THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 279, No. 22, Issue of May 28, pp. 23262–23266, 2004

Printed in U.S.A.

Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition*[S]

Received for publication, February 22, 2004, and in revised form, March 18, 2004
Published, JBC Papers in Press, March 23, 2004, DOI 10.1074/jbc.M401932200

Manoj Bhasin and Gajendra P. S. Raghava‡

From the Institute of Microbial Technology, Chandigarh 160036, India

Nuclear receptors are key transcription factors that regulate crucial gene networks responsible for cell growth, differentiation, and homeostasis. Nuclear receptors form a superfamily of phylogenetically related proteins and control functions associated with major diseases (*e.g.* diabetes, osteoporosis, and cancer). In this study, a novel method has been developed for classifying the subfamilies of nuclear receptors. The classification was achieved on the basis of amino acid and dipeptide composition from a sequence of receptors using support vector machines. The training and testing was done on a non-redundant data set of 282 proteins obtained from the NucleaRDB data base (1). The performance of all classifiers was evaluated using a 5-fold cross validation test. In the 5-fold cross-validation, the data set was randomly partitioned into five equal sets and evaluated five times on each distinct set while keeping the remaining four sets for training. It was found that different subfamilies of nuclear receptors were quite closely correlated in terms of amino acid composition as well as dipeptide composition. The overall accuracy of amino acid composition-based and dipeptide composition-based classifiers were 82.6 and 97.5%, respectively. Therefore, our results prove that different subfamilies of nuclear receptors are predictable with considerable accuracy using amino acid or dipeptide composition. Furthermore, based on above approach, an online web service, NRpred, was developed, which is available at www.imtech.res.in/raghava/nrpred.

The availability of sequence data for different genomes in recent years has increased the demand for computational tools that can recognize new proteins from this data. The recognition of nuclear receptors is crucial, because many of them are potential drug targets for developing therapeutic strategies for diseases like breast cancer and diabetes (2). Nuclear receptors are one of the most abundant classes of transcriptional regulators, which regulate diverse functions during reproduction, metabolism, and development. Nuclear receptors function as ligand-activated transcriptional factors, providing a direct link between the signaling molecules that control these processes and transcriptional responses (3). The nuclear receptors share a common structural organization. All nuclear receptors consist of six distinct regions or domains as follows: highly variable N-terminal and C-terminal regions (A/B and F domains) that contain one or more transactivation regions; a central, well conserved DNA binding domain (C); a non-conserved hinge region (D) that contains a nuclear localization signal (NLS), and a moderately conserved ligand binding domain (E) (4). The DNA binding domain (C region) of nuclear receptors consists of two zinc fingers, which act as a signature for this superfamily (5). The presence of these zinc fingers facilitates the recognition of nuclear receptors from a genome sequence using simple similarity-based search tools like BLAST and FASTA (6–7). On the other hand, the major limitation of these search tools is that they are not able to recognize the subfamilies of nuclear receptors. The nuclear receptors have been classified and assigned seven subfamilies according to the NucleaRDB database, which include thyroid and estrogen hormone-like receptors (1). However, classification of these subfamilies by using phylogeny-based or BLAST-based tools is difficult due to a scarcity of data for some subfamilies. Thus, there is a crucial need for methods that will enable the automated assignment of nuclear receptor subfamilies.

In this report, we have made an attempt to develop a method for recognizing the subfamilies of nuclear receptors. We were able to design a method for recognizing the four subfamilies of nuclear receptors: (i) thyroid hormone-like (TR, RAR, and ROR); (ii) HNF4-like (HNF4, RXR, TLL, Coup, and USP); (iii) estrogen-like (ER, ERR, GR, MR, PR, and AR); and (iv) Fushi tarazu-F1-like (SFI, FTF, and FTZ-F1). Sequences for the other three subfamilies are not available in significant number (<10). The classification and assignment of nuclear receptors to various subfamilies was done on the basis of amino acid composition and dipeptide composition. Amino acid and dipeptide compositions are simplistic approaches for producing patterns of fixed length from the protein sequences of varying length (8). In the past, amino acid composition has been used to predict the structural class of domains and the subcellular localization of proteins (9–11). The dipeptide composition is also widely used to encapsulate the global information and give a fixed pattern length of 400. In the past, dipeptide composition has been used for predicting the subcellular localization of proteins (11) and for fold recognition (12, 13). In this study, support vector machines (SVMs)[1] were applied to classify nuclear receptors. SVMs are a relatively new type of statistical learning method that have proven to be particularly attractive for biological analysis due to their ability to handle large data sets and avoid overfitting. SVMs have been shown to perform well in multiple areas of biological analysis, including classifi-

* This work was supported by grants from the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology (DBT), Government of India. This report is IMTECH communication number 015/2004. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplementary Tables S1, S2, and S3.

