1. CONVERSION OF PROTEIN SEQUENCES INTO NUMERICAL EXPRESSIONS

In order to use protein sequences in various bioinformatics and artificial intelligence applications, these sequences need to be converted into numerical expressions. In this direction, there are various categories of protein mapping techniques. There are six different categories: character-based, physicochemical-based, evolution-based, structure-based, machine learning-based and algorithm-based.

1.1. Physicochemical-Based Protein Mapping Techniques

In this method, protein sequences are converted into numerical expressions based on the chemical information of proteins. The methods used in this direction are hydrophobicity, Meiler parameters, Atchley factors and EIIP methods.

1.1.1. Hydrophobicity Protein Mapping Technique

The biological function of a protein is determined by the protein sequence itself and the chemical properties of the protein sequence. Twenty amino acids have hundreds of physicochemical properties. The most important of these are polarisability, Van der Waals volume, ionisation constant, accessible solvent surface area and hydrophobicity. In this method, the hydrophilic and hydrophobic properties of each of the twenty amino acid side chains are considered. The hydrophobicity properties of amino acids are frequently used as an effective way to compare and analyse amino acid sequences. The hydrophobicity protein quantification technique consists of one-dimensional numerical data. Table 1 shows the hydrophobicity values of amino acids.

Table 1. Hydrophobicity values of amino acids

Amino acid code	Hydrophobicity value	Amino acid code	Hydrophobicity value
A	1.8	M	1.9
C	2.5	N	-3.5
D	-3.5	P	-1.6
E	-3.5	Q	-3.5
F	2.8	R	-4.5
G	-0.4	S	-0.8
H	-3.2	T	-0.7
I	4.5	V	4.2
K	-3.9	W	-0.9
L	3.8	Y	-1.3

In line with the values given in Table 1, a protein sequence of the form $P(N) = [MWWFYPC \ldots]$ is quantified by the hydrophobicity method as $C(N) = [1.9\ -0.9\ 2.8\ -1.3\ -1.6\ 2.5\ \ldots]$.

1.1.2. Atchley Factors

Atchley factors are a method proposed to overcome the metric problems frequently seen in protein sequences. Variance and covariance variables, which are frequently found in protein sequences, are digitised based on the alphabetical ordering of amino acid codes and information theory with the technologies developed in recent years. There are five factors in total in this method. These factors are secondary structure (iy), molecular volume (mh), polarity (p), protein electrostatic charge (ey) and codon density (ky). Since there are five factors, it is a five-dimensional method. Atchley factor values of amino acids are given in Table 2.

Table 2. Atchley factor values of amino acids

Amino acid code	p	Good.	mh	p	ey
A	-0.591	-1.302	-0.733	1.570	-0.146
R	1.538	-0.055	1.502	0.440	2.897
N	0.945	0.828	1.299	-0.169	0.933
D	1.050	0.302	-3.656	-0.259	-3.242
C	-1.343	0.465	-0.862	-1.020	0.255
E	1.357	-1.453	1.477	0.113	-0.837
Q	0.931	-0.179	-3.005	-0.503	-1.853
G	-0.384	1.652	1.330	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-0.078
I	-1.239	-0.547	2.131	0.393	0.816
L	-1.019	-0.987	-1.505	1.266	-0.912
K	1.831	-0.561	0.533	-0.277	1.648
M	-0.663	-1.524	2.219	-1.005	1.212
F	-1.006	-0.590	1.891	-0.397	0.412
P	0.189	2.081	-1.628	0.421	-1.392
S	-0.228	1.399	-4.760	0.670	-2.647
T	-0.032	0.326	2.213	0.908	1.313
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.260	0.830	3.097	-0.838	1.512
V	-1.337	-0.279	-0.544	1.242	-1.262

In line with the values given in Table 2, a protein sequence of the form $P(N) = [ARN \dots]$, Atchley factors with $C(N) = [[-0.591\ -1.302\ -0.733\ 1.570\ -0.146][1.538\ -0.055\ 1.502\ 0.440\ 2.897][0.945\ 0.828\ 1.299\ -0.169\ 0.933]]$

1.1.3. EIIP Protein Mapping Technique

The EIIP protein quantification method was first used to determine the interactions between DNA and proteins. In the first stage of the EIIP method, protein sequences were decomposed into signals by Fourier transform. After decomposition, power spectrum density values were calculated from the signals. Finally, each power spectrum value was assigned to the relevant amino acids and digitisation was performed. Table 3 shows the EIIP values of the amino acids.

