Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 28, S. 1540-1544, Kasım 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 28, pp. 1540-1544, November 2021
Copyright © 2021 EJOSAT
Research Article

Bioinformatic Analysis of Human Collagen Sequence Mutations on Osteogenesis Imperfecta

Gülsüm Tiraş[1], Esma Eryılmaz Doğan[2*]

[1]Selcuk University, Institute of Natural Sciences, Program of Biomedical Engineering, Konya/Turkey, (ORCID ID 0000-0001-8658-6580), gulsum-6342@outlook.com
[*2]Selcuk University, Faculty of Technology, Department of Biomedical Engineering, Konya/Turkey, (ORCID ID 0000-0001-6809-7513), eeryilmaz@selcuk.edu.tr
Email of the corresponding author

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(**DOI:** 10.31590/ejosat.1026083)

ATIF/REFERENCE: Tiraş G., & Eryilmaz Dogan, E., (2021). Bioinformatic Analysis of Human Collagen Sequence Mutations on Osteogenesis Imperfecta. *European Journal of Science and Technology*, (28), 1540-1544.

Abstract

Collagen has been implicated in a number of pathological conditions. When an amino acid in triple helix is replaced with other amino acids, the collagen structure is destroyed. The deterioration in the collagen structure causes various hereditary diseases and dysfunctions. In this study, the mutations on the alpha-1 chain of type I collagen, which is the most common in the human body, were examined using Python programming language. Based on the previous studies, brittle bone disease (OI) type 2 caused by mutations in type-I collagen alpha-1 chains, has been focused on. UniprotKB database were used for the mutations reported. The mutations obtained were combined in an alpha-I chain and it was seen that the most mutated amino acid was glycine (Gly). Since glycine amino acid affects the stability of the helix structure of the collagen alpha-I chain, it can be considered to influence collagen-induced diseases. The most frequently recurring mutations (glycine (G)> arginine (R), glycine (G)> serine (S), glycine (G)>aspartate (D)) were detected. As a result of comparison, increase in molecular mass, change in isoelectric point, decrease in hydropathy index, change in charge state and acid-base properties were observed. The effect of these changed features on brittle bone disease (OI) has been interpreted.

Keywords: Alpha-1 chain, Collagen, Extracellular matrix, Osteogenesis imperfecta, Python

[*] Corresponding Author: eeryilmaz@selcuk.edu.tr

1. Introduction

The human body is a perfect organism that has a complex structure but contains systems that work in harmony. The systems in the body work in connection with each other so that a person can live a healthy life. Each of them has a separate task and these tasks continue continuously within the system. The cell is the smallest part of the living thing in terms of structure and function. Protein, on the other hand, is an important building block in every cell in the human body. It is the most abundant substance in the body after water.

Collagen is an important protein found in large numbers in the human body. It is an important component of our skin, connective tissue, tendons, joints, nails, teeth, hair, cartilage, and bone. Collagen has many critical functions in our body. It is a type of protein found in human tissues and formed by fibroblasts. Its main task is to strengthen the connective tissue and maintain the integrity of our body. Providing elasticity and firmness to the skin, keeping our tasks.

Collagen is biocompatible and safe due to its various properties such as biodegradability, weak antigenicity, and self-assembly. Collagen can form fibers with extra strength and stability by self-assembly and cross-linking. It is a biomaterial frequently used in medical applications. Various diseases may occur as a result of change in the structure of collagen and mutations in genes encoding collagen. For this reason, it is necessary to better analyze collagen in the molecular level. Based on this information, research has been done on the mutation in collagen genes and the diseases that occur as a result. In this study, we investigated the change in physical properties and the number of amino acids following mutations in the structure of collagen. In our study, it is aimed to interpret the effect of mutations on collagen-borne diseases by interpreting the physical properties that change resulting of mutation by using Python programing language. For this purpose, mutations obtained from previous studies recorded in the database were used. Using the known physic al properties of amino acids, it was evaluated that the change resulting from mutation could influence the disease. The number of amino acids with the most mutations was calculated by examining all the obtained mutations.

2. Material and Method

2.1. Collagen Molecule

Collagen is the protein that forms the building blocks of the movement system, especially bones, cartilage, fibers, and joints. The main molecule of collagen is tropocollagen. Tropocollagenes are composed of procollagen produced inside the cell. Collagens are complex protein structures belonging to a large family of 28 members in humans. They form a triple helix with three different polypeptide chains commonly known as alpha chains. It has also been found that this triple helix has a unique "Gly-Xaa-Yaa" sequence repeat. The presence of glycine (Gly) in each sequence explains the stability of the helix due to its feature of being the smallest amino acid. Xaa and Yaa can be any amino acid but are mostly proline. Therefore, collagen is known to be a protein rich in glycine and proline[1]. The structure of collagen is shown in Figure 1.

