

Article

Unsupervised Learning for Feature Representation Using Spatial Distribution of Amino Acids in Aldehyde Dehydrogenase (ALDH2) Protein Sequences

Monika Khandelwal¹, Sabha Sheikh¹, Ranjeet Kumar Rout¹ , Saiyed Umer²  and Saurav Mallik^{3,4} 
and Zhongming Zhao^{3,5,*} 

¹ Department of Computer Science & Engineering, National Institute of Technology Srinagar, Hazratbal 190006, India; monika_01phd21@nitsri.net (M.K.); sabha_2021phacse007@nitsri.net (S.S.); ranjeetkumarrou@nitsri.net (R.K.R.)

² Department of Computer Science and Engineering, Aliah University, Newtown, Kolkata 700160, India; saiyed.umer@aliah.ac.in

³ Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; smallik@hsph.harvard.edu or sauravmtech2@gmail.com

⁴ Molecular and Integrative Physiological Sciences (MIPS), Harvard University, Boston, MA 02115, USA

⁵ Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

* Correspondence: zhongming.zhao@uth.tmc.edu; Tel.: +1-713-500-3631

Abstract: Aldehyde dehydrogenase 2 (ALDH2) enzyme is required for alcohol detoxification. ALDH2 belongs to the aldehyde dehydrogenase family, the most important oxidative pathway of alcohol digestion. Two main liver isoforms of aldehyde dehydrogenase are cytosolic and mitochondrial. Approximately 50% of East Asians have ALDH2 deficiency (inactive mitochondrial isozyme), with lysine (K) for glutamate (E) substitution at position 487 (E487K). ALDH2 deficiency is also known as Alcohol Flushing Syndrome or Asian Glow. For people with an ALDH2 deficiency, their face turns red after drinking alcohol, and they are more susceptible to various diseases than ALDH2-normal people. This study performed a machine learning analysis of ALDH2 sequences of thirteen other species by comparing them with the human ALDH2 sequence. Based on the various quantitative metrics (physicochemical properties, secondary structure, Hurst exponent, Shannon entropy, and fractal dimension), these fourteen species were clustered into four clusters using the unsupervised machine learning (K-means clustering) algorithm. We also analyze these species using hierarchical clustering (agglomerative clustering) and draw the phylogenetic trees. The results show that *Homo sapiens* is more closely related to the *Bos taurus* and *Sus scrofa* species. Our experimental results suggest that the testing for discovering medicines may be done on these species before being tested in humans to alleviate the impacts of ALDH2 deficiency.

Keywords: aldehyde dehydrogenase 2; ethanol metabolism; machine learning; physicochemical properties; secondary structure

MSC: 92B05



Citation: Khandelwal, M.; Sheikh, S.; Rout, R.K.; Umer, S.; Mallik, S.; Zhao, Z. Unsupervised Learning for Feature Representation Using Spatial Distribution of Amino Acids in Aldehyde Dehydrogenase (ALDH2) Protein Sequences. *Mathematics* **2022**, *10*, 2228. <https://doi.org/10.3390/math10132228>

Academic Editor: Zhisheng Shuai

Received: 16 May 2022

Accepted: 22 June 2022

Published: 25 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Aldehyde dehydrogenase 2 is an enzyme coded by the ALDH2 gene and is needed for alcohol detoxification. Aldehyde dehydrogenase 2 (ALDH2) proteins belong to the aldehyde dehydrogenase family. Aldehyde dehydrogenase is the second enzyme of the essential oxidative pathway of alcohol digestion. Two primary liver isozymes of aldehyde dehydrogenase, cytosolic and mitochondrial, can be differentiated by their subcellular localizations, kinetic properties, and electrophoretic mobilities. Almost all Caucasians have two primary isozymes, while around 50% of East Asians have the cytosolic isozyme

but not the mitochondrial isozyme. An astoundingly higher recurrence of intense liquor intoxication among East Asians than among Caucasians can be associated with the absence of a catalytically active type of the mitochondrial isozyme. The expanding exposure to acetaldehyde in people with the catalytically inactive type may also experience more significant susceptibility to various forms of cancer. ALDH2 removes acetaldehyde, a toxic product from ethanol breakdown [1]. ALDH2 converts acetaldehyde to acetate, which the body can easily digest. ALDH2 plays a crucial role in the pathogenesis of diabetes, cancer, neurodegenerative diseases, and cardiovascular diseases [2]. ALDH2 deficient people have a more significant risk of alcohol-related cancers such as breast cancer, liver cancer, neck and head cancer, colorectal cancer, and esophageal cancer [3].

Alcohol metabolism in humans includes two main enzymes, aldehyde dehydrogenase and alcohol dehydrogenase. First, the alcohol is oxidized to acetaldehyde by alcohol dehydrogenase (ADH), and then acetaldehyde is oxidized to non-toxic acetic acid by aldehyde dehydrogenase (ALDH) for evacuation. Amid various ALDH isoenzymes in humans, a mitochondrial enzyme, ALDH2, is the primarily effective enzyme to get rid of acetaldehyde [4]. Acetaldehyde is converted to acetate for people with the variant ALDH2*1. ALDH2 deficiency is also called alcohol flushing syndrome, a genetic condition that obstructs alcohol metabolism [5]. ALDH2 deficiency affects 8% of the world population, mainly in East Asia, affecting 36% of the population in East Asia. People carrying the mutant ALDH2*2 are more likely to have various types of cancer. People with ALDH2*2 variants turn red and might have other symptoms like dizziness, headache, heart palpitation, and hypertension after consuming alcohol [2,6]. ALDH2 dysfunction leads to various human diseases like diabetes, cancer [7], stroke, neurodegenerative diseases, and cardiovascular diseases. ALDH2*2 variant encodes a lysine (K) for glutamate (E) substitution at position 487 (E487K), named the ALDH2*2 allele [8].

Humans have 19 ALDH genes on distinct chromosomes, and among them, ALDH2 is important for ethanol metabolism. The ALDH2 gene is located at chromosome 12 in the locale of q24.2 [9]. Humans are exposed to acetaldehyde regularly via various sources such as cigarettes, foods, the environment, and beverages, but the highest exposure is due to alcohol consumption [10]. ALDH2 plays a crucial role in pathological and physiological processes. Utilization of extreme amounts of alcohol impacts worldwide DNA methylation [11,12]. DNA methylation gives biomarkers of alcohol consumption, so recognizing epigenetic biomarkers helps diagnose alcohol-related diseases [13].

In this study, we chose fourteen different species (*Homo Sapiens*, *Pongo abelii*, *Rattus norvegicus*, *Amblyraja radiata*, *Sus scrofa*, *Meleagris gallopavo*, *Xenopus tropicalis*, *Mus pahari*, *Arvicanthus niloticus*, *Cricetulus griseus*, *Danio rerio*, *Bos taurus*, *Grammomys surdaster*, and *Mus musculus*) and analyzed the ALDH2 sequence from thirteen non-human species concerning the human ALDH2 sequence and found out the degree of changeability by which the sequences vary from each other. Besides the phylogenetic analysis of these sequences, we conducted a comprehensive study based on physicochemical properties, Shannon entropy, Hurst exponent, secondary structure, and fractal dimension.

The main contributions of this study are as follows:

- A comprehensive analysis of ALDH2 gene sequences of various species, including *Homo sapiens* (human).
- Fractal dimension leading to the discovery of the self-similarity within ALDH2 sequences.
- Identification of the auto-correlation between sequences by the Hurst exponent.
- Phylogenetic analysis of ALDH2 sequences of fourteen species.

The rest of the paper is organized as follows. Section 2 will illustrate the dataset, feature representation, and methods used in this study. Section 3 will describe the result and discussion based on various parameters. Finally, Section 4 will summarize the work done in this paper.

2. Materials and Methods

This section will discuss the dataset used in this study and various features extracted from the ALDH2 sequences. We will also discuss the unsupervised machine learning methods used in this study.

2.1. Dataset

The ALDH2 protein sequences from fourteen species *Pongo abelii* (sumatran orangutan), *Homo sapiens* (human), *Rattus norvegicus* (brown rat), *Amblyraja radiata* (thorny skate), *Sus scrofa* (domestic pig), *Meleagris gallopavo* (wild turkey), *Xenopus tropicalis* (western clawed frog), *Mus pahari* (gairdner's shrew-mouse), *Arvicanthis niloticus* (african grass rat), *Cricetulus griseus* (chinese hamster), *Danio rerio* (zebrafish), *Bos taurus* (aurochs), *Grammomys surdaster* (african woodland thicket rat), and *Mus musculus* (house mouse) were obtained from the NCBI (National Center for Biotechnology Information) database [14]. The fourteen species with their respective length are given in Table 1.

Table 1. Fourteen species and their respective ALDH2 protein sequences.

