ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Neural Network Model for Prediction of PPI using Domain Frequency & Association Score Base Classification of Protein Pairs

Garima Srivastava* and Navita Srivastava
Dept. of Computer Science,
Awadhesh Pratap Singh University, Rewa, M.P., India
garima.gs17@gmail.com*, navita.srivastava@gmail.com

Gulshan Wadhwa
Dept. of Biotechnology, Ministry of Science & Technology
Govt. of India, New Delhi, India.
gulshan.dbt@nic.in

Abstract: Developing In Silico Computational Techniques to predict Protein Protein Interactions (PPIs) is one of the challenging research area for computational biologists. Experimental techniques for interaction prediction, lack accuracy and are error prone. Observing the limitations and low accuracy of existing methods, this work focuses on improvement in computational efficiency with increased performance for domain base prediction of PPIs. In the present paper, a novel approach of Domain Frequency Count (DFC) method with association score base classification, using feed forward back-propagation neural network, has been proposed for prediction of interacting protein pairs based on their domain characteristic features data. Results obtained are quite encouraging. When compared with the existing MLE, DT, CL NN, and RDF techniques [8][10] on similar datasets, it is observed that accuracy and sensitivity are increased by 8.91% and 5.70%, respectively.

Keywords: Neural Network, Protein-Protein Interaction, Domain-Domain Interaction, Domain Frequency Count, Association Score.

I. INTRODUCTION

Determining the protein-protein interaction (PPI) networks is a daunting task and has been the subject of extensive research in recent past. Despite the development of reasonably successful methods, serious technical difficulties still exist. Arguably, the small intersection between the two major experimental approaches, the Yeast Two-Hybrid systems (Y2H) and Coimuno Precipitation, best reflects these difficulties [1]. Moreover, the number of possible protein interactions within one cell is enormous, which is a potentially limiting factor for experimental analyses[25]. Various computational techniques have been developed so far, to observe or predict the PPI networks in biological systems[20]. A more recent and accurate method is the domain based prediction of protein interaction network. It is widely believed that protein interactions are, basically, caused by their domains[23]. The preliminary results of domain based methods have demonstrated their feasibility [5-10]. Hence motivated with the grand challenge of computational complexity of protein interaction prediction problem, its high significance in biomedical science[26], low accuracy of existing methods, and limitations of general assumptions, this work presents a novel approach to model the prediction framework of PPIs based on machine learning architecture of neural network with hybrid concepts of biological domain.

II. BACKGROUND

The domain based prediction of protein interaction network is performed in the context of domain-domain interactions at the primary sequence level. The seminal work of Sprinzak gave the Association Method [2], and was successful in predicting several interesting domain interactions. The method was tested with several machine learning applications, support vector machines [4], probabilistic learning [3]. Followed by Deng et. al. [5], a probabilistic model (MLE method) linking domain interaction with protein interaction was proposed. Ng et. al. [6] introduced an integrated data source approach to increase accuracy. Riley et. al. [7] was next to improve upon MLE method. They introduced a domain pair exclusion analysis (DPEA) approach to predict domain interactions. The random decision forests was also proposed to improve upon accuracy [8]. Guimarães et al. [9] proposed a model-free-based method employing Maximum Parsimony Explanation (PE) based optimization. Recently, Xue wen Chen and Mei Liu [10] presented two machine learning methods, decision trees and neural network for domain feature vector based PPI prediction. The authors compared their result with the existing MLE method and claimed accuracy of 73% (approx.) as compared to 60% of MLE. Though their results were promising, the constraints imposed on feature vector refrained several other aspects of the problem. The proposed method in this paper, improves upon the previous work [2-10] on domain base prediction of protein-protein interaction (PPI).

A. *Protein Domain:*

Protein domains are defined as evolutionary, functional, and structural units of proteins and different combinations of domains result in diverse range of proteins[26]. It is believed that once a set of domains is formed with sufficient functions to support the basic life form, it would be much easier and faster for the genome to produce various new proteins by duplication, divergence, and recombination [11]. A modest increase in the number of domains in interacting protein partners may directly translate into numerous new interactions. Moreover, it is shown that domains and their interactions determine the functions of proteins [12][21]. Therefore, it is essential to understand protein interactions at the domain level.