‡ To whom the correspondence should be addressed: Bioinformatics Center, Inst. of Microbial Technology, Sector 39A, Chandigarh 160036, India. Tel.: 91-172-2690557 or 2695225; Fax: 91-172-2690632 or 2690585; E-mail: raghava@imtech.res.in.

[1] The abbreviations used are: SVM, support vector machine; MCC, Matthew's correlation coefficient; RI, reliability index.

TABLE I
The number of sequences belonging to each nuclear receptor subfamily

Nuclear receptor subfamilies	No. of protein sequences
Thyroid hormone-like	114
HNF-4-like	72
Estrogen-like	75
Fusi-tarazu-like	21

cation of G-protein-coupled receptors (GPCRs) (14) and enzyme families (15), analysis of protein functions and types (16–17), and prediction of RNA-binding proteins (18). The overall accuracy of amino acid and dipeptide composition-based classifiers are 82.6 and 97.5% respectively. The Matthew's correlation coefficient (MCC) of the dipeptide composition-based classifier is 0.96, which is significantly higher in comparison to that of the amino acid composition-based classifier. MCC is a better parameter for evaluating the performance of a method, as it accounts for both over- and under-predictions. The performance of both classifiers has been estimated through a 5-fold cross-validation test. It was found that various subfamilies of nuclear receptors are correlated with amino acid or dipeptide composition, implying that the subfamilies of nuclear receptors are predictable to a highly accurate extent if good training data can be established. The method is available via the World Wide Web at www.imtech.res.in/raghava/nrpred.

MATERIALS AND METHODS

Data Set—The data for four subfamilies of nuclear receptors was obtained from the NucleaRDB data base available at www.receptors.org/NR/ (1). All entries not marked as fragments were extracted from the data base by the text-parsing method. The initial data set had 577 sequences belonging to four subfamilies of nuclear receptors. Redundancy was reduced so that no sequence had ≥90% sequence identity with any other sequence in the data set, using PROSET software (19). The final data set contains 282 sequences belonging to different subfamilies of nuclear receptors as shown in Table I.

Design and Implementation of the Prediction System—The prediction of subfamilies of nuclear receptors is a multi-class classification problem. In this case, the number of subfamilies of nuclear receptors was four. To handle this multi-class situation, we designed a series of binary SVMs. For N class classification, N SVMs were constructed. The ith SVM was trained with all samples of the ith subfamily being labeled as positive, and the samples of all other subfamilies being labeled as negative. The SVMs trained in this way were referred to as 1-v-r SVMs (9). In this classification approach, each of the unknown proteins will achieve four scores. An unknown protein will be classified into the subfamily that corresponds to the 1-v-r SVM with the highest output score.

Support Vector Machine—The SVMs were implemented using freely downloadable software, SVM_light, written by T. Joachims (20). This software enables the users to define a number of parameters as well as the choice of inbuilt kernel, such as a radial basis function (RBF) or a polynomial kernel (of given degree). In this study, all of the parameters of a kernel were kept constant, except for the regulatory parameter C. The experimentation was conducted by using various types of kernels such as polynomial and radial base function. The SVMs require a fixed number of inputs for training, thus necessitating a strategy for encapsulating the global information about proteins of variable length in a fixed length format. The fixed length format was obtained from protein sequences of variable length using amino acid and dipeptide composition.

Amino Acid Composition—Protein information can be encapsulated in a vector of 20 dimensions, using amino acid composition of the protein. In the past, this approach has been used for predicting the subcellular localization of proteins (9, 21). The amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 natural amino acids were calculated by using Equation 1,

$$\text{Fraction of aai} = \frac{\text{total number of amino acids of type i}}{\text{total number of amino acids in protein}} \quad \text{(Eq. 1)}$$

where i is an specific type of amino acid (aa).