Sequence ID	Species	Common Name of Species	ALDH2 Accession ID	Length (aa) *
S ₁	<i>Pongo abelii</i>	Sumatran orangutan	XP_024111823.1	436
S ₂	<i>Homo sapiens</i>	Human	NP_000681.2	517
S ₃	<i>Rattus norvegicus</i>	Brown rat	NP_115792.2	519
S ₄	<i>Amblyraja radiata</i>	Thorny skate	XP_032899607.1	516
S ₅	<i>Sus scrofa</i>	Domestic pig	NP_001038076.1	521
S ₆	<i>Meleagris gallopavo</i>	Wild turkey	XP_010718484.1	422
S ₇	<i>Xenopus tropicalis</i>	Western clawed frog	NP_001004907.1	521
S ₈	<i>Mus pahari</i>	Gairdner's shrew-mouse	XP_021042739.1	519
S ₉	<i>Arvicanthis niloticus</i>	African grass rat	XP_034345265.1	519
S ₁₀	<i>Cricetulus griseus</i>	Chinese hamster	XP_007625900.2	519
S ₁₁	<i>Danio rerio</i>	Zebrafish	XP_002662252	516
S ₁₂	<i>Bos taurus</i>	Aurochs	NP_001193787.1	520
S ₁₃	<i>Grammomys surdaster</i>	African woodland thicket rat	XM_028765895.1	519
S ₁₄	<i>Mus musculus</i>	House mouse	NP_033786.1	519

* aa: amino acid.

2.2. Feature Representation

This section will describe the various features used in this study. The features for all fourteen ALDH2 sequences were calculated using online web servers. We used the following three features to represent a protein sequence:

1. Physicochemical properties;
2. Statistical measures;
3. Secondary structure prediction.

2.2.1. Physicochemical Properties

Several physicochemical properties such as extinction coefficients, theoretical pI, molecular weight, instability index, aliphatic index, amino acid composition, negatively and positively charged residues, grand average of hydropathicity (GRAVY), and the atomic composition of carbon, hydrogen, nitrogen, oxygen, and sulfur for all fourteen species were calculated using the online web server ProtParam [15].

Instability index: The instability index estimates whether a protein is stable or unstable in a test tube. A protein with an instability index smaller than 40 is considered stable, and a value larger than 40 is considered unstable. It was estimated using the ProtParam server [15]. There are 400 dipeptides, and a dipeptide instability weight value (DIWV) was assigned to each dipeptide by Guruprasad et al. [16]. The instability index (II) for a protein sequence was calculated using the following equation:

$$II = \frac{10}{N} \sum_{i=1}^{N-1} DIWV(X_i X_{i+1}) \quad (1)$$

where N is the length of the protein sequence and $DIWV(X_i X_{i+1})$ is the instability weight value given to dipeptide $X_i X_{i+1}$.

Extinction coefficients: The extinction coefficients demonstrate how much light a protein consumes at a specific wavelength. It is beneficial to evaluate this coefficient when purifying a protein and estimate it by using ProtParam [15]. The molar extinction coefficient can be estimated from the protein sequence [17]. From the molar extinction coefficient, we can estimate the extinction coefficient (EC) of an essential protein in water using the equation:

$$EC = N(Y) * Ex(Y) + N(W) * Ex(W) + N(Cy) * Ex(Cy) \quad (2)$$

where $Ex(Y)$, $Ex(Cy)$, and $Ex(W)$ are the molar extinction coefficients of tyrosine, cysteine, and tryptophan, respectively. $N(Y)$, $N(W)$, and $N(Cy)$ are the number of tyrosine, tryptophan, and cysteine residues per molecule, respectively.

Aliphatic index: The aliphatic index for a protein is described as the respective volume collected by aliphatic side chains (leucine, isoleucine, valine, and alanine). It may be considered as a positive factor for the expansion of the thermostability of spheroproteins [15,18]. The aliphatic index (AI) of a protein sequence was calculated using the following equation:

$$AI = x_{ala} + a * x_{val} + b * (x_{ile} + x_{leu}) \quad (3)$$

where x_{ala} , x_{ile} , x_{leu} , and x_{val} are the mole fraction of alanine, isoleucine, leucine, and valine, respectively. a and b are coefficients of the relative volume of valine and isoleucine/leucine side chains to the alanine side chains, respectively.

GRAVY: The value of GRAVY for a protein/peptide was estimated as the sum of hydrophathy values [19] of all amino acids, divided by the length of the protein/peptide sequence. It was estimated using the online web server ProtParam [15].

Theoretical pI: The theoretical pI indicates the pH where the protein has a net zero charge, i.e., the negative and positive charges are the same. The ProtParam server [15] estimates the theoretical pI and molecular weight by using the pI/Mw tool. The theoretical pI is affected by the buffer size of the protein.

2.2.2. Statistical Measures

We used three statistical measures, i.e., Shannon entropy, Hurst exponent, and fractal dimension, to respectively find randomness, correlation, and self-similarity in a protein sequence.

Shannon Entropy (SE): It estimates the degree of complexity in an ALDH2 sequence and is calculated using the following equation [20]:

$$SE = - \sum_{i=1}^{20} p_i \log_2(p_i) \quad (4)$$

where p_i indicates the probability of amino acid i in a protein sequence [21].

Hurst exponent (HE): The Hurst exponent measures the smoothness and degree of similarity of a data set. It can be computed using rescaled range analysis (R/S analysis), whose value lies between 0 and 1 [22–24]. If the value of HE lies between 0 and 0.5, then it

indicates a negative autocorrelation, and if it lies between 0.5 and 1, it indicates a positive autocorrelation of a time series. If the value of HE is 0.5, then it indicates the randomness of a series, which means there is no correlation between the variable and its past values [25]. The HE of a sequence X_n is defined by the following equation:

$$\frac{R(n)}{S(n)} = \left(\frac{n}{2}\right)^{HE} \tag{5}$$

where:

$$S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - m)^2} \tag{6}$$

$$R(n) = \max(Y_1, Y_2, \dots, Y_n) - \min(Y_1, Y_2, \dots, Y_n) \tag{7}$$

$$Y_t = \sum_{i=1}^t (X_i - m) \quad \text{for } t = 1, 2, 3, \dots, n \tag{8}$$

$$m = \frac{1}{n} \sum_{i=1}^n X_i \tag{9}$$

The HE is evaluated by plotting the values of (R/S) versus n in a log–log plot. The slant of the best fitting line approximates the HE. The HE specifies the amount of self-similarity of a primary sequence. The HE for long-range dependence is between 0.5 and 1. An expanding value of HE specifies a surge in the amount of long-range dependency and self-similarity.

Indicator matrix and fractal dimension: Each protein sequence is encoded into indicator matrices [26,27]. Let S_N be a protein sequence with length N . The indicator function is defined as:

$$F : S_N \times S_N \rightarrow \{0, 1\} \tag{10}$$

such that the indicator matrix will be:

$$I(N, N) = \begin{cases} 1, & \text{if } s_i = s_j \\ 0, & \text{if } s_i \neq s_j \end{cases} \quad \text{where } s_i, s_j \in S_N \tag{11}$$

Here, $I(N, N)$ is a matrix with 0 and 1, giving a binary image of the protein sequence as a 2D dot-plot. The binary image can conceptualize the distribution of zeros and ones within the same sequence. It can be done by allocating a white dot to 0 and a black dot to 1. The fractal dimension (FD) from an indicator matrix can be estimated as the average number of $\sigma(n)$ of 1, randomly taken $n \times n$ from an $N \times N$ indicator matrix [28]. Using $\sigma(n)$, FD is given by the following equation:

$$FD = -\frac{1}{N} \sum_{n=2}^N \frac{\log(\sigma(n))}{\log n} \tag{12}$$

2.2.3. Secondary Structure Prediction

The online web-server CFSSP (Chou and Fasman Secondary Structure Prediction Server) was utilized to predict the secondary structure of ALDH2 sequences of all fourteen species [29]. This server predicts the protein sequences’ beta-sheet, alpha-helix, and turns.

2.3. Feature Extraction

Every ALDH2 sequence was represented using 22 features. These features were denoted by F_1 to F_{22} . The features from F_1 to F_{16} were computed using physicochemical properties of all fourteen sequences, and these properties were computed using the online web server ProtParam [15]. The features F_{17} , F_{18} , and F_{19} were calculated by using Shannon entropy, Hurst exponent, and fractal dimension, respectively. The features F_{20} to F_{22} were

computed by the secondary structure of protein sequences by using the online web server CFSSP [29]. The characterization of feature extraction is shown in Figure 1.

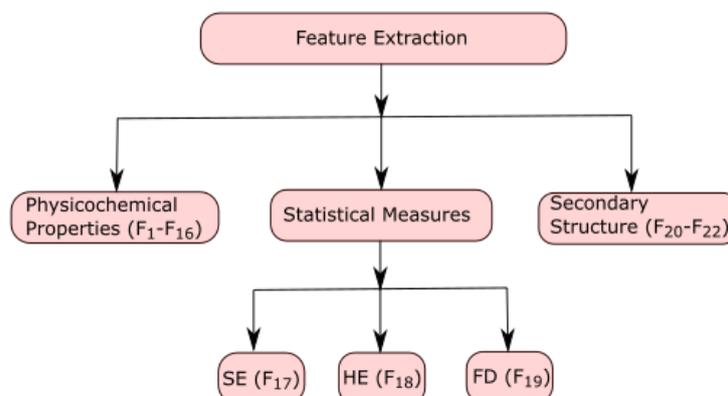


Figure 1. Characterization of features used to represent ALDH2 sequences. SE: Shannon entropy. HE: Hurst exponent. FD: fractal dimension.

