Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Increment of diversity with quadratic discriminant analysis – an efficient tool for sequence pattern recognition in bioinformatics

This article was published in the following Dove Press journal:
Open Access Bioinformatics
30 July 2010
Number of times this article has been viewed

Jun Lu[1]
Liaofu Luo[2]
Lirong Zhang[2]
Wei Chen[2]
Ying Zhang[1,2]

[1]Department of Physics, Inner Mongolia University of Technology, Hohhot 010051, China; [2]Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

Abstract: Accompanying the rapid development of genome research, how to correctly recognize functional sites and structural modes of DNA and protein sequences has become a great challenge to bioinformatics. Therefore, a great number of algorithms and tools have been proposed and applied to sequence pattern recognition. Increment of Diversity with Quadratic Discriminant analysis (IDQD) is one of the efficient computational tools. In this article we shall introduce the main points of IDQD method and review its application in DNA and protein sequence pattern recognition, and finally give some discussions on the prospects of the approach.

Keywords: Increment of Diversity, Quadratic Discriminant analysis, sequence pattern recognition

Introduction

The availability of genome 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 (called sequence pattern). The sequence pattern is formed under evolutionary pressure and functional needs. "From sequence to structure, then to function" is the basic logic for the expression of life information. Sequence is the starting point of the logic. In the past 20 years, many statistical models and recognition algorithms were introduced and employed in the sequence pattern recognition. Among them, Bayes probability model, Artificial Neural Network (ANN), Hidden Markov model (HMM) and Support Vector Machine (SVM) are the most famous examples widely used in sequence analysis. Focused on the problem of sequence pattern recognition, in 2003, Zhang and Luo for the first time proposed the algorithm of Increment of Diversity with Quadratic Discriminant analysis (IDQD) and successfully employed it in intron splice site recognition.[1] In the following years, our group has made more detailed investigations on this method and applied it extensively to various bioinformatics problems, for example, the promoter and transcriptional starts recognition,[2,3] the DNase I hypersensitive site recognition,[4] the protein classification,[5–7] the nucleosome positioning prediction.[8] All these studies proved that IDQD is an efficient tool for sequence pattern recognition and prediction.

IDQD algorithm can be divided into two parts, Increment of Diversity (ID) and Quadratic Discriminant analysis (QD). The diversity measure was firstly introduced and employed in biogeography.[9] In that study the geographical distribution of species (the absolute frequencies of the species in different locations) was used as a source of diversity. Now, we employ the measure to the sequence pattern recognition. A pattern

Correspondence: Liaofu Luo
Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
Email lolfcm@imu.edu.cn; lujun@imut.edu.cn

or a distribution of sequence segments is described by several feature variables. To recognize or predict a sequence, for example, to recognize intron splicing sites, to predict transcription starts in a DNA sequence, one should define a set of feature variables first and then synthesize them in a scheme to give a prediction. The diversity measure and ID provide a method to extract information from genome sequence and QD gives a nonlinear approach to integrate different kinds of information into a scheme to recognize the pattern.

An introduction to IDQD

Increment of diversity algorithm[9]

Suppose the peculiarities of a sample, a sequence or a group of sequences, are described by a set of numbers. The *i*-th peculiarity is expressed by number n_i. For example, n_i describes the number of certain base in given site of sequences. We call n_i the informational parameter of the sample ($i = 1, \ldots, s$). Define the diversity of sample X as

$$D(X) = D(n_1, n_2, \cdots, n_s) = N \log_2 N - \sum_{i=1}^{s} n_i \log_2 n_i \quad (1)$$

$$(N = \sum_i n_i)$$

Generally, to give a classification of sequence X we should compare it with some standard samples (called standard diversity source). Let the *i*-th peculiarity in standard source expressed by number m_i ($i = 1, \ldots, s$), where m_i is the sum of peculiarity *i* over standard samples (training set samples). The diversity of standard source S is defined by

$$D(S) = D(m_1, m_2, ..., m_s) = M \quad \log_2 M - \sum_{i=1}^{s} m_i \log_2 m_i \quad (2)$$

$$(M = \sum_i m_i)$$

Likewise, the total diversity of the system $X + S$, $D(X + S)$, can be defined in the same manner. The increment of diversity is defined by

$$ID(X, S) = D(X + S) - D(X) - D(S) \quad (3)$$

ID gives the relation of sequence X with standard source S. The smallest ID, the most intimate relation of X to S. It can be proved that ID changes from 0 to D(N,M), where

$$D(N,M) = (N + M) \log_2 (N + M) - N \log_2 N - M \log_2 M \quad (4)$$

Because the dimension of informational parameters s (called the dimension of ID) may be large enough, the ID algorithm contains the projection manipulation of sequence peculiarity information onto high-dimension space.

