BIOINFORMATICS

Filling the gap in gene prediction

Genome sequencing tells us the order of the As, Cs, Gs and Ts in a genome, but annotation programs make sense of it all by telling us where genes are and what they look like. Gene-finding programs are reasonably good at identifying protein-coding regions, but are less proficient at finding other potentially important sequences — such as *cis*-regulatory regions and non-coding exons — that lie upstream of the translational start site. Now, Davuluri and colleagues have filled this technical gap by developing a program that accurately recognizes promoters and first exons. Although the program was developed to annotate the human genome, the authors believe it will also prove useful for the genomes of other species.

The starting point in constructing any sequence prediction program involves 'training' the algorithm to recognize the type of sequence you want. Because most sequence annotations do not contain information about 5′ untranslated regions, the authors constructed their own data set of more than 2,000 genes for which first exons and promoters had been experimentally validated. Using these sequences, the algorithm 'learned' to recognize features ~500 bp either side of the first exon — defined as the region between a promoter and the first splice-donor site. The program — called first-exon finder or FirstEF – operates by finding every potential promoter and splice-donor site and then calculating the probability that the intervening sequence is a first exon. The power of FirstEF lies in its abilty to identify first exons that are associated with either CpG-rich or CpG-poor promoters, and to predict both coding and non-coding first exons. Two tests confirm the accuracy of FirstEF. When the algorithm was trained on 90% of the gene data set and then tested on the remaining 10%, it correctly predicted 84% of first exons. Its performance on the annotated genomic sequences of human chromosomes 21 and 22 (from the public consortium) was also quite impressive, whether it was asked to confirm experimentally validated first exons or to localize promoters upstream of annotated genes. FirstEF is the first and the only computational tool available at present that can predict first exons, especially non-coding ones.

The effort of annotating the human genome is likely to continue for many more years, but FirstEF has brought bioinformatics one step closer to its goal of defining the 5′ boundaries and non-coding regions of genes. Notably, FirstEF has estimated the percentage of CpG-related first exons to be 70%, and not 50% as was previously believed. And, if you like a challenge, the authors have made FirstEF's predictions — all 68,645 of them — from the working draft of the human genome available for scrutiny.

Tanita Casci

References and links

ORIGINAL RESEARCH PAPER Davuluri, R. V. *et al*. Computational identification of promoters and first exons in the human genome. *Nature Genet*. **29**, 412–417 (2001)

WEB SITE

Michael Zhang's lab: http://www.cshl.org/public/SCIENCE/zhang.html

IN BRIEF

MULTIFACTORIAL GENETICS

Identification of wound healing/regeneration quantitative trait loci (QTL) at multiple time points that explain seventy percent of variance in (MRL/MpJ and SJL/J) mice F_2 population.

Masinde, G. L. *et al. Genome Res*. **11**, 2027–2033 (2001)

To identify key genes involved in wound healing, Masinde *et al.* crossed two mouse strains together that have markedly different wound-healing rates. They then carried out genome-wide scans in F_2 mice at different time points after ear punching and found ten wound-healing quantitative trait loci (QTL), eight of which are new. Some QTL are involved at all stages of wound healing, and epistatic interactions between loci also occur.

COMPARATIVE GENOMICS

Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order.

Liu, H. *et al. Genome Res.* **11**, 2020–2026 (2001)

Arabidopsis and rice are distantly related, but their genomic similarities could help in transferring information, such as gene location, from the smaller, sequenced *Arabidopsis* genome to the larger rice genome. To investigate the degree of conserved gene order between these two species, Liu *et al.* compared 126 annotated and mapped rice BAC sequences with the *Arabidopsis* sequence. Gene order was conserved in some regions, but these are quite small, indicating that the *Arabidopsis* genome might not be useful for assembling the rice genome.

HUMAN GENETICS

Type 2 diabetes and three *Calpain-10* gene polymorphisms in Samoans: no evidence of association.

Tsai, H.-J. *et al. Am. J. Hum. Genet*. **69**, 1236–1244 (2001)

In 2000, three calpain-10 (*CAPN10*) single nucleotide polymorphisms (SNPs) were found to be associated with increased susceptibility to type 2 diabetes in Mexican Americans. Tsai *et al.* have now tested the association of these SNPs with the disease in Samoans, among whom it is highly prevalent. No association was found, perhaps because *CAPN10* is a susceptibility gene only in certain ethnic groups or because their sample size was too small to detect its effects.

TECHNOLOGY

Transgenic DNA introgressed into traditional maize landraces in Oaxaca, Mexico.

Quist, D. & Chapela, I. H. *Nature* **414**, 541–543 (2001)

Transgenic DNA has been found in wild strains of maize in the Oaxaca region of Mexico. Despite a moratorium in Mexico on the planting of transgenic maize since 1998, these authors have found clear evidence of transgenic DNA in 5 of 7 landraces tested, which they believe was introduced multiple times. This finding has implications for the maintenance of food and crop diversity.