jmb Communication

Metric to Distinguish Closely Related Domain Families Using Sequence Information

Elizabeth A. Proctor[1,2], Pradeep Kota[2,3], Stephen J. Demarest[4], Justin A. Caravella[4] and Nikolay V. Dokholyan[1,2,3,5]

[1] - ***Curriculum in Bioinformatics and Computational Biology,*** *University of North Carolina, Chapel Hill, NC 27599, USA*
[2] - ***Program in Molecular and Cellular Biophysics,*** *University of North Carolina, Chapel Hill, NC 27599, USA*
[3] - ***Department of Biochemistry and Biophysics,*** *University of North Carolina, Chapel Hill, NC 27599, USA*
[4] - ***Biogen Idec,*** *Cambridge, MA 02142, USA*
[5] - ***Center for Computational and Systems Biology,*** *University of North Carolina, Chapel Hill, NC 27599, USA*

Correspondence to Nikolay V. Dokholyan: *Curriculum in Bioinformatics and Computational Biology, University of North Carolina, 120 Mason Farm Road, Campus Box 7260, Chapel Hill, NC 27599–7260, USA.* dokh@unc.edu
http://dx.doi.org/10.1016/j.jmb.2012.11.031
Edited by A. Panchenko

Abstract

Engineered antibodies are emerging as a promising class of therapeutic biomolecules, as well as having applications in medical research. Knowledge on conserved functional and structural regions within antibody domains is imperative in order to rationally design stable and specific antibodies. Of particular interest for the design of therapeutics are antibody variable and constant domains, which are responsible for antigen binding and immune response. These antibody domains are part of the larger immunoglobulin (Ig) V-class and C-class families, respectively. We find that, although both classes belong to the Ig-fold superfamily, the sets of conserved residue positions and identities differ between these classes. We exploit these evolutionary differences to derive a metric based on sequence positional entropy that distinguishes C-class from V-class sequences utilizing only sequence information. By distinguishing different domain families using sequence information alone, we enable the application of domain-specific design strategies without the need for secondary or tertiary structural information.

Antibody therapeutics are emerging as an important class of drugs due to their tolerance by the human body and their amenability to design. However, the creation of antibody drugs is expensive and time consuming. Rational design strategies are often adopted in order to decrease the time and expense of antibody engineering.[1] A common strategy employed to obtain useful information that can further be applied to rational design of proteins is to compile evolutionary information from related protein families. Dokholyan and Shakhnovich have shown from first principles that residues crucial to function or structural stability will be conserved throughout evolution and are common to proteins that are closely related.[2] Therefore, key residues responsible for protein structural stability are not chosen for mutation during the design process, in order to preserve the integrity of the design scaffold. These residues are often buried in the core of the protein and do not participate in protein-binding interactions.[3] For example, two cysteines common to the immunoglobulin (Ig)-fold superfamily are highly conserved during evolution and are involved in a covalent disulfide linkage in the core of the domain, playing a crucial role in maintaining the tertiary structure of the fold.[4] We may therefore determine the identities of residues important for the structural, and hence functional, integrity of a relatively uncharacterized protein by using pre-existing information from evolutionarily related proteins. Here, we develop a metric based solely on sequence information to distinguish between evolutionarily related fold families. We apply our metric to differentiate sequences encoding Ig-fold C-class domains from those corresponding to Ig-fold V-class domains. Using previously constructed multiple sequence alignments of C- and V-class sequences (Proctor *et al.*, submitted),[5] we generate conservation profiles for each domain family using established techniques.[2] Briefly, we calculate amino-acid-type frequencies at each position along the sequence and obtain the probability $p_k(\sigma)$ for each amino acid type σ at each position k in the multiple sequence alignment. We then calculate the sequence positional entropy for each position k in the alignment as:

$$S(k) = -\sum_{\sigma} p_k(\sigma) ln(p_k(\sigma))$$

We observe that, in general, C-class domains have a stretch of positions with low sequence positional entropy (high conservation) in the N-terminus of the protein sequence (positions 20–30), while V-class domains are more highly conserved at the C-terminus of the protein sequence (positions 80–90) (Fig. 1a). By mapping the corresponding regions onto the C- and V-class domains from antibody crystal structures, we find that these sequence positions belong to residues that are in domain interfaces (Fig. 1b). This observation supports the hypothesis that these low-entropy positions are important for the selection of constant–variable domain partners and quaternary structure in

