Using Model Proteins to Quantify the Effects of Pathogenic Mutations in Ig-like Proteins*

Lucy G. Randles 1, Ilkka Lappalainen 2, Susan B. Fowler, Benjamin Moore 3, Stefan J. Hamill, and Jane Clarke 4

From the University of Cambridge, Department of Chemistry, Medical Research Council Centre for Protein Engineering, Lensfield Road, Cambridge CB2 1EW, United Kingdom

It has proved impossible to purify some proteins implicated in disease in sufficient quantities to allow a biophysical characterization of the effect of pathogenic mutations. To overcome this problem we have analyzed 37 different disease-causing mutations located in the L1 and IL2Rγ proteins in well characterized related model proteins in which mutations that are identical or equivalent to pathogenic mutations were introduced. We show that data from these models are consistent and that changes in stability observed can be correlated to severity of disease, to correct trafficking within the cell and to in vitro ligand binding studies. Interestingly, we find that any mutations that cause a loss of stability of more than 2 kcal/mol are severely debilitating, even though some model proteins with these mutations can be easily expressed and analyzed. Furthermore we show that the severity of mutation can be predicted by a ΔΔG evolution scale, a measure of conservation. Our results demonstrate that model proteins can be used to analyze disease-causing mutations when wild-type proteins are not stable enough to carry mutations for biophysical analysis.

The sequencing of the human and other genomes presents a number of challenges to biophysics: how far is it possible to predict structure and function of a protein from its sequence, and of particular relevance to this study, can one predict how changes in sequence will affect the biophysical properties of a protein?

Although a number of computational methods for predicting the effect of amino acid substitutions have been published recently (1–5) and the wealth of protein engineering results of mainly conservative mutations has provided insight into protein folding and stability (6, 7), biophysical studies of pathogenic mutations are still in their infancy. The effect of disease-associated mutations on the structure and kinetic/thermodynamic stability has been studied in only a few affected proteins including p53 (8, 9), Cu,Zn superoxide dismutase (10–13), and cystic fibrosis transmembrane conductase regulator (14). The small number of systems analyzed is due partly to intrinsic difficulties in expression, purification, and analysis of these proteins in vitro. Here we investigate the use of stable, well characterized, model proteins to predict the properties of disease-associated proteins in the same structural family.

As part of on-going studies to investigate the relationship between structure and the folding of proteins, we have studied a number of proteins from two superfamilies, fibronectin type III (fnIII) 5 and immunoglobulin (Ig) domains (Fig. 1). The folding of a number of these proteins has been studied in depth (15–24). Both superfamilies are very common, the current fnIII and I-set Ig domain alignments in the Pfam data base (version 19) include 8445 and 7041 sequences, respectively (25). Therefore a wealth of bioinformatics data exists on sequence variation within the superfamilies (and within different classes of Ig domains). The logical extension of this work was to investigate the effect of pathogenic mutations in these domains.

We chose to study the effect of pathogenic mutations in two proteins where point mutations in Ig and fnIII domains cause disease. Mutations in the L1CAM and IL2RG genes affect the two encoded proteins, L1 and the IL2Rγ, respectively. No experimental structures are available for these disease-related proteins. The extracellular portion of the L1 protein is predicted to be composed of six I-set Ig domains and five fnIII domains (26, 27). Missense mutations in all of the L1 domains cause X chromosome-linked recessive neurological disorders with a varying presentation both within and between families (28, 29). The pathogenic mutations located in the single fnIII domain on the extracellular part of the IL2Rγ protein are responsible for a rare X-linked severe combined immunodeficiency (30, 31). We intended to investigate the effect of these mutations but our attempts to express and purify individual wild-type L1 and IL2Rγ domains were unsuccessful. The effect of the pathogenic mutations on these domains could not be studied by standard biophysical methods. We therefore determined the effects of equivalent mutations on well characterized model protein systems, to investigate how well the effect of the mutations can be predicted and to determine the properties of an ideal model system.
**RESULTS**

**The Model Protein Systems**—The proteins chosen for our study had all been characterized to some extent in our laboratory. The model fnIII domains were both from human extracellular matrix proteins: TNfn3, the 3rd fnIII domain of human tenascin (33), and FNFn10 the 10th fnIII domain of human fibronectin (34, 35). These domains have \( \sim 25\% \) sequence identity but the same structure. Both have been studied extensively in terms of stability and of folding mechanism (18, 19, 23). Only one of the model Ig I-set proteins had been subject to an intensive mutagenic study, TI I27, the 27th Ig domain from the I-band of human cardiac titin (16, 36). The other proteins that we used had all been shown to be stable, folded, to behave in a two-state manner at equilibrium, and to be amenable to biophysical study. These were two other titin domains, for which no structural information was available but where there were hand-curated alignments (TI I28, TI I30), and TK (the C-terminal Ig domain of turkey myosin light chain kinase) for which a structure was available (37). The free energy of unfolding of the wild-type proteins was as follows (in kcal/mol): 6.6 TK, 7.5 TI I27, 3.9 TI I28, 5.0 TI I30, 6.7 TNfn3, 9.4 FNFn10 (15, 16, 19).

