Input Features

Composition of amino acids (420 features)

- AAComposition_{AA} –composition (count) of a given AA type {AA} in a given protein divided by protein's sequence length.(20 features)
- AAComposition_{AA}_{AA} – composition (count) of a given dipeptide {AA}_{AA} (pair of two consecutive amino acids) in a given protein divided by protein's sequence length. (400 features)

Physiochemical properties of proteins based on amino acid indices (448 features). These features are based on per AA values of 64 hydrophobicity and energy based indices collected from the AAIndex database (http://www.genome.jp/aaindex/); see Table 1 for the list of the considered indices:

- AAindex_{Index}_avg –average value of a given AA index {Index} over the whole input protein sequence. (64 indices = 64 features)
- AAindex_{Index}_{min,max}_{5,10,15} –The minimal/maximal average value of a given AA index {Index} among all sliding windows of sizes 5, 10, and 15 over the input protein chain. For chains shorter than a given window size, we use the window size equal the length of the sequence. (64 indices x 6 values per index = 384 features)

Other properties of proteins (4 features):

- pI –The isoelectric point of the input protein. (1 feature)
- AliphaticIndex –The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). This index is regarded as a positive factor for the increase of thermostability of globular proteins. (1 feature)
- InstabilityIndex –The instability index provides an estimate of the stability of a given protein (1 feature)
- NetCharge – Net charge of a given protein. (1 feature)

Disorder and sequence complexity (68 features). Features based on predictions of residues disorder performed by IUpred method (http://iupred.enzim.hu/), which includes predictions of both Short (IUpred_S) and Long (IUpred_L) disorder segments, and based on assignment of sequence complexity utilizing SEG algorithm (http://www.biology.wustl.edu/gcg/seg.html):

- PRprobability_{IUpredL, IUPredS}_avg – average value of probabilities/complexity values of a given predictor/algorithm {IUpredL, IUPredS, Complexity} over the whole protein sequence. (2 predictors = 2 features)
- PRprobability_{IUpredL, IUPredS}_{min,max}_{5,10,15} – The minimal/maximal average value of probabilities/complexity values of a given predictor/algorithm {IUpredL, IUPredS, Complexity} among all sliding windows of sizes 5, 10, and 15. For chains shorter than a given window size we use the window size equal the length of the sequence. (2 predictors x 6 values per index = 12 features)
- PRSegmentCount_{IUpredL, IUPredS, Complexity}_{0, 1}_{1-5, 6-10, 11-15, >15} – count of the number of short (1-5 residues)/medium (6-10 residues)/long (11-15 residues)/very long (over 15 residues) segments in the input protein for each binary prediction/complexity value {0, 1} of each predictor/algorithm{IUpredL, IUPredS, Complexity}. These counts were normalized by the total number of segments (for that predictor) in the protein. (3 predictors/algorithm x 2 predictions/assignments per predictor/algorithm x 4 segment sizes = 24 features)

- PRSegmentComposition_{IUpredL, IUPredS, Complexity}_{0, 1}_{1-5, 6-10, 11-15, >15} – count of the number of AAs in the input protein sequence that are in short (1-5 residues)/medium (6-10 residues)/long (11-15 residues)/very long (over 15 residues) segments for each binary prediction/complexity value {0, 1} of each predictor/algorithm {IUpredL, IUPredS, Complexity}. These counts were normalized by the length of the protein. (3 predictors/algorithm x 2 predictions/assignment per predictor/algorithm x 4 segment sizes = 24 features)
- PRLongestSegment_{IUpredL, IUPredS, Complexity}_{0, 1} – the length of the longest segment for each binary prediction/complexity value {0, 1} of each predictor/algorithm{IUpredL, IUPredS, Complexity}divided by the protein sequence length.(3 predictors x 2 predictions per predictor = 6 features)

Table 1. List of considered 64 hydrophobicity- and energy-based indices. The names are based to the nomenclature from the AAIndex1 database.

ARGP820101	BULH740101	CHAM820102	CIDH920105	EISD840101
EISD860101	EISD860102	EISD860103	FAUJ830101	GOLD730101
GUYH850101	HOPT810101	JANJ790102	JOND750101	KYTJ820101
LAWE840101	LEVM760101	MANP780101	MIYS850101	NOZY710101
OOBM770101	OOBM770102	OOBM770103	OOBM770104	OOBM770105
OOBM850103	OOBM850104	PONP800101	PONP800102	PONP800103
PRAM900101	RADA880101	RADA880102	RADA880103	RADA880104
RADA880105	ROBB790101	ROSM880101	ROSM880102	SIMZ760101
SWER830101	VHEG790101	WERD780102	WERD780103	WERD780104
YUTK870101	YUTK870102	YUTK870103	YUTK870104	ZIMJ680101
PONP930101	WILM950101	WILM950102	WILM950103	WILM950104
KUHL950101	JURD980101	WOLR790101	KIDA850101	COWR900101
BLAS910101	CASG920101	ENGD860101	FASG890101	