Recent Progresses in Identifying Nuclear Receptors and Their Families

Xuan Xiao[1,2,*], Pu Wang[1] and Kuo-Chen Chou[3,4]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China; [2]Information School, ZheJiagn Textile & Fashion College, NingBo, 315211; [3]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia; [4]Gordon Life Science Institute, Belmont, Massachusetts, United States of America

Abstract: Nuclear receptors (NRs) are members of a large superfamily of evolutionarily related DNA-binding transcription factors. They regulate diverse functions, such as homeostasis, reproduction, development and metabolism. As nuclear receptors bind small molecules that can easily be modified by drug design, and control functions associated with major diseases (e.g. cancer, osteoporosis and diabetes), they are promising pharmacological targets. According to their different action mechanisms or functions, NR superfamily has been classified into seven families: NR1 (thyroid hormone like), NR2 (HNF4-like), NR3 (estrogen like), NR4 (nerve growth factor IB-like), NR5 (fushi tarazu-F1 like), NR6 (germ cell nuclear factor like), and NR0 (knirps or DAX like). With the avalanche of protein sequences generated in the postgenomic age, Scientists are facing the following challenging problems. Given an uncharacterized protein sequence, how can we identify whether it is a nuclear receptor? If it is, what family even subfamily it belongs to? To address these problems, many cheminformatics tools have been developed for nuclear receptor prediction. The current review is mainly focused on this field, including the functions, computational methods and limitations of these tools.

Keywords: Pseudo amino acid composition, physical-chemical property matrix, NR-2L, iNR-PhysChem, covariant discriminant, chou's invariance theorem, web-server.

INTRODUCTION

In the field of molecular biology, nuclear receptors (NRs) are a class of proteins found within cells that are responsible for sensing steroid and thyroid hormones and certain other molecules. Nuclear receptors are ligand-inducible transcription factors that regulate processes, such as homeostasis, differentiation, embryonic development and organ physiology. As nuclear receptors are involved in almost all aspects of human physiology and are implicated in many important diseases including cancer, diabetes and osteoporosis, understanding of these receptors has major implications for human biology and for the development of new drug treatments [1, 2]. Nuclear receptors are targets for pharmaceutical industries with similar importance as the G protein-coupled receptors (GPCRs), ion channels and kinases [3].

NRs are modular proteins composed of six distinct regions (See Fig. **1**, from http://en.wikipedia.org/wiki/File: Nuclear_Receptor_Structure.png) that correspond to functional and structural domains [4, 5]. The N-terminal region (A/B domain) is highly variable, and contains at least one constitutionally active transactivation region (AF-1) and several autonomous transactivation domains (AD); The most conserved region is the DNA-binding domain (DBD, C domain), which notably contains the P-box, a short motif responsible for DNA-binding specificity on sequences typically containing the AGGTCA motif, and is involved in dimerization of nuclear receptors. Between the DNA-binding and ligand-binding domains is a less conserved region (D domain) that behaves as a flexible hinge between the C and E domains, and contains the nuclear localization signal (NLS). The largest domain is the moderately conserved ligand-binding domain (LBD, E domain), which is responsible for many functions, mostly ligand induced, notably the AF-2 transactivation function, a strong dimerization interface, another NLS, and often a repression function. Nuclear receptors may or may not contain a final domain in the C-terminus of the E domain, the F domain, whose sequence is extremely variable and whose structure and function are unknown [6].

The importance of NRs has prompted a rapid accumulation of the relevant data from a great diversity of fields of research: sequences, expression patterns, 3-D (three-dimensional) structures, protein-protein interactions, target genes, physiological roles, mutations, etc. If searching in the comprehensive protein database UniProt (Release 2012_10) with the query words "nuclear hormone receptor family", you will obtain 7,605 results, from witch you can access the information of protein attributes, comments, ontologies and so on. Databases that revolve around a single protein family can help researchers in using all data needed for their research, while relieving them of the onerous tasks related to the retrieval of many data from different sources [7]. The NucleaRDB is a data source that holds many different data types in a well organized and easily accessible form [8]. The data are validated, internally consistent and updated regularly. The NucleaRDB provides access to the data via various interfaces, which depending on the users' needs, are suited either for automated access or interactive usage [9].

*Address correspondence to this author at the Computer Department, Jing-De-Zhen Ceramic Institute, China; Tel/Fax: 086-0798-8499229;
E-mails: jdzxiaoxuan@163.com or xxiao@gordonlifescience.org

1873-5294/13 $58.00+.00

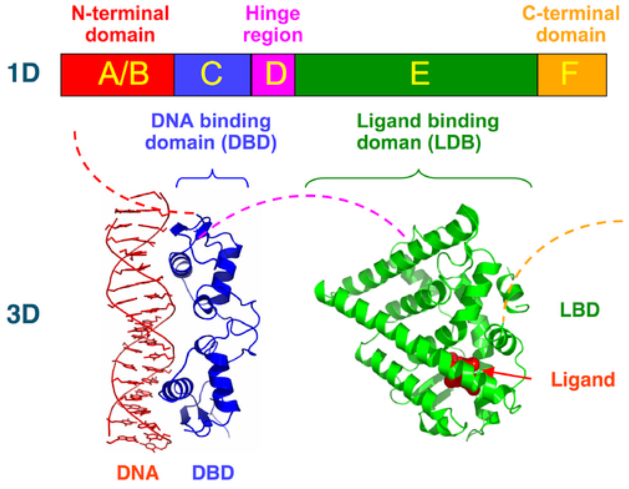


Fig. (1). Structural organization of nuclear receptors.

These accumulated data are very helpful for data mining and knowledge discovery. Since the function of a NR is closely correlated with which family or subfamily it belongs to, facing the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop automated methods for rapidly and effectively identifying NRs and their types according to their sequences information alone, because the knowledge thus acquired may benefit both basic research and drug development. Actually, some efforts have already been made in this regard. The current review is mainly focused on this field, including the functions, computational methods and limitations of the bioinformatics tools designed.

