Research Statement

Omar De la Cruz C.

My research consists of developing and applying sophisticated statistical techniques in the biological sciences, especially in genetics.

One of the main themes of my work is the search for and exploitation of geometric structure in biological data sets. This line of inquiry is very promising, due to the confluence of the following trends:

- The high-dimensional nature of the data makes it difficult to apply traditional statistical methods, unless the data are found to lie in or near a low-dimensional structure (which might not be a linear subspace).
- The redundancy provided by the large number of measurements per unit can often be used to learn the geometric structure underpinning the data. Sometimes this structure is what we want to know (as in the cell cycle problem described below); other times, we want to adjust for its influence (as in the case of population stratification in genome-wide association studies, see below).
- Theoretical methods for learning geometric and topological structures from data are being very actively developed in the Mathematics, Computer Science, and Statistics communities.

My work on the cell cycle problem led in a natural way to the use of geometric methods, and to theoretical work on the methods themselves.

Contributions to the theory of geometric data analysis and kernel methods: One geometric approach is based on the study of point clouds (considering each unit as a point, with coordinates given by all the measurements performed on the unit). The assumption is that these points are samples from an unknown geometric object (e.g., an embedded manifold) plus a random noise component. Assuming the noise is full-dimensional, I introduced the notion of *modal ridge*: the generalized ridges of the density function induced by the model. These ridges provide an approximation of the original manifold, which is biased, but useful if the noise component is not too large.

Kernel methods are another approach to analyzing data with non-linear underlying geometric structure. By considering a sampling model for the units, kernels can be identified with some operators on Hilbert spaces; we study these operators and their adjoints (in a fashion similar to the "duality diagram") to show how most kernel methods are closely related to each other, to locally linear methods, and to graph Laplacian methods: all of them implicitly define a notion of smoothness (reminiscent of a Sobolev norm).

Also, the sampling model allowed us to analyze in detail ideal cases. In particular, we can describe the kernel eigenvectors when the underlying dimension is 2 or higher using a tensor product of one-dimensional kernel models, whose eigendecomposition is well understood.

Gene expression and the cell cycle: Recent advances make it possible to measure the expression levels of most genes from a single cell. If we have this information for enough cells harvested at random from a growing population (say, an embryo, a culture, or a tumor) then each cell corresponds to a point in observation space, and these points cluster around a closed curve

—a loop— which traces the idealized course of the cell cycle. Estimating that curve and the corresponding synthetic time-stamps for the cells allow us to establish or confirm the annotation of cycle-regulated genes (in species other than yeast), with a quantifiable level of confidence, as well as adjusting for the influence of the cell cycle in single-cell-per-microarray studies (making possible, for example, a more accurate detection of cell subpopulations within a tumor, or the measurement of treatment effects at the single cell level).

I studied single-cell microarray data from mouse retinal progenitor cells and was able to confirm annotations for many cycle regulated genes; the results also suggested intriguing hypothesis for further study, like the possible role of intermediate filaments in interkinetic nuclear migration during the development of the retina.

Population structure and association studies: Genetic association studies are hampered by the presence of population structure, since detected associations between particular alleles and disease status can be the result of demographic effects (if the disease is more prevalent in a subpopulation) rather than causality. Several methods are currently used to infer population structure from the collected genomic information itself; one of those methods involves using principal components analysis (PCA) to find a representation of the population in a low-dimensional linear subspace. However, there is no reason a priori to expect that this leads to the best description of the population.

We proposed an iterative method with a Markov kernel and the corresponding kernel PCA; the kernel is interpretable as part of a more general version of the Hardy–Weinberg equilibrium.

Region-specific p-values for genome-wide association studies: In the setting of genome-wide association studies, we proposed a method for assigning a measure of significance to pre-defined functional regions of the genome. This way, evidence for association between a particular functional unit and a disease status can be obtained not just by the presence of a strong signal from a SNP within it, but also by the combination of several simultaneous weaker signals.

This approach has at least three advantages: First, moderately strong signals from different SNPs are combined to obtain a much stronger signal for the region, therefore increasing power. Second, in combination with methods that provide information on untyped markers, it leads to results that can be readily combined across studies and platforms that use different SNPs. Third, the results are easy to interpret, since they refer to functional regions that are likely to behave as a unit in their phenotypic effect.

Applied analysis of gene expression data: Besides the development and application of new methods, I am interested in the use of existing methods for applied data analysis, especially when the methods need to be modified to fit an unusual situation.

The raw data from high-throughput biological assays need sophisticated pre-processing (often in a case by case basis) before meaningful analyses can be done.

I performed the microarray normalization and data analysis for two studies characterizing the putative olfactory receptor (OR) genes that are actually active in the olfactory epithelium: one in humans, and the other in chimpanzees. The difficulty was that a large proportion of the genes in the custom array were expected to have higher expression levels in olfactory epithelium, and this made it difficult to normalize the different arrays. The chimp study provided a signature of evolutionary constraint on a subset of OR genes that are expressed ectopically.