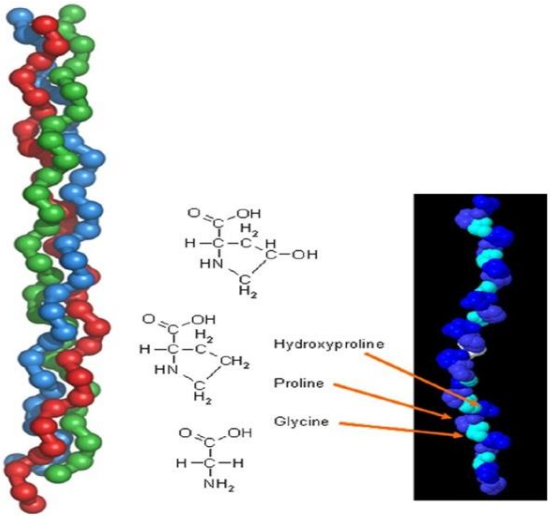

Fig. 1 The structure of collagen [2].

Collagen is one of the most abundant proteins in the human body. It is classified according to its structure. The common types are called fibril-forming collagen; among these, type I, type II, type III, type V, and type XI are found most often in body tissue. Where fibril forming collagens is found in the human body;

Type I: It is the most abundant type of collagen. It is usually found in connective tissue such as skin, bone, dermis, ligament, cornea, and tendon.

Type II: It is found in cartilage, embryonic epithelium mesenchymal transitions, corneal epithelium, notochord, nucleus pulposus of intervertebral discs and vitreous body.

Type III: Usually found often with type I collagen in the skin, vessel wall, intestines, and reticular fibers.

Type V: It is mostly found with type I collagen, which is found in bone, lung, cornea, and fetal membranes.

Type XI: Usually type XI and type II collagens coexist in the vitreous body, intervertebral disc, and cartilage [3].

2.1.1. Osteogenesis Imperfecta

Osteogenesis imperfecta (Ol) is a genetic disorder of increased bone fragility, low bone mass, and other connective tissue manifestations. Most patients with a clinical diagnosis of osteogenesis defect have a mutation in one of the two genes encoding the alpha chains of collagen type 1 (COL1A1 and COL1A2) [4]. The severity of the disease ranges from mild (OI type I) to perinatally fatal (OI type II) and exhibits both autosomal dominant and recessive inheritance patterns [5].

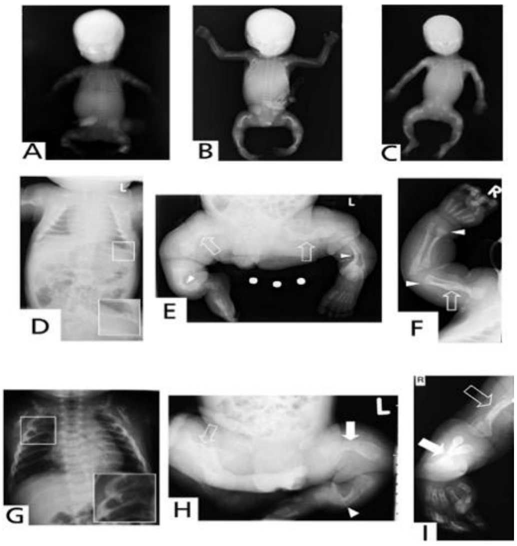

Fig. 2 Radiographs of patients with Osteogenesis imperfecta (OI) due to collagen type I mutations [6].

2.1.2. Python Programming and PyCharm

The type and the count of each amino acid in collagen alpha-1 chain were calculated through the PyCharm text editor using codes written in Python programming language. PyCharm is an integrated development environment used in computer programming specifically for the Python language. Python is an object oriented, interpretative, modular and interactive high level programming language.

2.1.3. UniProtKB/SWISS-PROT

UniProt (http://www.uniprot.org/) is a freely accessible protein sequence and functional information database, derived from many introductory genome sequencing projects. It contains a large amount of information about the biological functions of proteins.

