PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Highly covarying residues have a functional role in antibody constant domains

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, North Carolina
[2] Program in Molecular and Cellular Biophysics, University of North Carolina, Chapel Hill, North Carolina
[3] Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina
[4] Biogen Idec, Weston, Massachusetts
[5] Center for Computational and Systems Biology, University of North Carolina, Chapel Hill, North Carolina

ABSTRACT

The ability to generate and design antibodies recognizing specific targets has revolutionized the pharmaceutical industry and medical imaging. Engineering antibody therapeutics in some cases requires modifying their constant domains to enable new and altered interactions. Engineering novel specificities into antibody constant domains has proved challenging due to the complexity of inter-domain interactions. Covarying networks of residues that tend to cluster on the protein surface and near binding sites have been identified in some proteins. However, the underlying role these networks play in the protein resulting in their conservation remains unclear in most cases. Resolving their role is crucial, because residues in these networks are not viable design targets if their role is to maintain the fold of the protein. Conversely, these networks of residues are ideal candidates for manipulating specificity if they are primarily involved in binding, such as the myriad interdomain interactions maintained within antibodies. Here, we identify networks of evolutionarily-related residues in C-class antibody domains by evaluating covariation, a measure of propensity with which residue pairs vary dependently during evolution. We computationally test whether mutation of residues in these networks affects stability of the folded antibody domain, determining their viability as design candidates. We find that members of covarying networks cluster at domain-domain interfaces, and that mutations to these residues are diverse and frequent during evolution, precluding their importance to domain stability. These results indicate that networks of covarying residues exist in antibody domains for functional reasons unrelated to thermodynamic stability, making them ideal targets for antibody design.

Proteins 2013; 81:884–895.
© 2012 Wiley Periodicals, Inc.

Key words: protein design; protein domains; protein evolution; multiple sequence alignment; covariation networks

INTRODUCTION

Antibody design is of great interest in the pharmaceutical field because of its usefulness to the development of novel therapeutics targeting many human diseases.[1–3] Antibodies bind desired ligands with high affinity and specificity, and can be engineered to recognize specific targets. However, full-length antibodies cannot reach some tissues or binding surfaces that are accessible to smaller molecules.[4] Single antibody domains may be more easily engineered for higher stability and increased circulation, thus making for better therapeutics.[5–8] Therefore, the engineering of individual antibody domains and antibody fragments is of particular interest for targeted therapeutic design.

Antibodies are glycoproteins belonging to the immunoglobulin (Ig) superfamily. Members of the Ig superfamily are characterized by their β-sandwich topology comprising two β-sheets, each forming a Greek-key folding motif.[9] Apart from antibodies, Ig domains are commonly found in cell adhesion molecules (CAMs), receptor tyrosine kinases (RTKs), immunomodulatory proteins, major histocompatibility complexes (MHCs), muscle proteins, and many other protein families with diverse functions.[10] Despite the remarkable variations in sequence and function,

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Nikolay V. Dokholyan, 120 Mason Farm Rd., Ste. 3097, Chapel Hill, NC 27599-7260. E-mail: dokh@unc.edu

Received 7 September 2012; Revised 5 December 2012; Accepted 14 December 2012
Published online 27 December 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24247

© 2012 WILEY PERIODICALS, INC.

the members of the Ig superfamily are grouped into only four folding "sets" (V, C1, C2, and I) in the database of structural classification of proteins (SCOP).[11] Antibodies typically comprise two heavy chains and two light chains made of variable (V-class) and constant (C-class) Ig-domains.

C-class Ig-domains are particularly of interest in therapeutic design because of their native ability to both bind diverse antigens and convey information to the immune system.[8] These domains interact with the immune system to determine the mechanism used to destroy the bound antigen, and also interact with V-class domains, which determine the specificity of the antibody.[12] Depending on the type of C-class domain in the antibody, the antibody may bind to various cell receptors to trigger phagocytosis, lysis, or degranulation.[13,14] Although V-class Ig-domains are responsible for specificity in antigen binding, C-class Ig-domains would make ideal scaffolds for the design of therapeutics to incite specific immune reactions. The engineering of C-class domains to directly bind antigens is another promising avenue for therapeutic design.[7] The rational design of any protein requires knowledge of the interactions necessary for the protein's stability and function, in order to guide the choice of residues subject to mutagenesis.[15] One method for gaining insight into these interactions is to examine the evolutionary variation of C-class domain sequences, such as the residues that are conserved in the domain. Examination of a set of diverse but structurally related sequences for residues that are highly conserved can reveal those positions critical for structural stability and protein function.[16] In addition to individual conserved residues, networks of covarying residues have been found in the interfaces between the variable and constant regions of Ig domains.[17,18] The residues in domain-domain interfaces are critical for determining the binding properties of antibodies, and hence, knowledge of correlations between such residues is invaluable for the design of novel antibody therapeutics. During the evolution of protein sequences, those residues that undergo mutation together in a set of structurally similar proteins are likely to be involved in the diversification of key interactions. Therefore, evolutionary information involving both individual residues and residue networks is crucial in order to uncover conservation and mutation patterns driven by selection for protein stability and function.

Two motivating forces exist for residues to be evolutionarily conserved, whether alone or as part of a covarying network: either the identity of the residue is crucial for thermodynamic stability, allowing the structure of the protein to be held intact to perform its function while not participating in aberrant interactions, or the residue contributes directly to the function of the protein outside of thermodynamic stabilization, such as participating in recognition for binding or an enzymatic reaction, which is necessary or beneficial to the cell. Without one of these two justifications, the survival of the organism would not be affected by the mutation or conservation of the residue, and the identity of the position would vary widely, independent of all other positions. Conserved residues, those residues exhibiting low sequence entropy over evolution, have long been hypothesized to be crucial to thermodynamic stability of the protein core.[19,20] Covarying residues, conversely, are by definition never among the most highly conserved individually, but consist of pairs of mutations that tend to occur together, and one is rarely present without the other over evolution. Although in some cases covarying residues have been shown to predict structural contacts,[21,22] the importance and meaning of distal covarying residues has been long debated and never resolved.[23–26]

In this study, we identify both individually conserved residues and evolutionarily-conserved networks of covarying residues from a diverse alignment of C-class sequences. We find that highly conserved and highly covarying positions have vastly different properties and contributions to the protein, although they tend to be near each other in sequence. Our structural and thermodynamic analyses confirm previous findings that highly conserved positions tend to be responsible for the stability of the fold. However, we show that the protein can be energetically stabilized by decoupling covarying pairs, suggesting that thermodynamics alone cannot justify covariation. This finding is further supported by the observation that highly covarying residues are often among the least-conserved positions in the C-class subfamily. We further observe that highly covarying residues appear to cluster in binding interfaces, where they form networks of interlinking covariation relationships, crucial information for design efforts. This finding supports the hypothesis that highly covarying positions appear to be important for binding and functional properties not related to thermodynamic stability.

