Rainier, S., Herrera, J. M., & McCormick, A. M. (1983) *Arch. Biochem. Biophys. 225*, 818–825.

Ross, A. C. (1982) *J. Biol. Chem. 257*, 2453–2459.

Saari, J. C., & Bredberg, L. (1982) *Biochim. Biophys. Acta 716*, 266–272.

Siegenthaler, G., Saurat, J. H., & Ponec, M. (1990) *Biochem. J. 268*, 371–378.

Strickland, S. S., & Sawey, M. J. (1980) *Dev. Biol. 78*, 76–85.

Srivastava, D. K., & Bernhard, S. A. (1986) *Science 234*, 1081–1086.

Williams, J. B., & Napoli, J. L. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4658–4662.

Williams, J. B., & Napoli, J. L. (1987) *Biochem. Pharmacol. 36*, 1386–1388.

Williams, J. B., Pramanik, B. C., & Napoli, J. L. (1984) *J. Lipid Res. 25*, 638–645.

Williams, J. B., Shields, C. O., Brettel, L. M., & Napoli, J. L. (1987) *Anal. Biochem. 160*, 267–274.

Yost, R. W., Harrison, E. H., & Ross, A. C. (1988) *J. Biol. Chem. 263*, 18693–18701.

Approaches to Predicting Effects of Single Amino Acid Substitutions on the Function of a Protein[†]

Hal B. Zabin, Martin P. Horvath, and Thomas C. Terwilliger*

Department of Biochemistry and Molecular Biology, The University of Chicago, 920 East 58th Street, Chicago, Illinois 60637

Received September 18, 1990; Revised Manuscript Received January 11, 1991

ABSTRACT: The relative activities of 313 mutants of the gene V protein of bacteriophage f1, assayed in vivo, have been used to evaluate two approaches to predicting the effects of single amino acid substitutions on the function of a protein. First, we tested methods that only depend on the properties of the wild-type and substituting amino acids. None of the properties or measures of the functional equivalence of amino acids we tested, including the frequency of exchange of amino acids among homologous proteins as well as changes in side-chain size, hydrophobicity, and charge, were found to be more than weakly correlated with the activities of mutants. The principal reason for this poor correlation was found to be that the effect of a particular substitution varies considerably from site to site. We then tested an approach using the activities of several mutants with substitutions at a site to predict the activity of another mutant, and we find that this is a relatively good indicator of whether the other mutant at that site will be functional. A predictive scheme was developed that combines the weak information from the models depending on the properties of the wild-type and substituting amino acids with the stronger information from the tolerance of a site to substitution. Although this scheme requires no knowledge of the structure of a mutant protein, it is useful in predicting the activities of mutants.

It is not possible, at present, to predict with certainty whether a protein differing from a natural protein by an amino acid substitution is likely to be functional, even if detailed structural information on the wild-type protein is available. Nevertheless, it is widely thought that chemically conservative amino acid substitutions, such as one in which a leucine is replaced with an isoleucine, will generally not affect the function of a protein, while other less conservative replacements are more likely to affect protein function. Although this idea has never been systematically tested, it is supported by the observation that amino acid side chains that are chemically similar to one another are much more likely to be found at corresponding positions in structurally or functionally related proteins than dissimilar ones (Dayhoff et al., 1978). For example, leucine and isoleucine are found to substitute for one another more than 10 times as frequently as valine and arginine.

We have set out to evaluate two simple approaches for predicting whether a protein with an amino acid substitution is likely to be functional. One approach makes the assumption that the activity of a protein with an amino acid exchange relative to the wild-type protein is related to some measure of the functional equivalence of the wild-type and substituting amino acid side chains. An approach of this type has been implicitly used in many schemes to align amino acid sequences (Gribskov et al., 1987; Pearson & Lipman, 1988), to predict structures and functions of proteins (Chou & Fasman, 1978; Garnier et al., 1978; Eisenberg et al., 1982; Kidera et al., 1985), and to design new proteins with structures similar to natural ones (DeGrado, 1988). It seems certain that a mutant protein which differs from the wild type by an exchange that occurs very frequently among homologous proteins is more likely to be functional than one which differs by an infrequently observed exchange. It is not clear, however, whether this difference is sufficiently large to be of any use in predicting whether a mutant protein will be functional. A second approach incorporates information on the tolerance of each site in a protein to substitution. For a number of proteins, some sites are much more tolerant of amino acid substitutions than others (Loeb et al., 1989; Bowie & Sauer, 1989; Bowie et al., 1990; Kleina & Miller, 1990). In this second type of model,

[†] We acknowledge with gratitude support from an NSF predoctoral fellowship to M.P.H., from a Lucille P. Markey Foundation predoctoral fellowship to H.B.Z., and from NIH Grant GM 38714 and NSF Presidential Young Investigator Award DMB 8657754 to T.C.T. and generous gifts of matching funds from the Cancer Research Foundation, from the Duchossis Foundation, from Bristol-Myers, and from Merck, Sharp, & Dohme Research Laboratories.

* To whom correspondence should be addressed.

the tolerance of a site to substitution, evaluated from several substitutions at the site, is used to predict whether a new mutation at that site would lead to a functional protein.

To evaluate models for the effects of amino acid substitutions on the function of a protein, we have applied an in vivo activity assay to mutants of the gene V protein of bacteriophage f1. The gene V protein of bacteriophage f1 is a single-stranded nucleic acid binding protein (Coleman & Oakley, 1980; Kowalczykowski et al., 1981; Baas, 1985). The X-ray crystal structure of the gene V protein has been determined; it is a dimer with two identical subunits, each containing 87 amino acid residues (Brayer & McPherson, 1983). Some features of the solution structure of the f1 gene V protein and of the related gene V protein from bacteriophage IKe are now apparent from 2-D NMR analysis as well and suggest that some revision of the crystal structure may be needed (de Jong et al., 1989a; van Duynhoven et al., 1990). The gene V protein is required for propagation of bacteriophage f1. The protein binds cooperatively and stoichiometrically to the single-stranded intermediate in phage f1 DNA replication, coating the DNA, inhibiting complementary strand synthesis, and allowing packaging of the single-stranded DNA into phage particles (Salstrom & Pratt, 1971; Alberts et al., 1972; Pretorius et al., 1975; Cavalieri et al., 1976; Baas, 1985). The gene V protein also binds specifically to a translational operator sequence at the 5′ end of the bacteriophage f1 gene II message, inhibiting its translation (Model et al., 1982; Yen & Webster, 1983; Fulford & Model, 1988; Michel & Zinder, 1989; Zaman et al., 1990). Some of the residues in the bacteriophage f1 and IKe gene V proteins responsible for binding to single-stranded nucleic acids have been tentatively identified (Anderson et al., 1975; Bayne & Rasched, 1983; Brayer & McPherson, 1983; King & Coleman, 1987; Dick et al., 1988; de Jong et al., 1989b) and appear to lie on one surface of the protein. In a dimer of gene V protein, these surfaces are thought to complex antiparallel single strands of DNA (Brayer & McPherson, 1983).

We recently constructed a set of codon-specific mixtures of gene V mutants in a plasmid-based system (Zabin & Terwilliger, 1991) and a small number of individual mutants by directed mutagenesis (Sandberg & Terwilliger, 1991). We have isolated a collection of single amino acid mutants from the codon-specific libraries of mutants, tested their activities of both sets of mutants in vivo, and used these measured activities to evaluate schemes for the prediction of the function of gene V protein mutants.

Materials and Methods

Recombinant DNA. Wild-type gene V protein and mutant proteins were expressed from derivatives of plasmid pTT18 (Terwilliger, 1988b) by induction with IPTG.[1] Plasmid pTT18 contains a gene encoding the wild-type bacteriophage f1 gene V protein but differing in DNA sequence from the wild-type gene V at 45 nucleotide positions so as to incorporate 9 unique restriction sites (Terwilliger, 1988b). The gene V in plasmid pTT18 is under the control of the strong *tac* promoter (Amann et al., 1983). The construction of codon-specific mixtures of gene V mutants has been recently described (Zabin & Terwilliger, 1991). Each of these mixtures contains plasmids with gene V sequences that differ from the gene in pTT18 at a single codon. Plasmids from single codon specific mixtures were used to transform the *lac* I^Q *Escherichia coli* strain K561 (Davis et al., 1985). Single colonies were isolated, and the sequence of the gene V in the region of interest was determined. In most cases, the sequence of the remainder of the gene was also obtained. From the 639 isolates sequenced we found mutants with 332 different DNA sequences encoding single amino acid substitutions. Previously, we had constructed 29 mutant gene V plasmids encoding single amino acid substitutions by directed mutagenesis (Sandberg & Terwilliger, 1991; Terwilliger, unpublished work). The set of 361 plasmids isolated with the two approaches encodes a total of 317 different protein sequences. Except as noted, all mutant genes encoding the same protein sequence are treated as a single mutant.

