Database,
cytogenetics,
automatic analysis.

Iwona KOSTORZ [1], Marek SIKORA [1], Tomasz STECLIK [1]

DATABASE STRUCTURE FOR STORING AND AUTOMATIC ANALYSIS OF TEXT GTG AND FISH CYTOGENETIC RESULTS

The paper presents research on storing text cytogenetic tests data conducted with GTG and FISH methods for browsing and automatic processing of its string results. It presents analysis of the text cytogenetic test result and a two proposal structures of the databases (the storing data and the analytical one) used for anonymised text GTG and FISH cytogenetic test results storing and their processing. Presented structures were queried with known chromosome aberrations for T-cell acute lymphoblastic leukaemia (T-ALL), B-cell acute lymphoblastic leukaemia (B-ALL) and both T-ALL and B-ALL in the proposed database. Additionally, the analytical database was filled with the set of Mittelman database results and browsed with non-standard queries. Conducted tests demonstrated that the proposed analitical structure is useful for GTG and FISH cytogenetic test results text mining.

1. INTRODUCTION

Presented research is a part of the project ”Personalization of childhood acute lymphoblastic leukemia treatment in Poland”. It is a development of innovative diagnostics of childhood acute lymphoblastic leukemia that should enable more targeted therapies and lead towards improved treatment outcome.

The project includes research on molecular diagnostics of acute lymphoblastic leukemia (ALL), the most frequent malignancy in childhood, aiming towards treatment personalization. The project also aims at collection of biologic material for mass analyses, including systematic banking of cells and nucleic acids. Next generation sequencing are employed for development of innovative methodology for monitoring of treatment effectiveness.

One of the project's objectives is to create an advanced Information Technology System for the medical personnel and patients. This system will contain a unified database with clinical, genetic and molecular data from all patients as well as an information portal for patients.

The main aim of the research presented in the paper was to propose the structure of the database to store the results of cytogenetic tests collected during the research project and structure of database designed for automatic analysis of text cytogenetic test results.

Reports of the cytogenetic tests results are usually stored as a complex string of characters in accordance with the newest edition of the International System for Human Cytogenomic

[1] Institute of Innovative Technologies EMAG, Poland, Katowice, ul. Leopolda 31

Nomenclature (ISCN 2016) [8]. It contains standard nomenclature that is used to describe any genomic rearrangement identified by currently used techniques: karyotyping, fluorescence in situ hybridization, microarray, various region specific assays, and DNA sequencing. Each result of cytogenetic test contains a set of information about a karyotype designation, an uncertainty in chromosome or band designation, an order of chromosome abnormalities in the karyotype, normal variable chromosome features, numerical chromosome abnormalities, structural chromosome rearrangements, chromosome breakage and results additional technics of chromosome analysis as in situ hybridization (FISH) [4], microarrays-based, region-specific assays and sequence-based assays.

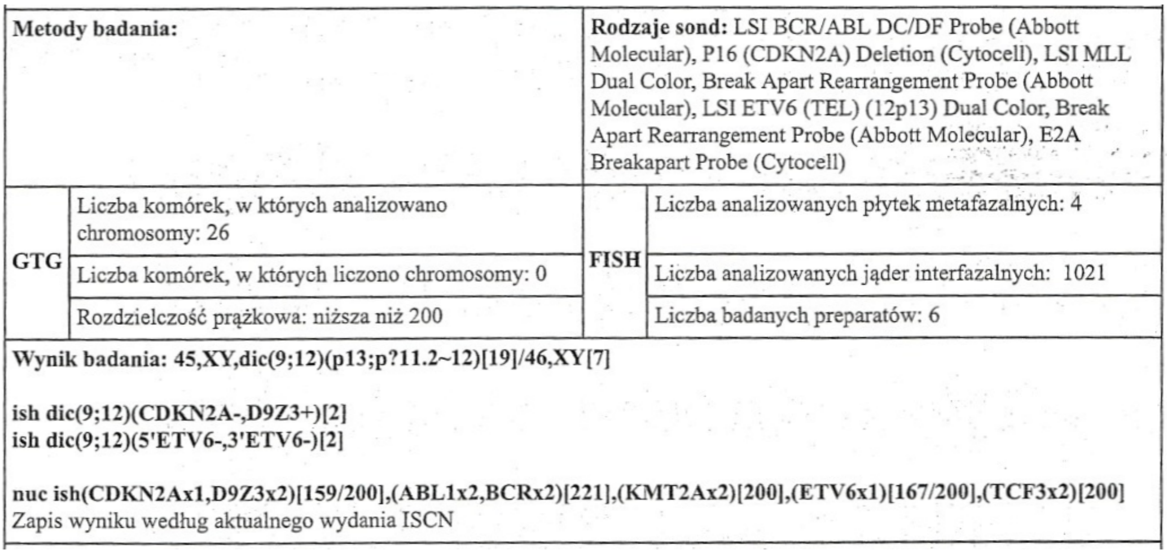
The results that were analysed contained information about karyotype and chromosome abnormalities. Cytogenetic tests were carried out using GTG banding method (G-bands treated with trypsin and stained with Giemsa stain) [7]. It also contained the results of fluorescence in situ hybridization.

