PHENOTYPING NEURONS WITH PATTERN RECOGNITION OF MOLECULAR MIXTURES

ABSTRACT

Phenotyping cells and tracking their functional states are key tasks in cell biology and molecular medicine. Current cell classification methods are idiosyncratic to specific fields and based on *ad hoc* discovery of presumed univariate markers. We propose a general theory of phenotyping based on broadly distributed multivariate markers as the metrics of classification and standard pattern recognition algorithms as the method of class discovery. We present a real-world test case based on the vertebrate retina and demonstrate that pattern recognition methods can extract singular populations of neurons from complex heterocellular arrays: populations visualized solely as elements in a micromolecular N-space. The applications of this computational approach to cell phenotyping range from phylogenetics to drug discovery to environmental monitoring

1 INTRODUCTION

Vertebrate nervous systems are arrays of $\approx 10^3$-10^4 different neuronal phenotypes. Mammalian retinas, regardless of species or eye size, are composed of 55-60 classes of retinal neurons [1] in copy numbers ranging from 50 to 500,000/mm^2. But this is a tabulation based on decades of piece-wise morphologic work and no metric or analytical method has comprehensively parsed all classes. Tracking cell classes is essential to functional analysis of any cell system. The paucity of candidate probes and persistent uncertainties regarding phenotyping criteria, [2-6] have impeded discovery of solutions. Even so, Famiglietti [3] expressed the consensus view "…that 'natural' cell types emerge as distinct clusters of points in parametric space." He and others expected that the dimensions of this space would be morphologic rather than chemical. However, morphology emerges from a molecular space.

Modern molecular biology has formalized cellular biochemistry as serial interacting compartments: genome → transcriptome → proteome → cytosome. In theory, patterns of transcribed genes and expressed proteins comprise a given cell's *macromolecular phenotype*, prompting searches for cell-specific mRNA or protein signals. The preliminary successes of gene [7] and protein [8] microarrays for screening tissues or purified cell cohorts have not been replicated in complex heterocellular tissues, even though a cell is a pre-printed array. Concurrent probes of multiple genes or proteins with single-cell resolution are still difficult to implement. Neuroscientists have sought univariate *macromolecular* markers for phenotyping, hoping to slowly assemble probe libraries to track multiple cell classes [9]. The strategy has defects: univariate probes need not exist; probe discovery is *ad hoc*; and no ground truth exists for probe validation. Part of the problem is diversity: each mammalian cell expresses some 1000-5000 proteins encoded from a set of ≈ 33,000 transcriptional units [10]. For which proteins should we screen and how? Finally, the expected number of phenotypes in heterocellular tissues is often unknown.

An alternative strategy involves abandoning the search for univariate probes to exploit the fact that every cell also possesses a simpler *micromolecular* mixture of 100-200 major metabolic reactant monomers (amino acids, carboxylyates, nucleic acids, etc.). While virtually none are univariate markers, their steady-state values vary across known cell classes, leading to the obvious notion that intrinsic N-dimensional signatures may prove to be the comprehensive classifiers. Furthermore, extrinsic tracer molecules can be embedded in data sets as surrogates for detection of expressed of ion channels, receptors, and transporters [11]. Sensitive, robust, immunoglobulin (IgG) probe libraries for micromolecular mixtures have been developed [11-14] and are applicable all cell types in all tissues and taxa with constant fidelity. We here describe the use of classical unsupervised pattern recognition [15-17] as a comprehensive and general tool for segmentation and class discovery, independent of morphology. We also show that previously invisible structural attributes (e.g. spatial patterning of classes) become test statistics for validation when unmasked by classification.

2 MATERIALS & METHODS

There are three steps to phenotyping complex cell populations: (1) generating arrays of high-resolution targets serially probed with a designed IgG library; (2) acquiring stable, calibrated registered image data; (3) clustering multichannel data and exploring classifications. We describe a general case based on analysis of a 2D heterocellular array: the mammalian retinal ganglion cell layer.

