I often introduce myself as an accidental genomicist. My earlier undergraduate and graduate school training had been in pure physics, communication engineering, computer science, and mathematics. My research career has been structured around the theme of bringing rigorously developed efficient computational approaches into applied areas such as very large-scale integrated circuit (VLSI design), computer languages and their compilers (PQCC, Production Quality Compiler-Compiler), computer-aided verification, logic, control theory, robotics, game theory, and mathematical finance.

In 1997, I saw an opportunity to apply algorithmic and statistical ideas to create accurate genome-wide physical maps of organisms, starting from single genomic DNA molecules. In a very productive collaboration with two very creative scientists, D.C. Schwartz, a chemist, and T.S. Anantharaman, a computer scientist, I developed an integrated optical mapping platform that could create an ordered restriction map of a whole-genome from a small amount of genomic material and in reasonable amount of time. With many innovations in the statistical design of experiments, image processing, statistical data-analysis, and more reliable chemistry, it was possible to produce the first working single molecule mapping technology for clones [Jing et al., *PNAS*, 1998], whole microbial genomes [Lin et al., *Science*, 1999], and contiged multi-enzyme BAC maps of difficult-to-sequence regions of human genome (e.g., DAZ locus of Y-chromosome) [Giacalone et al., *Genome Research*, 2000].

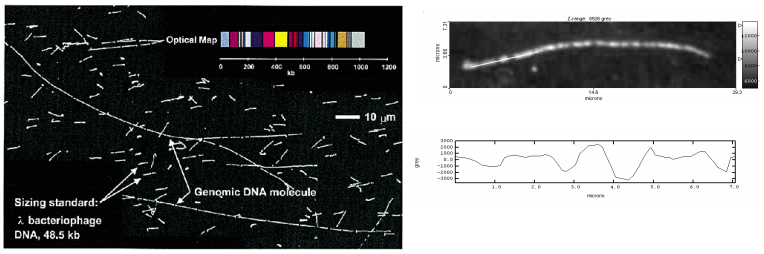

Figure 1: *(Far Left, circa 1999) Digital fluorescence micrograph and map of a typical genomic DNA molecule. P. falciparum molecule digested with NheI is shown with its corresponding optical map. Comparison with the consensus optical map shows this molecule to be an intact chromosome 3. (Left, circa 2005) . Fluorescent image of lambda DNA, stained with YOYO-1, after in situ partial digestion by HaeIII restriction endonuclease. Below, the intensity contour across the two sites is indicated by the white line. (Images from J. Reed et al. 2005, under preparation.)*

The basic technology is simple to explain: Uncloned DNA (directly extracted from cells by lysing) is randomly sheered into 0.5-2.0 mega base pieces and attached to a charged glass substrate, where it is first cleaved by the restriction enzyme, then stained with a fluorescent dye. The restriction enzyme cleavage sites show up as breakages in the DNA under a fluorescent microscope. Tiled images of the surface are collected automatically using a fluorescent microscope with a computer controlled *x-y-z* sample translation stage. The approximate size of a restriction fragment is estimated based on the integrated fluorescent intensity relative to that of a standard DNA fragment (typically some small cloned piece of DNA, e.g., Lambda Phage Clones) that has been added to the sample.

The underlying chemistry, while unruly and noisy, was tamed by our system with proper experimental design and a sophisticated Bayesian statistical algorithm ("Gentig"), created in collaboration with Anantharaman. A key insight to control the computational complexity came from a probabilistic analysis that indicated a "0-1 Law" at play: In a general setting, the computational problem is NP-complete and the probability of computing the correct map is close to zero, but if experiments were designed properly (e.g., choice of restriction enzymes, digestion rate, errors in sizing the restriction fragment lengths, optical chimerisms, etc.) then one could guarantee that the correct map was computed with arbitrarily high probability in sub-quadratic time [Ananthraman et al., *Jnl. Comp. Bio.*, 1997]. In an elegant synergy, where chemistry stumbled, algorithmics rescued it, and vice versa. As the technology moved on to assume a commercial life at a start-up (OpGen in Wisconsin), I began to seek a deeper role for computing in deciphering the principles of biology.

