DNA pattern recognition using canonical correlation algorithm

BK Sarkar[1] and Chiranjib Chakraborty[2,*]

[1]*Department of Physics, School of Basic & Applied Sciences, and* [2]*Department of Bioinformatics, School of Computer Sciences, Galgotias University, Greater Noida, India*

**Corresponding author (Email, drchiranjib@yahoo.com)*

We performed canonical correlation analysis as an unsupervised statistical tool to describe related views of the same semantic object for identifying patterns. A pattern recognition technique based on canonical correlation analysis (CCA) was proposed for finding required genetic code in the DNA sequence. Two related but different objects were considered: one was a particular pattern, and other was test DNA sequence. CCA found correlations between two observations of the same semantic pattern and test sequence. It is concluded that the relationship possesses maximum value in the position where the pattern exists. As a case study, the potential of CCA was demonstrated on the sequence found from HIV-1 preferred integration sites. The subsequences on the left and right flanking from the integration site were considered as the two views, and statistically significant relationships were established between these two views to elucidate the viral preference as an important factor for the correlation.

[Sarkar BK and Chakraborty C 2015 DNA pattern recognition using canonical correlation algorithm. *J. Biosci.* **40** 709–719] **DOI** 10.1007/s12038-015-9555-z

1. Introduction

Progresses in genome sequencing technology enable scientists to search some common sequence elements in the DNA sequence. For example, DNA-binding proteins composed of DNA-binding domains are likely to bind related DNA sequences (Dickerson 1983; Pabo and Sauer 1984). Searching for functionally similar DNA sequences is very important to find some particular pattern in the sequence. In order to identify common patterns, several search approaches can be found in the literature. Hertz and Stormo (1999) modelled a succession of mechanisms to conclude alignments of multiple sequences. Based on the greedy algorithm they described log-likelihood scoring systems for searching alignments of functionally related sequences. Guha Thakurta and Stormo (2001) implemented Gibbs sampling method to explore subjective binding site patterns in DNA sequences. In Gibbs sampling method a stochastic variant of expectation maximization is determined (Lawrence *et al.* 1993; Neuwald *et al.* 1995) and a predetermined motif is substituted by another one that possesses a higher score, thus permitting escape from local optima. An algorithm based on palindromic behaviour between the frequencies of bases was formulated to explore the integration sites in HIV-1 consensus sequences (Holman and Coffin 2005; Wu *et al.* 2005). But in such stochastic motif detection algorithm, it sometimes results in non-identical outputs in multiple runs of the simulation, keeping the input unchanged. Another drawback is that they are single modal techniques that can only deal with data from a single view. However, multimodal data from the semantic groups is prevalent in practice. Therefore, handling such multimodal data at the same time is a fundamental and practical problem. One effective method to address this issue is 'canonical correlation analysis' (CCA) (Hotelling 1936).

CCA compares two sets of multidimensional variables (in this case, a set of DNA sequence and a set of pattern) to examine the correlation between them (Hotelling 1936). CCA looks for correlated functions which are covariates of two different sets having some relation (Kettenring 1971; Johnson and Wichern 1992; Hardoon *et al.* 2004; Tenenhaus and Tenenhaus 2011). The attainability of such correlated

Keyword. Canonical correlation analysis; DNA sequence; pattern recognition

Published online: 28 September 2015

functions of two semantic sets is likely to persist due to a causal factor accountable for the correlation. CCA aims to discover the correlation between a linear combination of the variables in one set and another linear combination of the variables in the other set by projecting them onto a lower-dimensional space and maximum correlated (Kettenring 1971; Breiman and Friedman 1985; Yu *et al.* 2006; Iaci *et al.* 2010). The advantage of CCA is evident: First, it keeps the operative discriminant information of multiple modalities; secondly, it also removes the information redundancy to a certain limit. Thus, CCA has received more attentions in pattern recognition (Liang *et al.* 1995; Zhou and Shen 2009; Lei *et al.* 2010; Jing *et al.* 2011; Yuan *et al.* 2011). In this article, we describe an algorithm for recognizing a pattern in DNA sequence based on multivariate statistical method, canonical correlation analysis or CCA and here we analysed DNA pattern recognition using canonical correlation algorithm.

2. Methodology

2.1 *Genomic sequence dataset*

We have taken *Homo sapiens* beta hemoglobin coding sequences (HBB) for proposed simulation. A hemoglobin molecule contains two alpha hemoglobin and two beta hemoglobin chains. The alpha as well as beta chains is coded by separate genes. The alpha hemoglobin gene is found on chromosome 16, and the beta hemoglobin gene is found on chromosome 11. A sequence of 444 nucleotide bases from beta hemoglobin sequence of the human genome is downloaded from GenBank with accession number NM_000518.4 from 51 to 494.

2.2 *Canonical correlation analysis*

Canonical correlation analysis is to find two sets of basis vectors for two sets of variables which on projections upon their respective basis vectors offer maximal correlation (Al-Kandari and Jolliffe 1997; Kursun *et al.* 2011). The underlying assumption is that the basis vectors X and Y will exist for variables x and y, $\{x(n) \in \mathbb{R}^{\mathrm{P}}, y(n) \in \mathbb{R}^{\mathrm{Q}}; n = 1, 2, \cdots, \mathrm{N}\}$ in such a way that the transformed projection upon X and Y are mutually maximized (Hotelling 1936). Thus, canonical correlation analysis finds which directions account for much of the covariance between two data sets.

Canonical correlation analysis computes two projection vectors, $W_x \in \mathbb{R}^P$, $W_y \in \mathbb{R}^Q$ to find the correlation between x and y, such that, the correlation is given as

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E\left[W_x^T x y^T W_y\right]}{\sqrt{E\left[W_x^T x x^T W_x\right] E\left[W_y^T y y^T W_y\right]}}$$
$$\rho = \frac{W_x^T C_{xy} W_y}{\sqrt{W_x^T C_{xx} W_x W_y^T C_{yy} W_y}} \quad (1)$$

C_{xy} is the covariance matrix of x and y. C_{xx} and C_{yy} are the dispersion matrices of x and y, respectively.

