Online database and bioinformatics toolbox to support data mining in cancer cytogenetics

Michael Baudis
University of Florida Shands Cancer Center and Division of Pediatric Hematology/ Oncology University of Florida, Gainesville, FL, USA

BioTechniques 40:269-272 (March 2006)
doi 10.2144/000112102

Oncogenomic screening in malignant neoplasias has led to the description of oncogenetic mechanisms and, recently, to the first successful targeted drug development approaches (1). Individual genomic abnormalities are used as diagnostic markers or for the individual prediction of clinical aggressiveness (2). However, most malignancies show nonrandom aberration patterns that may reflect the cooperation of multiple onco- and tumor suppressor genes, according to the multistep model of oncogenesis (3). The complexity of those changes warrants the application of advanced data mining methods for the development of oncogenomic models.

A number of cytogenetic and molecular genetic techniques describe chromosomal imbalances or changes in the regional DNA content of tumor cells. Historically, the microscopic inspection of stained metaphase spreads (4) had been most widely applied, and still is the reference method, in many clinical applications. Comparative genomic hybridization (CGH) (5) permits the detection of genomic imbalances from tumor samples with more than 50% tumor cell content as well as from archival material (6). Recently, array or matrix CGH (7,8) has started to overcome the limited spatial resolution (9) of metaphase CGH.

An intriguing concept for oncogenomic data mining is the combination of the accumulated cytogenetic data with the molecular cytogenetic data from metaphase and array-based CGH experiments. However, complex annotation formats are used for the description of experimental results. The standards for cytogenetic banding and reverse in situ hybridization (ISH) (e.g., CGH) have been defined in the International System for Cytogenetic Nomenclature (ISCN) (10). The results of genomic microarray experiments usually are stored according to the minimal information about a microarray experiment (MIAME) guidelines (11).

The largest publicly accessible resource for molecular cytogenetic screening data in oncology is the Mitelman Database of Chromosome Aberrations in Cancer (cgap.nci.nih.gov/Chromosomes/Mitelman), which describes more than 46,000 samples analyzed by metaphase banding. Utilization of this data has been limited by the lack of a format amenable to data mining procedures, though valuable studies have been published by the database maintainers (12). Another resource is the National Center for Biotechnology Information (NCBI) spectral karyotyping (SKY)/CGH database (www.ncbi.nlm.nih.gov/sky/skyweb.cgi) (13). It provides well-structured clinical and experimental information for the included cases, but due to the reliance of the NCBI site on voluntary data submission it is, with currently 1006 included experiments, quantitatively limited. Recently (13), the Mitelman database and the SKY/CGH project have been integrated into NCBI's Entrez Cancer Chromosomes site (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cancerchromosomes) and now offer band-specific search capabilities. By far, the largest collection of case-specific CGH data are presented through the Progenetix web site (www.progenetix.net) (14), on which this article is focused.

