SCIENTIFIC DATA

Data Descriptor: Dataset of eye disease-related proteins analyzed using the unfolding mutation screen

Caitlyn L. McCafferty[1] & Yuri V. Sergeev[1]

A number of genetic diseases are a result of missense mutations in protein structure. These mutations can lead to severe protein destabilization and misfolding. The unfolding mutation screen (UMS) is a computational method that calculates unfolding propensities for every possible missense mutation in a protein structure. The UMS validation demonstrated a good agreement with experimental and phenotypical data. 15 protein structures (a combination of homology models and crystal structures) were analyzed using UMS. The standard and clustered heat maps, and patterned protein structure from the analysis were stored in a UMS library. The library is currently composed of 15 protein structures from 14 inherited eye diseases including retina degenerations, glaucoma, and cataracts, and contains data for 181,110 mutations. The UMS protein library introduces 13 new human models of eye disease related proteins and is the first collection of the consistently calculated unfolding propensities, which could be used as a tool for the express analysis of novel mutations in clinical practice, next generation sequencing, and genotype-to-phenotype relationships in inherited eye disease.

Design Type	disease state design ● molecular modeling objective
Measurement Type(s)	protein unfolding
Technology Type(s)	computational modeling technique
Factor Type(s)	disease
Sample Characteristic(s)	Homo sapiens

[1]Ophthalmic Genetics and Visual Function Branch, National Eye Institute, NIH, Bethesda, Maryland 20892, USA. Correspondence and requests for materials should be addressed to Y.V.S. (email: sergeevy@nei.nih.gov).

Background & Summary

For globular proteins, primary protein structure dictates the folds and interactions that occur between amino acids in the structure[1,2]. Genetic mutations lead to protein misfolding and in many cases disease[3]. Protein secondary structure is stabilized by hydrogen bonds from the amide and carboxyl groups of amino acids[4]. The side chains of the amino acids interact in a variety of ways to create the protein tertiary structure (hydrophobic and disulfide interactions). In folding, the protein structure goes through a series of trails and errors to identify the most thermodynamically stable conformation[4]. Therefore, correctly folded proteins have long term stability in biological systems.

The role of missense mutations in inherited disease is not well understood. Disease-causing missense mutations occur when a change at the DNA level causes an amino acid in the protein sequence to be substituted with another, changing the interactions between amino acids and occasionally lead to protein misfolding. Currently, many inherited diseases are caused by missense mutations leading to misfolding of proteins in the cell[3–7].

UMS is an *in-silico* scan to evaluate the destabilizing effects of multiple point mutations derived from the protein atomic model. It may be used as a tool to analyze the complicated relationship between missense mutations, protein folding, and disease[5]. UMS reads a protein structure file (PDB file) and predicts the unfolding effect for a list of every possible missense mutation that may occur in the protein structure including an identity mutation. For each mutation, UMS calculates an unfolding propensity, derived from the Gibbs free energy equation, to describe whether the mutation will lead to protein misfolding. The output of UMS is a mutational matrix, standard heat map, clustered heat map, and patterned structure. UMS has the ability identify critical residues in the protein structure, which give insight into the most significant residues to protein stability and function. UMS also may explain mutations that can lead to both increased and decreased enzymatic activity, identify trends of residues relating to stability, and predict the severity of missense mutations in disease and their relation to disease phenotype.

Currently there exist a number of protein stability predictors[6–11]. There are also programs that work to predict the functional consequences of missense mutations[12–16]. UMS provides various benefits and advances over current mutant screening techniques. Given that UMS is derived from the atomic structure level and thermodynamics rather than sequence conservation, it has the ability to predict the effect of *de novo* missense mutations[17]. The unfolding propensity is determined using the linear extrapolation model from the normalized sigmoidal unfolding curve obtained experimentally[18]. This data classifies the effects of the missense mutations and uses a universal value so that unfolding propensities from different protein structures may be compared.

The 3 maps and mutation matrix are designed to make this large dataset readable for investigators with different backgrounds such as geneticists, clinicians, biochemists, pharmacologists or protein engineers, and those who may not have any preliminary experience in homology modeling and calculations of protein stability.

Residue depth has been used to describe the protein interior and predict fold types[19–21]. It has been hypothesized that the conservation of 'deep' residues is related to folding requirements and function[20]. Relationships exist between highly conserved residues in structural neighbors of the same fold type, and their mean residue depth in the reference structure[21]. There are programs that use residue depth as a parameter to predict protein structural models using fold recognition[19].

Here, we are reporting the library of 181,110 mutations from 15 proteins from inherited eye disease analyzed with UMS. This analysis includes the preparation of 13 homology models of human proteins. The UMS program has been subjected to intensive validation using the Protherm database and 3 proteins from retinal disease (rhodopsin, complement factor H, and RPE65)[5]. We present 10 new homology models for human proteins related to retinal diseases that have been verified using the internal control. In addition, we provide new maps for each of the 15 proteins for prediction and express analysis of missense mutations. Finally, this study targets a number of new diseases that have not yet been studied using UMS.