The maximum value of ρ with respect to W_x and W_y is the maximum canonical correlation or simply canonical correlation (CC):

$$\rho(x;y) = \max_{W_x, W_y} \frac{W_x^T C_{xy} W_y}{\sqrt{W_x^T C_{xx} W_x W_y^T C_{yy} W_y}} \quad (2)$$

To maintain the invariance of ρ subject to the scaling of vectors W_x and W_y, CC can be expressed as the following optimization problem:

$$\max_{W_x, W_y} \; W_x^T C_{xy} W_y$$
$$\text{subject to } W_x^T C_{xx} W_x = W_y^T C_{yy} W_y = 1 \quad (3)$$

Assuming C_{yy} as nonsingular, one can obtain W_{x} by solving the optimization problem

$$\max_{W_x} \; W_x^T C_{xy} C_{yy}^{-1} C_{yx} W_x$$
$$\text{subject to } W_x^T C_{xx} W_x = 1 \quad (4)$$

The eigenvectors corresponding to top eigenvalues are determined from the following generalized eigenvalue problem:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} W_x = \rho^2 W_x \quad (5)$$

The eigenvalues ρ^2 are the squared canonical correlations. The eigenvectors, W_{x} and W_{y} are the normalized canonical correlation basis vectors. The maximum number of canonical correlations are restricted to the minimum dimensionality of x and y. For example, if the dimensionality of x and y is 10 and 6 respectively, the maximum number of canonical correlations is 6.

2.3 *Nucleotide density*

CCA analyzes were done by introducing a w-wide window sliding in base-by-base manner along the sequence between position i = 1 to 444 region of *Homo sapiens* beta

hemoglobin gene (HBB). Window sliding is to find the density of nucleotide {A, C, G, T} in the sequence. The following definitions are helpful for the calculation of nucleotide density in the sequence.

Definition 1. Nucleotide Signature: For a sequence, the nucleotide signature S_k is the mapping with $b_k \in$ {A, C, G, T} where *i*-th bit in S_{ki}, is corresponding to the presence or absence of b_k.

Example 1. Consider a sequence, S = 'AACTCG'. The signatures of A, C, G, T in the sequence is S_A, S_C, S_G, S_T:

$$S_A = [1\ 1\ 0\ 0\ 0\ 0],\ S_C = [0\ 0\ 1\ 0\ 0\ 0],\ S_G = [0\ 0\ 0\ 0\ 0\ 1],\ S_T = [0\ 0\ 0\ 1\ 0\ 0]$$

Definition 2. Nucleotide Density: A sequence x[*n*] is transformed through mapping of the sequence into the output sequence y[*n*] via a weighted window *b* by means of the convolution summation as

$$y[n] = \sum_i b_i x[n-i] \tag{6}$$

b is independent of *x*[*n*] and *y*[*n*], where *n* is the base position. *y*[*n*] is the response of the transformation to input signal *x*[*n*]. The output is computed as a weighted, finite term sum, of previous and present input.

Example 2. Weighted output of S_A with the weighted window b = [0.2 0.1 0.3 0.4] is as follows:

$$S_A = [1\ 1\ 0\ 0\ 0\ 0]$$

$$y_A[n] = \sum_0^3 b_k S_A[n-k] \text{ with } b_0 = 0.2,\ b_1 = 0.1, b_2 = 0.3, b_3 = 0.4.$$

$$\Rightarrow y_A[n] = b_0 S_A[n] + b_1 S_A[n-1] + b_2 S_A[n-2] + b_3 S_A[n-3]$$

y_A = [0.2 0.3 0.4 0.7 0.4 0]; Similarly for other nucleotide viz., C, G, T, the output is obtained as

$$y_C = [0.2\ 0.0\ 0.2\ 0.1\ 0.5\ 0.5];\ y_G = [0.0\ 0.0\ 0.0\ 0.0\ 0.0\ 0.2];\ y_T = [0.0\ 0.0\ 0.0\ 0.2\ 0.1\ 0.3].$$

For nucleotide density calculation, evenly distributed window of unit value is considered. As explained, the output of the convolution summation represents the nucleotide density along the sequence.

3. Results

The validity and applicability of the proposed CCA are assessed on the basis of the simulated data. We have considered a predetermined pattern of DNA sequence as GGCCTGGCTCACCTGG having nucleotide density generated as $\{x(n) \in \mathbb{R}^4,\ n = 1, 2, \cdots, 16\}$, such that they have a relationship with *Homo sapiens* beta hemoglobin gene *HBB* (NM_000518.4 from 51 to 494) of length 444 bp. *HBB* sequence has nucleotide density distribution generated as $\{y(i) \in \mathbb{R}^4,\ i = 1, 2, \cdots, 444\}$. Here four dimension vector space is considered as an attribute of four nucleotides (A, C, G, T). Figure 1 shows the nucleotide density distribution along the pattern sequence (henceforth called as pattern set). The pattern shows enriched G nucleotide zone at the two end positions. The intermediate region of the pattern sequence (~ position 9 to 13) is enhanced with C nucleotide. Similarly the nucleotide density along the *HBB* sequence (henceforth called as test set) is shown in figure 2. A pattern matrix P (=x') of 4×16 dimension is constructed such that the (*k*,*i*) entry of P, includes the nucleotide density of *k*-th nucleotide, $b_k \in$ {A, C, G, T} at *i*-th position in the sequence. Similarly, test matrix of *Homo sapiens* beta hemoglobin sequence is prepared as Q (= y') of 4×444 dimension.

For the recognition of the pattern, we make 'sliding' of the pattern sequence of length 16 sites along the test sequence one-by-one positions. At each position, we perform the canonical correlation analysis between P and Q. In figure 3, the correlation between first two pairs of canonical variates (U_1, V_1 and U_2, V_2) are shown at base pair position, *i*=1. Figure 4 shows the correlation between first two pairs of canonical variates at position *i*=429. Analogous to the scatter plot of canonical variates over the test sequence starting at *i*=1, correlation between first two pairs of canonical variates at position *i*=429 also shows the same type of scattering.