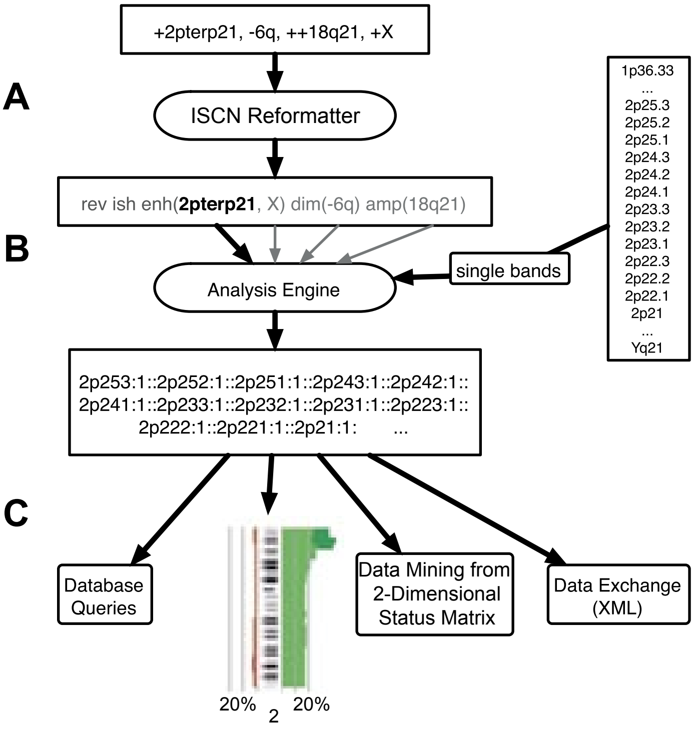

A
B
C
+2pterp21, -6q, ++18q21, +X
ISCN Reformatter
rev ish enh(2pterp21, X) dim(-6q) amp(18q21)
Analysis Engine
single bands
1p36.33
...
2p25.3
2p25.2
2p25.1
2p24.3
2p24.2
2p24.1
2p23.3
2p23.2
2p23.1
2p22.3
2p22.2
2p22.1
2p21
...
Yq21
2p253:1::2p252:1::2p251:1::2p243:1::2p242:1::
2p241:1::2p233:1::2p232:1::2p231:1::2p223:1::
2p222:1::2p221:1::2p21:1: ...
Database Queries
Data Mining from 2-Dimensional Status Matrix
Data Exchange (XML)
20%
2
20%

Figure 1. Cytogenetic data transformation, using comparative genomic hybridization (CGH) data as example. (A) The various International System for Cytogenetic Nomenclature (ISCN)-related annotation formats found in the literature are transformed to standard reverse in situ hybridization (ISH) ISCN. (B) Contiguous aberration intervals are checked for their inclusion of chromosomal bands. (C) A band-specific status annotation format serves as basis for data representation and analysis.

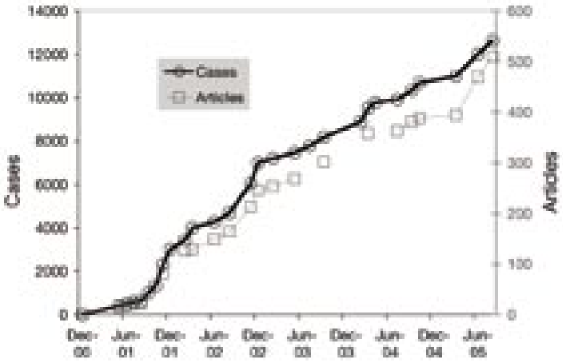


Figure 2. Expansion of the Progenetix database. The thick line and open circles indicate the case numbers (left ordinate). The open boxes depict the number of included publications (right ordinate). The abscissa gives a linear time scale.

The Progenetix project was initiated in December 2000. The main inclusion criterion was the complete description of the genomic status of each tumor specimen in a peer-reviewed article. Data sampling methods included copying of ISCN annotations from publication files or online supplements and transcription of data from printed matter. For some array CGH data, pseudo-reverse ISH annotations were generated (e.g., based on the Bioconductor DNAcopy package; www.bioconductor.org). For 72 articles, experimental results were provided by the authors of the original publications.

