

**Written Statement of Dr. Stanley Burt for the Senate Committee on Commerce, Science, and transportation Subcommittee on Technology, Innovation, and Competitiveness to be held July 19, 2006.**

Over the past couple of decades many important milestones in biology have been obtained. These include completing the genomic sequences from several mammalian genomes, including human, and a producing draft sequences for several additional genomes. Also, new technologies allowing for simultaneous measurement of mRNA expression levels for thousands of transcripts and application of this method to RNA samples from tumors and normal tissues have identified many genes whose expression is influenced by cancer and other disease states. Further improvements in this technology and other discoveries have led to chip technologies capable of simultaneously monitoring the entire genome for person-to-person variations including both single nucleotide polymorphisms (SNPs) and other larger alterations such as deletions and duplications. Other data derived from experiments that measure microRNA expression levels and transcription factor binding studies have also contributed to the extensions of this technology. Finally, measurements of protein expression levels using high throughput mass-spectrophotometry and chip based tissue and antibody arrays have given biologists the ability to correlate changes in mRNA expression with changes in protein expression levels that may contribute to the disease process or at least be markers for these altered physiological states.

New instrumentation in microscopy has allowed for simultaneous monitoring of cells responses to drugs or other agents in parallel. Further, con-focal imaging techniques have allowed for multiple slices of the same fields to be examined in detail so that a three dimensional image of a specimen can be reconstructed. Whole-animal imaging is also being used to study drug distribution throughout an animal's tissue. Other methodologies allowing for higher levels of protein expression and purification are being leveraged to allow for more direct biochemical metrics of an enzymes function to be collected. High throughput binding technologies can be used to determine affinities of proteins with cofactors and drugs. Better docking software applications now exist to screen some of these interactions in-silico. Newer, more sensitive and reliable methods have been developed out to identify protein-protein interactions. The biomedical literature has also grown dramatically as all of these new methods and the data associated with each of them has increased.

Taken together, these new methods and the need to process and analyze the data produced by them have resulted an explosion in the need for high performance computing in biology and medicine. This need requires both increased capacity, as the sheer volume of data generated is considerable, and also increased capability. One of the

confounding problems associated with the needs analysis of this problem is that there does not appear to be any single solution to the problem. Because of the diversity in the algorithmic requirements for analysis of each of these data types, no particular computer hardware seems suited for all of the problems. Thus, some of the problems are embarrassingly parallel, meaning they are ideally suited to a cluster environment. Good examples of this type of application would be comparing fragments of one genome to another, where each computation is entirely independent of the other. Advances in microarray plating technology now allows for increased spot density. This translates into a tremendous increase in the amount of data from a single experiment, at a significantly reduced cost. In addition, since these experiments only produce useful results when they are run for many samples (e.g. tumor and normal tissues) a greater volume of data is produced. This is leading to the need for the biologist to have access to computers with more memory and higher processor speeds to allow the data to be analyzed in a reasonable time. Already, the ABCC has received requests from cancer biologists for help with genomic analysis of promoters, control regions, miRNAs, better annotation of the genome and comparison of genomes, understanding of fragile sites, sites of chromosome translocation, and the relationship to cancer of segmental duplications (see pg. x). In addition, the new 500K SNP chips are flooding researchers with data that requires big computers to process, store, and interpret. Cancer biologists want new methods to look at the data and estimate haplotypes and look for interaction among many loci. This and all of the abovementioned challenges require the use of HPC resources.

Another area in which cluster computing can be useful is in biomarker discovery. Aside from prevention, diagnostic tools to detect cancer at an early stage are of great benefit to patients. Great efforts are being made to identify biomarkers from gene or protein expression profiles. One tool being used to find biomarkers is mass spectroscopy, which can identify proteins and their fragments based on their size and electrical charge. In mass spectroscopy experiments thousands of spectral peaks are produced. These peaks are then used to find biomarkers for proteins. Because there are so many data points that are trying to be fit to few markers, this can lead to false results because the problem is over determined. In order to avoid this mistake, one needs to perform thousands of calculations to develop a consistent set of models to find the proper biomarker. The ABCC does this by using methods that converge on a model that has the same biomarkers in each solution, thereby guaranteeing a biologically relevant answer. This procedure can benefit from hundreds of processors, but large memory is not needed since each calculation is independent of the others. The ABCC has been successful in finding biomarkers for bladder cancer and colorectal cancer. Hopefully, these markers, which are derived from urine and human serum, will translate into efficient, inexpensive screens that can be used for early detection of these cancers.