Dipeptide Composition—The dipeptide composition was used to transform the variable length of proteins to fixed length feature vectors. Dipeptide composition has been used earlier by Grassmann *et al.* (12) and Reczko and Bohr (13) for the development of fold recognition methods (12–13). We adopted the same dipeptide composition-based approach in developing an SVM-based method for predicting subcellular localization of eukaryotic proteins (11). The dipeptide composition gave a fixed pattern length of 400. Dipeptide composition encapsulates information about the fraction of amino acids as well as their local order. It was calculated using Equation 2,

$$\text{Fraction of dep(i)} = \frac{\text{total number of dep(i)}}{\text{total number of all possible dipeptides}} \quad \text{(Eq. 2)}$$

where dep(i) is one dipeptide i of 400 dipeptides.

Evaluation of Performance—The performance all classifiers was evaluated through 5-fold cross validation. In 5-fold cross validation, the data set was partitioned randomly to five equally sized sets. The training and testing of each classifier was carried out five times using one distinct set for testing and the other four sets for training. The performance of classifiers was evaluated by measuring accuracy and the MCC for each subfamily of nuclear receptors, as described by Hua and Sun (9) and shown below in Equations 3 and 4,

$$\text{Accuracy(x)} = \frac{p(\text{x})}{\exp(\text{x})} \quad \text{(Eq. 3)}$$

$$\text{MCC(x)} = \frac{p(\text{x})n(\text{x}) - u(\text{x})o(\text{x})}{\sqrt{(p(\text{x}) + u(\text{x}))(p(\text{x}) + o(\text{x}))(n(\text{x}) + u(\text{x}))(n(\text{x}) + o(\text{x}))}} \quad \text{(Eq. 4)}$$

where x can be any subfamily of nuclear receptors, exp(x) the number of sequences observed in subfamily x, p(x) the number of correctly predicted sequences of subfamily x, n(x) the number of correctly predicted sequences not of subfamily x, u(x) the number of under-predicted sequences, and o(x) the number of over-predicted sequences.

Reliability Index (RI)—The determination of prediction reliability is important when using machine learning techniques to assign subfamilies of nuclear receptors. The reliability index (RI) was assigned on the basis of difference (Δ) between highest and second highest value of SVMs in multi-class classification (22–23). RI provides an insight into the accuracy/reliability of prediction. Equation 5, shown below,

$$\text{RI} = \begin{cases} \text{INT } (\Delta^*5/3) + 1 & \text{if } 0 \le \Delta < 4 \\ 5 & \text{if } \Delta \ge 4 \end{cases} \quad \text{(Eq. 5)}$$

demonstrates how the RI was defined for each sequence.

RESULTS AND DISCUSSION

Artificial intelligence-based techniques such as SVM and the neural network are elegant approaches for the extraction of complex patterns from biological sequence data. These techniques are highly successful for residue state prediction, where fixed window/pattern length is used (24). The major limitation of artificial intelligence techniques is that they need pattern/input units of fixed length. In this study, amino acid composition and dipeptide composition were used to transform the variable lengths of proteins to fixed length patterns. The classifiers were developed using the support vector machines, because it was shown in the past that SVM is better at classifying the biological data in comparison with the artificial neural network (25–26).

The results of 5-fold cross validation of amino acid composition-based and dipeptide composition-based classification are summarized in Table II. The accuracy of the five different sets used in the evaluation of the classifier during 5-fold cross-validation are shown in Table S1 of the supplementary material (at www.imtech.res.in/raghava/nrpred/help.html#bottom and in the on-line version of this article). The information about the number of true positive (P), true negative (N), false positive (O), and false negative (U) sequences predicted in each set used during 5-fold cross-validation are illustrated in Table S2 of the supplementary material. Table II illustrates that the best results achieved kernel parameters. The overall accuracy and MCC of the amino acid composition-based classifier for classifying the four subfamilies of nuclear receptors was 82.6% and 0.74, respectively. It proved that subfamilies of nuclear

TABLE II
The prediction accuracy (ACC) and MCC of both amino acid and dipeptide composition-based classifiers with the radial basis function (RBF) type of kernel function

Nuclear receptors	Amino acid composition RBF kernel ($\gamma = 500$; C = 5)[a]		Dipeptide composition RBF kernel ($\gamma = 3$)[a]	
	ACC	MCC	ACC	MCC
	%		%	
Thyroid hormone-like (TR, RAR, ROR)	87.7	0.75	100	0.98
HNF4-like (HNF4, RXR, TLL, COUP, USP)	68.0	0.62	95.8	0.96
Estrogen-like (ER, ERR, GR, MR, PR, AR)	89.3	0.83	98.7	0.96
Fushi tarazu-F1-like (SF1, FTF, FTZ-F1)	80.9	0.89	85.3	0.92
Overall	82.6	0.74	97.5	0.96

[a] Results were obtained through 5-fold cross-validation.