For the searching of pattern on the *HBB* sequence, we evaluate first and second correlation (r_1 and r_2) base-by-base, for the first and second component pair, respectively.

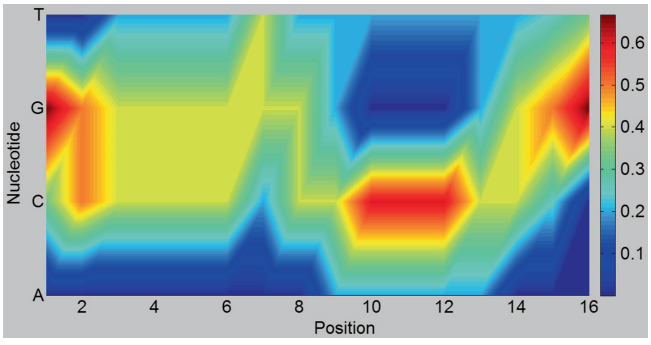

Figure 1. Distribution of the nucleotide density along base position on the pattern sequence.

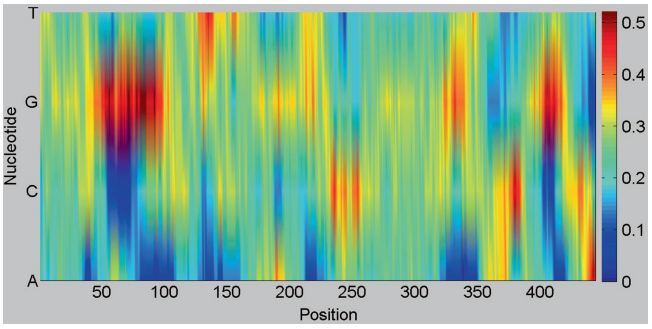

Figure 2. Distribution of the nucleotide density along base position on the *Homo sapiens* beta hemoglobin sequence.

We have calculated values of r_1 and r_2 with window size, w=3, 5 and 10. The variations of correlations between pattern and test sequence for the entire 444 bp region are shown in figures 5–7 with different window sizes. Scatter plot of first two pairs of canonical variates at 223 bp position is shown in figure 8. We notice that the values of first two pairs of canonical variates at 223 bp position maintain a high correlation among themselves.

For further application of CCA we have applied the method of pattern recognition in case of retroviral integration sites in the human genome (Schröder *et al.* 2002; Wu *et al.* 2003; Mitchell *et al.* 2004). Schröder *et al.* analysed the integration site in the human genome by human immunodeficiency virus type 1 (HIV-1). They infected human lymphoid and generated 689 HIV-1 clones in human SupT1 cells, which are deposited in GenBank with accession no. BH609398 to BH610086. We took these sequences from GenBank and consequently, and used the NCBI BLAST program to find the integration sites to the human genome. The BLAST program takes HIV-1 clone as a query sequence and searches it against the entire database of sequences maintained at NCBI. The output is obtained as 'hits' and these are combined into a Seq-annotation. Among all these hits, we considered DNA/assembled sequences that were found in the chromosome region. The sequences were aligned to their integration site, and each base position was numbered according to distance from integration site. The numbering of the base position started from the first genomic base 3′ to the viral integration point, labelled as offset 0 and flanked by 500 bases 5′. The end side is denoted as offsets –500 through –1. Similarly base positions are flanked by 500 bases 3′ on the other side as offsets 1 to 500. As before, the nucleotide density distribution was generated for the sequence. Figure 9 shows canonical correlations along the sequence between –500 and +500 base positions based on 1-gram word calculations with different window width. Figure 10 depicts canonical correlations along the sequence based on 2- gram word calculations with different window width.

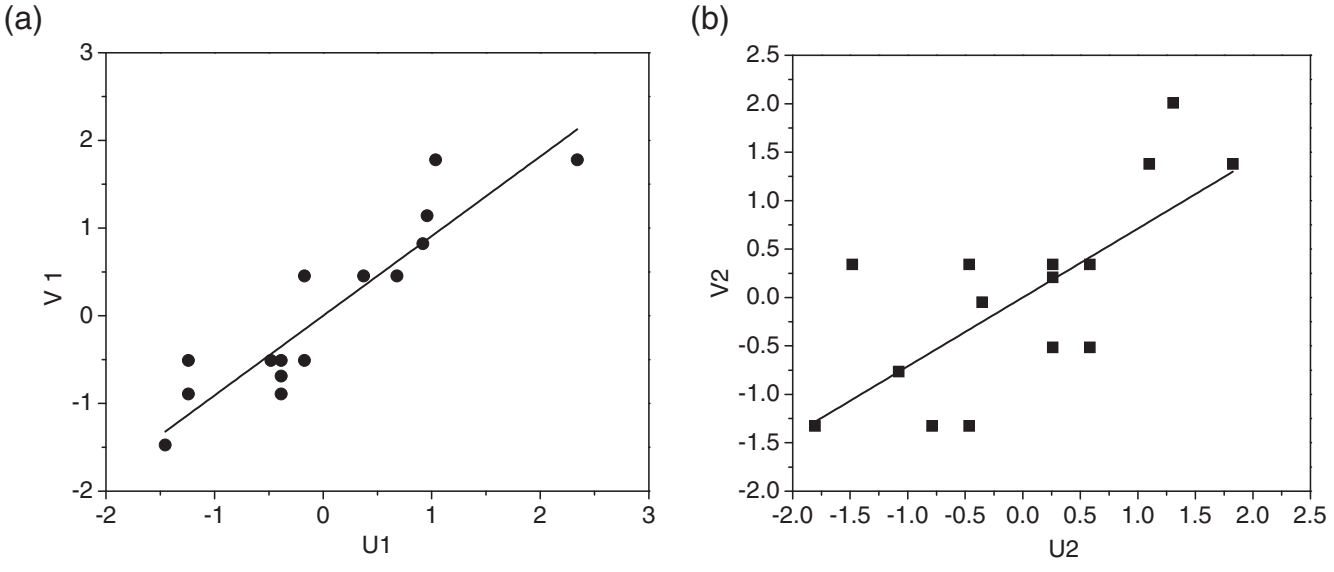

Figure 3. (**a**) Scatter plot of first pair of canonical variates at base position n=1 with regression (solid line). (**b**) Scatter plot of second pair of canonical variates at base position n=1.

