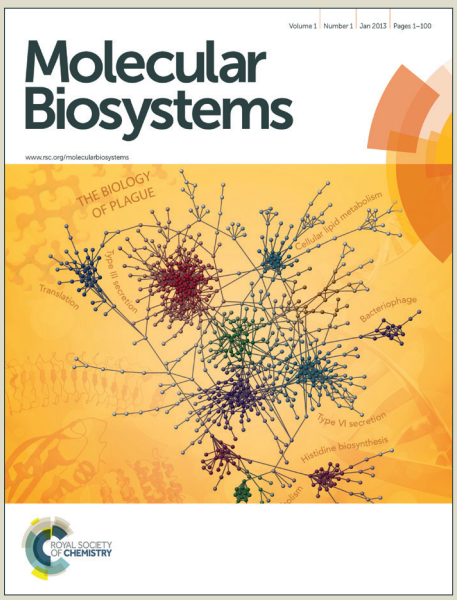
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences

Wei Chen[1,3*], Hao Lin[2,3,4*], Kuo-Chen Chou[1,3,5*]

[1] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China;
[2] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China;
[3] Gordon Life Science Institute, Boston, Massachusetts, USA;
[4] Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65201, USA;
[5] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

* Corresponding authors

E-mail addresses of all authors

Wei Chen: chenweiimu@gmail.com; wchen@gordonlifescience.org
Hao Lin: hlin@uestc.edu.cn; linha@missouri.edu; hlin@gordonlifescience.org
Kuo-Chen Chou: kcchou@gordonlifescience.org

Corresponding authors' mail addresses

WC: Department of Physics, Hebei United University, Tangshan 063000, China;
HL: School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054 China;
KCC: Gordon Life Science Institute, Boston, MA 02478, United States of America

ABSTRACT

With the avalanche of DNA/RNA sequences generated in the post-genomic age, it is urgent to develop automated methods for analyzing the relationship between the sequences and their functions. Towards this goal, a series of sequence-based methods have been proposed and applied to analyze various character-unknown DNA/RNA sequences in order for in-depth understanding their action mechanisms and processes. Compared with the classical sequence-based methods, the pseudo nucleotide composition or PseKNC approach developed very recently has the following advantages: (1) it can convert length-different DNA/RNA sequences into dimension-fixed digital vectors that can be directly handled by all the existing machine-learning algorithms or operation engines; (2) it can contain the desired features and properties according to the selection or definition of users; (3) it can cover considerable sequence pattern information, both local and global. This minireview is focused on the concept of pseudo nucleotide composition, its development and applications.

1. INTRODUCTION

The explosive growth of genomic sequences provides an unprecedented opportunity to explore genetic variability and biological function of organisms from a very fundamental point. In genome analysis, the information is generally given by a statistical distribution of sequence segments. So far, many methods have been proposed to decode the complicated genomes or DNA/RNA therein [1-3]. However, most of the existing methods were merely based on the nucleic acid composition or some short-range or local sequence order effect without taking into account the physicochemical properties of nucleotide and the long-range or global sequence order effect.

DNA/RNA sequences consist of four nucleotides (A, C, G and T/U). Thus, according to the formula given in [4], for a sequence of only 60 nucleotides, the number of different combinations of the four nucleotides would be $4^{60} = 10^{60\log 4} > 1.3289 \times 10^{36}$. Actually, the length of DNA/RNA sequences is much longer than 60, and hence the number of different combinations will be 1.3289×10^{36}. For such an astronomical number it is impracticable to construct a reasonable training dataset to statistically cover all the possible different sequence-order patterns. Besides, DNA/RNA sequences vary widely in length, which poses an additional difficulty for incorporating the sequence-order information in both the benchmark dataset construction and algorithm formulation. Facing such a dilemma, can we find an approach to partially incorporate the sequence-order effects?

Actually, similar problem also occurred in dealing with protein/peptide sequences. To address it, the pseudo amino composition [4, 5] or PseAAC [6, 7] was proposed. In PseAAC, a series of correlation factors along a protein/peptide chain is introduced to approximately reflect its sequence order effect. Ever since the concept

of PseAAC was proposed in 2001 [5], it has rapidly penetrated into nearly all the areas of computational proteomics, as reflected by the fact that in the papers [8-160], their titles contain either "pseudo amino acid composition" [5] or "PseAAC" [115], clearly indicating they were the key approaches for all these studies in various areas of computational proteomics. Recently, PseAAC was selected as one of the key topics in a special issue for drug development and biomedicine [161]. Its impact to medicinal chemistry was also reported in a recent review article [162] in a special issue [163].

Stimulated and encouraged by the successes of PseAAC in dealing with protein/peptide sequences, the pseudo nucleotide composition or PseKNC was introduced to predict methylation status of human DNA sequences [164], identifying recombination spots [145, 165], predicting promoters [166, 167], identifying translation initiation site in human genes [168], identifying splicing sites [169], predicting CpG island methylation status [170], predicting DNase I hypersensitive sites [171], predicting nucleosome positioning in genomes [172, 173], identifying microRNA precursor [174, 175], and predicting DNA methylation sites [176].

The present minireview is to summarize the progresses of using PseKNC to deal with DNA/RNA sequences in developing various methods for genome analysis, with the focus on those for which a publically accessible web-server has also been established.

2. PSEUDO NUCLEOTIDE COMPOSITION

2.1. Concept and Formulation

Similar to the formulation where a protein/peptide sequence is denoted by **P** [72], here we use **D**/**R** to represent a DNA/RNA sequence; i.e.

$$\mathbf{D}\,/\,\mathbf{R} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\mathrm{N}_4\mathrm{N}_5\mathrm{N}_6\mathrm{N}_7\cdots\mathrm{N}_{\mathrm{L}} \tag{1}$$

where L represents the length of a DNA/RNA sequence or the number of its constituent nucleic acid residues, and

$$\mathrm{N}_i \in \{\mathrm{A\ (adenine),\ C\ (cytosine),\ G\ (guanine),\ T\ (thymine)/U(urine)}\} \quad (i = 1,\ 2,\ \cdots,\ \mathrm{L}) \tag{2}$$

denotes the nucleic acid residue at the i-th sequence position, and $\in$ is a symbol in the set theory meaning "member of". Also, similar to the general form of PseAAC [72] for protein/peptide sequences, the general form of PseKNC can be formulated as

$$\mathbf{D}/\mathbf{R} = \left[\ \phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_{\mathrm{Z}}\ \right]^{\mathbf{T}} \tag{3}$$

where **T** is the transposing operator, the subscript Z is an integer, and its value and the components ϕ_u $(u = 1,\ 2,\ \cdots)$ will depend on how to extract the desired features

and properties from the DNA/RNA sequence (cf. **Eq.1**). The form of **Eq.3** can cover all the existing modes of PseKNC, as illustrated below.

When

$$\begin{cases} \phi_u = f_u^{\text{K-tuple}} \\ \text{Z} = 4^{\text{K}} \end{cases} \tag{4}$$

where $f_u^{\text{K-tuple}}$ $(u = 1,2,\cdots,4^{\text{K}})$ is the u-th component of the K-tuple nucleotide composition for a DNA/RNA sequence (cf. **Eq.1**), we immediately obtain the formulation of the K-tuple nucleotide composition (see **Eq.6** of [177]).

When

$$\phi_u = \begin{cases} \dfrac{f_u^{\text{K-tuple}}}{\sum_{i=1}^{4^{\text{K}}} f_i^{\text{K-tuple}} + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 4^{\text{K}}) \\ \dfrac{w\theta_{u-4^{\text{K}}}}{\sum_{i=1}^{4^{\text{K}}} f_i^{\text{K-tuple}} + w\sum_{j=1}^{\lambda}\theta_j} & (4^{\text{K}}+1 \le u \le 4^{\text{K}}+\lambda) \end{cases} \tag{5}$$

we obtain the formulation of the type-1 PseKNC (see Eq.11 of [177]). The component of θ_j in **Eq.5** is defined by

$$\begin{cases} \theta_1 = \dfrac{1}{\text{L}-\text{K}}\sum_{i=1}^{\text{L}-\text{K}}\Theta_{i,i+1} \\ \theta_2 = \dfrac{1}{\text{L}-\text{K}-1}\sum_{i=1}^{\text{L}-\text{K}-1}\Theta_{i,i+2} \\ \theta_3 = \dfrac{1}{\text{L}-\text{K}-2}\sum_{i=1}^{\text{L}-\text{K}-2}\Theta_{i,i+3} \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{\text{L}-\text{K}-\lambda+1}\sum_{i=1}^{\text{L}-\text{K}-\lambda+1}\Theta_{i,i+\lambda} \end{cases} \quad (\lambda < \text{L}-\text{K}) \tag{6}$$

where θ_1 is called the first-tier correlation or coupling factor introduced to reflect the sequence order correlation between all the most contiguous K-tuple nucleotides along the DNA sequence (**Fig.1a**); θ_2 is the second-tier correlation factor used to reflect the sequence order correlation between all the second most contiguous K-tuple nucleotides (**Fig.1b**); θ_3 the third-tier correlation factor used to reflect the sequence order correlation between all the third most contiguous K-tuple nucleotides (**Fig.1c**); and so forth. The number λ is an integer used to reflect the correlation rank (or tier) and hence must be smaller than (L-K).

When

$$\phi_u = \begin{cases} \dfrac{f_u^{\text{K-tuple}}}{\sum_{\text{i=1}}^{4^{\text{K}}} f_{\text{i}}^{\text{K-tuple}} + w\sum_{\text{j=1}}^{\lambda\Lambda}\tau_j} & (1 \le u \le 4^{\text{K}}) \\ \dfrac{w\tau_{u-4^K}}{\sum_{\text{i=1}}^{4^{\text{K}}} f_{\text{i}}^{\text{K-tuple}} + w\sum_{\text{j=1}}^{\lambda\Lambda}\tau_j} & (4^K+1 \le u \le 4^{\text{K}} + \lambda\Lambda) \end{cases} \quad (7)$$

we obtain the formulation of the type-2 PseKNC (see Eq.15 in [177]). The component of τ_j in **Eq.7** is defiled by (cf. **Fig.2**)

$$\begin{cases} \tau_1 = \dfrac{1}{\text{L}-\text{K}-1}\sum_{i=1}^{\text{L}-\text{K}-1} \text{J}_{i,i+1}^{1} \\ \tau_2 = \dfrac{1}{\text{L}-\text{K}-1}\sum_{i=1}^{\text{L}-\text{K}-1} \text{J}_{i,i+1}^{2} \\ \cdots\cdots\cdots\cdots \\ \tau_\Lambda = \dfrac{1}{\text{L}-\text{K}-1}\sum_{i=1}^{\text{L}-\text{K}-1} \text{J}_{i,i+1}^{\Lambda} \qquad \lambda < (\text{L}-\text{K}) \\ \cdots\cdots\cdots\cdots \\ \tau_{\lambda\Lambda-1} = \dfrac{1}{\text{L}-\text{K}-\lambda}\sum_{i=1}^{\text{L}-\text{K}-\lambda} \text{J}_{i,i+\lambda}^{\Lambda\text{-}1} \\ \tau_{\lambda\Lambda} = \dfrac{1}{\text{L}-\text{K}-\lambda}\sum_{i=1}^{\text{L}-\text{K}-\lambda} \text{J}_{i,i+\lambda}^{\Lambda} \end{cases} \quad (8)$$

It is instructive to point out that with the general formulation of **Eq.3**, PseKNC can cover much more properties, features and their intrinsic patterns, such as distance-pair composition [174], secondary structure status derived from Vienna RNA structure server [178], and the corresponding long-range or global order information used recently in identifying microRNA precursors [174, 175, 179, 180].

