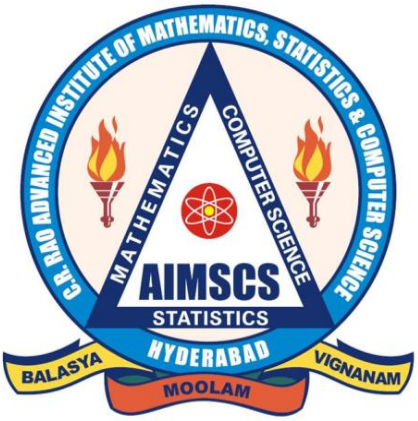
CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report



Author (s): **Saumyadipta Pyne, Sharon X. Lee, Geoffrey J. McLachlan**

Title of the Report: **Nature and Man: The Goal of Bio-security in the Course of Rapid and Inevitable Human Development**

Research Report No.: **RR2015-02**

Date: **April 17, 2015**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Nature and Man: The Goal of Bio-security in the Course of Rapid and Inevitable Human Development

Saumyadipta Pyne[1,2], Sharon X. Lee[3], Geoffrey J. McLachlan[3]

[1]*CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India*

[2]*Public Health Foundation of India*

[3]*Department of Mathematics, University of Queensland, Australia*

SUMMARY

The current course of human development, along with its driving forces such as globalization and urbanization, appears to be rapid and inevitable. Hence ensuring bio-security against emerging and re-emerging diseases is among the top challenges in this day and age. The present paper will discuss some of the projects that are being conducted internationally towards realizing the aim of ensuring bio-security through a variety of research approaches, many of which were developed in the past decade. In particular, we focus on the research methodology developed by the authors to enable parametric modeling of immunologic profiles of individual subjects, which can be further extended to large groups and populations.

Keywords: flow cytometry, mixture models, clustering of cells, skew mixture models, EM algorithm

1.INTRODUCTION

The year 2014 will be long remembered for its widespread Ebola virus disease outbreak that took place across multiple countries in West Africa. It was, in fact, the deadliest outbreak till date, claiming more than 8,000 lives between December 2013 and January 2015, the burden of deaths being the highest in Liberia, Sierra Leone and Guinea. (The corresponding figure exceeded 10,500 deaths by April 2015.) Among the suspects of origin and transmission of the Ebola virus (formerly known as Zaire virus) were fruit bat species, which may have been a natural host serving as the reservoir of the virus in the wild. While the bat-to-human transmission could take place along multiple routes, it has, nonetheless, rightly brought back to focus the critical issue of growing human activity and deforestation in previously untouched forests. In other words, the outbreak points to a much larger and deeper problem that lies at the crossroads of socio-economics, environment and human health.

Unfortunately, the global trend in emerging (and re-emerging) infectious diseases has seen almost steadily increasing incidence since 1940, even if controlled for effects of higher reporting. Clearly not much of this phenomenon is either novel or entirely unexpected, if viewed in the light of anthropogenic factors. Not long ago, in 2002, the SARS outbreak in China was also traced to hunting and trading of bats that are natural reservoirs of SARS-like coronaviruses. It has been found that about 60% of known human infectious diseases, and more than 75% of newly emerging infections have zoonotic origin, i.e. they originate in vertebrate animals before being transmitted to humans. Almost 80% of the viruses and 50% of the bacteria that infect humans are of zoonotic

This paper is based on the 36th Dr. VG Panse Memorial Lecture by SP delivered at the 68th Annual Conference on Statistics and Informatics in Agricultural Research organised by the Indian Society of Agricultural Statistics at ICAR-IASRI, New Delhi, on January 29, 2015.

Corresponding author: Saumyadipta Pyne

E-mail address: spyne@crraoaimscs.res.in

origin. In fact, more than 70% of the known zoonotic pathogens have originated from wildlife species. A conservative estimate puts the number of viruses present in vertebrate species at 1 million, which suggests that more than 99.9% of the viruses that lie out there are currently unknown to humans. As the frequency of emergence of new pathogens continues to rise, so does the importance of societies need to understand and be prepared to systematically address the challenge of emerging infectious diseases at different levels.

To address such a complex issue of global significance, a meeting of experts titled, "Unhealthy Landscapes: How Land Use Change Affects Health", was convened at the 2002 biennial meeting of the International Society for Ecosystem Health (6–11 June 2002, Washington DC, USA). It was noted that "human-induced land use changes are the primary drivers of a range of infectious disease outbreaks and emergence events and also modifiers of the transmission of endemic infections" (Patz et al., 2004) Often such land use changes could be rapid, and may include deforestation, biodiversity loss, habitat encroachment, urbanization, and other activities. "These changes," noted Patz et al., "in turn cause a cascade of factors that exacerbate infectious disease emergence, such as forest fragmentation, pathogen introduction, pollution, poverty, and human migration. These are important and complex issues that are understood only for a few diseases." The larger overall scenario, needless to say, is much more challenging and less well understood.