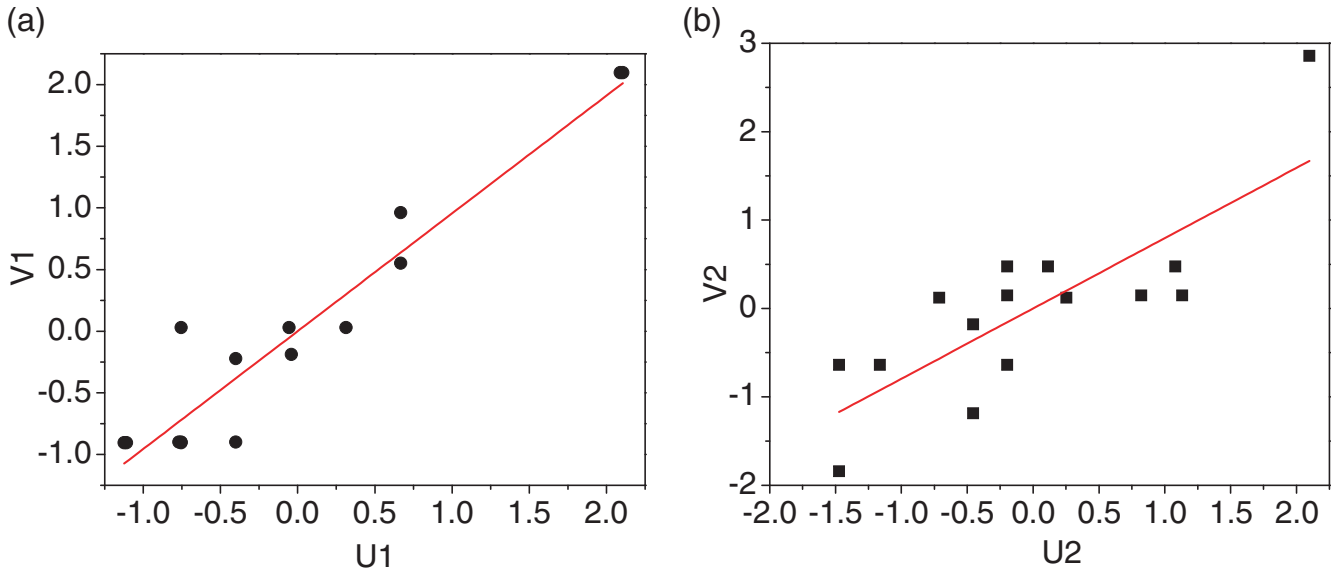

Figure 4. Scatter plot of (**a**) first pair and (**b**) second pair of canonical variates at base position n=429.

4. Discussion

We used CCA in two cases for pattern recognition, one in the presumed model investigation in *HBB* sequence and another in the case of integration site searching in the human genome by human immunodeficiency virus type 1 (HIV-1). In the first case, a comparison was done with the test set (*HBB* sequence) and the training set (ideal pattern). At two extreme points (i = 1 and 429) of the *HBB* sequence (figures 3 and 4), it shows that the pair of canonical variates follow a scattered clustering which clearly indicates that the pattern is not existing at these positions. Figures 5–7 show the first and second correlation dispersion along the sequence with window width of 3, 5 and 10 units. It is evident

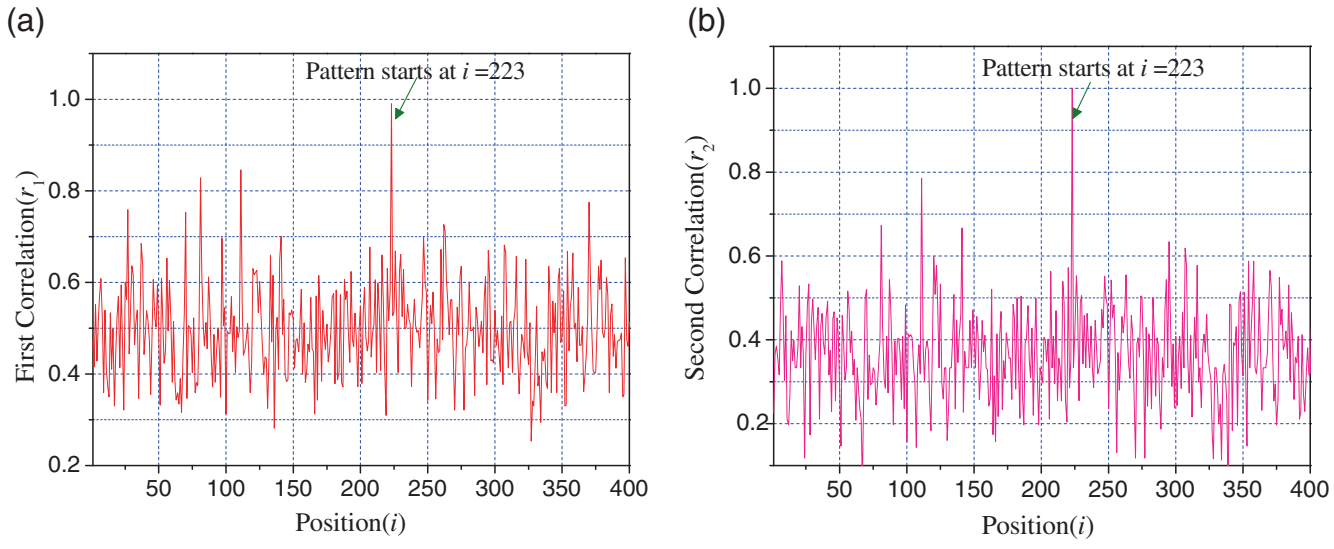

Figure 5. Canonical correlations on test set along the 1-429 base with a peak at n=223 (correlation of 0.991 (**a**) and 1.0 (**b**)) with window=3.

