Protein Encodings

-Shreyasi Ghosh

1. Taylor’s Venn Diagram

- It is based on 10 physiochemical groups. They are:
  a. Hydrophobic
  b. Positive
  c. Negative
  d. Polar
  e. Charged
  f. Small
  g. Tiny
  h. Aliphatic
  i. Aromatic
  j. Proline
- The 20 natural amino acids belong the above group.

- The amino acids are encoded as binary vectors of length 10.

$$TVD_p(t) = \begin{cases} 1 & \text{if } t \in p \\ 0 & \text{if } t \notin p \end{cases}, \qquad t \in A$$

Where p is the property, A is the set of 20 natural amino acids, N is the sequence length

All sequences must have same length.

Example

Amino Acid: A

Binary Vector: 1000000000

TVD is 1 as t belongs to hydrophobic group.

2. Secondary Structure Elements Binary (SSEB)

- This method represents each amino acid depending on the secondary structure element as a vector of 3 binary digits.
  - Helix(001)
  - Sheet(010)
  - coil(100)
- Vector Length is 3N, N is the sequence length.

All the sequences must have same length.

Example:

Amino Acid:	A	B	C	D
Secondary Structure:	Helix	Sheet	Coil	Helix
Binary Vector:	001	010	100	001

3. Secondary Structure Elements Content (SSEC)

- This method calculates the frequency of each secondary structure element type (helix, sheet, coil) found in the peptide.

$$SSEC(e) = \frac{N(e)}{N}, \qquad e \in Helix, Sheet, Coil$$

  Where N(s) is the number of times the element e appears in the sequence , N is the sequence length.

- Vector length is 3

Example

Peptide Sequence: AABBCCHHH

So, N(helix)=3(H appears 3 times)

N(Sheet)=2(B appears two times)

N(coil)=4(A and C appears two times each)

According to the given formula,

SSEC(helix)=3/9

SSEC(sheet)=2/9

SSEC(coil)=4/9

4. Secondary Structure Probabilities Bigram(SSBP)

- This method sums over the multiplication of probabilities for each of the **combination** between structural elements **among pairs** of amino acids separated by n residues.
- Initially n was set to 1 , meaning the calculation was for adjacent amino acids.

$$SSPB(e,f) = \frac{1}{N}\sum_{i=1}^{N-n} P_i(e) * P_{i+n}(f), \qquad e,f \in \{helix, coil, sheet\}$$

Where $P_i(e)$ and $P_{i+n}(f)$ are probabilities of amino acids at i and i+n in the sequence having e and f and N is the sequence length.

- Vector length is 9.

Example

Amino Acid: A B C D E

Probabilities: (0.3, 0.4, 0.3), (0.2, 0.6, 0.2), (0.5, 0.3, 0.2), (0.1, 0.7, 0.2), (0.4, 0.2, 0.4)

These probabilities indicate each amino acid belonging to Helix,Coil,Sheet.

According to the formula:

N=5

$SSPB(H,C)=\frac{1}{5}(P_1(H).P_2(C)+P_2(H).P_3(C)+P_3(H).P_4(C)+P_4(H).P_5(C))$

Likewise, we need to calculate for other pairs

5. Secondary Structure Probabilities Auto-Covariance(SSPAC)

- This method sums the multiplication of the probabilities for **each** structural element among the pairs of amino acids separated by n residues
- n ranges from 1 to N.

$$SSPAC(n,e) = \frac{1}{L}\sum_{i=1}^{L-n} P_i(e) * P_{i+n}(e), \qquad 1 \leq n \leq N, e \in helix, coil, sheet$$

Where $P_i(e)$ and $P_{i+n}(e)$ are the probabilities of the amino acids at positions i and i + n in the sequence having the element e, N is the maximum value for the separation between residues, and L is the sequence length.

- Vector Length is 3N

Example:

Peptide Sequence: ABCDEFGHIJKLMN

Probabilities for Helix (P(H)):

(0.2, 0.4, 0.3, 0.6, 0.8, 0.1, 0.7, 0.5, 0.4, 0.3, 0.6, 0.2, 0.9, 0.3, 0.5)

N=6,

According to the given formula,

For n=1

SSPAC(1,H)=1/14((0.20*0.4)+(0.4*0.3)+.....+(0.3*0.5))

This calculation has to be continued till n=6.

6. Disorder

- This method reads the probability values per amino acid and adds them to the vector.
- Vector length is N, where N is the sequence length.

  All sequences must have the same length.