What could be done towards the objective of ensuring bio-security both at the human-animal-environment interfaces, and overall, in the course of rapid and inevitable human development? This is a "big" question that is currently being asked in many major forums across the world. Recently, the concept of "One Health" has received much attention led by the tripartite initiative of WHO, FAO of the UN, and the World Organization for Animal Health. Importantly, the need for a concerted, inter-disciplinary approach has been felt in public health research, bringing together teams of epidemiologists, clinicians, veterinarians, microbiologists, wildlife biologists, ecologists, statisticians and public health officials. Apparently, a global strategy seems to be emerging for monitoring and bio-surveillance, partly driven also by concerns of bioterrorism. However, the coverage or efficiency of public health surveillance systems world-wide is still far from uniform.

The question is how do we move systematically towards the aim of ensuring bio-security? Here we will discuss some projects that are being conducted internationally by different laboratories and organizations using a variety of research approaches, many of which were developed in the past decade. The aim here is not to provide a comprehensive review of either the approaches or the projects undertaken, but to get a general understanding of the flavor of certain scientific platforms and analytical methods that could be brought to bear in demonstrating this global effort. Many public health researchers and policy-experts are currently involved in designing programs and strategies to be prepared for emerging (and re-emerging) diseases. Systematic identification of new hotspots of epidemics, modeling and mapping of ecological niches, characterization of human-animal interfaces by exposure types and frequency, creation of host-pathogen databases, identification of key taxonomic groups, and phylodynamic analysis of host, epidemiological and molecular data – all lead to a pool of valuable information that could be useful for prediction and prevention of future zoonotic outbreaks. As an example of this approach, the PREDICT project of the Emerging Pandemic Threats program launched by United States Agency for International Development (USAID) in 2009 aims to facilitate predictive modeling for the identification of the most likely regions, hosts and human-animal interfaces for forthcoming emergence of zoonoses.

A project such as PREDICT aims to collect timely and reliable data based on internet surveillance of reports of unusual health events occurring in countries with hotspots. Further, it conducts analysis to test if the pathogen is likely to emerge and spread in the social systems that exist in those hotspots. PREDICT combines risk modeling with targeted wildlife field sampling for selected locations, interfaces and host taxa. In this effort, it is aided by inter-disciplinary experts, computerized data collection and analysis, and

active partnership with local and national governments. The program currently collaborates with 20 African, Asian and South American countries, and in just a few years, it has detected hundreds of novel viruses in the hundreds of thousands of samples collected from tens of thousands of animals from these locations (Morse et al., 2012).

While programs such as PREDICT can demonstrate the benefits of local capacity-building efforts as part of international efforts to counter zoonotic threats, it also underscores the need for statistical and computational capability to work with massive datasets of high volume, velocity, variety and veracity — the so-called BIG-DATA characteristics. To create integrative models for forecasting, the researchers need to consider multi-sectorial data on a large number of parameters. These include socio-economic parameters such as population density, mixing patterns, migration, trade, agricultural practices, sanitation, age groups, diet, vaccination history, drug and antibiotic use, cultural norms, occupational exposures, nutritional and immunological status, etc. Additional information about the interface includes wildlife diversity, human-wildlife contact frequency, similarity of host species, similarity of microbial species present in host, ease of evolvability of the pathogens, host-pathogen co-evolution, and so on. Sophisticated spatio-temporal models of ecological niches are being developed in different parts of the world to facilitate timely and data-driven administrative policy and decision-making (Eubank et al., 2004).

In addition to yielding practical benefits on the ground, advanced platforms for novel pathogen discovery are revolutionizing diverse areas of biotechnology, medicine and agriculture. The real game-changer over the past decade has been high-throughput sequencing (HTS) technologies. These generate massive datasets not only based on DNA and RNA sequences of a specific organism, but also of collective microbial communities that are studied for metagenomics analysis. As less than 1% of all microbes could be cultivated inside labs, culture-independent metagenomics is increasingly the technology of choice in characterizing new pathogens. Further, the human microbiome data are revealing new insights on host-pathogen interactions, as well as effects of diet, antibiotics and environment. Another emerging field, of viral metagenomics, is not only becoming popular, it can now be performed at the astonishing granularity of single virus level. It is difficult nowadays to imagine a sophisticated pathogen discovery program without the support of an efficient and reliable HTS based genomic and metagenomic pipeline. Not surprisingly, it is not so much the constraints of data acquisition that are of concern to HTS labs across the globe today as the challenge of efficient bioinformatics and biostatistical data analysis (Firth and Lipkin 2013).