In Vivo Function of Mutant Gene V Proteins. The wild-type gene V protein, when expressed at a very high level in *E. coli*, strongly inhibits the growth of the cells (Terwilliger, 1988a). An assay for function of mutant proteins was developed based on this inhibition of growth. *E. coli* cells, strain K561, were transformed with derivatives of plasmid pTT18 encoding a single amino acid substitution mutant of gene V, were plated on 2YT plates (Miller, 1972) containing ampicillin (50 μg/mL), and were grown overnight at 37 °C. Isolated colonies were picked and transferred into 100 μL of 2YT media. Three microliters of the suspensions, containing about 1000 viable cells, were spotted onto 2YT plates containing ampicillin (150 μg/mL) with and without IPTG (1 mM). The higher concentration of ampicillin (150 μg/mL) used in these plates was included to ensure that cells containing very few or no copies of the pTT18-derivative plasmids would not form colonies. The plates were incubated at 37 °C for 8–20 h, after which the growth on each of the two plates was observed and the differences were noted. Repeated observations on the same sets of plates indicated that the differences between the appearances of the plates did not change appreciably over this range of observation times. We defined three categories of activity of the mutant proteins based on the appearance of lawns of *E. coli* that expressed them: an activity of 2 corresponded to essentially no growth when gene V protein was expressed, as for the fully active wild-type protein; an activity of 1 corresponded to reduced growth in the presence of the gene V protein, as for a partially active protein; and an activity of 0 corresponded to equal growth with and without expression of the gene V protein, as for an inactive mutant. At least 2, and an average of 6, measurements on separate transformants were made for each mutant, and the activities were averaged.

Elsewhere, we tested whether 11 gene V protein mutants that were nonfunctional or partially functional in our assays were expressed at high levels (Terwilliger et al., manuscript in preparation). We found that the mutant proteins all were present in *E. coli* expressing them at levels at least as high as the wild-type protein, so the loss of activity of these mutants was not simply due to low expression levels or rapid degradation. We also observed that missense gene V mutants that do not inhibit *E. coli* growth generally do not support growth of bacteriophage f1, suggesting that our assay monitors properties closely related to the biological function of the protein. These results suggested that our activity assay is a reasonable measure of whether a mutant gene V protein, once expressed, will or will not form a stable protein that can perform certain functions of the wild-type gene V protein.

Results and Discussion

Functional Activities of Mutant Gene V Proteins. To determine which of the proteins encoded by the mutant gene V plasmids were functional, we developed a rapid assay that makes use of the relatively slow growth of *E. coli* expressing wild-type gene V protein compared to those not producing this

[1] Abbreviations: IPTG, isopropyl β-D-thiogalactopyranoside; w.t., wild type.

protein (see Materials and Methods). A principal function of the gene V protein in bacteriophage f1 DNA replication is to prevent complementary strand synthesis, and the gene V protein can bind nonspecifically to single-stranded DNA or RNA (Salstrom & Pratt, 1971; Alberts et al., 1972; Pretorius et al., 1975; Cavalieri et al., 1976; Baas, 1985). The growth inhibition caused by expression of the gene V protein in *E. coli* is therefore probably due to interference with host DNA replication and RNA translation. Mutant proteins that, when expressed, strongly inhibited *E. coli* growth were considered active (activity = 2). Those that did not visibly inhibit *E. coli* growth were considered inactive (activity = 0). Those that inhibited *E. coli* growth, but less than did the wild-type gene V protein, were considered partially active.

Overall, 35% of the 317 different single amino acid substitution mutants studied here were highly active in this assay (activity ≥ 1.5), 26% were partially active (0.5 ≤ activity < 1.5), and 38% were essentially inactive (activity < 0.5). We estimated the uncertainties in our activity measurements by calculating the standard deviation of measurements for each mutant (σ_A) and by comparing the values obtained for the same mutants on different occasions (σ_B). The overall reproducibility of our measurements (σ_A) was 0.3 unit on our scale of 0–2. The reproducibility of measurements carried out in different experiments (σ_B) was slightly higher, 0.4 unit, suggesting that there may be a small systematic variation in the assay from day to day. Finally, the standard deviation of activity measurements obtained for proteins with the same amino acid sequence, but which were encoded by different codons (σ_C), was 0.3 unit, indicating that codon usage had little effect on our assay. Four of the 317 unique mutants we studied gave variable results (σ_A > 0.7 unit) and were not included in the analyses that follow. Table I lists the activities of the 313 unique single amino acid substitution mutants for which we obtained consistent results. A substitution of Ile 2 with arginine, for example, leads to a completely inactive protein (activity = 0.0), while a substitution of the same site with proline leads to a partially functional protein (activity = 0.8).

Comparison of Activities of Mutant Proteins with the Frequency That the Corresponding Amino Acid Exchange Occurs among Homologous Proteins. We used our measurements of the activities of mutants of the gene V protein to evaluate several approaches for predicting the effects of amino acid substitutions on protein function. First, we examined the utility of an approach that considers only the identities of the wild-type and substituting amino acids, and not the context of the mutation. We anticipated, for example, that proteins differing from the wild type by amino acid exchanges that occur frequently between homologous proteins would be more active, as a whole, than those differing by exchanges that are infrequent. To test this, we compared the activities of gene V missense mutants with the frequency of occurrence of the amino acid exchange that differentiates them from the wild-type protein (Figure 1). We use the number of occurrences of an amino acid exchange among the 1572 exchanges examined by Dayhoff (1978) as a measure of the frequency of exchange of those amino acids.

As expected, those exchanges that are found with high frequency among homologous proteins do lead to a functional protein more often than those that are infrequently observed (Figure 1). This effect, however, is remarkably small compared to the variation in activities of mutants within any particular frequency range. For example, of the 29 mutants with an exchange that was not found at all in 1572 exchanges, 11 had activities of less than 0.5, 10 had activities from 0.5 to 1.5, and 8 had activities of 1.5 or greater. For the 18 mutants with an exchange that occurs more than 50 times in 1572 exchanges, 3 were inactive, 1 was partially active, and the remaining 14 were highly active. Table IIA lists the 8 examples of mutants that were not found at all among the 1572 exchanges examined by Dayhoff (1978) yet are fully active. Evidently these substitutions can lead to fully functional proteins in at least some cases. Table IIB lists the mutant proteins that differ from the wild type by an exchange that occurs more than 50 times in 1572 exchanges, yet which are inactive. There are 3 mutants in this class, each involving an exchange between glycine and alanine or serine and alanine. Although such exchanges are common, in these cases they lead to nonfunctional proteins. Presumably the wild-type residues at these sites fulfill specific functions that cannot be carried out by the substituting residues.

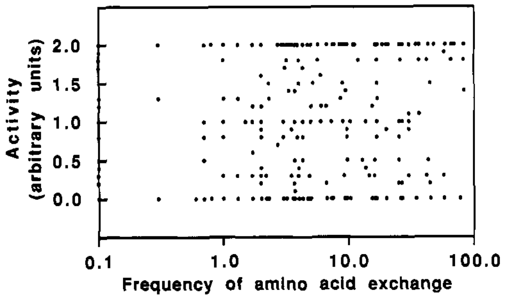

FIGURE 1: Activities of gene V protein mutants with amino acid substitutions compared to the frequency of occurrence of the corresponding substitution between homologous proteins (Dayhoff, 1978). Frequencies of exchange are shown on a logarithmic scale, and exchanges with frequencies of zero [i.e., that were not observed in the analysis of Dayhoff (1978)] were given values of 0.1. Some points superimpose in this figure. Activity measurements are based on the inhibition of *E. coli* growth upon expression of mutant gene V proteins (see Materials and Methods).

Overall, we find that even though some exchanges are many times as frequent as others, commonly found substitutions lead to active proteins only slightly more often than infrequent substitutions. The frequency of exchange among homologous proteins as an estimate of the functional equivalence of pairs of amino acids therefore has limited utility by itself in predicting whether a mutant protein will be active.

