Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Structural visualization of sequential DNA data*

Xiao-hong MAO[§1], Jing-hua FU[§2], Wei CHEN[†‡2], Qian YOU[3], Shiao-fen FANG[3], Qun-sheng PENG[2]

([1]*The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310013, China*)

([2]*State Key Lab of CAD & CG, Zhejiang University, Hangzhou 310058, China*)

([3]*Department of Computer and Information Science, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, IN 46202, USA*)

[†]E-mail: chenwei@cad.zju.edu.cn

Received Apr. 11, 2010; Revision accepted July 5, 2010; Crosschecked Jan. 31, 2011

Abstract: To date, comparing and visualizing genome sequences remain challenging due to the large genome size. Existing approaches take advantage of the stable property of oligonucleotides and exhibit the main characteristics of the whole genome, yet they commonly fail to show progression patterns of the genome adjustably. This paper presents a novel visual encoding technique, which not only supports the binning process (phylogenetic analysis), but also allows the sequential analysis of the genome. The key idea is to regard the combination of each k-nucleotide and its reverse complement as a visual word, and to represent a long genome sequence with a list of local statistical feature vectors derived from the local frequency of the visual words. Experimental results on a variety of examples demonstrate that the presented approach has the ability to quickly and intuitively visualize DNA sequences, and to help the user identify regions of differences among multiple datasets.

Key words: Genome sequence, Sequential visualization, Bio-information visualization
doi:10.1631/jzus.C1000091 **Document code:** A **CLC number:** TP391.1; R394.3

1 Introduction

To study the differences and similarities among different organisms, biologists have focused on the intrinsic properties of their corresponding genome sequences. Particularly, one genome sequence contains many chromosomes which consist of four chemical bases (Adenine (A), Thymine (T), Cytosine (C), Guanine (G)) and their attached nucleotides. The genome sequence has special properties in terms of the frequencies of k-nucleotides ($1 < k \leq 6$) (Zhou *et al.*, 2008). In genetics, A pairs with T and C pairs with G. Thus, the reverse complement of a DNA sequence is its reverse, complement, or reverse-complement counterpart. For instance, the reverse and reverse complement of a five-nucleotide ACAGT are TGACA and ACTGT, respectively. The relative combined frequency of ACAGT and ACTGT over a sequence $< y_1, y_2, ..., y_N >, y_i \in \{\text{A,C,G,T}\}$ is given by

$$\frac{1}{N-5+1}\sum_{i=1}^{N-5+1}(\delta_{y_i,\text{A}}\delta_{y_{i+1},\text{C}}\delta_{y_{i+2},\text{A}}\delta_{y_{i+3},\text{G}}\delta_{y_{i+4},\text{T}} + \delta_{y_i,\text{A}}\delta_{y_{i+1},\text{C}}\delta_{y_{i+2},\text{T}}\delta_{y_{i+3},\text{G}}\delta_{y_{i+4},\text{T}}), \quad (1)$$

where $\delta_{a,b} = \begin{cases} 1, & a = b, \\ 0, & a \neq b. \end{cases}$

Traditionally, scientists have sought to study different organisms by analyzing individual homologous genes, markers, single-nucleotide polymorphisms (SNPs) , and other features from the genome sequences. These features, however, require comprehensive domain knowledge, not to mention that the biological meanings of many segments in a genome sequence are still unclear. For instance, the evolu-

‡ Corresponding author

§ The two authors contributed equally to this work

* Project supported by the National Natural Science Foundation of China (Nos. 60873123 and 60903085), the National Basic Research Program (973) of China (No. 2010CB732504), the Natural Science Foundation of Zhejiang Province (No. Y1080618), and the Open Project Program of the State Key Lab of CAD & CG, Zhejiang University, China (No. A0905)

tionary history of a particular feature might be quite different from the evolutionary history of the organism in which it can be observed (Savva *et al.*, 2003). In addition, the size of a long DNA sequence can be as large as several billion base pairs (bp) (e.g., the DNA of the human), making the identification, comparison, and analysis of individual features very laborious and often erroneous.

Statistical analysis has played an important role in biological and medical applications. Yet, unsupervised differentiation of a set of genome sequences overwhelms the efficiency of statistical analysis, particularly because there are a very limited number of features that have been biologically identified among different organisms. Brute-force statistical computing can be very time-consuming and may easily lead to results that are difficult to interpret.

The power of visualization is driven by the fact that humans are able to detect patterns much better with visual representations than with categorical (e.g., human beings and animals) or textual ones (e.g., A, C, G, T). Existing color mapping approaches map individual DNA units (nucleotides, gene segments, or genes) into color, texture, or shape. Although local information is preserved, the degree of semantics in a larger range (e.g., the entire genome sequence) is unavailable. Fig. 1a displays a barcode representation of the 13th chromosome of the *Mus musculus* (a lab mouse) dataset.

A more effective solution would be constructing a multi-resolution representation and providing level-of-detail views to the underlying data. The recently developed MizBee system (Meyer *et al.*, 2009) allows biologists to explore many kinds of conserved synteny relationships with linked views at genome, chromosome, and block levels. MizBee is particularly effective when every feature region has been identified, but it can display only hierarchical structures when no prior knowledge on the feature regions is available. One screenshot of the MizBee system is shown in Fig. 1b. Shah *et al.* (2004) adopted the volume rendering technique to visualize sequential DNA data in the 3D space, which avoids the constraint of the 2D screen space. However, they failed to provide insights into the intrinsic properties of DNA data. Fig. 1c displays one result of their method.