METHODS

Selection of sequences for C-class domain library

We obtain a set of C-class sequences compiled from multiple publicly available databases: PDB, DDBJ, GenBank, EMBL, SWISS-PROT, and others compiled by NCBI, as described in Ref. [17]. We search the ASTRAL database for known C-class Ig fold structures, produce a structure-based sequence alignment, and utilize a Hidden Markov Model (HMMER software package,[27] as in Ref. [17]) to identify additional sequences. We filter the resulting 10,332 sequences for sequence identity of less than 85% in order to remove redundancy and avoid bias from over-represented sub-families,[28,29] and remove sequences with incomplete or obsolete annotation. We also remove sequences with length outside of one

standard deviation (11 residues) of the mean sequence length (99 residues), because these sequences affect the positional sequence entropy by introducing many gaps into the alignment. The resulting sequence database consists of 729 diverse, representative C-class sequences.

C-class sequence alignment

From the constructed C-class sequence database, we identify by sequence similarity the sequences of various major subfamilies including CH1, CH2, CH3, and CL. We select a representative structure for each identified subfamily. We then align the sequences using ClustalW.[30] In order to improve the alignment generated based on sequence alone, we use secondary structure information to guide the sequence alignment. We use this alignment of representative structures to create a profile for the sequences of their respective subclasses, assigning an identical set of gaps within each subclass. We use PSIPRED[31] to predict the secondary structure of the remaining, unclassified sequences in the sequence database, and use this information to align the remaining sequences to the existing profile.

Construction of sequence profile

Using the aligned C-class database, we construct a sequence profile for the C-class sequences using methods adapted from Dokholyan and Shakhnovich.[16] First, we calculate the frequency of each amino acid type (21 including a gap type) at each sequence position, from which we obtain the probability $p_k(\sigma)$ for each amino acid type σ at each position k in the alignment. Using these probabilities, we determine the consensus sequence for the aligned database and calculate the sequence positional entropy for each position k in the alignment:

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

Calculation of φ- and Ψ-values

We determine covariation of residue pairs by the calculation of their correlation coefficient from the multiple sequence alignment represented by one φ-value for each pair as described elsewhere:[17,32]

$$\varphi(x_i, y_j) = \frac{(x_i y_j \times \bar{x}_i \bar{y}_j) - (x_i \bar{y}_i \times \bar{x}_i y_j)}{\sqrt{(x_i y_j + \bar{x}_i y_j)\times(x_i \bar{y}_j + \bar{x}_i \bar{y}_j)\times(x_i y_i + x_i \bar{y}_j)\times(\bar{x}_i y_j + \bar{x}_i \bar{y}_j)}},$$

where, for example, $x_i\bar{y}_j$ is the number of times that amino acid type x is present in position i and amino acid type y is absent from position j. We perform these calculations using Python scripts developed in-house. These calculations produce a symmetric four-dimensional matrix, with dimensions of sequence position and amino acid type for each of the two covarying residues. In order to eliminate bias due to small sets of related sequences, we discard covarying pairs that are observed fewer than 10 times in our alignment.[32] To examine positional covariation, which is independent of amino acid type, we create a two-dimensional matrix by averaging across the amino acid type dimensions:

$$\psi(i, j) = \frac{1}{a}\sum_{x=1}^{a}\sum_{y=1}^{a}\phi(x_i, y_j),$$

where a represents the number of amino acid types.

Statistics

To demonstrate the statistical significance of the positional covariation, we randomly shuffle the columns of the C-class alignment, such that the entropy of each position is unchanged, but the correlations across positions are randomized, and re-calculate the Ψ-values. We perform this analysis 10,000 times, and construct a distribution of Ψ-values for each covariational pair. We then calculate the Z-score, defined as the number of standard deviations of the real Ψ-value from the average of the randomized distribution. In the most relevant cases (Supporting Information Table S2), the Ψ-value that we obtain for the pair in the C-class alignment lies outside the distribution, and the Z-score is so large that a meaningful P value cannot be calculated.

Computational mutagenesis of the CH1 domain

We use Eris[33–35] to computationally mutate residues featuring high covariation (Results) to all possible amino acids in order to determine the most energetically favorable combinations. During these simulations, we extensively sample all rotameric and subrotameric orientations for the side chains of all residues in the CH1 domain. In this process, we maintain the backbone as static in order to conserve the overall fold of the domain. Because this process is stochastic, we perform 1000 independent trials and obtain the amino acid sequence from each trial. From these 1000 designed sequences we determine the consensus sequence and sequence positional entropy at each position as described above.

Identification and analysis of covariation networks

We represent the CH1 domain as a weighted graph of sequence positions, with each node representing a sequence position and each edge representing the covariation relationship between the respective nodes. In this representation, the edge weight is equal to the covariation value between the two node residues. We apply various cutoffs to covariation values for the formation of an edge and observe the effect in the size of the largest subgraph component. We calculate the degree of each node