According to a comprehensive review [10] and demonstrated by a series of recent publications [11-15] , to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (1) construct or select a high quality benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform crossvalidation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly web-server for the predictor that is accessible to the public. The organization of content will be the same with above.

BENCHMARK DATASET

The nuclear receptor superfamily has been classified and assigned seven families according to the NucleaRDB database, which are (1) NR1: Thyroid hormone like (Thyroid hormone, Retinoic acid, RAR-related orphan receptor, Peroxisome proliferator activated, Vitamin D3-like), (2) NR2: HNF4-like (Hepatocyte nuclear factor 4, Retinoic acid X, Tailless-like, COUP-TF-like, USP), (3) NR3: Estrogen like (Estrogen, Estrogen-related, Glucocorticoid-like), (4) NR4: Nerve Growth factor IB-like (NGFI-B-like), (5) NR5: Fushi tarazu-F1 like (Fushi tarazu-F1 like), (6) NR6: Germ cell nuclear factor like (Germ cell nuclear factor) and (7) NR0: Knirps like (Knirps, Knirps-related, Embryonic gonad protein, ODR7, Trithorax) or DAX like (DAX, SHP). The dataset constructed in [16] only covered four NR families: (i) NR1; (ii) NR2; (iii) NR3; and (iv) NR5, and the sequences for the other three families are not considered because of the relatively small number (<10). The final dataset includes 282 proteins obtained from the NucleaRDB database. Furthermore, [17] classified the NRs into subfamily degree. The data was also taken from nucleaRDB database (Jul 2004 release). All putative/orphan sequences and fragments were excluded and redundancy was reduced so that pair-wise sequence identity is relative low. About 400 protein sequences compose the initial data set. According to pharmacological knowledge, these NRs belong to various subfamily components. Any subfamily that contained less than 6 proteins was dropped for further consideration and NRs were divided into 19 subfamilies. The dataset in [18] covers all the NR families, what's more, the NR0 family was divided into NR0A (Knirps like) and NR0B (DAX like), so there were 465 sequences belonging to eight types. [19] proposed a technique to identify nuclear receptors among six families and the NR0 family is not included in the dataset. As improvement to the previous work, the authors also created a negative dataset (i.e., non-NR protein sequences) so that the NRs could be distinguished from the non-NRs. The GPCR protein sequences were taken as the negative samples. The dataset of NRs used in [20] was also extracted from the nucleaRDB. The authors grouped all protein sequences by CD-HIT with the cluster identity threshold of 0.9 to ensure that no sequence had ≥90% sequence similarity to any sequences in the dataset, families with too few nuclear receptors were also excluded from the dataset for statistical significance. The final dataset contains 345 proteins belonging to four families: NR1, NR2, NR3 and NR5. The authors in [21] and [22] constructed a stricter benchmark dataset. The initial data set had 727 sequences covering all the seven families of nuclear receptors. To avoid any homology bias, a redundancy cutoff was imposed with the program CD-HIT to winnow those sequences which have ≥60% pairwise sequence identity to any other in a same subset except for the NR6 because it contained only 5 nuclear receptor protein sequences [23]. If the redundancy-cutoff operation was also executed on this class, the samples left would be too few to have any statistical significance. The final data set contains 159 sequences belonging to different families. Meanwhile, in order to train a statistical predictor with the ability to distinguish NR proteins from non-NR proteins, non-NR datasets are also constructed by collecting thousands of non-NR proteins from the UniProt at http://www.uniprot.org/ according their annotations in the "Keyword" field. After redundancy-cutoff operation as above, 500 non-NR proteins with low-redundancy are randomly selected to form the training dataset.

MATHEMATICAL EXPRESSION FOR PROTEIN SEQUENCE

The most straightforward formulation for a protein sample $\mathbf{P}$ of L amino acids is its entire amino acid sequence; i.e.,

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\cdots\mathrm{R}_L \tag{1}$$

where R_1 represents the 1st residue, R_2 the 2nd residue, …, R_L the L-th residue, and they each belong to one of the 20 native amino acids: A, C, D, E, F, G, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. In order to identify its attribute(s), the sequence-similarity-search-based tools, such as BLAST [24], was utilized to search the protein database for those proteins that have high sequence similarity to the query protein $\mathbf{P}$. Subsequently, the attribute(s) of the proteins thus found were used to deduce the attribute(s) for the query $\mathbf{P}$. Unfortunately, this kind of straightforward sequential model, although quite intuitive and able to contain the entire information of a protein sequence, failed to work when the query protein $\mathbf{P}$ did not have significant sequence similarity to any attribute-known proteins.

Thus, various non-sequential or feature vector models were proposed in hopes to establish some sort of correlation or cluster manner by which to enhance the prediction power.

Among the discrete models for a protein or protein sample, the simplest one is its amino acid (AA) composition or AAC [25]. According to the AAC-discrete model, the protein $\mathbf{P}$ of Eq. **1** can be formulated by [26]

$$\mathbf{P} = [f_1 \quad f_2 \quad \cdots \quad f_{20}]^{\mathrm{T}} \tag{2}$$

where $f_i \ (i = 1,\ 2,\ \cdots,\ 20)$ are the normalized occurrence frequencies of the 20 native amino acids in protein $\mathbf{P}$, and T the transposing operator. Many methods for predicting protein attributes were based on the AAC-discrete model (see, e.g., [25, 27-29]). However, as we can see from Eq. **2**, if using the ACC model to represent the protein $\mathbf{P}$, all its sequence-order effects would be lost, and hence the prediction quality might be considerably limited.

To avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC) [30, 31] was proposed , as formulated by

$$\mathbf{P} = [p_1 \ p_2 \ \cdots \ p_{20} \ p_{20+1} \ \cdots \ p_{20+\lambda}]^{\mathrm{T}} \tag{3}$$

where

$$p_u = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} p_j}, & (1 \le k \le 20) \\ \dfrac{wp_{k-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} p_j}, & (20+1 \le k \le 20+\lambda) \end{cases} \tag{4}$$

where $f_i \ (i = 1,\ 2,\ \cdots,\ 20)$ are the same as in Eq. **2**, and $p_j \ (j = 1, 2, \cdots, \lambda)$ are the pseudo amino acid components, and w the weight factors.

Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and proteins-related systems [32], such as identifying bacterial virulent proteins [33], predicting supersecondary structure [34], predicting protein subcellular location [35-37], predicting membrane protein types [38], discriminating outer membrane proteins [39], identifying antibacterial peptides [40], identifying allergenic proteins [41], predicting metalloproteinase family [42], predicting protein structural class [43], identifying GPCRs and their types [44], identifying protein quaternary structural attributes [45], predicting protein submitochondria locations [46], identifying risk type of human papillomaviruses [47], identifying cyclin proteins [48], predicting GABA(A) receptor proteins [49], classifying amino acids [50], among many others (see a long list of papers cited in the References section of [10]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [51, 52], as well as other biological samples (see, e.g., [53, 54]). Because it has been widely and increasingly used, recently two powerful softwares, called ‘PseAAC-Builder’ [55] and ‘propy’ [56], were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server ‘PseAAC’ [57] built in 2008.

For nuclear receptor families prediction [16], two kinds of PseAAC were actually adopted that are sequence-order-correlated factors [58], and hydrophobicity-correlated factor [59]. The results thus obtained indicate that pseudo-amino acid composition could indeed provide more information about a protein sequence, resulting in the improvement of prediction accuracy.

On the other hand, the dipeptide composition [60], which is actually also a different mode of PseAAC [55, 56], was also used in [16] for sequence encoding. Traditional dipeptide (amino acid pair) composition was used to capture the local-order information of a protein sequence, which gives a fixed pattern length of 400. The fraction of each dipeptide was formulated as

$$\text{Fraction of dip}(u) = \frac{\text{Total number of dip}(u)}{\text{Total number of all possible dipeptides}} \tag{5}$$

where $\mathrm{dip}(u) \ (u = 1, 2, \cdots, 400)$ is the u-th dipeptide.

Meanwhile, the simplified 4-tuple residues composition was also used encode the amino acid sequences [17], where the sequence alphabet was first reduced from 20 amino acids to six categories of biochemical similarity: [I, L, M, V], [F, W, Y], [H, K, R], [D, E], [N, P, Q, T] and [A, C, G, S] [61]. After this reduction, there were 6^4=1296 possible substrings of length 4. The researchers extracted and counted the occurrences of these substrings from a NR sequence string in a sliding window fashion. For a given protein sequence, the 4-tuple residues composition is simply an integral vector of length 1296, in which each bit indicates the counts the corresponding length-4 substring occurs in the protein.

The Fourier transform (FT) has been commonly used in bioinformatics [62-64] because the frequency content of signals is often of great importance. It is a good method in capturing the essence of data. In ref. [18], three kinds of substitution models: hydrophobicity model, electron-ion interaction potential model [65] and c-p-v model [66], representing three principal properties of hydrophobicity, electronic property and bulk respectively, were used to transform the NR sequences into numerical sequences. Then the Fast Fourier transform (FFT) was used to transform proteins of variable length into fixed length vectors. The power spectrum or power spectral density, a measurement of the power at various frequencies was extracted by using 512-point FFT. But note that, as pointed out by Liu *et al.* [67], only the low-frequency parts of Fourier spectrum were used to represent the PseAAC components. This is because "the high-frequency components are more noisy and hence only the low-frequency components are more important, just like the case of protein internal motions where the low-frequency components are functionally more important [68-70].

In the paper by Gupta *et al.* [19], seven physicochemical properties of amino acids are used to transform the symbolic sequences into numeric sequence and depicts the variation of physicochemical properties of protein sequences. Then based on the Maximal Overlap Discrete Wavelet Transforming (MODWT) [71], followed by proposing a novel feature extracting method based on Maximal Overlap Discrete Wavelet Transforming (MODWT) [71]. Furthermore, the dimension of the proposed feature vector is equal to 35. This helps in building faster and memory efficient classifier than dipeptide composition based approach.

Based on physicochemical characters of amino acids, Gao *et al.* [20] proposed an optimal pseudo amino acid composition to represent proteins for predicting the families of nuclear receptors. Six physicochemical characters of amino acids were adopted to generate the protein sequence features via the web server PseAAC at http://chou.med.harvard.edu/bioinf/PseAA. The optimal values of the rank of correlation factor and the weighting factor about PseAAC were determined to get the appropriate descriptor of proteins that leads to the best performance.

By incorporating various physicochemical and statistical features derived from the protein sequences, such as amino acid composition, dipeptide composition, complexity factor [72], and low-frequency Fourier spectrum components into PseAAC [10], Wang *et al.* [21] developed a predictor called NR-2L predictor for identifying nuclear receptor subfamilies based on their sequence information alone.

In a latest paper Xiao *et al.* [22] proposed a web-server predictor called iNR-PhysChem for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. In their approach, a novel mode of PseAAC was formulated to represent protein sequences by a physical-chemical matrix via a series of auto-covariance and cross-covariance transformations. A total of ten physical-chemical (PC) properties were adopted for encoding the 20 native amino acids. Thus the protein P of Eq. **1** can be formulated with a $10 \times L$ physical-chemical property matrix as given by

$$\mathbf{P} = \begin{bmatrix} \mathrm{PC}^1(\mathrm{R}_1) & \mathrm{PC}^1(\mathrm{R}_2) & \cdots & \mathrm{PC}^1(\mathrm{R}_L) \\ \mathrm{PC}^2(\mathrm{R}_1) & \mathrm{PC}^2(\mathrm{R}_2) & \cdots & \mathrm{PC}^2(\mathrm{R}_L) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{PC}^{10}(\mathrm{R}_1) & \mathrm{PC}^{10}(\mathrm{R}_2) & \cdots & \mathrm{PC}^{10}(\mathrm{R}_L) \end{bmatrix} \qquad (\mathbf{6})$$

where $\mathrm{PC}^1(\mathrm{R}_1)$ is the value of the 1st PC property for residue R_1, $\mathrm{PC}^2(\mathrm{R}_1)$ is the value of the 2nd PC property for residue R_1, and so forth.