2.1.4. Amino Acids

Amino acids, the basic building blocks of proteins are organic compounds containing amine (–NH2) and carboxyl (–COOH) functional groups and a side chain (R group) specific to each amino acid [7, 8]. The basic elements of an amino acid are carbon (C), hydrogen (H), oxygen (O), and nitrogen (N), but some amino acids have other elements in their side chains. About 500 naturally occurring amino acids are known (although only 20 appear in the genetic code) and can be classified in many ways [9]. They can be classified according to polarity, ph level and side chain group. The class and the structure of 20 types of amino acids is given in Figure 3.

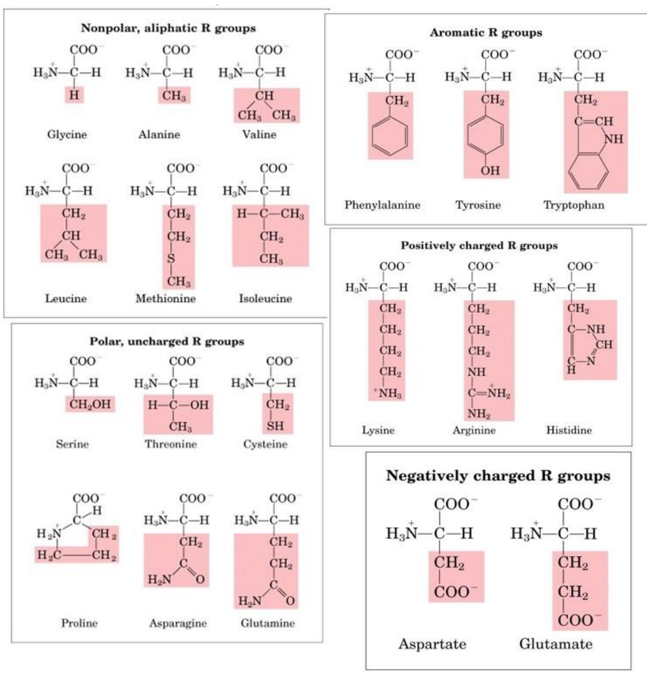

Fig. 3 Structure and classification of amino acids [10].

2.1.5. Isoelectric Point (pI)

The isoelectric point (pI, pH (I), IEP) is the pH at which a molecule carries no net electrical charge or is electrically neutral on the statistical average [11]. Many molecules are zwitterions containing both positive and negative charges. The net charge on the molecule is governed by the pH of the surrounding environment. The isoelectric points of amino are given in Table 1.

Table 1 Isoelectric points of amino acids [12].

Amino acids	*pI (Isoelectric Point)*
Glycine(G)	5,97
Alanine(A)	6,01
Arginine(R)	10,76
Asparagine(N)	5,41
Aspartic acid(D)	2,77
Cysteine(C)	5,07
Glutamic acid(E)	3,22
Glutamine(Q)	5,65
Histidine(H)	7,59
Isoleucine(I)	6,02
Leucine(L)	5,98
Lysine(K)	9,74
Methionine(M)	5,74
Phenylalanine(F)	5,48
Proline(P)	6,48
Serine(S)	5,68
Threonine(T)	5,87
Trptophan(W)	5,89
Tyrosine(Y)	5,66
Valine(V)	5,97

2.1.6. Hydropathy Index

While hydropathy (hydrophobicity vs hydrophilicity or lipophilic vs lipophilicity) is generally characterized by numbers from -7.5 to 3.1, hydrophobicity is a measure of how strongly side chains are repelled from water. The more positive the number, the more likely the amino acid residue will not be in an aqueous environment [13].

Table 2 Hydropathy index values of amino acids [14].

Amino Acid	Hydropathy Index
Lysine(K)	-3.9
Arginine(R)	-4.5
Histidine(H)	-3.2
Glutamic Acid(E)	-3.5
Glutamine(Q)	-3.5
Aspartic Acid(D)	-3.5
Asparagine(N)	-3.5
Trptophan(W)	-0.9
Tyrosine(Y)	-1.3
Serine(S)	-0.8
Threonine(T)	-0.7
Proline(P)	-1.6
Glycine(G)	-0.4
Alanine(A)	1.8
Methionine(M)	1.9
Cysteine(C)	2.5
Phenylalanine(A)	2.8
Leucine(L)	3.8
Valine(V)	4.2
Isoleucine(I)	4.5

