Digest Journal of Nanomaterials and Biostructures Vol. 4, No. 4, December 2009, p. 775 - 782

BINARY CLASSIFICATION OF UNCHARACTERIZED PROTEINS INTO DNA BINDING/NON-DNA BINDING PROTEINS FROM SEQUENCE DERIVED FEATURES USING ANN

AMIYA KUMAR PATEL[*], SEEMA PATEL[a], PRADEEP KUMAR NAIK[b]

Division of Biotechnology, Majhighariani Institute of Technology and Science (MITS), At- Sriram Vihar, Bhujbala, Po- Kolnara, Rayagada, (Pin – 765017), Orissa, India

[a]*Department of Computer Science, Sundargarh Engineering College, At/Po-Kirei, Distt.-Sundargarh, (Pin- 770073), Orissa, India*

[b]*Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Distt.-Solan, (Pin- 173 215), Himachal Pradesh, India*

The problem for predicting DNA binding and non-DNA binding proteins from protein sequence information is still an open problem in bioinformatics. It is further becoming more important as the number of sequenced information grows exponentially over time. Sequence similarity matrices are a useful approach to provide functional annotation, but its use is sometime limited, prompting the development and use of machine learning methods. We implemented a novel approach for predicting the DNA binding and non-DNA binding proteins from its amino acid sequence using artificial neural network (ANN). The ANN used in this study is a feed-forward neural network with a standard back propagation training algorithm. Using 62 sequence features alone, we have been able to achieve 72.99% correct prediction of proteins into DNA binding/non-DNA binding (in the set of 1000 proteins). For the complete set of 62 parameters using 5 fold cross-validated classification, ANN model revealed a superior model (accuracy = 72.99%, Q_{pred} = 73.952%, sensitivity = 81.53% and specificity = 72.54%).

(Received July 26, 2009; accepted October 8, 2009)

Keywords: DNA binding proteins, binary classification, sequence derived features, Artificial neural network.

1. Introduction

The rapid progress in genome analysis has made available the complete genome sequences of many organisms. Subsequent annotation of the genes enabling their function to be inferred from sequence homology is an important next step in the post-sequence analysis of genome. In that regard, X-ray crystallographic and NMR spectroscopic analyses of DNA binding proteins have provided valuable information about the general features of protein-DNA interactions. However, it is both time-consuming and costly. Particularly, the number of newly found protein sequences is now increasing rapidly. For instance, the number of total sequence entries in SWISS-PROT was only 3939 in 1986; recently, it was expanded to 272212 (increasing by more than 69 times in just two decades!) according to release 53.2 (26 June 2007) of SWISS-PROT (http://www.expasy.org/sprot/relotes/relstat.html). With such a sequence explosion, it has become vitally important to develop an automated and fast method to differentiate the DNA binding protein from the non-DNA binding proteins.

[*] Corresponding author: amiya_gene@yahoo.com

It is generally accepted that protein structure is determined by its amino acid sequence [1] and that the knowledge of protein structures plays an important role in understanding their functions. To understand the roles relating amino acid sequence to three-dimensional protein structure is one of the major goals of contemporary molecular biology. A priori knowledge of protein to bind with DNA has become quite useful from both an experimental and theoretical point of view. A generic approach to this problem consists of transferring the annotation from sequences of known DNA binding proteins to uncharacterized proteins [2]. The transfer mechanism might be subdivided in two steps: (i) to establish the list of known DNA binding proteins with significant sequence similarity to the uncharacterized sequence; (ii) to select the known sequence(s) from which the annotation is transferred [3]. The first step is usually performed with sequence alignment tools such as FASTA [4] or BLAST [5]. When sensitivity is critical, alternative tools such as PSI-BLAST [6] and hidden Markov models [7] can be used. Finding homologous proteins can also be accomplished using alignment independent sequence comparison tools, which have been developed to overcome the limitation arising from the assumption of contiguity between homologous segments [8,9]. However, annotating the uncharacterized protein sequence as DNA binding and non-DNA binding proteins requires highly automated computational methods linking experimental data. These methods must be able to discriminate the distinct catalytic function encapsulated in the protein's structure or in its primary sequences. To this end, the machine learning methods (MLMs) seem to be best suited for the task. MLMs also have a certain degree of flexibility regarding data inputs, allowing them to expand progressively to meet the requirements of rapidly accumulating mountain of data generated from genomics research. The most often used methods of MLMs are support vector machine (SVM), artificial neural network (ANN), hidden Markov model (HMM), decision tree (DT) and so on. Among these, ANNs are particularly attractive due to its ability for pattern recognition [10] to handle large or small datasets, large input spaces [11] and its greater accuracy compared to simple BLAST or HMM methods [12,13]. Currently, there is no reliable systematic way for recognizing DNA binding proteins. For example, Luscombe and Thronton [14] analyzed amino acid conservation and the effects of mutations on the binding capacity within protein-DNA complexes. Pabo and Nekludova [15] developed geometrical models for characterizing side chain-base interactions and in related studies. Nadassy et al. [16] analyzed the importance of the interface surface area between the protein and DNA for protein-DNA recognition.