Predictive Values of Various Measures of the Functional Similarities of Amino Acid Side Chains. Although the frequency of exchange of two amino acids among homologous proteins is not very well correlated with the activity of a particular protein with that exchange, it was possible that some other properties of the amino acids would be useful in predicting the activities of mutant proteins. Properties we examined included whether the substituting amino acid had very different allowed backbone conformations than the wild-type amino acid (proline and glycine substitutions), whether amino acids with β-branched side chains were added or removed, and the changes in side-chain hydrophobicity, volume, and charge. We constructed simple functions ($g_{k,n}$; see Appendix and legend to Table III) based on each property and used them in a model to predict the activities of the mutants in our study (eq A5). We then determined whether inclusion of each of these functions improved the correspondence between the observed activities of mutants and those calculated from the model. As discussed in the Appendix, the relative improvement in the mean square deviation between observed and calculated activities, ν (eq A12), due to including some function $b_k g_{k,n}$ in the model, is an estimate of the relative contribution of that function to the observed mean square variation in the activities of mutants. For example, the frequency of exchange of amino acids among homologous proteins is, as discussed above, weakly correlated with the activities of mutants with corresponding substitutions. Table III shows that when all sites are considered, the agreement between observed activities of mutants and activities calculated from a model given by a constant term and a term proportional to the frequency of exchange is best when the constant of proportionality (b_k) is 0.012. The value of the coefficient b_k corresponds to predicting that a mutant with a frequency of exchange of 50 would have an activity about 0.6 unit higher than one with a frequency of exchange of 1. Including the frequency of exchange in the prediction of activities of mutants reduces the mean square deviation between observed and predicted activities from 0.68 unit^2 to 0.64 unit^2. After subtraction of the estimated mean square deviation due to errors in measurement (0.16 unit^2; see Appendix), this corresponds to a relative improvement in mean square deviation between observed and predicted activities, ν, of 7.7% (Table III).

We carried out this analysis for all sites as a group as well as separately for substitutions at residues that are buried in the interior of the protein and for those that are exposed to the solvent according to the crystal structure of the gene V protein (Brayer & McPherson, 1983). The results were similar in the three cases (Table III). Of the nine properties of this type that we tested, the two that had the most predictive value were the frequency of exchange of the wild-type and substituting amino acids among homologous proteins, discussed above, and whether the substituting residue was a glycine or proline. There are 40 mutants with glycine or proline substitutions that we characterized, and the average activity of these mutants is 0.40, compared to an average of 0.96 for all 313 mutants in this analysis. Consequently, the best estimate of the activity of a mutant with a substitution to a glycine or proline is about 0.6 unit lower than the average activity for all mutants (b_k = -0.60, Table III). On the basis of the improvement in the correspondence between observed and predicted activities when this property is included in the model, we estimated (Table III) that substitutions of glycine or proline contribute about 8% of the mean square variation in the activities of mutants.

The magnitude of the change in hydrophobicity was somewhat correlated with the activities of mutants, particularly those at buried sites (Table III). This factor contributes about 4% of the mean square variation in activities overall and about 13% of the mean square variation in activities at buried sites. Other properties, including introducing or removing a β-branched amino acid and changing the side-chain volume or charge, had little predictive value for gene V protein mutants.

Even when all three of the properties of the amino acids that we found most useful were combined together, the predictive value of these properties was very low (Table III, set 1). We estimate that the effects measured by the frequency of exchange of amino acid side chains among homologous proteins, effects of introducing glycine or proline residues, and effects of changing the hydrophobicity of the residue, all together, contribute only about 17% of the mean square variation in the activities of mutants.

Dependence of Effect of a Substitution on Context. Given that none of the properties of the amino acids that we have examined are good predictors of whether two amino acids can effectively substitute for each other in a particular protein, it would be useful to know if any measures of the functional equivalence of two amino acids that would be any more useful exist. It seemed possible that the location, within a protein, of the amino acid that is substituted might have a dominant effect on the activity of the mutant protein. If the context of a mutation does have such a large effect, then no measure of the similarities of the amino acids would be very useful in predicting the activities of mutant proteins. We tested this possibility by examining the activities of mutants with the same substitution at different sites within the gene V protein.

In our sample of 313 mutant proteins, there are 84 examples of substitutions that we observed at more than one site, and we found that their activities did vary considerably from site to site. Table IV lists the activities of a subset of these, the 22 substitutions we obtained at three or more sites. Substitution of leucine with alanine, for example, leads to an active protein if the exchange is made at residue 28, but inactive proteins if the exchange is made at residue 37 or 76. In contrast, all three substitutions of tyrosine with phenylalanine led to fully active proteins, and all three substitutions of glycine with serine residues led to inactive proteins. Overall, the standard deviation in the activities measured for the same 84 substitutions at different sites was 0.69 unit. This is quite remarkable because the standard deviation of the activities of all 313 mutants we studied was 0.82 unit. The location, or context, of a mutation, therefore, has nearly the same influence on the function of the mutant as all effects combined.

We can use the standard deviation in activity measurements for the same substitutions at different sites to estimate the relative contribution of all properties associated with the nature of the amino acid substitution to the mean square variation in activities of mutants. The variance of activities of the same amino acid substitution at different sites is a measure of the best possible value of the variance that could be obtained after correction for the nature of the mutation using eq A5. This leads to the conclusion, according to eq A12 with s_0 = 0.82, s = 0.69, and σ_{obs} = 0.40, that about 38% of the mean square variation on the activities is due to the nature of the amino acid substitution. Our best model (above) accounted for about 17% of the mean square variation in activities, so somewhat better models could be developed, but none would account for more than about a third of the total variation in activities. Even if much more information on the effects of amino acid substitutions on function were available from many proteins, it would therefore still not be possible to devise a procedure to predict accurately the effects of arbitrary amino acid substitutions on the function of a protein without information on the context of the substitution. Although the average effect of a particular substitution could be determined accurately, a great deal of variation would be found at different sites.

Tolerance to Substitution at Surface and Buried Sites. The strong dependence of the effect of an amino acid substitution on location could be due to several causes. One possibility is that many sites in the interior of the protein are less tolerant to substitution than those that are on the surface (Bowie et al., 1990; Reidhaar-Olson & Sauer, 1990). Table V lists the average activities of mutants with substitutions at buried and surface sites in the gene V protein, classified according to the polarity of the wild-type residues and of the substituting residue. At buried sites, substitutions from apolar to polar residues generally led to inactive proteins, as expected, while substitutions from apolar to apolar residues sometimes led to functional proteins and sometimes did not (Table V). For surface sites, substitutions that changed the polarity of the residue were slightly less active than those that maintained polarity. Overall, interior sites in the gene V protein are somewhat less tolerant than surface sites in that polar residues at interior sites generally lead to nonfunctional proteins. Nevertheless, some surface sites are tolerant of substitution

Table I: Activities of Gene V Mutants[a]

