Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

Regular Article

Entropic Analysis Reveals Unique Features in Anti-cancer Drugs

Chih-Yuan Tseng
Department of Oncology, University of Alberta, Edmonton, Alberta, Canada
MDT Canada, Edmonton, Alberta Canada
Correspondence, Email: rtseng@mdtcanada.ca

Received 2 March 2015; Accepted 19 March 2015; Published 10 April 2015

Abstract
Small molecule drugs are designed to bind to specific targets including proteins, enzymes, DNA, or RNA to inhibit/regulate their functions. Basically, these compounds all share some common drug-like properties such as Lipinski's rules of five. In this study, we propose to utilize entropic component analysis designed to tackle variable selection problems to explore further whether anti-cancer drugs designed for all types of cancers possess unique properties different from other types of drugs. We found out number of aromatic rings, partition coefficient (LogP) and number of hydrogen bond donor are those key factors. These results may provide a guideline to design better antic-cancer drugs.

Keywords: Maximum entropy, variable selection, anticancer drugs.

1. Introduction

Small molecule drugs normally are chemical compounds designed to bind to specific proteins, enzymes, DNA, or RNA that are identified as potential targets to inhibit/regulate their functions for treating diseases. Basically, chemical compounds must satisfy drug-like properties in order to be considered as lead compounds and become drugs. Two criteria, Lipinski's rules of five [1] and Absorption, Distribution Metabolism, Excretion and Toxicity properties (ADMET), which involve hundreds of physical, chemical pharmaceutical factors, are commonly considered. An obvious question one can ask is are all of those factors really required attentions while we are designing drugs. Particularly, we are more interested in determining whether there are common properties that define drugs as anti-cancer drugs despite the knowledge on biological targets used for the design and if there are what are they.

By answering these questions, we can expect to gain insights regarding the crucial chemical, physical and pharmaceutical properties of anti-cancer drugs. Furthermore, these understandings may be utilized as additional guideline to design better drugs specifically targeting cancers.

Since these questions are similar to variable selection problems, we propose to utilize entropic component analysis (ECA), which has been developed and demonstrated its applicability in various subjects involving multivariate problems in physics [2] and geology [3,4] to answer these questions in this study. We first briefly discuss the ECA method and compare to other conventional methods in next section. For our purpose, drugs data from DrugBank database [5, 6] will be the sources for this study. Based on the ECA, we find out there exists key factors that define drugs as anti-cancer drugs. These factors are number of aromatic rings, partition coefficient (LogP) and number of hydrogen bond donor.

2. Methods and materials

2.1 Entropic component analysis

Theory and approach. As mentioned earlier, the questions raised in Introduction can be considered as a variable selection problem. One common approach to tackle variable selection problems is through converting it into a model selection problem, which has been well studied in the past including the hypothesis testing based on the p-value method and statistical significance tests. Yet there exists some difficulties of practical use of these conventional methods for model selection. These difficulties include the uncertainty regarding the selection of the "right" significance level, which is usually set empirically and the hypothesis test for model selection is not designed for multi-model and large samples. A detailed review has been addressed by Raftery [7,8].

To be, a strategy to solve variable selection problems commonly involves the answers of two questions: *what is the form of the models that are optimally associated with variables*; and *what is the selection criterion*. The first question is usually partially solved by a process of trial and error. For example, a logistic model has been found to be suitable for binary response problems [9,10]. Regarding the second question, two criteria have been proposed and widely used, Akaike information criterion $AIC = -l\sum_{i=1}^{l} 1/l \log L_{y^i=1}\left(X^i, \hat{\beta}\right)$ [11] and Bayesian information criterion $BIC = -l\sum_{i=1}^{l} 1/l \log L_{y^i=1}\left(X^i, \hat{\beta}\right) + N/2 \log l$ [7, 12-14], where we assume *N* variables, $X^i = \left\{x_j^i\right\}$, labeled by subscript *j=1⋯N* and superscript *i=1⋯l* labels the observations, being considered for problems of interests and *l* corresponding outcomes or dependent variables, $\hat{Y} = \left\{y^i\right\}$. For binary-output problems, the positive outcome of the *i*th observation is denoted by $y^i = 1$ and the negative outcome is $y^i = 0$. Furthermore, $L_{y^i=1}\left(X^i, \hat{\beta}\right)$is the likelihood function and $\hat{\beta} = \left\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_N\right\}$ are maximum likelihood estimates (MLE) [15]. There have been some other criteria are proposed to provide better solutions including C information criterion [16], a generalization of BIC and AIC, and the minimum entropy analysis (MiEA) [2-4].

