Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain

Boris Steipe[1]†, Britta Schiller[1], Andreas Plückthun[2] and Stefan Steinbacher[1]

[1] *Abteilung Strukturforschung, Max-Planck Institut für Biochemie*
Am Klopferspitz, D-82152 Martinsried, Germany

[2] *Biochemisches Institut, Universität Zürich*
Winterthurerstr. 190, CH-8057 Zürich, Switzerland

Immunoglobulin variable domains are generally thought of as well conserved platforms providing the base for antigen binding loops of highly varying sequence and structure. However, domain evolution must ensure a balance between optimizing antigen affinity and the requirements of a stable, cooperatively folding domain. Since random mutations can carry a significant penalty for domain stability, constraints are imposed both on the repertoire of germline sequences and on somatic amino acid replacements during affinity maturation. Analyzing these constraints in the conceptual framework of statistical mechanics, we have been able to predict stabilizing mutations in the McPC603 V_κ domain from sequence information alone with better than 60% success rate. The validity of this concept not only has far reaching implications for antibody engineering but may also be generalized to engineer other proteins for higher stability.

Keywords: protein stability; immunoglobulin variable domain; point mutation; stability prediction; canonical ensemble

Theories of protein folding and protein stability are linked through the finding that the native state is the structure with the lowest free energy. Quantitative predictions of the adopted structure and the stability of the native state have been confounded by the difficulties involved in calculating the small difference of many large enthalpic and entropic contributions, which drives the transition from the unfolded to the native state (Matthews, 1987). Thus the rational engineering of either structure or stability remains an unsolved problem (Shi *et al.*, 1993). The alternative approach of screening random point mutations (Rollence *et al.*, 1988; Chen & Baldwin, 1989; Turner *et al.*, 1992) succeeds in only one out of 10^3 to 10^4 mutations (Risse *et al.*, 1992; Arase *et al.*, 1993) and screening procedures are restricted to proteins with some identifiable function such as enzymatic activity. We have addressed the problem of stabilizing an immunoglobulin domain, with an approach that analyses the immunoglobulin sequence database (Kabat *et al.*, 1991) in the conceptual framework of statistical mechanics.

Clusters of sequence variability in well delineated regions of the antibody molecule provide the molecular basis for the capacity of the immune system to respond to a large variety of antigenic challenges (Wu & Kabat, 1970). Antibodies thus appear to possess conserved structural frameworks with a repertoire of variable, antigen binding loops (Davies & Metzger, 1983; Chothia & Lesk, 1987). The genes of immunoglobulin variable domains have diverged by multiple gene duplications and mutations. Selected genes are subjected to an accelerated, local evolution that optimizes the capacity of the antibody to bind to antigen structures selectively and with high affinity (Tonegawa, 1983; Berek & Milstein, 1989; French *et al.*, 1989). In this process of affinity maturation, the whole of the domain coding sequence is randomly mutated and B-cells producing improved antigen receptors are selected and propagated (Berek & Ziegner, 1993). Even though selection for antigen binding plays a dominant role for any specific sequence, the measure of fitness in this process must be a composite of factors including antigen affinity, domain stability, assembly and interaction of the heavy and the light-chain, variable and constant domains, protease resistance and competence for export and secretion; the relative contribution of each factor differs from position to position of the domain.

Our approach is based on the hypothesis that the immunoglobulin repertoire approximates a canonical ensemble of sequences, each derived from one of a set of germ-line sequences and selected in a process

† Author to whom all correspondence should be addressed.