This paper introduces a different approach that eliminates the need to have knowledge of the underlying DNA sequences. The main goal is to reveal

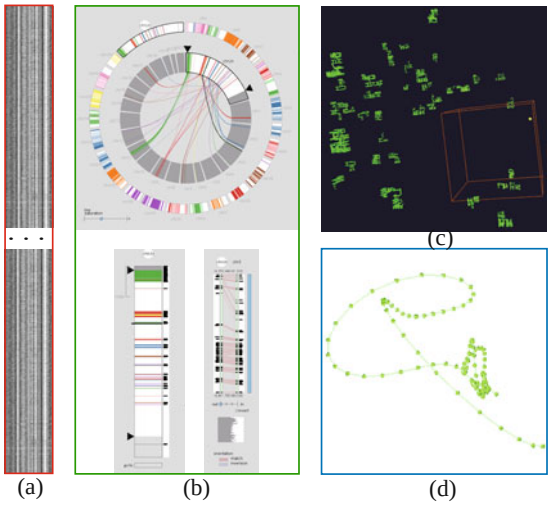

Fig. 1 Visualizing the 13th chromosome of the *Mus musculus* (a lab mouse) with the barcode representation (Zhou *et al.*, 2008) (a), two kinds of fish Stickleback and Tetraodon with the MizBee system (Meyer *et al.*, 2009) (b), the DNA data with the volume rendering technique (Shah *et al.*, 2004) (c), and the Zebra dataset with our approach (d)

the global and structural characteristics and thereby enable effective exploration before detailed analysis. We encode every k-nucleotide as a visual word, and regard a long DNA sequence as a list of visual words. By computing a sequence of local statistical feature vectors based on the word list, and projecting the sequence into a recognizable low-dimensional space, an abstractive, global, and curve-based representation is obtained. Visualizing and interacting with the new representation favor complex reasoning processes, like characterizing the primary content of the DNA data and comparing different sequences from varied organisms. Fig. 1d shows one of our results.

Note that our approach is designed not to investigate the individual nucleotides or sequence segments, but to draw a pre-screened overview. Preliminary results demonstrate that this study would have significant importance on many genetics-related problems like the identification of genetic material transferred from other organisms or through virus invasions, and phylogenetic analysis of genomes.

2 Related work

2.1 Genome-wide analysis

In the past two decades, the intrinsic properties of genome sequences have attracted considerable

attention. One interesting work is the discovery of the periodicity property of DNA sequences across the prokaryotic and eukaryotic genomes (Trifonov and Sussman, 1980). For example, Karlin *et al.* (Karlin and Burge, 1995; Karlin *et al.*, 1997; 1999) have studied numerous genome properties based on the frequencies of k-nucleotides. They have also observed that the bi-nucleotide relative quantity is normally stable across a genome, measured on 50 000 bp fragments. Zhou *et al.* (2008) verified that for each genome, the majority of its short sequence fragments have highly similar k-nucleotide distributions. These observations built a solid foundation for our approach.

With more and more whole-genome sequences becoming available, scientists have begun to perform analyses on a genome-wide scale. Usually, the whole genome of an organism preserves more reliable properties than one feature region (gene, SNP, marker, promoter region, coding region, non-coding region, etc.). Recent studies include the genome-wide association study (GWAS) of the National Institutes of Health (NIH), genome-wide expression patterns analysis (Eisen *et al.*, 1998), and genome-wide phylogenic analysis (Herniou *et al.*, 2001; Bourque and Pevzner, 2002; Savva *et al.*, 2003). These studies focused on either the whole-genome DNA sequences (Herniou *et al.*, 2001) or the feature regions (markers, genes), and provided reliable and inspiring results.

2.2 Visualization of DNA data

Several toolkits for visualizing genome sequences are publicly available. Some of them afford whole-genome browsing (Hallin *et al.*, 2008; Zhou *et al.*, 2008). Among them, one popular tool is the Mapviewer (www.ncbi.nlm.nih.gov/prjects/mapview), which provides special browsing capabilities for a subset of organisms in Entrez Genomes. Specifically, it allows users to view and search an organism's complete genome, display chromosome maps, orderly zoom into greater levels of details, and jump to the sequence data in a region of interest.

Alternatively, many works focus on certain aspects of DNA data (Shah *et al.*, 2004; Meyer *et al.*, 2009). For instance, Meyer *et al.* (2009) presented a multi-scale synteny browser, which nicely integrates three levels of the feature regions, and is capable of showing multiple types of relationships on multiple scales.

The challenge for genome data visualization lies in the fact that only limited ranges of the data are clear to scientists, and that the size of a genome sequence is typically huge. Traditionally, a DNA dataset is represented in a 1D space, and visualized as either a long colored map (Zhou *et al.*, 2008) or a radial graph (Meyer *et al.*, 2009). To afford the global visualization without loss of local details, a DNA dataset can be arranged into a 3D space, allowing interactive volume visualization for DNA data (Shah *et al.*, 2004). Our approach interprets the genome sequence from another viewpoint. To the best of our knowledge, this is the first work that explores and visualizes the genome sequence on the scale of the whole-genome sequence, based on the frequencies of the k-nucleotides.