So far karyotype text mining research was conducted in the field of aid in drug repurposing efforts, a the large - scale analysis of karyotype data [2] [3]. Simulations of reconstructing cancer karyotypes from short read data were also performed [5]. No research was found on the design of databases for analysing the text results of cytogenetic tests based on GTG and FISH analysis.

2. MATERIALS AND METHODS

As a part of the advanced Information Technology System for the medical personnel and patients elaborated in the project ”Personalization of childhood acute lymphoblastic leukemia treatment in Poland” two database structures were proposed. The first of these was designed to store the results of cytogenetic tests reports collected during the research project. The second one was designed for automatic analysis of cytogenetic tests results including GTG and FISH results.

The set of results came from one cytogenetic laboratory. All analysed data was fully anonymised. Generated cytogenetic test results were compliant with ISCN 2016. An exemplary fragment of the original test report is shown in the Fig 1.



Metody badania:		**Rodzaje sond:** LSI BCR/ABL DC/DF Probe (Abbott Molecular), P16 (CDKN2A) Deletion (Cytocell), LSI MLL Dual Color, Break Apart Rearrangement Probe (Abbott Molecular), LSI ETV6 (TEL) (12p13) Dual Color, Break Apart Rearrangement Probe (Abbott Molecular), E2A Breakapart Probe (Cytocell)	
GTG	Liczba komórek, w których analizowano chromosomy: 26	**FISH**	Liczba analizowanych płytek metafazalnych: 4
	Liczba komórek, w których liczono chromosomy: 0		Liczba analizowanych jąder interfazalnych: 1021
	Rozdzielczość prążkowa: niższa niż 200		Liczba badanych preparatów: 6
Wynik badania: 45,XY,dic(9;12)(p13;p?11.2~12)[19]/46,XY[7] **ish dic(9;12)(CDKN2A-,D9Z3+)[2]** **ish dic(9;12)(5'ETV6-,3'ETV6-)[2]** **nuc ish(CDKN2Ax1,D9Z3x2)[159/200],(ABL1x2,BCRx2)[221],(KMT2Ax2)[200],(ETV6x1)[167/200],(TCF3x2)[200]** Zapis wyniku według aktualnego wydania ISCN			

Fig. 1. An exemplary original fragment of cytogenetic test report.

The report on cytogenetic analysis of a blood or bone marrow sample contains the information about its origin (blood or marrow), its properties and it also contains specific information about methods and molecular probes used for analysis. Then the parameters of used methods are presented.

Parameters of GTG banding:

- number of fully analysed metaphases,
- number of metaphases with counted chromosomes,
- band level resolution.

