EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

Original article:

A NOVEL HYBRID METHOD OF B-TURN IDENTIFICATION IN PROTEIN USING BINARY LOGISTIC REGRESSION AND NEURAL NETWORK

Mehdi Poursheikhali Asghari[a], Sayyed Hamed Sadat Hayatshahi, Parviz Abdolmaleki[*]

Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Gisha, P.O. Box 14115/175, Tehran, Iran

[a] first author: Mehdi Poursheikhali Asghari; E-mail: mehdi.poursheikhali@modares.ac.ir
[*] corresponding author: Parviz Abdolmaleki, Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Gisha, P.O. Box 14115/175, Tehran, Iran. Tel: +98 21 82883404 Fax: +98 21 88009730 E-mail: parviz@modares.ac.ir

ABSTRACT

From both the structural and functional points of view, β-turns play important biological roles in proteins. In the present study, a novel two-stage hybrid procedure has been developed to identify β-turns in proteins. Binary logistic regression was initially used for the first time to select significant sequence parameters in identification of β-turns due to a re-substitution test procedure. Sequence parameters were consisted of 80 amino acid positional occurrences and 20 amino acid percentages in sequence. Among these parameters, the most significant ones which were selected by binary logistic regression model, were percentages of Gly, Ser and the occurrence of Asn in position i+2, respectively, in sequence. These significant parameters have the highest effect on the constitution of a β-turn sequence. A neural network model was then constructed and fed by the parameters selected by binary logistic regression to build a hybrid predictor. The networks have been trained and tested on a non-homologous dataset of 565 protein chains. With applying a nine fold cross-validation test on the dataset, the network reached an overall accuracy (Q_{total}) of 74, which is comparable with results of the other β-turn prediction methods. In conclusion, this study proves that the parameter selection ability of binary logistic regression together with the prediction capability of neural networks lead to the development of more precise models for identifying β-turns in proteins.

Keywords: β-turns, binary logistic regression, neural networks, secondary structure prediction, sequence parameters

INTRODUCTION

Protein secondary structure prediction is a preceding step to the more complicated tertiary structure prediction (Richardson, 1981). Among many structural elements, tight turns play an important role in protein folding and stability. They are classified as δ-turns, γ-turns, β-turns, α-turns and π-turns, depending on the number of residues forming the turn (Chou, 2000). β-turns are the most existing type of tight turns in proteins and include almost 25 % of all residues in globular proteins (Kabsch and Sander, 1983). They consist of four consecutive residues defined by positions i, i+1, i+2 and i+3. The distance between C_α (i) and C_α (i+3) is less than 7 Å (Chou, 2000). According to the Ø, ψ angles of the residues i+1 and i+2, β-turns can be classified into 9 different types: I, I′, II, II′, IV, Via1, Via2, VIb and VIII.

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

Both from structural and functional points of view, β-turns play important biological roles in proteins. They tend to be found at solvent-exposed surfaces and therefore involve in molecular recognition processes between proteins, as well as in interactions between peptide substrates and receptors (Rose et al., 1985). β-turn formation is a determining stage during the process of protein folding. Also, β-turns are responsible for the compact globular shape of proteins since they have the ability for reserving the alignment of protein chain (Takano et al., 2000). Hence, the development of a precise method for analysis and prediction of β-turns (according to the amino acid sequences) would be useful for protein folding studies as well as for predicting the overall three-dimensional structure of proteins.

Many efforts have been made for analysis and prediction of β-turns in proteins. They can be divided into two categories: statistics-based and machine learning-based methods. The majority of statistics-based methods used positional preferences of amino acids in β-turns (Lewis et al., 1973; Chou and Fasman, 1974; Wilmot and Thornton, 1988; Hutchinson and Thornton, 1994; Zhang and Chou, 1997; Fuchs and Alix, 2005) .The second category includes neural network (NN) (McGregor et al., 1989; Shepherd et al., 1999; Kaur and Raghava, 2003, 2004; Kirschner and Freshman, 2008; Petersen et al., 2010) as well as support vector machine (SVM) (Cai et al., 2003; Pham et al., 2003, 2005; Zhang et al., 2005; Zheng and Kurgan, 2008; Hu and Li, 2008 ; Liu et al., 2009; Meissner et al., 2009; Kountouris and Hirst, 2010; Shi et al., 2011; Tang et al., 2011) approaches. To compare main methods of β-turn prediction, Kaur and Raghava (2002) have made an evaluation on the benchmark data set. They showed that neural network approach by Shepherd et al. (1999) presented the best prediction performance among other evaluated methods. In a previous study based on a hybrid approach, we employed the multinomial logistic regression as well as neural networks for analysis and identification of β-turn types (Asgary et al., 2007). More recently, Zheng and Kurgan (2008) used SVM for β-turn prediction getting the performance of their method is the highest among all. As a result, machine learning methods (especially SVM approach) can be considered as the most accurate ones for prediction of β-turns.

