Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Natural/random protein classification models based on star network topological indices

Cristian Robert Munteanu a, Humberto González-Díaz b,*, Fernanda Borges c, Alexandre Lopes de Magalhães a

a REQUIMTE-University of Porto, Faculty of Science, Chemistry Department, University of Porto 4169-007, Portugal
b Unit of Bioinformatics and Connectivity Analysis (UBICA), Institute of Industrial Pharmacy, Faculty of Pharmacy, University of Santiago de Compostela, Campus Universitario Sur, 15782 Santiago de Compostela, Spain
c Faculty of Pharmacy (FFUP), Organic-Chemistry Department, University of Porto 4050-047, Portugal

Article info
Article history:
Received 7 May 2008
Received in revised form 9 July 2008
Accepted 15 July 2008
Available online 22 July 2008

Keywords:
Protein structure
Graph theory
Random proteins
Python applications
GDA
S2SNet

Abstract
The development of the complex network graphs permits us to describe any real system such as social, neural, computer or genetic networks by transforming real properties in topological indices (TIs). This work uses Randic's star networks in order to convert the protein primary structure data in specific topological indices that are used to construct a natural/random protein classification model.

The set of natural proteins contains 1046 protein chains selected from the pre-compiled CulledPDB list from PISCES Dunbrack’s Web Lab. This set is characterized by a protein homology of 20%, a structure resolution of 1.6 Å and R-factor lower than 25%. The set of random amino acid chains contains 1046 sequences which were generated by Python script according to the same type of residues and average chain length found in the natural set.

A new Sequence to Star Networks (S2SNet) wxPython GUI application (with a Graphviz graphics back-end) was designed by our group in order to transform any character sequence in the following star network topological indices: Shannon entropy of Markov matrices, trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau–Broto indices, Balaban distance connectivity index, Kier–Hall connectivity indices and Randic connectivity index. The model was constructed with the General Discriminant Analysis methods from STATISTICA package and gave training/predicting set accuracies of 90.77% for the forward stepwise model type.

In conclusion, this study extends for the first time the classical TIs to protein star network TIs by proposing a model that can predict if a protein/fragment of protein is natural or random using only the amino acid sequence data. This classification can be used in the studies of the protein functions by changing some fragments with random amino acid sequences or to detect the fake amino acid sequences or the errors in proteins. These results promote the use of the S2SNet application not only for protein structure analysis but also for mass spectroscopy, clinical proteomics and imaging, or DNA/RNA structure analysis.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

One of the widely used methods for the predicting of the protein properties is quantitative structure activity relationship (QSAR) (Devillers and Balaban, 1999). Graph theory can be used to obtain macromolecular descriptors named topological indices (TIs). The branch of mathematical chemistry dedicated to encode the DNA/protein information in graph representations by the use of the TIs has become an intense research area with interesting works of Liao (Liao and Wang, 2004a, b; Liao and Ding, 2005; Liao et al., 2006), Randic, Nandy, Balaban, Basak, and Vracko (Randic, 2000; Randic et al., 2000; Randic and Basak, 2001; Randic and Balaban, 2003), Bielinska-Waz team (Bielinska-Waz et al., 2007) or our group (Perez et al., 2004; Aguero-Chapin et al., 2006). Using graphic approaches to study biological systems can provide useful insights, as indicated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Andraos, 2008; Chou, 1989; Chou and Forsen, 1980, 1981; Chou and Liu, 1981; Chou et al., 1983; King and Altman, 1956; Kuzmic et al., 1992; Myers and Palmer, 1985; Zhou and Deng, 1984), protein folding kinetics (Chou, 1990), inhibition kinetics of processive nucleic acid polymerases and nuclease (Althaus et al.,
1993a, b, c, 1994a, b, 1996; Chou et al., 1994), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1993, 1994), base frequencies in the anti-sense strands (Chou et al., 1996), and analysis of DNA sequence (Qi et al., 2007). Moreover, graphical methods have been introduced for QSAR study (Gonzalez-Diaz et al., 2006, 2007b; Prado-Prado et al., 2008) as well as utilized to deal with complicated network systems (Diao et al., 2007; Gonzalez-Diaz et al., 2007a, 2008). Recently, the “cellular automaton image” (Wolfram, 1984, 2002) has also been applied to study hepatitis B viral infections (Xiao et al., 2006a), HBV virus gene missense mutation (Xiao et al., 2005b), and visual analysis of SARS-CoV (Gao et al., 2006; Wang et al., 2005), as well as representing complicated biological sequences (Xiao et al., 2005a) and helping to identify protein attributes (Xiao and Chou, 2007; Xiao et al., 2006b).

