Mathematics in the deluge of DNA data

A more efficient way of decoding DNA information using spaced seeds

Introduction

Human beings are evolutionary products of nature and environment. Here, the nature is the DNA and the environment includes food, air and many other inputs to individuals. To fully understand ourselves, we need to know our DNA, a string of three billion nucleotides of four kinds, abbreviated as A, G, C, and T. Human cells contain two copies of the DNA, one inherited from the father and one from the mother. The whole DNA (that is genome) is broken up into 23 "modules", called chromosomes, and within each chromosome are sparsely scattered genes. Each cell uses the genes to determine which proteins should be made to perform specific cellular functions (such as metabolic reactions and transducing chemical signals) at a particular time point. The Human Genome Project (http://www.genome.gov) spent $2.7 billion on completing the first draft of the human genome from 1991 to early 2000.

Challenges In Decoding DNA Information

Decoding DNA information creates formidable computational challenges in data storage, transfer and analysis. Thirty years ago, there were little DNA sequences and computers were slow. Presently, computational speed is improving but at a rate much less than sequencing machines. Under Moore's Law, computer processors doubled in speed every 18 months. In comparison, between 2008 and 2014, DNA sequence data has grown about three- to fivefold per year.

Change (mutation) in DNA never stops in living organisms. A mutation may delete a nucleotide, insert a nucleotide, or replace one nucleotide by another. Large-scale mutations also occur, which lead to one segment of DNA duplicated or relocated at a different position, from time to time. Mutation leads to gain and loss of genes and can affect the fate of a living organism during evolution.

Since DNA has a hidden structure, it is traditionally analysed in a comparative manner. For example, when the genomes of different species are compared, one will identify the conserved regions where different species have almost identical nucleotide sequences and the volatile regions where each species has a different nucleotide composition. Conserved regions are likely genes and other interesting functional units that a species cannot afford to lose. Therefore, the sequences of a conserved region in different species which descend from a common ancestor are called homologs. Since convergent evolution that leads different genes to have the same sequence occurs rarely, highly conserved regions are sought to infer genes or other interesting functional units.

Mathematics Behind Homolog Search Tools

When genomes are compared, most regions are dissimilar. Any exact algorithm for homolog search is time-consuming, as it wastes a huge amount of time in dissimilar regions. Consider human and mouse genomes, each having over three billion nucleotides. An exact algorithm may take years to compare them!

Presently, widely-used computer programs for homolog search are all designed based on the so-called filtration approach. To determine potential locations of conserved regions in the query DNA sequences, a filtration-based program first records down the occurrences of all 4^ℓ words of a fixed length ℓ (say 11) in each sequence. A pair of positions *(k, j)* is called a *seed hit* if there is a word of length ℓ that occurs in the *k*-th position in one sequence and the *j*-th position in another. The program then extends each seed hit in both directions to look for long conserved regions. Since the number of seed hits reported in the first step is significantly small, compared to all possible pairs of positions, a filtration-based program improves in speed by an order. Such a program usually takes 10 to 20 days to compare human and mouse genomes on a desktop computer.

What about the performance of the filtration-based programs? Clearly, the filtration approach sacrifices sensitivity for speed. The larger the word length ℓ is, the less seed hits are reported in the first place and hence the more homologous sequences are likely missing.

Seed Selection

Next, how to select a seed? The BLAST program (http://www.ncbi.nlm.nih.gov/) is the official homolog search program for searching DNA and protein sequence databases at the National Center for Biotechnology Information, USA. When it was launched in 1990, the BLAST used a perfect word match of length 11 as a seed hit, meaning 11 consecutive identical nucleotides found in the two input sequences will trigger the extension process. Requesting 11 consecutive base matches seems a good choice. However, mathematics analysis suggests otherwise! That is, it is much better to use identical nucleotides in a set of non-consecutive positions, called a spaced seed, to trigger the extension process. For example, we may request to have identical nucleotides in the 1st, 2nd, 4th and 6th positions in a window of length 6. Such a spaced seed is written