According to the concept of auto-covariance (AC), the correlation of the same PC property between two subsequences separated by λ amino acids can be formulated as

$$\mathrm{AC}(i,\lambda) = \sum_{j=1}^{L-\lambda} (\mathrm{PC}^i(\mathrm{R}_j) - \overline{\mathrm{PC}^i})(\mathrm{PC}^i(\mathrm{R}_{j+\lambda}) - \overline{\mathrm{PC}^i}) \Big/ (L-\lambda) \qquad (i = 1, 2, \ ..., \ 10) \qquad (\mathbf{7})$$

where $\lambda < L$ and $\overline{\mathrm{PC}}^i$ represents the mean value of the *i*th horizontal line in Eq. **6**, as given by

$$\overline{\mathrm{PC}}^i = \sum_{j=1}^{L} \mathrm{PC}^i(\mathrm{R}_j) / L \qquad (\mathbf{8})$$

As we can see from Eq. **7**, using auto-covariance on the physical-chemical property matrix of Eq. **6**, $10 \times \lambda$ auto-covariance components can be generated.

On the other hand, according to the concept of cross-covariance (CC), the correlation between two subsequences with each belonging to a different PC property can be formulated by

$$\mathrm{CC}(i1, i2, \lambda) = \sum_{j=1}^{L-\lambda} (\mathrm{PC}^{i1}(\mathrm{R}_j) - \overline{\mathrm{PC}^{i1}})(\mathrm{PC}^{i2}(\mathrm{R}_{j+\lambda}) - \overline{\mathrm{PC}^{i2}}) \Big/ (L-\lambda) \qquad (\mathbf{9})$$
$$(i1 = 1, 2, ..., 10; \ i2 = 1, 2, ..., 10; \ i1 \neq i2)$$

Hence, using cross-covariance on the physical-chemical property matrix of Eq. **6**, we can generate $10 \times 9 \times \lambda$ cross-covariance components.

Accordingly, a total of $(10 \times \lambda + 10 \times 9 \times \lambda) = 100 \times \lambda$ components can thus be generated from Eq. **6**. Thus, in this study the PseAAC for protein P is expressed as

$$\mathbf{P} = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_u & \cdots & \psi_{100\times\lambda} \end{bmatrix}^{\mathbf{T}} \qquad (\mathbf{10})$$

where ψ_u is the *u*-th components generated by operating the above auto-covariance and cross-covariance on the physical-chemical property matrix of Eq. **6**.

OPERATION ALGORITHM OR PREDICTION ENGINE

There have been various and mature prediction engines or classification algorithms in the field of machine learning, and they often are used directly in the area of bioinformatics, such as Covariant Discriminant (CD) [26, 73-77], Support Vector Machine (SVM) [78, 79], and K-Nearest Neighbor (KNN) [80, 81], among many others.

The CD algorithm is a very elegant predictor based on the Mahalanobis distance [82-86]. It is instructive to point out that it may cause divergent problem when using the CD prediction engine to deal with those systems in which the components of constituent feature vectors are normalized. To avoid the divergent problem, the dimension-reduced procedure [31, 76, 77] based on the Chou's Invariance Theorem [26, 87] is needed. For more information about Chou's Invariance Theorem and its applications, see a Wikipedia article at http://en.wikipedia.org/wiki/Chou%27s_invariance_ theorem.

The KNN classifier is quite popular in pattern recognition community owing to its good performance and simple-to-use feature. According to the KNN rule [88-90], named also as the "voting KNN rule", the query sample should be assigned to the subset represented by a majority of its K nearest neighbors, as illustrated in Fig. **5** of [10]. There are many different definitions to measure the "nearness" for the KNN classifier, such as Euclidean distance, Hamming distance [91], and Mahalanobis distance [26, 82, 92]. A state-of-the-art technique in dealing with the parameter K for the KNN approach is the so-called "ensemble KNN classifier" established by fusing many individual KNN classifiers with different values of K. The ensemble KNN classifier has proved to be very powerful for predicting subcellular locations of proteins [93, 94] as well as their many other attributes (see, e.g., [95-98]). The detailed formulation of how to construct a powerful ensemble classifier by fusing many basic individual KNN classifiers, see a comprehensive reviews [81] as well as Eqs.17-23 of [10]. Fuzzy KNN classification method [99] is a special variation of the KNN classification family. Instead of roughly assigning the label based on a voting from the K nearest neighbors, it attempts to estimate the membership values that indicate how much degree the query sample belongs to the classes concerned. Obviously, it is impossible for any characteristic description to contain complete information, which would make the classification ambiguous. In view of this, the fuzzy principle is very reasonable and particularly useful under such a circumstance. Fuzzy KNN classification method has been used successfully for NR prediction [21].

SVMs are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The original SVM algorithm was proposed by Vladimir Vapnik and the current standard incarnation (soft margin) was described by Corinna Cortes and Vladimir Vapnikhttp://en.wikipedia.org/wiki/Support_ vector_machine - cite_note-0#cite_note-0 [100]. This algorithm first maps data into a high-dimensional feature space, and then establishes a hyperplane as the decision-making surface, which maximizes the boundary between two classes. The actual mapping is achieved through a kernel function, making it easy to implement and fast to compute. In principle, SVM is a two-class classifier. With the recent improvements, the SVM can directly cope with multi-class classification problem via the one-against-all or pairwise approach. Of the freely accessible SVM toolkits, the LIBSVM package is one of the most famous, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/ libsvm/. SVM has been used successfully for NR prediction [16-20, 22].

