Computing for Cancer: the role of distributed computing in cancer research

Samuel Krausz, Meds 2009
Jordan Glicksman, Meds 2010

Computers have always been used in medicine for patient management, personal research and communication. Recently, the utilization of massively parallel processing, brought about by distributed computing grids, has allowed for cancer researchers to develop drugs and test therapies at speeds never before available. The use of distributed computing in testing molecules for their cancer protein binding potential, automated tissue microarray analysis and tumour growth modeling are all currently promising research areas. Despite some inherent challenges, distributed computing may one day be used extensively in all areas of medical research.

This article has been reviewed by Dr. Eric Wong and by Dr. Wayne Weston and by Dr. Victor Han.

The Internet may be synonymous with browsing and email, but its utility is much greater. The Internet is often used to garner information, be it a patient searching for facts regarding a recent diagnosis, or a physician trying to find the latest published research on treatment for an obscure disease. A conceptual link which is not immediately made, however, is the applicability of computers and the Internet in producing potential breakthroughs in cancer research.

Cancer research has traditionally focused on exploring epidemiology, practical experiments in molecular biology, clinical trials of potential medications or therapies and, more recently, an exploration of immunotherapy and gene therapy. It is a direct extension of cancer research’s foray into molecular biology and genetics which has ushered in the use of computers in the oncological arena.

History of distributed computing

Since the birth of computers over fifty years ago, the raw speed of computation has increased by a factor of over one million.\(^1\) Despite this increase in computational power, personal computers are still considered extremely slow in terms of requirements for solving highly complex scientific problems.\(^1\) Cancer research is one such area where the complexity of molecular and biological interactions practically precludes the use of even modern personal computers in any but the simplest experiments.

One method of solving this problem emerged in the 1980’s, and involved clustering multiple individual computers to form what was termed a ‘supercomputer’.\(^1\) While these combined computers have powers greatly exceeding those of individual units, the supercomputer must remain in a dedicated location, and both physical and economical constraints limit the system’s size.\(^1\) This static model of a supercomputer eventually evolved with the proliferation of the personal computer to ordinary individuals. With the subsequent advent of the Internet, a means of connecting individual computers located thousands of kilometers apart was feasible, and the birth of ‘distributed computing’ or ‘GRID computing’, was made. Before long, it was realized that most of the current 400 million computers’ processors sit idle for long periods of time, thus providing a potentially enormous computational resource. With the use of distributed computing, the formation of a supercomputer consisting of thousands or millions of individual computers was possible and finally ushered the way towards truly useful cancer research using computers.

Current implementations of distributed computing projects involve both a central server and client software downloaded by individual users.\(^2\) A simplified explanation of the system revolves around the server breaking apart a complex project and distributing chunks to individual clients. The client computers process these chunks, during idle processor time or when a screensaver is running, and the forward the completed work to the central server. The server combines these with other completed chunks and distributes further work to the clients.\(^3\)
While many scientific projects attempted to make use of distributed computing, cancer researchers quickly discovered the value this system could have in solving some previously insurmountable molecular biology problems.

**Searching for anti-cancer drugs**

One project, funded by the National foundation for Cancer Research and run by Oxford University and United Devices, aims to discover drugs suitable for killing cancerous cells or retarding their growth. Normally, it takes between 12-15 years from a molecule’s discovery to go through the process of effectiveness and safety testing and regulatory approval to finally reach the consumer market. Almost half this time is spent screening potentially therapeutic compounds, a time which this project hopes to reduce significantly.

In essence, the Cancer Project engages in a game of reduction and simplification. The project started with a list of over 1 billion compounds found in commercial catalogues and combinatorial chemical libraries. The list was reduced to compounds with drug-like properties, namely ones with suitable molecular masses and solubilities, which yielded a total of 35 million compounds. With this list, an exchange of various functional groups in each molecule produced 100 derivatives for each compound, producing a final collection of 3.5 billion potential anti-cancer drugs.

Twelve cancer-related proteins, with established active sites, were subsequently chosen for testing. Each of the 3.5 billion compound was to be tested against these active sites, in order to discover which would bind and potentially interfere with the given cancer protein’s method of action.