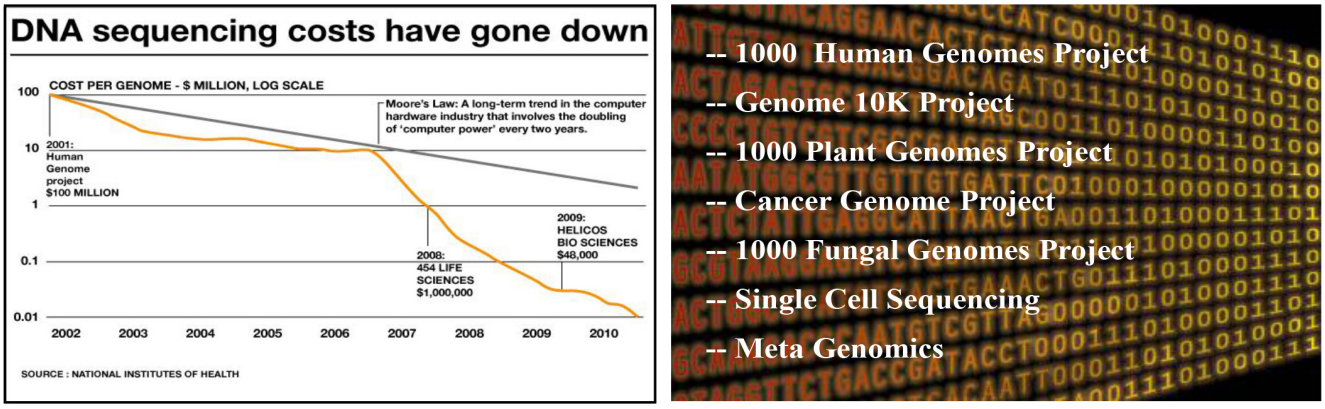


Figure 1: DNA sequencing costs.

11*1*1, where 1s and *s denote match and don't care positions, respectively.

In homolog search, we identify regions that are highly conserved, meaning nucleotides are identical in the most corresponding positions in the input genomes and what nucleotides in which positions is less important. Representing a match position by 1 and mismatch position by 0, we simply model a conserved region as a 0-1 random sequence in which 1 appears independently in each position with a specified probability p, where p is called the model parameter, corresponding to the sequence identity and being in the range from 60% to 70% in protein coding genes.

Under this simple stochastic model, any spaced seed with w match positions (called seed weight) hits a position with probability p^w in a random sequence, where p is the model parameter. Therefore, all spaced seeds of the same weight have the same mean number of hits in a 0-1 random sequence of L, p^wL. This sounds like it does not matter which seed is used. However, in the nature of homology search, two overlapping seed hits lead to the same extension. Therefore, the number of disjoint hits makes much more sense in choosing a seed.

Indeed, the consecutive seed $C=\overbrace{111\ldots1}^{w}$ has a smaller mean number of disjoint hits than almost all the spaced seeds of the same weight. In a 0-1 random sequence with model parameter p, the mean distance between one hit and the next is $1/p^w$ for C. But, an occurrence of C would provide a huge "head start" to the next occurrence. If the next bit is 0, then the next occurrence is disjoint from the current one but, if a bit 1, a new overlapping hit has already taken place. If X_C denotes the mean distance between two consecutive disjoint hits of C, by the law of total probability,

$$(1-p)(1+X_C) + p = 1/p^w,$$

or, equivalently,

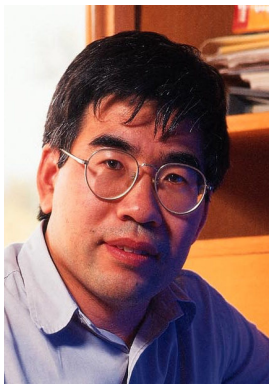
$$X_C = (1-p^{w+1})/(p^w(1-p))-1.$$

When p = 0.7 and w = 11, X_C =167.58. This suggests on the average, there is a hit of C in a window of size 167.58 in a long random sequence.

Because of the complex relationship between overlapping hits for a spaced seed S, the mean distance X_S between two consecutive disjoint hits of S does not have a simple closed formula. For S=111*1**1*1**11*111, X_S is equal to the ratio of the determinants of two 128 by 128 matrices. Computation shows that X_S is about 108.6 when p = 0.7. This implies that S has 30% more disjoint hits than the consecutive seed C in a long random 0-1 sequence with model parameter p = 0.7. Although genomic sequences do not quite fit the simple model used for above reasoning, using the spaced seed, the PatternHunter program hit 50% more homologous regions than the BLAST program with the consecutive seed of the same weight when the DNA sequences of 70% identity were compared.

The transition from consecutive seed to optimised spaced seed takes 10 years. Presently, the spaced seed has been a standard technique for designing computer programs for homolog search, genome alignment, and mapping short read into a reference genome. Applying mathematics to understand DNA is an essential first step in uncovering the molecular mechanisms behind human diseases.

ZHANG Louxin received his Ph.D. from the University of Waterloo and is currently an associate professor at the Department of Mathematics, NUS. His research mainly focuses on building mathematical models and designing algorithms that allow biologists to better analyse DNA sequences and to decipher gene regulation.



References

[1] J Ma, B., Tromp, J., Li, M., "PatternHunter – faster and more sensitive homology search", *Bioinformatics*, 18, 440-445, 2002.

[2] Zhang, L.X., "Superiority of spaced seeds for homology search", *IEEE-ACM Trans. Comput. Biol. Bioinform.*, 4, 496-505, 2007.

[3] Tran, N.H., Choi, K.P., Zhang, L.X., "Counting motifs in the human interactome", *Nature Commun.*, 4, art. no. 2241, doi:10.1038/ncomms3241.