Parameters of FISH technique:

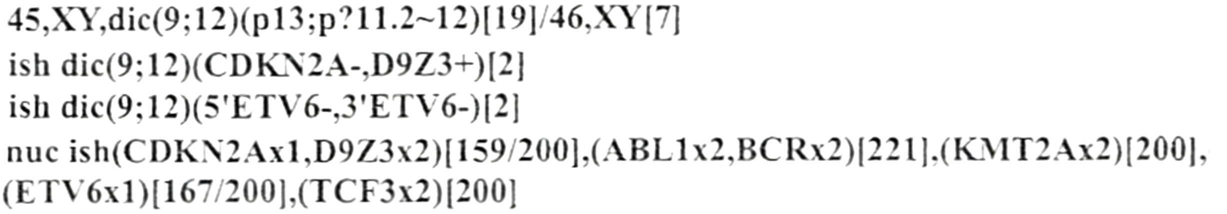
- total number of examined metaphases,
- total number of examined interphase nuclei,
- total number of examined microscopic slides.

Finally, the text result of cytogenetic test and its verbal interpretation is given at the end of the report.

2.1. ANALYSIS OF CYTOGENETIC TEST RESULTS REPORT

The first part of the report regarding the analysed sample refers to information related to the quality of the sample and the test methods used. This information is presented once in the report. The second one represents text record of the cytogenetic test result.

The result of is formatted as a complex string. Each karyotype starts a with total number of chromosomes and followed by a comma sex chromosomes. Subsequently, chromosome abnormalities are described. Sex chromosome aberrations are presented first. Then, the FISH test results are presented. The result string of cytogenetic test is presented in the Fig.2.



```
45,XY,dic(9;12)(p13;p?11.2~12)[19]/46,XY[7]
ish dic(9;12)(CDKN2A-,D9Z3+)[2]
ish dic(9;12)(5'ETV6-,3'ETV6-)[2]
nuc ish(CDKN2Ax1,D9Z3x2)[159/200],(ABL1x2,BCRx2)[221],(KMT2Ax2)[200],
(ETV6x1)[167/200],(TCF3x2)[200]
```

Fig. 2. The result of the cytogenetic test saved in accordance with ISCN 2016

The aspects of GTG and FISH results syntax that were mainly took into account in the research are presented below.

Different analysed clones are separated by a one slant line (/) as shown in the Fig.2. In the case of karyotyping of persons after bone marrow transplantation, the karyotype of both the recipient and the donor is determined during the study. These karyotypes are separated by a double slant lines (//).

Present aberrations (numerical or structural chromosome rearrangements, chromosome breakage etc.) are listed and separated with comas. A plus (+) or minus (-) signs can be put before a chromosome or an aberration. These signs indicate additional or missing chromosomes. It also may mark normal or abnormal chromosomes.

Aberrations or ranges of chromosome or region band may be indicated as uncertainty. Uncertainty can be marked in several positions with the question mark (?) and approximate sign (∼) respectively. The question mark denotes questionable identification of a chromosome or it structure. The approximate sign indicates intervals or boundaries of a chromosome segment.

The FISH results presenting observations on normal chromosomes are written by the symbol ish followed by the symbol of structural abnormality and then by the chromosome, region band,

