Review Article

# Computational Immunology – From Bench to Virtual Reality

Cliburn Chan,[1]MBBS, PhD, Thomas B Kepler,[1]PhD

**Abstract**

**Drinking from a fire-hose is an old cliché for the experience of learning basic and clinical sciences in medical school, and the pipe has been growing fatter at an alarming rate. Of course, it does not stop when one graduates; if anything, both the researcher and clinician are flooded with even more information. Slightly embarrassingly, while modern science is very good at generating new information, our ability to weave multiple strands of data into a useful and coherent story lags quite far behind. Bioinformatics, systems biology and computational medicine have arisen in recent years to address just this challenge. This essay is an introduction to the problem of data synthesis and integration in biology and medicine, and how the relatively new art of biological simulation can provide a new kind of map for understanding physiology and pathology. The nascent field of computational immunology will be used for illustration, but similar trends are occurring broadly across all of biology and medicine.**

**Ann Acad Med Singapore 2007;36:123-7**

**Key words: Mathematical models, Medical informatics, Scientific visualisation, Simulation, Systems biology**

## Introduction

The pace at which data is being generated in both the basic and clinical sciences has been growing exponentially in recent years, driven by improvements in automation, robotics and high-throughput assays. In addition, we often have information about the same phenomenon from multiple different types of studies – in immunology for example, it is not uncommon to have histology, flow cytometry and functional assays for the investigation of a single immune response. More data are easily accessible online, either as electronic manuscripts on PubMed, or from an ever growing list of specialised online databases.

The ability to generate ever more detailed information about the organism is a spectacular achievement of reductionistic biology, and is justly celebrated. The Human Genome Project is providing a detailed parts list for the genome,[1] while the HapMap study is elucidating the diversity of the human genome ;[2] similar efforts are being conducted (or have been completed) for important model organisms. High-throughput techniques like microarray (or gene chip) analysis and proteomics are doing the same for mRNA and proteins, respectively. Systematic mouse gene knockouts are providing insight into the functional importance of each gene.[3] In the field of immunology, technical advances in instrumentation have allowed us to track the expression of multiple cell surface and intracellular molecules simultaneously both in vitro (polychromatic flow cytometry) and in vivo (fluorescent in situ hybridisation), the secretion of multiple cytokines and chemokines (Luminex xMAP® assays), and even the movement and interactions of immune cells in vivo (2 photon video-microscopy).

This runaway success of reductionistic science has given rise to a new problem – how do we make optimal use of the data available?[4] The most obvious challenge is simply how to extract useful information from a high-throughput study. At another level, the challenge is to integrate information from *different* experimental approaches to characterise the same biological phenomenon. An even more ambitious programme would be to unify a phenomenon at multiple scales of description. This review explores how statistics, mathematics, physics and computer science are helping us to tell a coherent story about how the immune response works; our attempt to put the Humpty Dumpty of immunology together again (Fig. 1).

[1] Center for Computational Immunology, Department of Biostatistics and Bioinformatics, Department of Immunology, Institute of Statistics and Decision Sciences, Duke University, Durham, USA

Address for Correspondence: Dr Cliburn Chan, Duke University, Department of Biostatistics and Bioinformatics, Box 2734 DUMC, 2424 Erwin Road, Hock Plaza G033, Durham NC 27705, USA.
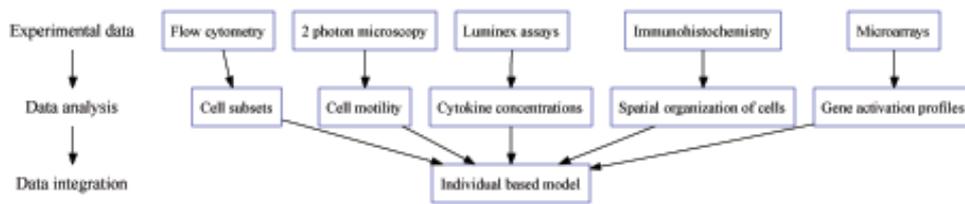Email: cliburn.chan@duke.edu

Fig. 1. A model integrates information from diverse experimental sources, and is the basis for a synthetic, reconstructionist approach to understanding the immune response.

