MEDGENET

Medical Genomics and Epigenomics Network

H2020-TWINN-2015-692298

1st on-line book of abstracts WP1

December 2016

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 692298.
This report reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

Table of Contents

1 Introduction

The "Medical Genomics and Epigenomics Network" (MEDGENET) project is a TWINNING H2020 project which aims to create a well-educated taskforce of biomedical researchers, who will contribute to the development of new genomics and bioinformatics tools and their application in clinical practice, and enable establishing best practices for performing innovative and high quality applied research.

In order to achieve the stated aim various workshops are organized in the project. In order to share the ideas discussed in these workshops, and therefore, widen their impact beyond the workshop participants ***books of abstracts*** will be published on the MEDGENET website in Section Study Materials. This is ***the 1st online book of abstracts*** that presents abstracts from the workshops that took place in the Work Package 1"Knowledge transfer in medical genomics, gene expression control and bioinformatics" in year 2016.

The Work Package 1 focuses on educating and training new generation of genomics and bioinformatics scientists as well as transfer of methodological proficiency to sustain technological platform for genomic research at CEITEC MU. It aims to help to develop expertise for gene expression regulation studies and performing functional studies on important signalling pathways in disease development, microenvironment interactions and resistance.

Tree workshops were organized in the Work Package 1 in year 2016:

- The 6th Educational workshop of the IgCLL group workshop at Uppsala University, Sweden
- Structural Validation Workshop at CEITEC MU, the Czech Republic
- Workshop on Grid and Cloud technologies for high-performance high-throughput bioinformatics at INAB-CERTH, Greece

Abstracts form these workshops are presented in the next chapters.

2 6th IgCLL Educational Workshop Immunoglobulin Gene Analysis in CLL: From Prognosis and Prediction to Follow-up

6th IgCLL Educational Workshop Immunoglobulin Gene Analysis in CLL: From Prognosis and Prediction to Follow-up took place on 22nd and 23rd September 2016. The workshop, organised in Uppsala in September, focused on recent developments in immunogenetics and MRD and novel analytical strategies relevant for both research and clinical diagnostics/prognostication. Its aim was bringing together basic research and clinical practice.

This two-day event consisted of several sessions including lectures by experts in the field, computer-based practical exercises, and panel sessions which provided participants with an opportunity to share ideas, knowledge, and experiences.

2.1 KEYNOTE LECTURES ABSTRACTS

Chair: **Richard Rosenquist** (SE)

Dimitar Efremov (IT) - *B cell receptor signaling pathways in health and disease*

The lecture provided the introduction into the B cell receptor (BCR) signaling function in the context of normal but also leukemic conditions. As the BCR mediates cellular processes such as B cell proliferation, differentiation, anergy etc., it is not surprising that BCR signalling is involved in development of B cell malignancies. Therefore, also BCR signaling inhibitors were discussed as they have been introduced to the clinical testing and actual use. Deep understanding of the BCR pathway action is necessary prerequisite for better management of B cell malignancies in future.

Lesley Ann Sutton (SE) **-** *Immunogenetics in CLL: the last 20 years*

Dr. Sutton recapitulated the most important immunogenetic discoveries from the last 20 years that formed current knowledge of chronic lymphocytic leukemia (CLL) biology. Initially, it was discovered that certain IGHV genes are overrepresented in CLL, which suggests antigen involvement in CLL cell selection and the disease initiation. Such presumption was further supported through gene expression profiling, CLL cell phenotype assessment, identification of certain IG heavy and light chain combinations, and identification of stereotyped receptors. All these phenomena were described in detail.

