Self-organizing approach for automated gene identification in whole genomes

Alexander N. Gorban[1,2], Andrey Yu. Zinovyev[1,2],* and Tatyana G. Popova[1]
[1] *Institute of Computational Modeling RAS, 660036 Krasnoyarsk, Russia and*
[2] *Institut des Hautes Etudes Scientifiques, France*
(Dated: January 30, 2002)

An approach based on evolutionary consideration and very simple and clear idea of distinguished coding phase in explicit form for identification of protein-coding regions in whole genome has been proposed. For several genomes the optimal window length for averaging GC-content function and calculating codon frequencies has been found. It is shown that the structure of distribution of triplet frequencies in sliding window has symmetric bullet-like form with 4 clusters. Self-training procedure based on clustering in multidimensional space of triplet frequencies is proposed.

PACS numbers: 87.15.Cc, 87.10.+e, 87.14.Gg

Since inventing sequencing techniques one of the most interesting problems in DNA analysis is to find principles of computational (automated) distinguishing of coding (exons) and non-coding (junk and introns) regions in DNA.

Now most of the computational approaches for identification of coding regions in DNA have following limitations [1]: they need a training set of already known examples of coding and non-coding regions, they work with a comparably short subsequence of DNA rather than whole sequence and they are able to recognize mainly protein-coding regions.

Recently some approaches appeared which promise to be free of these limitations. In the works of Yeramian E. [2, 3] DNA sequence is considered as a linear chain of strong (GC-bond) and weak (AT-bond) hydrogen bonds. Calculating partition function one can obtain a thermal DNA stability map (a plot of probability of every DNA basepair to be disrupted). With appropriate temperature chosen, the map in some cases shows believable correlation with the arrangement of coding regions in DNA. This fact was exploited with some success to identify coding regions in Plasmodium falsiparum in some non-standard for gene-finders situations.

Another promising approach is to partition the DNA sequence into homogenious in some sense subsequences and to find such a way of partitioning that corresponds to the "coding – non-coding" one. In the works of Bernaola-Galvan et al.[4] method of entropic segmentation was formulated that uses difference in codon compositions in coding and non-coding regions. The hypothesis is that codon composition in coding regions is different from junk because of well-known fact of biasing in codon usage.

Methods of gene finding use a variety of numerical characteristics reflecting statistical regularities in some subsequence (window) of DNA. Inphase hexamers seem to be the most effective single measure (see, for example, [5]). Calculation of inphase hexamers are based on division of given subsequence into non-overlapping triplets and counting for every triplet dicodon occurences starting from the first, second and third position in the triplet.

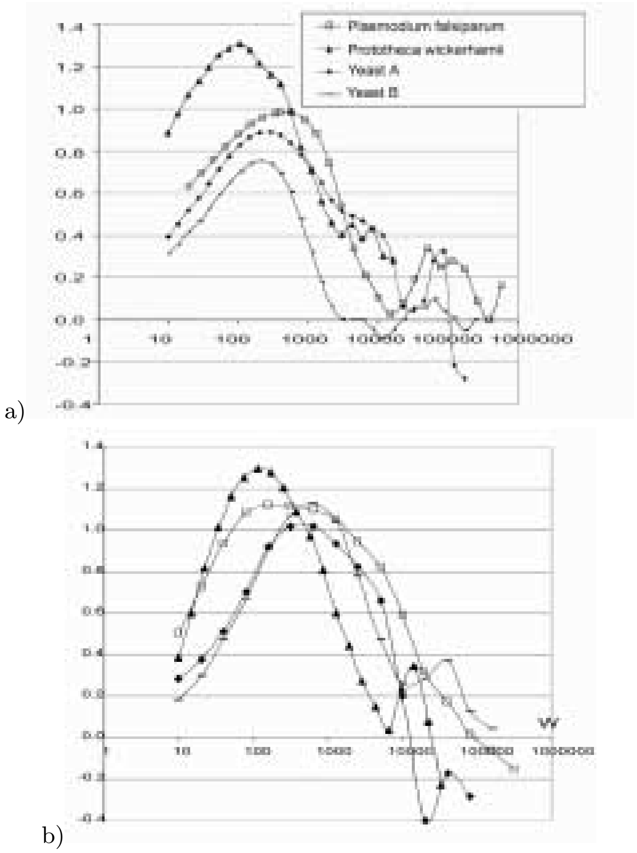

FIG. 1: Effectiveness of two measures (local GC-concentration (a), mixing entropy (b)) for several genomes. Bimodal character of graphs can be explained: first maximum is the difference of coding and non-coding regions themselves, second is statistical difference of long regions (isochores)

In this work we introduce a method for identification of protein-coding regions in DNA that uses notion of *distinguished coding phase*. We try to explain the reasons for measures which use in a way the idea of coding phase (including such measures as inphase hexamers, assymetry,

*Electronic address: gorban@icm.krasn.ru entropy etc., see [5, 6] for definitions) to be useful in the identification of protein coding regions. Using the idea in the *explicit* form we formulate self-training procedure for identification of protein-coding regions for whole genomes.

Let's take arbitrary subsequence of DNA (below we think of both DNA strands as of one chain, not touching problem of possible genes overlapping). Divide it into non-overlapping triplets of nucleotides and count frequencies of all possible triplets. We will call set of the frequencies *triplet usage.* There are three different ways of calculating triplet usage, starting from the first, second and third basepair in the window. So, we have 3 distributions of triplet frequencies $f^{(1)}_{ijk}$, $f^{(2)}_{ijk}$, $f^{(3)}_{ijk}$, $i, j, k \in \{A, C, G, T\}$. We will call them *triplet usage in three different phases.* Consider also mixed distribution $f^{(s)}_{ijk} = \frac{1}{3}(f^{(1)}_{ijk} + f^{(2)}_{ijk} + f^{(3)}_{ijk})$.

Suppose first that given subsequence is protein coding and homogeneous (without introns). Therefore one of the three distributions (phases) is the real codon distribution. We will call this set of triplet frequencies *distinguished coding phase.* Distribution of triplets in coding phase is strictly conserved in the process of evolution, because inserting or deleting one basepair leads to frameshift and all coding region become senseless. It means that three phases of triplet usage are not equivalent in protein coding region from evolutionary viewpoint, and, therefore, expected to be different.

The situation is fairly opposite in non-coding regions. Now operations of deleting and inserting a basepair in sequence are allowed, because it does not lead to crucial changes in DNA properties. Therefore all three distributions are equivalent and expected to be mixed and equal to $f^{(s)}_{ijk}$.

We think that such evolutionary viewpoint on the nature of differences in triplet distributions in coding and non-coding regions is not well understood. No matter what is the hypothetical nature of junk (or exons) is, we can determine if a window of DNA is coding, detecting presence of coding phase (one of the simplest method is proposed below). It gives the possibility of constructing self-training techniques for automatic gene recognizing. We underline here that the idea is not the same as in the methods which use the experimental fact of *codon bias.* Instead we use three observations: 1) information in protein-coding regions is coded by triplets; 2) this information is conserved during the evolutionary process; 3) junk regions are not very sensitive to deleting and inserting basepairs. These assumptions lead to the conclusion that protein-coding region must have distingushed coding phase in three different triplet distributions. And presence of the coding phase can be detected.

