EMLIB: a C++ Library to manage transcripts and genomic variations

Cereda M[1,2], Sironi M[1], Pozzoli U[1]

[1] Bioinformatics Lab, Scientific Institute IRCCS E.Medea, Bosisio Parini (LC)
[2] Department of Theoretical Physics, University of Torino, Torino

Motivation

The last decade has seen an explosive growth in biological data related both to multiple species and to multiple individuals. The Genomic Era has produced immense quantities of nucleotide sequences. Massive resequencing projects are now increasing and changing on a daily basis these informations through the identification of sequence variations within distinct genomes. With all these data at hand the need for remarkable computing resources has become extremely important, and the design of flexible and reusable software into libraries has become the benchmark for a more effective analysis. Nevertheless, in order to fathom the complexity of entire genomes the software must also offer the possibility to manipulate data in a straightforward and easy way. The design of a library which allows programmers to handle sequence data and annotations, to assess the effects of genomic variations and to perform large scale analysis is a brand-new challenge in the sea of bioinformatics tools. Another important issue is the demand for a library which would not stick the user to source-specific format(s) of the data. In this work, we propose EMLIB, a C++ library able to retrieve gene informations and build objects representing transcripts or groups of transcripts (i.e. genes) independently from the data source. Unlike other already developed C++ libraries, our attention is particularly directed toward the characterization of genomic variations. In our opinion this feature will become crucial in the near future: the dramatic growth of resequencing projects will require methods able to deal with individual sequences making the notion of “reference sequence” obsolete.

Methods

For an efficient, convenient and faster analysis the most suitable programming language is C++. In order to be more effective and manage the complexity which arises in the design of large and complex software we choose an object-oriented approach. In order to achieve independence from data sources we left the classes constructors undefined. We provided a list of base classes describing mutations, transcripts and features respectively; most of these are built as heterogeneous containers. The core of the project is represented by two base classes, em_transcript and em_variation, which are designed to store the corresponding biological data: transcripts and sequence variations (base substitutions, insertions and deletions). These two classes are tightly connected allowing the user to apply variation(s) to a given transcript (or group of transcripts). The library also contains classes to manage informations about groups of both transcripts and mutations. In addiction, it provides classes to stock and query position-dependent informations that can be used to characterize transcripts features.

Results

We created a novel hierarchy of classes useful to define transcripts, to manage sequence alterations and to calculate position-specific quantitative features in a “variation dependent” way. Algorithms built upon this class hierarchy can easily be used to measure or estimate quantitative differences between distinct “versions” of the same transcript or gene (i.e. mutations, SNPs alleles, haplotypes). A number of ready-to-use classes are helpful to calculate and store quantitative information along the sequences like, for example, Positional Weight Matrices (PWMs) or frequency counts. As an instance, PWM classes can be used to represent splice site strength, possible transcription factor targets, miRNA targets and so on. When a transcript object is “mutated” all its features are coherently re-calculated making very easy and straightforward to quantify the effect of variations on such characteristics. Finally, it is quite easy to derive new classes representing features not yet implemented in our library. EMLIB provides an intuitive and powerful environment to gain insights about the effect of genomic variations.

Contact : matteo.cereda@bp.lnf.it