The sequence pattern classification can be formulated by ID algorithm. For a c-classification problem one can form c standard sources S_l ($l = 1, \ldots, c$) which are derived from some set of sequence peculiarities of c classes of samples in training set. Then the standard diversity measures $D(X)$ for sequence X to be classified and $D(S_l)$ for the l standard sources can be obtained from Eq (1) and (2), respectively. The increment of diversity $ID(X,S_l)$ ($l = 1, 2 \ldots, c$) is deduced from $D(X)$, $D(S_l)$ and $D(S_l + X)$ by use of Eq (3). The decision rule is

$$X \in l \ if \ \mathrm{ID}(X, S_l) = min \ \{\mathrm{ID}(X, S_1), \mathrm{ID}(X, S_2), \ldots, \mathrm{ID}(X, S_C)\} \quad (5)$$

When there exits r sets of sequence peculiarities, we have r feature variables ID_1 to ID_r forming an r-dimensional vector and need to integrate it by quadratic discriminant analysis to give a decision.

Integrating IDs by quadratic discriminant analysis

Consider a 2-classification problem at first. For a sequence X to be classified into two sets (positive set ω_1 and negative set ω_2), the discriminant function is defined by

$$\xi = \ln p(\omega_1 \mid X) - \ln p(\omega_2 \mid X) \quad (6)$$

where $p(\omega_i|X)$ means conditional probability. According to Bayesian Theorem,

$$p(\omega_l \mid X) = p(\omega_l)\, p(X \mid \omega_l) / p(X) \ (l = 1, 2) \quad (7)$$

where $p(\omega_l)$ is the priori probability, proportional to the size of set l. Inserting (7) into (6), we obtain

$$\xi = \ln \frac{p(\omega_1)}{p(\omega_2)} + \ln \frac{p(X \mid \omega_1)}{p(X \mid \omega_2)} \quad (8)$$

Set the feature vector of sample X being $\boldsymbol{R}_X = (ID_1, \ldots, ID_r)$. Assume normal distribution of feature variables in two sets

$$p(X \mid \omega_l) = \frac{1}{Z_l} \exp\left(-\frac{1}{2}(\mathbf{R}_X - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\mathbf{R}_X - \boldsymbol{\mu}_l)\right) \quad (9)$$

$$Z_l = (2\pi)^{r/2} \left|\Sigma_l\right|^{1/2}$$

where $\boldsymbol{\mu}_l$ ($l = 1, 2$) (r-dimensional vector) and Σ_l ($l = 1, 2$) ($r \times r$ matrix) are the mean and covariant of feature variables over positive and neg ative sets respectively, $|\Sigma_l|$ is the determinant of matrix Σ_l. Inserting (9) into (8), we obtain[10,11]

$$\xi = \ln\frac{p(\omega_1)}{p(\omega_2)} - \frac{1}{2}((\mathbf{R}_X - \mathbf{m}_1)^T \Sigma_1^{-1}(\mathbf{R}_X - \mathbf{m}_1) - (\mathbf{R}_X - \mathbf{m}_2)^T \Sigma_2^{-1}(\mathbf{R}_X - \mathbf{m}_2)) - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \quad (10)$$

This result can easily be generalized to the classification of more than two-groups. For any two, say i and j, in c classes one can deduce a discriminant function between class i and j,

$$\xi_{ij} = g_i(\mathbf{R}_X) - g_j(\mathbf{R}_X) \quad (11)$$

where

$$g_l(\mathbf{R}_X) = \log P_l - \frac{1}{2}\delta_l - \frac{1}{2}\log|\Sigma_l|,$$

$$\delta_l = (\mathbf{R}_X - \boldsymbol{\mu}_l)^T \Sigma_l^{-1}(\mathbf{R}_X - \boldsymbol{\mu}_l)\ (l = 1,2,\ldots,c) \quad (12)$$

