Automated Clustering Analysis of Immunoglobulin Sequences in Chronic Lymphocytic Leukemia Based on 3D Structural Descriptors

Immunoglobulins (Igs) are crucial for the defense against pathogens, but they are also important in many clinical and biotechnological applications. Their characteristics, and ultimately their function, depend on their three-dimensional (3D) structure; however, the procedures to experimentally determine it are extremely laborious and demanding. Hence, the ability to gain insight into the structure of Igs at large relies on the availability of tools and algorithms for producing accurate Ig structural models based on their primary sequence alone. These models can then be used to determine structural and eventually functional similarities between different Igs. An example of such a task is the clustering of Igs based on their structure to determine meaningful common features such as the possible existence of common molecular targets (antigens). Several approaches have been proposed in order to achieve an optimal solution to this task yet their results were hindered mainly due to the lack of efficient clustering methods based on the similarity of 3D structure descriptors. Here, we present a novel workflow for robust Ig 3D modeling and automated clustering. We validated our protocol in chronic lymphocytic leukemia (CLL), where the clonotypic Igs are critically implicated in the disease ontogeny and evolution. Indeed, immunogenetic studies on the clonotypic Igs have strongly implicated antigen selection in the pathogenesis of CLL, while also providing robust prognostic information. In the present study, we used the structure prediction tools PIGS and I-TASSER for creating the 3D models and the TM-align algorithm to superpose them. The innovation of the current methodology resides in the usage of methods adapted from 3D content-based search methodologies to determine the local structural similarity between the 3D models. The Fast Point Feature Histograms descriptors derived from the structurally aligned parts are used to compute a distance matrix, which is then used as input for the clustering procedure. Clustering analysis on the data is performed through the application of the agglomerative and density-based clustering approaches. The first method is unsupervised whereas the second belongs to the semi-supervised type, i.e. requires a predefined number of clusters. To evaluate the quality of the herein described workflow, we performed a supervised analysis of 125 Ig 3D models originating from 5 CLL stereotyped subsets i.e. subgroups sharing (quasi) identical Igs, namely subsets #1, #2, #4, #6, #8. The reasoning behind this choice was that (i) homologous Ig primary sequences can be reasonably anticipated to be reflected in overall similar 3D structures, hence providing a reference for evaluating the developed workflow; and, (ii) these subsets are well characterized at both the clinical and biological levels. Subset size distribution was as follows: subset #1 (IGHV clan I/IGKV1(D)-39), n=37; subset #2 (IGHV3-21/IGLV3-21), n=43; subset #4 (IGHV4-34/IGKV2-30), n=22; subset #6 (IGHV1-69/IGKV3-20), n=12; and, subset #8 (IGHV4-39/IGKV1(D)-39), n=11. Overall, we obtained a high level of clustering accuracy i.e. Ig 3D model clusters matched to a very high degree the subsets defined by Ig primary sequence similarity. In detail, 5 Ig 3D model clusters were produced by: (i) cluster 1 containing 37/37 (100%) subset #1 models and one (8.3%) subset #6 model, (ii) cluster 2 containing 43/43 (100%) subset #2 models, (iii) cluster 3 containing 21/22 (95.5%) subset #4 models, (iv) cluster 4 containing 11/12 (91.7%) #6 models, and, (v) cluster 5 containing 11/11 (100%) subset #8 models along with a single (4.5%) subset #4 model (subsets #4 and #8 concern IgG CLL, in itself a rarity for CLL). These findings support that the innovative workflow described here enables robust clustering of 3D models produced from Ig sequences from patients with CLL. Furthermore, they indicate that CLL classification based on stereotypy of Ig primary sequences is likely also verified at the Ig 3D structural level. Studies are ongoing for both addressing the minor discrepancies observed here and producing the unsupervised 3D clustering of the Igs from a large series of both stereotyped and non-stereotyped CLL cases.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Center for Research and Technologies Hellas, Uppsala University, Masaryk University, Papanikolaou Hospital, Carlsberg Research Center, Nikea General Hospital, Vita-Salute San Raffaele University, Feinstein Institute for Medical Research
Number of pages: 1
Pages: 4365
Publication date: 2016
Peer-reviewed: Yes

Publication information
Journal: Blood
Volume: 128
Issue number: 22
ISSN (Print): 0006-4971
Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2