The actual work presents for the first time a natural/random protein classification using only the chain sequence and amino acid connectivity protein structural data. The data are transformed into sequence and connectivity Star Graph’s TIs, which are then used as input for a statistical linear method in the construction of a simple classification model.

2. Materials and methods

2.1. Protein set

Two sets of proteins are compared in the new classification model: a set (Nat) of 1046 natural protein chains as defined in the pre-compiled CulledPDB list from PISCES Dunbrack’s Web Lab (Wang and Dunbrack, 2003) and a second (Rnd) with the same size formed by random amino acid sequences generated with Python scripts (Rossum, 2006). The natural set is characterized by a homology of 20%, a structure resolution of 1.6 Å and R-factor lower than 25%. The random set is composed by the same standard amino acid types and the average length of the chains is the same as that of the natural set. Python scripts are used to download PDB files from the PDB data bank (Berman et al., 2000) and to create the correspondent DSSP file with the DSSP application (Kabsch and Sander, 1983). The chain sequences were extracted with a Python script from these DSSP files and were filtered with our Prot-2S Web Tool (http://www.requimte.pt:8080/Prot-2S/) by removing the chains that contain non-standard amino acid (usually labelled X).

2.2. Star graph

Each protein can be considered as a real network where the amino acids are the vertices (nodes), connected in a specific sequence by the peptide bonds. The graph is the abstract representation of the network and is a collection of N vertices and the connections between them. The star graph is a special case of trees with N vertices where one has got N−1 degrees of freedom and the remaining N−1 vertices have got one single degree of freedom (Harary, 1969). In addition, as a general property, there is a unique path between any pair of vertices. For proteins, each of the 20 possible branches (“rays”) of the star contains the same amino acid type and the star centre is a non-amino acid vertex.

The same protein can be represented by different forms which are associated to distinct distance matrices (Randic et al., 2007). If the vertices do not carry a label, the sequence information will be lost; for that reason, the best method is to construct a standard star graph where each amino acid/vertex holds the position in the original sequence and the branches are labelled by alphabetical order of the three-letter amino acid code (Randic et al., 2007).

In the present study we are using the alphabetical order of one-letter amino acid code. The standard star graph for a random virtual decapeptide (ACADCEFDGH) is illustrated in Fig. 1. If the initial connectivity in the protein chain is included, the graph is embedded (Fig. 2). In order to compare the graphs, it is necessary to transform the graphical representation in connectivity matrix, distance matrix and degree matrix. In the case of the embedded graph, the matrices of the connectivity in the sequence and in the star graph are combined. These matrices and the normalized ones are the base for the TIs calculation.
2.3. TIs for star graph

The protein chain sequences are transformed into Star Graph representations and then characterized by several TIs using our new Sequence to Star Networks (S2SNet) application. S2SNet is a wxPython (Noel Rappin, 2006) GUI application with Graphviz (Koutsofios, 1993) as a graphics back-end. The user of this interactive tool is able to choose the level of calculations, such as: embedded graph, additional weights for each amino acid, Markov normalization, power of the matrix connectivity, the input files (files with sequences, groups and weights), the output files, the level of details (files for summary and detailed results) and the type of graph visualization (dot, neato, fdp, twopi, circo). In particular, the calculations presented in this work are characterized by embedded and non-embedded TIs, no weights, Markov normalization and power of matrices/indices \((n)\) up to 5. The summary file contains the following TIs (Todeschini and Consonni, 2002):

- Shannon entropy of the \(n\) powered Markov matrices \((S_{hn})\):
  \[
  S_{hn} = \sum_{i} p_i \log(p_i),
  \]
  (1)

  where \(p_i\) are the \(n\) elements of the \(p\) vector, resulted from the matrix multiplication of the powered Markov normalized matrix \((n_i \times n_i)\) and a vector \((n_i \times 1)\) with each element equal to \(1/n_i\);

- The trace of the \(n\) connectivity matrices \((Tr_n)\):
  \[
  Tr_n = \sum_{i} (M^n)_{ii},
  \]
  (2)
  where \(n = 0–\text{power limit}\), \(M = \text{connectivity matrix (}i \times i\) dimension\); \(ii = \text{ith diagonal element}\);