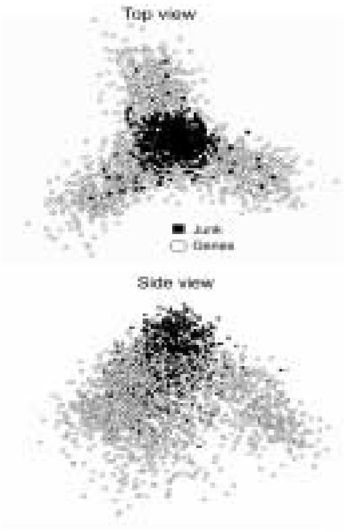

FIG. 2: Distribution of triplet frequencies in the subspace spanned by first three principal components for P.Wickerhamii. "Top view" means projection on the plane perpendicular to the symmetry axis of the bullet-like structure, "side view" is projection on a plane containing the axis.

It is worth noticing, that distributions $f^{(1)}_{ijk}$, $f^{(2)}_{ijk}$, $f^{(3)}_{ijk}$ are projections of distribution of pentamers f_{ijklm}, $i, j, k, l, m \in \{A, C, G, T\}$ which counted from every third position starting from the first basepair in a window. It means that information contained in distribution of pentamers seems to be sufficient for prediction of coding regions with the same accuracy as using hexamers (but requires shorter subsequence to evaluate frequencies).

Another interesting note is that GC-concentration in a window is the linear function of the frequencies of triplet distribution in any phase. It means that in the space of triplet (or pentamer, hexamer etc.) frequencies gradient of this functional determines a distinguished direction along what the separation of coding and non-coding windows is good. Really, for many genomes coding regions are GC-rich comparing to non-coding. We will show below that the difference in GC-concentration between coding and non-coding regions is most contrast at the scales comparable to the average gene length in genome.

To implement the simple idea of distinguished phase as a procedure for identification of protein coding regions in DNA, first we investigated dependence of effectiveness of two simple measures on the length of sliding window. It was done on the known genome anotations and it was shown that the dependence has bimodal character and is not very strong in some range of window lengths.

Suppose every nucleotide in one strand of DNA to be "coding" or "non-coding". Then assume that this property depends on some measure calculated over the whole window with length W, centered in the position of the nucleotide. We can evaluate the effectiveness of this measure for separation of coding subset **G** and non-coding subset **J** in the set of all nucleotides. Let $A_W(i)$ be our measure calculated for the i-th position and

$$\Delta_W = \frac{\frac{1}{W}(\sum_{i\in\mathbf{G}} A_W(i) - \sum_{i\in\mathbf{J}} A_W(i))}{\sqrt{DA_W(i)}}$$

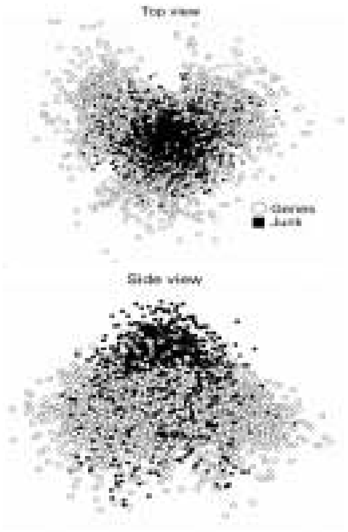

FIG. 3: Distribution of triplet frequencies in the space of first three principal components for S.Cerevisiae III

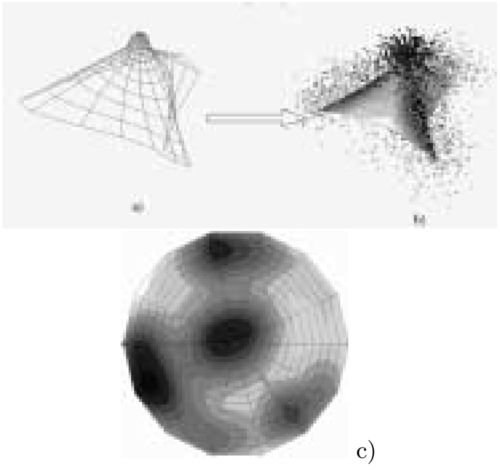

FIG. 4: Two-dimensional visualization of distribution density using method of elastic map: a) form of constructed elastic map; b) position of the map in data space (projection on the first three principal components); c) resulting picture of estimation of density distribution.

be a measure of effectiveness, where D is dispersion of $A_W(i)$ over the whole set $\{\mathbf{G}, \mathbf{J}\}$. In figures 1(a), 1(b) dependence of Δ_W on W for two measures and several genomes is shown. First (fig.2(a)) is local concentration of GC-bonds in a window. Second (fig.2(b)) is so called "mixing entropy" $S_M = \frac{1}{3}(3S - S^{(1)} - S^{(2)} - S^{(3)})$, where $S = -\sum_{ijk} f_{ijk}^{(s)} \ln f_{ijk}^{(s)}$, $S^{(m)} = -\sum_{ijk} f_{ijk}^{(m)} \ln f_{ijk}^{(m)}$. It is clear that sets $\mathbf{J}$ and $\mathbf{G}$ can be separated with confidence (with enormous number of points we have difference of two mean values more than one standard deviation) and effectiveness of S_M measure seems to be better then GC-average. An optimal window length for calculating the measures is about 400 bp for S.Cerevisiae and P.Falsiparum genomes and about 120 bp in the case of short mitochondrial genome.

Using these values we constructed a finite set of points in 64-dimensional space of triplet frequencies, each point corresponds to the frequencies distribution $f^{(1)}$ of non-overlapping triplets with phase 1 (starts from the first basepair in window). Then coordinates of points in the set $X = \{x_i\}, i = 1...N$ were normalized on unit standard deviation:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j},$$

where x_{ij} is the j-th coordinate of the i-th point and $\bar{x}_j$, σ_j are mean value and standard deviation of the j-th coordinate.

The set of normalized vectors $\tilde{x}_i$ was projected into the subspace spanned by the first three principal components of the distribution and visualized showing known separation for coding and non-coding nucleotides (see fig.2,3). The distribution has bullet-like structure with a kernel corresponding to the non-coding regions (where there is no distinguished phase) and three tails which correspond to the three possible shifts of real codon distribution to the phase of test triplets in a window.

To visualize density of the distribution more advanced technology was used called "method of elastic maps"(see [7–10]). The method of elastic maps just like self-organizing maps (SOMs) [11] constructs point approximation to the non-linear principal 2D-surface using minimization of elastic energy functional that consist of three parts describing "node – data points" attraction and energies of stretching and binding of the net of nodes with predefined topology. More isometric than in SOM net of nodes allows to construct piece-wise linear 2D-manifold and to project data points in a piece-wise linear fashion onto it, then using the manifold as a 2D screen for visualization purposes. In our case we initialized the net on the 2D-hemisphere put into multidimensional space. After that it was deformed using algorithm of construction elastic net for the optimal approximation and half-tone image was used to visualize the resulting density of projections of data points (more precisely, its non-parametric estimation). The distribution of data points has 4 clusters (fig.4), corresponding to the non-coding regions (central cluster) and protein-coding (three peripherial clusters).