2.2. Physicochemical Property of Nucleotides

In **Eq.6** the correlation function is given by

$$\begin{cases} \Theta_{i,i+j} = \dfrac{1}{\Lambda}\sum_{\xi=1}^{\Lambda}\left[\text{H}_\xi\left(\text{R}_i\text{R}_{i+1}\cdots\text{R}_{i+\text{K-1}}\right) - \text{H}_\xi\left(\text{R}_{i+j}\text{R}_{i+j+1}\cdots\text{R}_{i+j+\text{K-1}}\right)\right]^2 \\ i = 1,\ 2,\ \cdots, \text{L}-\text{K}+1; \quad j = 1, 2, \cdots, \lambda; \quad \lambda < \text{L}-\text{K} \end{cases} \quad (9)$$

where R_i and all the other symbols of its kind can be any valid nucleic acid A, C, G, or T/U (cf. **Eq.1**); $H_\xi\left(R_iR_{i+1}\cdots R_{i+K-1}\right)$ is the numerical value of the ξ-th physicochemical property for the K-tuple nucleotide $R_iR_{i+1}\cdots R_{i+K-1}$ in a DNA/RNA sequence, and $H_\xi\left(R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}\right)$ the corresponding value for the K-tuple nucleotide $R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}$, while Λ is the total number of the correlation functions counted. As we can see from **Eq.9**, the different physicochemical properties considered are taken into account via a parallel manner, and hence the type-1 PseKNC belongs to the parallel type [51].

In **Eq.8**, the correlation function is given by

$$\begin{cases} J^{\xi}_{i,\,i+m} = H_\xi\left(R_iR_{i+1}\cdots R_{i+K-1}\right)\cdot H_\xi\left(R_{i+m}R_{i+m+1}\cdots R_{i+m+K-1}\right) \\ \xi = 1,\,2,\,\cdots,\,\Lambda; \quad m = 1,\,2,\,\cdots,\,\lambda; \quad i = 1,\,2,\,\cdots,\,L-K-\lambda \end{cases} \tag{10}$$

where H_ξ has exactly the same meaning as defined in **Eq9**; i.e., it is associated with the physicochemical properties of the K-tuple nucleotide concerned. As we can see from **Eq.10**, the different physicochemical properties considered are taken into account via a series manner, and hence the type-2 PseKNC belongs to the series type [4]. For both the parallel and series types, there are many choices to select the desired physicochemical properties, as will be further discussed later.

2.3. Optimizing the Parameters of PseKNC

As it is, there are three uncertain parameters in the PseKNC formulation. The 1st one is K, which reflects the sequence pattern within the segment of K nucleotides for DNA/RNA sequences. As we can see from **Eqs.3-4**, the dimension of PseKNC increases rapidly with the value of K. For example, when $K = 7$, i.e., the sequence order information considered is confined within a segment of seven nucleotides, the dimension of the corresponding PseKNC would be $Z = 4^7 = 16{,}384$. Such a high dimension will cause the high-dimension disaster [181] as reflected by the following disadvantages: (i) the overfitting problem that will make the predictor with a serious bias and extremely low capacity for generalization; (ii) the information redundancy or noise that will bring about the error of misrepresentation resulting in very poor prediction accuracy; and (iii) unnecessarily increasing the computational time. Thus, in practical application, the scope of K is generally set at 2, 3, or 4. Accordingly, the K-tuple nucleotide approach can only be used to reflect the short-range or local sequence order information for DNA/RNA sequences. The second parameter in PseKNC formulation is λ, which stands for the number of the total counted tiers of the long-range correlations along a DNA/RNA sequence, and hence is used for the global sequence pattern information. As shown in **Figs.1-2**, the maximum number allowed for λ in a DNA/RNA sequence (**Fig.1**) is $(L-K)$. The third parameter in PseKNC formulation is the weight factor *w*, which is used to

adjust the weight between the local and global sequence pattern effects (cf. **Eq.5** or **Eq.7**), and its value is within the range of $0 \le w \le 1$.

In a brief way, the three parameters and their roles in PseKNC can be formulated as

$$\begin{cases} \text{K}, & \text{number of nearest nucleotides for local pattern} \\ \lambda, & \text{number of correlation tiers for global pattern} \\ w, & \text{weight to adjust the effcts of K and } \lambda \end{cases} \quad (11)$$

and their values are determined via a optimization process with respect to various concrete problems and hence the final results will be case by case.

3. APPLICATIONS

Genome analysis can help in-depth understanding many complicated biological processes in cell network, and reveal the molecular mechanisms of genetic diseases or disorders. Therefore, it is highly demanded from both basic research and drug development to develop sequence-based tools to timely perform genome analyses on uncharacterized DNA/RNA sequences. Listed below are some recent reports of using the PseKNC approach to conduct genome analyses.

3.1. Identify Recombination Spots

Meiosis and recombination are the two opposite aspects coexisting in a DNA system (**Fig.3**). They are two indispensible aspects for cell reproduction and growth. The process of meiosis is a special type of cell division by which the genome is divided in half to generate daughter cells for participating in sexual reproduction, while the process of recombination is to produce single-strand ends that can invade the homologous chromosome. Therefore, a combination of the two processes plays a very important role in driving the evolution via generating natural genetic variations. Interestingly, rather than in a random manner across a genome, the recombination occurs with higher probability in some genomic regions called "hotspots", while with lower probability in those called "coldspots". Identification of recombination spots (hotspots) may provide very useful information for in-depth understanding the reproduction and growth of cells. Using the concept and approach of PseKNC, two predictors were developed for identifying the recombination spots in DNA. One is called "iRSpot-PseDNC" [165] and the other called "iRSpot-TNCPseAAC" [145]. Also, a publically accessible web-server for each of the two predictors has been established. Their website addresses are given in **Table 1**. Particularly, their success rates in identifying the recombination spots are remarkably higher than the method in which only the k-mar frequencies were used to incorporate the short-range or local sequence order effects [182].

3.2. Identify Nucleosome Positioning

The basic unit of eukaryotic chromatin is nucleosome, which contains a 147

bp core DNA [183] tightly wrapped in 1.67 left-handed superhelical turns around a histone octamer (**Fig.4)**. Participating in many cellular activities, nucleosomes play significant roles in these biological processes. Facing the explosive growth of genome sequences discovered in the postgenomic age, it is highly demanded to develop high throughput tools for rapidly and effectively identifying the nucleosome positioning sequences. Using the concept and approach of PseKNC, two web-server predictors, named "iNuc-PhysChem" [172] and "iNuc-PseKNC" [173], were developed to identifying nucleosome positioning in genome. Also, their web-servers have been established, as listed in **Table 1**. Again, the predictors thus obtained have remarkably outperformed the previous ones (see, e.g., [184, 185]).

3.3. Predict Promoters

Promoter is a region of DNA that determines the transcription of a particular gene. Based on the discrete wavelets transform and PseKNC, Zhou et al. [166] develop a method to predict promoters. The sigma-54 promoters (**Fig.5**) are unique in prokaryotic genome and responsible for transcripting carbon and nitrogen-related genes. Recently, using PseKNC and incremental feature selection technique, Lin et al. [167] developed a predictor, called "iPro54-PseKNC", for identifying sigma-54 promoters in prokaryote. Its web-server is given in **Table 1** as well. Compared with the corresponding previous work [186] without using PseKNC, iPro54-PseKNC [167] is much more accurate and catch the real features of sigma-54 promoters.

3.4. Identifying Translation Initiation Sites

Translation is a key process for gene expression, by which the information carried by the messenger RNA (mRNA) is decoded by ribosome complex to produce a specific protein (or peptide) chain according to the rules specified by the genetic code. Translation proceeds in four phases: (1) initiation, (2) elongation, (3) translocation, and (4) termination [187]. As illustrated in **Fig.6**, during the first initiation process, a proper start position on the mRNA will be identified. The region at which the translation initiated is called the Translation Initiation Site (TIS). Timely identification of TIS is very important for conducting in-depth genome analysis. It is by the genetic translation process that the information carried by the messenger RNA (mRNA) is decoded by ribosome complex to produce a specific protein (or peptide) chain according to the rules specified by the genetic code. Recently, by means of the PseKNC approach, a predictor called "iTIS-PseTNC" was developed for identifying TIS site in human genes [168]. The corresponding web-server has also been established, and its website address is given in **Table 1**. The success rates achieved by iTIS-PseTNC [168] are higher than those by StartScan [188]. To our best knowledge, the latter was the best predictor in identifying the human TIS prior to the appearing of iTIS-PseTNC [168].

3.5. Predict DNA Methylation Status and Sites

Predominantly occurring on cytosine within a CG dinucleotide, DNA methylation is a covalent modification of DNA catalyzed by DNA methyltransferase enzyme (DNMT) (**Fig.7**). The DNA methylation sites are occupied by various proteins, including methyl-CpG binding domain (MBD) proteins; the MBD-containing proteins can recruit varieties of histone deacetylase (HDAC) complexes and chromatin remodeling factors, causing chromatin compaction and transcriptional repression as well. By either impeding the binding of transcriptional proteins to the gene or bonding to the MBD, DNA methylation may affect the transcription of genes. It plays a significant role for epigenetic gene regulation in life development; it also plays a crucial role in developing nearly all types of cancer. Therefore, knowledge of DNA methylation sites is important for both basic research and drug development. To meet such demand, a web-server predictor called "iDNA-Methyl" [176] was developed, and its web-site address is given in **Table 1**. Meanwhile, by using the PseKNC approach as well, a method for predicting the methylation status of human DNA sequences [164] and a method for predicting the CpG island methylation status [170] were also developed. It is interesting to note that the model trained on the data from $CD4^+$ T lymphocyte cell was also applicable to other human tissues/cell types and that the predictive accuracy by the model of using PseKNC [170] are higher than that by only using the trinucleotide composition.

3.6. Detect Splicing Sites

In eukaryotic genomes, exons that code for proteins are typically interrupted by introns, the non-coding regions of genes. The borders between exons and introns are called splice sites (**Fig.8**). A splice site can be located at either the upstream or the downstream part of an intron. For the former, it is called the 5' splice site or donor site; for the latter, it is called the 3' splice site or acceptor site. The vast majority of the donor and acceptor sites are canonical or regular splice sites that are characterized by the presence of the GT and AG, respectively. During RNA splicing, both the donor and acceptor sites will be recognized by a large macromolecule called spliceosome that is comprised of more than 300 proteins and five small nuclear RNAs (snRNAs U1, U2, U4, U5, and U6) [189]. Once the splice sites are recognized, the spliceosome will remove introns through two sequential transesterification reactions (**Fig.8)**. Removing introns from precursor messenger RNA (pre-mRNA) so that exons can be joined together to form mature mRNA is an essential step of gene expression. Therefore, to better understand the splicing process and mechanism, it is important to accurately detect the splice sites in the genome. To address this, a predictor called "iSS-PseDNC" [169] was developed for detecting the splice sites by incorporating the six DNA local structural properties into the PseKNC formulation. Of the six properties, three are local translational (slide, shift, and rise) and three local angular (roll, tilt, and twist). Meanwhile, it was observed that the accuracy based on the dinucleotide composition alone is lower than that based on the pseudo dinucleotide composition [169], once again indicating the importance of global sequence order effects incorporated in PseKNC [177, 190].

3.7. Identify DNase I hypersensitive Sites

DNase I hypersensitive sites (DHSs) are those chromatin regions in genome where the structural density is loose and exposing, making them accessible by DNase I enzyme to degrade the structure. Therefore, DHSs are sensitive to the cleavage, and they are associated with a wide variety of regulatory DNA elements. Knowledge about the locations of DHS is helpful for deciphering the function of non-coding genomic regions. In view of this, a method was proposed to identify DHS with PseKNC [171].

