Performance of Protein Disorder Prediction Programs on Amino Acid Substitutions

Ali, Heidi; Urolagin, Siddhaling; Gurarslan, Omer; Vihinen, Mauno

Published in:
Human Mutation

DOI:
10.1002/humu.22564

2014

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):
Ali, H., Urolagin, S., Gurarslan, O., & Vihinen, M. (2014). Performance of Protein Disorder Prediction Programs on Amino Acid Substitutions. *Human Mutation*, *35*(7), 794-804. https://doi.org/10.1002/humu.22564

Total number of authors:
4

Performance of protein disorder prediction programs on amino acid substitutions

Heidi Ali[1], Siddhaling Urolagin[2], Ömer Gurarslan[1], and Mauno Vihinen[1,2,3]

[1]Institute of Biomedical Technology, FI-33014 University of Tampere, Finland, and BioMediTech, Tampere, Finland

[2]Department of Experimental Medical Science, Lund University, SE-22184 Lund, Sweden

[3]Tampere University Hospital, Tampere, Finland

Corresponding author: Professor Mauno Vihinen, Department of Experimental Medical Science, Lund University, BMC D10, SE-22184 Lund, Sweden, mauno.vihinen@med.lu.se, tel +46 72 5260022, fax +46 46 2220296

ABSTRACT

Many proteins contain intrinsically disordered regions, which may be crucial for function, but on the other hand be related to the pathogenicity of variants. Prediction programs have been developed to detect disordered regions from sequences and used to predict the consequences of variants, although, their performance for this task has not been assessed. We tested the performance of protein disorder prediction programs in detecting changes to disorder caused by amino acid substitutions. We assessed the quality of 29 protein disorder predictors and versions with 101 amino acid substitutions, whose effects have been experimentally validated. Disorder predictors detected the true positives at most with 6% success rate and true negatives with 34% rate for variants. The corresponding rates for the wild type forms are 7 and 90%. The analysis revealed that disorder programs cannot reliably predict the effects of substitutions, consequently the tested methods, and possibly similar programs, cannot be recommended for variant analysis without other information indicating to the relevance of disorder. These results inspired us to develop a new method, PON-Diso (http://structure.bmc.lu.se/PON-Diso), for disorder related amino acid substitutions. With 50 % success rate for independent test set and 70.5% rate in cross validation it outperforms the evaluated methods.

KEYWORDS: protein disorder, method evaluation, bioinformatics, protein disorder prediction, disorder prediction program, protein flexibility, amino acid substitution, disease-causing variants

Introduction

Proteins usually fold to their distinct three-dimensional structures. Exceptions to this are intrinsically disordered proteins (IDPs) and regions, which do not have stable secondary and/or tertiary structure under physiological conditions. These proteins have several essential functions, which link with their disordered structural state. Protein disorder is conserved in evolution and eukaryotic proteins contain more disordered regions than prokaryotic ones [Schlessinger et al., 2011]. Ordered and disordered regions have distinctive sequence patterns, both of which differ from those for random sequences. Thus, protein disorder is a distinct structural state, not just an extremely flexible form of regular structures [Schlessinger et al., 2011]. Disordered regions contain less hydrophobic amino acids than ordered regions [Chouard, 2011] and many such regions have an increased aggregation tendency, because hydrophilic residues can cluster together leading to the formation of aggregates. To avoid protein aggregation, natural selection has favored certain patterns in disordered regions.

Disordered regions appear in the functional sites of proteins [Dunker et al., 2001; Ward et al., 2004], for example, in proteins involved in transcription, translation, cell signaling, alternative splicing, signal transduction, and molecular recognition of partner molecules [Dunker et al., 2002; Dyson and Wright, 2002; Liu et al., 2002; Ward et al., 2004; Uversky et al., 2005; Romero et al., 2006].

Disordered regions are very flexible but may not always be disordered. Ligand binding, for instance, requires ordered structure [Sugase et al., 2007]. However, in some cases, the structure remains disordered while binding [Mittag et al., 2010].Because of their inherent flexibility, disordered segments can interact with large numbers, even hundreds, of partners [Oldfield et al., 2008], and affect protein properties such as stability [Vihinen, 1987].

Disorder prediction methods are based on known instances, their locations in the protein sequences, and physicochemical properties of amino acids. Disorder prediction programs are mainly based on machine learning approaches such as neural networks, for instance, DisEMBL, DISpro, Regional Order Neural Network (RONN), and support vector machines (SVMs), such as DISOPRED2, POODLE-L, POODLE-S, and Spritz.

Flexibility (and the opposite rigidity) determines the possibility of a motion in a molecule or part of it [Gohlke and Thorpe, 2006]. Knowledge about flexibility or rigidity can, for example, simplify the task of modeling protein dynamics. The difference between disorder and flexibility is that disordered regions miss partially or completely secondary or tertiary structure, whereas flexible regions have high mobility visible in large B-factor values in crystal structures rather than missing structure. Disordered proteins and regions are highly flexible [Teilum et al., 2009], and disordered proteins generally contain more flexible amino acids [Radivojac et al., 2004] as defined by an amino acid propensity scale [Vihinen et al., 1994]. Flexible residues differ from regular and rigid ones in their tendency to form secondary structures, be solvent accessible, and by having different amino acid distributions [Schlessinger and Rost, 2005]. D^2 (disorder in disorders) concept signifies that protein disorder is highly abundant in many human diseases [Uversky et al., 2008]. Variants, including amino acid substitutions, alter the degree of protein disorder [Zhang et al., 1995; Guy et al., 2008]. These changes may relate to pathogenicity [Thusberg and Vihinen, 2009] due to having impact, for example, on protein stability [Fisher and Stultz, 2011], aggregation [Hartl and Hayer-Hartl, 2009], and hydropathy [Williams et al., 2001].

Recently,Thusberg et al. (2011) assessed the performance of variation effect prediction programs for tolerance, and Khan and Vihinen (2010) and Potapov et al. (2009) for stability. Deiana and Giasanti (2010) studied the performance of disorder prediction programs in the recognition and disease association of the disordered region, but nobody has investigated their effect on amino acid substitutions.

We and others have applied disorder prediction programs to analyze the disease mechanisms of single amino acid substitutions [Thusberg and Vihinen, 2006; Thusberg and Vihinen, 2007; Radivojac et al., 2008; Li et al., 2009; Mort et al., 2010]. Hu et al. (2011) and Vacic and Iakoucheva (2012) have speculated about the disorder relevance of amino acid changes and their frequency based on the predictions of disordered regions. Here, we tested for the first time with experimentally tested cases how well the prediction programs detect changes in protein disorder caused by amino acid substitutions. We noticed that the evaluated programs, and possibly similar ones, do not suit well for this task and if you use them to investigate residue changes, we recommend applying them together with other prediction programs. After noticing this, we developed a dedicated prediction method.

Methods

Datasets and Test Cases

Information about the effects of amino acid substitutions on protein disorder cannot be found in any database. Therefore, text mining of literature (PubMed abstracts) was performed to obtain experimentally verified cases of the effects of variants on protein disorder or order. MeSH terms proved useless for the elimination of irrelevant literature. The literature mining tools included OvidSP (http://ovidsp.uk.ovid.com/), QUOSA Information Manager (http://www.quosa.com/), and Biomart (http://www.biomart.org/). As clearly defined search terms were missing, numerous keyword combinations were applied. Three groups of keywords were generated and applied in combinations (Box 1). One keyword at a time from each group was combined with logical "AND". $ indicates character wild card.

All the keyword combinations were generated to search in Ovid Medline In-Process and Other Non-Indexed Citations 1950 to present dataset. Altogether, 255 articles were found. The relevance of the obtained articles was confirmed manually.

Disorder Prediction Programs

The effects of variants on disorder were studied by 29 versions of 19 programs (Table 1) including Anchor [Dosztanyi et al., 2009], DisEMBL [Linding et al., 2003a], DISOclust [McGuffin, 2008], DISOPRED2 [Ward et al., 2004], DISpro [Cheng et al., 2005b], FoldIndex [Prilusky et al., 2005], GeneSilico Metadisorder [Kozlowski and Bujnicki, 2012], GlobPlot [Linding et al., 2003b], iPDA [Su et al., 2007], IUPred [Dosztanyi et al., 2005], MetaPrDOS [Ishida and Kinoshita, 2008], multilayered fusion-based disorder predictor (MFDp) [Mizianty et al., 2010],OnD-CRF [Wang and Sauer, 2008], POODLE-I [Hirose et al., 2010], POODLE-L [Hirose et al., 2007], POODLE-S [Shimizu et al., 2007a], POODLE-W [Shimizu et al., 2007b], PrDOS [Ishida and Kinoshita, 2007], RONN [Yang et al., 2005], Softberry PDISORDER, SPRITZ Vullo et al., 2006], and WinDiso [Holladay et al., 2007]. These methods were chosen due to their availability and accessibility.

All the evaluated disorder programs were based on the amino acid sequence information. The output varies—some programs provided disorder prediction for each residue either by classification or numeric value, whereas others generate graphs. The default parameters of the programs were utilized. Protein wild-type fasta format sequence and variant sequence were utilized as the input.

From the output of each prediction program, the disorder classification O (meaning order) or D (meaning disorder) was collected for each variant position, for both the wild-type and variant form. Of the disorder prediction programs, only WinDiso did not provide such classification; hence, all the numerical values below zero were considered as ordered and values above disordered.

Anchor

ANCHOR [Dosztanyi et al., 2009] is based on the pairwise energy estimation approach similar to IUPred [Dosztanyi et al., 2005]. It uses basic biophysical properties of disordered binding regions and estimated energy calculations [Meszaros et al., 2007].

Datasets for training and testing the method were 46 complexes of short-disordered and long-globular proteins, 28 complexes of long-disordered and long-globular proteins, and 553 monomeric globular proteins as a negative dataset.

