BRIEF COMMUNICATION

2D random walk representation of *Begonia* × *tuberhybrida* multiallelic loci used for germplasm identification

I. WIESNER* and D. WIESNEROVÁ

Institute of Plant Molecular Biology, Group of Molecular Biodiversity, Biology Centre AS CR,
Branišovská 1160/31, CZ-37005 České Budějovice, Czech Republic

Abstract

In this study, we wanted to inspect whether the evolutionary driven differences in primary sequences could correlate, and thus predict the genetic diversity of related marker loci, which is an important criterion to assess the quality of any DNA marker. We adopted new approach of quantitative symbolic DNA sequence analysis called DNA random walk representation to study multiallelic marker loci from *Begonia* × *tuberhybrida* Voss. We described significant correlation of random walk-derived digital invariants to genetic diversity of the marker loci. Specifically, on the 3D-contour plot of multivariate principal component analysis (PCA), we revealed statistical correlation between the first two PCA factors and the number of alleles per marker locus. Based on that correlation, we suggest that DNA walk representation may predict allele-rich loci solely from their primary sequences, which improves current design of new DNA germplasm identificators.

Additional key words: bioinformatics, information entropy, Markov chain, primary sequence analysis, principal component analysis.

Mathematical descriptors of DNA sequences and their use in rationalizing biological properties encoded in DNA follow from the structure-property similarity principle and represent a newly emerging field of bioinformatics (Nandy *et al.* 2006).

Visual methods illustrate how DNA sequences are read and provide alternative approach to understanding of the underlying genomic information. By handling purine and pyrimidine bases as the elements of a discrete evolutionary-derived time signal, various digital signal processing techniques can be employed for bioinformatic analysis of DNA sequences (Bai *et al.* 2007). Using DNA walk representations and by applying the frequency domain transformations, non-random behavior of primary sequences is revealed. Such DNA walk representation can be used to extract useful non-trivial characteristics from sequence data, which are then applicable to evolutionary sequence comparisons or delineation of yet-unknown genetic function and diversity (Berger *et al.* 2004).

There are several methods to convert DNA sequences into the digital sequences with various statistical approaches like DNA walk digitization with root-mean-square fluctuation, information entropy, Fourier transform or wavelet analysis (Peng *et al.* 1992, Tsonis *et al.* 1996, Dodin *et al.* 2000). These techniques were developed for the multidimensional mapping from the 2-D Cartesian space to quite complex 6-D space (Berger *et al.* 2004). This digitization approach opened quite new way for the quantitative analysis of evolutionary and functional links to yet unrevealed genetic function solely on the information encoded within the primary DNA sequence. Any set of sequences may be analyzed this way whether they are orthologs or be related only by their application like a group of genetic markers, which is also the case of our study.

DNA markers and germplasm identificators are for a long time routinely used in identification and genetic analysis of crop germplasm collections (Griga *et al.* 2007, Cordeiro *et al.* 2008). Therefore, it is of the general

Received 15 July 2008, *accepted* 9 April 2009.

Abbreviations: PCA - pricipal component analysis.

Acknowledgements: This study was supported by the program of Targeted Research DP3-PP2 of the National Research Program, project No. 1QS500510566 of the Academy of Sciences of the Czech Republic, and by the project AV CR No. VZ-AVOZ 50510513. We thank Jan Wiesner for technical assistance.

* Corresponding author: e-mail: wiesner@umbr.cas.cz

economic interest to understand the hidden nature of marker polymorphism (Li *et al.* 2008). Especially, the question frequently arises about how genetic characteristics of marker loci like their observed heterozygosity (H_o) or allelic composition might be related to their DNA primary sequence.

Here, we adopted new approach of quantitative symbolic DNA sequence analysis called DNA random walk, on the multiallelic marker loci which we have isolated recently for effective identification of *Begonia* × *tuberhybrida* Voss. germplasm (Wiesner and Wiesnerova 2008). In this study, we wanted to inspect whether the evolutionary driven differences in primary sequences could correlate, and thus predict, the genetic diversity of related marker loci, which would be an important criterion to assess the quality of any DNA marker.

