Demonstrating that Epigenomic Modifications are Cell-Type Specific Through Clustering and Identifying Biologically Relevant Domains

ADRIANA SPERLEA and DOUGLAS ARNESON, University of California Los Angeles

General Terms: Epigenomics

Additional Key Words and Phrases: Chromatin modifications, PCA, Clustering, F-Score, Purity

ACM Reference Format:

Adriana Sperlea and Douglas Arneson. 2015. Demonstrating that Epigenomic Modifications are Cell-Type Specific Through Clustering and Identifying Biologically Relevant Domains. *ACM Trans. Appl. Percept.* 2, 3, Article 1 (May 2010), 11 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The recent availability of epigenomic data sets such as histone modifications or DNA methylation, coupled with the growing popularity of applying machine learning techniques to genomic data have enabled the computational imputation of genome-wide cross-sample epigenetic marks. In the paper by Ernst and Kellis 2015, the authors developed a method called ChromImpute that accurately predicts a variety of epigenomic data in 127 human cell lines originating from a diverse set of tissues. Previous work has shown that although histone modifications are not cell type specific at promoter and transcription factor binding sites, they show a high level of cell type specificity at enhancer sites, which they are strongly associated with. Thus, it is expected that similar cell types will exhibit similar epigenetic marks. Here, we show that the imputed data from Ernst and Kellis 2015 preserves the cell-type specific aspect of chromatin modifications by running a principal component analysis (PCA) on the dataset from their paper, and then clustering the resulting distribution of cell lines. Certain histone modifications were more cell type specific and thus more accurately assigned similar types to the same cluster. We attempted to gain insight into the biological function of the genomic regions that proved to be most important to the PCA in a certain chromatin modification, and particularly to those that were present across marks by cross-checking these regions with known annotations of the human genome. These regions fell predominantly in non-coding regions, but exhibited a potential association with enhancer sites.

STATSM254 Final Report

DOI: http://dx.doi.org/10.1145/0000000.0000000

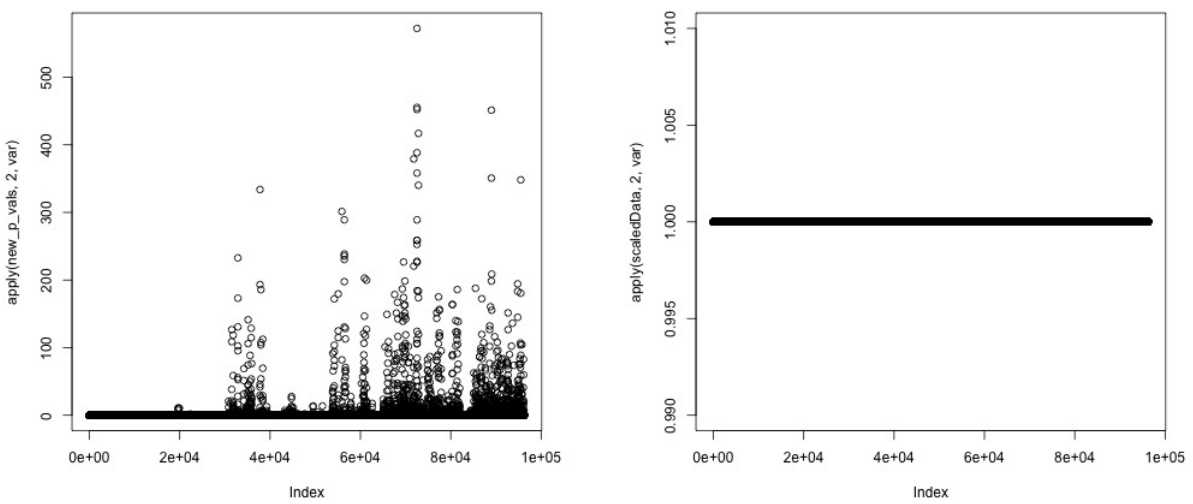

Fig. 1: H3Kme27ac variance of p-values across chromosome 21 before and after normalization. After normalization, the p-values show an uniform distribution. See Supplementary Materials for the variance of the other chromatin modifications before and after normalization.