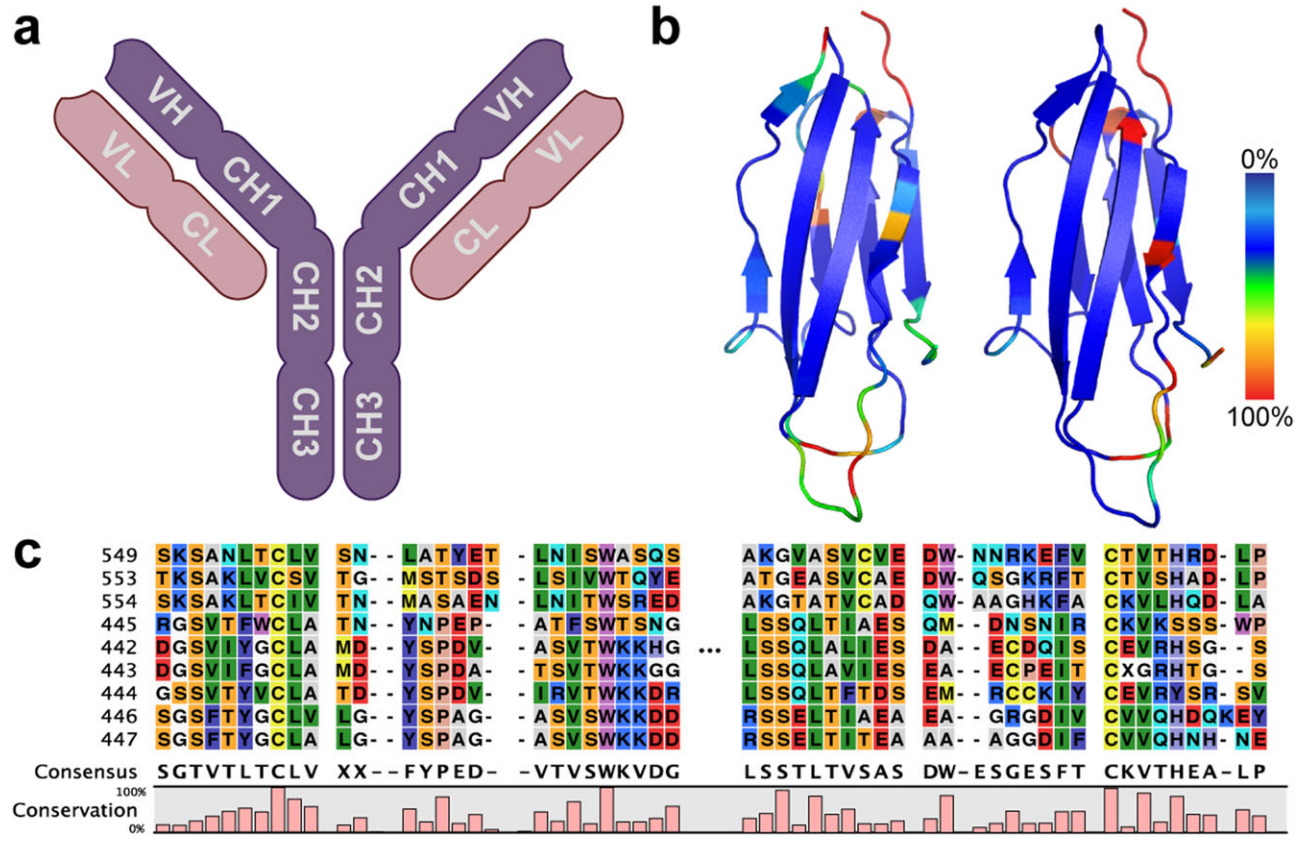

Figure 1

Database of C-class Ig-like sequences. (**a**) Cartoon of the spatial organization of constant and variable domains in an immunoglobulin molecule. (**b**) Structure of a CH1 domain (PDB ID: 1A5F) colored based on the percentage of gaps in the C-class alignment before (left) and after (right) incorporation of secondary structural constraints in building the sequence alignment. (**c**) An excerpt from the C-class sequence alignment constructed by imposing secondary structural constraints. The excerpt shows the conserved cysteines in the C-class Ig domains.

at each cutoff of interest, or how many covariation connections are formed by a certain residue at the given cutoff.

RESULTS

Aligned C-class sequence database

To use evolutionary information to aid in the design of stable antibodies, we construct an aligned database of diverse C-class Ig fold sequences (Methods). Due to significant diversity, both in terms of length and sequence composition, multiple sequence alignment (MSA) of the C-class sequence database performed using ClustalW results in common alignment artifacts such as dominance of gaps and misalignment of known evolutionarily conserved residues.[36,37] To improve the alignment, we generate an MSA by incorporating secondary structure information into ClustalW to guide the alignment. Inclusion of secondary structure information, whether from a representative structure or predicted using PSI-PRED (Methods), resolves these issues resulting in a significant improvement in the quality of the alignment [Fig. 1(b)]. Our final alignment has 160 positions. Sixty-six positions are occupied by gaps in the consensus sequence [Fig. 1(c), Supporting Information). The alignment consists of 729 annotated C-class sequences, 247 (~34%) of which are assigned to a subfamily of immunoglobulin constant domains [CH1, CH2, CH3, or CL; Fig. 1(a)] and associated with a representative structure. These sub-families comprise 48, 47, 82, and 82 sequences, respectively. The remaining 482 sequences are not explicitly annotated as belonging to one of these subfamilies, and feature relatively low sequence identity to those associated with subfamilies. To estimate the diversity of the constructed database of C-class sequences, we grouped the sequences based on the originating species. The C-class sequence database represents 24 different species of mammals, reptiles, and amphibians, with at least 10 sequences from each species. Out of the 729 sequences in the database, 41 sequences were not classified as belonging to any particular species. These results suggest that the C-class

