BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 651–655
doi:10.1093/bioinformatics/bts027

Sequence analysis

Advance Access publication January 12, 2012

Application of canonical correlation analysis for identifying viral integration preferences

Ergun Gumus[1,*], Olcay Kursun[1], Ahmet Sertbas[1] and Duran Ustek[2]

[1]Department of Computer Engineering, Istanbul University, Istanbul 34320 Turkey and [2]Department of Genetics, Institute for Experimental Medicine, Istanbul University, Istanbul 34093 Turkey

Associate Editor: John Quakenbush

ABSTRACT

Motivation: Gene therapy aims at using viral vectors for attaching helpful genetic code to target genes. Therefore, it is of great importance to develop methods that can discover significant patterns around viral integration sites. Canonical correlation analysis is an unsupervised statistical tool that is used to describe the relations between two related views of the same semantic object, which fits well for identifying such salient patterns.

Results: Proposed method is demonstrated on a sequence dataset obtained from a study on HIV-1 preferred integration regions. The subsequences on the left and right sides of the integration points are given to the method as the two views, and statistically significant relations are found between sequence-driven features derived from these two views, which suggest that the viral preference must be the factor responsible for this correlation. We found that there are significant correlations at $x=5$ indicating a palindromic behavior surrounding the viral integration site, which complies with the previously reported results.

Availability: Developed software tool is available at http://ce.istanbul.edu.tr/bioinformatics/hiv1/

Contact: egumus@istanbul.edu.tr

Received on November 15, 2011; revised on December 26, 2011; accepted on January 9, 2012

1 INTRODUCTION

Advances in genome sequencing technologies provided scientists determine target sites of viruses in human genome. This concept has brought along an innovative step called 'gene therapy'. In gene therapy, connection parts of viruses to genome, which are called as LTRs (Long Terminal Repeats), are used for linking helpful genetic code to target gene. However, in order to understand linking characteristics of such a vector to a target gene, first, a cell of host genome must be infected with a virus having proper LTR. This methodology has been applied to one of the most popular gene therapy challenges, HIV-1 (Human Immunodeficiency Virus Type-1). Schroder *et al.* (2002) infected a human lymphoid cell line with HIV-1-based vector and obtained 524 chimeric sites. After transcriptional profiling they showed a correlation between gene activity and viral insertion preference. A hot spot, covering a region of 2.5 Kb, was reported to include only 1% of these sites.

After sequencing phase, virus-specific patterns (motifs) can arise in chimeric sequences. In order to detect such common patterns, a numerous types of motif search approaches have been proposed. Hertz *et al.* (1999) described a series of components to determine alignments of multiple sequences. They derived log-likelihood scoring schemes and then implemented a greedy algorithm for finding alignments of functionally related sequences. In another study, GuhaThakurta *et al.* (2001) presented a Gibbs sampling-based (Lawrence *et al.*, 1993) method to discover and weigh binding site patterns in DNA sequences. Holman *et al.* (2005) and Wu *et al.* (2005) manually identified symmetrical/palindromic behavior between the frequencies of bases on each side of integration sites in HIV-1 consensus sequences reported by Schroder *et al.* (2002). In this study, in order to find such common sequence patterns automatically, we propose a method based on the well-known statistical tool called 'Canonical Correlation Analysis' (CCA).

CCA seeks correlated functions (covariates) of two different, but related, views (i.e. two sets of related variables; Hardoon *et al.*, 2004). The availability of such correlated functions of the two views of the same semantic object is likely to exist due to an underlying factor responsible for the correlation. In relation to the problem of finding patterns around viral integration sites, we are given DNA sequences targeted by the virus as the semantic object; two nearby windows placed on each sequence can be used as the two views, where the fragments falling in windows are represented by some sequence-driven features such as base/dimer frequencies or a more sophisticated one such as moment descriptors; and the correlation of the two functions over these views is expected to be due to the viral preference as the factor responsible for this correlation.