A systematic understanding gained through sustained and continuous investigation of viral genetic diversity is likely to become a key aspect of advanced biosurveillance efforts in global preparedness for disease outbreaks. In particular, the rapid evolvability of RNA viruses, given their high mutation rates (and missing error-correcting mechanisms), allow such evolution to take place in time scales of human observation. The field of phylodynamics, therefore, seeks to combine data from phylogenetics and epidemiological dynamics. This combined approach can allow more effective bio-surveillance, and prediction of the epidemiological impact, of recently emerged or evolved viruses. Towards this, simultaneous collection of data of different types, such as (a) spatio-temporal epidemic dynamics, (b) viral genome sequences, (c) contact networks of susceptible host individuals, and (d) the immune history of the individuals in contact networks, are modeled together "for understanding both the dynamics of epidemic spread and the evolutionary pressures that shape virus diversity" (Holmes and Grenfell, 2009).

In the following sections, we will discuss about statistical and computational methods that were developed by the authors and collaborators over the past decade to model and analyze detailed immunologic profiles of subjects — both in clinical settings or otherwise. Objective and automated characterization of population immunophenotypes based on high-throughput platforms such as clinical flow cytometry and rigorous data modeling can help in addressing a variety of issues regarding the computational bio-security (see Table 1). Systematic immunologic profiling data

generated by planned programs (e.g., The Human Immunology Project Consortium established in 2010 by NIAID of NIH, USA) can immensely benefit from our analytical platforms.

Flow cytometry uses a panel of p flurophore-conjugated antibodies to measure the expression of p specific markers for each individual cell in a given sample, such as a subject's whole blood. Cells (i.e., p-dimensional points) with similar expressions form clusters (or "cell populations"), which together define the immuno-phenotypic profile of a subject from whom the sample was obtained. Thus, a mixture of cell populations can be modeled by a mixture of p-variate probability distributions. Under different conditions (such as during an infection), the expressions of the cell populations may be different, which can then be detected via the altered values of model parameters. In particular, by introducing the use of multivariate skewed and heavy-tailed probability distributions, and their finite mixture models, to characterize the multivariate expressions of different cell populations, the speaker and his coworkers were able to model immunologic profiles with precision and rigor (Pyne et al., 2009; Frühwirth-Schnatter and Pyne, 2010; Ray and Pyne, 2012; Azad et al., 2012; Rossin et al., 2011; Ho, Pyne and Lin, 2012; Ho, Lin, Chang, Haase, Huang and Pyne, 2012).

New algorithms for fitting finite mixture models of multivariate skew distributions were developed for parametric modeling of high-dimensional immuno-phenotypic data obtained from human subjects, as demonstrated in the next section. The skewness parameter can capture interesting phenomena that may occur in the tail populations such as immunologic state transitions in profiles or altered cell signaling during an infection. By joint clustering and matching (JCM) of cell populations with our robust multi-level JCM models, cell populations (i.e., clusters) were matched across subjects allowing the models parameters to be compared across different classes, time points, etc. Further, the JCM model parameters can shed light on the diversity of immunologic states and profiles, not only of individual subjects, but also of large and heterogeneous cohorts, by accounting for the challenging issue of subject-specific variation (Pyne et al., 2014). Each JCM modeled sample's every parameter – including specific cell population size, shape, location and variation in marker-space – is output for downstream analysis. For instance, the same could be used to classify human sub-populations of interest, such as those with potential vulnerability or resistance for specific diseases.

2.METHODS

2.1 The JCM Methodology

The JCM methodology provides a powerful framework for modeling a cohort of (cytometric) samples, where the characteristics of the entire cohort or class of samples can be described by a flexible parametric template. It also enables simultaneous clustering and matching of cell populations across samples, with the ability to accommodate subtle inter-sample variations. JCM adopts a two-level hierarchical approach where (at the lower level) each sample is modeled by a finite mixture model with flexible component distributions, and (at the higher level) these components are linked to an overall template through a random-effects model (REM) that accounts for inter-sample variation. Under this setting, each sample can be conceptualized as an instance of the template, possibly transformed with a flexible amount of variation. A brief description of the JCM model is given below.