Table I. : Number of Principal Components Retained in Each Epigenetic Mark

Epigenetic Mark	DNase	H3K4me1	H3K4me3	H3K9ac	H3K9me3	H3K27ac	H3K27me3	H3K36me3	RNAseq
Number of principal components retained	3	5	3	4	5	4	3	5	6

The number of principal components retained in the analysis of each epigenetic mark.

2. PRINCIPAL COMPONENT ANALYSIS

The imputed data from Ernst and Kellis 2015. consists of 9 different chromatin modifications in 127 different cell-types, with a corresponding p-value for each bin of 500 base pairs in the genome indicating the likelihood of having that particular mark at that location. In order for our problem to remain computationally tractable we restricted our analysis to chromosome 21, though data is available for the entire human genome. For each of these 9 chromatin modifications we normalized the data such that the p-values for every location in the genome have mean 0 and variance 1 (Figure 1).

After normalization, we ran a principal component analysis on each of the 9 chromatin modifications. Our results show that upwards of 30 principal components were necessary to explain 85% of the variance which is a common threshold used in the literature, but by analyzing the scree plots of the variance of each component we were able to restrict the number of important components further. We used the broken-stick method Cangelosi and Goriely 2007 to quantify the point in the turn in the scree graph where retaining an additional principal component would be of no benefit (Figure 2). The number of principal components retained in each epigenetic mark are available in Table I.

3. CLUSTERING

Subsequent to data normalization and principal component selection, an exploratory visualization of the transformed data was conducted to determine how best to proceed (Figure 1). This visualization

Table II. : F-Scores and Purity Scores

	DNase	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9ac	H3K9me3	RNAseq
F-Score	0.52380	0.697478	0.4895104	0.5545454	0.5726495	0.4171779	0.713178	0.3972602	0.603351
Purity	0.769230	1	0.777777	1	1	0.8181818	1	1	1

Add any caption you'd like here.

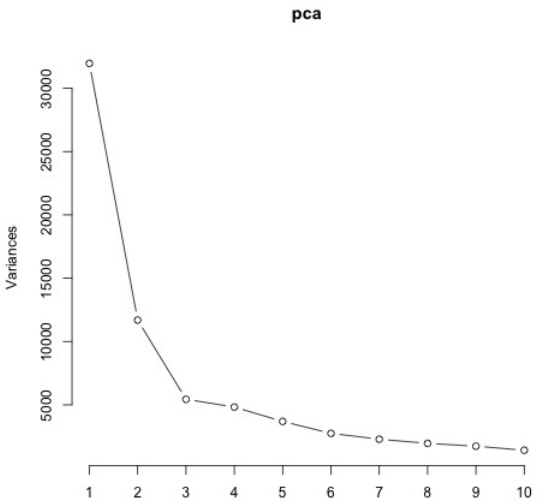

Fig. 2: Scree plot of the proportion of variance explained by each principal component for the DNase assay. The broken stick method selects 3 principal components, because the eigenvalue of the 4th component is not larger than the value given by the broken-stick distribution.

revealed that although some tissue groups were tightly clustered (indicated by red circles), many of the points on the plot were not well clustered, or would intermingle with points from a different tissue group.

Our initial schema was to leverage a k-means approach to identify clusters within the data (in the number of PC dimensions as determined by the skree plots). This approach was based on the hypothesis that the variance from certain histone modifications would allow classification of different tissues groups in the form of clusters [cite]. To facilitate this procedure, the originally furnished dataset was subset to eliminate particular tissue groups that made the task less feasible. The ENCODE2012 and Other tissue groups had representative cell types across multiple tissue types. These samples would be difficult to reclassify under the appropriate tissue group, and not all fit directly into the already established subcategories. Thus, the samples that comprised these two tissue groups were removed from the collection of data. Although this solution made the data more manageable, further reduction was required. The objective function of k-means (Equation 1) has the intrinsic property of giving more weight to larger clusters; this property can be seen in Figure 2. To correct for this cluster size bias, smaller tissue groups, those with less than 5 samples per group, were also removed (mesench, neurosph, thymus, and sm. muscle).