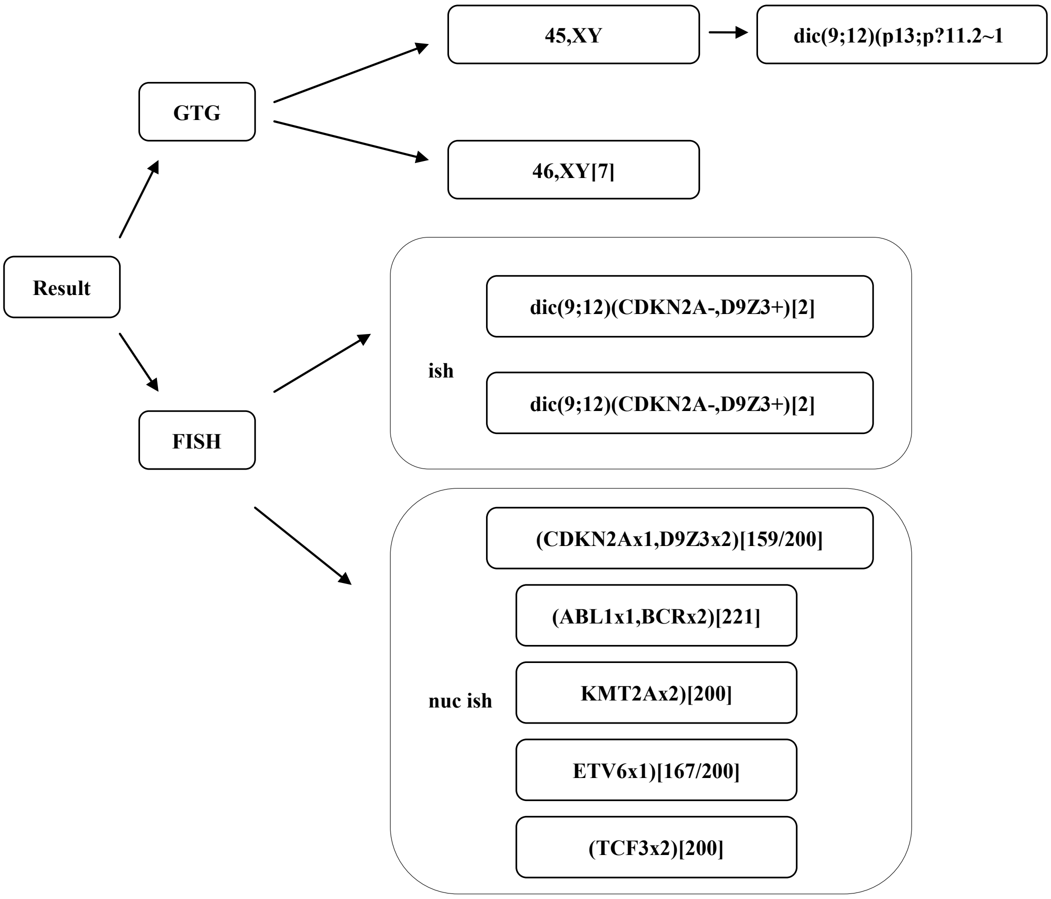


Fig. 3. The generic structure of the text GTG and FISH cytogenetic test result

or sub-band designation of the locus or loci tested. After that a probe or clone name or accession number or GDB (Genome Database) D-number can be given. After this section a status of locus can be specified. Gene name following with the multiplication sign (x) describes a number of signals seen, their uncertainty and their relative positions.

Full description of the text cytogenetic test result syntax is presented in [8].

Analysis of the text results of the cytogenetic tests showed that its structure is similar to an asymmetrical tree due to the variety of conducted tests and diagnostic probes used. That implicates high probability of lack of some tree elements or their presence on different levels of the tree.

