

PROSPECTING FOR LIVE CELL BIOIMAGING PROBES WITH CHEMINFORMATIC ASSISTED IMAGE ARRAY (CAIA)

Maria M. Posada(^{1§†}), Kerby Shedden (^{2§†}), Young-Tae Chang(³), Qian Li(³) and Gus R. Rosania(^{1†})*

(1) Department of Pharmaceutical Sciences and (2) Department of Statistics, University of Michigan, Ann Arbor, MI 48109; (3) Department of Chemistry, New York University, New York, NY 10003; [†] Michigan Alliance for Cheminformatic Exploration; [§] contributed equally; *corresponding author –email: grosania@umich.edu;

ABSTRACT

Cheminformatic Assisted Image Array (CAIA) is a data mining and visualization tool linking chemical structures to microscope images of cells incubated with prospective bioimaging probes. In a CAIA, machine vision can be used to calculate the quantitative contribution of the chemical features of candidate probes to the apparent image features. In turn, image arrays can be constructed and sorted based on the features of molecules or images, so that the contribution of chemical features on visual features of the images can be readily discerned, across a large data set. By enabling visualization of complex multidimensional relationships between chemical structures and visual signals, CAIA can facilitate the search for new classes of fluorescent probes of cell structure and function.

Index terms: Machine vision; visualization; chemistry; imaging; microscopy; computer applications.

1. INTRODUCTION

The development of a cheminformatic-plus-imaging toolkit for prospecting bioimaging probes facilitates the search for molecules that exhibit extraordinary intracellular localization patterns, as well as interesting or unique chemical features. For conventional high throughput or high content screening, image analysis and cheminformatic tools are available to search for “hits” based on a molecule’s ability to influence a well-established cellular response—for example, stimulating or inhibiting the translocation of a transcription factor from cytoplasm to the nucleus [1]. In conventional cheminformatics, there are well-validated QSAR techniques to find relationships between chemical structure of a molecule and its pharmacological activity [1, 2]. However, statistical or computational techniques to relate the chemical structure of small molecule bioimaging probes to the spatial features of the probes’ visual signal still remain to be developed.

Automated microscopic imaging instruments [1, 3] can be used to acquire data from live, adherent cells incubated

with fluorescent, cell-permeant small molecules. With large datasets, informatic tools are essential to manage, visualize and analyze quantitative chemical structure-fluorescence signal relationships [1, 2]. Here, we demonstrate how CAIA can be used to organize microscope images into arrays linking probe chemical structure with quantitative image features. Previously, a combinatorial library of 1500 styryl molecules [4, 5] was synthesized and screened with a high content screening instrument [6]. Using the DNA-specific Hoescht dye as a reference marker [6], machine vision techniques were used to extract three basic features associated with a fluorescence signal localized to discrete sites within the nucleus of each cell in each image of the CAIA: total intracellular intensity; nuclear to cytoplasmic ratio; and, coefficient of variation of the pixels in the nuclear region of each cell [6]. Based on statistical regression, the images have been sorted in a way that directly links to regressed contribution of each building block to the measured features of the fluorescence signal (Figure 1). In this manner, global trends in structure-localization relationships between each building block and the subcellular visual features become apparent upon visual inspection, helping discover new bioimaging probes [4].

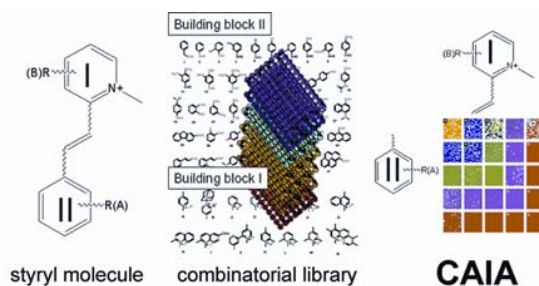


Figure 1. Styryl molecules are made of two building blocks (I, pyridinium or quinolinium; and II, an aromatic ring system). CAIAs can be constructed so that they directly map out to the building blocks of the styryl molecules.