2.2 SESSION I – IG SEQUENCE ANALYSIS

Chair: **Kostas Stamatopoulos** (GR)

Frederic Davi (FR) - *Mechanisms of immunoglobulin diversity*

Anton Langerak (NL) - *From the patient to the sequence: Primers,*

PCR, Detection of clonality, Sequencing

Veronique Giudicelli (FR) - *"IMGT tools for Interpretation of IG sequences and NGS*

repertoires: 2016 novelties"

Andreas Agathangelidis (GR) - BcR IG *Stereotypy: Identifying CLL subsets*

The main focus of first day programme was on immunoglobulin (IG) sequence analysis: Starting with theory, detailed lecture on IG / BCR structure and their assembly mechanisms, including semantics of immunoglobulin genes and their rearrangements, was provided by Frederic Davi. Anton Langerak continued with methodology of IG sequencing and sequence interpretation – from pick of material and nucleic acid, through PCR methodology, to clonality determination and sequencing. At this occasion, results of the survey from the previous workshop on preferred methodologies among workshop participants were presented. In the next two presentations, Veronique Giudicelli introduced features of ImMunoGeneTics Information system widely used for analysis of IG rearrangements and Andreas Agathangelidis provided theoretical bases of stereotyped BCR subsets in CLL together with ways of their identification.

2.3 SESSION II – IG PRACTICAL EXERCISES

Vasilis Bikos (CZ) **-** Hands-on exercises I

Vasilis Bikos (CZ) - *Sequence interpretation and results*

The theoretical part was followed by hands-on exercises lead by Vasilis Bikos. In this session of the workshop the participants will be given a number of selected sequences of immunoglobulin (Ig) heavy and light chain rearrangements. The purpose of this is for the participants to comprehensively understand the workflow of the analysis of Ig gene rearrangements by utilizing the specified bioinformatics tool IMGT/ V-QUEST. The session will include hands-on exercises and the participants will be given an answering sheet, in order to record the molecular features of each of the rearrangements, namely gene assignment, percentages of gene identity and CDR3 region characteristics, as well as comment on the general information that may be provided via this analysis.

Ultimately, the session will conclude with a tutorial presentation guiding through the analysis of the sequences. Due to the careful selection sequence analysis matters such as functionality of the gene rearrangement, proper identification of the somatic hypermutation status, double IG V-gene assignment, existence of insertions/deletions, recognition of CLL subset cases, absence of CDR3 anchors and proper sequence alignment will be highlighted and discussed. Moreover and where appropriate solutions on how to work around problems with sequence analysis both on the bench (amplification repeat, propositions on primer sets, usage of the appropriate genomic material etc.) and on the bioinformatics level will be discussed, as well as matters concerning the interpretation of the results at the level of a report to the clinician.

2.4 SESSION III – IG GENES FOR CLINICAL DECISION-MAKING

Richard Rosenquist (SE) **-** *CLL genetics and prognostication*

Paolo Ghia (IT) - *Targeting signaling pathways for treating CLL*

Round Table Discussion/Debate:

IG genes or genetic aberrations for prognostication in CLL?

Panel: Fred Davi (FR)**, Kostas Stamatopoulos** (GR)**, Paolo Ghia** (IT)**, Richard Rosenquist** (SE)

The day was concluded with two lectures about clinical aspects of CLL impacted by current findings in CLL biology. In the first one, dealing with CLL genetics and prognostication, Richard Rosenquist presented new findings in genomic CLL landscape characterization broadened by technical possibilities of high-throughput sequencing, also including studies of its temporal dynamics, intraclonal composition, and associations with BCR stereotyped subsets. Following that, he put in contrast our current approach to CLL prognostication and prognostic guidelines, and outlined future directions involving next-generation sequencing (NGS) prognostic gene panels. The latter lecture by Paolo Ghia focused on new CLL drugs targeting signaling pathways: Bcl2 inhibitors affecting apoptotic pathway and BCR signaling inhibitors – from their characterization and biological context to results of clinical trials. In the case of BCR signaling inhibitors, differences in BCR signaling depending on BCR structure were highlighted. The session was wrapped up with round table discussion addressing a question of IG genes vs. genetic aberrations for CLL prognostication.

2.5 SESSION V – IMMUNOGENETICS IN THE NGS ERA

Chair: **Ton Langerak** (NL)

Anna Vardi (GR) - *Technical considerations for NGS analysis of immunoglobulin gene repertoires*

Anastasia Hadzidimitriou (GR) - *IMGT/High V-QUEST tool*

Nikos Darzentas (CZ) - Bioinformatics: *State-of-art tools for NGS immunogenetics*

Kostas Stamatopoulos (GR) - *Challenges with immunogenetics in the NGS era*

Round Table Discussion/Debate:

Sanger versus NGS for IG gene analysis within clinical routine: what is feasible?