With a bias that I shared with many of my colleagues from physics, mathematics and computer science, I had assumed that once the genome was sequenced, properly annotated, and made available for instant downloads through a browser, we would be free to return to our natural home disciplines, since the rest would be simple and mechanical. In a computational version of Crick's central dogma: the biologists could start with draft versions of genomic sequences, readily constructed through shot-gun assembly; next, annotate them with

statistical-learning algorithms to find regulatory elements, genes, exon-intron boundaries, etc.; find the transcribed and spliced mRNAs; computationally determine the amino-acid sequences, proteins and protein-protein-interactions; and finally, reconstruct the regulatory, metabolic and signaling networks, from which we could discover how different phenotypes arise. In a similar computational version of Wilsonian new synthesis, we could compare the genomes of different species, or polymorphisms within a population to compute phylogenetic relations and maps of haplotypic variations (HapMaps), and be able to understand the biological diversity, evolutionary processes and selective pressures on the phenotypes. Naively, if this picture were true, we could declare victory and assume to now possess the book of life—to be browsed anytime at leisure. Furthermore, in this setting, the most critical tools for biologists would come from better bioinformatics environments and systems biology. With a team of about 20 computer scientists, mathematicians and biologists, I focused on two computational tools: Valis and Simpathica. As we continued along this direction, there would occur a number of events to drastically change our collective perception of biology. We would find that biology is infinitely less banal, and demands more attentive and patient examiners.

In collaboration with S. Paxia, I designed Valis to be a flexible, efficient, and highly scalable software system that can be used in biological applications, in the same manner that MatLab is used by the engineering community, or Mathematica, by the physical science community. Valis is a rapid-prototyping tool, an efficient software environment (supporting scripting languages such as Perl, Python, Javascript, VisualBasicScript, etc. and very high level languages such as R, Octave, Lisp, Setl, etc.) with a novel scalable free-format database [Paxia et al., *IEEE Computer*, 2002]. Initial applications of Valis have been in the areas of genomics and systems biology, sequence validation with physical maps from *optical mapping* or *array mapping*, investigation of lesions in breast and prostate cancer genomes, systems biology models of apoptosis processes, purine metabolism, mitosis and meiosis in *C. elegans*, etc.

The Simpathica system, developed in collaboration with M. Antoniotti and N. Ugel, allows biologists to construct and simulate models of metabolic and regulatory networks and analyze their behavior. Such pathways can be drawn on the screen through a visual programming environment, in a specialized XML format (SBML), or by modularly combining existing building blocks representing biochemical reactions and modulations of their effects [Antoniotti et al., *Cell Biochem. and Biophys.*, 2003 & *Theo. Comp. Sci.*, 2004]. The Simpathica tool allows a user to query and reason about the temporal evolution of these pathways in plain English. With collaborators at Walter-Reed and Cold Spring Harbor Laboratory, we have used Simpathica to explore a *Caspase cascade* model for apoptosis, and other proapoptic processes.

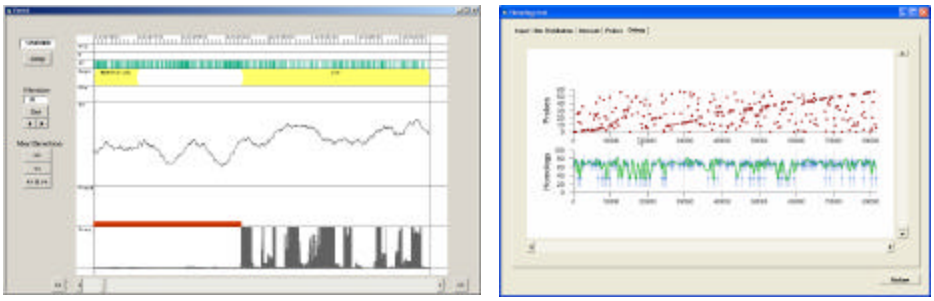

***Figure 2**. (Far Left) Valis, displaying the AT positions, repeats and coding region annotations, strand melting free energy (?G), segmental duplication copy numbers and frequencies of 20bp-mers in a human genomic regions. (Left) The Bayesian homology algorithm PRIZM in **Valis**. M. genitalium (Y-axis) and M. pneumoniae (X-axis), were compared using 300 probes from six iterations, taking about 5 seconds The microbial genomes share about 75% homology.*