residue	exposure (%)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ile 2	55	–	0.0	–	–	–	–	–	–	–	–	–	–	–	–	0.8	–	–	–	–	–
Lys 3	30	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.8	–	–	–	–
Val 4	11	–	0.0	–	–	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	–	–
Glu 5	18	–	–	–	1.8	–	–	–	2.0	–	–	–	–	–	–	0.0	–	–	–	–	0.0
Ile 6	2	–	0.0	–	–	–	0.0	0.3	–	0.0	–	0.2	–	–	–	0.0	0.0	0.5	–	–	2.0
Lys 7	27	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.9	–	–	–
Pro 8	8	–	2.0	–	–	–	–	–	–	2.0	–	–	–	–	2.0	–	–	–	2.0	–	–
Ser 9	9	–	–	–	0.5	–	–	–	–	–	–	–	1.8	–	–	0.0	–	2.0	–	–	0.5
Gln 10	49	0.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0	–	–
Ala 11	67	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	–	–	–	0.2	–	–
Gln 12	69	–	–	–	–	–	–	0.3	2.0	–	2.0	2.0	–	2.0	–	0.0	2.0	–	–	–	–
Phe 13	4	1.2	–	–	–	2.0	–	–	–	–	2.0	2.0	–	–	–	–	0.4	2.0	–	–	2.0
Thr 14	40	–	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	–	–	–	–	–
Thr 15	39	–	1.6	–	–	–	–	–	0.0	–	–	–	2.0	–	–	–	–	–	–	–	–
Arg 16	75	–	–	–	0.8	0.0	–	–	–	1.0	–	–	–	–	–	–	–	–	–	–	–
Ser 17	67	–	–	–	–	–	–	–	1.5	–	–	–	–	–	–	–	–	2.0	–	–	–
Gly 18	34	–	–	–	0.0	0.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.0
Val 19	59	–	0.4	0.0	–	–	–	–	0.0	0.0	–	–	–	–	–	0.0	–	2.0	–	–	–
Ser 20	52	–	–	–	–	–	–	–	–	–	0.2	0.0	–	–	–	–	–	–	–	–	–
Arg 21	71	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	–	–	–	–	–	–
Gln 22	83	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.0	–	–	2.0
Gly 23	22	–	–	–	–	–	–	–	–	–	0.5	–	–	–	–	–	0.0	–	–	–	–
Lys 24	76	–	–	–	–	–	–	–	–	–	–	0.8	–	–	–	0.0	–	–	1.0	–	2.0
Pro 25	70	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	–	2.0	–	–	–	–
Tyr 26	89	–	1.3	–	–	–	–	–	–	–	–	1.3	–	–	1.8	–	–	–	–	–	–
Ser 27	46	–	–	–	–	–	–	2.0	–	–	–	–	–	–	–	–	–	2.0	–	–	–
Leu 28	35	2.0	–	–	–	–	–	–	–	1.5	–	–	–	–	–	0.0	–	–	–	–	–
Asn 29	54	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	0.7	0.5	1.3	–	0.0	0.3
Glu 30	18	–	0.0	–	1.4	–	–	–	–	–	–	–	–	0.0	–	0.0	1.3	–	–	–	1.4
Gln 31	11	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.0	–	–	–
Leu 32	11	–	2.0	–	–	–	–	–	–	–	0.5	–	–	–	–	2.0	0.0	–	2.0	–	–
Cys 33	2	–	–	–	–	–	–	–	–	–	–	0.0	–	0.4	–	0.0	–	0.0	–	0.0	–
Tyr 34	45	–	–	–	–	–	–	–	–	–	–	–	–	–	2.0	0.2	0.3	–	–	–	–
Val 35	1	1.8	–	–	–	2.0	–	–	0.9	–	1.8	0.3	–	2.0	0.3	–	0.0	–	1.0	–	–
Asp 36	12	1.2	–	–	–	–	–	2.0	0.3	2.0	–	0.3	–	–	–	0.3	–	–	–	–	–
Leu 37	8	0.0	0.3	–	–	–	–	1.0	0.0	–	0.8	–	–	–	–	0.0	0.0	–	–	–	0.9
Gly 38	8	–	–	–	2.0	–	–	–	–	1.3	–	–	1.0	–	–	–	0.2	0.0	–	–	–
Asn 39	51	–	–	–	–	–	–	–	–	–	–	0.5	–	1.2	–	–	2.0	2.0	–	–	–
Glu 40	78	–	0.0	1.4	–	–	–	–	–	1.5	2.0	1.0	–	1.0	–	–	–	1.7	–	–	–
Tyr 41	79	2.0	2.0	–	–	–	2.0	–	1.9	0.0	–	–	–	1.0	1.8	–	–	–	–	–	–
Pro 42	42	–	–	–	–	–	–	–	–	–	–	1.0	–	–	–	–	–	1.0	–	–	–
Val 43	43	–	–	–	–	–	–	–	–	–	–	1.0	–	–	–	–	1.7	–	–	–	–
Leu 44	44	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0	0.0	1.5	–	–	–
Val 45	10	–	–	–	–	–	–	0.2	–	–	2.0	1.1	–	–	0.0	–	0.0	–	–	–	–
Lys 46	26	–	–	–	–	–	0.3	–	–	–	–	–	–	–	–	–	1.0	–	–	–	–
Ile 47	5	1.2	–	–	–	1.2	–	0.9	0.0	–	–	2.0	–	1.8	2.0	0.0	0.8	–	0.0	–	2.0
Thr 48	29	–	–	–	–	–	–	–	0.0	–	–	–	–	–	–	0.0	–	–	–	–	2.0
Leu 49	1	–	1.0	–	–	–	0.0	–	–	0.0	–	–	–	–	–	0.0	0.0	–	–	–	–
Asp 50	16	–	1.7	–	–	–	–	–	0.5	–	–	–	–	–	1.3	–	–	–	–	–	0.0
Glu 51	23	2.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Gly 52	74	1.9	–	–	–	–	2.0	2.0	–	–	–	–	1.6	–	–	–	–	1.2	–	–	–
Gln 53	41	2.0	1.8	–	–	2.0	–	1.8	–	–	–	–	–	–	0.5	–	2.0	–	–	–	2.0
Pro 54	65	–	–	–	–	–	–	–	–	–	1.0	–	2.0	–	–	–	2.0	–	–	–	–
Ala 55	38	–	–	–	–	–	–	–	–	–	–	0.0	–	–	–	0.0	–	–	–	–	–
Tyr 56	14	–	–	–	–	0.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ala 57	43	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.0	–	–	–	–	1.8
Pro 58	45	–	0.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Gly 59	48	0.0	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0	0.0	–	1.0	–	–
Leu 60	45	–	–	–	–	–	–	–	–	–	–	–	2.0	–	–	–	2.0	–	–	–	2.0
Tyr 61	17	–	0.0	–	–	0.0	–	–	–	–	–	–	–	–	–	0.5	0.0	–	0.0	–	–
Thr 62	3	–	–	–	–	–	2.0	–	–	–	–	–	–	–	–	1.0	2.0	–	–	–	–
Val 63	19	–	–	–	–	–	–	–	–	0.8	2.0	–	–	–	–	–	–	–	–	–	–
His 64	18	–	–	–	–	–	–	–	–	–	–	1.8	–	1.8	1.0	–	–	–	–	–	–
Leu 65	57	–	2.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.3	–	–
Ser 66	30	–	–	–	–	–	–	–	–	–	–	–	–	0.9	–	–	–	–	–	–	–
Ser 67	2	0.0	0.4	–	–	–	0.0	–	–	–	0.0	–	–	–	–	0.2	–	2.0	–	–	0.8
Phe 68	9	–	0.5	–	–	–	–	–	–	0.3	1.0	–	–	–	–	0.8	–	0.3	–	0.0	0.8
Lys 69	33	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.5	1.5	–	–	–	–
Val 70	30	1.1	–	–	–	1.4	–	–	–	–	–	–	–	–	–	1.0	–	–	–	–	–
Gly 71	54	–	–	–	–	1.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.8
Gln 72	61	–	–	–	–	–	–	–	–	–	–	2.0	–	–	–	–	–	–	–	–	–
Phe 73	52	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.0	–	–
Gly 74	15	–	–	–	–	1.0	1.0	–	–	–	–	–	1.2	–	–	0.3	–	0.0	–	–	–
Ser 75	24	–	–	–	–	–	–	–	–	–	–	1.8	–	–	–	–	–	–	–	–	–
Leu 76	31	0.0	–	–	–	–	–	–	–	–	–	–	–	–	0.0	–	0.0	–	–	–	–
Met 77	4	–	–	2.0	–	–	–	–	–	–	2.0	2.0	–	–	–	–	–	–	–	–	2.0

Table I (Continued)

residue	exposure (%)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ile 78	47	–	–	–	–	–	–	–	–	–	–	–	–	–	1.5	–	–	–	–	–	–
Asp 79	50	0.0	–	–	–	–	–	–	–	1.0	–	0.2	–	–	–	1.0	–	–	–	–	–
Arg 80	20	–	–	–	–	–	–	–	–	0.0	–	0.6	–	–	0.0	–	1.0	–	–	–	0.0
Leu 81	16	–	–	–	–	–	–	–	–	–	1.0	–	–	–	–	–	–	–	–	–	–
Arg 82	25	0.0	–	–	–	–	–	–	–	–	–	–	–	–	0.0	0.0	–	0.0	–	–	–
Leu 83	7	–	–	0.1	–	–	–	–	–	0.3	–	–	–	–	–	–	–	–	–	0.0	–
Val 84	13	2.0	–	–	2.0	–	–	2.0	–	–	–	–	–	2.0	–	–	–	–	–	–	–
Pro 85	49	–	–	–	–	–	–	–	0.0	–	–	0.0	–	–	0.0	–	1.0	–	–	–	–
Ala 86	27	–	2.0	2.0	0.0	–	–	–	0.3	–	–	–	–	2.0	–	–	–	2.0	–	–	2.0
Lys 87	64	2.0	2.0	–	2.0	–	–	–	–	–	2.0	–	–	–	–	–	–	2.0	–	–	–

[a] For each substitution characterized, the activity of the mutant protein based on our in vivo assay, on a scale of 0 (inactive) to 2 (fully active), is indicated. The wild-type residue at each site and the percentage of the surface area of the residue in the crystal structure of the wild-type protein (Brayer & McPherson, 1983) accessible to a solvent water molecule (Kabsch & Sander, 1983; Richards, 1977) are also listed. We note that there is some uncertainty in the crystal structure (de Jong et al., 1989a; van Duynhoven et al., 1990). A dash (–) indicates that the activity for that particular mutant was not determined.

