Proteus: A Random Forest Classifier that Predicts Disorder-to-Order Transitioning Binding Regions in Intrinsically Disordered Proteins

Sankar Basu[1], Fredrik Söderquist[1], Björn Wallner[1]*

[1]Bioinformatics Division, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

* To whom correspondence should be addressed

Abstract

The focus of the computational structural biology community has taken a dramatic shift over the past one-and-a-half decade from the classical protein structure prediction problem to the possible understanding of intrinsically disordered proteins (IDP) or proteins containing regions of disorder (IDPR). The current interest lies in the unraveling of a disorder-to-order transitioning code embedded in the amino acid sequences of IDPs overtaking the well established sequence to structure paradigm. Disordered proteins are characterized by enormous amount of structural plasticity which makes them promiscuous in binding to different partners, multi-functional in cellular activity and atypical in folding energy landscapes resembling partially folded molten globules. Also, their involvement in several human diseases including cancer, cardiovascular, and neurodegenerative diseases makes them both attractive as drug targets, as well as important for a biochemical understanding of the diseases. The study the structural ensemble of IDPs is rather difficult, in particular for transient interactions. When bound to a partner the IDPRs adapt to an an ordered structure in the complex. The residues that undergo this disorder-to-order transition are called protean residues, and the first step in understanding the interaction with a disordered partner would be to predict the residues that are responsible for the interaction and will undergo disorder-to-order transition, i.e. the protean residues. There are a few available methods which predict these protean segments given their amino acid sequences, however, their performance reported in the literature leaves clear room for improvement. In this background, the current study presents 'Proteus', a random-forest-based protean predictor that predicts the likelihood of a residue to undergo disorder-to-order transition upon binding to a partner protein. The prediction is based on features that can be calculated using the amino acid sequence alone. Proteus compares favorably with existing methods predicting twice as many true positives as the second best method (55% vs. 27%) at a much higher precision on an independent data set. The current study also shades some light on a possible 'disorder-to-order' transitioning consensus, untangled, yet embedded in the amino acid sequence of IDPs. Some guidelines have also been suggested to proceed for a real-life structural modeling of an IDPR using Proteus.

Keywords: Intrinsic Disorder, Protean, Random-Forest, disorder-to-order Transition, Topography length

Short Title: Prediction of Protean Segments by Proteus

Software Availability: https://github.com/bjornwallner/proteus

Introduction

After extensive research over one and a half decades, it is now evident that many functional proteins lack well-folded 3D structures which could either be intrinsically disordered proteins (IDPs) or could contain intrinsically disordered protein regions (IDPRs) [1–4]. In contrast to the classical view of protein folding [5], where a nascent cytoplasmic polypeptide chain immediately begins to fold into a stable three-dimensional globule (within the limits of their essential dynamics [6]) even while being synthesized [7,8], these proteins are born disordered [3] and remain either completely or partially unstructured throughout their entire life span. It is only when they interact with functionally relevant suitable binding partners, they switch to stable ordered structures [4].

They are highly abundant in nature and involved in a number of functions within living cells, most of which belong to the non-classic (non-enzyme) type [9,10]. They possess remarkable binding promiscuity [4] in a wide range of intermolecular interactions, complementing the functional repertoire of ordered proteins, likewise to the phenomena of enthalpy-entropy compensation [11]. The promiscuity is primarily manifested by their ability to interact specifically with structurally unrelated partners and thereby gaining different structures upon binding. It is highly likely that these peculiar characteristics are attributed by their non-native-like rough and relatively flat energy landscapes [12,13], whereupon the favored conformations closely resemble to the partially folded molten-globules [13] and also the ability to keep necessary amount of disorder even in the bound form [4]. Considering this flexible nature, they have been referred to be part of the 'edge of chaos' systems [14], serving as a bridge between well-ordered and chaotic systems – extremely critical in the context of a living cell. In addition to these peculiar biophysical and folding attributes, they are also of considerable biomedical interests due to their functional importance. In fact, their existence in a biologically active form without adapting to unique 3D-structure contradicts the traditional notion of “one protein–one structure–one function” paradigm [1]. In particular, they are involved in regulation, signaling, and control, where high specificity / low-affinity interactions [15] are crucial. Recent studies have also highlighted their multifarious key activities as molecular rheostats, molecular clocks, in tissue specific and alternative splicing of m-RNA, transport of r-RNA and as protein and RNA-chaperons [16]. Also, the unique structural feature of this 'intrinsic disorder' enables IDPs to participate in both one-to-many and many-to-one signaling [2]. The promiscuity in binding also suggests that not only misfolding [17], but also misidentification or mis-signaling [2] in biomolecular recognition could serve as the root cause of some extremely complex human diseases [3] including cancer, diabetes, amyloidoses, cardiovascular and neurodegenerative diseases [18].