## Analysing High-throughput Data

In experimental immunology, one of the earliest and certainly the most ubiquitous high-throughput methods is flow cytometry. Flow cytometry is widely used in clinical settings as well. For example, it is used by haematologists and pathologists to diagnose and subtype leukaemias. Since most clinicians are familiar with flow cytometry, it will be used as an example to illustrate the process of analysing high-throughput data.

In flow cytometry, cells tagged with one or more fluorescent dyes stream down a capillary tube in single file, and light from a laser is used to activate the tagged fluorochromes. Since each fluorochrome emits light of a specific colour when activated, the density of tagged molecules on each cell can be estimated. Typically, tens or hundreds of thousands of cells are streamed in a single session, giving a distribution of emitted colour intensities representing the distribution of tagged molecules in that cell population. Flow cytometers also report the amount of laser light that passes through each cell as the *forward scatter* (which gives an indication of the size of the cells) and the amount of laser light scattered transversely by each cell as the *side scatter (*which gives an indication of the internal complexity of the cell).[5]

Not surprisingly, flow cytometry has gradually been increasing in complexity. Originally, flow cytometry experiments would tag 1 or 2 cell surface molecules, allowing visual analysis using a histogram (1D) or dot plot (2D) to identify different subsets by demarcating discrete regions known as *gates*. Recently, the number of fluorescent markers that can be used to measure different cell parameters simultaneously has been rising rapidly, to approximately 20 colours at present.[6] In practice, the data is typically analysed by gating interesting regions sequentially using dot plots, a highly tedious process which can take several hours per data set.

Quite apart from the tedium, there are at least 2 problems with manual analysis. First, the number of possible different sequences of dot plots with $k$ colours grows rapidly with $k$. A simple combinatorial calculation shows that there are

$$\begin{bmatrix} k \\ 2 \end{bmatrix} \times 2^{k-2} \times (k-2)! \quad \text{possible ways to do sequential}$$

bivariate dot plots when there are $k$ colours. This implies that any possible gating sequence must be highly arbitrary since it is impossible to check all possible sequences for the "optimal" one. Equally troubling, the demarcation of gating regions is done by eye, and thus highly operator-dependent. Different operators using cell samples from the same experiment can easily get different subset population densities simply because the gating regions chosen are slightly different at each stage, even if they follow the same marker sequence at each stage.

Surprisingly, relatively little work appears to have been done on automating the extraction and summary of cell subsets from multiple colour flow cytometry data.[7,8] Our approach is to consider the data as being generated from some unknown multi-dimensional probability distribution, and using a non-parametric statistical method to estimate the density at each point in $n$-dimensional space. Interesting features are then sought in the estimated density – the simplest is probably the *mode* corresponding to a local maximum in density. Cell subsets are identified by being in the basin of attraction of a particular mode, and are separated by density valleys between different modes. One way to visualise the algorithm is to consider each point simply climbing up its slope in $n$-dimensional space to the top of the local hill (the mode). Since this operates directly in $n$-dimensional space, it does not face the problem of combinatorial explosion in possibilities, unlike the sequential dot plot strategy. Also, since the basins of attraction of a mode define the cell subset, there is no requirement for manual gating with its inherent inaccuracies.