Table II: Substitutions That Occur Infrequently yet Lead to Fully Active Proteins or That Occur Frequently and Lead to Inactive Proteins[a]

(A) Infrequently Observed Mutations That Lead to Active Gene V Proteins		
Pro 8 → Trp	Tyr 41 → Gln	His 64 → Met
Phe 13 → Cys	Asp 50 → Arg	Met 77 → Asn
Tyr 41 → Gly	Gln 53 → Cys	
(B) Frequently Observed Mutations That Lead to Inactive Gene V Proteins		
Gly 59 → Ala		
Ser 67 → Ala		
Ala 86 → Gly		

[a] (A) Mutants with a substitution that does not appear among the 1572 exchanges considered by Dayhoff (1978) that have an activity greater than 1.5 unit. (B) Mutants with a substitution that occurs 50 or more times in the exchanges considered by Dayhoff (1978) with an activity of less than 0.5 unit.

while others are not (Table I). This low tolerance to substitution at buried sites and variable tolerance to substitution at surface sites is in general agreement with observations reported by Reidhaar-Olson and Sauer (1990) in which buried sites and some surface sites in the N-terminal domain of λ repressor are relatively intolerant of substitution, but many surface sites are tolerant of substitution. Although some surface sites and a few buried sites in the gene V protein are tolerant of mutation, the exposure of a side chain to solvent is not the most important factor determining the activity of a mutant. When surface residues are considered alone, the standard deviation of activities of mutants with a particular amino acid substitution, at various sites, is 0.69 unit, the same value as found for substitutions at all sites. This means that specific interactions made by amino acid side chains, and not simply the solvent accessibility of the site, dominate the dependence of activity of a mutant on the location of the mutation within the gene V protein.

Context-Dependent Properties. We tested two context-dependent properties for their contributions to the variation in activities of mutants (Table III). As discussed above, the exposure of a side chain in the wild-type protein to solvent was somewhat correlated with the activities of mutants at that site. We estimated the mean square contribution of this property to the variation in activities of mutants (Table III) and found it to be about 3% of the total. On the other hand, the turn or β sheet conformational parameters (Chou & Fasman, 1978) of amino acids being substituted at turns or β sheet sections of the gene V protein, respectively, were not correlated with the activities of the mutant proteins (Table III). When the exposure of residues to solvent was combined with the three most useful properties of the amino acids being exchanged (see above), the contribution of these properties to the mean square variation in activities from mutant to mutant was estimated to be about 21% of the total (Table III, set 2).

Sensitivity of Sites in the Gene V Protein to Mutation. A second possible reason for the variation in activity of mutants with the same substitutions at different sites would be a variation in the properties of the sites themselves. The sensitivities of different sites to mutation might vary considerably even among those sites that are buried or those that are exposed. A large variation in tolerance to substitution has been observed previously for sites in several other proteins (Loeb et al., 1989; Bowie et al., 1990; Kleina & Miller, 1990). We find a substantial variation in tolerance to substitution at various sites in the gene V protein. At some sites, certain substitutions lead to functional proteins and other substitutions do not. Substitutions of Glu 5 by aspartate or glycine, for example, lead to functional proteins, while substitutions of the same residue by proline or valine lead to inactive proteins (Table I). On the other hand, there are some sites, such as Ile 6 or Val 19, where nearly all substitutions lead to nonfunctional proteins, and other sites, including Pro 8 and Lys 87, where nearly all substitutions lead to active proteins. The variation in tolerance to substitution from site to site indicates that the characteristics of the site of a substitution are indeed very important in determining whether a particular mutant gene V protein will be functional. These characteristics might include, for example, specific interactions such as salt bridges or hydrogen bonds with the remainder of the protein or with DNA, or the geometry of the space available for the side chain at that site, as well as the polarity of the site.

Use of the Tolerance of Each Site to Mutation To Predict the Activity of a Mutant Protein. If some sites are much more tolerant of substitution than others, then it might be possible to predict the effect of a particular amino acid substitution on the function of a protein on the basis of the effects of other substitutions at the same site. We found that this approach was effective. First we simply used the average activity of all other mutants at each site to predict the activity of a particular mutant at that site (Table III, parameter 12). The optimal value of the parameter b_1 was 0.67 overall (Table III and Appendix), indicating that the best estimate of the activity of a mutant is quite close to that of all other mutants at the same site. We estimate (Table III) that the tolerance of individual sites alone contributes about 30% of the mean square variation in activities. This contribution is considerably greater than the combined contributions of all properties of the amino acids that we tested, and almost as large as our estimate of the contributions of all factors based on the identities of the amino acids that are exchanged.

Table III: Estimates of the Contributions of Various Properties to the Variation in Activity of Mutants[a]

	all sites		buried sites		exposed sites	
property of substitution from amino acid *i* to *j*	b_k	ν (%)	b_k	ν (%)	b_k	ν (%)
	Individual Properties					
(1) frequency of exchange	0.012	7.7	0.012	9.0	0.011	6.8
(2) amino acid *i* is Gly or Pro	0.044	0.0	0.590	4.7	−0.131	0.0
(3) amino acid *j* is Gly or Pro	−0.595	8.2	−0.707	9.8	−0.562	7.4
(4) introduce β-branched amino acid	0.220	1.9	0.175	1.1	0.205	1.6
(5) remove β-branched amino acid	−0.195	0.7	−0.014	0.0	−0.284	1.3
(6) $\Delta\Delta G^{tr}$	0.007	0.0	0.133	5.9	−0.046	0.7
(7) $\|\Delta\Delta G^{tr}\|$	−0.153	4.2	−0.263	12.8	−0.112	1.9
(8) $\|\Delta V\|$	−0.002	0.0	−0.000	0.0	−0.002	0.2
(9) $\|\Delta Q\|$	−0.021	0.0	−0.045	0.0	−0.047	0.0
(10) exposure of side chain	0.005	3.4	0.017	0.0	0.007	3.9
(11) conformational preference of amino acid	0.163	0.3	0.382	2.6	0.082	0.0
(12) tolerance of site to substitution	0.670	30.0	0.811	44.0	0.588	22.3
	Combinations of Properties					
set 1: properties 1, 3, and 7		16.8		24.3		14.0
(1) frequency of exchange	0.008		0.006		0.009	
(3) amino acid *j* is Gly or Pro	−0.611		−0.673		−0.580	
(7) $\|\Delta\Delta G^{tr}\|$	−0.121		−0.219		−0.082	
set 2: properties 1, 3, 7, and 10		21.1		23.9		18.3
(1) frequency of exchange	0.008		0.007		0.009	
(3) amino acid *j* is Gly or Pro	−0.601		−0.670		−0.546	
(7) $\|\Delta\Delta G^{tr}\|$	−0.131		−0.220		−0.085	
(10) % exposure of side chain	0.006		0.023		0.007	
set 3: properties 1, 3, and 12		44.7		62.1		35.0
(1) frequency of exchange	0.010		0.012		0.008	
(3) amino acid *j* is Gly or Pro	−0.519		−0.484		−0.540	
(12) tolerance of site to substitution	0.700		0.841		0.608	

[a] Properties were tested individually or in combination as indicated, by refining the parameters b_k in models based on eq A5 or A15. The relative contribution of the property or properties in each model to the mean square variation in the activities of mutants, ν, was calculated as described in the Appendix (eq A12). The functions based on these properties ($g_{k,n}$) that we used were all constructed so that their mean values, averaged over all the mutants considered in determining the parameters b_k, were zero. The functions, before subtraction of their mean values, were as follows. For property 1, the function used was simply the frequency of exchange of each pair of amino acids among homologous proteins (Dayhoff, 1978). For properties 2 and 3, the function had a value of 1 if the wild-type or substituting amino acids, respectively, were glycine or proline and a value of zero otherwise. The functions for properties 4 and 5 had values of 1 if the substitution introduced or removed, respectively, a β-branched amino acid and a value of zero otherwise. The function for property 6 corresponded to the difference in free energies of transfer, $\Delta\Delta G^{tr}$, of the two amino acid side chains from octanol to water, in units of kcal/mol (Fauchère & Pliška, 1983). The function for property 7 corresponded to the absolute value of this difference, $|\Delta\Delta G^{tr}|$. The function for property 8, $|\Delta V|$, was the absolute value of the change in volume of the amino acids, in units of Å^3 (Richards, 1974, 1977), and the function for property 9, $|\Delta Q|$, was the absolute value of the change in charge of the side chain, in units of electrons. Property 10 was the percentage of the surface area of the residue in the wild-type protein exposed to solvent (Brayer & McPherson, 1983; Kabsch & Sander, 1983). Property 11 was the conformational parameter (Chou & Fasman, 1978) for the substituting amino acid for either β structure or for β turns, depending on the type of secondary structure at the site in the wild-type protein (Brayer & McPherson, 1983). Property 12 was the tolerance of the site to substitution, as described in the Appendix. For all analyses except those involving the tolerance of the site to substitution, there were 92 mutants considered at buried sites and 221 mutants at surface sites, for a total of 313 mutants. For those involving the tolerance of the site to substitution, only sites with at least three mutants were considered. In these analyses there were 92 mutants considered at buried sites and 169 at surface sites, for a total of 261 mutants. For the calculations involving the tolerance of the site to substitution, the square root of the estimated mean square variation in activities of mutants at all sites was 0.83, for buried sites it was 0.82, and for surface sites it was 0.84. The mean of all activities for this set of mutants was 0.96. The square root of the estimated variance calculated with eq A2 after fitting the model on the basis of properties 1, 3, and 12 was 0.67 for all sites, 0.60 for buried sites, and 0.72 for surface sites.

We note that, in this simplified model, we do not differentiate between sites where all mutants are partially active and sites where some are fully active and others are inactive. In both cases, the predicted activity at the site would be about equal to the overall average activity.

On the basis of the good correlation between the tolerance of individual sites to substitution and the activities of mutants at those sites, and the weaker correlations of properties of the amino acids and these activities, we constructed and tested a model that combines both types of effects. In this model, the activity of a mutant is predicted on the basis of both a set of properties of the amino acids and the tolerance of the site to mutation. The model included the tolerance of each site to mutation, the frequency of exchange of the wild-type and substituting amino acids among homologous proteins, and whether the substituting amino acid is a glycine or proline. The results of this modeling are listed in Table III (set 3) and are shown graphically in Figure 2. We estimate that the model accounts for about 45% of the mean square variation in the activities of the gene V protein mutants. Figure 2 shows that most of those mutants that are predicted to be active do have high activities and most of those that are predicted to

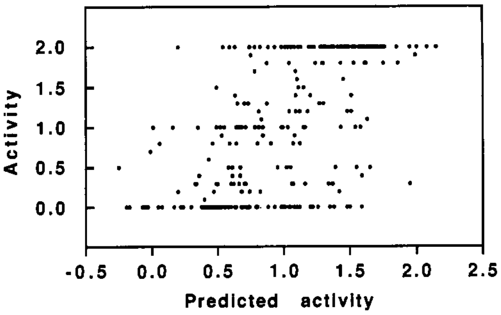

FIGURE 2: Prediction of the activities of gene V protein mutants based on the activities of other mutants at that site, the frequency of exchange of the wild-type and substituting amino acids (Table I), and whether the substituting amino acid is a glycine or proline (see Materials and Methods). Some points superimpose in this figure. Only mutants at sites where three or more mutants have been characterized are included.

Table IV: Activities of Mutants with a Single Substitution at Different Sites[a]

substitution	sites	activities
Leu → Ala	28, 37, 76	2.0, 0.0, 0.0
Leu → Ile	32, 37, 81	0.5, 0.8, 1.0
Leu → Pro	28, 32, 37, 44, 49	0.0, 2.0, 0.0, 0.0, 0.0
Leu → Ser	32, 37, 44, 49, 60, 76	0.0, 0.0, 0.0, 0.0, 2.0, 0.0
Leu → Arg	32, 37, 49, 65	2.0, 0.3, 1.0, 2.0
Leu → His	28, 49, 83	1.5, 0.0, 0.3
Val → Ala	35, 70, 84	1.8, 1.1, 2.0
Val → Leu	35, 43, 45	0.3, 1.0, 1.1
Val → Ile	35, 45, 63	1.8, 2.0, 2.0
Val → Pro	4, 19, 70	0.0, 0.0, 1.0
Val → Ser	35, 43, 45	0.0, 1.7, 0.0
Ile → Pro	2, 6, 47	0.8, 0.0, 0.0
Pro → Leu	25, 42, 85	0.0, 1.0, 0.0
Pro → Ser	25, 54, 85	2.0, 2.0, 1.0
Gly → Cys	18, 71, 74	0.8, 1.8, 1.0
Gly → Thr	38, 52, 74	0.0, 1.2, 0.0
Gly → Ser	23, 38, 59	0.0, 0.2, 0.0
Gly → Lys	38, 52, 74	1.0, 1.6, 1.2
Ser → Thr	9, 17, 27, 67	2.0, 2.0, 2.0, 2.0
Lys → Ser	3, 46, 69	0.8, 1.0, 1.5
Tyr → Arg	26, 41, 61	1.3, 2.0, 0.0
Tyr → Phe	26, 34, 41	1.8, 2.0, 1.8

[a] Substitutions that we have obtained at three or more sites are listed. The sites at which substitutions were obtained are indicated, and the activities of mutants with these substitutions are indicated in the same order.

Table V: Activities of Mutants (Mean ± SEM) as a Function of Solvent Accessibility and Hydrophobicities of Substituted Residues[a]

location of residue	polarity of w.t. amino acid	polarity of substituting residue	no. of mutants	av act. of mutants
buried	apolar	apolar	33	1.2 ± 0.1
		polar	40	0.4 ± 0.1
	polar	apolar	5	1.1 ± 0.4
		polar	19	1.1 ± 0.2
surface	apolar	apolar	20	1.2 ± 0.2
		polar	42	0.9 ± 0.1
	polar	apolar	60	0.9 ± 0.1
		polar	94	1.1 ± 0.1

[a] Residues are classified by exposure to external solvent in the crystal structure of the gene V protein (see footnote to Table I), where buried residues are those with less than 10% of their surface exposed to a solvent water molecule (Richards, 1977) and the remainder are classified as surface residues. Apolar residues are defined here as cysteine, methionine, isoleucine, leucine, valine, phenylalanine, tyrosine, and tryptophan (Faucheře & Pliška, 1983).

be inactive have low activities. This means that a model of this type may be useful in predicting whether a mutation with unknown activity will be active, as long as some information is available on the activities of other mutations at the same site. As the prediction illustrated in Figure 2 accounts for only about half the mean square variation in activities of gene V protein mutants, however, it is clear that the tolerance of a site to substitution and the general characteristics of the wild-type and substituting amino acids at a site are often not sufficient to make an accurate prediction of the activity of a mutant. In particular, mutants with predicted activities near the overall mean activity are frequently either fully active or fully inactive. As our model does not take into account the detailed interactions made by an amino acid in the protein, it cannot generally predict the activity of a mutant at a site where some mutants are active and others are not and the variation in activity is due to specific interactions made by each residue. We expect that detailed structural information will ultimately allow a much better prediction of the activity of a mutant than any general methods such as we have developed. Improved models could incorporate information on the interactions both within the protein and between protein and substrate and could take into account, in detail, the physical properties that lead to the effects we have modeled.

It is somewhat remarkable that our most effective model for predicting the activities of gene V protein mutants does not use the available knowledge of the structure of the protein. Additional terms based on the exposure of a site to solvent or the secondary structure of the polypeptide chain at that site do not substantially improve the prediction. Presumably, this is in part because the sensitivity of a site to mutation is itself related to the exposure of the site to solvent (Bowie et al., 1990; Reidhaar-Olson & Sauer, 1990) and the secondary structure preferences of the amino acids have smaller average effects on activities than those of the factors we have included.

The observations we have made on the gene V protein are probably applicable to many other proteins as well. For example, we applied our analysis to a collection of 1716 single amino acid substitution mutants of the lac repressor (Kleina et al., 1990) and to a collection of 336 mutants of the HIV-1 protease (Loeb et al., 1989) that have recently been characterized. We found that, just as in the case of the gene V protein, the frequency of exchange of amino acids among homologous proteins is very poorly correlated with the activities of mutants with these exchanges in both proteins, accounting for less than 9% of the mean square variation in activities in both cases. Also as in the gene V protein, the tolerance of each site in the two proteins to substitution is a very good predictor of the activity of a mutant at that site, accounting for about 60% of the mean square variation in activities of the mutants of the lac repressor and 40% of the mean square variation in activities of the HIV-1 protease mutants.

Conclusions

On the basis of our analysis of the activities of mutants in the gene V protein, we conclude that neither the frequency of exchange of amino acids between homologous proteins nor any other measure of the properties of the amino acids will be particularly useful by themselves in predicting whether a protein with an amino acid substitution will be functional. The context of a mutation plays such a strong role in determining the activity of a mutant that such a prediction would be very uncertain. A reasonable prediction of the effects of amino acid substitutions can be made, however, by considering the sensitivity to substitution of the site of a mutation. Combined with information on the properties of the amino acid being substituted, this approach can yield a useful estimate of the activity of a mutant protein.