3.8. Identify microRNA Precursors

Being a small non-coding RNA molecule, the microRNA (miRNA) plays an important role in transcriptional and post-transcriptional regulation of gene expression. So far over 1000 miRNAs have been encoded by the human genome, and hence they are widely, although still poorly characterized, deemed as important regulators. They are also involved in many other important biological processes, such as translation of mRNAs, affecting stability, and negatively regulating gene expression in post-transcriptional processes (**Fig.10**). It has been observed in many cancers and other disease states that there exist abnormal expressions of miRNAs, implying that they are deeply involved in these diseases, particularly in carcinogenesis. Accordingly, discriminating the real pre-miRNAs from the false ones (such as hairpin sequences with similar stem-loops) is important for both basic research and miRNA-based therapy. Base on the concept of PseKNC, two predictors were developed for identifying the real pre-miRNAs or true microRNA precursors: One is called "iMcRNA" [175], and one called "iMiRNA-PseDPC" [174]. The corresponding web-servers have also been established, and their website addresses are given in **Table 1**.

Because the PseKNC approaches have been increasingly used, recently three flexible web-servers have been established [177, 190, 191], and their website addresses are given below

$$\begin{cases} \text{PseKNC} & \text{http://lin.uestc.edu.cn/pseknc/default.aspx} \\ \text{repDNA} & \text{http://bioinformatics.hitsz.edu.cn/repDNA/} \\ \text{PseKNC-General} & \text{http://lin.uestc.edu.cn/server/pseknc} \end{cases} \quad (12)$$

where the PseKNC web-server [177] contains 38 and 12 built-in physicochemical properties for dinucleotides and trinucleotides, respectively, which can be selected by users to generate their desired modes of PseKNC for DNA sequences; the repDNA web-server [191] can generate various modes of PseKNC for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effect; the PseKNC-General web-server allows users to select their desired ones from more than 100 built-in physicochemical properties to generate PseKNC for both DNA and RNA sequences, and it also allows users to calculate PseKNC with the properties defined by their own.

4. SOME REMARKS ON COMPUTATIONAL GENOME ANALYSIS

4.1. Metrics for Measuring the Prediction Quality

In conducting genome analysis, we are often facing a binary (two-lass) classification problem; i.e., for a given segment or site of DNA/RNA sequence, whether its outcome is positive or negative such as those listed in the 1st or 2nd column of the following equation

$$\begin{cases} \text{Recombination hotspot} & \text{Recombination coldspot} \\ \text{Nucleosome} & \text{linker} \\ \text{Promoter} & \text{Non-promotor} \\ \text{Translational initiation site} & \text{Non-TIS} \\ \text{Methylation site} & \text{Non-methylationsite} \\ \text{Splicing site} & \text{Non-splicing site} \\ \text{Hypersensitive site} & \text{Non-hypersensitive site} \\ \text{True pre-miRNA} & \text{False pre-miRNA} \end{cases} \tag{13}$$

To this kind of binary classification problem, the following set of metrics were often used to measure the prediction quality

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN+FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \\ \mathrm{MCC} = \dfrac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}} \end{cases} \tag{14}$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient [192]. The metrics formulated in **Eq.14** is not easy-to-understand for most experimental scientists, and hence here we would prefer to use the following formulation as done in a series of recent publications (see, e.g., [133, 134, 146, 148, 149, 193]):

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+}, & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-}, & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-}, & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-} \right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+} \right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-} \right)}}, & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (15)$$

where N^+ is the total number of the positive samples investigated while N_-^+ the number of positive samples incorrectly predicted to be of negative sample; N^- the total number of the negative samples investigated while N_+^- the number of the negative samples incorrectly predicted to be of positive sample. According to **Eq.15** we can easily see the following. When $N_-^+ = 0$ meaning none of the positive samples was mispredicted to be negative, we have the sensitivity $\mathrm{Sn} = 1$; while $N_-^+ = N^+$ meaning that all the positive samples were mispredicted negative, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the negative sample was mispredicted, we have the specificity $\mathrm{Sp} = 1$; while $N_+^- = N^-$ meaning all the negative sample were incorrectly predicted to be of positive sample, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of the positive samples and none of the negative samples was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the positive samples and all the negative samples were mispredicted, we have the overall accuracy $\mathrm{Acc} = 0$. The Matthews correlation coefficient MCC is usually used for measuring the quality of binary classifications. When $N_-^+ = N_+^- = 0$ meaning that none of the positive samples and none of the negative samples was mispredicted, we have $\mathrm{MCC} = 1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $\mathrm{MCC} = 0$ meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have $\mathrm{MCC} = -1$ meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using **Eq.15** to examine a predictor for its four metrics, particularly for its Mathew's correlation coefficient.

Note that, of the four metrics in **Eq.14** or **15**, the most important for a predictor are Acc and MCC: the former stands for its overall success rate, and the latter for its stability; in contrast, the other two metrics Sn and Sp are used only for its the partial

success rates at different angles. Therefore, it is usually by optimizing the Acc and MCC to determine the three parameters of **Eq.11**.

Also, it should be pointed out that the metrics given in **Eqs.14-15** are valid only for the single-label systems; for the multi-label systems (see, e.g., [194-201], we should use a set of more complicated metrics as defined by **Eq.16** of a review paper [202].

4.2. Optimizing Imbalanced Training Datasets

Many existing predictors developed for conducting genome analysis were trained by a skewed training dataset in which the number of the negative samples is overwhelmingly larger than that of the positive ones. For example, in the benchmark dataset used to train the predictor iMcRNA [175] for identifying the microRNA precursors, there were 1,612 true pre-miRNAs or positive samples and 8,489 false pre-miRNAs or negative samples. In other words, the original benchmark dataset is very imbalanced: the size of the negative subset is more than five times the size of the positive subset. Although this might reflect the real world where the false pre-miRNAs or non-miRNA precursors are always the majority compared with the true pre-miRNAs or miRNA precursors, a predictor trained with such a skewed benchmark dataset would have the consequence that many miRNA precursors might be mispredicted as non-miRNA ones [203]. Actually, what is really most intriguing for basic research and drug development is the information of the positive samples. Therefore, it is important to find an effective approach to optimize the unbalanced benchmark dataset and minimize the consequence of this kind of misprediction.

In the study of identifying the microRNA precursors [175], to balance the size of negative subset with that of positive one, the authors adopted the random selection treatment; i.e., they randomly picked 1,612 samples from 8,489 non-miRNA precursors to form the negative subset, and make both positive and negative subsets have a same size.

Recently, a more effective treatment, the so-called "optimizing imbalanced training datasets" (OITD) treatment was introduced [176, 204]. The OITD treatment consists of the following two steps.

The first step is a subset-reducing operation by removing some redundant negative samples from the negative subset via the following three procedures: (1) for each sample in the training dataset, find its three nearest neighbors; (2) if the sample belongs to the negative subset and, of its three nearest neighbors, at lease two belong to the positive subset, then remove the sample from the negative training dataset; (3) if, however, it belongs to the positive subset, then removed should be those of its nearest neighbors that belong to the negative subset. For example, in the study of identifying DNA methylation sites [176], originally there were 787 positive samples and 1,639 negative samples. It was via the subset-reducing operation that 522 negative samples were removed making the negative subset only contain

$(1{,}639 - 522) = 1{,}117$ samples.

The second step is a subset-expanding operation by creating some hypothetical samples for the positive subsets via the linear interpolation scheme, which can be likened to the seed-propagation approach in [205] and the Monte Carlo sampling approach in [206, 207] for expanding the positive subsets. In the aforementioned study [205], it was via the subset-expanding operation that 330 positive hypothetical samples were created and added into the positive subset making it also contain $(787 + 330) = 1{,}117$ samples.

As we can see from above, after the OITD treatment, the original skewed training dataset constructed for studying DNA methylation sites [176] has become a perfectly balanced one with both the positive and negative subsets having a same size of 1,117 samples.

It is instructive to point out that the training dataset generated via the OITD treatment may contain some hypothetical data that are not experiment-confirmed samples. Will it affect the objectivity in evaluating the quality of a predictor trained by such a dataset? The answer is absolutely no. This is because the data obtained via the OITD treatment are only used for training a predictor but not for testing it. All the cross-validations must be carried out strictly based on, and only on, the experiment-confirmed data. When doing cross-validation to test the predictor, only the original experiment-confirmed data will be used. In other words, none of hypothetical samples, such as the 330 hypothetical positive samples created by the subset-expanding operation in [176], will be used for testing the predictor in counting its score. But on the other hand, all the experiment-confirmed samples, including those removed during the subset-reducing operation such as the 522 negative samples in [176], will be used as tested data for its score-counting. To realize this kind of testing procedure, a special cross-validation, the so-called "target-jackknife" test was introduced, as elaborated in [176].

The advantage of using the OITD treatment to optimize the imbalanced training datasets is quite obvious, as shown in **Table 2**. It can be seen from the table that, for a same prediction model, the one trained with the dataset treated by OIDT is remarkably superior to the one without such a treatment, particularly in the overall success rate (Acc) and Mathew Correlation Coefficient (MCC). The former stands for the overall accuracy of a predictor, and the latter for its stability.

5. CONCLUSIONS AND FUTURE EFFORTS

Being an extension of pseudo amino acid composition or PseAAC used to formulate the samples of protein/peptide sequences, the pseudo K-tuple nucleotide composition or PseKNC can be used to formulate the samples of DNA/RNA sequences. As demonstrated by a series of reports [8-160], PseAAC has been widely used in many areas of computational proteomics, it is anticipated that PseKNC will be increasingly and widely used in various areas of computational genetics and

genomics as well.

Particularly, with the development of RNA sequencing technology, more and more non-coding RNA transcripts (such lncRNAs) will be available. In our future work, we will apply the PseKNC to this realm.

As pointed out in [7, 208], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful analysis methods and predictors, we shall make efforts to also establish a web-server for predicting the CpG island methylation status [170], detecting the splicing sites [169], and identifying DNase I hypersensitive sites [171], as mentioned in Section 3.

In conducting genome analyses, we are often facing the highly imbalanced or skewed datasets, in which the negative samples are overwhelmingly larger than the positive ones. The method Optimizing Imbalanced Training Datasets or OITD treatment developed very recently is a very effective approach to deal with this kind of problems. We shall adopt the OITD treatment to improve the web-serves listed in **Table 1** as well as to develop new methods for genome analyses.

ACKNOWLEDGEMENTS

The authors wish to thank the three anonymous reviewers, whose constructive comments were very helpful for strengthening the presentation of this review article. This work was supported by the National Nature Scientific Foundation of China (Nos. 61100092, 61202256), the Nature Scientific Foundation of Hebei Province (No.C2013209105).

Table 1. List of web-servers developed with PseKNC for genome analysis and their website addresses.

Server's name	Target	Website address
iRSpot-PseDNC[a]	Recombination spot	http://lin.uestc.edu.cn/server/iRSpot-PseDNC
iRspot-TNCPseAAC[b]	Recombination spot	http://www.jci-bioinfo.cn/iRSpot-TNCPseAAC
iNuc-PhysChem[c]	Nucleosome positioning	http://lin.uestc.edu.cn/server/iNuc-PhysChem
iNuc-PseKNC[d]	Nucleosome positioning	http://lin.uestc.edu.cn/server/iNuc-PseKNC
iPro54-PseKNC[e]	Promoter	http://lin.uestc.edu.cn/server/iPro54-PseKNC
iTIS-PseTNC[f]	Translation initiation site	http://lin.uestc.edu.cn/server/iTIS-PseTNC
iDNA-Methyl[g]	DNA methylation site	http://www.jci-bioinfo.cn/iDNA-Methyl
iMcRNA[h]	MicroRNA precursor	http://bioinformatics.hitsz.edu.cn/iMcRNA/server
iMiRNA-PseDPC[i]	MicroRNA precursor	http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/

[a] See [165].
[b] See [145].
[c] See [172].
[d] See [173].
[e] See [167].
[f] See [168].
[g] See [176].
[h] See [175].
[i] See [174].

