics, and they can be found from websites such as ExPASy which gathers existent prediction programs for protein secondary structures (PSIPRED, Jpred), membrane helices (MEMSAT-SVM), disordered regions (DISOPRED, GlobPlot), domains (CHOPnet) and motifs (MOTIF) [46, 47, 48, 49]. Most of programs mentioned above are based on machine learning algorithms. In the supervised learning, the program learns by itself from acquired dataset, as input-output pairs. The PDB provides abundant datasets of structural and functional features of protein for machine learning [13].

Hidden Markov Model (HMM)

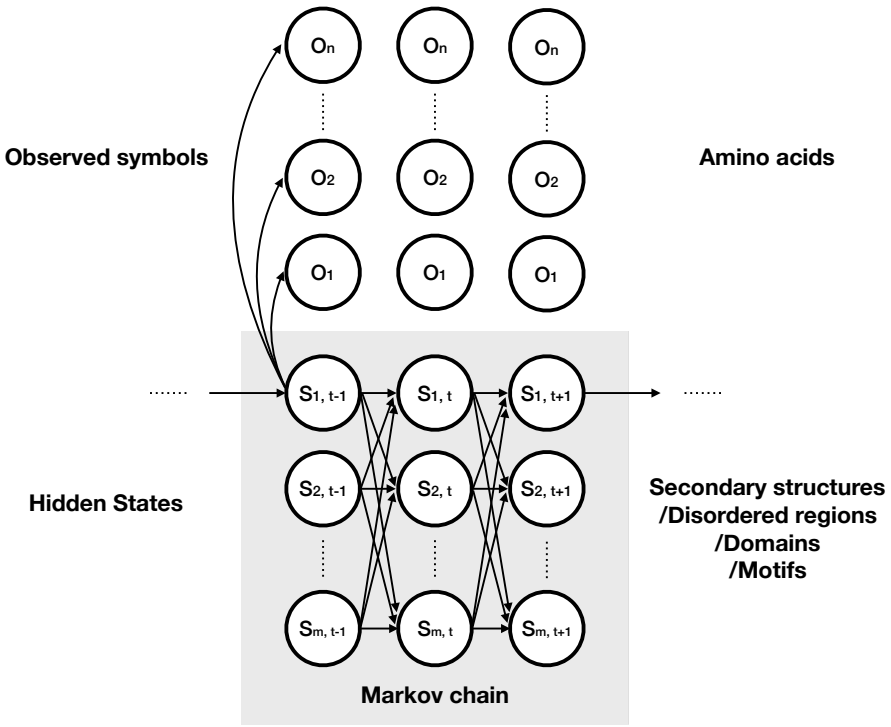

Figure 4 A scheme of Hidden Markov Model.

This section introduces a mathematical model, Hidden Markov Model (HMM), which is an algorithm used for predicting the secondary structure of proteins. Markov model refers to a mathematical model in which a probability of present step depends only on the probability of previous step. For example, we assume that today's weather (q_i) is only affected by yesterday's weather (q_{i-1}), not by the weather before yesterday ($q_{i-2}, q_{i-3}, q_{i-4}$...). The probability of weather today ($P(q_i)$) is to be described by following expression.

$$P(q_i|q_1, q_2, ..., q_{i-1}) = p(q_i|q_{i-1}) \tag{1}$$

The weather can be clear, rainy, or snowy, q regards three states ($m = 3, s_1, s_2, s_3$) and the number of possible states at the tth step is 3. In the the Markov process, the probability of transition from one state (s_i) to another state (s_j), for example, in case of $m = 3$, $a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, a_{31}, a_{32}$ and a_{33} are called the transition probabilities which can be expressed in a $m \times m$ matrix (A). That is, through the transition probability, we can predict the weather at tth step in Markov process.