One use of such a predictive scheme would be in deciding where in a protein sequence to place a fluorophore, a prosthetic group, or a modified amino acid. If a collection of mutants for a protein were available with several mutants at each site, sites that are very tolerant of substitution could be identified and these would be suitable locations for introducing bulky or polar groups. Another potential use is in sequence alignment procedures, which currently depend on some measure of the likelihood that pairs of amino acids will substitute for each other among homologous proteins (Gribskov et al., 1987; Pearson & Lipman, 1988). It seems possible that amino acids at sites that are tolerant of substitution in a protein are more likely to be different from the amino acids at corresponding sites in homologous proteins than the amino acids at sites that are sensitive to substitution. Accordingly, if the tolerance to substitution of each site in a protein of interest is known, then other proteins that are structurally or functionally related to this protein might be identified by sequence comparisons in which the alignment of amino acids at sensitive sites is weighted more heavily than at sites that are tolerant. Alternatively, the observed or predicted activities of each of the 20 amino acids at each site could be used as weighting factors in profile-matching alignment procedures (Gribskov et al., 1987). In either case, the additional information on which amino acids are allowed at each site provided by a predictive scheme such as that presented here is, in effect, structural and functional information that might allow a more certain identification of related protein sequences than is currently obtainable.

Acknowledgments

We thank P. Schlunk, W. Sandberg, and H. Liang for helpful discussions and H. Tarczy, S. Choi, J. Oh, R. Parish, and Bionetics Research for DNA sequencing.

Appendix

Error Estimates. We evaluated uncertainties in measurements by estimating the standard deviation of measurements in two ways. First, to examine the reproducibility of our measurements, we evaluated the standard deviation (σ_A) of all the activity measurements pertaining to each mutant protein, averaged over all mutations (Bevington, 1969). Next, to determine whether there were systematic variations in our measurements from day to day, we compared the activities of mutants measured on more than one day. We estimated the standard deviation (σ_B) of these measurements, averaged over all mutations, using the relation

$$\sigma_B{}^2 = \frac{1}{(N - m)}\sum_{i=1}^{m}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_i)^2 \qquad \text{(A1)}$$

where y_{ij} is the jth measurement of the value of the activity of the ith mutant, $\bar{y}_i$ is the mean value of the activity of the ith mutant overall, based on N_i measurements, and there are a total of N measurements made on m different mutants (Bevington, 1969).

To determine whether codon usage had an effect on our measurements, we calculated the standard deviation of the activities of mutants with the same protein sequences, but different DNA sequences (σ_C) using eq A1. We did this by substituting y_{ij} in eq A1 with the value of the activity of the mutant with the jth different DNA sequence encoding the ith protein sequence. In this case, $\bar{y}_i$ is the mean value of the activity of mutants with the ith protein sequence, based on N_i different DNA sequences, and there are a total of N measurements made on m different mutants.

Estimating the Contribution of an Arbitrary Function of an Amino Acid Substitution to the Activity of a Mutant. The utility of each model (see Results and Discussion) for predicting the activity of mutant proteins was evaluated by estimating the variance, s^2, between the observed activities and the activities calculated by using the model:

$$s^2 = \frac{1}{(N - K)}\sum_{n=1}^{N}(A_n{}^{\text{obs}} - A_n{}^{\text{model}})^2 \qquad \text{(A2)}$$

where $A_n{}^{\text{obs}}$ and $A_n{}^{\text{model}}$ are the measured and predicted activities of mutant n, there are a total of N measurements of the activities of mutants, and there were K coefficients used to fit the model to the observed activities (Bevington, 1969).

We used this variance (s^2) to estimate the contribution of properties included in the model to the activities of mutants. To do this, we first constructed a simplified description of a possible relationship between the amino acid substitution at a site and the activity of the resulting mutant. We will consider this simplified description to be the "true" relationship in this analysis. Then we develop a model which has elements that are correct, that is, that are part of the "true" relationship, as well as elements that are not. This model corresponds to a model that might be developed in practice. Finally, we calculate the expected values of the coefficients in the model, and from these values, we show how the variance, s^2, depends on that part of the model that is correct.

Initially, we assume that the activity of a mutant depends only on the amino acid substitution, and not on the site at which it is made. For purposes of this analysis let the "true" activity of the nth mutant (A_n) be a combination of some linearly independent functions ($f_{k,n}$) of the properties of the wild-type and substituting amino acids in this mutant:

$$A_n = A_0 + \sum_{k=1}^{m} a_k f_{k,n} \qquad \text{(A3)}$$

Here A_0 is the mean value of activities overall, and a_k are coefficients expressing the relative contributions of the m properties to the activity of a mutant with this amino acid substitution. To simplify our analysis, the functions $f_{k,n}$ are defined in such a way that they have mean values of zero and are uncorrelated with each other.

In practice, were are given N measurements of activities of mutants, $A_n{}^{\text{obs}}$, assumed to be uniformly distributed among the possible amino acid substitutions. If there are some errors in measurement, the measured activity of the nth mutant ($A_n{}^{\text{obs}}$) is given by

$$A_n{}^{\text{obs}} = A_n + z_n \qquad \text{(A4)}$$

where the "true" activity of this mutant is A_n, given by eq A3, and the error in measurement is z_n. The mean value of z_n is assumed to be zero, and its variance is given by $\sigma_{\text{obs}}{}^2$. We assume that the number of measurements, N, is very large so that the mean value of the activity, A_0 (eq A3), can be accurately determined from the observations of $A_n{}^{\text{obs}}$, and we assume that $\sigma_{\text{obs}}{}^2$ can be determined independently from repetitions of measurements.

Next, suppose that we have a model that we would like to test for its utility in predicting the values of activities, and that this model has the same form as the "true" relationship between properties of the amino acids and activities of mutants (eq A3), but that the functions in the model ($g_{k,n}$) are not exactly equal to the true functions ($f_{k,n}$), and not all of the "true" functions are included in the model. The model to be tested could then be written as

$$A_n{}^{\text{model}} = A_0 + \sum_{k=1}^{K} b_k g_{k,n} \qquad \text{(A5)}$$

where $g_{k,n}$ are functions that are potentially related to the activity of a mutant with substitution n, the coefficients b_k are parameters reflecting the contribution of each function $g_{k,n}$ to these activities, and K functions $g_{k,n}$ are included in the model. We assume that the number of functions in the model (K) is small compared to the number of measurements made (N). For our analysis, we assume that the "true" functions, $f_{k,n}$, and the functions in the model, $g_{k,n}$, are related by

$$g_{k,n} = f_{k,n} + e_{k,n} \qquad \text{(A6)}$$

where $e_{k,n}$ are functions reflecting the errors in the formulation of the model functions, $g_{k,n}$, and are assumed to have mean values of zero and to be independent of the true functions, $f_{k,n}$. The parameters b_k in eq A5 are determined by minimizing the mean square difference between predicted ($A_n{}^{\text{model}}$) and observed ($A_n{}^{\text{obs}}$) activities for all mutants examined by using eq A2.

As the functions $g_{k,n}$ are independent and N is very large, we can calculate the expected values of the coefficients b_k.

First, substituting eq A5 into eq A2 and determining the values of the coefficients b_k that minimize the variance, s^2, lead to

$$b_k \approx \sum_{n=1}^{N} (A_n^{\text{obs}} - A_0) g_{k,n} / \sum_{n=1}^{N} g_{k,n}^2 \qquad \text{(A7)}$$

The expected relationship between the coefficients in the model b_k (eq A7) and the true coefficients a_k (eq A3) can be determined by substituting A_n^{obs} in eq A7 with eq A3 and A4, and $g_{k,n}$ with eq A6. Noting that the functions $f_{k,n}$ and $e_{k,n}$ are independent, this yields

$$b_k \approx a_k \alpha_k \qquad \text{(A8)}$$

where

$$\alpha_k = \frac{\langle f_k^2 \rangle}{\langle f_k^2 \rangle + \langle e_k^2 \rangle} \qquad \text{(A9)}$$

Here $\langle f_k^2 \rangle$ and $\langle e_k^2 \rangle$ are the overall mean square values of the functions $f_{k,n}$ and $e_{k,n}$, respectively. Equation A8 indicates that the coefficients (b_k) in the model (eq A5) will be smaller than the actual coefficients (a_k) in eq A3 by an amount depending on the ratio of the parts of the functions in the model ($g_{k,n}$) that are correct, $\langle f_k^2 \rangle$, to those that are incorrect, $\langle e_k^2 \rangle$.