Table 2. A comparison of success rates for predictors trained by the datasets before and after the OITD treatment.

Network system	Status of training dataset	Success rates [a]			
		Acc (%)	MCC	Sn (%)	Sp (%)
Drug-GPCR	Before OIDT treatment [b]	85.50	0.6775	80.00	88.30
	After OIDT treatment [c]	**90.32**	**0.8066**	97.58	86.69
Drug-channel	Before OIDT treatment [d]	87.27	0.7233	86.30	87.76
	After OIDT treatment [c]	**88.78**	**0.7643**	91.98	87.17
Drug-enzyme	Before OIDT treatment [e]	91.03	0.8039	90.81	91.14
	After OIDT treatment [c]	**92.56**	**0.8429**	95.99	90.84
Drug-NR	Before OIDT treatment [f]	89.15	0.7519	79.07	94.19
	After OIDT treatment [c]	**93.02**	**0.8453**	91.86	93.60

[a] The rates reported in this table were obtained by the rigorous cross-validations on the original experiment-confirmed datasets.
[b] The drug-GPCR system is for studying the interaction between drugs and G-protein coupled receptors (GPCRs) in cellular networking. Before the OITD treatment, the training dataset used for the predictor iGPCR-Drug [129] was very unbalanced, with 620 positive samples and 1,240 negative samples.
[c] See [204] for the details of the corresponding training dataset obtained after the OITD treatment.
[d] The drug-channel system is for studying the interaction between drugs and ion channels. Before the OITD treatment, the training dataset for the predictor iCDI-PseFpt [130] contained 1,372 positive samples and 2,744 negative samples.
[e] The drug-enzyme system is for studying the interaction between drugs and enzymes. Before the OITD treatment, the training dataset for the predictor iEzy-Drug [209] contained 2,719 positive and 5,438 negative samples.
[f] The drug-NR system is for studying the interaction between drugs and nuclear receptors. Before the OITD treatment, the training dataset for the predictor iNR-Drug [210] was with 2,719 positive samples and 5,438 negative samples.

FIGURE LEGENDS

Figure 1. A schematic drawing to show the 1st or parallel type of PseKNC. Panel (**a**) reflects the correlation mode between all the most contiguous K-tuple nucleotides, panel (**b**) that between all the second-most contiguous K-tuple nucleotides, and panel (**c**) that between all the third-most contiguous K-tuple nucleotides. **P_1** represents the first K-tuple nucleotide, i.e., $R_1R_2...R_K$, along the DNA/RNA sequence; **P_2** the second K-tuple nucleotide $R_2R_3\cdots R_{K+1}$; **P_3** the third K-tuple nucleotide $R_3R_4\cdots R_{K+2}$; and so forth. $L^* = L - K$ is the maximum number allowed for the K-tuple nucleotides in a L-bp long DNA/RNA sequence.

Figure 2. A schematic drawing to show the 2nd or series type of PseKNC. Panel (**a1/a2**) reflects the correlation mode between all the most contiguous K-tuple nucleotides, panel (**b1/b2**) that between all the second-most contiguous K-tuple nucleotides, and panel (**c1/c2**) that between all the third-most contiguous K-tuple nucleotides, and so forth. See the legend of Fig.1 and the text for further explanation.

Figure 3. An illustration to show the process of meiosis and recombination in a DNA system.

Figure 4. A schematic illustration to show the basic structure of nucleosome. Each nucleosome consists of approximately 147 base pair of DNA wrapped 1.67 turns around a histone octamer. See the text for further explanation

Figure 5. A schematic illustration to show the basic structure of σ^{54} promoter and its biological process.

Figure 6. A schematic map to show the initiation region of translation process initiates.

Figure 7. A schematic drawing to show the process of DNA methylation. Catalyzed by DNA methyltransferase (DNMT), a methylation group (Me) is binding to the base cytosine (C) via a covalent bond.

Figure 8. A schematic drawing to show the pathways of RNA splicing. (**a**) The 2'OH of the branchpoint nucleotide within the intron (solid line) carries out a nucleophilic attack on the first nucleotide of the intron at the 5' splice site (GU) forming the lariat intermediate. (**b**) The 3'OH of the released 5' exon then performs a nucleophilic attack at the last nucleotide of the intron at the 3' splice site (AG). (**c**) Joining the exons and releasing the intron lariat.

Figure 9. A schematic drawing to show DNase I hypersensitive sites in chromatin.

Figure 10. An illustration to show biogenesis of miRNAs and the process mRNA

degradation. MiRNA genes are transcribed by RNA polymerase II, resulting in the primary transcripts termed as pri-miRNAs, which are typically 60-70 nucleotides. The pri-miRNAs are processed by the enzyme drosha to release the hairpin-shaped intermediates (pre-miRNAs), followed by being exported into the cytoplasm by exportin V and Ran-GTP cofactor, and then cleaved by the enzyme dicer to yield miRNA/miRNA* duplexes.