In the year 2002, Kaur and Raghava suggested that combining a statistics method with a machine learning method may provide substantially better results than either one alone (Kaur and Raghava, 2002). In the present study, we followed their recommendation by combining the binary logistic regression as statistical method with the neural network as machine learning one for identification of β-turns.

The binary logistic regression method, which has not been applied for β-turn analysis so far, is useful when the presence or absence of a characteristic or an outcome based on a set of predictor variables is needed to be predicted. It is similar to a linear regression model but is suited to models with dichotomous dependent variable (Hosmer and Lemeshow, 2000). We used binary logistic regression to select the most effective set of parameters which then were fed into a well-established neural network. In this way, we increased the accuracy and reliability of neural networks, in β-turns identification.

MATERIALS AND METHODS

The dataset

Our dataset consisted of 565 non-homologous protein chains (Table 1). These protein chains were selected using the PDB-REPRDB server (Noguchi et al., 2001). In this dataset, no two proteins have more than 25 % sequence uniformity. All proteins have reported X-ray structures with 2.0 Å resolution or better. The program PROMOTIF (Hutchinson and Thornton, 1996) was employed to identify β-turns in the proteins. Sequence parameters including 80 amino acid positional occurrences as well as 20 amino acid percentages (of existence) in β-turn sequences were generated

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

using IF and COUNTIF functions of EXCEL software (2003), respectively.