2.4. Methods

Based on the physicochemical properties, Shannon entropy, fractal dimension, secondary structure, and Hurst exponent, the fourteen ALDH2 sequences of different species were clustered using hierarchical and K-means clustering. The employed feature representations are physicochemical properties, secondary structure, and statistical measures. These features have different importance in extracting the information from ALDH2 sequences. Thus, there is no standard method to apply to the extracted information from the sequences such that the generic information can be retrieved and that information can differentiate different species. Due to the diverse importance of extracted importance, the problem is applying unsupervised machine learning techniques for extracting informative patterns. The well-known unsupervised machine learning technique is clustering, and for clustering, K-means and hierarchical clustering are primarily used in practice. The motivation behind using K-means clustering is to group the specific species into its proper class. K-means clustering is used to find different clusters based on common characteristics of unlabeled data. Further analysis has been carried out by using phylogeny trees. The motivation behind using hierarchical clustering is to find the evolutionary relationships among these species.

K-means clustering: It is a well-known unsupervised machine learning algorithm that partitions the unlabeled data into different groups or clusters based on similar characteristics and common patterns. According to a distance measure such as Euclidean distance, the data points in a cluster should be similar and dissimilar to those in different clusters. K-means clustering is an iterative method that partitions the data points into K clusters (pre-defined) where each data point should lie in only one cluster [20]. Let X_1, X_2, \dots, X_n represent n data points. C_1, C_2, \dots, C_K denote K clusters and $\mu_1, \mu_2, \dots, \mu_K$ represent centroids of K clusters.

The K-means algorithm steps are as follows:

1. Decide the value of K , i.e., the desired number of clusters (in this study, $K = 4$).
2. Randomly select K data points as centroids ($\mu_1, \mu_2, \mu_3, \mu_4$).
3. Repeat the following steps until there is no change to the centroids.
4. Find the Euclidean distance between all centroids and data points.
 $d(X_i, \mu_j)$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, K$.
5. Assign every data point to the nearest centroid (cluster), i.e., $X_i \in C_j$ and j is given by:

$$j = \min_j d(X_i, \mu_j) \quad \text{for } j = 1, 2, \dots, K \quad (13)$$

- For each cluster, compute the centroid by averaging all data points belonging to that cluster.

$$\mu_j = \frac{1}{|C_j|} \sum_{\forall X_i \in C_j} X_i \quad \text{for } j = 1, 2, \dots, K \quad (14)$$

Hierarchical clustering: Hierarchical clustering is a technique that partitions the objects into homogeneous groups based on the similarity between objects [30]. Unlike K-means clustering, hierarchical clustering does not depend on K (the number of clusters). In this paper, we used agglomerative hierarchical clustering, which starts by taking each object as a separate cluster and repeatedly follows the two steps below until there is a single cluster:

- Find the two closest clusters based on similarity measured by the Euclidean distance matrix.
- Combine the two closest clusters to form a new cluster.

The hierarchical clustering output is shown as a dendrogram.

3. Results

All fourteen species were clustered based on the physicochemical properties, secondary structure, Shannon entropy, fractal dimension, and Hurst exponent. All fourteen species were clustered using the K-means algorithm.

In this work, three feature representation techniques were considered for extracting information from each sequence. Here, physicochemical properties derived 16 different features, statistical techniques measured 3 different features, and secondary structure techniques computed 3 different features. Hence, each sequence is represented by a 22-dimensional feature vector. The values of these features are different. Thus, to preserve these features' dominance property, each feature will have an equal contribution to finding the information from ALDH2 sequences. Thus, to make these features useful, a Z-score data normalization technique was employed to scale them in [0, 1] intervals. Now, these normalized features are differently employed for clustering analysis of ALDH2 sequences. In this work, Euclidean distance calculation has been employed for clustering analysis to find the similarity between two clusters. Hence, the raw score with the largest magnitude has been preserved using Z-score normalization, while the dissimilarity between the clusters was performed using Euclidean distance. For each species, a distance matrix was calculated using the Euclidean distance:

$$d(S_1, S_2) = \sum_{j=1}^{22} (p_j - q_j)^2 \quad (15)$$

Here, p_j and q_j represent the j th feature for the species S_1 and S_2 , respectively. The distance matrix for all fourteen species is given in Figure 2.

After finding the distance matrix, we applied K-means clustering. The first step is to find the value of K. We used the elbow method to find the optimal value of K (numbers of clusters) [31]. For every value of K starting from 2, we computed WCSS (Within-Clusters Sum of Squares), i.e., the sum of squared distance among every sample and its nearest centroid of a cluster. As the value of K increases, the sum of squared distance decreases. To find the optimal value of K, we chose the value of K at the elbow point where WCSS starts decreasing linearly. Thus from Figure 3, we selected the value of K as 4. The following equation calculates WCSS:

$$WCSS = \sum_{k=1}^K \sum_{\forall S_j \in C_k} (S_j - Cen_k)^2 \quad (16)$$

where S_j represents samples in a cluster and C_1, C_2, \dots, C_k denote clusters. K is the total number of clusters. $Cen_1, Cen_2, \dots, Cen_k$ represent the centroids for the respective clusters.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0.3129	0.8186	0.7196	0.9575	0.0842	0.7797	0.8211	0.8202	0.7551	0.7675	0.8860	0.7570	0.8191
2	0.3129	0	0.5062	0.4117	0.6446	0.3613	0.4710	0.5088	0.5079	0.4447	0.4568	0.5730	0.4468	0.5068
3	0.8186	0.5062	0	0.1193	0.1498	0.8658	0.0748	0.0068	0.0031	0.0742	0.0617	0.0745	0.0742	0.0011
4	0.7196	0.4117	0.1193	0	0.2685	0.7729	0.0602	0.1195	0.1193	0.0451	0.0576	0.1936	0.0451	0.1193
5	0.9575	0.6446	0.1498	0.2685	0	1	0.2166	0.1493	0.1494	0.2238	0.2110	0.0754	0.2234	0.1496
6	0.0842	0.3613	0.8658	0.7729	1	0	0.8323	0.8687	0.8677	0.8058	0.8179	0.9304	0.8080	0.8665
7	0.7797	0.4710	0.0748	0.0602	0.2166	0.8323	0	0.0731	0.0735	0.0342	0.0305	0.1442	0.0309	0.0743
8	0.8211	0.5088	0.0068	0.1195	0.1493	0.8687	0.0731	0	0.0052	0.0747	0.0620	0.0744	0.0744	0.0062
9	0.8202	0.5079	0.0031	0.1193	0.1494	0.8677	0.0735	0.0052	0	0.0744	0.0617	0.0743	0.0742	0.0020
10	0.7551	0.4447	0.0742	0.0451	0.2238	0.8058	0.0342	0.0747	0.0744	0	0.0130	0.1487	0.0033	0.0742
11	0.7675	0.4568	0.0617	0.0576	0.2110	0.8179	0.0305	0.0620	0.0617	0.0130	0	0.1361	0.0125	0.0617
12	0.8860	0.5730	0.0745	0.1936	0.0754	0.9304	0.1442	0.0744	0.0743	0.1487	0.1361	0	0.1485	0.0744
13	0.7570	0.4468	0.0742	0.0451	0.2234	0.8080	0.0309	0.0744	0.0742	0.0033	0.0125	0.1485	0	0.0742
14	0.8191	0.5068	0.0011	0.1193	0.1496	0.8665	0.0743	0.0062	0.0020	0.0742	0.0617	0.0744	0.0742	0

Figure 2. Distance matrix of all species derived using Euclidean distance. The numbers on the top and left denote the protein sequence ID (see Table 1).