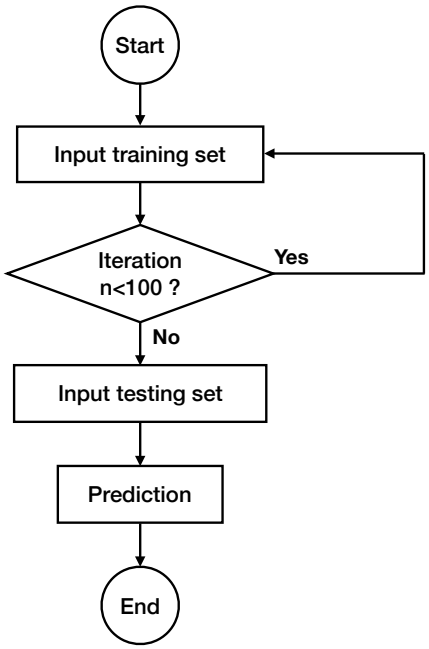

Figure 5 A flow chart of the prediction algorithm.

The HMM follows the basic theory of Markov model, but it is distinguished by the fact that each state is hidden (see Figure 4). For example, when we study the climate in a period, say 10 years ago, but the available information we know is only the consumption of beer in that specific period. If the weather refers to the hidden state, then the daily beer intake per person is called the observable (also known as symbol). The probability that a particular observable (o_j) is observed in a state (s_i) is called the emission probability (b_{ij}). The emission probability can be represented by a $m \times n$ matrix (B) with n, the number of observables. That is, observables (the beer consumption) are used to estimate the hidden state (the weather) at a certain t step (the period) even when the observable is not along the Markov chain.

Thus, datasets of proteins from the PDB allow the HMM to treat states such as secondary structures, IDRs or motifs as hidden states. Updating A and B by reading all state-symbol pairs makes the algorithm learn by itself (supervised learning) in the prediction program.

Prediction of protein secondary structure

Table 2 Parameters used for the prediction algorithm.

Parameters	Training set	Testing set
Population	111	17
Length of protein (min)	26	35
Length of protein (max)	498	461

In this work, following hidden states and symbols were decided to implement the algorithm that predicts the secondary structure of the protein. Three hidden states are α-helix (h), β-sheet (e) and random coil (c), and 20 symbols are 20 amino acids, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. For example, a polypeptide has a primary structure as AADDKKKPTDD, and it has a corresponding secondary structure as ccchhhhcccc. 111 peptides are used as a training set and the results were evaluated from 17 peptides, a testing set [50]. Specific information of the training and testing sets, such as Maximum and minimum length of proteins, are listed in Table 2.

In training process, A and B are updated. Updating is repeated 100 times and this is described as an iteration in Figure 5. After 100 iterations, the testing set is evaluated by counting correctly and incorrectly predicted states. The statistics result from counting them provides accuracy of each structure.

Accuracy is calculated by a percentage ratio of the number of correctly predicted state to total number of states. There are four different accuracies for counting predicted states: Q is the accuracy including all secondary structures, h, e and c. Q_h is an accuracy calculated by a percentage ratio of the number of correctly predicted α-helix to total number of α-helix in polypeptides. Q_e is for β-sheet and Q_c is for random coil.

Table 3 Accuracy of protein secondary structure prediction by HMM. Q_3

-	Q	Q_h	Q_e	Q_c
Accuracy (%)	42.4	2.4	41.3	60.5

The implemented algorithm here shows Q as 42.4%. Especially, Q_h is 2.4% as the lowest, and Q_c is 60.5% as the highest, that is, the accuracy of predicting structure is distinctively different from each other (see Table 3).

Q is not high enough for using this algorithm for proper prediction, but it implies HMM is applicable in other usage such as prediction of domains, IDRs or motifs by using databases of them instead of secondary structures. That is, the prediction algorithm implemented here has a potentiality of various bioinformatical applications in a regard of analysing structure and function of proteins.

Conclusion

The basic structure of a protein is the sequence of amino acids, primary structure, and a protein has secondary, tertiary and quaternary structure with different functional characteristics. The presence and arrangement of certain amino acids or ions are related with the function and regulation of some proteins, and examples of these can be found in various proteins. Furthermore, a mathematical algorithm can be applicable for studying protein structures with a help of database. That is, the association of the structure and function of proteins can be researched with various study areas such as structural biology and bioinformatics.