Table 1: The PDB (Protein Data Bank) codes of 565 protein chains

1NWZA, 1P9GA, 1MUWA, 1V6PA, 1G66A, 1IQZA, 1L9LA, 1OK0A, 1GVKB, 4LZT_, 1NKIA, 1GQVA, 1UG6A, 1P1XA, 1EB6A, 1LNIB, 1A6M_, 2ERL_, 1EXRA, 1R2MA, 1C7KA, 1QTWA, 1CC8A, 1LWBA, 1PSRB, 1SFSA, 1UWCA, 1N62B, 1C5EA, 1QL0A, 1P5FA, 1SU8A, 1YFQA, 1RG8A, 1KT6A, 1I1XA, 1K7CA, 1SAUA, 1TUKA, 1P6OB, 1OE3A, 1J0OA, 1HEUA, 1HG7A, 1M1NB, 1M1NA, 1C9OA, 1CZPA, 1WKQA, 1T1EA, 1IFC_, 1ZBYA, 1I6TA, 1JBC_, 1ARB_, 2BMOA, 1TU9A, 1UAIA, 1X6IB, 1AMM_, 2SN3_, 2PTH_, 1V8HA, 1HXHB, 1GNLA, 1MEXL, 1LXZA, 1QFTA, 1WDPA, 1MQKH, 1VLBA, 1QKSA, 1C52_, 1RTTA, 1G61A, 1JR0F, 1OX0A, 1K3YA, 1LQ9A, 1WVFA, 1WKRA, 1SG4A, 7FD1A, 1BXAA, 1RRO_, 1I0DA, 1S2PA, 1ISPA, 1F41A, 1ULRA, 1XEOA, 1V70A, 1NYKA, 1VF8A, 1VH5A, 1U1WB, 1SJWA, 2LISA, 1FD3A, 1F1GA, 1BSMA, 1TZVA, 1MQOA, 1THM_, 1H2WA, 1LLFA, 1GK8I, 1EZGA, 1L6RA, 1GG6B, 1GG6C, 1Y8AA, 1RKQA, 1QK8A, 1N13B, 3VUB_, 1N13E, 1YGE_, 1YRCA, 1G8AA, 1NYCA, 1V37A, 1H2RL, 1FP2A, 1H2RS, 1JY2O, 1QH4A, 1XSZA, 1NOFA, 1MXRA, 1PKHA, 1PP0B, 1E6UA, 1YYDA, 1O9RA, 1HZTA, 1QH5A, 1C1KA, 1V4PA, 1C8CA, 1O82A, 1JU2A, 1SMOB, 1R7JA, 1GQIA, 1YXLA, 1ZD8A, 1XNB_, 1HBZA, 1HT6A, 1OFZA, 1G6SA, 1OFNB, 1NKGA, 1QWOA, 1F1UA, 1DJ0A, 1CYO_, 1V5DA, 1KQ3A, 1OC2B, 1OFWA, 1UV4A, 1DQZA, 2BKVB, 1IQQA, 1ISUA, 1MLA_, 1I1NA, 1UI0A, 1M2AA, 3EZMA, 1GUTA, 1WHI_, 1GD0B, 1P1MA, 1V5VA, 1AH7_, 1QB7A, 1NP4A, 1I0RB, 1ROCA, 1F46B, 1AGI_, 1S67L, 1ZHLB, 1KPF_, 1IRQB, 1IO7A, 1LLMC, 1KRHA, 1OMRA, 1Y1NA, 1YPYA, 1YME_, 1T6CA, 1USGA, 1RV9A, 1DLWA, 1SZNA, 2CBA_, 3GRS_, 2BEMA, 1QQJA, 1NA3A, 1EDQA, 1ZG4A, 1U11B, 1U8YB, 1JKEC, 1UGIA, 1KQ1H, 1SXRA, 1JOVA, 1IZ7A, 1RP0A, 1BFD_, 3PTE_, 1Y7BA, 1U8VA, 1E6YE, 1HFEL, 1HFES, 1KWGA, 1LAM_, 1S9RA, 1RU4A, 2SIL_, 1T4BA, 1KEIA, 1YYAA, 1M44B, 1T92A, 1B3AA, 1JI1A, 1SMD_, 1FS7A, 1MRP_, 1L9XA, 1CUOA, 1Q0GA, 1NC5A, 1QGIA, 1EG9B, 1CG5B, 1WS8B, 1KNT_, 1QWKA, 1O26B, 1YW5A, 1MD6A, 1W0PA, 1NM8A, 1U0EA, 1IT2A, 1GY6A, 1U84A, 1FWXA, 1HTRP, 1H63A, 1IWDA, 1G8KA, 1KAPP, 1X3KA, 1I9DA, 1YO3A, 3STDA, 1EZWA, 1DHN_, 1US6B, 1WUBA, 3CHY_, 1E3UB, 1DOSA, 1B4KA, 1CZFA, 1EU3A, 1Y7YA, 1T0BH, 1AYX_, 1TXGA, 1Q8FA, 1BKPB, 1DP0A, 1KNB_, 2MHR, 1CSS_, 1VCLA, 1JHDA, 1YOCB, 1MOLA, 1GOF_, 1MTYD, 1PMI_, 1PB1A, 1MTYB, 1NE9A, 1AGJA, 1MTYG, 1UQRD, 1PBYA, 1WY2A, 1PXZA, 1NVMG, 1NVMB, 1V58A, 1IS6A, 1B2PA, 1YGTA, 1V9FA, 1WAB_, 1MXIA, 1W4XA, 1R12A, 1FVAB, 1FZQA, 1NAQA, 1SL8A, 1MK4A, 1TXLA, 1QSTA, 1Q2HB, 1P3QU, 1SF9A, 1DJEA, 1GXYA, 1B5FA, 1LTUA, 1K6WA, 1T0TV, 1PZ3A, 1UMKA, 2BBKL, 1CQXA, 1NPYB, 1CHD_, 1CV8_, 1PM4A, 1WD3A, 1XX1A, 7YASA, 1UDH_, 1KU8A, 1UJ8A, 1GTFA, 1GNUA, 1H4PA, 1R8NA, 1XKRA, 1ZAIA, 1C8KA, 1R4PB, 1QAZA, 1AXN_, 1KHIA, 2BLFB, 1SVDA, 2NACA, 1NTFA, 2HVM_, 1PAMA, 1PT7A, 1MH9A, 1HPI_, 1VD5A, 1FN9A, 1ML4A, 1J1QA, 1GBG_, 1H0HB, 1LENC, 1C44A, 1LENB, 1RYIA, 1RJDC, 1LD8A, 1JUEA, 1DOZA, 1TML_, 1GBS_, 1RLHA, 1XXOA, 1LY2A, 1ULKA, 1J48A, 1EPTB, 1EPTA, 1QJDA, 1OWLA, 1VFRA, 1G5TA, 1GDEA, 1LJ5A, 1MML_, 1L8FA, 1C02A, 1XTYA, 2SPCA, 1CMBA, 1V54H, 1FOBA, 1KN3A, 1R9DA, 1GIQA, 1IG0B, 1IQCA, 1RWZA, 1G8EA, 3EIPA, 1V77A, 1H8UA, 1X8DA, 1GND_, 1DMR_, 1V73A, 1RY9A, 1IYEA, 1EUHA, 1WNHA, 1QBA_, 1L9FB, 1IWBA, 1XS5A, 1M5SB, 1EX2A, 1VLS_, 1J7DA, 1GAKA, 1DYR_, 1ONRA, 1XX2A, 1WF3A, 1QGDA, 2PII_, 1JFRA, 1OAOC, 1KV9A, 1ODSA, 1GYCA, 1H7WD, 1TG7A, 1ON3E, 1UA4A, 1BGVA, 1B8AA, 1QGJA, 1BUDA, 1KEKA, 1R8CA, 1UYPA, 1YKDA, 1CPO_, 1M0SA, 1QB5D, 1PNKB, 1F20A, 1SFTA, 1TM2A, 1ESGB, 1G0SB, 1PNKA, 1K12A, 2CHSA, 1YEA_, 1TZYH, 1Q16A, 1IVUA, 1Q16B, 1EU8A, 1JK7A, 1J9JA, 1T06A, 1JD5A, 1AGQD, 1MZL_, 1OTFA, 1EJJA, 1U1IA, 1UJ1B, 1B94A, 1Y7PB, 1QNXA, 1I5ZB, 1IAKB, 1F5MB, 1F4QB, 1PU5A, 1STMA, 1DUTA, 1C8UA, 1DZFA, 1FW1A, 1QU1F, 1XXUA, 1Z6OM, 1FUA_, 1K7HA, 1EKJG, 1KBLA, 1KX5C, 1LTSC, 1QHDA, 1FO6A, 1KLXA, 1GVNB, 1GVNC, 1NFVI, 1IHBA, 1YNHB, 2HRVA, 1HCZ_, 7NN9_, 1BF2_, 1RQHA, 1CVRA, 2EBN_, 1TL2A, 1OBPA, 1ESL_, 1ECFB, 2HPDA, 1FP3A, 1TVFA, 1NSYA, 1APYB, 1TGJ_, 1PMMA, 1VPNB, 1IGS_, 1UGQA, 1SR4C, 1UOK_, 2PIA_, 1AOCA, 1WPBA, 1EVXA, 1YF9B, 1M4RA, 1R5TA, 1CDCA, 1CB6A, 1MW3A, 1OGSA, 2PGD_, 1H8ED, 2PLC_, 1XCLA, 1XRJB, 1N3FA, 1COZA, 1VC1A, 1JS1X, 1A8P_, 1JWIB, 1A1X_, 1HHSA, 1FNF_, 1NIJA, 1KCMA, 1OQWA, 1R7LA, 1LPJA, 1NFJA, 1K32A, 1IIPA.