Using this fact the procedure for unsupervised prediction of protein coding regions can be formulated. One can prepare distribution of triplet frequencies just as we did it (using some suboptimal value of window length) and then cluster it for 4 clusters, using appropriate clustering algorithm. It gives separation of all nucleotides into non-coding (0-phase) and protein-coding (1,2,3-phase).

We used simplest method of K-means [12] for clustering and found that separation of nucleotides in investigated genomes relates to the known data with accuracy from 65% up to 85% (calculating accuracy as percentage of correctly predicted nucleotides - coding and non-coding). Though these results are comparable with performance of gene-finders used in real practice [1],

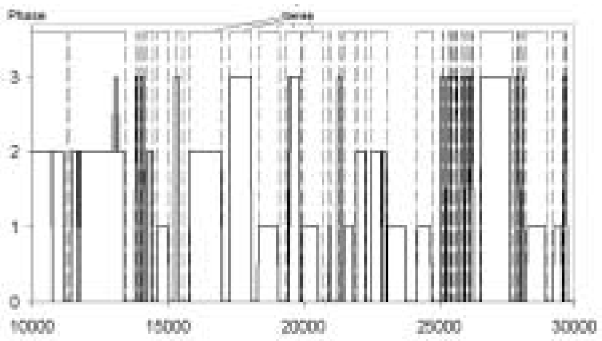

FIG. 5: Prediction of protein-coding regions using clustering in the space of triplet frequencies for P.Wickerhamii. X-axis is basepair position in sequence, Y-axis is number of cluster (coding phase).

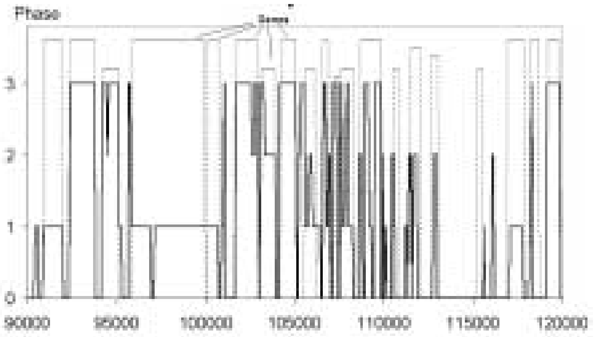

FIG. 6: Prediction of protein-coding regions using clustering in the space of triplet frequencies for S.Cerevisiae III. X-axis is basepair position in sequence, Y-axis is number of cluster (coding phase). Dotted line shows positions of ORFs, the height of bar corresponds to the confidence of gene presence (highest bars are experimentally discovered genes.)

more advanced techniques for clustering promise better results. Fragments of the resulting graphs of phase (that is actually cluster number) of sliding window (calculated through every 3 bp) are shown in fig.5,6.

The quality (given in percentage of correctly predicted nucleotides) of presented algorithm is not striking. But we emphasize here that it is based on the single (and very clear) idea of existence of coding phase and it does not use any learning dataset. The only parameter of the algorithm is the width of sliding window. Its value determines the characteristic length of the coding regions to be detected. Combining this approach with other methods should give results which will satisfy requirements of molecular biologists.

Thus, in this paper it was demonstrated that simple notion of existence of distinguished coding phase in three possible distributions of triplets in a window of DNA can serve as a good detector of coding information. Visualization of set of sliding windows in the space of triplet frequencies shows symmetric bullet-like structure. Linear dimensions of the structure are determined by the amplitudes of two measures: local GC-concentration and mixing entropy.

These two measures have maximum of their effectiveness for separating coding and non-coding regions in the same quite wide range of window lengths (determined by average length of exons). As average mixing entropy measure is more effective, but it can separate only protein-coding regions, while effectiveness of GC-concentration does not depend on the type of the coding region.

Visualization shows that distribution of windows of DNA in triplet frequencies space forms 4 clusters (central one for junk region, where there is no coding phase, and 3 side ones for three possible phase shifts). Though this clustering is not very compact, it can be used for gene-finding without any learning dataset.

Acknowledgments

Our efforts were inspired by Misha Gromov (IHES). We are thankful to Alessandra Carbone (IHES) and Iliya Karlin (ETH, Zurich) for stimulating discussion and help.

[1] J.-M. Claverie, Human Molec. Genetics **6**, 1735 (1997).

[2] E. Yeramian, Gene **255**, 139 (2000).

[3] E. Yeramian, Gene **255**, 151 (2000).

[4] P. Bernaola-Galvan, I. Grosse, P. Carpena, and et.al., Phys.Rev.Lett. **85**(6) (2000).

[5] J. Fickett, Computers Chem. **20**(1), 103 (1996).

[6] A. Gorban, A. Zinovyev, and T. Popova, *Statistical approaches to automated gene identification without teacher*, Institut des Hautes Etudes Scientifiques Preprint, IHES/M/01/34 (2001).

[7] A. Gorban and A. Rossiev, Journal of Computer and System Sciences International **38**(5), 825 (1999).

[8] A. Zinovyev, *Visualisaton of Multidimensional Data* (Krasnoyarsk State Technical University Press, Russia, 2000).

[9] A. Gorban, A. Zinovyev, and A. Pitenko, Informatsionnie technologii, Moscow. **6**, 26 (2000).

[10] A. Gorban and A. Zinovyev, *Visualization of data by method of elastic maps and its application in genomics, economics and sociology*, Institut des Hautes Etudes Scientifiques Preprint, IHES/M/01/36 (2001).

[11] T. Kohonen, *Self-Organizing Maps* (Berlin - Heidelberg, 1997).

[12] J. Hartigan, *Clustering Algorithms* (John Wiley & Sons, Inc., New York, 1975).