Acknowledgement and author's information

This paper is aimed to submit a report for a course, FY3490. Related python codes and files can be downloaded from link below.
https://github.com/hyejeonc/protein-review [51]

References

1. Whitford, D.: Proteins: structure and function (2013)
2. Becker, W.M., Kleinsmith, L.J., Hardin, J., Bertoni, G.P.: The world of the cell (2003)
3. Thomas, P., Fenech, M.: A review of genome mutation and alzheimer's disease. Mutagenesis **22**(1), 15–33 (2007)
4. Osterman, A., Overbeek, R.: Missing genes in metabolic pathways: a comparative genomics approach. Current opinion in chemical biology **7**(2), 238–251 (2003)
5. Medical Definition of Protein. https://www.medicinenet.com/script/main/art.asp?articlekey=15380
6. Peters, T.: Proteins: Structure and function. david whitford. chichester, west sussex, england: John wiley & sons ltd., 2005, 542 pp., paperback, 65.00. isbn 0-471-49894-7. Clinical Chemistry **51**(11), 2220–2221 (2005)
7. Orders of protein structure. https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure. Accessed: 2019-08-23
8. Sprang, S.R.: G protein mechanisms: insights from structural analysis. Annual review of biochemistry **66**(1), 639–678 (1997)
9. Wüthrich, K.: Nmr with proteins and nucleic acids. Europhysics News **17**(1), 11–13 (1986)
10. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J.M., Gaub, H.E.: Reversible unfolding of individual titin immunoglobulin domains by afm. science **276**(5315), 1109–1112 (1997)
11. Dubochet, J., Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A.W., Schultz, P.: Cryo-electron microscopy of vitrified specimens. Quarterly reviews of biophysics **21**(2), 129–228 (1988)
12. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide protein data bank. Nature Structural & Molecular Biology **10**(12), 980 (2003)
13. RCBS PDB. http://www.rcbs.org. Accessed: 2019-07-31
14. Edsall, J.T., Blanchard, M.H.: The activity ratio of zwitterions and uncharged molecules in ampholyte solutions. the dissociation constants of amino acid esters. Journal of the American Chemical Society **55**(6), 2337–2353 (1933)
15. Abela, J., Michael, J.: Topics in evolving transformation systems [microform]. (2019)
16. Sælensminde, G., Halskau, Ø., Jonassen, I.: Amino acid contacts in proteins adapted to different temperatures: hydrophobic interactions and surface charges play a key role. Extremophiles **13**(1), 11 (2009)
17. Pakula, A.A., Sauer, R.T.: Genetic analysis of protein stability and function. Annual review of genetics **23**(1), 289–310 (1989)
18. What Is the Difference Between a Peptide and a Protein? https://www.britannica.com/story/what-is-the-difference-between-a-peptide-and-a-protein. Accessed: 2019-07-30
19. Pauling, L., Corey, R.B., Branson, H.R.: The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences **37**(4), 205–211 (1951)
20. Rao, S.T., Rossmann, M.G.: Comparison of super-secondary structures in proteins. Journal of molecular biology **76**(2), 241–256 (1973)
21. https://proteinstructures.com/Structure/Structure/amino-acids.html. https://proteinstructures.com/Structure/Structure/amino-acids.html. Accessed: 2019-07-30
22. Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.-M., Lovell-Badge, R., Goodfellow, P.N.: A gene from the human sex-determining region encodes a protein with homology to a conserved dna-binding motif. Nature **346**(6281), 240 (1990)
23. Hosaka, H., Nakagawa, A., Tanaka, I., Harada, N., Sano, K., Kimura, M., Yao, M., Wakatsuki, S.: Ribosomal protein s7: a new rna-binding motif with structural similarities to a dna architectural factor. Structure **5**(9), 1199–1208 (1997)
24. Branden, C.I., Tooze, J.: Introduction to protein structure (2012)
25. Ryan, R.F., Darby, M.K.: The role of zinc finger linkers in p43 and tfiiia binding to 5s rrna and dna. Nucleic acids research **26**(3), 703–709 (1998)
26. Krishna, S.S., Majumdar, I., Grishin, N.V.: Structural classification of zinc fingers: survey and summary. Nucleic acids research **31**(2), 532–550 (2003)
27. Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A., Wright, P.E.: Three-dimensional solution structure of a single zinc finger dna-binding domain. Science **245**(4918), 635–637 (1989)
28. Brown, R.S., Sander, C., Argos, P.: The primary structure of transcription factor tfiiia has 12 consecutive repeats. FEBS letters **186**(2), 271–274 (1985)
29. Laity, J.H., Lee, B.M., Wright, P.E.: Zinc finger proteins: new insights into structural and functional diversity. Current opinion in structural biology **11**(1), 39–46 (2001)
30. Leucine zipper. https://commons.wikimedia.org/wiki/File:CREB_protein.png. Accessed: 2019-08-23
31. Grigorescu, A.A., Rosenberg, J.M.: Dna sequence recognition by proteins (2004)
32. Pollard, T.D., Earnshaw, W.C., Lippincott-Schwartz, J., Johnson, G.: Cell biology e-book (2016)
33. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., *et al.*: Conformations of immunoglobulin hypervariable regions. Nature **342**(6252), 877 (1989)
34. Chothia, C., Levitt, M., Richardson, D.: Helix to helix packing in proteins. Journal of molecular biology **145**(1), 215–250 (1981)
35. Levitt, M., Chothia, C.: Structural patterns in globular proteins. Nature **261**(5561), 552 (1976)
36. Peptide Bond and structrue : Hemoglobin. https://commons.wikimedia.org/wiki/File:225_Peptide_Bond-01.jpg. Accessed: 2019-07-28
37. 1GZX. https://www.rcsb.org/structure/1GZX. Accessed: 2019-07-30

