INTEGRATING MULTIPLE IMMUNOGENETIC DATA SOURCES FOR FEATURE EXTRACTION AND MINING MUTATION PATTERNS: THE CASE OF CHRONIC LYMPHOCYTIC LEUKEMIA SHARED MUTATIONS

IOANNIS KAVAKIOTIS[1,*], ALIKI XOCHELLI[2], ANDREAS AGATHANGELIDIS[3], GRIGORIOS TSOUMAKAS[1], NICOS MAGLAVERAS[2,4], KOSTAS STAMATOPOULOS[2], ANASTASIA HADZIDIMITRIOU[2], IOANNIS VLAHAVAS[1], IOANNA CHOUVARDA[2,4]

[1] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] Institute of Applied Biosciences, CERTH, Thessaloniki, Greece
[3] Division of Molecular Oncology and Department of Onco-Hematology, San Raffaele Scientific Institute, Milan, Italy
[4] Laboratory of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece
[*]Correspondence to: ikavak@csd.auth.gr

Introduction: The aim of this work is to extract features and create high quality datasets through integration of multiple information resources for somatic hypermutation (SHM) analysis in the clonotypic immunoglobulin (IG) receptors of patients with Chronic Lymphocytic Leukemia (CLL). This can set the basis for an in-depth investigation of a series of as yet unanswered biological questions, through data mining analysis, which is clinically relevant given the great prognostic value of SHM in CLL (Damle et al, 1999). The virtue of the proposed approach is illustrated via the case of "towards analysis" which is our attempt to identify potential developmental transformation or movement of IG gene germlines towards other IG gene germlines through SHM.

Methods:

A. Input Datasets

The analysis uses three different dataset inputs for the feature extraction process.

1. IMGT/HighV-QUEST Output

The first dataset is a collection of results obtained from IMGT/HighV-QUEST alignment analysis (Giudicelli et al., 2011) in a single run of a set of patient cases i.e. sequences. This collection consists of several files containing information about the analyzed sequences such as identified genes, nucleotide mutations, amino acid changes and various related statistics (Brochet et al., 2008; Giudicelli et al., 2011).

2. Reference Dataset

The reference dataset consists of the amino acid and nucleotide germline sequences of *Homo sapiens* IGHV genes obtained from IMGT/GENE-DB (Giudicelli et al., 2005). These are organized in a hierarchical manner of alleles-genes-subgroups-clans.

3. Classification of Patient Sequences to Stereotyped Subsets

The third dataset is the result of clinicobiological database quering that classifies various types of clinical and biological data, including the sequence of the clonotypic IG genes and relevant characteristics, such as shared IG motifs, defining subsets with stereotyped B-cell receptor (BcR) (Stamatopoulos et al., 2007). The latter is an example of how clinicobiological contextual information can distinguish groups of patients with regards to their unique biological features and clinical behavior. It can also be helpful for data mining depending on the question under investigation.

B. Data Preprocessing

The first step in feature extraction process is the data preprocessing step, whose aim is twofold. First, to integrate the different data sources and, second, to ensure the data quality.

1. Data Integration

The analysis is patient-orientated and, therefore, the key behind the data integration is the patient unique ID in the patient related datasets. The first step of data integration is the parsing of the IMGT/HighV-QUEST output files and Stereotyped Subset characterization file. Information obtained for each sequence includes: PatientID, functionality of the IGHV-IGHD-IGHJ gene rearrangement (productive/unproductive), closest germline VGene and Allele, germline percentage identity, the nucleotide and amino acid gapped sequence according to IMGT numbering (Lefranc et al., 2003) and the list of nucleotide mutations, amino acid changes and their class identity or change.

2. Filtering Integrated Data

In this step several filters have been developed firstly to ensure the high data quality and secondly to choose the appropriate subsets/subgroups for further analysis.

a. Ensuring Data Quality

In this category belong filters that exclude unqualified sequences such as sequences that have sequence ambiguities or unproductive IGHV-IGHD-IGHJ gene rearrangement sequences.

b. Subgroups of Analyzed Sequences

Moreover, filters have been developed in order to direct the analysis to a specific subgroup of the analyzed sequences. They concern the selection of sequences that belong to specific stereotyped subsets, have specific range of IGHV gene germline percentage identity, carry the same IGHV gene. A further focus is on specific VH domain subregions (e.g. heavy variable CDR1).

C. Shared Mutation Analysis Feature Extraction

In this study we introduce a new term, namely "Shared Mutation (SM)". We refer to a mutation as shared if the nucleotide introduced by SHM is present as germline at the same position. Moreover, in this analysis we refer to the closest VGene germline as "Sequence Before the Mutation (Bef)" and to the patient sequence as

the "Sequence After the Mutation (Af)". Finally, we call the germline with the shared mutations "Towards Germline" (Tow) because the shared mutation indicates a potential movement from the Bef sequence towards this germline. In this part of the feature extraction process we examined all mutations of the dataset in order to determine for each patient sequence, all shared and non-shared mutations. The result of this feature extraction process is the production of three datasets, described below.