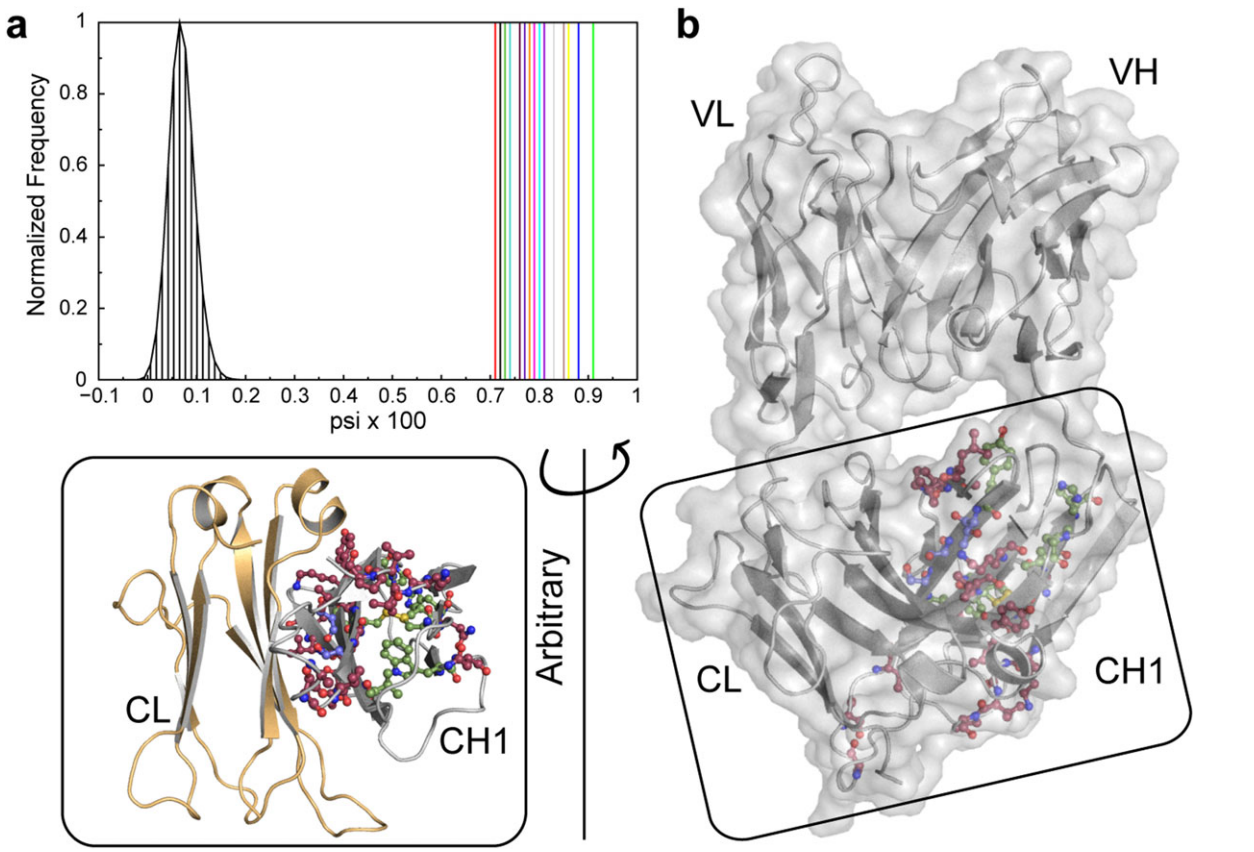

Figure 2

Conserved and covarying residues in C-class Ig-domains. (**a**) Histogram of Ψ-values of the top 24 highly covarying residue pairs in the C-class database after performing randomized shuffling of residues in the corresponding positions. Lines represent Ψ-values before shuffling. Large separation between the histogram and the actual values indicates significance of the Ψ values. (**b**) Heavy chain Fab fragment of a monoclonal anti-E-selectin antibody (PDB ID: 1A5F), comprised of a CH1 domain and a CL domain. Conserved residues are highlighted in green and covarying residues are in red. Inset (lower left) shows the CH1-CL interface with CL domain shown in orange. The most highly covarying residues are in the interface between CH1 and CL domains and are colored blue.

database is diverse and the sequences are not biased toward any species.

To identify highly conserved residues within the sequence database, we calculate the entropy at each sequence position of the C-class database (Methods). Sequence positions having low positional entropy ($S < 1$; Methods) across multiple domain sub-classes are likely to be conserved amino acids responsible for domain stability. The two cysteine residues known to be conserved in C-class Ig-fold domains[36,37] have very low positional entropy (Supporting Information Fig. S1, Table S1). We also observe instances of conserved residues in two prolines (here positions 1 and 26), Trp33, Val78, His80, and Phe24 (Supporting Information Fig. S1, Table S1; Fig. S3 for consensus sequence numbering). These residues form the structural core of the domain, indicating that they are important for C-class Ig-fold stability (Fig. 2).[36–38] These observations indicate that the constructed sequence database captures evolutionary information pertinent to C-class Ig-fold sequences with varying evolutionary functions.

Covariation in residue pairs in C-class sequences

To study residue networks within the C-class domain family, we calculate the covariation of pairs of residues within the alignment as the φ-value,[17] which we calculate based on both sequence position and residue identity, forming a four-dimensional matrix (Methods). The advantage of using φ-values, as opposed to other similar measures of covariation, is that the φ-value represents the correlation coefficient not only between two sequence positions, but between specific amino acid types at those positions.[32] Φ-values therefore provide additional insight into which particular amino acids are strongly correlating at these evolutionarily-linked sites, which is of interest from a design perspective. To obtain a more general view

of covariation within the domain, we reduce the dimensionality of this matrix to compute Ψ-values, indicative of the average covariation over all residue types for any two given sequence positions (Methods). We sort all possible sequence position pairs by their Ψ-value, and select for further analysis the most highly covarying pairs in which both residues in the pair have sequence positional entropy >2.0. This procedure resulted in 24 covariation pairs comprising 20 distinct sequence positions (Supporting Information Fig. S1, Table S2). To determine the significance of these results, we randomly shuffle the columns of the C-class alignment, such that the entropy at each position is unchanged, but the correlations across positions are randomized, and re-calculate the Ψ-values (Methods). We find that the Ψ-values of the highly covarying pairs of positions are statistically significant as indicated by their *Z*-scores [Fig. 2(a), Supporting Information Table S2). Thus, the set of highly covarying residues and the set of highly conserved residues are mutually exclusive.

Structural location of highly covarying residues

We choose the IgG CH1 domain as a model for *in silico* evaluation of the structural and thermodynamic importance of highly covarying residues. The interactions of the CH1 domain are critical for association of the antibody heavy and light chains. The CH1 domain is intrinsically unfolded, and is part of a quality-control mechanism for the proper secretion of fully assembled, heterotetrameric IgGs.[39] The stability of the folded domain is derived through its interactions with light chain CL domains. Therefore, we anticipate that the domain interfaces of CH1 feature regions of high conservation and covariation that provide elements of both protein stability and protein function. Very few antibody engineering efforts have centered on the CH1 domain, even though it has high potential for generating new and novel antibody platforms. For these many reasons, we feel that the CH1 domain is an ideal model domain for applying our C-class observations.

To determine if highly covarying residues have preferred locations in the structure of the domain, we map the 20 most highly covarying sequence positions to the structure of a CH1 domain (PDB ID 1A5F, CH1 domain) [Fig. 2(b), Kabat numbering in Supporting Information Fig. S2]. We find that the most highly conserved residues tend to co-localize with the highly covarying positions, often alternating in the sequence. We observe that conserved residues form the core of the domain, while the covarying residues are surface-exposed. Indeed, in full-length Fab, we find that the most highly covarying residues participate in interactions with the CL (light chain) domain, facilitating binding [Fig. 2(b)]. The three most highly covarying positions, which participate in 6 to 9 of the top 24 highest-covariation pairs (Supporting Information Table S2), are directly in contact with the CL domain (Fig. 2). To show the transferability of this phenomenon to other types of C-class sequences, we map highly covarying positions to the CL domain of the same Fab structure. Although sequence positioning of the CL domain in the C-class alignment differs significantly from that of the CH1 domain, we observe that the highly covarying positions in the CL domain also are located in the CH1-CL interface [Fig. 2(b)]. Therefore, we conclude that highly covarying residues are located in binding interfaces, suggesting that highly covarying residues and the networks that they form are involved in binding interactions. We further note that only 8 of the 24 highly covarying pairs are in structural contact (C_β-C_β distance <7.5 Å), indicating that the majority of these pairs do not directly structurally influence each other.