However, it requires prior information generated from some ad hoc prior modeling rules that suits people's need to properly use either AIC or BIC (refer to [2] for more discussion). We proposed an entropic analysis strategy based on the MiEA and Covariance (CV) analysis, which we term entropic component analysis (ECA), to overcome these shortcomings.

The ECA consists of two steps, (1) ranking the variables and (2) determining the preferred variables or key factors. Regarding ranking the variables, we apply Bayes's theorem to convert the variable selection problem into a probability model selection problem first. The probability model best represents our state of knowledge for positive outcome of the system associated with the variables X^i and some prior knowledge denoted by $P(y^i=1)$ is given by $P(y^i=1|X^i,\hat{\beta})$ = $P(y^i=1)L_{y^i=1}(X^i,\hat{\beta})$. Given *N* variables, there are 2^N-2 different combinations (sets) S_i of variables $X_{S_i} \in X$. Similarly, one can define submodels by $P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})$. Thus selecting the variables $X_{S_i} \in X$ is identical to selecting the corresponding $P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})$. The submodels are ranked according to the relative entropy of the submodels and a reference model $\mu(y^i=1|X^i)$ (refer to [2,4] for more discussions),

$$S[P,\mu]=-\sum_{i=1}^{l}P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})\log\frac{P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})}{\mu(y^i=1|X^i_{S_i})}\le 0$$

(1)

Because it is usually difficult to determine the underlying functions of the systems as the reference $\mu(y^i=1|X^i)$, a uniform distribution, $\mu(y^i=1|X^i)=m$, which indicates our complete ignorance of the system's outcomes, is chosen. One can then rewrite Equation (1) as

$$S[P,\mu]=\begin{array}{l}-\sum_{i=1}^{l}P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})\log L_{y^i=1}(X^i_{S_i},\hat{\beta}_{S_i})\\ +\log l+l\log m\le 0\end{array}$$

(2)

by substituting $P(y^i=1|X^i,\hat{\beta})$ = $P(y^i=1)L_{y^i=1}(X^i,\hat{\beta})$ and a prior probability $P(y^i=1)=1/l$ to denote our complete ignorance of weightings for each observation into $S[P,\mu]$. The maximum relative entropy max $S[P,\mu]=0$ indicates the model $P(y^i=1|X^i_{S_i},\hat{\beta}_{S_i})$ is identical to the uniform distribution, and contains no information. Conversely, a model with the minimum relative entropy indicates the maximum amount of information relevant to the system has been codified into it. Within the family of submodels, the decreasing order of $S[P,m]$ presents an increasing preference of the submodels. Furthermore, the full model associated with all independent variables $P_{\text{full}}(y^i=1|X^i)$ should contain the information mostly. Thus we define the explanatory power $P_o=S[P,m]/S[P_{\text{full}},m]\times 100\%$ to quantify the ability of submodels respect to full model to explain the outcome, where $S[P_{\text{full}},m]$ can be treated as the unit power.

Because the Bayesian interpretation of probability treats probability as a meaningful scale to represent the degree of knowledge one has to describe the system, the explanatory power defined based on this concept also presents a scale for ranking and is a descriptive statement about the extent of information codified in models [7,17]. When the explanatory power is zero, the corresponding model contains no information. The higher the explanatory power of the corresponding model, the more information it has to interpret the outcome.

