Towards an Integrated Framework for Clinico-Biological Data Management and Analysis: the Case of Chronic Lymphocytic Leukemia

Evangelia Minga, Athanasios Gkoufas, Anna Vardi, Evangelia Stalika, Anastasia Hadzidimitriou,
Kostas Stamatopoulos, Nicos Maglaveras, *Sr Member, IEEE,* and Ioanna Chouvarda, *Member, IEEE*

Abstract—**This paper addresses the challenges of gathering, analyzing and employing clinically meaningful biological information, for the advancement of translational medicine. In this respect we propose an approach for the integrated management and analysis of large quantities of clinico-biological data, including high-throughput sequencing data. The proposed concepts are applied in the case of Chronic Lymphocytic Leukemia, a paradigmatic age-related hematologic cancer, and the steps for data integration and information analysis are illustrated via the CLL-BIND framework.**

I. Introduction

In recent years, molecular biology and genetics have entered the era of massive data production, exploiting "blind screening" methodologies. Extracting the information that might prove relevant to clinical purposes from within large pools of data is maybe the most challenging step of modern biological research. Thus, the rationale behind the proposed work is the need for clinicogenomic integration at large. The ability to organize, validate and analyze large sets of clinicobiological data coming from different sources and methodologies is a cornerstone for such integration. On these grounds, we proceed to the development of a standardized clinic-biological data management and analysis system. Taking advantage of our expertise and long-standing interest in hematological malignancies, we propose Chronic Lymphocytic Leukemia (CLL), the most common malignancy of the elderly in the West, as a test case for the proposed framework named CLL-BIND.

Key to the understanding of lymphoid malignancies is the study of the repertoire of adaptive immunity receptors – immunoglobulins (IG) on B cells and T cell receptors (TR) on T cells, respectively. Thanks to molecular and cellular processes taking place during normal B and T cell maturation, the probability that two independent B- or T-cell clones carry exactly the same IG or TR gene rearrangement by chance alone is virtually negligible. Therefore, the IG or TR gene rearrangement(s) carried by each B- or T-cell clone, respectively, can be justifiably viewed as its corresponding molecular signature. Thus, not paradoxically, immunogenetic analysis of antigen receptors has proven instrumental for both diagnostic and research of lymphoid malignancies, including CLL, but also for the study of normal immune responses, immunodeficiency and autoimmunity.

In CLL, immunogenetics and immunoinformatics have been particularly rewarding fields. The use of certain IGHV (Immunoglobulin Heavy Variable) genes in the clonotypic IG rearrangements is non-random, while the load of somatic hypermutations (SHM) in the IGHV genes is an established independent prognostic factor [1-2]. Perhaps more importantly, the identification of quasi-identical ("stereotyped") IGs among different patients in 30% of all CLL cases provided evidence for the involvement of common antigen(s) in CLL ontogeny. Furthermore, it led to a molecular classification of the disease based on shared IG sequence features, paving the way for compartmentalized clinical and biological research [3].

Currently, the advent of high-throughput, next-generation sequencing (NGS) technologies holds the promise of an unprecedented depth of immunogenetic analysis which may reform our views of immune repertoire constitution. Presently, limitations and pitfalls exist in every aspect of the analytical procedure, leading to a less than truly comprehensive and/or reliable appreciation of the IG/TR gene repertoires in various settings [4]. Bioinformatics challenges include not only the computational burden of analyzing many thousands or millions of complex rearranged IG/TR sequences, but also the comprehensive and consistent annotation of sequences and their features, as well as truly representative repertoire results, meaningful clonotype detection and quantitation, and end-user-friendly but unambiguous visualization.

We are already in the process of collecting clinical and biological data (including immunogenetic information) from patients with CLL within our broad network of collaborative European centers in order to ultimately attain the aim of obtaining personalized profiles and approaches in this as yet incurable disease, linked to their antigen receptor sequence signatures. Comparing identical categories of data coming from different medical centers, we have come up against

Manuscript received July 30, 2013.