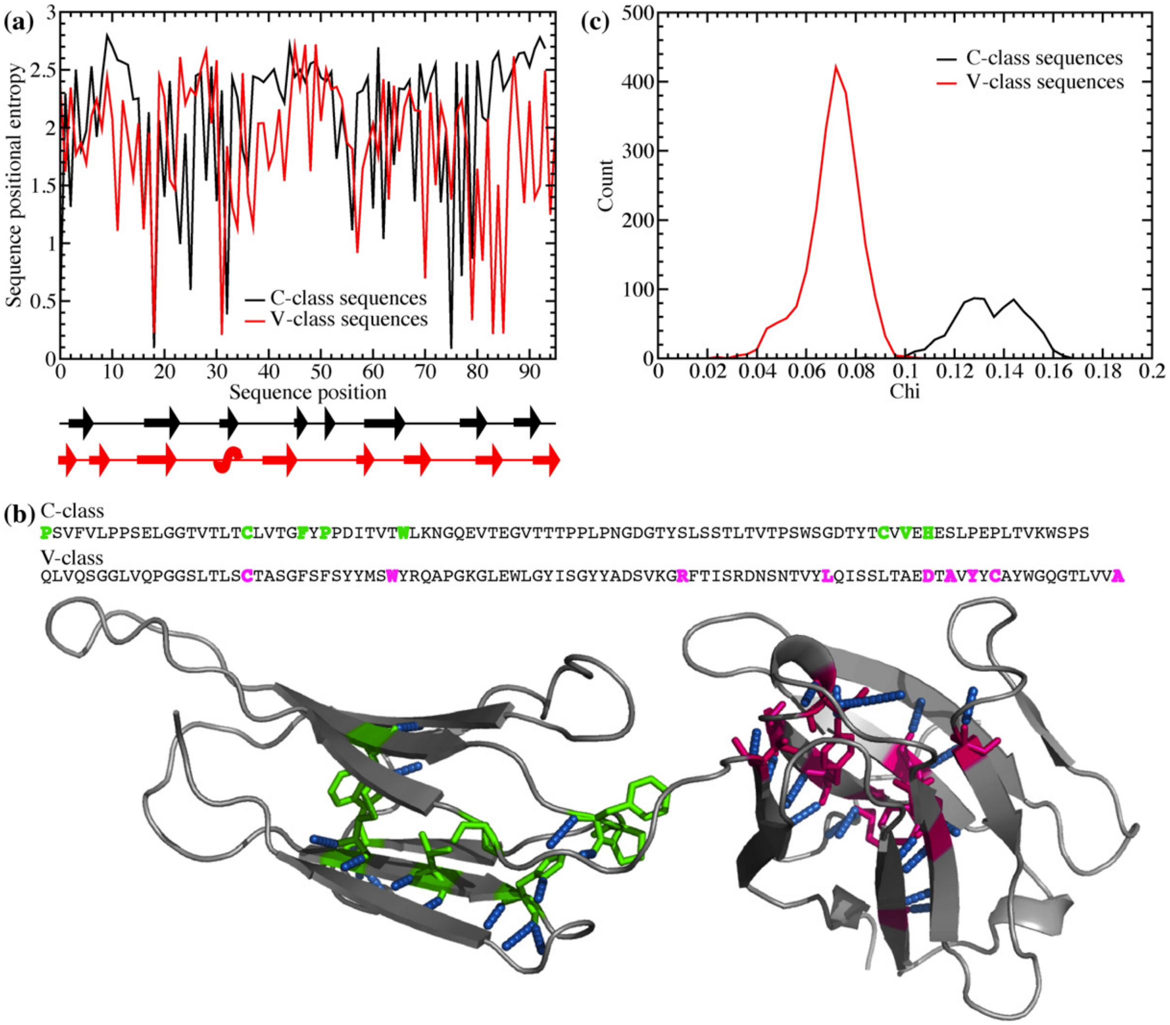

Fig. 1. Differences in sequence conservation allow distinguishing C-class and V-class Ig-fold sequences. (a) Sequence positional entropy at each position in the C- or V-class alignments. We show entropies only of those positions for which the consensus (majority) residue for the alignment is not a gap. We display for reference the corresponding secondary structural elements along the sequence of the domain. (b) The consensus sequences of the C-class (top) and V-class (bottom) databases, with gaps removed, and the structure of a heavy-chain Fab fragment of a monoclonal anti-E-selectin antibody (Protein Data Bank ID: 1A5F) composed of a CH1 domain and a VH domain. We highlight conserved ($S<1$) residues in green (C-class) or hot pink (V-class). Polar contacts of conserved residues are shown in blue. (c) Distributions of metrics calculated for C- and V-class sequences using the C-class database as reference. The two distributions are distinct, with very little overlap, demonstrating the evolutionary differences in the two domain families. We obtain similar results when scoring using the V-class database as reference, with the V-class distribution shifted to higher scores than the C-class.

antibodies. Many of the conserved residues are also integral parts of the beta-strands responsible for domain tertiary structure. The observed differences in C- and V-class conservation profiles clearly demonstrate that different design strategies are needed for the various domain classes within the Ig superfamily. However, the classification of Ig domain sequences into families is challenging, even in cases with a solved structure. Well-established classification systems such as SCOP and CATH do not agree in the classification of many Ig-like domains,[6] causing difficulties for the rational optimization and redesign of novel Ig-like domains.