2.2 Finite Mixture modeling

At the lower level, JCM adopts a finite mixture model to characterize a sample. A cell population within a sample is modeled by a component distribution of the mixture model. More specifically, let $\boldsymbol{y}_j$ be the vector containing the measurements of the p markers on the jth cell of a sample, where $j = 1, \ldots, n$. Here n denotes the total number of cells in the sample under consideration. Suppose there are g populations in this sample, then the distribution of $\boldsymbol{y}_j$ can be modeled by a g-component mixture model, given by

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \, f(\boldsymbol{y}_j; \boldsymbol{\theta}_h), \qquad (1)$$

where the mixing proportions π_h are non-negative and sum to one. In the above, $f(\cdot; \boldsymbol{\theta}_h)$ denotes a

- How to mathematically characterize a "normal"subject's immuno-phenotypic profile?
- How to model a diverse range of population immuno-phenotypes?
- How to group the population immuno-phenotypes into meaningful classes?
- How to map the immuno-phenotypic classes over geographical space and time?
- How to assign correct classification to new or rare immuno-phenotypes?
- How to rapidly detect any departure or outlier from a given range of normal profiles?
- How to represent population profiles in databases to enable mining and fast queries?
- How to find associations between population genotypes and immuno-phenotypes?
- How to identify a vulnerable or resistant sub-population based on such associations?
- How to systematically estimate the risks and parameters of potential outbreaks in a given population?

Table 1: Issues to be addressed by statistical modeling of population immuno-phenotypes.

component density with parameters specified by $\boldsymbol{\theta}_h$ $(h = 1, \ldots, g)$, and $\boldsymbol{\Psi}$ is a vector consisting of all the unknown parameters of the mixture model.

On specifying a parametric form $f(\boldsymbol{y}_j; \boldsymbol{\theta}_h)$ for each component density the mixture model (1) can be fitted by maximum likelihood (ML) via the Expectation Maximization (EM) algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (2008).

The model (1) provides a convenient method for clustering the cells into g clusters. A probabilistic clustering can be obtained in terms of the fitted posterior probabilities of component membership. For outright clustering, the *maximum a posteriori* (AMP) rule can be applied, which assigns a cell to the component with the highest posterior probability.

2.3 Skewed Component Distributions

Observing that the clusters of cells are typically asymmetrically distributed as well as having heavy tails, JCM adopts multivariate skew distributions as component densities. In particular, a skew version of the t-distribution known as the skew t (ST) distribution is employed, given by

$$\begin{aligned} f(\boldsymbol{y}_j; \boldsymbol{\theta}_h) &= 2\, t_p(\boldsymbol{y}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \nu_h) \\ &\quad T_1\left(\frac{q_{jh}}{\lambda_h}; 0, 1, \nu_h + p\right), \end{aligned} \tag{2}$$

where

$$\begin{aligned} q_{jh} &= \boldsymbol{\delta}_h^T \boldsymbol{\Sigma}_h^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_h) \sqrt{\frac{\nu_h + p}{\nu_h + d_h(\boldsymbol{y}_j)}}, \\ \lambda_h^2 &= 1 - \boldsymbol{\delta}_h^T \boldsymbol{\Sigma}_h^{-1} \boldsymbol{\delta}_h, \\ d_h(\boldsymbol{y}_j) &= (\boldsymbol{y}_j - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_h), \end{aligned}$$

$t_p(\boldsymbol{y}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \nu_h)$ denotes the p-variate t-density with location vector $\boldsymbol{\mu}_h$, scale matrix $\boldsymbol{\Sigma}_h$, and degrees of freedom ν_h, and $T_p(\cdot)$ denotes its corresponding distribution function. With the extra parameter of skewness $\boldsymbol{\delta}_h$, the ST distribution can flexibly handle non-symmetric distributional shapes. Parameter estimation for the mixture model (1) with (2) as component densities can be carried out using the EM algorithm as described in Pyne et al. (2009) and Wang et al. (2009). Further discussions of skew t-mixture models and parameter estimation can be found in Lee and McLachlan (2014) and the references therein.

2.4 Multi-Level modeling

At the upper level, JCM links the individual mixture models for each sample obtained from the lower level through a random effects model (REM). Briefly, the random effects term in the REM specify how the sample-specific component densities vary from an overall representative template of the cohort. More specifically, the REM in JCM governs the relationship between the loca-

tion vector $\boldsymbol{\mu}_{hk}$ of sample k and the batch location vector $\boldsymbol{\mu}_h$, as specified by an affine transformation

$$\boldsymbol{\mu}_{hk} = \boldsymbol{a}_{hk} \circ \boldsymbol{\mu}_h + b_{hk}, \tag{3}$$

where $\circ$ denotes the Hadamard product, and $\boldsymbol{a}_{hk}$ and b_{hk} are RE terms for scaling and translation, respectively. They are given by

$$\begin{aligned} \boldsymbol{a}_{hk} &\sim N_p(\mathbf{1}_p, \boldsymbol{A}_h), \\ b_{hk} &\sim N_1(0, B_h), \end{aligned} \tag{4}$$

respectively. In effect, the sample specific component location parameter $\boldsymbol{\mu}_{hk}$ can be viewed as a realization of $\boldsymbol{\mu}_h$, that is, $\boldsymbol{\mu}_{hk}$ has a normal distribution with mean $\boldsymbol{\mu}_h$ and variance that depends on $\boldsymbol{A}_h$ and B_h. The fitting of the JCM model (3) with (4), (3), and (1) can be implemented using the EM algorithm. Further details can be found in Pyne et al. (2014) and Lee et al. (2014).