Genetic diversity, *i.e.* number of alleles per a locus for analyzed 18 *Begonia* marker loci, was obtained previously (Wiesner and Wiesnerova 2008) from the genetic analysis of the panel of 62 ornamental cultivars and breeding genotypes of *Begonia* × *tuberhybrida* Voss. maintained at the *BEGOBIO* germplasm collection of Biology Centre AS CR České Budějovice, and in the Begonia collection of *Sempra Flora*, Holice, Czech Republic. Analyzed DNA primary sequences of all 18 marker loci were published in GenBank under the following accession numbers: D83opa01 (3 alleles, EF606692); D89opa02 (2 alleles, EF606693); D67opa03 (3 alleles, EF606694); D62opa04 (4 alleles, EF606695); D40opa05 (3 alleles, EF606696); D31opa07 (2 alleles, EF606698); D46opa11a (3 alleles, EU035986); D46opa11b (4 alleles, EU035986); D57opa13 (2 alleles, EF606698); D43opa14 (4 alleles, EF606699); D48opa16 (2 alleles, EF606700); D22opa18 (2 alleles, EF606701); D49opa19a (4 alleles, EU035987); D49opa19b (6 alleles, EU035987); D91opab05a (2 alleles, EU035990); D91opab05b (2 alleles, EU035990); D77opx05 (2 alleles, EU035988); D80opz20 (4 alleles, EU035989).

Statistical computations like single-factor *ANOVA* and multivariate principal component analysis (PCA) were performed using *Statistica 8.0* (*StatSoft Inc.*, Chicago, USA). Necessary calculations for DNA random walk representation were performed using the script written by authors in *Perl 5.8* equipped with *BioPerl* toolkit (Stajich *et al.* 2002).

In order to generate digital random walk representation of DNA primary sequences of 18 marker loci from *B.* × *tuberhybrida* genome (Wiesner and Wiesnerova 2008), we divided according to Bai *et al.* (2007) four bases A, T, C, G into the three transform classes t (where t = RY or t = MK or t = WS) with [-1, +1] digital mapping defined as follows:

1) If t = RY transform (purine ↔ pyrimidine) and RY refers to:
R = {A, G}
Y = {C, T}
then R-Y DNA walk is defined as:
Y_i^{RY} = +1, if X_i^{RY} ε R; or
Y_i^{RY} = -1, if X_i^{RY} ε Y,
where X_i^{RY} ε seq (X_i) denotes a base X_i from the i-th position of DNA sequence seq (X_i)

2) If t = M-K transform (amino ↔ keto group) and MK refers to:
M = {A, C}
K = {G, T}
then M-K DNA walk is defined as:
Y_i^{MK} = +1, if X_i^{MK} ε M; or
Y_i^{MK} = -1, if X_i^{MK} ε K,
where X_i^{MK} ε seq (X_i) denotes a base X_i from the i-th position of DNA sequence seq (X_i)

3) If t = W-S transform (weak ↔ strong H-bonds) and WS refers to:
W = {A, T}
S = {G, C}
then W-S DNA walk is defined as:
Y_i^{WS} = +1, if X_i^{WS} ε W;
or Y_i^{WS} = -1, if X_i^{WS} ε S,
where X_i^{WS} ε seq (X_i) denotes a base X_i from the i-th position of DNA sequence seq (X_i) .