**Structural Alignments**—It was necessary to use model systems to study pathogenic mutations because there is no three-dimensional structure. Based on structural alignments between disease-associated proteins and their model proteins, identical or equivalent mutations were created in the model protein domains. In the case of an identical mutation, the model and disease-associated protein domains both have the same residue at a site, an equivalent mutation was made where the substitution was as close as possible to that of the disease-causing mutation. 26 positions in the disease-related proteins were probed by creating a total of 41 mutations in the model proteins. When possible, the mutation was made in more than one model protein affecting corresponding positions. The mutations made are reported in Table 1.

**Effects of the Mutations on Thermodynamic Stability**—Some of the mutant proteins could not be purified in soluble form (supplemental Table 1). The thermodynamic stability of the wild-type and all soluble mutant proteins was determined by chemical denaturation. Most of the mutant proteins were less stable than the wild type with measured changes in free energy of unfolding (\( \Delta \Delta G \)) ranging from \(-0.2\) to \(5.1\) kcal/mol (Table 1).

The mutations clearly fall into two classes. The mutations affecting buried positions destabilized the model proteins significantly, many were unfolded with \( \Delta \Delta G > \Delta G \) of wild type. These very destabilized proteins were also highly susceptible to forming dimers and higher aggregates during purification. Most probably, the mutant proteins that were insoluble are destabilized to such an extent that they are no longer folded or, at best, only marginally stable. In contrast, proteins with muta-
**TABLE 1**

Thermodynamic properties of all mutations modeled

Mutations that decrease the stability by ≤2 kcal/mol are in plain font. Mutations that decrease protein stability by >2 kcal/mol are shown in bold. Mutations that decrease the stability by more than the WT stability (cause the protein to be unfolded or are insoluble) are shown in italics. Mutations marked * are at positions where the structure alignment is not certain (see "Discussion").

| Disease causing mutationa | Model protein | Mutation made in model proteinb | Residue typeb | ∆∆Gc,d,<d|
|---------------------------|--------------|---------------------------------|---------------|-----------|
| L1 Ig domains             |              |                                 |               |           |
| L120V (1)                 | TK           | L121V (FG)                      | Turn (surface) | -0.1      |
| G1215 (1)                 | TK           | G1225 (G)                       | Surface       | > 6.6 (unfolded) |
| I1795 (2) Buried          | TI I27       | L145 (C)                        | Buried        | 3.9       |
| I1796 (2)                 | TI I27       | H25 (C)                         | Buried        | 4.1       |
| I1796 (2)                 | TI I28       | R47D (C,D)                      | Insoluble     |           |
| R184Q (2)*                | TI I28*      | R47D (C,D)                      | Loop          | 1.5       |
| R184N (2)*                | TI I28*      | R47D (C,D)                      | Loop          | 2.0       |
| H212Q (2)                 | TI I27       | Q74H (F)                        | Surface       | -0.2      |
| P240L (3)                 | TK           | P242 (A)                        | Buried        | Unknown   |
| C269Y (3)                 | TI I28       | C241 (B)                        | Buried        | Insoluble |
| C269Y (3)                 | TI I30       | C231 (B)                        | Buried        | > 5.0 (unfolded) |
| G268D (3)                 | TK           | G67D (B)                        | Buried        | > 6.6 (unfolded) |
| E309K (3)                 | TK           | K112E (F)                       | Surface       | -0.1      |
| E309K (3)                 | TI I27       | E70K (F)                        | Surface       | 0.2       |
| P533R (4)                 | TK           | P42R (A)                        | Buried        | Unknown   |
| W335R (4)                 | TK           | F44R (A)                        | Buried        | > 6.6 (unfolded) |
| W335R (4)                 | TI I28       | F44R (A)                        | Buried        | Insoluble |
| G370R (4)*                | TI I27       | G38R (C)                        | Surface       | 1.8       |
| N408I (4)                 | TK           | N1101 (FG)                      | Turn (buried) | 4.0       |
| G452R (5)                 | TK           | G67R (B)                        | Buried        | > 6.6 (unfolded) |
| L482P (5)                 | TK           | L100P (E)                       | Buried        | Unknown   |
| L482P (5)                 | TI I27       | L58P (E)                        | Buried        | > 7.5 (unfolded) |
| C497Y (5)                 | TI I30       | C731 (E)                        | Buried        | > 5.0 (unfolded) |
| L1 fnIII domains          |              |                                 |               |           |
| K655E (7)                 | TNfn3        | K39E (C)                        | Surface       | 1.2       |
| V768I (8)                 | FNfn10       | V50I (C')                       | Buried        | 1.4       |
| V768F (8)                 | FNfn10       | V50F (C')                       | Buried        | 2.8       |
| V768F (8)                 | TNfn3        | L50F (C')                       | Buried        | 1.3       |
| Y784C (8)                 | FNfn10       | Y68C (F)                        | Buried        | 3.1       |
| Y784C (8)                 | TNfn3        | Y68C (F)                        | Buried        | 4.9       |
| L935P (10)                | FNfn10       | L19P (B)                        | Surface       | 4.0       |
| L935P (10)                | TNfn3        | L19P (B)                        | Surface       | > 6.7 (unfolded) |
| P941L (10)*               | TNfn3        | P25L (BC)                       | Loop (buried) | 2.0       |
| L2R fnIII domain          |              |                                 |               |           |
| A156V*                    | TNfn3        | A4V (A)                         | Buried        | 2.4       |
| A156L*                    | TNfn3        | A4L (A)                         | Buried        | > 6.7 (unfolded) |
| L162H                     | TNfn3        | V10H (A)                        | Buried        | 3.9       |
| L162R                     | TNfn3        | V10R (A)                        | Buried        | 5.1       |
| L172Q*                    | TNfn3        | E20P (B)                        | Buried        | > 6.7 (unfolded) |
| L172Q                     | TNfn3        | E20Q (B)                        | Buried        | > 6.7 (unfolded) |
| C182R*                    | TNfn3        | E29D (BC)                       | Loop (buried) | 2.8       |
| R226C                     | TNfn3        | R75C (FG)                       | Surface       | 0.6       |
| R226H                     | TNfn3        | R75H (FG)                       | Surface       | 1.1       |

a The L1 domain and protein strand in which the mutation is located are shown in parentheses.

b A buried residue has a side chain with solvent accessibility of <10%.

c The value of ∆∆G was calculated from the following equation: ∆∆G = ∆Gm - ∆GmWT, where ∆Gm and ∆GmWT are the mean values of wild type and mutants, and ∆Gm is the concentration of denaturant where half the molecules are folded and half unfolded for wild-type (WT) and mutant (MUT) proteins. ∆GmWT and m were obtained from the fit of the equilibrium denaturant curves to a standard two-state equation (32). We used the following mean m-values: TNfn3 (1.22 kcal/mol m⁻¹), FNfn10 (6.49 kcal/mol m⁻¹) (19); 127 (2.5 kcal/mol m⁻¹) (16); 128 (1.29 kcal/mol m⁻¹), and TK (2.7 kcal/mol m⁻¹). The error in ∆∆G is ±0.1 kcal mol⁻¹.

d "Unfolded" refers to proteins that were expressed solubly but after measuring the fluorescence profiles were found to be unfolded; "Insoluble" mutations were those where protein was expressed but remained in the cell pellet after centrifugation (no attempt was made to solubilize them) and "Unknown" refers to proteins that could be expressed in only very small quantities of monomer (from 4 liters), which prevented measurements being made from them. (Further details of each mutant are available in supplemental Table 1).

**DISCUSSION**

Assessing the Models—The model proteins chosen were mostly well characterized in our laboratory and were predicted to have close structural similarity with the disease protein domains. In general, the more stable the model protein the more informative the data. Most of the mutations made in the most stable models were folded: FNfn10 (∆G = 9.4 kcal/mol, 4/4 folded), TI I27 (∆G = 7.5 kcal/mol, 4/5 folded), and TNfn3 (∆G = 6.7 kcal/mol, 10/14 folded). In contrast, the least stable model TI I28 could not tolerate any mutation of buried residues (∆G = 3.9 kcal/mol, 2/5 folded, both loop mutations). Originally we considered TK (∆G = 6.6 kcal/mol) to be the best model in terms of sequence similarity to the target Ig domain proteins (hence it was the host of many mutations). However, it was intolerant of most mutations in the core, only one of seven buried mutations being stable enough to be folded. Although an unfolded protein shows that the mutation is severe, it is not a useful quantitative measure. The results suggest that more pathogenic mutations could have been made in the more stable model TI I27. TI I27 responded to mutation in the same way as TK with both the same mutation and with equivalent mutations.
Modeling the Effects of Pathogenic Mutations

Hydrophobic to hydrophobic substitutions: effect of change of residue size. Δ size reflects the number of heavy atom groups added (+) or deleted (−) on mutation, e.g. Ala to Val is an increase of 2. The TNfn3 mutants A4V and A4L are displayed as open circles (see “Discussion”). A, ΔΔG as a function of change of residue size. B, ΔΔG as a function of absolute size change.

We conclude that it is best to use the most stable model possible to give usable stability data, and it may be better to make an equivalent mutation to that causing disease in a more stable protein than the identical mutation in a less stable one. The results show that where the same or the equivalent mutation was made in at least two different domains, equivalent results were usually observed. An example is the L1 mutation I179S. Here a buried hydrophobic residue was replaced by a polar residue in the relatively hard to align C’D loop. The mutations V82S and L41S in TK and TI I27, respectively, cause a loss of −4 kcal/mol. This is a large loss of stability and consistent with the hydrophobic residue being significantly buried. TI I28 with the equivalent L42S mutation could not be purified solubly. The free energy of unfolding of wild-type TI I28 is only 3.9 kcal/mol. A destabilization of 4 kcal/mol would be expected to cause the protein to be significantly unfolded and therefore likely to be insoluble.

Two fnIII domains were used as model proteins (TNfn3 and FNfn10). After starting the work on the L1 fnIII domains, we observed (in other work) that FNfn10 behaved “atypically” and that mutations (particularly in the peripheral regions) were not as destabilising as might be predicted from packing considerations (38). We see this same effect in our results to some extent. The identical mutations Y68C and L19P are significantly more destabilising in TNfn3 than in FNfn10 (by ≈2 kcal/mol). This serves to emphasize that an ideal model protein should be extensively characterized before being used in this kind of work. As a result mutations to the model IL2Rγ fnIII domain were made in TNfn3 only.

Analysis of the Unfolding Data—As expected, mutations to surface residues do not alter protein stability significantly, but those at buried sites alter the stability by varying degrees, many resulting in the complete unfolding of the protein. Buried residues in loops and turns were more tolerant to mutation than buried residues in strands. Table 1 shows these results and several trends can be seen.