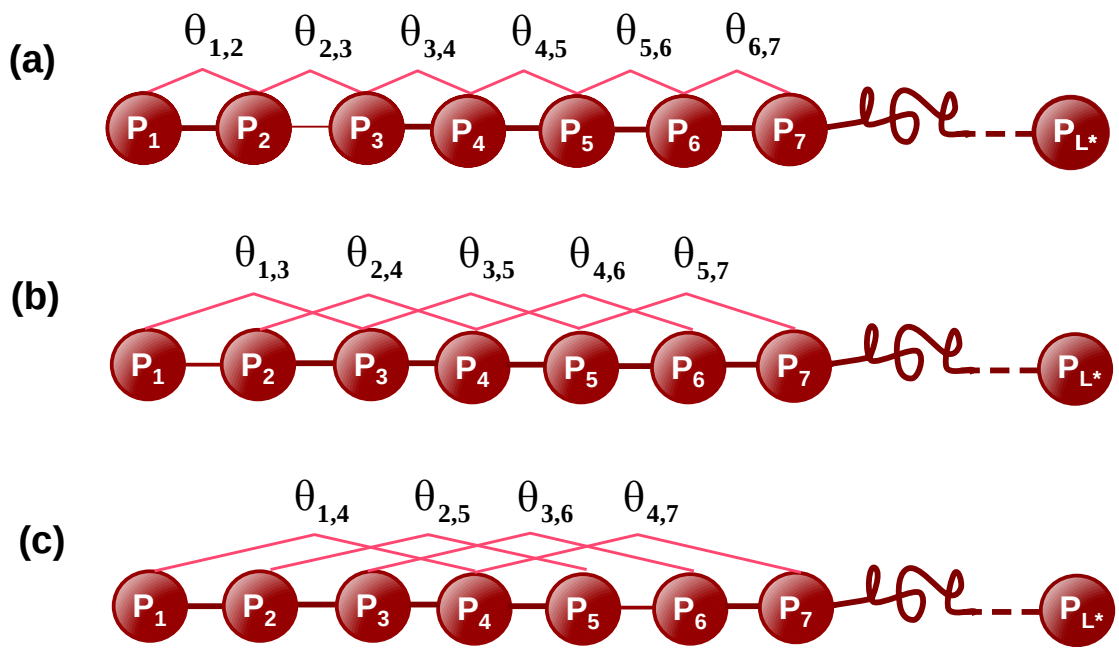


Figure 1

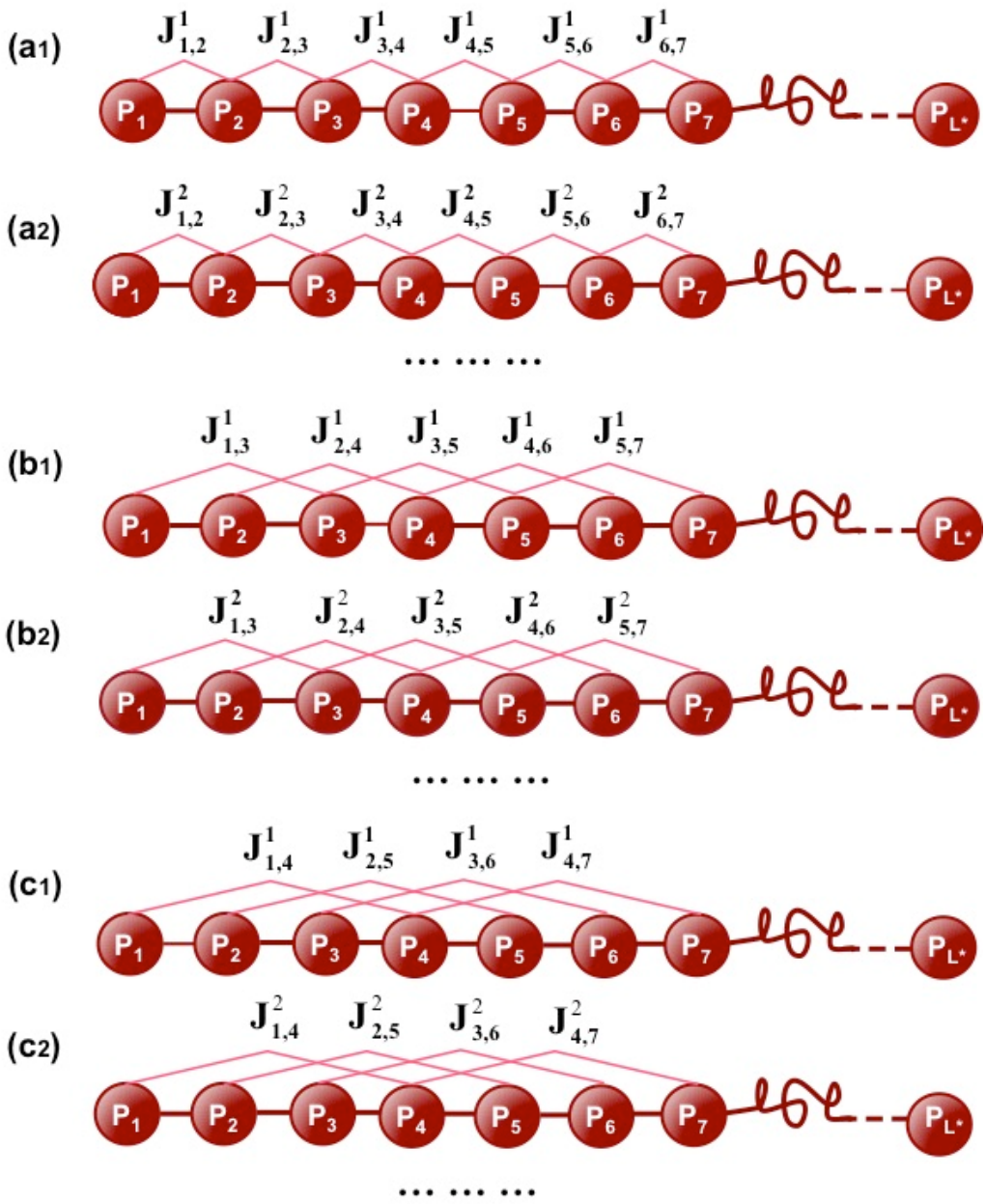


Figure 2

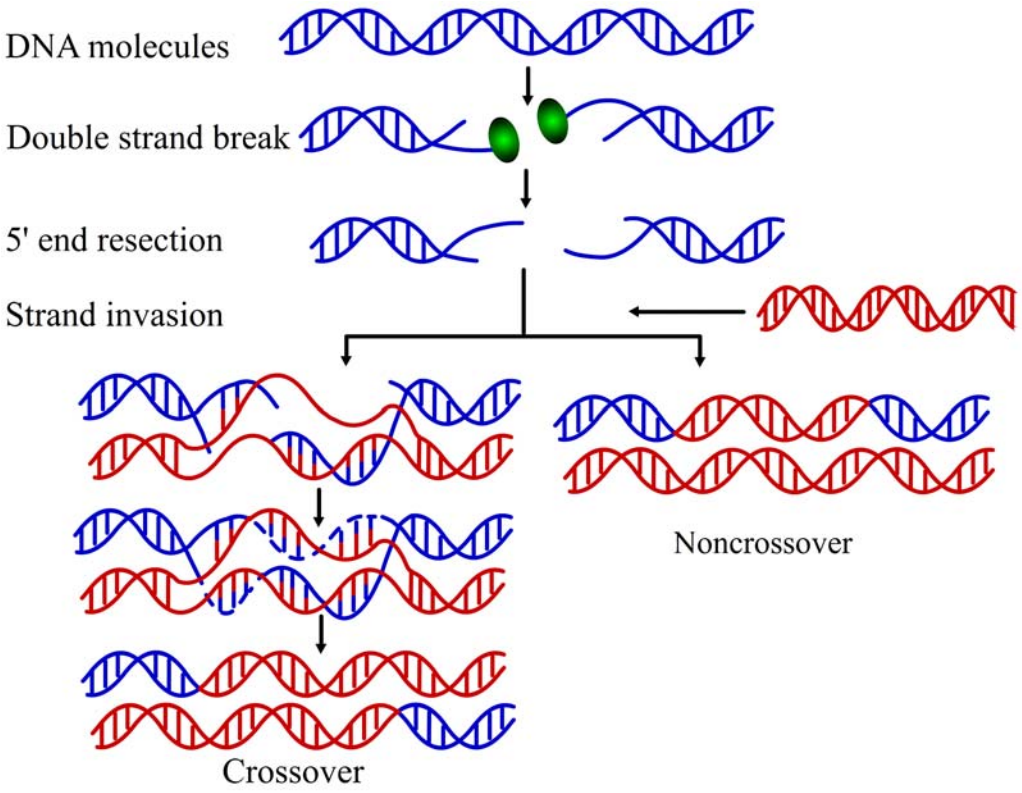


Figure 3

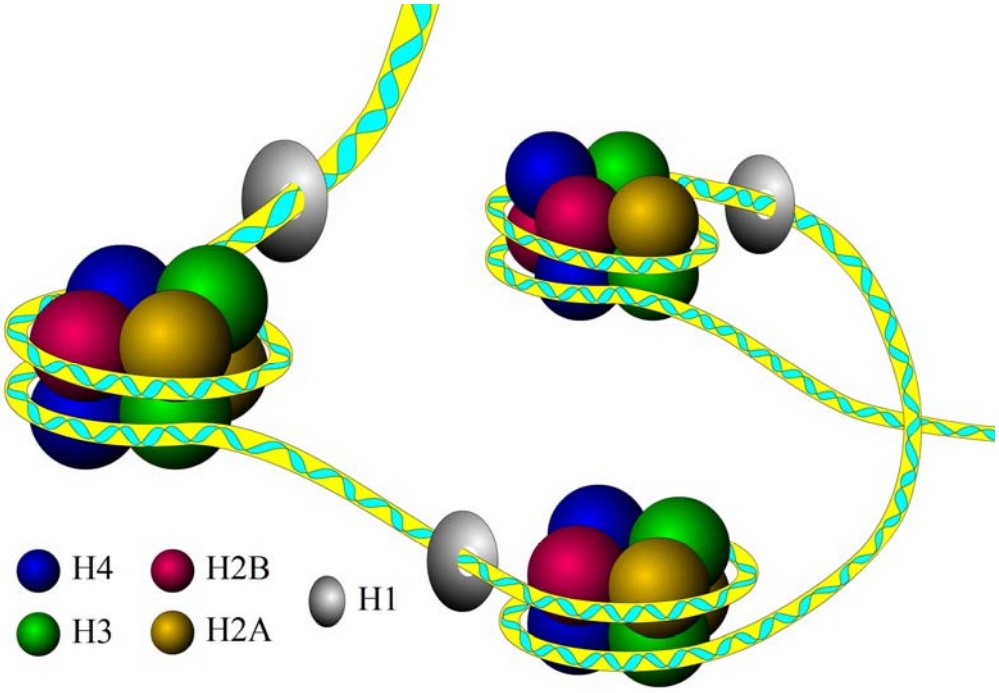


Figure 4

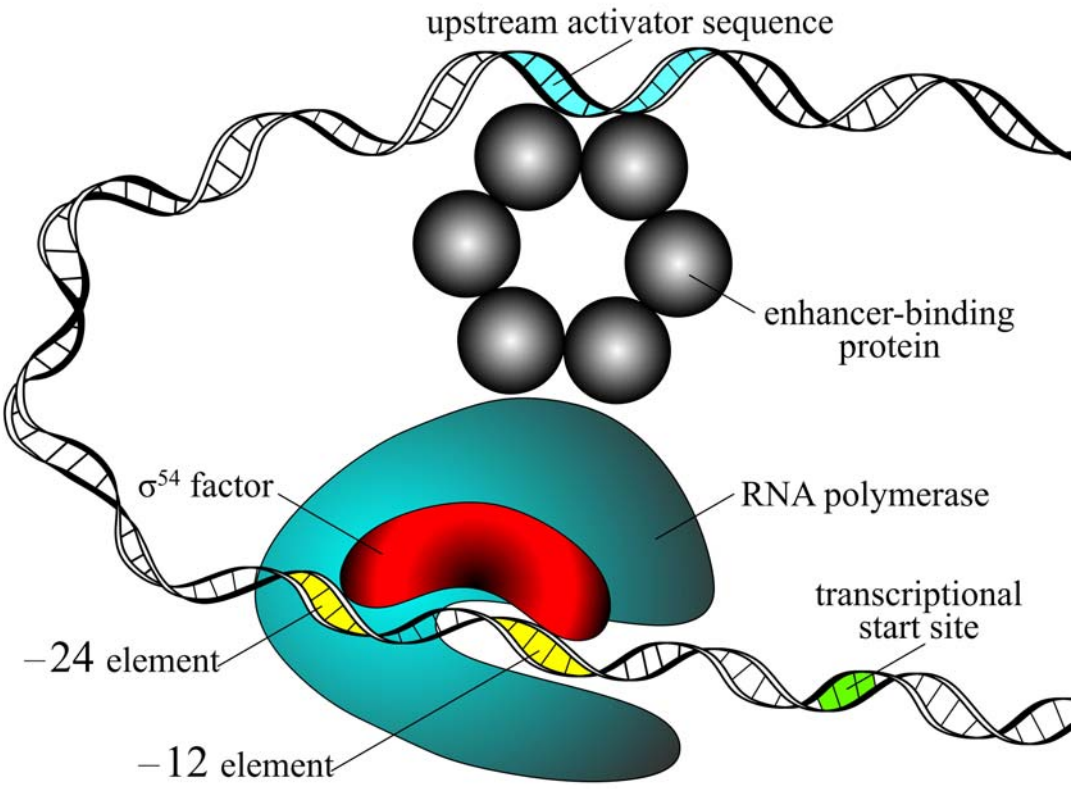


Figure 5

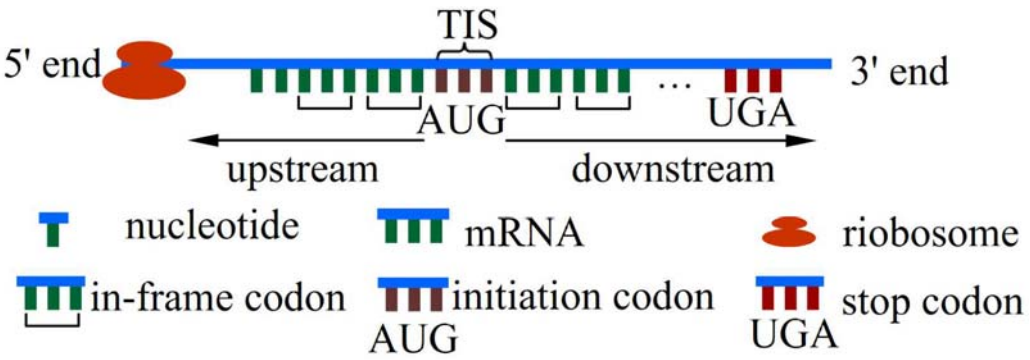


Figure 6

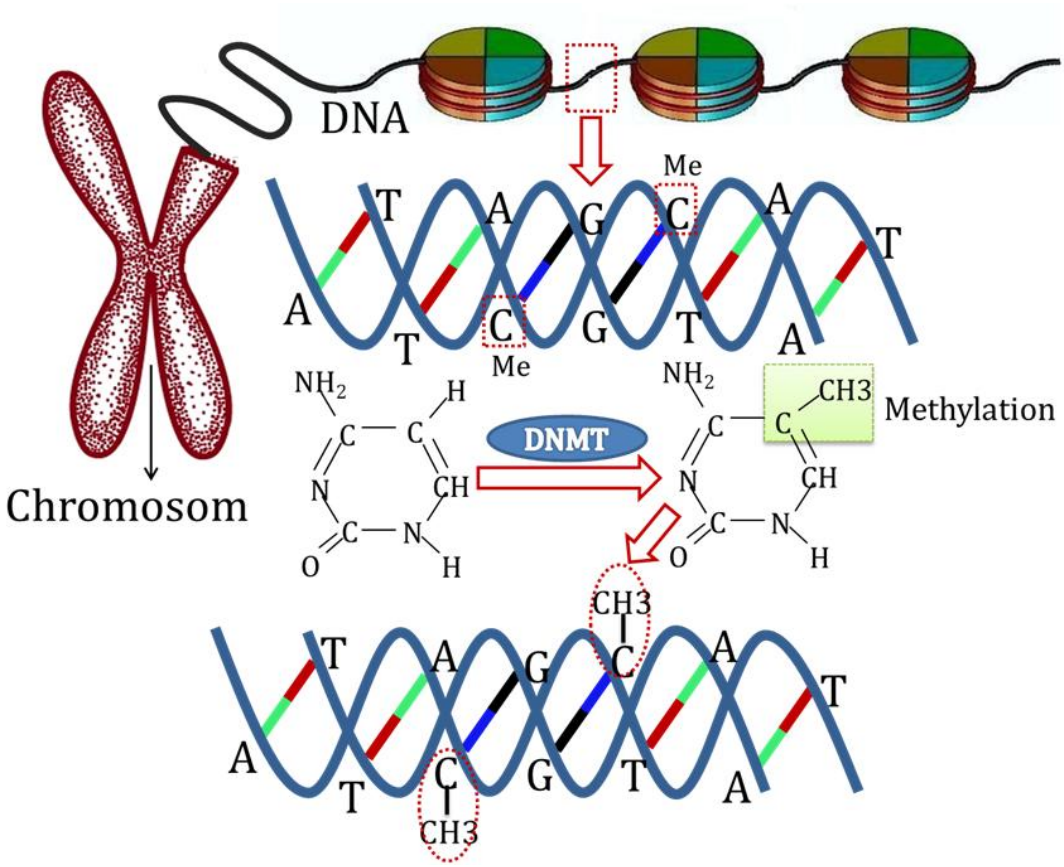


Figure 7

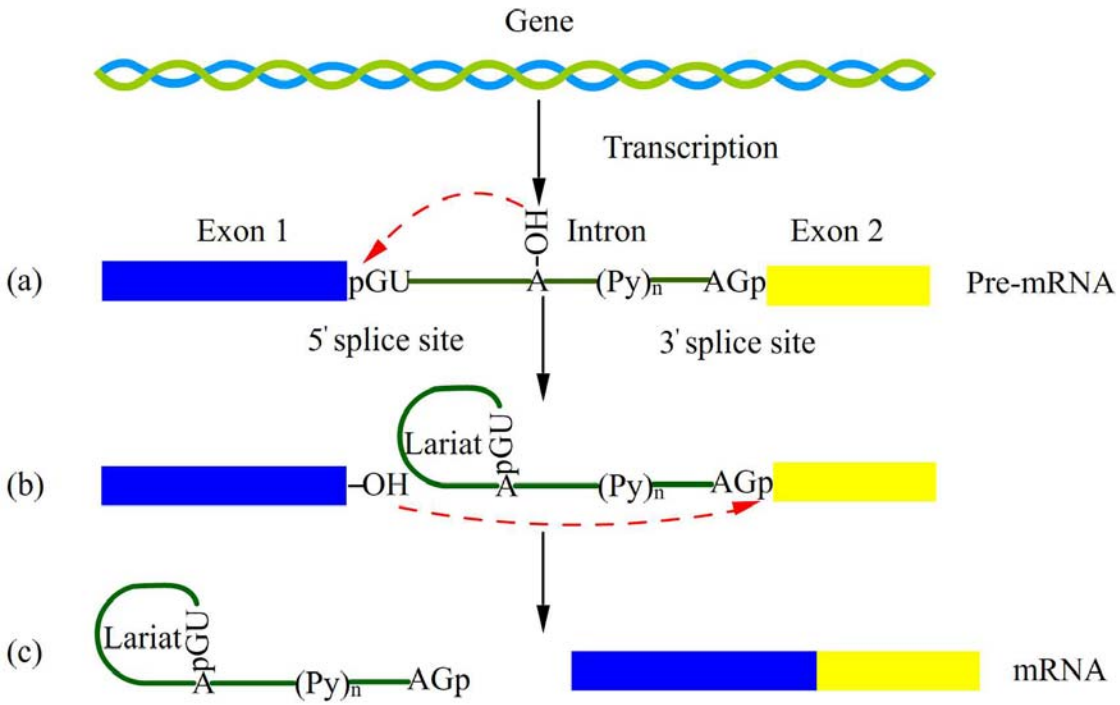


Figure 8

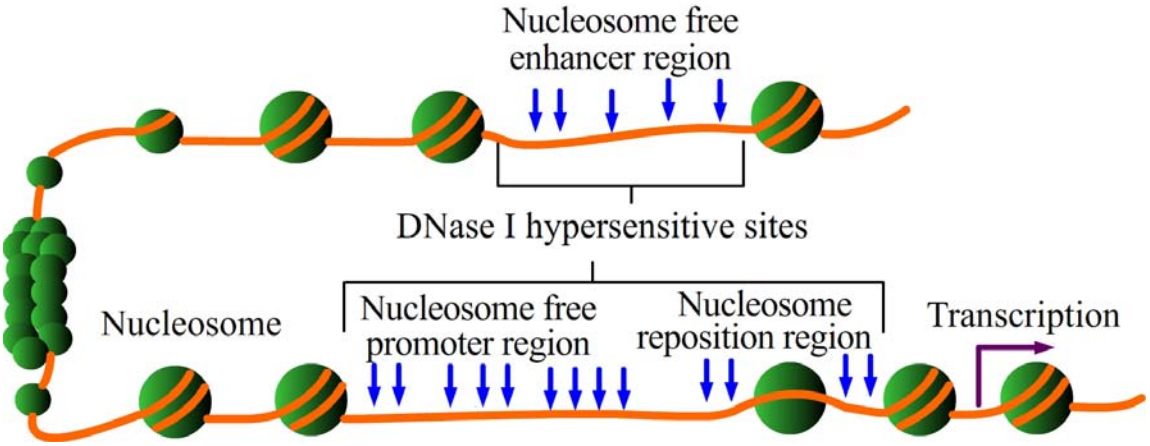


Figure 9

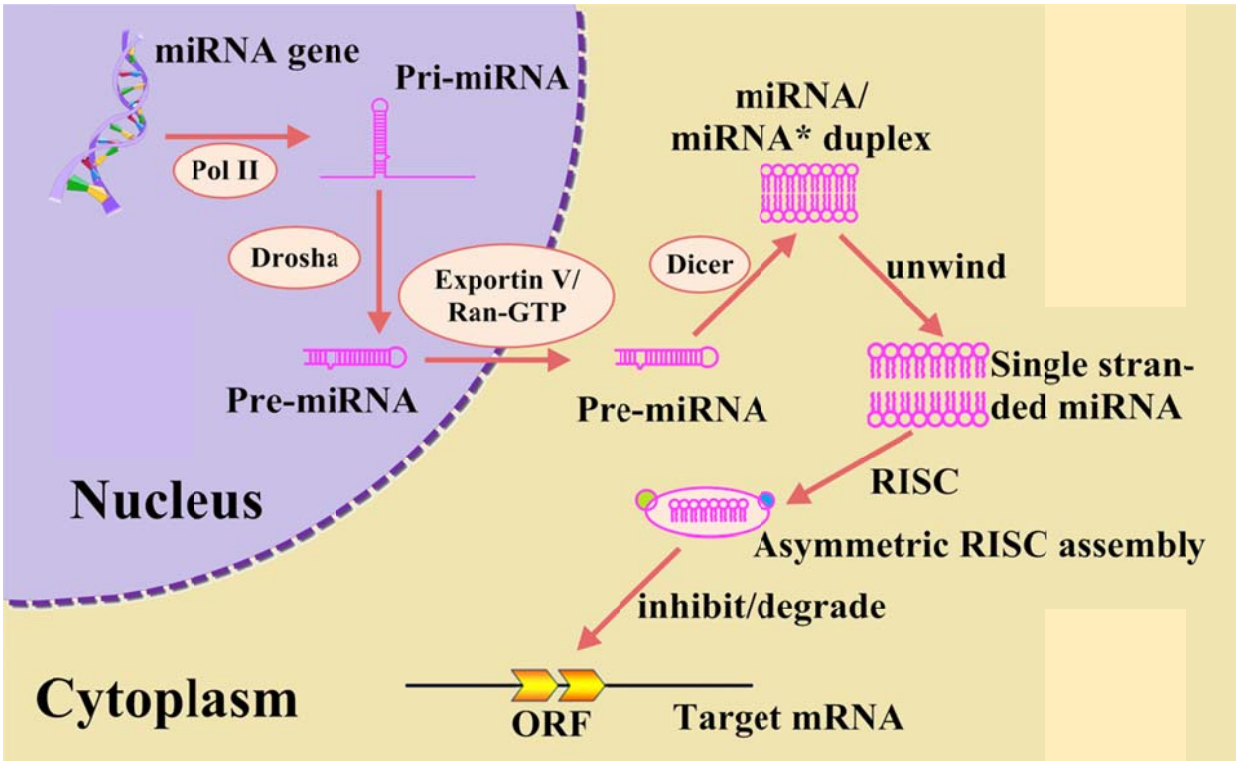


Figure 10

REFERENCES

1. A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. M. Claverie and O. Gascuel, *Nucleic acids research*, 2008, 36, W465-469.
2. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li and W. S. Noble, *Nucleic acids research*, 2009, 37, W202-208.
3. G. Reinert, D. Chew, F. Sun and M. S. Waterman, *Journal of computational biology : a journal of computational molecular cell biology*, 2009, 16, 1615-1634.
4. K. C. Chou, *Bioinformatics*, 2005, 21, 10-19.
5. K. C. Chou, *PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60)*, 2001, 43, 246-255.
6. H. B. Shen and K. C. Chou, *Anal. Biochem.*, 2008, 373, 386-388.
7. S. X. Lin and J. Lapointe, *J. Biomedical Science and Engineering (JBiSE)*, 2013, 6, 435-442.
8. Z. M. Guo, *Master Thesis, Bio-X Life Science Research Center, Shanghai Jiaotong University*, 2002.
9. K. C. Chou and Y. D. Cai, *Proteins: Struct., Funct., Genet.*, 2003, 53, 282-289.
10. K. C. Chou and Y. D. Cai, *Journal of Cellular Biochemistry (Addendum, ibid. 2004, 91, 1085)*, 2003, 90, 1250-1260.
11. Y. X. Pan, Z. Z. Zhang, Z. M. Guo, G. Y. Feng, Z. D. Huang and L. He, *J. Protein Chem.*, 2003, 22, 395-402.
12. K. C. Chou and Y. D. Cai, *J. Cell. Biochem.*, 2004, 91, 1197-1203.
13. M. Wang, J. Yang, G. P. Liu, Z. J. Xu and K. C. Chou, *Protein Engineering, Design, and Selection*, 2004, 17, 509-516.
14. Y. D. Cai and K. C. Chou, *Journal of Proteome Research*, 2005, 4, 967-971.
15. Y. Gao, S. H. Shao, X. Xiao, Y. S. Ding, Y. S. Huang, Z. D. Huang and K. C. Chou, *Amino Acids*, 2005, 28, 373-376.
16. H. Liu, J. Yang, M. Wang, L. Xue and K. C. Chou, *The Protein Journal*, 2005, 24, 385-389.
17. H. B. Shen and K. C. Chou, *Biochemical & Biophysical Research Communications (BBRC)*, 2005, 334, 288-292.
18. H. B. Shen and K. C. Chou, *Biochem Biophys Res Comm. (BBRC)*, 2005, 337, 752-756.
19. Y. D. Cai and K. C. Chou, *J. Theor. Biol.*, 2006, 238, 395-400.
20. S. Mondal, R. Bhavna, R. Mohan Babu and S. Ramakumar, *J. Theor. Biol.*, 2006, 243, 252-260.
21. H. B. Shen, J. Yang and K. C. Chou, *J. Theor. Biol.*, 2006, 240, 9-13.
22. S. Q. Wang, J. Yang and K. C. Chou, *J. Theor. Biol.*, 2006, 242, 941-946.
23. X. Xiao, S. H. Shao, Y. S. Ding, Z. D. Huang and K. C. Chou, *Amino Acids*, 2006, 30, 49-54.
24. X. Xiao, S. H. Shao, Z. D. Huang and K. C. Chou, *J. Comput. Chem.*, 2006, 27, 478-482.
25. S. W. Zhang, Q. Pan, H. C. Zhang, Z. C. Shao and J. Y. Shi, *Amino Acids*, 2006, 30, 461-468.