2.3 Visualization of sequential information

Sequential information exists everywhere in our daily life, including video sequences (Goldman *et al.*, 2006), animal movement (Grundy *et al.*, 2009), motion animation (Assa *et al.*, 2008), and time-varying volume datasets (Lu and Shen, 2008). The sequential information can be encoded with visual encoding schemes, and visualized following data analysis, abstraction, and projection techniques. Document visualization, for example, has been the focus of much research for years (Havre *et al.*, 2001; Fortuna *et al.*, 2005; Blei and Lafferty, 2006). Existing software systems, such as IN-SPIRE (http://in-spire.pnl.gov), Thomson's RefViz (www.refviz.com), and the Science Topic Browser (Blei and Lafferty, 2007), are examples of non-sequential visualization of a number of documents via consideration of the distance relation between individual documents. The recently proposed methods, such as Lowbow, aim at visualizing sequential semantic progression within a single document. Similar ideas were proposed in Mao *et al.* (2007).

3 Structural and sequential visualization

Our approach is driven by the observation that each genome has a stable distribution of the combined frequency for each k-nucleotide and its reverse complement (Zhou *et al.*, 2008). Studying the combined frequency, we can effectively describe the

semantic transition, critical variations, as well as the appearance of repeated patterns.

The input is a genome sequence, and needs preprocessing before the analysis and visualization. Researchers have proved that the sequences of several dozens of kilobases in length can preserve the features of the whole genome (Deschavanne *et al.*, 1999). Thus, we can retrieve a long segment from an input dataset for analysis. The main visualization pipeline consists of three components (Fig. 2): visual encoding, spectrum-based representation, and low-dimensional embedding.

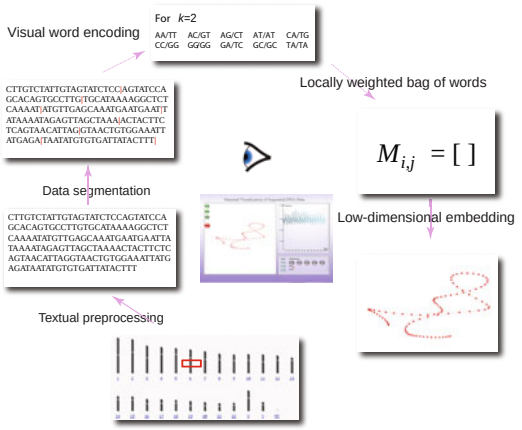

Fig. 2 Overview of our visualization approach. The main visualization pipeline consists of three components: visual encoding, spectrum-based representation, and low-dimensional embedding

3.1 Encoding visual words from nucleotides

A nucleotide is a molecule that is composed of three functional groups: a base (one of four chemicals: A, T, G, and C), a sugar, and a phosphate. The set of nucleotides encodes the intrinsic property in terms of genetics, and forms the building blocks of DNA. On the other hand, the order of the bases in DNA determines the genetic code. Thus, by definition we can make an analogy between the nucleotides in a genome sequence and the words in a document (Fig. 3).

Because a nucleotide itself cannot form an effective feature region, we map a group of nucleotides into a word in a document. Additionally, based on the argument that a genome sequence has a stable distribution of the combined frequency for each k-nucleotide and its reverse complement (Zhou *et al.*, 2008), we make a fundamental assumption that the property-encoding words of DNA are of equal length. This means k is constant in a DNA sequence.

Looping over all four kinds of chemicals for all possible nucleotide combinations, yields 4^k combinations of nucleotides. Because the combined frequency of a k-nucleotide and its reverse complement is more stable than that of the k-nucleotide (Zhou *et al.*, 2008), we couple each combination and its reverse complement. Specifically, when k is even, there are $4^{k/2}$ palindromes (compliment and inverse equal the original k-nucleotide). Thus, the number of all possible visual words is $4^k/2$ when k is odd, and $(4^k + 4^{k/2})/2$ when k is even. For instance, when k is 2, there are 10 visual words, namely AA/TT, AC/GT, AG/CT, AT/AT, CA/TG, CC/GG, CG/CG, GA/TC, GC/GC, TA/TA. Note that a visual word is composed of two parts, and its frequency is the sum of the frequencies of both parts.

3.2 Sequential representation based on Lowbow

Given the similarity between documents and DNA sequences in terms of their discrete properties, we modify a recently developed document visualization approach for effectively representing a DNA sequence based on the visual encoding of the

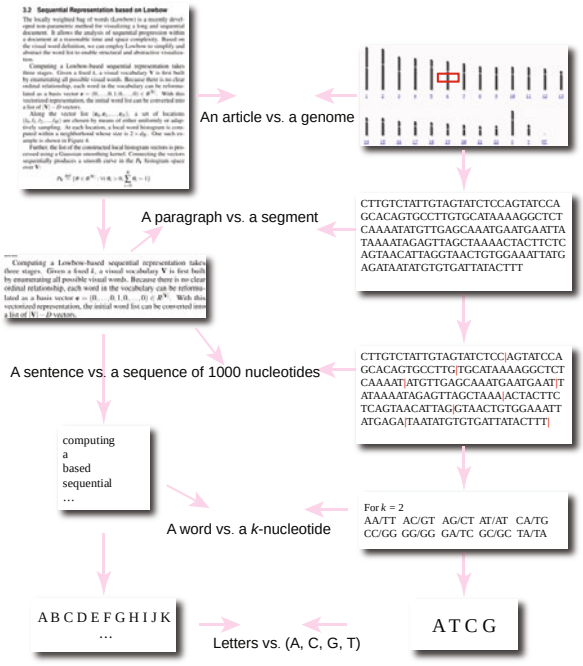

Fig. 3 A conceptual analogy between a document (a word sequence) and a genome sequence

nucleotide.