RESULTS AND WEB-SERVERS

The existing methods for identifying NRs and their types can basically be summerized as follows.

The **NRpred** [16] is a SVM based tool for the classification of nuclear receptors on the basis of amino acid composition or dipeptide composition. The overall prediction accuracy of amino acid composition and dipeptide composition based methods is 82.6% and 97.2% respectively by 5-fold cross-validation. The method can classify the nuclear receptors among the following four families: NR1, NR2, NR3 and NR5. The web-server for Nrpred is available at www.imtech.res.in/raghava/nrpred.

The method developed by Cai and Li [17] was based a 4-tuple residue composition and SVM that can be used to classify the 19 subfamilies of NRs. It was reported by the authors [17] that about 96% success rate was observed by a 5-fold cross-validation test. Unfortunately, no web-server was provided for their method and hence its usage is limited.

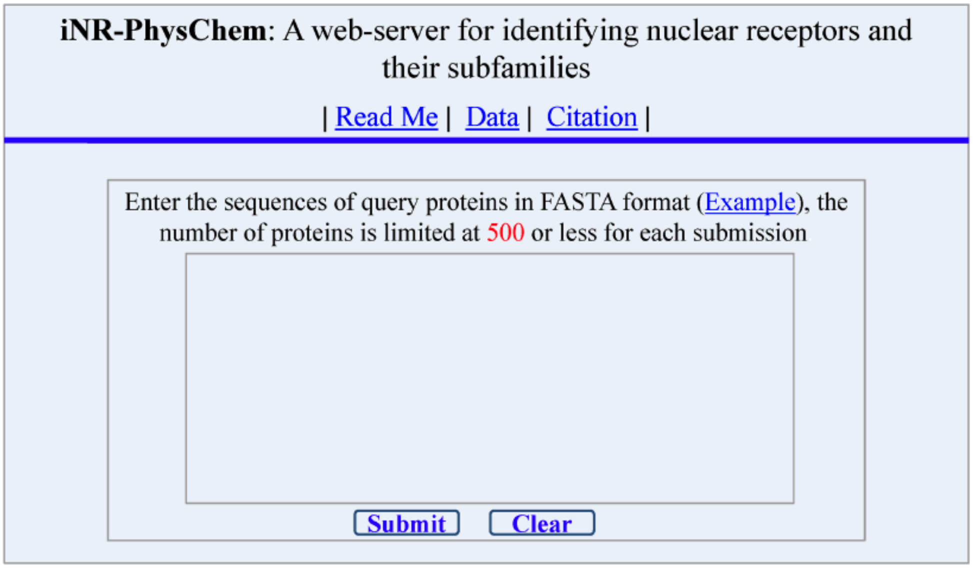
The method developed in [18] was based on a combination of Fast Fourier transform with SVM; it can be used to classify NRs among among NR1, NR2, NR3, NR4, NR5, NR6, NR0A and NR0B. The overall jackknife success rate reported by the authors was 95.3%. As stated by the authors in their paper [18], the corresponding web-server was provided at http://chem.scu.edu.cn/blast/Pred-NR. Unfortunately, it is no longer working at present.

In 2007 Gupta *et al.* [19] proposed a predictor for indentfying NRs and their families based on the wavelet variance of seven important physicochemical properties of amino acids. Tested by a 10-fold cross-validation, these authors reported a success rate of 99.91% in indentifying NRs from non-NRs, and by 5-fold cross-validation, a success rate of 96.19% in indentifying NRs among their five falilies. Again, no web-server was provided.

Two years later, by combining the optimal pseudo amino acid composition and SVM technique Gao *et al.* [20] reported an overall accuracy of 99.6% in indentifying the families of NRs among NR1, NR2, NR3 and NR5. The auuracy rate was derived from both 5-fold cross-validation and jackknife tests, indicating their method is quite competitive with that of **Nrpred** [16].

All the aforementiond methods each have their own merits and did play a role in stimulating the development of this area, but they all have the following shortcomings. (1) The datasets constructed to train the predictors cover very limited NRs families. For instance, the datasets constructed in [16, 20] only cover four families. (2) The cutoff threshold set by them to remove homologous sequences was 90%, meaning that the benchmark dataset thus constructed would allow inclusion of those proteins which have up to 90% pairwise sequence identity to others. (3) The existing web-server could not filter the irrelevant sequences, and all the input sequences would be assumed belonging to NRs regardless and hence might generate meaningless outcome, and hence their application value is quite limited. To improve this kind of situations, the following two predictors were developed recently.

Wang *et al.* [21] developed a two-level predictor, called **NR-2L**, by which one can identify a query protein as a NR or non-NR based on its sequence information alone. If it is, the prediction will be automatically continued to further identify the query protein among all the seven families of NRs. The identification was made by the Fuzzy K nearest neighbor classifier based on the pseudo amino acid composi-


iNR-PhysChem: A web-server for identifying nuclear receptors and their subfamilies

Enter the sequences of query proteins in FASTA format (Example), the number of proteins is limited at 500 or less for each submission

Fig. (2). User interface of the nuclear receptor predictor **iNR-PhysChem**.

tion. The overall success rates, achieved by the jackknife test on a low redundancy (≦60) dataset, were about 93% and 89% for the first and second level predictions, respectively. **NR-2L** is freely accessible at http://www.jci-bioinfo.cn/NR2L.

Shortley afterwords, based on the benchmark dataset constructed in [21], Xiao *et al.* [22] developed a predictor called **iNR-PhysChem** for identifying NRs and their families via Physical-Chemical Property Matrix. The overall jackknife success rate in identifying NRs or non-NRs achieved by **iNR-PhysChem** was over 98%, and the corresponding overall success rate in identifying NRs among their seven families was over 92%. The powerful predictor is freely accessible to the public at http://www.jci-bioinfo.cn/iNR-PhysChem and the top page of its interface is shown in (Fig. **2**).