DisEMBL

DisEMBL [Linding et al., 2003a] is based on three neural network prediction programs. Results are supplied as disordered by loops/coils definition, disordered by hot-loops definition, and disordered Remark-465 (regions deficient of electron density in structure) definition. Loops and coils are defined by DSSP [Kabsch and Sander, 1983], hot-loops are the regions with high B-factors, and Remark-465 are absent from the PDB [Berman et al., 2000] X-ray structures.

Training set for coils prediction included one chain from each SCOP superfamily [Gough et al., 2001]. For the loops training set, secondary structure assignments were obtained from DSSP for representatives of SCOP family members. B-factors from regions of regular secondary structure were applied for normalization by establishing chain-specific cutoffs for discriminating between ordered and disordered regions. Loop regions with B-factors above the 90% quantum were considered as disordered. The training set of 1,547 sequences for prediction of missing coordinates (Remark-465) included only one chain per SCOP protein family.

DISOclust

DISOclust [McGuffin, 2008] has two steps: the prediction of the per-residue error in multiple fold recognition models followed by a simple analysis of the conservation of per-residue error across all models.

The first prediction program applies coil predictions from PSIPRED [Jones, 1999] and labels all the coil regions as disordered. The second program counts missing residues in multiple fold recognition models. The count is utilized to estimate the probability of disorder by calculating the number of times the residue was missing in a model divided by the total number of models.

DISOPRED2

DISOPRED2 [Ward et al., 2004] is a SVM classifier. During training, residues with missing atomic coordinates in 715 high resolution X-ray structures for ordered proteins with less than 25% sequence identity were defined as disordered. Different combinations of binary-encoded amino acid sequences and PSIPRED [Jones, 1999; Jones et al., 2008] secondary structure predictions and PSIBLAST [Altschul et al., 1997] for symmetric windows of 15 positions were exploited to train the prediction program.

DISpro

DISpro [Cheng et al., 2005b] is a 1D recursive neural network trained with a nonredundant dataset of disordered regions in protein X-ray structures from PDB. Amino acids with missing ATOM records were considered as disordered. Then, homologous protein chains were filtered out with UniqueProt [Mika and Rost, 2003]. Secondary structure and relative solvent accessibility were predicted with Sspro [Pollastri et al., 2002b] and ACCpro [Pollastri et al., 2002a], respectively. The filtering procedures resulted in a set of 723 nonredundant disordered chains.

FoldIndex

FoldIndex [Prilusky et al., 2005] is a modified version of the [Uversky et al. 2000] algorithm applied to disorder prediction. This algorithm is based on the local hydropathy values and net charge of the sequence analyzed by a sliding window technique. The charge/hydropathy program predicts fully unstructured domains (random coils) by applying global sequence composition (hydrophobicity versus net charge). Regions that have a low hydrophobicity and high net charge are predicted to be either loops or unstructured regions.

GeneSilico Metadisorder

First, GeneSilico Metadisorder [Kozlowski and Bujnicki, 2012] harvests predictions from primary disorder prediction programs POODLE-L, iPDA, IUPred, DISpro, POODLE-S, IUPRED, SPRITZ, PrDOS, RONN, DISOPRED2, and DisEMBL. Then, it weights the outputs of the individual programs according to their accuracies. Three separate datasets were utilized for training and benchmarking the programs. The first set was 1,147 proteins from PDB database filtered by resolution <2 Å , *R*-factor <0.2, length 50–1,000 amino acids, and sequence identity <20%. The second set was 566 proteins composed from Disprot (version 3.6) and CASP7 [Bordoli et al., 2007] targets. The last set was 122 targets from CASP8 [Noivirt-Brik et al., 2009].

GlobPlot

GlobPlot [Linding et al., 2003b] exploits the Russell/Linding scale of disorder (propensities for secondary structures and random coils) to predict regions with the propensity for globularity. GlobPlot is based on a running sum of the propensities for amino acids. It identifies interdomain segments containing linear motifs, and apparently ordered regions that do not contain any recognized domain.

iPDA

iPDA [Su et al., 2007] integrates DisPSSMP2 [Hsu et al., 2011] with several other sequence-based prediction programs to investigate the functional role of disordered regions. DisPSSMP2 employs position-specific scoring matrices (PSSMP) for amino acid physicochemical properties. The predicted information includes sequence conservation, secondary structure, sequence complexity, and hydrophobic clusters.

IUPred

IUPred [Dosztanyi et al., 2005] predicts regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner. It is based on energy resulting from interresidue interactions and estimated from local amino acid composition. The algorithm discriminates between globular and disordered proteins. Globular proteins form a number of interresidue interactions, providing stabilizing energy to overcome the entropy loss during folding, whereas disordered regions have less of these interactions.

MetaPrDOS

MetaPrDOS [Ishida and Kinoshita, 2008] is a metapredictor that combines results from PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-s. The training set was collected from PDB and 10-fold cross-validation was utilized to optimize the training parameters for the SVM program.

MFDp

MFDp [Mizianty et al., 2010] combines three SVMs, which predict short-, long-, and generic-disordered regions. In addition to SVMs, it utilizes the sequence, sequence profiles, predicted secondary structure, solvent accessibility, backbone dihedral torsion angles, residue flexibility, and B-factors. MFDp was trained with 514 protein sequences that included residues from disordered regions of all sizes in 309 proteins from the DisProt database [Sickmeier et al., 2007] and 205 X-ray structures from the PDB [Berman et al., 2000].

OnD-CRF

Predicting order and disorder by using conditional random fields (OnD-CRF) [Wang and Sauer, 2008] applies CRFs, which employs features generated from the amino acid sequence and secondary structure prediction. The OnD-CRF training dataset, derived from PDB, contained 215,612 residues, of which 13,909 were defined as disordered, since they missed coordinates in PDB entries. The features were extracted only from the amino acid sequence and from the secondary structure predicted by SSpro [Cheng et al., 2005a]. A sliding window was optimized as nine amino acids with 10-fold cross-validation.

POODLE

POODLE (Prediction Of Order and Disorder by machine LEarning) is a set of machine learning-based programs for predicting protein disorder from amino acid sequences. POODLE provides three disorder predictions according to the length of the target disordered segment. POODLE-L [Hirose et al., 2007] (L stands for a long-disordered region) and POODLE-S [Shimizu et al., 2007a] (S for a short-disordered region) are based on SVMs, which exploit 10 kinds of physicochemical properties of amino acids.

POODLE-L predicts long-disorder regions, longer than 40 consecutive amino acids. The negative training set for POODLE-L was from PDB and the positive from Uversky’s article [2000] and Dis-Prot [Sickmeier et al., 2007]. POODLE-S is a group of seven SVM prediction programs each responsible for a specific region of the whole sequence. It predicts shorter disorder regions. POODLE-W [Shimizu et al., 2007b] (W for a wholly disordered region) is based on the spectral graph transducer, which embraces a semisupervised learning and was trained on sequences with known structure. It predicts which proteins are mostly disordered. The training set was collected from DisProt [Sickmeier et al., 2007]. POODLE-I [Hirose et al., 2010] (I for integration) integrates three programs.

PrDOS

PrDOS [Ishida and Kinoshita, 2007] consists of two prediction programs. One is based on the local amino acid sequence and the other on template proteins (or homologous proteins for which structural information is available). The first tool applies SVM and the second PSI-BLAST. The final prediction is done as the combination of the two programs.

RONN

RONN [Yang et al., 2005] is based on neural networks and utilizes sequence alignments. Bio-Basis Function Neural Network (BBFNN) [Thomson et al., 2003], which was originally developed for a sequence alignment-based detector of protease cleavage sites, was utilized.

The training set for RONN, which included 872 entries after filtering, was collected from Molecular Structure Database [Boutselakis et al., 2003]. Experimentally defined structures were utilized excluding multicomponent complexes. The training set was formed of 891 ordered regions and 530 disordered regions.

Softberry PDISORDER

Softberry PDISORDER (Softberry) is based on the combination of neural network, linear discriminant function, and acute smoothing procedure. It uses a window of 31 residues.

Spritz

Spritz [Vullo et al., 2006] is based on two SVM prediction programs: one recognizes short and the other long regions of disorder. The long-disordered regions prediction program was trained with 45 completely disordered sequences from DISPROT and 45 ordered fragments from PDBselect25. The filtered and balanced set contained 293 sequences corresponding to 34,159 residues, 17,001 of which were classified as belonging to the long regions of disorder.

Short-disordered sequence fragments were compiled from Protein Data Bank (PDB) and contained chains with at most 20 disordered amino acids sharing no more than 25% sequence similarity. The final set contained 1,017 sequences corresponding to 278,600 residues, 8,824 of which were classified as belonging to the short regions of disorder.

WinDiso

WinDiso [Holladay et al., 2007] is a weighted window SVM predictor similar to DISOPRED2 [Ward et al., 2004]. Its training and testing data included X-ray crystallographic data from domains in 1,912 families in the first fivefold classes of the Structural Classification of Proteins (SCOP) version 1.67 [Andreeva et al., 2004].

Novel Method

PON-Diso was developed to predict the effects of amino acid substitutions on protein disorder. The classifier is based on machine learning technique called random forest (RF) classifier [Breiman, 2001]. The RF classifier is built on two sets of features: those selected from AAindex [Kawashima and Kanehisa, 2000], and evolutionary sequence conservation features.