1. *Shared Mutations Position Dataset (SMPD)*

Each entry of the SMPD is a SM with a specific germline i.e. the Tow. The features that have been produced include for each entry: PatientID, patient's stereotyped subset, Tow germline, number of shared mutations with the Tow germline, and positional information of the mutation and amino acid properties based on the IMGT scientific chart (Lefranc et al., 2009). This results in a dataset with 34 features per sequence.

2. *Non-Shared Mutations Position Dataset (nSMPD)*

We define as "non-Shared Mutation" a mutation that resulted in a nucleotide that cannot be found in any germline sequence at this particular position. This dataset contains almost the same information as SMPD, only replacing the amino acid properties information about the Tow Germline with a column that indicates the functionality subgroup where the new property can be found.

3. *Germlines with Shared Mutations Dataset (GSMD)*

The last dataset of the Shared Mutation Analysis contains a list of the Tows for each patient sequence ordered by the number of the Shared Mutations.

Results: The Case of Towards Analysis

We have previously reported distinctive SHM patterns amongst CLL cases utilizing the IGHV4-34 gene, especially subsets with stereotyped BCR (Murray et al. 2008). With this analysis we sought for differentiation trends concerning shared mutations. The analysis is focused on IGHV4-34-expressing subsets #4, #11, #16, #29 and #201.

The integrated dataset was based on the following 3 datasets: 1) the alignment results obtained from IMGT/HighV-QUEST output for a set of 16,528 CLL cases, 2) IMGT/GENEBD Version 3.1.0 (4 April 2014) reference genes, and 3) the results of querying the clinicobiological database for CLL (http://www.imgt.org/CLLDBInterface/query).

In order to ensure data quality we discarded through filtering all unproductive rearrangement sequences, incomplete and upstream regions and sequences that had genotype errors. Following, we selected only sequences that were classified to the above mentioned stereotyped subsets (#4 – 156 cases; #11 – 16 cases; #16 – 41 cases; #29 – 39 cases; and #201 – 43 cases).

For our analysis, we constructed five datasets, one for each subset. For the towards analysis we used only the GSMD. In towards analysis each patient sequence was represented by its closest germline which in our case was IGHV4-34 alleles. For each allele, every Tow germline was scored by the following formula:

$$Score=SM*(SM/M)*(1/SMP)*(1/TVGM)$$

where SM is the number of shared mutations with this Tow germline., M is the number of mutations in this patient sequence, TVGM (Total VGene Mutations) is the total number of mutations in patient sequences with the same closest germline and SMP is the number of Tow germlines that have the same number of Shared Mutations in this patient sequence.

Subsets 16, 4, 29 and 201 showed an even distribution of shared mutations with germline IGHV genes belonging to IGHV clan II (as IGHV4-34 gene) vs other clans. More specifically, 43.2 Vs 56.8% for subset #16; 41.1 Vs 58.8% for subset#4; 45.2 Vs 54.8% for subset #29, 52.8 Vs 47.2% for subset #201. In contrast, subset #11 showed a statistically important trend (z-test; $p<0.0001$) towards IGHV genes of other clans (13.5 Vs 86.5%) suggesting a distinct ontogenetic pathway.

Discussion - Conclusion: We herein present an approach for an integrated feature extraction framework that can set the basis for exploratory analysis or a fully automated computational data mining approach, the latter shortly demonstrated via a scenario for Tow exploration. Future steps include the wider exploitation of the proposed feature set in automated knowledge discovery procedures accompanied by expert's validation.

References:

Brochet, X. et al., (2008) *IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis*. Nucl. Acids Res. 36, W503-508

Damle R. N., et al, (1999) *Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia*, Blood, vol 94 no 6, pp.1840-7.

Giudicelli, V. et al., (2005*) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes*.Nucleic Acids Res., 33: D256 - D261. PMID: 15608191

Giudicelli, V., et al., (2011) *IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences*. Cold Spring Harb Protoc. 2011 Jun 1;2011(6). pii: pdb.prot5633.

Lefranc, M.-P. et al., (2003) *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains* Dev. Comp. Immunol., 27, 55-77

Lefranc, M.-P. et al., (2009*) IMGT®, the international ImMunoGeneTics information system*® Nucl. Acids Res, 37, D1006-D1012;

Murray F, et al., (2008) Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. Blood. 2008 Feb 1;111(3):1524-33.

Stamatopoulos K, et al., (2007) *Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations*. Blood. 2007 Jan 1;109(1):259-70.