Evangelia, Minga, Anna Vardi, Evangelia Stalika, Anastasia Hadzidimitriou, and Kostas Stamatopoulos are with the Institute of Applied Biosciences (INAB), Center for Research and Technology Hellas (CERTH), Thessaloniki, Greece (eva.minga@gmail.com, kostas.stamatopoulos@gmail.com).

Ioanna Chouvarda, Athanasios Gkoufas and Nicos Maglaveras are with the Lab of Medical Informatics, Aristotle University of Thessaloniki, Greece. (phone: +302310999247; fax: +302310999263; e-mail: ioannach@auth.gr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

redundancies, semantic inconsistencies, problematic registrations effectively masking a large amount of information and, thus, hindering a reproducible, complete and in-depth analysis.

In this context, our aim is to develop: (i) a robust, sophisticated and sustainable database system which will allow collection of multi-originating data while at the same time providing integration and validation mechanisms for eliminating data inconsistency and redundancy; and, boost the performance of data searching, data sharing and statistical analysis, also helping clinical research by correlating different categories of data, and (ii) a bioinformatics pipeline, specifically designed for immunogenetic analysis in different biological scenarios, based on the output of the International imMunoGeneTics system (IMGT), the global reference in immunogenetics and immunoinformatics. In this paper, we present the basic steps taken so far towards the CLL-BIND integrated framework.

We aim at the establishment and promotion of an innovative and sustainable system for handling large sets of data that will lead to the development of common conceptual and methodological frameworks for every-day clinical and laboratory practice, including the integration of biological information in a manner that is usable by researchers and clinicians alike [5].

II. Methods

Although not yet fully integrated as an automate procedure, the CLL-BIND framework, consisting of the core databases, the data management functions, and the bioinformatics analysis pipeline, is depicted in Fig 1. More details about the different parts are given in the next subsections.

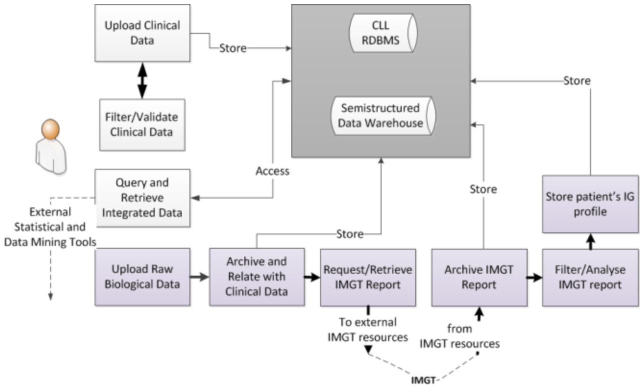

Upload Clinical Data
Store
CLL RDBMS
Filter/Validate Clinical Data
Store
Semistructured Data Warehouse
Access
Query and Retrieve Integrated Data
Store
Store patient's IG profile
External Statistical and Data Mining Tools
Store
Upload Raw Biological Data
Archive and Relate with Clinical Data
Request/Retrieve IMGT Report
Archive IMGT Report
Filter/Analyse IMGT report
To external IMGT resources
from IMGT resources
IMGT

Fig.1 overview of CLL-BIND framework.

A. Clinical and Biological Data Description

The clinico-biological data in focus of the foreseen integrated system include the following types:

1) Clinical data, comprised of the clinical setting at the time of diagnosis, prognostic and predictive markers, medical history related to diagnosed disease, follow-ups and treatment information provided by medical doctors; and, also additional events, composing all the necessary and significant information to form an accurate and complete representation of the clinical course of the disease.

2) Laboratory data coming from different methodologies/laboratories in each medical center and information gathered from laboratory analysis, followed by interpretation from authorized supervisors in each lab.

3) *NGS Data.* The test set of data analyzed involved IG and TR rearrangements of CLL patients sequenced on the Illumina NGS platform; however the presented bioinformatics pipeline is applicable to all platforms.