All these factors pile up to raise an increasing demand for greater structural knowledge on IDPs, presenting a tough challenge to crystallographers owing to their inherent disorder and thereby providing a realistic scope for computational model building tightly coupled with realistic confidence estimates. According to the most popular description of IDPRs, only a subset of them can undergo the 'disorder-to-order' transitions and thereby adapt an ordered structure only via binding to a suitable protein partner, giving birth to the concept of 'folding coupled with binding' [19]. There are different terms in the literature to address these segments but the most popular is perhaps the term 'protean' [19] borrowed from Greek mythology, meaning 'ever-changeable' or 'mutable'. Thus, in an attempt to solve the structure of an IDPR, a computational structural biologist should first aim to predict the potential 'disordered' regions followed by the prediction of the potentially 'mutable' protean regions. Important to note that, due to intrinsic disorder, these regions in an isolated X-ray structure are presented as 'missing electron density' regions and, in principle, should only appear as available 3D coordinates in a complex form, bound to a competent molecular partner. In fact this is one of the more established definitions of the 'protean' segments amongst a few similar ones.

Thus, it is highly challenging to decipher the root cause of intrinsic disorder from pure sequence-based investigation given the limitation of available structural data. Concerted efforts have been devoted in that line which include formulation of statistical mechanical potentials describing sequence-derived elasticity (or plasticity) [20], and even proposition of an alphabet of intrinsic disorder [21]. Machine learning algorithms have been extensively used to develop knowledge-based Predictors which can not only predict the disordered regions [22–26], but also the 'protean' segments [26–29]. However, the 'protean prediction' part can still be regarded as at a very early stage, offering much room for improvement. In this background, the current study not only attempts to shade some light on a possible yet unexplored 'sequence consensus' of such 'disorder-to-order' transitions, but also presents a random forest classifier, namely 'Proteus', which predicts the potential protean (or protein binding) segments from the amino acid sequence of an IDP. Proteus compares favorably to the existing predictors.

Materials and Methods

Training dataset

Two databases containing proteins with annotated protean segments were pulled together to build the final training dataset: IDEAL and MoRF. IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature [19]) contains 557 proteins with experimentally verified protean segments. However, only 203 of 557 proteins in this database actually contain protean segments. The rest are intrinsically disordered proteins where no protean segments has yet been experimentally verified and thus serve as negative examples in training. MoRF comes from MoRFpred [28], one of the existing classifiers. It contains 840 proteins, and all of them has at least one protean segment. More importantly, all members of MoRF has direct structural evidence from the PDB. Members from IDEAL and MoRF will henceforth be referred to as 'ProS' and 'MoRF' respectively, and the combined dataset as 'PnM'. The details of all datasets have been enlisted in **Table 1.**

Independent Validation Benchmark

The proposed methodology is somewhat limited in the number of available targets in form of an independent validation set and uses the same 9 proteins that were used in the DISOPRED3 benchmark [26]. DISOPRED3 initially culled 29 chains by database annotations and scientific reports which were then reduced to this small set of 9 proteins, as the rest of the chains were found to be used in the training datasets of the competing methods, ANCHOR [27], MorFPred [28] and MFSPSSMPred [29].

