A Missing Link – Molecular–Cytogenetic Data for Oncogenetic Data Mining

Keywords: Comparative Genomic Hybridization, data mining, online database, molecular cytogenetics

Author: Dr. med. Michael Baudis

Contact: Dr. Michael Baudis
Dept. of Pathology
Stanford University Medical Center
300 Pasteur Drive
Lane L210
Stanford, CA 94305
fon ++1 (650) 723 5304
fax ++1 (650) 498 6222
mbaudis@stanford.edu
A Missing Link – Molecular–Cytogenetic Data for Oncogenetic Data Mining

Michael Baudis

Abstract

Ever since the description of the t(9;22) as the cytogenetic hallmark in the majority of Chronic Myelogeneous Leukemias (CML), the analysis of chromosomal aberrations has been an important and rewarding tool in oncogenetic research as well as in clinical oncology. Although the Human Genome Project and related efforts have led to the deciphering of the genomic sequence, and although the growing field of expression array analysis relies on genome based information management, no systematic effort has been undertaken to integrate the published information about cytogenetic aberrations into oncogenetic data-mining efforts. Here, I will detail objectives, methods and current status of the progenetix.net online CGH data collection, and discuss present and future applications of the dataset.

Introduction

Human malignancies develop through the accumulation of genetic defects, finally giving rise to a fully transformed, malignant clone. Of those defects, some may represent themselves through numerical or structural chromosomal aberrations. Those aberrations may either point towards a genetic dosage effect (deletion of regions harboring tumor suppressor genes, amplification of regions harboring proto-oncogenes), a deregulated expression through translocation of regulatory sequences or a deregulation through generation of fusion genes with constitutively activated functional properties.

For several decades, the analysis of chromosomal abnormalities in malignant cells has been an important tool for the evaluation of malignant tumors, in basic research as well as in clinical practice. Detection of distinct structural and numerical aberrations has led to the detection of genes relevant for the neoplastic transformation in a variety of malignant entities. In some instances, this knowledge could be translated in targeted therapeutic approaches, the most impressive being the successful introduction of tyrosine kinase inhibitors for blocking the constitutively activated abl kinase in Chronic Myelogeneous Leukemias (Druker et al. 2001).
Another target of cytogenetic analysis has been the identification of malignancy related signatures in cytogenetic aberration patterns (Mitelman 1982). Of the attempts to collect and classify chromosomal aberrations in malignant neoplasias, the most widely recognized is the catalog by F. Mitelman and coworkers (Mitelman 1994; online access through http://cgap.nci.nih.gov/Chromosomes/Mitelman) and classification projects arising from that (Pandis et al. 1996).

Cytogenetic aberration data derived from "traditional" banding experiments harbor some limitations concerning the usage in large scale data mining procedures. Although the technique allows to report a whole genomic aberration status, in a relevant number of cases karyotypes are only partially resolved, mostly through occurrence of complex marker chromosomes or other chromosomal segments of unknown origin (e.g. homogeneously staining regions, HSRs, or double minute chromosomes, DMINs). Also, the technique requires the collection of actively dividing tumor cells, which limits the application in many solid tumors and may lead to artifacts introduced through the use of mitogen supported short term cultures.

Several years ago, the molecular–cytogenetic technique of Comparative Genomic Hybridization (CGH; Kallioniemi, Kallioniemi et al. 1992; du Manoir, Speicher et al. 1993) was introduced. Although CGH lacks the ability to detect chromosomal translocations, the usage of a defined hybridization matrix and the possibility of semi–automated, quantitative image analysis as well as the option to use frozen or paraffin embedded archival tumor tissue as starting material made CGH the choice for genomic imbalance screening.

With the exception of some regions rich in repetitive sequences, e.g. at chromosomal centromeres, and with the possibility of missing aberration below the spatial resolution of few Mega bases (Bentz et al. 1998), CGH experiments give a reliable overview of the relative abundance of all genomic regions in the genome of the analyzed cell population. The consistent coverage of the whole genome makes it possible to compare CGH results through different experiments.

Until recently, the mining of the large number of CGH publications has been hindered through the various annotation methods. Usually, the result of those experiments is communicated either in various text formats related to the International System for Cytogenetic Nomenclature (Mitelman 1995) or graphically, using chromosomal ideograms for the representation of chromosomal gains and losses.
Few attempts have been made to collect and represent CGH data on a publicly available platform, the most prominent being the AMBA CGH database at the Charite, Berlin (http://amba.charite.de/cgh/) and the CGH and SKY database at the NCBI (http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi). Although both of those allow the inspection of actively contributed cases, they fail to provide a generalized overview of published CGH cases, and also offer no accessible source format that can be easily used in genomic data mining procedures.