Methods

Protein preparation

A library of 15 different inherited eye disease related proteins was created for analysis. Figure 1 demonstrates the outline of the stages of analysis used. The proteins included in the dataset, pdb names, and their corresponding diseases are shown in Table 1. The human proteins were taken from the RCSB database[22] or prepared using homology modeling. CYP1B1, IRBP, LRAT, NYX, RDH5, RDH8, RDH12, REP-1, RHO, RPE65, TIMP3, WDR36, and domains 4, 5, 14 and 17 from CFH are homology models. While the remaining crystal structures of CFH domains, CRYAB, and CRYBB1 were obtained from the protein data bank[22]. The pdb files used for the protein analysis are available on the server.

Internal control

After the homology models were created they were run through the internal control program. The internal control program for the analysis of unfolding propensities is explained in depth in McCafferty & Sergeev[5]. In this work the internal control was adjusted from UMS to calculate the difference in the free energy of the side chain rotamers for the same amino acid. The output of the internal control program was used to either select the best protein models or determine if more refinement of the structures was
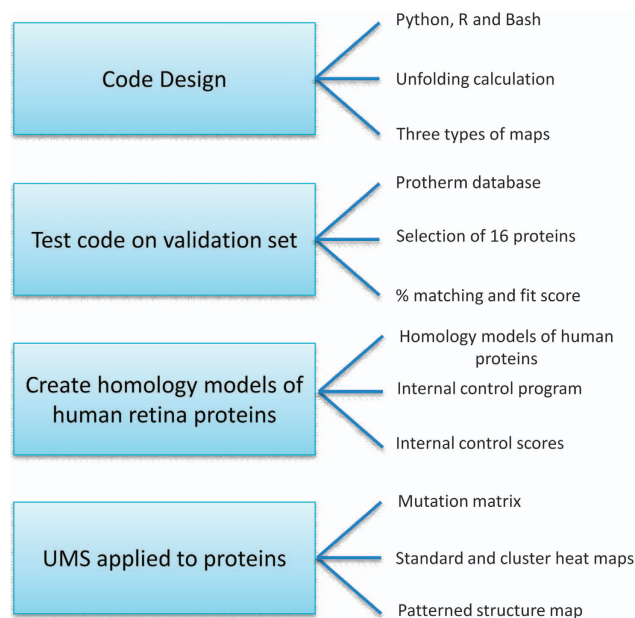
Figure 1. The schematic illustrates the design and testing process used in developing the dataset. The first step was designing the program. The program was written in Python, R and Bash. The purpose of the program was to calculate the unfolding propensity for each missense mutation in a proteins structure. Following, an effective display of the data was pivotal. Next the program was validates using data for 16 proteins obtained from the Protherm database. After the validation human models of proteins related to inherited eye disease were built using homology model. The quality of the structures was tested using the internal control program. Finally UMS was applied to these proteins to produce the mutation matrix, standard and clustered heat maps, and the patterned structure.

Protein	PDB	Mutations Analyzed	Disease	Standard HM Size	Cluster HM Size	Patterned Structure Size
CFH	CFH.pdb	24,320	Age-related Macula Degeneration (AMD)	613 kB	780 kB	1.9 MB
RHO	RHO.pdb	13,920	Retinitis Pigmentosa (RP)	426 kB	477 kB	1.2 MB
TIMP3	TIMP3.pdb	3,940	Sorsby's fundus distrophy	401 kB	430 kB	334 MB
REP-1	REP-1.pdb	12.54	Choroideremia	487 kB	576 kB	559 kB
LRAT	LRAT.pdb	7,320	Leber's Congenital Amaurosis (LCA)	424 kB	478 kB	599 kB
RDH5	RDH5.pdb	12,700	Fundus Albipunctatus	464 kB	556 kB	1.1 MB
RDH8	RDH8.pdb	11,880	Myopia, AMD, LCA	453 kB	539 kB	967 kB
RDH12	RDH12.pdb	6,000	LCA, RP, Cone-Rod Dystrophy (CRD)	408 kB	452 kB	499 kB
RPE65	RPE65.pdb	21,320	LCA	544.7 kB	702 kB	2.1 MB
IRBP (domain 1)	IRBP-1.pdb	25,260	arRP	431 kB	477 kB	522 kB
CYP1B1	CYP1B1.pdb	10,160	Glaucoma	470 kB	541 kB	850 kB
CRYBB1	1oki.pdb	7,340	Cataracts	429 kB	481 kB	506 kB
CRYAB	3l1g.pdb	1,900	Cataracts	387 kB	402 kB	322 kB
NYX	NYX.pdb	9,620	Congenital stationary night blindness	463 kB	532 kB	772 kB
WDR36	WDR36.pdb	14,600	Primary Open Angle Glaucoma (POAG)	514 kB	617 kB	1.4 MB