TABLE III
The performance of different pseudo-amino acid composition-based approaches in classifying the subfamilies of the nuclear receptors

The pseudo-amino acid composition was derived as described by Chou and Cai (27). The rank specifies the sequence order-coupling mode along the sequence. For example, the 1st rank reflects the coupling mode between most contiguous residues, and the 5th ranks specify the coupling mode between all of the 5th most contiguous residues. The rank "up to 5th" means it was developed using pseudo-amino acid composition generated by merging five sequence-order-correlated factors with amino acid composition (amino acid composition + (1st + 2nd + 3rd + 4th + 5th) sequence order factors. ACC denotes accuracy.

Pseudo-amino acid composition (sequence-order-correlated factor (a) + amino acid composition)										
Rank	Thyroid-like		HNF-like		EST-like		FUS-like		Overall	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Up to 5th	87.7	0.81	80.5	0.76	93.3	0.86	80.9	0.82	86.8	0.81
Up to 10th	87.7	0.81	80.5	0.72	93.3	0.88	80.9	0.87	86.8	0.81
Up to 15th	88.6	0.79	79.1	0.74	94.6	0.90	80.9	0.89	87.1	0.81
Up to 20th	89.4	0.82	81.9	0.77	94.6	0.90	80.9	0.87	88.2	0.83
Up to 25th	90.3	0.82	84.7	0.78	96.7	0.93	76.1	0.86	89.5	0.84
Up to 30th	92.1	0.85	86.1	0.84	97.3	0.93	76.1	0.84	90.7	0.86
Pseudo-amino acid composition (hydrophobicity-correlation factor (b) + amino acid composition)										
Rank	Thyroid-like		HNF-like		EST-like		FUS-like		Overall	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Up to 5th	88.6	0.75	65.3	0.60	88.0	0.79	57.1	0.68	80.1	0.71
Up to 10th	87.7	0.75	62.5	0.60	90.7	0.79	66.7	0.72	80.5	0.72
Up to 15th	90.4	0.78	66.6	0.60	88.0	0.80	61.9	0.74	81.5	0.73
Up to 20th	88.6	0.78	70.8	0.66	90.7	0.81	71.4	0.81	83.3	0.75
Up to 25th	85.9	0.76	72.2	0.66	90.6	0.82	71.4	0.75	82.5	0.74
Up to 30th	86.8	0.74	65.2	0.60	89.3	0.79	71.4	0.84	80.8	0.72
Pseudo-amino acid composition (a + b + amino acid composition)										
Rank	Thyroid-like		HNF-like		EST-like		FUS-like		Overall	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Up to 5th	86.8	0.79	80.5	0.73	90.6	0.83	71.4	0.81	85.0	0.79
Up to 10th	88.6	0.76	63.8	0.61	93.3	0.83	66.7	0.75	81.0	0.73
Up to 15th	91.2	0.78	76.8	0.66	90.6	0.85	66.7	0.81	85.5	0.77
Up to 20th	90.3	0.80	77.7	0.74	94.6	0.89	76.1	0.86	87.2	0.81
Up to 25th	87.7	0.77	72.2	0.67	92.0	0.84	76.1	0.86	84.0	0.76
Up to 30th	89.7	0.78	69.4	0.65	92.0	0.84	76.1	0.86	84.1	0.76

receptors correlated with amino acid composition and can be easily distinguished on this basis.

Recently, Chou and Cai demonstrated that the order of sequence along with amino acid composition (pseudo-amino acid composition) improved the accuracy of the classification of proteins (27, 28). Therefore, we also used pseudo-amino acid composition to classify nuclear receptors with better accuracy and reconfirm the role of sequence order on classification accuracy.

