Bioinformatics of sequence variation

Lehväslaiho, Heikki

EMBL Outstation, European Bioinformatics Institute,
Wellcome Trust Genome Campus, Hinxton, Hinxton Hall,
Cambridge, CB10 1SD, United Kingdom
heikki@ebi.ac.uk

Mutations have been recorded as long as modern genetics has existed. This long history has caused fuzzy and variable usage of terminology. Recently, the technological advances in disease gene characterization have prompted human geneticists to form consortiums to collect and maintain lists of gene specific mutations, each defining its own terms. The HUGO Mutation Database Initiative was set up to unify the approaches [1]. Large scale efforts of creating Single Nucleotide Polymorphisms (SNPs) are the most recent addition to the scene.

I will describe an approach to collect all publicly available sources of sequence variation and define common data structures that allows unified querying and viewing of data [2]. The tool used is Thure Etzold's SRS [3] which has the advantages of a programming language suited for describing semi-structured text files, ICARUS, and multiple levels of abstraction of data to attain uniform data structure for querying. The ease of handling databases has enabled me to simultaneously analyze the data well enough to give feedback[4] and offer a service to the community [5].

Currently, about 20,000 non-redundant mutations from single locus databases, 4000 SNPs and over 30,000 nucleotide level and over 13,000 amino acid level variants are included in the system. I will describe the data structures identified and the on-going processes of cross-referencing and data validation.

[1] Cotton RGH, McKusick VA, Scriver CR (1998) *Science* **279**, 10–11.
http://ariel.ucs.unimelb.edu.au/~cotton/mdi.htm

[2] Lehväslaiho H, Ashburner M, Etzold T (1998) *Trends Genet* **14**, 205–206.
http://www2.ebi.ac.uk/mutations/publications/TIG98.html

[3] Etzold T, Ulyanov A, Argos P (1995) *Methods Enzymol* **266**, 114–128.

[4] Scriver CR, Nowacki PM, Lehväslaiho H (1999) *Human Mutation* **13**, 334–350.

[5] http://www2.ebi.ac.uk/mutations/, http://srs.ebi.ac.uk/