26. G. P. Zhou and Y. D. Cai, *PROTEINS: Structure, Function, and Bioinformatics*, 2006, 63, 681-684.
27. Y. L. Chen and Q. Z. Li, *J. Theor. Biol.*, 2007, 248, 377–381.
28. Y. S. Ding, T. L. Zhang and K. C. Chou, *Protein & Peptide Letters*, 2007, 14, 811-815.
29. H. Lin and Q. Z. Li, *Biochem. Biophys. Res. Commun.*, 2007, 354, 548-551.
30. H. Lin and Q. Z. Li, *Journal of Computational Chemistry*, 2007, 28, 1463-1466.
31. P. Mundra, M. Kumar, K. K. Kumar, V. K. Jayaraman and B. D. Kulkarni, *Pattern Recognition Letters*, 2007, 28, 1610-1615.
32. J. Y. Shi, S. W. Zhang, Q. Pan, Y.-M. Cheng and J. Xie, *Amino Acids*, 2007, 33, 69-74.
33. T. L. Zhang and Y. S. Ding, *Amino Acids*, 2007, 33, 623-629.
34. Y. Diao, D. Ma, Z. Wen, J. Yin, J. Xiang and M. Li, *Amino Acids*, 2008, 34, 111-117.
35. Y. S. Ding and T. L. Zhang, *Pattern Recognition Letters*, 2008, 29, 1887-1892.
36. Y. Fang, Y. Guo, Y. Feng and M. Li, *Amino Acids*, 2008, 34, 103-109.
37. X. Jiang, R. Wei, T. L. Zhang and Q. Gu, *Protein & Peptide Letters*, 2008, 15, 392-396.
38. X. Jiang, R. Wei, Y. Zhao and T. Zhang, *Amino Acids*, 2008, 34, 669-675.
39. F. M. Li and Q. Z. Li, *Amino Acids*, 2008, 34, 119-125.
40. F. M. Li and Q. Z. Li, *Protein & Peptide Letters*, 2008, 15, 612-616.
41. H. Lin, *J. Theor. Biol.*, 2008, 252, 350-356.
42. H. Lin, H. Ding, F. B. Feng-Biao Guo, A. Y. Zhang and J. Huang, *Protein & Peptide Letters*, 2008, 15, 739-744.
43. J. Y. Shi, S. W. Zhang, Q. Pan and G. P. Zhou, *Amino Acids*, 2008, 35, 321-327.
44. X. Xiao, W. Z. Lin and K. C. Chou, *J. Comput. Chem.*, 2008, 29, 2018-2024.
45. X. Xiao, P. Wang and K. C. Chou, *J. Theor. Biol.*, 2008, 254, 691-696.
46. G. Y. Zhang and B. S. Fang, *J. Theor. Biol.*, 2008, 253, 310-315.
47. S. W. Zhang, W. Chen, F. Yang and Q. Pan, *Amino Acids*, 2008, 35, 591-598.
48. S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao and Q. Pan, *Amino Acids*, 2008, 34, 565-572.
49. T. L. Zhang, Y. S. Ding and K. C. Chou, *J. Theor. Biol.*, 2008, 250, 186-193.
50. C. Chen, L. Chen, X. Zou and P. Cai, *Protein & Peptide Letters*, 2009, 16, 27-31.
51. K. C. Chou, *Current Proteomics*, 2009, 6, 262-274.
52. H. Ding, L. Luo and H. Lin, *Protein & Peptide Letters*, 2009, 16, 351-355.
53. Q. B. Gao, Z. C. Jin, X. F. Ye, C. Wu and J. He, *Anal. Biochem.*, 2009, 387, 54-59.
54. D. N. Georgiou, T. E. Karakasidis, J. J. Nieto and A. Torres, *J. Theor. Biol.*, 2009, 257, 17-26.
55. Z. C. Li, X. B. Zhou, Z. Dai and X. Y. Zou, *Amino Acids*, 2009, 37, 415-425.
56. H. Lin, H. Wang, H. Ding, Y. L. Chen and Q. Z. Li, *Acta Biotheoretica*, 2009, 57, 321-330.
57. J. D. Qiu, J. H. Huang, R. P. Liang and X. Q. Lu, *Anal. Biochem.*, 2009, 390, 68-73.
58. X. Xiao, P. Wang and K. C. Chou, *J. Appl. Crystallogr.*, 2009, 42, 169-173.
59. Y. H. Zeng, Y. Z. Guo, R. Q. Xiao, L. Yang, L. Z. Yu and M. L. Li, *J. Theor. Biol.*, 2009, 259, 366–372.
60. M. Esmaeili, H. Mohabatkar and S. Mohsenzadeh, *J. Theor. Biol.*, 2010, 263,

203-209.
61. Q. B. Gao, X. F. Ye, Z. C. Jin and J. He, *Anal. Biochem.*, 2010, 398, 52-59.
62. Q. Gu, Y. Ding, T. Zhang and Y. Shen, *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*, 2010, 27, 500-504.
63. Q. Gu, Y. S. Ding and T. L. Zhang, *Protein & Peptide Letters*, 2010, 17, 559-567.
64. K. K. Kandaswamy, G. Pugalenthi, S. Moller, E. Hartmann, K. U. Kalies, P. N. Suganthan and T. Martinetz, *Protein and Peptide Letters*, 2010, 17, 1473-1479.
65. T. Liu, X. Zheng, C. Wang and J. Wang, *Protein & Peptide Letters*, 2010, 17, 1263-1269.
66. H. Mohabatkar, *Protein & Peptide Letters*, 2010, 17, 1207-1214.
67. L. Nanni, S. Brahnam and A. Lumini, *J. Theor. Biol.*, 2010, 266, 1-10.
68. X. H. Niu, N. N. Li, F. Shi, X. H. Hu, J. B. Xia and H. J. Xiong, *Protein and Peptide Letters*, 2010, 17, 1466-1472.
69. J. D. Qiu, J. H. Huang, S. P. Shi and R. P. Liang, *Protein & Peptide Letters*, 2010, 17, 715-722.
70. S. S. Sahu and G. Panda, *Computational Biology and Chemistry*, 2010, 34, 320-327.
71. Y. C. Wang, X. B. Wang, Z. X. Yang and N. Y. Deng, *Protein & Peptide Letters*, 2010, 17, 1441-1449.
72. K. C. Chou, *J. Theor. Biol.*, 2011, 273, 236-247.
73. H. Ding, L. Liu, F. B. Guo, J. Huang and H. Lin, *Protein & Peptide Letters*, 2011, 18, 58-63.
74. J. Guo, N. Rao, G. Liu, Y. Yang and G. Wang, *Journal of Computational Chemistry*, 2011, 32, 1612-1617.
75. M. Hayat and A. Khan, *J. Theor. Biol.*, 2011, 271, 10-17.
76. L. Hu, L. Zheng, Z. Wang, B. Li and L. Liu, *Protein and Peptide Letters*, 2011, 18, 552-558.
77. X. Jingbo, Z. Silan, S. Feng, X. Huijuan, H. Xuehai, N. Xiaohui and L. Zhi, *Journal of Theoretical Biology*, 2011, 284, 16-23.
78. B. Liao, J. B. Jiang, Q. G. Zeng and W. Zhu, *Protein & Peptide Letters*, 2011, 18, 1086-1092.
79. H. Lin and H. Ding, *J. Theor. Biol.*, 2011, 269, 64-69.
80. J. Lin and Y. Wang, *Protein & Peptide Letters*, 2011, 18, 1219-1225.
81. J. Lin, Y. Wang and X. Xu, *African Journal of Biotechnology*, 2011, 10, 16963-16968.
82. X. L. Liu, J. L. Lu and X. H. Hu, *Protein & Peptide Letters*, 2011, 18, 1244-1250.
83. M. Mohammad Beigi, M. Behjati and H. Mohabatkar, *Journal of Structural and Functional Genomics*, 2011, 12, 191-197.
84. J. D. Qiu, S. B. Suo, X. Y. Sun, S. P. Shi and R. P. Liang, *Journal of Molecular Graphics & Modelling*, 2011, 30, 129-134.
85. R. Shi and C. Xu, *Protein and Peptide Letters*, 2011, 18, 625-633.
86. M. Shu, X. Cheng, Y. Zhang, Y. Wang, Y. Lin, L. Wang and Z. Lin, *Protein & Peptide Letters*, 2011, 18, 1233-1243.
87. D. Wang, L. Yang, Z. Fu and J. Xia, *Protein & Peptide Letters*, 2011, 18, 684-689.
88. W. Wang, X. B. Geng, Y. Dou, T. Liu and X. Zheng, *Protein and Peptide Letters*, 2011, 18, 480-487.

89. X. Xiao and K. C. Chou, *Current Bioinformatics*, 2011, 6, 251-260.
90. X. Xiao, P. Wang and K. C. Chou, *Molecular Biosystems*, 2011, 7, 911-919.
91. R. Zia Ur and A. Khan, *Protein & Peptide Letters*, 2011, 18, 872-878.
92. D. Zou, Z. He, J. He and Y. Xia, *J. Comput. Chem.*, 2011, 32, 271-278.
93. J. Z. Cao, W. Q. Liu and H. Gu, *Protein and Peptide Letters*, 2012, 19, 1163-1169.
94. C. Chen, Z. B. Shen and X. Y. Zou, *Protein & Peptide Letters*, 2012, 19, 422-429.
95. P. Du, X. Wang, C. Xu and Y. Gao, *Anal. Biochem.*, 2012, 425, 117-119.
96. G. L. Fan and Q. Z. Li, *J. Theor. Biol.*, 2012, 304, 88-95.
97. G. L. Fan and Q. Z. Li, *Amino Acids*, 2012, 43, 545-555.
98. M. Hayat and A. Khan, *Protein & Peptide Letters*, 2012, 19, 411-421.
99. L. Q. Li, Y. Zhang, L. Y. Zou, Y. Zhou and X. Q. Zheng, *Protein & Peptide Letters*, 2012, 19, 375-387.
100. B. Liao, Q. Xiang and D. Li, *Protein & Peptide Letters*, 2012, 19, 1133-1138.
101. W. Z. Lin, J. A. Fang, X. Xiao and K. C. Chou, *PLoS One*, 2012, 7, e49040.
102. L. Liu, X. Z. Hu, X. X. Liu, Y. Wang and S. B. Li, *Protein & Peptide Letters*, 2012, 19, 439-449.
103. S. Mei, *J. Theor. Biol.*, 2012, 293, 121-130.
104. S. Mei, *J. Theor. Biol.*, 2012, 310, 80-87.
105. L. Nanni, S. Brahnam and A. Lumini, *Amino Acids*, 2012, 43, 657-665.
106. L. Nanni, A. Lumini, D. Gupta and A. Garg, *IEEE/ACM Trans Comput Biol Bioinform*, 2012, 9, 467-475.
107. X. H. Niu, X. H. Hu, F. Shi and J. B. Xia, *Protein & Peptide Letters*, 2012, 19, 940-948.
108. Y. F. Qin, C. H. Wang, X. Q. Yu, J. Zhu, T. G. Liu and X. Q. Zheng, *Protein & Peptide Letters*, 2012, 19, 388-397.
109. L. Y. Ren, Y. S. Zhang and I. Gutman, *Protein & Peptide Letters*, 2012, 19, 1170-1176.
110. X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang and R. P. Liang, *Molecular BioSystems*, 2012, 8, 3178-3184.
111. J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang and X. Gao, *Comput Biol Med*, 2012, 42, 564-574.
112. X. Yu, X. Zheng, T. Liu, Y. Dou and J. Wang, *Amino Acids*, 2012, 42, 1619-1625.
113. X. W. Zhao, Z. Q. Ma and M. H. Yin, *Protein & Peptide Letters*, 2012, 19, 492-500.
114. R. Zia Ur and A. Khan, *Protein & Peptide Letters*, 2012, 19, 890-903.
115. D. S. Cao, Q. S. Xu and Y. Z. Liang, *Bioinformatics*, 2013, 29, 960-962.
116. T. H. Chang, L. C. Wu, T. Y. Lee, S. P. Chen, H. D. Huang and J. T. Horng, *Journal of Computer-Aided Molecular Design*, 2013, 27, 91-103.
117. Y. K. Chen and K. B. Li, *J. Theor. Biol.*, 2013, 318, 1-12.
118. G. L. Fan and Q. Z. Li, *J. Theor. Biol.*, 2013, 334, 45-51.
119. D. N. Georgiou, T. E. Karakasidis and A. C. Megaritis, *The Open Bioinformatics Journal*, 2013, 7, 41-48.
120. M. K. Gupta, R. Niyogi and M. Misra, *SAR QSAR Environ Res (SAR AND QSAR IN ENVIRONMENTAL RESEARCH)*, 2013, 24, 597-609.
121. C. Huang and J. Yuan, *Biosystems*, 2013, 113, 50-57.