AAindex Feature Selection

Feature selection was performed among the characteristics available in AAindex, which contains three databases of altogether 685 physicochemical and biochemical properties of amino acids. Each entry contains numerical values for the amino acid types. There is a wide range of indices including hydropathy and secondary structural element propensities. The AAindex 1 features have a numerical index for each amino acid, whereas AAindex 2 and AAindex 3 features contain amino acid substitution matrices. Six-hundred seventeen features were left after eliminating incomplete features using AAindex1; the differences between indices for the variant and wild-type residue were calculated. The values for differences between the normal and variant amino acid were taken from substitution matrices for features in AAindex 2 and AAindex 3. Feature ranking based on Gini importance by RF yielded two most significant features.

Evolutionary Features

Sequences homologous to the query are collected with PSI-Blast [Altschul et al., 1997] and sequences with greater than 90% identity with the query are discarded (Supp. Fig. S1). Then, the homolog sequences are clustered with USEARCH [Edgar, 2010] so that each group has at least 90% similarity among sequences. A consensus sequence is calculated for each cluster with USEARCH. A multiple sequence alignment (MSA) is generated in an iterative procedure. Each consensus sequence is globally aligned with the query/MSA using MUSCLE [Edgar, 2004] and then the Jensen–Shannon conservation (JSC) [Capra and Singh, 2007] score is computed. The JSC scores are normalized to the length of the MSA, and then the consensus sequence with the highest JSC score is aligned with the query/MSA, and JSC score is calculated for the resulting MSA. A logical test on JSC score is performed to check whether the score for the current iteration is greater or in the range of 5% of the absolute deviation to the previous JSC score. The procedure of adding sequences to the MSA is continued as long as the test result remains true. At the final step, PSSM, a $l \times 20$ matrix, is calculated for the MSA. l is the length of the MSA. Pca in the PSSM represents the probability of amino acid a substitution at position c, calculated from general formula of Henikoff and Henikoff (1996) as:

$$P_{ca} = \frac{N_c}{(N_c + B_c)} \times g_{ca} + \frac{B_c}{(N_c + B_c)} \times f_{ca}.$$

Here N_c is the number of sequences in the MSA, g_{ca} is the sequence-weighted frequency. The gaps in the alignment at position c with frequency g_{c-} are distributed to all 20 amino acids a, by incrementing count g_{ca} by 1/20 of g_{c-}. f_{ca} represents pseudocount, which is calculated based on substitution probabilities (Tatusov et al., 1994), B_c is the total number of pseudocounts. When the probability is greater than 0.5, the variation is considered to change the order/disorder status, otherwise it is considered as tolerated. From the PSSM, two scores for the probability and the mean probability at the variation position, the JSC score of the finalMSA, and the decision based on probability are collected and used as features for the RF classifier. To speed up the process of calculating the MSA, after every 10 iterations, consensus sequences having JSC score below a certain threshold are discarded. The threshold was set based on empirical study of known variations.

PON-Diso Classifier

The PON-Diso classifier using *R* interface (http://cran.r-project.org/web/packages/randomForest/index.html) was built based on the selected six features, two AAindex features and four evolutionary features. The number of variables randomly sampled at each split was set to three and the number of trees grown as 500. The training of RF was carried out using samples of disorder-related effects on amino acid changes (Supp. Table S1). Since there were only three cases of D to O change, they were excluded and the remaining 98 variants were used to train the RF. AWeb server was installed to run PON-Diso. It is available at http://structure.bmc.lu.se/PON-Diso/. PON-Diso provides the user with aWeb-based interface to submit a protein and variants to predict the effect of amino acid substitutions. Users can either submit one or more protein identifiers (Ensembl ID, RefSeq ID, or Swiss-Prot ID) or protein sequences with their corresponding variations.

Results

We examined the usability and reliability of protein disorder programs on predicting changes to protein structural disorder/order caused by amino acid substitutions. For this purpose, we performed literature search and collected a dataset of 101 residue changes in 31 proteins and then evaluated them with 29 versions in 19 protein disorder prediction programs.

Protein structures are dynamic and in constant movement. Folded protein can partially unfold and refold back to normal conformation (Fig. 1). There are numerous structural stages and possible transitions between them [e.g., Chiti and Dobson, 2006]. For the assessment of the predictor performance, we collected variants for which information about their misfolding was available. Unfolded protein has four possible destinies; it will be either refolded, degraded, aggregated, or sequestered. Since aggregation is a distinct process for which a number or specific prediction tools are available, including Aggrescan [Conchillo-Sole et al., 2007], PASTA [Trovato et al., 2007], Tango [Fernandez-Escamilla et al., 2004], and others, we concentrated on misfolding-related cases, which disorder predictors should be able to address.

Test Set

With extensive text mining, we identified from literature 101 cases (Supp. Table S1) in 31 proteins in which the effects of the amino acid substitutions on disorder are known. Search for the test cases was demanding because the effects of substitutions on disorder have been studied in few articles only. We optimized the search key words and then applied them in combination to mine PubMed abstracts. Finally, we curated the obtained results manually. The dataset is available from VariBench [Nair and Vihinen, 2013], a database for variation effect datasets, along with Variation Ontology (VariO) annotations [Vihinen, 2014].

The variant effects have been studied by structural methods including X-ray crystallography, NMR spectroscopy, small-angle X-ray scattering, circular dichroism or fluorescence spectrometry, or by immunological analyses to reveal unfolded protein response caused by accumulation of unfolded proteins to endoplasmic reticulum.

Test Set Features

Protein order/disorder is not affected by 39 of the variants, 31 of them retain the ordered and eight retain the disordered state. These cases are considered as the negative test set, TNs (true negatives). Sixty-two variants affect protein structural order. These cases are considered as the positive test set or TPs (true positives). In 59 cases, the variant increases the structural disorder and only in three cases decreases it. The test is biased to the cancer variants containing 36 cases of BRCA1 variants.

We employed PON-P portal (http://bioinf.uta.fi/PON-P) [Olatubosun et al., 2012] to submit the test set to eight investigated disorder prediction programs. The other programs were used from the Web services of the programs. In the test set, amino acid residue distribution is biased because the original residues contain altogether 12 arginines, 11 aspartic acids, and nine alanines, whereas the variant residues include 12 alanines, nine serines, and eight asparagines. The most common residue substitutions are four cases from cysteine to serine and from isoleucine to valine. TP variants changing from disorder to order had two aspartic acids as original residues and two alanines as altered residues. Fifty-nine TP variants from order to disorder contain eight alanines, six arginines, and five isoleucines, serines, and tryptophans as original residues, whereas the substituted residues include six alanines and five leucines and prolines. The biggest number of cases in the changes from order to disorder was three cases from alanine to proline. Due to the low number of cases, more detailed statistical analysis of distributions was not feasible.

Additionally, we studied the effect of the residue types by organizing the variant and original amino acids in to six groups according to the physicochemical properties: hydrophobic (C, F, I, L, M, V, W, and Y), positively charged (H, K, and R), negatively charged (D and E), conformational (G and P), polar (N, Q, and S), and A and T [Shen and Vihinen, 2004]. The change from order to disorder is conserved in the variant in the same physicochemical group in nine cases and changes in 50 cases. All three changes from disorder to order changed the physicochemical group.

Performance of Protein Disorder Prediction Programs

The results for the capability of the disorder prediction programs to predict the type of the wild-type residues at the variant sites are in Table 2. The success rate is not very high. Disorder prediction programs are clearly better predicting negative cases not involved in disorder than positive cases having disorder. The success rate for the negative cases varied between 48.5% and 90.1%, whereas results for positive cases vary between 0% and 6.9%. DisEMBL version for loops/coils definition and POODLE-I are the best prediction programs for positive cases, but have high false-positive rates of 44.6% for DisEMBL and 10.9% for WinDiso. Their TN rates are among the poorest in this validation with 48.5% (DisEMBL) and 78.2% (POODLE-I). Dispro and POODLE-S are the best predictors for the negative cases with the false-negative rate of 5.9%. The best overall disorder predictors for wild-type residues are Dispro and POODLE-S with high TP and TN rate of 91.1%.

Results for the analysis of the effects of variants on protein disorder are in Table 3. The conclusion is that the programs have a relatively low success rate in detecting the changes caused by amino acid substitutions. These programs are especially poor in predicting the change from disorder to ordered structure and order to disorder, as the success rate in these cases varied from 0% to 5.9%. On the other hand, the predictions of negative cases, which indicate no change in disorder, had somewhat higher accuracy: the success rate varies between 21.8% and 33.7%. However, there is a massive false-negative rate (56.4%–70.3%).

OnDCRF and Spritz short disorder have the highest number of TPs, yet the value is only 5.9%. All the other programs have TP rate maximum of 5% and many have only 0%. The best FP results, 0%, were obtained by DISOclust, Dispro, iPDA, MetaPrDos, POODLES for missing residues, POODLE-W, and PrDos, whereas the worst was 7.9% for WinDiso. Nevertheless, the good FP results are not because of reliable predictions as Dispro, MetaPrDos, POODLE-W, and PrDos did not predict any of the variants to increase disorder, whereas iPDA predicted just one variant, and DISOclust and POODLE-S (for missing residues) for three variants.

Results for TN rate are significantly better ranging from 21.8% for DisEMBL to 33.7% for OnDCRF and Spritz long disorder. Even these results are due to the overprediction of negative cases, FN values ranging from 56.4% to 70.3%.

Picking the best program is impossible as the performance figures differ for positive and negative cases. Although certain programs obtain good FP characteristics, that is achieved with the cost of very high FN, because virtually no cases are predicted to affect the structural order state. OnDCRF could be considered as the best overall program, but TP of 5.9% and TN of 33.7% are far from practically applicable range. A sequence profile-based program RONN may have suffered from the close similarity of the wild-type and variant sequence used. However, its results are not worse than for the majority of the predictors. Softberry PDISORDER could not predict all the cases missing two.