To gauge the power of a tool like Valis, one may consider the Valis implementation of PRIZM software, a homology algorithm I designed to compare whole genomes quickly. This new homology algorithm (unpublished), using a Bayesian scheme, can efficiently compute homologous regions between two genomes even when the homology level drops to a value around 75% (Figure 3). These algorithms incorporate background knowledge about genome evolution (using various priors: noninformative improper prior, exponential, Gamma and Kummer-Gamma priors, and priors based on Juke-Cantor one parameter and Kimura's two parameters models of evolution). Accordingly, they can achieve remarkable efficiency. (It takes about 30 seconds to compare a rat-mouse chromosome.)

On may ask: What more does biology need from computer science? A short menu will contain: genome assembler and browser, genome alignment and comparison tool, genome annotator, microarray database and

clustering tools, and tools to simulate biochemical pathways and find their homeostatic states (flux-balanced-analysis). Will that be all? If not, what is to be done?

Around 2000, a chance meeting with M. Wigler, a cancer biologist, at Cold Spring Harbor Laboratory led to a deep collaborative relationship. The focus of this collaboration was to invent a system on top of an available off-the-shelf technology (e.g., spotted or oligo-nucleotide arrays) to cheaply map and compare genomes with a respectable resolution (about a marker every 30Kb). The main application of this technology would be in comparing tumor and healthy genomes in cancer to understand copy-number variations. While it was easy to see that a single-molecule technology like optical mapping could eventually surpass this array-based technology, the array-CGH technology had the advantage of being widely used. These technologies now exist in my lab and Wigler-lab: linear-mapping, ROMA and NarCGH-software.

At the invitation of M. Wigler, and with encouragement from B. Stillman and J. Watson, I joined Cold-Spring Harbor (CSHL) as a half-time faculty (the other half being Courant). My colleagues at CSHL, most prominently Wigler, turned out to be a group of excellent mentors for a biology-challenged not-so-young technologist. For me, it was a chance to learn about biology directly from many excellent practitioners. Instead of focusing on a body of knowledge that biology had accumulated, I began to think about how biologists think about biology and how it differs from my own training. For instance, in the foreseeable future, it is hard to argue that there will be a clear way to determine if the current body of biological facts is sufficient to explain phenomenology. As various phenomenological inconsistencies become apparent, new experimental techniques, new models, and new biomedical applications arise quickly in rapid and sporadic paradigm shifts.

A concrete example of these ideas appear in the context of onco-genomics, the genomic models of cancer evolution, and possible relation of such models to a wider body of biomedical knowledge, as in aging and development. A study designed along this line motivates the following three specific aims for my future research:

***Aim 1:* Develop inexpensive and accurate single-cell single molecule technologies to perform haplotype sequencing, methylation patterns and chromosomal aberrations in whole human genomes. Develop inexpensive and accurate single-cell technologies to characterize cDNAs, quantitate gene-expression profiles, alternative splicing and RNA stability. Apply this technology to characterize various classes of tumor-cells and cells in a poly-clonal tumor.**

***Aim 2*: Using the data created in Aim 1, develop dynamical models of genome evolution, changes in methylation patterns, changes to alternative splicing, and micro RNA to explain their role in cancer and aging.**

***Aim 3:* Using the models of evolution created in Aim 2, formulate a mathematical predictive model of stem cell clonality to explain their role in cancer initiation and aging.**