Target Function

The 1-0 binary status for each amino acid residue in the sequence (as assigned by the corresponding training dataset) serves as the target function in training the classifier, Proteus. Protean and non-protean residues are denoted by 1 and 0 respectively, meaning positive and negative examples in training.

Data Clustering and Cross-validation Benchmark

To avoid training and testing on similar examples, BLASTclust was used to cluster the protein sequences in the combined dataset 'PnM'. Sequences with a pairwise similarity of at least 30% over at least 50% of the sequence length (-S 30 -L 0.5) were clustered in the same group (fold). This resulted in 774 clusters, with the largest cluster containing 38 proteins, and 253 clusters containing more than one protein. $1/3^{rd}$ of all ProS sequences were found to be similar to at least one MoRF sequence and vice-versa.

The five groups were built while keeping together all proteins in the same cluster. As the proteins

vary largely in length, it was not possible to maintain equality in size (in terms of the number of amino acid residues) amongst the five-folds. Instead, care was taken to keep the number of target proteins consistent amongst the groups: 280 in four of them and 279 in the fifth. The number of examples varied from 158,651 for the smallest to 218,870 amino acid residues for the largest group, whereas the proportion of positive examples (i.e., predicted 'disordered' residues) ranged from 1.42 to 2.23%. It should be noted, that, each of the 5-fold cross-validation training set has been trained on $4/5^{th}$ of the whole training data exclusive of the corresponding test examples ($1/5^{th}$).

Another independent BLASTClust was run with identical constraints on the combined training and independent validation test to ensure that they shared no homologs. One of the ProS proteins clustered with one protein from the independent validation set and was thus removed from the combined training set, PnM.

Random Forest Classifier

The Random Forest Classifier (RFC) module part of the Python machine learning package scikit-learn [30] was used for training. Every decision tree in the forest get to vote for the binary classification of every example, and the examples are classified positive according to the majority vote. In parallel, Extremely Randomized Trees (ERT), a more randomized variant of the Random Forest was also used in training, which is not only slightly cheaper to compute, but also, makes the classifier more resistant to over-fitting (i.e., lower variance). However, in the course of lowering the variance, ERT increases the bias slightly, owing to the heavy randomization involved in the classifier causing it to miss low-relevance features.

Evaluation Measures

In binary classification, there are four possible outcomes when predicting (i.e. classifying) an example: (i) True Positive (TP): a positive example correctly classified as positive; (ii) True Negative (TN): a negative example, correctly classified as negative; (iii) False Negative (FN): a positive example incorrectly classified as negative; and (iv) False Positive (FP): a negative example incorrectly classified as positive. Based on these four possible outcomes and their corresponding counts, the following evaluation measures were calculated.

Precision

Precision, also known as specificity or the Positive Predicted Value (PPV), measures how many examples classified as positive were actually positive, and, calculated by the ratio, TP / (TP + FP).

Recall

Recall (or coverage) measures how many positive examples were correctly classified as (true) positives. It is also called the 'True Positive Rate' (TPR) and calculated by the ratio, TP / $\sum P$, where $\sum P$ is the total outcome positives, i.e., $\sum P$ = TP + FN.

F1-score

F1-score is the harmonic mean between PPV and TPR and could be interpreted as a trade-off between PPV and TPR. It is defined by the following equation: $F1=2PPV\times TPR/(PPV+TPR)$.

Matthews Correlation Coefficient

Another direct evaluation measure is the Matthews Correlation Coefficient (MCC) ranging between -1 (perfect inverse prediction) to +1 (perfect prediction) and calculated as: $MCC=((TP\times TN)-(FP\times FN))/((TP+FP)(TP+FN)(TN+FP)(TN+FN))^{1/2}$ and was used in conjugation with the F1-score to estimate the overall performance of the predictor.

Tuning Training Parameters

Decision tree depth

In general the deeper the tree, the more complex patterns it can fit. However, this can easily lead to over-fitting. Thus, finding an optimal tree depth is important. The number of leaves (N) in a decision tree is an exponential function of the leaf-depth (d) of the tree ($N = f(d) = 2^d$) which increases by successive binary branching, leading to a geometric sequence (N = 1, 2, 4, 8, 16, 32 …).

