ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 154

6th International Conference on Measurement, Instrumentation and Automation (ICMIA 2017)

A novel coding scheme of nuclear receptor subfamilies prediction

Chun-cai XIAO[1, a]

[1]School of Electromechanical Engineering Xinyu University, Jiangxi Xinyu 338004, China

[a]xiaochuncai123@yeah.net

Keywords: Nuclear receptor, Pseudo amino acid composition, Fuzzy K-nearest neighbor.

Abstract. Nuclear receptors (NRs) are important transcriptional regulators in animals. They regulate different functions, such as lipid, reproduction, carbohydrate metabolism, fibrosis and metabolism. NRs form a category of phylogenetically evolutional proteins and have been separated into diverse subfamilies on account of their domain function. But for predicting the subfamilies, the preliminary step is to distinguish whether the protein sequence is a nuclear receptor or non-nuclear receptor. The sample with a pseudo amino acid (PseAA) composition representation of the protein sequence so as to incorporate a plentiful amount of protein sequence pattern information in order to increase the prediction precision for the classification. This article, which is based on the value of hydrophobicity, hydrophilicity, side-chain mass for sequence, we put forward a new percentage of method to predict types from protein sequences of subfamilies. Three percentages are on the base of the physical and chemical properties were collected from each of the protein sequences are made for their PseAA. It could testify by means of the jackknife cross-check method that the total successful rate are over 95%. The experimental results indicate that bioinformatics based on theory methodology can simplify and make experimental studies more intuitive.

Introduction

The NRs are important role in regulating crucial gene expression for cell growth, differentiation and homeostasis[1]. As NRs bind the hormone response element that can be easily changed by drug design, and control functions related to major diseases (e.g. liver injury, liver fibrosis and liver cancer), they are promising pharmacological target genes[2,3]. NRs are classified into a category that includes receptors for steroid receptor, vitamin D and thyroid hormone and so on[4]. In recent years, It was found that there was an obvious corelevance between functional classes of proteins and amino acid composition. Researchers put the algorithm developme nts as the goal followed by predicting the functional type of a protein according to amino acid composition alone[5]. It failed to contain any help of protein sequence information in a protein despite the amino acid composition model is very simple. The method of PseAA composition was raised to avoid the sequence information losing when using the amino acid composition model[6]. The method of PseAA composition was first put forward to improve the performance prediction of membrane protein type[7]. The expression of a known protein sample through a set of discrete decimal numbers, where the first 20 numbers show the 20 components of the traditional amino acid composition while other numbers show the PseAA composition, as defined by its expression way.

Method

The data set of containing redundancy built by Wang and Xiao[8] and was used to indicate the present method for the purpose of facilitating the comparison between the other methods. It is extremely hard to discover its characteristic pattern detailedly when the sequence is interminable[9]. In order to overcome this limitation, we turn to the percentage of method came out of the amino acid sequence with the help of the physical and chemical properties of amino acid. There were various kinds of physical and chemical properties of amino acids, which were given in Table 1. The other significant aspect we thought over seriously the hydrophilic amino acid, which influences the composition of protein sequences and is broadly used in a lot of magazines. Kellis et al. discover that the main reasons is that the hydrophobic of side-chains for the folding of proteins[10]; Chou

apply the side-chain mass and the hydrophilic values of amino acids to make up the pseudo amino acid composition to improve the quality of prediction for protein cellular properties[11]. All of these results give effective assistances for the applying hydrophobic, hydrophilic and side-chain mass of protein to make up percentage.