Substitution of a Buried Hydrophobic by a Polar or Charged Residue—All but one (V10R, TNfn3) of the replacements of a buried by a charged residue were unfolded or produced insoluble aggregates. It seems probable that the V10R mutation is tolerated, as it affects an edge strand, and the charge at the end of a long hydrophobic side chain need not be buried. Replacement of a buried hydrophobic residue with non-charged polar residues appears to be more tolerable, enabling stability measurements to be taken from V82S (TK), L41S (TI I27), and V10H (TNfn3) but note that these are also in edge strands and not deeply buried in the core. I20Q in TNfn3 in the central B-strand causes the protein to be unfolded.

Substitutions at the Surface—Changing the charge or swapping a hydrophobic residue on the surface of a protein had, in most cases, a negligible effect on stability, with only G122S (TK) showing a distinctly lower stability than WT.

Substitutions in Loops—Substitutions in loops generally had an effect intermediate between surface and buried residues, being destablizing but not so radical as to make the protein unfold completely.

Residue-specific Effects: Substitutions Involving Proline Residues—Pro residues are relatively inflexible and are only found in the edge strand of β-sheets (where they introduce a bulge) and in loops in the Ig-like proteins. Positions where Pros were introduced were all originally Leu. Most of these Pros were introduced into buried positions within a sheet and had significant effects on protein stability. Only the surface mutation L19P, made in the very stable model protein FNfn10 in a region known to be tolerant of mutation (38), was expressed in sufficient quantities to allow the stability to be determined and this was highly destabilized. Similarly, removing a Pro from a buried position within an edge strand (P42L and P42R in TK) was highly destabilizing (neither protein could be expressed and purified in sufficient quantities). Consistently, substitution of a buried Pro in a more flexible loop region by the hydrophobic residue Leu (P25L in TNfn3) destabilizes the protein but only to a similar extent as most of the other loop mutations.
Residue-specific Effects: Substitutions Involving Cysteine Residues—Two pathogenic mutations were reported that are predicted to remove a disulfide bond in the parent Ig domain of L1 (C264Y and C497Y). None of the model proteins had a disulfide bond; however, the data suggest that substitution of the small hydrophobic sidechain of Cys by the large aromatic Tyr (which also results in burial of a hydroxyl group) is in itself sufficient to destabilize the model proteins TI I28 and TI I30 to the extent that they are unfolded (TI I28 is insoluble). Substitution by Cys of the conserved Tyr (Y784C) that forms the “tyrosine corner” ubiquitous in Ig-like proteins is highly destabilizing (39). This substitution involves loss of hydrogen bonding interactions within the E-F loop of the domain and packing interactions from the aromatic ring. Substitution of a surface Arg by Cys has little effect on stability (R226C in IL2Rγ).

Residue-specific Effects: Substitution of Glycine Residues—Gly is generally a highly conserved residue, it has no side chain and so has the particular ability to occupy a large region of Ramachandran φ,ψ space. All substitutions of Gly were destabilizing, in particular where the substituted polar residue (Ser, Asp, or Arg) is predicted to be buried in the mutant protein.

Relating Genotype to Phenotype—Since we have only a small sample of mutations in IL2Rγ, this discussion on correlating the biophysical effects of pathogenic mutations is limited to mutations of L1 only. Correlating a particular genotype to phenotype for many of the L1 mutations is complex. Most mutations that have been identified so far are present in only one family, and the clinical effects found within the same family can be varied (40–44). An example of this is the mutation I179S found in families that display all ranges of phenotypes, spastic paraplegia, MASA syndrome, and hydrocephalus (45). This variation may be explained by other factors unique to each individual, such as better chaperone systems, that allow a threshold amount of marginally stable protein to be correctly folded and trafficked in one patient and not in another. Another difficulty in correlating the genotype to the disease phenotype is lack of clinical data. If the phenotype is particularly severe a fetus may die in the earliest stages of development. Similarly, the less severe MASA syndrome is generally less well documented; spasticity may be attributed to a number of other conditions and defects in the LICAM gene may not be tested for.

Mutations That Have the Largest Effects on Protein Stability Are Related to More Severe Disease Phenotypes—One early study (41), however, attempted to divide known mutations into less severe (MASA) and more severe (HSAS) phenotypes. This study noted that a more severe phenotype correlates with mutations affecting buried positions more than those at surface positions. Interestingly, our mutations show a clear correlation between severity and effect on stability (Fig. 3). There is a “cut-off” between 1 and 2 kcal/mol. Above this cut-off all mutations are associated with severe phenotypes, below this milder phenotypes are reported. Consistent with this observation, mutation of residue Val-768 to Ile (mildly destabilizing) is reported to result in a milder form of disease (with some male carriers being apparently unaffected), while patients with the mutation V768F (moderately destabilizing) are reported to, in general, have the more severe phenotype (42, 46).