PON-Diso Classifier

A new predictor, PON-Diso, was developed (Supp. Fig. S1) for the analysis of effects of amino acid substitutions. Its performance was evaluated using an independent test set and fivefold cross-validation. First, 10 variants (10% of the dataset) were selected and used only as independent test set to assess the performance of the final predictor. The remaining cases were partitioned into five groups by randomly stratifying proteins and variations so that all the variations in a protein were always in one set.

Feature selection was performed five times by using the different combinations of partitions, one of the partitions was always used for testing. Gini index from RF was used for ranking features resulting in five sets of selected features. A union set of these features was created. To dimensionally reduce the union set, second feature selection using RF classifier-based

Gini indices was performed giving rise to two most significant AAindex features. These features are FINA910103 (AAindex code) helix termination parameter at position j-2,j-1,j [Finkelstein et al., 1991] and KOSJ950102 context-dependent optimal substitution matrices for exposed beta [Koshi-Goldstein, 1995].

For cross-validation, five RF classifiers were trained on four partitions of the training set using the two most significant AAindex features and evolutionary features. The average classification rate on remaining partition was computed to be 70.4% for the five combinations. On independent test, the classification rate of the final PON-Diso method is 50%. This performance is clearly better than for any of the tested disorder methods and can be used for the analysis of amino acid substitutions. Due to the small number of cases, it is not possible to use all the measures recommended for reporting the performance of predictors of this kind [Vihinen, 2012, 2013], instead classification rates are shown for percentages of cases correctly classified.

Discussion

We evaluated the applicability and reliability of disorder prediction programs for predicting protein disorder changes caused by amino acid substitutions. The test set included 101 amino acid substitutions. We developed a novel method for predicting the changes in disorder caused by amino acid substitutions. The tested disorder prediction methods are for detecting longer disordered regions in protein sequences, whereas PON-Diso is to our knowledge the only one dedicated for this purpose.

The performance of the disorder prediction programs on the wild-type residues of variation sites was relatively low (Table 2). Moreover, we noticed that the programs that were good at predicting negative cases often had the high false-negative rate, whereas good positive predictors had the high false-positive rate. If a program predicts all the cases to belong to one class, it inevitably has a high false-positive/negative rate. Our novel method predicted the independent test cases 50% correct, and thus outperforms other methods in this evaluation. However, due to the small size of the test set possibly causing random effects, we consider the crossvalidation result of 70.5% to be more reliable and to better describe the performance of the method.

From the perspective of utilizing disorder prediction programs to predict the effects of amino acid variants, the tested programs performed poorly. Disorder prediction methods are available for many different lengths, but not for very short regions. None of these programs have been trained for amino acid substitutions. Even the study of length-dependent predictor performance excluded short 1–3 residue regions [Peng et al., 2005; Yue et al., 2006] as was done, for example, in CASP9. Comparison of the short- and long-disorder region predictors of Spritz showed that short disorder is slightly better in predicting positive cases, 5.9%, as compared with long disorder, 1.0% (Table 3). As our test set indicates, single amino acid substitutions can have profound effects on protein order/disorder.

Some disorder prediction programs have been evaluated previously in Critical Assessment of Techniques for Protein Structure prediction (CASP challenges): CASP5 [Melamud andMoult, 2003], CASP6 [Jin and Dunbrack, 2005], CASP7 [Bordoli et al., 2007], CASP8 [Noivirt-Brik et al., 2009], and CASP9 [Monastyrskyy et al., 2011]. CASP9 had 117 proteins (98 X-ray and 19 NMR structures) that were used to test 32 participants, 22 servers, and 10 human expert groups. A residue was considered being in a disordered state if it did not have spatial coordinates or displayed a high conformational variability across different X-ray chains or NMR models. This dataset was biased toward shorter disordered regions. After eliminating short segments (more than four amino acids considered as noise), the assessment set included 26,075 residues of which 2,417 were classified as disordered. The best performing methods on the CASP9 set were PrDOS2, DisoPred3C, MULTICOM-refine, Zhou-Spine-D, Zhou-

Spine-DM, CBRC Poodle, biomine dr pdb, biomine dr pdb c, GSmetaDisorderMD, GSmetaserver, OnD-CRF, Mason, and McGuffin. The accuracies of these methods varied between 0.661 (biomine dr pdb c) and 0.754 (PrDOS2), whereas MCC varied between 0.274 (OnD-CRF) and 0.508 (DisoPred3C). Due to the small size of our dataset, the calculations of parameters as accuracy and MCC and ROC curve would not have been meaningful, although strongly recommended for larger datasets [Vihinen, 2012; Vihinen, 2013].

We could not evaluate all the disorder prediction programs published. No Ordered Regular Secondary Structure (NORSp) [Liu and Rost, 2003] and MD [Schlessinger et al., 2009] are part of Predict-Protein service and thus not freely available, whereas SEG [Wootton, 1994] offers the low complexity region information in a form that was not possible to compare with other programs. Similarly, results of MULTICOM were difficult to interpret. Scooby-domain [George et al., 2005] was difficult to interpret for this purpose. Dismeta (http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/), DNDisorder [Eickholt and Cheng, 2013],DRIP-PRED (http://www.sbc.su.se/_maccallr/disorder/cgi-bin/submit.cgi), Hydrophobic Cluster Analysis [Gaboriaud et al., 1987], IUPforest [Han et al., 2009], MeDor [Lieutaud et al., 2008], PreDisorder [Deng et al., 2009], and SPINE-D [Zhang et al., 2012] did not work when we tested them.

Due to the large number of the methods, all of them could not be tested. However, we included all the possible programs fromCASP9. Many of the CASP9 methods combined a prediction program and human expertise as Zhou-Spine-D is the combination of Spine-D prediction and Zhou's expertise.

Conclusion

Our main conclusion is that general disorder prediction programs are not applicable to detect the changes in disorder caused by amino acid substitutions. Therefore, we do not recommend utilizing the evaluated tools as the only predictors for variation effects. Nonetheless, they might prove useful if employed together with other types of protein effect predictions [Thusberg and Vihinen, 2006; Thusberg and Vihinen, 2007; Mort et al., 2010] or applied in predictions with the wide spectrum of features [Li et al., 2009]. However, even in these cases, the emphasis of disorder predictions should be low except for PON-Diso predictions.

Acknowledgments

The authors gratefully acknowledge financial support from all the sponsors.

Disclosure statement: The authors declare no conflict of interest.

References

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl Acids Res 25:3389-3402.

Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. 2004. SCOP database in 2004: Refinements integrate structure and sequence family data. Nucl Acids Res 32:D226-D229.

Ayuso-Tejedor S, García-Fandiño R, Orozco M, Sancho J, Bernadó P. 2011. Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. J Mol Biol 406:604-619.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucl Acids Res 28:235-242.

Bordoli L, Kiefer F, Schwede T. 2007. Assessment of disorder predictions in CASP7. Proteins Struct Funct and Bioinf 69:129-136.

Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. 2003. E-MSD: The European Bioinformatics Institute Macromolecular Structure Database. Nucl Acids Res 31:458-462.

Breiman L. 2001. Random forests. Mach Learn 45:5-32 .

Brocca S, Scaronamalíková M, Uversky VN, Lotti M, Vanoni M, Alberghina L, Grandori R. 2009. Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins Struct Funct and Bioinf 76:731-746.

Buckle AM, Henrick K, Fersht AR. 1993. Crystal structural analysis of mutations in the hydrophobic cores of barnase. J Mol Biol 234:847-860.

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics 1;23(15):1875-82.

Cheng J, Randall AZ, Sweredoski MJ, Baldi P. 2005a. SCRATCH: A protein structure and structural feature prediction server. Nucl Acids Res 33:W72-W76.

Cheng J, Sweredoski M, Baldi P. 2005b. Accurate prediction of protein disordered regions by mining protein structure data. Data Mining and Knowledge Discovery 11:213-222.

Chiti F, Dobson CM. 2006. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75:333-366.

Chouard T. 2011. Breaking the protein rules. Nature 471:151-153.

Conchillo-Sole O, de Groot N, Aviles F, Vendrell J, Daura X, Ventura S. 2007. AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8:65.

de Almeida SF, Fleming JV, Azevedo JE, Carmo-Fonseca M, de Sousa M. 2007. Stimulation of an unfolded protein response impairs MHC class I expression. The Journal of Immunology 178:3612-3619.

Deiana A, Giansanti A. 2010. Predictors of natively unfolded proteins: Unanimous consensus score to detect a twilight zone between order and disorder in generic datasets. BMC Bioinformatics 11:198.

Deng X, Eickholt J, Cheng J. 2009. PreDisorder: Ab initio sequence-based prediction of protein disordered regions. BMC Bioinformatics 10:436.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433-3434.

Dosztányi Z, Mészáros B, Simon I. 2009. ANCHOR: Web server for predicting protein binding regions in disordered proteins. Bioinformatics 25:2745-2746.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. Biochemistry (NY) 41:6573-6582.

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. 2001. Intrinsically disordered protein. J Mol Graph Model 19:26-59.

Dyson HJ, Wright PE. 2002. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12:54-60.

Edgar R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput Nucleic Acids Res. 32(5):1792-1797.
Edgar R. C. 2010. Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461.

Eickholt J, Cheng J. 2013. DNdisorder: Predicting protein disorder using boosting and deep networks. BMC Bioinformatics 14:88.

Fefeu S, Biekofsky RR, McCormick JE, Martin SR, Bayley PM, Feeney J. 2000. Calcium-induced refolding of the calmodulin V136G mutant studied by NMR spectroscopy: Evidence for interaction between the two globular domains. Biochemistry (NY) 39:15920-15931.

Feng S, Zhao T, Zhou H, Yan Y. 2007. Effects of the single point genetic mutation D54G on muscle creatine kinase activity, structure and stability. Int J Biochem Cell Biol 39:392-401.

Fernandez-Escamilla A, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotech 22:1302-1306.