Another complementary approach we are taking is to develop interactive 3D graphical visualisations of flow cytometry data. By using animations to illustrate dynamics, for example, tracking the paths of test points as they hill climb from selected or random positions simultaneously in several 3D graphics windows, we hope that better intuition for the structure of point clusters in multi-dimensional flow data will emerge. The basic idea is that most, if not all, interesting features (for example, corresponding to cell subsets) are likely to be lower dimensional structures embedded in a high dimensional space. If so, then an interactive 3D graphical application may provide clues as
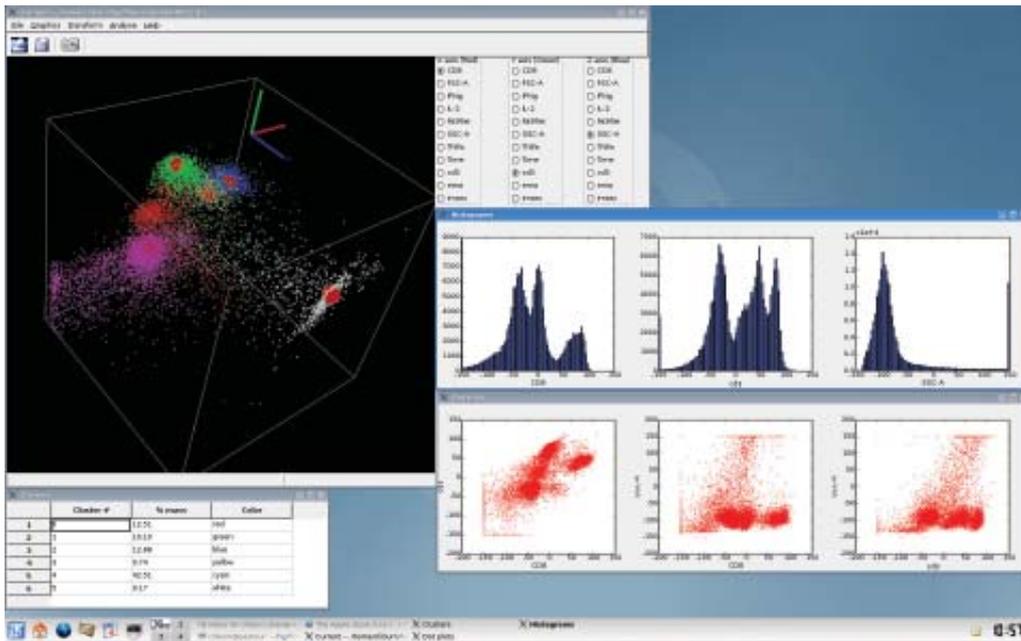
Fig. 2. Screen shot of software prototype for automatic identification and extraction of cell subsets from multi-colour flow cytometry data.

to how best to project the data onto a *meaningful* lower dimensional space, which would be much simpler to analyse. A screen shot of a tool to visualise and extract such cell subsets automatically that we developed is shown in Figure 2.

### Relating Data to Mechanism with Mathematical Models

While the above approach to analysing flow data may allow the automated identification of cell subsets, it provides little insight into what governs the shape of the underlying probability distribution, and how the distribution changes under different experimental or clinical conditions. Generally, when we wish to understand the underlying mechanisms, we construct a *mathematical* model in which we simplify biological reality (often drastically!) by removing details, which we do not believe contribute to the phenomenon we are trying to explain. The model is then solved, or if it proves to be intractable, numerically simulated and parameters systematically varied to see their effects. Model construction and experimentation is ideally an interactive process – models should suggest new experiments, and experiments should suggest model improvements.

A simple example related to flow cytometry is to explain why the distributions of cell surface molecules appears to be highly skewed (non-normal), typically requiring log-transformation before gating. One of the simplest possible models for the density of a particular molecule x on a cell surface is that it is synthesised at a constant rate, and

removed from the cell surface linearly. This can be modelled as the following *ordinary differential equation*

$$\frac{dx}{dt} = s - rx$$

where $x$ is the density of molecule $x$ on the cell surface, $s$ is the rate of synthesis and $r$ is the rate of removal from the cell surface. At equilibrium, the rate of change of $x$ is 0, and so the equilibrium density of $x$ is simply $s/r$. In a sample of cells, we would expect the values of $s$ and $r$ to vary between cells – a simple assumption is that both $s$ and $r$ are normally distributed. Then the distribution of equilibrium densities of the molecule is a random variable given by the ratio of 2 normally distributed random variables, which has a right skew as is typically observed. While the model is almost trivial, the fact that it predicts a skewed distribution suggests that we should not be too surprised by the observed skew in real data.

More interestingly, suppose we had data from separate experiments that tracked how the rate of synthesis of molecule $x$ varies when some cytokine $c$ is added to the culture medium. This information can be incorporated into the model by making $s$ a function of $c$, giving

$$\frac{dx}{dt} = s(c) - rx$$

For simplicity, suppose the relation is linear with $s = kc$, where $k$ is some constant. Now, if we had control and test cell samples cultured without and with cytokine $c$ run through a flow cytometer, we could use the differences in
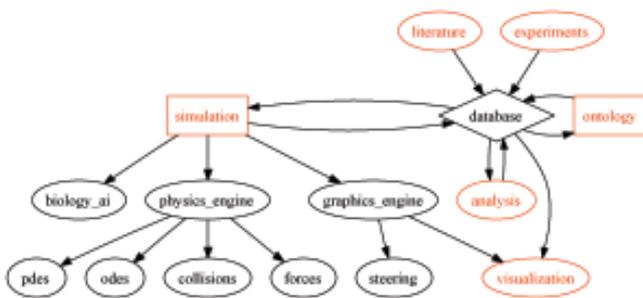
Fig. 3. Framework for data analysis and immune simulation.