CONCLUSION AND PERSPECTIVES

Nuclear receptors have been deemed to play an important role in many pathological processes, such as cancer, diabetes, rheumatoid arthritis, asthma or hormone-resistance syndromes. In the post genomic age, with the increasing number of NRs being discovered and stored in biological database, it is feasible as for us to develop high throughput tools for reliably identifying NRs and their families. The information thus obtained will be very useful for both basic research and drug development. During the last decade or so, many efforts have been made in this regard [16-22]. To make the predicted results scientifically more reliable and practically more useful, the futhure efforts in this regard should be focused on the following aspects.

(1) With more experimental data available in future, the benchmark dataset used to train and test the predictor should be further improved from two different aspects. One is to enhance the coverage scope by including more classes of NR families, and the other is to impose more stringent cutoff threshold to reduce the redundancy and homologous bias.

(2) There are still many rooms to improve the formulations for protein sequences by catching the key characters of NRs and their different families via the general form of PseAAC [10], such as by incorporating the gene ontology, functional domain, and evolution informations. The grey model approach [101-104] is also quite promising in this regard.

(3) It is worthwhile to try various prediction engines that have proved quite efficient in other areas, such as Covariance Determinant (CD) [52], GIA (grey incident degree)-nearest neighbor [103], random forest algorithms [101, 105, 106], as well as various ensemble classifier approaches [81, 93, 107-121].

(4) Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods [122], it is important to provide a web-server for each new predictor, making it accessible to the public and user-friendly, i.e., most experimental scientists can use it to acquire their desired data without the need to follow the detailed mathematics.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No. 31260273 and No. 6121027), the Province National Natural Science Foundation of JiangXi (No.2010GZS0122, No.20114BAB211013 and No. 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490), the LuoDi plan of the Department of Education of JiangXi Province(KJLD12083), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008). The funders had no role in study design,

data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] Altucci, L.; Gronemeyer, H. Nuclear receptors in cell life and death. *Trends Endocrinol. Metab.*, **2001**, *12*, 460-468.

[2] Moore, J. T.; Collins, J. L.; Pearce, K. H. The nuclear receptor superfamily and drug discovery. *Chem. Med. Chem.*, **2006**, *1*, 504-523.

[3] Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.*, **2002**, *1*, 727-730.

[4] Huang, P.; Chandra, V.; Rastinejad, F. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. *Annu. Rev. Physiol.*, **2010**, *72*, 247-272.

[5] Kumar, R.; Thompson, E. B. The structure of the nuclear hormone receptors. *Steroids*, **1999**, *64*, 310-319.

[6] Robinson-Rechavi, M.; Escriva Garcia, H.; Laudet, V. The nuclear receptor superfamily. *J. Cell Sci.*, **2003**, *116*, 585-586.

[7] Folkertsma, S.; van Noort, P.; Van Durme, J.; Joosten, H. J.; Bettler, E.; Fleuren, W.; Oliveira, L.; Horn, F.; de Vlieg, J.; Vriend, G. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J. Mol. Biol.*, **2004**, *341*, 321-335.

[8] Horn, F.; Vriend, G.; Cohen, F. E. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.*, **2001**, *29*, 346-349.

[9] Vroling, B.; Thorne, D.; McDermott, P.; Joosten, H. J.; Attwood, T. K.; Pettifer, S.; Vriend, G. NucleaRDB: information system for nuclear receptors. *Nucleic Acids Res.*, **2012**, *40*, D377-380.

[10] Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.*, **2011**, *273*, 236-247.

[11] Xiao, X.; Wang, P.; Lin, W. Z.; Jia, J. H.; Chou, K. C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **2013**, *436*, 168-177.

[12] Jiang, Y.; Huang, T.; Lei, C.; Gao, Y. F.; Cai, Y. D.; Chou, K. C. Signal propagation in protein interaction network during colorectal cancer progression. *Biomed. Res. Int., openly accssible at http://www.hindawi.com/journals/bmri/**2013**/287019/*, **2013**, *2013*, 1-9.

[13] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.*, **2013**, *9*, 634-644.

[14] Chou, K. C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Mol. Biosyst.*, **2013**, 9, 1092-1100.

[15] Xu, Y.; Ding, J.; Wu, L. Y.; Chou, K. C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **2013**, *8*, e55844.

[16] Bhasin, M.; Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.*, **2004**, *279*, 23262-23266.

[17] Cai, J.; Li, Y. In *Advances in Neural Networks – ISNN **2005***; Wang, J., Liao, X.-F., Yi, Z., Eds.; Springer Berlin Heidelberg, **2005**; Vol. 3498, p 680-685

[18] Guo, Y. Z.; Li, M.; Lu, M.; Wen, Z.; Wang, K.; Li, G.; Wu, J. Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. *Amino Acids*, **2006**, *30*, 397-402.

[19] Gupta, R.; Mittal, A.; Singh, K.; Sharma, P.; Sharma, A. *Information Technology, (ICIT **2007**). 10th International Conference on*, **2007**; p 68-73.

[20] Gao, Q. B.; Jin, Z. C.; Ye, X. F.; Wu, C.; He, J. Prediction of nuclear receptors with optimal pseudo amino acid composition. *Anal. Biochem.*, **2009**, *387*, 54-59.

[21] Wang, P.; Xiao, X.; Chou, K. C. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS One*, **2011**, *6*, e23505.

[22] Xiao, X.; Wang, P.; Chou, K. C. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PLoS One*, **2012**, *7*, e30869.

[23] Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **2006**, *22*, 1658-1659.

[24] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*, 3389-3402.

[25] Nakashima, H.; Nishikawa, K.; Ooi, T. The Folding Type of a Protein Is Relevant to the Amino Acid Composition. *J. Biochem.*, **1986**, *99*, 153-162.

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

[27] Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.*, **1999**, *12*, 107-118.

[28] Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **1998**, *17*, 729-738.

[29] Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins*, **2003**, *50*, 44-48.