B. *Domain-Domain Interaction:*

The analysis of various types of domain-domain interactions is an extremely important task for protein interaction and function annotation. There are four basic types of domain-domain interactions (DDI) that can explain protein-protein interactions (PPI). These are illustrated in Figure 1.

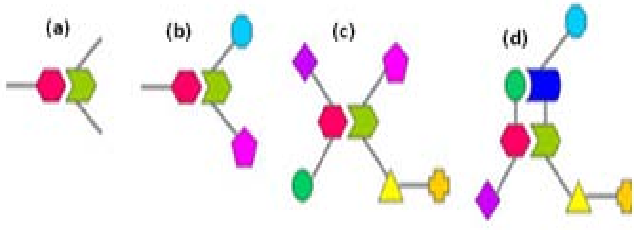

Figure 1 : Types of Domain Domain Interaction. Different shapes represent different domains

a. Singlet-to-Singlet: Interactions between two single-domain proteins (Figure 1a).
b. Singlet-to-Multiplet: Interaction between single-domain and multi-domain proteins (Figure 1b).
c. Multiplet-to-Multiplet via Single Binding: Interaction between two multi-domain proteins by a single binding between one of many domains from each protein (Figure 1c).
d. Multiplet-to-Multiplet via Multiple Bindings: Interaction between two multi-domain proteins by multiple bindings between two or more domains from each protein (Figure 1d).

This work includes multiplet DDIs generated by combinations of input neurons mapping in the neural network model. With this, the model is able to incorporate every domain information in the protein.

III. MATERIALS AND METHODS

A. Implementation Platform:

Hardware: Intel Pentium µp, 2.53 GHz ; RAM – 4 GB ; Hard Disk – 320 GB; Software: Operating System – Windows XP, Language – MATLAB (R2009b)

B. Feature Vector Construction:

The representation of an appropriate feature space that describes the training data is essential for any supervised machine learning system. The likelihood of two proteins to interact with each other is associated with their structural domain composition [13][24]. The PPI prediction problem can be formulated as a two-class classification problem, where, each pair of belongs to either the 'interaction' class (i.e. two proteins interact with each other) or 'non-interaction' class (i.e. two proteins do not interact). Every protein pair is characterized by their respective domains. Thus, each pair is represented by a vector of features where each feature corresponds to a domain.

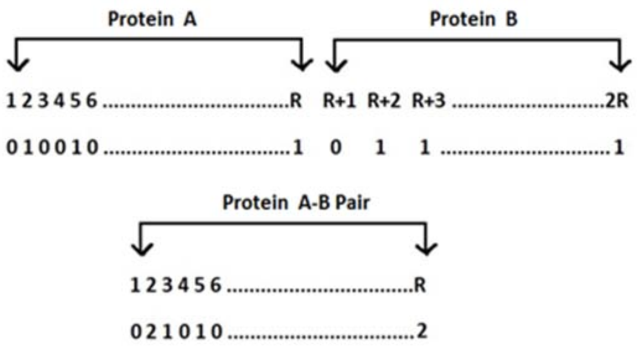

Figure 2 : Feature Vector for interacting protein pair A-B by DFC Method

The feature vector for each protein M is formulated as: P_M= [D_1 ,D_2 ,….. D_R] , where, each domain is labeled with a number between 1 and R. Each protein is represented by a vector of R binary numbers and each binary number is associated with the Rth domain. For example, if a protein has a domain with label 5, then the 5th number of the feature vector is 1, otherwise 0 (Figure 2- Protein A and Protein B). P_M= [0,0,0,0,1,0,0, … …0,1, 1], Mth protein has (present) domain label 5th,7th, (r-1)th, rth ; rest of the domains are absent.

a. Domain Frequency Count:

The construction of feature vector for each protein pair X, is done by a *novel Domain Frequency Count (DFC)* method, concatenating independent feature vectors of proteins. If a domain indexed as 2 is found in both the proteins (A&B) in an interacting pair, the 2nd entry is replaced by a frequency count 2 (Figure 2-AB Pair). If a domain is found either in protein A or protein B (domain1 or domain2), that entry is replaced by 1, otherwise 0. Hence attributes can assume values={0,1,2}.

