Neural Network Modeling in Dithiothreitol Reduction and Ion Treatment of Recombinant Human Insulin Obtained from the Circular Dichroism (CD) Spectral Information

Soroush Sardari* and Saeed Soltani

Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute, #69 Pasteur Ave., Tehran, 13164, Iran

Abstract: In this study, induced changes in the secondary structure of the human insulin were carried out by addition of various reagents causing modification in the disulfide bond such as dithiothreitol (DTT) three dimensional structure of insulin. CD spectra were taken accordingly and the spectra recorded. There are methods to predict and estimate spectral changes of a peptide molecule, however there is no method to process CD spectral data and correlate them with that of inducing factor. Artificial intelligence backpropagation algorithm, as a strong model building tool was used here for prediction and data mining. Therefore, artificial neural network (ANN) methodology was used to build a model to study the effect of selected biochemical factors in the downstream process of a recombinant peptide.

Keywords: Recombinant human insulin, Dithiothreitol reduction, Artificial intelligence, Circular dichroism, Downstream processing.

INTRODUCTION

Insulin is the major anabolic hormone in all 'higher' organisms involved in regulating the uptake of glucose by cells of the body, amino acid synthesis and the conversion of carbohydrate into triacylglycerols [1]. Human insulin was shown to contain 51 amino acids, arranged in two chains (an acidic A-chain of 21 residues and a basic B-chain of 30 residues). Insulin is a globular protein containing two chains; its structure as a monomer in solution resembles a crystallographic T-state protomer [2, 3].

At the production of insulin, the upstream process includes the bioreactor application, for which an estimated 30 hour time at 370 C^o in a 40 m^3 reactor would suffice. The downstream would follow by isolation of the inclusion bodies after removal of the cell fragments. Following that, further purifications and oxido-reductive modifications on recombinant human insulin are applied. During such steps the secondary structure of insulin is affected. It is necessary for S-S bonds to be opened prior to final refolding. Such a process requires sulfur atoms to be converted into –SH functionality [4]. Despite the fact that the efficiency of protein isolation from inclusion bodies is high, such manipulations are crucial steps in the whole downstream process, especially for scale-up and manufacturing of recombinant protein from inclusion bodies. Oxidation breakdown can produce the higher amount of properly folded fusion protein feeding the later digestion process. This makes it vulnerable to higher scrutiny and optimizations. Insight into the informational content of protein sequences has been obtained from genetic analysis of allowed and disallowed sequences [5] and from development of knowledge-based algorithms for evaluating the fit between a given sequence and a known structural template [6]. However, complementary insights into direction and extent of downstream in-process factors affecting structural features that contribute to the native state have not been provided before. Yet the construction of incorrectly folded models and their analysis using empirical force fields have been applied [7]. This approach allows direct visualization of competing "threads" but is largely restricted to computer simulation domain [8]. There are partial folds in relation to energy landscapes and possible mechanisms of insulin fibrillation. Circular dichroism (CD) is an instrumentation technology for elucidating the secondary structure of the proteins and peptides. It also allows the detection and quantitation of the chirality of molecular structures. CD is a variant of absorption spectroscopy which measures the difference in absorption of left and right polarized light in the ultraviolet (UV) band by a medium or sample. Although the peptide bond is planar and hence symmetric, there is usually an asymmetric alpha-carbon on either side; hence the peptide bond transitions interact to give a CD signal which is very sensitive to secondary structure. In the far UV region (180-250 nm) the CD of a protein is primarily that of the amide chromophores along the backbone. With the introduction of an optically active sample, a preferential absorption is seen during one of the polarization periods and the intensity of the transmitted light now varies during the modulation cycle. The variation is directly related to the circular dichroism of the sample at that wavelength. Successive detection is performed at various wavelengths leads to the generation of the full CD spectrum. CD uses very little sample (200 µl of 0.5 mg/ml solution in standard cells), it is non-destructive and relative changes due to influence of environment on sample (pH, denaturants, temperature etc.) can be monitored very accurately. CD has an important role in the structural determinants of proteins. The real power of CD is in the analysis

*Address correspondence to this author at the Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute, #69 Pasteur Ave., Tehran, 13164, Iran; Tel: (98-21) 6640-5535; Fax: (98-21) 6646-5132; E-mail: sardari@pasteur.ac.ir; ssardari@hotmail.com

1875-0362/09

of structural changes in a protein upon some perturbation, or in the comparison of the structure of an engineered protein to the parent protein. CD is rapid and can be used to analyze a number of candidate proteins from which interesting candidates can be selected for more detailed structural analysis like NMR or X-ray crystallography. This method proved to be an effective downstream method optimizing assistant while having the capability of being routinely used as a complementary quality control (QC) instrument, which can identify the nature of peptide folding in each step.