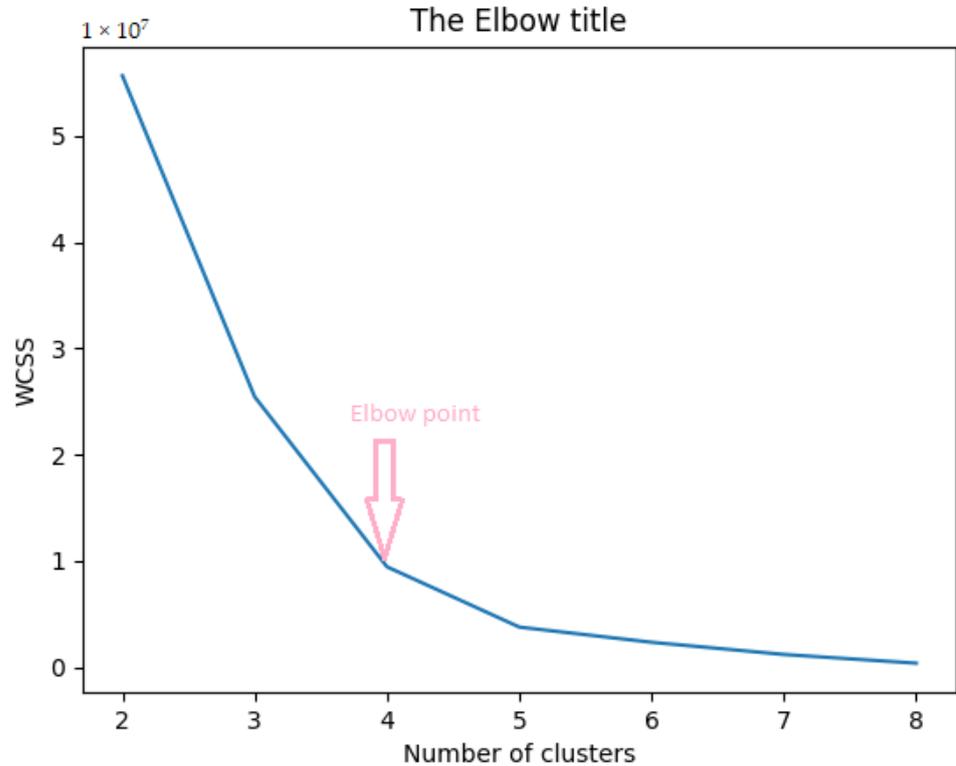


Figure 3. Elbow method to find the value of K (number of clusters) for K -means clustering of all fourteen species. WCSS denotes Within-Clusters Sum of Squares.

3.1. Experiment Based on Physicochemical Properties

Homogeneity was derived based on physicochemical properties using K-means clustering and hierarchical clustering among all fourteen species. The results of K-means clustering and phylogenetic trees are shown in Figures 4 and 5, respectively. The fourteen species were clustered into four clusters using the K-means algorithm based on physicochemical properties. The clusters of species {*Pongo abelii*, *Meleagris gallopavo*}, {*Sus scrofa*, *Bos taurus*} and {*Rattus norvegicus*, *Amblyraja radiata*, *Xenopus tropicalis*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Danio rerio*, *Grammomys surdaster*, *Mus musculus*} indicate that these species are closely related. From the clusters based on physicochemical properties, we observed that *Homo sapiens* was not clustered with any other species.

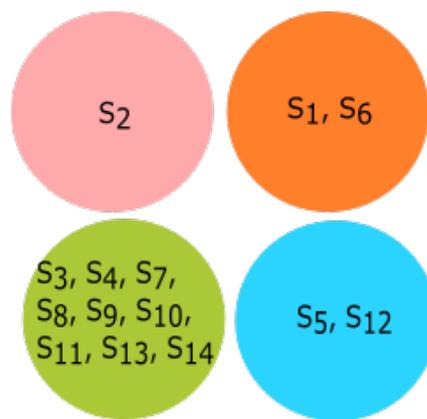


Figure 4. Clusters based on physicochemical properties using the K-means algorithm. Here, S_i indicates the species name as specified in Table 1.

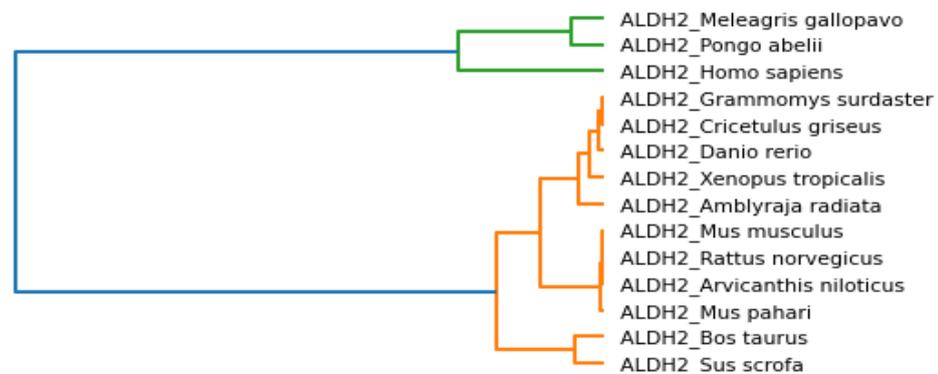


Figure 5. Hierarchical clustering phylogenetic tree based on physicochemical properties.

Phylogenetic trees represent evolutionary relations between species, and the associated clusters are shown in Figure 6. We observed that *Danio rerio*, *Grammomys surdaster*, *Cricetulus griseus*, *Xenopus tropicalis*, and *Amblyraja radiata* are closer based on the analysis of physicochemical properties. *Mus musculus*, *Rattus norvegicus*, *Arvicanthis niloticus*, and *Mus pahari* were grouped. It was a similar case for *Meleagris gallopavo* and *Pongo abelii*. Similarly, *Bos taurus* and *Sus scrofa* were placed nearby. *Homo sapiens* was placed separately but bore similarities to *Pongo abelii* and *Meleagris gallopavo*.

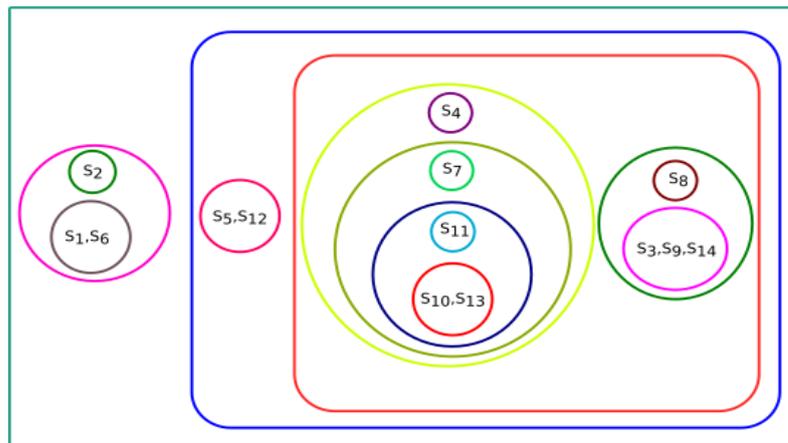


Figure 6. Clusters drawn from hierarchical clustering phylogenetic tree based on physicochemical properties. Here, S_i represents the species name as specified in Table 1.

3.2. Experiment Based on Secondary Structure

Secondary structures for all fourteen species were predicted using the online web server CFSSP (Chou and Fasman Secondary Structure Prediction Server) [29]. All of the fourteen species were clustered using K-means clustering based on their secondary structure, and clusters are shown in Figure 7. The groups of species {*Pongo abelii*, *Xenopus tropicalis*, *Danio rerio*}, {*Sus scrofa*, *Meleagris gallopavo*, *Bos taurus*}, and {*Homo sapiens*, *Rattus norvegicus*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Grammomys surdaster*, *Mus musculus*} are closely related. From Figure 7, we observed that *Amblyraja radiata* was not clustered with any other species. However, it was clustered with *Rattus norvegicus*, *Xenopus tropicalis*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Danio rerio*, *Grammomys surdaster*, and *Mus musculus* as per the physicochemical properties of ALDH2 sequences.

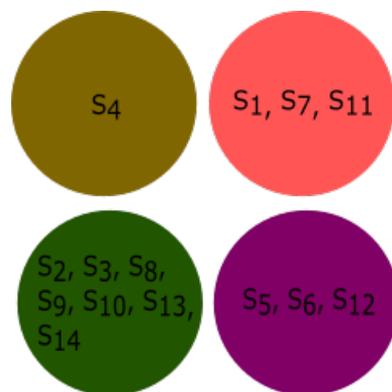


Figure 7. Clusters based on secondary structure using k-means algorithm. Here, S_i indicates the species name as specified in Table 1.

The phylogenetic tree based on the secondary structure and associated clusters are shown in Figures 8 and 9, respectively. We observed that *Mus pahari*, *Grammomys surdaster*, *Cricetulus griseus*, *Arvicanthis niloticus*, and *Rattus norvegicus* were closer according to the secondary structure analysis. *Homo sapiens* and *Mus musculus* grouped along with the above species. Again, the cases for *Meleagris gallopavo*, *Sus scrofa* and *Bos taurus* were similar. Similarly, *Danio rerio*, *Xenopus tropicalis* and *Pongo abelii* were close. However, *Amblyraja radiata* was placed separately because it did not show similarity based on the secondary structure.

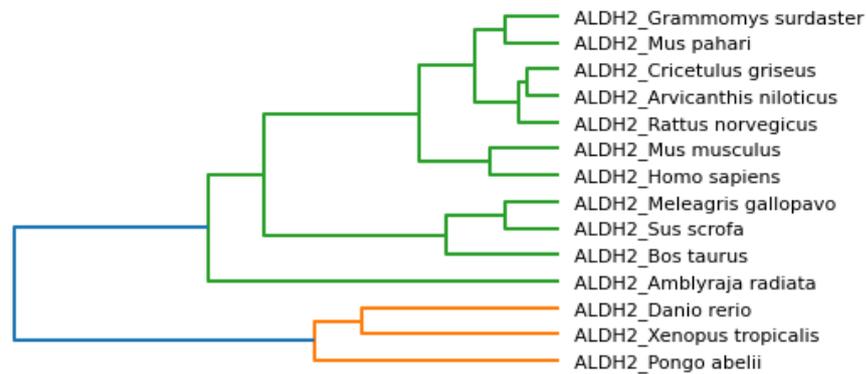


Figure 8. Hierarchical clustering phylogenetic tree based on secondary structure.