Table 1. UMS output for each protein. In the table each protein is listed with the PDB file name, the mutations analyzed the inherited disease it is associated with, and the size of each of the map files. The files for other 3 IRBP domains were similar in size (not shown in the Table). In total, 181,110 mutations were analyzed for the 15 proteins.
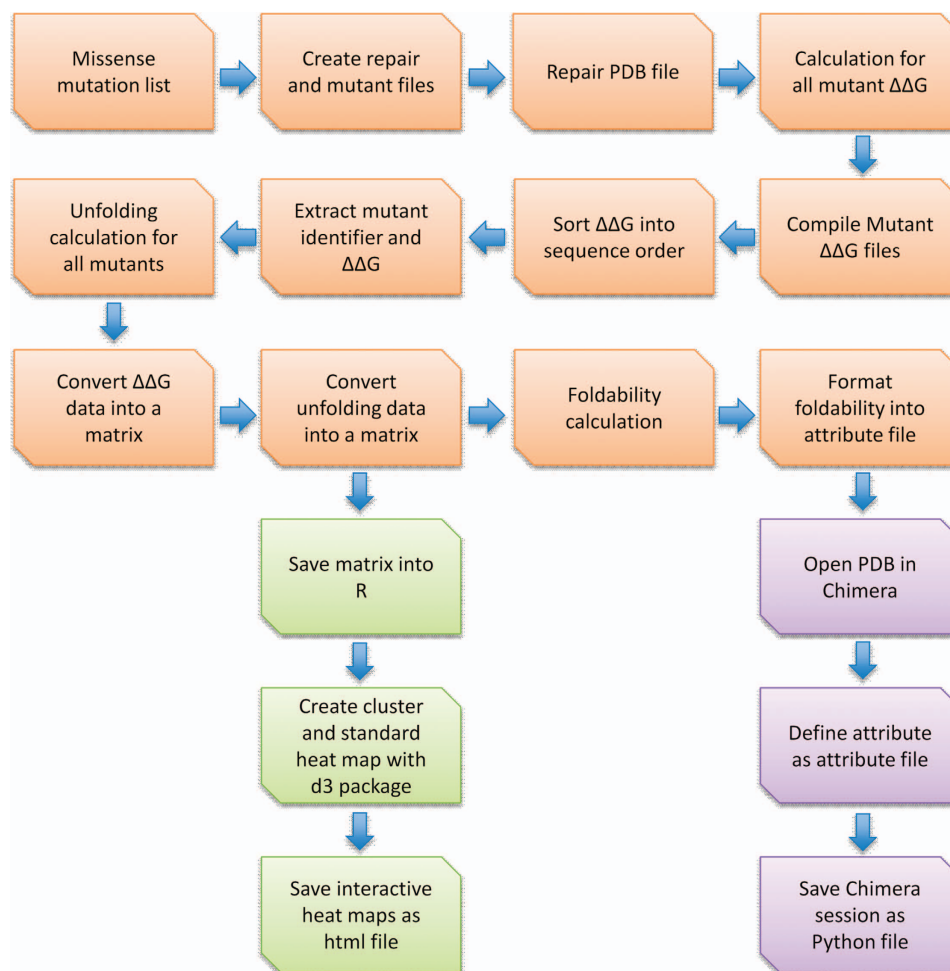
Figure 2. **The workflow diagram describes the process of the UMS program.** Each function written in the program to carry out the calculations is described. The orange squares represent the operations performed in python, the green squares in R, and the purple in Chimera.

required. In selecting the best structure, we looked for models with statistically significant data (*P* value < 0.05) and then looked for the smallest confidence interval with average close to 0. For those that required further refinement, structures were tested until they fit this requirement.

UMS program
A full description of the UMS program can be found in the methods section of McCafferty & Sergeev[5]. In summary, the program is written in Python, R, and Bash programing languages. The architecture of the program is designed to perform a full mutagenesis analysis as efficiently as possible by implementing a quick and space sensitive procedure. Figure 2 outlines the order of the functions created within the program. The unfolding propensity calculation is derived from the Gibbs free energy equation. The standard and clustered heat maps are produced using the d3heatmap package for R. The maps are interactive and allow specific rows or columns of interest to be selected. Specific region may also be selective for zoomed in view. The clustered maps use an agglomerative, hierarchical clustering method. The groupings are then mapped using a dendrogram. The final map used to convey UMS is the patterned foldability structure. The 3D structures were colored using the foldability value of the residues[5]. These foldability values can then be used to identify critical residues in the protein structure. The critical residues are considered to be essential to proper protein folding.