The locally weighted bag of words (Lowbow) is a non-parametric method (Mao *et al.*, 2007) for visualizing a long and sequential document. It allows the analysis of sequential progression within a document at a reasonable time and space complexity. Based on the visual word definition, we can employ Lowbow to simplify and abstract the word list to enable structural and abstractive visualization.

Computing a Lowbow-based sequential representation takes three stages. Given a fixed k, a visual vocabulary $\boldsymbol{V}$ is first built by enumerating all possible visual words. Because there is no clear ordinal relationship, each word in the vocabulary can be reformulated as a basis vector $\boldsymbol{e} = (0,\ldots,0,1,0,\ldots,0) \in \mathbb{R}^{|\boldsymbol{V}|}$.

An input DNA sequence is partitioned into a set of fragments, the size of which is set to be around 1000. The reason is that the combinations of k-nucleotides in every 1000 nucleotides can preserve some certain stable patterns (Zhou *et al.*, 2008). For each resulting fragment, a local word histogram is computed (see **v0**–**v9** in Fig. 4), which is actually a vector list with respect to the vocabulary. The size of each vector, $N(k)$, is the number of the uniquely combined k-nucleotides.

Along this vector list, a sequence of locations are chosen by means of either uniform or adaptive sampling, e.g., **v0**, **v3**, **v6**, **v9**. At each location, a neighborhood is chosen, whose size $2d_W$ is adjustable. To ensure the smooth transition along the sequence, the windows of two consecutive sampling points should be overlapping. In the example shown in Fig. 4, $d_W = 4$. Thereafter, the local histogram list in each window is joined and normalized into a uniform histogram vector (e.g., **U0**–**U3** in Fig. 4).

The list of the constructed histogram vectors is further processed using a Gaussian smoothing kernel. Connecting the vectors sequentially produces a smooth curve in the $P_{\boldsymbol{V}}$ histogram space over $\boldsymbol{V}$:

$$P_{\boldsymbol{V}} \overset{\text{def}}{=} \{\theta \in \mathbb{R}^{|\boldsymbol{V}|} : \forall i,\ \theta_i > 0, \sum_{i=0}^{N} \theta_i = 1\}.$$

For more details of the Lowbow approach, please refer to Mao *et al.* (2007).

3.2.1 Comparison to the barcode representation

The barcode representation (Zhou *et al.*, 2008) is a sequence of high dimensional vectors derived from a genome sequence. To begin with, the sequence is subdivided into a series of non-overlapping and equally-sized (a typical choice is 1000) fragments. Then, within each partitioned fragment, the combined frequency of each k-nucleotide and its reverse complement is computed. The set of the resulting frequencies in all fragments forms a list of $N(k)$-dimensional vectors. In this way, each element of the vector represents the frequency of the corresponding k-nucleotide within the corresponding fragment.

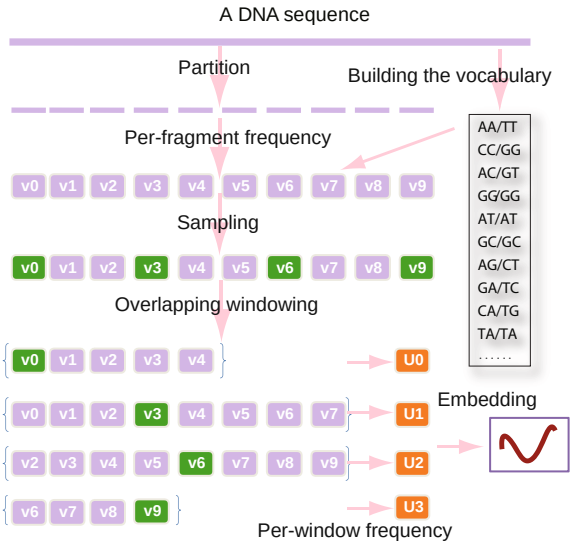

Fig. 4 Illustration of the Lowbow representation with an artificial dataset. Four overlapping windows are used to build four local histogram vectors U0–U3

Both our Lowbow-based representation and the barcode representation are spectrum-based, meaning that they characterize the data on the basis of the frequency. But Lowbow is different from the barcode in two aspects. First, Lowbow allows for adaptive sampling, while the barcode takes uniform sampling. Adaptive sampling may better characterize the data properties, e.g., capturing the boundaries between known feature regions. Second, the local neighbors in the Lowbow representation are overlapped, while the barcode representation takes a non-overlapping subdivision to the list. Overlapped windowing makes consecutive histograms closer than the non-overlapped windowing scheme does. Consequently, smoother progression is achieved.

3.3 Low-dimensional embedding

The dimensionality of each histogram vector is $|\boldsymbol{V}|$. A natural choice to display the vector list is using a dimension-reduction method to project the

high-dimensional points into the perceptible 2D or 3D space.

In practice, we employ the multi-dimensional scaling (MDS) technique (Hastie *et al.*, 2005) to convert the vector list into the 3D space. One main advantage of the MDS algorithm is that it seeks to preserve as much as possible in a low-dimensional space, the proximities found among points in a high-dimensional space. We use the Pearson distance to evaluate the differences among genomes, which proves useful in genome comparison (Schbath *et al.*, 1995). For more details of MDS approaches, please refer to Borg and Groenen (2003).