C. Neural Network Architecture:

The unique characteristic of the domain base data feature vector, prompts a challenge for machine learning algorithms. This is because the size of the feature space is equivalent to the total number of unique domains which is extremely large, in the range of thousands. Hence, being able to efficiently deal with huge matrices, a multilayer feedforward backpropagation neural network architecture (FB NN) is utilized. It predicts the protein-protein interactions using the protein functional domain information. It is tested on model organism Saccharomyces Cerevisiae (Yeast PPI) [16][17].

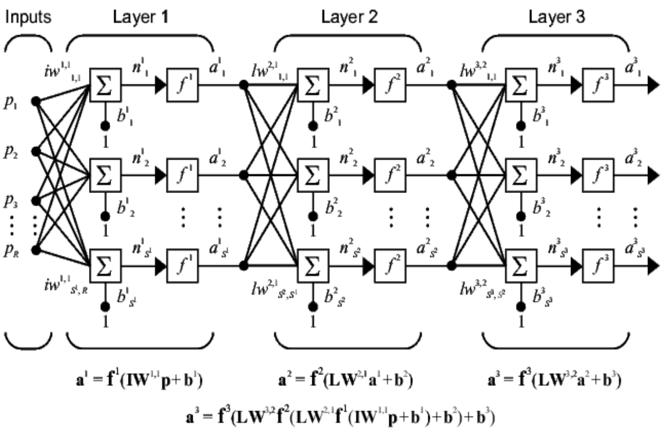

Figure 3 : A Multi-Layer Artificial Neural Network (ANN) for domain base PPI prediction.

In Figure 3, the input to the network are the **R** features of the domain feature vector. Each layer has a weight matrix **W**, a bias vector **b**, and an output vector **a**. The network shown above has $\mathbf{R^1}$ inputs, $\mathbf{S^1}$ neurons in the first layer, $\mathbf{S^2}$ neurons in the second layer, etc. A constant input 1 is fed to the bias for each neuron. Thus layer 2 can be analyzed as a one-layer network with $\mathbf{S^1}$ inputs, $\mathbf{S^2}$ neurons, and an $\mathbf{S^2 x S}^1$ weight matrix $\mathbf{W^2}$. The backpropagation learning refers to the manner in which the gradient is computed for the nonlinear multilayer network. There are a number of variations on the basic algorithm and parameters that govern the learning and optimization of weights in order to adapt for a correct prediction. A comparison is shown in the results in *section VI-C*.

IV. DATA SOURCES

PPI data for the yeast organism is a collection from the DIP, Salwinski et al., [19], Deng et al.[5] (Uetz et al. [15] and Ito et al. [16]), Schwikowski et al.[18], and Xenarios et al. [14]. This dataset given by Chen and Liu [10], has 9834 protein interaction pairs among 3713 proteins. It is separated evenly (4917 pairs each) into training and testing datasets. 8000 negative samples are randomly generated, separated into two halves. Both final training and testing datasets contain 8917 samples, 4917 positive and 4000 negative samples [8]. The protein domain information is same as reported in [8][10] with 4293 Pfam domains[17] defined by the set of proteins.

V. A NOVEL PREDICTION MODEL

The basic constituents of the proposed novel model is illustrated in the following points -

A. Simulation Technique:

Feed forward Back propagation Multilayer Neural Network FB NN can assume several combinations of domains. As such it overcomes the conventional limitation of single domain pair consideration. Here each domain, associated with an input neuron, may contribute to output of neural network depending on neuronal network weights

B. Input Domain Feature Vector:

String of discrete real values corresponding to domain frequency count in protein pair – (i) Every pair exist in two forms: Pi ⇌ Pj, Pi→Pj, ⟹Pj→Pi where, it considers both direct and reverse interactions, which improves upon the training of neural network, avoiding over-training. The previous methods over-trained the network, where lack of information leads to low accuracy and high rate of false predictions. (ii) Feature vector considers the frequency of domain appearance in proteins. This accounts for natural tendency of active domains to be present in higher frequency ratio so as to cause interactions. This also reduces input feature vector matrix to half of the storage requirement.