The objectives for the project described here were 1. the collection of results from CGH experiments representing the majority of human malignancies; 2. the development of a parsing engine for transformation of the ISCN karyotypes into a pseudo-binary report giving a status report (“no change – gain – loss”) for each chromosomal band in each included case; 3. the application of the transformed data for generating graphical as well as numerical representations of the database content, and 4. the generation of an interval based report format for integrating the genomic aberration data into data mining procedures.
Data and Methods

Data collection
For collection of the CGH data, PubMed was used to identify publications containing results of CGH experiments. Articles focussing on cell lines were excluded. Additional publications not showing up in the primary search efforts were found through references, or through information received from the article's authors. CGH data was included if the publication contained a listing of the included cases, with the minimal required information being band specific rev ish karyotype and clinical diagnosis. In few instances, aberration information was transcribed from chromosomal ideograms with case specific annotation. If case specific information was lacking, the corresponding author of the article was contacted. For disease classification, all cases were coded according to the WHO ICD–O–3 classification system.

Data parsing
For transformation of the CGH aberration information to a status information for predefined genomic intervals a dedicated parsing engine was implemented in Perl. As template for the genomic intervals, the mapping of chromosomal bands to the "Golden Path" from the UCSC genome database (T. Furey, version from Nov. 2002; http://genome.ucsc.edu) is used. Parsing is performed through a three step process (Figure 1):
1. normalization by transforming the raw rev ish related annotations to standard rev ish ISCN 1995 format
2. format validation by comparing the transformed annotation to allowed annotation patterns, and content validation by comparison of the annotated bands to a list of true band values
3. for each of the chromosomal bands from a band–list file of arbitrary resolution, the occurrence of informative annotation intervals including this band is evaluated

The transformed data is stored in an XML file, containing the status information for each aberrant band as well as additional information (ICD–O–3 code, PubMed identifier, case–name, diagnosis).

The main XML output file as well as additional information (publication data) is parsed for offline generation of HTML files representing various aspects of the data content.
Current status and outlook

Database content
As of 2003-03-28, progenetix.net contained a of 7372 cases, 5721 of which were showing chromosomal imbalances (78%). From all cases, a total number of 721972 aberrant bands was described, giving an average of 97.9 per case, or 11.4% of all bands using the 863 bands karyotype. The chromosomal band with the highest frequency of losses was 13q21.1 (13.1% of all cases). The highest percentage of gains was found in the region 8q23q24 (max. 16.7%). The highest number of distinct high level amplifications was reported for 12q15 (1.5% of all cases). The overall distribution pattern is depicted in figure 2.

Through the web interface, separate presentation groups are available for each included article, all ICD-O-3 entities and supergroups consisting of cases from related ICD-O-3 entities or representing clinically defined pictures (e.g. "uterine leiomyomatous tumors"). For each presentation group, the graphical representation of imbalances in ideogram format as well as the band specific listing of aberration percentages, with link to the corresponding location in the ENSEMBL database, are available. Currently, presentation groups with a casenumber between 20 and 500 contain also a graphical representation of the clustering of cases according to their aberration pattern.

Cluster analysis of CGH data
So far, several publications have described attempts to cluster chromosomal aberrations in malignant tumors for the identification of prognostically relevant signatures or the reconstruction of oncogenetically relevant pathways (Mattfeldt et al. 2001; Chibon et al. 2003). Those efforts have been limited to small sample sizes, or applied only low resolution input data (e.g. arm specific). In contrast, the complete progenetix.net dataset is available in a cluster ready file format, with the aberration status of each band from each of the currently 5721 informative cases as a data point in the precluster matrix. The cluster picture for the analysis of all cases for a reduced 393 bands resolution (Pearson correlation, clustering of cases, bands not clustered) is shown in Figure 3.

Especially due to the physical relation between neighboring bands, which is reflected in their related probability to be involved in chromosomal aberrations, a simple clustering approach may not represent the optimal solution in identifying significant aberration patterns. Vandesompele et al. (manuscript in preparation) used Principle Component Analysis (PCA) for detecting relevant marker bands in an analysis of Neuroblastoma CGH data prepared by using the progenetix.net engine. Through this approach,
subgroups with significant differences in the clinical behaviour of the cases could be delineated.