Finally, the mean square difference, s^2, that would be obtained if the observed activities were given by eq A4 and the model in eq A5 were fitted to them may be estimated by substituting eq A3–A6, A8, and A9 into eq A2. Considering the independence of the errors z_n and the functions $f_{k,n}$ and $e_{k,n}$ and the fact that the number of measurements N is very large relative to the number of coefficients in the model, K, this yields

$$s^2 \approx s_0^2 - \sum_{k=1}^{K} \{\alpha_k a_k^2 \langle f_k^2 \rangle\} \qquad \text{(A10)}$$

where s_0^2 is given by

$$s_0^2 \approx \sigma_{\text{obs}}^2 + \sum_{k=1}^{m} \{a_k^2 \langle f_k^2 \rangle\} \qquad \text{(A11)}$$

By use of eqs A3 and A4, s_0^2 may be shown to correspond to the overall mean square deviation, $\langle (A_n^{\text{obs}} - A_0)^2 \rangle$, of the activities of the N mutants from their average value, A_0. Equation A10 shows that the improvement in mean square deviation, $s_0^2 - s^2$, made by using the model with functions $\mathrm{g}_{\mathrm{k,n}}$ and refined coefficients $\mathrm{b_k}$, corresponds to the sum of the mean square values of the "true" terms ($a_k f_{k,n}$) corresponding to each of the terms ($b_k g_{k,n}$) in the model, weighted by the factors α_k in eq A9. We use this improvement in mean square deviation, $s_0^2 - s^2$,.as an estimate of the mean square value of the contribution of the terms included in the model to the variation in activity from mutant to mutant. In this analysis, we neglect the factors α_k in eq A9, as their values are unknown. The contribution of the terms in the model to the variance in activity from mutant to mutant could therefore be somewhat higher than our estimates.

Finally, we use eqs A10 and A11 to normalize the mean square contribution of the terms included in the model to the mean square contribution of all the "true" terms in eq A3. This relative mean square contribution (ν) is given by

$$\nu \approx \frac{s_0^2 - s^2}{s_0^2 - \sigma_{\text{obs}}^2} \qquad \text{(A12)}$$

In practice, to calculate the value of ν, we determine the value of s_0^2 from the mean square value of the difference between the activities of mutants and their mean, $\langle (A_n^{\text{obs}} - A_0)^2 \rangle$, the value of s^2 from eq A2, and the value of σ_{obs}^2 from repetitions of measurements made on different days (σ_{B}^2, eq A1).

A similar analysis was used to evaluate the contribution of site-specific effects to the activity of mutants, except that one function ($g_{K+1,n}$) was made to depend not just on the amino acid substitution but also on the values of activities of other mutants at the same site. To do this, it was assumed that the "true" activity of a mutant at site j is given by

$$A_n = A_0 + \sum_{k=1}^{m} a_k f_{k,n} + C_j \qquad \text{(A13)}$$

where C_j, the tolerance of the site to substitution, reflects the effect of properties of site j on the activity of mutants at that site. To predict the activities of mutant n at site j, the value of C_j is estimated from the measured activities of other mutants at that site, after correction for the activities expected from effects due the type of substitutions. Our estimate (D_{jn}) of the tolerance to substitution C_j is given by

$$D_{jn} = \frac{1}{N_j} \sum_{i=1}^{N_j} (A_i^{\text{obs}} - A_i^{\text{model}}) \qquad \text{(A14)}$$

where the summation is over all mutants i at site j except the mutant of interest, for a total of N_j measurements, and A_n^{model} is based on the properties of the wild-type and substituting amino acids using eq A5. Our prediction of the value of the activity of a mutant at site j (B_n^{model}) is then related to the estimates based on just the amino acid substitution (A_n^{model}) by

$$B_n^{\text{model}} = A_n^{\text{model}} + b_{K+1} D_{jn} \qquad \text{(A15)}$$

where the parameter b_{K+1} is included to reflect the possibility that the estimate D_{jn} of the tolerance of site j to substitution contains some errors. As in the case of predicting activities based on only the properties of the wild-type and substituting amino acid, the parameters b_k in eqs A5 and A15 are refined so as to minimize the mean square difference between calculated (B_n^{model}) and observed (A_n^{obs}) activities. According to eq A5, the expected activity of a mutant based on the type of substitution (A_n^{model}), used as a correction factor in eq A14, itself depends on the values of the refined parameters b_k. Accordingly, we alternately used eq A14 to determine D_{jn} and refined all $K + 1$ parameters b_k in eqs A5 and A15 until convergence.

References

Alberts, B., Frey, L., & Delius, H. (1972) *J. Mol. Biol. 68*, 139–152.

Amann, E., Brosius, J., & Ptashne, M. (1983) *Gene 25*, 167–178.

Anderson, R. A., Nakashima, Y., & Coleman, J. E. (1975) *Biochemistry 14*, 907–917.

Baas, P. D. (1985) *Biochim. Biophys. Acta 825*, 111–139.

Bayne, S., & Rasched, I. (1983) *Biosci. Rep. 3*, 469–474.

Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.

Bowie, J. U., & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 2152–2156.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990) *Science 247*, 1306–1310.

Brayer, G. D., & McPherson, A. (1983) *J. Mol. Biol. 169*, 565–596.

Cavalieri, S. J., Neet, K. E., & Goldthwait, D. A. (1976) *J. Mol. Biol. 102*, 697–711.

Chou, P. Y., & Fasman, G. D. (1978) *Adv. Enzymol. 47*, 45–148.

Coleman, J. E., & Oakley, J. L. (1980) *CRC Crit. Rev. Biochem. 7*, 247–289.

Davis, N. G., Boeke, J. D., & Model, P. (1985) *J. Mol. Biol. 181*, 111–121.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978) in *Atlas of protein sequence and structure* (Dayhoff, M. O., Ed.) pp 345–352, National Biomedical Research Foundation, Washington, DC.

DeGrado, W. F. (1988) *Adv. Protein Chem. 39*, 51–124.

de Jong, E. A. M., van Duynhoven, J. P. M., Harmsen, B. J. M., Konings, R. N. H., & Hilbers, C. W. (1989a) *J. Mol. Biol. 206*, 119–132.

de Jong, E. A. M., van Duynhoven, J. P. M., Harmsen, B. J. M., Tesser, G. I., Konings, R. N. H., & Hilbers, C. W. (1989b) *J. Mol. Biol. 206*, 133–152.

Dick, L. R., Sherry, A. D., Newkirk, M. M., & Gray, D. M. (1988) *J. Biol. Chem. 263*, 18864–18872.

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1982) *Nature 299*, 371–374.

Fauchère, J.-L., & Pliška, V. (1983) *Eur. J. Med. Chem.-Chim. Ther. 4*, 369–375.

French, S., & Robson, B. (1983) *J. Mol. Evol. 19*, 171–175.

Fulford, W., & Model, P. (1988) *J. Mol. Biol. 203*, 39–48.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 4355–4358.

Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., & Scheraga, H. A. (1985) *J. Protein Chem. 4*, 23–55.

King, G. C., & Coleman, J. E. (1987) *Biochemistry 26*, 2929–2937.

Kleina, L. G., & Miller, J. H. (1990) *J. Mol. Biol. 212*, 295–318.

Kowalczykowski, S. C., Bear, D. G., & Von Hippel, P. H. (1981) In *The Enzymes* (Boyer, P. D., Ed.) Vol. XIV, pp 373–444, Academic Press, New York.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., & Hutchison, C. A., III (1989) *Nature 340*, 397–400.

Michel, B., & Zinder, N. D. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 4002–4006.

Miller, J. H. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Model, P., McGill, C., Mazur, B., & Fulford, W. D. (1982) *Cell 29*, 329–335.

Pearson, W. R., & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 2444–2448.

Pretorius, H. T., Klein, M., & Day, L. A. (1975) *J. Biol. Chem. 250*, 9262–9269.

Reidhaar-Olson, J. F., & Sauer, R. T. (1990) *Proteins 7*, 306–316.

Richards, F. M. (1974) *J. Mol. Biol. 82*, 1–14.

Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng. 6*, 151–176.

Risler, J. L., Delorme, M. O., Delacroix, H., & Henaut, A. (1988) *J. Mol. Biol. 204*, 1019–1029.

Salstrom, J. S., & Pratt, D. (1971) *J. Mol. Biol. 61*, 489–501.

Sandberg, W. S., & Terwilliger, T. C. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 1706–1710.

Terwilliger, T. C. (1988a) *Gene 69*, 317–324.

Terwilliger, T. C. (1988b) *Gene 71*, 41–47.

van Duynhoven, J. P. M., Folkers, P. J. M., Stassen, A. P. M., Harmsen, B. J. M., Konings, R. N. H., & Hilbers, C. W. (1990) *FEBS Lett. 261*, 1–4.

Yen, T. S., & Webster, R. E. (1982) *Cell 29*, 337–345.

Zabin, H. B., & Terwilliger, T. C. (1991) *J. Mol. Biol.* (in press).

Zaman, G. J. R., Schoenmakers, J. G. G., & Konings, R. N. H. (1990) *Eur. J. Biochem. 189*, 119–124.