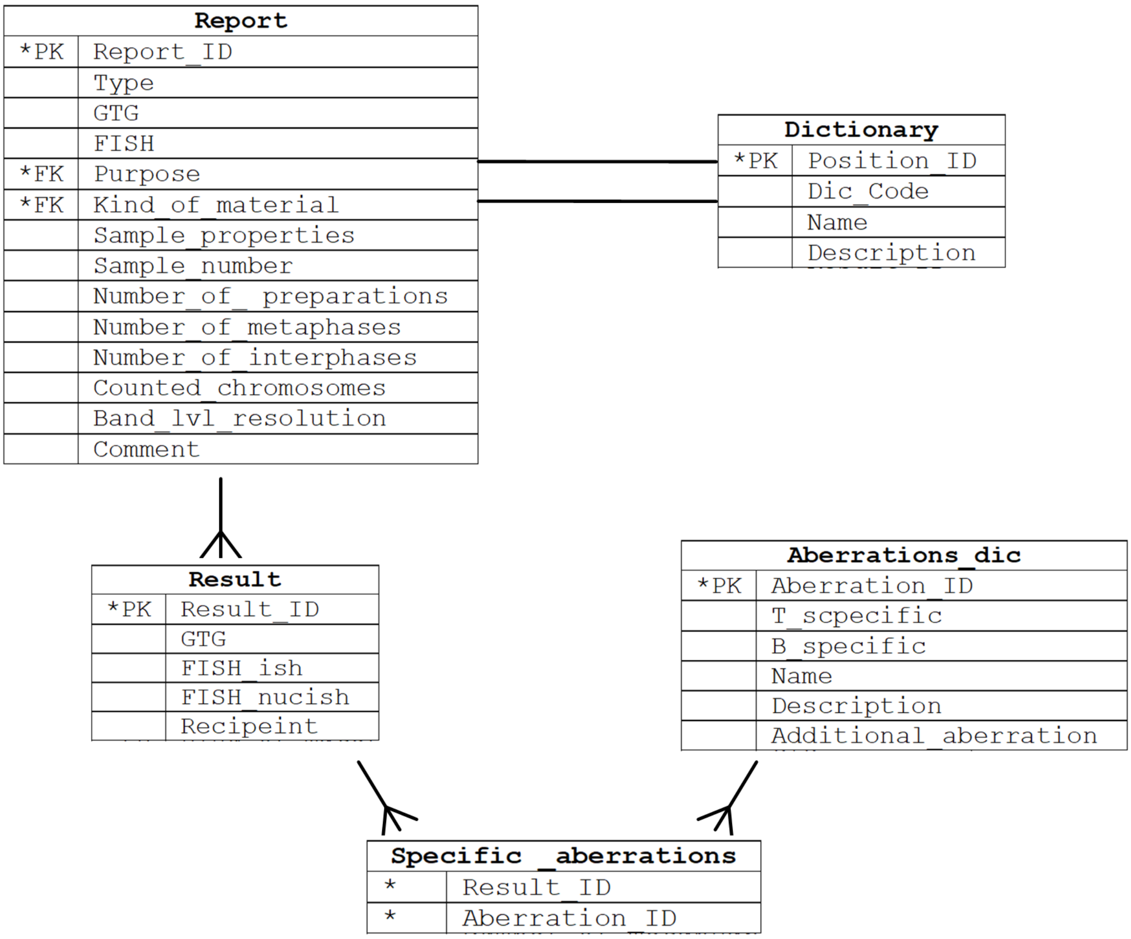
The tree consists of two main first level nodes: GTG results and FISH results. GTG section includes chromosome designation and its abnormalities originating from all analysed clones. FISH section may be divied into two techniques results: metaphase FISH (ish) and interpase FISH (nuc ish). The generic structure of the cytogenetic result text is shown in the Fig.3. Each chromosome or its aberration designation may involve characteristic syntax and its permissible uncertainty.

An important element of the structure is the occurrence of the results from different clones or different cells if the donor and recipient cells are present in the case of bone marrow transplantation.
Verbal interpretation of the result placed at the end of the report was not analysed.

2.2. PROPOSED STRUCTURES OF DATABASES

Prior to the development of database structures containing the results of cytogenetic studies it was assumed two structures of databases will be designed. The main one will be a structure that will store the results of cytogenetic research collected during the project and will be available to all authorized system users (doctors, authorized laboratory staff etc.).

The second structure will be an analytical structure and will enable the recording of the anonymised test result, reconstruction of its structure, sample and mass analysis.



Report	
*PK	Report_ID
	Type
	GTG
	FISH
*FK	Purpose
*FK	Kind_of_material
	Sample_properties
	Sample_number
	Number_of_ preparations
	Number_of_metaphases
	Number_of_interphases
	Counted_chromosomes
	Band_lvl_resolution
	Comment