4) *IMGT/HighV-QUEST processing and Immunogenetics Reports.* IMGT/HighV-QUEST [6] is the high-throughput tool for the analysis of thousands of IG and TR rearranged nucleotide sequences (up to 150,000 sequences) per run. IMGT software utilizes "anchor" aminoacids in conserved positions of the IG/TCR rearranged genes so as to align sequences to their respective germline genes and delineate the complementarity determining regions (CDR) and framework regions (FR), taking into account the extent of natural polymorphism. The standard output is a series of text files with information per sequence on the recognition of genes involved as well as the aminoacids and nucleotides of specific areas of interest.

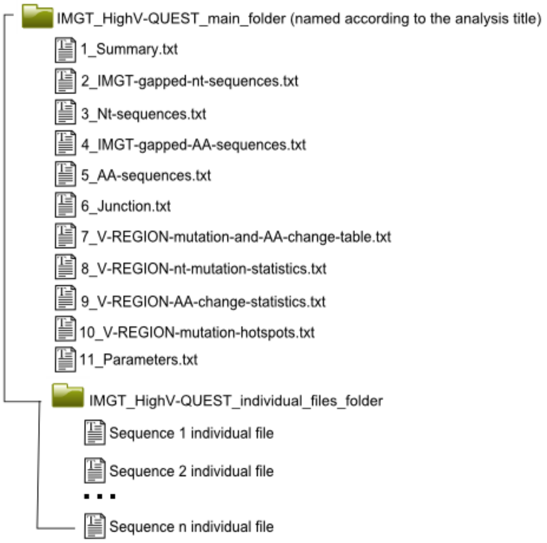

IMGT_HighV-QUEST_main_folder (named according to the analysis title)
1_Summary.txt
2_IMGT-gapped-nt-sequences.txt
3_Nt-sequences.txt
4_IMGT-gapped-AA-sequences.txt
5_AA-sequences.txt
6_Junction.txt
7_V-REGION-mutation-and-AA-change-table.txt
8_V-REGION-nt-mutation-statistics.txt
9_V-REGION-AA-change-statistics.txt
10_V-REGION-mutation-hotspots.txt
11_Parameters.txt
IMGT_HighV-QUEST_individual_files_folder
Sequence 1 individual file
Sequence 2 individual file
...
Sequence n individual file

Fig.2 The IMGTHighV-Quest structure of output report files that constitute the Immunogenetics information analysis (adopted from http://www.imgt.org/HighV-QUEST).

These files, as depicted in Fig. 2, contain the information that constitutes the basis for further immunogenetic processing described below, the results of which are stored in the clinicogenomic database framework.

B. Data Management and System Description

The core component of the system is a clinic-biological database. A relational database management system is used in order to develop, manage and monitor the database. The conceptual data model adopted in CLL-BIND is depicted in Fig. 3. Of special interest is the integration of Immunogenetics data interpretation in the clinicobiological database, as the patient's biological profile, consisting of a characterization of clonotypic IG/TR rearrangements, and

SHM/Intraclonal Diversification patterns (for IGs).

Additional tools are being developed to enhance data sharing, management and handling, as well as correlation and statistical analysis, including online data registration and error-correction tools, and a web interface with query forms and download sections to retrieve data and analyze results either online or offline. The semi-structured immunogenetic flat files are being stored in patient-specific file hierarchies linked with the main database. The system is being designed to easily be extended and also adapted/used for different purposes, always in the context of biomedical research.

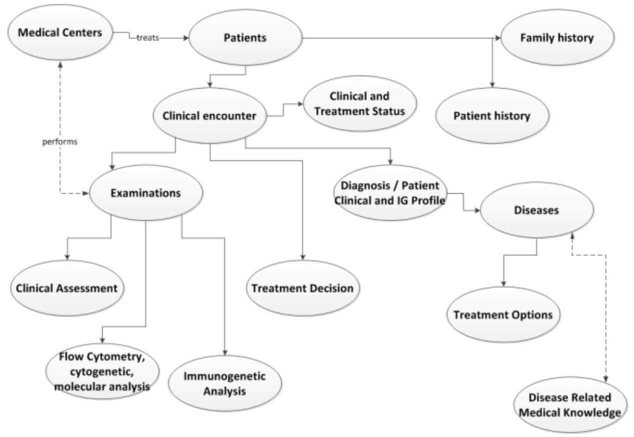


Fig. 3. Conceptual integrated clinicobiological data model.