0022-2836/94/280188-05 $08.00/0

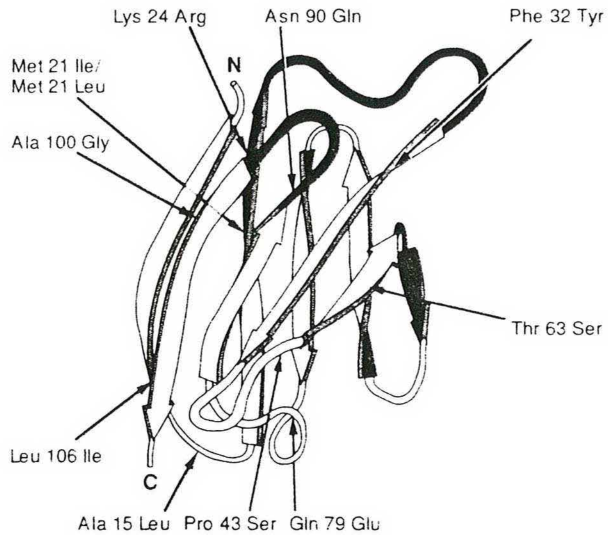

Figure 1. Distribution of analyzed point mutations over the structure of the V_κ domain of McPC603 (Steipe *et al.*, 1991). Amino (N) and carboxy (C) termini are labeled. Complementarity-determing regions are shaded dark. Mutations were made in CDR1 (Phe32Tyr), CDR3 (Asn90Gln), at the interface with the V_H domain (Pro43Ser and Ala100Gly), near the interface with the C_κ domain (Gln79Glu), in the hydrophobic core (Met21Ile, Met21Leu, Leu106Ile) and at solvent exposed positions (Ala15Leu, Lys24Arg, Thr63Ser).

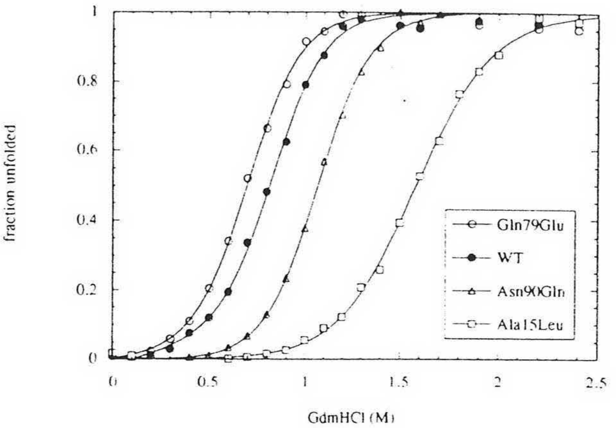

Figure 2. Equilibrium unfolding curves for Gln79Glu, WT, Asn90Gln and Ala15Leu plotted as fraction of unfolded protein *versus* concentration of denaturant. Mutant genes were constructed by site-directed mutagenesis of the WT gene in an expression plasmid derived from pASK30 (Skerra & Plückthun, 1989) by deletion of the V_H portion and addition of a His_5 tag to the V_κ C terminus (Lindner *et al.*, 1992). Proteins were expressed in *E. coli* and purified from the cellular periplasm using immobilized metal affinity chromatography with Zn^{2+} as the cation on an iminodiacetic acid matrix and eluting with a linear imidazole gradient (10 mM to 200 mM). Unfolding is monitored through measurement of a large fluorescence increase upon denaturation, caused by dequenching of the single tryptophan residue of the domain located in the vicinity of the single disulfide bond in the folded state. An amount (5 μl) of protein of a concentration of 0·2 mM in phosphate buffered saline (PBS) was incubated overnight at 20°C with 500 μl guanidinium·HCl solutions in PBS. Denaturant concentrations were from 0 to 5 M in intervals of 0·1 M. Fluorescence ($\lambda_{ex} = 284$ nm, $\lambda_{em} = 360$ nm) was integrated over a 3 s interval. Raw values were corrected for buffer background fluorescence and analyzed for post-unfolding baselines, cooperativity and free energy of folding, using a non-linear least-squares fit procedure. The coefficient of correlation assuming a two-state model typically exceeds 0·999, errors are consistently below 10%.

of random, independent mutations (Allen *et al.*, 1987) to be compatible with every aspect of antibody function. Such an ensemble is expected to have two properties: (1) the average level of domain stability is marginal; and (2) the ensemble is at a state of equilibrium with respect to sequence changes affecting stability. These properties follow from the fact that while destabilizing random mutations are highly probable, they are selectively neutral as long as the overall domain stability does not fall below a certain threshold; conversely, stabilizing random mutations are highly improbable but there is no positive selection above a certain threshold (Kimura, 1991; Gregoret & Sauer, 1993). The near-independence of mutations distributed over a protein domain has been well established for a number of systems under investigation (Gregoret & Sauer, 1993; Sandberg & Terwilliger, 1993; Serrano *et al.*, 1993; Zhang *et al.*, 1992; Pantoliano *et al.*, 1989; Wells, 1990). Our canonical sequence approximation states that the most probable distribution of amino acids at a specific position is given by Boltzmann's law, thus we can calculate a statistical "free energy" (Sippl, 1988) from the frequencies of observation. This statistical free energy quantitates selection on that position: the deviation of the observed distribution from randomness.