In the context of cancer genomics, phenomena such as copy number fluctuation studies in array-CGH data promises to give us many insights about various players in cancer initiation and progression: in particular, we expect a chromosomal segment that is systematically amplified to contain oncogenes and similarly, a chromosomal segment, hemizygously or homozygously deleted, to overlap with tumor suppressor genes. In collaboration with A. Rudra, R. Daruwala, I. Ionita, and Wigler, I have developed statistical algorithms to segment arrayCGH data (working over wide variety of technologies: BACarrays, ROMA and Affymetrix chips, with powerful background correction algorithms), assign LOD-scores to intervals, and determine precise locations of onco- and tumor-suppressor genes [Lucito et al., *Genome Research*, 2000 & Daruwala et al., *PNAS*, 2004]. But when one examines the data more carefully, there seem to be much more interesting dynamics in the way cancer genomes evolve, leaving behind tell-tale signs in the distributions of break-points and lengths of aberrant segments. The methods I have developed in the context of segmental duplications in mammalian genomes (in collaboration with Y. Zhou and E. Thomas et al.) [Zhou et al., *PNAS*, 2005, Thomas et al., *PNAS*, 2004] could be adapted with appropriate modifications to reveal these processes. The underlying mathematics, in this setting, employs a dynamic (Markov) model of genome evolution to enumerate possible genomic states (affected by recombination-like mechanisms or induced by fragile subregions), and the transitions among them.

In order to validate hypothesized mechanisms, it relies on various rate-parameters: (mutation rates in genome, mutation rates in LINEs, Alu & L1 amplification rates, etc. in evolutionary context). To understand genomic processes, involved in cancer, similar rate parameters must be extracted from multiple patient data and relate them to various plausible hypothetical mechanisms. For instance, one may hypothesize distribution of hot-spots (where duplication is initiated) and cold-spots (where duplication must be terminated as otherwise it might affect a vital gene) that modulate and are modulated by genome-evolution and selection processes. These questions are summarized in my Aim 2, and will be addressed through algorithms that manage a database of genomic information; cluster them according to population stratification, existing copy-number polymorphisms, disease class, disease-stage, etc.; statistically assign temporal order to important events in disease progression; extract rate parameters; create accurate mathematical (hybrid-system) models[1]; and validate (or refute) hypothesized mechanisms.

On the other hand, we also suspect that chromosomal aberrations may provide only a partial pictures as we know that epigenomics, microRNAs and alternative splicing also play a role and closely intertwine with genomic duplications, deletions and translocations. Extracting a more detailed picture of cancer from tumor genomes can only be imperfectly achieved with bioinformatics. A clearer picture emerges with access to an accurate, high-throughput single-molecule technology that can characterize DNA molecules of any size and any abundance at any scale (through sequence or densely occurring probe or restriction-enzyme based markers). While a similar technology for proteomics (e.g., mass-spec or protein arrays) would be an ideal partner, it will remain beyond our immediate scope. As I have already made some significant progress in single-molecule based technologies (unpublished), it is now possible to plan for a fast, accurate, high resolution, and yet, inexpensive platform to map genomic DNA and cDNA from mRNA. For instance, with my collaborator J. Reed at UCLA, I am developing a system to cleave cDNA's with four-cutters and measure the fragment lengths using AFM. With this approach, it is feasible to measure the abundance of mRNA in a single cell with high accuracy, recognize alternative splicing, and measure RNA-stability. By augmenting restriction enzyme cleavage process with LNA (Linked Nucleic Acid), PNA (Peptide Nucleic Acid), pcPNA (pseudo-complementary PNA) and TFO (Triplex Forming Oligo) based probe-hybridization, with my collaborators C. Cantor, V. Demidov, A. Lim, Anantharaman, and Reed, I have been planning methods and algorithms to create haplotypic whole-genome sequences [2]and methylation patterns. But four-cutter enzymes, several thousand eight-mer hybridization probes, haplotypic separation, etc. conspire to push the resulting computational problems into a new parametric setting, changing the landscape of 0-1 laws into a terra incognita where the old algorithmic techniques become ineffective; several fresh new ideas are needed: complex mixture models, nonlinear Kalman-Bucy filtering, complex geometric hashing, beam-search like heuristics, etc. But once these tools are available, it will be possible to characterize every single tumor cell in a poly-clonal tumor, organize them in a phylogeny, decipher genome evolution processes and enumerate selective pressure that sculpted that tumor.

