Sosik, H.M. and R.J. Olson. 2006. New technologies to acquire time series of phytoplankton community structure: submersible image-in-flow cytometry and automated taxonomic classification. Proceedings of Ocean Optics XVIII, 15 pp.

New Technologies to Acquire Time Series of Phytoplankton Community Structure: Submersible Image-in-Flow Cytometry and Automated Taxonomic Classification

Heidi M. Sosik and Robert J. Olson
Biology Department, Woods Hole Oceanographic Institution
Woods Hole, MA 02543-1049, USA

ABSTRACT

New sensor and analysis systems are making it possible to characterize taxonomic composition of plankton communities with unprecedented temporal resolution. For example, we have developed a series of automated, submersible instruments for single cell analysis. The latest in this series, Imaging FlowCytobot, uses a combination of flow cytometric and video technology to capture images of organisms and to measure chlorophyll fluorescence associated with each image. Extended unattended deployments are possible because Imaging FlowCytobot operation is automated, including anti-fouling measures and periodic standard analysis. Several-month trial deployments in Woods Hole Harbor have yielded millions of 1-μm resolution images of cells and chains in the size range ~10-200 μm. Ecological interpretation of these data requires automated image identification and enumeration. For this purpose, we have developed image processing and feature extraction methods, combined with a Support Vector Machine for classification into taxonomic categories. After training, a preliminary 22-category classifier provides 88% overall accuracy for an independent test set, with individual category accuracies ranging from 70% to 99%. We demonstrate application of this classifier to a nearly uninterrupted two-month time series of images acquired in Woods Hole Harbor. Our approach, which provides taxonomically resolved estimates of phytoplankton abundance every two hours, permits access to scales of variability from tidal to seasonal and longer.

INTRODUCTION

Many aspects of how natural phytoplankton communities are regulated remain poorly understood, in large part because traditional organism-level sampling strategies are not amenable to high frequency, long duration application. Methods such as conventional microscopic analysis are prohibitively labor intensive and time consuming, while newer and more rapid approaches, such as bulk optical measurements (e.g., chlorophyll fluorescence or light absorption), provide little or no information about taxonomic composition. The lack of appropriate sensor and analysis technologies and the challenges faced even to reach many ocean environments has meant that much of marine ecology is observation limited.

Working to overcome aspects of this limitation, we have developed a series of automated submersible flow cytometers capable of rapid, unattended analysis of individual plankton cells (and other particles) for long periods of time. The first such instrument, FlowCytobot, has proven capable of multi-month deployments (Olson et al. 2003) that provide new insights (e.g., Sosik et al. 2003). FlowCytobot, now in its fourth year of long-term deployment at the Martha's Vineyard Coastal Observatory[1], is optimized for analysis of pico- and small nanoplankton (~1-10 μm). Here we describe a new instrument and analysis system, Imaging FlowCytobot, designed to sample phytoplankton (and microzooplankton) in

[1] http://www.whoi.edu/mvco

the size range 10-100 μm. This is a critical development because phytoplankton in this size range, which include many diatoms and dinoflagellates, can be especially important in blooms events and as sources of new production.

IMAGING FLOWCYTOBOT DESCRIPTION

Imaging FlowCytobot uses a combination of flow cytometric and video technology to both capture images of organisms for identification and measure chlorophyll fluorescence and light scattering associated with each image. The submersible and autonomous aspects were patterned after successes with the original FlowCytobot (Olson et al. 2003), while the addition of cell imaging capability and a design with higher sample volumes are critical for the application to microplankton.

Imaging FlowCytobot (Fig. 1) utilizes a customized quartz flow cell (800 x 180 μm channel), with hydrodynamic focusing of a seawater sample stream in a sheath flow of filtered seawater to carry cells in single file through a red (635 nm) diode laser beam. The fluidics (Fig. 2A) and sampling system are similar to that of FlowCytobot except that, to prevent large particles from settling in the sample injection syringe, the syringe is mounted vertically rather than horizontally, and the flow through the flow cell is downward rather than upward. Each cell passing through the laser beam scatters laser light, and chlorophyll-containing cells emit red (680 nm) fluorescence, which are both collected by a 10X objective focused on the flow cell (Fig. 2B). Scattering signals that exceed a preset threshold trigger a xenon flashlamp strobe to emit a 1-μsec flash of light, which illuminates the flow cell after passing through a green bandpass filter (514 nm) and a condenser. A dichroic mirror behind the objective sends the green light to a high-resolution monochrome CCD camera (1380x1034 pixels) and a frame grabber board (Matrox Meteor II CL) captures the image. The red light from laser scattering (635 nm) and fluorescence (680 nm) excited by the laser are reflected to a second dichroic, which directs each wavelength to a separate photomultiplier module, whose signals are then integrated and digitized. The optical path is folded with mirrors on either side of the flow cell to conserve space. All control and data acquisition functions are carried out by a PC-104plus computer (700 MHz), with custom software for automated operation. Optics and electronics are contained in a pressure housing (12” diameter x 30” long, flushed with dry nitrogen to prevent condensation) (Fig. 1C). Images are processed in real time to locate and store only the region of interest within each frame and timing control ensures that the image and associated fluorescence and light scattering signals are correlated for each particle that triggers the system (Fig. 3).

Custom electronic circuitry was fabricated for triggering the flashlamp and camera, as well as for integrating the optical signals and determining optical signal duration. All other functions are carried out with commercially available PC-104plus boards (frame grabber, analog-to-digital conversion, and logic pulse generation for controls). Power is input at 36V DC (~100 W) and the instrument is designed for 2-way Ethernet-based communication during deployment.