4 Interactive exploration

We designed an interactive system that combines a global 3D view, a local histogram view, and a suite of visualization, manipulation, and comparison toolkits. The key feature of this system is the ability to produce visualizations of multiple DNA sequences, and to explore those datasets interactively, allowing a user to discover feature regions, detect possible semantic variations, and differentiate two datasets in real-time.

4.1 Interface design

The interface consists of three windows, namely the main view on the left, the histogram view on the top right, and the control panel on the bottom right. Fig. 5 shows a screen-shot of the main interface of our system.

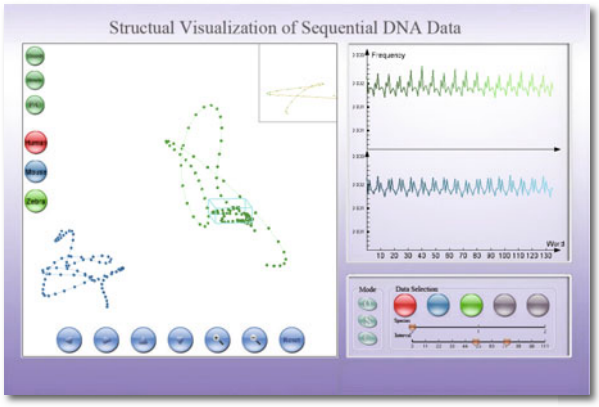

Fig. 5 A snapshot of our visualization system. The interface consists of three windows: main view (left), histogram view (top right), and control panel (bottom right)

The main view displays the 3D projection of the sequential representation. Two display modes are provided: point-based and curve-based (Figs. 6a and 6b). The latter is preferred because a curve represents the sequentiality of the dataset. The legends for different organisms are shown on the left side. We employ a qualitative, eight-element color map provided by the ColorBrewer (www.personal.psu.edu/cab38/ColorBrewer). The coloring scheme is consistent across the entire interface.

The histogram view shows the mean frequency with respect to the visual words of the selected regions of the underlying dataset. The x-axis sorts different words, and the y-axis records the count associated with each k-nucleotide. When two objects are shown in the main view, two histograms are displayed for comparison.

The control view combines a suite of exploration widgets. Three groups of controls are designed: the browsing controls (left), the choosing controls (top right), and the sliding widgets (bottom right).

4.2 Exploration widgets

Several simple user widgets are incorporated.

Exploration controls include the three buttons shown in the top left corner of the main view. The top two buttons control whether a thumbnail view or a detailed view is provided in the small rectangle region in the top right corner of the main view. The third one switches the display modes (point- or curve-based).

Viewing controls are shown in the bottom of the main view. Translating, rotating, resizing, and zooming are provided. These widgets are visible only when the mouse touches their regions.

Browsing controls control the browsing mode for a set of datasets, including individual browsing, pairwise browsing, and all-in-one comparative browsing (Figs. 6c and 6d).

Choosing controls determine which datasets are to be studied. Five buttons in the top right corner of the control view are used to specify five continuous organisms (from left to right) in the database. The first organism of the five is determined by the top sliding widget on the bottom right of the panel, and the region of interest of the selected organism is determined by the bottom sliding widget.

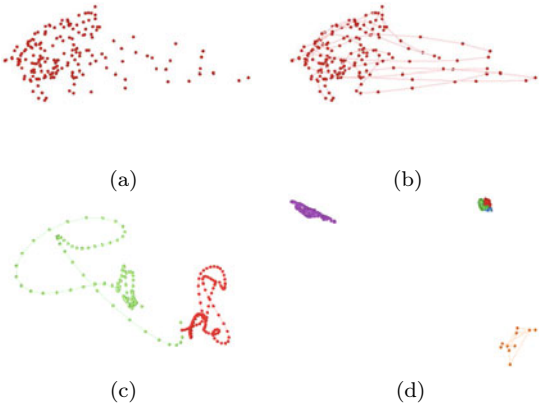

Fig. 6 The point mode (a) vs. the curve mode (b) for the Aeromonas_hydrophila_ATCC_7966 dataset (a kind of bacterium); (c) Pairwise browsing for the same bacterium in (a) and (b) and the Aeh1 dataset (a kind of virus); (d) All-in-one browsing for five datasets: Aeromonas_hydrophila_ATCC_7966 (purple), Aeh1 (orange), Human (red), Mouse (blue), and Zebra (green)

4.3 Adaptive and hierarchical exploration

Adaptive and hierarchical exploration in a large-sized dataset is always desirable to quickly obtain an overview and emphasize important parts. Our sequential representation by nature favors this exploration mode because it employs a Gaussian smoothing kernel and an adaptive sampling scheme. By varying the bandwidth of the smoothing kernel, it is able to capture the primary content of the underlying DNA sequences at various resolutions, as shown in Section 5. The adaptive sampling is enabled when there is some prior knowledge about the segmentation of feature regions. For instance, the simplest adaptive sampling is the segmentation according to the boundaries of DNA feature regions (coding regions, non-coding regions, promoting sequences, etc.). Figs. 7a and 7b show the effects of uniform sampling and adaptive sampling, the latter determined by the boundaries of coding regions.