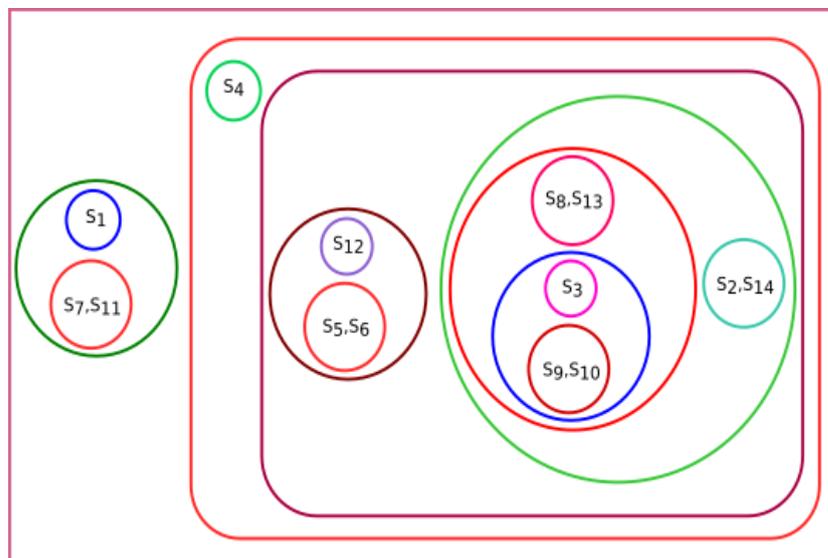


Figure 9. Clusters drawn from Hierarchical clustering phylogenetic trees based on secondary structure. Here, S_i represents the species name as specified in Table 1.

3.3. Experiment Based on Hurst Exponent

All fourteen species were clustered based on the Hurst exponent using the K-means algorithm. The clusters of species $\{Homo sapiens, Amblyraja radiata, Bos taurus, Danio rerio, Xenopus tropicalis\}$ and $\{Rattus norvegicus, Meleagris gallopavo, Mus pahari, Arvicanthis niloticus, Cricetulus griseus, Grammomys surdaster, Mus musculus\}$ are closely related according to the Hurst exponent. The species *Pongo abelii* formed a singleton cluster when the Hurst exponent was taken into consideration even though it was clustered with *Xenopus tropicalis* and *Danio rerio* as per the secondary structure and clustered with *Meleagris gallopavo* as per the physicochemical properties. *Sus scrofa* formed a singleton cluster. Still, it was grouped with *Bos taurus* as per the physicochemical properties and with *Meleagris gallopavo* and *Bos taurus* as per the secondary structure of ALDH2 sequences. The results of K-means clustering are shown in Figure 10.

The phylogenetic tree based on the Hurst exponent and associated clusters are shown in Figures 11 and 12, respectively. From Figure 11, we observed that *Danio rerio, Amblyraja radiata, Xenopus tropicalis, Bos taurus* and *Homo sapiens* were close. Similarly, the species *Arvicanthis niloticus, Mus pahari, Cricetulus griseus, Rattus norvegicus, Grammomys surdaster* were related to *Mus musculus* and *Meleagris gallopavo*. *Pongo abelii* was not grouped with other species.

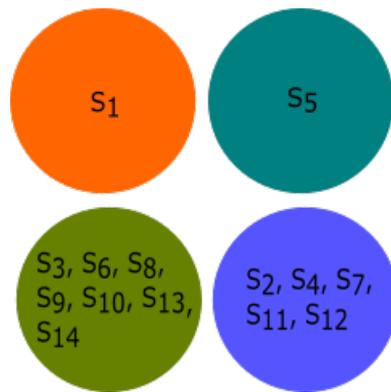


Figure 10. Clusters based on the Hurst exponent using the K-means algorithm. Here, S_i indicates the species name as specified in Table 1.

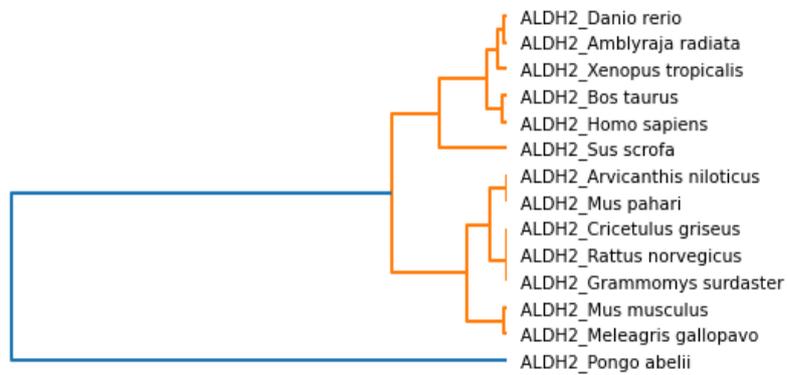


Figure 11. Hierarchical clustering phylogenetic tree based on Hurst exponent.

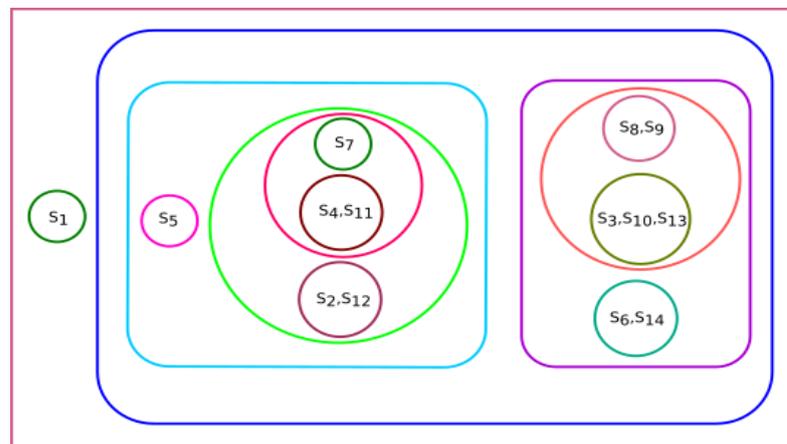


Figure 12. Clusters drawn from phylogenetic tree based on Hurst exponent. Here, S_i indicates the species name as specified in Table 1.

3.4. Experiment Based on Fractal Dimension

All species were clustered based on the fractal dimension using K-means clustering. The results of the four clusters are shown in Figure 13. The clusters of species {*Bos taurus*, *Homo sapiens*}, {*Amblyraja radiata*, *Mus pahari*, *Sus scrofa*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Rattus norvegicus*, *Grammomys surdaster*, *Mus musculus*}, and {*Meleagris gallopavo*, *Xenopus tropicalis*, *Danio rerio*} were placed together according to fractal dimension. *Pongo abelii* formed a singleton cluster and not grouped with other species.

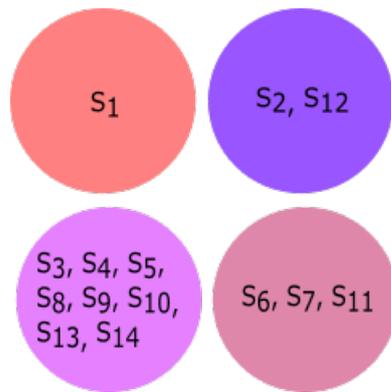


Figure 13. Clusters based on fractal dimension using K-means algorithm. Here, S_i indicates the species name as specified in Table 1.

The phylogenetic tree based on fractal dimension and associated clusters are shown in Figures 14 and 15, respectively. We found that *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Arvicanthis niloticus*, *Mus pahari*, and *Grammomys surdaster* were closely related to *Sus scrofa* and *Amblyraja radiata* and grouped. Similarly, *Danio rerio*, *Meleagris gallopavo*, and *Xenopus tropicalis* were placed together. Due to the fractal dimension similarity between *Bos taurus* and *Homo sapiens*, they were grouped. The species *Pongo abelii* was placed separately.



Figure 14. Hierarchical clustering phylogenetic tree based on fractal dimension.