As is routine in the original FlowCytobot, a programmable syringe pump/distribution valve system (Fig. 2A) allows us to inject antifouling and cleaning agents, as well as internal standards such as fluorescent microspheres that enable us to monitor instrument performance during extended operation. Experience to date suggests that a 1-2 day interval for automated antifouling measures (which take only a few minutes to execute) is sufficient to prevent bio-fouling from affecting deployment duration.

IMAGE ANALYSIS AND CLASSIFICATION

Imaging FlowCytobot can generate more than 10,000 high quality plankton (and/or detritus) images every hour, and it can do so every day for months. This precludes manual inspection for cell identification as a feasible tool for many applications. For this reason, we have developed automated procedures to classify a variety of image types, with emphasis on morphologically distinct taxonomic groupings and on accurate estimation of group abundance. The method involves five main steps: 1) image processing and feature (characteristics or properties) extraction, 2) feature selection to identify an optimal subset of characteristics for multi-category discrimination, 3) design, training, and testing of a machine learning algorithm for classification (on the basis of selected features as input), 4) statistical analyses to estimate category-specific misclassification probabilities for accurate abundance estimates and for quantification of uncertainties in abundance estimates following the approach of Solow et al. (2001), and 5) application of the resulting feature extraction, classifier algorithm, and statistical correction sequence to sets of unknown images.

Our present approach incorporates 22 explicit image categories (or classes, in the machine learning vernacular), plus one more for "other" or unidentifiable images (typically 5-15% of the total). Most categories are phytoplankton taxa at genus level (see Fig. 4 for complete list), but a few are more generic ("ciliates", a category of heterotrophic nanoplankton, and "detritus", a mixture of non-cellular material and debris). To complete steps 2) and 3) above, we compiled a set of 6600 images (300 per category) that we visually inspected and manually identified; these images were randomly split into "training" and "test" sets, each containing 150 images from each category. For step 4), we further inspected *every* image acquired during selected periods of natural sample analysis (~10,000 image subset of the time series described below) for manual identification, and subsequent determination of the matrix of inter-category misclassification probabilities under real sampling conditions.

There have been a variety of previous efforts dealing with related image classification problems in marine ecology, notably several lines of development towards identifying zooplankton images to taxa, in some cases as fine as to species (e.g., Davis et al. 1996; Tang et al. 1998; Grosjean et al. 2004; Hu and Davis 2006). As well, there is a lot of relevant work on other image processing and classification applications, such as face recognition and fingerprint recognition. Our approach builds upon some previous work in these areas, while addressing the particular combination of image characteristics and identification markers relevant for Imaging FlowCytobot measurements of nano- and micro-phytoplankton.

All imaging processing and feature extraction was done with the MATLAB (Mathworks, Inc.) software package, including the standard MATLAB image processing toolbox. We use standard MATLAB morphological processing and property functions to determine some features (e.g., length, area) of regions of interest (i.e., cells) in an image, but we have also used functions described in Gonzalez et al. (2004) and implemented in the accompanying "Digitial Image Processing for MATLAB" (DIPUM) toolbox. These include invariant moments, texture properties, and use of Fourier descriptors to get simplified region boundaries. Additional features we calculate include 1) co-occurrence matrix statistics (standard MATLAB toolbox) following the success of Hu and Davis (2005) with this technique for zooplankton images; 2) digital diffraction pattern sampling (custom MATLAB code), previously shown to be effective for fingerprint and other recognition problems

(Berfanger and George 1999); and 3) a few custom shape, symmetry, and other measures such as the number of line segment ends on the region perimeter (indicative of the presence of spines for instance). The resulting feature set has 210 independent quantities associated with each image. All features are transformed to have mean = 0 and standard deviation = 1 (based on distributions from the training set) before use in classification.

Many of the features we calculate require information about the boundary of the region of interest within an image, so preliminary image processing is critical for edge detection and boundary segmentation. We found that conventional edge detection algorithms are inadequate for reliable automated boundary determination over the range of image characteristics and plankton morphologies that we encounter with Imaging FlowCytobot. For this reason, we turned to a computationally intensive but effective approach based on phase congruency (Kovesi 1999) and implemented in MATLAB by Kovesi (2005).

Because inclusion of redundant or uninformative features can compromise the overall classifier performance, feature selection algorithms can be useful to choose the best features for presentation to a machine learning algorithm. We have used the "Greedy Feature Flip Algorithm" (G-flip), a greedy search approach for maximizing a margin-based evaluation function, as described by Gilad-Bachrach et al. (2004b) and available in a MATLAB implementation (Gilad-Bachrach et al. 2004a). With our current 22-category problem applied to the manually identified training set (150 image from each category), G-flip selection reduces our feature set from the original 210 elements down to 133.

For our multi-category classification problem, we use a Support Vector Machine (SVM) with a radial basis function kernel and 10-fold cross-validation (with the training set) for model (parameter) selection. After model selection, we train the SVM (fixed with the best model parameters) using the entire training set and, and then test overall performance with the independent test set of images. The algorithms we use have been implemented with a MATLAB interface as the LIBSVM package (Chang and Lin 2001). LIBSVM uses a one-against-one approach to the multi-category problem, as justified by (Hsu and Lin 2002), and includes an extension of the SVM framework to provide probability estimates for each classification, according to Wu et al. (2004). Application of our selected feature set and this SVM framework provides excellent results for many genera of phytoplankton (> 90% correct identifications on the test set), and overall classification accuracy across all 22 categories in the test set is 88% (Fig. 4).