Mutations That Decrease Stability by More than 1.5 kcal/mol Are Associated with Poor Surface Expression—In a further study, a clear relationship was shown between site of mutation and cell surface expression of L1. For proteins with mutations affecting buried residues, the degree of surface expression was significantly reduced compared with wild type. In contrast, the majority of surface mutations were correctly trafficked to the surface (47). Again, using our data we are able to reveal that this relates to the effect a mutation has on protein stability (Fig. 4, A and B). Mutations that reduce the stability of the model proteins more than ~1.5 kcal/mol are all associated with reduced cell surface expression in both CHO cell and COS-7 in the previous studies. Apparently if a single domain in this multidomain protein has compromised stability the result is poor trafficking within the cell. Mutations preventing L1 expression at the cell surface are generally more likely to result in early mortality (44). Effect of Mutation on Ligand Binding—L1 is known to have both cell signaling and cell adhesion functions and regulates this through binding to various ligands. When the role of several missense mutations expressed in the context of full-length L1 protein in ligand binding were analyzed in vitro, all mutations affecting buried residues were found to decrease ligand binding, whereas the effect of mutations at surface positions varied based on the ligand analyzed (48). We are now in a position to compare these results with our data concerning the effect of mutation on protein stability (Fig. 4, C and D).

Nearly all mutations that destabilized L1 constituent domains by more than 2 kcal/mol were associated with proteins with severely decreased ligand binding even though the majority of these mutations affected buried residues. This is consistent with our observation that moderately and highly destabilizing mutations caused the less stable of our model proteins to be unfolded. This result suggests that having just one domain unfolded in the full-length protein is sufficient to disrupt the
binding activity, even when the mutated domain is not itself involved in binding. Proteins destabilized by less than 2 kcal/mol are associated both with proteins that have wild-type binding activity and proteins that have decreased binding activity. Since most of these proteins have mutations of surface or loop residues, this is consistent with the hypothesis that substitution in L1 of sites involved with specific interactions will be pathogenic (26).

**FIGURE 4. Comparison of the effects on stability in individual domains with trafficking and binding in L1 protein.** Mutant L1 proteins have previously been assayed for localization with a cell and in binding assays (47, 48). These published results are compared with data presented in this work. The data for the mutant of Arg-184 are shown by open circles (A, C, and D) and by an arrow (B) (see "Discussion").

A. Localization of mutant proteins within CHO cells, compared with the effect of the same mutation on stability in the model Ig and fnIII domains. Mutations that decrease the stability of the model proteins by >1.5 kcal/mol are all related to proteins that fail to be expressed on the protein surface. Where ΔΔG < 1.5 kcal/mol proteins are mostly trafficked with similar efficiency to wild type.

B. Localization of proteins in COS-7 cells. Proteins were defined as being located on the surface "as wild type" or as being moderately or completely retained within the cell. Only mutants with a ΔΔG < 1.5 kcal/mol are translocated as wild type.

C. Homophilic and heterophilic binding. The capacity of mutant proteins to establish homophilic binding or binding to TAX-1 has been determined (47, 48). Mutants with ΔΔG > 2 kcal/mol nearly all have severely reduced binding activity. However, mutants with ΔΔG < 2 kcal/mol (mainly surface and loop mutations) can be classified as binding like wild type or as having reduced binding.
**Modeling the Effects of Pathogenic Mutations**

**Comparison with Qualitative Predictions**—Until this study, mutations in L1 have been qualitatively described as “likely to affect binding” or “surface mutations” and “likely to affect stability” or “buried residue mutations” based on structural alignments. Here we have been able to quantify the predictions. In general, our results are in good agreement with these previous qualitative predictions (26). An interesting exception is for mutants of Arg-184 (replaced by Gln or Trp), which result in a severe phenotype. Bateman et al. (26) suggested that Arg-184 must be involved in a buried salt bridge at this position, and stated that this Arg was highly conserved. However, mutation of Arg-184 to Gln or Trp in TI I28 is only characterized as “mildly destabilizing” (1.5 and 2 kcal/mol respectively), consistent with its position in a loop and inconsistent with the hypothesis that it is involved in a buried salt bridge. Sequence analysis also suggests that this residue is not, in fact, highly conserved. Can the severity of disease be explained? One possibility is that the severity of this mutation is due principally to loss of important binding activity, and indeed both homophilic binding and Tax-1 binding are severely reduced in R184Q (Fig. 4, C and D). This might be because Arg-184 is directly involved in binding or because mutation of a partly buried Arg-184 may be important in maintaining structure of a loop that is involved in binding (26). Alternatively, mutation of Arg-184 may be more destabilizing in L1 than in TI I28 because it has a different structural role. (We note that we do not have a structure for the model protein or the L1 protein at this position.) Cell surface expression is reduced in the R184Q mutant (Fig. 4, A and B). We have suggested that there is an approximate 1.5 kcal/mol cut-off for the relationship between destabilization and poor cell surface expression; it would need very little difference in the effect of the mutation between L1 and the model protein to tip the L1 mutant protein over this line.

Our finding that any mutation that decreases the stability of a single domain by >2 kcal/mol results in severe disease is likely to be due to lowered levels of folded protein in the cell; even where the protein is marginally stable this may result in inefficient trafficking. This cut-off value of 2 kcal/mol is particularly interesting. Similar results have been observed for p53; mutations that lower the stability by more than 3 kcal/mol are found to cause the core domain to be unfolded at 37 °C (55). In the Ig-like protein superoxide dismutase, mutations that cause a loss of stability of >2 kcal/mol are associated with severe pathologies (10). Furthermore, in a recent extensive study by Yue et al. (49) mutations assigned to disease were shown typically to destabilize the folded state by 2–3 kcal/mol. Although most protein domains are said to have a stability of 2–8 kcal/mol, this is based on biophysical measurements made at 25 °C (for experimental convenience), usually under conditions (pH, salt) where the protein is maximally stable. At body temperature it is likely that many proteins are marginally stable. In p53, for instance, the stability of the core domain is lowered from 7.3 to 3.0 kcal/mol by raising the temperature from 25 to 37 °C. (ΔΔG values are unaffected by changes in temperature; Ref. 55.)