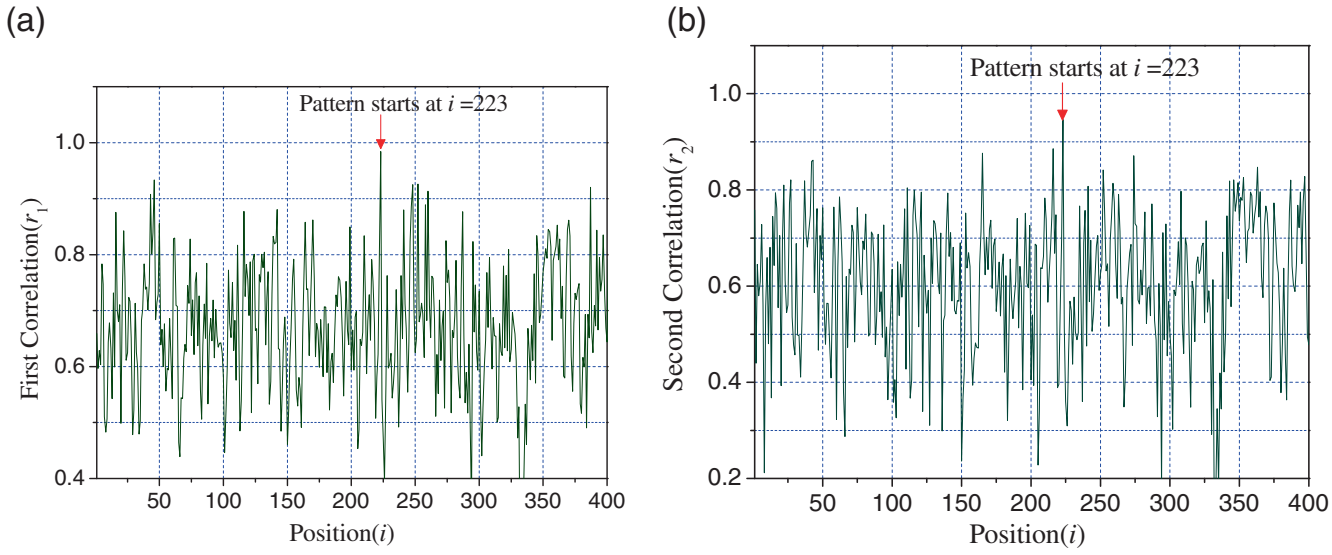

Figure 6. Canonical correlations on test set along the 1-429 base with a peak at n=223 (correlation of 0.984 (**a**) and 0.952 (**b**)) with window=5.

in each figure that the highest peak of the correlation distribution appears at the base position *i*=223 with a value >0.94, which indicates the presence of the pattern on the test sequence starting from the base position *i*=223. It is worthy to be mentioned that the peak height decreases with the increase of window size. Basically, window size of unity makes one-to-one nucleotide comparison for learning the correlated function. On the other hand, larger window size results in loss of significance of the correlations because it shows significant base preferences proximal to the site of assessment and little liking distant from it. Figure 8a–b provides a strong support for the existence of the pattern at the base position *i*=223. Scatter plot of two pair of canonical variates demonstrates good regression relation and validate the existence of the pattern on the test sequence at 223 base positions.

For the validation of the CCA method, we considered large data sets and performed a detailed analysis. We have applied CCA analyses to explore integration site in the human genome by human immunodeficiency virus type 1 (HIV-1). CCA was performed by making 'sliding' of 2*w*-wide window along the test sequence one-by-one position along *x*=–500 to +500 region, a total of 1001 base positions with 20 sequences kept in row-wise. Similar work is reported by Gumus *et al.* (2012). But they used two overlapping windows with an adverse gap between them for sequence dynamics study. In the present work one window was halved into two parts and overlapping, and was not considered for simplicity. Two halves of the window (each with width *w*) were taken as the input for the correlation calculations. The window sliding was to check whether the relationship at a particular position was sound enough to be more than the relationships found somewhere else. Correlation distributions with different window width are shown in figures 9 and 10. In the correlation analysis, the 20 HIV-1 sequence sets revealed significant preferences at offsets in the vicinity of the integration site. During integration, both the 3 ends of the viral DNA were introduced into the host DNA separated by few bases, being contingent on the viral species (Goodarzi *et al.* 1995, 1997).

To investigate the highly significant base preferences surrounding the integration sites of HIV-1, we implemented statistical validation. Twenty sequences in the vicinity of the integration point were considered. The dataset was randomized, and 11 samples were generated in each base position with offset from –5 to +5 flanking the viral insertion point. A randomization test was performed based on ANOVA to determine whether there was a base preference in the vicinity of the integration site. The randomization was to ensure that all bases are permitted at all locations; none is categorically required or even proscribed. In doing so, it was assumed that structural features ideal for interaction