As the final step towards abundance estimates in each category, we used manual analysis of all images in randomly selected field samples, combined with the above automated classifier, to produce a (23-by-23 element; 22 categories plus "other") matrix of classification probabilities, where the diagonal elements represent the probability of detection for each category and the off diagonal elements are misclassification probabilities for each possible combination of categories. We then used this information to correct abundance estimates for expected misclassification errors and to calculate approximate standard errors for the abundance estimates, according to Solow et al. (2001). In applying this approach, we take advantage of the probability estimates available from the LIBSVM classification results and utilize only initial classifications with relatively high certainty ($p > 0.75$); this leads to lower values of detection probability for some categories, but gives better overall performance (lower residuals relative to manual results) for corrected abundance estimates. For the manually identified field samples, this procedure results in ~80% of the classifier-based abundance estimates falling within 2 standard errors of the manual results and no significant biases between the manual and

classifier -based results (e.g., Fig. 5). The few cases with poorer performance are concentrated in one phytoplankton genera (*Chaetoceros*, which is challenging due to its morphological diversity) and in several relatively non-specific categories: "detritus", "nanoflagellates", and "other < 20 μm".

TIME SERIES RESULTS

We carried out trial deployments of Imaging FlowCytobot at the Wood Hole Oceanographic Institution dock during February-April of 2005. Here we present results from analyses of more than 1.5 million images collected over an 8-week period. Imaging FlowCytbot was connected to power and data communication systems analogous to those at the Martha's Vineyard Coastal Observatory, and all control and data acquisition was fully automated. Images were processed and classified as described above and cell concentrations were determined with 2-hour resolution.

Historical observations in waters near Woods Hole point to late winter-early spring as a period of transition in the phytoplankton community. Blooms of large-celled species and chain-forming diatoms are more commonly found in fall and winter than at other times of year (e.g., Lillick 1937; Riley 1947; Glibert et al. 1985). Our Imaging FlowCytobot observations capture this transition in unprecedented detail. In late February, the most abundant nano- and microphytoplankton (besides the mixed class of ~10-20 μm rounded cells that cannot be taxonomically discriminated from our images) were chain-forming diatom species, especially *Chaetoceros* spp., *Dactyliosolen* spp., and *Guinardia* spp., which were present at approximately 20 chains ml^{-1}, 15 chains ml^{-1}, and 10 chains ml^{-1}, respectively (other taxa were at levels of 3 cells ml^{-1} or less). By mid-March , the previously abundant diatom genera had declined by ~1-3 orders of magnitude, to near undetectable levels. The full 2-hour resolution time series emphasize the power of these observations for exploring ecological phenomena, such as species succession (Fig. 6). The dominant diatoms all declined over the 2-month sampling period, but they responded with very different temporal trajectories. For example, *Dactyliosolen* spp. and *Guinardia* spp. started at similar concentrations, but *Dactyliosolen* spp. declined roughly exponentially over the entire period, while *Guinardia* spp. persisted until a more rapid decline in early April (Fig. 6). The full time series are also rich with even higher frequency detail, such as fluctuations associated with the complexity of water masses and tidal currents in Woods Hole Harbor (e.g., Fig. 7).

FUTURE PROSPECTS

The prototype Imaging FlowCytobot instrument has proven reliable and effective for characterizing phytoplankton community structure with high taxonomic and temporal resolution. At the same time as we proceed with field deployments of the prototype for ecological studies, we anticipate continued efforts to improve and expand the capabilities of this instrument series. Fruitful avenues for research and development may include hardware and algorithm advances to facilitate rapid onboard image classification, efforts to reduce size and power constraints, and other design modifications to enable deployment on a wider range of oceanographic platforms and under a wider range of environmental conditions.

We have recently begun deployments of the existing Imaging FlowCytobot at the Martha's Vineyard Coastal Observatory, located in 15 m of water on the New England continental shelf near Woods Hole. This study site is where we have operated the original FlowCytobot (for pico- and small nanoplankton observations) for several years. With these two instruments, FlowCytobot and Imaging FlowCytobot,

now side-by-side, we can make high temporal resolution observations of the entire phytoplankton community, ranging from picoplankton to chain-forming diatoms, and do so for extended periods (months to years). We expect these kinds of observations will provide new insights into ecological processes and responses to environmental perturbations. Expansion of these capabilities in the context of a new generation of ocean observatory infrastructure, such as proposed through the NSF ORION[2] program, is a very exciting prospect.

ACKNOWLEDGMENTS

This research was supported by grants from NSF (Biocomplexity IDEA program and Ocean Technology and Interdisciplinary Coordination program) and by funds from the Woods Hole Oceanographic Institution (Ocean Life Institute, Coastal Ocean Institute, and Access to the Sea Fund). We are indebted to Alexi Shalapynok for expert assistance in the lab and field; to Melissa Patrician for hours of manual image classification; to Cabell Davis, Qiao Hu, Kacey Li, Mike Neubert, and Andy Solow for insights into image processing, machine learning, and statistical problems; and to the Martha's Vineyard Coastal Observatory operations team, especially Janet Fredericks, for logistical support.

REFERENCES

Berfanger, D. M., and N. George. 1999. All-digital ring-wedge detector applied to fingerprint recognition. Appl. Opt. 38: 357-369.

Chang, C.-C., and C.-J. Lin. 2001. LIBSVM -- A library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Davis, C. S., S. M. Gallager, M. Marra, and W. K. Stewart. 1996. Rapid visualization of plankton abundance and taxonomic composition using the Video Plankton Recorder. Deep-Sea Res. II 43: 1947-1970.

Gilad-Bachrach, R., A. Navot, and N. Tishby. 2004a. Large margin principals for feature selection. http://www.cs.huji.ac.il/labs/learning/code/feature_selection/.