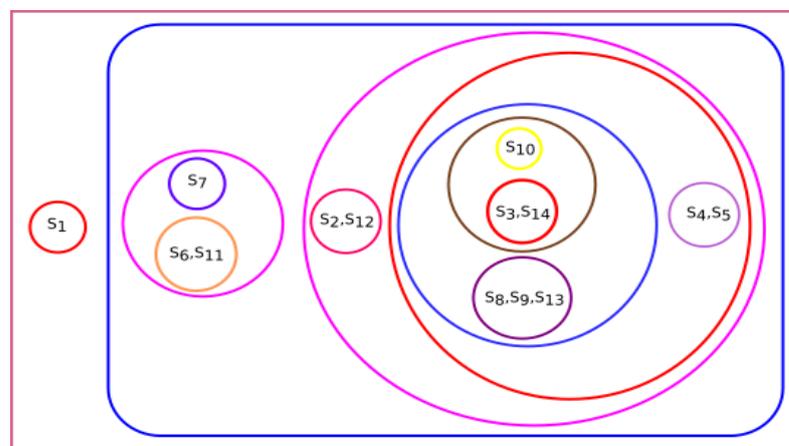


Figure 15. Clusters drawn from hierarchical clustering phylogenetic tree based on fractal dimension. Here, S_i indicates the species name as specified in Table 1.

3.5. Experiment Based on Shannon Entropy

All fourteen species were grouped based on their Shannon entropy. K-means clustering was used to cluster these species; the results are shown in Figure 16. The groups of species {*Amblyraja radiata*, *Danio rerio*}, {*Rattus norvegicus*, *Sus scrofa*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Grammomys surdaster*}, and {*Homo sapiens*, *Meleagris gallopavo*, *Xenopus tropicalis*, *Bos taurus*, *Mus musculus*} were closely related. We found that *Pongo abelii* was clustered uniquely when Shannon entropy was considered. It did not form any cluster with other species per the Hurst exponent and fractal dimension of ALDH2 protein sequences.

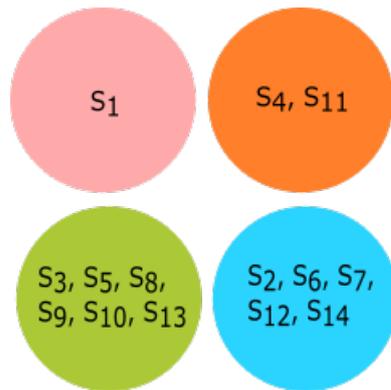


Figure 16. Clusters based on Shannon entropy using the K-means algorithm. Here, S_i represents the species name as specified in Table 1.

The phylogenetic tree based on Shannon entropy and its associated clusters are shown in Figures 17 and 18, respectively. In this scenario, *Cricetulus griseus*, *Rattus norvegicus*, and *Grammomys surdaster* are grouped under a single clade, and these species are closely related to *Sus scrofa*, *Mus pahari*, and *Arvicanthis niloticus*. Similarly, *Mus musculus* and *Xenopus tropicalis* are closely related to *Bos taurus*, *Homo sapiens*, and *Meleagris gallopavo*. The species *Pongo abelii* formed a singleton cluster.

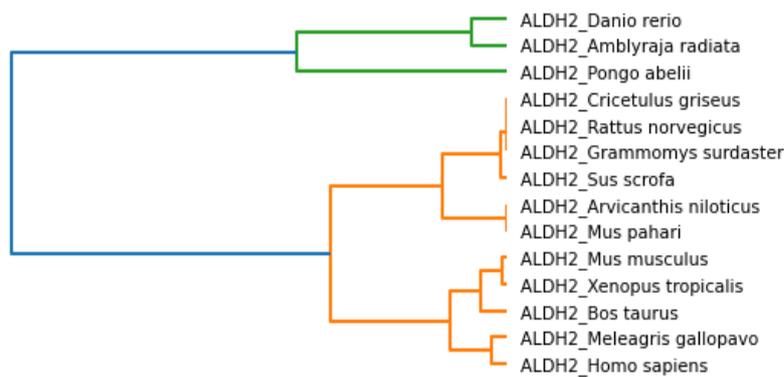


Figure 17. Hierarchical clustering phylogenetic tree based on Shannon entropy.

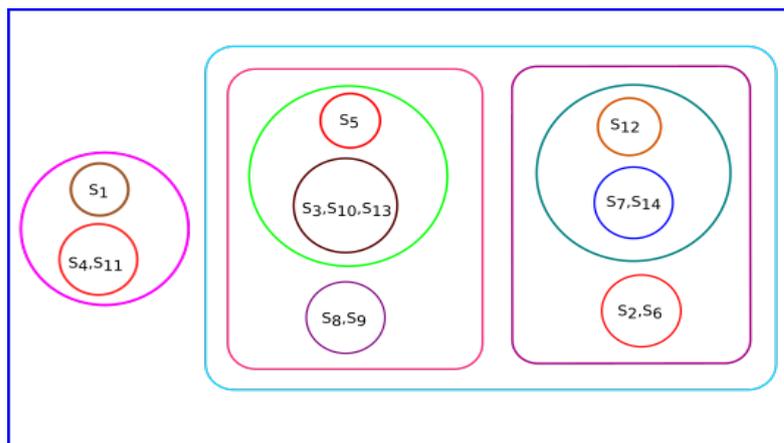


Figure 18. Clusters drawn from Hierarchical clustering phylogenetic tree based on Shannon entropy. Here, S_i indicates the species name as specified in Table 1.

4. Discussion

This paper comprehensively analyzed ALDH2 sequences of fourteen species, including *Homo sapiens* (human). The other species are *Pongo abelii*, *Rattus norvegicus*, *Amblyraja radiata*, *Sus scrofa*, *Meleagris gallopavo*, *Xenopus tropicalis*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Danio rerio*, *Bos taurus*, *Grammomys surdaster*, and *Mus musculus*.

ALDH2 is an enzyme needed for alcohol detoxification. ALDH2 removes acetaldehyde, a toxic product from ethanol breakdown [1]. According to recent research, the ALDH2 gene's genetic polymorphism may be highly connected with the risk of developing human cancers such as esophageal, colorectal, and liver cancer [32]. Alcohol-related cancers are much more likely to occur in people with ALDH2 deficiencies. Alcohol flushing syndrome, a hereditary disorder that affects alcohol metabolism, is another name for ALDH2 deficiency. ALDH2 deficiency affects 8% of the global population, with 36% of the population in East Asia being affected. People who have the mutant ALDH2*2 gene are more prone to develop different cancers [33]. A study by Zhang et al. [34] found out the role of ALDH2 and its underlying processes in the progression and occurrence of cancer. A biomarker for cancer stem cells called ALDH2 has been linked to cancer cells' growth, metastasis, and medication resistance. People with ALDH2 deficiency are linked with the chances of cerebral stroke and cardiovascular diseases [35]. Targeting ALDH2 may be a potential strategy to prevent stroke trauma and cancer. The proposed method carries out a feature generation process based on several quantitative metrics that best depict the evolutionary relationships among the species.

From the analysis of ALDH2 sequences of various species, from Figures 4, 7, 10, 13, and 16, we observed that the species *Rattus norvegicus*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, and *Grammomys surdaster* always belonged to the same cluster when K-means clustering was applied based on physicochemical properties, Shannon entropy, Hurst exponent, fractal dimension, and secondary structure. From Figures 10, 13 and 16, we found that some species formed singleton clusters; for example, *Pongo abelii* was not grouped with other species in the case of statistical measures, i.e., Shannon entropy, Hurst exponent, and fractal dimension. From Figures 7 and 10, we observed that *Amblyraja radiata* and *Sus scrofa* species formed a singleton cluster when clustering was done based on secondary structure and Hurst exponent, respectively. As shown in Figure 4, we found that *Homo sapiens* also formed a singleton cluster when all fourteen species were clustered based on physicochemical properties. A more detailed analysis of the fourteen sequences is shown in Table 2.

Table 2. Comprehensive analysis of ALDH2 sequences of fourteen species.