Here P_l means the total number of samples in the l-th class of training set. Eq (11) and (12) are in the same form as Eq(10). So, the general decision rule for c-classification is

$$X \in l \text{ if } g_l(\mathbf{R}_X) = max\{g_1(\mathbf{R}_X), g_2(\mathbf{R}_X), \ldots, g_c(\mathbf{R}_X)\} \quad (13)$$

In particular, for $c = 2$ classification, the decision rule is

$$X \in \begin{cases} positive, \text{ if } \xi > \xi_0 \\ negative, \text{ otherwise} \end{cases} \quad (14)$$

In the common use of quadratic discriminant analysis the threshold of ξ (denoted as ξ_0) is taken to be 0. However, due to the limited sizes of positive and negative sets and the large difference between them, the optimal threshold ξ_0 may not be 0.

Determination of threshold ξ_0

For a 2-classification problem, the choice of the best-fit ξ_0 depending on the size ratio of positive set to negative set should be determined empirically in principle. Consider N ($N >> 1$) samples stochastically taken from positive and negative set (N_+ samples in positive set and N_- samples in negative set) and classify them. For a sample X_i (i=1,2, …, N), we obtain ξ_i according to Eq(10). Then we arrange the $N\xi_i$'s in a decreasing order. The former $N\frac{N_+}{N_+ + N_-}$ samples (with larger ξ-values) should be predicted as positive ones and the latter $N\frac{N_-}{N_+ + N_-}$ should be predicted as negative ones. Thus the best threshold ξ_0 is deduced, which we denote as ξ_T. In general, as the size of positive set does not equal to that of negative set, it is evidently that $\xi_T \neq 0$.

One can calculate the prediction sensitivity Sn and specificity Sp for each assumed ξ_0 and plot receiver operating characteristic (ROC) curve – the relation between Sn and $1–Sp$ – for varying ξ_0. Set $y = Sn = TP/N_+$, $x = 1–Sp = FP/N_-$, where TP means true positive and FP means false positive. The slop of ROC curve is

$$\frac{dy}{dx} = \frac{\delta TP}{\delta FP}\frac{N_-}{N_+} = -\frac{\delta FN}{\delta FP}\frac{N_-}{N_+} \quad (15)$$

where FN means false negative, $N_+ = TP + FN$, $N_- = TN + FP$. The minus at the right hand of (15) means the variation of FN and FP is always in opposite direction as ξ_0 changing. It can be shown that $\delta FN = -\delta FP$ at $\xi_0 = \xi_T$ (the intersection of the distribution curves of positive and negative samples at ξ-axis). Therefore

$$\frac{dy}{dx} = \frac{N_-}{N_+}(\xi_0 = \xi_T) \quad (16)$$

We have demonstrated that the best ξ_0 value can be determined from the ROC curve at the point where the slope of ROC curve equals the negatives-to-positives ratio.

In fact, two groups of performance measures can be used to assess the accuracy on classification prediction. The first group includes sensitivity Sn, specificity Sp, false positive rate FPR, positive predictive value PPV and correlation coefficient CC, etc. Second group "single-number" performance measures include auROC (area under the curve Receiver Operator Characteristics) and auPRC (area under the curve Precision Recall Curves, PRC curve giving the relation between positive predictive and true positive). Since in our approach the only parameter is ξ_0, so in the latter performance measures the ROC curve and PRC curve can be plotted by use of all ξ_0 values taken from the full parameter space. These performance evaluations are independent of the parameter choice.

Examples

The biological signal occurs in a sequence can generally be two types: one is the k-mer content in a given region of sequence and the other is the consensus sequence segment at some particular sites. The former describes the compositional features of a sequence and the latter describes the base (or amino acid) dependencies at adjacent/non-adjacent positions of a group of sequences of same kind. We will give examples to show how the IDs are defined. For splice site recognition, suppose there are a consensus sequence of length m around the splice site GT or AG and a longer