Example

Peptide Sequence: ABCDE

Disorder Vector: (0.1, 0.4, 0.7, 0.2, 0.6)

7. Disorder Content (DisorderC)

- This method calculates the frequency of ordered and disordered residue in the sequence.
- Vector length 2

$$DisorderC(d) = \frac{N(d)}{N}, \qquad d \in order, disorder$$

Where N(d) is the number of ordered and disordered residues in the sequence, and N is the sequence length.

Example:

If we have 100 residues and 40 of them are classified as disordered, then DisorderC is 0.4

8. Disordered Binary (DisorderB)

- This methods encodes each amino acid as a binary vector of length 2
- If the residue is ordered, then the encoding is [1,0] else it is [0,1]
- Vector length is 2N, N is the sequence length.

Example:Peptide Sequence: ABC

B and C are disordered

Disorder Binary Encoding: [1, 0, 0, 1,0,1]

Torsional Angles (TA)

- This method adds the phi and psi values per amino acid to the vector
- Vector length is 2N,N is the sequence length.

Example:

Protein Sequence: A1,A2,A3

A1:Φ1,ψ1

A2:Φ2,ψ2

A3:Φ3,ψ3

Now, the vector would be

[Φ1,ψ1,Φ2,ψ2,Φ3,ψ3]

10. Torsional Angles Composition (TAC)

- This method converts the phi and psi values of each amino acid from degrees to radians.
- It also calculates the sines and cosines of these two angles, divides these values by the length of the sequence and adds four final values to the vector.

$$TAC(f,a) = \frac{1}{N}\sum_{i=1}^{N} f(\frac{a_i \pi}{180}), \qquad f \in \{sin, cos\}, a \in \{phi, psi\}$$

Where a_i is the phi or psi value for the amino acid at position in the sequence and N is the sequence length.

- Vector length is 4

Example:

For simplicity, let's assume the values of Φ,ψ for each amino acid

A1:$\Phi1=60^o$,$\psi1=-45^o$

A2:$\Phi2=-30^o$,$\psi2=0^o$

A3:$\Phi3=45^o$,$\psi3=30^o$

First convert the angles into radians and then according to the above formula

TAC(sin,Φ)=⅓(sin(Π/3) + sin(-Π/6)+sin(Π/4))

TAC (sin,ψ)=⅓(sin(-Π/4) + sin(0)+sin(Π/6))

Likewise, we have to calculate the values for TAC(cos,Φ) and TAC(cos,ψ) and construct the vector

[TAC(sin,Φ),TAC(cos,Φ),TAC (sin,ψ),TAC(cos,ψ)]

11. Torsional Angles Bigram(TAB)

- It converts phi and psi values per amino acid from degrees to radians, and calculates the sine and cosine of these two angles, so each amino acid has 4 associated values.
- Then each type of value is multiplied **as pairs** in the sequence separated by n residues, and finally divided by the sequence length
- The value of n was originally 1.

$$TAB(f,g,a,n) = \frac{1}{N}\sum_{i=1}^{N-n} f(\frac{a_i\pi}{180}) * g(\frac{a_{i+n}\pi}{180}), \qquad f,g \in \{sin,cos\}, a \in \{phi,psi\}$$

Where a_i and a_{i+n} are the phi or psi values for the amino acid at position i and i+n in the sequence, N is the sequence length.

- Vector length is 10

Example:

Considering there are 4 amino acids and their angles:

Φ=[-60º,-45º,30º,-15º,45º]

ψ=[45º,-30º,60º,-45º,-30º]

N = 5 (sequence length)

n = 2

TAB(sin, cos, phi, 2)

= 1/5 * [sin(-60 * π / 180) * cos(30 * π / 180) + sin(-45 * π / 180) * cos(-15 * π / 180) + sin(30 * π / 180) * cos(45 * π / 180)]

Similarly, other pairs are calculated.

13. Torsional Angles Autocovariance(TAAC)

- This method converts the phi and psi values per amino acid from degrees to radians, and calculates the sine and cosine of these two angles, so each amino acid has 4 associated values.
- Then, it sums the multiplication of each type of value among the pairs of amino acids separated by n residues, where n ranges from 1 to N.

$$TAAC(f,a,n) = \frac{1}{L}\sum_{i=1}^{L-n} f(\frac{a_i\pi}{180}) * f(\frac{a_{i+n}\pi}{180}), \qquad f \in \{sin, cos\}, a \in \{phi, psi\}, 1 \leq n \leq N$$

Where, a_i and a_{i+n} are the phi and psi values for the amino acids at position i and i+n in the sequence,N is the maximum value for the separation between residues and L is the sequence length.