Stabilization of the CH1 domain via rational mutagenesis of the covarying residues

To determine whether the C-class domain can be stabilized by selective mutation of the most highly covarying sequence positions, we trace all positions identified as the highest-covariation (Supporting Information Table S2) to a template structure of a CH1 domain (PDB ID 1A5F, CH1 domain with variable and CL domains removed[40]) and simultaneously exhaustively mutate only these most highly-covarying residues using the Eris suite[33–35] (Methods). As we note above, the CH1 domain is intrinsically unfolded in the absence of the CL domain, and thus we maintain the backbone as static while allowing the side chains to explore all rotameric and subrotameric states. Using Eris, we calculate the change in ΔG of the protein upon mutation for all possible amino acids at the specified positions simultaneously, and determine the most energetically favorable sequence. We perform 1000 independent iterations, and calculate the resulting sequence positional entropy at the high-covariation positions. We find that most of the highly covarying positions have low sequence positional entropy ($S < 1$) among the Eris-stabilized sequences, contrary to the high entropies ($S > 2$) that we observe in the database of naturally occurring C-class sequences (Table I). In addition, the most favorable residue (identified using Eris) in each position is usually similar but not identical to the corresponding residue in either the C-class database consensus sequence or the sequence of the CH1 template structure (PDB ID 1A5F, CH1 domain). This finding is supported by experimental evidence from directed evolution.[41] For example, position 153 is a serine in the C-class database consensus sequence ($S = 2.6$), a threonine in a consensus sequence of only CH1 domains in the C-class database ($S = 1.5$), an isoleucine in the 1A5F CH1 domain, and a valine ($S = 0.065$) in the Eris-stabilized sequences. This result indicates that, although there exists a

Table I
Entropy of Highest-Covariation Residues

Position	C-class entropy	C-class consensus residue	1A5F residue	Eris consensus residue	Eris entropy	Eris consensus residue (with CL domain/fixed CL domain)	Eris entropy (with CL domain/fixed CL domain)
16	2.5	F	Y	V	0.41	H/Y	1.8/0.13
36	2.6	L	T	G	0.051	G	0.11
38	2.5	G	S	T	0.22	T	0.54/0.21
40	2.2	V	V	T	0.18	T	0.88/0.68
43	2.1	T	G	G	0.0	G	0.0
45	2.0	L	L	L	0.28	L	0.43/0.080
47	2.4	T	K	V	0.37	V/I	1.6/1.0
65	2.4	L	N	N	0.0	N	0.0
95	2.3	P	P	D	0.29	Q	1.1/1.0
96	2.5	P	A	G	0.46	G	0.10/0.0079
97	2.5	L	V	E	0.71	G/I	1.0/0.76
98	2.4	P	L	E	0.14	E	0.19/0.13
110	2.2	S	T	V	0.90	I	1.0/0.24
112	2.4	S	S	V	0.70	L	1.1/0.58
114	2.6	T	S	I	1.4	G/S	0.87/0.27
134	2.5	V	N	V	1.0	V	1.0/0.99
152	2.5	W	K	E	1.0	E	0.70/0.55
153	2.6	S	I	V	0.065	V	0.15/0.069
154	2.7	P	V	E	1.3	E	0.94/0.92
155	2.6	S	P	A	0.75	A	0.77/0.64

We list the 20 sequence positions participating in the pairs listed in Table S2, with their positional entropies in natural C-class sequences and sequences stabilized on the 1A5F structure using the Medusa force field. We also list the consensus C-class and Medusa-stabilized residues at each position, in addition to the residue at that position in the 1A5F template structure. Position numbers are according to the consensus alignment of the C-class database. All highly-covarying residues have high entropies among natural C-class sequences, yet most have low entropies among the Medusa-stabilized sequences. This implies a functional role for these residues, since even though there exists a favored residue in the position, among natural sequences the residue identity varies, suggesting that thermodynamic stability is not the determining factor for residue identity at this position.

thermodynamically most stable residue at these most highly covarying positions, this residue is highly variable, suggesting that the most highly covarying residues are not involved in stabilization of the domain.

Mutations with favorable $\Delta\Delta G$ are not necessarily unique to covarying residues; many proteins are known to be only marginally stable in their wild-type form.[42] However, the fact that the residues that we examine are highly covarying with high statistical significance over evolution makes them unique and worthy of study. If the covariation relationship exists in order to preserve domain stability, we would expect decoupling of this covariation relationship, as in mutation to only one residue of the pair, to be destabilizing. However, we find the opposite effect. This finding implies that thermodynamic stability alone cannot account for covariation relationships, and another reason exists for the residues to mutate in tandem. Evolutionary relationships are created to preserve fitness by preserving function or structure. In this case, we eliminate the possibility of a stabilizing effect, and hypothesize that the covariation relationship between these residues may be important for function and binding interactions.

To determine whether the set of the most highly covarying residues is responsible for stabilization of the complex of the C-class domain with its binding partner, we use Eris to mutate the same positions of the CH1 domain exhaustively while in complex with the CL domain of the light chain. We obtain a nearly identical consensus sequence as to when we allow such mutation of CH1 in isolation (Table I), a surprising result because we expect the presence of the CL domain to influence residue choice. However, although the examined residues exist in the CH1-CL interface, in a beta-sandwich domain such as those studied here each individual residue is responsible not only for outside interactions, but also for intradomain interactions. The obtained mutations likely stabilize these intradomain interactions or optimize the protein backbone for the formation of the beta-sandwich fold. However, the entropy at these sequence positions among the various iterations is surprisingly higher than in the isolated protein, albeit less than that found in the natural C-class sequence database. When we treat the CL domain as a rigid object, not allowing the side chains to rearrange to accommodate mutations, the positional entropy decreases to approximately the same levels as in solution. The observation that higher sequence positional entropy is found in CH1 when bound to CL suggests that the presence of a binding partner confers greater flexibility in amino acid identity at the given positions. This greater flexibility allows multiple residues to be represented at these positions, as opposed to in the unbound protein where one or two residues are overwhelmingly favored. This observation suggests that highly covarying residues are not likely to be involved in energetic stabilization of the bound complex. Although some

conserved and covarying residues must certainly confer stability upon the bound complex, since CH1 is inherently disordered unless bound to CL,[39,43] decoupling of the covariation relationships by the individual mutation of each these highly covarying residues not only fails to destabilize the complex, but actually stabilizes it. This finding suggests a mechanism other than thermodynamic stability for the phenomena of covariation in C-class antibody domains.

To determine if we can assign a functional role to certain residues, we search for sequence positions at which the most stable residue type as identified using Eris is deselected in evolution. At two positions, the entropy in the Eris-stabilized sequences is high ($S > 1$), and the residues resulting in that position are of a different type from either the C-class consensus sequence or the 1A5F CH1 domain: position 114 is a threonine in the C-class consensus sequence ($S = 2.69$), a serine in the 1A5F CH1 domain, and an isoleucine ($S = 1.4$) in the Eris-stabilized sequences (Table I). This finding suggests that these residues are important for function, because stability alone dictates that they should have a different hydrophobicity than they have in evolution. The finding that the set of highly covarying residues can be further thermodynamically stabilized by forced decoupling of covariation using computational means suggests that the covariation relationships between these residues are conserved for functional reasons other than thermodynamic stability, and not for the stability of the domain. Variation of these residues could have important design implications, as the stability of the domain would not be adversely affected, but the functional properties are likely to change.