Table 3. EIIP values of amino acids

Amino acid code	EIIP value	Amino acid code	EIIP value
M	0.0823	Q	0.0761
W	0.0548	S	0.0829
F	0.0946	A	0.0373
Y	0.0516	N	0.0036
P	0.0198	G	0.0050
C	0.0829	R	0.0959
T	0.0941	I	0
H	0.0242	D	0.1263
V	0.0057	E	0.0058
L	0	K	0.0371

In line with the values given in Table 3, a protein sequence in the form of $P(N) = [MWWFYPC \dots]$ is digitised by the EIIP method as $C(N) = [0.0823\ 0.0548\ 0.0946\ 0.0516\ 0.0198\ 0.0829\ \dots]$.

1.2. Evolution-Based Protein Mapping Techniques

In evolution-based protein digitisation methods, protein sequences are digitised based on the evolutionary information of proteins. Evolutionary information of proteins is obtained from phylogenetic trees or sequence alignment. In the evolution-based protein digitisation category, two different digitisation methods, PAM250 (Point Accepted Mutation) and BLOSUM62), were evaluated.

1.2.1. BLOSUM Matrix

BLOSUM matrix is one of the most frequently used methods for aligning protein sequences in bioinformatics. BLOSUM matrices are used to score the alignments between evolutionarily different protein sequences. There are different types of matrices such as BLOSUM80, BLOSUM45. The numbers next to the matrices represent the similarity value. Similarities between the two evolutionarily closest organisms score high. Therefore, the higher the score in BLOSUM matrices, the greater the similarity. For example, the BLOSUM80 matrix is used when aligning the protein sequences of two organisms that are close in species. Conversely, the proteins of two organisms that are distant in species can be aligned with the BLOSUM45 matrix. However, when the distance or proximity of the organisms is not known exactly, the BLOSUM62 matrix is used. BLOSUM62 values of amino acids are given in Table 4.

Table 4. BLOSUM62 values of amino acids

Amino acid code	BLOSUM62 value	Amino acid code	BLOSUM62 value
A	4	L	4
R	5	K	5
N	6	M	5
D	6	F	6
C	9	P	7
Q	5	S	4
E	5	T	5
G	6	W	11
H	8	Y	7
I	4	V	4

In line with the values given in Table 4, a protein sequence of the form $P(N) = [ARNDCQ \dots]$ is digitised by BLOSUM62 method as $C(N) = [4\ 5\ 6\ 6\ 9\ 5 \dots]$.

1.2.2. PAM250 Matrix

The PAM250 matrix is a method used to determine similarities in protein sequences by scoring aligned peptide sequences. The values in the PAM250 matrix were obtained by comparing aligned protein sequences of known homology and identifying observed point mutations. PAM250 values of amino acids are given in Table 5.

Table 5. PAM250 values of amino acids

Amino acid code	PAM250 value	Amino acid code	PAM250 value
A	2	L	6
R	6	K	5
N	2	M	6
D	4	F	9
C	4	P	6
Q	4	S	3
E	4	T	3
G	5	W	17
H	6	Y	10
I	5	V	4

In line with the values given in Table 5, a protein sequence of the form $P(N) = [ARNDCQ \dots]$ is digitised by the PAM250 method as $C(N) = [2\ 6\ 2\ 4\ 4\ 4\ 4 \dots]$.

1.3. Structure-Based Protein Mapping Techniques

In structure-based protein digitisation methods, proteins are digitised based on the structure information of the proteins. This information comes from the primary, secondary, tertiary and quaternary structures of proteins. Although there are several methods in this category, they are used less frequently than the other categories. One of the reasons for this is the difficulty in obtaining protein structure information. Miyazawa energies and Micheletti potentials are the most commonly used methods in this category.

1.3.1. Miyazawa Energies Protein Mapping Technique

Miyazawa energies are a proposed method to determine the energies between residues in protein sequences. In this method, the contact energies of protein sequences are obtained by regression coefficients. In Miyazawa energies, energy values are divided into two terms: secondary structure energies and tertiary structure energies. Tertiary structure energies are obtained by summing the residue-to-residue contact energies of proteins. Secondary structure energies are calculated according to the interactions between atom chains and atom chain-side chain interactions. Miyazawa energy values of proteins are given in Table 6.