Our basic rational is as follows. Tracing an advection pathway for a particle dropped in a flow field is a perceptual task that can be carried out with the aid of a visual representation of the flow. The task

requires that an individual attempts to trace a continuous contour from some designated starting point in the flow until some terminating condition is realized. This terminating condition might be the edge of the flow field or the crossing of some designated boundary. If we can produce a neurologically plausible model of contour perception then this may be the basis of a rigorous theory of flow visualization efficiency.

Identify. Characteristics of an object.

Locate. Absolute or relative position.

Distinguish. Recognize as the same or different.

Categorize. Classify according to some property (e.g., color, position, or shape).

Cluster. Group same or related objects together.

Distribution. Describe the overall pattern.

Rank. Order objects of like types.

Compare. Evaluate different objects with each other.

Associate. Join in a relationship.

Correlate. A direct connection.

3.1 Conditions

The reproduction of the gestures was performed in the presence or absence of visual and auditory feedback, resulting in four (2×2) conditions.

(1) Visual and auditory feedback ($\mathrm{V} + \mathrm{A}$).
(2) Visual feedback, no auditory feedback (V).
(3) Auditory feedback, no visual feedback (A).
(4) No visual or auditory feedback (None).

The order of the four conditions was randomized across participants.

—*when* + *where* $\Rightarrow$ *what*: State the properties of an object or objects at a certain time, or set of times, and a certain place, or set of places.
—*when* + *what* $\Rightarrow$ *where*: State the location or set of locations.
—*where* + *what* $\Rightarrow$ *when*: State the time or set of times.

When conducting a user study, the goal for the study is to measure the suitability of the visualization in some sense. What is actually measured is a fundamental question that we believe can be handled by using the concepts of effectiveness, efficiency, and satisfaction. These three concepts are derived from the ISO standard of usability 9241-11.

> Extent to which a product can be used by specified users to achieve specified goals with *effectiveness*, *efficiency*, and *satisfaction* in a specified context of use.

The mechanisms of contour perception have been studied by psychologists for at least 80 years, starting with the Gestalt psychologists. A major breakthrough occurred with the work of Hubel and Wiesel [**??**] and from that time, neurological theories of contour perception developed. In this article, we show that a model of neural processing in the visual cortex can be used to predict which flow representation methods will be better. Our model has two stages. The first is a contour enhancement model. Contour enhancement is achieved through lateral connections between nearby local edge detectors. This produces a neural map in which continuous contours have an enhanced representation. The model

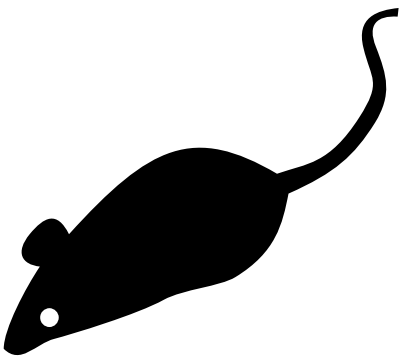

Fig. 3: Neurons are arranged in V1 in a column architecture. Neurons in a particular column respond preferentially to the same edge orientation. Moving across the cortex (by a minute amount) yields columns responding to edges having different orientations. A hypercolumn is a section of cortex that represents a complete set of orientations for a particular location in space.

or cortical processing we chose to apply is adapted from **?**. The second stage is a contour integration model. This represents a higher level cognitive process whereby a pathway is traced.

THEOREM 3.1. *For a video sequence of n frames, an optimal approach based on dynamic programming can retrieve all levels of key frames together with their temporal boundaries in O(n^4) times.*

We apply the model to a set of 2D flow visualization methods that were previously studied by **?**. This allows us to carry out a qualitative comparison between the model and how humans actually performed. We evaluated the model against human performance in an experiment in which humans and the model performed the same task.

Our article is organized as follows. First we summarize what is known about the cortical processing of contours and introduce Li's [**?**] model of the cortex. Next we show how a slightly modified version of Li's model differentially enhances various flow rendering methods. Following this, we develop a perceptual model of advection tracing and show how it predicts different outcomes for an advection path-tracing task based on the prior work of **?**. Finally we discuss how this work relates to other work that has applied perceptual modeling to data visualization and suggest other uses of the general method.