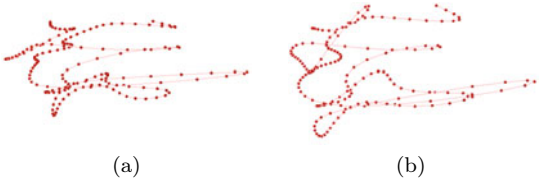

Fig. 7 Visualizations of the coding region of the Aeromonas_hydrophila_ATCC_7966 dataset by means of uniform (a) and adaptive (b) sampling

5 Results

We designed and implemented an interactive visualization and exploration system based on the Processing language (http://processing.org/). All results and the companion video were generated on a PC with an Intel 1.6 GHz CPU, 1.0 GB RAM. All the DNA sequences are available from the NCBI database (www.ncbi.nlm.nih.gov).

For each organism, multiple DNA segments can be randomly chosen from its genome sequence because of the stability property of the k-nucleotides. If there are multiple datasets, we regard them as an entirety to compute their sequential representations, and visualize them with different colors. The typical size of an individual sequence we visualized ranges from 1 million to 3 million bp.

Fig. 8 depicts the point-based rendering results for the Mouse (blue) and the Human (green) datasets, yielding two observations. First, the points of each dataset cluster together, demonstrating the stability property of k-nucleotides in each genome. Second, there are significant differences between the two clusters. These observations verify that our approach can well differentiate various organisms.

With our visual encoding scheme, the histogram view provides additional insights into the underlying datasets. Fig. 9 compares the histograms of two mammals, i.e., Human and Mouse, and two microorganisms, i.e., Bacterium Aeromonas_hydrophila_ATCC_7766 and Virus Aeh1. A quick view reveals that each histogram has a certain frequency distribution pattern. The results for the Human and the Mouse contain repetitive frequencies, while those of the Bacterium and the Virus exhibit large variations. This can potentially lead to a fingerprint for classifying different organisms.

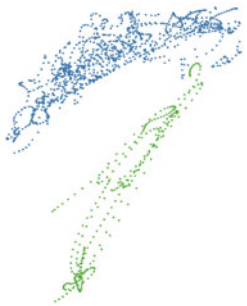

Fig. 8 Results for selected multiple segments of the Mouse (blue) and the Human (green) datasets

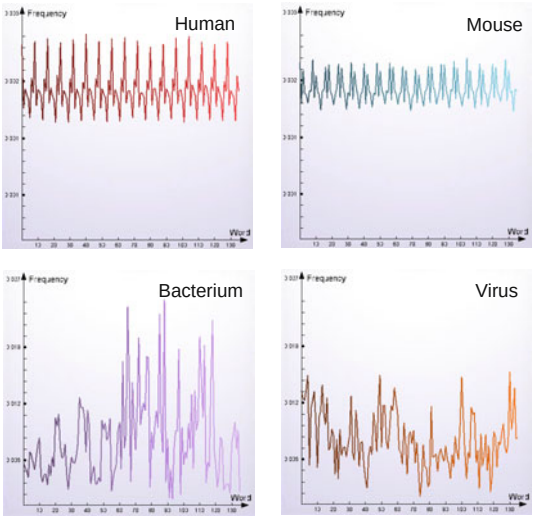

Fig. 9 Four histograms with respect to the visual words: Human, Mouse, Bacterium Aeromonas_hydrophila_ATCC_7766, and Virus Aeh1

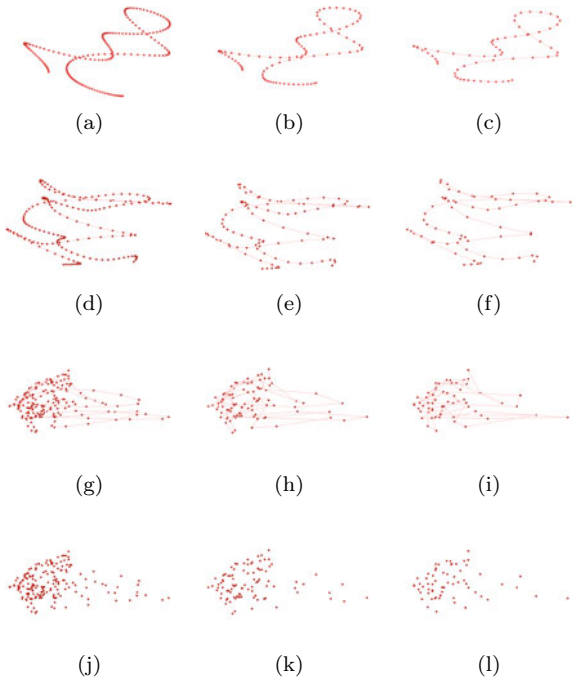

Fig. 10 Results with different Gaussian kernel sizes and sample region sizes. (a) 0.05, 25; (b) 0.05, 50; (c) 0.05, 75; (d) 0.025, 25; (e) 0.025, 50; (f) 0.025, 75; (g) 0.01, 25; (h) 0.01, 50; (i) 0.01, 75; (j) 0.01, 25; (k) 0.01, 50; (l) 0.01, 75. In (j–l), the point-based mode is employed for better visualization