Finkelstein AV, Badretdinov AY, Ptitsyn OB. 1991. Physical reasons for secondary structure stability: alpha-helices in short peptides. Proteins 10:287-299.

Fisher CK, Stultz CM. 2011. Protein structure along the order-disorder continuum. J Am Chem Soc 133:10022-10025.

Freeman L, Buisson M, Tarbouriech N, Van der Heyden A, Labbé P, Burmeister WP. 2009. The flexible motif V of Epstein-Barr virus deoxyuridine 5′-triphosphate pyrophosphatase is essential for catalysis. J Biol Chem 284:25280-25289.

Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. 1987. Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. FEBS Lett 224:149-155.

George RA, Lin K, Heringa J. 2005. Scooby-domain: Prediction of globular domains in protein sequence. Nucl Acids Res 33:W160-163.

Georgescauld F, Mocan I, Lacombe M, Lascu I. 2009. Rescue of the neuroblastoma mutant of the human nucleoside diphosphate kinase A/nm23-H1 by the natural osmolyte trimethylamine-N-oxide. FEBS Lett 583:820-824.

Gleghorn LJ, - Trump D, - Bulleid NJ. 2009. Wild-type and missense mutants of retinoschisin co-assemble resulting in either intracellular retention or incorrect assembly of the functionally active octamer. Biochem. J 425:275-83

Gohlke H, Thorpe MF. 2006. A natural coarse graining for simulating large biomolecular motion. Biophys J 91:2115-2120.

Gorbatyuk MS, Knox T, LaVail MM, Gorbatyuk OS, Noorwez SM, Hauswirth WW, Lin JH, Muzyczka N, Lewin AS. 2010. Restoration of visual function in P23H rhodopsin transgenic rats by gene delivery of BiP/Grp78. PNAS 107:5961-5966.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 313:903-919.

Guy JE, Wigren E, Svärd M, Härd T, Lindqvist Y. 2008. New insights into multiple coagulation factor deficiency from the solution structure of human MCFD2. J Mol Biol 381:941-955.

Han P, Zhang X, Norton RS, Feng Z. 2009. Large-scale prediction of long disordered regions in proteins using random forests. BMC Bioinformatics 10:8.

Han S, Liu Y, Chang A. 2007. Cytoplasmic Hsp70 promotes ubiquitination for endoplasmic reticulum-associated degradation of a misfolded mutant of the yeast plasma membrane ATPase, PMA1. J Biol Chem 282:26140-26149.

Hartl FU, Hayer-Hartl M. 2009. Converging concepts of protein folding *in vitro* and *in vivo*. Nat. Struct. Mol. Biol 16:574-581.

Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. 2007. POODLE-L: A two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics 23:2046-2053.

Hirose S, Shimizu K, Noguchi T. 2010. POODLE-I: Disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach. In silico biol. 10:185-91.

Henikoff JG, Henikoff S. 1996. Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci. 12(2):135-43.

Holladay NB, Kinch LN, Grishin NV. 2007. Optimization of linear disorder predictors yields tight association between crystallographic disorder and hydrophobicity. Prot Sci 16:2140-2152.

Hsu C, Chen C, Liu B. 2011. WildSpan: Mining structured motifs from protein sequences. Alg Mol Bio 6:6.

Hu Y, Liu Y, Jung J, Dunker AK, Wang Y. 2011. Changes in predicted protein disorder tendency may contribute to disease risk. BMC Genomics 12:S2.

Idowu SM, Gautel M, Perkins SJ, Pfuhl M. 2003. Structure, stability and dynamics of the central domain of cardiac myosin binding protein C (MyBP-C): Implications for multidomain assembly and causes for cardiomyopathy. J Mol Biol 329:745-761.

Iimura S, Umezaki T, Takeuchi M, Mizuguchi M, Yagi H, Ogasahara K, Akutsu H, Noda Y, Segawa S, Yutani K. 2007. Characterization of the denatured structure of pyrrolidone carboxyl peptidase from a hyperthermophile under nondenaturing conditions: Role of the C-terminal alpha-helix of the protein in folding and stability. Biochemistry (NY) 46:3664-3672.

Ishida T, Kinoshita K. 2007. PrDOS: Prediction of disordered protein regions from amino acid sequence. Nucl Acids Res 35:W460-W464.

Ishida T, Kinoshita K. 2008. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344-1348.

Jin Y, Dunbrack RL. 2005. Assessment of disorder predictions in CASP6. Proteins Struct Funct Bioinf 61:167-175.

Jones DT 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195-202.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong S, Fu B, Lin M, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321:1801-6.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577-2637.

Kasakov AS, Bukach OV, Seit-Nebi AS, Marston SB, Gusev NB. 2007. Effect of mutations in the ß5–ß7 loop on the structure and properties of human small heat shock protein HSP22 (HspB8, H11). FEBS Journal 274:5628-5642.

Kawashima S, Kanehisa M. 2000. AAindex: Amino acid index database. Nucl Acids Res 28:374-374.

Kim JH, Füzéry AK, Tonelli M, Ta DT, Westler WM, Vickery LE, Markley JL. 2009. Structure and dynamics of the Iron−Sulfur cluster assembly scaffold protein IscU and its interaction with the cochaperone HscB. Biochemistry (NY) 48:6062-6071.

Kishii R, Falzon L, Yoshida T, Kobayashi H, Inouye M. 2007. Structural and functional studies of the HAMP domain of EnvZ, an osmosensing transmembrane histidine kinase in *Escherichia coli*. J Biol Chem 282:26401-26408.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31:675-684.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng 8:641-645.

Kozlowski L, Bujnicki J. 2012. MetaDisorder: A meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13:111.

Krämer-Albers E, Gehrig-Burger K, Thiele C, Trotter J, Nave K. 2006. Perturbed interactions of mutant proteolipid Protein/DM20 with cholesterol and lipid rafts in oligodendroglia: Implications for dysmyelination in spastic paraplegia. J Neurosci 26:11743-11752.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1081.

Lakshminarasimhan M, Maldonado MT, Zhou W, Fink AL, Wilson MA. 2008. Structural impact of three parkinsonism-associated missense mutations on human DJ-1. Biochemistry 47:1381-92.

Lawless M, Mankan A, White M, O'Dwyer M, Norris S. 2007. Expression of hereditary hemochromatosis C282Y HFE protein in HEK293 cells activates specific endoplasmic reticulum stress responses. BMC Cell Biology 8:30.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744-2750.

Lieutaud P, Canard B, Longhi S. 2008. MeDor: A metaserver for predicting protein disorder. BMC Genomics 9 Suppl 2:S25.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003a. Protein disorder prediction: Implications for structural proteomics. Structure 11:1453-1459.

Linding R, Russell RB, Neduva V, Gibson TJ. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. Nucl Acids Res 31:3701-3708.

Liu J, Rost B. 2003. NORSp: Predictions of long regions without regular secondary structure. Nucl Acids Res 31:3833-3835.

Liu J, Tan H, Rost B. 2002. Loopy proteins appear conserved in evolution. J Mol Biol 322:53-64.

Liu Y, Wang Y, Wu C, Liu Y, Zheng P. 2009. Deletions and missense mutations of EPM2A exacerbate unfolded protein response and apoptosis of neuronal cells induced by endoplasm reticulum stress. Hum Mol Gen 18:2622-2631.

Liu Y, Lee SY, Neely E, Nandar W, Moyo M, Simmons Z, Connor JR. 2011. Mutant HFE H63D protein is associated with prolonged endoplasmic reticulum stress and increased neuronal vulnerability. J Biol Chem 286:13161-13170.

McGuffin LJ. 2008. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 24:1798-1804.

Melamud E, Moult J. 2003. Evaluation of disorder predictions in CASP5. Proteins Struct Funct Bioinf 53:561-565.

Mészáros B, Tompa P, Simon I, Dosztányi Z. 2007. Molecular principles of the interactions of disordered proteins. J Mol Biol 372:549-561.

Mika S, Rost B. 2003. UniqueProt: Creating representative protein sequence sets. Nucl Acids Res 31:3789-3791.

Mittag T, Marsh J, Grishaev A, Orlicky S, Lin H, Sicheri F, Tyers M, Forman-Kay JD. 2010. Structure/Function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 18:494-506.

Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L. 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26:i489-i496.

Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. 2011. Evaluation of disorder predictions in CASP9. Proteins Struct Funct Bioinf 79:107-118.

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD. 2010. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335-346.

Nair PS, Vihinen M. 2013. VariBench: a benchmark database for variations. Hum Mutat 34:42-49.

Narayana N, Phillips NB, Hua Q, Jia W, Weiss MA. 2006. Diabetes mellitus due to misfolding of a β-cell transcription factor: Stereospecific frustration of a schellman motif in HNF-1α. J Mol Biol 362:414-429.

Noivirt-Brik O, Prilusky J, Sussman JL. 2009. Assessment of disorder predictions in CASP8. Proteins 77:210-216.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: Integrated predictor for pathogenicity of missense variants. Hum Mutat 33:1166-1174.

Oldfield C, Meng J, Yang J, Yang MQ, Uversky V, Dunker AK. 2008. Flexible nets: Disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9:S1.

Pappachan A, Chinnathambi S, Satheshkumar PS, Savithri HS, Murthy MRN. 2009. A single point mutation disrupts the capsid assembly in sesbania mosaic virus resulting in a stable isolated dimer. Virology 392:215-221.

Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. 2005. Optimizing Long Intrinsic Disorder Predictors With Protein Evolutionary Information. J Bioinform Comput Biol 03:35-60.

Pollastri G, Baldi P, Fariselli P, Casadio R. 2002a. Prediction of coordination number and relative solvent accessibility in proteins. Proteins Struct Funct Bioinf 47:142-153.