C. Analysis Methodology

1) Data Quality in Clinical Data. Clinical data validation and consistency check are basic steps in complex and multi-centric databases. In the CLL-BIND framework, these are based on the following principles:

- Data validation, according to predefined types/lists
- Data consistency (language and typographical)
- Data integrity (reduction of internal incompatibilities)
- Semantic (term consistency, for example entities coming from WHO classification)
- Redundancy reduction (elimination of duplicates)
- Consistency of information, i.e. patient history, including, multiple follow-ups of the same patient, multiple experiments on the same sample/cells, and multiple molecular techniques on specific molecule
- Rules (constraints, e.g. date of treatment ≥ date of diagnosis, assumptions, e.g. diagnosis/patient ≥ 1)

2) Immunogenetic Data Filtering. Preprocessing and Filtering of the IMGT report data takes place to exclude potentially erroneous information. This includes a series of rules based on existing and ongoing experience as regards the NGS analyses, with characteristic examples:

- Filter out sequences with IGHV gene denoted as pseudogene or non-functional gene, especially when dealing with potentially functional repertoires
- Filter out non-productive/problematic sequences
- Filter out sequences where characteristic landmark aminoacids are not present e.g. CDR3 does not start with C or does not end with W (IG VH) or F (TR).
- Mark sequence groups with biologically questionable/ambiguous outcomes

3) IMGT/HighV-QUEST post-processing. Processing of IMGT semi-structured data files is necessary for the investigation of clonality, and characterization of patients' immunogenetic profile. The challenge here pertains to the large amount of data, as well as to the heterogeneity of data types, ranging from simple values, and strings, to DNA sequences, that need to be simultaneously handled, tackling big data issues. The two main axes of analysis include:

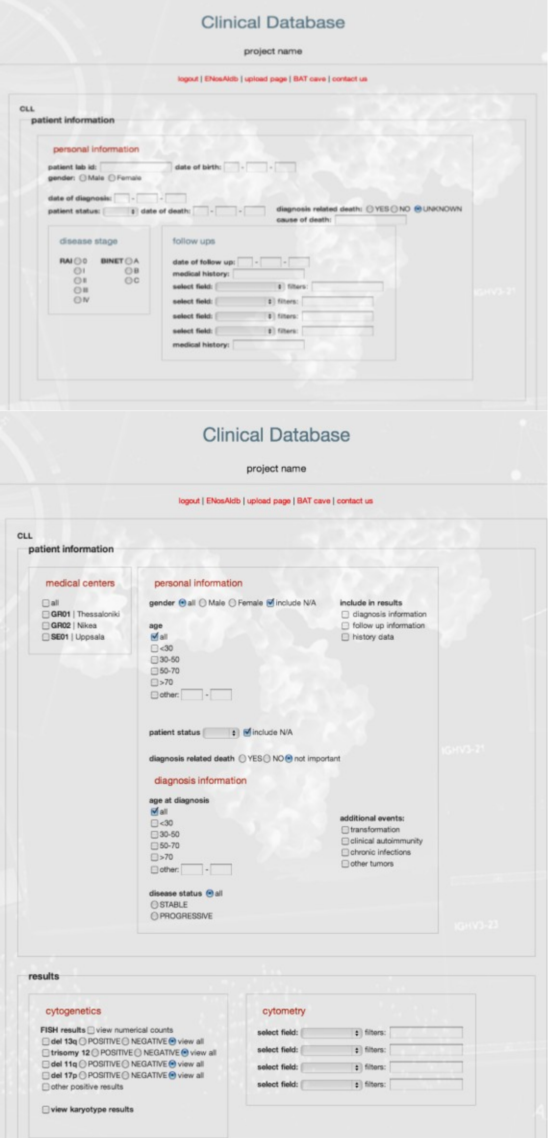
- Intra-subject repertoire [7] and statistical analysis of the IMGT reports, for the characterization of clonality/diversity. This part involves the identification of unique clusters of rearranged sequences (clonotypes) deriving from same germline genes and showing identity within the CDR3, inherently the most diverse and also the most critical part of the IG/TR in the process of interaction with antigen(s), and the distribution of clusters in a subject's set of samples.
- Inter-subject repertoire comparison and co-clustering of dominant clonotypes, aiming at the identification of shared (public) clonotypes, which alludes to recognition of and selection by common antigens.

