COGANGS - Comparative Genomics and Next Generation Sequencing

Executive summary

The COGANGS project is about developing a software suite, the COGANGS engine, where a large number of genomes can be used as knowledge input in gene regulation analysis – e.g. analysis of which factors influence gene regulation, how much impact they have on gene regulation, how they can be identified in the genome of interest, how different gene regulation factors influence each other, and how they work in combination. The software is able to provide completely new knowledge, and thus has tremendous value to life science researchers globally, including pharmaceutical companies, biotech companies, agricultural companies, biofuel companies, and research hospitals, as well as universities and governmental research organizations.

Challenge overview

In the last few years, new revolutionary technologies have been developed for DNA sequencing. These methods, commonly termed High-throughput sequencing (HTS), are significantly faster and much cheaper than the traditional sequencing technology (Sanger sequencing). The speed of sequencing, and thus the volume of data being generated are approx. 10,000 times faster than just a few years ago. The price per sequenced DNA base has been reduced accordingly.

These new technologies will in the near future create a vast amount of sequence data that can be exploited, among other things, to discover genetic causes of diseases and traits. Notably, the recently launched 1,000 Genomes project was started with the aim of sequencing more than 1,000 human beings, and very recently, a 10,000 genome project was announced (Genome 10K), with the aim of sequencing the full genomes of 10,000 different species.

Implementation of the initiative

The COGANGS project has been implemented in an international collaboration, in which private bioinformatics software developing companies and academic research groups worked together. The academic groups were asked to develop efficient algorithms for large scale data analysis, and the software writing companies finished the implementation of the software packages, tailoring them to the need of the market. The private companies were CLC bio A/S, BIOBASE GMBH, deCODE genetics and Novel Computing Systems in Biology LLC, and the academic institutes were the Rényi Institute and the University of Oxford. The project has been financed by FP7.

The problem

A genome is a very long string over the alphabet A, C, G and T. Part of the genome does not code any information, the remaining part of the genome encodes the proteins and different RNA sequences that describe the biochemical system of the organism. The key problem is to identify the functional motifs in the genomic sequence called Transcription Factor Binding Sites (or TFBSs for short), and to do this for the vast amount of genomes coming from the HTS projects.

This is a non-trivial bioinformatics task, that cannot be done in silico with 100% accuracy. The TFBSs are typically 6-12 nucleotide long sequences, and although less variable than their surrounding 'junk' DNA, still have considerable variation. However, the current statistical methods can be improved by improving the background models used in the statistical learning algorithms, and by using comparative methods. The basic principle of the comparative bioinformatics approach is that the function (in this case the TFBSs) is more conserved than the sequences themselves, and thus, if we compare the genomes of closely related species, then we might improve the accuracy of genome annotation.

The ideal solution to the problem takes one input genome, searches its related genomes from a database, compare the input genome to the closely related genomes, and use a combination of comparative and ab initio methods to predict where the TFBSs are in the given genome.

Results and achievements

The COGANGS consortium has implemented the COGANGS engine software package. The package has the following modules:

- TransFoot: this package implements a sophisticated statitical learning algorithm which is based on the fusion of two stochastic models. The first model is a continuous time Markov model that models the main mutations in the genomic sequences, namely, substitutions, insertions and deletions. The second model is a Hidden Markov Model that models transcription factors and 'junk' DNA. In the joint model, the ancestor of a set of sequences is generated by the Hidden Markov Model, and then this sequence evolves on a phylogenetic tree, according to the continuous time Markov model.

  The input of the package is a set of related sequences. The program takes these sequences, generates an evolutionary tree based on the estimated evolutionary distances of the sequences, then outputs a Bayesian estimation where the TFBSs are in the given sequences. Markov chain Monte Carlo methods are applied for sampling from the Bayesian distribution, and the Bayesian estimation is given based on these samples.

  The package uses the knowledge in the TRANSFAC database to build the Hidden Markov Models modelling the TFBSs.

- The phylogenetic segmentation package is an implementation of the so-called k-spannoid building algorithms. The k-spannoids are generalizations of evolutionary and spanning trees. They reasonably approximate phylogenetic trees while the computational complexity of many bioinformatics algorithms are much better on k-spannoids then on evolutionary trees. The package can be used when the number of related sequences is so big that the running time of algorithms using phylogenetic trees would be too long.
- The F-Match portal is a webservice for identifying statistically over-represented TFBSs in a set of sequences compared against a control set, assuming a binomial distribution of TFBS frequency. The program reads FASTA DNA sequence entries for the query and control sets. F-Match uses the Match algorithm and a library of positional weight matrices from TRANSFAC6.0 to scan the input sequences in the two sets for putative TFBSs then compares the experimental and control sets to identify the statistically over-represented TFBSs. The portal is publicly available from http://www.gene-regulation.com/pub/programs.html#fmatch

Lessons learned and replicability

The joint work between academy and private companies were very successful, however, it turned out during the implementation of the project that the time needed to implement a final software from prototypes might be easily underestimated. The project has been finished in 28 months instead of the predicted 24 months.

Contacts, references

Rényi Instiute, Hungarian Academy of Sciences,
H-1053 Budapest
Reáltanoda utca 13-15
Hungary

Project homepage:
http://www.2020-horizon.com/COGANGS-Comparative-Genomics-and-Next-Generation-Sequencing(COGANGS)-s6593.html