2.1 Probed Target Arrays

Standard glutaraldehyde quenched, resin embedded retinal samples were sectioned by ultramicrotomy into serial sample arrays of 40-250 nm thickness on multiwell patterned slides [12] and each well probed with one or more IgGs from a library targeting a spectrum of micro-

molecules. As a typical cell is ≈ 10 μm in diameter, serial 250 nm sections yield forty consecutive samples. Signals are visualized with wide-dynamic range, photostable silver detection from a basis set of six intrinsic channels detecting aspartate [D], γ-aminobutyrate (GABA, γ), glycine [G], glutamate [E], glutamine [Q] and taurine [τ] and one extrinsic channel (1-amino-4-guanidobutane, AGB) that reports the excitatory history of the neurons prior to fixation [11]. In the case illustrated here, the neurons were activated *in vitro* with 25 μM AMPA

.

2.2 Data Acquisition

Calibrated images of silver-intensified signals were captured as 8-bit frames under fixed gain, gamma and source irradiance with a DAGE CCD camera using standard brightfield imaging at a resolution of 243 nm/pixel. Complete image datasets ranged from 0.5-3.5 Gb. The key to extracting signatures is image registration. Code developed for planetary imaging (PCI Geomatics, Richmond, Canada) was used to mosaic individual frames into comprehensive channels and align all serial channels with 1st or 2nd order transforms. In principle, then, every pixel in a cell indexes an N-space micromolecular signature.

2.3 Classification and Exploration

Formal classification of N-dimensional datasets is essential to visualization of underlying cell populations. One cannot visually screen R image sets for correlations (R = 1 monochrome, 2 duochrome, 3 standard rgb trichrome), as the number of unique images U is combinatorial with the number of channels N [U = C(N,R)] and we don't know the dimensionality of classification *a priori*. Datasets were clustered using a number of methods, including K-means, isodata, and Narendra-Goldberg algorithms, using commercial (PCI Geomatics) and custom code (IDL, RSI, Boulder, Co). In practice, isodata clustering was effective, fast and produced results identical to the other methods. Cluster separability was evaluated by transformed divergence and calculated as probability of error p_e. In some cases, a derived cluster was clearly a superclass of structures and other methods such as histogram deconvolution were used to further segment the population. Structural datasets were visualized by remapping derived theme classes onto a single morphological channel, their signatures explored by use of superimposed bivariate 2N-plots[13], a scheme inspired by the parallel coordinate space described by Inselberg and Dimsdale [18]. Pairs of signals were displayed as class means bounded by 2 SD margins on axes spanneding 0.1–10 mM with logarithmic scaling. The [x,y] pairs were coded: [AGB, AGB] gray; [E, γ] orange; [D, Q] cyan; [G, τ] magenta. Class significance was tested in part by cells sizes and patterning order in cell distributions (Voronoi tiling, mean spacing distributions). Ordered patterning was gauged by conformity ratios (CR), the mean intercell spacing/spacing variance ratio, using the significance tables of Cook [19].

3 RESULTS

3.1 The Classified ganglion cell layer

A representative input dataset for pattern recognition is shown in Fig.1, with density-coded images. While it is clear that there are dramatic differences in signal patterning across images, no obvious strategy for segmentation in visually accessible 1, 2, or 3 space emerges. It is useful to note, at this point, that these data are unlike multispectral images in a very important way: the channels are already explicitly orthogonal. The glutamate content of cells has no representation in any other channel, thus approaches such as PCA have no obvious value for this problem. That does not mean that the glutamate content of a cell is uncorrelated with the presence of other molecular species, but those correlations are also not necessarily monotonic or time stationary.