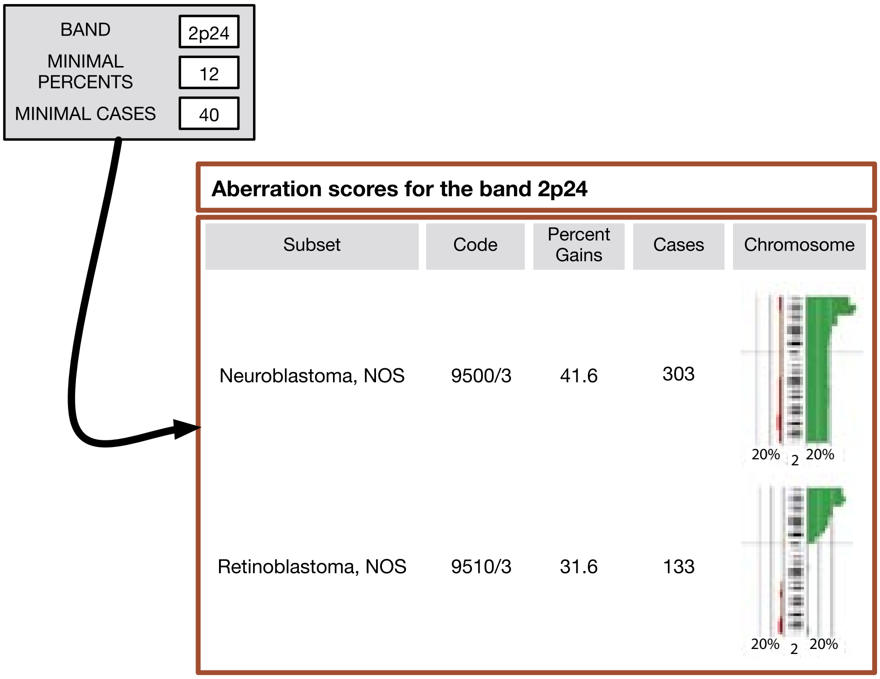
For the conversion of cytogenetic annotations, software was implemented in the Perl scripting language (www.isc.org/sources/devel/lang/perl.txt). Cytogenetic data are converted to standard ISCN 1995 format (Figure 1A) and automatically checked for syntax errors. Each band of a cytogenetic reference table with 862 bands resolution [currently University of California Santa Cruz (UCSC) May 2004 edition; hgdownload.cse.ucsc.edu/goldenPath/hg17/database/cytoBand.txt.gz] is evaluated for its inclusion in intervals derived from the text annotation, and the status (gain, loss, or high-level gain) is assigned accordingly (Figure 1B). The band status is annotated, and a two-dimensional band-specific status matrix file is generated (Figure 1C).

The minimal consistent amount of case-specific information is sampled from the literature. Diagnoses and topographies are recoded to the International Classification of Diseases in Oncology (ICD-O-3) format (15). Each case is referenced to the PubMed ID of its originating publication. For the web site generation, all different case entities (disease, locus, publication, custom group) are identified, and for each of them, specific overview pages are generated. These consist of a list of case-specific information, an ideogrammatic representation of genomic gains and losses, and a page showing the unsupervised clustering of cases according to their aberration pattern using XCluster (Gavin Sherlock; genetics.stanford.edu/~sherlock/cluster.html).

At the time of writing, 13,240 unique experiments published in 535 peer-reviewed articles have been included into the Progenetix database (Figure 2), representing 273 distinct neoplastic entities. The majority of those cases (12,179 or 92%) came from chromosomal CGH experiments.

Progenetix presents a unique case-specific structured overview of chromosomal imbalances for most neoplasias. After free registration, academic researchers are able to download the main database content, including the band-specific annotation data in an XML format. As an additional unique feature, the web site offers a query option for the relative aberration status of single bands in disease entities (Figure 3).

To allow users to convert, mine, and visualize their own molecular cytogenetic data sets, a version of the ISCN2matrix parser was implemented as a Perl CGI script. Users can upload a file containing data from multiple cases and generate chromosomal ideograms,



BAND	2p24
MINIMAL PERCENTS	12
MINIMAL CASES	40

Aberration scores for the band 2p24

Subset	Code	Percent Gains	Cases	Chromosome
Neuroblastoma, NOS	9500/3	41.6	303	
Retinoblastoma, NOS	9510/3	31.6	133	

Figure 3. A unique query option permits the search for tumor entities [as annotated by their International Classification of Diseases in Oncology (ICD-O-3) code] with a large number of imbalances involving a particular band. Given a suspected target gene, this feature allows the instantaneous identification of disease categories in which this gene could be deregulated based on frequent copy number changes. Here, the query for the *MYCN* locus on 2p24 shows the band to contain a local maximum for gains in neuroblastomas as well as in retinoblastomas.

cluster graphics, and XML files as described above.

Recently, the interval-specific aberration information from the Progenetix data set and the parsing software for CGH, as well as metaphase banding-based annotations, have shown their usefulness for the delineation of genomic aberration patterns with prognostic relevance (16) and for producing tumor type-specific combined genomic imbalance maps (17,18).

Large-scale data mining approaches based on tens of thousands of genomic profiles should lead to the identification of genomic signatures for a variety of neoplasias and the development of new diagnostic tools (e.g., disease-specific genomic arrays with low complexity). The integration of genomic aberration patterns will be of great benefit for the interpretation of expression array data, allowing for selection of genes with high probability of tumor-specific involvement. Additionally, the delineation of recurring genomic aberration patterns may become the basis for the development of smart target gene detection methods, using sequence similarity searches over commonly involved loci. Through the powerful combination of advanced data mining tools with unique data content, the Progenetix project should be useful for a new generation of oncogenomic data mining projects.