sequence (say, length L) containing compositional signals. Consider base-triplet compositional signal and base-pair conservative signal only. One can define $D_1(S)$ of standard source to describe base-triplet content by use of Eq (2) where m_i means the frequency of i-th triplet (i = 1,…,64) occurring in all L-long sequences in positive or negative training set, and define $D_1(X)$ by use of Eq (1) where n_i means the frequency of i-th triplet occurring in sequence X. Likewise, one can define $D_1(X+S)$ of the mixed system of X and standard source and obtain ID_1 following Eq (3). To describe the base pair dependencies at adjacent or non-adjacent positions nearby splice site one can define $D_2(S)$ of standard source by use of Eq (2) where m_i means the frequency of i-th base-pair correlation (i = 1,…, $m(m\text{-}1)/2$) occurred in all m-long sequences in positive or negative training set, and define $D_2(X)$ by use of Eq (1) where n_i means the frequency of i-th base-pair correlation in sequence X. Then one can define $D_2(X+S)$ of the mixed system and obtain ID_2. ID_1 and ID_2 form a 2-dimensinal feature vector ID. They should be integrated into a single discriminant function by QD through Eq (10), where R_X means 2-dimensinal feature vector ID for sample X, $\boldsymbol{\mu}_l (l = 1,2)$ is the 2-dimensional vector obtained from the average of ID over training positive or negative sets respectively and Σ_l (l = 1,2) the corresponding 2×2 matrix deduced from the covariant of feature variables.

Application of IDQD in DNA and protein sequence pattern recognition

IDQD algorithm is an efficient tool for sequence pattern recognition. From 2003 on, we have applied the algorithm to many examples of DNA and protein sequence recognition and prediction and achieved admirable results.

Application of IDQD in DNA sequence pattern recognition

The splice site recognition of eukaryotic genes is one of the important problems in computational biology. In 2003, Zhang and Luo for the first time applied IDQD algorithm to the recognition of constitutive splice sites of several species including human.[1] In that study, eight diversity sources were constructed based on the conservation of nucleotides at splicing sites and the features of base composition and base correlation around these sites. It includes the adjacent and nonadjacent base correlation in donor and acceptor consensus sequences, and the triplet content in 48 (80)-base-long sequences around splice sites. Then the quadratic discriminant vector constructed by eight IDs was deduced. The 3-fold cross-validation results were higher than the leading software Genesplicer.[12] In later years Zhang et al applied IDQD algorithm to the recognition of human alternative splice sites and obtained the overall accuracies of prediction 87.9% and 89.9% for donors and acceptors respectively (with the chosen threshold-2), higher than currently reported accuracy by use of other prediction methods.

From 2005 on, Lu and Luo applied IDQD algorithm to the recognition of human promoter and transcription start site.[2,3] In the neglect of nucleosome positioning and histone modification information and by use of DNA sequence information only they chose different k-mer frequencies as the diversity sources. They constructed 14 diversity sources by use of 6-mer frequencies in four 500bp segments as the main recognition information and 5-mer and 4-mer frequencies as the initiator sequence and DNA structure information, respectively. For each diversity source, two IDs were introduced, one from the comparison with positive set and the other from the comparison with negative set. They found that the employment of two-ID always improves the prediction. The result showed that both sensitivity and positive predictive value have achieved a value higher than 65% with positives/negatives ratio 1:58, higher than the upper limit deduced from eight leading algorithms as analyzed by Bajic.[13] The result of IDQD was also better than the SVM model ARTS[14] by comparison of auROC and auPRC in two models when the same data were used.

Recently, Chen and Luo employed IDQD algorithm to the prediction of DNase I hypersensitive sites (DHSs).[4] For DHSs prediction in K562, $CD4^+$ T, Hela and GM06990 cell lines, the average accuracies of 10-fold cross-validation test are 98.52%, 96.50%, 99.25% and 97.58%, respectively, and the mean areas under ROC curves (auROC) are all greater than 0.90. The results showed that the IDQD method is an effective tool for DHSs recognition.