38. Colombo, M.F., Rau, D.C., Parsegian, V.A.: Protein solvation in allosteric regulation: a water effect on hemoglobin. Science **256**(5057), 655–659 (1992)
39. Maheswari, D.K.U.: Self-assembly of Proteins. https://nptel.ac.in/courses/118106019/Module%201/Lecture%201/Lecture%201.pdf (accessed June, 2019)
40. Salgado, E.N., Radford, R.J., Tezcan, F.A.: Metal-directed protein self-assembly. Accounts of chemical research **43**(5), 661–672 (2010)
41. Merriam-webster Dictionary. https://www.merriam-webster.com/dictionary/self-assembly#h1
42. Wright, P.E., Dyson, H.J.: Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of molecular biology **293**(2), 321–331 (1999)
43. Wright, P.E., Dyson, H.J.: Intrinsically disordered proteins in cellular signalling and regulation. Nature reviews Molecular cell biology **16**(1), 18 (2015)
44. Sugase, K., Dyson, H.J., Wright, P.E.: Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature **447**(7147), 1021 (2007)
45. Csermely, P., Palotai, R., Nussinov, R.: Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends in biochemical sciences **35**(10), 539–546 (2010)
46. MOTIF Search. https://www.genome.jp/tools/motif/MOTIF.html. Accessed: 2019-07-30
47. PSIPRED 4.0 (Predict Secondary Structure) : UCL Department of Computer Science : Bioninformatics. http://bioinf.cs.ucl.ac.uk/psipred/. Accessed: 2019-07-30
48. GLOBPLOT 2 : Intrinsic Protein Disorder, Domain Globularity Prediction. http://globplot.embl.de/. Accessed: 2019-07-31
49. Liu, J., Rost, B.: Sequence-based prediction of protein domains. Nucleic acids research **32**(12), 3522–3530 (2004)
50. Hidden Markov Models. https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Protein+Secondary+Structure)
51. Repository of Author. https://github.com/hyejeonc