CCA has proved to be a powerful method in previous studies especially in GWAS (Genome Wide Association Studies). Peng *et al.* (2010), have proposed a canonical correlation based U statistic metric to detect single nucleotide polymorphism (SNP)-based gene–gene co-association in two sample case-control datasets. They explored its Type-I error rate through simulations on two real datasets. In another study, Naylor *et al.* (2010) have implemented CCA to detect genetic associations between SNPs and gene expression levels. By this methodology, they claim to reduce the amount of necessary comparisons. For large-scale genomic studies, Parkhomenko *et al.* (2009) suggested using a sparse implementation of CCA. They divided their dataset into small subsets of variables belonging to different types and performed variable selection by maximizing the correlation between these subsets. CCA-based

*To whom correspondence should be addressed.

analysis can be applied to protein datasets as well, such as for searching correlated functions between moment-based features, autocorrelation, composition, transition or distribution (Li *et al.*, 2006; Shi *et al.*, 2006). In this study, adding to this wide variety of its applications, we utilize CCA in searching viral integration patterns on the sequences reported by Schroder *et al.* (2002).

2 MATERIALS AND METHODS

2.1 Genomic sequence dataset

Schroder *et al.* (2002) reported a total number of 688 sequences that were deposited in GenBank with accession numbers from BH609398 to BH610085. These sequences are human parts of chimeric sequences oriented in same direction (5′–3′) relative to the virus. We determined chromosomal locations regarding 5′ sides (integration point) of these sequences by using online BLAST (Zhang *et al.*, 2000) function of MATLAB (Mathworks) called 'blastncbi'. Then from each side of detected locations, 500 bases belonging to human genome were downloaded using function 'getgenbank' that was also present in MATLAB. Downloaded sequences were aligned according to integration point $x=0$.

As reported in Schroder *et al.* (2002), some of these sequences did not yield high-quality matching scores to genome and got excluded from the study. Among remaining ones, we chose a total number of 231 sequences that match to genome only with 'plus/plus' strand. A list of accession numbers corresponding to the sequences used can be found at the web address given in 'Availability'.

2.2 CCA

CCA describes the linear relation between two multidimensional (or two sets of) variables as the problem of finding basis vectors for each set such that the projections of the two variables on their respective basis vectors are maximally correlated (Hardoon *et al*., 2004; Izenman *et al.*, 2008; Kursun *et al.*, 2011). These two sets of variables may correspond to different views of the same semantic object. For example, in searching the viral preference patterns, these two views can be taken as the left and right subsequences next to the integration point. Among the common examples from other application domains, audio versus video, binocular vision, text versus images in documents can be listed.

Let x and y be two u-dimensional real-valued variables of two sets ($x, y \in \mathrm{R}^u$), then the canonical correlation between x and y is defined as:

$$\rho(x;y)=\sup_{f,g}\mathrm{corr}(f^T x; g^T y) \tag{1}$$

where, corr(x;y) stands for Pearson's correlation coefficient. If we have N observations from each set then we can group them in X and Y as $X=[x_1, x_2, \ldots, x_N]^T$ and $Y=[y_1, y_2, \ldots, y_N]^T$. Here, we can calculate within-sets ($\mathbf{C}_{XX}$, $\mathbf{C}_{YY}$) and between-sets ($\mathbf{C}_{XY}=\mathbf{C}_{YX}{}^T$) covariance matrices as:

$$\mathbf{C}(X,Y)=\begin{pmatrix}X\\Y\end{pmatrix}\begin{pmatrix}X\\Y\end{pmatrix}^T=\begin{bmatrix}\mathbf{C}_{XX} & \mathbf{C}_{XY}\\ \mathbf{C}_{YX} & \mathbf{C}_{YY}\end{bmatrix} \tag{2}$$

We can calculate canonical correlation which is the maximized value of ρ with respect to basis sets f and g as seen in Equation 3.