Sequence ID	Summary
S_1^*	<i>Pongo abelii</i> forms a singleton cluster based on statistical measures, i.e., Shannon entropy, Hurst exponent, and fractal dimension.
S_2	<i>Homo sapiens</i> forms a singleton cluster based on physicochemical properties. <i>Homo sapiens</i> is closely related to <i>Rattus norvegicus</i> , <i>Mus pahari</i> , <i>Arvicanthis niloticus</i> , <i>Cricetulus griseus</i> , <i>Grammomys surdaster</i> , and <i>Mus musculus</i> based on secondary structure.
S_3	<i>Rattus norvegicus</i> is more closely related to <i>Mus pahari</i> , <i>Arvicanthis niloticus</i> , <i>Cricetulus griseus</i> , <i>Mus musculus</i> , and <i>Grammomys surdaster</i> .
S_4	Based on secondary structure, <i>Amblyraja radiata</i> forms a singleton cluster. Based on other features, <i>Amblyraja radiata</i> is closely related to <i>Danio rerio</i> .
S_5	<i>Sus scrofa</i> forms a singleton cluster based on Hurst exponent. Based on physicochemical properties and secondary structure, <i>Sus scrofa</i> is more closely related to <i>Bos taurus</i> .
S_6	<i>Meleagris gallopavo</i> is closely related to <i>Pongo abelii</i> based on physicochemical properties.
S_7	<i>Xenopus tropicalis</i> is related to <i>Danio rerio</i> based on physicochemical properties, secondary structure, and fractal dimension.
S_8	<i>Mus pahari</i> is more closely related to <i>Rattus norvegicus</i> , <i>Arvicanthis niloticus</i> , <i>Cricetulus griseus</i> , <i>Mus musculus</i> , and <i>Grammomys surdaster</i> .
S_9	<i>Arvicanthis niloticus</i> is more closely related to <i>Rattus norvegicus</i> , <i>Mus pahari</i> , <i>Cricetulus griseus</i> , <i>Mus musculus</i> , and <i>Grammomys surdaster</i> .
S_{10}	<i>Cricetulus griseus</i> is more closely related to <i>Rattus norvegicus</i> , <i>Mus pahari</i> , <i>Arvicanthis niloticus</i> , <i>Mus musculus</i> , and <i>Grammomys surdaster</i> .
S_{11}	<i>Danio rerio</i> is closely related to <i>Xenopus tropicalis</i> based on secondary structure, fractal dimension, and Hurst exponent. It is closely related to <i>Amblyraja radiata</i> based on physicochemical properties and Shannon entropy.
S_{12}	<i>Bos taurus</i> is more closely related to <i>Sus scrofa</i> according to physicochemical properties. <i>Bos taurus</i> is more closely related to <i>Homo sapiens</i> based on the statistical measures.
S_{13}	<i>Grammomys surdaster</i> is more closely related to <i>Rattus norvegicus</i> , <i>Mus pahari</i> , <i>Cricetulus griseus</i> , <i>Mus musculus</i> , and <i>Arvicanthis niloticus</i> .
S_{14}	<i>Mus musculus</i> is closely related to <i>Rattus norvegicus</i> , <i>Mus pahari</i> , <i>Arvicanthis niloticus</i> , <i>Cricetulus griseus</i> , and <i>Grammomys surdaster</i> .

* Here, S_i represents the species name as specified in Table 1.

In Section 3, we observed that all fourteen species were clustered based on various parameters, i.e., physicochemical properties, secondary structure, Hurst exponent, fractal dimension, and Shannon entropy. To conclude the final analysis, a consensus tree has been built from multiple phylogeny trees generated by hierarchical clustering. Phylogeny trees based on all parameters were combined to make a consensus tree. Figure 19 shows the majority consensus tree. The consensus tree shows that the *Homo sapiens* species is more closely related to the *Bos taurus* and *Sus scrofa* species.

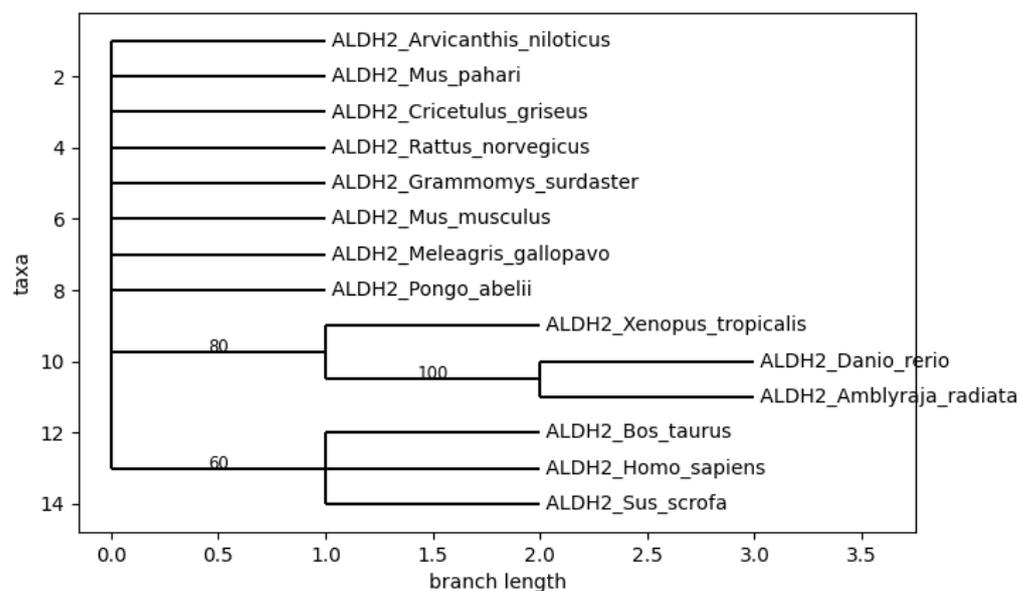


Figure 19. Majority consensus tree built from multiple trees generated using physicochemical properties, secondary structure, fractal dimension, Shannon entropy, and Hurst exponent.

A simple neighbor-joining tree with 1000-times bootstrap analysis was built for the phylogenetic analysis of the primary sequences. The bootstrap tree is shown in Figure 20. The tree was built using MEGA 11 (Molecular Evolutionary Genetics Analysis) software [36]. The consensus tree was compared with the bootstrap tree to find the evolutionary analysis of the proposed model. From Figure 20, we could observe that *Homo sapiens* is more closely related to the *Pongo abelii*, *Bos taurus*, and *Sus scrofa* species. Our results from Figure 19 show that *Homo sapiens* is closely related to *Bos taurus* and *Sus scrofa*, which can be verified from the standard results of the bootstrap analysis of the primary sequences. The hierarchical clustering phylogenetic tree results based on physicochemical properties shared similarities with standard bootstrap analysis. As shown in Figure 6, *Bos taurus* and *Sus scrofa* are closely related to each other and then also grouped with *Pongo abelii* and *Homo sapiens*.

The statistical parameters were used for feature extraction and for the classification point of view; they are also used in other papers [20,24,25] for biological evolutionary analysis of the species. In the near future, other available ALDH2 genes can be studied and will strengthen the observations reported here. Our results show that *Homo sapiens* is more closely related to the *Bos taurus* and *Sus scrofa* species, which can be seen from the consensus tree. Experimentally it has been concluded that the testing for discovering medicines may be done on these species before testing in humans to alleviate the impacts of ALDH2 deficiency. It was observed that they share evolutionary closeness.

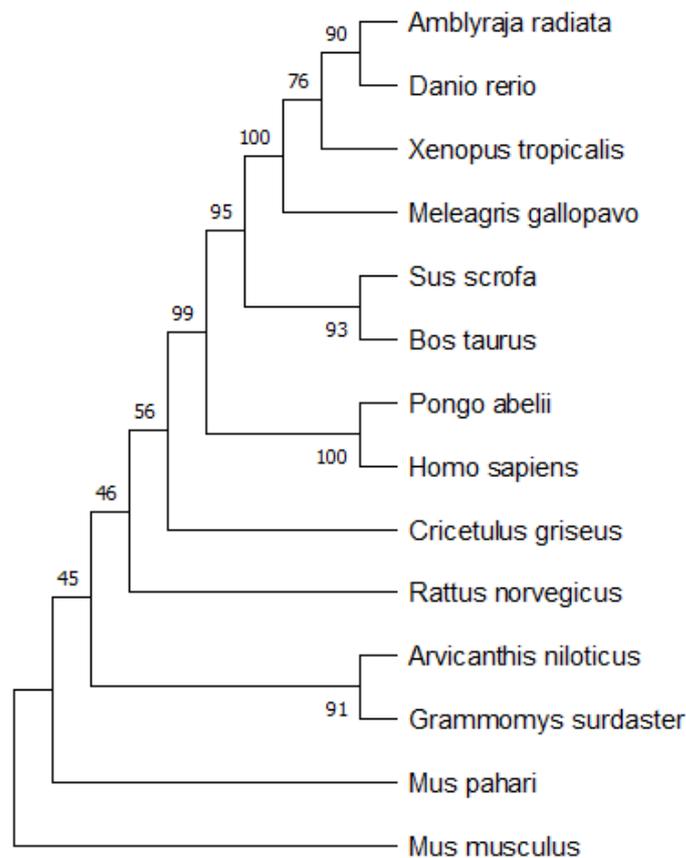


Figure 20. A neighbor-joining tree with 1000-times bootstrap analysis of the primary sequences by using MEGA 11 software [36].

5. Conclusions

This research aimed to develop an efficient alignment-free tool in protein sequence comparison and phylogenetic study. A comprehensive analysis of ALDH2 sequences of fourteen species, including *Homo sapiens* (human), was carried out. The proposed method performs a feature generation process based on the various quantitative metrics (physicochemical properties, secondary structure, Hurst exponent, Shannon entropy, and fractal dimension) properties of amino acids that best describe the evolutionary relationship among the species in these protein families. The results show that some species always belong to the same cluster even if we consider secondary structure, physicochemical properties, Shannon entropy, fractal dimension, or Hurst exponent. These species are *Mus pahari*, *Rattus norvegicus*, *Arvicanthis niloticus*, *Cricetulus griseus*, and *Grammomyssurdaster*. *Homo sapiens* shows similarity to *Amblyraja radiata*, *Xenopus tropicalis*, *Danio rerio*, and *Bos taurus* when all species are clustered with the Hurst exponent using the K-means algorithm. If we consider Shannon entropy, then *Homo sapiens* shows similarity to *Meleagris gallopavo*, *Mus musculus*, *Bos taurus*, and *Xenopus tropicalis*. *Homo sapiens* is clustered with *Bos taurus* in the case of fractal dimension, but if we consider secondary structure, then *Homo sapiens* forms a cluster with *Rattus norvegicus*, *Mus pahari*, *Arvicanthis niloticus*, *Cricetulus griseus*, *Grammomyssurdaster*, and *Mus musculus*. Using all phylogenetic trees generated by various features, i.e., physicochemical properties, secondary structure, and statistical measures, a consensus tree was built to summarize the results. *Homo sapiens* is more closely related to the *Bos taurus* and *Sus scrofa* species as supported by the consensus tree. Experimentally it has been concluded that the *Bos taurus* and *Sus scrofa* species are the best options for testing for discovering medicines before they are applied to humans to alleviate the impacts

of ALDH2 deficiency. The physicochemical properties clusters share similarities with bootstrap standard results.