On fitting JCM to a batch of sample, we obtain an individual model for each sample and a template model summarizing the overall characteristics of the batch. These fitted models are given in the form of a skew t-mixture distribution with parameters capturing useful features such as the location of each cluster, the proportion of cells in each cluster, the skewness of their distribution, and the degree of long/heavy-tailedness. This approach enables direct comparison across different batches or classes of samples, and facilitates objective classification of unlabelled samples. As the models are defined parametrically, a range of information-based measures can be applied to quantify their similarities or differences, for both within-class and inter-class variations. In the illustration to follow, we shall demonstrate this with the Kullback-Leibler (KL) distance (Kullback and Leibler, 1951) and Bhattacharyya distance (Bhattacharyya, 1943).

3.RESULTS

3.1 Immune Tolerance Network (ITN) dataset

A subset of 15 samples from the Immune tolerance Network (ITN) were analyzed. These are based on blood samples acquired from three different groups patients, five in each group. Each sample is analyzed using cytometry to measure the expression of five markers, namely CD3, CD4, CD8, CD69, and HLA-DR. The ITN dataset is available publicly from the BioConductor package flowStats. Together they can be used to determine the types and functions of different subpopulations of T cells such as Natural Killer (NK) cells, T-helper cells, etc. Markers such as HLA-DR can inform about T cell activation and signaling states.

3.2 JCM Template for ITN Classes

A JCM model is fitted to each of the three classes of samples. Briefly, the fluorescence channels were transformed before JCM was applied, and the FSC and SSC channels were not included in the analysis. JCM identified 5 components (i.e. common cell populations) in each of these classes. On comparing the templates fitted to these classes, a marked difference can be observed for one of the clusters. In particular, there is considerable difference in the distribution of the $CD4^+CD8^+$ populations (Figure 1), where all three classes are clearly different. While the overall variation between Class 2 and Class 3 is much less profound, they differ in the $CD4^-CD8^+$ populations. Notably, there is no observable differences across the three classes in terms of the $CD4^+CD8^-$ and $CD4^-CD8^-$ populations.

The within-class variation can be visually compared with overlay plots produced by JCM (Figure 2). As can be observed, the templates are providing a good representation of the batch, summarizing the main features of the samples in their respective classes. Again, considerable differences can be observed from the samples across different classes.

To gain a better appreciation of the differences among the three classes, Figure 3 shows the 3D contours of the individual components of the templates for the markers CD4, CD8, CD69, and HLA-DR. It can be observed that most of the components in Class 1 are distributed differently from Classes 2 and 3. This is also supported by quantitative comparisons of these templates. The KL and Bhattacharyya distances between each pair of templates are given in Table 2. With a KL distance of almost 4 for Class 1 versus Class

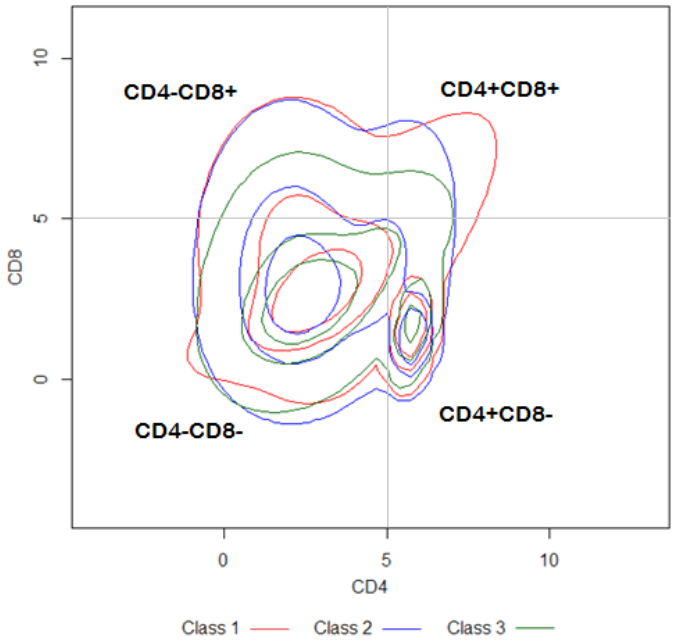


Figure 1: The ITN dataset (three classes each with five samples) was analysed with JCM. The parametric class templates from the 3 classes are overlaid and visualized along two dimensions: CD4 and CD8. The variation in the $CD4^{hi}CD8^{hi}$ subsets among the 3 classes can be observed (in the $CD4^{+}CD8^{+}$ quadrant).