Since the combined training dataset (PnM) contained ~850,000 examples (which is close to 2^{20}), the maximum depth was varied between 1 and 25 (**Supplementary Fig.S1**). Random Forest Classifier (RFC) gave slightly better results (MCC, MCC+F1 scores) than Extremely Randomized Trees (ERT), and a depth of 13 ($2^{13} = 8192$ possible leafs) yielded the highest MCC and F1 scores. Hence, all further experiments were designed using RFC with a maximum decision tree depth of 13.

Number of trees in the forest

Another important parameter is how many decision trees to use. In theory, the more trees the better, but there is a saturation in performance, beyond which the increase in performance is only marginal. Therefore, it is important to find the optimal number of trees to save computation time. As can be seen from the **Supplementary Fig.S2**, 50 decision trees yield a reasonable performance, which is only slightly increased (by ~5%) by using more trees. Therefore, using 50 trees was considered to be enough for the computationally expensive feature selection part. However, for the final selected combination of features 500 trees were used to achieve maximum performance.

Probability cut-off

The classifier needs a user-defined probability cutoff (P_{cut}) above which an example is classified as positive. P_{cut} was varied in the whole range of 0.0 to 1.0 and based on the performance (**Supplementary Fig.S3**), was set to 0.5 (majority vote). Therefore, if 50% or more decision trees voted for the particular example to be positive, it was classified as positive.

Amino Acid Propensity

Propensity (Pr) for a particular amino acid, X to occupy a particular 'class' (e.g. protean vs. disordered residues) has been calculated as the ratio of two probabilities (P) as: $Pr(X) = P(X)_{class}/P(X)_{full} = (N(X)_{class}/N(All)_{full}) / (N(X)_{full} / N(All)_{full})$ where 'full' stands for the entire training dataset and N denotes the raw count of amino acid(s) in the said 'class'. A propensity value of 1 represents no preference whereas a higher and lower value to that of 1 represents higher and lower preference of the amino acid to occupy the given class with respect to the baseline, usually taken to be the whole dataset.

Secondary Structural Content

PSIPRED [31] was used to predict the secondary structure in three classes (H: Helix, E: Strand, C: Coil). For each amino acid, the relative fraction of each of the three main secondary structural classes (H, E, C) were calculated for protean, non-protean, disordered and ordered sequences. The aim was to decipher if there was any preference in disorder vs. order sequences that might have propagated to protean segments during the 'disorder-to-order' transitions.

Design of the sequence-driven features

Consideration of local and global effects

Intrinsic disorder, which is essential for the 'protean' segments, is a function of both the local sequence and the global three dimensional fold of the protein. The design of features should give proper weights to both. However, it is highly non-trivial to take into account, the global effect of the

overall protein fold without actually attempting to build homology models for the predicted 'structured' regions, obviously in their bound form. This will not only be computationally costly but will also have a very low confidence associated with the built models, due to the lack of enough structural data. One alternative way to indirectly take into account the global constraints is to perform a homology search against all sequences and then convert the sequence into a profile. To this end, PSI-BLAST [32] was used to construct sequence profiles. In addition, PSI-PRED [31] was used to predict secondary structure of each amino acid residue of each input sequence and DISOPRED [26] to predict their disorder probability score. This is an implicit way to account for the possible global constraints in the designed features.

To describe the neighboring environment a sliding window of 15 residues centered around the current residue was considered in the design of most features. This will produce an average property of the said feature, taking into account the local sequence dependence associated with order-to-disorder transitions. The size of the window was optimized by trying different sizes in the range of 9-21. The optimal size agrees with average length of protean segments (**Fig.1**).