However, when the number of independent variables gets too large, ranking all possible submodels is unwieldy and slow. To determine the preferred variables or key factors and analyze the joint effects of different combinations of the variables efficiently, the following procedure is considered. It starts by evaluating the explanatory power ranking scheme of all single variables, and then one select the first key factor from the top of the ranking scheme. One can further pursue the second key factor, which gives maximum joint explanatory power along with the first. One can continue this process until a criterion based on CV analysis is reached. The CV of explanatory power of the variables, except for the first ranking variable, is defined as the ratio of standard deviation and mean of the explanatory power. When the CV of a step is less than or equal to a threshold value, for example 5% in this study, the explanatory power of these variables is 95% indistinguishable, and the process will be stopped. Thus, all of the first ranking variables determined prior to this step are considered to have explanatory power being distinguishable. They will be recognized as the key factors that mostly influence the outcomes.

Statistical analysis procedure for large number of samples. Additionally, it always requires assessing goodness of regression to ensure obtaining "right" models. Many statistical methods thus have been devised. Following the Monte Carlo aspect, which relies on repeated random sampling to compute and obtain statistical results, we propose to integrate assessment of goodness of fit and the variable selection in one analysis procedure.

The procedure starts from analysing a set of small subsets data randomly sampled from the raw data using the first step of ECA, single variable analysis. The goodness of MLE fitting will roughly be examined using variance matrix of regression parameters first. The fitting will be considered as an acceptable fitting when the following two conditions are met. First condition, the variance matrix reciprocal condition given a subset for regression is larger than a threshold value, which is chosen as 0.0001 here (MATLAB algorithm "RCOND" was used [18]). Furthermore, the number of subsets that satisfies the variance matrix reciprocal condition is larger than half of total number of subsets. Second condition, when the explanatory power CV and normalized frequency or probability of selected variables become consistent given a sets of the subset sizes or the number of subsets larger than threshold values, the MLE fitting is acceptable. Finally, we will only analyze intersections of the subsets for submodels that pass these two criteria. Once the preferred size and number of the subsets is empirically determined, it will then be utilized directly for rest calculations.

Some remarks on ECA, AIC and BIC. Some remarks on the proposed selection criterion and other two widely used ones are discussed here before we tackle our problem. There are several differences can be found from their analytical forms. The dominant contribution of the three criteria is the average likelihood functions. Both AIC and BIC average the likelihood function with the same weightings for each observation, while entropic criterion, Equation (2), assigns each observation a probability $P\left(y^i = 1 \middle| X^i_{S_i}, \hat{\beta}_{S_i}\right)$ as weightings. There are also several minor differences among these criteria. For example, BIC additionally takes the logarithm of sample size *l* into account while AIC does not. It indicates that BIC leans toward lower-dimensional models more than the AIC does. When sample size is large and is assumed to come from a Koopman-Darmois family, Schwarz has shown AIC cannot be asymptotically optimal [12]. Both the entropic criterion and BIC takes the logarithm of sample size log *l* into account. Namely, entropic criterion also can overcome AIC's issue in large sample cases. Yet BIC further includes the effect of the total number of variables. Finally, entropic criterion additionally includes effect from the reference distribution, log *m*. However, because *m* is chosen as a constant, it has no impact on ranking.

One may expect AIC, BIC, and entropic criterion to reveal similar ranking schemes from the above observations. However, because entropic criterion takes the differences described by $P\left(y^i = 1 \middle| X^i_{S_i}, \hat{\beta}_{S_i}\right)$ in each observation into account, while both AIC and BIC fail to do so, one can expect that the entropic criterion is more adaptive to data than the other two. Consequently, variables selected by the ECA depend more on the data qualities than those chosen by either AIC or BIC. When data qualities are ensured, one may expect that the better results from ECA and likely to accurately obtain key factors.