Networks of highly covarying residues in the CH1-CL binding interface

We observe that many of the residues identified as highly covarying participate in more than one high-covariation relationship, connecting covarying residues into networks that reach over the entire domain. Taking these relationships into account is important when choosing positions that may be mutated to produce a desired effect in the protein domain. Hence, knowledge of the covariation network (i.e., which residues are affected by the mutation of any one position) is imperative to design efforts. In order to explore the extent and location of the densest areas of these networks, we represent the covariation profile of the CH1 domain as a complete weighted graph, with the C_α atoms of the residues as nodes and the covariation of any pair of residues as the corresponding edge weight. We apply covariation cutoffs ranging from $\Psi = 0.001$ to $\Psi = 0.009$, in increments of 0.001, to the graph. For each cutoff, we remove edges with weights less than the cutoff, and determine the size of the largest component in the resulting subgraphs [Fig. 3(a)]. The graph remains as one large component until a cutoff of $\Psi = 0.002$, and is completely dissolved into individual nodes at a cutoff of $\Psi = 0.009$. We examine the covariation network defined using the graph with the edge cutoff for which the largest component contains approximately 50% of the total number of nodes in the graph, here $\Psi = 0.005$ [Fig. 3(b)]. We observe that the covariation network that we describe is comprised of nodes with varying degrees. Some nodes are connected to the network by only one edge, whereas some nodes have many connections, following a Poissonian distribution of node degree over the network [Fig. 3(c)]. There exists a subset of nodes, which we designate as hubs, having degree 25% greater than the next most connected node. These nodes have many connections to other nodes in the network, and retain their status as network hubs as we increase the cutoff for edge formation. These residues (C-class alignment positions 110, 112, and 114; 1A5F CH1 domain T63, S65, and S67) are members of the largest covariation network in the domain, and have strong covariation with several other residues in the network, as well as among each other. Furthermore, we note that these hub residues are located in the CH1-CL binding interface but covary with residues throughout the domain [Fig. 3(d)]. Thus, covariation networks are likely important to the functionality and binding properties of the C-class domains.

DISCUSSION

To apply rational design to the engineering of antibodies, we must have detailed knowledge of the structure and function of each antibody domain to guide our choice of mutation sites. For instance, in this study, we observe that mutations to key core residues significantly decrease the stability of the protein, while mutations to key interface residues affect the binding properties. The information derived from our analysis method of identifying conserved residues and covarying residue networks can potentially be applied to select key "hotspots" for antibody design. By analyzing the sets of conserved and covarying residues in C-class sequences, we uncover patterns and networks of conservation, the knowledge of which can steer the rational design of antibodies for improved stability and binding specificity.

Conserved residues, those with low positional entropy in a database of C-class Ig-domain sequences, often are vital for domain stability. Because of their importance in maintaining the structure of the domain, these residues are almost uniformly conserved across sequences comprising the C-class, and should not be mutated in the design process.[44] For example, the two conserved cysteines common to the Ig-fold superfamily form a disulfide bond in the core of the protein, linking the two β-sheets that make up the tertiary structure of the fold.[10] The majority