First, pseudo-amino acid composition was generated using amino acid composition with "sequence-order-correlated" factors. The different pseudo-amino acid compositions were generated using sequence-order-correlated factors of different orders (1st to 35th rank), as described by Chou and Cai (27). The overall performance of the classifiers generated by using pseudo-amino acid in classifying the four subfamilies of nuclear receptors is shown in Table III. The results depicted show that the overall MCC (0.86) and accuracy (90.7%) of the pseudo-amino acid based approach is better in comparison with that of the conventional amino acid composition-based approach. The results also demonstrate that accuracy as well as MCC increases up to the 30th rank (pseudo-amino acid composition generated by merging sequence-order-correlated factors from the 1st to the 30th rank with amino acid composition) and becomes nearly constant afterward. Thus, a remarkable improvement in predicting the subfamilies of nuclear receptors was achieved using pseudo-amino acid composition.

In a similar manner, a pseudo-amino acid composition based on the hydrophobicity correlation factor was generated to see the effect of biochemical properties on classification. Pseudo-amino acid composition based on this approach was derived up to the 30th rank from our data set. The calculation of the hydrophobicity-correlated factor was done by formulas from Chou and Cai (27). The hydrophobicity values of amino acids were taken from Argos *et al.* (29). The performance of pseudo-amino acid composition generated using the amino acid composition and hydrophobicity correlation factor is demonstrated

in Table III. The performance of the classifier based on this approach was slightly lower than that seen with the amino composition-based approach.

Finally, a pseudo-amino acid composition was generated using sequence-order factor and hydrophobicity-correlated factor with amino acid composition. The performance of classifier based on this approach up to the 30th rank is shown in Table III. The 30th rank reflects the coupling mode between all of the 30th most contiguous residues. The performance shown in Table III was obtained using 5-fold cross-validation. The best results (accuracy = 87.2 and MCC = 0.81) obtained using this approach were significantly better in comparison with those obtained with the amino acid composition-based approach. These results proved that pseudo-amino acid composition could provide more information about a protein sequence, resulting in the improvement of prediction accuracy.

On the other hand, the performance of the dipeptide composition-based approach was better as compared with that of the pseudo-amino acid composition-based approach. The most likely reason for the lesser performance of the pseudo-amino acid composition-based approach may be that it considers only identical pairs of amino acids and ignores non-identical pairs. The dipeptide composition based approach considers all of the contiguous pairs of amino acids whether they are identical or non-identical.

To further improve prediction accuracy, a classifier based on the conventional dipeptide composition of protein was developed. The dipeptide encapsulates the global information of the amino acid fraction and the local order of amino acids. Thus, dipeptide composition is a better feature as compared with amino acid composition alone. The overall accuracy and MCC of a dipeptide composition-based classifier were 97.6% and 0.97, respectively, which were significantly higher in comparison with the accuracy and MCC of the amino acid composition-based classifier. The detailed performance of the classifier on each set that was used for evaluation during 5-fold cross-validation is shown in Tables S1 and S2 of the supplementary material (www.imtech.res.in/raghava/nrpred/help.html#bottom and in the on-line version of this article). The overall accuracy of the dipeptide composition-based classifier was nearly 15% greater than that of the amino acid composition-based classifier. The overall MCC of the dipeptide composition-based classifier was 0.97, which was significantly higher than the MCC of the amino acid composition-based classifier. The detailed accuracy and MCC of dipeptide composition based classifiers in recognizing different subfamilies of nuclear receptors are shown in Table II. The results demonstrate that the thyroid hormone-like receptor subfamily is classified more accurately (accuracy = 100% and MCC = 0.98) in comparison with other subfamilies. The most likely reason for such results may be due to the large size of the data set related to thyroid hormone-like receptors. It is expected that the success rate for the other subfamilies can be further enhanced by improving the training data by adding more new proteins belonging to the subfamilies defined here.

These results also illustrated that the dipeptide composition-based classifier is able to recognize the subfamilies of nuclear receptors more accurately as compared with amino acid composition-based classifier alone. This suggests that different subfamilies of nuclear receptors are correlated dipeptide composition considerably.