Panel: Fred Davi (FR)**, Kostas Stamatopoulos** (GR)**, Ton Langerak** (NL)

The session covered the issue of using NGS methods in immunogenetics. Technical considerations for NGS analysis of IG gene repertoires, especially in areas of NGS experimental design, management of sequence errors and data mining tools, were discussed by Anna Vardi. State-of-art tools for IG NGS data analysis and synthesis were presented by two speakers, Nikos Darzentas and Veronique Giudicelli. The study of immunoglobulins and T cell receptors using next-generation sequencing is becoming increasingly mainstream, as it allows exploring immune repertoires and responses in their immense variability and complexity. Naturally, such experiments result in vast amounts of complex data, which makes their analysis and interpretation a highly convoluted task.
In response to this, an ARResT/Interrogate, a web-based, interactive application was developed, over the past few years and mainly in the context of the EuroClonality-NGS consortium - a number of members of which are also involved in the IgCLL group. ARResT/Interrogate can organize and filter large amounts of data by numerous criteria and user-provided metadata and their combinations, calculate several relevant statistics, and present results in the form of multiple interconnected visualizations. It has its own data annotation pipeline, It accepts and understands IMGT/HighV-QUEST output files with immunogenetics-specific functionalities, or it can load practically any data in tabular format.
The ARResT/Interrogate was presented as a continuation of a decade of work within the IgCLL group and in light of all the new challenges presented by next-generation sequencing, with the help of examples of the interface and its functionalities, general and specific use

cases, advice and insights and applications (clinical SOPs, R&D e.g. primer development and troubleshooting.
The session was concluded with a round table discussion that explored feasibility of NGS for IG gene analysis in clinical routine in comparison to conventional Sanger sequencing.

2.6 SESSION VI – IMMUNOGENETICS AND DISEASE FOLLOW-UP: TRACKING MRD

Chair: **Paolo Ghia** (IT)

Christiane Pott (DE) - *New developments in MRD monitoring via NGS in mature B cell malignancies*

Andy Rawstron (GB) - *Flow cytometric vs PCR MRD for CLL*

Andy Rawstron (GB) - *Guidelines for interpreting MRD flow data (hands-on computer session)*

Round Table Discussion/Debate:

MRD detection in CLL and B cell lymphomas: what does the future hold?

Panel: Andy Rawstron (GB)**, Paolo Ghia** (IT), **Christiane Pott** (DE)

Anastasia Hadzidimitriou (GR) - *Concluding remarks*

The meeting was concluded by Christiane Pott and Andy Rawstron with their talks on tracking minimal residual disease (MRD), an important predictor of disease progression in B cell malignancies. While dr. Pott shared her experience with high-throughput sequencing methods for MRD monitoring in mature B cell malignancies presenting new developments in the area, dr. Rawstron provided an overview of various methods used for MRD assessment, including flow cytometry and RT-qPCR comparing their advantages and disadvantages. He also provided MRD guidelines supplemented with computer-based practical step-by-step demonstration of flow-cytometric MRD analysis. Following round table discussion was dedicated to the future of MRD detection in CLL and B cell lymphomas.

3 Structural Validation Workshop

The Structural Validation Workshop was organised from 17th to 19th October 2016 at CEITEC MU in Brno. CEITEC MU lecturers together with invited experts from EMBL-EBI Dr. Sameer Velankar and Dr. Oliver Smart introduced the topic of 3D structure validation and its importance to participants. The workshop consisted of 8 sessions with theoretical lectures and practical exercises on validation of topology and geometry and data archiving.

3.1 SESSION I – INTRODUCTION AND WARM-UP

A goal of this session is to introduce the MEDGENET project and institutes, which organize the workshop. First, MEDGENET, its goals, participants, resources etc. will be described by Dr. Svobodova. Than, Central European Institute of Technology (CEITEC), its research programmes, core facilities, partners and goals will be introduced by CEITEC scientific director, prof. Koca. Afterwards, Dr. Svobodova will speak about Computational chemistry research (from CEITEC), its research activities and its results and know-how in the field of validation. Then, Dr. Velankar will introduce European Bioinformatics Institute (EBI) and describe its goals and activities. Last but not least, Dr. Velankar will speak about his research group, which is responsible for development and maintenance of Protein Data Bank Europe, including their validation tools and workflow.