- Vector length is 4N

Example:

Φ=[-60º,-45º,30º,-15º]

ψ=[45º,-30º,60º,-45º]

For n=1, calculating according to the above formula

L = 4 (sequence length)

N = 1 (maximum separation between residues)

TAAC(sin, phi, 1)

= 1/4 * [sin(-60 * π / 180) * sin(-45 * π / 180) + sin(-45 * π / 180) * sin(30 * π / 180) + sin(30 * π / 180) * sin(-15 * π / 180)]

Likewise, we calculate for other n values

14. Accessible Surface Area(ASA)

- This method reads the ASA values per amino acid and adds them to the vector.
- Vector length is N, N is the sequence length.

Example:

ASA1:20

ASA2:15

ASA3:25

ASA4:18

ASA5:22

Then the resulting vector is [20,15,25,18,22]

15. Bigram PSSM(BiPSSM)

- This method sums the product between the PSSM values of two residues in the sequence separated by n characters for two amino acid types and divides that sum by the sequence length
- The value of n was originally 1.

$$BiPSSM(t,u) = \frac{1}{N}\sum_{i=1}^{N-n} s_{i,t} * s_{i+n,u}, \qquad t,u \in A$$

Where A is the set of 20 natural amino acids, $s_{i,t}$ and $s_{i+n,u}$ are the scores in the PSSM matrix for the amino acids t and u at positions i and i + n respectively, and N is the sequence length.

- Vector length is 400

Example:

PSSM Matrix Scores:

Position	A	G
1	0.2	0.1
2	0.5	0.3
3	0.7	0.4
4	0.3	0.6
5	0.2	0.5

Peptide Sequence: AGGAA

BiPSSM=⅕((0.2*0.3)+(0.5*0.4)+(0.7*0.6)+(0.3*0.5))

16. PSSM Autocovariance (PSSMAC)

- This method calculates the autocovariance between two residues separated by n characters for a specific amino acid type.

$$\bar{s}_t = \frac{1}{N} sum_{i=1}^{N} s_{i,t}, \qquad t \in A$$

$$PSSMAC(t,n) = \sum_{i=1}^{N-n} \frac{(s_{i,t} - \bar{s}_t) * (s_{i+n,t} - \bar{s}_t)}{N-n}, \qquad t \in A$$

Where A is the set of the 20 natural amino acids, $s_{i,t}$ and $s_{i+n,t}$ are the scores in the PSSM matrix for the amino acid t at positions i and i+n, and N is the sequence length.

- Vector length is 400

Example:

PSSM Matrix Scores for A:

Position	A
1	0.2
2	0.5
3	0.7
4	0.3
5	0.2

Peptide Sequence: AGGAA

$\bar{S}_t$ = ⅕(0.2+0.5+0.7+0.3+0.2)=0.38

According to the formula,

PSSM AC(A,1)=¼((0.2-0.38).(0.5-0.38)+.....+(0.3-0.38).(0.2-0.38))

17. Pseudo PSSM(PPSM)

- This method finds the average for every amino acid type in the PSSM matrix.
- Then it calculates the correlation between residues separated by n characters per each amino acid type.
- All values in the PSSM matrix must be standardized by using the following formula:

$$s_{i,t} = \frac{s^0_{i,t} - \frac{1}{20}\sum_{j=1}^{20} s^0_{i,j}}{\sqrt{\frac{1}{20}\sum_{k=1}^{20}(s^0_{i,k} - \frac{1}{20}\sum_{j=1}^{20} s^0_{i,j})^2}}, \qquad t \in A$$

Where A is the set of the 20 natural amino acids, $s^0_{i,t}$ is the initial score in the PSSM matrix for the amino acid t at the row i, and $s^0_{i,j}$ and $s^0_{i,k}$ are the initial scores in the PSSM matrix for the row i, columns j and k.

$$\bar{s}_t = \frac{1}{N}\sum_{i=1}^{N} s_{i,t}, \qquad t \in A$$

$$\rho_t(n) = \frac{1}{N-n}\sum_{i=1}^{N-n} (s_{i,t} - s_{i+n,t})^2, \qquad t \in A$$

Where $s_{i,t}$ and $s_{i+n,t}$ are the standardized scores in the PSSM matrix for the amino acid t at rows i and i + n, and N is the sequence length.

- The PPSM vector is the concatenation of the 20 values for $\bar{s_t}$ and the 20 values of $\rho_t(n)$.
- Vector length is 40

Example

Position	A	B	C
1	1	-1	0
2	0	2	-1
3	-2	1	1

According the formula of standardization, we calculate $s_{1,A}$, then find the average of s_A.

Likewise, similar types of calculation are done for other types of amino acids.

Then we have to calculate $\rho_A(1)$ according to the above formula. Similar calculations are done for the other types of amino acids.

Then the PSSM vector is calculated.

THANK YOU