Table 1. The various kinds of physical and chemical characteristics of amino acids

Amino acid	Symbol	Hydrophobic	Hydrophilic	Side-chain mass
Alanine	A	0.62	-0.5	15.0
Cysteine	C	0.29	-1.0	47.0
Aspartic acid	D	-0.90	3.0	59.0
Glutamic acid	E	-0.74	3.0	73.0
Phenylalanine	F	1.19	2.5	91.0
Glycine	G	0.48	0.0	1.0
Histidine	H	-0.40	-0.5	82.0
Isoleucine	I	1.38	-1.8	57.0
Lysine	K	-1.5	3.0	73.0
Leucine	L	1.06	-1.8	57.0
Methionine	M	0.64	-1.3	75.0
Aspargine	N	-0.78	0.2	58.0
Proline	P	0.12	0.0	42.0
Glutamine	Q	-0.85	0.2	72.0
Arginine	R	-2.53	3.0	101.0
Serine	S	-0.18	0.3	31.0
Threonine	T	-0.05	-0.4	45.0
Valine	V	1.08	-1.5	43.0
Tryptophan	W	0.81	-3.4	130.0
Tyrosine	Y	0.26	-2.3	107.0

We have a hypothesis to make the meaning more clear, N amino acids make up a protein P, the following notation:

$$P = R_1 R_2 \cdots R_N \tag{1}$$

Where R_1 refers to the first amino acids, R_2 refers to the second amino acids, and so on. In order to represent a protein sequence from a English letter as a number. The superiority of adding the PseAA components is that they can include some key characteristics of a protein sequence by means of a discrete mode as stated above[12]. In this way, a protein sequence can be represented by a vector with the help of the Chou's PseAA composition, i.e.,

$$X = [x_1, x_2, \cdots, x_{20}, x_{21}, x_{22}, x_{23}]^T \tag{2}$$

Where T is the inversion operator.

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i}, (1 \le k \le 20) \\ \dfrac{p_{(k-20)}}{\sum_{j=1}^{20} p_i}, (21 \le k \le 23) \end{cases} \tag{3}$$

Where f_i (i =1, 2, …, 20) are the percentages of the 20 natural amino acids in each protein, arranged on the basis of their single alphabetic codes in alphabetical order, p_j (j=1, 2, 3) are come from percentage of protein sequences. Meanwhile the Fuzzy K-nearest neighbor algorithm was used to create the classification.

Fuzzy K-nearest neighbor classification algorithm is a special forms of the K-nearest neighbor algorithm family[13,14]. Replace a voting of assigning roughly the label in the nearest neighbors with

estimating the category values that distinguish how much relevancy the unknown sample pertain to the corresponding member, as briefed follows.

We presume that $\{P_1, P_2, \cdots, P_N\}$ is a group of vectors denoting N proteins in the training group which has been divided into M categories: $\{C_1, C_2, \cdots, C_M\}$, where C_i indicates the i-th category. Thus, for a unknown protein sequence P, its fuzzy sample value for the i-th category is described as follows:

$$\mu_i(P) = \frac{\sum_{j=1}^{K} \mu_i(P_j) d(P, P_j)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} d(P, P_j)^{-2/(\varphi-1)}} \tag{4}$$

Where K is the amount of the nearest neighbors calculated, $\mu_i(P_j)$ is the fuzzy sample statistic of the protein P_j to the i-th category (it is configured to 1 if the true marker of P_j is C_i, or else it is zero), $d(P, P_j)$ is the interval between the unknown protein sequence P and its j-th nearest protein sequence P_j in the training set, and $\varphi\ (>1)$ is the fuzzy factor to determine the weight of interval when calculate the contribution of each nearest neighbor to the sample value.

The Euclidean measure was adopted in this paper. After computing all the samples for a unknown protein sequence, it is distributed to the category which it has the best match sample value; i.e., the predicted category for the unknown protein sequence P should be

$$C_u = \mathbf{argmax}_i \{\mu_i(P)\} \tag{5}$$

where u is the thesis of i that maximizes $\mu_i(P)$.

Conclusions

The database have collected and harvested all the seven subfamilies of nuclear receptors marked with (1) NR1, (2) NR2, (3) NR3, (4) NR4, (5) NR5, (6) NR6 and (7) NR0. Among the methods for examining the effects of prediction, the independent sample test, secondary sample procedures test, and jackknife test, which are employed frequently to checkout the precision of a statistical forecasting technique[15], the jackknife test was regarded as the best objective in practical use. As a consequence, we employ the jackknife cross-validation to checkout our way. The success rates of jackknife test are got with the present percentage predictor in recognizing nuclear receptors are provided in Table 2.