Another problem that confronts biological computing and cancer research in particular is the sheer volume of data that must be collected, analyzed and compared. Data already exists in older databases in many places and in different formats. Part of the problem is already being approached by the NCI through its caBIG (Cancer Bioinformatics Grid) initiatives of NCICB and it involves identifying and leveraging information technologies that facilitate data interconnectivity, amongst other goals. In this regard, the development and enforcement of data exchange standards through caDSR and caCore are designed to bridge the gap between a clinicians and a bioinformaticists perspective of a set of genomic data. In order to analyze and house this data, there needs to be a computational infrastructure and visualization capabilities. Furthermore, while distributed databases are convenient for data maintenance, the National Security Agency has found that having all the data reside locally, where it can be called into computer memory, is essential for rapid data scanning. This will require HPC resources with large memory resources. Also, database consolidation is not enough. There needs to be development of methods for the construction of a knowledge base in which non-experts, especially clinicians, can query data from various sources. This will require a serious research effort in knowledge base development area, although some manufactures have obtained preliminary results in this area. In addition, because there are problems suited for both hardware configurations mentioned above, the data I/O infrastructure must also be able to be connected to both of these scenarios. Again, because of bandwidth issues resulting from the sheer volume of the data, this results in a need for new technologies in computer architectures.

Another complicating factor in data combination and analysis for biological research is that while massive storage and bandwidth have become relatively cheap and abundant, the data can not only be from different sources but it can represent experiments in different scales, from years to femtoseconds—time scales that go across orders of magnitude. This is a problem that is referred to as multiscale modeling, and it is a profound problem in computational science. Solving this problem will require a commitment of resources to advanced architecture development, more efficient algorithms, and clever data reduction.

I will now address some computational bottlenecks for a few areas that the National Cancer Institute has identified for their roadmap.

### **Nanotechnology:**

Nanoparticles typically have dimensions smaller than 100 nanometers, which are smaller than human cells. Nanometer devices smaller than 50 nanometers can easily enter most cells. Nanoscale devices can interact with biomolecules on both the cell surface and within the cells. Despite their small size, nanoscale devices can also hold tens of

thousands of small molecules such as a contrast agent or a multi-component diagnostic system capable of assaying a cell's metabolic state. This can provide a mechanism for detecting cancer at its earliest stages. Nanoscale constructs, such as dendrimers and liposomes, can provide customizable drug delivery to targeted cancer cells or tissues. This has already been demonstrated experimentally.

While nanoparticles have great promise, it also has to be demonstrated that they are not toxic to normal tissue. The ABCC is supporting the NCI's Nanoparticle Characterization Laboratory through modeling of bulk properties and calculation of atomistic properties. At the nanoscale, the physical, chemical, and biological properties of matter differ fundamentally and often unexpectedly from those of corresponding bulk material because of the quantum mechanical properties of atomic interactions which are influenced by material variations on the nanometer scale. Modeling of bulk properties such as surface charge or shape is not difficult. The calculation of atomic level quantities is a huge computational issue, even atomistic calculations on quantum dots are beyond our current capability, and will require large increases in HPC.

#### **DRUG DESIGN:**

Over the years there has been great success in drug design using HPC. Drug design is usually done against a protein target, such as an enzyme whose function one wants to inhibit. A great recent example is the discovery of Gleevec, an inhibitor of protein kinase activity, which brings about complete and sustained remission in nearly all patients in the early stages of chronic myeloid leukemia. If the structure of the protein is known, docking calculations can be performed. This usually involves docking thousands of molecules into an active site and scoring the resultant interaction. If the docking is done with rigid molecules, the calculations are fairly trivial. If, however, flexibility is allowed, and most proteins and ligands do flex, then the problem becomes enormously computationally expensive.

If the protein structure is not known, and the protein is not similar to another one, then one must perform *ab initio* structure determination. David Baker's group at Illinois took approximately 150 CPU days to determine the structure of the CASP6 target T0281. Also to do a docking interaction between two proteins took 15 CPU days. He makes particular note that his group is limited by computational power. Our group has studied the enzyme mechanism of many enzymes involved in cancer. For an enzyme named Ras, which is mutated in over 30% of known cancers, we modeled 1,622 atoms of the protein by molecular mechanics and only 43 atoms by quantum chemistry. These studies took several years and were bound by computational power. To calculate reaction surfaces normally takes several months of time on HPCs. Luthey-Schulten's group at Illinois did molecular dynamics simulations of Imidazole Glycerol Phosphate Synthase, an enzyme involved in making DNA and RNA. It took 10 hours, 12 hours, and 40 hours to animate

one nanosecond on three cluster machines (with different processor speeds). It takes many nanoseconds of simulation to just relax the systems to prepare for further simulations. It has been estimated that to go from nanoseconds to milliseconds will require an increase in computer capacity of approximately 1,000,000. This can only be achieved by the combination of improved hardware and software.

### **INTEGRATIVE BIOLOGY:**

Computer aided design of HIV protease inhibitors remains one of the most successful stories in modern biology. Although this was a remarkable achievement, the complexity of a single viral particle pales in comparison to characterizing the complete catalog of the cell (the proteome) and the full map of the interactions of the members of the proteome. For a subset of interactions of the proteome, the immunome, the combinatorial problem of treating all possible pairs in the immunome (1,000,000 of them) escapes the capacity of current computers.