2. METHODS

2.1 Acquisition and selection of images for analysis.

A styryl library was synthesized by condensing eight different pyridinium derivative groups (building block I; Figure 1) with 167 different aromatic aldehydes (building block II, Figure 1), as described [4, 5]. Using a 20X objective lens, live HeLa cells were screened with a Cellomics Kineticscan® instrument, equipped with an environmental-control chamber. Registered 12-bit images were acquired at 1sec and 200 msec exposure times, on FITC and TRITC channels. A 50 msec reference nuclear image was acquired with the Hoechst channel [6].

For analysis, images lacking a detectable fluorescence signal were excluded as follows. Probes with fluorescent signal were detected by comparing the relative change in the background pixel intensities in the 1 sec and 200 msec camera exposures, in the presence and absence of fluorescent probes. For each well, a nuclear mask was constructed by applying a threshold to the Hoechst channel image so as to define each nucleus [6]. Next, a cell mask was constructed by dilating the nuclear mask five pixels, and a background mask was constructed by taking the complement of the nuclear mask dilated by 10 pixels. Least squares regression was then used to compare background pixel intensities in the 1sec and 200 msec exposures, using only pixels below 4095 units in intensity. If the slope of 1 sec vs. 200 msec pixel intensities for the regression was less than or equal to the slope of negative control, unlabeled cells in a specific channel, the probe was deemed non-fluorescent on that channel.

Next, images with cellular fluorescence signal at or below the extracellular background fluorescence, or showing extensive saturation of pixel intensities, were also excluded from analysis. Specifically, images with fewer than 100 pixels in both the cell mask and nuclear mask, or with a ratio below 1.2 between the 75th percentile of pixels in the cell mask and the median of pixels in the background mask were eliminated. Last, images with more than 5000 pixels with intensity greater than 4095 (extensive saturation) were also excluded from analysis. Out of 13824 FITC and TRITC images obtained at 1s and 200 msec exposure times, and under steady state and efflux conditions, 2231 passed all the above filtering. The remaining images were screened for artifacts. Lastly, we removed compounds whose building block II in the combinatorial structure occurred in the image set in combination with less than 3 different building block I for the same experimental condition (either in the presence or absence of extracellular dye). This left 488 1 sec exposure images acquired in the presence of extracellular dye, and 540 efflux 1 sec exposure images acquired in the absence of extracellular dye.

2.2 Measurement of image features.

A cytoplasmic mask [6] was constructed by taking the intersection of the cell mask with the complement of the nuclear mask. Distinct objects were identified, and the

number of topological holes covering more than 5 pixels within each cytoplasmic mask was calculated. Cytoplasmic masks that did not have a single such hole and cytoplasmic masks touching the edge of the image were disregarded. A cytoplasmic mask was also disregarded if the nuclear region contained more than 500 or fewer than 100 pixels, if the cytoplasmic mask area contained fewer than 100 pixels, or if any pixel in the cytoplasmic mask had intensity 4095 (the maximum possible value for a 12-bit image). Next the coefficient of variation and integrated intensity were calculated for each cytoplasmic mask and nuclear mask object. Integrated intensities were background adjusted by subtracting the background median for the whole image times the number of pixels in the object, truncating at zero. The ratio of the background corrected integrated nuclear and cytoplasmic mask intensities was then calculated, henceforth called "nuclear-to-cytoplasmic ratio" or N/C ratio (if the background corrected cytoplasmic intensity was non-positive, the cell was disregarded). Integrated cellular intensities were calculated by summing the background corrected total intensities for the cytoplasmic mask and nuclear regions of each cell. The median values for total cellular intensity, nuclear CV, and N/C ratio were calculated across all objects in the image [6], and used as the primary image-level features for analysis.