Dictionary	
*PK	Position_ID
	Dic_Code
	Name
	Description

Result	
*PK	Result_ID
	GTG
	FISH_ish
	FISH_nucish
	Recipeint

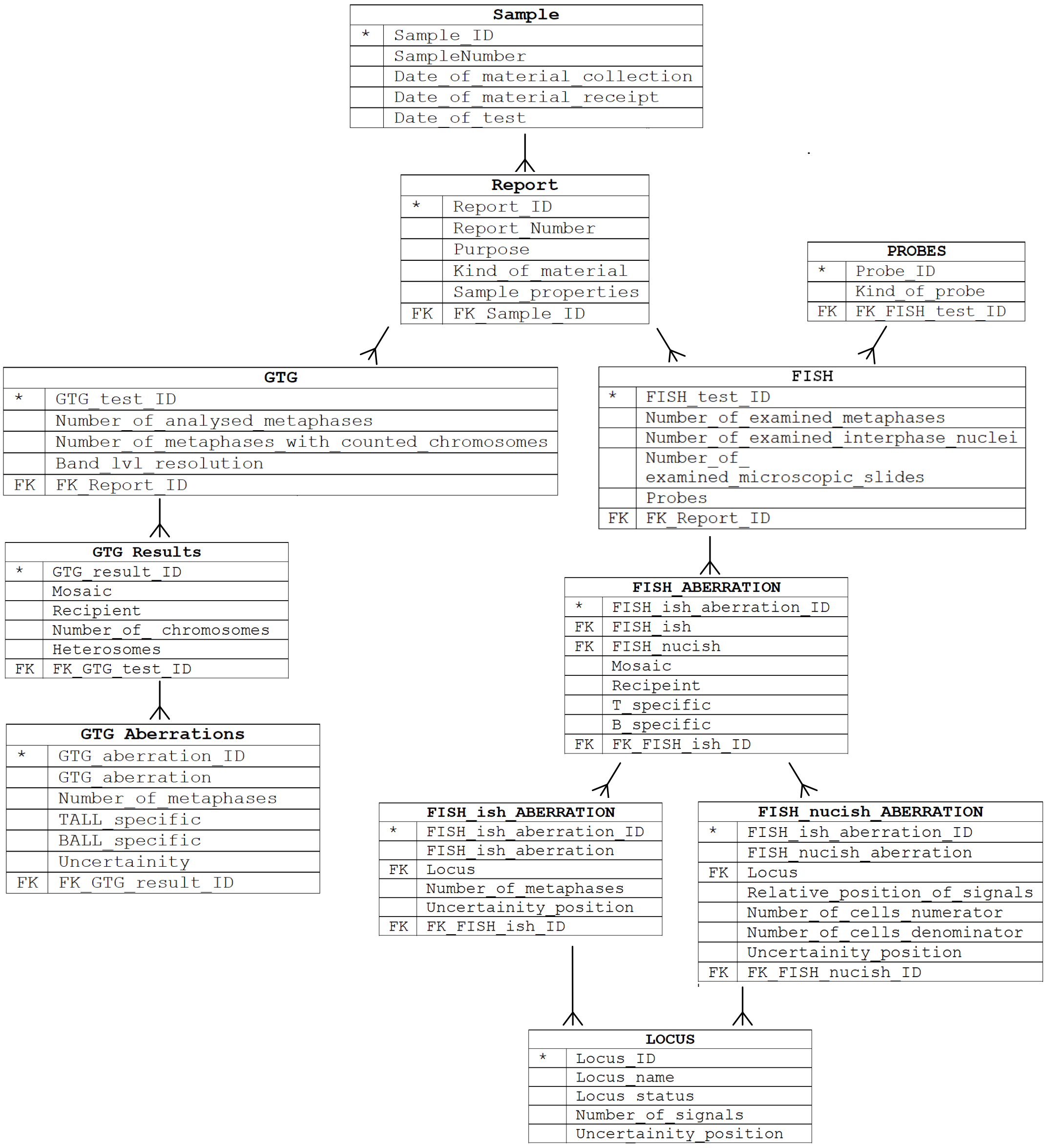
Aberrations_dic	
*PK	Aberration_ID
	T_scpecific
	B_specific
	Name
	Description
	Additional_aberration

Specific _aberrations	
*	Result_ID
*	Aberration_ID

Fig. 4. Database structure for storing data containing the cytogenetic results of the study.

Both database structures enable storing information about properties of the tested sample as well as characteristic parameters of the cytogenetic test methods used. The first structure of the database allows storing data containing the results of the study, taking into account the possibility of entering information whether in the result was an aberration characteristic of T-ALL, B-ALL or for both types of leukemia. Proposed structure of database is shown in the Fig. 4.

The second one is the foundation for creating queries for specific loci and its status, aberrations, chromosome or aberration uncertainty. In the case of FISH results the the number of signals and its relative positions is also stored. The second structure is shown in the Fig. 5 All these detailed information make possible to create multivariate custom queries.