We also studied how the patterns of the sequential visualizations would be influenced by four adjustable parameters: word length k, fragment size M, Gaussian kernel size g, and sample region size s. According to the statement in Zhou *et al.* (2008), the larger the fragment size M is, the more stable the frequencies of the k-nucleotides are. For $M \geq 1000$, the most stable frequencies are observed when k is four. Thus, we used $k = 4$ and $M = 1000$, and visualized a sequence (the Aeromonas_hydrophila_ATCC_7966 dataset) with different g and s (Fig. 10). Results in the same row (e.g., Figs. 10a–10c) reveal that larger s would result in smoother curves; results in the same column (e.g., Figs. 10b, 10e, and 10h) reveal that smaller g would produce more variations. Note that, smoother curves may better reflect the overall progression of the sequence, but might hide local characteristics. When g is very small, however, local details of the curves tend to occlude each other (e.g., Figs. 10g–10i), and point-based rendering results (e.g., Figs. 10j–10l) show clearer distributions of the sample points. In the following experiments, we set $g = 0.03$ and $s = 25$.

5.1 Case study: phylogenetic analyses of genomes

Phylogenies (evolutionary trees) can be estimated by comparing different organisms with our visualization approach. Fig. 11 shows the results for three eukaryotes (Human, Mouse, and Zebra) and two prokaryotes (Bacterium and Virus). From this figure, we can make three observations. First, the relative distances among these curves can partition the organisms into one eukaryote group and one prokaryote group: the three eukaryote curves are tightly clustered and are far away from the two prokaryote curves. Second, the general patterns of the curves are also consistent with the major two groups: eukaryote curves (enlarged in the highlighted circle) demonstrate more cyclic and rotating patterns than the curves of prokaryotes. Third, the point distribution along each curve reflects finer differences among individual genomes in one group. For example, the curve representing the Zebra has denser point clusters towards one end, while the curves representing the Mouse and the Human have more evenly distributed points. This is consistent with the fact that the Human (one family of Primates) is closer in phylogenies to the Mouse (one family of Rodentia) than to the Zebra (one family of Perissodactyla). Note that, the size of the Virus dataset is significantly smaller than those of other datasets, but the

spatial locations and the curve patterns are preserved, showing that our method is robust, regardless of the sequence size.

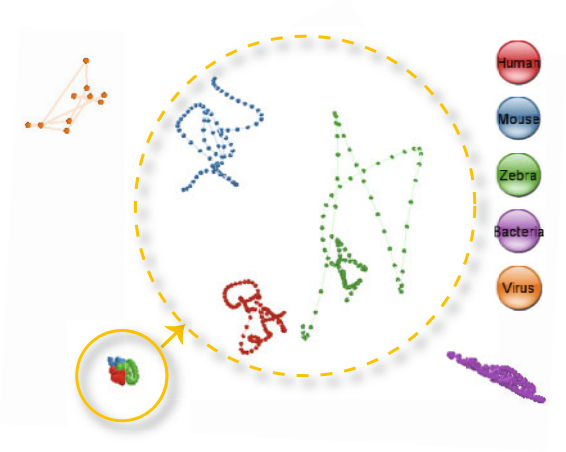

Fig. 11 Phylogenetic analyses with our approach. Organisms studied are the Human, Mouse, Zebra, Bacterium Aeromonas_hydrophila_ATCC_7766, and Virus Aeh1

Furthermore, we used another nine organisms to test our approach. The nine species are *Halobacterium* sp. NRC-1 (NC_002607), *Halorubrum lacusprofundi* ATCC 49239 (NC_012029), *Mycobacterium gilvum* PYR-GCK (NC_009338), *Leifsonia xyli* subsp. *xyli* str. CTCB07 (NC_006087), *Mycobacterium* sp. JLS (NC_009077), *Azoarcus* sp. BH72 (NC_008702), *Rhodococcus opacus* B4 (NC_012522), *Frankia* sp. Ccl3 (NC_007777), and *Rhodococcus jostii* RHA1 (NC_008268). Their relative distances are shown in Fig. 12, which corresponds well to their phylogenetic relationship (Table 1). Also, it is easy to attain the tree representation of the nine organisms (Fig. 12).

6 Conclusions

Genome sequences are long and sequential data, whose segmentation, classification, and comparison remain a quite challenging problem. The presented approach has the ability to quickly and intuitively visualize DNA sequences, and to help the user identify regions of differences among multiple datasets. It favors effective reasoning processes like discovering semantic patterns in a sequence and comparing different sequences from varied organisms. Specifically, we highlight the power of our method with two case studies, i.e., the phylogenetic analysis of genomes

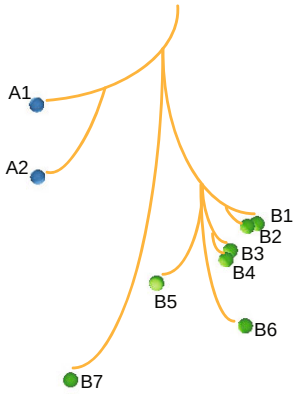

Fig. 12 Phylogenetic analyses with nine organisms. A1, A2, and B1–B7 correspond to indices in Table 1

and the identification of genetic material transferred through the virus invasions.

To preserve the stability property, the size of the nucleotides in a studied fragment should be larger than 1000. Thus, small-scale features such as the features in genes, SNPs, and promoter sequences cannot be easily identified with our approach, because the partitioning process and the overlapped window may destroy the region boundaries and corresponding features. A better solution is needed to address this problem. We believe that similar analytical concepts can be employed to categorize video sequences, time-varying volumetric datasets, and motion capture datasets by encoding distinctive feature vectors with well-defined visual words.