2.3 Statistical regression and CAIA assembly.

Because each building block I is combined to every building block II (see Figure 1), a quantitative image feature obtained from the microscopic images can be related to the individual building blocks, using a statistical regression approach [7]. In as much as there may be a simple additive contribution from each building block to a given image feature, CAIA helps visualize the effect of each building block I on every other building block II, and vice-versa. To evaluate the additive effect of styryl building blocks I and II on each image feature (Figure 1), we used linear regression. For a given image feature Y (one of total cellular intensity, N/C ratio, nuclear CV), the linear model $\log(Y) = \alpha(i) + \beta(j)$ was fit, where i is the index of building block I and j is the index building block II [7]. The regression design matrix was reduced to an orthogonal matrix consisting of the left singular vectors of the original design matrix having singular value greater than 0.1. This had the effect of reducing the number of regressors from 189 (the total number of A and B groups) to 173 for images acquired in the presence of extracellular dye and to 167 for images acquired in the absence of extracellular dye. Next, fitted values for $\log(Y)$ were calculated and the Pearson correlation coefficient between the fitted and observed log-scale image feature was used to measure the quality of the additive fit. Randomization was used to assess the whether the fit quality was better than expected by chance. The image feature data was randomly permuted and the whole fitting process was repeated 300 times for data acquired in the presence and

absence of extracellular dye, for each image feature. The proportion of the 300 randomized fits giving greater correlation between observed and fitted values than the actual data was used as an empirical p-value. The average of the fitted versus observed correlation coefficients for randomized data were also recorded. Regression coefficients from the reduced fit were mapped back to the full design space to produce estimates of the $\alpha(i)$ and $\beta(j)$ parameters for each image feature. Using the Miner3D visualization software (Dimension 5, Ltd; Slovakia, EU), CAIAs were arranged based on sorted $\alpha(i)$ and $\beta(j)$ values. Miner3D allows navigating through large image arrays sorted based on chemical or visual features, rapidly zooming into individual cells at high resolution, and then rapidly zooming out of the entire array at lower resolution. We note that images that were excluded from calculating the group scores (because of saturation, lack of signal or other artifacts) were included in the final CAIAs and flagged. These images serve as a test set to assess if the regressed intensity scores are predictive.

3. RESULTS

We found that blocks I and II additively encode the total cell-associated pixel intensity of the styryl probes within each cell. The correlation coefficient between the empirically determined structure scores (the sum of scores for the two building blocks in a particular molecule) and log transformed total cellular intensity (LTCI) was 0.64 for steady state and 0.60 for efflux. In 300 random permutations of LTCI, the average correlation between structure scores and randomized LTCI was 0.5 (steady state) and 0.49 (efflux). None of the 300 randomizations yielded correlations as high as the observed values of 0.64 and 0.60.

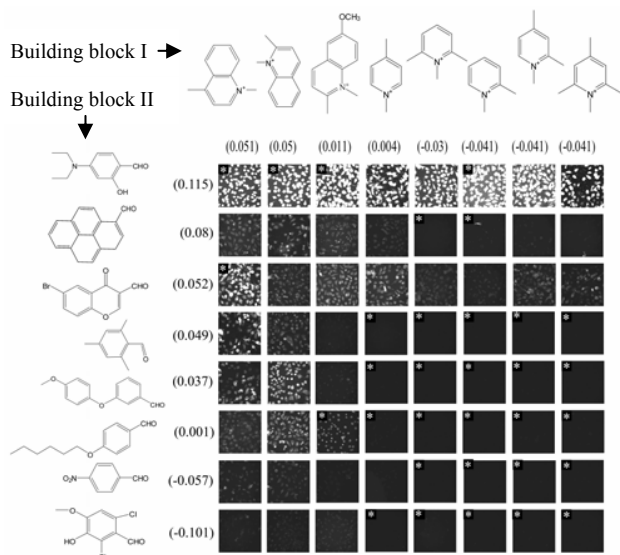


Figure 2. Total intensity regression CAIA. Numbers indicate the regressed contribution of each building blocks to total intensity.