Computational approaches, such as structural bioinformatics [9,10], molecular docking [11-13], molecular packing [14, 15], pharmacophore modeling [16], Mote Carlo simulated annealing approach [17], protein subcellular location prediction [18-20], protein structural class prediction [21], identification of membrane proteins and their types [22], identification of enzymes and their functional classes [23], identification of proteases and their types [24], protein cleavage site prediction [25-27], and signal peptide prediction [28, 29] can timely provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community. The present study was attempted to use artificial neural network (ANN) methodology to build a model to study the effect of selected biochemical factors in the downstream process of a recombinant peptide in hope to provide useful information for basic research and drug development. An artificial neural network (ANN) is composed of many processing elements (PE). Each processing element has inputs, transfer functions and output. Processing elements are connected with coefficients and are arranged in layers, i.e., input layer, output layer and hidden layers in between [30]. Application of ANN in pharmaceutical research is a new field with novel potentials to be discovered. A variety of areas have been described to benefit from such algorithm predictions that range from industrial design systems to optimal formulation prediction and SAR evaluations [31, 32]. In this study the influence of selected reaction condition and mixture ingredients used in the above-mentioned steps on recombinant human insulin folding and its optimization for human insulin production applying CD spectroscopy were investigated.

MATERIAL AND METHODS

The CD-spectra were obtained at 250 ºC on a Jasco J-810 spectropolarimeter (Tokyo, Japan). Protein (human insulin) concentration was 0.5 mg/ml unless otherwise indicated. The far-UV spectra are recorded using a step size of 1 nm and a bandwidth of 1.5 nm. Spectra are recorded in a cell with a path length of 0.1 mm, the cell path length for far-UV spectra (190-250 nm). The spectra were corrected for buffer contributions. Quartz cuvettes of 1 mm in path length at 22 °C were used. The CD-spectra were analyzed and the secondary structure content was calculated using the program Spectra Manager for Windows 95/NT, Spectra Analysis, Version 1.53.02 [Build 1], JASCO Corporation. Human insulin (Lilly-Eksir) solution was made in pure ware at a concentration of 0.5-1 mg/ml using dilute 0.1 N HCl to co-dissolve.

DTT and Salt Treatment

Aliquots from stock solutions were mixed to a final concentration of 0.5-1 mg/ml human insulin and treated with DTT (0.2 mM) and other salt solutions varying from 0.1 to 1 mM. In case of using buffer solutions, the pH was justified to 7 with 10-100 mM Tris-HCl. Samples were incubated for 60 minutes before the spectra were collected for the salts and varied for DTT, with the DTT being added immediately before measurements were taken. The absorbance of each sample was then measured every 3 minutes over the course of 30 minutes. Data were collected at 1-nm intervals over the range 240-190 nm for native protein, and for denatured protein with a collection time of 5 s/data point. To eliminate contributions from buffer, salt solutions were prepared in the same manner described above. To obtain correct spectra, their spectral data were subtracted from those of protein solutions

Heat Treatment

The effect of heat as an affecting factor in all the steps of a typical downstream process was measured using the heating module connected to the Jasco CD instrument. This module uses the circulating water adjusted to the desired temperature to operate.

Sulfitolysis

To study the factors affecting S-S bond, the sulfitolysis reaction was used. In this method, the concentration of protein was adjusted to 0.5mg/ml using the above mentioned buffer solution and Na_2SO_3 and $K_2S_4O_6$ were added to make concentration of 7.9 and 6 mg/ml respectively. The pH was adjusted to 9.1 using NaOH 2M solution and sample was incubated for 12h at room temperature [33]. The CD spectrum of the control solutions, buffers or the salt solutions, were subtracted from the main solution's spectrum.

ANN

A standard feed-forward network, with back propagation rule and with single hidden layer architecture was chosen applying the EasyNN, 8.01 (1999-2001). The wavelengths, insulin treatment concentration, were used as the inputs, while the kind of treatment was the output of the network architecture. There were assigned numerical values to various treatment conditions; these values were arbitrary to prevent text based learning and modeling. They ranged from -500 to +500 and would indicate the general category in each treatment. Insulin concentration, treatment amount and time of incubation, in addition to the CD degree values at each particular wavelength (240-190 nm) would comprise the input values. The number of neurons was kept minimum to avert an over-fitting problem, which is usually produced by more weights due to higher numbers of neurons in input and hidden layers. However, to produce the optimum architecture, powerful enough to model the functions and not create errors more than 0.05%, the total number of hidden layer neurons was varied from 10 to 50 and the hidden layers were from 1 to 3. The architectures were produced that met the error limit condition using least number of calculation cycles. Higher numbers of hidden layer did not improve the performance, yet can decrease the speed of calculation. This finding is in accordance with previous reports [34].