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

Model design

Initially, binary logistic regression serves as a non-linear model on the dataset to select significant parameters due to the *"Re-substitution Test"*. This test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. Certainly, a prediction algorithm cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating an identification method. When this test was implemented, each tetra peptide in the dataset concerned is in turn identified using the rule parameters derived from the same dataset, the so-called training dataset. After using the binary logistic regression in this manner, the NNs (which act non-linearly in the last stage of this hybrid procedure) were fed by the outputs of binary logistic regression to predict β-turns. The NN method has been trained and tested using 9-fold cross-validation techniques, whereby the dataset is divided into nine subsets (i.e. 8 subsets containing 62 protein chains, 1 subset containing 61 protein chains). The method has been trained on eight subsets and the performance was calculated on the remaining ninth subset. This procedure was repeated nine times, once for each subset. Actually, the *"Cross-validation Test"* can reflect the effectiveness of an identification method in practical application.

Binary logistic regression model

This model is used only when the dependent variable is dichotomous, that is, there are only two possible answers for the dependent variable. Let the dependent variable be Y. Since it is dichotomous, it takes on 0 or 1 for failure and success, respectively. The logistic regression model can be expressed as follows (Hosmer and Lemeshow, 2000):

$$\text{Log}\,[p/(1-p)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n,$$

where p is the probability of Y=1, β_0 is a constant, and $\beta_1 - \beta_n$ are unknown logistic regression coefficients of independent variables x_1-x_n (amino acid occurrences or percentages in β-turn sequence). The ratio p/1-p takes on values between 0 and plus infinity. Therefore, the logarithm of this ratio (logit) is a continuous variable that takes on values between minus infinity and plus infinity. Using this equation, the value of logit is determined. Then a cutoff should be taken to recode logit values into two possible states of dependent variable (i.e. non-β-turn sequence and β-turn sequence) (Hosmer and Lemeshow, 2000). The optimized cutoff value in this research was 0.5.

Several different options are available during the creation of logistic regression model. Independent variables can be entered into the model in the order specified by the researcher and logistic regression can test the fit of the model after each coefficient is added or deleted, called *"stepwise regression"*. Stepwise regression is used in the exploratory phase of research. We used the *Backward Wald (stepwise)* binary logistic regression routine in SPSS program to develop our model. This routine appears to be the preferred method of exploratory analysis, where the analysis begins with a full or saturated model and independent variables are eliminated from the model in an iterative process. The fit of the model is tested after the elimination of each independent variable to ensure that the model still adequately fits the data. When no more independent variables can be eliminated from the model, the analysis has been completed. The measure for model fitness in each step is an index called *-2 Log Likelihood*. In general, as the model becomes better, this index (-2LL) will decrease in magnitude. In fact, in backward Wald routine, the first step has the minimum value of -2 log likelihood and hence its reported result (i.e. *Parameter Estimates Table*) is the main output of binary logistic regression model (Hosmer and Lemeshow, 2000).