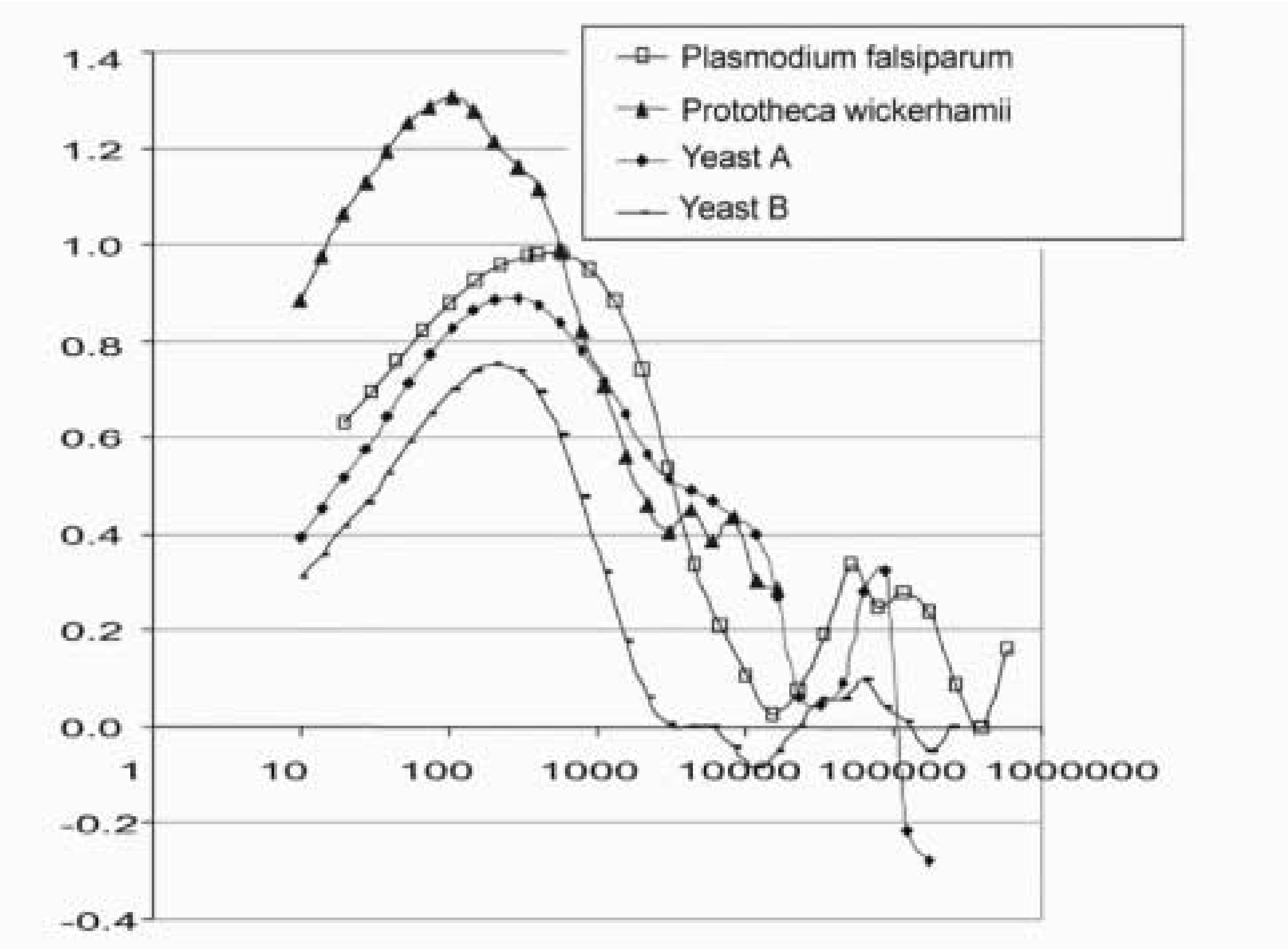
Plasmodium falsiparum
Prototheca wickerhamii
Yeast A
Yeast B
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
1
10
100
1000
10000
100000
1000000