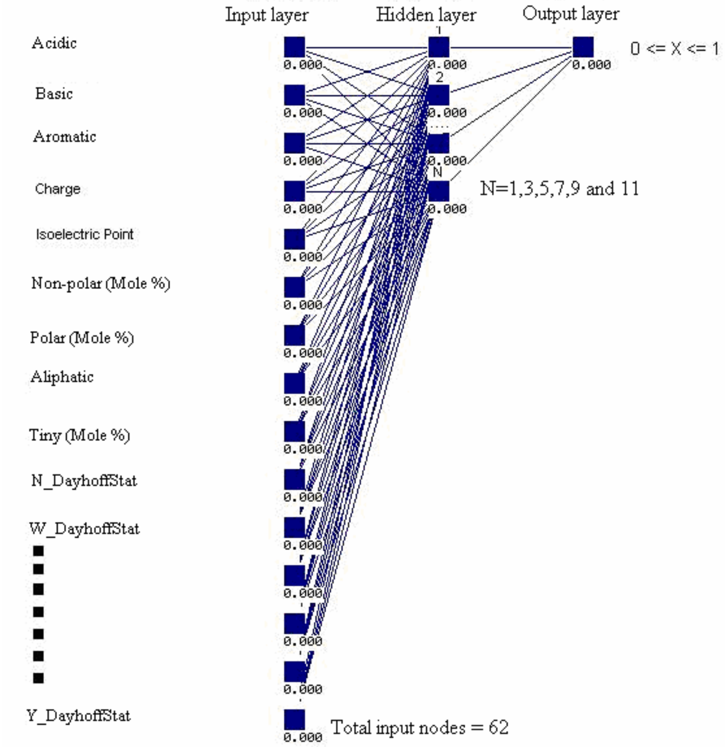

Fig. 1. Configuration of artificial neural network (ANN) used to develop binary primary sequence descriptor model.

The family of DNA binding proteins is one of the most populated and studies amongst the various genomes of bacteria, archea and eukaryotes. Most of the proteins, such as the eukaryotic and prokaryotic transcription factors contain independently folded units (domains) in order to accomplish their recognition with the contours of DNA. It is now clear that the majority of these DNA-binding scaffolds, which are in general relatively small, less than 100 amino acid residues, belong to large number of structural families with characteristic sequences and three-dimensional designs or conformations [17]. Determination of three-dimensional structure is the traditional approach to functional classification of proteins. However, as structure determination is still another problem for itself, the need for a faster method of classification is obvious [18].

Strategically, we have develop a neural network, fully automated computational method capable of recognizing and classifying uncharacterized proteins as DNA binding or non-DNA binding.

2. Materials and methods

2.1 Training data

To discriminate between the DNA binding and non-DNA binding proteins, a set of 1000 proteins consisting of 500 non-redundant DNA binding proteins and the same number of non-redundant non-DNA binding proteins were used for training and testing. The DNA binding proteins dataset used in this study was obtained from PDB database (http://www.rcsb.org). It consisted of almost equal number for each of four major classes of DNA binding proteins (125 zinc finger, 125 leucine zipper, 125 helix-turn-helix, 125 homeo box). The pairwise sequence identities in the datasets was less than 70% for both DNA binding and non-DNA binding protein classes.