C. Classification Rule:

The classification of protein pair as interacting or non interacting is based on association score cutoff threshold – The classification cutoff threshold is calculated as the score of association between domains of interacting proteins. As such false predictions are avoided to a great extent. The association method is based on over-represented domain pairs which occur more frequently in interacting protein pairs. It assumes that random association of frequent domain pairs are more likely to interact with each other [2].

a. Calculation of Association Score:

Let I_{mn} be the observed frequency of interacting protein pairs with one protein containing domain D_m and the other protein containing domain D_n. Let I_m and I_n be the frequencies of proteins containing domains D_m and D_n in all proteins, respectively. Then, the likelihood ratio defined as $A_{mn} = I_{mn} / (I_m I_n)$ is used to measure strength of association between domains D_m and D_n [22].

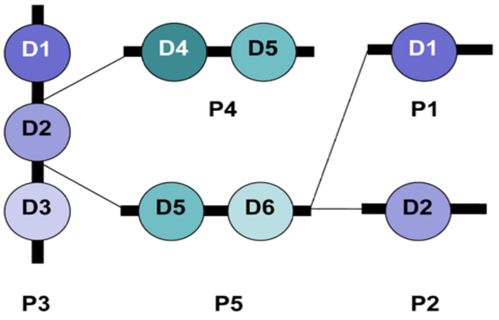

Figure 4 : An example of five proteins and their domains interacting with each other.

An artificial example of five proteins (P_1,P_2, P_3,P_4,P_5) and their six interacting domains (D_1,D_2,D_3,D_4,D_5,D_6) are shown in Figure 4. The edges between the nodes indicate interactions between the proteins. The adjacency matrix between the proteins and domains is given in Table 1. It shows the presence and frequency of six domains. The upper triangle of Table 2 shows the frequency of interacting protein pairs with one protein containing one domain and the other protein containing other domain. The lower triangle shows the calculated association scores A_{mn} for domain pairs[22].

Table 1: Adjacency matrix of interacting protein pairs and their domains.

Proteins	Domains					
	D1	D2	D3	D4	D5	D6
P1	1	0	0	0	0	0
P2	0	1	0	0	0	0
P3	1	1	1	0	0	0
P4	0	0	0	1	1	0
P5	0	0	0	0	1	1
I	2	2	1	1	2	1

The association scores for domain pairs in an interacting protein pair, signifies the bond strength and likelihood of interaction. Since domains are assumed to be main drivers for interaction[2][23], the scores of all domain pairs present in a protein pair is calculated and normalized to fall within a small range (between 0 to 1). This score is used in neural network output layer as threshold, to classify the protein pair as interacting, iff the output is greater than the threshold (1-score value) , and non-interacting, iff output is less than or equal to the threshold. Higher score gives a low cutoff, thus accepting pair with a greater chance to be a true positive and a low score results in high cutoff, where weak association penalizes protein pair to be classified as interacting.

Table 2 : Association Scores- The upper part above diagonal, gives the values of I_{mn}(frequency of interacting protein pairs) , $m \neq n$, and lower part below diagonal, gives the value of A_{mn}(association score)

Domain	Domains					
	D1	D2	D3	D4	D5	D6
D1	-	0	0	1	3	1
D2	0	-	0	1	3	1
D3	0	0	-	1	2	1
D4	0.5	0.5	1.0	-	0	0
D5	0.75	0.75	1.0	0	-	0
D6	0.5	0.5	1.0	0	0	-

In this way the cutoff value to decide upon the class of protein pair, directly depends on the probability of association of proteins.

D. Integrated Data Source:

Model Organism Saccharomyces Cerevisiae (Yeast PPI)- Because a large quantity of publicly available databases for Saccharomyces cerevisiae (yeast), it is arguably the best model organism for benchmark testing. The integrated dataset of Chen & Liu[10] combined protein interactions from multiple databases and experimental techniques. This generalizes input data over multiple sources making a strong train and test dataset.

VI. RESULTS AND DISCUSSION

A. Predicting PPI-Training and Testing:

In order to predict protein-protein interactions, the model needs to be trained first with the training data set. In the presented neural network structure, 8586 input nodes are fed to the hidden layer. Values of the hidden neurons are then fed to the output layer with 1 output node. The network is trained for different set of parameters. Testing is done with simulation of the network with the test dataset. The predicted outputs are compared with the target outputs to determine the accuracy of the prediction.