RESULT

1. Effect of Temperature

The effect of temperature on human insulin as monitored in two different wavelengths (194.8 and 208.1 nm) are

shown in Fig. **1a**) and **b**) for increasing temperature denaturing profiles and Fig. **1c**) and **d**) for decreasing temperature renaturing profiles

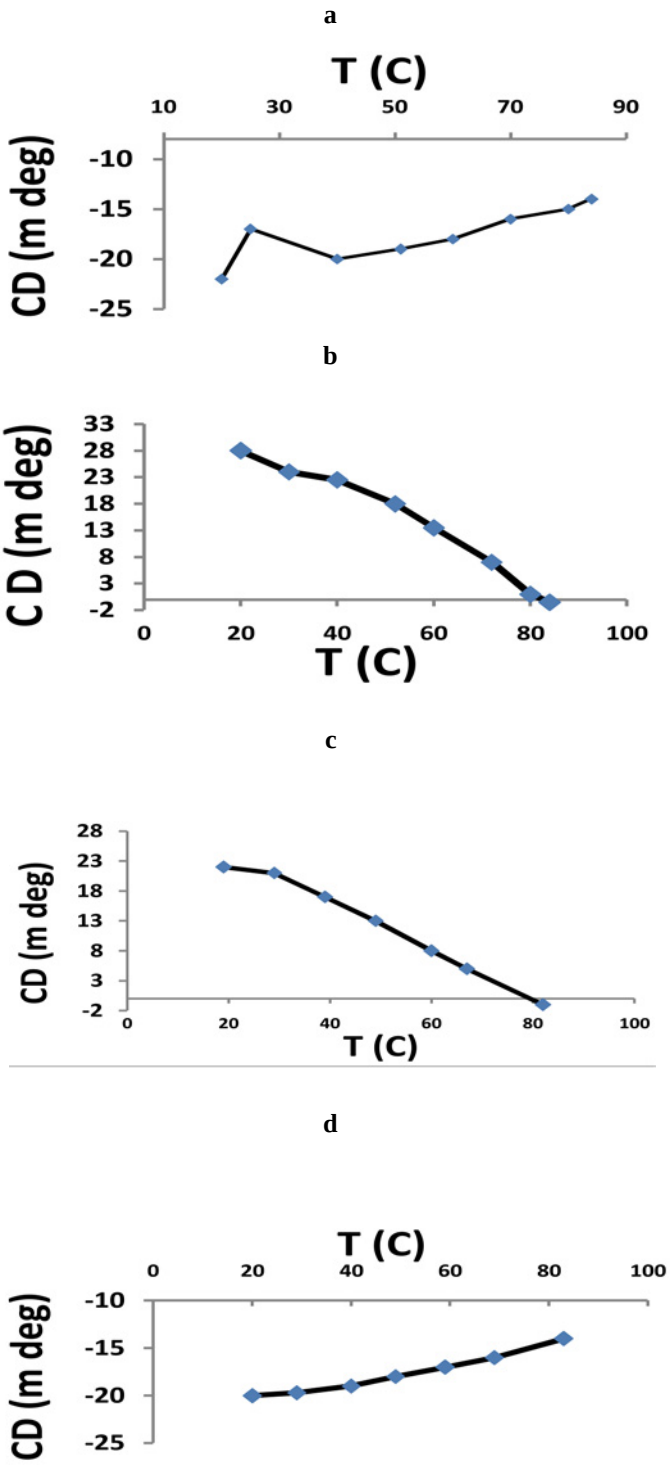

Fig. (1). a) Profile of insulin thermodynamic study; CD data at 208.1 nm with increasing temperature.

b) Profile of insulin thermodynamic study; CD data at 194.8 nm with increasing temperature.

c) Profile of insulin thermodynamic study for refolding; CD data at 194.8 nm with decreasing temperature.

d) Profile of insulin thermodynamic study for refolding; CD data at 208.1 nm with decreasing temperature.

2. Oxidative Folding Study

Since insulin is mainly composed of alpha helix as the dominant secondary structure, the effect of sulfitolysis reagent [34] is easily detected as indicated. Fractional content of insulin secondary structure elements in the oxidative un/refolding is shown in the Table **1**.