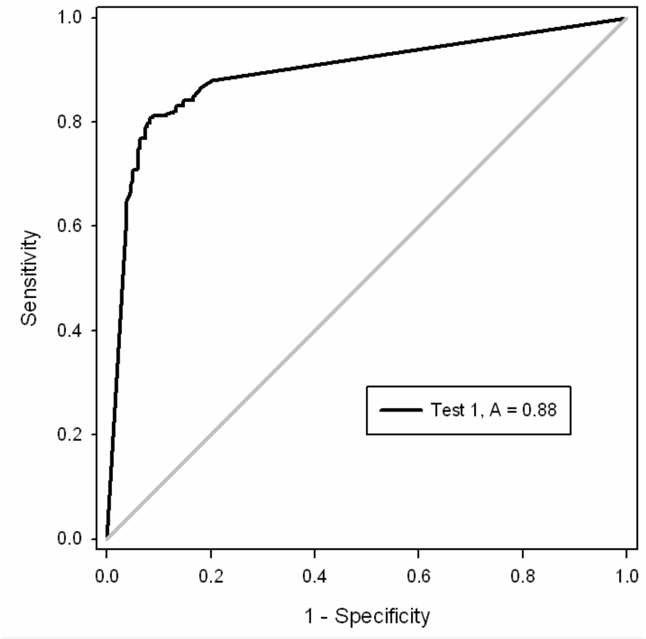

Fig. 2. ROC curves for binary ANN network systems.

2.2 Sequence derived parameters calculation and selection

A set of 62 parameters (Table 1) for each protein sequence were calculated using PEPSTAT (EMBOSS suite) ftp://emboss.open-bio.org/pub/EMBOSS [19] and used as input to ANN. The values of these 62 parameters independently calculated for DNA binding and non-DNA binding showed clear distinction between the two classes (Table 1). The normalized values were used to generate ANN models for binary prediction.

Table 1. 62 ‘PEPSTAT (EMBOSS)’ primary sequence descriptors used in the study.