B. Evaluation Criteria:

After building FB NN PPI prediction system, it is important to estimate how accurately the model will perform in practice.

The seven common measures of performance are –

a. Confusion Matrix TP, FP, TN, FN
b. Accuracy = (TP + TN) / (TP + TN +FP + FN)
c. Specificity = TN / (TN + FP)
d. Sensitivity (Recall) = TP / (TP + FN)
e. Precision = TP / (TP+FP)
f. Fmeasure=(2*precision*recall)/(precision+recall)
g. Mathew Corelation Coeff.= (TP*TN)-(FP*FN) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

C. Outputs and Inference:

The output results for various measures and comparative performance are reported in Table 3. It can be observed that the proposed neural network prediction model outperforms the previous methods of MLE (maximum likelihood estimation), DT(decision tree) , CL NN(neural network [10]), and RDF(random forests) [8][10]. The FP and FN rates are reduced with a greater number of true predictions. The accuracy and sensitivity is increased by 8.91% and 5.70%, respectively as compared to the best previous performance.

The proposed model is capable of utilizing all the possible interactions between domains, where every domains contributes to the prediction of protein-protein interactions with different weights. The experimental results have shown that the novel approach can predict PPIs with higher computational efficiency and increased performance The overall accuracy is reported in Table 3 for the best setting of the network parameters.

Table 3 : Comparison of various techniques with novel hybrid prediction model* on seven basic evaluation metrics. The accuracy reported for the novel method is best one for optimal parameter settings.

Evaluation Measures	Maximum Likelihood Estimation (MLE)	Decision Tree DT	Chen & Liu NN	Random Forest RDF	Feedforward Backpropagation Neural Network –DFC & AS score (Novel Prediction Model)*
True Positives- TP	3850	3899	3813	3923	4203
False Positive- FP	2499	1488	1368	1425	911
True Negative- TN	1501	2512	2632	2575	3089
False Negative-FN	1067	1018	1104	994	714
Specificity- SP	37.53%	62.80%	65.80%	64.38%	77.23%
Sensitivity- SN	78.30%	79.30%	77.55%	79.78%	**85.48%**
Precision	60.64%	72.38%	73.60%	73.35%	82.19%
F Measure	0.6835	0.7568	0.7552	0.7643	0.8380
Correlation Coeff.	0.1738	0.4281	0.4370	0.4480	0.6306
Accuracy	60%	71.89%	72.27%	72.87%	**81.78%**

VII. CONCLUSION

The proposed neural network prediction model is able to represent protein-protein interactions in a more comprehensive way. It is closer to the nature of biological interactions, where domain frequency count (*the DFC Method*) presents the characteristics of over-presented domains and association scoring indicates the bond energy of interactions. Moreover it increases computational efficiency where a highly sparse domain feature matrix is reduced to half of the storage requirement. Neural network learning produces a weight matrix adapted to the train dataset of input domain feature vector. The simulation of network, including improved biological features fed to input layer, produces increased accuracy 81.78% (8.91% gain) and sensitivity 85.48% (5.70% gain) for the unknown test data set. This is in contrast with a lower accuracy of the previous techniques of MLE, DT, CL NN and RDFF for domain based prediction [8][10].

With a comparative validation, the Novel Neural Network Prediction Model proves to be robust even for large datasets and multiple combination of domains giving the association score. Hence, the model can serve as the basis for analytical and statistical computation of protein interaction networks establishing several important concepts of active cell processes and system behaviour as a whole.

VIII. REFERENCES

[1] Sprinzak E, Sattath S, Margalit H, ‘How reliable are experimental protein-protein interaction data?’ J. Mol. Biol. , 327(5):919-923, 2003.

[2] Sprinzak E, Margalit H, ‘Correlated sequence-signatures as markers of protein-protein interaction.’ J Mol. Biol , 311(4):681-692, 2001.

[3] Gomez SM, Rzhetsky A, ‘Towards the prediction of complete protein–protein interaction networks.’ Pac Symp Biocomput. 413-424, 2002.