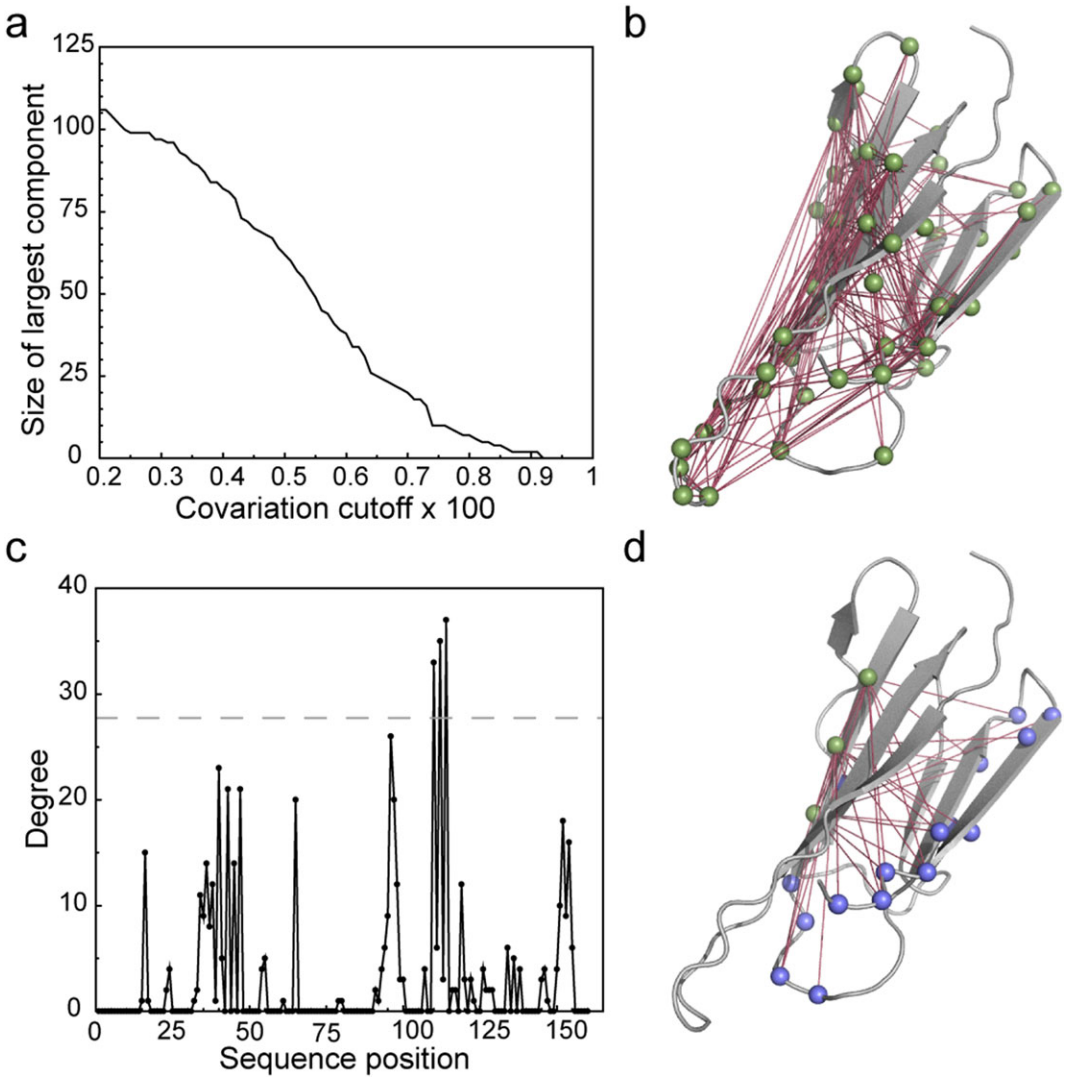

Figure 3

Covariation networks in C-class Ig-domains. (**a**) Size of the largest network component decreases when the cutoff in pairwise residue covariation is increased. (**b**) Covariation network with edges satisfying a Ψ-value covariation cutoff of 0.005, applied to the heavy chain Fab fragment of a monoclonal anti-E-selectin antibody (PDB ID: 1A5F). (**c**) Degree of each node in the covariation graph (cutoff: $\Psi = 0.005$). The top 25% of the nodes (above the gray line) are designated as hubs in the graph. (**d**) Covariation network hubs (residues 63, 65, and 67 in the structure of 1A5F CH1 domain) are shown as green spheres, with lines representing the edges connecting the hubs to other covarying residues (blue). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of the remaining highly conserved residues are hydrophobic, holding positions in the core of the fold forming tertiary interactions between β-strands [Fig. 2(b)]. In our example structure, two of the highly conserved residues are present in the interface between the CH1 and the VH domains of antibodies [Fig. 2(b), Supporting Information Table S1], and two are present in what would be the interface between the CH1 the CH2 domains. These residues are highly conserved across all C-class domains despite their various binding partners, implying that these residues contribute to the stability of the interdomain interaction rather than to the binding specificity.

Conversely, we find that residue pairs and networks showing high levels of covariation consist of poorly conserved residues. In order for a residue pair to exhibit high covariation, the pair must not only appear together often within the sequence database, but also must be absent together. A perfectly conserved residue is never absent, and so would exhibit poor covariation with any other residue, even another perfectly conserved residue. Therefore, we do not expect individual residues within highly covarying residue networks to be highly conserved. However, the fact that these residues correlate with each other so strongly implies that these residue networks

possess some evolutionary advantage, although they may not be necessary for the stability of the domain as a whole. This finding suggests that these covarying residues are instead important for function,[45] the binding specificity of the constant domains to their respective variable domains, light chains, or other constant domains, and may therefore be manipulated when designing antibodies for therapeutic use. Sparse networks of co-evolving residues have been shown to connect functional areas in other types of protein domains.[46,47] The knowledge of statistical coupling of active site residues to distal regions has been applied to the design of simple protein domains,[48] demonstrating the feasibility of rational design using the knowledge obtained from covariational analysis.

In some cases, pairs of covarying residues in the core of a protein may directly impact protein stability. In a study of covarying residues in SH3 domains, Larson *et al.* found compensatory residue covariations that stabilize the core of the SH3 domain.[32] These covarying residue pairs are in the hydrophobic core of the domain, were in most cases due to direct steric or electrostatic effects, and were not distal surface pairs such as those that we find in our study. When we examine the amino acid identity-averaged covariation measure, Ψ, we find that the most highly covarying positions also have high sequence positional entropy ($S > 2$) (Table I). However, in contrast to the high variation observed in evolution, exhaustive computational mutation of residues in these positions with exceptionally high covariation results in a set of energetically stable sequences, which at most positions has low entropy. The consensus sequence for this set of energetically stable sequences is at most positions identical or similar to C-class consensus or template structure (Table I). Even when accounting for solvation of the CH1-CL domain interface, two positions feature a stabilizing mutation of a different type from both the C-class consensus and the template structure residues (e.g. hydrophobic instead of polar), suggesting that the most stable residue in this position is actually deselected in evolution, strengthening the hypothesis that these positions are important to protein function.

The decrease in entropy observed at highly covarying sites upon exhaustive sampling of mutations could be a result of the single structure used in this study to evaluate the effect of mutation; the obtained sequences could be highly structure-specific. Positions with high entropy over evolution do not necessarily have high entropy in a single structure. However, such an interpretation would imply that the structures of the various C-class sequences vary widely at these locations, which would conflict with the known result that the conserved residues do not vary and in fact hold together the core of the protein, but are present in the same areas as the covarying residues.

When these same CH1 positions are exhaustively sampled for residue type in the presence of its binding partner, the CL domain, we obtain a nearly identical consensus sequence as when CH1 is alone in solution. However, the sequence positional entropies are in almost every case significantly higher in the presence of the CL domain than when CH1 is in isolation (Table I). This finding suggests that the presence of the CL domain makes several different residues probable at these positions, while in isolation a more definite candidate exists for the most thermodynamically stable sequence. When we do not allow the side chains of the CL domain to rearrange to accommodate mutations to CH1, the positional entropies decrease to approximately the levels found in isolation (Table I), suggesting that the nature of the inter-domain interactions allows for greater flexibility than do solvent interactions. Given the high structural similarity between C-class domains, and our finding that highly covarying residues are found in domain interfaces in both CH1 and CL domain structures, we expect similar results for other C-class domains.

We determine that highly covarying residue positions are not involved in stabilization of the domain using exhaustive Eris mutation (Results), which raises several interesting possibilities for the conservation of highly covarying networks. The criteria for evolutionary conservation or correlation indicate that if a behavior does not increase stability, then it must benefit function or prevent the loss of function in some other manner. In this case, highly covarying residues must have some other functional role not apparent from the study of stability of one complex alone, such as the specificity of binding or binding kinetics. Magliery *et al.* found that covariation of poorly-conserved positions in TPR domains permits the maintenance of a near-neutral surface charge distribution.[49] Another possibility is that the sequence of the CH1 domain is not thermodynamically optimal for the given CL domain, but it may be the best fit that will exclude other binders. Or, perhaps the most stable complex is not evolutionarily favorable because the domains must be allowed to dissociate for a functional role. Further experimental studies will shed light on which of these functional possibilities could be the justification for evolutionarily highly covarying residues.

The representation of highly covarying residues as a covariational network provides the opportunity to further examine the localization of these residues in the domain, as well as to more stringently define those residues most of interest for additional experimental studies. The finding that not only is there an overwhelming localization of highly covarying residues in the CH1-CL binding interface, but also that network connections are stronger in and between residues in the interface than to those outside of it, is notable and further argues for a functional purpose for covarying residues and their networks. Moreover, the hubs of this covariational network, or those residues having the most and the strongest covariation connections to other residues, are in this same interface: alternating residues located in the same strand.