4. CORTICAL PROCESSING OF CONTOURS

Visual information passes along the optic nerve from the retina of the eye where it is relayed, via a set of synaptic junctions in the midbrain lateral geniculate nucleus, to the primary visual cortex at the back or the brain (Visual Area 1 or V1). It has been known since the Hubel and Wiesel's work in the 60s that the visual cortex contains billions of neurons that are sensitive to oriented edges and contours in the light falling on the retina. Such neurons have localized receptive fields each responding to the orientation information contained within the light imaged in a small patch of retina. A widely used mathematical model of a V1 neuron's receptive field is the Gabor function [**?**]:

$$Gabor(u, v, \lambda, \theta, \phi, \sigma, \gamma) = e^{-\frac{u'^2+\gamma^2 v'^2}{2\sigma^2}} cos(2\pi \frac{u'}{\lambda} + \phi). \quad (1)$$

Hubel and Wiesel [**??**] found that neurons responding to similar orientations were clustered together in a structure they called a column which extended from the surface of the visual cortex to the white matter (see Figure 3). Later, they and other researchers discovered hypercolumn structures consisting

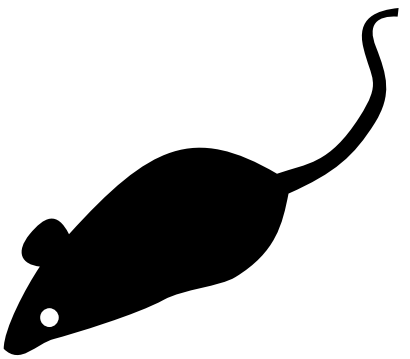

Fig. 4: Neurons whose receptive fields are aligned along a continuous contour mutually reinforce each other. They inhibit nearby neurons with a similar orientation sensitivity.

of thousands of neurons all responding to the same area of visual space and selecting for a range of orientations. Overall, V1 contains a topographic map of the visual field having the property that every part of the retinal image is processed in parallel for all orientations. These orientation selective neurons have provided the basis for all subsequent theories of contour and edge detection.

There remains the problem of how the output of orientation sensitive neurons, each responding to different parts of a visual contour, becomes combined to represent the whole contour. Part of the solution appears to be a contour enhancement mechanism. **?** examined the human's ability to perceive a contour composed of discrete oriented elements. They placed a contour composed of separated Gabor patches, among a field of randomly orientated Gabor patches. Contours were detected when the patches were smoothly aligned. They were not detected when there was misalignment. This work suggests that there is some manner of lateral coupling among the visual elements involved in perceiving the Gabor patches in the contour. These researchers have suggested that similarly oriented aligned contours mutually excite one another, while they inhibit other neurons that are nearby (Figure 4).

5. LI'S V1 MODEL

Based on the observed organization of the neurons in the visual cortex by Hubel and Wiesel [**??**] and the experimental evidence by **?**, Zhaoping Li constructed a simplified model of the behavior of V1 neurons and examined the model's ability to integrate contours across multiple V1 neurons. The model is introduced briefly here, and described in more detail in **?**. In Li's model, the cortex is approximated by a set of hypercolumns arranged in a hexagonal grid. Each hexagonal cell has 12 orientation-selective neuron pairs oriented in 15-degree increments. One of the main simplifications embodied in Li's model is that it fails to incorporate the way the mammalian visual systems scales with respect to the fovea. Real neural architectures have much smaller receptive fields near the fovea at the center of vision than at the edges of the visual field. The neurons in each hex cell were grouped into excitatory and inhibitory pairs responding to an edge of a particular orientation at that location. Thus there were a total of 24 neurons per cell. The firing rates of both the inhibitory and excitatory neurons were modeled with real values. The neuron pairs affected neighboring neuron pairs via a transfer function that depended on the alignment of the edge selectivity orientations. Neuron pairs that were aligned with one another exhibited an excitatory effect on each other, while pairs that were not aligned inhibited each other. Finally, Li's model also contains feedback pathways for higher-level visual areas to influence individual neurons.