Residue depth and informational entropy
In addition, two new descriptors, residue depth and informational entropy, were included as described below. First descriptor, the residue depth of an amino acid in the protein structure, is described as the distance of an atom from the solvent accessible surface[23]. The Biopython package was used to calculate the residue depth for each of the wild type residues in the protein structure[24]. The Biopython package uses the MSMS program for the surface calculation[25], the residue depth is then presented as the average
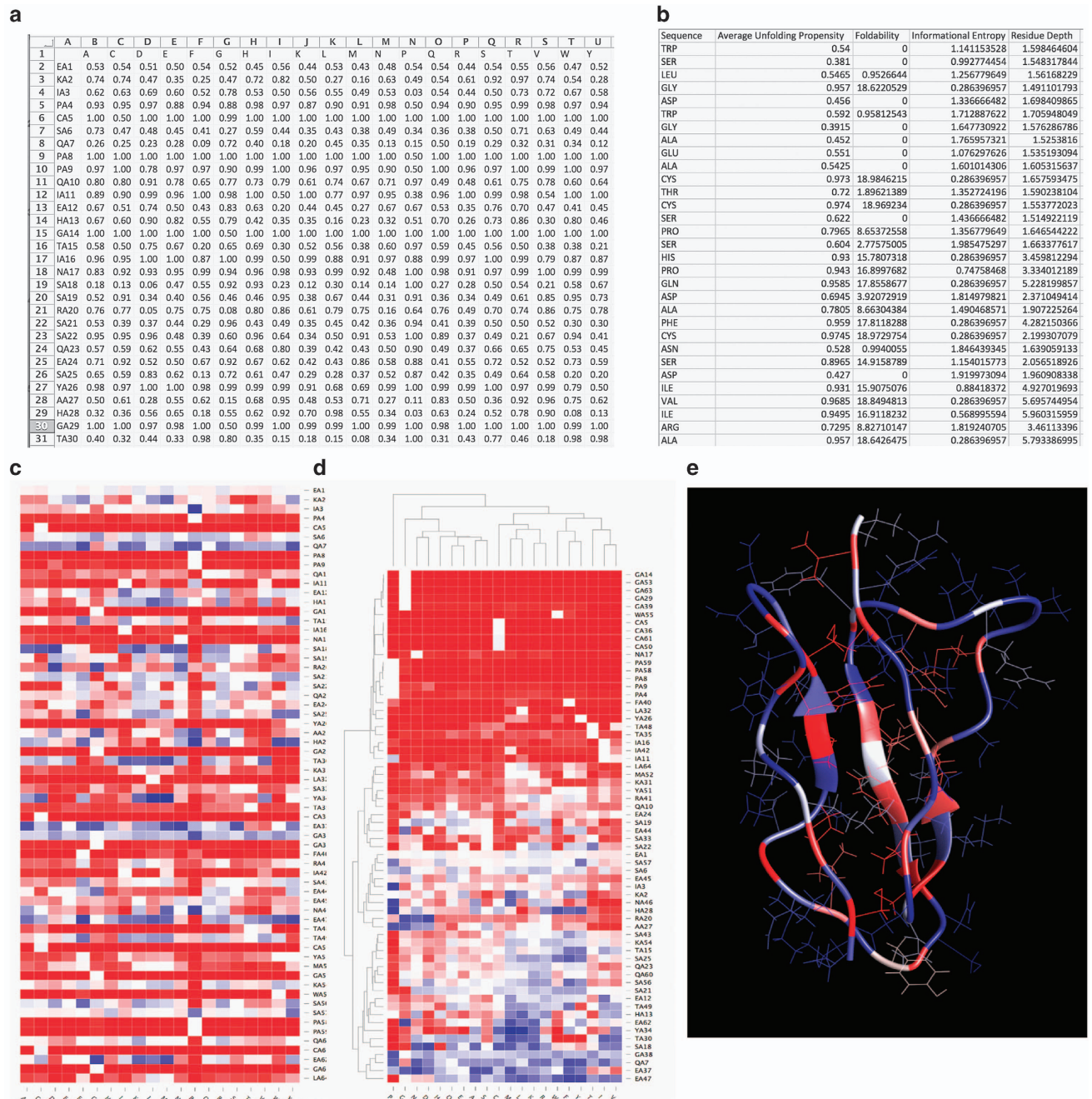
Figure 3. The fragments of data types included for each protein. (**a**) The mutation matrix. The file is in the text format and can be opened in Excel or a text editor. The wild type residues are along the Y-axis and the mutations across the X-axis with the unfolding propensities listed in each box. (**b**) The set of data descriptors for the protein. The first column denotes the wild type sequence, second column the average unfolding propensity at the location, third the foldability, fourth the information