density to estimate the value of $k$. Alternatively, if we had an estimate of $k$ available, we could simulate the resulting distributions of $x$ with different values of $c$, and these predictions can then be tested with flow cytometry experiments.

This is a particularly simple example of how one can use formal models and numerical simulations to try to integrate data from different types of experiments. Similar approaches involving the use of statistical and/or mathematical models are also being developed to optimally extract knowledge from other immunological assays – giving us the pieces we are trying to put together again.

**Integrating Data to Reconstruct Immune Responses**

We are interested in the integration of data from microarrays, cytokine assays, flow cytometry, immuno-histochemistry and 2-photon video-microscopy to reconstruct the early immune events at the site of injection and draining lymph node following injection of an antigen with adjuvant. Our first problem is simply to understand who does what, when and where. The trouble is that each assay only gives information about a small aspect of the immune response we are interested in – how do we relate the functional behaviour of each cell with its location in space and time? After that, how does the behaviour of individual cells become the integrated immune response we are interested in? We are in the position of the 6 blind men of the Indian fable who are trying to understand an elephant.

Our approach is to reconstruct the system we are trying to understand as nested sets of mechanistic models. Mathematical models have been described above – here we are interested in models that can be mapped easily to biological mechanism (hence *mechanistic*), and furthermore, are constructed in such a way that a more complex model can be put together from simpler models of its components (hence *nested*).

In our formulation, a reconstructionist model for an immune response consists of 3 main parts – models for the tissue environment, cell types and soluble factors. The
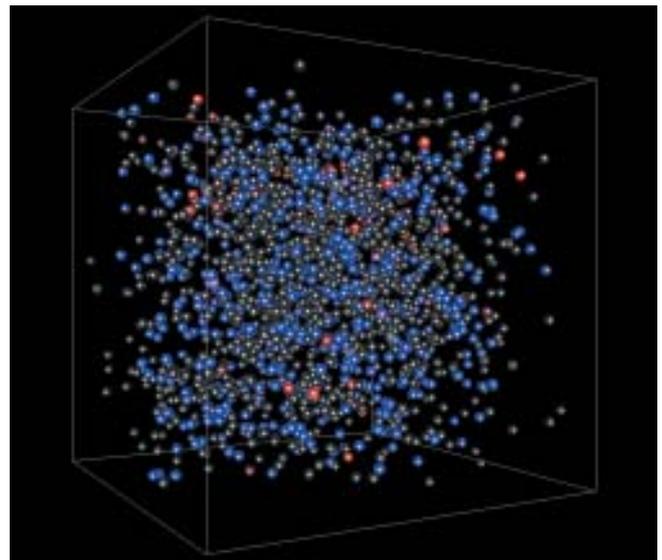


Fig. 4. Screen shot of immune simulation in progress showing aggregation of macrophages in response to an LPS stimulus after 32 hours. Blue and red colours indicate relative activation of the tumour necrosis factor (TNF) and soluble tumour necrosis factor receptor (sTNFR) shedding pathways respectively, while naïve cells and exhausted cells are light and dark grey respectively.

environment model accounts for the tissue structure, as well as stromal cells that play important roles in an immune response but are not modelled in detail. Each cell type is a composite model with sub-models governing different effector functions like motility and cytokine secretion. Soluble factors representing antigen, cytokines and chemokines are modelled as concentration matrices governed by numerical algorithms for reaction-diffusion processes.

While we use the standard approaches of applied mathematics (i.e., stochastic, ordinary and partial differential equations) to model the evolution of individual components, the overall simulation is designed with an object-oriented framework. As a result, we have an *individual-based* simulation, in which each cell of a particular type behaves as an autonomous agent in its environment, capable of sensing and responding appropriately to its physical environment, soluble factor concentrations, and every other cell. This allows us to maintain a faithful, albeit simplified, correspondence between model and biology, allowing our experimental colleagues to provide useful feedback and critique of the simulations.