3. Results and Discussion

Analysis of the mutated alpha-1 chains by Python

Collagen-borne diseases have not been investigated in molecular level using a programing language so far. In this study, analysis was made using the alpha-1 chain of Type 1 collagen, which is the most abundant type in our body. In the UniProt database, information is given about the mutated amino acids of collagen-induced diseases in the alpha-1 chain of Type 1 collagen. Osteogenesis imperfecta (OI) type II disease caused mutations mostly in the type 1 collagen alpha-1 chain was reviewed. Therewith, articles in the Uniprot database on OI Type II disease were examined. In the study of Bodian et al., mutations detected in 63 different people were reported [15]. All the mutated amino acids in that study were combined into a single collagen alpha-1 chain in order to calculate the type and frequent of each mutated amino acid. The alpha-1 chain of the collagen obtained from the database in FASTA format with the obtained mutated chain was defined in the PyCharm program. Python code was written to calculate the number of amino acids in both chains. A comparison was made to identify the most mutated amino acid. The most repetitive amino acid changes were detected in mutations obtained from different individuals, and the changes caused by these mutations in the alpha-I chain were investigated in the molecular terms. Characteristic molecular parameters such as physical properties, molecular weight, hydropathy index, isoelectric point, and pH were examined, and the contribution of changes in these criteria to intermolecular interactions and their effect on disease formation mechanism were interpreted.

Comparison of the amino acid sequences of the normal and the mutated alpha-I chain

The amino acid sequence in the normal collagen alpha-I chain and the mutated chain were obtained from different people and they were combined in one collagen alpha-I chain. The types and the numbers of each amino acids were calculated using the codes written in the Python programming language with the PyCharm text editor.

As a result of the calculations, it was seen that the highest change was in the number of glycine (G) amino acids, which is the most abundant one in the collagen structure and ensures the stability of the collagen helix structure [16]. The most obvious feature of glycine is that its molecular mass and the volume is the smallest amino acid. Since the excessive number of changes in glycine on the mutated chain, Glycine mutation will affect stability. Osteogenesis is considered as the replacement of glycine by different amino acids in the fragile bone structure formed in Imperfecta disease.

Studying repeated mutations

When looking at the mutations in the study examined, the most recurring mutations were determined as follows: These are aspartic acid (D) instead of glycine (G), arginine (R) instead of glycine (G), serine (S) instead of glycine (G). New alpha-I chains were created separately by selecting some of these mutations. In the alpha-I chain consisting of 1464 amino acids, the mutations of the 353rd row of glycine to aspartic acid, the 614th row of glycine to arginine, and the 884th row of glycine to serine were selected. Alpha-I chains containing mutations and, normal alpha-I chains were defined in FASTA format in PyCharm program, and the average isoelectric (pı) values and total molecular weights of alpha-I chains were calculated.

As a result of the mutations, the average isoelectric values of the alpha-I chains changed and the total molecular weight decreased. Considering the values that change resulting of a single mutation in the alpha-I chain, it can be said that these changes could have a significant effect on Osteogenesis imperfecta disease.

Comparison of the properties of glycine and amino acids substituted for glycine

In order to examine the changing properties of amino acids replacing glycine as a result of mutations, a table containing the physical properties of amino acids was created, inspired by the work of Nassa et al.

The properties of arginine, aspartic acid, and serine amino acids, which replace glycine, were examined. As a result of the comparison, an increase in molecular mass, change in isoelectric point, reduction in hydropathy index, were observed a series of changes in charge state and acidic-basic properties. The decrease in the hydropathy index indicates that the hydrophilic property is increased. It is seen that the amino acids replaced by uncharged glycine, aspartic acid (D), arginine (R), serine (S) are charged molecules. Since these changes affect the alpha-I chain structure, it is thought that the deterioration in the structure is related to Osteogenesis imperfecta disease.

One of the main goals of genome sequencing projects is to provide a better understanding of the genetics of the disease and to help identify the genes associated with the disease [17, 18]. Similar to our study, Flier et al. When his study was examined, it was observed that mutations in Type I procollagen genes caused osteoarthritis and osteoporosis, and the mutations in Type III procollagen genes caused aortic aneurysm. But while they are now convinced that the hypothesis they tested applies to some common diseases in which tissues are broken down, it is unclear how many millions of people with osteoporosis, osteoarthritis, or aortic aneurysm have mutations in their collagen genes [19]. Therefore, with advances in computational technology, it seems reasonable to expect such genetic and sequential analysis to be available at the point of diagnosis and treatment soon.