In our implementation, the mapping of the hexagonal grid to the image space was such that the hex centers were separated by 10 pixels. For the V1 neuron response, we used the Gabor function (Eq. (1)) with a wavelength, λ, of 21 pixels, a σ of 7 pixels, and an aspect ratio, γ, of 1.

6. STREAMLINE TRACING ALGORITHM

? compared the effectiveness of visualization techniques by presenting test subjects with the task of estimating where a particle placed in the center of a flow field would exit a circle. Six different flow-field visualization methods were assessed by comparing the difference between the actual exit numerically calculated and the estimation of the exit by the human subjects. Laidlaw et al.'s experiment was carried out on humans but, in our work, we apply this evaluation technique to humans as well as to our model of the human visual system and use a streamline tracing algorithm to trace the path of the particle.

We use the term streamline tracing to describe the higher level process that must exist for people to judge a streamline pathway. We call it streamline tracing because the task seems to require the user to make a series of judgments, starting at the center, whereby the path of a particle dropped in the center is integrated in a stepwise pattern to the edge of the field. Though many algorithms exist in the machine vision literature for contour tracing, we found these to be inappropriate for use in this application. Contour tracing algorithms are generally designed to trace out the boundary of some shape but a streamline tracing algorithm must also be able able to produce a streamline in a field of disconnected contours, such as is the case with the regular arrows. The streamline to be traced will often not follow a visible contour but instead be locate between contours, and will sometimes pass through areas devoid of visual elements. Thus we developed a specialized algorithm that is capable of tracing streamlines that do not necessarily correspond to the boundary of any shape but can pass between visual contours.

Perception is a combination of top-down and bottom-up processes. Bottom-up processes are driven by information on the retina and are what is simulated by Li's model [**?**]. Top-down processes are much more varied and are driven in the brain by activation from regions in the frontal and temporal cortex that are known to be involved in the control of pattern identification and attention [**?**]. All of the flow visualizations evaluated by **?**, except for LIC, contain symbolic information regarding the direction of flow along the contour elements (e.g. an arrowhead). In a perpetual/cognitive process this would be regarded as a top-down influence. At present our model does not deal with symbolic direction information but it does do streamline tracing once set in the right general direction.

Streamline tracing is a combination of top-down and bottom-up processes. Broadly speaking, top-down processes reflect task demands and the bottom-up processes reflect environmental information. In our case, the bottom-up information comes from the different types of visualization, while the top-down information is an attempt to model the cognitive process of streamline pathway tracing. Contour integration was modeled using the following iterative algorithm.

ALGORITHM 1: Iterative Algorithm

```
current_position ← center
current_direction ← up
current_position is inside circle
while current_position is inside circle, do
    neighborhood ← all grid hexes within two hexes from current_position
    for each hex in neighborhood, do
        for each neuron in hex do
            convert neuron_orientation to vector
            scale vector by neuron_excitation
            vector_sum ← vector_sum + vector
        end
    end
    normalize vector_sum
    current_position ← current_position + vector_sum
    current_direction ← vector_sum
    return current_position
end
```

The algorithm maintains a context that contains a current position and direction. Initially, the position is the center, and the direction set to upward. This context models the higher-order, top-down influence on the algorithm that results from the task requirements (tracing from the center dot) and the directionality which in our experiment was set to be always in an upwardly trending direction.

The algorithm traces the contour by repeatedly estimating the flow direction at the *current_position* and moving the position a small distance (.5 hex radii) in that direction. The flow direction is calculated from the neural responses in the local neighborhood of the *current_position*. The excitation of each neuron is used to generate a vector whose length is proportional to the strength of the response and whose orientation is given by the receptive field orientation. Because receptive field orientations are ambiguous as to direction (for any vector aligned with the receptive field, its negative is similarly aligned). The algorithm chose the vector most closely corresponding to the vector computed on the previous iteration. Vectors are computed for all neurons in hypercolumns within a 2-hexes radius of the current position; they are summed and normalized to generate the next *current_direction*.

Some changes were made from the method published by **?**. Previously, the algorithm considered only a single hex cell at each iteration of the algorithm. We found that this would occasionally cause unrealistically large errors in streamline tracing. For example, on visualizations with arrowheads, the neural network might yield a very strong edge orthogonal to the flow field positioned at the back of an arrowhead. If the algorithm considered only the edges at this point, it may make a significant error, despite the edges in nearby positions indicating the correct direction. We felt that creating an average over *neighborhood* was the more correct approach, and we found closer agreement with human performance with this change.