Developing this powerful technology will constitute my Aim 1. The constituent tasks interact with each other in a multi-dimensional design space: (1) Surface Chemistry, (2) Optical Systems, (3) Probe and Restriction Enzyme Chemistry, (4) Data Collection, Analysis and Assembly Algorithms, (5) Software Implementations, and (6) Computer Hardware Architecture. It will be necessary to develop each part independently, but flexibly, so that they can be mixed-and-matched with relative ease to be used for genomic mapping, comparative genomics, and expression profiling. The technical challenges in each of these areas are many, but not beyond the accumulated experience of our multi-disciplinary team. Given the limited available space, it is illustrative to examine just one example: optical systems, an apparently straightforward technology. It is necessary that, using fluorescent probes, the fluorescent microscopy images must be able to detect single fluorochromes in the image. This is possible using conventional wide field fluorescent microscopy, if we combine laser illumination to increase the signal with various techniques to reduce the background noise, including the use of barrier filters tuned to the fluorescent probe wavelength, low noise quartz cover slips and

[1] More general than traditional Vogelstein-maps.
[2] For a cost of about $700 and taking less than 24 hours, by our estimate.

slides [Funatsu et al., *Nature*, 1995]. Additional enhancement of the signal to noise ratio can be achieve by using TIRF (total internal reflection fluorescent) microscopy. Furthermore, the nature of these signals, using image calibration, chromatic aberrations of the lens system, point-spread function, etc., must be accurately modeled in the image-processing and subsequent statistical analysis software. Similar issues crop up in each of the other areas. Note that while my aim 1 represents a very ambitious goal, and has a long-term payoff, my subsequent aims can proceed in parallel with a simpler subset of technology (achievable within a year and a half) providing measurement accuracy adequately beyond currently available microarrays.

One may wonder whether these processes, we discern, are simply anomalous, only responsible for cancer, or perhaps, they are much more ubiquitous. One may imagine, like tumor cells, any cell in our body, undergoing cell division, is subject to the same evolutionary processes and same selective pressures. In particular, if stem cells, in response to injuries, undergo symmetric divisions, there may be pools of highly clonal stem cells in various niches accumulating over one's lifetime. Thus these processes may, in fact, hold a very general explanation for autoimmune diseases, aging and various forms of neuro-degenerations. With my graduate students M. Heymann and Y. Zhou, I have been developing a stochastic differential equation model that may explain qualitatively many interesting genomic events of this nature. A very interesting picture of development and aging may be seen by combining the genomic portraits of diverse groups of human cells taken over a wide-range of time scale: from spontaneously aborted fetuses to octogenarians, and everything in between.

I suspect that this view of aging and cancer may be somewhat idiosyncratic, and controversial, and in normal context of grant applications, lends only to a strategy of careful avoidance. Nonetheless, it has a quiet simplicity, already explains a disparate set of phenomena elegantly, can be mathematically dissected, and may in turn point to even more general biological principles. On the other hand, even if it fails, it would still leave, on its wake, many important concepts, technologies and tools for traditional biology.

In summary, it is our thesis that most important critical steps for biology now constitute parallel development of both faithful models and accurate measurements, a symbiosis between theory and experimentation—large-scale computation and high-throughput biotechnology—the "wet" and the "dry." A concrete, useful and elegant application of this framework emerges as we attempt to better understand aging, cancer and disease progression from a genomic perspective, but demands even more powerful single-molecule genomic technology wedded to a flexible and effective bioinformatics. The main thrust of this research will be to sharply focus a wide variety of technologies on to one biomedical problem of critical importance, and in the process create a new paradigm for understanding complex biological processes.

Finally, the scope of these problems is beyond a single individual, or even a single field. But my close collaborators and I have always thrived in an environment dealing with ill-posed non-obvious problems, problems requiring cross-disciplinary migrations, and kind mentoring and collaboration from other like-minded individuals across many disciplines. We wish to continue in that tradition.