- Harary number \((H)\):
  \[
  H = \sum_{i<j} (m_{ij}/d_{ij})w_{ij}^{nw},
  \]
  (3)
  where \(d_{ij}\) are the elements of the distance matrix, \(m_{ij}\) are the elements of the \(M\) connectivity matrix, \(w_{ij}\) are the weight elements and \(nw\) is a switch to select \((1)\) or not select \((0)\) weights calculations;

- Wiener index \((W)\):
  \[
  W = \sum_{i<j} d_{ij} w_{ij}^{nw},
  \]
  (4)

Table 1

| Model  | Star Graph type | Train | Cross-validation | Total |
|--------|-----------------|-------|-----------------|-------|
|        | % Nat | % Rnd | % Total | % Nat | % Rnd | % Total | % Nat | % Rnd | % Total |
| Forward | nE    | 86.50 | 96.17 | 91.33 | 83.52 | 96.95 | 90.25 | 85.76 | 96.37 | 91.06 |
|         | E     | 80.00 | 88.65 | 84.32 | 78.54 | 90.08 | 84.32 | 79.64 | 89.01 | 84.32 |
|         | nE and E | 85.86 | 96.17 | 91.01 | 81.99 | 98.09 | 90.06 | 84.89 | 96.65 | 90.77 |
| Backward | nE    | 86.11 | 96.68 | 91.40 | 83.52 | 98.09 | 90.82 | 85.47 | 97.04 | 91.25 |
|         | E     | 81.27 | 90.82 | 86.04 | 79.69 | 92.75 | 86.23 | 80.88 | 91.30 | 86.09 |
|         | nE and E | 86.75 | 97.19 | 91.97 | 84.67 | 98.47 | 91.59 | 86.23 | 97.51 | 91.87 |
| Best    | nE    | 86.75 | 96.68 | 91.71 | 81.52 | 98.09 | 90.82 | 85.95 | 97.04 | 91.49 |
|         | E     | 81.40 | 90.05 | 85.72 | 79.31 | 91.60 | 85.47 | 80.88 | 90.44 | 85.66 |

Fig. 3. Training cases against the residuals for the full set.
- Gutman topological index ($S_0$):

$$S_0 = \sum_{ij} \text{deg}_i \text{deg}_j w_{ij}^w / d_{ij},$$

(5)

where $\text{deg}_i$ are the elements of the degree matrix;

- Schultz topological index (non-trivial part) ($S$):

$$S = \sum_{i<j} (\text{deg}_i + \text{deg}_j) d_{ij} w_{ij}^w,$$

(6)

- Moreau-Broto, autocorrelation of topological structure ($ATS_n$, $n = 1$–power limit), only with weights included:

$$ATS_n = \sum_{ij} d_{ij}^n w_{ij},$$

(7)

where $d_{ij}^n$ are the elements of the pair distance matrix when the distance is $n$;

- Balaban distance connectivity index ($J$):

$$J = (\text{edges} - \text{nodes} + 2) \sum_{i<j} m_{ij} \sqrt{\sum_k d_{ik} \sum_k d_{kj} w_{ij}^w},$$

(8)

where $\text{nodes} + 1 = \text{AA} \text{ numbers/node number in the Star Graph+origin}$, $\sum d_{ik}$ is the node distance degree;

- Kier–Hall connectivity indices ($^X$):

$$^0X = \sum_{ij} w_{ij}^w / \sqrt{\text{deg}_i},$$

(9)

$$^2X = \sum_{i<j<k} m_{ij} m_{jk} w_{ik}^w / \sqrt{\text{deg}_i \text{deg}_j \text{deg}_k},$$

(10)

$$^3X = \sum_{i<j<k<m} m_{ij} m_{jk} m_{km} w_{im}^w / \sqrt{\text{deg}_i \text{deg}_j \text{deg}_k \text{deg}_m},$$

(11)

$$^4X = \sum_{i<j<k<m<o} m_{ij} m_{jk} m_{km} m_{mo} w_{io}^w / \sqrt{\text{deg}_i \text{deg}_j \text{deg}_k \text{deg}_m \text{deg}_o},$$

(12)

$$^5X = \sum_{i<j<k<m<o<q} m_{ij} m_{jk} m_{km} m_{mo} m_{oq} w_{io}^w / \sqrt{\text{deg}_i \text{deg}_j \text{deg}_k \text{deg}_m \text{deg}_o \text{deg}_q},$$

(13)

- Randic connectivity index ($^1X$):

$$^1X = \sum_{ij} w_{ij}^w / \sqrt{\text{deg}_i \text{deg}_j},$$

(14)

All these TIs will be used to construct a natural/random classification model by statistical methods.

2.4. Statistical analysis

General discriminant analysis (GDA) (Kowalski and Wold, 1982; Van Waterbeemd, 1995) from STATISTICA 6.0 package.
(StatSoft, Inc., 2002) has been chosen as the simplest and fastest method. In order to decide if a protein chain is classified as natural (if exists in the PDB database) or random, we added an extra dummy variable named Nat/Rnd (binary values of 0/1) and a cross-validation variable (CV). There are three often used cross-validation methods to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang, 1995). Through a crystal-clear analysis, Chou and Shen (2007, 2008) have shown that only the jackknife test has the least arbitrariness. Therefore, the jackknife test has been increasingly used by investigators to examine the accuracy of various predictors (Chen and Li, 2007a, b; Diao et al., 2007; Ding et al., 2007; Jiang et al., 2008; Jin et al., 2008; Li and Li, 2008; Lin, 2008; Lin et al., 2006, 2008; Wang et al., 2008; Xiao and Chou, 2007; Zhou et al., 2007; Zhang et al., 2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008). In the actual work, the independent data test is used by splitting the data at random in a training series (2008).

Using S2SNet methodology, as defined previously we can attempt to develop a simple linear QSAR, with the general formula

\[ \text{Nat/Rnd - score} = c_0 + \sum_{i=1}^{n} c_i T_i, \]

where \( c_0 \) is the continue score value for the \( \text{Nat/Rnd} \) classification, \( T_i \) are TIs described above, \( c_i - c_n \) are TIs coefficients, \( n \) is the number for the indices and \( c_0 \) is the independent term.

GDA models quality was determined by examining Wilk’s U statistics, Fisher ratio (F), p-level (p), and canonical regression coefficient (\( R_c \)). We also inspected the percentage of good classification, cases/variables ratios, and number of variables to be explored in order to avoid over-fitting or chance correlation. The forward, backward and best subset model types are tested for the embedded, non-embedded and both data.

3. Results

Eight variable selection methods were applied in order to find the best GDA equation which is able to discriminate between natural and random chain proteins. Eight models were constructed using embedded/non-embedded Star Graph TIs obtained with S2SNet application and forward, backward and best subset model types. The values obtained for the training/predicting accuracies are presented in Table 1.

The forward stepwise selection variable method conjugated with the \( \text{nE} \) and \( \text{ETI} \) provides the best results for our data set with values of correctly classified compounds of 91.01%, 90.06% and 90.77% for the training, cross-validation and full sets, respectively, and using a minimum number of 12 parameters (Eq. (15)). The embedded TIs have the name of the non-embedded ones plus “e” as suffix:

\[ \text{Nat/Rnd - score} = 0.1 + 4.85h0 + 254.9H + 1860.2W \]
\[ - 1931.05 + 39.4J - 139.2X0 \]
\[ - 73.0X3 + 146.7X4 - 159.3X5 \]
\[ - 6.6T4e + 7.1X2e, \]

\[ N = 2092, \quad R_c = 0.79, \quad U = 0.38, \quad F = 228.58, \quad p < 0.001, \]

where \( N \) is the number of studied protein sequences (\( \text{Nat+Rnd} \)), \( R_c \) is the canonical regression coefficient, \( U \) is the Wilk’s statistics, \( F \) is the Fisher’s statistics and \( p \) is the p-level (probability of error).

The present \( R_c \) value shows a high level of correlation between the input variables and the classification of proteins. Wilk’s \( U \) is used to measure the statistical significance of the discriminatory power of the model and has values from 1.0 (no discriminatory power) to 0.0 (perfect discriminatory power). The \( F \) value shows the statistical significance in the discrimination between groups, a measure of the extent to which a variable makes a unique contribution to a prediction of group membership. The values of the \( p \)-level of Fisher’s test for the GDA is less than 0.05 and show that the hypothesis of group overlapping with a 5% error can be rejected (Hua and Sun, 2001). The above results are typically

![Fig. 6. Distribution for GDA model residuals, chi-square and Kolmogorov-Smirnov tests.](image-url)
considered as excellent in the literature for LDA-QSAR models (Garcia-Garcia et al., 2004; Marrero-Ponce et al., 2004, 2005).

The parametrical assumptions such as normality, homoscedasticity (homogeneity of variances) and non-collinearity have the same importance in the application of multivariate statistic techniques to QSPR (Bisquerra Alzina, 1989; Stewart, 1998) as the correct specification of the mathematical form has. The validity and statistical significance of any model is conditioned by the above-mentioned factors.