Table 6. Miyazawa energy values of amino acids

Amino acid code	Miyazawa energy value	Amino acid code	Miyazawa energy value
A	-0.02	L	-0.32
R	0.08	K	0.30
N	0.10	M	-0.25
D	0.19	F	-0.33
C	-0.32	P	0.11
Q	0.21	S	0.11
E	0.15	T	0.05
G	-0.02	W	-0.27
H	-0.02	Y	-0.23
I	-0.28	V	-0.23

In line with the values given in Table 6, a protein sequence of the form $P(N) = [ARNDCQ \ldots]$ is numericalised with Miyazawa energies as $C(N) =$ [-0.02 0.08 0.10 0.19 - 0.32 0.21 ...].

1.3.2. Micheletti Potentials Protein Mapping Technique

Micheletti potentials, another structure-based protein quantification method, are based on the potential energy in the interactions that occur between proteins. The main idea of Micheletti potentials is to determine the most favourable protein interactions. Micheletti potential values of proteins are given in Table 7.

Table 7. Micheletti potential values of amino acids

Amino acid code	Micheletti potential value	Amino acid code	Micheletti potential value
A	-0.001461	L	-0.000782
R	0.009875	K	0.005109
N	-0.001962	M	0.031655
D	-0.000531	F	-0.013128
C	-0.002544	P	-0.003621
Q	0.006456	S	-0.000802
E	0.008438	T	0.003269
G	0.000990	W	0.131813
H	0.001314	Y	-0.007699
I	0.006801	V	0.001445

In line with the values given in Table 7, a protein sequence of the form $P(N) = [ARNDCNQ \ldots]$ is quantified by Micheletti potentials as $C(N) =$ [-0.001461 0.009875 - 0.001962 - 0.000531 - 0.002544 0.006456 ...].

1.4. Character-Based Protein Mapping Techniques

In character-based protein digitisation methods, protein sequences are digitised without the need for specific information. In the other category of protein digitisation methods, proteins are digitised based on structural information, evolutionary information or chemical information, while there is no such approach in character-based methods. In this category, two protein digitisation methods, CPNR and binary-coding, are discussed.

1.4.1. CPNR Protein Mapping Technique

The CPNR protein quantification method was first proposed and used to compare protein functions. In the proposed protein digitisation method, protein sequences are digitised according to codons. In the first step of the digitisation process, amino acids are decomposed into codon numbers. Then, prime numbers are assigned to amino acid sequences according to codon numbers. CPNR values of amino acids are given in Table 8.

Table 8. CPNR values of amino acids

Amino acid code	CPNR value	Amino acid code	CPNR value
M	1	Q	29
W	2	S	31
F	3	A	37
Y	5	N	41
P	7	G	43
C	11	R	47
T	13	I	53
H	17	D	59
V	19	E	61
L	23	K	67

In line with the values given in Table 8, a protein sequence of the form $P(N) = [MWWFYPC \dots]$ is digitised as $C(N) = [1\ 2\ 3\ 5\ 7\ 11 \dots]$ by CPNR protein mapping technique.

1.4.2. One-Hot Protein Mapping Technique

In the one-hot protein mapping technique, amino acids in protein sequences are represented by "0" and "1" values. It is the most widely used binary coding method. In this method, each of the twenty standard amino acids is digitised with a twenty-dimensional binary vector. In the first step of the digitisation process, the amino acids are sorted alphabetically. Then, the value of the amino acid in the *ith* row takes the value "1", while the others take the value "0". One-hot values of amino acids are given in Table 9.

Table 9. One-hot values of amino acids

Amino acid code	One-hot value	Amino acid code	One-hot value
A	10000000000000000000	M	00000000001000000000
C	01000000000000000000	N	00000000000100000000
D	00100000000000000000	P	00000000000010000000
E	00010000000000000000	Q	00000000000001000000
F	00001000000000000000	R	00000000000000100000
G	00000100000000000000	S	00000000000000010000
H	00000010000000000000	T	00000000000000001000
I	00000001000000000000	V	00000000000000000100
K	00000000100000000000	W	00000000000000000010
L	00000000010000000000	Y	00000000000000000001

In line with the values given in Table 9, a protein sequence of the form $P(N) = [AC \ldots]$, one-hot protein Mapping Technique with It is numerised as $C(N) = [10000000000000000000\ 01000000000000000000 \ldots]$.