[30] Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43*, 246-255.

[31] Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21*, 10-19.

[32] Chou, K. C. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. *Curr. Proteomics*, **2009**, *6*, 262-274.

[33] Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2012**, *9*, 467-475.

[34] Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.*, **2011**, *32*, 271-278.

[35] Zhang, S. W.; Zhang, Y. L.; Yang, H. F.; Zhao, C. H.; Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids*, **2008**, *34*, 565-572.

[36] Kandaswamy, K. K.; Pugalenthi, G.; Moller, S.; Hartmann, E.; Kalies, K. U.; Suganthan, P. N.; Martinetz, T. Prediction of Apoptosis Protein Locations with Genetic Algorithms and Support Vector Machines Through a New Mode of Pseudo Amino Acid Composition. *Protein Pept. Lett.*, **2010**, *17*, 1473-1479.

[37] Mei, S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.*, **2012**, *310*, 80-87.

[38] Chen, Y. K.; Li, K. B. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol*., **2013**, *318*, 1-12.

[39] Hayat, M.; Khan, A. Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. *Protein Pept. Lett.*, **2012**, *19*, 411-421.

[40] Khosravian, M.; Faramarzi, F. K.; Beigi, M. M.; Behbahani, M.; Mohabatkar, H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett.*, **2012**, *20*, 180-186.

[41] Mohabatkar, H.; Beigi, M. M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. *Med. Chem.*, **2013**, *9*, 133-137.

[42] Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics*, **2011**, *12*, 191-197.

[43] Sahu, S. S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.*, **2010**, *34*, 320-327.

[44] Zia Ur, R.; Khan, A. Identifying GPCRs and their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-scale Energy Representation and Position Specific Scoring Matrix. *Protein Pept. Lett.*, **2012**, *19*, 890-903.

[45] Sun, X. Y.; Shi, S. P.; Qiu, J. D.; Suo, S. B.; Huang, S. Y.; Liang, R. P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst.*, **2012**, *8*, 3178-3184.

[46] Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **2008**, *34*, 653-660.

[47] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **2010**, *263*, 203-209.

[48] Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **2010**, *17*, 1207-1214.

[49] Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **2011**, *281*, 18-23.

[50] Georgiou, D. N.; Karakasidis, T. E.; Nieto, J. J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2009**, *257*, 17-26.

[51] Chen, W.; Feng, P. M.; Lin, H.; Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **2013**, *41*, e68.

[52] Chen, W.; Lin, H.; Feng, P. M.; Ding, C.; Zuo, Y. C.; Chou, K. C. iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE*, **2012**, *7*, e47843.

[53] Li, B. Q.; Huang, T.; Liu, L.; Cai, Y. D.; Chou, K. C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS ONE*, **2012**, *7*, e33393.

[54] Huang, T.; Wang, J.; Cai, Y. D.; Yu, H.; Chou, K. C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. *PLoS ONE*, **2012**, *7*, e34460.

[55] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **2012**, *425*, 117-119.

[56] Cao, D. S.; Xu, Q. S.; Liang, Y. Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **2013**, *29*, 960-962.

[57] Shen, H. B.; Chou, K. C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **2008**, *373*, 386-388.

[58] Chou, K. C.; Cai, Y. D. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.*, **2003**, *90*, 1250-1260.

[59] Chou, K. C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **2000**, *278*, 477-483.

[60] Reczko, M.; Bohr, H. The DEF data base of sequence based protein fold class predictions. *Nucleic Acids Res.*, **1994**, *22*, 3616-3619.

[61] Taylor, W. R.; Jones, D. T. Deriving an amino acid distance matrix. *J. Theor. Biol.*, **1993**, *164*, 65-83.

[62] Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **2002**, *30*, 3059-3066.

[63] Nagarajan, N.; Keich, U. FAST: Fourier transform based algorithms for significance testing of ungapped multiple alignments. *Bioinformatics*, **2008**, *24*, 577-578.

[64] Nagarajan, V.; Kaushik, N.; Murali, B.; Zhang, C.; Lakhera, S.; Elasri, M. O.; Deng, Y. A Fourier transformation based method to mine peptide space for antimicrobial activity. *BMC Bioinformatics*, **2006**, *7 Suppl 2*, S2.

[65] Cosic, I. Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications. *IEEE Trans. Biomed. Eng.*, **1994**, *41*, 1101-1114.

[66] Grantham, R. Amino acid difference formula to help explain protein evolution. *Science*, **1974**, *185*, 862-864.

[67] Liu, H.; Wang, M.; Chou, K. C. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem. Biophys. Res. Commun.*, **2005**, *336*, 737-739.

[68] Chou, K. C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.*, **1988**, *30*, 3-48.

[69] Chou, K. C. Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem. Sci.*, **1989**, *14*, 212-213.

[70] Chou, K. C. Low-frequency motions in protein molecules: beta-sheet and beta-barrel. *Biophys. J.*, **1985**, *48*, 289-297.

[71] Percival, D. B.; Mofjeld, H. O. Analysis of subtidal coastal sea level fluctuation using wavelets. *J. Am. Stat. Assoc.*, **1997**, *92*, 868-880.

[72] Xiao, X.; Shao, S. H.; Huang, Z. D.; Chou, K. C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput. Chem.*, **2006**, *27*, 478-482.

[73] Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **1998**, *17*, 729-738.

[74] Chou, K. C.; Elrod, D. W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.*, **1998**, *252*, 63-68.

[75] Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins.*, **2001**, *44*, 57-59.

[76] Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.*, **2003**, *22*, 395-402.

[77] Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins.*, **2003**, *50*, 44-48.

[78] Chou, K. C.; Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **2002**, *277*, 45765-45769.

[79] Cai, Y. D.; Zhou, G. P.; Chou, K. C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **2003**, *84*, 3257-3263.

[80] Cai, Y. D.; Chou, K. C. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem. Biophys. Res. Commun.*, **2003**, *305*, 407-411.