6.1 Qualitative Evaluation

Four different flow visualization methods were used in our evaluation of the theory. These were implementations of four of the six used by **?**. We chose to investigate a regular arrow grid because it is still the most commonly used in practice and a jittered arrow grid because of the arguments that have been made that this should improve perceptual aliasing problems [**?**]. We added Line Integral Convolution (LIC) because of its widespread advocation by the visualization community [**?**] and head-to-tail aligned streaklets because of Laidlaw et al.'s finding that is was the best and the theoretical arguments in support of this method [**?**]. Note that Laidlaw et al. used Turk and Banks algorithm to achieve aligned

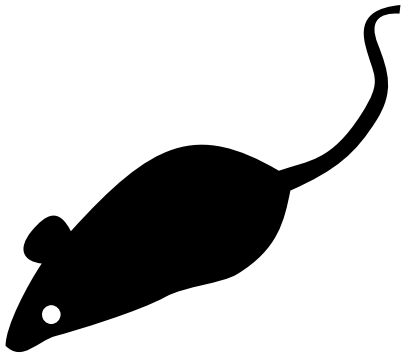

Fig. 5: Regular arrows.

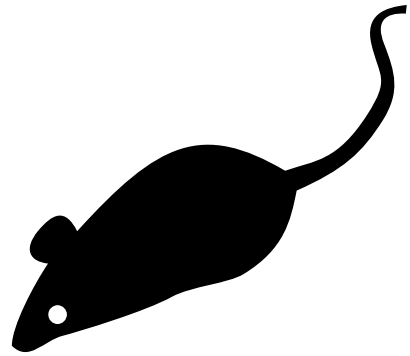

Fig. 6: Jittered arrows.

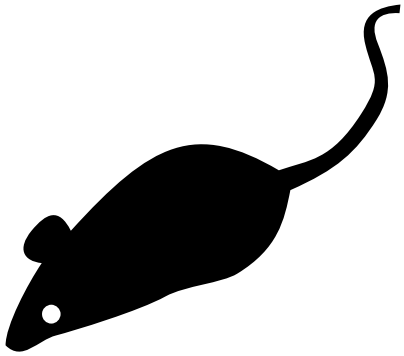

Fig. 7: Closeup of neural response to arrowheads.

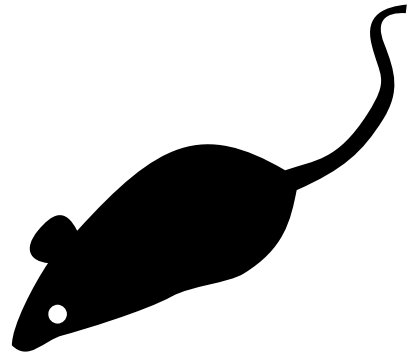

Fig. 8: Closeup of neural response to aligned streaklets.

arrows on equally spaced streamlines while we used Jobard and Lefer's [**?**] method to achieve the same effect and we used streaklets without an arrowhead [**?**].

V1 is known to have detectors at different scales. However, to make the problem computationally tractable we chose only a single scale for the V1 and designed the data visualizations with elements scaled such that they were effectively detected by the gabor filter used by the model. The widths of the arrows and streaklets were chosen to be smaller than the central excitatory band of the gabor filter. This allowed the edge to be detected even if not precisely centered on the receptive field of the neuron. The spatial frequency of the LIC visualization is defined by the texture over which the vector field is convoluted. Our texture was created by generating a texture of random white noise of one-third the necessary size and scaling it up via. interpolation. The resulting spacial frequency of the LIC visualization was of a scale that was effectively detected by the gabor filters of the model.

6.1.1 *Regular Arrows (Figure 5).* This visualization is produced by placing arrow glyphs at regular spacings. The magnitude of the vector field is indicated by the arrow length, and the flow direction by the arrow head. The grid underlying the regular arrows is apparent to humans, but the edge weights of the model show no obvious signs of being negatively affected. In fact, the regularity ensures that the arrows are well spaced, preventing any false edge responses that might be produced by the interference of multiple arrows. We can expect that nontangential edge responses will be produced by the arrowheads and these will lead to errors in the streamline advection task.