1.4.3. Whole Number Protein Mapping Technique

In this method, amino acid codes are sorted alphabetically and after sorting, integer numbers from 1 to 20 are assigned to these codes. Table 10 shows the integer values of amino acids.

Table 10. Integer values of amino acids

Amino acid code	Integer value	Amino acid code	Integer value
A	1	L	11
R	2	K	12
N	3	M	13
D	4	F	14
C	5	P	15
Q	6	S	16
E	7	T	17
G	8	W	18
H	9	Y	19
I	10	V	20

In line with the values given in Table 10, a protein sequence in the form of $P(N) = [ARNDCQ \ldots]$ is digitised as $C(N) = [1\ 2\ 3\ 4\ 5\ 6 \ldots]$ by integer protein mapping technique.

1.5. Algorithm-Based Protein Mapping Techniques

In algorithm-based protein mapping techniques, amino acid sequences are converted into numerical expressions based on a specific algorithmic structure. As in character-based approaches, no specific information is required. However, the biggest advantage of this technique is that it performs a fast digitisation process. The protein mapping techniques evaluated in this category are AVL-tree and FIBHASH methods.

1.5.1. AVL-Tree Protein Mapping Technique

In the AVL-tree protein mapping technique, amino acids were first sorted alphabetically and added to the AVL-tree. After all amino acids were added to the AVL-tree, necessary operations were performed and the tree was stabilised. Figure 1 shows the addition of amino acids to the AVL-tree.

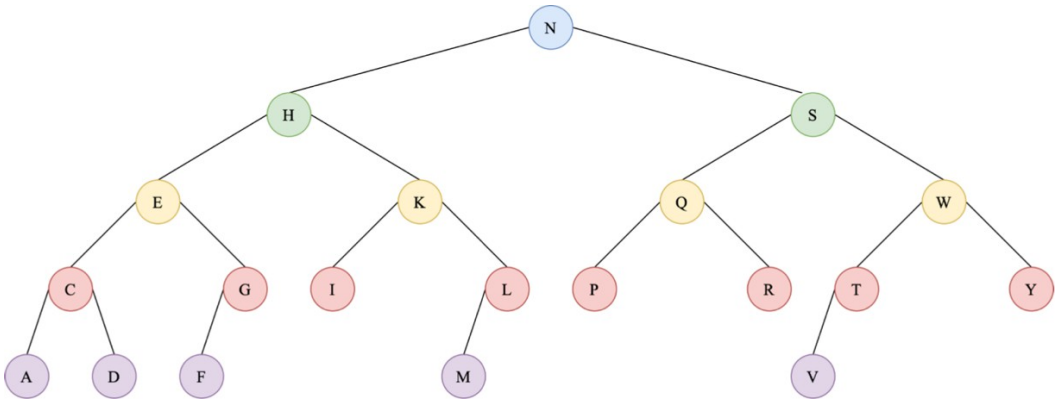


Figure 1. Addition of amino acids to the AVL-tree

After being added to the tree, the depth values of each amino acid were calculated and the amino acid sequences were converted into numerical expressions. AVL-tree values of amino acids are given in Table 11.

Table 11. AVL-tree values of amino acids

Amino acid code	Integer value	Amino acid code	Integer value
A	4	L	3
R	3	K	2
N	0	M	4
D	4	F	4
C	3	P	3
Q	2	S	1
E	2	T	3
G	3	W	2
H	1	Y	3
I	3	V	4

In line with the values given in Table 11, a protein sequence of the form $P(N) = [ARNDCQ \ldots]$ is digitised by AVL-tree protein mapping technique as $C(N) = [4\ 3\ 0\ 4\ 3\ 2 \ldots]$.

1.5.2. FIBHASH Protein Mapping Technique

In the FIBHASH protein mapping technique, Fibonacci numbers were assigned to amino acid sequences in the first step. Since there are twenty amino acids in nature, values up to the twentieth number of the Fibonacci series were considered. However, after a certain series, since the size of the numbers increases, these values were assigned to the hash table, which is a table frequently used in algorithm analysis. In this way, the data was ensured to be between 1 and 20. The values of the amino acid sequences coming to the same index were transferred to new indexes by linear probing. Table 12 shows the FIBHASH values of the amino acids.