Table 1. Content of Insulin Secondary Structure Elements in the Oxidative Un/Refolding (the Values are in Percentage)

Time (min) / Structure	1	3	6	9	12	15	20
Alpha	18.6	19.3	14.9	13.7	11.6	10.2	9.1
Beta	44.7	38.2	46.1	44.0	44.4	46.9	44.5
Turn	8.4	11.6	9.1	8.6	9.4	7.2	8.6
Random	28.3	31.0	30.0	33.7	34.6	35.7	37.8

3. Kinetics of S-S Bond Reduction

DTT is a compound that has been used commonly in the downstream process for reducing the disulfide bonds. The numbered values are indicated in the Table **2**. As shown in the Fig. (**2**), the progress of time domain on the completion of the reaction and separation of S-S bonds can be monitored by CD.

Table 2. The Secondary Elemental Component of Insulin Treated with DDT at Various Time Intervals (the Values are in Percentage)

Secondary Structure (%) / Insulin Treatment	Alpha	Beta	Turn	Random
-	42.6	31	13.9	12.5
DTT	8.2	40.8	10.6	40.4
Na_2SO_3, $K_2S_4O_6$	18.3	37.4	0	44.3

4. Effect of Reducing Agents

Reduction of insulin by agents like DTT would cause conversion of S-S bonds to –SH. In these circumstances, chain A and B are separated and the only dominant structure would be beta. In fact, this has been indicated in our study, obtaining 44.5 % of beta after 20 min at pH=8 (Tris, DTT 5mM, and Insulin 1mg/ml). This study was completed in conditions with reduced buffer concentration. Although at the beginning, the level of beta sheets was quite different at 2 and 0.5 mM of DTT, the yield was similar at the end of 20 min period (data not shown). In all instances, alpha contribution was not less than 8%.

The progress of disulfide bond reduction was observed by electrospray ionisation and Fourier transform ion cyclotron resonance mass spectrometry. Circular dichroism was used to monitor conformational changes of reduced proteins and of their unreduced counterparts [35].

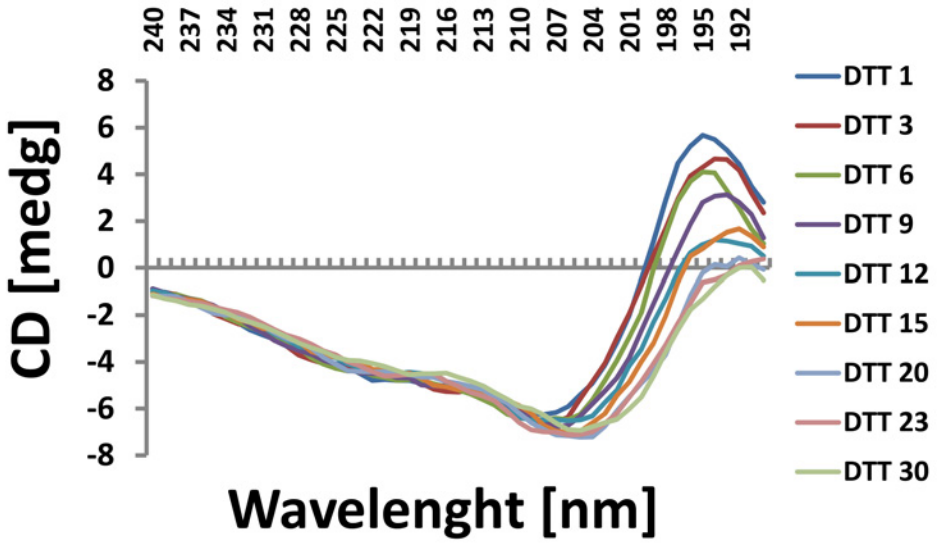

Fig. (2). Kinetics of S-S bond reduction at different DTT incubation times (minute).

5. Effect of pH on the Structure

Acidity has potential effects on folding and the molecular reactions. By using the insulin solution at different pH value and measuring the CD absorption spectra at two different wavelengths, 194.8 and 208.1 nm, it became evident that alpha helix is more prevalent at pH=5. While pH=10 and above have the least alpha helix contribution, similar values were obtained for pH=7 and pH=9. This has been shown in Fig. (**3**).