	Sample
*	Sample_ID
	SampleNumber
	Date_of_material_collection
	Date_of_material_receipt
	Date_of_test

	Report
*	Report_ID
	Report_Number
	Purpose
	Kind_of_material
	Sample_properties
FK	FK_Sample_ID

	PROBES
*	Probe_ID
	Kind_of_probe
FK	FK_FISH_test_ID

	GTG
*	GTG_test_ID
	Number_of_analysed_metaphases
	Number_of_metaphases_with_counted_chromosomes
	Band_lvl_resolution
FK	FK_Report_ID

	FISH
*	FISH_test_ID
	Number_of_examined_metaphases
	Number_of_examined_interphase_nuclei
	Number_of_ examined_microscopic_slides
	Probes
FK	FK_Report_ID

	GTG Results
*	GTG_result_ID
	Mosaic
	Recipient
	Number_of_ chromosomes
	Heterosomes
FK	FK_GTG_test_ID

	FISH_ABERRATION
*	FISH_ish_aberration_ID
FK	FISH_ish
FK	FISH_nucish
	Mosaic
	Recipeint
	T_specific
	B_specific
FK	FK_FISH_ish_ID

	GTG Aberrations
*	GTG_aberration_ID
	GTG_aberration
	Number_of_metaphases
	TALL_specific
	BALL_specific
	Uncertainity
FK	FK_GTG_result_ID

	FISH_ish_ABERRATION
*	FISH_ish_aberration_ID
	FISH_ish_aberration
FK	Locus
	Number_of_metaphases
	Uncertainity_position
FK	FK_FISH_ish_ID

	FISH_nucish_ABERRATION
*	FISH_ish_aberration_ID
	FISH_nucish_aberration
FK	Locus
	Relative_position_of_signals
	Number_of_cells_numerator
	Number_of_cells_denominator
	Uncertainity_position
FK	FK_FISH_nucish_ID

	LOCUS
*	Locus_ID
	Locus_name
	Locus_status
	Number_of_signals
	Uncertainity_position

Fig. 5. Database structure for cytogenetic result text analysis.

Additionally, the second database was filled with data records originated from Mittelman Database of Chromosome Aberrations and Gene Fusions in Cancer [9]. The set of acute lymphoblastic leukemia was used for test searching.
In the research the known chromosome aberrations for T-cell acute lymphoblastic leukaemia, B-cell acute lymphoblastic leukaemia and both T-ALL and B-ALL was used [1] [10] [6] [11]. Used aberration are shown in the Table 1.
Both structures were queried for existing known (T-ALL, B-ALL and both) aberrations. Furthermore, the second database was queried for uncertainty in known aberrations, relative positions of interphase FISH signals, all unknown aberrations, and prognostic loci abnormalities. Conducted tests demonstrated the the proposed structure is very useful for GTG and FISH cytogenetic test results text mining.

Table 1. The sets of known chromosome aberrations for T-ALL, B-ALL and both T-ALL and B-ALL that were used).

Known aberrations for B-ALL	Known aberrations for T-ALL
• KMT2A (MLL) – t(4;11)(q21;q23) – t(6;11)(q27;q23) – t(9;11)(p23;q23) – t(10;11)(p12;q23) – t(11;19)(p13;q23) • BCR-ABL1 – t(9;22)(q34;q11.2) • RUNX1 (AML1) – t(12;21)(p13;q22) • ETV6 (TEL) – t(12;21)(p13;q22) • TCF3 (E2A) – t(1;19)(q23;p13) – t(17;19)(q22;p13) • other abberations – t(8;14)(q24;q32) – t(2;8)(p12;q24) – t(8;22)(q32;q11.2) – del(6)(q21q25) – dic(9;20)(p13;q11.2)	• TCRA/D – t(1;14)(p32;q11.2) – t(1;14)(p34;q11.2) – t(5;14)(q35;q11.2) – t(10;14)(q24;q11.2) – t(11;14)(p13;q11.2) – t(11;14)(p15;q11.2) • TCRB – t(1;7)(p32;q34) – t(1;7)(p34;q34) – t(7;9)(q34;q32) – t(7;9)(q34;q34.3) – t(7;11)(q34;p13) – t(7;11)(q34;p13) – t(7;9)q34;p13.2) • TLX1 – t(10;14)(q24;q11.2) • TLX3 – t(5;14)(q35;q11.2) • the other abberations – del(1)(p32) – t(5;14)(q35;q32) – t(8;14)(q24;q11.2) – del(9)(p21) – t(9;12)(p24;p13) – t(9;12)(q34;p13) – t(10;11)(p12)(q14)
Known aberrations for both B-ALL and T-ALL	
• del(11)(q23) • del(12)(p12) • i(17)(q10)	