ACKNOWLEDGMENTS

The author is indebted to all individuals who contributed their otherwise not accessible original data. A list of contributors can be found on the Progenetix web site. Alejandra Ellison-Barnes is thanked for helping with data transcription from printed matter.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Druker, B.J. and N.B. Lydon.** 2000. Lessons learned from the development of an abl tyrosine kinase inhibitor for chronic myelogenous leukemia. J. Clin. Invest. *105*:3-7.
2. **Dohner, H., S. Stilgenbauer, A. Benner, E. Leupolt, A. Krober, L. Bullinger, K. Dohner, M. Bentz, and P. Lichter.** 2000. Genomic aberrations and survival in chronic lymphocytic leukemia. N. Engl. J. Med. *343*:1910-1916.
3. **Vogelstein, B. and K.W. Kinzler.** 1993. The multistep nature of cancer. Trends Genet. *9*:138-141.
4. **Crossen, P.E.** 1972. Giemsa banding patterns of human chromosomes. Clin. Genet. *3*:169-179.
5. **Kallioniemi, A., O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, and D. Pinkel.** 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science *258*:818-821.
6. **Speicher, M.R., S. du Manoir, E. Schrock, H. Holtgreve-Grez, B. Schoell, C. Lengauer, T. Cremer, and T. Ried.** 1993. Molecular cytogenetic analysis of formalin-fixed, paraffin-embedded solid tumors by comparative genomic hybridization after universal DNA-amplification. Hum. Mol. Genet. *2*:1907-1914.
7. **Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter.** 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer *20*:399-407.
8. **Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, et al.** 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. *20*:207-211.
9. **Bentz, M., A. Plesch, S. Stilgenbauer, H. Dohner, and P. Lichter.** 1998. Minimal sizes of deletions detected by comparative genomic hybridization. Genes Chromosomes Cancer *21*:172-175.
10. **Mitelman, F. (Ed.).** 1995. International System for Cytogenetic Nomenclature. Karger, Basel.
11. **Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, et al.** 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*:365-371.
12. **Hoglund, M., A. Frigyesi, T. Sall, D. Gisselsson, and F. Mitelman.** 2005. Statistical behavior of complex cancer karyotypes. Genes Chromosomes Cancer *42*:327-341.
13. **Knutsen, T., V. Gobu, R. Knaus, H. Padilla-Nash, M. Augustus, R.L. Strausberg, I.R. Kirsch, K. Sirotkin, and T. Ried.** 2005. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. Genes Chromosomes Cancer *44*:52-64.
14. **Baudis, M. and M.L. Cleary.** 2001. Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics *17*:1228-1229.
15. **Fritz, A., C. Percy, A. Jack, L.H. Sobin, and M.D. Parkin (Eds.).** 2000. International Classification of Diseases for Oncology (ICD-O), 3rd ed. World Health Organization, Geneva.
16. **Vandesompele, J., M. Baudis, K. De Preter, N. Van Roy, P. Ambros, N. Bown, C. Brinkschmidt, H. Christiansen, et al.** 2005. Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. J. Clin. Oncol. *23*:2280-2299.
17. **Mao, X., R.A. Hamoudi, I.C. Talbot, and M. Baudis.** Allele-specific loss of heterozygosity in multiple colorectal adenomas: towards the integrated molecular cytogenetic map II. Cancer Genet. Cytogenet. (In press).
18. **Mao, X., R.A. Hamoudi, P. Zhao, and M. Baudis.** 2005. Genetic losses in breast cancer: toward an integrated molecular cytogenetic map. Cancer Genet. Cytogenet. *160*:141-151.

Received 18 September 2005; accepted 9 November 2005.

Address correspondence to Michael Baudis, Division of Pediatric Hematology/Oncology, 1600 SW Archer Rd., ARB-R4-186, Gainesville, FL 32610-100296, USA. e-mail: mbaudis@ufl.edu