Neural network model

The neural network (NN) was utilized as a robust non-linear predictor in hybrids with the binary logistic regression. In this way, the selected variables from binary logistic regression model were used as input

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

nodes of neural network. This is supposed to decrease the number of input nodes, simplify the network architecture and shorten the time needed for model building. We used feed-forward back propagation networks with a single hidden layer. Using such algorithm, the parameters related to the training cases were fed into the networks. The final outputs estimated by the networks were compared with the actual type of cases; generating a mean of the sum of square error (MSE). This quantity was propagated back into the networks to adjust the randomly chosen weights. The training cases were then tested with new weights and the procedure repeated. Through such process, the MSE was minimized.

Three layer networks were used in this study. Each unit in the input layer was fed by one independent variable which has been selected by binary logistic regression model. The output layer included two units which represented 1 0 and 0 1 for β-turn and non-β-turn sequence, respectively.

The MSE was utilized as an index of network efficiency in determining the optimized number of hidden units (Hayatshahi et al., 2005). To do so, the number of hidden units was changed in every network in order to develop networks producing the minimal MSE. At last, after such optimizing process, the number of units for hidden layer reached 8.

The final neural network structure was consisted of 33 units in input layer, 8 units in hidden layer and 2 units in output layer. The activation function of hidden layer units was *logsig*. We used the *Quasi-Newton* training function in this research. This training function is prior to simple 'batch' gradient-descent and lead to significantly better solutions requiring fewer training steps. Besides, this method does not suffer from the specification problem of the learning rate parameter which is crucial for the performance of the gradient-descent method (Likas and Stafylopatis, 2000).

Training has been done for 1000 epochs for nine networks. The value of the learning rate parameter has been set to 0.2. The software employed to build neural networks was in-house written in the MATLAB programming language.

Performance measures

Four different parameters have been used to measure the performance of prediction methods. These four parameters can be derived from the four scalar indices: TP (true positives: number of correctly classified β-turns), TN (true negatives: number of correctly classified non-β-turns), FP (false positives: number of non-β-turns incorrectly classified as β-turns) and FN (false negatives: number of β-turns incorrectly classified as non-β-turns). Using the following formulas which have been previously reported in the published material, we calculated these parameters for the output of binary Logistic Regression and NN models.

(1) $$Q_{total} = \left(\frac{TP+TN}{TP+TN+FP+FN}\right)\times 100$$

which is the fraction of correctly predicted β-turns and non-β-turns among all predictions.

(2) $$Q_{predicted} = \left(\frac{TP}{TP+FP}\right)\times 100$$

which is the percentage of correctly predicted β-turns.

(3) $$Q_{observed} = \left(\frac{TP}{TP+FN}\right)\times 100$$

which is the percentage of observed β-turns that are correctly predicted.

(4) Matthews correlation coefficient (MCC): We used MCC as a more robust measure to evaluate the reliability of the established method (Matthews, 1975). The MCC is defined by

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The MCC is a limited number tween -1 and 1. If there is no relationship between the predicted values and the actual values, the MCC should be 0 or very low (the predicted values are not better than random numbers). In contrast, the MCC value would increase as the strength of the

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

relationship between the predicted values and actual values increases. It is obvious that a perfect fit gives a coefficient of 1.0. Furthermore, the higher MCC indicates the better performance of the prediction for the model.

Statistical analysis was performed using SPSS 13 for Windows (SPSS Inc., Chicago, USA).

RESULTS

Binary logistic regression analysis

Binary logistic regression model was runned on the dataset using the *re-substitution test*. One of output tables of binary logistic regression model was *"Omnibus Tests of Model Coefficients"* (Table 2). This table reports significance levels by the traditional chi-square method and is an alternative to the "Hosmer-Lemeshow Test" (Hosmer and Lemeshow, 2000). Among 24 steps, the only positive chi-square value (i.e. significant) can be seen in step 1 which is 6429.548 (P-value= 0.000) (Table 2). Since the probability of the first step chi-square was less than the significance level (0.05), the existence of a relationship between independent variables (100 structural parameters) and the dependent variable (β-turn and non-β-turn sequences) was supported.

Also, according to *"Model summary"* table in the output of binary logistic regression model (Table 3), we see that *-2 log likelihood index* has its minimum value in the first step of the model (47450.373). The lowest value of this index indicates the best step of the model (Hosmer and Lemeshow, 2000). Therefore, the first step was recognized as the reference step.