Reliability Index—To bring further confidence to the user about the reliability of the prediction, the reliability index of both amino acid composition-based prediction and dipeptide composition-based prediction was also calculated. The RI assignment provides information about the certainty of prediction for a particular sequence. The reliability index was assigned according to the difference between the highest and the second highest 1-v-r SVM output score (9, 11). If a sample was predicted to have a large positive score for a class of nuclear receptors (away from the average score), the sample had a greater probability of belonging to that class. The reliability index is the key tool for considering receptors with high prediction accuracy (30). The curves shown in Fig. 1 answer the question of how reliable is the prediction for sequences labeled with a particular reliability index. For example, in the case of amino acid composition-based classifiers, the expected accuracy of sequences with RI = 5 is 100%, with 32.2% of the sequences of the whole data set having RI = 5. The expected accuracy at different RI values is shown by plotting a curve between the expected accuracy and the RI (Fig. 1*A*). A similar curve is also plotted for dipeptide composition-based classifiers (Fig. 1*B*). The dipeptide composition-based classifier predicted 84.2% sequences with RI > = 3. These sequences (RI > = 3) are nearly 100% correctly predicted. These results suggest that our method is able to predict the subfamilies of nuclear receptors with high accuracy. Thus, this classifier will complement the existing similarity search-based methods like BLAST and FASTA in recognizing the nuclear receptor proteins with high accuracy.

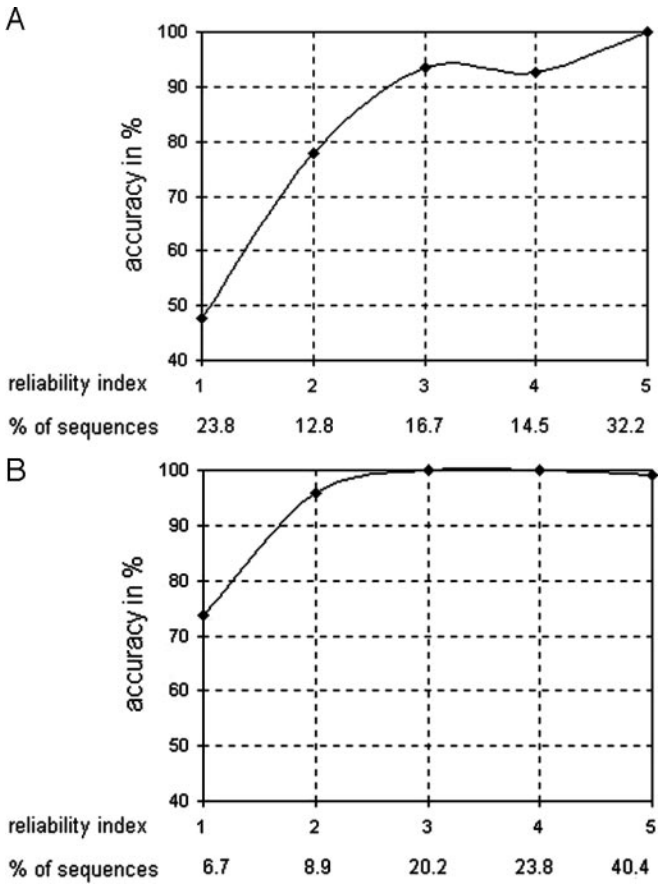

FIG. 1. *A*, expected accuracy of an amino acid composition-based classifier with a reliability index equal to a given value. The fraction of sequences that are predicted at a given reliability index are also shown on the *x*-axis. *B*, expected accuracy of a dipeptide composition-based classifier with a reliability index equal to a given value. The fraction of sequences that are predicted at a given reliability index are also shown on the *x* axis.

Conclusion—We have developed the NRpred server for recognizing the subfamilies of nuclear receptors proteins. This method, in association with a similarity search tool, can be used for automated annotation of genomic data. The study also

proves that there is a direct correlation between the features of the proteins (amino acid and dipeptide composition) and the subfamilies of nuclear receptors. The establishment of such methods will speed up the pace of identifying subfamilies of nuclear receptors and, thus, will facilitate drug discovery for inflammatory diseases or osteoporosis.

Description of Server—NRpred runs as a CGI server, written in PERL and operating under Solaris 420R. The interface of the server is straightforward and intuitive. The server accepts the protein sequence in any standard format like EMBL, GCG, FASTA, or in plain text format. The server uses the Readseq program to read the input sequence. The server provides the option of prediction either on the basis of amino acid composition or dipeptide composition. After analysis, the result will be displayed in a user-friendly format. The result provides information about the predicted subfamily of nuclear receptors, its reliability index, and the expected accuracy. The server and related information is available from www.imtech.res.in/raghava/nrpred.

