Introduction

Protein tertiary structure is the interactions of the helices and sheets [1]. An ordered protein transforms to a disordered protein when this tertiary structure starts to deform. This lack of ordered structure provides a flexible and
random-coil composition to the proteins. Therefore, they interact physically and functionally with their target partners. Hence, disordered proteins become a part of cellular regulation processes such as cell signaling, transcription, translation and chromatin remodeling functions [2–4]. In addition, the conservation state of their open structure in case of interfering with their target, being able to adopt different structure on different partners, their interaction rate enhancing features and finally their proteolytic sensitivity make disordered proteins important part of proteomics field. Furthermore, there is a strong relationship between disordered protein structure and many diseases. For instance, collection of disorder protein structures causes synucleinopathies, neurodegenerative diseases, cardiovascular diseases, diabetes, neural diseases [5–7]. Determination of protein structures as ordered-disordered can have been done by protein crystallography, electron microscopy, nuclear magnetic resonance (NMR) spectroscopy, small-angle X-ray and neutron scattering [8]. However, experimentally pursuing these processes are costly and time-consuming. For this reason, fast identification and classification techniques have continuously been improving day by day to effectively classify the proteins whether they are ordered or disordered [9–12].

One of the most used classification method is the support vector machine which is commonly used in bioinformatics [13–16]. Support vector machine (SVM) can be extended to the nonlinear structure as kernel. In this study kernel SVM is investigated for ordered-disordered proteins.

In literature, linear SVM is also used in bioinformatics problems [17, 18]. However, in real-world data may not be distributed normally. Kernel SVM is very useful method in order to classify nonlinear data [19]. SVM used for identifying the positions in the amino-acid sequence [20, 21]. SVM also used for prediction of protein secondary content [22].

In this study, linear and kernel SVM will be applied for classify the ordered-disordered proteins. This is very important problem in bioinformatics because of the significance of ordered-disordered proteins.

Materials and methods

Support vector machine (SVM)

Classifying data is one of the most important tasks in machine learning. SVM is one of the basic machine learning techniques. It is a robust method to classify the data by using its class label. The idea of SVM is to create a hyper plane in between data sets to indicate which class it belongs to, as shown in Figure 1.

SVMs can separate the classes using the labeled training data with a hyper-plane in a high dimensional space which have maximum distance to the nearest training data point of any class.

In order to perform SVMs as a classification technique, SVMs separate a given known set of (−1, +1) labeled training data via a hyper-plane that is the largest distance from the positive samples and negative samples [23]. It is also possible to classify nonlinear dataset by using the kernel method. In kernel method, data is mapped into nonlinear feature space F and the hyper-plane found by the SVM in feature space F.

Linear support vector machine

Training set is selected for SVM in a two-class classification problem which can be separated linearly.

\[
\begin{align*}
\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_k, y_k\}
\end{align*}
\]

The following equations give us inequalities belonging to the optimum hyper-plane [24]:
\[ w \cdot x_i + b = \pm 1 \]

The aim is to find the best hyper-plane that is the one that represents the largest separation, or margin, between the two classes as seen in Figure 2.

Points which create the hyper-plane, the so-called support vectors and referred as follows:

\[ w \cdot x_i + b = \pm 1 \]

The following optimization problem can be solved to maximize the boundary of optimum hyper-plane.

\[ \min \left[ \frac{1}{2} ||w||^2 \right] \]

We can describe the decision functions as below:

\[ w \cdot x_i - b \geq 1 \] for \( x_i \) belongs to first class

\[ w \cdot x_i - b \leq -1 \] for \( x_i \) belongs to second class

The limitations can be expressed as follows:

\[ y_i \cdot w \cdot x_i + b \geq 0 \]

The optimization problem can be solved by using Lagrange Equations and the following obtained equality [17]:

\[ L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{k} \alpha_i y_i(w \cdot x_i + b) + \sum_{i=1}^{k} \alpha_i \]

Consequently, the decision function can be written as follows:

\[ f(x) = \text{sign} \left[ \sum_{i=1}^{k} \alpha_i y_i (x \cdot x_i) + b \right] \]

Kernel support vector machine

Linear was introduced by Vapnik in 1963. Then, ‘Kernel Method’ was introduced by Bernhard E. Boser and Vladimir Vapnik to solve nonlinear classification problems. ‘Kernel Method’ uses ‘Kernel Trick’. Kernel trick provides bridge linearity to nonlinearity [26]. Training data \( X \) is mapped into a high-dimensional feature space \( F \) by using ‘Kernel Function’. Next, the classification rule is transformed into the following form [27]:

\[ f(x) = \text{sign} \left[ \sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b \right] \]