Figure 1a

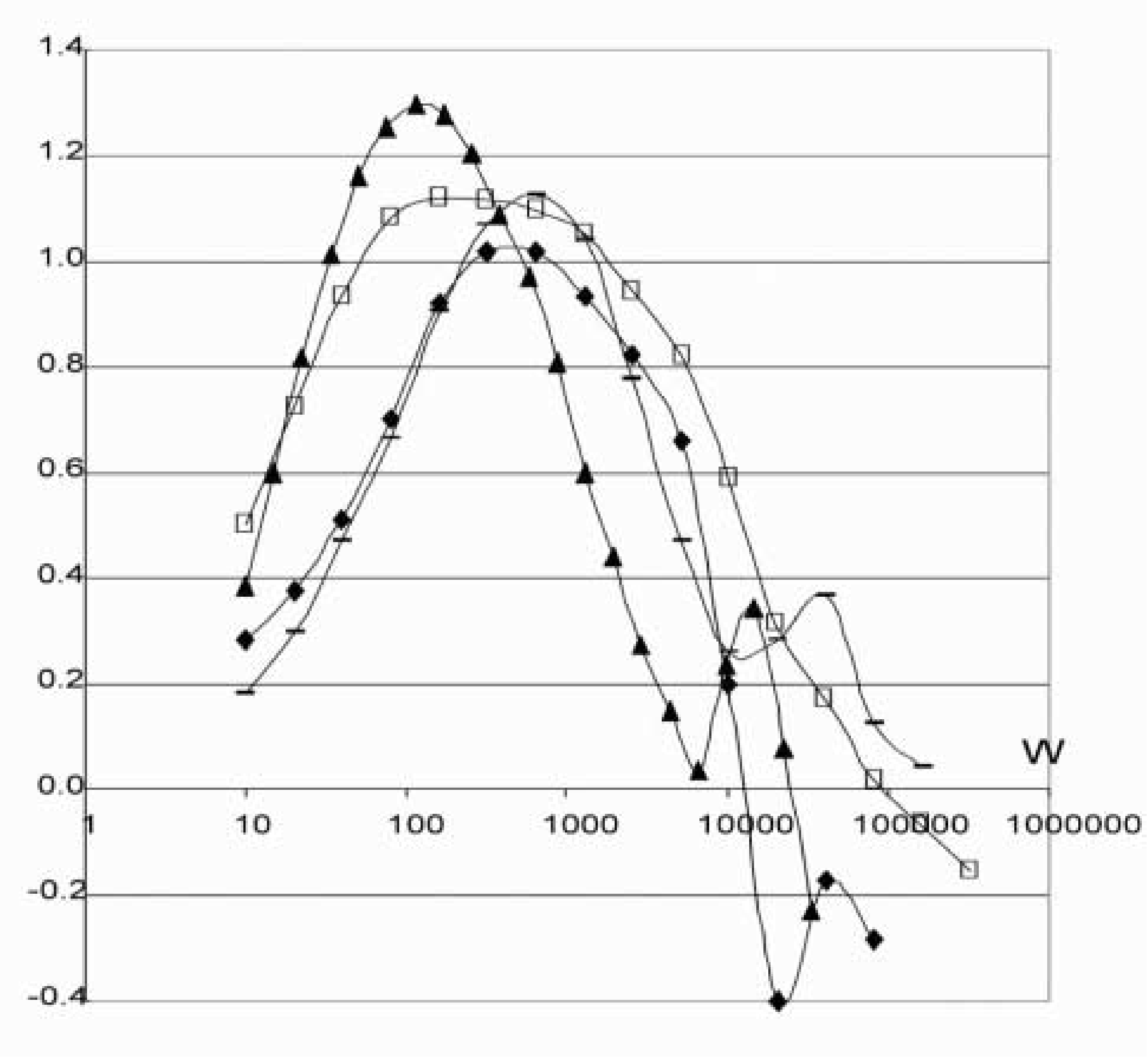

Figure 1b

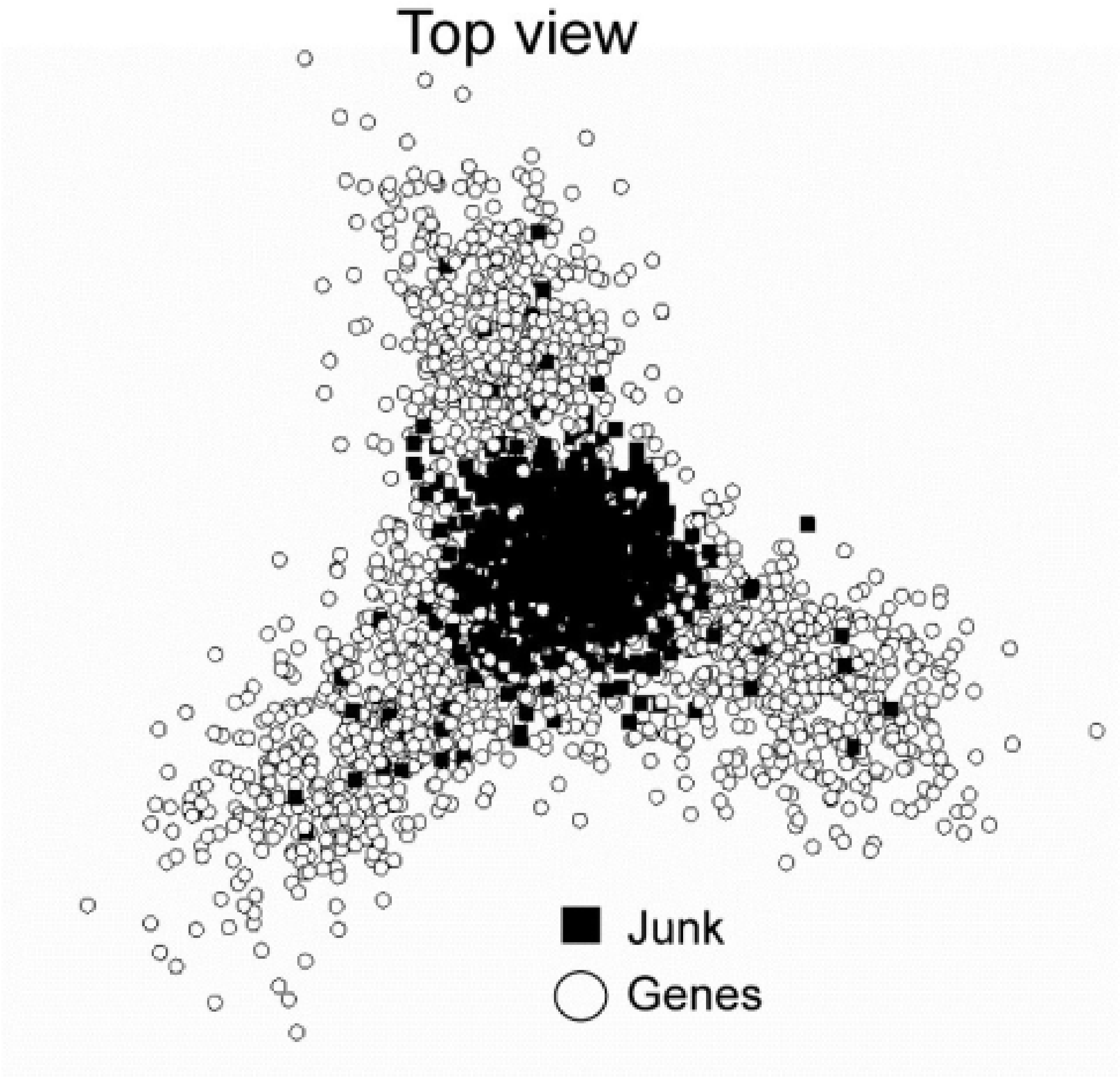

Figure 2a

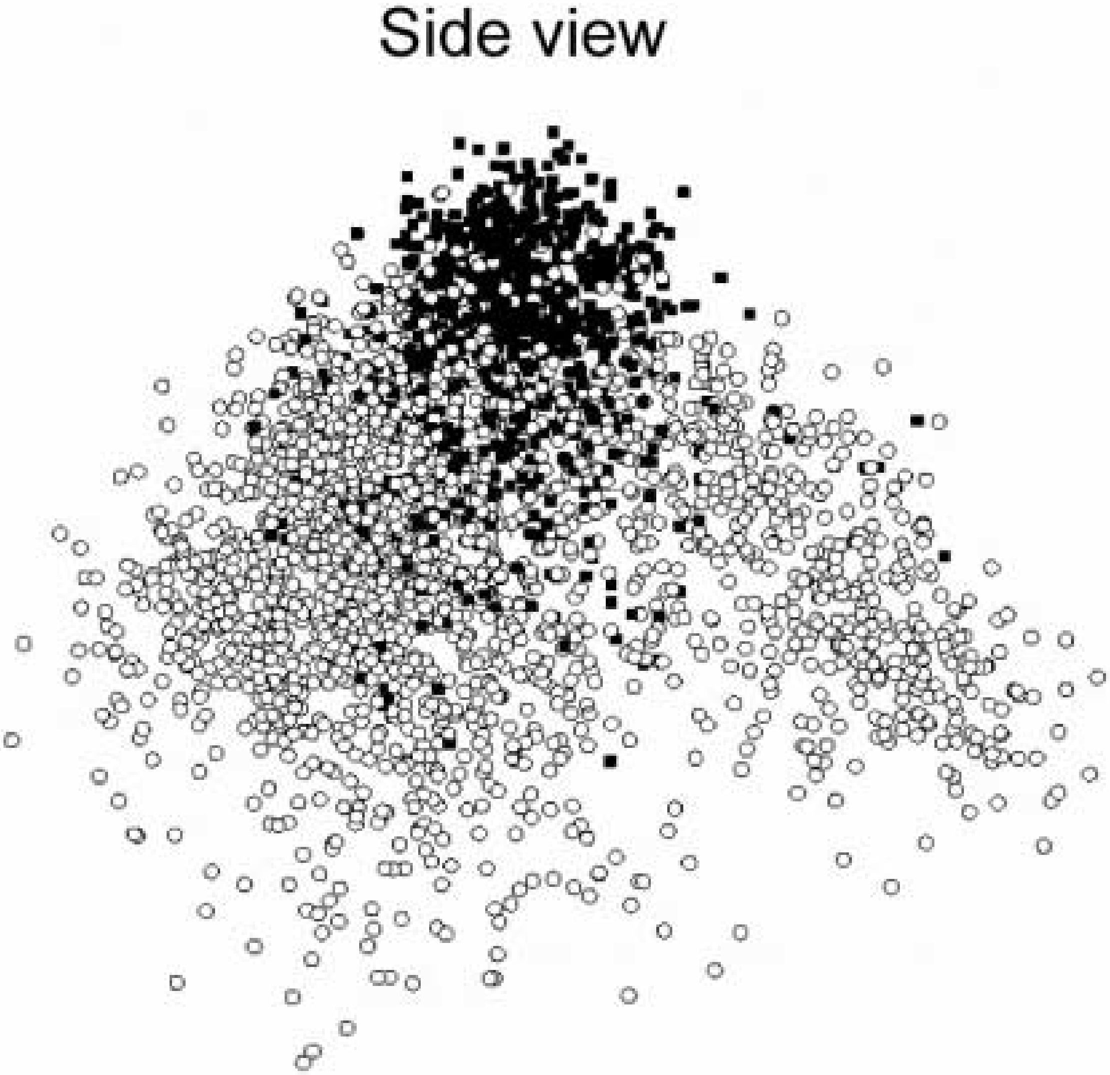