2.2 Drug data

Since the DrugBank database collects detailed drug properties including comprehensive chemical, physical and pharmaceutical and corresponding biological targets [5,6], the DrugBank data are considered for this study. Particularly, we only consider 4072 out of 7759 drugs that are designed to target proteins for various diseases including all types of cancer. The data acquired from DrugBank was pre-processed into a two-dimensional matrix array for the studies. One dimension records all types of drugs properties. Here we consider 9 properties as listed in Table 1. The other one represents the dependent variable, which indicates whether this drug has anticancer function in this study. Note that there are 95 % of 4072 drugs are non-anticancer drugs denoted by $y^i = 0$ and for anticancer drugs are labelled by $y^i = 1$.

3. Results and Discussions

3.1 Descriptive statistics of all 4072 drugs

Before applying ECA to identify key factors that define drugs as anti-cancer drugs, we study the descriptive statistics of all 4072 drugs data and clustered them in three ways. First, when all 4072 data are considered, Fig. 1 shows the histograms of each individual factor. The mean values and standard deviation of nine factors are listed in Table 1. According to Table 1, there are 70% of drugs does not have hydrogen bond donor and the maximum number of hydrogen bond donor is 5. There are around 11~12% of drugs have 5 hydrogen bond acceptor and the maximum number is around 25. Regarding partition coefficient, logP, it ranges from -2 to 2 equally likely. Regarding solubility, logS, there are 20% of drugs has value 2. For polar surface area (PSA), 2.25 % of drugs have value around 50. The percentages of drugs that have PSA differs from 50 are all less than 1.5%.

There are 11% of drugs has 4 rotatable bonds. For number of rings, drugs that have at least 2 rings are around 20% of total drugs studied. There are 35 % of drugs do not have aromatic rings. Finally, molecular weights of drugs range from 5 mg to 400 mg. The molecular weight that the most drugs (1.3 %) have is 47 mg.

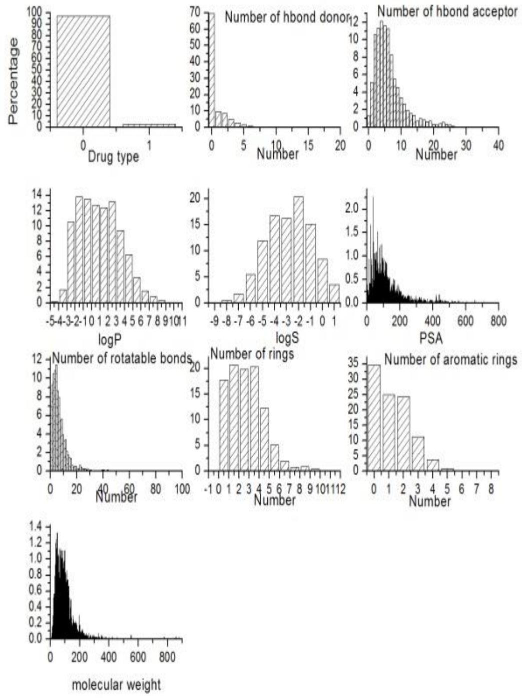
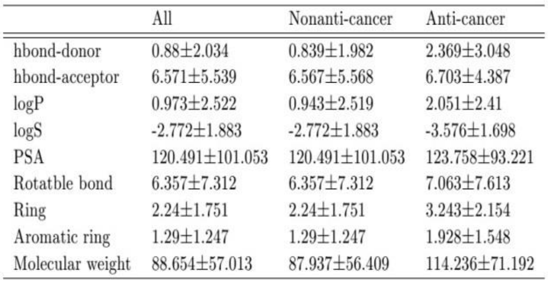


Figure 1. Histogram of 9 properties of 4072 drugs. The top left panel simply shows majority of 4072 drugs are not anti-cancer drugs.