Table 2: Omnibus tests of model coefficients

Sig.	Chi-square	Step
0.000	6429.548	1
0.928	-0.008	2
0.943	-0.005	3
0.905	-0.014	4
0.921	-0.010	5
0.875	-0.025	6
0.752	-0.100	7
0.637	-0.222	8
0.613	-0.255	9
0.588	-0.294	10
0.534	-0.386	11
0.623	-0.242	12
0.450	-0.571	13
0.443	-0.590	14
0.358	-0.845	15
0.251	-1.319	16
0.212	-1.558	17
0.196	-1.669	18
0.157	-2.003	19
0.125	-2.358	20
0.115	-2.485	21
0.119	-2.430	22
0.307	-1.042	23
0.143	-2.149	24

Table 3: Model summary

-2 Log likelihood	Step
47450.373	1
47450.381	2
47450.387	3
47450.401	4
47450.411	5
47450.435	6
47450.536	7
47450.758	8
47451.013	9
47451.307	10
47451.694	11
47451.936	12
47452.507	13
47453.096	14
47453.941	15
47455.260	16
47456.819	17
47458.488	18
47460.491	19
47462.849	20
47465.334	21
47467.764	22
47468.807	23
47470.955	24

Table 4 shows *parameter estimates (β), standard errors, Wald statistic, p-values* and *corresponding odds ratios* for selected parameters among 100 ones, for β-turns in contrast to non-β-turns as reference group of the backward-Wald binary logistic regression procedure. This information is related to the first step of the model. Among sequence parameters, 13 amino acid percentages and 35 amino acid positional occurrences were found to be significant in determining β-turns and non-β-turns.

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

Table 4: Parameter estimates, indicating statistical results for the significant parameters in the output of binary logistic regression procedure (β-turns in contrast to non-β-turns (as reference group))

odds ratios {Exp (β)}	P-values	Wald statistics	Standard errors	Parameter estimates (β)	Variable
0.047	0.000	935.030	0.100	- 3.068	step1 constant
2.864	0.000	13.645	0.285	1.052	Arg (%)
3.962	0.000	25.076	0.275	1.377	Asn (%)
4.217	0.000	30.435	0.261	1.439	Asp (%)
3.309	0.002	9.141	0.396	1.197	Cys (%)
2.246	0.006	7.496	0.295	0.809	Gln (%)
43.762	0.000	272.766	0.229	3.779	Gly (%)
3.334	0.000	12.545	0.340	1.204	His (%)
0.414	0.002	9.414	0.288	- 0.882	Ile (%)
0.411	0.001	11.659	0.260	- 0.889	Leu (%)
3.165	0.000	18.933	0.265	1.152	Lys (%)
3.598	0.000	21.408	0.277	1.280	Pro (%)
4.522	0.000	34.775	0.256	1.509	Ser (%)
3.540	0.000	23.368	0.262	1.264	Thr (%)
1.648	0.000	27.067	0.096	0.499	Asn (i)
2.094	0.000	66.749	0.090	0.739	Asp (i)
1.413	0.014	6.054	0.140	0.346	Cys (i)
0.788	0.030	4.710	0.110	- 0.238	Gln (i)
0.623	0.000	31.460	0.084	- 0.473	Gly (i)
1.353	0.001	10.729	0.092	0.302	Leu (i)
1.368	0.003	8.711	0.106	0.313	Phe (i)
2.031	0.000	55.227	0.095	0.708	Pro (i)
1.461	0.000	17.597	0.090	0.379	Ser (i)
1.203	0.048	3.916	0.094	0.185	Thr (i)
1.416	0.000	15.038	0.090	0.348	Ala (i+1)
1.273	0.019	5.544	0.103	0.242	Arg (i+1)
1.214	0.050	3.841	0.099	0.194	Asn (i+1)
1.441	0.000	15.693	0.092	0.366	Asp (i+1)
0.665	0.008	7.117	0.153	- 0.409	Cys (i+1)
2.141	0.000	65.065	0.094	0.761	Glu (i+1)
0.609	0.000	33.945	0.085	- 0.496	Gly (i+1)
1.244	0.021	5.355	0.094	0.218	Leu (i+1)
1.597	0.000	24.534	0.094	0.468	Lys (i+1)
3.691	0.000	195.250	0.093	1.306	Pro (i+1)
1.291	0.005	7.718	0.092	0.256	Ser (i+1)
1.534	0.000	15.889	0.107	0.428	Arg (i+2)
4.223	0.000	216.433	0.098	1.440	Asn (i+2)
3.367	0.000	166.228	0.094	1.214	Asp (i+2)
1.308	0.018	5.626	0.113	0.268	Gln (i+2)
1.945	0.000	43.474	0.101	0.665	Glu (i+2)
2.347	0.000	97.801	0.086	0.853	Gly (i+2)
2.166	0.000	39.088	0.124	0.773	His (i+2)
1.365	0.002	9.590	0.100	0.311	Leu (i+2)
1.453	0.000	13.718	0.101	0.373	Lys (i+2)
1.855	0.000	30.815	0.111	0.618	Phe (i+2)
1.674	0.000	28.287	0.097	0.515	Ser (i+2)
1.675	0.000	27.589	0.098	0.516	Thr (i+2)
1.563	0.003	8.631	0.152	0.447	Trp (i+2)
1.833	0.000	29.820	0.111	0.606	Tyr (i+2)

EXCLI Journal 2012;11:346-356 – ISSN 1611-2156
Received: March 05, 2012, accepted: July 04, 2012, published: July 05, 2012

Percentages of *Gly, Ser, Asp, Asn, Pro, Thr, His, Cys, Lys, Arg* and *Gln*, respectively, have the most positive parameter estimate values among percentages of other amino acids. Therefore the probability of the sequence to be β-turn increases as increasing their values. On the other hand, percentages of *Ile* and *Leu*, respectively, have the most negative parameter estimate values among percentages of other amino acids and hence the probability of the sequence to be non-β-turn will increase as increasing values.