Author Contributions: Conceptualization, M.K. and R.K.R.; Formal analysis, M.K., S.S. and R.K.R.; Funding acquisition, Z.Z.; Investigation, R.K.R., S.M. and Z.Z.; Methodology, M.K., S.S., R.K.R., S.U. and S.M.; Writing—review & editing, S.U., S.M. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Z.Z. was partially supported by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734) and the Precision Health Chair Professorship fund. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the entire manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data set used in this study can be found at NCBI (National Center for Biotechnology Information) database [14].

Acknowledgments: We thank all colleagues and researchers within our department of both the labs for their valuable group discussion and suggestion on this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yin, S.J. Alcohol dehydrogenase: enzymology and metabolism. *Alcohol Alcohol.* **1994**, *2*, 113–119.
2. Chen, C.H.; Ferreira, J.C.B.; Gross, E.R.; Mochly-Rosen, D. Targeting aldehyde dehydrogenase 2: New therapeutic opportunities. *Physiol. Rev.* **2014**, *94*, 1–34. [CrossRef] [PubMed]
3. Chang, J.S.; Hsiao, J.R.; Chen, C.H. ALDH2 polymorphism and alcohol-related cancers in Asians: A public health perspective. *J. Biomed. Sci.* **2017**, *24*, 1–10. [CrossRef] [PubMed]
4. Klyosov, A.A.; Rashkovetsky, L.G.; Tahir, M.K.; Keung, W.M. Possible role of liver cytosolic and mitochondrial aldehyde dehydrogenases in acetaldehyde metabolism. *Biochemistry* **1996**, *35*, 4445–4456. [CrossRef] [PubMed]
5. Chang, C.; Ho, T.; Huang, I.; Wu, J. Say No to Glow: Reducing the Carcinogenic Effects of ALDH2 Deficiency. The Oral Cancer Foundation (blog), September 2019. Available online: <https://oralcancernews.org/wp/say-no-to-glow-reducing-the-carcinogenic-effects-of-aldh2-deficiency/> (accessed on 16 January 2022).
6. Jackson, B.C.; Holmes, R.S.; Backos, D.S.; Reigan, P.; Thompson, D.C.; Vasiliou, V. Comparative genomics, molecular evolution and computational modeling of ALDH1B1 and ALDH2. *Chem.-Biol. Interact.* **2013**, *202*, 11–21. [CrossRef]
7. Mallik, S.; Zhao, Z. Graph-and rule-based learning algorithms: A comprehensive review of their applications for cancer type classification and prognosis using genomic data. *Briefings Bioinform.* **2020**, *21*, 368–394. [CrossRef]
8. Jin, S.; Chen, J.; Chen, L.; Histen, G.; Lin, Z.; Gross, S.; Hixon, J.; Chen, Y.; Kung, C.; Chen, Y.; et al. ALDH2 (E487K) mutation increases protein turnover and promotes murine hepatocarcinogenesis. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 9088–9093. [CrossRef]
9. Kim, J.S.; Kim, Y.J.; Kim, T.Y.; Song, J.Y.; Cho, Y.H.; Park, Y.C.; Chung, H.W. Association of ALDH2 polymorphism with sensitivity to acetaldehyde-induced micronuclei and facial flushing after alcohol intake. *Toxicology* **2005**, *210*, 169–174. [CrossRef]
10. Uebelacker, M.; Lachenmeier, D.W. Quantitative determination of acetaldehyde in foods using automated digestion with simulated gastric fluid followed by headspace gas chromatography. *J. Autom. Methods Manag. Chem.* **2011**, *2011*, 907317. [CrossRef]
11. Mallik, S.; Odom, G.J.; Gao, Z.; Gomez, L.; Chen, X.; Wang, L. An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Briefings Bioinform.* **2019**, *20*, 2224–2235. [CrossRef]
12. Mallik, S.; Mukhopadhyay, A.; Maulik, U. RANWAR: Rank-based weighted association rule mining from gene expression and methylation data. *IEEE Trans. Nanobiosci.* **2014**, *14*, 59–66. [CrossRef] [PubMed]
13. Liu, C.; Marioni, R.E.; Hedman, Å.K.; Pfeiffer, L.; Tsai, P.C.; Reynolds, L.M.; Just, A.C.; Duan, Q.; Boer, C.G.; Tanaka, T.; et al. A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **2018**, *23*, 422–433. [CrossRef] [PubMed]
14. O’Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufu, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef] [PubMed]
15. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*; Springer: Berlin, Germany, 2005; pp. 571–607.
16. Guruprasad, K.; Reddy, B.B.; Pandit, M.W. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng. Des. Sel.* **1990**, *4*, 155–161. [CrossRef]
17. Gill, S.C.; Von Hippel, P.H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **1989**, *182*, 319–326. [CrossRef]

18. Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **1980**, *88*, 1895–1898.
19. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [[CrossRef](#)]
20. Hassan, S.; Ghosh, S.; Attrish, D.; Choudhury, P.P.; Aljabali, A.A.; Uhal, B.D.; Lundstrom, K.; Rezaei, N.; Uversky, V.N.; Seyran, M.; et al. Possible transmission flow of SARS-CoV-2 based on ACE2 features. *Molecules* **2020**, *25*, 5906. [[CrossRef](#)]
21. Khandelwal, M.; Rout, R.K.; Umer, S. Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm. In Proceedings of the 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 27–28 January 2022; pp. 268–272.
22. Qian, B.; Rasheed, K. Hurst exponent and financial market predictability. In *IASTED Conference on Financial Engineering and Applications*; Proceedings of the IASTED International Conference: Cambridge, MA, USA, 2004; pp. 203–209.
23. Hurst, H.E. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* **1951**, *116*, 770–799. [[CrossRef](#)]
24. Das, J.K.; Choudhury, P.P.; Chaudhuri, A.; Hassan, S.S.; Basu, P. Analysis of purines and pyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat. *Sci. Rep.* **2018**, *8*, 9974. [[CrossRef](#)]
25. Rout, R.K.; Hassan, S.S.; Sindhwani, S.; Pandey, H.M.; Umer, S. Intelligent classification and analysis of essential genes using quantitative methods. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–21. [[CrossRef](#)]
26. Rout, R.K.; Ghosh, S.; Choudhury, P.P. Classification of Mer Proteins in a Quantitative Manner. *Int. Comput. Appl. Eng. Sci.* **2014**, *4*, 31–34.
27. Rout, R.K.; Pal Choudhury, P.; Maity, S.P.; Daya Sagar, B.; Hassan, S.S. Fractal and mathematical morphology in intricate comparison between tertiary protein structures. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2018**, *6*, 192–203. [[CrossRef](#)]
28. Cattani, C. Fractals and hidden symmetries in DNA. *Math. Probl. Eng.* **2010**, *2010*, 507056. [[CrossRef](#)]
29. Kumar, T.A. CFSSP: Chou and Fasman secondary structure prediction server. *Wide Spectr.* **2013**, *1*, 15–19.
30. Day, W.H.; Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1984**, *1*, 7–24. [[CrossRef](#)]
31. Kodinariya, T.M.; Makwana, P.R. Review on determining number of Cluster in K-Means Clustering. *Int. J.* **2013**, *1*, 90–95.
32. Li, R.; Zhao, Z.; Sun, M.; Luo, J.; Xiao, Y. ALDH2 gene polymorphism in different types of cancers and its clinical significance. *Life Sci.* **2016**, *147*, 59–66. [[CrossRef](#)]
33. Wang, L.S.; Wu, Z.X. ALDH2 and cancer therapy. In *Aldehyde Dehydrogenases*; Springer: Berlin, Germany, 2019; pp. 221–228.
34. Zhang, H.; Fu, L. The role of ALDH2 in tumorigenesis and tumor progression: Targeting ALDH2 as a potential cancer treatment. *Acta Pharm. Sin. B* **2021**, *11*, 1400–1411. [[CrossRef](#)]
35. Xu, H.; Zhang, Y.; Ren, J. ALDH2 and stroke: A systematic review of the evidence. In *Aldehyde Dehydrogenases*; Springer: Berlin, Germany, 2019; pp. 195–210.
36. Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022–3027. [[CrossRef](#)]