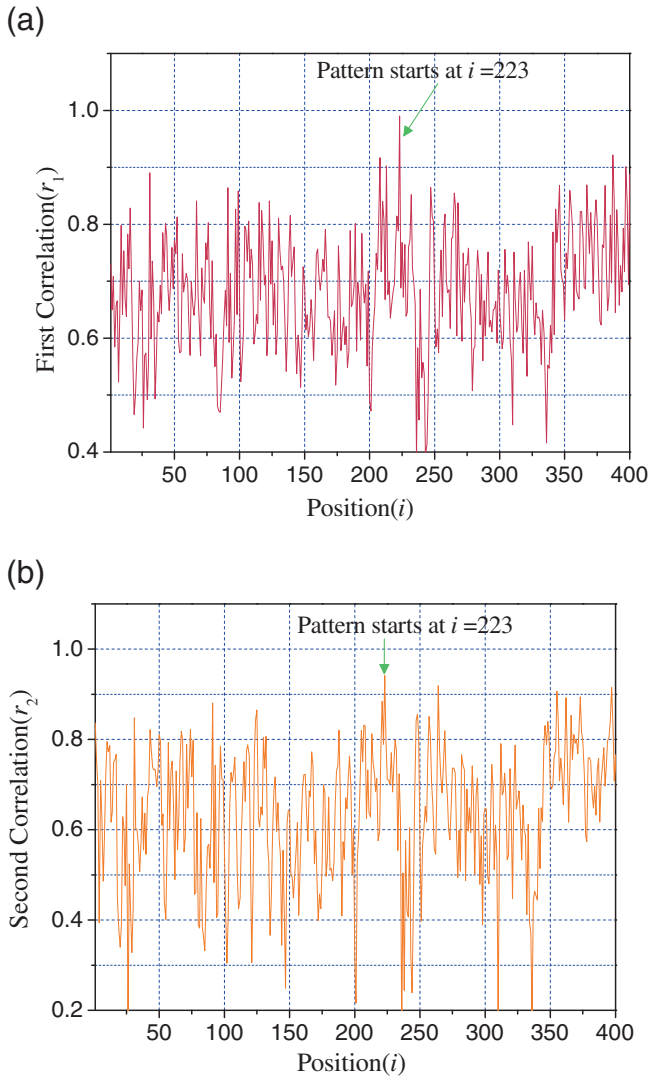

Figure 7. Canonical correlations on test set along the 1-429 base with a peak at n=223 (correlation of 0.982 (**a**) and 0.941 (**b**)) with window=10.

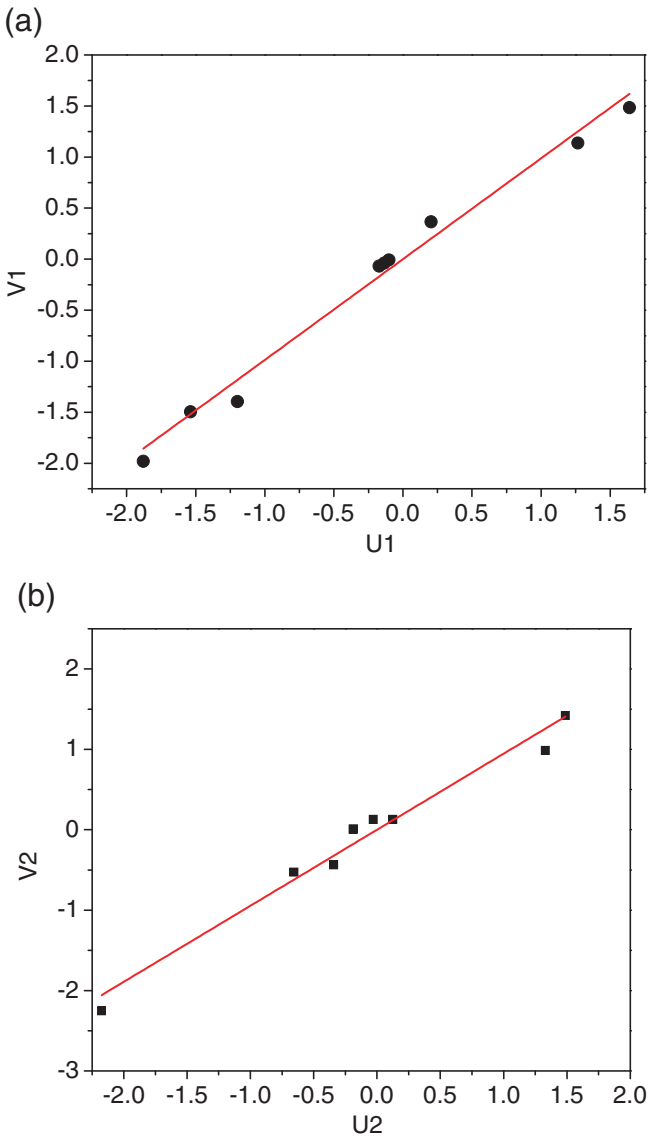

Figure 8. Scatter plot of (**a**) first pair and (**b**) second pair of canonical variates at base position n=223.

with integration site is determined by the DNA primary sequence. However, ANOVA determined F-distribution with *P*-values to check whether correlations (viral preference patterns) still exist in different sequences. The F-distribution with *P*-values is shown in table 1.

It should be kept in our mind that CCA maximizes the correlations between two sets of features obtained from the same semantic pattern, and subsequently extracts the effective discriminant feature. So, CCA is an unsupervised linear abstraction from two sets. For further improvement of discriminative power of CCA features, one can use supervised algorithms (Sun *et al.* 2005, 2008; Peng *et al.* 2010), for example, generalized CCA (GCCA) (Sun *et al.* 2005), and discriminant CCA (DCCA) (Sun *et al.* 2008). These algorithms apparently enhance multimode recognition rates from different viewpoints. Nevertheless, CCA approaches only deal with two sets of features, which cannot accurately describe the correlations among more than two sets of features.