The results of isodata clustering are summarized in Fig. 2, which is a refined theme map of classifications. Raw theme maps include an unavoidable error that plagues anatomists. A misalignment kerf is built up around each structure as one sections though these spheroids, resulting in a ring of poorly correlated signals around each cell,

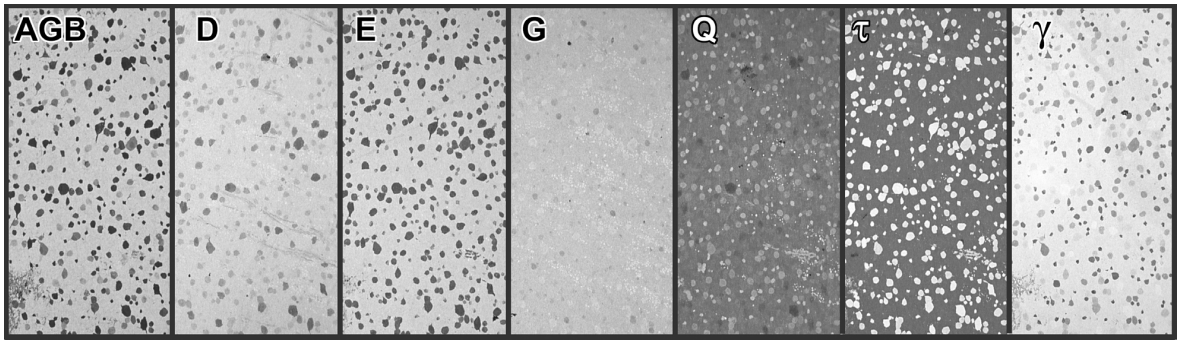

Figure 1. Excitation (AGB) and intrinsic (D E G Q τ γ) micro molecular signals in the rabbit ganglion cell layer visualized in an array of seven registered serial 250 nm sections. Each spot represents a single neuron with sizes ranging from 8-35 μm in diameter.

corrupting size assessments [12]. Refined theme maps are built by using one channel as a structuring object image (e.g. glutamate) and mapping the aligned theme class to each cell/object. Isodata clustering alone results in 10 natural ganglion cell superclasses/classes and 3 natural amacrine cell classes. Superclass 1 is obviously a mixture of different size groups and can be segregated into classes 1a,b,c by size histogram deconvolution [13]. Another superclass demonstrated clear bimodality in D and γ space, yet the subclusters were not separable by our criterion p_e. In this case we deconvolved the signal histograms and assigned cells to classes based on a winner-take-all criterion. The result of this chemical separation was the emergence of two size groupings of cells: class 5 with a diameter of 16.2 ± 2.5 μm and class 9 with a diameter of 34.0 ± 2.1 μm, significantly different at $p < 0.01$ by t-test. Thus, in the end, we achieve 14 ganglion cell and 3 amacrine cell classes, consistent with the dye injection, photofilling and ballistic dye imaging data of Rockhill et al. [20].

3.2 Independent Tests of Classification Significance

Classification is blind to structure in our implementation, though it needn't be. But how do we know any of these classifications are real? How do we know that they are natural classes or functional biologic entities? Fortunately some simple tests emerge. For example, we already knew from other work that starburst amacrine cells represent 30% of the ganglion cell layer, that they are the smallest of the cells, are patterned with a CR of 2, have a robust γ signature and are extremely AMPA sensitive. Class 14 of our clustering results finds these cells based only on the 7-space molecular classification, but reveal them to be 32.7% of all cells, the smallest cells at 8.0 μm, patterned with a CR of 2.3, and the most AMPA sensitive cell in the cohort. More importantly, patterns of cells previously unseen emerge. Fig. 3 shows the distribution of class 6 ganglion cells that we have identified as a specific physiologic type based on comparisons with published data. These and other classes were non-randomly patterned with CR values >3. And all emergent classes turned out to be homogenous size groups that were significantly different from most other classes. Thus population fractions, hidden patterns, and hidden size groups emerge from classical pattern recognition methods applied solely to molecular data sets.