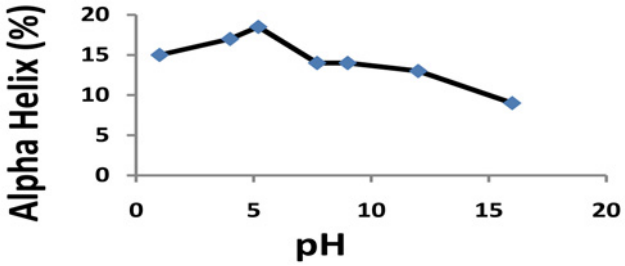

Fig. (3). Alpha helix content change according to change in pH; CD correlation data at 194.8nm.

6. Artificial Neural Network

The network architecture used was 54:32:7:1, and those are the node numbers for input layer, and two hidden layers and the output. The average error was 0.0059%. The epochs were 2047 to reach a level below the desired learning error. It could be possible to give text strings for the solution conditions in the building of the architecture, however, in general, better results would be reached if the real values were used and the output of treatment condition would be given numerical codes. The insulin solutions with various concentrations were also given different identification codes.

Sensitivity

To obtain the sensitivity, the inputs are all set to the median values and then each in turn is increased from the lowest value to the highest value. The change in the output is measured as each input is increased from lowest to highest to establish the sensitivity to change. Sensitivity Analysis is not the same as Input Importance. Sensitivity Analysis is a measure of how the outputs change when the inputs are changed. It is a method for measuring the cause and effect relationship between the inputs and outputs. Input Importance is a measure of how each input will influence the next layer in the network. Sensitivity Analysis shows how much an output changes when the inputs are changed. The highest values are related to insulin concentration, treatment concentration, and incubation treatment time. For the wavelengths, values between 190 and 203 nm, show a more significant impact. The correlation coefficient (R^2) between the predicted and the actual treatment type was 0.94, which is within the acceptable range concerning the versatility of information and size of the network (Fig. **4**). The acceptable neural network prediction results have been reported to be accurate within error of 5.5% [36]. Leave-n-out validation was carried out on the network. When n=1, it is also called jackknife cross-validation. Table **3** shows the Leave-n-out validation data. In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [37]. In the independent dataset test, although none of the proteins to be tested occurs in the training dataset used to train the predictor, the selection of proteins for the testing dataset could be quite arbitrary unless it is sufficiently large. This kind of arbitrariness may directly affect the conclusion. For instance, a predictor yielding higher success rate than the others for a testing dataset might fail to remain so when applied to another testing dataset [38]. For the subsampling test, the practical procedure often used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with the subsampling examination as such is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset (see Eq.50 of [39]). Therefore, any practical result by the sub-sampling test only represents one of many possible results, and hence cannot

avoid the arbitrariness either. In the jackknife cross-validation, each of the samples in the benchmark dataset is in turn singled out as a tested sample and the predictor is trained by the remaining samples. During the jackknifing process, both the training dataset and testing dataset are actually open, and a sample will in turn move from one to the other. The jackknife cross-validation can exclude the memory effects during entire testing process and also the result thus obtained is always unique for a given benchmark dataset. Therefore, of the above three examination methods, the jackknife test is deemed the most objective [38], and has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors [40-47].

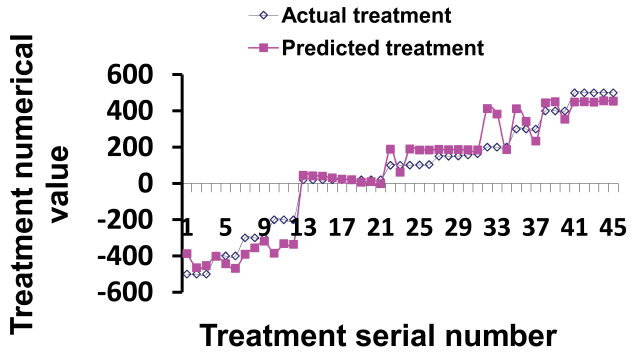

Fig. (4). The correlation between actual and predicted insulin downstream treatment indicators. The correlation coefficient (R^2) between the predicted and the actual treatment type was equal to 0.94.

Table 3. Leave-n-Out Validation Model Data

	Sum of Absolute Errors in Learning	
Left "n"	**Cycle 1**	**Cycle 10**
1	11.3131	12.4941
4	12.4989	12.7422

DISCUSSION

In the early days of insulin production, chains A and B were expressed and purified separately and then chemically transformed to the active insulin molecule by an oxidative step. Today, recombinant proinsulin is produced as a fusion protein and processed by several steps into active insulin. Optimized production strains of *E.coli* synthesize up to 40% of their cell mass as proinsulin fusion protein. Thus a 40 m^3 bioreactor can produce about 100 g of pure recombinant human insulin, which is about 1% of the annual world demand [35].