2, and for Class 2 versus Class 3, compared to 0.20 for Class 2 versus Class 3, it indicates that Class 1 is significantly different to the other two classes. The Bhattacharyya distances also revealed similar results, where the differences between Class 1 and Classes 2 and 3 were relatively high compared to that between Class 2 and Class 3.

	Class 1	**Class 2**	**Class 3**
Class 1	-	3.99	3.90
Class 2	0.82	-	0.20
Class 3	0.78	0.04	-

Table 2: The between-class variation can be observed from the KL distances (upper right elements) and Bhattacharyya distances (lower left elements) between each pair of class templates. Notably, Class 1 is evidently different from Classes 2 and 3, whereas the latter two classes were quite similar to each other.

4.DISCUSSION

In the above sections, we demonstrated a methodology for mathematical characterization of a subject's individual and a cohort's overall immunophenotypic profile. The parametric mixture modeling makes it particularly effective to both represent the data about populations in terms of statistically well-defined parameters as well as to use the same for identifying outlier profiles. A robust understanding of population-level diversity in profiles and the corresponding baselines under so-called normal conditions holds the key for systematic outbreak detection.

By mapping the distribution of immunologic states in a diverse population, whether human or livestock, over geographic space and time, it might be possible to identify regions that are vulnerable for future outbreaks of emerging and re-emerging infectious diseases. Immunological, serological and virological surveillance can be applied to monitor hotspots of such diseases. New technological platforms such as highly multiparametric cytometry, tetramer assays, HTS, and many others are currently being used. A planned, integrative approach to computational bio-security will need to combine various technological and analytical expertise (Figure 4).

Fortunately, the recent Ebola outbreak did not become a global pandemic, as the cases were mostly localized within the less urbanized centres in the three worst affected countries (and thanks to the selfless efforts of many health workers). However, it served as a stark reminder of our vulnerability as we live in highly dynamic societies today – in which many forces are continuously and interactively at work – globalization, urbanization, migration, global trade and travel – and therefore, the importance of prepared-ness to handle any threat to bio-security cannot be overemphasized.

ACKNOWLEDGEMENTS

SP was supported by Ramalingaswami Fellowship of DBT, DRDE, and MoS&PI India. The work of SXL and GJM was supported by an Australian Research Council Discovery Grant.

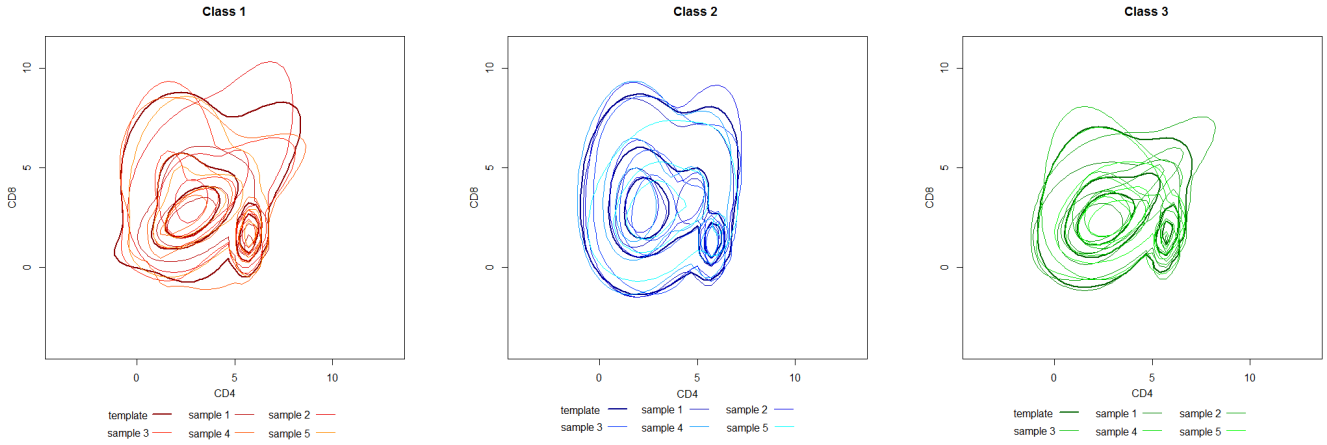


Figure 2: The within-class variation and summarization in the ITN dataset may be observed. For each of the 3 classes, the JCM-fit mixture model of each sample is shown along with the overall class template in bold.