entropy, fifth the residue depth. (**c**) The standard heat map. Again, the wild type residues are along the Y-axis and the mutations across the X-axis. The heat map is html file format and is interactive (specific rows/columns/regions may be selected) the red represents the most severe mutations while the blue is the most stabilized. (**d**) The cluster heat map. The file is also html format; here the data is grouped according to an agglomerative hierarchical clustering technique. (**e**) The patterned structure. The structure is a python file that is opened in Chimera. The structure is colored according the foldability values of the residues to highlight the critical residues in the structure.
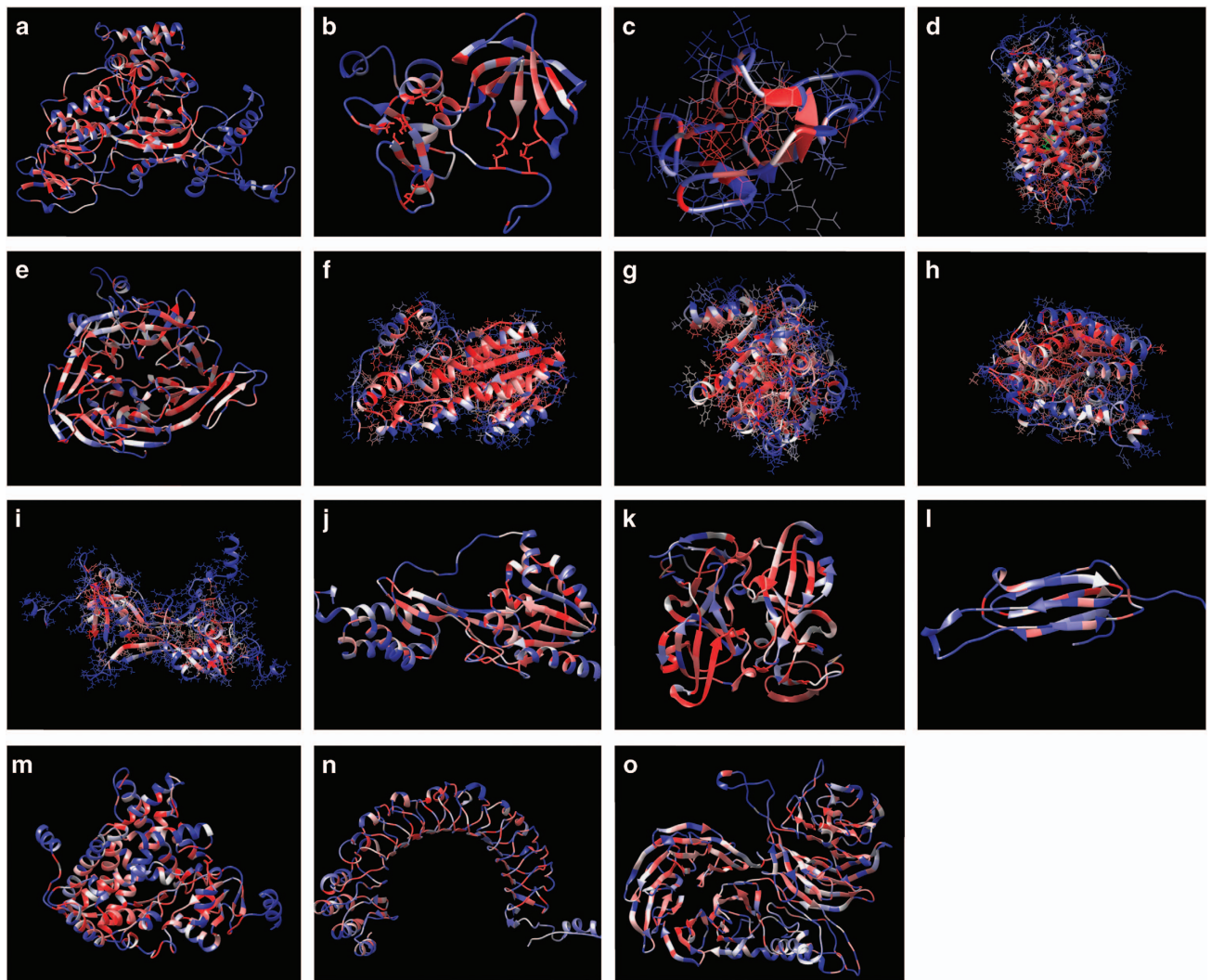
Figure 4. **The protein ribbon structures colored by predicted foldability.** The structures of proteins related to inherited eye disease. (**a**) REP-1, (**b**) TIMP3, (**c**) sushi domain of CFH, (**d**) rhodopsin, (**e**) RPE65, (**f**) RDH8, (**g**) RDH5, (**h**) RDH12, (**i**) LRAT, (**j**) IRBP, (**k**) CRYBB1, (**l**) CRYAB, (**m**) CRY1B1, (**n**) NYX, (**o**) WDR36.