2.3. DATABASE OPERATIONAL PERFORMANCE

The presented database is the first database integrating research results and case reports from all over the country, available to Polish hematologists. Due to the fact that in Poland the average number of registered cases of leukemia in children is over 200 per year. It is predicted that over the next 10 years the size of the database will not exceed 4000 cases. Each query does not exceed 1 ms using an average class server or desktop computer.

3. SUMMARY AND CONCLUSIONS

Presented structures of the databases were developed for two purposes. Mainly, it was designed for storing cytegenetic results data obtained in the project. In this aspect database is useful for doctors, scientists and laboratory staff. In addition, the database allows to search for information about known chromosome aberrations in the examined population. The second structure of database was designed for results reconstruction and analysis having regard to the syntax of karyotypes and FISH observations records. It allows to use of custom SQL non-statndard queries for fast data searching and its mass analysis.

4. AKNOWLEDGMENTS

This work was supported by the Polish National Centre for Research and Development within the programme Programme Prevention and Treatment of Civilization Diseases - STRATEGMED III, grant number STRATEGMED3/304586/5/NCBR/2017 (PersonALL).

BIBLIOGRAPHY

[1] Atlas of Genetics and Cytogenetics in Oncology and Haematology. 2018. http://atlasgeneticsoncology.org/Anomalies/Anomliste.htmlB-ALL [Online; accessed 2018-11-02].

[2] ABRAMS, ZACHARY B. PEABODY A. L. H. N. A., PAYNEA P. R. O. Text mining and data modeling of karyotypes to aid in drug repurposing efforts. Stud Health Technol Inform., 2015.

[3] ABRAMS Z. B. A translational bioinformatics approach to parsing and mapping iscn karyotypes: A computational cytogenetic analysis of chronic lymphocytic leukemia (cll). 2016.

[4] BARTLETT J. M. S. Fluorescence in situ hybridization. Molecular Diagnosis of Cancer: Methods and Protocols, 2004. Humana Press, Totowa, NJ, pp. 77–87.

[5] EITAN, R. SHAMIR R. Reconstructing cancer karyotypes from short read data: the half empty and half full glass. BMC Bioinformatics, 2017. p. 18(1):488.

[6] HOELZER D., GOKBUGET N., OTTMANN O., PUI C.-H., RELLING M. V., APPELBAUM F. R., VAN DONGEN J. J., SZCZEPANSKI T. Acute lymphoblastic leukemia. ASH Education Program Book, 2002, Vol. 2002. American Society of Hematology, pp. 162–192.

[7] JORDE, L. CAREY J. B. M. Medical genetics. 2010. MOSBY ELSEVIER.

[8] MCGOWAN-JORDAN, J. SIMONS A. S. M. Iscn 2016 an international system for human cytogenomic nomenclature (2016)reprint of: Cytogenetic and genome research 2016. 2016. Karger Verlag.

[9] MITELMAN F J. B., (EDS.) M. F. "mitelman database of chromosome aberrations and gene fusions in cancer (2018). 2018. http://cgap.nci.nih.gov/Chromosomes/Mitelman [Online; accessed 2018-10-02].

[10] SOSZYNSKA, K. MUCHA B. D. R. S. K. D. E. K. A. W. M. H. O. The application of conventional cytogenetics, fish, and rt-pcr to detect genetic changes in 70 children with all. Annals of Hematology, 2008, Vol. 87. pp. 991–1002.

[11] SZCZEPANSKI, T. HARRISON J. V. D. C. Genetic aberrations in paediatric acute leukaemias and implications for management of patients. The lancet oncology, 2010, Vol. 11. pp. 880–9.