To address this issue, we define a metric that can distinguish Ig domain families using only sequence information. We use the recently constructed sequence alignments for C- and V-class sequences as our test cases to demonstrate the applicability of our metric (Proctor *et al.*, submitted).[5] Because the sequences of C- and V-class domains may be distinguished statistically by the conservation of different positions, as we demonstrate above, we are motivated to derive a metric that will estimate the likelihood of a given sequence belonging to either domain family. The need for such a metric is highlighted by the fact that it is not possible to distinguish C- and V-class sequences based purely on sequence length despite differences in the number of strands in these two domains. Excluding those cases for which a three-dimensional structure is available, obtaining sufficiently accurate secondary structure information to distinguish the one beta-strand difference is a complicated undertaking. Differentiating C- from V-class sequences is further complicated even assuming accurate knowledge of secondary structural elements because, due to variations in sequence length, one cannot be certain in an unknown sample whether one has obtained a complete sequence, especially in mixed populations or impure samples. In addition, although V- and C-class domains appear together in antibodies, they can be found separately in other proteins and complexes and, thus, cannot be distinguished simply based on their order of appearance on a chain. We construct a scoring metric χ as a measure of likelihood for a query sequence to be a member of a domain family, represented by a pre-aligned database of sequences. We utilize Clustal W[7] for our alignments, but any method of multiple sequence alignment can be used. We first calculate the most commonly occurring amino acid (or gap) at each sequence position to obtain a consensus sequence from the multiple sequence alignment of the domain family under consideration. We align the query sequence with the consensus sequence and sum the positional probabilities for each amino acid of the query sequence, including gaps. We define the metric χ as the average positional probability for the query sequence:

$$\chi = \frac{1}{N}\sum_{k} p_k(\sigma_k)$$

where N is the number of positions in the alignment and $p_k(\sigma_k)$ is the probability of finding amino acid σ at position k. In order to determine the likelihood of an unclassified sequence to be a member of a given domain family, we first compute the χ metrics for each individual sequence in the domain family alignment. We calculate the mean and standard deviation of the distribution of the individual χ metrics for the domain family. We compare the χ metric of the query sequence to this distribution by calculating the Z-score of χ_{query} according to the domain family χ distribution. A sequence with a χ metric Z-score absolute value of greater than 2 has very low probability of belonging to the domain family. For example, with an absolute value cutoff in Z-score of 2, none of the V-class sequences are falsely classified as C-class, while excluding only 2% of all C-class sequences. When the C-class sequences are measured against the V-class profile, an absolute value cutoff in Z-score of 2 results in only 3% of total C-class sequences falsely classified as V-class, while excluding only 0.9% of all V-class sequences. This metric is not specific for any certain type of protein or nucleotide sequences and may be applied to any system for which a pre-aligned database of sequences has been constructed.

Using our metric derived from the probability of each given residue at its respective position, we are able to distinguish V-class sequences from C-class sequences. We score each sequence in the C-class and V-class databases against the set of C-class sequences and obtain distributions of scores (Fig. 1c). The C-class and V-class distributions have very little overlap, which amounts to only 0.7% of C-class sequences and 6% of V-class sequences, demonstrating the ability of positional sequence entropy and a statistical representation of sequences to distinguish different families of sequences.

In conclusion, in this study, we find that the conservation profile of residues varies between families, specifically between C-class and V-class Ig-folds. We evaluate and compare patterns of conserved residues in C-class and V-class Ig-fold sequences and use this information to develop a metric for distinguishing C-class and V-class Ig domains based on sequence information alone. The conservation profile specific to each family can be used as a type of "fingerprint" to improve rational design strategies in order to enable family-specific engineering of antibodies or other Ig-fold domains.

This finding paves the way for novel design techniques in Ig domains. Given an uncharacterized Ig-fold sequence, sequence databases of various domain families will provide reference conservation profiles for comparison and classification for design purposes. The highest-scoring conservation profile will provide a framework for stabilization and redesign of intra-domain interactions by avoiding residues that are highly conserved across the domain family and targeting mutations to functionally relevant sites.

Acknowledgements

The authors would like to thank Dr. Dikran Aivazian and Dr. Alexey A. Lugovskoy for their helpful discussion and suggestions and Mr. Norman Wang for his work on the C- and V-class databases.

Received 31 August 2012;
Received in revised form 20 November 2012;
Accepted 23 November 2012
Available online 3 December 2012

Keywords:
domain classification;
conservation profile;
sequence entropy;
protein design;
therapeutic design

Present address: S. J. Demarest, Eli Lilly, 10300 Campus Point Drive, Suite 200, San Diego, CA 92121, USA.

References

1. Carter, P. J. (2006). Potent antibody therapeutics by design. *Nat. Rev., Immunol.* **6**, 343–357.
2. Dokholyan, N. V. & Shakhnovich, E. I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307.
3. Di Nardo, A. A., Larson, S. M. & Davidson, A. R. (2003). The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641–655.
4. Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily—domains for cell surface recognition. *Annu. Rev. Immunol.* **6**, 381–405.
5. Wang, N., Smith, W. F., Miller, B. R., Aivazian, D., Lugovskoy, A. A., Reff, M. E. *et al.* (2009). Conserved amino acid networks involved in antibody variable domain interactions. *Proteins*, **76**, 99–114.
6. Hadley, C. & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**, 1099–1112.
7. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H. *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.