High-Dimensional Calligraphic Visualization of DNA and Protein Sequences

Ruth West
University of North Texas
COI/CVAD/CAS/COE
ruth.west@unt.edu

JP Lewis
Victoria University
Computer Graphics
zilla@computer.org

Jeff Burke
UCLA Center for Research in
Engineering Media & Performance
jburke@remap.ucla.edu

Eitan Mendelowitz
Smith College, Computer Science
emendelo@email.smith.edu

Cheryl Kerfeld
DOE Joint Genome Institute
ckerfeld@lbl.gov

Abstract

We present a visualization of nucleic acid and amino acid sequenes based on an ideographic and pictographic language reminiscent of Chinese calligraphy or Sanskrit writing. Motivated by the desire to move beyond the traditional representation of genes as long strings of 'ACTG' nucleotide lists, we visualize characteristics of genes, and their protein products, as the strokes and radicals of a non-phonetic alphabet. Stroke curvature, width, pressure, and the brush profile are varied according to the physical and chemical qualities of each nucleotide and amino acid sequence. This visualization is not "problem or hypothesis driven" rather it is an aesthetically motivated, high-dimensional, and holistic mapping inspired by the parallels between the way in which protein structure specified by DNA reflects its function in an organism and the manner in which form and visual structure in pictographic languages is directly connected to their meaning.

Keywords: Information Vsiualization, Bioinformatics, Fine Art

1 Introduction

Traditionally, DNA and protein sequences within genomic databases are presented as long strings of letters, many of which span hundreds of thousands of characters. Internal organization, structural elements, features of biological interest, and patterns among sequences are difficult, if not impossible, to discern simply by reading these lengthy text strings. Our approach explores the use of calligraphic forms to provide a spatially-compact visualization and afford the simultaneous-display of many genes, enabling the recognition of similarity and homology via pattern recognition (Fig. 1). Conventional tools cannot easily obtain this holistic view in a visceral, human-readable format. The visualizaiton was developed for Ecce Homology, an interactive art work offering viewers an encounter wth genomic biology(http://insilicov1.org).

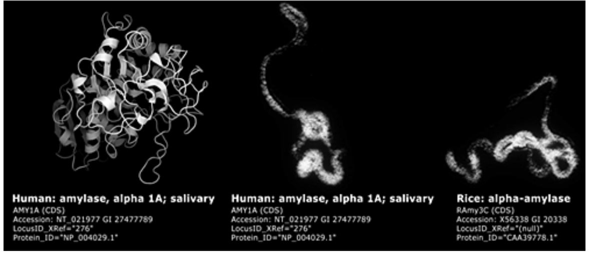
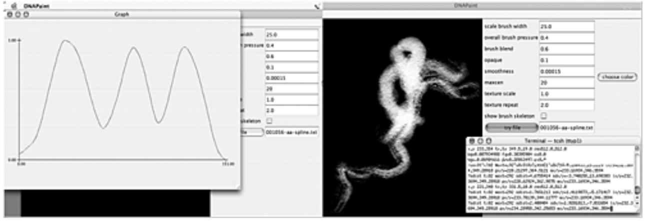


Figure 1. Calligraphic visualization of human and rice amylase proteins.

2 Calligraphic DNA & Protein Visualization

Motivated by the desire to move beyond the traditional representation of genes as long strings of 'ACTG' nucleotide lists, we visualize characteristics of genes as the strokes and radicals of a non-phonetic alphabet inspired by an ideographic and pictographic language reminiscent of Chinese calligraphy or Sanskrit writing. Stroke curvature, width, pressure, and the brush profile are varied according to the physical and chemical qualities of each nucleotide and amino acid sequence. To create each character, a brush stroke generator (BSG) written in C++ reads the nucleotide sequences from GenBank files [1] for the genes that are to be visualized. These are the sole source of input data for the calligraphic strokes—each gene is processed by the same algorithm to generate a unique character. For a given gene, the amino acid sequence expressed by its nucleotides is determined by the BSG and then sent to the BetaTPred2 server [2] for secondary structure and turn prediction. This combined information is used by the BSG to determine the basic shape of the character. Each character is thus fundamentally a two-dimensional structure prediction for a gene, and is represented at this point in the process by a spline output as a text file.



scale brush width 25.0
overall brush pressure 0.4
brush blend 0.6
opaque 0.1
smoothness 0.00015
maxcen 20
texture scale 1.0
texture repeat 2.0
show brush skeleton □
try file 001056-aa-spline.txt

Figure 2. A sample brush profile curve and brush stroke rendering interface

The BSG also calculates pKa, the mass to volume ratio, and hydrophobic effect (side chain burial) along the amino acid sequence, using a hamming window to smooth neighbouring data points slightly. The sharpness of each turn in the spline is refined using the pKa data. Finally, the mass-to-volume ratio and hydrophobic effect are used to create a separate data file that controls the width and brush pressure of each stroke.

The BSG is also capable of generating splines for the individual introns and exons in the nucleotide sequences themselves. The latter become radicals in the resulting composite calligraphic character. (The nucleotide spline generation uses bending and curvature calculations from [3].)

The output splines — amino acid sequence shape, pressure, width and optionally color and intron/exon sequence shapes — are used to drive a separate Java-brush stroke renderer (Fig 2). A naturalistic rendering of gene characters is achieved by modelling a brush depositing ink as it is drawn across a textured sheet of paper. In addition to the spline input, this process accepts a number of parameters for the brush and paper models that can be adjusted to achieve different 'looks'.

3 Conclusion

We have implemented an aesthetically motivated holistic calligraphic gene visualization that incorporates multidimensional biological (genomic and protein) data.

Referernce

[1] NCBI, Sample Genbank Record, http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html.

[2] Kaur, H. and Raghava, G.P.S. (2002) BetaTpred: Prediction of beta-turns in a protein using statistical algorithms. Bioinformatics 18:498-9. http://www.imtech.res.in/raghava/betatpred/

[3] Goodsell, D.S. & Dickerson, R.E. (1994) "Bending and Curvature Calculations in B-DNA" Nucl. Acids. Res. 22, 5497-5503.