[4] Bock J R, Gough D A, ‘Predicting protein–protein interactions from primary structure.’ Bioinformatics, 17(5):455-460, 2001.

[5] Deng M, et. al. ‘Inferring domain-domain interactions from protein-protein interactions.’ Genome Res. 12(10):1540-1548, 2002.

[6] Ng S K, Zhang Z, Tan S H, ‘Integrative approach for computationally inferring protein domain interactions.’ Bioinformatics, 19(8):923-929, 2003.

[7] Riley R, Lee C, Sabatti C, Eisenberg D, ‘Inferring protein domain interactions from databases of interacting proteins.’ Genome Biol, 6(10):R89, 2005.

[8] Chen Xue-Wen, Liu Mei, ‘Prediction of protein–protein interactions using random decision forest framework’, Bioinformatics, 21(24):4394–4400, 2005.

[9] Guimaraes K S, Jothi R, Zotenko E, Przytycka T M, ‘Predicting domain-domain interaction using a parsimony approach.’ Genome Biol, 7(11):R104, 2006.

[10] Chen Xue-Wen, Liu Mei, ‘Domain-Based Predictive Models for Protein-Protein Interaction Prediction’, Hindawi Publishing Corp. EURASIP Journal on Applied Signal Processing, article ID 32767:1–8, 2006.

[11] Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A., ‘Evolution of the protein repertoire.’ Science, 300(5626): 1701-1703, 2003.

[12] Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A., ‘Structure, function and evolution of multidomain proteins.’, Curr Opin Struct Biol, 14(2): 208-16, 2004.

[13] Oyama T., Kitano K., Satou K., T.Ito, ‘Extraction of knowledge on protein-protein interaction by association rule discovery’, Bioinformatics, 18(5):705-714, 2002.

[14] Xenarios I., Eisenberg D., ‘Protein Interaction Databases’,Curr. Opinion. Biotech., 12: 334–339, 2001.

[15] Uetz P, Giot L, Cagney G, Mansfield T A, et al., ‘A comprehensive analysis of protein- protein interactions in Saccharomyces cerevisiae.’ Nature, 403(6770):623-627, 2000.

[16] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, ‘A comprehensive two-hybrid analysis to explore the yeast protein interactome.’ Proc Natl Acad Sci USA, 98(8):4569-4574, 2001.

[17] Bateman A, et al., ‘The Pfam Protein Families Database.’ Nucleic Acids Res D138-141, 2004.

[18] Schwikowski B., Uetz P., and Fields S., ‘A network of protein-protein interactions in yeast,’ Nature Biotech. 18(12):1257–1261, 2000.

[19] Salwinski, L. et al., ‘The Database of Interacting Proteins: 2004 update.’ Nucleic Acids Res., 32 (Database issue), D449–D451,2004.

[20] Srivastava G., Srivastava N. , ‘Application of Machine Learning Techniques to study Biological Networks’, The IUP Journal of Information Technology, ICFAI University Press, 7(3):7-20, 2011.

[21] Liu M., Chen X.W., Jothi R., ‘Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks’, Bioinformatics 25: 2492-2499, 2009.

[22] Feng Jianfeng, Fu Wenjiang, Sun Fengzhu, ‘Frontiers in Computational and Systems Biology’, Springer-Verlag London Limited, 160-162, 2010.

[23] Zaki, N.M. , ‘Protein-protein interaction prediction using homology and inter-domain linker region information’, International Conference of Systems Biology and Bioengineering, Imperial College, LNCS, Springer, 635-645, 2009.

[24] Roslan R., Othman R. M., Shah Z. AS., Kasim S., et al. ‘Incorporating multiple genomic features with the utilization of interacting domain patterns to improve the prediction of protein–protein interactions’, Information Sciences, 180:3955–3973, 2010.

[25] Khan A.U., Baig M.H., Wadhwa G., ‘Molecular docking analysis of new generation cephalosporins interactions with recently known SHV-variants’, Bioinformation 5(8):331, 2011.

[26] Shrivastava S., Srivastava G., Srivastava N., ‘Parallel Recombinative Simulated Annealing Technique for PSP using the Subsequence Secondary Structures : A Novel Protein Fold Lock Model’, International Journal of Advanced Research in Computer Science, 2(5): 46- 51, 2011

.