Occurrences of *Asp, Pro, Asn, Ser, Cys, Phe, Leu* and *Thr* respectively in position *i* of sequence have the most positive logit coefficients (or parameter estimates) among occurrences of other amino acids in the same position and support of β-turn sequence. Vice versa, occurrences of *Gln* and *Gly* respectively in position *i* of β-turn sequence have the most negative logit coefficients among occurrences of other amino acids in that position and support of non-β-turn sequence.

In analysis of position *i+1* of sequence, occurrences of *Pro, Glu, Lys, Asp, Ala, Ser, Arg, Leu* and *Asn*, respectively, have the highest logit coefficients and occurrences of *Cys* and *Gly* have the most negative logit coefficients among others. Obtained results can be interpreted like above-mentioned sentences.

Ultimately, the final position of sequence which is highlighted in the table is *i+2*. Occurrences of *Asn, Asp, Gly, His, Glu, Phe, Tyr, Thr, Ser, Trp, Arg, Lys, Leu* and *Gln*, respectively, have the highest parameter estimates among others and hence they support the sequence to be β-turn.

The result of re-substitution {Self-Consistency} test was evaluated by the performance measures. The results shown in Table 5 are obtained according to the output of the model.

Neural network

We fed our neural networks with 33 significant parameters selected in the re-substitution binary logistic regression procedure to a build two-stage hybrid model. The number of units in the hidden layer was optimized in networks regarding the least MSE rate (refer you to *Materials and Methods section*). The final MSE rate was 0.13 (with the eight units in the hidden layer) which was the lowest among different examined numbers of units in the hidden layer. Ultimately, we ended to an optimal neural network architecture with 33 input units and a single hidden layer with 8 units for our binary prediction (i.e. be β-turn or non-β-turn). A nine fold cross-validation procedure was used for prediction of β-turns. The performance of the model was evaluated by averaging the mentioned measures over nine sets.

The prediction results using neural networks are presented in Table 5. With applying a nine fold cross-validation test on the dataset, it would be found that the network reached an overall accuracy (Q_{total}) of 74. Also, the network yielded $Q_{predicted}$ value of 45 and $Q_{observed}$ value of 30. Ultimately, the value of MCC for the network was 0.21.

Table 5: Prediction results of our two-stage hybrid procedure

Test	MCC	$Q_{observed}$	$Q_{predicted}$	Q_{total}
Re-substitution (Binary logistic regression)	**0.25**	**21**	**59.4**	**77.2**
Cross-validation (neural networks)	**0.21**	**30**	**45**	**74**

DISCUSSION

An important advantage for using logistic regression model is its capability of determining weights of each selected significant parameter, which highlights its priority of importance in clarifying the sequence-structure relationship. This study shows that this parameter selection ability of binary logistic regression in combination to the prediction ability of neural networks leads to the development of more precise models.

Binary logistic regression analysis showed that only 48 structural parameters among 100 ones were significant in identification of β-turns. Percentage of *Gly* in sequence was the most important parameter with the parameter estimate (β) value of 3.779 and the odds ratio value of 43.762. The second major parameter was the percentage of *Ser* with the parameter estimate and odds ratio values of 1.509 and 4.522, respectively. The third important parameter was the occurrence of *Asn* in positions *i+2* of the sequence. Consequently, these three parameters have the highest effect on the constitution of β-turn sequence, among others. On the other hand, the percentage of *Leu* in sequence had the most negative parameter estimate value (i.e. –0.889). Thus this parameter has the highest effect on the constitution of non-β-turn sequence, among others.

In conclusion, our research highlighted the efficiency of using the statistical model of binary logistic regression as a preprocessor in determining effective parameters. Besides, the optimal structure of neural network can be simplified by a preprocessor in the first stage of hybrid approach, which in turn causes decreasing the time needed for neural network training procedure in the second stage.

ACKNOWLEDGEMENTS: We thank Professor Anoshirvan Kazemnejad and Mr Samad Jahandideh for their helpful comments and discussions. We also thank the Department of Biophysics (from the Tarbiat Modares University, Iran) for financial support.

REFERENCES

Asgary MP, Jahandideh S, Abdolmaleki P, Kazemnejad A. Analysis and identification of β-turn types using multinomial logistic regression and artificial neural network. Bioinformatics 2007;23:3125-30.