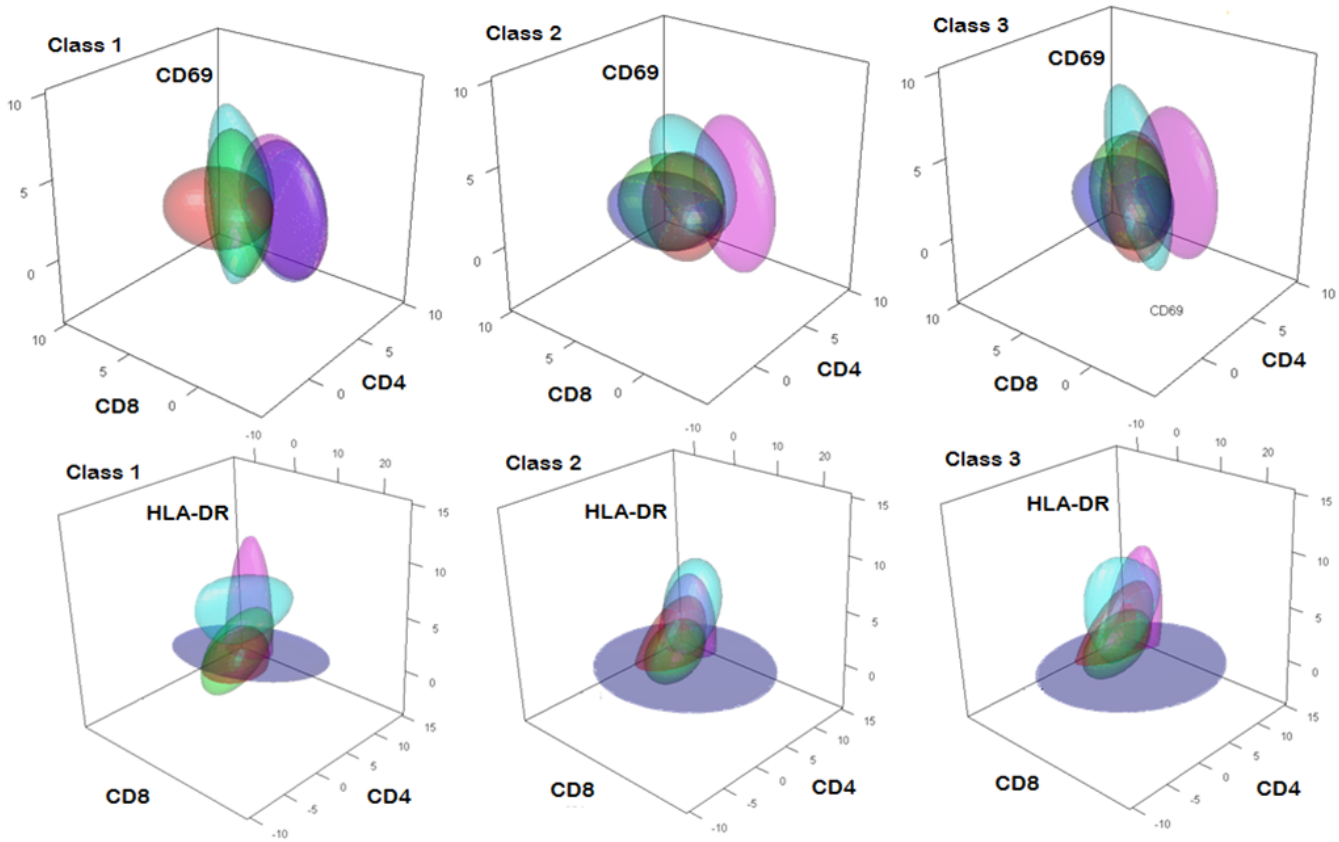


Figure 3: The JCM-fit class templates of the ITN dataset are shown in 3D contours of individual components, each representing a specific cell population or cluster as defined by expressions of antibody markers in 5-dimensional marker-space.

