CODE: Comparative Genomics of Disease Genes.

Pedro Cruz[1 2], Vincenza Maselli[1], Remo Sanges[1], Elia Stupka[1]

[1] **TIGEM** (Telethon Institute of Genetics and Medicine) Via Pietro Castellino 111, 80131, NAPOLI, Italy
[2] cruz@tigem.it

Keywords: comparative genomics, disease genes, non-coding sequences, genomics

Abstract

The CODE project aims to provide a comparative genomics analysis of curated disease genes families. It integrates experimental and human verified information into automated gene-centric pipelines, which regularly map disease genes and related features across available sequenced metazoan genomes. Of particular interest to the outcome of the project is the semi-automated annotation of non-coding sequences (ncRNA, promoters, enhancers and splice regulators). Considerable attention is paid to the evolutionary clues provided by the analysis in particular when model animals are concerned. Finally, the establishment of a community portal, complementary to the existent international projects, will disseminate the results of the research and augment the annotation of disease genes.

Introduction

With the increase in available genomes from metazoan organisms, the ability to assign functions, predict evolution and study the regulation of conserved genes has been greatly enhanced. The tools and approaches used in large international projects have focused for a long time on wide genome annotation or painstaking annotation based on publications. The CODE project aims at contextualizing the information available from a wide variety of sources, develop algorithms to predict and understand the role of non-coding sequences, dynamically integrate disease information and mutational data arising from several individual projects.

Development of CODE

Implementation and Integration

The CODE project is based on complementary, widely supported, open source community projects such as BioPerl [1] and Ensembl [2]. Figure 1 gives a conceptual framework for the data integration aspects in the CODE project. More than providing contextual linking between the several informational resources, the CODE provides validation and allows for standard or innovative algorithms to be applied independently of formats.

Approach and Validation

The project is based on the insights provided by the studies conducted on individual genes over the last couple of years. On a first stage all locus information for a particular gene is gathered, and through concurrent processes of automated (using information present in Ensembl, Swiss-Prot, Genbank, SMART, etc) and manual analysis of relevant publications, families of genes are validated and all related information is stored. An evolutionary analysis is performed resulting in a phylogenic view of the gene family. The structure of both genes and proteins (domains, functional regions, sequence) are analyzed to obtain a consensus framework, which characterizes the family. Flaking regions are scouted for non-coding conserved elements and their putative role determined as far as possible. Information potentially relevant for disease is contextualized at this point, including mutations, SNP's, known gene isoforms. This information is then used to search the less annotated genomes. The families looked at initially are those of close expertise, with validation of results obtained from laboratory experiments and manual curation.

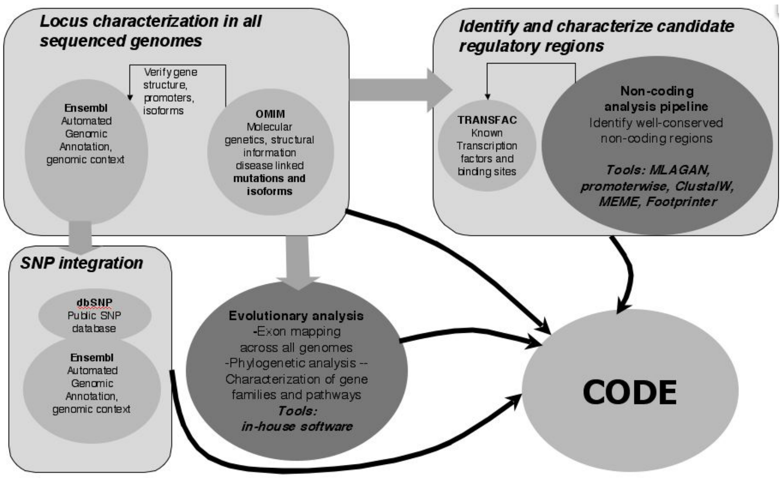
Locus characterization in all sequenced genomes
Verify gene structure, promoters, isoforms
Ensembl Automated Genomic Annotation, genomic context
OMIM Molecular genetics, structural information disease linked mutations and isoforms
Identify and characterize candidate regulatory regions
TRANSFAC Known Transcription factors and binding sites
Non-coding analysis pipeline Identify well-conserved non-coding regions
Tools: MLAGAN, promoterwise, ClustalW, MEME, Footprinter
SNP integration
dbSNP Public SNP database
Ensembl Automated Genomic Annotation, genomic context
Evolutionary analysis -Exon mapping across all genomes -Phylogenetic analysis -- Characterization of gene families and pathways
Tools: in-house software
CODE

Figure 1. A conceptual framework for the data integration aspects in the CODE project

Conclusion

As the project evolves by iterative stages and through a variety of gene families of medical interest and for which good quality information is available, it will certainly become a useful tool for researchers working on disease genes and model organisms.

Acknowledgments

This project is supported by a grant from the Telethon Foundation.

References

[1] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002: The Bioperl toolkit: Perl modules for the life sciences.
Genome Res. 2002 Oct;12(10):1611-8.

[2] Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E. Ensembl 2002: accommodating comparative genomics Nucleic Acids Res. 31(1):38-42 (2003)