Application of IDQD in protein sequence pattern recognition

The secondary structure is the basis of spatial structure of a protein. Chou and Fasman were the first authors to formulate an approach to the empirical prediction of protein secondary structures based on amino acid sequence. The successful score was about 50% or larger. After 40 year's effort, most prediction accuracy of various methods still paces up and down at the level of 75% or higher. So the dinosaurs of secondary structure prediction are still alive and any 1% improvement of the prediction will be a great progress. In 2008, Feng

and Luo developed the IDQD algorithm and used it to the secondary structure prediction and attained a higher accuracy.[5] Based on tetra-peptide signals of structures they proposed tetra-peptide-based increment of diversity with quadratic discriminant analysis (TPIDQD). As predicting the structure of the central residue for 21-residue fragments in the noted CB513 dataset the three-state overall per-residue accuracy (Q3) is about 80% in the 3-fold cross-validated test. The results show the efficiency of IDQD method and indicate the importance of tetra-peptide signals as the protein folding code in the protein structure prediction.

Based on N-terminal amino acid sequence peculiarities the method of IDQD also predicted the subcellular localization of proteins of four plant categories and three nonplant categories and obtained accuracies 87.4(± 0.5)% and 91.2(± 0.2)% respectively in 5-fold cross-validation test better than other published methods.[7]

Specific examples of applications of IDQD method

Example 1

Antimicrobial peptides are crucial components of the innate host defense system of most living organisms. Since their discovery in 1925,[15] hundreds of antimicrobial peptides have been identified and some of them have been successfully used for defense against both animal and human pathogens.[16,17] Thus, recognizing the biological activities of these peptides is critically important for the design of novel therapeutic agents. However, theoretical classification of antimicrobial peptides according to biological activities is not yet explored.

In 2009, Chen and Luo successfully applied the IDQD method to the classification of antimicrobial peptides for the first time.[6] By use of amino acid and dipeptide composition as the diversity source, 37 antiviral/HIV, 41 anticancer/tumor, 389 antibacterial and 177 antifungal peptides were classified with accuracies of 94.38%, 89.36%, 90.32% and 95.46%. These results are superior to that of support vector machine, Table 1.

Example 2

Protein phosphorylation is one of the most important reversible post-translational modifications (PTMs). Zhang, Luo and Lu used the method of IDQD to predict the phosphorylation sites recognized by three kinase families CK2, PKA and PKC and obtained 7-fold cross-validation test accuracy (Ac) 86%, 90% and 85% respectively for these three kinds of phosphorylation sites. The results are comparable with or better than other published top software. The results are summarized in Table 2.

Table 1 Comparative result for the classification of antimicrobial peptides by the jackknife test*

Methods	Antimicrobial Peptides	Sn (%)	Sp (%)	Acc (%)	MCC
IDQD	Antiviral/HIV	95.23	83.33	94.38	0.85
	Antifungal	84.07	90.91	89.36	0.76
	Anticancer/tumor	89.29	83.33	90.32	0.80
	Antibacterial	90.91	90.48	95.46	0.85
SVM	Antiviral/HIV	71.43	75.00	90.58	0.65
	Antifungal	81.48	78.57	87.50	0.71
	Anticancer/tumor	78.57	75.86	86.52	0.67
	Antibacterial	81.82	85.71	91.67	0.78

Abbreviations: *Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, matthew's correlation coefficient.

$$Sn = \frac{TP}{TP + FN}, SP = \frac{TP}{TN + FP}, Acc = \frac{TP + TN}{TP + FN + TN + FP},$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

TP, true positive; FN, false negative; TN, true negative; FP, false positive.

Example 3

H2A.Z, the variant of H2A, is involved in diverse biological functions, such as gene activation, chromosome segregation, heterochromatin silencing and cell cycle progression. In view of this, differentiating histone variant containing nucleosomes with canonical nucleosomes will provide novel insights into the full understanding of gene regulation.

We extracted 2000 H2A.Z and 2000 H2A containing nucleosomes with lowest P-values from human $CD4^+$ T cells and constructed a training set.[19] Based on the presence (or absence) of 20 histone methylations (H2BK5me1, H3K27me1, H3K36me1, H3K36me3, H3K79me1, H3K79me2, H3K79me3, H3K9me3, H3K9me1, H3R2me1,

Table 2 Comparative result of prediction for phosphorylation site by 7-fold cross-validation test

Method	Kinase family	Acc (%)	Sn (%)	Sp (%)	MCC
IDQD	CK2	86.14	78.32	91.14	0.71
	PKA	90.12	89.33	90.63	0.80
	PKC	84.63	76.94	90.28	0.68
SVM*	CK2	91.47	83.90	96.43	0.82
	PKA	89.98	88.32	91.11	0.79
	PKC	82.90	78.71	85.79	0.65

Notes: *The results are given by Kim and colleagues[18] for comparison.