Circular dichroism (CD) is a valuable spectroscopic technique for studying protein structure in solution because many common conformational motifs, including alpha-helixes, L-pleated sheets, poly-L-proline II-like helices and turns, have characteristic far UV (190-250 nm) CD spectra. CD has an important role in the structural determinants of proteins. However, the effort expended in determining secondary structure elements is usually not worth it because it is somewhat unreliable. The real power of CD is in the analysis of structural changes in a protein upon some perturbation, or in comparison of the structure of an engineered protein to the parent protein.

Modern secondary structure determination by CD are reported to achieve accuracies of 0.97 for helices, 0.75 for beta sheet, 0.50 for turns, and 0.89 for other structure types [48]. For many proteins and peptides, the values of secondary structure are in direct correlation with their activity. The secondary structure would affect the three dimensional structure of the molecule and that would in turn, result in change in interaction of the peptide with its target receptor and pharmacological profile of the drug. Temperature as seen in Figs. (**1a-d**) has great impact on the unfolding and folding of insulin. As temperature is also part of the downstream process, it was included in the study to better consider the different conditions under which the peptide molecule is produced.

Incubation of proteins or peptides containing disulfide bonds (S-S) with DTT would result in cleavage of S-S bonds producing approximately equimolar amounts of free thiols (-SH). The effect of S-S bond cleavage of human insulin and the fusion protein on some of its structural properties, including conformation, were investigated. As shown in Table **2**, the effect of incubation period is incremental and is represented in the modification of secondary structures observed.

Cleavage of S-S bonds may decrease the solubility of human insulin and also shifts its isoelectric point to lower pH values. S-S bond cleavage resulted in changes in shape and hydrodynamic volume of the protein, increasing the specific viscosity, with cleavage of up to 3 S-S bonds. Both UV difference spectral measurements indicated that conformational flexibility increases with S-S bond cleavage. CD spectra of the fusion protein at native condition and under denaturing conditions (sulfitolyzed) were taken. Data were collected at 1-nm intervals over the range 260-200 nm for native protein, and 260-211 nm for denatured protein. Secondary structure estimations by far UV-CD suggested a gradual decrease in (alpha-helical and beta sheet) content of the protein with progressive cleavage of its S-S bonds. However, fully S-S bond cleaved protein maintained some (alpha-helical and random coil) structure. The analysis of the CD spectrum has shown the presence of approximately 35-37% alpha-helical structure (32% for insulin and 39% for proinsulin, [35] indicating a proinsulin-like structure of the S-sulfonated FP molecule, stable even in 7.5 M urea in spite of the absence of disulfide bonds in the S-sulfonated molecule. Sulfitolysis of the protein also decreases its (RP C18-binding) ability. In addition to the S-S bond content of the protein, in our study, the protein folding showed to depend on the reaction medium composition (including additives), and the presence of impurities like accompanied cell components, concentrations of protein, and temperature. This method can be used as a suitable IP-QC tool to determine the direction of downstream processing effectively.

CONCLUSIONS

In conclusion, there are various applications in the fields of ANN, including clinical data evaluation [49, 50] and drug development and molecular studies [30]. This study came up to show the importance of CD and power of ANN in modeling data from spectropolarimetric analysis of peptides in settings including the pharmaceutical industry. It is known that the CD data are valuable in the comparative experiments but not to determine the absolute shape and secondary structure of peptides and proteins. Therefore, modeling of CD data by supervised ANN methods further strengthens the ability of CD in predicting the information needed in research and production pharmaceutical peptides and proteins, which is particularly important in the case of limited access to *in vivo* facilities and saving time. Such applications can accompany the databases included in the CD instruments to facilitate the methodologies involved in molecular three dimensional structure determinations.

ACKNOWLEDGEMENT

This work was sponsored by a grant from Medical Biotechnology Network of Iran.

REFERENCES

[1] Q.X. Hua, S.Q.Hu, B.H. Frank, W. Jia, Y.C. Chu, S.H. Wang, G.T. Burke, P.G. Katsoyannis and M.A. Weiss, "Mapping the functional surface of insulin by design: structure and function of a novel A-chain analogue", *J. Mol. Biol*., vol. 264, pp. 390-403, 1996.

[2] J. Badger, M.R. Harris, C.D. Reynolds, A.C. Evans, E.J. Dodson, G.G. Dodson and A.C. North, "Structure of the pig insulin dimer in the cubic crystal", *Acta Cryst*., vol. 47, pp.127-136, 1991.