---. 2004b. Margin based feature selection - theory and algorithms. ACM International Conference Proceeding Series, Proceedings of the twenty-first international conference on Machine learning 69: 337-343.

Glibert, P. M., M. R. Dennett, and J. C. Goldman. 1985. Inorganic carbon uptake by phytoplankton in Vineyard Sound, Massachusetts. II. Comparative primary productivity and nutritional status of winter and summer assemblages. J. Exp. Mar. Biol. Ecol. 86: 101-118.

Gonzalez, R. C., R. E. Woods, and S. L. Eddins. 2004. Digital Image Processing Using MATLAB. Prentice Hall. Upper Saddle River, NJ.

Grosjean, P., M. Picheral, C. Warembourg, and G. Gorsky. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES Journal of Marine Sciences 61: 518-525.

Hsu, C.-W., and C.-J. Lin. 2002. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13: 415-425.

Hu, Q., and C. Davis. 2005. Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. Mar. Ecol. Prog. Ser. 295: 21-31.

[2] http://www.orionprogram.org/

---. 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar. Ecol. Prog. Ser. 306: 51-61.

Kovesi, P. 1999. Image features from phase congruency. Videre: A Journal of Computer Vision Research, MIT Press 1: 1-26.

Kovesi, P. D. 2005. MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia. http://www.csse.uwa.edu.au/~pk/research/matlabfns/>.

Lillick, L. C. 1937. Seasonal studies of the phytoplankton off Woods Hole, Massachusetts. Biol. Bull. mar.biol. Lab.,Woods Hole 73: 488-503.

Olson, R. J., A. A. Shalapyonok, and H. M. Sosik. 2003. An automated submersible flow cytometer for pico- and nanophytoplankton: FlowCytobot. Deep-Sea Res. I 50: 301-315.

Riley, G. A. 1947. Seasonal fluctuations of the phytoplankton population in New England Coastal Waters. J. Mar. Res. 6: 114-125.

Solow, A., C. Davis, and Q. Hu. 2001. Estimating the taxonomic composition of a sample when individuals are classified with error. Mar. Ecol. Prog. Ser. 216: 309-311.

Sosik, H. M., R. J. Olson, M. G. Neubert, A. A. Shalapyonok, and A. R. Solow. 2003. Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. Limnol. Oceanogr. 48: 1756-1765.

Tang, X. and others. 1998. Automatic plankton image recognition. Artificial intelligence review 12: 177-199.

Wu, T.-F., C.-J. Lin, and R. C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5: 975-1005.

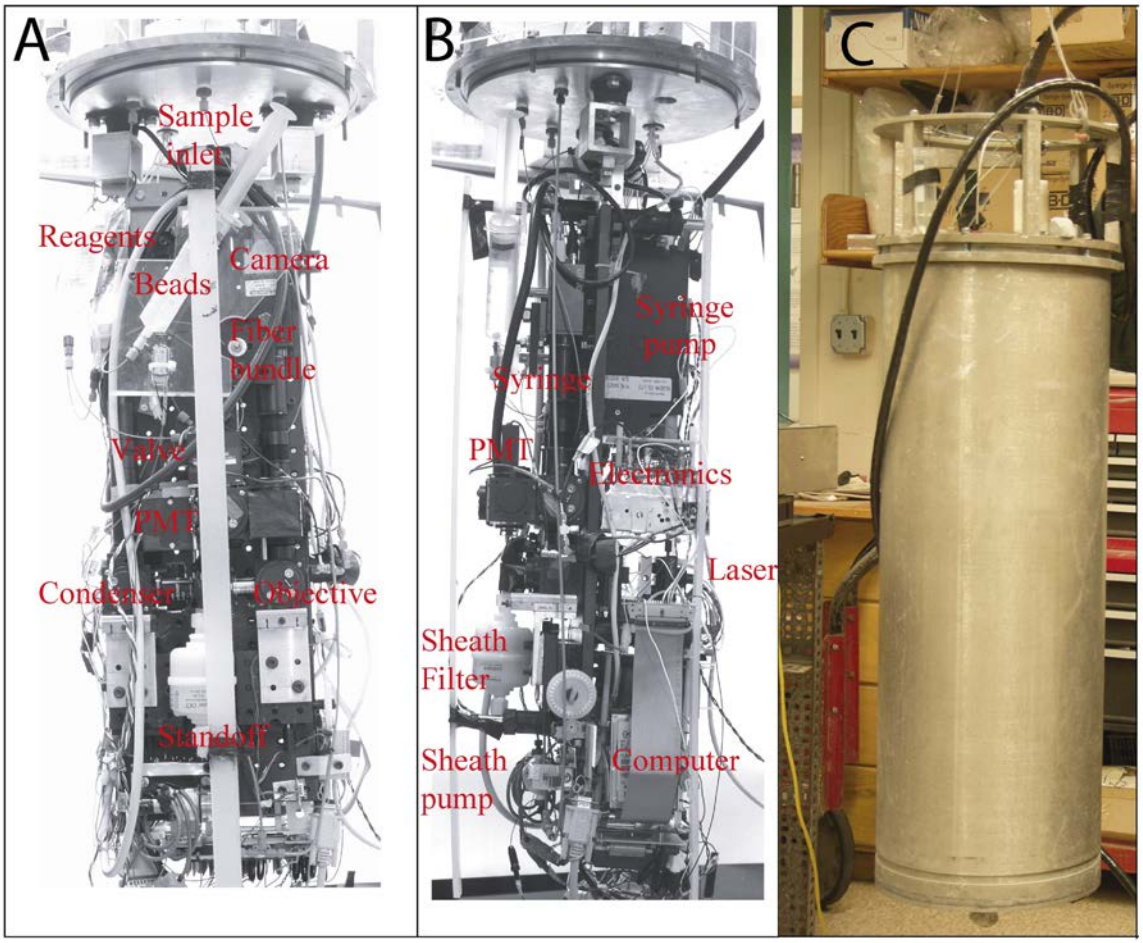

Figure 1. Three views of Imaging FlowCytobot. Removed from underwater housing: A) front view of fluidics components and B) side view with optical breadboard in center, fluidics components mounted on left and electronics on right. C) In its pressure housing ready for deployment. Flow cell (not visible) is located between the condenser and objective lenses (view A). Three plastic standoffs (one of which is easily visible running vertically near the center in view A) prevent contact of the components and housing during installation.