Pollastri G, Przybylski D, Rost B, Baldi P. 2002b. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins Struct Funct Bioinf 47:228-235.

Popelkova H, Betts SD, Lydakis-Symantiris N, Im MM, Swenson E, Yocum CF. 2006. Mutagenesis of basic residues R151 and R161 in manganese-stabilizing protein of photosystem II causes inefficient binding of chloride to the oxygen-evolving complex. Biochemistry (NY) 45:3107-3115.

Potapov V, Cohen M, Schreiber G. 2009. Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. PEDS 22:553-560.

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex(C): A simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21:3435-3438.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. 2004. Protein flexibility and intrinsic disorder. Prot Sci 13:71-80.

Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD. 2008. An integrated approach to inferring gene-disease associations in humans. Proteins Struct Funct and Bioinf 72:1030-1037.

Rellos P, Ali M, Vidailhet M, Sygusch J, Cox TM. 1999. Alteration of substrate specificity by a naturally-occurring aldolase B mutation (Ala337->Val) in fructose intolerance. Biochem J 340: 321-7.

Roboti P, Swanton E, High S. 2009. Differences in endoplasmic-reticulum quality control determine the cellular response to disease-associated mutants of proteolipid protein. J. Cell Sci 122:3942-3953.

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. PNAS 103:8390-8395.

Rowling PJE, Cook R, Itzhaki LS. 2010. Toward classification of BRCA1 missense variants using a biophysical approach. J Biol Chem 285:20080-20087.

Roybal CN, Marmorstein LY, Vander Jagt DL, Abcouwer SF. 2005. Aberrant accumulation of fibulin-3 in the endoplasmic reticulum leads to activation of the unfolded protein response and VEGF expression. IOVS 46:3973-3979.

Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. 2009. Improved disorder prediction by combination of orthogonal approaches. PLoS ONE 4:e4433.

Schlessinger A, Rost B. 2005. Protein flexibility and rigidity predicted from sequence. Proteins Struct Funct and Bioinf 61:115-126.

Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. 2011. Protein disorder — a breakthrough invention of evolution? Curr Opin Struct Biol 21:412-418.

Shan B, McClendon S, Rospigliosi C, Eliezer D, Raleigh D. 2010. The cold denatured state of the C-terminal domain of protein L9 is compact and contains both native and non-native structure. J Am Chem Soc. 132:4669-77.

Shemetov AA, Gusev NB. 2011. Biochemical characterization of small heat shock protein HspB8 (Hsp22)–Bag3 interaction. Arch Biochem Biophys 513:1-9.

Shen B, Vihinen M. 2004. Conservation and covariance in PH domain sequences: Physicochemical profile and information theoretical analysis of XLA-causing mutations in the btk PH domain. PEDS 17:267-276.

Shimizu K, Hirose S, Noguchi T. 2007a. POODLE-S: Web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics 23:2337-2338.

Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T. 2007b. Predicting mostly disordered proteins by using structure-unknown protein data. BMC Bioinformatics 8:78.

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. 2007. DisProt: The database of disordered proteins. Nucl Acids Res 35:D786-D793.

Smith SE, Granell S, Salcedo-Sicilia L, Baldini G, Egea G, Teckman JH, Baldini G. 2011. Activating transcription factor 6 limits intracellular accumulation of mutant a1-antitrypsin Z and mitochondrial damage in hepatoma cells. J Biol Chem 286:41563-41577.

SoftBerry - PDISORDER:, [http://linux1.softberry.com/berry.phtml?topic=pdisorder&group=programs&subgroup=propttopic=pdisorder&group=programs&subgroup=propt

Steichen JM, Kuchinskas M, Keshwani MM, Yang J, Adams JA, Taylor SS. 2012. Structural basis for the regulation of protein kinase A by activation loop phosphorylation. J Biol Chem 287:14672-14680.

Stopa JD, Chandani S, Tolan DR. 2011. Stabilization of the predominant disease-causing aldolase variant (A149P) with zwitterionic osmolytes. Biochemistry (NY) 50:663-671.

Su C, Chen C, Hsu C. 2007. iPDA: Integrated protein disorder analyzer. Nucleic Acids Res 35:W465–W472.

Sugase K, Dyson J,H., Wright P,H. 2007. Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447:1021-1025.

Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA 91, 12091-12095.

Teilum K, Olsen J, Kragelund B. 2009. Functional aspects of protein flexibility. Cell Mol Life Sci 66:2231-2247.

Thomson R, Hodgman TC, Yang ZR, Doyle AK. 2003. Characterizing proteolytic cleavage site activity using bio-basis function neural networks. Bioinformatics 19:1741-1747.

Thusberg J, Vihinen M. 2007. The structural basis of hyper IgM deficiency – CD40L mutations. PEDS 20:133-141.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358-368.

Thusberg J, Vihinen M. 2006. Bioinformatic analysis of protein structure-function relationship: Case study of leucocyte elastase (ELA2) missense mutations. Hum Mutat 27:1230-1243.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703-714.

Trovato A, Seno F, Tosatto SCE. 2007. The PASTA server for protein aggregation prediction. PEDS 20:521-523.

Uversky VN, Gillespie JR, Fink AL. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins Struct Funct Bioinf 41:415-427.

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: Introducing the D2 concept. Annu Rev Biophys 37:215-246.

Uversky VN, Oldfield CJ, Dunker AK. 2005. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Rec 18:343-384.

Vacic V, Iakoucheva LM. 2012. Disease mutations in disordered regions-exception to the rule? Mol BioSyst 8:27-32.

Vihinen, M. 1987. Relationship of protein flexibility to thermostability. Protein Engin 1: 477-480.

Vihinen M, Torkkila E, Riikonen P. 1994. Accuracy of protein flexibility predictions. Proteins 19(2):141-9.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13:S2.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275-282.

Vihinen, M. 2014. Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24:356-364.

Vullo A, Bortolami O, Pollastri G, Tosatto SCE. 2006. Spritz: A server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucl Acids Res 34:W164-168.

Wang L, Sauer UH. 2008. OnD-CRF: Predicting order and disorder in proteins using conditional random fields. Bioinformatics 24:1401-2.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635-645.

Webb ME, Lobley CMC, Soliman F, Kilkenny ML, Smith AG, Blundell TL, Abell C. 2012. Structure of *Escherichia coli* aspartate [alpha]-decarboxylase Asn72Ala: Probing the role of Asn72 in pyruvoyl cofactor formation. 68:414-417.

Wigren E, Bourhis J, Kursula I, Guy JE, Lindqvist Y. 2010. Crystal structure of the LMAN1-CRD/MCFD2 transport receptor complex provides insight into combined deficiency of factor V and factor VIII. FEBS Lett 584:878-882.

Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. 2001. The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. Pac Symp Biocomput 6:89-100.

Wootton JC. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. Comput Chem 18:269-285.

Worrall EG, Worrall L, Blackburn E, Walkinshaw M, Hupp TR. 2010. The effects of phosphomimetic lid mutation on the thermostability of the N-terminal domain of MDM2. J Mol Biol 398:414-428.

Yang ZR, Thomson R, McNeil P, Esnouf RM. 2005. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369-3376.

Yue P, Melamud E, Moult J. 2006. SNPs3D: Candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.

Zhang J, Zhang F, Ebert D, Cobb MH, Goldsmith EJ. 1995. Activity of the MAP kinase ERK2 is controlled by a flexible surface loop. Structure 3:299-307.

Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. 2012. SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn 29:799-813.

Table 1: The evaluated protein disorder prediction programs.

Method	URL
Disorder prediction programs	
Anchor	http://anchor.enzim.hu/
DisEMBL	http://dis.embl.de/
DISOclust	http://www.reading.ac.uk/bioinf/DISOclust/
DISOPRED2	http://bioinf.cs.ucl.ac.uk/disopred
DISpro	http://scratch.proteomics.ics.uci.edu/
FoldIndex	http://bioportal.weizmann.ac.il/fldbin/findex
GeneSilico Metadisorder	http://genesilico.pl/metadisorder/
GlobPlot	http://globplot.embl.de/
iPDA	http://biominer.bime.ntu.edu.tw/ipda/
IUPred	http://iupred.enzim.hu/
MetaPrDOS	http://prdos.hgc.jp/cgi-bin/meta/top.cgi
MFDp	http://biomine-ws.ece.ualberta.ca/MFDp
OnDCRF	http://babel.ucmp.umu.se/ond-crf/
POODLE	http://mbs.cbrc.jp/poodle/
PrDOS	http://prdos.hgc.jp/cgi-bin/top.cgi
RONN	http://www.strubi.ox.ac.uk/RONN/
Softberry PDISORDER	http://linux1.softberry.com/berry.phtml?topic=pdisorder&group=programs&subgroup=propt
Spritz	http://protein.cribi.unipd.it/spritz/
WinDiso	http://prodata.swmed.edu/disorder/disorder_prediction/predict.cgi

Table 2: The performance of the disorder prediction programs on wild type residues of proteins

