Encoders(2)

Shreyasi Ghosh

SSEB

- For every amino acid,this tool categorizes it into 3 secondary structures (Helix, Sheet,Coil)
  - Helix(100)
  - Sheet(010)
  - Coil(001)

Vector Length is 3N, N is the sequence length.

Amino acids having C as secondary structure:
A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,Y

Amino acids having S as secondary structure:V,E,A,L,Y

Amino acids H as secondary structure:A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,,W,Y

Example

Example:

Amino Acid:	A	R	N	D
Secondary Structure:	Sheet	Coil	Coil	Helix
Binary Vector:	010	001	001	100

SSEC

- For every amino acid,this tool categorizes it into 3 secondary structures (Helix,Coil,Sheet)
- Then, the tool calculates the frequency of each element type (Helix,Coil,Sheet) found in the entire sequence.

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

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

Example:

Amino Acid: A R N D

Secondary Structure: Sheet Coil Coil Helix

Number of times Helix appeared in the sequence is 1

Number of times Coil appeared in the sequence is 2

Number of times Sheet appeared in the sequence is 1

Total length of the sequence is 4

SSEC(H)=Number of times Helix appeared in the sequence/Total length of the sequence=1/4

SSEC(C)=2/4=1/2

SSEC(S)=1/4

SSEB

- For every amino acid,this tool categorizes it into 3 secondary structures (Helix,Coil,Sheet) with their corresponding probabilities.
- Then,the tool makes combinations of Helix(H), Coil(C),Sheet(S) among each pair of amino acid .The possible combinations are as follows: HH,HC,HS,CC,CH,CS,SS,SH,SC.
- The above values are then divided by the length of the sequence

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

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, n is the residue gap.

Vector length is 9 because of 9 possible combinations.

Example

Amino Acid: A R N 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

$SSPB(H,C)=\frac{1}{5}(P_A(H).P_R(C)+P_R(H).P_N(C)+P_N(H).P_D(C)+P_D(H).P_E(C))$

Likewise, we need to calculate for other pairs

SSPAC

- For every amino acid,this tool categorizes it into 3 secondary structures (Helix,Coil,Sheet) with their corresponding probabilities.
- Then, the tool generates bi-mers of the sequence.
- Then, the tool sums the multiplication of the probabilities of secondary structures (Helix,Coil,Sheet) between the bi-mers of amino acid.
- The above values are then divided by the length of the sequence

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

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

Amino Acid: A R N 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

$$SSPAC(1,H)=\frac{1}{5}(P_A(H).P_R(H)+P_R(H).P_N(H)+P_N(H).P_D(H)+P_D(H).P_E(H))$$

$$=((0.3*0.2)+(0.2*0.5)+(0.5*0.1)+(0.1*0.4))/5$$

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: ARN

A:Φ1,ψ1

R:Φ2,ψ2

N:Φ3,ψ3

Now, the vector would be

[Φ1,ψ1,Φ2,ψ2,Φ3,ψ3]

TAC

- This tool assigns each amino acid phi and psi values and then converts them into radians
- Then,it calculates the sum of all the sine of phi values,cosine of phi over all the amino acids,sine of psi values,cosine of psi values over each amino acid
- Lastly, all the above values are divided by the length of the sequence.

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

Vector length is 4

Example

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

A:$\Phi 1=60^{o},\psi 1=-45^{o}$

R:$\Phi 2=-30^{o},\psi 2=0^{o}$

N:$\Phi 3=45^{o},\psi 3=30^{o}$

According to the above formula

$TAC(\sin,\Phi)=\frac{1}{3}(\sin(\Pi/3)+\sin(-\Pi/6)+\sin(\Pi/4))$

$TAC(\sin,\psi)=\frac{1}{3}(\sin(-\Pi/4)+\sin(0)+\sin(\Pi/6))$

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)]

TAB

- This tool assigns each amino acid phi and psi values and then converts them into radians
- Then,the tool makes combinations of phi and psi values with the help of sine and cosine between of bi-mers of amino acid
- Lastly, each of the above values are divided by the length of the entire sequence.

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

Vector length is 10

Example

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

A:Φ1=60^o,ψ1=-45^o

R:Φ2=-30^o,ψ2=0^o

N:Φ3=45^o,ψ3=30^o

According to the above formula

TAB(sin,cos,phi,1)= ⅓(sin(π/3)*cos(-π/6)+sin(-π/6)*cos(π/4))

So the possible pairs calculated for TAB are:

- phi_sin-phi_sin
- phi_sin-phi_cos
- phi_sin-psi_sin
- phi_sin-psi_cos
- phi_cos-phi_cos
- phi_cos-psi_sin
- phi_cos-psi_cos
- psi_sin-psi_sin
- psi_sin-psi_cos
- psi_cos-psi_cos

TAAC

- This tool assigns each amino acid phi and psi values and converts them into radians
- Then,the tool makes dimers of the amino acid and calculates their phi_Sin,phi_Cos,psi_Sin,psi_Cos values.
- Lastly, all the above values gets divided by the total length of the sequence.