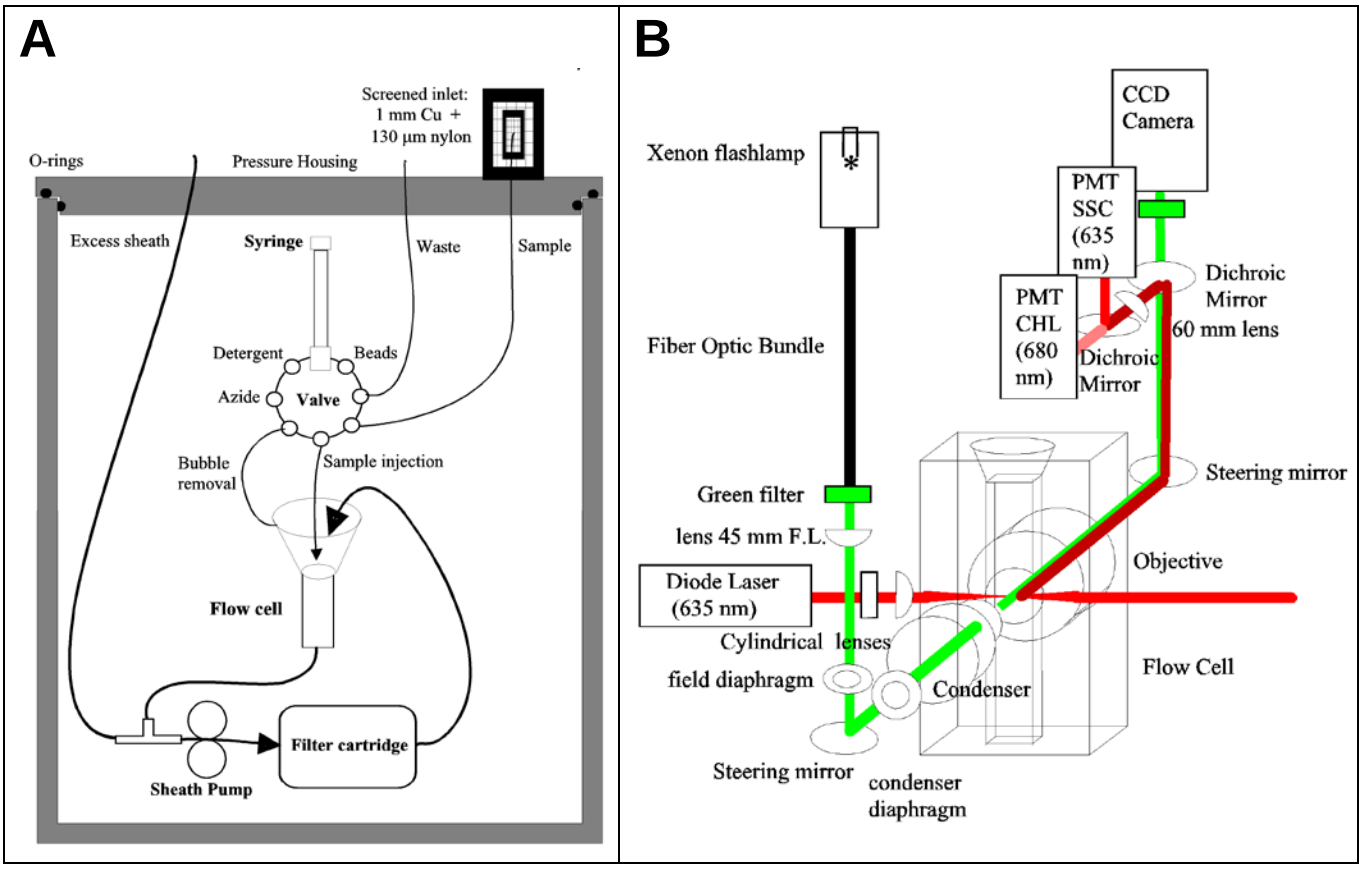

Figure 2. Schema of A) fluidics system and B) optical layout for Imaging FlowCytobot (not to scale). [Panel A shows the fluid path for sample seawater coming from outside the instrument pressure housing, proceeding through the programmable syringe valve to the sample syringe, and then back through the valve to the flow cell and ultimately to the filter cartridge for sheath recycling. Panel B shows the optical paths of the red laser and xenon flashlamp on the 2 excitation sides of the flow cell and the signal pick up paths (light scattering, chlorophyll fluorescence, and camera image) on the side opposite the strobe path (and 90 degrees from the laser path).]