6.1.1.1 *Jittered arrows (Figure 6).* This visualization is similar to the regular arrows, but the arrows are moved a small random distance from the regular locations. While composed of the same basic elements as the regular grid, we see instances where nearby arrows interfere with each other and produce edge responses nontangential to the flow direction. Also, as with gridded arrows, the arrowheads will excite neurons with orientation selectivity nontangential to the flow. This can be seen in Figure 7. In this figure, we can see orthogonal neural excitation to each side of the upper arrow, caused by the back edge of the arrowhead (blue circles). We can also see excitation caused by the interference of two arrows at the bottom right (green circle). These nontangential responses are much stronger than those found in the aligned streaklets visualization (Figure 8).

7. DISCUSSION

The overall agreement between the pattern of results for human observers and the V1-based model provides strong support of the perceptual theory we outlined in the introduction. The aligned arrows style of visualization produced clear chains of mutually reinforcing neurons along the flow path in the representation, making the flow pathway easy to trace as predicted by theory.

The fact that LIC produced results as good as the equally spaced streamlines was something of a surprise, and this lends support to its popularity within the visualization community. While it did not produce as much neuron excitation as the aligned arrows method, this was offset by the lack of nontangential edge responses produced by glyph-based visualizations. However, its good performance was achieved only because our evaluation method ignored the directional ambiguity inherent in this method. **?** found this method to be the worst and there is little doubt that had we allowed flow in any direction, up or down, human observers would have found pathways with close to 180 degrees of error half of the time.

The performance of both the model and the human test subjects is likely to be highly dependent on the underlying vector field used. As described in Section 5.1.6, the vector field was generated by interpolating between an 8x8 grid of random, but generally upward pointing vectors. A consequence of this is that when adjacent vectors in this grid point somewhat toward each other, the vector field forms an area of convergence. This convergence area tends to funnel neighboring streamline paths together, reducing error in streamline tracing (Figure 5 is an example of this). Thus, the overall accuracies of both the model and human subjects may be higher than might be might be observed using a vector field without such convergence zones.

We were surprised that the computer algorithm actually did better at the task than human observers. One reason for this may have been that humans would have to make saccadic eye movements to trace a path, whereas the computer did not. For the patterns we used, it is likely that the observers had to make fixations on several successive parts of a path, and errors may have accumulated as they resumed a trace from a previous fixation. Nevertheless, we feel that the algorithm could easily be adjusted to make it give results closer to human subjects. A more sophisticated approach would be to simulate eye fixations.

The model we applied is a considerable simplification over what actually occurs. It only uses the simplest model of the simplest orientation sensitive neurons, and fails to include cortical magnification, among other shortcomings. Real cortical receptive fields are not arranged in a rigid hexagonal grid as they are in Li's model. Furthermore, the neurons of V1 respond to many frequencies, however our model only uses one in its present form. In addition, besides the so-called simple cells modeled by **?**, other neurons in V1 and V2 called complex and hypercomplex cells all have important functions. For example, end-stopped cell respond best to a contour that terminates in the receptive field and understanding these may be important in showing how the direction of flow along a contour can be unambiguously shown. Moreover, visual information is processed through several stages following the

primary cortex, including V2, V4 and the IT cortex. Each of these appears to abstract more complex, less localized patterns. Researchers are far from having sufficient information to model the operations of these stages all of which may have a role in tracing contours. Nevertheless, the results are compelling and there are advantages in having a relatively simple model. We have plans to add some of these more complex functions in future versions of the model.

8. TYPICAL REFERENCES IN NEW ACM REFERENCE FORMAT

A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only output the series if the volume number is given [**?**] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**], a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**?**], [**?**], [**?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], 'YYYYb'-test for prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Boris / Barbara Beeton: multi-volume works as books [**?**] and [**?**].