Figure 2b

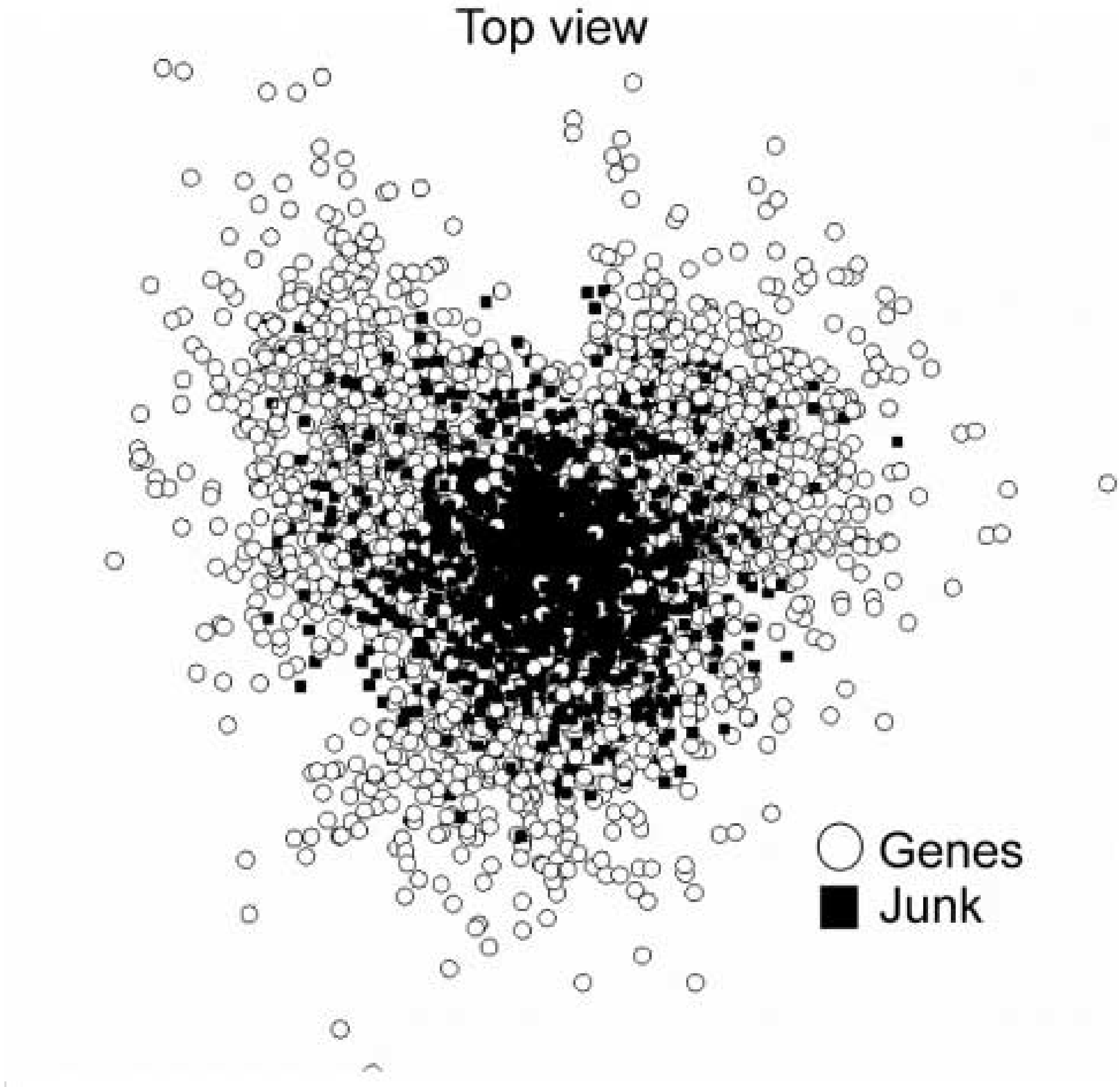

Figure 3a

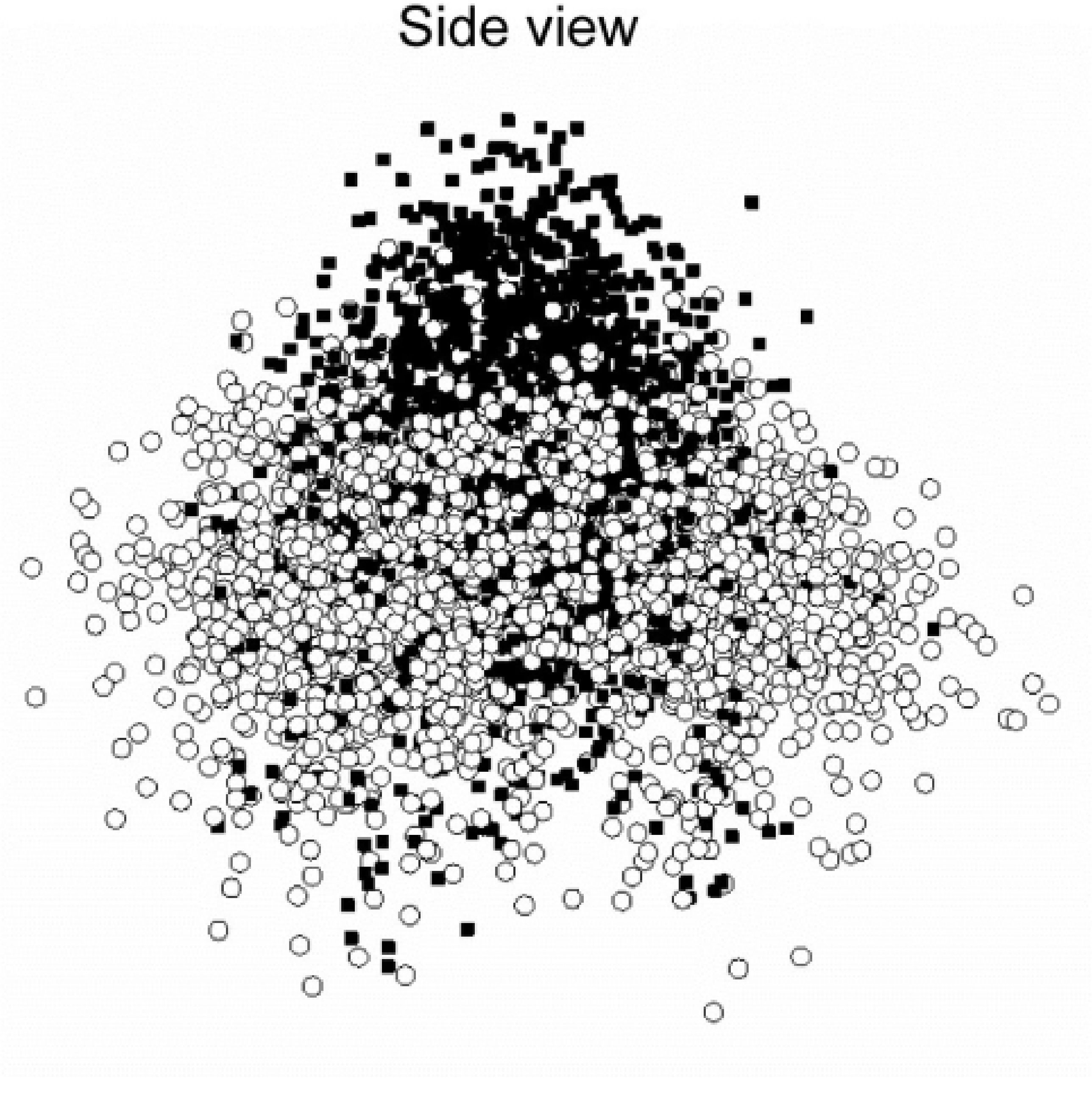

Figure 3b