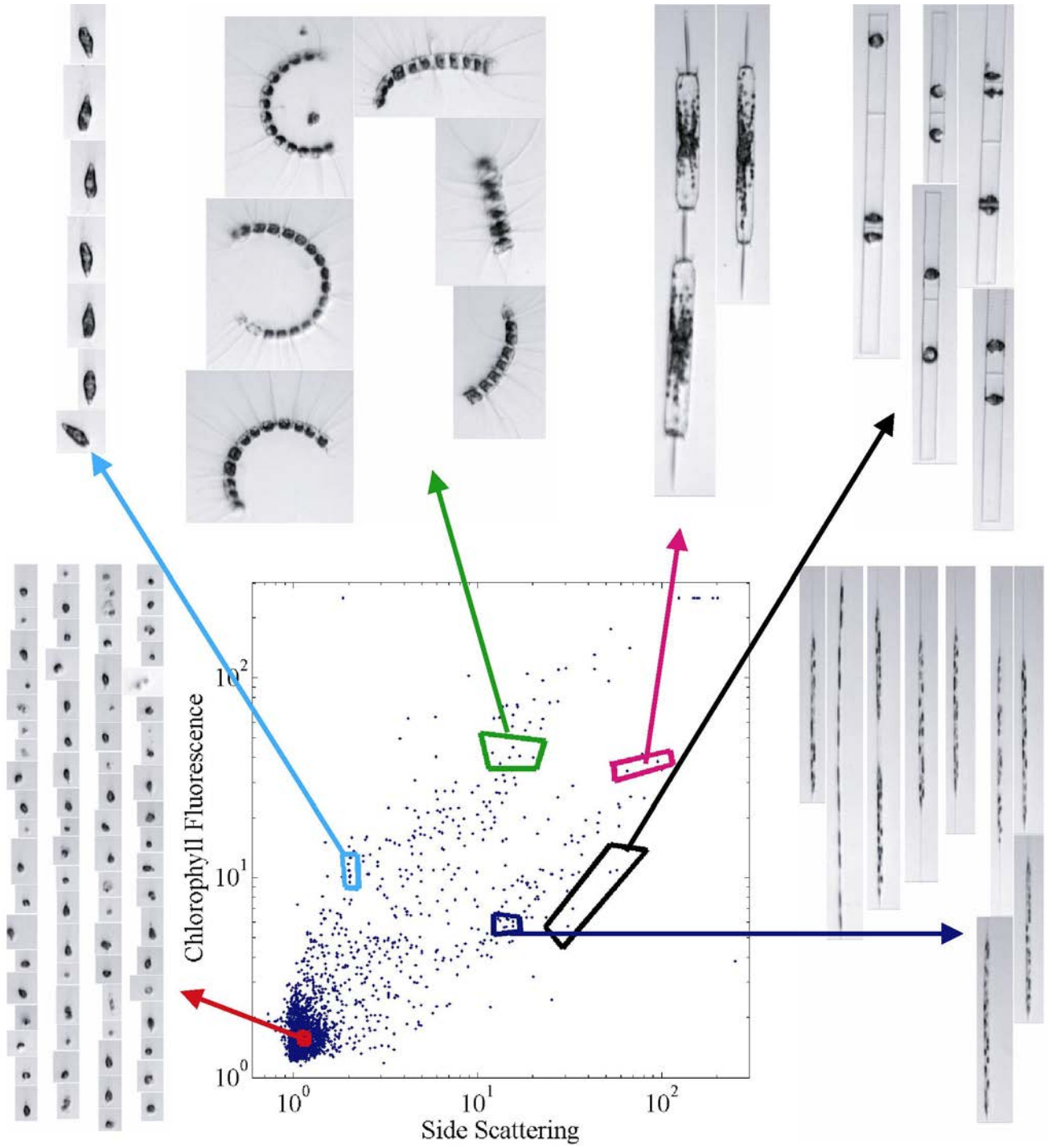

Figure 3. Flow cytometric measurements of individual cell side scattering and chlorophyll fluorescence, and selected images of phytoplankton cells in seawater from the WHOI dock (Dec. 2004), analyzed by Imaging FlowCytobot (triggered by chlorophyll fluorescence). Each dot in the fluorescence-scattering signature has an associated image, only a few of which are shown here. All images are displayed at the same scale; the smallest cells are ~5 μm. Different regions in the fluorescence-scattering signature contain different species, but population boundaries are indistinct. Clockwise from lower left: mixed "small cells",* Euglena *sp.,* Chaetoceros *spp.,* Ditylum *sp.,

Dactyliosolen *sp*., Rhizosolenia *spp*.