Cai YD, Liu XJ, Li YX, Xu XB, Chou KC. Prediction of β-turns with learning machines. Peptides 2003;24:665-9.

Chou KC. Prediction of tight turns and their types in proteins. Anal Biochem 2000;286: 1-16.

Chou PY, Fasman GD. Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. Biochemistry 1974;13:211-22.

Fuchs PFJ, Alix AJP. High accuracy prediction of β-turns and their types using propensities and multiple alignments. Proteins 2005;59:828-39.

Hayatshahi SHS, Abdolmaleki P, Safarian S, Khajeh K. Non-linear quantitative structure-activity relationship for adenine derivatives as competitive inhibitors of adenosine deaminase. Biochem Biophys Res Com 2005;338:1137-42.

Hosmer DW, Lemeshow S. Applied logistic regression (Groves RM, Kalton G, Rao JNK, Schwarz N, Skinner C, eds.). New York: Wiley, 2000.

Hu X, Li Q. Using support vector machine to predict beta- and gamma-turns in proteins. J Comput Chem 2008;29:1867-75.

Hutchinson EG, Thornton JM. A revised set of potentials for β-turn formation in proteins. Protein Sci 1994;3:2207-16.

Hutchinson EG, Thornton JM. PROMOTIF: A program to identify and analyze structural motifs in proteins. Protein Sci 1996;5:212-20.

Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-637.

Kaur H, Raghava GPS. An evaluation of β-turn prediction methods. Bioinformatics 2002;18:1508-14.

Kaur H, Raghava GPS. Prediction of β-turns in proteins from multiple alignments using neural network. Protein Sci 2003;12: 627-34.

Kaur H, Raghava GPS. A neural network method for prediction of β-turn types in proteins using evolutionary information. Bioinformatics 2004;20:2751-8.

Kirschner A, Freshman D. Prediction of beta-turns and beta-turn types by a novel bidirectional Elman-type recurrent neural network with multiple output layers (MOLEBRNN). Gene 2008;422:22-9.

Kountouris P, Hirst GD. Predicting β-turns and their types using predicted backbone dihedral angles and secondary structures. BMC Bioinformatics 2010;11:407.

Lewis PN, Momany FA, Scheraga HA. Chain reversals in proteins. Biochem Biophys Acta 1973;303:211-29.

Likas A, Stafylopatis A. Training the random neural network using quasi-Newton methods. Eur J Oper Res 2000;126:331-9.

Liu L, Fang Y, Li M, Wang C. Prediction of beta-turn in protein using E-SSpred and support vector machine. Protein J 2009;28: 175-81.

Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta 1975;405:442-51.

McGregor MJ, Flores TP, Sternberg MJE. Prediction of β-turns in proteins using neural network. Protein Eng 1989; 2: 521-526.

Meissner M, Koch O, Klebe G, Schneider G. Prediction of turn types in protein structure by machine-learning classifiers. Proteins 2009;74:344-52.

Noguchi T, Matsuda TH, Akiyama Y. PDB_REPRDB: A database of representative protein chains from the Protein Data Bank (PDB). Nucleic Acids Res 2001;29: 219-20.

Petersen B, Lundegaard C, Petersen TN. NetTurnP - Neural network prediction of Beta turns by use of evolutionary information and predicted protein sequence features. PloS One 2010;5(11).

Pham TH, Satou K, Ho TB. Prediction and analysis of β-turns in proteins by support vector machine. Genome Inform 2003;14: 196-205.

Pham TH, Satou K, Ho TB. Support vector machines for prediction and analysis of beta and gamma-turns in proteins. J Bioinform Comput Biol 2005;3:343-58.

Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167-339.

Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. Adv Protein Chem 1985;37:100-9.

Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of β-turns in proteins using neural networks. Protein Sci 1999;8:1045-55.

Shi X, Hu X, Li S, Liu X. Prediction of β-turn types in protein by using composite vector. J Theor Biol 2011;286:24-30.

Takano K, Yamagata Y, Yutani K. Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. Biochemistry 2000;39:8655-65.

Tang Z, Li T, Liu R, Xiong W, Sun J, Zhu Y et al. Improving the performance of β-turn prediction using predicted shape strings and a two-layer support vector machine model. BMC Bioinformatics 2011; 12:283.

Wilmot CM, Thornton JM. Analysis and prediction of the different types of β-turns in proteins. J Mol Biol 1988;203:221-32.

Zhang CT, Chou KC. Prediction of β-turns in proteins by 1-4 & 2-3 correlation model. Biopolymers 1997;41:673-702.

Zhang Q, Yoon S, Welsh WJ. Improved method for predicting β-turn using support vector machine. Bioinformatics 2005;21: 2370-4.

Zheng C, Kurgan L. Prediction of beta-turns at over 80 % accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics 2008;9:430.