Table 2: Success rates are got by jackknife test in recognizing the seven subfamilies of nuclear receptors.

Subfamily	Number of proteins	amino acid composition	the PseAA composition
NR0	18	61.11%	77.78%
NR1	281	97.86%	97.15%
NR2	163	92.64%	95.09%
NR3	185	98.92%	99.46%
NR4	30	93.33%	93.33%
NR5	45	82.22%	84.44%
NR6	5	80.00%	100%

Using the percentages are on the base of physical and chemical properties of protein sequence as their PseAA components can contain availably the relevant protein sequence patterns, producing a higher total success rate in forecasting types of seven subfamilies. These types of interrelated patterns are contained in a cluster of tanglesome sequences. Therefore, the types of subfamily are just

a example for verification. It is promoting significance to indicate that the present innovation method can also be adopted to forecast a range of other protein properties, such as the G protein-coupled receptor functional class, protein structural classes, and so on.

References

[1] Bhasin M, Raghava G P. Classification of nuclear receptors based on amino acid composition and dipeptide composition.[J]. Journal of Biological Chemistry, 2004, 279(22):23262-6.

[2] Su M G, Huang C H, Lee T Y, et al. Incorporating amino acids composition and functional domains for identifying bacterial toxin proteins.[J]. Biomed Research International, 2015, 2014(2014):972692.

[3] Altucci, Lucia, Gronemeyer, et al. Nuclear receptors in cell life and death[J]. Trends in Endocrinology & Metabolism Tem, 2001, 12(10):460-8.

[4] Altucci L, Gronemeyer H. Altucci L and Gronemeyer HNuclear receptors in cell life and death. Trends Endocrinol. Metab. 12: 460-468[J]. Trends in Endocrinology & Metabolism, 2002, 12(10):460-468.

[5] Chou K C. Progress in protein structural class prediction and its impact to bioinformatics and proteomics.[J]. Current protein & peptide science, 2005, 6(5):423-36.

[6] Liu L, Hu X Z, Liu X X, et al. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions.[J]. Protein & Peptide Letters, 2012, 19(4):439-449.

[7] Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins Structure Function & Bioinformatics, 2001, 43(3):246.

[8] Shen H B, Chou K C. Signal-3L: A 3-layer approach for predicting signal peptides.[J]. Biochemical & Biophysical Research Communications, 2007, 363(2):297-303.

[9] Pu W, Xuan X, Kuo-Chen C. NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features[J]. Plos One, 2011, 6(8):e23505.

[10] Ramanathan K, Shanthi V, Rao S. Contribution of unconventional C-H…O bonds to the structural stability of Antimicrobial peptides[J]. Interdisciplinary Sciences: Computational Life Sciences, 2009, 1(4):263-271.

[11] Xiao X, Chou K C. Using Pseudo Amino Acid Composition to Predict Protein Attributes Via Cellular Automata and Other Approaches[J]. Current Bioinformatics, 2011, 6(2):251-260.

[12] Chou K C, Shen H B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms[J]. Nature Protocols, 2008, 3(2):153-62.

[13] Hayat M, Khan A. Discriminating outer membrane proteins with Fuzzy K-nearest Neighbor algorithms based on the general form of Chou's PseAAC[J]. Protein & Peptide Letters, 2012, 19(4):411.

[14] Castillo O, Melin P. Hybrid intelligent system for cardiac arrhythmia classification with Fuzzy K-Nearest Neighbors and neural networks combined with a fuzzy system[J]. Expert Systems with Applications An International Journal, 2012, 39(3):2947-2955.

[15] Ding H, Lin H, Chen W, et al. Prediction of protein structural classes based on feature selection technique[J]. Interdisciplinary Sciences: Computational Life Sciences, 2014, 6(3):235-240.