Computational Biology & Bio-Informatics

LSU Department of Computer Science

The recent mapping of the human genome was a monumental achievement both for biology and computer science. This achievement was brought about by years of cooperation between biologists and computer scientists, both at Celera Genomics and in the government sponsored Human Genome Project. Although the mapping of the human genome represents a major milestone, it certainly is not the final goal of this cooperative effort. There is still an enormous amount of work to be done in understanding the human genome, understanding the evolution of the human species, understanding our relationship with other species, etc. Because DNA sequences, RNA sequences and protein sequences are all essentially forms of information, computer science (which focuses on the manipulation, storage and organization of information) will continue to have fruitful interaction with biology in the years to come.

The Governor's Information Technology Initiative seeks to increase the competitiveness of Louisiana universities in the area of information technology. One of the sub-areas slotted for enhancement at LSU is *biological computing*. We interpret this broadly to mean any interdisciplinary endeavor on the boundary between biology and computer science. In this proposal, we focus on the research disciplines that are closer to computer science. The Department to Biology is preparing a proposal which will focus on disciplines closer to the biology side. We feel that for such an interdisciplinary effort to succeed, it is important to have a spectrum of research, with pure computer science on one end, and pure biology on the other. The hope is that by having researchers with interests across this spectrum, the communication gap between biologists and computer scientists will be bridged.

We focus on two disciplines within computer science. We describe them, and then explain their relevance to the problems faced by biologists. We then specify a plan for initiating research programs in these areas at LSU. The two disciplines are:

Algorithms & Data Structures An algorithm is a well specified finite procedure for computing some desired result or value. The design of efficient (in terms of time or space) algorithms is critical to making information technology practical. The organization of information in data structures is often critical to the design of efficient algorithms. The emphasis in this area is on the mathematical analysis of theoretical

models of computation, primarily when the theoretical analysis gives some insight to observed behavior,or is able to predict behavior.

Artificial Intelligence This area focuses on the problem of getting computers to behave in a more human fashion, and on the solution of difficult problems using heuristic (approximate) algorithms. While there are many problems in computer science that can be solved efficiently (i.e. quickly), there is a large class of problems known as *NP-hard problems*. Our current understanding is that none of these NP-hard problems can solved in a way significantly more efficient than simply trying all possible solutions. This brute force method of solution is often totally impractical. One of the main focuses of artificial intelligence is on finding approximate solutions to such problems, in a reasonable amount of time.

Both of these areas are relevant to biologists, as we shall explain.

The study of efficient algorithms for biological applications is termed *computational biology* within the computer science algorithms community. There are a large number of algorithmic problems within this area, as witnessed by the fact that several annual conferences specializing in computational biology have been established in recent years (for example Combinatorial Pattern Matching (CPM) & Research in Computational Biology (RECOMB)). An example of such an algorithmic problem is the longest common subsequence problem. A string of characters β is a subsequence of another string of characters α if some characters of α can be deleted to obtain β. For instance, 'ace' is a subsequence of 'abcde'. The longest common subsequence of two strings of characters α and β is the longest string γ which is a sub-sequence of both α and β. In some sense, the longest common subsequence measures the similarity between two strings. The longest common subsequence problem was studied by computer scientists in the algorithms community as early as the 1970's, the result being a number of efficient algorithms for its solution. This problem is closely related to certain types of sequence alignment problems which arise in biology, where one wishes to compare two (or more) strands of DNA. Much work has been done on finding efficient solutions to such string comparison problems arising from biology, but there is still much to be done.

There has also been a great deal of work on biological problems within the artificial intelligence community. It turns out that a number of problems which biologists would like to solve are NP-hard. An example is the following: Suppose we are given the DNA sequences for several different species. We wish to create an evolutionary tree which relates the species, in which

in some sense is most likely, given some probabilistic model of how DNA changes during evolution. There has been a great deal of work on finding such trees, but since finding them often turns out to be computationally difficult, we often are will to settle for finding a tree which is close to the most likely tree. Techniques developed in the artificial intelligence community are extremely useful in this context.

To initiate research in these areas, we recommend that 4 new faculty members be hired in the above areas. These faculty would have partial appointments in both computer science and biology. To ensure that both the interests of computer science and biology are met, we recommend a hiring process involving a joint biology/computer science committee. These hirings will put LSU on the map in terms of computational biology, and serve to foster future cooperation between computer science and biology at LSU.