Sequence derived parameters	**DNA binding**		**Non-DNA binding**		**Sequence derived parameters**	**DNA binding**		**Non-DNA binding**	
	Max	**Min**	**Max**	**Min**		**Max**	**Min**	**Max**	**Min**
Charge	0.207588	0.00182	0.20947	0.00419	N_mole%	0.7186	0.1200	0.9.91	0.2300
Isoelectric point	0.11811	0.09159	0.1209	0.09186	N_DayhoffStat	0.1671	0.0987	0.2114	0.1078
A_mole%	0.104656	0.0427	0.1288	0.03857	P_mole%	0.9572	0.3450	3.6556	0.5680
A_DayhoffStat	0.29032	0.019	0.33257	0.027	P_DayhoffStat	0.1841	0.0089	0.703	0.02908
B_Mole%	0.275	0.024	0.376	0.036	Q_mole%	0.585	0.0871	1.5106	0.1098
B_DayhoffStat	0.928	0.494	0.979	0.41	Q_DayhoffStat	0.15	0.0098	0.3873	0.0129
C_Mole%	0.18828	0.02881	0.21186	0.03	R_mole%	1.0682	0.0088	2.1256	0.0187
C_DayhoffStat	0.2189	0.0335	0.2464	0.045	R_DayhoffStat	0.218	0.02389	0.434	0.0452
D_Mole%	0.1989	0.0017	0.0902	0.0011	S_mole%	0.9035	0.1796	0.2034	0.0012
D_DayhoffStat	0.0292	0.001	0.0109	0.0009	R_DayhoffStat	0.129	0.0257	0.3148	0.0389
E_Mole%	1	0.00659	2.0339	0.0089	T_mole%	1.0497	0.3091	1.4352	0.1203
E_DayhoffStat	0.3448	0.02154	0.7013	0.0154	T_DayhoffStat	0.1721	0.0507	0.2353	0.0092
F_Mole%	0.8147	0.0154	1.206	0.0015	V_mole%	0.15	0.04484	0.17647	0.0289
F_DayhoffStat	0.1481	0.0152	0.2193	0.0652	V_DayhoffStat	0.2273	0.0679	0.2674	0.0546
G_Mole%	1.018	0.0147	1.8615	0.0254	W_mole%	0.4598	0.00245	0.4839	0.0254
G_DayhoffStat	0.1697	0.0215	0.3102	0.0145	W_DayhoffStat	0.3537	0.0021	0.3722	0.0215
H_Mole%	0.9195	0.1277	1.0044	0.0596	X_mole%	0.4562	0.025	0.3262	0.0254
H_DayhoffStat	0.2554	0.0355	0.279	0.0101	X_DayhoffStat	0.5263	0.0562	0.3215	0.025
I_Mole%	0.25	0.00769	0.36923	0.00530	Y_mole%	0.6135	0.0159	2.4615	0.0521
I_DayhoffStat	0.2976	0.0092	0.4396	0.006	Y_DayhoffStat	0.1804	0.0154	0.724	0.00987
K_Mole%	0.6513	0.00894	1.0271	0.021	Z_mole%	0.2222	0.0089	0.3262	0.0154
K_DayhoffStat	0.3257	0.0456	0.5136	0.0598	Z_DayhoffStat	0.894	0.1256	0.265	0.03652
L_Mole%	1	0.2077	1.0377	0.0089	Tiny Mole%	0.6	0.15569	0.6389	0.16239
L_DayhoffStat	0.2222	0.0462	0.2306	0.0564	Small Mole%	0.75	0.4012	0.77119	0.32479
M_Mole%	1.018	0.0591	2.0455	0.00115	Aliphatic Mole%	0.31481	0.14808	0.32903	0.02542
M_DayhoffStat	0.1542	0.00213	0.3099	0.0002	Acidic	0.1159	0.02569	0.13489	0.03654
Charged Mole %	0.19444	0.03139	0.19101	0.0321	Basic	0.1869	0.04521	0.12365	0.02564
Basic mole%	0.2628	0.0424	0.2581	0.0021	Charged	0.30125	0.00586	0.25634	0.01245
Acidic Mole%	0.5169	0.0456	1.2346	0.0268	Aliphatic	0.25692	0.04521	0.1667	0.00698
Aromatic Mole%	0.24521	0.04918	0.29231	0.08541	Polar Mole%	0.54479	0.15	0.68182	0.13846
Non-polar Mole%	0.85	0.45521	0.86154	0.31818	Aromatic	0.10256	0.04956	0.1558	0.07852

2.3 Fivefold cross-validation

A limited fivefold cross-validation was used to test the predictability of ANN model. Here the dataset was randomly divided into five subsets, each containing equal number of protein sequences. Each set was a balanced set that consisted 50% of DNA binding and 50% of non-DNA binding proteins. The dataset was divided into training and testing sets. The training set consisted of five subsets. The network was validated for minimum error on testing set to calculate the performance measure for each fold of validation. This process was repeated five times to test for each subset. The final prediction result was averaged over five testing sets.

2.4 ANN model for prediction of DNA binding/non-DNA binding proteins using sequence derived features

In this study, we had used standard back-propagation ANN configuration consisting of 62 inputs and 1 output node in order to discriminate between DNA binding and non-DNA binding proteins from the testing sets (Figure 1). For each sequence in the training and testing sets, we had transformed 62 network input parameters into the normalized values varying from 0 to 1. Similarly, the output parameters from the ANN were normalized to [0:1] range. The numbers of nodes in the hidden layer were varied from 1 to 11 in order to find out the optimal network that allowed most accurate separation of DNA binding and non-DNA binding proteins in the testing sets (Table 2). During the learning phase, a value of 1 was assigned for the DNA binding sequence and 0 for non-DNA binding sequence. For each configuration of the ANN 119 independent training runs were performed to evaluate the average predictive power of the network. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for DNA binding and non-DNA binding proteins respectively. Thus, a protein sequence from the testing set was considered correctly predicted by the ANN only when its output values ranged from 0.9 to 1.0. For each non-DNA binding protein of the testing set, the corrected prediction was assumed if the corresponding ANN output lies in between 0 to 0.1. Thus, all network output values ranging from 0.2 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined).

Table 2. Parameters of specificity, sensitivity, accuracy and positive predictive values for prediction of DNA binding and non-DNA binding proteins by the artificial neural network with the varying number of hidden nodes.