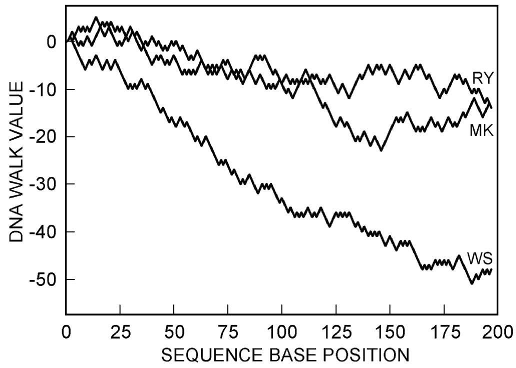

Fig. 1. RY, MK, and WS DNA random walk curves of D49opa19b locus. Along the sequence base position (X_i), *i*-th DNA walk value Y_i^t (Y_i^t ε {-1;+1}), and t ε {RY;MK;WS}) directs the next step. RY refers to purine↔pyrimidine transform; R = {A,G}; Y = {C,T}; MK refers to amino↔keto group transform; M = {A,C}; K = {G,T}; WS refers to weak↔strong H-bond transform; W = {A,T}; S = {G,C}.Y_i^t DNA walk values around zero values indicate random sequence (here, in the sequence of the locus D49opa19b, RY walk is close to randomness) while increasing negative distance from zero level during DNA walk along X_i indicates significant bias in sequence (here WS walk) given by the prevalence of GC bases (Y_i^{WS} = -1) over AT bases (Y_i^{WS} = +1).

Then, the generation of X_i^{RY}, Y_i^{MK}, and Y_i^{WS} values was called DNA random walk of a point jumping one unit of the negative or the positive direction (y-axis) successively along DNA sequence base after base ($X_{1,} X_{2\,...} X_{i-1,} X_i$ on x-axis) with starting point [x_0;y_0] = 0 (see Fig. 1 for the example).

It was proven earlier (Bai *et al.* 2007) that digitized sequence Y^t produced by [-1,+1] digital mapping, *i.e.* Y^{RY}, Y^{MK} or Y^{WS} derived from primary DNA sequence is also homogenous Markov chain, which means that any Y_i^t mapping value is uniquely determined from its mother primary DNA sequence and have no relationship with other Y_i^t values.

By the above definition of three types of digital mapping we could generate three different random walk curves corresponding to the same DNA primary sequence according to three mappings t ~ RY, MK, and WS. Fig. 1 illustrates 2D DNA random walk representation of the *Begonia* D49opa19b locus with six alleles.

Because this way generated random walk curve represents a homogenous Markov chain, we could further quantified these walk curves by the unidimensional invariant of average information content expressed as the transition information entropy H_p:

$Hp = -1/L \, \Sigma \, p_i \ln(p_i)$

where p_i -the transition probability in a digitized sequence, L - the length of [-1,+1] transformed sequence. Individual transition probabilities p_i: [-1 → -1], [-1 → +1], [+1 → -1], [+1 → +1] were then calculated as the following conditional probabilities:

$P^{RY}_{[-1 \to -1]} = P\{Y_{i+1} = -1 | Y_i = -1\}$

$P^{RY}_{[-1 \to +1]} = P\{Y_{i+1} = -1 | Y_i = +1\}$

$P^{RY}_{[+1 \to -1]} = P\{Y_{i+1} = +1 | Y_i = -1\}$

$P^{RY}_{[+1 \to +-1]} = P\{Y_{i+1} = +1 | Y_i = +1\}$

Transition probabilities of MK and WS transforms were calculated in the similar way.

In addition, we used mean square deviation $D(Y^t)$ of a random walk for each of three Y^t transforms (t = RY, MK or WS) as another numerical invariant to characterize digitized DNA sequences:

$D(Y^t) = [1/L \, \Sigma \, (Y_i^t - \mu(Y_i^t))]^{½}$

where $\mu(Y_i^t)$ is the mean of DNA random walk t over all the length of primary sequence and L is the length of [-1,+1] transformed sequence.

Combining all 12 one-step transition probabilities [RY(-1,-1); RY(-1,1); RY(1,-1); RY(1,1); MK(-1,-1); MK(-1,1); MK(1,-1); MK(1,1); WS(-1,-1); WS(-1,1); WS(1,-1); WS(1,1)] with three transition entropies [H_p(RY); H_p(MK); H_p(WS)], and three mean square deviations [Dev(RY); Dev(MK); Dev(WS)] of random walk sequence representation, each of 18 primary sequences of *Begonia* marker loci could be represented by 18 invariants written in the form of 18-component vectors.