References

Assa, J., Cohen-Or, D., Yeh, I.C., Lee, T.Y., 2008. Motion overview of human action. *ACM Trans. Graph.*, **27**(5):480-489. [doi:10.1145/1409060.1409068]

Blei, D.M., Lafferty, J.D., 2006. Dynamic Topic Models. Proc. 23rd Int. Conf. on Machine Learning, p.113-120. [doi:10.1145/1143844.1143859]

Blei, D.M., Lafferty, J.D., 2007. Modeling Science. Available from http://www.cs.cmu.edu/~lemur/science

Borg, I., Groenen, P., 2003. Modern multidimensional scaling: theory and applications. *J. Educat. Meas.*, **40**(3):277-280. [doi:10.1111/j.1745-3984.2003.tb01108.x]

Bourque, G., Pevzner, P.A., 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**(1):26-36.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, **16**:1391-1399.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**(25):14863-14868. [doi:10.1073/pnas.95.25.14863]

Table 1 Taxonomic ranks for the nine organisms

	Genome	Taxonomic rank						
		Kingdom	Phylum	Class	Order	Family	Genus	Species
A1	NC_002607	KA	PE	CH	OH	FH		
A2	NC_012029	KA	PE	CH	OH	FH		
B1	NC_009338	KB	PA	CA	OA	FA	GC	SM
B2	NC_009077	KB	PA	CA	OA	FA	GC	SM
B3	NC_012522	KB	PA	CA	OA	FA	GC	SN
B4	NC_008268	KB	PA	CA	OA	FA	GC	SN
B5	NC_006087	KB	PA	CA	OA	FA	GM	SM
B6	NC_007777	KB	PA	CA	OA	FA	GF	SF
B7	NC_008702	KB	PP	CB	OR	FR		

Fortuna, B., Grobelnik, M., Mladenic, D., 2005. Visualization of text document corpus. *Informatica*, **29**:497-502.

Goldman, D.B., Curless, B., Seitz, S.M., Salesion, D., 2006. Schematic storyboarding for video visualization and editing. *ACM Trans. Graph.*, **25**(3):862-871. [doi:10.1145/1141911.1141967]

Grundy, E., Jones, M.W., Laramee, R.S., Wilson, R.P., Shepard, E.L.C., 2009. Visualisation of sensor data from animal movement. *Comput. Graph. Forum*, **28**(3):815-822. [doi:10.1111/j.1467-8659.2009.01469.x]

Hallin, P., Binnewies, T., Ussery, D., 2008. The genome blastatlas—a genewiz extension for visualization of whole-genome homology. *Mol. BioSyst.*, **4**(5):363. [doi:10.1039/b717118h]

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. *Math. Intell.*, **27**(2):83-85. [doi:10.1007/BF02985802]

Havre, S., Hetzler, E., Perrine, K., Jurrus, E., Miller, N., 2001. Interactive Visualization of Multiple Query Results. Proc. IEEE Information Visualization, p.105-112.

Herniou, E., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Copy, J.S., O′Reilly, D.R., 2001. Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol.*, **75**(17):8117-8126. [doi:10.1128/JVI.75.17.8117-8126.2001]

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**(7):283-290. [doi:10.1016/S0168-9525(00)89076-9]

Karlin, S., Zhu, Z., Karlin, K.D., 1997. The extended environment of mononuclear metal centers in protein structures. *PNAS*, **94**(26):14225-14230. [doi:10.1073/pnas.94.26.14225]

Karlin, S., Brocchieri, L., Mrazek, J., Campbell, A.M., Spormann, A.M., 1999. A chimeric prokaryotic ancestry of mitochondria and primitive eukaryote. *PNAS*, **96**(16):9190-9195. [doi:10.1073/pnas.96.16.9190]

Lu, A., Shen, H., 2008. Interactive Storyboard for Overall Time-Varying Data Visualization. IEEE Pacific Visualization Symp., p.143-150. [doi:10.1109/PACIFICVIS.2008.4475470]

Mao, Y., Dillon, J., Lebanon, G., 2007. Sequential document visualization. *IEEE Trans. Visual. Comput. Graph.*, **13**(6):1208-1215. [doi:10.1109/TVCG.2007.70592]

Meyer, M., Munzner, T., Pfister, H., 2009. MizBee: a multiscale synteny browser. *IEEE Trans. Visual. Comput. Graph.*, **15**(6):897-904. [doi:10.1109/TVCG.2009.167]

Savva, G., Dicks, J., Roberts, I.N., 2003. Current approaches to whole genome phylogenetic analysis. *Brief. Bioinform.*, **4**(1):63-74. [doi:10.1093/bib/4.1.63]

Schbath, S., Prum, B., de Turckheim, E., 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, **2**(3):417-437. [doi:10.1089/cmb.1995.2.417]

Shah, N., Dillard, S.E., Weber, G.H., Hamann, B., 2004. Volume Visualization of Multiple Alignment of Large Genomic DNA. Springer-Verlag, p.325-342.

Trifonov, E.N., Sussman, J.L., 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *PNAS*, **77**(7):3816-3820. [doi:10.1073/pnas.77.7.3816]

Zhou, F., Olman, V., Xu, Y., 2008. Barcodes for genomes and applications. *BMC Bioinform.*, **9**:546. [doi:10.1186/1471-2105-9-546]