In order to predict the effects of point mutations on stability, two approximations are introduced. The first approximation is that the database of sequences of immunoglobulins (Kabat *et al.*, 1991) represents such a canonical ensemble of sequences. For the purposes of this study we have averaged over all V_κ sequences of the database in order to avoid introducing further assumptions about sequence evolution in the immune system. We expect deviations to arise from significant sampling errors of the database (e.g. bias for the capacity to bind small haptens; bias towards a few intensively studied sequence families; errors from species-specific differences). The second approximation is that selection is only for domain stability. We expect systematic overestimation to arise from the other global factors, common to all domains, that influence the process of selection on some positions. However, no position of the domain can be freely mutated without any effect on stability and even if selective pressure is towards a different factor, severely disruptive mutations are never allowed. Since antigen binding imposes specific constraints on individual domains, the effects of the requirements for antigen binding should average out over the range of all observed sequences and the predic-

Table 1

Comparison of predictions for stabilizing point mutations with experiment

Domain	f_{WT}	f_{mut}	ΔG^{P}_{fold} (kJ mol^{-1})	Experiment	Prediction
WT			−13·5		
Ala15Leu	0·084	0·349	−19·2	+ +	+ +
Asn90Gln	0·040	0·873	−17·9	+ +	+ +
Phe32Tyr	0·054	0·734	−15·1	+	+ +
Leu106Ile	0·263	0·666	−15·0	+	+
Thr63Ser	0·132	0·798	−14·7	+	+ +
Met21Ile	0·190	0·662	−14·5	+	+ +
Lys24Arg	0·188	0·514	−12·8	0	+
Met21Leu	0·190	0·119	−12·2	−	−
Ala100Gly†	0·256	0·547	−13·6	0	+
Pro43Ser†	0·172	0·568	−12·8	0	+ +
Gln79Glu‡	0·234	0·721	−11·8	−	+

f_{WT} and f_{mut} are relative frequencies of wild-type (WT) and mutant (mut) amino acids at the respective positions of aligned V_κ sequences from the computer formatted immunoglobulin sequence database (Kabat *et al.*, 1992), kindly provided by Dr H. Perry. Predictions for effects on stability were estimated as $\Delta\Delta G_{fold} = -RT\ln(f_{mut}/f_{WT})$. No attempt was made to include further specific information in order to correct for database bias, unequal representation of amino acids in the genetic code or differences in codon transition probabilities (Betz *et al.*, 1993). ΔG^{P}_{fold} is the experimentally determined free energy of folding in phosphate buffered saline solution (pH 7·4), absence of denaturant, 20°C. Rating of effects was: 0 to 5% change in stability as compared to WT: 0, 6 to 20% change in stability: + or −, more than 20% change in stability: + +. The following mutations are considered to be correctly predicted: Ala15Leu, Asn90Gln, Phe32Tyr, Leu106Ile, Thr63Ser, Met21Ile and Met21Leu.

† Interacts with V_H domain.

‡ Interacts with C_κ domain.

tion should hold true also in the complementarity determining regions.

The V_κ domain of the anti-phosphorylcholine antibody McPC603 provides an ideal experimental model to test this hypothesis. We have previously expressed this domain recombinantly in *Escherichia coli* (Skerra & Plückthun, 1987; Lindner *et al.*, 1992), crystallized it (Glockshuber *et al.*, 1990) and determined the three-dimensional structure at high resolution (Steipe *et al.*, 1991). Equilibrium unfolding studies in guanidinium hydrochloride (GdmHCl) demonstrate that the isolated V_κ domain is only marginally stable with a ΔG of folding of −13·5 kJ mol^{-1}; i.e. assuming a two-state equilibrium (Pace, 1990) every 250th molecule is unfolded under native conditions. We have calculated the amino acid frequency distributions for all positions of all V_κ sequences in the Database (Kabat *et al.*, 1992). For this domain we have predicted ten stabilizing point mutations as replacements of non-consensus amino acids by the most common amino acid of the respective positions (Figure 1).

We have constructed these point mutations by site-directed mutagenesis and measured the domain stability (Figure 2). We find that our approach reliably predicts the effect of mutations on stability (Table 1). Six out of ten mutations were correctly predicted as stabilizing, three are found to have no significant effect and only one (Gln79Glu) is experimentally found to be significantly destabilizing. One additional mutation (Met21Leu) was predicted and confirmed as destabilizing. Random mutations would be expected to produce a large excess of destabilizing mutations. The effects on stability of the mutations are equally well predicted in frame-