Mutations to residues within the covariational network affect the entire binding interface, and the effects can spread across the entire domain. However, as discussed earlier, these residues have high entropy across diverse C-class Ig-fold protein domains with varying functions and are not conserved for the stability of the protein. These residues are located in a binding interface, and yet they change their identity so readily over evolution without affecting stability, which implies that they are both amenable to mutation and that mutation may be related to the different binding properties adopted by the different members of the domain family. Hence, they may likely be mutated to modulate the function of the protein.

ACKNOWLEDGMENTS

The authors thank Dr. Srinivas Ramachandran, Dr. Dikran Aivazian, and Dr. Alexey A. Lugovskoy for their helpful discussion and suggestions, and Mr. Norman Wang for his initial work on the project.

REFERENCES

1. Kaneko E, Niwa R. Optimizing therapeutic antibody function: progress with Fc domain engineering. BioDrugs 2011;25:1–11.
2. Caravella JA, Wang D, Glaser SM, Lugovskoy A. Structure-guided design of antibodies. Curr Comput Aided Drug Des 2010;6: 128–138.
3. Chan AC, Carter PJ. Therapeutic antibodies for autoimmunity and inflammation. Nat Rev Immunol 2010;10:301–316.
4. Gong R, Wang Y, Feng Y, Zhao Q, Dimitrov DS. Shortened engineered human antibody CH2 domains: increased stability and binding to the human neonatal receptor. J Biol Chem 2011;286:27288–27293.
5. Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. Nat Biotechnol 2005;23:1126–1136.
6. Demarest SJ, Glaser SM. Antibody therapeutics, antibody engineering, and the merits of protein stability. Curr Opin Drug Discov Dev 2008;11:675–687.
7. Dimitrov DS. Engineered CH2 domains (nanoantibodies). MAbs 2009;1:26–28.
8. Demarest SJ, Rogers J, Hansen G. Optimization of the antibody C(H)3 domain by residue frequency analysis of IgG sequences. J Mol Biol 2004;335:41–48.
9. Bork P, Holm L, Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309–320.
10. Williams AF, Barclay AN. The immunoglobulin superfamily—domains for cell surface recognition. Annu Rev Immunol 1988;6: 381–405.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
12. Schroeder HW, Jr, Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol 2010;125(2Suppl 2):S41–S52.
13. Heyman B. Complement and Fc-receptors in regulation of the antibody response. Immunol Lett 1996;54:195–199.
14. Woof JM, Burton DR. Human antibody-Fc receptor interactions illuminated by crystal structures. Nat Rev Immunol 2004;4:89–99.
15. Michaelson JS, Demarest SJ, Miller B, Amatucci A, Snyder WB, Wu X, Huang F, Phan S, Gao S, Doern A, Farrington GK, Lugovskoy A, Joseph I, Bailly V, Wang X, Garber E, Browning J, Glaser SM. Antitumor activity of stability-engineered IgG-like bispecific antibodies targeting TRAIL-R2 and LTbetaR. MAbs 2009;1:128–141.
16. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001;312:289–307.
17. Wang N, Smith WF, Miller BR, Aivazian D, Lugovskoy AA, Reff ME, Glaser SM, Croner LJ, Demarest SJ. Conserved amino acid networks involved in antibody variable domain interactions. Proteins 2009;76:99–114.
18. Jordan JL, Arndt JW, Hanf K, Li G, Hall J, Demarest S, Huang F, Wu X, Miller B, Glaser S, Fernandez EJ, Wang D, Lugovskoy A. Structural understanding of stabilization patterns in engineered bispecific Ig-like antibody molecules. Proteins 2009;77:832–841.
19. Di Nardo AA, Larson SM, Davidson AR. The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. J Mol Biol 2003;333:641–655.
20. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29:7133–7155.
21. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
22. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–340.
23. Noivirt O, Eisenstein M, Horovitz ,A. Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel 2005;18:247–253.
24. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44:7156–7165.
25. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
26. Pritchard L, Bladon P, M O Mitchell J, J Dufton M. Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng 2001;14:549–555.
27. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
30. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–2948.
31. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
32. Larson SM, Di Nardo AA, Davidson AR. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 2000;303:433–446.
33. Yin S, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. Nat Methods 2007;4:466–467.
34. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comput Biol 2006;2:e85.
35. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. Structure 2007;15:1567–1576.
36. Rose DR, Przybylska M, To RJ, Kayden CS, Oomen RP, Vorberg E, Young NM, Bundle DR. Crystal structure to 2.45 A resolution of a monoclonal Fab specific for the Brucella A cell wall polysaccharide antigen. Protein Sci 1993;2:1106–1113.
37. Amzel LM, Poljak RJ. Three-dimensional structure of immunoglobulins. Annu Rev Biochem 1979;48:961–997.
38. van Zelm MC, Smet J, van der Burg M, Ferster A, Le PQ, Schandene L, van Dongen JJM, Mascart F. Antibody deficiency due to a missense mutation in CD19 demonstrates the importance of the conserved tryptophan 41 in immunoglobulin superfamily domain formation. Hum Mol Genet 2011;20:1854–1863.

39. Feige MJ, Groscurth S, Marcinowski M, Shimizu Y, Kessler H, Hendershot LM, Buchner J. An unfolded CH1 domain controls the assembly and secretion of IgG antibodies. Mol Cell 2009;34:569–579.
40. Rodriguez-Romero A, Almog O, Tordova M, Randhawa Z, Gilliland GL. Primary and tertiary structures of the Fab fragment of a monoclonal anti-E-selectin 7A9 antibody that inhibits neutrophil attachment to endothelial cells. J Biol Chem 1998;273:11770–11775.
41. Bershtein S, Goldin K, Tawfik DS. Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol 2008;379:1029–1044.
42. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332: 449–460.
43. Rothlisberger D, Honegger A, Pluckthun A. Domain interactions in the Fab fragment: a comparative evaluation of the single-chain Fv and Fab format engineered with variable domains of different stability. J Mol Biol 2005;347:773–789.
44. Demarest SJ, Chen G, Kimmel BE, Gustafson D, Wu J, Salbato J, Poland J, Elia M, Tan X, Wong K, Short J, Hansen G. Engineering stability into *Escherichia coli* secreted Fabs leads to increased functional expression. Protein Eng Des Sel 2006;19:325–336.
45. Xu Y, Tillier ER. Regional covariation and its application for predicting protein contact patches. Proteins 2010;78:548–558.
46. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
47. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
48. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437:512–518.
49. Magliery TJ, Regan L. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. J Mol Biol 2004;343:731–745.