HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research

Attimonelli M.(1), Accetturo M.(2), Santamaria M.(1), Lascaro D.(1), Scioscia G.(3), Pappadà G.(3), Tommaseo-Ponzetta M.(2)

(1) Department of Biochemistry and Molecular Biology, University of Bari, Bari (2) Department of Zoology, University of Bari, Bari (3) Java Technology Center, IBM Semea Sud , Bari

Motivation

Population genetics studies based on the analysis of mtDNA and mitochondrial disease studies have produced a huge quantity of sequence data and related information. These data, classified as RFLPs, mtDNA SNPs, pathogenic mutations, HVS1 and HVS2 sequences, and complete mtDNA sequences, are distributed in databases differently organised:: MITOMAP [1], HVRBASE [2], mtSNPs [3] and mtDB [4]. The two latter databases more or less report frequency data associated with the mitochondrial SNPs, while MITOMAP simply associates the mtSNP to the different phenotypes. HmtDB, stores human complete mitochondrial genomes annotated with variability data estimated through the application of specific algorithms implemented in an automatically running Variability Generation Work Flow (VGWF). Another Work Flow, called Classification Work Flow (CWF), is implemented to perform the automatic classification of newly sequenced genomes. The aims of HmtDB are (1) to collect and integrate all human mitochondrial genomes publicly available, (2) to produce and provide the scientific community with site-specific nucleotidic and aminoacidic variability data estimated on all available human mitochondrial genome sequences through the automatic application of VGWF, (3) to allow researchers to analyse their own complete or partial mitochondrial genomes in order to automatically detect the nucleotidic variants respect to the revised Cambridge Reference Sequence (rCRS) and to predict their haplogroup paternity. At present, 1255 genomes classified according to their continental origin are stored in HmtDB.

Methods

HmtDB Workflows The execution of both CWF and VGWF requires the application to multialigned sequences of the variability estimation bioinformatics methods SiteVar [5] and SiteVarProt [6], running on DNA and Protein sequences revised in collaboration with David Horner so as to adapt to mitochondrial dynamics. The SiteVar algorithm has been improved by assigning different scores to transitions and transversions. The SiteVarProt program has been transformed in MitVarProt where the Blosum-like index has been replaced with mtRev indeces [7] . A score for gapped sites has been introduced in both the methods. CWF CWF procedure is performed on a single human mitochondrial sequence (the Query sequence), and is aimed at predicting its haplogroup paternity by comparing the Query sequence against the rCRS sequence, thus detecting its mtSNPs patterns. A Genome Card of the analysed genome is generated and displayed to the user reporting the predicted haplogroup paternity. VGWF VGWF procedure is performed both on the entire content of the database, and on continent-specific subsets. It is applied every time the content of the database is consistently updated, as the site-specific variability software performs statistical estimates whose results significantly change only in case of a consistent change of the starting data (number of the sequences). In order to make these results more statistically significant, also bootstrap values are estimated. The simulated variability values provide an idea of the statistical significance of the dimension of the real samples: the closer the simulated value to the real value, the more “ideal” the dimension of the real sample stored in the database.

Results

HmtDB is a bioinformatic platform allowing the storage, query and analysis of human mitochondrial sequences. It is available on the web through a login procedure; free registration is required. It is organised in a relational database, storing the human mitochondrial complete genomes, data related to the sample and the subject from which the mtDNA was extracted, and the results of the variability analyses performed through the automatically running VGWF implemented in the resource itself. Five macro-functions have been designed : (1) the browsing of the database; (2) the analysis of a new human mt genome in order to automatically classify it according to the updated mt haplogroup classification; (3) the browsing of the previously performed analyses; (4) the query of the database through (a) a simple text search form or (b) a multi criteria form made of pop-up menu and free text retrieval windows; (5) the submission of a new mt human genome. Functions 4 and 5 have been designed but not yet implemented. The results obtained by the application of the SiteVar program to the continent-specific datasets have shown that nucleotidic mitochondrial sites presenting discriminating variability values in a particular geographic area respect to the rest of the world can be considered good population markers. Thus, VGWF can be considered as a haplogroup predicting tool contributing to the completion and refinement of the mt haplogroup classification and capable of discovering new haplogroups and sub haplogroups.

Contact email: m.attimonelli@biologia.uniba.it

URL: http://www.hmdb.uniba.it/

Supplementary Information: References [1] Brandon MC, Lott MT, Nguyen KC, Spolim S, Navathe SB, Baldi P, Wallace DC. MITOMAP: a human mitochondrial genome database--2004 update. Nucleic Acids Research. 2005. 33 (Database Issue): D611-613. [2] Handt O, Meyer S, von Haeseler A. Compilation of human mtDNA control region sequences. Nucleic Acids Research. 1998. 26(1): 126-129. [3] Tanaka M, Takeyasu T, Fuku N, Li-Jun G, Kurata M. Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. Ann N Y Acad Sci. 2004. 1011:7-20 [4] http://www.genpat.uu.se/mtDB/ [5] Pesole G, Saccone C. A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. Genetics. 2001. 157(2): 859-865. [6] Horner DS, Pesole G. The estimation of relative site variability among aligned homologous protein sequences. Bioinformatics. 2003. 19(5):600-606. [7] Adachi J, Hasegawa M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 1996. 42(4): 459-68.