H3R2me2, H3K27me2, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me2, H4K20me1, H4K20me3, H4R3me2),[19] all the nucleosomes in the training set were recoded in a 20-dimensional space by 1 (or 0). For example, if all of the 20 histone methylations appeared, the nucleosome would be represented as "11111111111111111111". By using IDQD algorithm, we for the first time obtained an accuracy of 89.30 for the classification of H2A.Z and H2A containing nucleosomes in the 5-fold cross validation test. These results demonstrate the effectiveness of the IDQD model.

Discussions and prospects on IDQD method

Remarks on IDQD method

The Maximum Information Principle (MIP) is a fundamental principle in nonequilibrium statistical theory. The principle states that the information content (Shannon information quantity) of any nonequilibrium system tends towards a maximum under a set of constraint conditions. Nucleic acid sequence is a typical nonequilibrium system, where bases often undergo the mutation stochastically due to the inherent perturbation in the microenvironment. Simultaneously, the stochastic mutation happens under some functional constraints. The base mutation is a fast-varying variable while the functional constraint evolves relatively slow. So the base occurrence in nucleic acid sequence takes a stable distribution due to stochastic mutation under functional constraints. Luo et al proposed that the MIP can be introduced as a guiding principle in the genetic language research and bioinformatics of nucleic acid sequence.[20] Jin and Luo developed the MIP formalism and applied it to the prediction of gene splicing sites.[21] Due to information maximization in the course of the evolution the Shannon information quantity of k-mer frequency of a DNA sequence segment around some functional sites (for example, splice sites, transcriptional starts, etc) always takes a stable value. Therefore, Shannon information is a key quantity which can be utilized in the bioinformatics analysis. The diversity measure is essentially a measure of information content. In fact, the diversity of any sequence X, Eq (1), is equal to the Shannon information quantity of the sequence multiplied by sequence length N. So, the ID analysis is essentially the information-theoretic analysis. This is the very reason why ID method is so successful in the bioinformatics study. Another point is: when several feature variables (IDs) are integrated into one discriminant function we start from Bayes theorem and use quadratic discriminant analysis. As long as the number of samples both in positive and negative sets is large enough one may always assume the distributions of feature variables in two sets are of Gaussian type. In this case QD is a good formalism applicable for information integration (see Generalization of QD section).

Extracting information for pattern recognition

In IDQD method the sequence pattern information is extracted by use of the diversity measure and the increment of diversity (ID). In fact, other measures for extracting information can be designed.[22] Let us study two kinds of Information Deviation Measure (IDM): the first is based on Laxton's diversity[9] and called ID_L and the second is based on Kullback's information gain,[23] called ID_K. The definition of ID_L has been given in Eqs (1) – (3). Suppose that the probability distribution of the character in standard source and in the sample to be predicted is $p_i = m_i / \sum m_i = m_i / M$ and $p_i' = n_i / \sum n_i = n_i / N$, respectively. The probability in the mixed set including standard source and the predicted sample is $q_i = (m_i + n_i) / (M + N)$. The Information Deviation ID_L of the inquired sequence from the standard source can be expressed as

$$\begin{aligned} ID_L &= -(M+N)\sum_i q_i \log q_i + M\sum_i p_i \log p_i + \\ &\quad N\sum_i p_i' \log p_i' \\ &= -N\left(\sum_i q_i \log q_i - \sum_i p_i' \log p_i'\right) - \\ &\quad M\left(\sum_i q_i \log q_i - \sum_i p_i \log p_i\right) \end{aligned} \tag{17}$$

Easily shown that $ID_L \geq 0$ and the smallest ID_L, the most intimate relation of the inquired sequence to the standard source. ID_L is essentially the increment of diversity defined by Laxton.[9] The meaning of ID_L can be understood as follows. For Boltzmann distribution the free energy E_i is proportional to log p_i. So $-M\sum_i p_i \log p_i$ describes the average energy of standard set, $-N\sum_i p_i' \log p_i'$ describes the average energy of inquired sequence and $-(M+N)\sum_i q_i \log q_i$ the average energy of the mixed system. Therefore ID_L gives the energy difference as the inquired sequence merged into standard source. Our computational experience shows ID_L is a sensitive parameter to measure if the inquired sequence belongs to the standard set. This is because of a large amount of terms appeared under the summation of Eq (17) (for example, if the character is the hexamer frequency, there are 4^6 terms in the summation) and the ID_L algorithm containing the projection manipulation of sequence

information onto high-dimension space. So, we are able to use ID_L to evaluate the detailed difference of any sample with the standard set and find the optimal hyperplane for the classification of samples in multi-dimensional space.