Even with the small list of proteins being investigated, the vast number of potential molecules and bindings was daunting. Moreover, in order to reduce error and improve quality and reliability, each molecule needed to be tested multiple times and ranked in terms of binding potential and binding energies. The sheer scale of the projected evolution into the largest computational chemistry project ever undertaken.

One year after the project’s launch however, over 1.5 million volunteer computers were recruited, in over 200 countries, producing a virtual supercomputer which had processed over 100,000 years worth of CPU time. Calculations were proceeding at a rate of 15,000 molecules screened every second and eventually came up with a list of 800,000 potential molecules in 2002.

While this ‘shortlist’ of molecules is only a fraction of the original 3.5 billion, it is still far from being a feasible amount testable against cancer. The project has since continued to a second phase, where distributed computing, together with software called LigandFit, will determine, and more accurately prioritize, the suitability of these molecules for actual drug development. As of this year, the second phase of the project has involved over 1.9 million volunteers and has computed almost 43 years worth of computational time, but is not yet complete.

While this project’s progress seems highly encouraging, there are still many difficulties in the approach. First, the active site on target proteins may not always be previously known. While brute-force computational searches for binding sites based on ligand-protein energy of interaction can be attempted, this is not feasible. The amount of calculations necessary for this type of search far outnumbers the capabilities of even the largest of distributed computing systems.

An efficient automated method of discovering these active sites must be developed, or else searches must be limited to small parts of the molecule instead of complete ligands. A second problem does not involve computation power, but rather social aspects of the system. As distributed computing requires large amount of volunteer computers, a given computational power cannot be guaranteed. Individuals may lose interest and uninstall the program or competing research may foster greater support and thus eliminate potential participants from the available pool. Furthermore, questions of security and viruses, of paramount importance to today’s consumers, must be addressed before widespread acceptance is achievable. There are methods of combating these risks, but they must be used effectively and transparently to the end-user. Efforts at recruitment, either through appropriate advertisement or incentives, must be considered if distributed computing is to be used successfully for future projects. Also, this type of investigation is hypothesis generating and not hypothesis testing in nature. Thus, while research such as the Cancer Project may eventually narrow down a small list of drugs to be tested, drug company involvement, governmental regulation and most importantly, in vivo testing, must all be completed before successful drugs can be produced and used.
Other distributed computing cancer projects

Distributed computing, in its fight against cancer, is not simply limited to screening potential drugs. A novel use of this technology is being developed utilizing IBM’s World Community Grid (a distributed computing environment) together with a new tool called tissue microarrays (TMA). TMA’s allow researchers to determine the specific cancer type and stage of a given tissue sample and systematically determine which therapies may be effective. A specific treatment can then be utilized based on the presence of a given biomarker. While TMA’s are arguably quite useful, their major limitation revolves around the subjective interpretation of the array by observers. By digitizing the specimens, computers can be utilized for this assessment, although the analysis is computationally very complex. While few biomarkers have been examined to date, a large database of these markers, coupled with the power of a distributed system, could allow parallel analysis of hundreds of arrays together with simultaneous experiments on them. This increased speed in computation and analysis, which would not be possible on an individual computer, could allow for the discovery of minute changes in measurable factors and thus facilitate future research in cancer biology and drug discovery.

Other cancer-related research projects utilize distributed computing in various ways. Some, like Parabon’s Compute-against-Cancer, aim to analyze patient responses to chemotherapy and thereby mitigate potentially adverse effects. Others, like Integrative Biology (IB), started in 2004, aim to develop all-encompassing models for future research. IB is a project aiming to build a secure and resilient distributed infrastructure which will be used to develop complex models of cancerous tumour growth. Eventually, it is hoped this system will allow for the understanding of cancer’s biological mechanisms completely. This system will utilize high-performance computers, large databases and complex visualization systems to achieve this goal. While the technical methods of distributed computation are similar to those used in the Cancer Project, the computer network consists solely of researchers and super-computers at specific facilities, not necessarily volunteer computational cycles provided by individuals.

It is evident that distributed, or GRID, computing holds tremendous potential in the field of cancer research. It is important to note that this technology is not limited to oncology, and has prospective use in helping develop cures for many diseases ranging from AIDS, to Alzheimer’s. The ever-increasing speed of both personal computers and the Internet will undoubtedly benefit future research utilizing this technology and may, one day, make distributed computing a method of choice in developing better therapy for a wide variety of diseases.

References