Table 12. FIBHASH values of amino acids

Amino acid code	Integer value	Amino acid code	Integer value
A	1	L	15
R	10	K	14
N	7	M	9
D	3	F	5
C	2	P	16
Q	17	S	11
E	4	T	18
G	8	W	19
H	13	Y	20
I	6	V	12

In line with the values given in Table 12, a protein sequence of the form $P(N) = [ARNDCQ \ldots]$ is digitised as $C(N) = [1\ 10\ 7\ 3\ 2\ 17 \ldots]$ by FIBHASH protein mapping technique.

Application Questions

1. Convert the protein sequence WVTTLEHLKGPHAQI into a numerical expression using the hydrophobicity protein mapping technique.

2. Convert the protein sequence given as LAQM into a numerical expression using the Atchley factors protein mapping technique.

3. Convert the protein sequence given as DFKFWNYQMGRQNRI into a numerical expression using EIIP protein mapping technique.

4. Convert the protein sequence given as GHDDKVGAMDMELPA into a numerical expression using BLOSUM62 protein mapping technique.

5. Convert the protein sequence given as DLQKKFPWLYYWLRGC into a numerical expression using the PAM250 protein mapping technique.

6. Convert the protein sequence given as FKGDEMTHKEKPFSY into a numerical expression using the Miyazawa energies protein mapping technique.

7. Convert the protein sequence given as HERGKVCPFMQEIMA into a numerical expression using the Micheletti potentials protein mapping technique.

8. Convert the protein sequence CTTVMIPQHGNAAHF into a numerical expression using CPNR protein mapping technique.

9. Convert the protein sequence given as WNCQH into a numerical expression using a hot protein mapping technique.

10. Convert the protein sequence given as SFWYFLGQSCSWLKE into a numerical expression using the integer protein mapping technique.

11. Convert the protein sequence YAGFQPQNMLLGPIQ into a numerical expression using the AVL-tree protein mapping technique.

12. Convert the protein sequence given as GVSPADAESCIIMQY into a numerical expression using FIBHASH protein mapping technique.

13. Develop a new method based on red-black trees to digitise a protein sequence given as PMWRITIYFKQQAKIP. The method to be developed will only cover the given protein sequence. Firstly, the amino acid sequences will be sorted alphabetically and placed in the tree accordingly. Finally, the depth value of each amino acid code in the tree will be determined and the digitisation process will be completed.

14. Develop a new Lucas series based method to quantify a protein sequence given as YAHTGVPSNHNILVWNNGVSKKRNE. The method to be developed will only cover the given protein sequence. Firstly, the amino acid sequences will be sorted alphabetically and Lucas numbers will be assigned to the amino acid codes. Since the size of the Lucas numbers is large, the hash table will be used and the digitisation process will be completed. In case of any conflict in the table, quadratic probing will be used.

Solutions

1. The protein sequence given as WVTTLEHLKGPHAQI is converted into a numerical expression by hydrophobicity protein mapping technique as follows:

C(N) = [-0.9 4.2 -0.7 -0.7 -0.7 3.8 -3.5 -3.5 -3.2 3.8 -3.9 -0.4 -1.6 -3.2 1.8 -3.5 4.5]

2. The protein sequence given as LAQM is converted into a numerical expression by Atchley factors protein mapping technique as follows:

C(N) = [[-1.019 -0.987 -1.505 1.266 -0.912] [-0.591 -1.302 -0.733 1.570 -0.146] [0.931 -0.179 -3.005 -0.503 -1.853] [-0.663 -1.524 2.219 -1.005 1.212]]

3. The protein sequence given as DFKFWNYQMGRQNRI is converted into a numerical expression by EIIP protein mapping technique as follows:

C(N) = [0.1263 0.0946 0.0371 0.0946 0.0548 0.00360.0516 0.0761 0.0823 0.0050 0.0959 0.0761 0.0036 0.0959 0]

4. The protein sequence given as GHDDKVGAMDMELPA is converted into numerical expression by BLOSUM62 protein mapping technique as follows:

C(N) = [6 8 6 6 6 5 4 6 4 5 6 4 5 6 5 5 5 4 7 4]

5. The protein sequence given as DLQKKFPWLYYWLRGC is converted into numerical expression by PAM250 protein mapping technique as follows:

C(N) = [4 6 4 4 5 9 6 17 6 10 10 17 6 6 5 4]

6. The Miyazawa energies of the protein sequence given as FKGDEMTHKEKPFSY are converted into numerical expression by protein mapping technique as follows:

C(N) = [-0.33 0.30 -0.02 0.19 0.15 -0.25 0.05 -0.02 0.30 0.15 0.30 0.11 -0.33 0.11 -0.23]