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

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

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

A:Φ1=60^{o},ψ1=-45^{o}

R:Φ2=-30^{o},ψ2=0^{o}

N:Φ3=45^{o},ψ3=30^{o}

TAAC(sin,phi,1)=⅓(sin(π/3)*sin(-π/6)+sin(-π/6)*sin(π/4))

Similarly TAAC has to be found out for other combinations like phi_Cos,psi_Sin,psi_Cos

ASA

- This tool assigns Accessible Surface Area Values to each amino acid.
- Then, the above values gets divided by the total length of the sequence.

Vector length is N, N is the sequence length.

Example

Sequence: ARND

ASA_A:0.35

ASA_R:0.01

ASA_N:0.064

ASA_D:0.23

Then the resulting vector is [0.35,0.01,0.0064,0.23]

BiPSSM

- This tool will assign PSSM scores to each amino acid with respect to the 20 naturally occuring amino acid
- This tool calculates the summation of products of the pssm values between each pair of amino acids , separated by n residues.
- Lastly,the values are divided by the length of the sequence.

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

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

Example: n=1

PSSM Matrix Scores:

Position	1 A	2G
1A	0.2	` 0.1
2R	0.5	0.3
3N	0.7	0.4
4D	0.3	0.6
5H	0.2	0.5

Peptide Sequence: ARNDH

BiPSSM(A,A)=⅕((0.2*0.5)+(0.5*0.7)+(0.7*0.3)+(0.3*0.2))

Likewise, we calculate for (A,G),(G,A),(G,G)

PSSMAC

- This tool assigns PSSM scores to each amino acid with respect to the 20 naturally occuring amino acid
- Then, it calculates the average of each of the 20 naturally occurring amino acid (column in the PSSM)
- This tool calculates the summation of products between the pssm values of each amino acid after subtracting the mean of the each of the 20 naturally occurring amino acid
- Lastly,the values are divided by the length of the sequence.

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

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

here 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 = 20

Example

Example:

PSSM Matrix Scores

Position	A	G
1A	0.2	0.1
2R	0.5	0.3
3N	0.7	0.4
4D	0.3	0.6
5H	0.2	0.5

Peptide Sequence: ARNDH

Mean of A=0.38 ,Mean of G=0.38

According to the formula,

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

PPSM

- This tool assigns PSSM scores to each amino acid with respect to the 20 naturally occuring amino acid
- Then, it calculates the average of each of the 20 naturally occurring amino acid (column in the PSSM)

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

- Every value in the PSSM is standardized using the given formula

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

- This tool calculates the summation of squares of the difference between the pssm values of each amino acids , separated by n residues.

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

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.

Example

Example:

PSSM Matrix Scores

Position	A	G
1A	0.2	0.1
2R	0.5	0.3
3N	0.7	0.4
4D	0.3	0.6
5H	0.2	0.5

Peptide Sequence: ARNDH

Each value in the PSSM matrix is standardized and the summation of squares of the difference between the pssm values of each amino acids is calculated according to the above formulas.

TVD

- This tool assigns each amino acid to the 10 physiochemical groups. The physiochemical groups are as follows:

  Hydrophobic, Positive, Negative,Polar,Charged, Small, Tiny,Aliphatic,Aromatic,Proline

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

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

All the sequences must have same length

Amino acids falling under the category of Hydrophobic group: A,C,F,G,H,I,K,L,M,T,V,W,Y

Amino acids falling under the category of Positive group: H,K,R

Amino acids falling under the category of Negative group: D,E

Amino acids falling under the category of Polar group:D,E,H,K,N,Q,R,S,T,W,Y

Amino acids falling under the category of Charged group:D,E,H,K,R

Amino acids falling under the category of Small group:A,C,D,G,N,P,S,T,V

Amino acids falling under the category of Tiny group:A,G,S

Amino acids falling under the category of Aliphatic group: I,L,V

Amino acids falling under the category of Aromatic group: F,H,W,Y

Amino acids falling under the category of Proline group: P

Example

Sequence: ARN

TVD:[10000110000101100000001010000]

A falls under Hydrophobic,Small,Tiny category

R falls under Positive,Polar,Charged category

N falls under Polar, Small category