In addition to the core *biological* modules which describe tissue environments, cells and cellular processes, the simulation framework also includes *physics* modules which handle reaction-diffusion equations, collisions and forces between objects, *visualisation/steering* modules that allow user interaction and control, and *informatics* modules for

data storage and retrieval. The overall schema is shown in Figure 3.

Readers interested in the technical aspects of model construction, validation and analysis can refer to a recent review of our work,[9] in which we simulate the regulation of an inflammatory response by soluble tumour necrosis factor receptors. Typical visualisations of the simulation process are shown in Figure 4.

This project to construct a virtual world of the immune response is still in its early days, and much remains to be learnt about how to convert data from biological experiments into algorithms that accurately describe the behaviour of immune cells in time and space, but the promise of the approach is that it provides a way for clinicians, experimental and theoretical scientists to work together to unravel the secrets of the immune system. We believe that it is also an ideal tool for learning about the immune response, especially when the simulation is projected into a virtual reality environment in which one can experience actually being *in* an immune response, which is being implemented with our colleagues in the Duke Immersive Virtual Reality (DIVE) facility.

## Conclusion

Engineers have a saying that one only understands what one can build, and we are attempting to understand immunology by building a model of an immune response, one cell type at a time. Due to the complexity of the biology, such models will inevitably be gross simplifications of what is really happening during an immune response. Still, such individual-based simulations are much more faithful to the biology than traditional ODE or PDE models. An important feature of such simulations is that they are accessible to the clinician or experimentalist, since the description is at the level of familiar biological entities – tissues, cells, cytokines and chemokines. Being able to visualise and manipulate the simulation in 3D is useful too, since interesting or aberrant system behaviour is often immediately visually obvious, allowing us to discover bugs quickly or pursue leads for further investigation.

Models are just formal stories for our current understanding of reality, not very different in spirit from descriptions of canonical diseases found in medical textbooks. It seems inevitable that formal and computational models will become more and more important to medical science as this is the most powerful way to integrate the exponentially growing amount of experimental data, so as to better understand complex physiological or pathological phenomena.[10,11] Ideally, a substantive quantitative component will be integrated into the medical curriculum to equip doctors with the knowledge necessary to exploit such models. However, the day when doctors will be cross-trained in both the formal and medical disciplines seems far away still.

We believe that developing realistic biological simulations can facilitate communication and provide non-intimidating exposure to the usefulness of models in biology and medicine. When the simulation maps well to the more intuitive biological knowledge tacitly known by experimentalists and clinicians, it provides a bridge between 2 very different worlds, and allows both communities to work together to tackle complex physiological systems. It is our hope that, together, we can attain useful insights into our fabulously complex immune systems that would be difficult to come by otherwise, and so improve our ability to design new experiments, new diagnostic tools, and new therapies.

REFERENCES

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-45.

2. International Human Genome Sequencing Consortium. A haplotype map of the human genome. Nature 2005;437:1299-320.

3. Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, Collins FS, et al. The knockout mouse project. Nat Genet 2004;36:921-4.

4. Kitano H. Systems biology: a brief overview. Science 2002;295:1662-4.

5. Herzenberg LA, Tung J, Moore MA, Herzenberg LA, Parks DR. Interpreting flow cytometry data: a guide for the perplexed. Nat Immunol 2006;7:681-5.

6. Perfetto SP, Chattopadhyay PK, Roederer M. Seventeen-colour flow cytometry: unravelling the immune system. Nat Rev Immunol 2004;4:648-55.

7. Roederer M, Hardy RR. Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56-64.

8. Roederer M, Moore W, Treister A, Hardy RR, Herzenberg LA. Probability binning comparison: a metric for quantitating multivariate distribution differences. Cytometry 2001;45:47-55.

9. Kepler TB, Chan C. Spatiotemporal programming of a simple inflammatory process. Immunol Rev 2007 (In press).

10. Ho RL, Bartsell LT. Biosimulation software is changing research. Biotechnol Annu Rev 2004;10:297-302.

11. Smye SW, Clayton RH. Mathematical modelling for the new millennium: medicine by numbers. Med Eng Phys 2002;24:565-74.