APPENDIX

With closest point to a given set of lines we intend the point having the minimum Euclidean distance with respect to those lines. Typically, this problem is formulated using Plücker coordinates. Instead, here we compute this point by solving the problem in a closed form, since the resulting matrices are not ill-conditioned in our case. More precisely, by indicating the set of n lines with

$$L = \left\{ l_i = O_i + t\vec{d_i} | \, t \in R \right\} \quad i = 1 \dots n, \tag{2}$$

where O_i is the origin of the ith line and $\vec{d_i}$ is the corresponding direction (normalized), we found the closest point by minimizing

$$p = \arg\min_x \sum_{i=1}^{n} d(x, l_i). \tag{3}$$

The distance $d(x, l_i)$ can be written as

$$d(x, l_i)^2 = (x - O_i) \left[\mathbf{I} - \vec{d_i} \vec{d_i}^T \right] (x - O_i). \tag{4}$$

The minimization is obtained by substituting (4) in (3), and imposing the derivative to zero. After some simple algebra, we obtain the final formulation:

$$p = \left[n\mathbf{I} - \sum_{i=1}^{n} \vec{d_i} \vec{d_i}^T \right]^{-1} \sum_{i=1}^{n} \left[\mathbf{I} - \vec{d_i} \vec{d_i}^T \right] O_i. \tag{5}$$

REFERENCES

Jason Ernst and Manolis Kellis. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues, *Nature Biotechnology*, **33**, 364-376.

Richard Cangelosi and Alain Goriely. (2007) Component retention in principal component analysis with application to cDNA microarray data, *Biology Direct*, **2**, 2.

Received February 2009; revised July 2009; accepted October 2009

Online Appendix to: Demonstrating that Epigenomic Modifications are Cell-Type Specific Through Clustering and Identifying Biologically Relevant Domains

ADRIANA SPERLEA and DOUGLAS ARNESON, University of California Los Angeles

A. ANALYSIS OF INVALID TRIALS

A.1 Results

Invalid trials were previously defined as those trials in which the subject pressed the space bar to end the trial without first bringing the virtual finger to a stop. The number of invalid trials for each subject is presented by feedback condition in Figure 12. Due to the irregular distribution of the data, no significance test was run. However, the figure shows two notable features. First, Subject 6 had more invalid trials than any other subject. Second, more invalid trials occurred under the proprioceptive-only (NV+P) feedback condition than any other.

A.2 Discussion

Although the number of invalid trials is not directly related to task performance, we now consider any trends that may be seen in this information. No statistical tests were done with this data, but some inferences can be drawn from the invalid trial counts in Figure 12. The only obvious trend is that the NV+P condition appears to have the most invalid trials, which is the case for all but two subjects. In the post-experiment survey, one subject commented on this trend, saying that with only proprioceptive motion feedback it was hard to tell if the finger was moving or not. This might be a result of a larger threshold for absolute motion detection for proprioceptive feedback than for visual feedback. This difficulty in stopping the finger did not appear to affect the ease of use ratings provided by subjects, as no correlation was observed with invalid trial counts.

It is interesting to note that the no-feedback condition (NV+NP) had fewer invalid trials than the proprioceptive-only condition (NV+P), especially in light of the findings of Ghez et al. [1990] that deafferented individuals tend to display endpoint drift in non-sighted targeted reaching movements (equivalent to NV+NP condition) while neurologically normal individuals do not (equivalent to NV+P condition). A notable difference between our study and the study by Ghez et al. is the availability of kinesthetic feedback from the thumb pressing on the force sensor, which indicates the magnitude of the applied force, that is, the movement command in our study. Thus, under the no-feedback condition, subjects could use this information to learn to apply grasping forces within the dead zone to stop finger movement. When motion feedback is available, subjects are likely focusing more on the feedback than on the forces applied, since the feedback allows them to achieve better accuracy. Thus, at the end of a trial, subjects are most likely using this feedback as an indicator of zero velocity rather than attending to the applied force. When visual feedback is available, it is easy to determine whether the finger is moving or not; however, when only proprioceptive feedback is available, the finger can be moving

DOI: http://dx.doi.org/10.1145/0000000.0000000

slowly without the subject being aware of its motion. This explanation would result in a larger number of failed trials for the NV+P condition than for any other, as observed.