7. The Micheletti potentials of the protein sequence given as HERGKVCPFMQEIMA are converted into numerical expressions by protein mapping technique as follows:

C(N) = [0.001314 0.008438 0.009875 0.000990 0.005109 0.001445 -0.002544 -0.003621 -0.013128 0.031655 0.006456 0.008438 0.006801 0.031655 -0.001461]

8. The protein sequence given as CTTVMIPQHGNAAHF is converted into numerical expression by CPNR protein mapping technique as follows:

C(N) = [11 13 13 13 19 1 53 7 29 17 43 41 41 37 37 37 17 3]

9. The protein sequence given as WNCQH is converted into a numerical expression by a hot protein mapping technique as follows:

C(N) = [00000000000000000010 00000000000100000000 01000000000000000000 00000000000001000000 00000100000000000000]

10. The protein sequence given as SFWYFLGQSCSWLKE is converted into a numerical expression by integer protein mapping technique as follows:

C(N) = [16 18 14 19 14 11 8 6 16 5 16 18 11 12 7]

11. The protein sequence given as YAGFQPQNMLLGPIQ is converted into numerical expression by AVL-tree protein mapping technique as follows:

C(N) = [3 4 3 3 4 2 3 2 3 2 0 4 3 3 3 3 3 3 2]

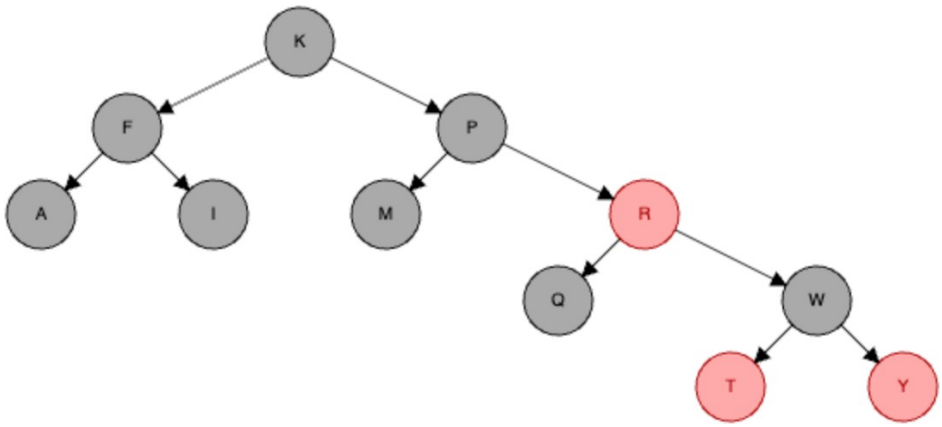
12. The protein sequence given in the form of GVSPADAESCIIMQY is converted into a numerical expression by FIBHASH protein mapping technique as follows:

C(N) = [8 12 11 16 1 3 1 4 11 2 6 6 9 17 20]

13. A protein sequence given as PMWRITIYFKQQAKIP is first sorted alphabetically. The sorting process is as follows:

A, F, I, K, M, P, Q, R, T, W and Y

Then red-black is added to the tree respectively. After the addition process, the final state of the tree is as follows. Then the depth value is calculated for each node and the digitisation process is completed.



C(N) = [1 2 3 2 2 2 2 4 2 4 1 0 3 2 0 2 1]

14. A protein sequence given as YAHTGVPSNHNILVWNNGVSKKRNE is first sorted alphabetically. After the sorting process, the sequences are as follows: A, E, G, H, I, K, L, N, P, R, R, S, T, V, W, Y

Then the numbers in the Lucas series are added to each codon and the following values are obtained:

A = 2, E = 1, G = 3, H = 4, I = 7, K = 11, L = 18, N = 29, P = 47, R = 76, S = 123, T = 199, V = 322, W = 521, Y = 843

Then the hash table is created and values are assigned. Since there are 15 amino acids in total, the size of the hash table will be 15. Overlaps are removed by quadrature probing. Accordingly, the value required for digitisation is as follows:

W = 0, E = 1, A = 2, G = 3, H = 4, L = 5, P = 6, I = 7, T = 8, Y = 9, R = 10, K = 11, S = 12, V = 13, N = 14

The last digitisation process is carried out as follows:

C(N) = [9 2 4 8 3 13 6 12 14 4 14 7 5 13 0 14 14 3 13 12 11 11 10 13 1]