Wilhelm Johannsen Centre for Functional Genome Research

University of Copenhagen

A method for prediction of candidate disease genes and regions by integration of positional, functional and clinical data

Mads Hjorth, Søren Holstebroe, Iben Bache, Merete Bugge, Zeynep TŸrmer, Niels Tommerup

Wilhelm Johannsen Centre for Functional Genome Research, The Panum Institute, University of Copenhagen, Blegdamsvej 3, Copenhagen N, DK-2200, Denmark;
Correspondence: Mads Hjorth, madsh@medgen.ku.dk

Introduction

Publications on functional aspects of genes, gene products and genomes, and relationships with human and animal biology and diseases are not only increasing exponentially, but are also dispersed to an expanding number of novel journals, including on-line journals. Also, the amount of data related to sequences, gene expression patterns, protein-protein interactions, etc, and ultimately whole biological systems, which can be addressed via the Internet, increasese rapidly. As a result, scientists will use an ever increasing time in hyperspace, jumping from site to site; it will not only be difficult to keep up with all relevant information on gene structure, function and dysfunction, but it will also be increasingly time consuming to link available data into a proper biological context, including relevance for human disease.

One solution will be to develop automatic systems that can integrate,evaluate and filter data resources, ideally to dispose of the vast excess of irrelevant data, and to present the relevant data in an extracted, condensed, easy-to-read, yet flexible format. We have used Mendelian Cytogenetics Network database (MCNdb) [1,2], with a content of more than 2800 disease-associated balanced chromosomal rearrangements (DBCRs), associated with more than 6000 chromosomal breakpoints and 8100 trait associations, as an ideal starting point for prototyping the idea of an automated association system.

Method

Each trait in an MCNdb case is queried against OMIM, to create a list of textually associated chromosomal positions (Figure 1). Also, OMIM are queried by the positional data from the involved breakpoint regions, creating another list of disorders and associated traits (Figure 2). These two lists are then compared. Positive hits are presented at various levels, ranging from candidate chromosomal regions to specific candidate genes and suggestions for FISH probes are given (Figure 3).

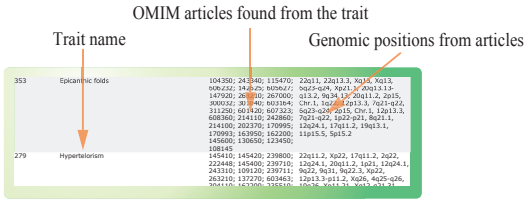

Figure 1. *Example of result of OMIM searches for traits from an example case with the karyotype of 46,XY,inv(12)(p13q24)de novo and a phenotype including macrocephaly, mental retardation, epicanthic folds and hypertelorism.*

The query against OMIM was found to give best results when text in the clinical synopsis field was given more weight then other information in each OMIM entry. To provide efficient, fine grained and tunable search of the text content of OMIM we have create a local field-based index using the Lucene search engine from the Apache Jakarta Project.

Got cases with DBCR?

The objectives of Mendelian Cytogenetics Network (MCN) are to collect data on Disease Associated Balanced Chromosomal Rearrangements (DBCRs), both retro- and prospectively; to facilitate the dissemination of these data by an online database containing the information on breakpoints and clinical features; identify those DBCRs with the highest disease potential in order to facilitate cost-efficient clinical reexamination and generation of material (cell lines, chromosome suspensions for FISH mapping, DNA); distribute standardized FISH probes for the molecular mapping of DCBRs; initiate collaborative studies of specific disorders; provide facilities and contacts for cell line production, FISH-mapping and molecular cloning of DBCRs for those laboratories who may wish it.

visit us at www.mcndb.org

Figure 2. *Example of result of OMIM searches for breakpoints in the example case. In the example we find 23 hits where a trait and a breakpoint relates to the same OMIM article.*

Currently we are using a simple overlapping algorithm to find hits, but in prototype development we are using distance messurements to allow for uncertainties in the cytogenetic description of the breakpoints. We have found when mapping breakpoints that the 'true' molecular position of the breakpoint maybe off-set with one or more subbands.

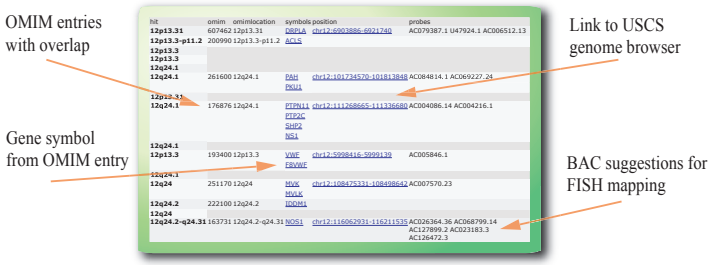

Figure 3. *Final result for the example case. For the 12p13 breakpoint 3 candidate genes with 4 known BACs are suggested and for the 12p24 breakpoint 8 candidate genes with 10 BACs are suggested.*

OMIM entries holding information about gene symbols are queryed against the list of BACs that are cytogenticaly mapped to the human genome which is available through the genome browser at UCSC [8]. The result of the total analysis is about a handfull candidate genes for each breakpoint with suggestions for probes for finding the breakpoint using FISH.

Discussion

To evaluate the correctness of the suggestions for probes spanning the involved breakpoints we have run two sets of cases with DBCR where breakpoints are known on a molecular level from FISH mapping. The first set is of 43 cases from the Wilhelm Johannsen Center for Functional Genome Research (WJC) [7] and the other is 13 cases avaliable online on June 1st 2004 from the Developmental Genome Anatomy Project (DGAP) [8]. The above method correctly identifies Prader Willi 15q, Charcot-Marie-Tooth 17p and Hirschsprung disease 2q loci in cases from the above sets when the syndrome is removed from the case description with only one or two false positives. Turner et. al. 2003 [3] suggest 'relative enrichment ratio' as a measure of correctness for disease gene prediction methods. We have not yet established a method to determin this ratio for our method, but preliminary results suggest that it is comparable to more complex metods described by Turner et. al. 2003[3], Freudenberg and Propping 2002[4], van Driel et. al. 2003[5] and Perez-Iratxeta et. al. 2002[6].

References

1. Mendelian Cytogenetics Network Database (MCNdb). http://www.mcndb.org
2. Bugge M, et. al. Disease-associated balanced chromosome rearrangements: a resource for large scale genotype-phenotype delineation. *J Med Genet* 37:858-865, 2000
3. Turner S, et. al. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biology* 4:R75, 2003
4. Freudenberg J, Propping P. A similarity-bases method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 18:S110-15, 2002
5. van Driel MA, et. al. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. Jou. Hum. Gen* 11:57-63, 2003
6. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using datamining. *Nature Genetics* 31:316-19, 2002
7. Wilhelm Johannsen Center for Functional Genome Research. http://www.wjc.ku.dk
8. Developmental Genome Anatomy Project (DGAP). http://www.bwhpathology.org/dgap

Acknowledgement

The Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation.

This poster can be obtained as a PDF-file at
www.wjc.ku.dk

Acknowledgement:
Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation