Biochimica et Biophysica Acta 1830 (2013) 3373–3381

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

The Global Sequence Signature algorithm unveils a structural network surrounding heavy chain CDR3 loop in *Camelidae* variable domains

Damjana Kastelic [a,b,1], Nicolas Soler [c], Radovan Komel [b], Denis Pompon [a,*]

[a] Université de Toulouse, INSA, UPS, INP, LISBP, INRA UMR792, CNRS UMR5504 Toulouse, France
[b] Medical Centre for Molecular Biology, Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Slovenia
[c] MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 13 February 2013
Accepted 15 February 2013
Available online 26 February 2013

Keywords:
Camelidae variable domain
Co-evolution
Complementarity determining region 3 (CDR3)
V_H
V_HH
Multiple sequence alignment

ABSTRACT

Background: A large fraction of camelid (camels and llamas) antibodies is composed of heavy chain-only homodimers, able to recognise antigens with their variable domain. Events in somatic assembly and maturation of antibodies such as hypermutations and rearrangement of variable loops (CDRs — complementary determining regions) and selection among a wide range of framework variants are generally considered to be random processes.
Methods: An original algorithmic approach (Global Sequence Signature—GSS) was developed, able to take into account multiple functional and/or local sequence properties to detect scattered evolutionary constraints into sequences.
Results: Using the GSS approach, we show that the length of the main hypervariable loop (CDR3) is linked to the nature of 19 surrounding residues on the scaffold. Surprisingly, the relation between CDR3 size and scaffold residues strongly depends on the considered species, illustrating either significant differences in selection mechanisms or functional constraints during antibody maturation.
Conclusions: Combined with the statistical coupling analysis (SCA) approach at the level of scaffold residues, this study has unravelled a robust interaction network on antibody structure surrounding the CDR3 loop.
General significance: In addition to the general applicability of the GSS algorithm, which can bring together functional and sequence data to locate hot spots of constrained evolution, the relationship between CDR3 and scaffold discussed here should be taken into account in protein engineering when designing antibody libraries.

1. Introduction

In most species, the antigen binding site of immunoglobulins (Ig) comprises six hypervariable loops supported by a structurally highly conserved β-sheet scaffold formed by the heavy (V_H) and light (V_L) chain variable domains. A striking exception in mammals is found in *Camelidae* (camels, dromedaries, llamas, alpacas, guanaco and vicuñas) where a considerable fraction (up to 75%) of functional antibodies is composed of heavy chain-only homodimers [1]. Their antigen-binding site consists of a single unpaired variable domain (referred to as V_HH) that can bind antigen with specificities comparable to those of the intact V_H/V_L structure. The high stability [2], small size (15 kDa) and monomeric nature of V_HH attract interest for protein engineering [3].

During their development, B cells ensure diversity of the antibody population by random combinatorial associations of V, D and J gene segments from their germ-line DNA. After immunisation, antibody affinity is further increased by somatic hypermutations. Little is known about how this increase in affinity can result from small structural changes confined on the periphery of the antigen binding site through these mutations [4]. Single amino acid substitutions distant from the antigen binding site are also capable of altering specificity or even abolishing antigen binding [5], underlining a spatial link between antigen binding residues and distant sites in the variable domain. Such an interaction network implies a process of co-evolution of the residues involved in order to maintain the antibody's structural and functional integrity.

Classical approaches for sequence mining in gene families first of all involve global and local sequence alignment methods [6,7]. In all cases these methods rely on sequence comparison involving a predefined homology matrix which scores the similarity of one amino acid to another. Multiple sequence alignments (MSA) allow in a first step quantifying conservation of each position in the sequences. Some important residues

Abbreviations: CDR, complementarity determining region; FR, framework; GSS, Global Sequence Signature analysis; Ig, immunoglobulin G; MSA, multiple sequence alignment; MDS, multidimensional scaling; PCA, principal component analysis; SCA, statistical coupling analysis; V_H, heavy chain variable domain of the conventional antibody; V_HH, variable domain of a heavy chain antibody; V_L, light chain variable domain

* Corresponding author at: Université de Toulouse, INSA, UPS, INP, LISBP, 135 Avenue de Rangueil, 31077 Toulouse, France. Tel.: +33 567048806; fax: +33 561559400.
E-mail address: denis.pompon@insa-toulouse.fr (D. Pompon).
[1] Present address: Division of Functional Genome Analysis, German Cancer Research Centre (DKFZ), Heidelberg, Germany.

0304-4165/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.02.014

might however display a low degree of conservation due to co-variation with one or more other residues within the sequence, and thus remain hidden to this kind of analysis. The general problem is in this case to distinguish a conservation/co-variation signal between sites due to functional or structural constraints as opposed to noise from random events, finite sequence sampling or phylogenetic history (heterogenous divergence). Diverse methods have been employed to identify such correlated positions from a set of pre-aligned sequences (reviewed in [8,9]). A majority of them aims to search for coevolving pairs of residues, by evaluating the dependence between substitutions or amino acid compositions at two different sites in a MSA through various scoring functions derived from: correlation coefficients [10], mutual information [11], chi-squared test [12,13], alignment perturbation [14], maximum likelihood [15] or other statistics. Apart from the sequences themselves, the scoring functions described in these methods use a limited amount of data relative to the studied family (i.e.: amino acid properties and phylogenetic or structural data). Here we present an original procedure using qualitative or quantitative knowledge (properties) related to a set of pre-aligned homologous sequences in order to identify the positions playing a critical role in these properties. Its application to camelid antibody variable domains unveils a network of coevolving positions demonstrating a tendency of spatial coupling, which maps to functionally important sites onto 3D structure.

2. Results

2.1. Global Sequence Signature (GSS) analysis

Considering a particular property related to a set of sequences (for example: charge, solubility, affinity, length of a segment within a sequence, presence/absence of a motif, etc.), the idea is to assume that every residue in a particular sequence contributes in its own extent to a global signature which, once determined, will enable predictions about how a sequence will behave. The global (GSS) score of a sequence relative to a property precisely quantifies the imprinting of the signature over the whole sequence. Two or more GSS scores per sequence, relative to different properties, can be calculated and plotted one vs another so as to group sequences into clusters and assess correlations between the different properties. By examining the nature of the most frequent residues at each position among clusters, it becomes then possible to spot the major contributing positions to one or several properties. The analysis carried out by the GSS algorithm on a MSA can be summarised in the following five steps:

1) Initial ranking: the algorithm needs as an input a set of aligned sequences (the learning set) for which a numerical value (quantitative observable of the property) can be assigned to each sequence. This learning set can constitute only a subpart of the available aligned sequences. A property can correspond to any type of biochemical measurement or alternatively be evaluated from some intrinsic feature in the considered sequences (for example presence of hallmarks, charge, hydrophobicity profiles, etc.). In our analysis of camelid V_H/V_HH sequences, this initial ranking of the sequences is referenced as "local scoring" since we choose as properties local features in the sequences themselves. The distribution of local scores is centred on zero by applying a suitable offset.
2) Construction of the tables: the amino acid distribution at each position of the ranked learning set is examined in order to build a scoring table for each type of residue at that position (i.e.: residue types more often found in highly ranked sequences for a given criteria will get a high score and vice versa). This is analogous to a substitution matrix in classical alignment procedures but there is one table per position in this case.
3) GSS scoring: the positional scoring tables can then be used to calculate back a GSS score for each sequence of the input MSA relative to a given property. This score is obtained by summing the table scores obtained in the previous step over all the constituent residues in a sequence and applying a suitable normalisation procedure (see methods).
4) Clustering: By repeating the GSS scoring process considering an alternative property, sequences can be mapped on 2D graphs so that the GSS scores are displayed relative to each other. In this way, it becomes possible to observe clustering of the sequences and potentially highlighting correlations between properties.
5) Extraction of significant positions: each cluster containing a set of aligned sequences, the last step is to compare the frequencies of observation of the different amino acids between clusters using a chi-squared test at each position. Those exhibiting a significant difference of composition are selected and validated using complementary statistical tools (see methods).

The whole procedure is automated and available through an online web server at the following address: http://gss.mrc-lmb.cam.ac.uk.

2.2. The GSS analysis evidenced interplay between the CDR3 loop and the framework sequence of llama immunoglobulin

As mentioned, positions containing specific features in the sequences themselves (local signatures) can be considered as internal properties to be correlated to all other remaining residues by the GSS analysis algorithm. In the present study, we used as properties two local features of camelid immunoglobulin sequences:

(i) *The length of CDR3 region* (whose sequence was excluded from the GSS scoring calculation), calculated as the number of amino acids found between alignment positions 95 and 103, according to Kabat numbering [16,17], and equivalent to positions 100–126 in tested multiple sequence alignments. Within V_H and V_HH, the complementary determining region 3 is the major contributor in determining the specificity and affinity of the antigen-binding site, due to its enhanced diversity and central position [18–20]. The longer CDR3 of V_HH compared to V_H increases the antigen binding surface area, thus compensating for the absence of antigen interactions provided by V_L [18,21,22].
(ii) *The number of V_HH hallmark residues.* V_HH hallmarks are four amino acid substitutions at positions 37, 44, 45 and 47 (Kabat numbering scheme for immunoglobulin variable domain) of the second framework (FR2) known to be discriminative between V_H and V_HH [1]. In V_H, these positions are occupied with the highly conserved tetrad VGLW and form a hydrophobic patch that interacts with the V_L, while in V_HH those positions are changed to the slightly more hydrophilic patch of residues F/Y/I37, E/Q44, R45 and G/S/L/F47.

The local score for CDR3 length was simply calculated as the CDR3 length of a given sequence (the distribution of length is then centred on zero, see methods). The local score relative to the number of V_HH hallmark residues was computed for each sequence by summing hallmark positions as follows: plus one for each of F/Y/I37, E/Q44, R/C45, G/S/L/F47 (V_HH hallmarks) and minus one for each of V37, G44, L45, W47 (V_H hallmarks). As for the CDR3 length property, the four hallmark positions were excluded from the GSS scoring process step to avoid trivial self-correlation.

2.2.1. Sequence mapping: local versus global (GSS) scoring

The set of used aligned sequences for llama and camels (SM1 and SM2 files: MSA_Lama_glama.txt and MSA_Camelus_dromedarius.txt, available on the website) was composed of experimentally obtained immunoglobulin heavy chain variable domain of llama (*Lama glama*) sequences from a cDNA library [23], to which was added a larger set of llama immunoglobulin sequences available from GenBank (sequences with extremely long hypervariable loops were excluded). Before GSS analysis, the relation between the two described local properties: CDR3