4 CONCLUSIONS

Pattern recognition methods are diverse and have become increasingly sophisticated. We demonstrate here that standard pattern recognition strategies suffice to attain a comprehensive classification of a neuronal cohort, for the first time in the history of neuroscience. While the implications of this approach for retinal scientists are immense, the theoretical implications transcend those. Micromolecular mixtures are, in fact, the tangible manifestation of a functioning proteome, of systems relations among cells, of

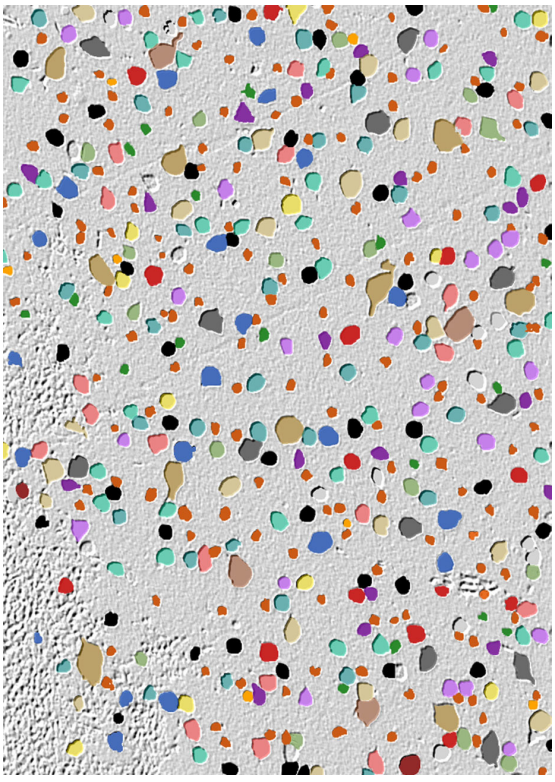

Figure 2. Refined theme map of the rabbit retinal ganglion cell layer derived from 7 molecular channels. Each maps a unique class or superclass of ganglion cells based solely on N-dimensional signatures.

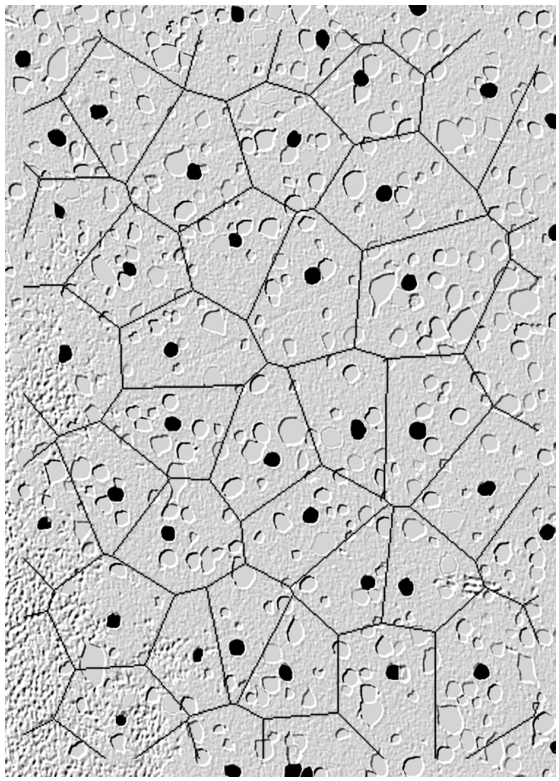

Figure 3. Class 6, OFF center sustained β ganglion cells based on morphology and patterning, surrounded by their Voronoi domains. This class is invisible in any combination of raw image channels.

environmental cues. We have established these methods to be applicable to neural disease models, tracking cell migrations and signature transformations [21, 22], as well as somatic disease models [23]. Though micromolecular probe libraries are not yet widely available, that is a minor barrier to application of these methods. The lack of experience of biologists with the mathematics and computational strategies underlying pattern recognition, however, are formidable impediments. Three solutions appear on the horizon. First, the "rediscovery" of pattern recognition by molecular biologists for analysis of gene clusters offers a portal for transfer of concepts to anatomists. Second, closer working relations are emerging between academic computer science and biological imaging groups. Finally, computer scientists might be willing to teach their colleagues about less challenging methods in classification and segmentation, even as they develop new tools.

5 REFERENCES

[1] R. H. Masland, "Neuronal diversity in the retina," *Current Opinion in Neurobiology*, vol. 11, pp. 431-6, 2001.

[2] R. W. Rodieck and R. K. Brening, "On classifying retinal ganglion cells by numerical methods," *Brain, Behavior and Evolution*, vol. 21, 1982.