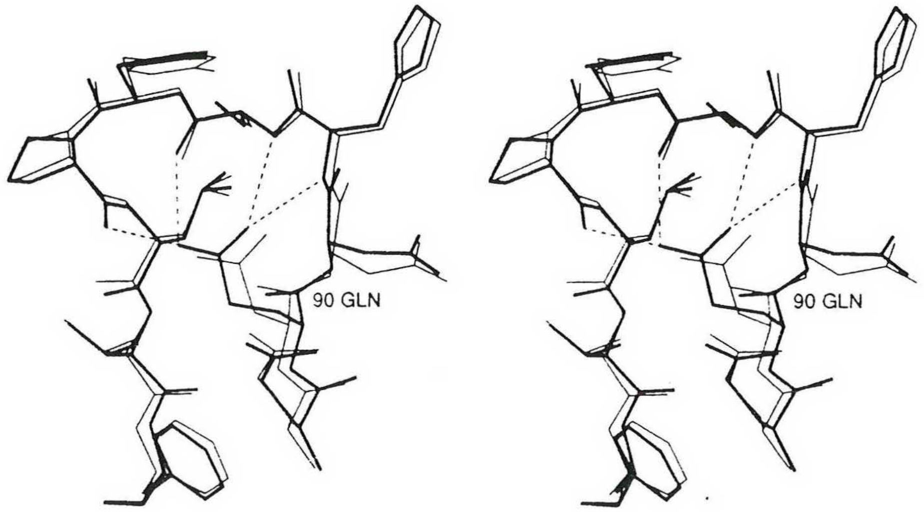

Figure 3. Crystal structures of WT (thin) and Asn90Gln mutant (bold) determined at 2.0 resolution. Residues 89 to 98 are shown, comprising the third complementarity determining region. Hydrogen bonds of Gln90 are shown with broken lines. Note that the position of the carboxamide group changes only slightly, the added methylene group is accommodated with a subtle conformational change propagated into the backbone structure of the loop. Energetic consequences can thus be expected to arise from multiple small changes in conformational energy as well as changes in free energies of four hydrogen bonds. The energetic consequences of such distributed small effects cannot be accurately predicted, even with an analysis of the three-dimensional structure. Methods of structure determination were as published (Steipe *et al.*, 1991). Diffraction data were analyzed by difference Fourier calculations ($F_o - F_c$). All atoms within a sphere of 6.0 around the modelled carboxamide group were initially excluded from the calculations.

work regions and complementarity-determining regions, except where constraints from inter-domain interactions play an important role. Including structural information about domain interactions, i.e. excluding residues Pro43, Gln79 and Ala100 from the prediction, improves the success rate of prediction to 7 out of 8. However, the strategy of simply choosing the most prevalent amino-acid in every position is found to carry a penalty of less than 2 kJ mol^{-1} at worst.

No obvious preference for a single mechanism of stabilization is observed. We have analyzed the following mutants crystallographically: Met21Ile, Met21Leu, Phe32Tyr, Pro43Ser, Thr63Ser, Asn90Gln and Ala100Gly. In all cases the structural effects are small, localized and distributed over a number of atoms. As an example, the structure of Asn90Gln is shown (Figure 3).

As the canonical sequence approximation is successful even if only sequence information is used for the prediction, it can be extended to other protein families, provided that sequence divergence is sufficiently small to make co-variation of residues or subdomains unlikely. We have obtained correct predictions from frequency ratios of 3 : 1 (Met21Leu) or even 2 : 1 (Met21Ile, Leu106Ile), suggesting that the minimum number of sequences needed for meaningful predictions can be correspondingly small.

We would like to acknowledge the participation of our students A.-B. Vogt, H.-E. Stöffel and C. Krasel in some of the experimental work presented here. Special thanks go to Drs R. Glockshuber, R. Jaenicke and A. Skerra for valuable discussions and to Drs R. Berendes, P. Reinemer and M. T. Stubbs for comments on the manuscript.

References

Allen, D., Cumano, A., Dildrop, R., Kocks, C., Rajewsky, K., Rajewsky, N., Roes, J., Sablitzky, F. & Siekevitz, M. (1987). Timing, genetic requirements and functional consequences of somatic hypermutation during B-cell development. *Immunol. Rev.* **96**, 5–22.

Arase, A., Yomo, T., Urabe, I., Hata, Y., Katsube, Y. & Okada, H. (1993). Stabilization of xylanase by random mutagenesis. *FEBS Letters*, **316**, 123–127.

Berek, C. & Milstein, C. (1988). The Dynamic nature of the antibody repertoire. *Immunol. Rev.* **105**, 5–26.