of the atom depth for each wild type residue in the native protein sequence. The other descriptor, informational entropy, also known as Shannon entropy, quantifies the uncertainty of the source of the information. Therefore, greater informational entropy relates to a greater degree of randomness amongst the mutations for a certain location[26]. For example, if the average unfolding propensity for two locations on the structure were both 0.5, the location where all unfolding propensities were 0.5 would have a lower informational entropy that those that were split between 1.0 and 0. A script was created using Python to calculate the informational entropy of the unfolding propensities. The equation of informational entropy used was $\sum P(R = x) \times \log_2\left(\frac{1}{P(R=x)}\right)$. Both parameters were added to the library to aid the user in studying the relationship between folding and depth within the structure and in analyzing the data provided by UMS, respectively.

Code availability
The code is available on the *Figshare* (Data citation 1).

Data Records
The UMS library for 15 proteins from inherited eye disease is available on the *Figshare* (Data citation 1). Table 1 presents the PDB file names for each of the proteins included in the study. For each of the proteins analyzed there are five separate files available to describe the data. Figure 3 displays examples of what each of these files looks like. The first is the mutation matrix.

The mutation matrix is available in the protein_matrix.txt format. This can be opened using a standard text editor as well as in Microsoft Excel. The mutation matrix is ideal for an investigator who
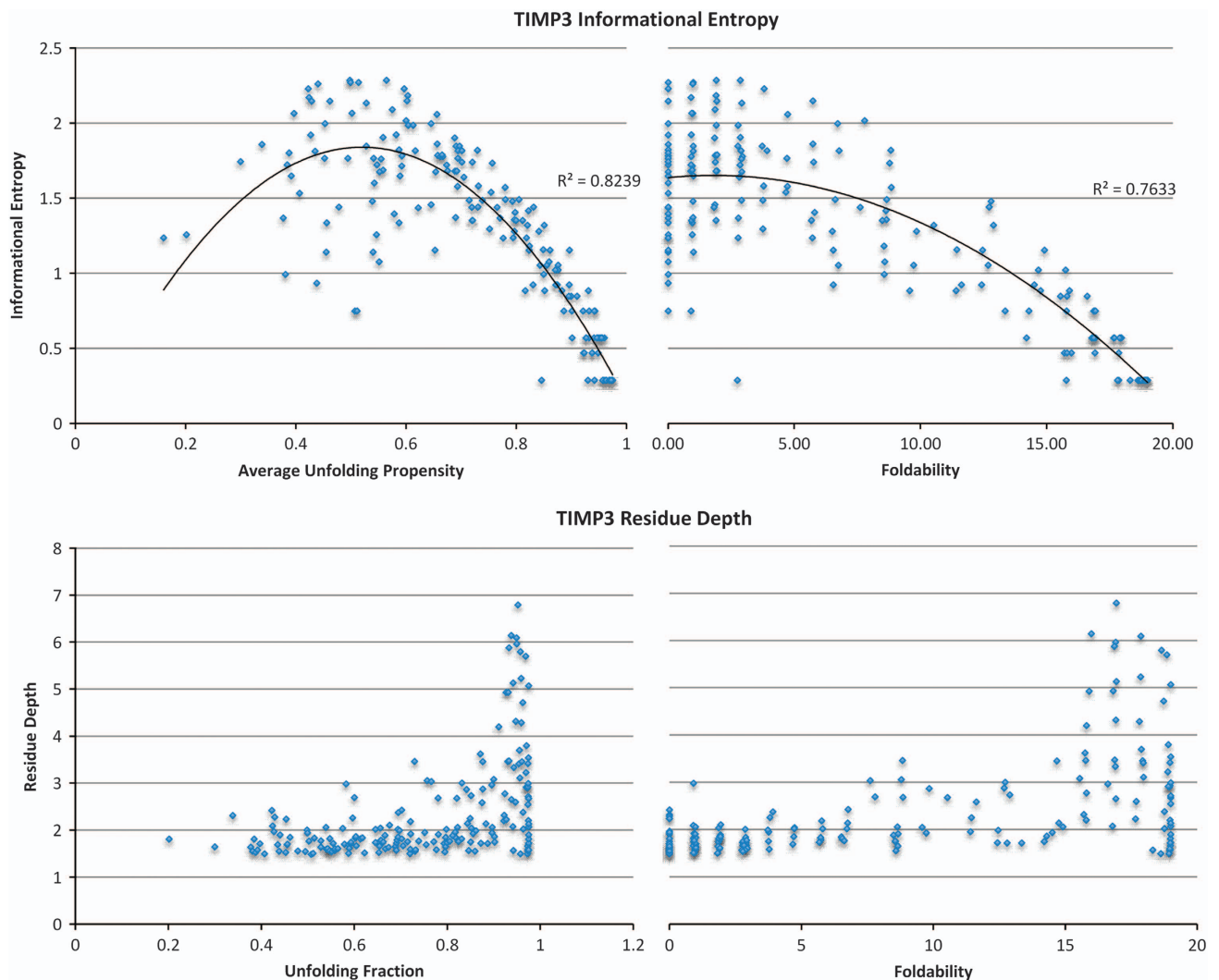
Figure 5. An example of the potential uses for the protein data descriptors. TIMP3 average unfolding propensities and foldability are plotted against the information entropy and residue depth for each wild type location.

wants access to the raw data. Since the file can be opened in a number of programs the user has the ability to analyze and manipulate the data however he/she pleases.

The standard and cluster heat maps are available in the protein_standard.html protein_cluster.html format, respectively. The size of each of these maps for the proteins is shown in Table 1. The interactive maps provide facilitated identification of the unfolding propensities even if the protein being studied is large. The html file format allows for the maps to open as a webpage. From here rows and/or columns of interest may be selected. Specific regions may also be highlighted for a zoomed in view. Dragging the mouse over the mutation of interest will reveal the corresponding unfolding propensity. The unfolding propensity ranges from 0–1, where 0 is the most thermodynamically stabile protein, 0.5 is folding-unfolding equilibrium, and 1 is a completely unfolded protein. The standard heat maps are ideal for situations where a specific mutation-unfolding propensity is desired, for example, a clinical setting. Here, each unfolding propensity can be accessed easily. The map can be downloaded and saved for easy reference for a specific patient mutation. The clustered heat map groups residues based on similarity and may be used in studying trends in the protein structure. For example, in structures with disulfide bonds, cysteine residues are clustered together based on the similar destabilization they undergo. This allows us to see the residues that undergo a number of severe mutations or the mutations that have the most harmful effects. A pharmacologist can use this map to identify stabilizing mutations to develop new drugs.