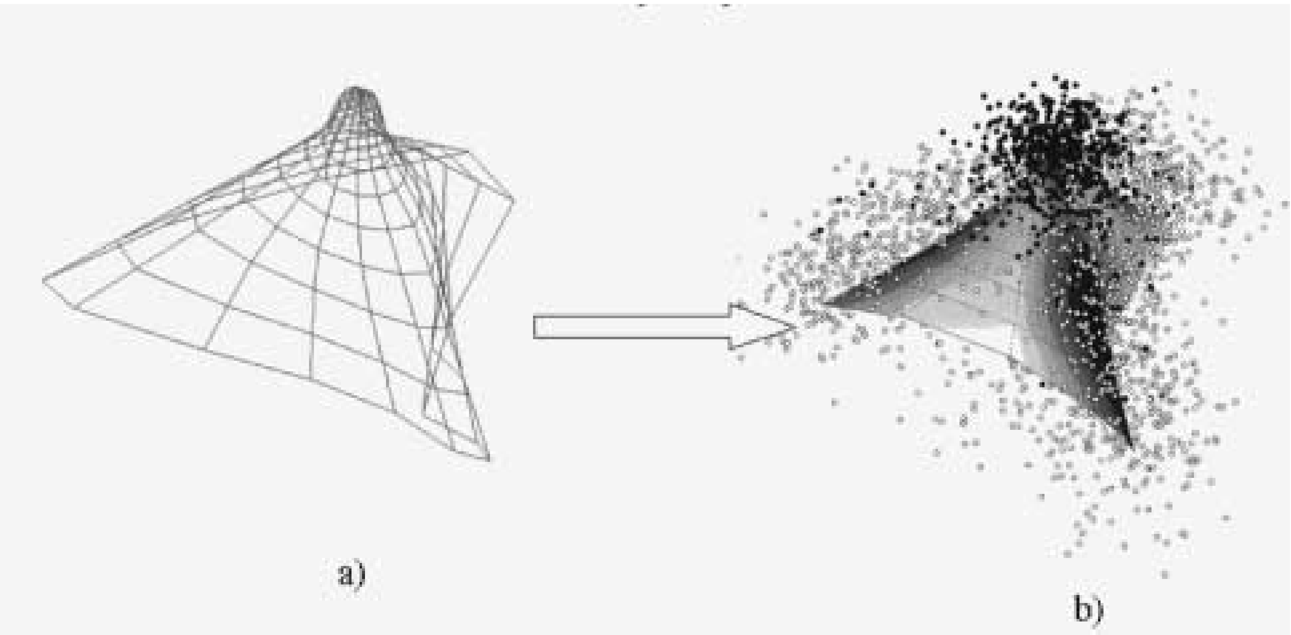

Figure 4ab

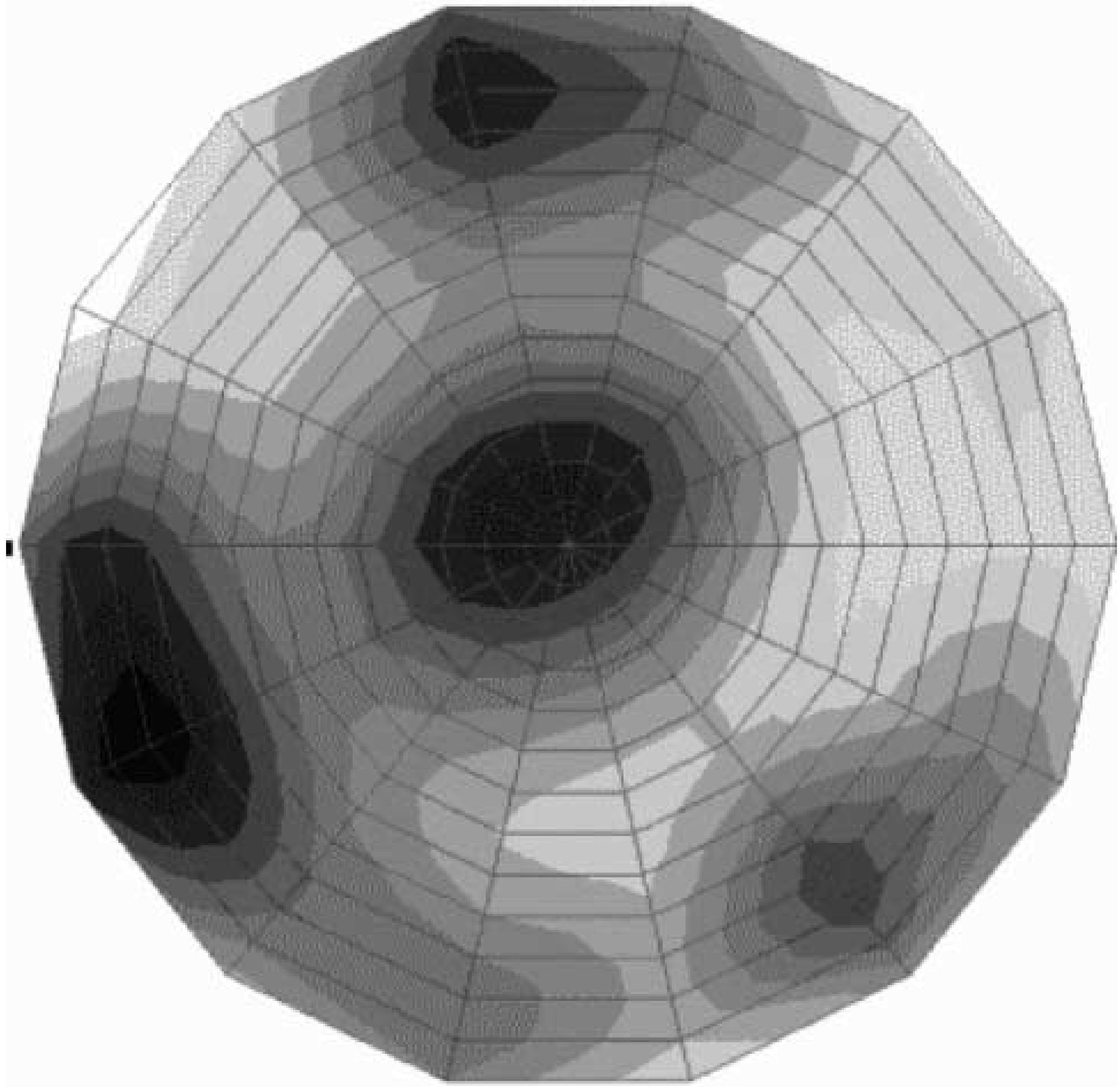

Figure 4c

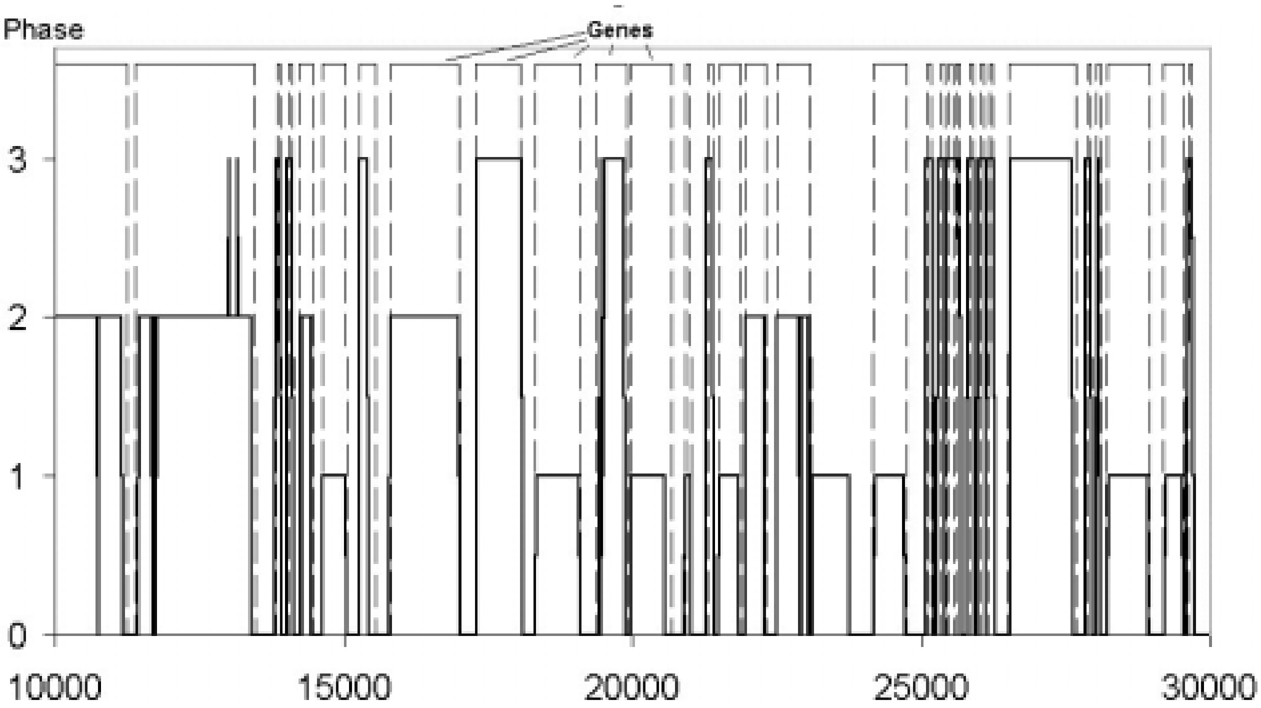