\( K \) is a symmetric and positive definite function. In this study, ‘Gaussian Kernel’ function used:

\[ k(x, y) = e^{-\frac{|x-y|^2}{c}} \]

The kernel function used for determine the optimum parameters. The problem can be solved using Lagrange equations [17]:

\[ L(\alpha) = \sum_{i=1}^{k} \alpha_i - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \lambda \sum_{i=1}^{k} \alpha_i \]

then the decision function found as below:

\[ f(x) = \text{sign} \left[ \sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b \right] \]

In this study, SVM was used with radial basis function (RBF) as the kernel function.

Dataset

The sequences of the disordered and ordered proteins were extracted from DisProt [28]. Our data set consists of 114 protein sequences; 57 of them ordered and 57 of them disordered. Sixty percent of data is randomly selected for training, and, the remaining is used for validation processes.

The protein dataset is transformed \( i^{th} \) amino acid composition by:

\[ F(x) = f_1(x), f_2(x), ..., f_{20}(x) \]

where \( F(x) \) is the vector contained each frequency of 20 amino acid types [29].

Therefore, the amino acid frequencies were calculated as follows. The percentage of the amino acid residue \( i \) in a protein \( x \) is defined by:

\[ f_i(x) = 100 \cdot \frac{n_i}{N} \]

where \( n_i \) is the frequency of amino acid \( i \) and \( N \) is the number of amino acid residues in the protein \( x \) [30].

Oldfield and Dunker [31] have referred that the flexibility and structural instability of disordered proteins are encoded by their amino acid sequences. Xue et al. [32] have concluded that amino acid composition of the disordered proteins has a difference from the other proteins including ordered ones. Amino acid sequence data is a significant key in the prediction of disordered proteins. Some amino acids have high frequencies in disordered protein than the ordered proteins. In addition, the common feature of
disordered proteins is compositional bias, and, the bias supports hydrophilic amino acids besides keeps hydrophobic residues out. Hereby, Arg, Gln, Glu, Lys, Pro and Ser are more abundant, on the other hand, Cys, Ile, Leu, Phe, Trp, Tyr, and Val are more inconsiderable in disordered proteins. As a result, there is a strong relationship between amino acid composition and disordered proteins were identified by Hansen et al. [33], Romero et al. [34] and Oldfield and Dunker [31].

Results

Linear and kernel SVM are applied to the dataset. It is clearly seen from Figure 3, linear SVM method has made some misclassifications. On the other hand, kernel SVM classifier gives better result than linear SVM as shown in the Figure 4. It is clearly seen from Tables 1 and 2, linear SVM gives worse performance than kernel SVM. It can be seen from Tables 3 and 4, overall prediction rate obtained by linear and kernel SVM in identifying the ordered-disordered proteins are 86.54% and 94.23%, respectively. As shown from results, kernel SVM gives the best classifying scheme for ordered-disordered protein sequences because of its appropriate mathematical background for linear and nonlinear data structure. Consequently, kernel SVM method is a very robust method to identify protein structures as ordered-disordered.

Discussion

In bioinformatics, proteins are large molecular structures that are composed of one or more chains of amino acids
with a variety of shapes, size and chemical properties. Different types of protein sequences have specific biochemical function [1]. Disordered proteins are one of the functionally important classes of proteins. In biochemistry, “active proteins” e.g. enzymes are known with their unique three dimensional folded structures. However, the disordered proteins are strictly different and they do not have unique three dimensional structure. Moreover, although they have not well defined three dimensional structures, they are considered as active proteins. Up to now, biological activities such as cell cycle control, regulation, sensing, etc. [35–40] of many disordered proteins have been reported. Moreover, the scientists consider that disordering provides proteins more flexibilities than those of the ordered proteins have. They can interact more ligands [2].

Disorder structure is mostly observed in proteins implicated in cell signaling, transcription and chromatin remodeling functions. Therefore, classification of ordered-disordered protein sequences is very essential process to express some diseases [41, 42] and herewith drug discovery studies.

Experimental classification is costly and time-consuming because of great amount of raw sequences. Therefore, computational solutions have become a useful tool for analysis [9–12]. The prediction methods have to give more accurate and rapid solutions. However available methods [43–46] can allow observing only one protein sequence whether ordered or disordered, and furthermore, they have remarkable slow execution time. So, in this study, more than one sequence of ordered-disordered proteins was classified by using the linear SVM classifier and the kernel classifier with fast execution time. At the same time, kernel SVM gives the high prediction accuracy, therefore, it can be referred that it is a very competitive, robust and fast classification method to identify proteins in terms of their ordered-disordered structures.

