CHAPTER 5

Literature and Genome Data Mining for Prioritizing Disease-Associated Genes

Carolina Perez-Iratxeta,* Peer Bork and Miguel A. Andrade

Abstract

The first step in understanding the molecular biology of an inherited disease is to identify which gene or genes are carrying variants. This process starts with locating the mutations in a chromosomal band, as narrow as possible, and follows with the manual analysis of all the genes mapping in this region. Usually this is not an easy task, but it can be facilitated by complementary computational approaches that evaluate all genes in a region of interest. We present here a method that combines literature mining, gene annotations, and sequence homology searches to prioritize candidate genes involved in a given genetic disorder. The method progresses in two steps. Firstly, we compute associations of molecular and phenotypic features as taken from MEDLINE. Secondly, for a disease with a given phenotype and linked to a chromosomal region, sequence homology based searches are carried on the chromosomal region to identify potential candidates that are scored using the precomputed associations. The scoring of associations between biological concepts using links across databases can be extended to other databases in Molecular Biology and to nondisease phenotypes.

Introduction

Some inherited mutations affecting one or more genes can produce exceptionally grave disorders that affect a high proportion of the population. Well known examples are asthma, diabetes or cancer. Finding out which genes contribute to the phenotypes can open the floor for better therapies, proper diagnosis and prognosis, and even prevention in some cases. Finding genes related to other more rare inherited pathologies has also a high biological and medical importance, because their identification may provide us with new insights about molecular mechanisms, and propitiate medical advances in related areas.

Many genes associated with (mostly monogenic) diseases have been identified and characterized in the past. To date, around 1200 of these are stored in the OMIM database (http://www.ncbi.nlm.nih.gov/omim/). The usual procedure to identify the molecular basis of a monogenic disease is to start by positioning the mutation in the genome by linkage analysis using data from families of affected individuals. The result is a more or less narrow cytogenetic location that is later screened for mutations in genes mapping to the region, often manually selected based on gene function and possible relation to the disease phenotype.

Complex diseases are much harder to position. Weaker linkage correlation signals to loci and the lack of homogeneity within the affected (usually very large) population produce imprecise association to several and larger chromosomal regions. Alternative experimental ways, as the use of polymorphisms, are ongoing (for review see ref. 6).

**Corresponding Author: Carolina Perez-Iratxeta—Ottawa Health Research Institute. Box 411, 501 Smyth Road, Ottawa, Ontario K1H 8L6. Canada. Email: cperez-iratxeta@ohri.ca

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

Typically, after positional cloning, researchers have to face manual analysis of tens to hundreds of candidate genes, trying to establish possible mechanistical relationships between the ethiology of the disorder and the function of those genes. Even in the case of monogenic diseases, where positional cloning may result in a reasonably narrow band, it can happen that the affected gene has not yet been characterized and functionally annotated, or, in the worst case, has not even been predicted as a gene. Moreover, many diseases are still far from being understood. This means that the molecular biology underlying the particular phenotype is not known, and consequently not described in the biomedical literature.

With the advent of functional genomics approaches, alternative or complementary methodology is frequently used on large sets of genes, for example gene expression analysis using DNA microarrays.[7] Yet, the interpretation of such results is difficult.

We have proposed a computational approach that helps to overcome the three major hampering factors mentioned above, namely, the large size of the cytogenetic bands, the presence of uncharacterized genes therein, and poorly known molecular mechanisms that prevent a straight-forward expert-based selection of candidates.[13]

Our system mines the existing literature and current knowledge about genes, and maps this information to the completed sequence of the human genome. The procedure starts with the phenotype associated to a disease and tries to relate this to molecular functions of the genes in the region. It then scores this information based on a corpus of precomputed links taken from more than 10 Million abstracts in the MEDLINE database. Then it compares the region of interest against annotated proteins by homology search using BLASTX[1a] and produces a ranking according to their scored associations.

Mapping Symptoms to Gene Functions

The first step in our method is to find out automatically which gene functions could be associated to a particular disease phenotype. Both the disease phenotype and the gene function can be summarized with a few keywords describing their main features. Our method is mainly based on detecting the associations between phenotype keywords and gene function keywords.

Our first source of information are the abstracts of literature reports stored in MEDLINE. Each MEDLINE reference is manually annotated with keywords, commonly around a dozen, at the National Library of Medicine (http://www.nlm.nih.gov/). These keywords are organized as an ontology called MeSH (Medical Subjects Headlines, http://www.nlm.nih.gov/mesh/). The MeSH terms are hierarchically organized in eight main categories. The 'C' category, corresponds to 'Diseases'. Then, given a disease, we take as its keywords the MeSH C terms annotated in the MEDLINE references dealing with that disease.

For the genes we use the RefSeq gene database of annotated and validated genes[15] and Gene Ontology[1] as the keyword system. Gene Ontology terms (GO terms, http://www.geneontology.org) constitute an ontology that has become very popular for functional annotation in molecular biology databases. We take as gene keywords all the GO terms associated to a gene in the RefSeq database.

To estimate the degree of relatedness between every MeSH C term, representing a symptom, and every GO term, representing a gene feature, a simple approach could consist of counting how often a given MeSHC term appears in any of the MEDLINE references linked to those gene entries in the RefSeq database annotated with a given GO term.

However, most of the papers linked to RefSeq genes are dealing more with the biochemical characterization of the gene than with clinical matters. Even in the whole MEDLINE there is not enough literature about molecular medicine to permit us to relate symptoms directly to molecular functions. We solve this problem by taking into account that genes relate to phenotypes by means of molecules. Accordingly, we enhance the signal strength of the relations by using an intermediate association step through another MeSH category: the D category consisting of 'Chemicals & Drugs' (see Fig. 1). Then, firstly we count all cooccurrences of MeSH C and MeSH D terms in references of the whole MEDLINE. For example, the MeSH terms 'Brain Ischemia' (C) and 'Glutamic Acid' (D) are mentioned together frequently in MEDLINE