Non syndromic Hereditary Hearing Loss (HHL) bioinformatic workbench

Leo P[1], Scioscia G[1], Accetturo M[1], Creanza TM[1], Santoro C[1], Tria G[1], Giordano A[1], Battagliero S[1]

Motivation

One of the main bioinformatic challenges in studying molecular biodiversity about genetic complex disease is concerned with the need to build an integrated vision about all the molecular aspects of the targeted pathology. In an ideal scenario, to build such a vision, it is needed to collect and curate all relevant data as well as develop/activate many customized analysis tools and pipelines. In practice, we are very far from this scenario having useful data spread in a number of public large data sources and public databases and the difficulty to activate specific algorithms/workflows in integrated/combined way, acting on all data sources. We argue that specialized bioinformatic pathology-based “workbenches", customized in terms of data and analysis/analytics ability, could better support research teams in their daily activity and speed-up the generation insights. We experimented such a scenario in building a bioinformatic workbench dedicated to the Hereditary Hearing Loss (HHL) that aims to provide a unified and high specialized environment for supporting researchers, as a one-stop in-silico workbench, in integrating, searching, analysing and discovery new genes, and in perspective other biomarkers, involved in the HHL. HHL is a good test-bed for our purposes since it is a largely genetically heterogeneous disorder which covers more than 70% of all hereditary hearing loss cases. The causes of HHL are complex and not completely understood on the genetic front. At present 51 genes are known as responsible, if mutated, of this phenotype (disease genes), while several linkage studies over the years have shown that the number of chromosomal regions involved in HHL is much greater: for some of them the genes causing HHL have not been identified yet, while for others it cannot be excluded the presence of more than one disease gene (as DFNA3 that harbors GJB2 and GJB6). Information and data about HHL, as for any other complex diseases, are at present scattered throughout a number of information resources, in the case of HHL the only tentative to gather them being the “official” HHL home page, where only lists of the disease genes and the susceptibility loci, although not complete and/or updated, are reported.

[1] IBM GBS BAO Advanced Analytics Services and MBLab, Via P. Leonida Laforgia, 14 - 70125 Bari. Italy

Methods

The HHL Workbench integrates data and bioinformatic tools in a web-based environment based also on open-source components, such as the Joomla content management system. Currently the “gene” is the main conceptual entity we considered to organize activities in the workbench. Data are collected and filtered from public and private sources and arranged around two classes: “disease genes” and “candidate genes”. The workbench provides a collection of data related to HHL coming from a number of data sources: literature as well as molecular databases extracted from public and private sources as well as a set of dedicated analytics tools to operate on the workbench data in integrated way. The HHL workbench has been thought in a re-usability perspective, as a prototype to provide a bioinformatic analysis pipeline to perform gene scoring on other genetic diseases, laying the groundwork to study the molecular biodiversity of other complex disorders.

Results

The HHL workbench is an integrated platform to store and analyse data about HHL. Data include the complete and updated lists of disease genes and susceptibility loci. They are accompanied by exhaustive bibliography, the chromosomal location and a brief description of the gene/locus of interest. The lists have been drawn accurately searching the literature for information about the disease. A human dynamic karyotype in the homepage drives to explore the knowledge into the workbench by localizing and displaying individual features on the chromosomes. A first set of bioinformatic tools are available to extract genes from a given locus or to directly query and download all gene expression experiments for a given gene form NCBI or extract all GeneRIF abstracts. A syntactic and semantic search engine, based Gene Ontology, supports information seek tasks, acting on more than 100k pubmed abstracts related to the considered genes and accumulated into the workbench. In addition a number of specific bioinformatic pipelines, able to work on the gene comparison level, are also available which aim to support gene prioritization studies. In particular a Gene Semantic Similarity Measure that performs the prioritization of candidate genes based on their Gene Ontology annotations, and a Gene Textual Similarity Measure, that ranks candidate genes on the basis of the textual profiles of their GeneRIFs. In the future other similarity measures based on different data sources will be integrated in the workbench, carrying to a comprehensive gene prioritization tool to provide important indications about where to look for new HHL causative mutations.

Contact e-mail

pietro_leo@it.ibm.com