Hidden Nodes	Accuracy	Specificity	Sensitivity	Q(Pred)
1	0.5869	0.6523	0.7423	65.23
3	0.6213	0.6452	0.5013	72.13
5	0.5522	0.5864	0.5123	55.23
7	0.6976	0.6878	0.7535	68.32
9	0.6435	0.6020	0.7632	65.18
11	0.6235	0.6425	0.7123	69.25

2.5 Performance measures

The prediction results of ANN model developed in the study were evaluated using the following statistical measures.

1. Accuracy of the methods: The accuracy of the prediction for neural network models were calculated as follows:

$$Q_{ACC} = \frac{P+N}{T}, \text{ where } T = (P + N + O + U)$$

Where *P* and *N* refer to correctly predicted DNA binding and non-DNA binding, and *O* and *U* refer to over and under predictions respectively.

2. Sensitivity (Q_{sens}) and specificity (Q_{spec}) of the prediction methods is defined as:

$$Q_{sens} = \frac{P}{P+U}$$

$$Q_{spec} = \frac{N}{N+O}$$

3. Q_{pred} (Probability of correct prediction) is defined as:

$$Q_{pred} = \frac{P}{P+O} \times 100$$

3. Results

3.1 Predictability of DNA binding proteins with sequence derived features

The ANN model (62-7-1) was trained with the sequence derived features (62 parameters) calculated using PEPSTAT. Applying a fivefold cross-validation test using five datasets, we found that the network was reached an overall accuracy of 72.99 ± 6.86%. The prediction results were presented in Table 3. The other performance measures were: Q_{pred} = 73.952 ± 13.123%, sensitivity = 81.53 ± 6.73% and specificity = 72.54 ± 6.39%. The predicted output for non-DNA binding proteins was in the range of 0.0 to 0.1 and for DNA binding proteins, it was 0.9 to 1.0 (Table 3). This illustrated that 0.1 and 0.9 cut-offs values provided adequate separation of two bioactive classes using ANN.

Table 3. Results of DNA binding/non-DNA binding prediction methods, using fivefold cross-validation.

Fivefold cross validation	Accuracy	Specificity	Sensitivity	Q(Pred)	Prediction range (DNA binding)	Prediction range (Non-DNA binding)
C1	0.8.20	0.8632	0.7271	85.12	0.6726 – 1.00	0.00 – 0.5240
C2	0.7430	0.7791	0.8580	70.61	0.5079 – 1.00	0.00 – 0.5658
C3	0.7002	0.6024	0.8001	71.61	0.4257 – 1.00	0.00 – 0.5386
C4	0.7140	0.6567	0.8901	62.28	0.3592 – 1.00	0.00 – 0.6486
C5	0.6906	0.7259	0.8015	80.14	0.4748 – 1.00	0.00 – 0.5836
Mean	0.7299 ± 0.0686	0.7254 ± 0.0639	0.8153 ± 0.0673	73.952 ± 13.123		

3.2 Evaluation of prediction accuracy

From a practical point of view, the most important aspect of a prediction method is its ability to make correct predictions. As prediction methods are never perfect, one always faces the dilemma of choosing between making few false-positive predictions and having a high sensitivity, that is correctly identifying as many positive examples as possible. This tradeoff can be visualized as what is known as the receiver output characteristic (ROC) curve, in which the sensitivity is plotted as a function of 1-specificity by varying the score threshold used for making positive predictions (Figure 2). The performance of the network was evaluated by calculating the area under the ROC curve. The area under the curve was 0.88; revealing a better discrimination of network system.

4. Discussion

The functional properties of uncharacterized protein sequences are usually determined either by biochemical analysis of eukaryotic and prokaryotic genomes or by microarray analysis. These experimental methods are both time-consuming and costly. With the explosion of protein entries in databanks, we are challenged to develop an automated method to quickly and accurately determine the functional attribute for a newly found protein sequence: is it a DNA binding or a non-DNA binding protein? If it is, to which subfamily class does it belongs to? The answers to these questions are important because they may help deduce the mechanism and specificity of the query protein, providing clues to the relevant biological function. Although it is an extremely complicated problem and might involve the knowledge of three-dimensional structure as well as many other physicochemical factors, some quite encouraging results were obtained by a computational method established on the basis of amino acid composition alone [20]. Since the amino acid composition of a protein does not contain any of its sequence-order information, a logical step to further improve the method is to incorporate the sequence-order information into the predictor. To realize this, the most straightforward way is to represent the sample of a protein by its entire sequence, the so-called sequential form.