**Making Predictions for the Effect of a Mutation on the Stability of a Protein**—Our aim was to be able to predict the effect a mutation will have. The mutations we made could be approximately divided into four classes: those that had little or no effect on stability (ΔΔG < 0.5 kcal/mol) were mildly (>2 kcal/mol), moderately (>2–4 kcal/mol), and highly destabilizing (>4 kcal/mol). Below the 2 kcal/mol cut-off mutations could be expressed in all model protein systems and are likely to be associated with milder forms of disease (see above). Many of the highly destabilizing mutations (>4 kcal/mol) caused the model proteins to be unfolded, to aggregate, or resulted in failure of soluble expression.

**Residue Conservation as a Predictive Tool**—It has been previously noted that mutations of “conserved” residues are likely to be more pathogenic than mutation of non-conserved residues (50). This is likely to be because, apart from “active site residues” highly conserved for function, most conserved residues are important for structure and so are likely to be buried. A commonly used scale is sequence entropy, a measure of the variability at a given site (51). We compared sequence entropy with the observed effect of mutation on stability (Fig. 5A). Although in general the more conserved sites (those with lowest sequence entropy) are associated with the higher experimental ΔΔG measurements (r = 0.65), there is a large scatter in the plot, and it does not seem this relationship would give a good prediction of the effect a mutation might have on stability. A deeper analysis reveals why. This scale does not take account of what specific substitution is made. For instance, mutation of a buried hydrophobic residue is likely to be only mildly destabilizing if substituted by a hydrophobic residue of similar size but significantly destabilizing if replaced by a charged residue, yet in terms of sequence entropy both these substitutions would have the same score.

We have used a different scale for conservation (ΔΔGevolution), which takes account specifically what residue is mutated and what it is substituted to,

\[
\Delta \Delta G_{\text{evolution}} = \ln \frac{f^{\text{WT}}}{f^{\text{MUT}}} \tag{1}
\]

where \(f^{\text{WT}}\) and \(f^{\text{MUT}}\) are the frequencies of the wild-type and mutant residues at a given site in the protein. When we compare our results on the effect of a mutation with ΔΔGevolution, we find a remarkably good correlation (Fig. 5B, \(r = 0.85\)). We find that substitutions with a ΔΔGevolution of <2 all result in a low ΔΔGexperiment (<2 kcal/mol), whereas nearly all mutations associated with a ΔΔGevolution above 4 are highly destabilizing (>4 kcal/mol), note that the similarity between the numerical value of ΔΔGevolution and ΔΔGexperiment is purely coincidental! Thus the ΔΔGevolution scale, taking into account not simply variability at a site but also the nature of the substitution made, appears to be a good predictor of the effect the mutation will have on stability, at least in terms of whether the mutation will be mildly or severely destabilizing. (Note that the mildly destabilizing category includes those mutations that have no significant effect on stability.) Similar correlation has previously been shown for conservative mutations and in surveys of residues found at the N-caps of helices (52–54). However, our results clearly show that the correlation exists also for non-conservative mutations.

Such a prediction tool will depend on the quality of the alignments available and the number of sequences in these align-
Modeling the Effects of Pathogenic Mutations

In the calculation of $\Delta G_{\text{evolution}}$ we used multiple sequence alignments taken from the Pfam database. There are a large number of I-set and fnIII domain sequences in the Pfam database, with the average pairwise sequence identity being 24 and 20%, respectively. As the number of sequences decreases or if the sequences do not diverge significantly, $\Delta G_{\text{evolution}}$ may fail to predict the effect of mutation.

Making and Testing Predictions—Based on our analysis we revisited a number of L1 pathogenic mutants which we had not previously analyzed (Table 2). For each pathogenic mutation we made a prediction of the effect of mutation based on two methods. First we compared the mutation to others we had already made, considering position in the protein and the nature of the substitution. We then made a prediction of the effect of the mutation based on the “biophysical knowledge” and placed the substitution into one of our four suggested classes: no effect ($<0.5$ kcal/mol), mildly destabilizing ($<2$ kcal/mol), moderately destabilizing ($2-4$ kcal/mol), and highly destabilizing ($>4$ kcal/mol). Second, we determined the $\Delta G_{\text{evolution}}$ and made a prediction based on that. Most of the predictions were consistent whether we used our biophysical knowledge approach or our $\Delta G_{\text{evolution}}$ prediction method. There are four interesting exceptions (in italics in Table 2) resulting, we suggest, from artifacts thrown up in the evolutionary analysis. Two of these (Y194C and S674C) involve substitution of an exposed residue by cysteine. Since substitutions on the surface cause little destabilization we predict that they will only be mildly destabilizing using our biophysical approach. However, because proteins rarely have Cys residues on the surface, as they may cause inappropriate crosslinking in vivo, this is reflected in a low $f^{\text{MUT}}$ and thus in the relatively high $\Delta G_{\text{evolution}}$ at these sites. Thus we suggest the $\Delta G_{\text{evolution}}$ score is overestimating the effect this substitution will have on stability. Similarly, substitution of a partially exposed Arg by Cys in a loop region (R473C) results in a higher predicted $\Delta G_{\text{evolution}}$ than we would predict using our biophysical knowledge. The other exception involves placement of Pro at a surface site in a $\beta$-sheet (R632P). The $\Delta G_{\text{evolution}}$ in this case is artificially lowered because the site is variable, i.e. apart from Pro, many substitutions are acceptable, so that the frequency of the parent residue ($f^{\text{WT}}$) is low. This lowers the $\Delta G_{\text{evolution}}$ leading to an underestimate of the effect of the mutation.