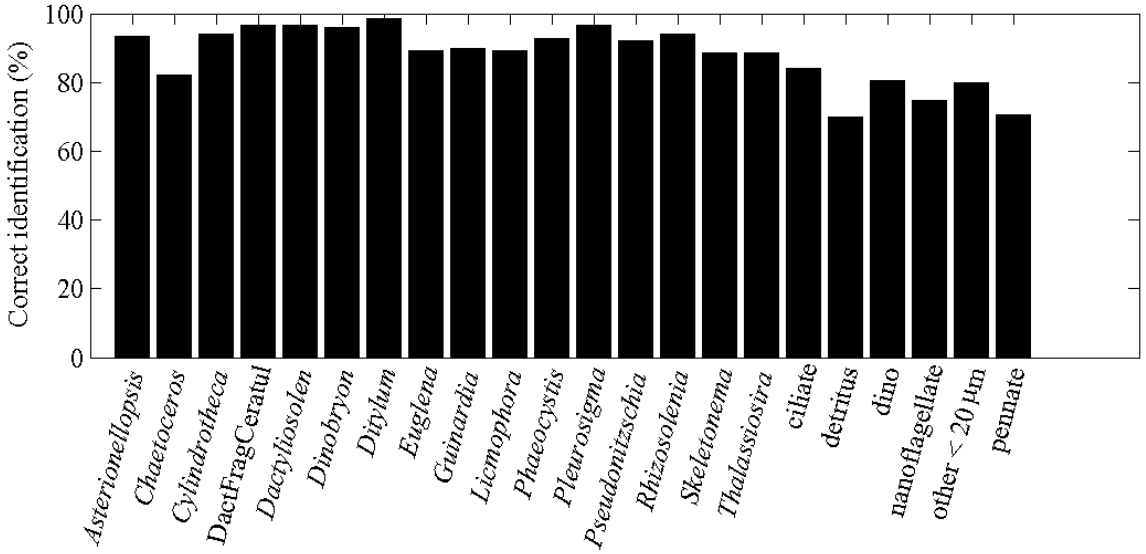

Figure 4. Automated classification results for 22 categories in the independent test set of images (i.e., images not used for classifier development). Values shown here represent the percentage of images manually placed in each category that were also placed there by the automated classifier. The class labeled "DactFragCeratul" represents a mixture of the morphologically similar species* Dactyliosolen fragilissimus *and* Ceratulina *spp.; "ciliate", "dino", "nanoflagellate", and "pennate" represent mixed species of ciliates (heterotrophs), dinoflagellates > ~20 µm, other small flagellated cells, and pennate diatoms, respectively; "detritus" is non-cellular material of various shapes and sizes and "other < 20 µm" are small cells that cannot be taxonomically identified from the images; all other labels are genus names. Percent correct classification ranges from 70-99% for the different categories, with an average of 88%.

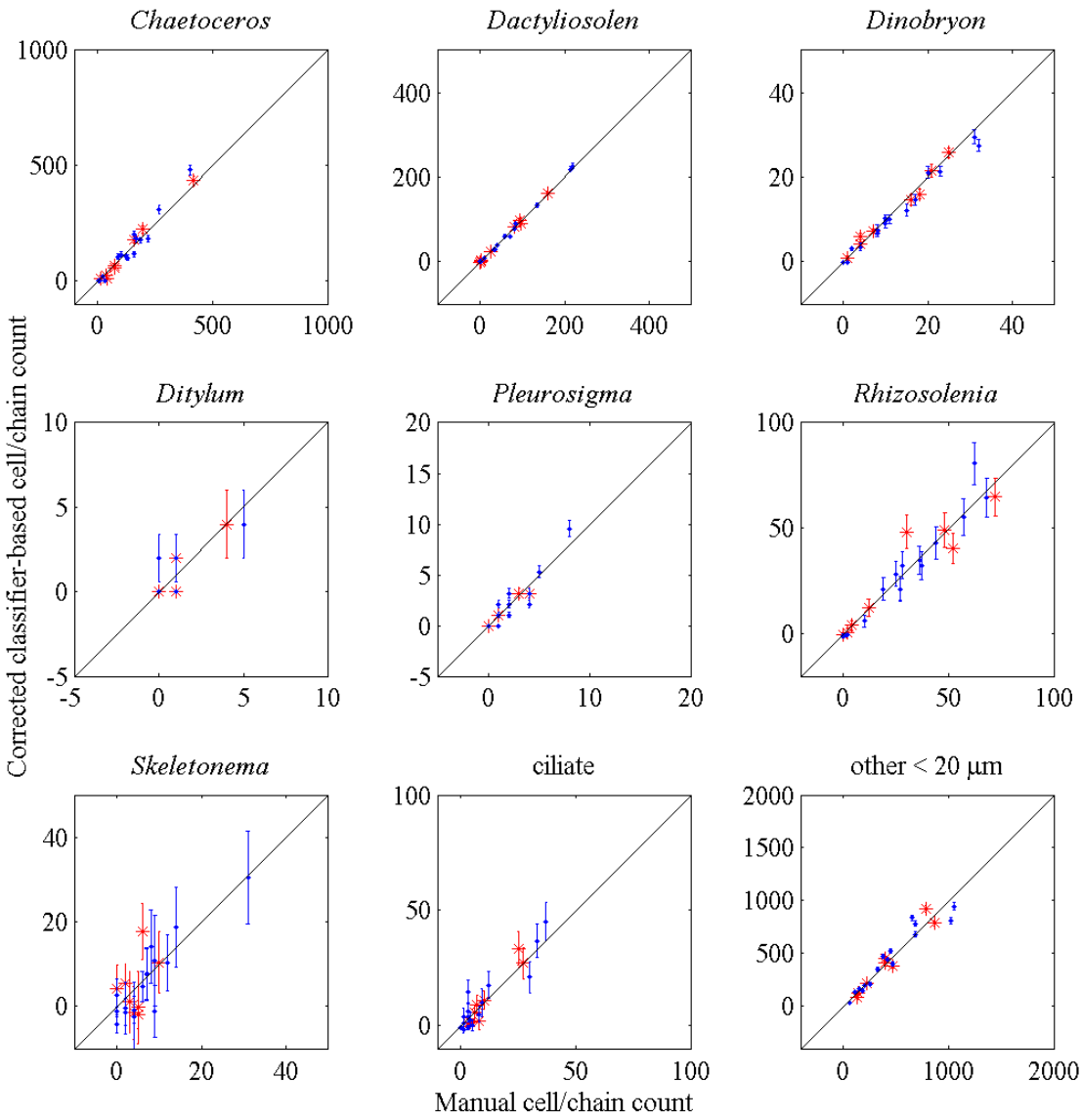

Figure 5. Comparison between classifier-based counts (after correction for classification probabilities) and manual counts (from visual inspection of images) for selected categories chosen to span the range of relative abundance and classification uncertainties evident across all 22 categories and across the range of natural sample conditions encounterd during the period of the Feb-Apr 2005 time series shown below in Fig. 6. All points are shown with error bars indicating +/- 1 standard error for the classifier estimates. Red points indicate samples used in the generation of the classification probability matrix, while blue points are completely independent samples, each selected randomly from within week-long intervals of the full time series. All examples show no overall bias between manual and classifier results. Standard errors for some taxa as small as the plot symbols, while in a few cases they are a substantial fraction of the estimated value. This latter circumstance arises when there are very low counts present in samples (e.g.,* Ditylum*) or when individual images are difficult to classify with high accuracy (e.g.,* Skeletonema*).