$$\rho(X;Y)=\sup_{f,g}\frac{f^T\mathbf{C}_{XY}g}{\sqrt{f^T\mathbf{C}_{XX}f\;\;g^T\mathbf{C}_{YY}g}} \tag{3}$$

By simply solving the eigenproblem given in Equations 4 and 5, we can determine projection vectors f and g that results in the maximum correlation of $\rho(X;Y)=\sqrt{\lambda}$.

$$\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}f=\lambda f \tag{4}$$

$$g=\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}f \tag{5}$$

2.3 Sequence features

Four sets of sequence features are used: Base Frequency (BF), Dimer Frequency (DF), Mean Base Position (MBP) and Variance of Mean Base Positions (VMBP) (Deng *et al.*, 2011; Shi *et al.*, 2006). We have an alphabet of bases α = {A, C, G, T} and K sequences each with length L_k. Base Frequencies (b_i^k) corresponding to each sequence can be calculated as:

$$b_i^k=\frac{1}{L_k}\sum_{j=1}^{L_k}X_{i,j}^k,\quad i=1,2,3,4\ \text{ and }\ k=1,\ldots,K \tag{6}$$

where j denotes base position in each sequence, and

$$X_{i,j}^k=\begin{cases}1 \text{ if } \alpha_i \text{ is present at position } j \text{ in sequence } k\\ 0 \text{ if } \alpha_i \text{ is not present at position } j \text{ in sequence } k\end{cases}$$

In a same manner Dimer Frequencies (d_i^k) with alphabet β = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT} can be calculated as:

$$d_i^k=\frac{1}{L_k-1}\sum_{j=1}^{L_k-1}Y_{i,j;j+1}^k,\quad i=1,\ldots,16\ \text{ and }\ k=1,\ldots,K \tag{7}$$

where j denotes dimer position in each sequence, and

$$Y_{i,j}^k=\begin{cases}1 \text{ if } \beta_i \text{ is present at position}[j:j+1] \text{ in sequence } k\\ 0 \text{ if } \beta_i \text{ is not present at position}[j:j+1] \text{ in sequence } k\end{cases}$$

As for the moment-based features, after finding base positions $X_{i,j}^k$, Mean Base Position (m_i^k) and Variance of Mean Base Positions (v_i^k) can be calculated as:

$$m_i^k=\frac{1}{S_i^k}\sum_{j=1}^{L_k}X_{i,j}^k j,\quad i=1,2,3,4\ \text{ and }\ k=1,\ldots,K \tag{8}$$

$$v_i^k=\frac{1}{S_i^k}\sum_{j=1}^{L_k}X_{i,j}^k(j-m_i^k)^2, i=1,2,3,4\ \text{ and }\ k=1,\ldots,K \tag{9}$$

where j denotes base position and $S_i^k=\sum_{j=1}^{L_k}X_{i,j}^k$.

While forming feature vectors of base (or similarly, dimer frequencies), due to the linear dependency among them (their sum is equal to 1.0), we omit one (e.g. the last) feature from the set. Thus, we obtain $b^k=[b_1^k b_2^k b_3^k]$ and $d^k=[\,d_1^k d_2^k \ldots d_{15}^k]$ for BF and DF, respectively.

CCA is expected to find a function (a set of weights for the projection vector), $A=[a_1 a_2 \ldots a_{N-1}]^T$, corresponding to base/dimer frequencies $D=[d_1 d_2 \ldots d_{N-1}]^T$ in the left window that correlates with another function $B=[b_1 b_2 \ldots b_{N-1}]^T$ of base/dimer frequencies $E=[e_1 e_2 \ldots e_{N-1}]^T$ in the right window:

$$D^T A \approx E^T B \tag{10}$$

Real weights $\overline{A}$ and $\overline{B}$ can be calculated from CCA functions A and B as seen in Equation 11.

$$\overline{a}_i=\frac{1-a_i}{\sqrt{\sum_{i=1}^{N}(1-a_i)^2}}$$

$$\overline{b}_i=\frac{1-b_i}{\sqrt{\sum_{i=1}^{N}(1-b_i)^2}} \tag{11}$$

where, $i=1\ldots N$ and assuming $a_N=b_N=0$ and N is the number of features ($N=4$ for bases and 16 for dimers).