In our study, a simple linear mathematical form of the model has been chosen in the absence of prior information. Figs. 3 and 4 show that the training cases against the residuals did not
present any characteristic pattern (Dillon and Goldstein, 1984). The protein nos. 632 and 864 are the only two cases not shown in Fig. 4 because the corresponding raw residuals are clear distinct from the whole set, ca -7. They correspond to 1QWN, chain A (1014 AAs) and 1JZ8, chain A (1011 AAs). One possible reason for the apparent different statistical behaviour could be the limitation of the model when the length of the chains is greater than 1000 amino acids. It is possible that the star net TIs for large proteins become similar to the TIs of the random proteins.
A different and better threshold for the a priori classification probability can be estimated by means of the receiver operating characteristics (ROC) curve (James and Hanley, 1982). As the Fig. 5 clearly shows, one can see that the model is not a random, but a truly statistically significant classifier, since the area under the ROC curve (for both training = 0.98 and validation = 0.96) is significantly higher than the area under the random classifier curve random = 0.5 = diagonal line (Morales Helguera et al., 2007).

The validity of the GDA models depends on the normal distribution of the sample used as well as the homogeneity of their variances. Thus, we carried out two significant tests for normality, chi-square and Kolmogorov–Smirnov tests, and we have found significant statistical differences (p < 0.01) on the respective values (chi-square, d). These results allow us to reject the hypothesis of normal distribution of the sample under study (Fig. 6) (Stewart, 1998).

The heteroscedasticity of a large set can be detected with the simple graphical method based on the examination of the residuals of the variable included in the model. Fig. 7(a and b) shows that the NAG/Rnd GDA model variables against the residuals plots do not present any pattern, which indicates that homoscedasticity assumption is fulfilled (Stewart, 1998).

Due to the robustness of the GDA multivariate statistical techniques, the predictive ability and interference reached by using the proposed model should not be affected (see Fig. 8).

4. Discussion

This study extends for the first time the classical TIs to protein Star Network TIs by proposing a model that can predict if a chain protein is natural or random. The results prove for the first time the excellent predictive ability (90.77%) of the simple and fast Star Network TIs and GDA statistics linear models in the case of natural/random protein model. This classification can help the study of the protein function by changing some fragments with random amino acid sequences or can detect the fake amino acid sequences or the errors in proteins. The S2SNet application can be very useful to calculate the protein Star Network TIs, which can be the base of a model for any other protein property. S2SNet can also be used for mass spectroscopy, clinical proteomics and imaging or DNA/RNA structure analysis.

Acknowledgments

Cristian R. Munteanu thanks the FCT (Portugal) for support from Grant SFRH/BPD/24997/2005. González-Díaz Humberto acknowledges an Isidro Parga Pondal research contract supported by Xunta de Galicia, University of Santiago de Compostela (Spain).

References

Aguero-Chapin, G., Gonzalez-Díaz, H., Molina, R., Varona-Santos, J., Uriarte, E., Gonzalez-Díaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett. 580, 723–730.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993a. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J. Biol. Chem. 268, 6119–6124.

Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993b. The guanidine U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993c. Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6550–6554.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1994a. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Experientia 50, 23–28.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1994b. Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-950152E. Biochemistry 33, 2663–2670.

Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 86, 342–357.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bielinska-Waz, D., Nowak, W., Waz, P., Nandy, A., Clark, T., 2007. Distribution of 2D-graphs as descriptors of DNA sequences. Chem. Phys. Lett. 443, 408–413.

Bisquerra Alzina, R., 1989. Introducción conceptual al análisis multivariante: Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD. PPU, Barcelona.

Chen, Y.L., Li, Q.Z., 2007a. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J. Theor. Biol. 248, 377–381.

Chen, Y.L., Li, Q.Z., 2007b. Prediction of the subcellular location of apoptosis proteins. J. Theor. Biol. 245, 775–783.

Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 264, 12074–12079.

Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35, 1–24.

Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.

Chou, K.C., Forsen, S., 1981. Graphical rules of steady-state reaction systems. Can. J. Chem. 59, 737–755.

Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J. Theor. Biol. 91, 637–654.

Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–18.

Chou, K.C., Shen, H.B., 2008. Cell-Plow: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Protocols 3, 153–162.

Chou, K.C., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1993. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1994. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1995. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1996. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1997. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1998. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 1999. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 2000. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.

Chou, K.C., Zhang, C.T., 2001. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 32, 6548–6554.