[81] Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[82] Pillai, K. C. S. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons. This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics: New York, **1985**; Vol. 5.

[83] Chou, K. C.; Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **1994**, *269*, 22014-22020.

[84] Liu, W.; Chou, K. C. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *J. Protein Chem.*, **1998**, *17*, 209-217.

[85] Chou, K. C.; Liu, W.; Maggiora, G. M.; Zhang, C. T. Prediction and classification of domain structural classes. *Proteins*, **1998**, *31*, 97-103.

[86] Chou, K. C.; Maggiora, G. M. Domain structural class prediction. *Protein Eng.*, **1998**, *11*, 523-538.

[87] Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*, 275-349.

[88] Geva, S.; Sitte, J. Adaptive nearest neighbor pattern classification. *IEEE Trans. Neural Netw.*, **1991**, *2*, 318-322.

[89] Denoeux, T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans Syst. Man. Cybern*, **1995**, *25.*, 804-813.

[90] Keller, J. M.; Gray, M. R.; Givens, J. A. A fuzzy k-nearest neighbours algorithm. *IEEE Trans Syst. Man. Cybern*, **1985**, *15*, 580-585.

[91] Mardia, K. V.; Kent, J. T.; Bibby, J. M. Multivariate Analysis: Chapter 11 Discriminant Analysis; Chapter 12 Multivariate analysis of variance; Chapter 13 cluster analysis (pp. 322-381); Academic Press: London, **1979**.

[92] Mahalanobis, P. C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India.*, **1936**, *2*, 49-55.

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

[94] Chou, K. C.; Shen, H. B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science, **2010**, 2, 1090-1103; doi:10.4236/ns.2010.210136). *Nat. Protoc.*, **2008**, *3*, 153-162.

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

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

[97] Chou, K. C.; Shen, H. B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential

evolution information. *Biochem. Biophys. Res. Comm.*, **2008**, *376*, 321-325.

[98] Shen, H. B.; Chou, K. C. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *J. Proteome Res.*, **2009**, *8*, 1577–1584.

[99] Keller, J. M.; Gray, M. R.; Givens, J. A. A fuzzy K-nearest neighbor algorithm. *IEEE Trans Syst. Man. Cybern*, **1985**, *SMC-15*, 580-585.

[100] Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.*, **1995**, *20*, 273-297.

[101] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS One*, **2011**, *6*, e24756.

[102] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. *PLoS One*, **2012**, *7*, e49040.

[103] Lin, W. Z.; Xiao, X.; Chou, K. C. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng. Des. Sel.*, **2009**, *22*, 699-705.

[104] Xiao, X.; Lin, W. Z.; Chou, K. C. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J. Comput. Chem.*, **2008**, *29*, 2018-2024.

[105] Kandaswamy, K. K.; Chou, K. C.; Martinetz, T.; Moller, S.; Suganthan, P. N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.*, **2011**, *270*, 56-62.

[106] Pugalenthi, G.; Kandaswamy, K. K.; Chou, K. C.; Vivekanandan, S.; Kolatkar, P. RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. *Protein Pept. Lett.*, **2012**, *19*, 50-56.

[107] Nanni, L.; Lumini, A. An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics*, **2006**, *22*, 1207-1210.

[108] Naveed, M.; Khan, A. GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble. *Amino Acids*, **2012**, *42*, 1809-1823.

[109] Ding, Y. S.; Zhang, T. L. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recogn. Lett.*, **2008**, *29*, 1887-1892.

[110] Jiang, X.; Wei, R.; Zhao, Y.; Zhang, T. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. *Amino Acids*, **2008**, *34*, 669-675.

[111] Afridi, T. H.; Khan, A.; Lee, Y. S. Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition. *Amino Acids*, **2012**, *42*, 1443-1454.

[112] Hayat, M.; Khan, A.; Yeasin, M. Prediction of membrane proteins using split amino acid and ensemble classification. *Amino Acids*, **2012**, *42*, 2447-2460.

[113] Kedarisetti, K. D.; Kurgan, L. A.; Dick, S. Classifier ensembles for protein structural class prediction with varying homology. *Biochem. Biophys. Res. Commun.*, **2006**, *348*, 981-988.

[114] Li, L.; Zhang, Y.; Zou, L.; Li, C.; Yu, B.; Zheng, X.; Zhou, Y. An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity. *PLoS ONE*, **2012**, *7*, e31057.

[115] Lin, J.; Wang, Y.; Xu, X. A novel ensemble and composite approach for classifying proteins based on Chou’s pseudo amino acid composition. *Afr. J. Biotechnol.*, **2011**, *10*, 16963-16968.

[116] Nanni, L.; Lumini, A. Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins. *Amino Acids*, **2008**, Accepted Mar-27-**2008**.

[117] Nanni, L.; Lumini, A. An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence. *Amino Acids*, **2008**, *35*, 573-580.

[118] Nanni, L.; Lumini, A. A Further Step Toward an Optimal Ensemble of Classifiers for Peptide Classification, a Case Study: HIV Protease. *Protein Pept. Lett.*, **2009**, *16*, 163-167.

[119] Nanni, L.; Lumini, A. Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins. *Amino Acids*, **2009**, *36*, 409-416.

[120] Peng, C. R.; Lu, W. C.; Niu, B.; Li, Y. J.; Hu, L. L. Prediction of the functional roles of small molecules in lipid metabolism based on ensemble learning. *Protein Pept. Lett.*, **2012**, *19*, 108-112.

[121] Shen, H. B.; Chou, K. C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **2006**, *22*, 1717-1722.

[122] Chou, K. C.; Shen, H. B. Review: recent advances in developing web-servers for predicting protein attributes (doi: 10.4236/ns.2009.12011). *Nat. Sci.*, **2009**, *2*, 63-92 (openly accessible at http://www.scirp.org/journal/NS/)

Received: February 26, 2013 **Revised: March 28, 2013** **Accepted: April 08, 2013**