2.4 Proposed method for identifying viral integration preferences

In application of CCA for identifying viral integration preferences, a pair of (nearby) windows, as seen in Fig. 1, is placed on fragment bunches. To represent the sequence fragments in these windows, sequence-driven features listed in Section 2.3 are extracted. Then the maximal correlation between functions of these two views is computed using CCA. Here, we introduce two parameters: *window-width* (w) and gap between windows.

Adjusting optimal window width is a heuristic task that we examined with various values in Section 3. For example, in search of small patterns, small values for w are needed to be used. On the other hand, choosing w too big may cause same repetitive regions fall into adjacent windows. This would yield a high but dummy correlation especially for basic sequence features like base or dimer frequencies (due to low frequency components). Also setting the gap too big might lead to lose all local dependencies, yielding no correlation.

CCA analyses were performed by placing two w-wide windows with a distance of gap and sliding them in base-by-base manner along $x=-500$ to $+500$ region, a total of 1000 bases (excluding $x=0$). This sliding is to find all interesting points in the sequences and also for checking whether the correlations at a particular point is well above the correlations found elsewhere, thus, serving as a way of statistically testing the correlation found in one point with 1000 others. First, on each possible location x, a separate CCA was trained to find maximally correlated functions of used features between the two input windows placed on the training set. Then, the extracted functions were tested on test set to check whether the correlation was truly present.

For developing the analysis programs, we used MATLAB with its bioinformatics toolbox and CCA implementation of (Borga *et al.*, 2001).

3 EXPERIMENTAL RESULTS

In its current implementation, the proposed method requires two predetermined parameters (windows size and gap between windows) and utilizes four sets of sequence-driven features (see 'Materials and Methods' section for details). Depending on the viral preference pattern of interest, these parameters need to be adjusted and the correct type of features set (or feature set combinations) should be chosen. The genomic sequence dataset of (Schroder *et al.*, 2002) was used in order to demonstrate the potential of the proposed method in identifying viral integration preferences. For statistical validation, the dataset was randomly split into two halves (with 116 and 115 fragments) to obtain the training and the test sets. For 5×2 cross-validation, we repeated this process five times and used each half once for training and once for testing. The training set was used for learning the correlated functions and the test set was used to check whether these correlations (viral preference patterns) still hold on unseen fragments.

As the dataset has been used in previous work reporting short segments of weak palindromic behavior near the integration sites (Holman *et al.*, 2005; Wu *et al.*, 2005), we used small window sizes (w) and negative gaps. Using a negative gap means the two windows overlap and the non-overlapping portions of the windows would account for the palindromic behavior. For such a small dataset and short window-width considered, the best choices of sequence-features were base and dimer frequencies. Using just the base frequencies, as seen in Table 1 and in Figs 2 and 3, for small windows with $w=2$ and $w=4$, high and significant correlations of 0.74 and

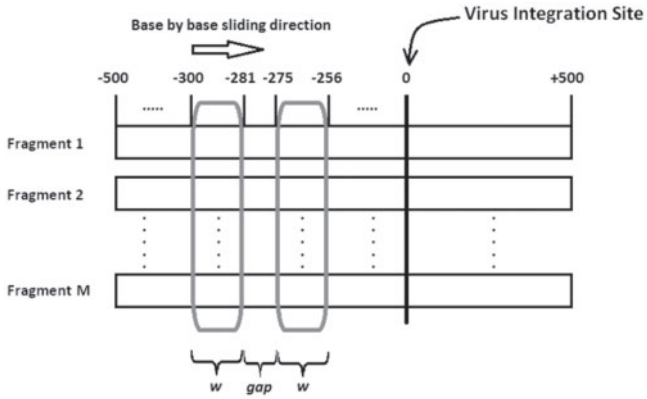

Fig. 1. A sample iteration of CCA in sliding window manner ($w=20$, gap $=5$, window-pair is placed at both sides of center base position $x=-278$).