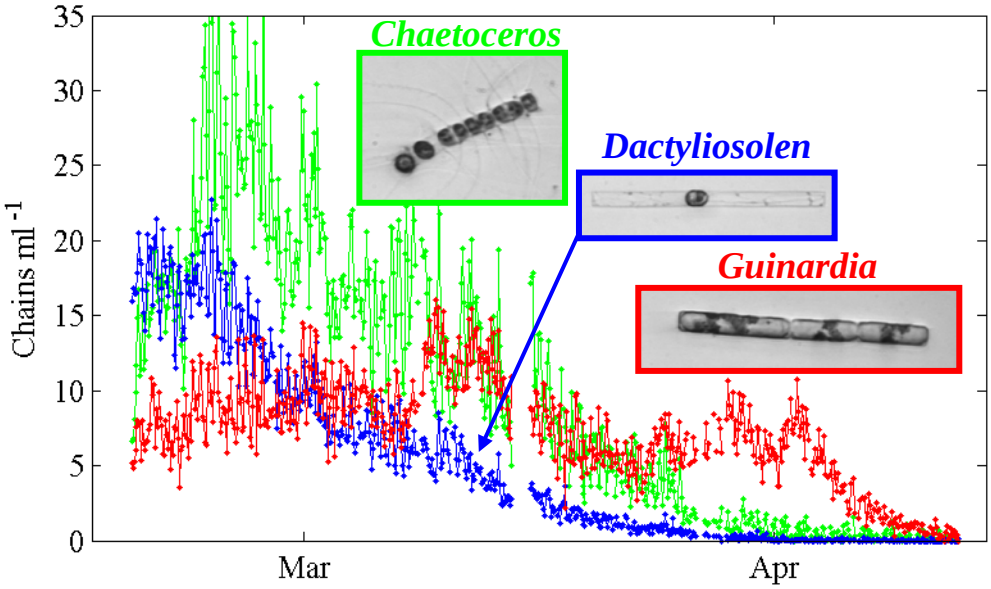

Figure 6. Time series of diatom chain abundances observed during a two-month (mid-February to mid-April) deployment of Imaging FlowCytobot in Woods Hole Harbor during 2005. Automated image classification was used to separate contributions of morphologically distinct genera; shown here are* Chaetoceros *spp.,* Dactyliosolen *spp., and* Guinardia *spp. While all three taxa declined to nearly undetectable levels by mid-April, they exhibited different temporal patterns, with* Dactyliosolen *disappearing earliest and* Guinardia *persisting longest and then declining most abruptly (in the first two weeks of April).

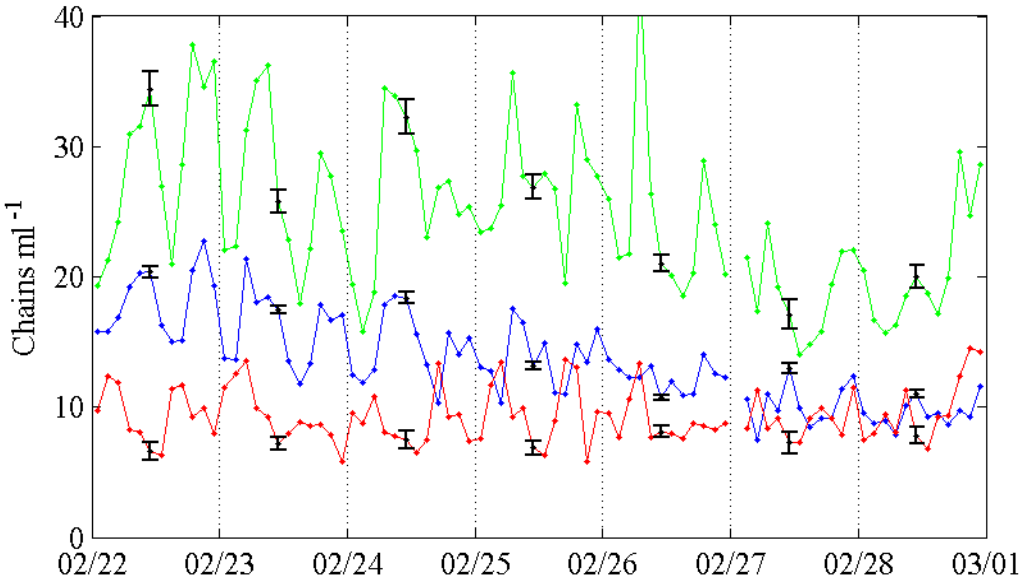

Figure 7. Expanded view of a single week (end of February) of the time series in Fig. 6, emphasizing the ability of Imaging FlowCytobot observations to capture variability related to water mass changes with semi-diurnal tidal flows in Woods Hole Harbor. Relative standard errors for the abundance estimates (shown only once per day for clarity) are small compared to the variations

evident at tidal frequencies. As in Fig. 6, green, blue, and red lines correspond to abundances of* Chaetoceros *spp.,* Dactyliosolen *spp., and* Guinardia *spp., respectively.