3.2 SESSION II – VALIDATION OF TOPOLOGY – OVERVIEW

Dr. Radka Svobodová (CEITEC MU)

Motivation – why to validate

Topology and geometry validation

Topology validation – common analyses

Software tools MotiveValidator and ValidatorDB

Validation of biomacromolecular structures obtained by NMR and X-ray crystallography has become a very important topic, because some published structures have been found to contain serious errors. The biomacromolecular structure validation can be divided into two main directions – topology validation and geometry validation. Topology validation tests, if the validated biomacromolecule has correct 2D structure. Specifically, if it contains all the atoms, which chemically correct biomacromolecule should have and if they are connected by the bonds properly. Geometry validation evaluates, if the biomacromolecule has a correct 3D structure. For example, if the bond length, bond angles and torsion angles are OK. A different methodology for validation of ligands was developed later - validation of annotation.

First proposal topology validation was published by Lütteke et al. and implemented in Pdb-care. It contained basic validation analyses and it was focused only on carbohydrates. Afterwards, a rich set of validation analyses, which covers main issues observed in the topology of ligands was developed and implemented in software tools MotiveValidator and ValidatorDB. MotiveValidator is a tool that allow to validate individual ligand or sets of

ligands and introduces several advanced validation analyses. ValidatorDB provides a database of weekly updated validation results for all ligands from Protein Data Bank and introduces further useful validation analyses.

3.3 SESSION III – VALIDATION OF TOPOLOGY – PRAXIS

Demo examples

Practical exercises

The course participants performed the following exercises:

- Demo exercise: Detecting of problematic ligands in nipah G attachment glycoprotein
- Validation of sucrose in Plant Photosystem I – missing atoms
- Validation of beta-carotene (BCR) in Plant Photosystem I – missing rings
- Demo exercise: Validation of maltose (MAL) ligands in Protein Data Bank – missing atoms and rings
- Detect all maltose (MAL) ligands in Protein Data Bank which have missing atoms and rings.
- Detection of atom substitution in biotin (BTN) from 50S Complex (PDB ID 1kqs)
- Detection of chirality problems in all sialic acids (SIA) from Protein Data Bank
- Demo exercise: Detection of chirality problems in testosterone derivatives

Note: Solution of the demo exercises was shown by the lectors.

3.4 SESSION IV – VALIDATION OF GEOMETRY – OVERVIEW

Dr. Sameer Velankar, Dr. Oliver Smart (EMBL-EBI)

Geometry validation – common analyses

Summary validation criteria

PDB Validation reports

Geometry validation evaluates whether the structure (geometry) is correct. Specifically, it compares the geometrical properties of a residue (e.g., atom distance, bond length, bond angle, torsion angle) with a correct (tabular) value of this property. If the difference is higher than a defined tolerance, an error is reported. The geometry validation results are summarized in main geometry validation criteria:

- **Rfree:** A measure of the fit of the model to a small subset of the experimental data, which was not used in model refinement.
- **Clashscore:** This score is derived from the number of pairs of atoms in the model that are unusually close to each other. It is calculated by MolProbity and expressed as the number or such clashes per thousand atoms.
- **Ramachandran outliers:** A residue is considered to be a Ramachandran plot outlier if the combination of its φ and ψ torsion angles is unusual. The Ramachandran outlier score for an entry is calculated as the percentage of Ramachandran outliers with respect to the total number of residues in the entry for which the outlier assessment is available.
- **Sidechain outliers:** Protein sidechains mostly adopt certain (combinations of) preferred torsion angle values (called rotamers or rotameric conformers), much like their backbone torsion angles (as assessed in the Ramachandran analysis). Validation

considers the sidechain conformation of a residue to be an outlier if its set of torsion angles is not similar to any preferred combination. The sidechain outlier score is calculated as the percentage of residues with an unusual sidechain conformation with respect to the total number of residues for which the assessment is available.