Disorder prediction program	TP	FP	TN	FN	TP%	FP%	TN%	FN%	TP%+TN%	Number of predicted variants
Anchor	0	10	82	9	0	9.9	81.2	8.9	81.2	101
DisEMBL[a]	7	45	49	0	6.9	44.6	48.5	0	55.4	101
DisEMBL[b]	4	15	77	5	4	14.9	76.2	4.9	80.2	101
DisEMBL[c]	0	4	90	7	0	4	89.1	6.9	89.1	101
DISOclust	2	17	77	5	2	16.8	76.2	4.9	78.2	101
DISOPRED2	1	3	89	8	1	3	88.1	7.9	89.1	101
Dispro	1	3	91	6	1	3	90.1	5.9	91.1	101
FoldIndex	1	16	78	6	1	15.8	77.2	5.9	78.2	101
GeneSilico	6	8	83	4	5.9	7.9	82.2	3.9	88.1	101
GlobPlot	1	11	83	6	1	10.9	82.2	5.9	83.2	101
iPDA	4	13	77	3	4.1	13.4	79.4	3	83.5	101
IUPred	3	5	87	6	3	5	86.1	5.9	89.1	101
MetaPrDOS	1	5	89	6	1	5	88.1	5.9	89.1	101
MFDp	2	6	84	5	2.1	6.2	86.6	5.1	88.7	101
OnDCRF	4	9	81	3	4.1	9.3	83.5	3	87.6	101
POODLE-I	7	11	79	4	6.9	10.9	78.2	3.9	85.1	101
POODLE-L	6	7	83	5	5.9	6.9	82.2	4.9	88.1	101
POODLE-S[d]	1	4	90	6	1	4	89.1	5.9	90.1	101
POODLE-S[e]	1	3	91	6	1	3	90.1	5.9	91.1	101
POODLE-W	0	6	88	7	0	5.9	87.1	6.9	87.1	101
PrDOS	1	5	89	6	1	5	88.1	5.9	89.1	101
RONN	5	8	81	4	5.1	8.2	82.7	4	87.8	101

Softberry PDISORDER	5	12	77	5	5.1	12.1	77.8	5	82.9	99
Spritz[f]	1	3	89	8	1	3	88.1	7.9	89.1	101
Spritz[g]	1	5	89	6	1	5	88.1	5.9	89.1	101
WinDiso	6	27	64	0	6.2	27.8	66	0	72.2	101

[a] Disorder by Loops/coils definition

[b] Disorder by Hot-loops definition

[c] Disorder by Remark-465 definition

[d] High B-factor residues

[e] Missing residues

[f] Short disorder

[g] Long disorder

Table 3: The performance of the disorder prediction methods for variants. True positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are indicated in the numbers of instances and in percentages.

	TP	FP	TN	FN	TP%	FP%	TN%	FN%	TP%+TN%	Number of predicted variants
Anchor	1	3	26	71	1.0	3.0	25.7	70.3	26.7	101
DisEMBL[a]	4	6	22	69	4.0	5.9	21.8	68.3	25.7	101
DisEMBL[b]	2	7	28	64	2.0	6.9	27.7	63.4	29.7	101
DisEMBL[c]	0	1	32	68	0.0	1.0	31.7	67.3	31.7	101
DISOclust	3	0	29	69	3.0	0.0	28.7	68.3	31.7	101
DISOPRED2	0	1	30	70	0.0	1.0	29.7	69.3	29.7	101
Dispro	0	0	33	68	0.0	0.0	32.7	67.3	32.7	101
FoldIndex	0	2	29	70	0.0	2.0	28.7	69.3	28.7	101
GeneSilico	1	4	33	63	1.0	4.0	32.7	62.4	33.7	101
GlobPlot	3	2	27	69	3.0	2.0	26.7	68.3	29.7	101
iPDA	1	0	32	68	1.0	0.0	31.7	67.3	32.7	101
IUPred	1	3	31	66	1.0	3.0	30.7	65.3	32.7	101
MetaPrDOS	0	0	33	68	0.0	0.0	32.7	67.3	32.7	101
MFDp	0	5	32	64	0.0	5.0	31.7	63.4	32.7	101
OnDCRF	6	4	34	57	5.9	4.0	33.7	56.4	39.6	101
POODLE-I	0	3	30	68	0.0	3.0	29.7	67.3	29.7	101
POODLE-L	1	2	30	68	1.0	2.0	29.7	67.3	30.7	101
POODLE-S[d]	1	1	33	66	1.0	1.0	32.7	65.3	33.7	101
POODLE-S[e]	3	0	33	65	3.0	0.0	32.7	64.4	35.6	101
POODLE-W	0	0	30	71	0.0	0.0	29.7	70.3	29.7	101

PrDOS	0	0	33	68	0.0	0.0	32.7	67.3	32.7	101
RONN	1	3	31	66	1.0	3.0	30.7	65.3	31.7	101
Softberry PDISORDER	0	2	30	67	0.0	2.0	30.3	67.7	30.3	99
Spritz [f]	6	3	30	62	5.9	3.0	29.7	61.4	35.6	101
Spritz [g]	1	3	34	63	1.0	3.0	33.7	62.4	34.7	101
WinDiso	5	8	27	61	5.0	7.9	26.7	60.4	31.7	101

[a] Disorder by Loops/coils definition

[b] Disorder by Hot-loops definition

[c] Disorder by Remark-465 definition

[d] High B-factor residues

[e] Missing residues

[f] Short disorder

[g] Long disorder

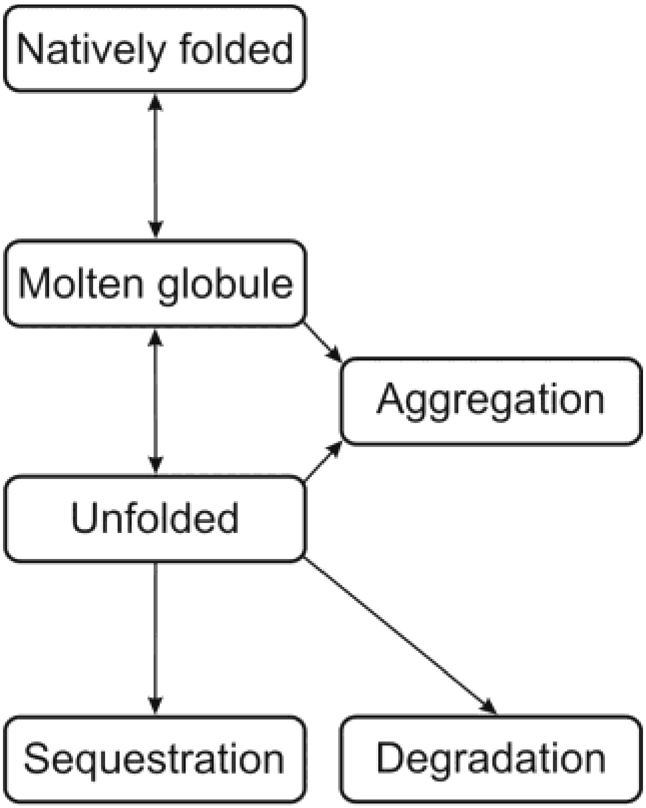


Figure 1. A diagram of the protein misfolding and the following steps.

Box 1: Keyword sets used in the text mining of the literature for cases where the change in protein disorder is caused by amino acid substitutions

First keyword set

1. disordered
2. unstructured
3. unfolded
4. (protein or intrinsic) disorder

Second keyword set

1. missense
2. point mutation
3. (single nucleotide or single-nucleotide) polymorphism
4. SNP$

Third keyword set

1. decrease$ or increase$ or reduce$ or become$
2. less or greater
3. high$ or low$
4. more disordered

Supplementary Table 1: Data set of 101 misfolding related amino acid substitutions in 31 proteins. Wild type indicates the order/disorder status in the original amino acid and variant protein for the variation.

Protein[a]	Organism	Amino acid substitution[b]	UniProt entry	Article reference	PMID	Wild-type	Variant	Experimental evidence[c]
α(1)-Antitrypsin	*Mus musculus*	E341K	P22599	Smith et al., 2011	21976666	O	D	immunoassays
Aldolase B, fructose	*Homo sapiens*	A150P	P05062	Stopa et al., 2011	21166391	O	D	CD
bisphophate (ALDOB)		A338V	P05062	Rellos et al., 1999	10229688	O	D	CD
Aspartate decarboxylase (ACD)	*E. coli* K12	N72A	P0A790	Webb et al., 2012	22505409	D	O	X-ray
Barnase	*Bacillus*	I98V	P00648	Buckle et al., 1993	8254677	O	O	X-ray
	velezensis	I123V				O	O	
		I135V				O	D	
		L136V				O	O	
		I143V				O	O	
Breast cancer 1, early	*H. sapiens*	M1652I	P38398	Rowling et al., 2010	20378548	O	O	fluorescence spectroscopy,
onset (BRCA1)		M1663L				O	O	equilibrium denaturation
		M1663K				O	O	
		L1664P				O	O	

		V1665M				O	O	
		A1669S				O	O	
		D1692N				O	O	
		D1692Y				O	D	
		R1699L				O	O	
		R1699Q				O	O	
		R1699W				O	D	
		G1706A				O	O	
		A1708E				O	D	
		S1715C				O	D	
		S1715N				O	D	
		S1715R				O	D	
		V1736A				O	D	
		R1751Q				O	O	
		L1764P				O	D	
		I1766S				O	D	
		T1773S				O	O	
		M1775R				O	O	

		D1778N				O	O	
		L1780P				O	D	
		M1783T				O	D	
		C1787S				O	O	
		G1788V				O	O	
		G1788D				O	D	
		P1806A				O	O	
		V1808A				O	D	
		V1833M				O	D	
		W1837G				O	D	
		W1837R				O	D	
		S1841N				O	D	
		A1843P				O	D	
		Y1853C				O	D	
Calmodulin (CALM1)	*H. sapiens*	V137G	P62158	Fefeu et al., 2000	11123919	O	D	NMR
Coat protein	*Sesbania*	W170E	Q9EB06	Pappachan et al., 2009	19643453	O	D	X-ray
	mosaic virus	W170K				O	D	
Creatine kinase, muscle (CKM)	*H. sapiens*	D54G	P12277	Feng et al., 2007	17030001	D	O	CD, fluorescence

								spectroscopy
Cyclin-dependent- kinase inhibitor	*S. cerevisiae*	S198E	P38634	Brocca et al., 2009	19280601	D	D	CD
SIC1		S198A	P38634			D	D	
EGF-containing fibulin-like extra-cellular matrix protein 1 (EFEMP1)	*H. sapiens*	R345W	Q12805	Roybal et al., 2005	16249470	O	D	immunoassays
Deoxyuridine 5'-triphosphate pyrophosphatase	Epstein-Barr virus B95-8	C4S	P03195	Freeman et al., 2009	19586911	O	D	X-ray
		D131N	P03195			O	O	
ENVZ	*E. coli* K12	A193L	P0AEJ4	Kishii et al., 2007	17635923	O	D	CD
Epilepsy, progressive myoclonus	*H. sapiens*	E28L	O95278	Liu et al., 2009	19403557	O	D	immunoassays
type 2A, Lafora disease (laforin)		W32D				O	D	
(EMP2A)		F84L				O	D	
		R108C				O	D	
		R171H				O	D	
		T194I				O	D	
		C266S				O	D	
		G279S				O	D	
		Q293L				O	D	
		Y294N				O	D	