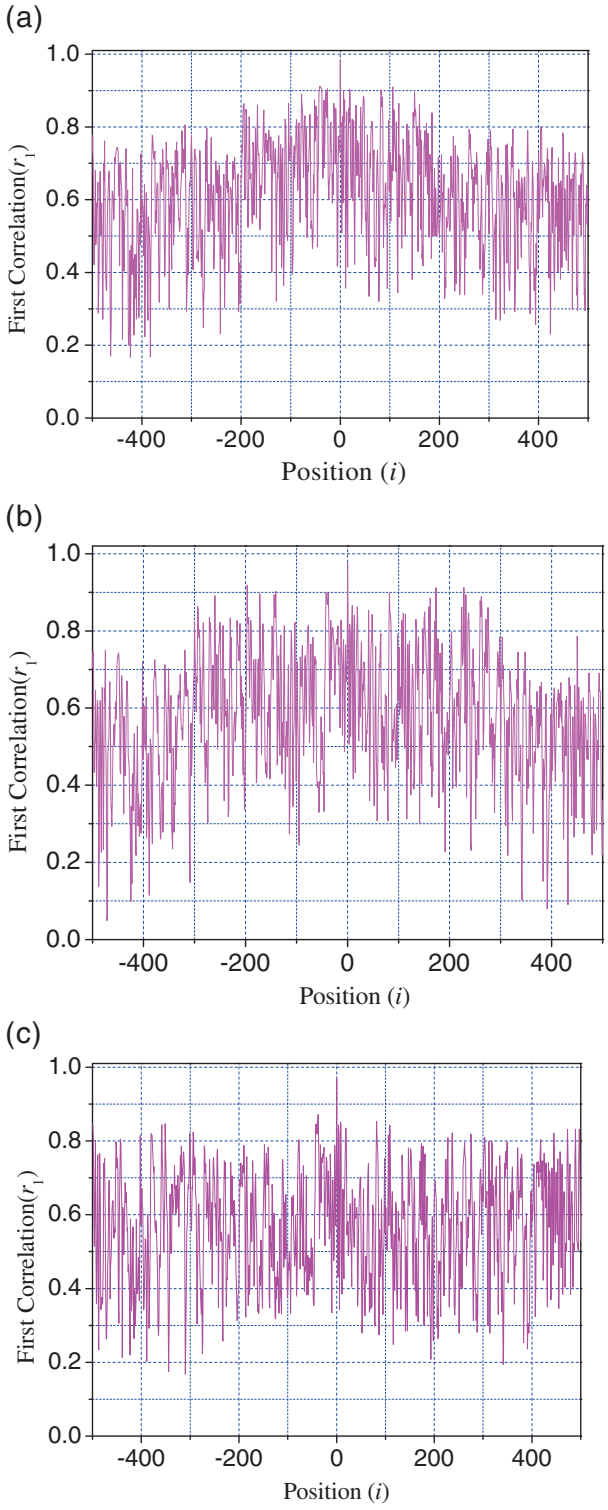

Figure 9. (**a**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.984). Calculation is done based on 1-gram word with window width 2w=6. (**b**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.978). Calculation is done based on 1-gram word with window width 2w=10. (**c**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.968). Calculation is done based on 1-gram word with window width 2w=14.

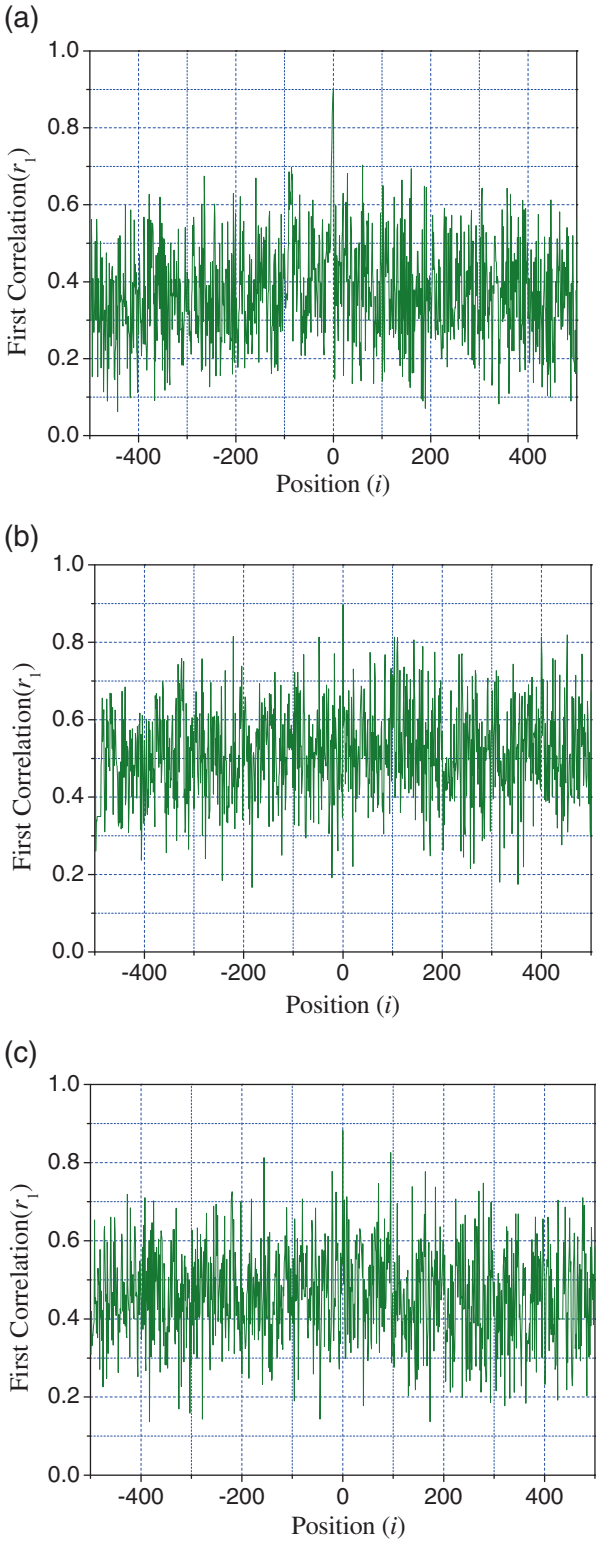

Figure 10. (**a**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.903). Calculation is done based on 2-gram word with window width 2*w*=6. (**b**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.897). Calculation is done based on 2-gram word with window width 2*w*=10. (**c**) Canonical correlations along the sequence between –500 and +500 base position with a peak at integration site (correlation of 0.882). Calculation is done based on 2-gram word with window width 2*w*=14.