Besides the generation of processing results, an important issue is the facilitation of access to summary and detailed information per subject, applying different selection criteria.

III. Results

Following the conceptual model depicted in Fig.2, a database has been created with the PostgreSQL, a powerful, open source object-relational database management system. Excel files were specifically designed, and have been used as templates for data collection by different centers, at the preparation stage before import into the database. A mechanism for the creation of new templates was setup via an online google document with all necessary information of the clinicobiological database. A web interface has been developed, to be made available for users to upload templates with clinical data (see Fig. 4). At this stage, clinical data consistency is supported by the following mechanism: a configuration file connects the templates with the database, a file with constraints not to be violated in each template and a file with synonyms for the translation of frequently used values to predefined values of the database. Tools to export clinical data or filter/query information online are being created and used through a web interface.

As regards the immunogenetic analysis step, a pipeline of Python scripts has been developed, mainly based on Pandas library for big data analytics, and handling of large semi-structured data. Pandas (http://pandas.pydata.org/) is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language, enabling the implementation of an entire data analysis workflow in Python without having to switch to a more domain-specific language like R (although binding with R is possible). This workflow can be both seen as an independent NGS analytics solution, e.g. via KNIME (http://www.knime.org/), or can be launched via CLL-BIND framework, in order to enable processing and visualization of the results (see Fig 5), and more importantly, embodiment of the NGS-based immunogenetic profile in the clinicobiological database. Methods from other clinicogenomic approaches will be taken into account [8].



Clinical Database

project name

logout | ENosAldb | upload page | BAT cave | contact us

CLL

patient information

personal information

patient lab id: date of birth: - -

gender: ○ Male ○ Female

date of diagnosis: - -

patient status: date of death: - - diagnosis related death: ○ YES ○ NO ● UNKNOWN

cause of death:

disease stage

RAI ○ 0 ○ I ○ II ○ III ○ IV BINET ○ A ○ B ○ C

follow ups

date of follow up: - -

medical history:

select field: filters:

select field: filters:

select field: filters:

select field: filters:

medical history:

Clinical Database

project name

logout | ENosAldb | upload page | BAT cave | contact us

CLL

patient information

medical centers

□ all
□ GR01 | Thessaloniki
□ GR02 | Nikea
□ SE01 | Uppsala

personal information

gender ● all ○ Male ○ Female ☑ include N/A

include in results
□ diagnosis information
□ follow up information
□ history data

age
☑ all
□ <30
□ 30-50
□ 50-70
□ >70
□ other: -

patient status ☑ include N/A

diagnosis related death ○ YES ○ NO ● not important

diagnosis information

age at diagnosis
☑ all
□ <30
□ 30-50
□ 50-70
□ >70
□ other: -

additional events:
□ transformation
□ clinical autoimmunity
□ chronic infections
□ other tumors

disease status ● all
○ STABLE
○ PROGRESSIVE

results

cytogenetics

FISH results □ view numerical counts
□ del 13q ○ POSITIVE ○ NEGATIVE ● view all
□ trisomy 12 ○ POSITIVE ○ NEGATIVE ● view all
□ del 11q ○ POSITIVE ○ NEGATIVE ● view all
□ del 17p ○ POSITIVE ○ NEGATIVE ● view all
□ other positive results

□ view karyotype results

cytometry

select field: filters:

select field: filters:

select field: filters:

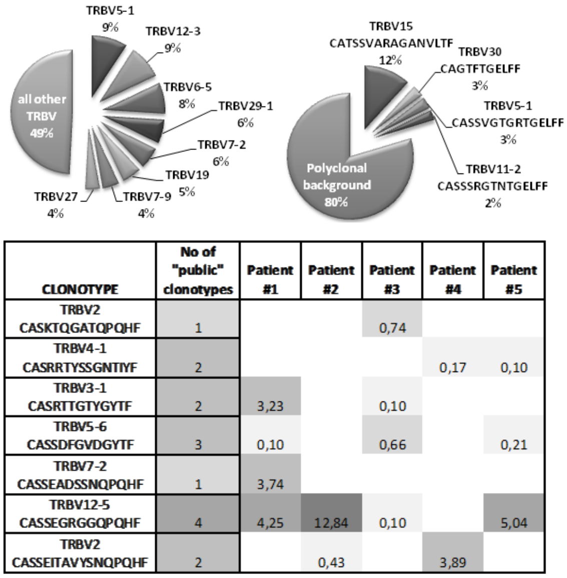
select field: filters:

Fig. 4. Two views of the currently developed web interface (top) data submission, and (bottom) query and data retrieval.

IV. Conclusion

The general concepts and the working version of CLL-BIND framework are presented in this paper. This is a work in progress, towards the standardization of the procedures and workflows [9], as well as their extensibility and user-friendliness. In parallel, new challenges arise, for example, involving grid technologies for more efficient handling of computational needs, as well as big data management, processing and visualisation challenges.

There is an important implication of the CLL-BIND integration paradigm in personalised medicine: thorough characterization of patients combining clinical and biological information and organization of such information in a readily available manner is a prerequisite for sophisticated subgrouping and efficient design of tailored therapeutic interventions.



CLONOTYPE	No of "public" clonotypes	Patient #1	Patient #2	Patient #3	Patient #4	Patient #5
TRBV2 CASKTQGATQPQHF	1			0,74		
TRBV4-1 CASRRTYSSGNTIYF	2				0,17	0,10
TRBV3-1 CASRTTGTYGYTF	2	3,23		0,10		
TRBV5-6 CASSDFGVDGYTF	3	0,10		0,66		0,21
TRBV7-2 CASSEADSSNQPQHF	1	3,74				
TRBV12-5 CASSEGRGGQPQHF	4	4,25	12,84	0,10		5,04
TRBV2 CASSEITAVYSNQPQHF	2		0,43		3,89	

Fig. 5. NGS results visualization. Top) The TRBV gene repertoire of a subject, and the dominant T cell clonotypes (>1% of the total repertoire) of a subject. Bottom) Inter-subject comparison and identification of shared T cell clonotypes, using color coding. In the patient columns, numbers correspond to the % frequency of the respective clonotype in the patient's total repertoire.

References

[1] F Fais, F Ghiotto, S Hashimoto, et al., ‘Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors’, *J Clin Inves.* vol 102, 1998, pp. 1515-25.

[2] RN Damle, T Wasil, F Fais, et al, ‘Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia’, *Blood*, vol 94 no 6, 1999, pp.1840-7.

[3] A Agathangelidis, N Darzentas, A Hadzidimitriou, et al, ‘Stereotyped b-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with complications for targeted therapies’, *Blood*, vol 119 no 19, 2012, pp. 4467-75.

[4] DA Bolotin, IZ Mamedov, OV Britanova, et al., ‘Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms’, *Eur J Immunol.* vol 42, 2012, pp. 1-11.

[5] V Marx, ‘Biology: The big challenges of big data’, *Nature*, vol. 498, no. 7453, 2013 , pp. 255-60.

[6] E Alamyar, P Duroux, MP Lefranc, V Giudicelli.’ IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS’, *Methods Mol Biol.*, vol 882, 2012, pp. 569-604.

[7] PD Baum, V Venturi, DA Price, ‘Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors’, *Eur J Immunol*. 2012, pp. 2834-9.

[8] L, Martin, A, Anguita N. Graf et al, ACGT: advancing clinico-genomic trials on cancer - four years of experience. Stud Health Technol Inform. 2011;169:734-8.

[9] P Lindenbaum, S Le Scouarnec, V Portero, R Redon, ‘Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME’, *Bioinformatics,* vol 27, no 22, 2011, pp. 3200-1.