Berek, C. & Ziegner, M. (1993). The maturation of the immune response. *Immunol. Today*, **14**, 400–404.

Betz, G. A., Neuberger, M. S. & Milstein, C. (1993). Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes. *Immunol. Today*, **14**, 405–411.

Chen, L. H. & Baldwin, T. O. (1989). Random and site-directed mutagenesis of bacterial luciferase: investigation of the aldehyde binding site. *Biochemistry*, **28**, 2684–2689.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.

Davies, D. R. & Metzger, H. (1983). Structural basis of antibody function. *Annu. Rev. Immunol.* **1**, 87–117.

French, D. L., Laskov, R. & Scharff, M. D. (1989). The role of somatic hypermutation in the generation of antibody diversity. *Science*, **244**, 1152–1157.

Glockshuber, R., Steipe, B., Huber, R. & Plückthun, A. (1990). Crystallization and preliminary X-ray studies of the V_L domain of the antibody McPC603 produced in *Escherichia coli*. *J. Mol. Biol.* **213**, 613–615.

Gregoret, L. M. & Sauer, R. T. (1993). Additivity of mutant effects assessed by binomial mutagenesis. *Proc. Natl. Acad. Sci., U.S.A.* **90**, 4246–4250.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, 5th edit., U.S. Dept. Health and Human Services, Bethesda, MD.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1992). *Distribution Files of the Fifth Edition of Sequences of Proteins of Immunological Interest*.

Kimura, M. (1991). Recent development of neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 5969–5973.

Lindner, P., Guth, B., Wülfing, Ch., Krebber, K., Steipe, B., Müller, F. & Plückthun, A. (1992). Purification of native proteins from the cytoplasm and periplasm of *Escherichia coli* using IMAC and histidine tails: a comparison of proteins and protocols. *Methods*, **4**, 41–56.

Matthews, B. W. (1987). Genetic and structural analysis of the protein stability problem. *Biochemistry*, **26**, 6885–6888.

Pace, C. N. (1990). Measuring and increasing protein stability. *Trends Biochem. Tech.* **8**, 93–98.

Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardmann, K. R., Rollence, M. L. & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in free energy of unfolding. *Biochemistry*, **28**, 7205–7213.

Risse, B., Stempfer, G., Rudolph, R., Schumacher, G. & Jaenicke,R. (1992). Characterization of the stability effect of point mutations of pyruvate oxidase from *Lactobacillus plantarum*: protection of the native state by modulating coenzyme binding and subunit interaction. *Protein Sci.* **1**, 1710–1718.

Rollence, M. L., Filpula, D., Pantoliano, M. W. & Bryan, P. N. (1988). Engineering thermostability in subtilisin BPN' by *in vitro* mutagenesis. *Crit. Rev. Biotechnol.* **8**, 217–224.

Sandberg, W. S. & Terwilliger, T. C. (1993). Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 8367–8371.

Serrano, L., Day, A. G. & Fersht, A. R. (1993). Step-wise mutation of barnase to binase. *J. Mol. Biol.* **233**, 305–312.

Shi, Y.-Y., Mark, A. E., Wang, C.-X., Huang, F., Berendsen, H. J. C. & van Gunsteren, W. F. (1993). Can the stability of protein mutants be predicted by free energy calculations? *Protein Eng.* **6**, 289–295.

Sippl, M. J. (1988). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.

Skerra, A. & Plückthun, A. (1988). Assembly of a functional Fv fragment in *Escherichia coli*. *Science*, **240**, 1038–1041.

Steipe, B., Plückthun, A. & Huber, R. (1991). Refined crystal structure of a recombinant immunoglobulin domain and a complementarity-determining region-1 grafted mutant. *J. Mol. Biol.* **225**, 739–753.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature (London)*, **302**, 575–581.

Turner, S. L. Ford, G. C., Mountain, A. & Moir, A. (1992). Selection of a thermostable variant of chloramphenicol acetyltransferase (Cat-86). *Protein Eng.* **5**, 535–541.

Wells, J. A. (1990). Additivity of mutational effects in proteins. *Biochemistry*, **29**, 8509–8517.

Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implication for antibody complementarity. *J. Exp. Med.* **132**, 211–250.

Zhang, X.-J., Baase, W. A. & Matthews, B. W. (1992). Multiple alanine replacements within α-helix 126–134 of T4 lysozyme have independent, additive effects on both structure and stability. *Protein Sci.* **1**, 761–776.

Edited by A. R. Fersht

(Received 10 December 1993; accepted 18 April 1994)