Following Kullback, the deviation of the probability distribution p_i' from the standard distribution p_i is described by $\sum\{p_i' \log(p_i'/p_i)\} \geq 0$.[23] We introduce the total Kullback distance of inquired sequence from source p_i and from mixed system q_i,

$$KD_1 = \sum_i \{p'_i \log \frac{p'_i}{p_i} + p'_i \log \frac{p'_i}{q_i}\} \quad (18)$$

Simultaneously we introduce the mutual Kullback distance between p_i and q_i,

$$KD_2 = \sum_i \{(p_i - q_i) \log \frac{p_i}{q_i}\} = \sum_i (p_i \log \frac{p_i}{q_i} + q_i \log \frac{q_i}{p_i}) \quad (19)$$

The Information Deviation ID_K of the inquired sequence from standard set is defined by

$$ID_K = NKD_1 + MKD_2 \quad (20)$$

Equation (20) can be comparable with Equation (17). $ID_K \geq 0$. The smallest ID_K, the most intimate relation of the inquired sequence to standard source. Our computational experience shows ID_K is also a sensitive parameter to measure the deviation of the inquired sequence from the standard set.

Generalization of QD

In some cases of bioinformatics applications the distribution of feature vector ID may not be of Gaussian type. The generalization of QD can be deduced as follows. Instead of Eq (10) we have

$$\xi = \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln P(x|\omega_1) - \ln P(x|\omega_2) \quad (21)$$

If the r components of vector ID are stochastic variables independent of each other, then

$$p(\boldsymbol{x} \mid \omega_l) = \prod_i^r p_i(x_i|\omega_l) \qquad (l = 1,2)$$

The ratio of $P(\boldsymbol{x}/\omega_1)$ *to* $P(\boldsymbol{x}/\omega_2)$ in Eq (21) can be obtained through comparison with $P_i(\text{positives}|\omega_1)$ and $P_i(\text{negatives}|\omega_2)$, namely through direct statistics of the distribution of feature vector IDs in positive samples and negative samples. However, the r components of vector ID are generally not independent each other. In this case we make linear transformation of ID components to diagonalize the covariance matrix Σ_l. Set

$$S_l \Sigma_l S_l^{-1} = \Xi_l \quad S_l^+ = S_l^{-1} \quad (22)$$

where Ξ_l is diagonal. Since Σ_l is real and symmetric, the unitary transformation S_l does exist. That is, by introducing different representations the feature vector in old representation x, ID(x, set-l), is related to that in new representation y, ID(y, set-l), by linear transformation

$$S_l \mathbf{ID}(\boldsymbol{x}, \text{set-}l) = \mathbf{ID}(\boldsymbol{y}, \text{set-}l) \quad (23)$$

In y-representation the r components of vector ID are independent. Now both the probability functions $p(\mathbf{y}|\omega_1)$ of positives and $p(\mathbf{y}|\omega_2)$ of negatives in y-representation can be expressed as the product of r factors,

$$p(\boldsymbol{y} \mid \omega_l) = \prod_i^r p_i(y_i \mid \omega_l) \qquad (l = 1,2) \quad (24)$$

It is easily to prove the normal distribution remains unchanged under unitary transformation S_l. For a sample $\boldsymbol{x}$ to be classified, through transformation S_l we deduce the feature vector in y-representation, $S_l \mathbf{ID}(\boldsymbol{x}) = \mathbf{ID}(\boldsymbol{y}_l)$ Note that two y-representations corresponding to transformations S_1 and S_2 are generally different. Finally, as Eq (21), by using probability functions of positives and negatives in y-representation we obtain the decision parameter ζ

$$\xi = \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln P(\mathbf{y}|\omega_1) - \ln P(\mathbf{y}|\omega_2) \quad (25)$$

The central point of the method is the unitary transformation (U-transformation) made for calculating probability distribution. As a natural generalization of IDQD the method is called IDUD[22] which can be utilized in cases where the normal distribution does not hold. We have shown that both IDQD and IDUD are efficient tools for sequence pattern recognition in many examples and expect that they can be applied broadly to genome and proteome analysis.