Next, spatial analysis of subcellular fluorescence localization focused on discriminating signal localization in the cell nucleus vs. cytoplasm. In the CAIA framework, this analysis succeeds to the extent that relative pixel intensity in the cell nucleus compared to the cytoplasm is additively encoded by building blocks I or II. Using the CAIA approach, the contribution of each building block I and II to the log transformed nuclear/cytoplasmic fluorescence ratio (LNCR) was calculated across the entire library. The correlation coefficient between the optimal structure scores and LNCR was 0.68 for steady state and 0.65 for efflux. In 300 random permutations of LNCR, the average correlation between structure scores and randomized LNCR was 0.5 (steady state) and 0.49 (efflux), and none of the 300 randomizations yielded correlations as high as the observed values of 0.68 and 0.65. Assembling the CAIA based on the LNCR feature reveals the expected trend, with the cells harboring the brightest nuclei (high N/C ratio) at the upper left hand corner, and cells harboring the darkest nuclei (low N/C ratio) at the bottom right (Figure 3). Upon close examination, at least four images of the top row of the LNCR CAIA show cells bright in the center (nucleus) relative to the periphery, while at least four images on the two bottom rows of the CAIA show cells with a dark center (nucleus) and a bright cytoplasmic fluorescence at the periphery.

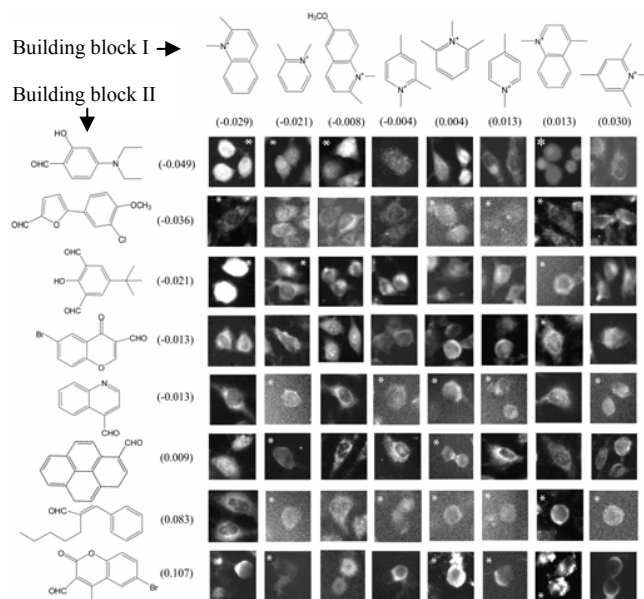


Figure 3. N/C ratio regression CAIA. Numbers indicate the regressed contribution of each building block to N/C ratio.

To determine signal localization to specific intranuclear features such as nucleoli, we used the coefficient of variation of pixels in the nucleus as a measure of the degree of spatial variability associated with the nuclear signal. Presumably, images showing bright and dark spots of fluorescence within the nucleus are indicative of localization of probe at discrete sites, yielding a high coefficient of variation measurement

over the nuclear region of the images. Conversely, images showing diffuse fluorescence signal throughout the nucleus reflect probes that are homogeneously distributed throughout the nucleus. Using the CAIA framework, the correlation between the fitted structure score and log transformed nuclear CV (LNCV) was 0.66 under steady state conditions and 0.53 under efflux conditions. In 300 random permutations of LNCV, the average correlation between structure scores and randomized LNCV was 0.5 (steady state) and 0.49 (efflux). None of the 300 randomizations yielded correlations as high as the actual steady state correlation, but 23 of the 300 randomizations of efflux data yielded correlations exceeding 0.53. Evidently, CV additivity is lost if cells are placed in efflux conditions, consistent with the decrease in nuclear staining described above. Like in the nuclear-to-cytoplasmic ratio CAIA (Figure 3), a relationship between the measured nuclear CV values and the fluorescence localization pattern of the probes within each nucleus was apparent when the sorted CAIA was visually inspected (data not shown).