Note that there are a number of positions where the structure is variable so that we neither had a good candidate for substitution in our model proteins nor did we have a reliable estimate for $\Delta G_{\text{evolution}}$. In some sites (labeled ND in Table 2) none of the sequences in the data base had the pathogenic residue at this site, so that the $\Delta G_{\text{evolution}}$ could not be determined. This suggests that these substitutions will be highly destabilising, consistent with our predictions based on our biophysical knowledge in all cases.

To test our predictions, we made four further mutant proteins, in positions where the pathogenic mutation had an identical or similar residue to the L1 protein. The results are given in Table 2 (in bold). In all cases our predictions were accurate (including our predictions that the $\Delta G_{\text{evolution}}$ scale would be...
### TABLE 2
Pathogenic mutations in L1 domains not included in the initial study

Predictions are based on the experimental results. Where the prediction from biophysical analysis and \( \Delta \Delta G_{\text{evolution}} \) differs is shown in italics. Where a model protein mutation was made to test the prediction is shown in bold. NRA, no reasonable alignment \( \Delta \Delta G_{\text{evolution}} \) could not be determined; ND, not done, no aligned sequence had pathogenic residue at this position.

| Mutant | Comments. | Prediction of change in stability based on biophysical analysis (mutation made, position in structure)\(^{(a)}\) | Prediction of change in stability based solely on \( \Delta \Delta G_{\text{evolution}} \) | Possible model substitutions Mutant made \( (\Delta \Delta G \text{ kcal/mol}) \) |
|--------|-----------|-------------------------------------------------|------------------------------------------|-------------------------------------------------|
| **L1 Ig domains** | | | | |
| A59V | Replacement of small buried hydrophobic by larger in B strand. Change of stereochemistry. Plot of change of size suggests destabilisation of \( \approx 2 \text{ kcal/mol} \). | \( \Delta \Delta G_{\text{evolution}} \) for A-V substitution \( \leq 0.5 \text{ kcal/mol} \) | Replace V65 in TK by L \( \Delta \Delta G = 1.6 \) |
| Y194C | Replacement of surface residue large hydrophobic by smaller in E (edge) strand. Suggests mild destabilisation \( \leq 2 \text{ kcal/mol} \). | \( \Delta \Delta G_{\text{evolution}} \) \( = 2.5 \) 2-4 kcal/mol | No equivalent |
| D202Y | In region of uncertain structure E-F loop. (Unlikely to be buried unless D is in salt bridge). Suggests will change loop structure, mildly destabilising. \( \leq 2 \text{ kcal/mol} \). | | NRA |
| I219T | In region of uncertain structure (F-G loop). Suggests will change loop structure, only mildly destabilising \( < 2 \text{ kcal/mol} \). | | NRA |
| R386C | In D-E loop region of uncertain structure. Note may put C on surface. Suggests little effect on stability \( < 0.5 \text{ kcal/mol} \). | | NRA |
| L391P | Proline substitution of buried hydrophobic in middle of E-strand. Suggests highly destabilising \( > 4 \text{ kcal/mol} \). | ND, Highly destabilising | L60 to P in I27 |
| A415P | Proline substitution of surface residue in G-strand. Suggests highly destabilising \( > 4 \text{ kcal/mol} \). | ND, Highly destabilising | A82 to P in I27 |
| V421D | Buried hydrophobic to charged residue in G-strand. Suggests highly destabilising \( > 4 \text{ kcal/mol} \). | \( \Delta \Delta G_{\text{evolution}} \) \( = 7.1 > 4 \text{ kcal/mol} \) | V132 to D in TK |
| A426D | Surface residue in A-strand. Suggests no significant effect on stability \( < 0.5 \text{ kcal/mol} \). | \( \Delta \Delta G_{\text{evolution}} \) \( = -0.5 \leq 1 \text{ kcal/mol} \) | No true equivalent – is close to Pro in L1 |
| R473C | Replacement of charged residue at start of D strand in “loop-like” region. | \( \Delta \Delta G_{\text{evolution}} \) \( = 6.5 > 4 \text{ kcal/mol} \) | I28 R47 to C |

