DATA MINING OF INDIVIDUAL HUMAN DNA SEQUENCES

Advances in biotechnology have spawned a new scientific field merging biology and computer science, known as bioinformatics. Roughly speaking, bioinformatics is the use of statistical and data mining techniques on large data sets from the biological sciences, primarily which arise from the DNA sequences of various organisms. DNA is itself a long chainlike molecule consisting of a sequence of pieces called nucleotides. Naturally, of particular interest is the human genome, or DNA sequence, which was the focus of an international public undertaking called the Human Genome Project (HGP), that completed its first final draft of the sequence in 2003. In 2006 Human Genome Project researchers announced the completion of a detailed and vetted specification of the DNA sequence for all 24 human chromosomes. (The intermediate three years were needed for the correction of some significant errors and the completion of gaps.) A parallel endeavor was run by the private company Celera Genomics, headed by scientist and businessman J. Craig Venter. Celera continues to keep some of its information proprietary. The public undertaking, run by the National Human Genome Research Institute (NHGRI) along with multiple other research institutes in the US Department of Energy (DOE) and NIH and various institutions abroad determined the roughly 99.9% of the 3 billion nucleotide sequence common to all human beings, to an accuracy of over 99.99% per nucleotide.

Both the international and private sector projects were conducted using DNA from numerous sources of a racially diverse background. The many small segments of DNA that vary among individuals, known as polymorphisms, were also identified by the HGP using statistical techniques to compare amongst sources. Preexisting data mining techniques have been essential in gaining insight into the various physiological effects of these polymorphisms. It turns out that most polymorphisms consist of singly nucleotide deviations, and are actually without physiological effect. Interestingly, only a minority of the 0.01% of DNA that varies between individuals contributes to the physiological and racial diversity of humanity. Evidently, an even smaller minority of polymorphisms affect or determine an individual’s susceptibility to various diseases and genetic disorders.

One of the main analyses of the universal genome sequence is the ongoing effort to determine what subsequences correspond to the production of particular proteins and expression of particular genes. There are two main approaches to this gene finding endeavor which are currently in widespread use. One is exemplified by a program known as Gene Recognition and Assembly Internet Link (GRAIL), which employs a neural network model that has been trained on known genes and is then used to predict which sequences may correspond to as yet unknown ones. The most recent incarnations, such as GRAIL-EXP, incorporate measures that are sensitive to context dependent information, such as techniques for pattern matching to specific databases. The second approach to gene finding employs a hidden Markov model. The model defines a set of states corresponding to short DNA sequences (such as nucleotide triplets known as codons). A long DNA sequence is then used as input to drive a
succession of state transitions, with the probability of a particular transition having been established by training data that reflects the regularity of known genes.\footnote{More detailed descriptions of both analyses can be found in MITRE, 2000.}

Of particularly urgent interest is the identification of sequences which either suppress or contribute to the development of various forms of cancer. As of 2001, the inventory of known cancer causing “oncogenes” included more than 100 dominant and 30 recessive members, and those numbers are certainly enlarging. The search for oncogenes and conversely tumor-suppressing genes is largely guided by biological theory resulting from years of cancer research prior to the availability of genome sequencing. As is the case with the identification of genome alterations that result in known genetic disorders, the identification of cancer related genes could be greatly benefitted by amassing the numerous genome sequences of similarly effected individuals. Statistical methods could then be used for cross comparison to unaffected individuals.

With each additional experimental validation of a predicted gene, the information available for use in the next gene finding procedure increases. The use of data crunching in this manner promises many future discoveries, and has already yielded a great deal of useful information in the way of gene identification. Notably, while these statistical algorithms are widely used by biologists, knowledge of the actual methodologies employed by the statistical algorithms is quite limited. Since these statistical methods are not greatly informed by biologists’ knowledge of the mechanisms of gene expression, it remains a promising challenge to find talented individuals with expert knowledge of both the relevant biology and statistics.