The patterned structure is available as a python file, protein.py. This file is to be opened using UCSF Chimera (http://www.cgl.ucsf.edu/chimera/). Once opened in Chimera, residues may be identified by placing the mouse over the area of interest. Table 1 also displays the size of the python files to be read by

CFH Domain	Avg ± Stdev	P-value	C.I.
S1	0.056 ± 0.12	4.53e-4	(-0.294, 0.181)
S2	-0.050 ± 0.11	7.59e-4	(-0.261, 0.160)
S3	-0.030 ± 0.07	1.91e-3	(-0.172, 0.113)
S4	-0.030 ± 0.04	5.17e-6	(-0.116, 0.057)
S5	-0.026 ± 0.05	1.12e-4	(-0.118, 0.067)
S6	-0.040 ± 0.10	1.40e-3	(-0.229, 0.149)
S7	-0.036 ± 0.09	2.88e-3	(-0.210, 0.138)
S8	-0.031 ± 0.06	2.80e-4	(-0.155, 0.092
S9	-0.024 ± 0.04	7.05e-6	(-0.096, 0.049)
S10	-0.053 ± 0.12	8.84e-4	(-0.294, 0.188)
S11	-0.051 ± 0.10	1.80e-4	(-0.241, 0.140)
S12	-0.044 ± 0.11	2.44e-3	(-0.252, 0.163)
S13	-0.036 ± 0.10	7.59e-3	(-0.226, 0.154)
S14	-0.033 ± 0.05	8.02e-6	(-0.131, 0.066)
S15	-0.044 ± 0.16	0.0350	(-0.358, 0.271)
S16	-0.047 ± 0.08	7.50e-6	(-0.206, 0.112)
S17	-0.043 ± 0.08	1.02e-4	(-0.195, 0.110)
S18	-0.047 ± 0.06	2.08e-6	(-0.174, 0.080)
S19	-0.017 ± 0.05	0.0153	(-0.114, 0.081)
S20	-0.035 ± 0.07	2.21e-4	(-0.165, 0.096)

Protein	Avg ± Stdev	P-value	C.I.
CFH			
RHO	-0.047± 0.14	7.69e-10	(-0.319, 0.225)
TIMP3	-0.033± 0.07	7.38e-10	(-0.172, 0.107)
REP-1	-0.018± 0.07	3.90e-10	(-0.161, 0.124)
LRAT	-0.043± 0.14	5.85e-9	(-0.311, 0.225)
RDH5	-0.043± 0.12	2.16e-18	(-0.280, 0.193)
RDH8	-0.050± 0.12	3.30e-21	(-0.292, 0.193)
RDH12	-0.043± 0.08	2.86e-19	(-0.193, 0.107)
RPE65	-0.033± 0.21	7.98e-7	(-0.452, 0.387)
IRBP			
CYP1B1	-0.054± 0.14	2.70e-16	(-0.336, 0.228)
CRYBB1	-0.053±0.13	2.72e-14	(-0.303, 0.197)
CRYAB	-0.043± 0.11	2.05e-4	(-0.256, 0.170)
NYX	-0.056± 0.19	3.7e-10	(-0.430, 0.319)
hWDR36	-0.048±0.13	2.03e-22	(-0.303, 0.206)
RS1	-0.058±0.14	1.43e-50	(-0.323, 0.208)

CFH Domain	Avg ± Stdev	P-value	C.I.
D1	-0.339±0.07	6.51e-16	(-0.172, 0.105)
D2	-0.032 ± 0.11	1.78e-7	(-0.248, 0.183)
D3	-0.040 ± 0.15	2.61e-6	(-0.334, 0.253)
D4	-0.033 ± 0.07	5.94e-16	(-0.166, 0.010)

Figure 6. Internal control data. The internal control mutates each residue to itself in order to test the stability of the structure. It is expected that each of these mutations will have a $\Delta\Delta G$ of 0. For each protein the average $\Delta\Delta G$ for all of the mutations is recorded as well as the standard deviation. The p-value is also calculated to check to statistical relevance and finally a 95% confidence interval to show the range of the $\Delta\Delta G$ values.

Chimera. This is a 3D map that shows the most critical residues in the protein structure to proper folding. For the particular residue position, foldability could then be used to differentiate between areas that experienced multiple severe mutations, those that experienced a few, and those that had none. Foldability is a more descriptive parameter over simply finding the average in that it can successfully tally all severe mutations that are occurring at a certain location without being influenced by those less severe. Figure 4 shows all of the patterned structures of the inherited eye disease related proteins that were analyzed.

Finally, the data descriptor file is again available in the protein_descriptor.txt format. In each column this file contains the native protein sequence, average unfolding propensities, foldability, informational entropy and residues depth (in this order). The descriptor file provides the user some innovation to analyze the data as he/she pleases. Figure 5 shows an example of how one may use this file to analyze the TIMP3 protein. The average unfolding propensity and foldability are plotted against the informational entropy and residue depth.