122. C. Huang and J. Q. Yuan, *J. Theor. Biol.*, 2013, 335, 205-212.
123. H. Lin, C. Ding, L. F. Yuan, W. Chen, H. Ding, Z. Q. Li, F. B. Guo, J. Hung and N. N. Rao, *International Journal of Biomathematics*, 2013, 6, Article Number: 1350003.
124. B. Liu, X. Wang, Q. Zou, Q. Dong and Q. Chen, *Molecular Informatics*, 2013, 32, 775-782.
125. H. Mohabatkar, M. M. Beigi, K. Abdolahi and S. Mohsenzadeh, *Medicinal Chemistry*, 2013, 9, 133-137.
126. Y. F. Qin, L. Zheng and J. Huang, *Int. J. Quantum Chem.*, 2013, 113, 1660-1667.
127. A. N. Sarangi, M. Lohani and R. Aggarwal, *Protein Pept Lett*, 2013, 20, 781-795.
128. X. Wang, G. Z. Li and W. C. Lu, *Protein & Peptide Letters*, 2013, 20, 309-317.
129. X. Xiao, J. L. Min, P. Wang and K. C. Chou, *PLoS ONE*, 2013, 8, e72234.
130. X. Xiao, J. L. Min, P. Wang and K. C. Chou, *J. Theor. Biol.*, 2013, 337C, 71-79.
131. N. Xiaohui, L. Nana, X. Jingbo, C. Dingyan, P. Yuehua, X. Yang, W. Weiquan, W. Dongming and W. Zengzhen, *J. Theor. Biol.*, 2013, 332C, 211-217.
132. H. L. Xie, L. Fu and X. D. Nie, *Protein Eng Des Sel*, 2013, 26, 735-742.
133. Y. Xu, J. Ding, L. Y. Wu and K. C. Chou, *PLoS ONE*, 2013, 8, e55844.
134. Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng and K. C. Chou, *PeerJ*, 2013, 1, e171.
135. B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang and K. C. Chou, *PLoS ONE*, 2014, 9, e106691.
136. P. Du, S. Gu and Y. Jiao, *International Journal of Molecular Sciences*, 2014, 15, 3495-3506.
137. Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani and H. Mohabatkar, *J. Theor. Biol.*, 2014, 341, 34-40.
138. G. S. Han, Z. G. Yu and V. Anh, *J. Theor. Biol.*, 2014, 344, 31-39.
139. M. Hayat and N. Iqbal, *Computer methods and programs in biomedicine*, 2014, 116, 184-192.
140. C. Jia, X. Lin and Z. Wang, *Int J Mol Sci*, 2014, 15, 10410-10423.
141. L. Kong, L. Zhang and J. Lv, *J. Theor. Biol.*, 2014, 344, 12-18.
142. L. Li, S. Yu, W. Xiao, Y. Li, M. Li, L. Huang, X. Zheng, S. Zhou and H. Yang, *Biochimie*, 2014, 104, 100-107.
143. S. Mondal and P. P. Pai, *J. Theor. Biol.*, 2014, 356, 30-35.
144. L. Nanni, S. Brahnam and A. Lumini, *J. Theor. Biol.*, 2014, 360C, 109-116.
145. W. R. Qiu, X. Xiao and K. C. Chou, *Int J Mol Sci*, 2014, 15, 1746-1766.
146. W. R. Qiu, X. Xiao, W. Z. Lin and K. C. Chou, *Biomed Res Int*, 2014, 2014, 947416.
147. R. Xu, J. Zhou, B. Liu, Y. A. He, Q. Zou, X. Wang and K. C. Chou, *Journal of Biomolecular Structure & Dynamics (JBSD)*, 2014, doi: 10.1080/07391102.2014.968624.
148. Y. Xu, X. Wen, X. J. Shao, N. Y. Deng and K. C. Chou, *Int. J. Mol. Sci.*, 2014, 15, 7594-7610.
149. Y. Xu, X. Wen, L. S. Wen, L. Y. Wu, N. Y. Deng and K. C. Chou, *PLoS ONE*, 2014, 9, e105018.
150. J. Zhang, P. Sun, X. Zhao and Z. Ma, *J. Theor. Biol.*, 2014, 363, 412-418.
151. J. Zhang, X. Zhao, P. Sun and Z. Ma, *Int J Mol Sci*, 2014, 15, 11204-11219.
152. L. Zhang, X. Zhao and L. Kong, *J. Theor. Biol.*, 2014, 355, 105-110.

153. H. Ding, E. Z. Deng, L. F. Yuan, L. Liu, H. Lin, W. Chen and K. C. Chou, *BioMed Research International (BMRI)*, 2014, 2014, 286419.
154. A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal and A. Sattar, *J. Theor. Biol.*, 2015, 364, 284-294.
155. C. Huang and J. Q. Yuan, *Protein Pept Lett*, 2015.
156. J. Jia, Z. Liu, X. Xiao and K. C. Chou, *J. Theor. Biol.*, 2015, 377, 47-56.
157. Z. U. Khan, M. Hayat and M. A. Khan, *J. Theor. Biol.*, 2015, 365, 197-203.
158. B. Liu, J. Xu, S. Fan, R. Xu, J. Jiyun Zhou and X. Wang, *Molecular Informatics*, 2015, 34, 8-17
159. M. Mandal, A. Mukhopadhyay and U. Maulik, *Med. Biol. Eng. Comput*, 2015, 53, 331-344.
160. B. Liu, J. Chen and X. Wang, *Molecular genetics and genomics : MGG*, 2015, DOI: 10.1007/s00438-015-1044-4, doi:10.1007/s00438-00015-01044-00434.
161. W. Z. Zhong and S. F. Zhou, *Intenational Journal of Molecular Sciences*, 2014, 15, 20072-20078.
162. K. C. Chou, *Medicinal Chemistry*, 2015, 11, 218-234.
163. G. P. Zhou, *Medicinal Chemistry*, 2015, 11, 216-216.
164. X. Zhou, Z. Li, Z. Dai and X. Zou, *Talanta*, 2011, 85, 1143-1147.
165. W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Nucleic Acids Res.*, 2013, 41, e68.
166. X. Zhou, Z. Li, Z. Dai and X. Zou, *J. Theor. Biol.*, 2013, 319, 1-7.
167. H. Lin, E. Z. Deng, H. Ding, W. Chen and K. C. Chou, *Nucleic Acids Res.*, 2014, 42, 12961-12972.
168. W. Chen, P. M. Feng, E. Z. Deng, H. Lin and K. C. Chou, *Anal. Biochem.*, 2014, 462, 76-83.
169. W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Biomed Research International (BMRI)*, 2014, 2014, 623149.
170. P. Feng, W. Chen and H. Lin, *Genomics*, 2014, 104, 229-233.
171. P. Feng, N. Jiang and N. Liu, *The Scientific World Journal*, 2014, 2014, 740506.
172. W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo and K. C. Chou, *PLoS ONE*, 2012, 7, e47843.
173. S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen and K. C. Chou, *Bioinformatics*, 2014, 30, 1522-1529.
174. B. Liu, L. Fang, F. Liu, X. Wang and K. C. Chou, *Journal of Biomolecular Structure & Dynamics (JBSD)*, 2015, DOI: 10.1080/07391102.2015.1014422, http://dx.doi.org/10.1080/07391102.07392015.01014422.
175. B. Liu, L. Fang, F. Liu, X. Wang, J. Chen and K. C. Chou, *PLoS ONE*, 2015, 10, e0121501.
176. Z. Liu, X. Xiao, W. R. Qiu and K. C. Chou, *Anal. Biochem.*, 2015, 474, 69-77.
177. W. Chen, T. Y. Lei, D. C. Jin, H. Lin and K. C. Chou, *Anal. Biochem.*, 2014, 456, 53-60.
178. I. L. Hofacker, *Nucleic Acids Res.*, 2003, 31, 3429-3431.
179. B. Liu, F. Liu, L. Fang, X. Wang and K. C. Chou, *Molecular Genetics and Genomics*, 2015, in press.
180. B. Liu, F. Liu, X. Wang, J. Chen, L. Fang and K. C. Chou, *Nucleic Acids Res.*, 2015, doi:10.1093/nar/gkv458.
181. T. Wang, J. Yang, H. B. Shen and K. C. Chou, *Protein & Peptide Letters*, 2008, 15,

915-921.
182. G. Liu, J. Liu, X. Cui and L. Cai, *J. Theor. Biol.*, 2012, 293, 49-54.
183. T. J. Richmond and C. A. Davey, *Nature*, 2003, 423, 145-150.
184. Z. Zhang, Y. Zhang and I. Gutman, *J. Biomol. Struct. Dyn.*, 2012, 29, 1081-1088.
185. X. Zhao, Z. Pei, J. Liu, S. Qin and L. Cai, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 2010, 18, 777-785.
186. V. Rangannan and M. Bansal, *J. Biosci.*, 2007, 32, 851-862.
187. R. J. Jackson, C. U. Hellen and T. V. Pestova, *Nat Rev Mol Cell Biol*, 2010, 11, 113-127.
188. Y. Saeys, T. Abeel, S. Degroeve and Y. Van de Peer, *Bioinformatics*, 2007, 23, i418-423.
189. A. A. Hoskins and M. J. Moore, *Trends Biochem. Sci.*, 2012, 37, 179-188.
190. W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang and K. C. Chou, *Bioinformatics*, 2015, 31, 119-120.
191. B. Liu, F. Liu, L. Fang, X. Wang and K. C. Chou, *Bioinformatics*, 2015, 31, 1307-1309.
192. J. Chen, H. Liu, J. Yang and K. C. Chou, *Amino Acids*, 2007, 33, 423-428.
193. W. R. Qiu, X. Xiao, W. Z. Lin and K. C. Chou, *Journal of Biomolecular Structure and Dynamics (JBSD)* 2014, doi:10.1080/07391102.2014.968875.
194. K. C. Chou and H. B. Shen, *Anal. Biochem.*, 2007, 370, 1-16.
195. Z. C. Wu, X. Xiao and K. C. Chou, *Molecular BioSystems*, 2011, 7, 3287-3297.
196. K. C. Chou, Z. C. Wu and X. Xiao, *Molecular Biosystems*, 2012, 8, 629-641.
197. W. Z. Lin, J. A. Fang, X. Xiao and K. C. Chou, *Molecular BioSystems*, 2013, 9, 634-644.
198. X. Xiao, P. Wang, W. Z. Lin, J. H. Jia and K. C. Chou, *Anal. Biochem.*, 2013, 436, 168-177.
199. L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng and K. C. Chou, *PLoS ONE*, 2012, 7, e35254.
200. K. C. Chou and H. B. Shen, *Nature Protocols*, 2008, 3, 153-162.
201. K. C. Chou and H. B. Shen, *Natural Science*, 2010, 2, 1090-1103.
202. K. C. Chou, *Molecular Biosystems*, 2013, 9, 1092-1100.
203. Y. Sun, A. K. Wong and M. S. Kamel, *International Journal of Pattern Recognition and Artificial Intelligence*, 2009, 23, 687-719.
204. X. Xiao, J. L. Min, W. Z. Lin, Z. Liu, X. Cheng and K. C. Chou, *Journal of Biomolecular Structure & Dynamics (JBSD)* 2014, doi: 10.1080/07391102.07392014.07998710.
205. C. T. Zhang and K. C. Chou, *J. Protein Chem.*, 1995, 14, 583-593.
206. C. T. Zhang and K. C. Chou, *Biophys. J.*, 1992, 63, 1523-1529.
207. K. C. Chou, *J. Biol. Chem.*, 1993, 268, 16938-16948.
208. K. C. Chou and H. B. Shen, *Natural Science*, 2009, 1, 63-92
209. J. L. Min, X. Xiao and K. C. Chou, *BioMed Research International (BMRI)*, 2013, 2013, 701317.
210. Y. N. Fan, X. Xiao, J. L. Min and K. C. Chou, *Intenational Journal of Molecular Sciences*, 2014, 15, 4915-4937.