All non anti-cancer drugs. Next, when all 3906 non anti-cancer drugs are analyzed, the descriptive statistics are shown in Fig. 2. The mean values and standard deviation of nine factors are listed in the "non anti-cancer" column in Table 1. The histograms are almost identical to the one in Fig. 1. The means and standard deviations as shown in Table 1 are identical to the values from all drugs for all factors except the number of hydrogen bond donor and acceptor, logP and molecular weight. This result is merely a consequence of non anti-cancer drugs are the dominant contributors.

All anti-cancer drugs. Finally, the histograms of nine factors when we just consider all anti-cancer drugs are given in Fig. 3 and the means and standard deviations are listed in anti-cancer column in Table 1. The histograms and mean values of all properties are quite different from non anti-cancer drugs.

Table 1. Descriptive statistics of 9 properties of 4072 drugs.



	All	Nonanti-cancer	Anti-cancer
hbond-donor	0.88±2.034	0.839±1.982	2.369±3.048
hbond-acceptor	6.571±5.539	6.567±5.568	6.703±4.387
logP	0.973±2.522	0.943±2.519	2.051±2.41
logS	-2.772±1.883	-2.772±1.883	-3.576±1.698
PSA	120.491±101.053	120.491±101.053	123.758±93.221
Rotatble bond	6.357±7.312	6.357±7.312	7.063±7.613
Ring	2.24±1.751	2.24±1.751	3.243±2.154
Aromatic ring	1.29±1.247	1.29±1.247	1.928±1.548
Molecular weight	88.654±57.013	87.937±56.409	114.236±71.192

hbond: hydrogen bond; PSA: polar surface area

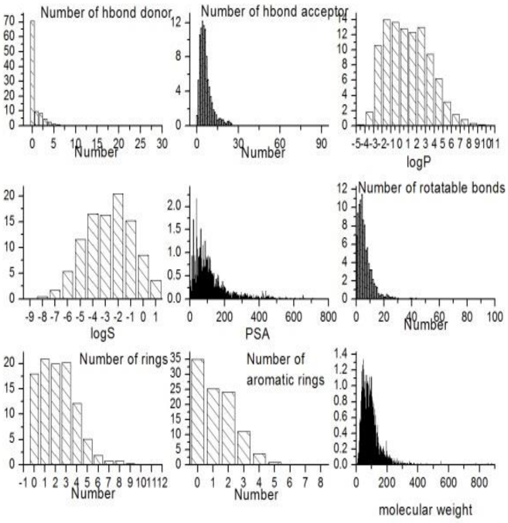


Figure 2. Histogram of the properties of all non anti-cancer drugs.

Summary. The descriptive statistics of nine properties as shown here shows that properties such as number of hydrogen bond donor, logP and molecular weight have distinct differences between non anti-cancer and anti-cancer drugs. Yet this descriptive statistics is still insufficient to answer the question that which chemical properties are key features for a drug to have anti-cancer functions.

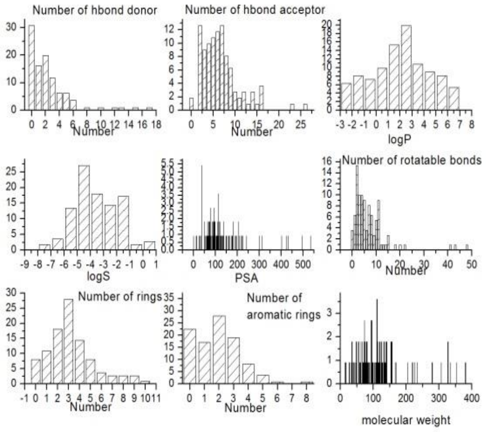


Figure 3. Histogram of the properties of all anti-cancer drugs.