Technical Validation
Validation set criterion (UMS reference)
The validation set for the UMS program was composed of 16 proteins. The proteins were selected from the ProTherm database (http://www.abren.net/protherm/)[27] based on available experimental thermodynamic data. Proteins with single mutations whose $\Delta\Delta G$ values were determined using fluorescence from denaturants and CD were selected. Specifically, tryptophan fluorescence data for chemical unfolding/refolding in the presence of urea or Gdm-HCl. Proteins with a large number of mutations with thermodynamic data were ideal for the validation. Finally, the proteins needed to have an available PDB file on the Protein Data Bank (http://www.rcsb.org/pdb/)[22]. Based on this criteria the 16 proteins selected for the validation set were: T4 Lysozyme (PDB id: 2LZM), Staphylococcal Nuclease (1STN), Protein L (1HZ6), Barnase (1BNI), Ribonuclease T1 Isozyme (1RN1), Gene V Protein (1VBQ), Chymotrypsin Inhibitor 2 (2CI2), Acyl-Coenzyme A (2ABD), Tyrosine-Protein Kinase (1FMK), Acylphosphatase

(1APS), Alpha Spectrin (1AJ3), Dihydrofolate Reductase (1RK4), Ribosomal Protein (1RIS), Tryptophan synthase (1WQ5), ARC Repressor (1ARR), and Azurin (5AZU). From the $\Delta\Delta G$ values for each of the experimental mutants from the Protherm database the unfolding propensity was calculated. The percent matching and a Fit-Score were used to evaluate the quality of the output from UMS.

Homology model validation
As mentioned in the Methods section an internal control program was designed to validate the homology models used. The results from the internal control are shown in Fig. 6. CFH and IRBP were divided in to their 20 and 4 domains (respectively) in the analysis. The data for each of the proteins fit our criteria for being statistically significant and having small confidence intervals with averages close to 0.

Usage Notes
All of the protein structures that are included in the dataset are human structures. The homology models, while not crystal structures, have been thoroughly tested for stability and represent models of the human proteins. We aim to eventually create a website of proteins that will constantly be updated and take requests for proteins of interest.

References

1. Eisenhaber, F., Persson, B. & Argos, P. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Critical reviews in biochemistry and molecular biology* **30,** 1–94 (1995).
2. Baker, D. A surprising simplicity to protein folding. *Nature* **405,** 39–42 (2000).
3. Bross, P. *et al.* Protein misfolding and degradation in genetic diseases. *Human mutation* **14,** 186–198 (1999).
4. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884–890 (2003).
5. McCafferty, C. L. & Sergeev, Y. V. *In silico* Mapping of Protein Unfolding Mutations for Inherited Disease. *Scientific Reports* **6,** 37298 (2016).
6. Parthiban, V., Gromiha, M. M. & Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic acids research* **34,** W239–W242 (2006).
7. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant 2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research* **33,** W306–W310 (2005).
8. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic acids research* **33,** W382–W388 (2005).
9. Dosztanyi, Z., Magyar, C., Tusnády, G. & Simon, I. SCide: identification of stabilization centers in proteins. *Bioinformatics* **19,** 899–900 (2003).
10. Magyar, C., Gromiha, M. M., Pujadas, G., Tusnady, G. E. & Simon, I. SRide: a server for identifying stabilizing residues in proteins. *Nucleic acids research* **33,** W303–W305 (2005).
11. Kurgan, L., Cios, K. & Chen, K. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC bioinformatics* **9,** 226 (2008).
12. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1,000 mutations. *Journal of molecular biology* **320,** 369–387 (2002).
13. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7,** 248–249 (2010).
14. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31,** 2745–2747 (2015).
15. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome research* **13,** 2129–2141 (2003).
16. De Baets, G. *et al.* SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Research* **40,** D935–D939 (2011).
17. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences* **112,** E5189–E5198 (2015).
18. Myers, J. K., Nick Pace, C. & Martin Scholtz, J. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Science* **4,** 2138–2148 (1995).
19. Liu, S., Zhang, C., Liang, S. & Zhou, Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins: Structure, Function, and Bioinformatics* **68,** 636–645 (2007).
20. Pintar, A., Carugo, O. & Pongor, S. Atom depth in protein structure and function. *Trends in biochemical sciences* **28,** 593–597 (2003).
21. Pintar, A., Carugo, O. & Pongor, S. Atom depth as a descriptor of the protein interior. *Biophysical journal* **84,** 2553–2561 (2003).
22. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28,** 235–242 (2000).
23. Chakravarty, S. & Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **7,** 723–732 (1999).
24. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
25. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38,** 305–320 (1996).
26. Coifman, R. R. & Wickerhauser, M. V. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on* **38,** 713–718 (1992).
27. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic acids research* **32,** D120–D121 (2004).

Data Citation
1. McCafferty, C. L. & Sergeev, Y. V. *Figshare* https://dx.doi.org/10.6084/m9.figshare.c.3291326 (2016).

Additional Information

How to cite this article: McCafferty, C. L. and Sergeev Y. V. Dataset of eye disease-related proteins analyzed using the unfolding mutation screen. *Sci. Data* 3:160112 doi: 10.1038/sdata.2016.112 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.