Analyses of larger subsections of DNA sequences are also highly useful in cross comparisons between the DNA sequences of different people and even different organisms. In a recent book describing his work on the HGP as director of the NHGRI, Francis Collins describes how a large branch for the mammalian kingdom within the evolutionary tree can be deduced purely from comparing DNA sequences.\footnote{A visualization of this branch can be found in Collins, p. 128} These techniques offer promising confirmations of various theories of evolutionary dissension based on fossil records and in many cases may be useful for accurately choosing between competing theories. More subtle but similar DNA comparison techniques are increasingly utilized by population geneticists to determine the evolutionary history of the various races of people. According to Collins, it is already clear from the study of DNA variations that all humans descend from a common set of about 10,000 founders who lived 100 to 150 thousand years ago.

A very common algorithm for DNA sequence comparison is called BLAST, an acronym for Basic Local Alignment Search Tool, which over years of widespread use has undergone many refinements since its initial development in 1990. BLAST is probably the most widely used piece of bioinformatics software. Given a query sequence, the algorithm compares it against all entries in the available sequence database, aligning against a test sequence while allowing for the possibility of insertions or deletions of small segments. Matches are recorded and annotated along with an estimation of the likelihood that such similarity could have merely occurred by chance.

Perhaps the most exciting aspect for the future of bioinformatics is the continuing series of successes in the aim of making genome sequencing a reasonably inexpensive endeavor. The initial draft of the DNA sequence of the human genome cost an estimated $300 million, though estimates for the
cost of the final draft (including the cost of all technology that made the endeavor possible) are as high as $3 billion. Yet by February of 2006, genome scientists completed a draft of the genome sequence of the rhesus macaque for $22 million, and by the end of the same year another full mammalian genome sequence was expected to be completed at an estimated cost of only $100,000. New technology for sequencing is advancing so quickly that many believe that a $1,000 price tag is not unrealistic for the near future. As early as 2003, the J. Craig Venter Science Foundation announced a $500,000 prize as an award for the achievement. The ante has since been upped by various foundations, but even the $70 million put up by the NIH pales in comparison to the financial reward that would be reaped by any company with the capacity to market such a service.

The effect that this fast approaching technology will have on medical practices is hard to overstate. Potentially millions of individuals would pay such a low price to sequence their own or their children’s DNA, provided this could also come with an analysis of the presence of known cancer or disorder causing genes. The costs that could be saved by preventative measures targeting the eventual symptoms of genetic disorders such as diabetes, or cancers of numerous sorts, support the argument that insurance companies may be willing to cover the cost of such tests, especially for newborns and children. Naturally, the security of such medical information is an important issue that many privacy advocates have already started to address publicly. With such a technology, it is conceivable that governments, whether in the US or elsewhere, may compel their citizens to undergo such testing or even submit their information to national databases. As recently as February, 2008 CNN reports that the FBI is “gearing up to create a massive computer database of people’s physical characteristics” for the purposes of identifying criminals. But you don’t have to be a criminal to enter the database: more than half of the checks the FBI runs involve criminal background checks for people applying for jobs in government or in services involving children, the elderly, or other vulnerable persons. Is a DNA sequence database far behind? The existence of such a national database raises many constitutional questions in the US, but the undeniable and universal benefit to law enforcement makes it highly probable that such a regime will be adopted somewhere in the world, even if not here.

Increasing and improving the applications of statistical and data mining techniques in this area is sure to be a tremendous focus in the future. Indeed, national databases of residents’ DNA sequences represent an application of data mining at some of the largest scales ever attempted. It is fascinating to think how the aggregation of such information could lead, through computerized statistical analyses, to tremendous new insights into the as yet unknown connections between our gene sequences and characteristics such as rare diseases and immunities, other physical traits, or even our personalities. The limits to our current computational power do not seem to be an impediment, especially as many algorithms such as those used in GRAIL or BLAST show a high potential for parallelization, i.e. the possibility to be greatly accelerated by the use of multiple (parallel) processors, as many researchers have already begun to exploit to great effect. We may only have to wait a few years to get a taste of what amazing changes the future DNA sequencing holds for us.
NOTES