References

1. Lesk AM. Introduction to bioinformatics. New York: Oxford University Press, 2005.
2. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–33.
3. Khan SH, Kumar R. An overview of the importance of conformational flexibility in gene regulation by the transcription factors. J Biophy 2009;2009:1–9.
4. Sandhu KS. Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. J Mol Recognit 2009;22:1–8.
5. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–84.
6. Li J, Feng Y, Wang X, Li J, Liu W, Rong L, Bao J. An overview of predictors for intrinsically disordered proteins over 2010–2014. Int J Mol Sci 2015;16:23464–62.
7. Uversky VN. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. Front Aging Neurosci 2015;7:1–6.
8. Snyder DA, Chen Y, Denissova NG, Acton T, Aramini JM, Ciano M, et al. Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. J Am Chem Soc 2005;127:16505–11.
9. Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol 2007;7:251–13.
10. Wang L, Sauer UH. OnD-CRF: predicting order and disorder in proteins conditional random fields. Bioinformatics 2008;24:1401–2.
11. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21:3369–76.
12. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–39.
13. Wei Z, He J, Harrison R, Tai P, Pan Y. Clustering support vector machines for protein local structure prediction. Expert Syst Appl 2007;32:518–26.
14. Zhang G, Ge H. Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. Comput Biol Chem 2013;46:389–422.
15. Chen C, Tian Y, Zou X, Cai P, Mo J. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 2006;243:444–8.
16. Pugalenthi G, Kumar KK, Suganthan PN, Gangal R. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochem Biophys Res Commun 2008;367:630–4.
17. Cai CZ, Wang WL, Sun LZ, Chen YZ. Protein function classification via support vector machine approach. Math Biosci 2003;185:111–22.
18. Cai YD, Liu XJ, Xu X, Chou KC. Prediction of protein structural classes by support vector machines. Comput Chem 2002;26:293–6.
19. Saruta K, Hirai Y, Tonaka K, Inove E, Okayasu T, Mitsuoka M. Predictive models for yield and protein content of brown rice using support vector machine. Comput Electron Agric 2013;99:93–100.
20. Lorena AC, de Carvalho AC. Protein cellular localization prediction with support vector machines and decision trees. Comput Biol Med 2007;37:115–25.
21. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–8.
22. Chen C, Tian Y, Zou X, Cai P, Mo J. Prediction of protein secondary structure content using support vector machine. Talanta 2007;71:2069–73.
23. Gürağşka GE, Hakli H, Uguz H. Support vector machines classification based on particle swarm optimization for bone age determination. Appl Soft Comput 2014;24:597–602.
24. Guyon I, Weston J, Barnhill S. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
25. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
26. Shawe-Taylor J, Cristianini N. Kernel methods for pattern recognition, 1st ed. Cambridge, Newyork, USA: Cambridge University Press, 2004.
27. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906–14.
28. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, et al. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–40.
29. Kandemir-Cavas C, Nasibov E. Classification of apoptosis proteins by discriminant analysis. Turk J Biochem 2012;37:54–61.
30. Cedano J, Aloy P, Pérez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
31. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem 2014;83:553–84.
32. Xue B, Oldfield CJ, Dunker AK, Uversky VN. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett 2009;583:1469–74.
33. Hansen JC, Lu X, Ross ED, Woody RW. Intrinsic protein disorder, amino acid composition, and histone terminal domains. J Biol Chem 2006;281:1853–6.
34. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequence. Neural Networks 1997;1:90–5.
35. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res 2009;19:929–49.
36. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–82.
37. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker A.K. Intrinsic disorder and functional proteomics. Biophys J 2007;92:1439–56.
38. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J Proteome Res 2007;6:1899–916.
39. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 2007;6:1882–98.
40. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 3. Ligands, posttranslational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res 2007;6:1917–32.
41. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2Concept. Annu Rev Biophys 2008;37:215–46.
42. Mulligan VK, Chakrabartty A. Protein misfolding in the late-onset neurodegenerative diseases: common themes and the unique case of amyotrophic lateral sclerosis. Proteins 2013;81:1285–303.
43. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics 2004;20:2138–9.
44. McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 2008;24:1798–804.
45. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 2010;26:i489–96.
46. Walsh I, Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Res 2011;39:W190–6.