Figure 5

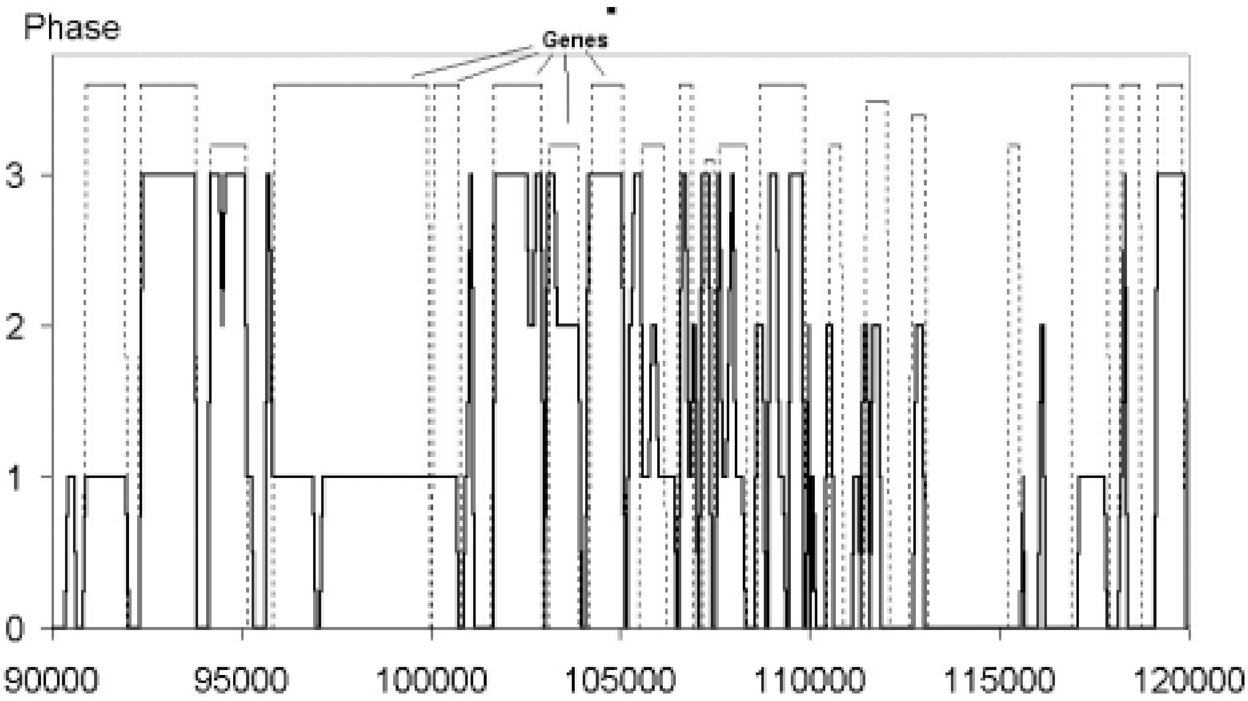
Phase
Genes
3
2
1
0
90000
95000
100000
105000
110000
115000
120000

Figure 6