Any information from experimental biologists regarding the different subfamilies of nuclear receptors is most welcome. In the future, more information about different subfamilies of nuclear receptors will be collected to establish a high quality, large data set, because the performance of any knowledge-based method is dependent on the quality and quantity of data. This data set will be used to further update and increase the performance of method. The users are encouraged to give their feedback about any experimental conformation or falsification of the predictions.

Acknowledgments—We are thankful to Dr. Girish Varshney and Sanjoy Paul for carefully reading the manuscript.

REFERENCES

1. Horn, F., Vriend, G., and Cohen FE. (2001) *Nucleic Acids Res.* **29,** 346–349
2. Robinson-Rechavi, M., and Laude, V. (2003) *Methods Enzymol.* **364,** 95–118
3. Robinson-Rechavi, M., Escriva Garcia, H., and Laudet, V. (2003) *J. Cell Sci.* **116,** 585–586
4. Laudet, V and Gronemeyer, H. (2002) *The Nuclear Receptors FactsBook*, Academic Press, London
5. Moras, D and Gronemeyer, H. (1998) *Curr. Opin. Cell Biol.* **10,** 384–391
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410
7. Pearson, W. R., and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2444–2448
8. Shepherd, A. J., Gorse, D., and Thornton, J. M. (2003) *Proteins* **50,** 290–302
9. Hua, S., and Sun, Z. (2001) *Bioinformatics* **17,** 721–728
10. Chou, K. C., and Cai, Y. D. (2002) *J. Biol. Chem.* **277,** 45765–45769
11. Bhasin, M., and Raghava, G. P. S. (2004) *Nucleic Acids Res.*, in press
12. Grassmann, J., Reczko, M., Suhai, S., and Edler, L. (1999) in *Proceeding of the Seventh International Conference on Intelligent Systems for Molecular Biology*, *Heidelberg, Germany, August 6–10, 1999* (Lengauer, T., Schneider, R., Bork, P., Brutlag, D. L., Glasgow, J. I., Mewes, H.-W., and Zimmer, R., eds) pp. 106–112, American Association for Artificial Intelligence, Menlo Park, CA
13. Reczko, M., and Bohr, H. (1995) *Nucleic Acids Res.* **22,** 3616–3619
14. Karchin, R., Karplus, K., and Haussler, D. (2002) *Bioinformatics* **18,** 147–159
15. Cai, C. Z., Han, L. Y., Ji, Z. L., and Chen, Y. Z. (2004) *Proteins* **55,** 66–76
16. Cai, Y. D., Ricardo, P. W., Jen, C. H., and Chou, K. C. (2004) *J. Theor. Biol.* **226,** 373–376
17. Cai, C. Z., Wang, W. L., Sun, L. Z and Chen, Y. Z. (2003) *Math. Biosci.* **185,** 111–122
18. Han, L. Y., Cai, C. Z., Lo, S. L., Chung, M. C., and Chen, Y. Z. (2004) *RNA* **10,** 355–368
19. Brendel, V. (1992) *Math. Comput. Model.* **16,** 37–43
20. Joachims, T. (1999) in *Advances in Kernel Methods–Support Vector Learning* (Schölkopf, B., Burges, C. J. C., and Smola, A. J., eds) pp.169–184, MIT Press, Cambridge, MA
21. Reinhardt, A., and Hubbard, T. (1998) *Nucleic Acids Res.* **26,** 2230–2236
22. Rost, B and Sander, C (1993) *J. Mol. Biol.* **232,** 584–599
23. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016
24. Krogh, A., and Riis, S. K. (1996) *Advances in Neural Information Processing System 8* (Touretzky, D. S., Mozer, M. C., and Hasaselmo, M. E., eds) pp. 917–923, MIT Press, Cambridge, MA
25. Ward, J. J., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2003) *Bioinformatics* **19,** 1650–1655
26. Bhasin, M., and Raghava, G. P. (2004) *Bioinformatics* **20,** 421–423
27. Chou, K. C., and Cai, Y. D. (2003) *J. Cell. Biochem.* **90,** 1250–1260
28. Chou, K. C (2001) *Protiens* **43,** 246–255
29. Argos, P., Rao, J. K., and Hargrave, P. A. (1982) *Eur. J. Biochem.* **128,** 565–575
30. Hua, S., and Sun, Z. (2001) *J. Mol. Biol.* **308,** 397–407