1. Discrete wavelet transform

Discrete Wavelet Transform (DWT) is employed to extract the effective information characteristics of protein amino acid sequences, which was first proposed by Nanni. The Wavelet Transform (WT) is defined as the projection of a signal $f(t)$ onto the wavelet function, as follows:

$$T(m,n) = \sqrt{\frac{1}{m}} \int_0^t f(t)\psi(\frac{t-n}{m})\, dt \tag{1}$$

where m and n are scale variable and translation variable, respectively. $\psi(\frac{t-n}{m})$ denotes the analyzing wavelet function. $T(m,n)$ is the transform coefficients.

The illustration of the discrete wavelet transform is as follows:

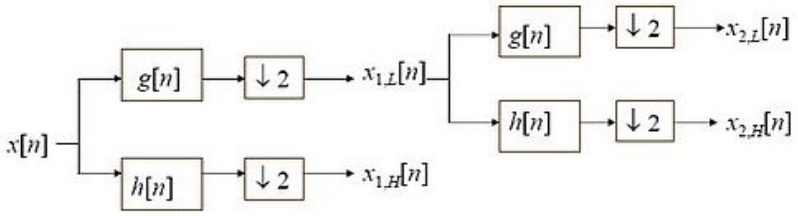

Figure 1: The framework of Discrete wavelet transform.

where x[n] is the discrete signal, with n is the length of discrete signal,g[n] indicates low pass filter and h[n] g indicates high pass filter, then the coefficients can be calculated as follows:

$$x_{1,L}\,[n] = \sum_{k=1}^{N} x\,[k]\, g\,[2n-k] \tag{2a}$$

$$x_{1,H}\,[n] = \sum_{k=1}^{N} x\,[k]\, h\,[2n-k] \tag{2b}$$

where g is the low pass filter and h is the high pass filter. So, $y_{low}\,[n]$ represent the approximation coefficients and $y_{high}\,[n]$ are the detailed coefficients.

Just like some previous methods, we use 4-level discrete wavelet transform to process the PSSM matrix. For each level, we obtain the approximate and detailed coefficients of each column. Then, we extract the maximum, minimum,

mean and standard deviation values of both approximate and detailed coefficients, and first five discrete cosine coefficients of the approximate coefficients. Totally, there are 4+4+5 features of each level for one of 20 column dimensions.

Finally, we extract the feature vector $F_{PSSM-DWT}$ with $(4+4+5)\times 4\times 20 = 1040$ dimensions.