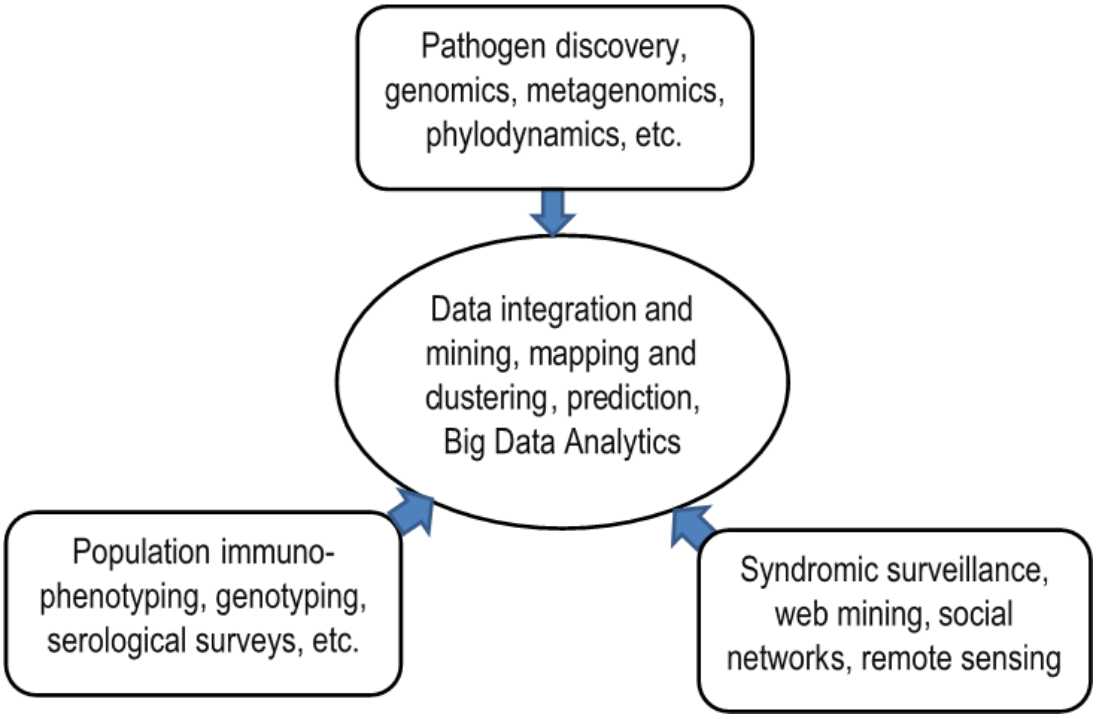


Figure 4: An integrative model is useful for computational bio-security.

REFERENCES

Azad, A., Pyne, S. and Pothen, A. (2012). Matching phosphorylation response patterns of antigen-receptor-stimulated t cells via flow cytometry. *BMC Bioinformatics* **13**, S10.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* **35**, 99–109.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society , Series B* **39**, 1–38.

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z. and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature* **86**, 180–184.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11**, 317–336.

Ho, H. J., Lin, T. I., Chang, H. H., Haase, H. B., Huang, S. and Pyne, S. (2012). Parametric modeling of cellular state transitions as measured with flow cytometry different tissues. *BMC Bioinformatics* **13**, (Suppl 5): S5.

Ho, H. J., Pyne, S. and Lin, T. I. (2012). Maximum likelihood inference for mixtures of skew student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing* **22**, 287–299.

Holmes, E. C. and Grenfell, B. T. (2009). Discovering the phylodynamics of rna viruses. *PLoS Computational Biology* **5**, 317–336.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing* **24**, 181–202.

Lee, S. X., McLachlan, G. J. and Pyne, S. (2014). Supervised classification of flow cytometric samples via the joint clustering and matching procedure. *arXiv:1411.2820 [q-bio.QM]* .

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. 2nd edn. Hokoben, N. J.. Wiley-Interscience.

Morse, S. S., Mazet, J. A. K., Woolhouse, M., Parrish, C., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I. and Daszak, P. (2012). Prediction and prevention of the next pandemic zoonosis. *Lancet* **380**, 1956–1965.

Patz, J. A., Daszak, P., Tabor, G. M., Aguirre, A. A., Pearl, M., Epstein, J., Wolfe, N. D., Kilpatrick, A. M., Foufopoulos, J., Molyneux, D., Bradley, D. J. and Working Group on Land Use Change and Disease Emergence (2004). Unhealthy landscapes: Policy recommendation on land use change and infectious disease emergence. *Environmental Health Perspectives* **112**, 1092–1098.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L. and Mesirow, J. P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences USA* **106**, 8519–8524.

Pyne, S., Lee, S. X., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, . T., Ng, S. K., Hafler, D., Levy, R., Nolan, G. P., Mesirov, J. and McLachlan, G. J. (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLOS ONE* **9**, e100334.

Ray, S. and Pyne, S. (2012). A computational framework to emulate the human perspective in flow cytometric data analysis. *PLOS ONE* **72**, 235693.

Rossin, E., Lin, T.-I., Ho, H. J., Mentzer, S. and Pyne, S. (2011). A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues. *Bioinformatics* **27**, 2746–2753.

Wang, K., Ng, S. K. and McLachlan, G. J. (2009). Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data In *DICTA 2009 (Conference of Digital Image Computing: Techniques and Applications, Melbourne)*. H. Shi, Y. Zhang, M. J. Bottema, B. C. Lovell and A. J. Maeder (Eds.). IEEE Computer Society. Los Alamitos, California. pp. 526–531.