[3] E. V. Famiglietti, "New metrics for analysis of dendritic branching patterns demonstrating similarities and differences in ON and ON-OFF directionally selective retinal ganglion cells," *Journal of Comparative Neurology*, vol. 324, pp. 295-321, 1992.

[4] R. J. T. Wingate, T. Fitzgibbon, and I. D. Thompson, "Lucifer yellow, retrograde tracers and fractal analysis characterize adult ferret retinal ganglion cells," *Journal of Comparative Neurology*, vol. 323, pp. 449–474, 1992.

[5] J. E. Cook, "Getting to grips with neuronal diversity: what is a neuronal type?." in *Development and organization of the retina*, B. Finlay, Ed. New York: Plenum, 1998, pp. 91-120.

[6] R. H. Masland and E. Raviola, "Confronting complexity: strategies for understanding the microcircuitry of the retina," *Annual Review of Neuroscience*, vol. 23, pp. 249-84, 2000.

[7] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature Genetics*, vol. 21, pp. 10-14, 1999.

[8] G. MacBeath and S. L. Schreiber, "Printing proteins as microarrays for high-throughput function determination," *Science*, vol. 289, pp. 1760-1763, 2000.

[9] S. Haverkamp and H. Wassle, "Immunocytochemical Analysis of the Mouse Retina," *Journal of Comparative Neurology*, vol. 424, 2000.

[10] FANTOM and RIKEN, "FANTOM Consortium (106 authors) and the RIKEN Genome Exploration Research Group Phase I & II Team (26 authors): Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs," *Nature*, vol. 420, pp. 563 - 573, 2002.

[11] R. E. Marc, "Mapping glutamatergic drive in the vertebrate retina with a channel-permeant organic cation," *Journal of Comparative Neurology*, vol. 407, pp. 47-64, 1999.

[12] R. E. Marc, R. F. Murry, and S. F. Basinger, "Pattern recognition of amino acid signatures in retinal neurons," *Journal of Neuroscience*, vol. 15, pp. 5106-29, 1995.

[13] R. E. Marc and B. W. Jones, "Molecular phenotyping of retinal ganglion cells," *Journal of Neuroscience*, vol. 22, pp. 413-427, 2002.

[14] R. E. Marc and D. A. Cameron, "A molecular phenotype atlas of the zebrafish retina," *Journal of Neurocytology*, vol. in press, 2002.

[15] G. B. Ball and D. J. Hall, "Clustering technique for summarizing multivariate data," *Behavioral Science*, vol. 12, pp. 153–155, 1967.

[16] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*. New York: Wiley, 1973.

[17] J. T. Tou and R. C. Gonzales, *Pattern Recognition Principles*. New York: Addison-Wesley, 1974.

[18] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multidimensional geometry," 1990.

[19] J. E. Cook, "Spatial properties of retinal mosaics: an empirical evaluation of some existing measures," *Visual Neuroscience*, vol. 13, pp. 15-30, 1996.

[20] R. L. Rockhill, F. J. Daly, M. A. MacNeil, S. P. Brown, and R. H. Masland, "The diversity of ganglion cells in a mammalian retina," *Journal of Neuroscience*, vol. 22, pp. 3831-3843, 2002.

[21] B. W. Jones, C. B. Watt, J. M. Frederick, W. Baehr, C. K. Chen, E. M. Levine, A. H. Milam, M. M. LaVail, and R. E. Marc, "Retinal remodeling triggered by photoreceptor degenerations," *Journal of Comparative Neurology*, vol. in press, 2003.

[22] R. E. Marc, R. F. Murry, S. K. Fisher, K. A. Linberg, and G. P. Lewis, "Amino acid signatures in the detached cat retina," *Investigative Ophthalmology & Visual Science*, vol. 39, pp. 1694-702, 1998.

[23] J. C. Jean, Y. Liu, L. A. Brown, R. E. Marc, E. Klings, and M. Joyce-Brady, " gamma -Glutamyl transferase deficiency results in lung oxidant stress in normoxia," *Am J Physiol Lung Cell Mol Physiol*, vol. 283, pp. 766-776, 2002.