Table 1. F-distribution and *P*-values (in parentheses) using 1-gram word calculation along offset from –5 to +5 flanking the viral insertion point at 0 base position

	Base position										
2*w*	–5	–4	–3	–2	–1	0	1	2	3	4	5
6	7.795 (0.185)	12.081 (0.111)	10.281 (0.135)	7.496 (0.193)	8.723 (0.164)	52.812 (0.021)	10.692 (0.129)	7.105 (0.205)	16.672 (0.075)	14.894 (0.086)	7.204 (0.202)
10	3.847 (0.045)	5.659 (0.015)	4.868 (0.023)	2.686 (0.101)	3.105 (0.074)	20.114 (0.0005)	4.037 (0.041)	2.797 (.093)	4.187 (0.036)	3.849 (0.045)	2.841 (0.092)
14	5.883 (0.013)	6.677 (0.008)	5.482 (0.016)	3.286 (0.066)	4.455 (0.038)	21.917 (0.0003)	5.751 (0.014)	3.588 (0.054)	4.391 (0.033)	4.294 (0.034)	3.144 (0.072)

5. Conclusions

We present the canonical correlation analysis applied on a sequence dataset to recognize a predetermined pattern on the sequence. CCA is an unsupervised numerical tool to find correlated functions over different sets of variables, and in its use of DNA sequence, the two sets can be sequence structures constructed from a pattern and a target sequence. The pattern having a relationship with *HBB* target sequence was considered for CCA test. *HBB* sequence of 444 bp length was deliberated as test sequence. CCA finds correlations between two observations of the same semantic pattern and test sequence. We found that there were significant correlations at 223 bp position, confirming the presence of the pattern starting from 223 bp position in the test sequence. Window width, which is a parameter for the determination of the nucleotide density, plays a crucial role in CCA design. For better understanding of small pattern recognition, small a window width is very useful. The standard CCA is very useful to find the linear relationship, and it is very sensitive to outliers due to its high dependence on the correlation coefficient. CCA very efficiently addressed integration mechanism in the human genome by HIV-1. The analysis has revealed very significant base preferences in the vicinity of the integration sites of HIV-1.

Acknowledgements

We acknowledge the support by Galgotias University through computing resources provided by the High-Performance Computing Facility.

References

Al-Kandari NM and Jolliffe IT 1997 Variable selection and interpretation in canonical correlation analysis. *Commun. Statist. Simulat. Comput.* **26** 873–900

Breiman L and Friedman JH 1985 Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **80** 580–598

Dickerson RE 1983 The DNA helix and how it is read. *Sci. Am.* **249** 94–111

Guha Thakurta D and Stormo GD 2001 Identifying target sites for cooperatively binding factors. *Bioinformatics* **17** 608–621

Gumus E, Kursun O, Sertbas A and Ustek D 2012 Application of canonical correlation analysis for identifying viral integration preferences. *Bioinformatics* **28** 651–655

Goodarzi G, Im GJ, Brackmann K and Grandgenett D 1995 Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase. *J. Virol.* **69** 6090–6097

Goodarzi G, Chiu R, Brackmann K, Kohn K, Pommier Y and Grandgenett DP 1997 Host site selection for concerted integration by human immunodeficiency virus type-I virions in vitro. *Virology* **231** 210–217

Hardoon DR, Szedmak S and Shawe-Taylor J 2004 Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16** 2639–2664

Hertz GZ and Stormo GD 1999 Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** 563–577

Holman AG and Coffin JM 2005 Symmetrical base preferences surrounding HIV-1, avian sarcoma/leucosis virus, and murine leukemia virus integration sites. *Proc. Natl. Acad. Sci. USA* **102** 6103–6107

Hotelling H 1936 Relations between two sets of variates. *Biometrika* **28** 321–377

Iaci R, Sriram T and Yin X 2010 Multivariate association and dimension reduction: a generalization of canonical correlation analysis. *Biometrics* **66** 1107–1118

Jing XY, Li S, Lan C, Zhang D, Yang JY and Liu Q 2011 Color image canonical correlation analysis for face feature extraction and recognition. *Signal Process.* **91** 2132–2140

Johnson RA and Wichern DW 1992 *Applied multivariate statistical analysis* 3rd edition (New-Jersey: Prentice Hall)

Kettenring JR 1971 Canonical analysis of several sets of variables. *Biometrika* **58** 433–451

Kursun O, Alpaydin E and Favorov O 2011 Canonical correlation analysis using within-class coupling. *Pattern Recogn. Lett.* **32** 134–144

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC 1993 Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214

Lei G, Zhou JL, Li X and Gong X 2010 Improved canonical correlation analysis and its applications in image recognition. *J. Comput. Inf. Syst.* **6** 3677–3685

Liang KH, Krus DJ and Webb JM 1995 K-fold crossvalidation in canonical analysis. *Multivar. Behav. Res.* **30** 539–545

Mitchell RS, Beitzel BF, Schröder AR, Shinn P, Chen H, Berry CC, Ecker JR and Bushman FD 2004 Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2** 1127–1137

Neuwald AF, Liu JS and Lawrence CE 1995 Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4** 1618–1632

Pabo CO and Sauer RT 1984 Protein-DNA recognition. *Annu. Rev. Biochem.* **53** 293–321

Peng Y, Zhang D and Zhang J 2010 A new canonical correlation analysis algorithm with local discrimination. *Neural. Process. Lett.* **31** 1–15

Schröder AR, Shinn P, Chen HC, Berry JR, Ecker JR and Bushman F 2002 HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110** 521–529

Sun QS, Liu ZD, Heng PA and Xia DS 2005 A theorem on the generalized canonical projective vectors. *Pattern Recogn.* **38** 449–452

Sun T, Chen S, Yang J, and Shi P 2008 A novel method of combined feature extraction for recognition, in *Eighth IEEE International Conference on Data Mining, ICDM'08, IEEE,* pp 1043–1048

Tenenhaus A and Tenenhaus M 2011 Regularized generalized canonical correlation analysis. *Psychometrika* **76** 257–284

Wu X, Li Y, Crise B and Burgess SM 2003 Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300** 1749–1751

Wu X, Li Y, Crise B, Burgess SM and Munroe DJ 2005 Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79** 5211–5214

Yu S, Yu K, Tresp V and Kriegel HP 2006 Multi-output regularized feature projection. *IEEE Trans. Knowl. Data Eng.* **18** 1600–1613

Yuan YH, Sun QS, Zhou QA and Xia DS 2011 A novel multiset integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recogn.* **44** 1031–1040

Zhou XC and Shen HB 2009 Regularized canonical correlation analysis with unlabeled data. *J. Zhejiang Univ. Sci. A*. **10** 504–511