[3] W.F. Heath, R.M. Belagaje, G.S. Brooke, R.E. Chance, J.A. Hoffmann, H.B. Long, S.G. Reams, C. Roundtree, W.N. Shaw and L.J. Slieker, "Human proinsulin, a novel insulin agonist and intermediate in the synthesis of biosynthetic human insulin", *J. Biol. Chem*., vol. 267, pp. 419-425, 1992.

[4] M. J. Raftery, "Selective detection of thiosulfate-containing peptides using tandem mass spectrometry", *Rapid Commun. Mass. Spectrom*., vol. 19, pp. 674-82, 2005.

[5] M.H. Hecht, "De novo design of β-sheet proteins", *Proc. Natl. Acad. Sci. USA*, vol. 91, pp. 8729-8730, 1994.

[6] A. Marchler-Bauer, S.H. Bryant, "A measure of success in fold recognition", *Trends Biochem. Sci*., vol. 22, pp. 236-240, 1997.

[7] A.R. Kinjo, A. Kidera, H. Nakamura and K. Nishikawa "Physicochemical evaluation of protein folds predicted by threading", *Eur. Biophys. J*., vol. 30, pp.1-10, 2001.

[8] Q.X. Hua, S.H. Nakagawa, W. Jia, S.Q. Hu, Y.C. Chu, P.G. Katsoyannis and M.A. Weiss, "Hierarchical protein folding: asymmetric unfolding of an insulin analogue lacking the A7-B7 interchain disulfide bridge", *Biochemistry*, vol. 40, pp.12299-12311, 2001.

[9] K. C. Chou, "Structural bioinformatics and its impact to biomedical science", *Curr. Med. Chem*., vol. 11, pp. 2105-2134, 2004.

[10] K. C. Chou, "Molecular therapeutic target for type-2 diabetes", *J. Proteome Res*., vol. 3, pp. 1284-1288, 2004.

[11] K. C. Chou, D. Q. Wei and W. Z. Zhong, "Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS, (Erratum: ibid., 2003, Vol. 310, 675)", *Biochem. Biophys. Res. Commun*., vol. 308, pp. 148-151, 2003.

[12] Y. Li, D. Q. Wei, W. N. Gao, H. Gao, B. N. Liu, C. J. Huang, W. R. Xu, D. K. Liu, H. F. Chen and K. C. Chou, "Computational approach to drug design for oxazolidinones as antibacterial agents", *Med. Chem*., vol. 3, pp. 576-582, 2007.

[13] J. F. Wang, D. Q. Wei, C. Chen, Y. Li and K. C. Chou, "Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design", *Protein Pept. Lett*., vol. 15, pp. 27-32, 2008.

[14] K. C. Chou, G. Nemethy and H. A. Scheraga, "Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices", *J. Am. Chem. Soci*., vol. 106, pp. 3161-3170, 1984.

[15] K. C. Chou, G. M. Maggiora, G. Nemethy and H. A. Scheraga, "Energetics of the structure of the four-alpha-helix bundle in proteins", *Proc. Natl. Acad. Sci. USA*, vol. 85, pp. 4295-4299, 1988.

[16] S. Sirois, D. Q. Wei, Q. S. Du and K. C. Chou, "Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points", *J. Chem. Inf. Comput. Sci*., vol. 44, pp. 1111-1122, 2004.

[17] K. C. Chou, "Energy-optimized structure of antifreeze protein and its binding mechanism", *J. Mol. Biol*., vol. 223, pp. 509-517, 1992.

[18] K. C. Chou, H. B. Shen, "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization", *Biochem. Biophys. Res. Commun*., vol. 347, pp. 150-157, 2006.

[19] K. C. Chou, H. B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites", *J. Proteome Res.*, vol. 6, pp. 1728-1734, 2007.

[20] K. C. Chou, H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers", *J. Proteome Res*., vol. 5, pp. 1888-1897, 2006

[21] K. C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space", *Proteins: Struct. Funct. Genet*., vol. 21, pp. 319-344, 1995.

[22] K. C. Chou and H. B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM", *Biochem. Biophys. Res. Commun*., vol. 360, pp. 339-345, 2007.

[23] H. B. Shen and K. C. Chou, "EzyPred: A top-down approach for predicting enzyme functional classes and subclasses", *Biochem. Biophys. Res. Commun*., vol. 364, pp. 53-59, 2007.

[24] K. C. Chou and H. B. Shen, "ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information", *Biochem. Biophys. Res. Commun*., vol. 376, pp. 321-325, 2008.

[25] K. C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins", *J. Biol. Chem*., vol. 268, pp. 16938-16948, 1993.

[26] K. C. Chou, "Prediction of HIV protease cleavage sites in proteins", *Anal. Biochem*., vol. 233, pp. 1-14, 1996.