**Mapping of CGH data to genomic intervals**

Due to the mapping of the underlying band matrix to the genomic intervals of the bands ("Golden Path" location), the chromosomal aberration information from selected subsets can be compared to the list of significantly altered clones from expression array experiments. For examplifying the viability of such an approach, the clones reported as significant delimiters from a SAM (significance analysis of microarrays; Tusher et al. 2001) analysis of expression array experiments in soft tissue tumors (Nielsen TO, West RB et al. 2002, data provided by M. van de Rijn and R. West) were remapped to their chromosomal location using the tools provided through the Stanford Microarray Database. For the different entities, the density of those relevant clones was drawn as the relative abundance of significant clones per chromosomal band. Figure 4 shows how the frequency of clones significant for the MFH signature reflects the chromosomal aberration pattern for chromosome 1 in MFH. Though certainly not being a statistically sound method, this simple model already points towards the connection between chromosomal imbalances and deregulation on the expression level.

**Future extensions**

The current database content was derived from 262 publications containing original CGH results. Those articles represent about 45% of the CGH articles presumably containing valid (unique, no cell line report, prenatal or xenograft analysis) CGH data, which have been found through literature search. One could estimate the 7372 database cases to represent approximately 30–40% of all valid CGH cases reported so far in the bio–medical literature. Besides the inclusion of upcoming CGH publications, the expansion of the database content will be driven by requesting data from authors of publications currently representing data in a format not suitable for direct database input (result summaries etc.).

Recently, various techniques applying the principle of CGH to array based experiments have been introduced ("Matrix–CGH", Solinas–Toldo et al. 1997; genomic arrays, Pollack et al. 1999). Due to the higher resolution and potentially better standardization of the spotted hybridization targets compared to chromosomal CGH, one can expect a rapidly growing number of publications in this field. Due to the "Golden Path" adjusted mapping of the traditional CGH data in progentix.net, the inclusion of array based experiments into the database should be feasible once certain quality criteria have been established (definition of "true" aberrant regions from a range of clones etc.).
Besides the quantitative extension of the dataset, for the future qualitative improvements are planned as well. On is the introduction of additional coding schemes, e.g. allowing the selection of cases according to the tumor localization. Also, active search options are to be evaluated.

**Use the link!**
Although we have moved into what some voices call the "post-genomic" era, and although a large number of bioinformatics tools have been developed, especially for the field of expression array analysis, there has been a lack of effort to relate this information to the treasure trove of (molecular–) cytogenetic information from human neoplasias collecte in the context of clinical tumor analysis. progenetix.net has been developed to serve as a link to include a large subset of those data into oncogenetic data-mining efforts.

Suggested targets for extensive data analysis include:
1. identification of disease specific genomic "hot spot" regions
2. matching of aberration patterns to disease entity coding schemes, in search for relevant signatures
3. filtering of target genes from expression array experiments according to genomic information
4. relating of genomic "hot spot" regions for reconstructing oncogenetic pathways
It should be possible to envision other applications as well.

The complete dataset is available for download from the progenetix.net website, with no restrictions for academic research purposes.
Selected Literature


Figure Captions

**Figure 1**
Flow chart describing the collection and processing of CGH data for the progenetix.net online database.

**Figure 2**
Graphical representation of all chromosomal imbalances reported in 7372 tumor specimens. Red: losses, green: gains, dark green: high level amplifications.

**Figure 3**
Clustering of 5721 informative cases according to their similarity in aberration pattern.

**Figure 4**
An ideogrammatic representation of chromosomal imbalances involving chromosome 1 in MFH (a, 104 cases from the literature) is compared to the density of clones relevant for distinguishing MFH cases from a SAM analysis of expression array experiments (b). The preferred involvement of bands 1q21q23 may point towards genes functionally involved in specific disease determination (for technical reasons, different banding resolutions were used here)
Figure 1

<table>
<thead>
<tr>
<th>BAND</th>
<th>LOSS</th>
<th>GAIN</th>
<th>AMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p36.32</td>
<td>8.7</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.31</td>
<td>8.7</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.23</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.22</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.21</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.13</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.12</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p36.11</td>
<td>8.3</td>
<td>8.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1p35.3</td>
<td>21.5</td>
<td>9.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

...
Figure 3

5721 informative cases

1p36.3 -> Yq12