3.2 Entropic component analysis

ECA convergence. The strategy of the ECA is to analyze the subset of raw data randomly selected rather than the whole data set. Therefore, we need to determine the minimum size of subset and numbers of trials randomly sampled from the raw data given the subset size required to obtain consistent analysis results. First, we fix the number of sampling to 1,000 and gradually increase the subset size from 10 % to 50 % of the size of raw data. Table 2 shows the results. In this table, the second column list the name of property that was selected to be primary factor to define a drug as an anti-cancer drug with the probability indicated in the third column. Given this property, ECA continually select second property as listed in the fourth column and the probabilities of these two properties being key factors are shown in the fifth column. The results indicate when the subset set size is larger than 20 % of the size of the raw data, the first three key factors selected by the ECA are identical. Furthermore, the normalized frequencies (denoted by Fre) of the selected key factors in 1,000 trials are gradually increased when the subset size is increased. From these results, one can conclude that the minimum size of subset required is 20 % of the size of the raw data. Note that after including the fourth (number of hydrogen bond acceptor) and fifth (logS) factors into the model, the frequency remains the same as the frequency when only three factors are considered.

Table 2. ECA prediction with consideration of different subset size.

Size(%)	1st	Fre (%)	2nd	Fre (%)	3rd	Fre (%)
10	Rot. bonds	28	PSA	31	logP	25
20	Aro. rings	46	logP	75	H-b d.	83
30	Aro. rings	57	logP	79	H-b d.	86
40	Aro. rings	61	logP	82	H-b d.	87
50	Aro. rings	67	logP	82	H-b d.	89

Second, we fixed the subset size to 40 % of the size of raw data and gradually increase the number of trials sampled from the raw data from 500 to 1,500. The results are listed in Table 3. Table 3 shows the same results for all cases except the normalized frequency is slightly smaller in the case with the number of trials being 1,500 than 1,000 as in Table 2. Nevertheless, the results indicate the number of trials has little impact on the ECA. Again, after including the fourth (number of hydrogen bond acceptor) and fifth (logS) factors into the model, the frequency remains the same as the frequency when only three factors are considered.

Table 3. ECA prediction with consideration of various number of sampling procedures.

Trial no	1st	Fre (%)	2nd	Fre (%)	3rd	Fre (%)
500	Aro. rings	62	logP	82	H-b d.	87
650	Aro. rings	62	logP	82	H-b d.	87
1000	Aro. rings	61	logP	82	H-b d.	87
1250	Aro. rings	62	logP	82	H-b d.	87
1500	Aro. rings	61	logP	82	H-b d.	87

Therefore, based on these two studies, we simply chose 40% of the raw data size as the minimum subset size, namely, 1,600, and randomly sampled raw data 1,000 trials for further investigating the key factors for anti-cancer drugs.

Number of aromatic rings, logP and number of hydrogen bond donor are key properties of anti-cancer drugs. Now we focus on the results obtained from using subsets with size being 40% of the raw data size in Table 1. The results indicate there is 61% chance that the number of aromatic rings being the primary key factor to define a drug as an anti-cancer drug. Given this property, it further indicates that there is 82 % chance that both the number of aromatic rings and logP are key factors. However, since there is only about 20% increase in probability (from 61% to 82%), it suggests that the logP only plays minor role. After including the third property, number of hydrogen bond donor, the probability is slightly increased to 87%. It suggests the effects of number of hydrogen bond donor are less than the logP. However, the probability remains 87% when either fourth, fifth or both properties are included. This result indicates that the effects of both factors are indistinguishable. From these results, therefore, one can conclude that the number of aromatic rings, logP and the number of hydrogen bond donor are three properties preferred over other factors are key features to define drugs as anti-cancer drugs. Moreover, ECA also indicates the preference of these three properties, the number of aromatic rings is preferred of logP and logP is preferred over the number of hydrogen bond donor.

4. Summary

We have proposed utilizing entropic component analysis to analyze chemical, physical and pharmaceutical properties of existing drugs collected in DrugBank database. Particularly, we are interested in determining whether there are common properties that make anti-cancer drugs unique and what are they despite biological targets. Our studies reveal there exists such common properties, which are number of aromatic rings, LogP and number of hydrogen bond donor. This study suggests that these three properties may provide a guideline to design compounds to have better potency in inhibiting/regulating cancer targets.