4. Conclusions and Recommendations

In this study, changed properties were observed as a result of the mutation of a single amino acid in the collagen alpha-I chain with the calculations made and the properties examined. These properties affect the stability, charge status, hydrophilicity, molecular mass, isoelectric point, acidic-basic properties of the entire alpha-I chain. When repeated mutations were examined, it was observed that amino acids substituting glycine were loaded, and that decreased hydropathy index, increased hydrophilic property, increased molecular mass, and changed isoelectric point and acidic basic properties. In addition, when looking at all mutations, it is meaningful that the most change occurs in the glycine amino acid. It is understood that the fragile bone structure, most prominent feature of patients with osteogenesis imperfecta, is that glycine, which provides stability and strength in the collagen structure, is replaced by different amino acids by mutations. Since the most obvious feature of glycine, which provides the stability of the collagen helix structure, is its molecular mass and therefore having the smallest amino acid volume. these changes are thought to cause Osteogenesis imperfecta, also known as glass bone disease.

Recommendations

When the molecular level studies of collagen in the literature are examined, it is seen that although there are sufficient number of experimental studies, theoretical computational studies are quite limited and inadequate. However, with the rapid development of computer technologies and programming in recent years, it is possible to contribute to the detailed analysis of the collagen family at the molecular level, by bioinformatics studies in addition to experimental studies. In addition, disruptions in the collagen structure led to different diseases and dysfunctions. Osteogenesis imperfecta is only one of these diseases. In this study, only Osteogenesis imperfecta type II, which is the most severe and fatal version of the disease, has been focused on. This approach can be applied in different types of the disease. Different bioinformatic analysis can be performed by detecting mutations in other types of osteogenesis imperfecta and different diseases originating from collagen.

References

1. Cen, L., et al., *Collagen tissue engineering: development of novel biomaterials and applications.* Pediatric research, 2008. **63**(5): p. 492-496.
2. Ileana, R. and K. Francois, *Biopolymers for application in photonics.* NBI-technologies, 2014(4).
3. Gelse, K., E. Pöschl, and T. Aigner, *Collagens—structure, function, and biosynthesis.* Advanced drug delivery reviews, 2003. **55**(12): p. 1531-1546.
4. Rauch, F. and F.H. Glorieux, *Osteogenesis imperfecta.* The Lancet, 2004. **363**(9418): p. 1377-1385.
5. Bodian, D.L., et al., *Predicting the clinical lethality of osteogenesis imperfecta from collagen glycine mutations.* Biochemistry, 2008. **47**(19): p. 5424-5432.
6. van Dijk, F.S., et al., *Lethal/severe osteogenesis imperfecta in a large family: a novel homozygous LEPRE1 mutation and bone histological findings.* Pediatric and Developmental Pathology, 2011. **14**(3): p. 228-234.
7. Arjadi, R., et al., *A systematic review of online interventions for mental health in low and middle income countries: a neglected field.* Global Mental Health, 2015. **2**.
8. Bahadıroğlu, S., et al., *ORGANİZMAMIZ İÇİN NİASİN.*
9. Wagner, I. and H. Musso, *New naturally occurring amino acids.* Angewandte Chemie International Edition in English, 1983. **22**(11): p. 816-828.
10. Aryal, S., *Amino Acids- Properties, Structure, Classification and Functions.* August 9, 2018.
11. Fforde, A., *From plan to market: The economic transition in Vietnam*. 2019: Routledge.
12. Idrees, M., et al., *Multimodal role of amino acids in microbial control and drug development.* Antibiotics, 2020. **9**(6): p. 330.
13. Biro, J., *Amino acid size, charge, hydropathy indices and matrices for protein structure analysis.* Theoretical Biology and Medical Modelling, 2006. **3**(1): p. 15.
14. Mitaku, S., T. Hirokawa, and T. Tsuji, *Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces.* Bioinformatics, 2002. **18**(4): p. 608-616.
15. Bodian, D.L., et al., *Mutation and polymorphism spectrum in osteogenesis imperfecta type II: implications for genotype–phenotype relationships.* Human molecular genetics, 2009. **18**(3): p. 463-471.
16. Nassa, M., et al., *Analysis of human collagen sequences.* Bioinformation, 2012. **8**(1): p. 26.
17. Green, E.D., *The human genome project and its impact on the study of human disease.* The genetic basis of human cancer, 1997: p. 33-64.
18. Lander, E.S., et al., *Erratum: Initial sequencing and analysis of the human genome: International Human Genome Sequencing Consortium (Nature (2001) 409 (860-921)).* Nature, 2001. **412**(6846): p. 565-566.
19. Prockop, D.J., *Mutations in collagen genes as a cause of connective-tissue diseases.* New England Journal of Medicine, 1992. **326**(8): p. 540-546.