All data were then represented by 18 × 18 matrix, the panel of 18 random walk invariants generated for each of 18 multiallelic locus sequence. Using this matrix, we first verified overall statistical significance of differences between these 18 loci according to new random walk invariants using *ANOVA F*-statistics.

Single-factor *ANOVA* was performed with the factorization according to the number of alleles per a locus in that as the low-allelic were classified the loci with 0 - 3 alleles while the high-allelic were those loci with 4 - 6 alleles per a locus. *ANOVA* prerequisite of homogeneity of variances was confirmed by Levene's test, which allowed to continue with *ANOVA* itself, which confirmed the differences between loci described by 18-component vectors ($F_{(1,16)} = 9.5886$, $P = 0.0069$). We found that especially the invariants of RY and WS transition probabilities significantly differ when compare between low-allelic *versus* high-allelic group of loci.

In the next step, we calculated sequence dissimilarity (distance) matrix and applied the multivariate method of principal component analysis (PCA) to compare 18 marker loci by DNA walk representation in more detail. Results of PCA analysis are given in Fig. 2, where PCA factor 1 (x-axis) explained 25.0 %, and factor 2 (y-axis) 18.9 % of total variability among *Begonia* marker loci represented by 18 random DNA walk invariants.

We further analyzed possible interpretations of complex PCA factor 1 and PCA factor 2 in relation to original 18-component vector. Even if interpretation of individual invariants is often difficult in multivariate space, we revealed the statistically significant correlations. Particularly, we found that PCA factor 1 was correlated to WS mapping (*i.e.* classification between weak H-bond, W = {A,T}, and strong H-bonds, S = {G,C}). Mean square deviation, Dev(WS) was correlated to PCA factor 1 with negative correlation coefficient $r = -0.8032$ significant at $P = 0.0154$. This indicates that the higher is PCA factor 1, the lower is AT content or the higher is GC content in a primary sequence.

In similar way, we revealed that PCA factor 2 was correlated to vector components related to MK mapping (i.e. classification between amino group M = {A,C}, and keto group K = {G,T}. Namely, mean square deviation, Dev(MK) was correlated to PCA factor 2 with positive correlation coefficient $r = +0.6997$ significant at $P = 0.0092$. This indicates that the higher is PCA factor 2, the higher is also AC content or the lower is GT content in a primary sequence.

In order to reveal possible correlations of newly obtained random walk invariants to genetic diversity of the marker loci, we constructed 3D-contour plot (Fig. 2). 3D-contour plot of PCA factor 1 (*x*-axis), PCA factor 2 (*y*-axis), and number of alleles per a locus (*z*-axis) revealed statistical correlation between PCA factors and number of alleles per a locus. This important relationship was defined by multiple correlation coefficient $R(z/xy) = +0.7602$, statistically significant at $P = 0.0137$.

Contour plot of the number of alleles per marker locus (Fig. 2) may be interpreted within the framework of two ultimate situations. First, if smooth linear relationship would exist, then we would see the continuous color gradient from green to dark red on the contour plot without any local (relative) extremes. In the opposite situation where no relationship would ever exist, we would see that the contour lines can not be constructed at all using distance weighted least squares approximation method.

Clearly, our actual contour plot (Fig. 2) represents the case in between of both extremes. On the left lower corner of the contour plot (lower values of factor 1 and 2), there are located low-allelic loci while we can see the trend of high-allelic loci to be located on the upper right corner of this contour plot. The degree of nonlinearity in this relationship is indicated by the presence of local minima (light green areas) and local maxima (dark red areas) of the number of alleles per a marker locus.