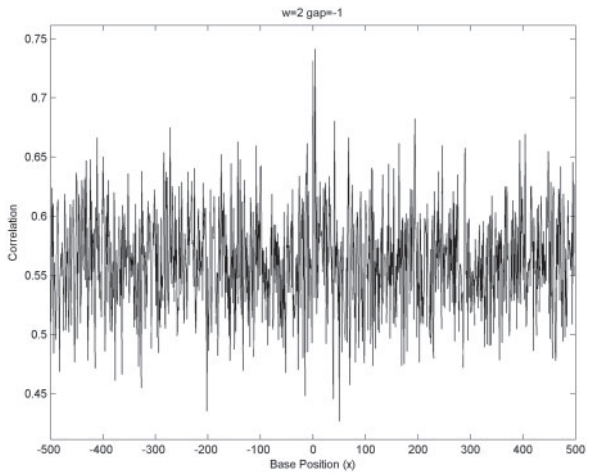

Fig. 2. Canonical correlations on test set along the $[-500, +500]$ region ($w=2$, gap $=-1$) with a peak at $x=5$ (correlation of 0.74 and $P=0.00003$).

Table 1. Canonical correlations on test set (and *P-value*s of z-test) using base frequencies at the short $[-5, +6]$ segment flanking the viral insertion point at $x=0$

w	gap	$x=-5$	$x=-4$	$x=-3$	$x=-2$	$x=-1$	$x=+1$	$x=+2$	$x=+3$	$x=+4$	$x=+5$	$x=+6$
2	−1	0.55 (0.9419)	0.56 (0.9162)	0.54 (0.7570)	0.49 (0.1290)	0.54 (0.7417)	0.73 (0.00008)	0.64 (0.0434)	0.60 (0.3197)	0.53 (0.5205)	0.74 (0.00003)	0.61 (0.2103)
4	−3	0.81 (0.2926)	0.76 (0.2196)	0.81 (0.3332)	0.85 (0.0096)	0.85 (0.0069)	0.86 (0.0035)	0.83 (0.1005)	0.89 (0.00002)	0.82 (0.2319)	0.86 (0.0015)	0.84 (0.0189)
6	−5	0.87 (0.7834)	0.86 (0.7944)	0.92 (0.0020)	0.89 (0.1753)	0.84 (0.1237)	0.88 (0.3194)	0.85 (0.3965)	0.87 (0.9663)	0.89 (0.1135)	0.90 (0.0374)	0.90 (0.0356)
8	−7	0.90 (0.5590)	0.93 (0.0492)	0.91 (0.5261)	0.91 (0.4056)	0.90 (0.9607)	0.93 (0.0206)	0.91 (0.6340)	0.93 (0.0257)	0.92 (0.2772)	0.93 (0.0303)	0.93 (0.0590)

The z-test uses all the correlations in the longer segment of $[-500, +500]$. Cells with $P<0.05$ are highlighted.

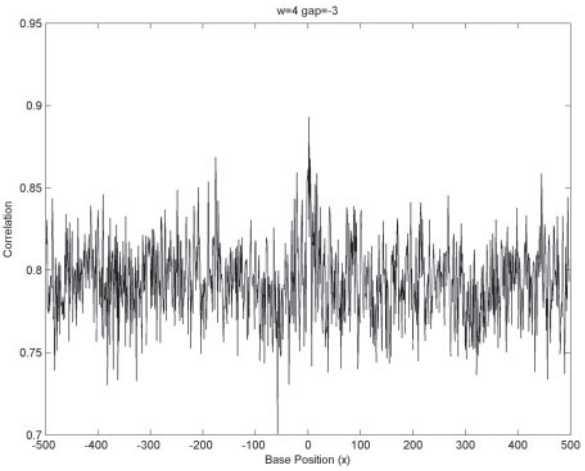

Fig. 3. Canonical correlations on test set along the [−500, +500] region ($w=4$, gap $=-3$) with a peak at $x=3$ (correlation of 0.89 and $P=0.00002$).