The results demonstrated that the developed ANN-based model for binary prediction of DNA binding/non-DNA binding proteins is adequate and can be considered an effective tool for '*in silico'* screening. The results also demonstrated that the sequence derived parameters readily accessible from the protein sequences only, can produce a variety of useful information to be used '*in silico*'; clearly revealed an adequate and good predictive power of the developed ANN model. There is strong evidence, that the introduced sequence features do adequately reflect the structural properties of proteins. The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of DNA binding proteins. This observation is not surprising considering that the calculated parameters should cover a very broad range of proprieties of bound atoms and molecules related to their size, polarizability, electronegativity, compactness, mutual inductive and steric influence and distribution of electronic density, etc.

The ANN model with 62 input-nodes, 7 hidden-nodes and 1 output nodes was able to classify the uncharacterized protein sequence into DNA binding and non-DNA binding protein with an accuracy of 72.99%. Presumably, accuracy of this approach operating by the sequence derived features can be improved even further by expanding the parameters or by applying more powerful classification techniques such as Support Vector Machines or Bayesian Neural Networks. Use of merely statistical techniques in conjunction with the sequence parameters would also be beneficial, as they will allow interpreting individual parameter contributions into "DNA binding likeness".

5. Conclusion

The results of the present work demonstrate that the sequence derived features with binary-ANN classification system (62-7-1) appear to be a very fast protein classification mechanism providing good results, comparable to some of the current efforts in the literature. The developed ANN-based model for classification proteins into DNA binding and non-DNA binding can be used as a powerful tool for filtering out DNA binding proteins from the proteome databases.

References

[1] C. B. Anfinsen, Science **181**, 223 (1973).

[2] M. A. Andrade, C. Sander, Curr. Opin. Biotechnol. **8,** 675 (1997).

[3] P. D. Karp, Bioinformatics **14,** 753 (1998).

[4] W. R. Pearson, Methods Enzymol. **183,** 63-98 (1990).

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, J. Mol Biol. **215,** 403 (1990).

[6] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Nucleic Acids Res. **25,** 3389 (1997).

[7] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, D. Haussler, J. Mol. Biol. **235,** 1501 (1994).

[8] S. Vinga, J. Almeida, Bioinformatics **19,** 513 (2003).

[9] J. K. Vries, R. Munshi, D. Tobi, J. Klein-Seetharaman, P. V. Benos, I. Bahar, Appl. Bioinformatics **3,** 137 (2004).

[10] K. Harpreet, G. P. S. Raghava, Protein Science **12,** 923 (2003).

[11] A. Narayanan, E. C. Keedwell, B. Olsson, Appl. Bioinformatics **1(4),** 191 (2002).

[12] M. Bhasin, G. P. S. Raghava, Nucleic Acids Research **32,** W383-W389. [doi:10.1093/nar/gkh001] (2004a).

[13] M. Bhasin, G. P. S. Raghava, Nucleic Acids Research **32,** W414-W419 (2004b).

[14] N. M. Luscombe, J. M. Thornton, J. Mol. Biol. **320,** 991-1009 (2002).

[15] C. O. Pabo, L. Nekludova, J. Mol. Biol. **301,** 597–624 (2000).

[16] K. Nadassy, S. J. Wodak, J. Janin, Biochemistry **38,** 1999–2017 (1999).

[17] C. Branden, J. Tooze, Introduction to Protein Structure, Garland Publishing Co., New York (1991).

[18] E. N. Baker, V. L. Arcus, J. S. Lott, Appl. Bioinformatics **2,** s3-s10 (2003).

[19] P. Rice, I. Longden, A. Bleasby, Trends in Genetics **16 (6),** 276—277 (2000).

[20] K. C. Chou, D. W. Elrod, J. Proteome Res. **2,** 183–190 (2003).