[27] H. B. Shen and K. C. Chou, "HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins", *Anal. Biochem*., vol. 375, pp. 388-390, 2008

[28] K. C. Chou and H. B. Shen, "Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides", *Biochem. Biophys. Res. Commun*., vol. 357, pp. 633-640, 2007.

[29] H. B. Shen and K. C. Chou, "Signal-3L: a 3-layer approach for predicting signal peptide", *Biochem. Biophys. Res. Commun*., vol. 363, pp. 297-303, 2007.

[30] S. Sardari and D. Sardari, "Applications of artificial neural network in AIDS research and therapy", *Curr. Pharm. Des*., vol. 8, pp. 659-670, 2002.

[31] R.C. Rowe and R.J Roberts. "Artificial intelligence in pharmaceutical product formulation: neural computing and emerging technologies", *PSTT*, vol. 1, pp. 200-205, 1998.

[32] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research", *J. Pharm. Biomed. Anal*., vol. 22, pp.717-727, 2000.

[33] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge, 1996, pp. 416.

[34] R.V. Tikhonov, S.E. Pechenov, I.A. Belacheu, S.A. Yakimov, V.E. Klyushnichenko and E.F. Boldireva, "Recombinant human insulin, VIII. Isolation of fusion protein±*S*-sulfonate, biotechnological precursor of, human insulin, from the biomass of transformed *Escherichia coli* cells", *Protein Exp. Purif*., vol. 21, pp. 176-182, 2001.

[35] M. Scigelova, P.S. Green, A.E. Giannakopulos, A. Rodger, D.H.G. Crout and P.J. Derrick, "A practical protocol for the reduction of disulfide bonds in proteins prior to analysis by mass spectrometry", *Eur. J. Mass. Spectrom*., vol. 7, pp. 29-34, 2001.

[36] Li M.M., X. Fan and K. Tickle. "Principal component analysis and neural networks for analysis of complex spectral data from ion backscattering", in the 25th Artificial Intelligence and Applications conferences. Innsbruck, 2006, pp. 17-21.

[37] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes", *Crit. Rev. Biochem. Mol. Biol*. vol. 30, pp. 275-349, 1995.

[38] K. C. Chou and H. B. Shen, "Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms", *Nat. Prot*., vol. 3, pp. 153-162, 2008.

[39] K. C. Chou and H. B. Shen, "Recent progresses in protein subcellular location prediction", *Anal. Biochem*., vol. 370, pp. 1-16, 2007.

[40] X. B. Zhou, C. Chen, Z. C. Li and X. Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes", *J. Theor. Biol.* vol. 248, pp. 546-551, 2007.

[41] G. Y. Zhang and B. S. Fang, "Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition", *J. Theor. Biol*., vol. 253, pp. 310-315, 2008.

[42] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", *J. Theor. Biol*., vol. 252, pp. 350-356, 2008.

[43] G. Y. Zhang, H. C. Li and B. S. Fang, "Predicting lipase types by improved Chou's pseudo-amino acid composition", *Protein Pept. Lett*., vol. 15, pp. 1132-1137, 2008.

[44] X. Jiang, R. Wei, T. L. Zhang and Q. Gu, "Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy", *Protein Pept. Lett*., vol. 15, pp. 392-396, 2008.

[45] F. M. Li and Q. Z. Li, "Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach", *Protein Pept. Lett*., vol. 15, pp. 612-616, 2008.

[46] H. Lin, H. Ding, F. B. Feng-Biao Guo, A. Y. Zhang and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using chou's pseudo amino acid composition", *Protein Pept. Lett*., vol. 15, pp. 739-744, 2008.

[47] T. Wang, J. Yang, H. B. Shen and K. C. Chou, "Predicting membrane protein types by the LLDA algorithm", *Protein Pept. Lett*., vol. 15, pp. 915-921, 2008.

[48] W.C. Johnson, "Protein secondary structure and circular dichroism: a practical guide", *Proteins*, vol. 7, pp. 205-214, 1990.

[49] E.Y.K. Ng, U. R. Acharya, L.G. Keith, S. Lockwood, "Detection and classification of breast cancer using neural classifiers with first warning thermal sensors", *Inf. Sci*., vol. 177, pp. 4526-4538, 2007.

[50] E.Y.K. Ng, S.C. Fok, "A framework for early discovery of breast tumor using thermography with artificial neural network", *Breast J*., vol. 9, pp. 341-343, 2003.

Received: August 26, 2008 Revised: November 17, 2008 Accepted: November 18, 2008