4. DISCUSSION

CAIA enables visual inspection of relationships between chemical structures of candidate imaging probes and patterns of subcellular signal distribution. In the present example, CAIA is used to sort the images along two axes based on the differential contribution of building blocks I and II (Figure 1) to the intracellular staining patterns. CAIA organizes images in a manner that directly maps out to the chemical structure of the fluorescent molecules, so relationships between chemical structure and subcellular distribution can be readily visualized and sorted, based on the contribution of the different building blocks of the molecules to the signal. The observed relationships are suggestive of specific interactions between probe structure and subcellular organelles, which can be validated in subsequent, biochemical and higher resolution imaging studies, to discover new classes of bioimaging probes [4].

Previously, quantitative-structure localization relationships were discovered in a styryl library by applying statistical regression approach to visual calls made by a human observer [7]. Automated acquisition and detection of subcellular localization patterns from microscopic image data has many advantages over manual, visual inspection [3, 8]. Machine vision yields quantitative measures of visual features and is scalable, adaptable and more objective, reproducible and less cumbersome than visual calls [3, 8]. To facilitate inspection of quantitative structure-localization relationships, CAIA allows the viewer to interact with the images in an intuitive manner, allowing visual inspection of trends occurring across large data sets that may not be so obvious when the images are viewed sequentially.

In conclusion, CAIA facilitates visual analysis of large datasets of images obtained from fluorescent small molecules localizing to different subcellular compartments. Prior to this work, machine vision techniques for classifying subcellular distribution patterns of proteins in cellular organelles have been developed [3, 8]. However, these techniques remain to be applied to small molecule fluorescent probes possessing less specific localization features. CAIA constitutes the first step in that direction.

5. ACKNOWLEDGMENT

This research was funded by NIH grants P20HG003890, R01GM078200 and NSF grant CHE0449139 to KS, GRR and YTC, respectively.

6. REFERENCES

- [1] E.A. Vaisberg, D. Lenzi, R.L. Hanzen, B.H. Keon, and J.T. Finer "An infrastructure for high throughput microscopy: instrumentation, informatics and integration", *Methods Enzymol.* 414, pp. 484-512, 2006.
- [2] E.X. Esposito, A.J. Hopfinger, and J.D. Madura "Methods for applying the quantitative structure-activity relationship paradigm" *Methods. Mol. Biol.* 275 pp 131-214, 2004.
- [3] E. Glory and R.F. Murphy "Automated subcellular location determination and high throughput microscopy" *Dev. Cell* 12(1), pp 7-16. Jan 2007.
- [4] Q. Li, Y. Kim, J. Namm, A Kulkarni, G.R. Rosania, Y.H. Ahn, and Y.T. Chang, "RNA-selective, live cell imaging probes for studying nuclear structure and function," *Chemistry & Biology*, 13(6) pp. 615-623, Jun. 2006.
- [5] G.R. Rosania, J.W. Lee, L. Ding, H.S. Yoon, and Y.T. Chang, "Combinatorial approach to organelle-targeted fluorescent library based on the styryl scaffold" *J Am Chem Soc* 125(5), pp. 1130-1131, Feb 5, 2003.
- [6] V.Y. Chen, S.M. Khersonsky, K. Shedden, Y.T. Chang, and G.R. Rosania "System dynamics of subcellular transport." *Mol. Pharm.* 1(6), pp. 414-425, Nov-Dec 2004.
- [7] K. Shedden, J. Brumer, Y.T. Chang and G.R. Rosania, "Chemoinformatic analysis of a supertargeted combinatorial library of styryl molecules," *J. Chem. Inf. Comput. Sci.*, 43(6), pp. 2068-2080, Nov-Dec 2003.
- [8] X. Chen and R.F. Murphy "Automated interpretation of subcellular location patterns" *Int. Rev. Cytol.* 249, pp 193-227, 2006.