In total 342 features, in seven different feature groups, were used and described in detail below (**Table 2**)

Feature Group 1: Sequence Profiles (Features: 1-300)
Considering the influence of the local sequence to disorder, it is likely to find empirical trends (over and under-representations) in the distribution of amino acids in protean compared to non-protean regions. In other words, certain amino acids might preferentially occur in the protean segments but not others. This was represented by Position Specific Scoring Matricies (PSSM) constructed by running three iteration (-j 3) of PSI-BLAST [32] against UniRef90 [33] with an inclusion E-value threshold of 10^{-3} (-h 0.001). The PSSM contains scores for each of the 20 possible amino acid substitutions in each position, representing the amino acid mutability at any given position. The higher the score, the higher the probability that these amino acids occurs at that position. To improve convergence, the raw PSSM scores were linearly scaled to [0.0, 1.0] based on the maximum and minimum values observed for each amino acid in the whole training set. To account for the local sequence bias a 15 residues window of the PSSM was used centered around the current residue, giving 300 (15x20) features in total for each residue.

Feature Group 2: Amino Acid Conservation (Feature: 301)
The conservation score is derived by PSI-BLAST [32] from the PSSM matrix, and, as the name suggests, conceptually, it is complementary to that of 'mutability'. Numerically, it is a modified Shannon Entropy [34] term representative of the heterogeneity of amino acid substitutions for a given position in the input sequence. Again, to take care of the neighboring environment, the conservation score was averaged over a 15 residue window. In contrast to all other feature groups, this group consists of only a single value.

Feature Group 3: Amino Acid Composition (Features: 302-321)
This feature group describes the individual concentration of all amino acids, in a 15 residue long window, i.e. 20 features in all. Representing a coarse-grain estimation of the amino acid properties in the local neighborhood around the central residue.

Feature Group 4: Amino Acid Properties (Features: 322-330)
Physiochemical properties of amino acids might serve crucial consensus for disorder and disorder to order transitions. In contrast to the 'amino acid composition group' above, Polarity, Charge, Hydrophobicity and Molecular Weight were explicitly described, in a 15 residue sliding window. Polarity were divided into polar, non-polar, acidic-polar or basic-polar, and Charge into positive, negative, and neutral [35]. Hydrophobicity was described using the Kyte Doolittle scale [36] For

each of these seven features, the corresponding counts were averaged over the 15 residue window.

Feature Group 5: Predicted Secondary Structure (Features: 331-333)
Secondary structural propensities of individual amino acids in the close neighborhood of a residue might have major influence on disorder and might serve as a discriminative feature between protean and non-protean fragments. For example, if this likelihood continually keeps altering between helices to sheets along the sequence, the resultant main-chain trajectory would potentially only keep wobbling giving rise to an unstructured region. The other possibility is of course having most residues as predicted 'random coils'. The probabilities of each amino acid residue in a sequence to form one of the three main secondary structures (Helix, Strand, Coil) were predicted by PSI-PRED[31] and averaged over a 15 residue sliding window, serving as three distinct features.

Feature Group 6: Predicted Disorder Probability (Features: 334-340)
The probability for disorder was predicted using DISOPRED [26]. The disorder prediction score from DISOPRED is a confidence estimate (or probability) for a residue in a protein sequence to be disordered. It is defined in the range [0, 1] and DISOPRED assigns the disordered status to a residue if the score is greater than 0.5. The disorder prediction score, averaged over the 15 residue window centered on the current residue was directly used as the first feature in this group. In addition, to describe the local properties of the disorder prediction, the length of disordered and ordered segments and the start and end positions relative to the total sequence length were also used. In detail, if the score is greater than 0.5, the positions on either side of the current residue where the score drops below 0.5 are identified, from this the length, start and stop positions of the segment can be calculated. This was performed for residues predicted to be disordered (score > 0.5) and for residues predicted to be ordered (score < 0.5), resulting in 7 (1+3+3) features and depending on the predicted disorder of the segment three of the seven features will always remain zero.

Feature Group 7: Disorder Topography (Features: 341-342)
Disorder topography measure the topography of peaks and valleys in the predicted disorder score graph (**Supplementary Fig.S4**). Each residue is classified as being part of a peak (1), valley (-1) or neither (0). A residue is part of a peak if on both sides there exists another residue with a score at least 10% lower than the current residue. Likewise, a residue is part of a valley if there are residues with disorder scores at least 10% higher than the current residue. If a residue is neither at a peak or in a valley it is classified as neither. In addition, the length of the current peak or valley residue are also calculated and used as a separate feature. Thus, the disorder topography feature consist of the peak/valley/neither classification and the length of the current peak/valley.