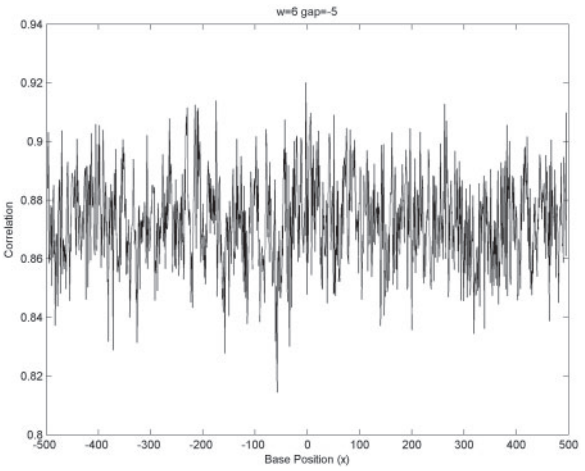

Fig. 4. Canonical correlations on test set along the [−500, +500] region ($w=6$, gap $=-5$) with a peak at $x=-3$ (correlation of 0.92 and $P=0.0020$).

0.89 are present at $x=5$ and $x=3$, respectively. Here, the symbol x denotes the center of the palindromic pattern; that is, one window is placed from bases $\lfloor x-w+1-\frac{\text{gap}}{2} \rfloor$ to $\lfloor x-\frac{\text{gap}}{2} \rfloor$ and the other window is placed from bases $\lfloor x+1+\frac{\text{gap}}{2} \rfloor$ to $\lfloor x+w+\frac{\text{gap}}{2} \rfloor$. For example, for $w=2$ and gap $=-1$, we have one window on the $[x-1, x]$ base-interval and the other one on $[x, x+1]$. We avoided the use of high degree k-mers because of their high dimensionality; e.g. there are 64 trimers, which can be considered high-dimensional given that there are only 116 samples in the training set). Moreover, when searching for larger patterns with larger windows, the moment features can also be utilized.

In order to make sure that these correlations found are not ordinary elsewhere, we applied z-test (Sprinthall, 2003) on them against the correlations obtained from all possible positions from $x=-500$ to $x=+500$. Increasing the window width results in loss of significance of the correlations (as seen in Fig. 4); implying that the symmetrical behavior has a short span as stated in (Holman

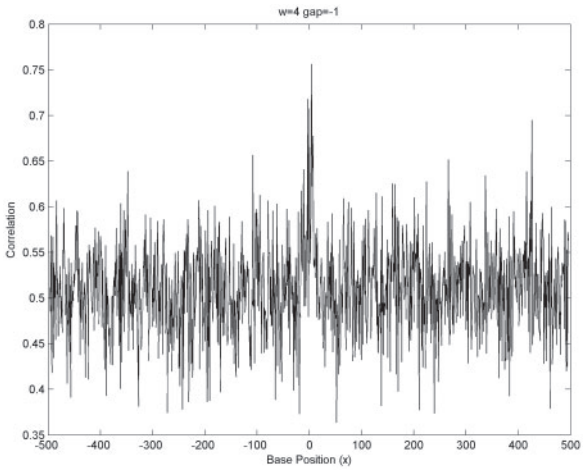

Fig. 5. Canonical Correlations on test set along the [−500, +500] region using dimer frequencies ($w=4$, gap $=-1$) with a peak at $x=5$ (correlation of 0.75 and $P=0.000001$).

et al., 2005; Wu *et al.*, 2005), roughly $2w+\text{gap}=7$ bases long. Using the dimer frequencies, shown in Fig. 5, with $w=4$ and gap $=-1$, again a high and significant correlation of 0.75 was obtained at $x=5$ ($P=0.000001$). Our experiments for searching longer patterns, using all features, including moments as well, could not identify any significant findings in this dataset.