The sequence pattern recognition is a fundamental problem of bioinformatics analysis. We have proposed a unified computational approach – IDQD and IDUD – which

includes 1) the information extraction through Information Deviation (ID) of the inquired sequence from standard set, and 2) the information integration through Quadratic Discriminant analysis (QD) in case of feature variables obeying multi-dimensional normal distribution or its generalization, U-transformation Discriminant analysis (UD) in more general cases.

Acknowledgments

This work was supported by Science Project of Inner Mongolia University of Technology ZD200917 and National Natural Science Foundation of China, No.90403010.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Zhang LR, Luo LF. Splice site prediction with quadratic discriminant analysis using diversity measure. *Nucleic Acids Res*. 2003;31(21):6214–6220.
2. Lu J, Luo LF. Prediction for human transcription start site using diversity measure with quadratic discriminant. *Bioinformation*. 2008;2(7):316–321.
3. Lu J, Luo LF. Predicting human transcription starts by use of diversity measure with quadratic discriminant. *Computation in Modern Science and Engineering: Proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007), AIP Conf Proc*. 2007;963:1273–1277.
4. Chen W, Luo LF, Zhang LR, Lin H. Recognition of DNase I hypersensitive sites in multiple cell lines. *International Journal of Bioinformatics Research and Applications*. 2009;5(4):378–384.
5. Feng YE, Luo LF. Use of tetra peptide signals for protein secondary-structure prediction. *Amino Acids*. 2008;35(3):607–614.
6. Chen W, Luo LF. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. *Journal of Microbiological Methods*. 2009;78(1):94–96.
7. Jia Y, Zhao JD, Lu J. The recognition of subcellular localization of proteins based on their N-terminal amino acid sequence. *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on. 16/06/2008*; doi: 10.1109/ICBBE.2008.64.
8. Chen W, Luo LF, Zhang LR. The organization of nucleosomes around splice sites. *Nucleic Acids Res*. 2010;doi:10.1093/nar/gkq007.
9. Laxton RR. The measure of diversity. *J Theor Biol*. 1978;70(1):51–67.
10. McLachlan GJ. *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley and Sons; 1992.
11. Zhang MQ. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci U S A*. 1997;94(2):565–568.
12. Pertea M, Lin XY, Salzberg SL. Gene splicer: a new computational method for splice site prediction. *Nucleic Acids Res*. 2001;29(5):1185–1190.
13. Bajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. *Nature Biotechnology*. 2004;22(11):1467–1473.
14. Sonnenburg S, Zien A, Rätsch G. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*. 2006;22(14):e472–e480.
15. Gartia A. Sur un remarquable exemple d'antagonisme entre deux souches de colibacille. *Comput Rend Soc Biol*. 1925;93:1040–1041.
16. Snelling AM. Effects of probiotics on the gastrointestinal tract. *Curr Opin Infect Dis*. 2005;18(5):420–426.
17. Kirkup BC. Bacteriocins as oral and gastrointestinal antibiotics: theoretical considerations, applied research, and practical applications. *Curr Med Chem*. 2006;13(27):3335–3350.
18. Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*. 2004;20(17):3179–3184.
19. Zhang Y, Shin H, Song JS, Lei Y, Liu XS. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*. 2008;9:537.
20. Luo LF, Bai GY. Maximum information principle and evolution of nucleotide sequences. *J Theor Biol*. 1995;174(2):131–136.
21. Jin HY, Luo LF, Zhang LR. Using estimative reaction free energy to predict splice sites and their flanking competitors. *Gene*. 2008;424(1–2):115–120.
22. Luo LF, Lu J. Sequence pattern recognition in genome analysis. *Computation in Modern Science and Engineering: Proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007), AIP Conf Proc*. 2007;963:1278–1281.
23. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Star*. 1951;22(1):79–86.