3.5 SESSION V – VALIDATION OF GEOMETRY – OVERVIEW

Demo examples

Practical exercises

The course participants performed the following exercises:

- Demo exercise: Obtaining structure validation data for a rhodostomin. E.g., the following data: How many clashes are in the structure? Which atoms in a clash are the closest? How many bond length outliers are in the structure? How many bond angle outliers are in the structure? Which bond angle outlier is the highest?
- Obtaining structure validation data for a mutant of cytochrome P450cam. E.g., How many clashes are in the structure? Are there some atoms closer than 1 Å? How many bond length outliers are in the structure? Are there any bond length outliers > 0.3 Å? How many bond angle outliers are in the structure? Which bond angle outlier > 20°?
- Demo exercise: Understanding PDB validation report summary criteria for oxy-hemoglobine in methanol

Note: Solution of the demo exercises was shown by the lectors.

3.6 SESSION VI – VALIDATION OF GEOMETRY – VALIDATION PERSPECTIVES

Dr. Radka Svobodová (CEITEC MU) – *Ligand validation – current state*

Vladimír Horský, Dr. Radka Svobodová (CEITEC MU) – *Validation trends*

Dr. Sameer Velankar, Dr. Oliver Smart (EMBL-EBI) – *Future plans and directions in validation*

Summary of the current state: The analysis of quality of all ligands in Protein Data Bank showed, 8% of ligands are incomplete and further 9% of ligands have chirality errors. Therefore, more than 80% of ligand structures are correct. Experimental and approved drugs showed as molecular classes having markedly higher quality than average ligands. As expected, carbohydrates proved as problematic from the quality point of view, since they have markedly higher percentage of C chirality errors. This problem is further accented for mannose derivatives. Polycyclic ligands demonstrate similar results as carbohydrates, just they have slightly lower percentage of C chirality. The most problematic ligands are organometals, showing outstanding validation problems in most of validation criteria.
Validation trends in ValTrendsDB: ValTrendsDB is a database which shows the relations between the available quality criteria and a rich set of biomacromolecular properties (e.g., size, number of atoms, resolution, number of ligands, year of publication, ligand flexibility). Among other things, the relations between two biomacromolecular properties can also be seen. Therefore, ValTrendsDB contains information about almost 1500 trends (i.e., meaningful pairings between quality criteria and a property or between two properties). Potential relationships were explored in two steps. First, potential relationships between factors (i.e., quality criteria or structure properties) were estimated using box plot-like graphs,

with one factor being sorted into intervals and another factor being represented with an arithmetic average value for each interval. The discovered relationships were then validated using the statistical method of correlation analysis. The factor value distribution was very skewed, which warranted the use of weighted averages and non equidistant intervals. Extra care was taken with to semi-automatic algorithms that divide data into intervals so that the inferred conclusions were sound. The strength of relationships was assessed by the value of Pearson coefficients.

3.7 SESSION VII – FUTURE OF DATA ARCHIVING

Dr. Sameer Velankar (EMBL-EBI) – *Future of archiving*

Dr. Sameer Velankar (EMBL-EBI) – *Big data and challenges*

Dr. David Sehnal (CEITEC MU) – *Data delivery and visualization*

Recent advances in Cryo-EM and other structure determination techniques enabled a rapid increase in the complexity and size of biomacromolecular structures available in the Protein Data Bank. A great challenge is to deliver and visualize this data to the user in a reasonable time. Until now, Java-based applets such as Jmol have served this purpose. However, the lack of support for Java applets in the web browsers and introduction of HTML5 and WebGL web frameworks provide an opportunity to develop better solutions to address this problem.
Therefore, a comprehensive solution for both efficient delivery of large data sets and interactive visualization of macromolecular structures is in development. It contains the following parts:

- The CoordinateServer provides functionality for dynamic selection of a subset of structural data (backbone, ligand binding sites, etc.) that is of interest to the user. The amount of transferred data is further reduced using binary representation of the standard mmCIF data format called BinaryCIF. The CoordinateServer thus provides a complete solution for efficient transfer of data to user interface.
- LiteMol is an interactive browser based 3D viewer, designed using HTML5 and WebGL technologies that provides all key structural visualization modes. The viewer also integrates PDBe REST API services to show biomacromolecular annotations (validation, sequence annotation). LiteMol can use the standard mmCIF files or the CoordinateServer for fast data access. Thanks to integration with the CoordinateServer, LiteMol can generate an interactive 3D display for the largest currently available structure, the HIV capsid with 2.4 million atoms, in less than a minute.