4 CONCLUSIONS

We present a novel application of CCA for investigating common patterns near viral integration sites in genome. CCA is an unsupervised statistical tool that can find correlated functions over different sets of variables, in its application to genome, the two sets can be sequence features extracted from adjacent windows placed on target sequences. The existence of such covariate functions over these separate windows is expected to be due to the viral preference as the factor responsible for this correlation. In particular, in this study, CCA is demonstrated and tested on the integration site sequences given by Schroder *et al.* (2002). The method successfully finds significant viral preferences near the integration point similar to results reported by Wu *et al.* (2005). More specifically, it has been found that at $x=5$, there are significant correlations (with $P<0.05$) for various values of window size and gap parameters.

In the proposed method, there are three major parameters that affect the correlation between two windows: (i) sequence features used: base/dimer frequencies, moment descriptors, etc.; (ii) the window width (small widths for short patterns and vice versa); and (iii) gap between windows (positive for non-overlapping and negative for overlapping windows). Depending on the dataset used, the optimal choices for these parameters may change. For example, in order to search for short patterns, base or dimer frequencies is more suitable with small window width and gap as in the HIV-1 dataset used in our study. On the other hand, for structural similarities where correlations between more distant areas are important (Shi *et al.*, 2006) moment features may be preferred.

Especially with better features such as trimers and moment-based features, the proposed method can be used in searching more complex and longer patterns. New sequence features can be easily

integrated into the software, which is made publicly available at the web address given in 'Availability'. String kernels (Liao *et al.*, 2003) can also be utilized for extending the analysis to the non-linear case, which can be viewed as an application of KCCA (Kernel CCA) to this problem. As future work, we will also incorporate more advanced statistical tests into the software.

Conflict of Interest: none declared.

REFERENCES

Borga,M. and Knuttson,H. (2001) A canonical correlation approach to blind source separation. *Technical Report LiU-IMT-EX-0062*, Department of Biomedical Engineering, Linköping University, Sweden.

Deng,M. *et al*. (2011) A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One,* **6**, e17293 PMID 21399690.

GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.

Hardoon,D. *et al*. (2004) Canonical correlation analysis: an overview with application to learning methods. *Neural Computation,* **16**, 2639–2664.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Holman,A.G. and Coffin,J.M. (2005) Symmetrical base preferences surrounding HIV-1, avian sarcoma/leucosis virus, and murine leukemia virus integration sites. *Proc. Nal Acad. Sci. USA*, **102**, 6103–6107.

Izenman,A.J. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics,* 1st edn. Springer, New York.

Kursun,O. *et al*. (2011) Canonical Correlation Analysis Using Within-class Coupling. *Pattern Recogn. Lett.*, **32**, 134–144.

Lawrence,C.E. *et al*.(1993) Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Li,Z.R. *et al*. (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **34** (Suppl. 2): W32–W37.

Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, **10**, 857–868.

MATLAB version 7.10.0. Natick, Massachusetts: The Math Works Inc., 2010.

Naylor,M.G. *et al*. (2010) Using canonical correlation analysis to discover genetic regulatory variants. *PLoS ONE*, **5**, e10395.

Parkhomenko,E. *et al*. (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–34.

Peng,Q. *et al*. (2010) A gene-based method for detecting gene-gene co-association in a case-control association study. *Eur. J. Hum. Genet.*, **18**, 582–587.

Schroder,A.R. *et al*. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell,* **110**, 521–529.

Shi,J. *et al*. (2006) Prediction of protein subcellular localizations using moment descriptors and support vector machine. In *Lecture Notes in Computer Science*, vol. 4146/2006, Springer, Heidelberg, pp. 105–114.

Sprinthall,R.C. (2003) *Basic Statistical Analysis,* 7th edn. Pearson, Boston.

Wu,X. *et al*. (2005) Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.,* **79**, 5211–5214.

Zhang,Z. *et al*. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.,* **7** (1–2), 203–214.