\(^{(a)}\) Predictions are based on the experimental results.
### TABLE 2—CONTINUED

| Mutant  | Comments. | Prediction of change in stability based on biophysical analysis (mutation made, position in structure) \(^a\) | Prediction of change in stability based solely on \(\Delta\Delta G_{\text{evolution}}\) | Possible model substitutions Mutant made (\(\Delta\Delta G\) kcal/mol) |
|---------|-----------|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------|
| S542P   | Proline substitution of surface residue at end of B-strand / start of B-C loop. | \(\Delta\Delta G_{\text{evol}} = 3.5\) 2-4 kcal/mol | S26 to P in 127  
\(\Delta\Delta G = 3.5\) | |
| D598N   | F-G loop no change in stereochemistry.  Suggests no change/ v small change in \(\Delta G\). | | NRA | No equivalent |
| L1 fnIII domains | Proline substitution of surface residue in B-strand. | \(\Delta\Delta G_{\text{evol}} = 3.6\) 2-4 kcal/mol | L19 to P in  
TNfn3 / FNfn10  
Already done  
\(\Delta\Delta G > 4\) | |
| R632P   | Suggests highly destabilising (> 4 kcal/mol) | | | |
| S674C   | Surface replacement in E strand.  Note puts C on surface.  Suggests little effect on stability (< 0.5 kcal/mol). | \(\Delta\Delta G_{\text{evol}} = 3.6\) 2-4 kcal/mol | T56 to C in  
FNfn10  
\(\Delta\Delta G = -0.6\) | |
| A691D   | Buried charge in F strand.  Suggests highly destabilising (> 4 kcal/mol) | \(\Delta\Delta G_{\text{evol}} = 8.0\) > 4 kcal/mol | A74 to D in  
FNfn10 | |
| A691T   | As above but less destabilising (2-4 kcal/mol).  In a region that is hard to predict structure but in a GPG motif which suggests structural constraints. | \(\Delta\Delta G_{\text{evol}} = 2.7\) 2-4 kcal/mol | A74 to T in  
FNfn10 | |
| G698R   | Buried hydrophobic to polar (but not charged) residue in B-strand.  Suggests highly destabilising (> 4 kcal/mol). | | | |
| I733N   | In a BC loop.  Predict mildly destabilising (< 2 kcal/mol). | | | |
| M741T   | | | | |
| R751P   | Proline substitution of surface residue in C-strand.  Suggests highly destabilising (> 4 kcal/mol). | \(\Delta\Delta G_{\text{evol}} = 5.6\) > 4 kcal/mol | R33 to P in  
FNfn10 or  
E33 to P in  
TNfn3 | |
| V752M   | Buried residue in C strand.  Smaller to larger hydrophobic, change of stereochemistry.  Suggests moderately destabilising (2-4 kcal/mol). | \(\Delta\Delta G_{\text{evol}} = 3.4\) 2-4 kcal/mol | I34 to M in  
FNfn10 | |
| Y1070C  | In D-E loop, hard to predict, unknown structure.  Loop position suggests mildly destabilising (< 2 kcal/mol). | | | |

\(^a\) Definitions: no/small effect, <0.5 kcal/mol; mildly destabilizing, < 2 kcal/mol; moderately destabilizing, 2–4 kcal/mol; highly destabilizing, >4 kcal/mol.
Modeling the Effects of Pathogenic Mutations

inaccurate where surface Cys substitutions are involved). We also analyzed pathogenic mutations affecting another all β-domain, p53. Although the limited numbers of p53 sequences prevented the use of the ΔΔG_{evolution} approach, it was possible to correctly predict the effect of 12 out of 15 mutations using our biophysical knowledge (see supplemental data). This suggests that the strategy we propose is a powerful predictive tool.

CONCLUSIONS

This study reveals that model protein systems can be used to make accurate quantitative predictions of the effect that a pathogenic mutation can have on the stability of a protein which causes disease. These predictions can be based on both biophysical insights gained from the model protein systems or on a ΔΔG_{evolution} scale, which takes into account both the nature of the residue that is substituted as well as the nature of the substitution.

REFERENCES

1. Sunyaev, S., Ramensky, V., Koh, I., Lathe, W., III, Kondrashov, A. S., and Bork, P. (2001) *Hum. Mol. Genet.* 10, 591–597
2. Wang, Z., and Moult, J. (2001) *Hum. Mutat.* 17, 263–270
3. Ng, P. C., and Henikoff, S. (2003) *Nucleic Acids Res.* 31, 3812–3814
4. Chasman, D., and Adams, R. M. (2001) *J. Mol. Biol.* 307, 683–706
5. Guerois, R., Nielsen, J. E., and Serrano, L. (2002) *Biochemistry* 41, 6293–6302
6. Bateman, A., Coin, L., Durbin, R., and Finn, R. D. (2002) *Struct. Fold Des.* 7, 1145–1153
7. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Bioinformatics* 18, 140–147
8. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Proteins* 49, 223–230
9. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
10. Chou, K. W., and Fasman, G. D. (1978) *Biochemistry* 17, 5294–5303
11. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
12. Bateman, A., Coin, L., Durbin, R., and Finn, R. D. (2002) *Struct. Fold Des.* 7, 1145–1153
13. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Bioinformatics* 18, 140–147
14. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Proteins* 49, 223–230
15. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
16. Chou, K. W., and Fasman, G. D. (1978) *Biochemistry* 17, 5294–5303
17. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
18. Chou, K. W., and Fasman, G. D. (1978) *Biochemistry* 17, 5294–5303
19. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
20. Chou, K. W., and Fasman, G. D. (1978) *Biochemistry* 17, 5294–5303
21. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Bioinformatics* 18, 140–147
22. Park, D. H., Park, S. Y., Lee, J. W., and Chou, K. W. (2002) *Proteins* 49, 223–230
23. Gu, S. M., Orth, U., Zankl, M., Schroder, J., and Gal, A. (1997) *J. Mol. Biol.* 263, 683–706
24. Chou, K. W., and Fasman, G. D. (1978) *Biochemistry* 17, 5294–5303