Results and Discussion

Propagation of sequence consensus during disorder-to-order transitions

To understand the possible relationship between sequence-derived properties of protean segments and the associated disorder, sequence-derived properties like amino acid propensities and secondary structural content were individually studied in disorder vs. order and protean vs. non-protean regions. This knowledge will serve not only to explore and understand certain plausible empirical trends in the designed features, but also as a guide in determining which features are more discriminative, and which can act like filters.

All characteristics were investigated in (i) protean vs. non-protean as well as in (ii) disordered vs. ordered sequences (as predicted by DISOPRED). The aim was to identify any pattern that might be responsible for disorder-to-order transitions of the protean segments.

In other words, the focus was to collect the most discriminative trends in the disordered vs. ordered regions that were also maintained in the protean vs. non-protean segments. Hence, they could be interpreted as properties propagating during the disorder-to-order transitions, implicit in the protean segments. However, since the 'disorder vs. order' classification is clearer and more distinct, it is expected that the trends for 'disorder vs. order' should be more prominent than the 'protean vs. non-protean' trends.

Amino acids preference in protean and disorder residues

The first and most fundamental characteristic investigated was the propensity of amino acids in disorder/ordered and protean/non-protean residues. The predicted disordered regions show drastic under-representations of hydrophobic amino acids compared to predicted ordered regions (**Fig.2A**). Even among the distribution of hydrophobic amino acids, there is an unmistakable trend with respect to the size of the hydrophobic side-chain. The gradual increase in the propensity of the hydrophobic side-chains in the predicted ordered regions appears directly proportional to their side-chain volume (Ala → Val → Leu → Ile → Phe → Tyr → Trp) (**Fig.2B**); whereas in the predicted disordered regions, the relationship appears inversely proportional. This trend is perfectly consistent with the notion of hydrophobic core formation within ordered protein tertiary structures [37], and on the other hand, bulky aromatics (Phe, Tyr, Trp) should be unfavorable in disordered regions, due to their potential incompatibility with regard to side-chain volume and entropy. The other noticeable features are the significant over-representation of cysteines in ordered regions with a concomitant under-representation in disordered regions, again consistent with the idea of fold stabilization by disulfide bridges [38] in the ordered, structured proteins, which must be avoided during the natural design of intrinsic disorder. On the other hand, prolines are significantly over-represented in disordered regions compared to ordered, which is consistent with their ability to break regular secondary structures [39], especially helices [40]. Even if found in regular secondary structures (β-sheets for example), proline needs additional structural constraints from pre-prolines (e.g., glycine rescue) to get stabilized [41]. In line with these observations, proline has been identified as the most disorder promoting amino acid residue [21].

The other well-known residue, responsible for backbone flexibility, glycine [39] was also found to be over-represented in disordered compared to ordered regions. This is in accord with the well-established idea that proline and glycinces are general indicators of entropic elasticity [20] and hence control self-organization of elastomeric proteins (e.g., amyloid fibrils) [42]. In fact, recent studies have formulated correlation functions of elasticity in terms of coiling propensity based on sequences rich in proline and glycines in disordered proteins [20].

The other noticeable difference was seen for serine, again a small and polar amino acid, significantly over-represented in disordered and under-represented in ordered regions. Indeed, serine-rich proteins in bacterial enzymes like kinases [43] and eukaryotic splicing factors [44] have been reported to be part of intrinsically disordered proteins. The other polar (Thr, Asn, Gln) and charged (Asp, Glu, Lys, Arg) amino acids were found to have similar or slightly higher propensities in disordered compared to ordered sequences. These results agree well with the previously proposed alphabet of intrinsic disorder [21].