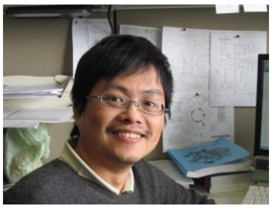
Acknowledgement
This research was enabled in part by support provided by Westgrid (www.westgrid.ca) and J. Tuszynski. The author appreciates technical support from P. Winter to retrieve and pre-process data from DrugBank.

Conflict of Interests
The author declares no conflict of interests regarding the publication of this article.

References

[1] Lipinski, C.A.; Lombardo, F.; Dominy, B. W. and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv*. Rev., **1997**, 23, 3–25.
[2] Tseng, C.-Y. Entropic Criterion for Model Selection. *Physica A*, **2006**, 370, 530-538.
[3] Tseng, C.-Y. and Chen, C.-C.* (2011) Entropic Component Analysis and Its Application in Geological Data. *Computers & Geosciences*, **2011**, 37, 1777-1782.
[4] Chen, C.-C.; Tseng, C.-Y. and Dong, J.-J. New Entropy-based Method for Variables Selection and Its Application to the Debris-flow Hazard Assessment. *Engineering Geology*, **2007**, 94, 19-26.
[5] Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B. and Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. **2008**, 36, D901-6.
[6] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. and Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *J. Nucleic Acids Res*., **2006**, 34, D668-72.
[7] Raftery, A.E. Bayesian model selection in social research. *Sociol. Methodol*. **1995**, 25, 111-196.
[8] Raftery, A. E.; Madigan, D. and Hoeting, J.A. Bayesian model averaging for regression models. *J. Am. Stat. Assoc*., **1997**, 92, 179–191.
[9] Cox, D.R. and Snell, E.J. Analysis of Binary Data second ed. **1989**, Chapman and Hall, New York
[10] DeMaris, A. A Tutorial in Logistic Regression. *J. Marriage and Family*, **1995**, 57, 956-968.
[11] Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **1974**, 14, 716–723.
[12] Schwarz, G. Estimating the dimension of a model. *Ann. Stat*., **1978**, 6, 461–464.
[13] Kieseppa, I. A. Statistical model selection criteria and Bayesianism. *Philos. Sci*., **2000**, 68, S141–152.
[14] Forbes, F.; Peyrard, N. Hidden Markov random field model selection criteria based on mean field-like approximations. *IEEE Trans. Pattern Anal. Mach. Intell*., **2003** 25, 1089–1101.
[15] Johnson, V. E. and Albert, J. H. Ordinal Data Modeling. Springer, New York. **1999**.
[16] Rodriguez, C. The ABC of model selection: AIC, BIC, and the new CIC. In: Knuth K.H., Abbas A.E., Morris R.D., Castle J.P. (Ed.) Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings, **2005**, 803, Melville, New York, 80–87.
[17] Jaynes, E.T. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK. **2003**.
[18] Anderson, E., Z.; Bai, C.; Bischof, S.; Blackford, J.; Demmel, J.; Dongarra, J.; Du Croz, A.; Greenbaum, S. and Hammarling, A. McKenney, and D. Sorensen, LAPACK User's Guide, Third Edition, SIAM, Philadelphia, **1999**.

Author's Biosketch



Dr. Tseng works at MDT Canada Inc. His laboratory is focusing on foundation and applications of entropic inference in biomedical sciences. Particularly, his research interests include foundation of theoretical statistical mechanics, protein folding dynamics, biological signal analysis, aptamer design, drug discovery methods in cancer, pharmacokinetic and pharmacodynamic modelling and simulation. He is co-founder of MDT Canada Inc. For more visit www.mdtcanada.ca or contact at rtseng@mdtcanada.ca.