### **SYNZYMES:**

There is great interest both in academia and industry for the creation of artificial enzymes that are much smaller but duplicate the enzymatic activity the large natural ones. Because they are smaller, they can be tethered to other molecules or nanoparticles, such as dendrimers or liposomes, and delivered to a particular targeted area such as a tumor cell. The ABCC staff has experience in this area. We modeled a particular inorganic catalyst know as Mn-salen, which is used commercially in the chemical industry for epoxidation reactions. After studying this reaction, we were able to convert this catalyst into one having biological activity and could act as a free radical scavenger. This could be useful for traumatic injuries, strokes, or even for cancer. However, this falls into the same category as enzyme mechanism studies and the calculations take months and months to perform. Complete characterization of these reactions took several years running on fast HPCs.

### **Specialized Hardware:**

Being able to take advantage of specialized HPC resources and software written for those resources can lead to dramatic increases in time to solution. In one instance, the ABCC staff in a research partnership with several NCI biologists investigated how to rapidly scan for microsatellites (tandem repeats). Tandem repeats are groups of DNA nucleotides

ranging from two to sixteen bases that are expanded in several diseases. For example, in normal people there is a pattern of DNA nucleotides, CAG that is expanded 10-35 times. In Huntington's this same pattern is expanded between 36-121 times. In the past finding these repeats were found in a heuristic and probabilistic manner on conventional computers.

Using specialized hardware such as bit matrix multiply and pop count, which had been requested by the NSA to be incorporated machines they were using in order to perform rapid pattern matching, we, along with industrial programmers, were able to drastically reduce the time to find all tandem repeats on chromosomes and the entire human genome. To scan a chromosome of approximately 150 million bases took 2 seconds. To scan the entire human genome took 2 minutes. We discovered 47 potential disease sites, 8 of which could be associated with cancer, and we more than doubled the known numbers of repeats. We also used this specialized hardware search for another genomic feature, known as segmental duplications, which are associated with diseases. This involved finding clusters of DNA bases approximately tens of thousand bases long that are separated by approximately 1 million bases from another cluster of bases that are the complimentary complement of the original DNA base cluster. When these complimentary clusters find each other during replication they combine and huge sections of the genome are excised. We could not have done this without these specialized hardware features.

We have also used FPGAs, which are reprogrammable hardware and support the custom computing needs that are characteristic of data-intensive problems. We programmed the FPGA for a powerful sequence alignment algorithm known as Smith-Waterman. The Smith-Waterman alignment method is a powerful algorithm for aligning sequences in which there may be gaps and one is trying to find the "best" alignment. This algorithm is widely used in the biological community but is particularly computationally demanding. We obtained speed-ups of over a thousand fold. However, the difficulty is that the programming of FPGAs is not a trivial task, and one that would not be normally within the expertise of a biologist. However, FPGAs offer great promise because there are expected to be huge increases in performance on these types of machines.

### **Recommendations:**

It has been said that biology will be the science of the 21st century. Due to the complexity of biology, the sheer volume of data, the fact that the environment of a cell, (particularly for cancerous cells) must be taken into account means that biology must be tackled using a systems biology approach. This means that teams of scientists such as biologists, computer scientists, mathematicians, physicists, and chemists should work on these problems in conjunction. In order to do this, it will require cross training to have a

meaningful dialogue. I believe that in order for the United States to remain competitive we should devote funding to education and training in the above disciplines. We also need to find mechanisms to encourage young people to enter the scientific field. I have seen for several years the lack of U.S. citizens applying for jobs in the ABCC. I believe that this reflects the national trend.

As biology matures the use of HPC in biological research will grow. There is clearly a need for large memory assets coupled with fast processors. I believe that cluster computing will still have its place, but as the problems grow in size and complexity, the need for HPC resources will be inevitable. One can already see this trend in Europe where several national centers have made purchases of Blue Jean machines, and others have made investments in large memory machines.

There needs to be funding of new computer architectures, specialized hardware, faster interconnects, etc. One area of funding that is especially important is software development. One thing holding back HPC development is that the software available today is not written for HPC machines. Sometimes software engineers spend considerable time to port non-parallel applications to parallel machines without much increase in speed or efficiency. We are running “old” software on newer architectures. Along with developing new software, research into new compilers must be encouraged.

I also recommend that the United States fund several centers for Integrative Computational Technology for Systems Biology. These centers would provide for the integration of biology with strong computational infrastructures and analytic tools. These centers need to provide intuitive, visual interfaces for biologists with real-time interactive data analysis. These centers could also serve as training facilities and facilitate communications between scientists of diverse backgrounds, disciplines, and expertise within a common framework. These centers would also facilitate the interplay between discovery and hypothesis-driven science. Several other countries are already creating such centers.

Maintaining a leadership role is vital for the economic health of the United States. We need to maintain our leadership in HPC in order to have the advantage in intellectual property, which is connected to our economic well-being. Support for our HPC industry is vital. Countries such as Japan, China, and India are making substantial investments in HPC. We need to do the same.

The need for supporting HPC extends across all of the hard science disciplines. I hope that I have been able to show in this statement that the increased need for this support is arising from biology. A recent Rand report entitled “The Global Technology Revolution” was prepared for the National Intelligence Council. In this report it summarizes how the future will be determined by the intersection of IT and biology, and the industries such as nanomaterial, materials, and biotechnology that are spun from this intersection. Clearly, the future is in this area. We should make the investment now.

Stanley K. Burt, Ph.D.