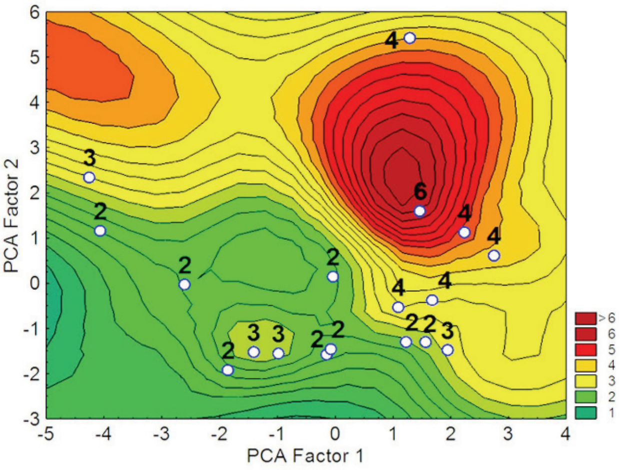

Fig. 2. 3D contour plot of the number of alleles per a locus against principal component analysis (PCA) factor 1 and 2. Contour plot was constructed by distance weighted least squares approximation. Legend coding refers to color areas to which the loci with identical number of alleles per a locus are mapped.

From these findings we can conclude that the DNA walk representation, as the novel method of DNA primary sequence digitization, yielded good quantitative measures of evolutionary-driven distance (dissimilarity) between the primary sequences of marker loci and that these measures (invariants) as the individual characteristics of a primary sequence could be statistically significantly correlated to the degree of allelism in related *Begonia* marker loci.

Similarly to our results, it was possible, using DNA walk representation method of sequence digitization, to successfully sort out allelic sequences of neuraminidase alleles of H5N1 virus either into benign or to highly pathogenic group, and to differentiate clearly mutational changes related to malignant allele variants (Nandy *et al.* 2007).

Our study thus contributes to the growing up general confirmation that sequence digitization by DNA walk representation method is the useful approach for identification of functional structures hidden in primary sequence lettering. Further, we suggest that DNA walk representation may function as an efficient pre-scanning procedure, which can predict allele-rich genomic loci as highly informative DNA markers solely using the information from their primary sequence.

References

Bai, F-L., Liu, Y-Z., Wang, T-M.: A representation of DNA primary sequences by random walk. - Math. Biosci. **209**: 282-291, 2007.

Berger, J.A., Mitra, S.K., Carli, M., Neri, A.: Visualization and analysis of DNA sequences using DNA walks. - J. Franklin. Inst. Eng. Appl. Math. **341**: 37-53, 2004.

Cordeiro, A.I., Sanchez-Sevilla, J.F., Alvarez-Tinaut, M.C., Gomez-Jimenez, M.C.: Genetic diversity assessment in Portugal accessions of *Olea europaea* by RAPD markers. - Biol. Plant **52**: 642-647, 2008.

Dodin, G., Pierre, V., Levoiretal, P.: Fourier and wavelet transform analysis: a tool for visualizing regular patterns in DNA sequences. - J. theor. Biol. **206**: 323-326, 2000.

Griga, M., Horáček, J., Klenotičová, H.: Protein patterns associated with *Pisum sativum* somatic embryogenesis. - Biol. Plant. **51**: 201-211, 2007.

Li, H., Zhang, S.G., Gao, J.M., Wang, C.G., Zhang, Y., Qi, L.W., Chen, L., Song, W. Q.: Development of a sequence characterized amplified region (SCAR) marker associated with high rooting ability in *Larix*. - Biol. Plant. **52**: 525-528, 2008.

Nandy, A., Harle, M., Basak, S.C.: Mathematical descriptors of DNA sequences: development and applications. - Arkivoc **9**: 211-238, 2006.

Nandy, A., Basak, S.C., Gute, B.D.: Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. - J. Chem. Inform. Model. **47**: 945-951, 2007.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E.: Long-range correlations in nucleotide sequences. - Nature **356**: 168-170, 1992.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne. B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E.: The Bioperl toolkit: Perl modules for the life sciences. - Genome Res. **12**: 1611-1618, 2002.

Tsonis, A.A., Kumar, P., Elsner, J.B., Tsonis, P.A.: Wavelet analysis of DNA sequences. - Phys. Rev. E **53**: 1828-1834, 1996.

Wiesner, I., Wiesnerova, D.: Sequence characterized markers from *Begonia* × *tuberhybrida* Voss. - Eur. J. hort. Sci. **73**: 244-247, 2008.