		P301L				O	D	
Flavodoxin	*Anabaena* sp. PCC 7119	F99N	P0A3E0	Ayuso-Tejedor et al., 2011	21216251	O	O	SAXS
Heat shock 22 kDa protein 8	*H. sapiens*	K137E	Q9UJY1	Kasakov et al., 2007	17922839	D	D	fluorescence spectroscopy
(HSPB8)		K141E	Q9UJY1	Shemetov and Gusev, 2011	21767525	D	D	fluorescence spectroscopy
Hemochromatosis (HFE)	*H. sapiens*	H63D	Q30201	Liu et al., 2011	21349849	O	D	immunoassays
		C282Y	Q30201	de Almeida et al., 2007	17339458	O	D	immunoassays
		G20A	P20823	Lawless et al., 2007	17650303			immunoassays
HNF-1 homeobox A (HNF-1A)	*H. sapiens*	G20R		Narayana et al., 2006	16930618	O	D	X-ray, NMR, CD
		D39A	Q1R8K3			O	D	
Iron-sulfur cluster assembly scaffold protein	*E. coli* UTI89/UPEC	D81Y	Q8NI22	Kim et al., 2009	19492851	D	O	NMR
Multiple coagulation factor	*H. sapiens*	D89A	Q8NI22	Wigren et al., 2010	20138881	O	D	X-ray
deficiency (MCFD2)		D122V	Q8NI22		20138881	O	D	
		D129E	Q8NI22		20138881	O	O	

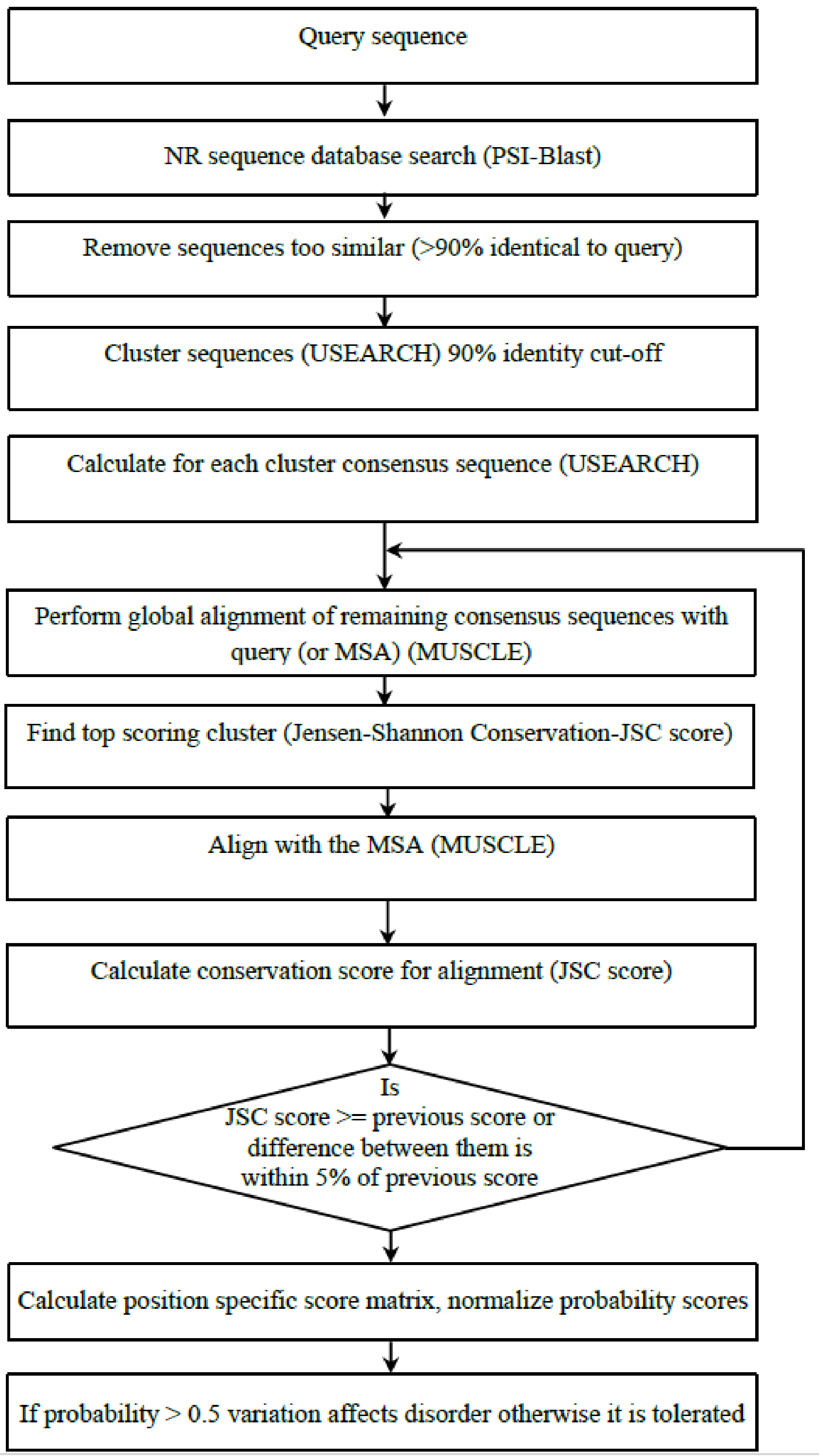
		Y135N	Q8NI22	Guy et al., 2008	18590741	O	O	NMR
		I136T	Q8NI22	Wigren et al., 2010	20138881	O	O	X-ray
		F99N	P0A3E0	Guy et al., 2008	18590741	O	D	NMR
Myosin-binding protein C, cardiac	*H. sapiens*	R654H	Q14896	Idowu et al., 2003	12787675	O	D	NMR
(MYBPC3)		N755K	Q14896			O	O	
NME/NH23 nucleoside diphosphate kinase 1 (NME1)	*H. sapiens*	S120G	P15531	Georgescauld et al., 2009	19186179	O	D	CD, fluorescence spectoscopy
Oncogene, E3 ubiquitin protein ligase (MDM2)	*H. sapiens*	S17D	Q00987	Worrall et al., 2010	20303977	D	D	fluorescence spectoscopy
Oxygen-evolving enhancer protein 1, chloroplastic	*Spinacia oleracea*	R235D	P12359	Popelkova et al., 2006	16503666	D	D	CD
		R235G	P12359			D	D	
		R245G	P12359			D	D	
Parkinson protein 7 (PARK7)	*H. sapiens*	M26I	Q99497	Lakshminarasimhan et al., 2008	18181649	O	D	X-ray
		A104T				O	D	
		E163K				O	D	
Plasma membrane [H^+] ATPase (PMA1)	*S. cerevisiae*	D378S	P05030	Han et al., 2007	17631501	O	O	immunoassays
Protein kinase, cAMP-dependent, catalytic, alpha (PRKACA)		R195A	P17612	Steichen et al., 2012	22334660	O	D	X-ray

Proteolipid protein (PLP1)	*H. sapiens*	W163L	P60201	Roboti et al., 2009	19825935	O	O	immunoassays
		I187T		Krämer-Albers et al., 2006	17093095	O	D	immunoassays
		A242V		Roboti et al., 2009	19825935	O	D	immunoassays
		G246A				O	O	
Pyrrolidone carboxyl peptidase	*Pyrococcus furiosus*	A199P	O73944	Iimura et al., 2007	17309236	O	D	NMR
Retinoschisin 1 (RS1)	*H. sapiens*	C59S	O15537	Gleghorn et al., 2009	19849666	O	O	immunoassays
		C110Y				O	D	
Rhodopsin (RHO)	*Rattus norvegicus*	P23H	P51489	Gorbatyuk et al., 2010	20231467	O	D	immunoassays
Ribosomal protein L9 (rPL9)	*H. sapiens*	I101A	P32969	Shan et al., 2010	20225821	O	D	CD, NMR

[a] Systematic HGNC names for human proteins.

[b] Numbering according to the reference sequence.

[c] Methods employed: CD, circular dichroism; NMR, nuclear magnetic resonance spectroscopy; SAXS, small angle X-ray scattering; X-ray, X-ray crystallography.


Query sequence
NR sequence database search (PSI-Blast)
Remove sequences too similar (>90% identical to query)
Cluster sequences (USEARCH) 90% identity cut-off
Calculate for each cluster consensus sequence (USEARCH)
Perform global alignment of remaining consensus sequences with query (or MSA) (MUSCLE)
Find top scoring cluster (Jensen-Shannon Conservation-JSC score)
Align with the MSA (MUSCLE)
Calculate conservation score for alignment (JSC score)
Is JSC score >= previous score or difference between them is within 5% of previous score
Calculate position specific score matrix, normalize probability scores
If probability > 0.5 variation affects disorder otherwise it is tolerated

Supp. Figure S1. Flow chart for calculating evolutionary features.