3.8 SESSION VIII – FUTURE OF DATA ARCHIVING

Demo examples

Practical exercises

Closing remarks

The course participants performed the following exercises:

- Basic visualization: Visualization of retinoic-acid-binding protein (1cbs). Comparison of different visualization modes: C-alpha trace, Balls and sticks and VDW Balls
- Visualization of surface: Visualization a surface of a catalytic Step I spliceosome (PDB ID 5gmk)

- Visualization of electron density: Visualization electron density data for Catalytic Step I spliceosome (PDB ID 3pah) and its adrenaline ligand.
- Visualization of validation data: Visualization validation data for Human methemoglobin indicative of fiber formation (PDB ID 3odq).
- Vizualization of assemblies. Visualization an assembly for Aquareovirus virion (PDB ID 3k1q).
- Vizualization HIV-1 capsid (3j3q), comparison of the following ways:
    - Visualization directly (just type its PDB ID to LiteMol)
    - Visualization from file, downloaded to disk
    - Visualize using Coordinate Streeming.

4 Workshop on Grid and Cloud technologies for high-performance high-throughput bioinformatics

The ***workshop on Grid and Cloud technologies for high-performance high-throughput bioinformatics*** was a one-day workshop that took place at the Institute of Applied Biosciences (INAB) of the Center for Research and Technology Hellas (CERTH), Greece on December 15th. A number of presentations and hands-on exercises at the workshop introduced the characteristics and constraints of the computing and storage resources available to European Researchers through EGI, and provide a detailed training on deploying standard bioinformatics pipelines (including NGS data analysis) through the available resources.

4.1 SESSION I – HPC RESOURCES

Kostas Stamatopoulos (CERTH) – *Welcome and Introductions*

Yin Chen (EGI Foundation) – *The EGI e-infrastructure*

Yin Chen (EGI Foundation), Fotis Psomopoulos (CERTH/AUTH) – *Clouds, VMs and other resources from EGI for bioinformatics training*

Yin Chen (EGI Foundation), Fotis Psomopoulos (CERTH/AUTH) – *Basic tools to exploit the EGI FedCloud (hands-on exercises)*

First in the presentation from Dr Yin Chen the EGI Foundation was introduced. The EGI Foundation (also known as EGI.eu) coordinates the EGI infrastructure on behalf of the participants of the EGI Council: national e-infrastructures and European Intergovernmental Research Organisations (EIROs). The EGI Foundation is not-for-profit and was established in Amsterdam in 2010 under Dutch law. The foundation is lead by Yannick Legré (Managing Director) and is governed by the EGI Council. Day-to-day operations are supervised by the Executive Board.

A hands on training followed on the direct connection and usage of the EGI computational cloud. This training involved access training infrastructure, localization and spawning of the virtual machine in the EGI computational cloud. Data connection and processing was also explained specifically mounting of the external data storage and creating of a permanent data storage on the VM.

4.2 SESSION II – NGS DATA ANALYSIS ON EGI FED CLOUD

Kimmo Mattila (CSC), Fotis Psomopoulos (CERTH/AUTH) – *NGS Data Analysis on EGI FedCloud: The Chipster platform*

Fotis Psomopoulos (CERTH/AUTH) – *Hands-on exercises for RNA-Seq and Variant Calling*

Kostas Stamatopoulos (CERTH) – *Closing, Final Remarks*

In the second session specific bioinformatics tools for NGS analysis were presented with their usage shown on the EGI foundation cloud. Specifically a Jupyter notebook application was

introduced which is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.

Consecutively a versatile data analysis platform with interactive visualizations and workflows called Chipster was presented. It offers a comprehensive collection of analysis tools for next generation sequencing (NGS), microarray and proteomics data. The NGS functionality covers analysis from quality control and alignment to downstream applications such as pathway analysis and motif detection. An exercise to perform both DNA variant call analysis and RNAseq expression analysis was included.