But as mentioned earlier, the focus of the current work was to identify patterns that were not only discriminative in disorder vs. order sequences but were also maintained in protean vs. non-protean sequences and therefore might form crucial consensus in the understanding of disorder-to-order transitions. However as expected, the patterns in protean vs. non-protean sequences were not as prominent as in disorder vs. order sequences (**Fig. 3**). The collection of all (non-ProS + non-MoRF) sequences served as the (non-protean) baseline which raised a value of ~1.00 (+/- 0.01) for the baseline propensities of all the amino acids (**Fig. 3B**). This was not surprising since the bulk majority of the training dataset contained negative examples (non-protean sequences). Similar to

ordered residues, all large hydrophobic residues (Leu, Ile, Phe, Tyr, Trp) were found to be over-represented in the protean segments (**Fig. 3A**) and at the same time these residues were drastically under-represented in disordered regions (**Fig. 2A**). The same is true for all charged residues (Glu, Asp, Lys, Arg) that acquired much larger propensities compared to what they had in disordered sequences, and also noticeably higher than ordered sequences in general (**Fig. 2B and Fig. 3A**). The results clearly indicate that both large-hydrophobic and charged residues get preferentially selected during the 'disorder-to-order' transitions (via binding). In other words, not all disordered regions undergo the same transition, rather, there is a preferential selection of sequences containing large hydrophobic and charged residues leading to stabilization through hydrophobic and salt bridge interactions at the protein-protein interface. This is in accord with the general notion of stability upon binding in protein-protein interfaces where both shape and electrostatic complementaries are crucial for binding [45,46].

Finally, as for disorder residues cysteines are clearly under-represented in protean residues as well, reflecting the fact the stability of protean residues should not involve disulfide bridges (at the cost of massive loss of plasticity). However, in contrast to disordered residues both proline and glycine are under-represented in protean residues, indicating these residues do not undergo disorder-to-order transition, but rather remain disordered.

Secondary structure preference in protean and disorder residues

It is also important to conceptualize the secondary structural trends during the course of disorder-to-order transitions. The relative content of coil (C), including loops and turns are higher than helix (H) and strands (E) in all classes of sequences ranging from disorder to order and from protean to non-protean. But when comparing between two opposite class (e.g. disordered vs. ordered), it is the relative increment in (H+E)/C that is interesting. On that note, ordered sequences naturally have far greater regular secondary structures (H+E) amounting to ~50% of the whole population than disorder sequences (H+E: ~15%; C: ~85%) (**Fig.4**). As expected, the relative low fraction (~15%) of helices and strands in disorder residues has a definite rise upon the disorder-to-order transitions in protean segments (H+E:~40%), which is roughly the same as in non-protean sequences (**Fig.5**). Recall that the large majority of the non-protean sequences are in fact the usual ordered sequences and the subset of disordered sequences that get ordered are only the leftover minority. Among the regular secondary structures, helices appear to be more prevalent in protean (~32%) than non-protean segments (~27%) whereas beta-strands seem to be slightly more preferred in non-protean (~10%) compared to protean segments (~5%).

Indecisiveness in adapting a particular secondary structure class from sequence

Another property investigated based on secondary structure is the indecisiveness of an amino acid sequence in adapting a particular secondary structure. This was based on the assumption that protean segments, when disordered in isolation, might keep on altering the choice to adapt a particular secondary structure (H, E or C) along their main-chain trajectory and thereby end up being unstructured. Given the current lack of structural data for these sequence, PSIPRED [31] was used to predict secondary structure to try to shade some light on the above hypothesis. A measure for the indecisiveness or randomness in secondary structure prediction called Altscore was defined as the average number of transitions (H→C, C→E etc.) for each protean and non-protean segment. Regions with an Altscore value of 'zero' were omitted for both protean and non-protean regions, since they will only add noise to any potential signal. Focusing on the regions with Altscore> 0, the frequency distribution (**Fig.6**) clearly discriminate between protean and non-protean classes with a wider spread being obtained for the protean class followed by a peak-shift towards higher values (0.1 compared to 0.05 for non-protean). The results indicate that the intrinsic disorder associated with the unbound protean segments potentially suffers from the indecisiveness of the main-chain trajectory to adapt a particular secondary structure.

Both the above observations, (i) the reappearance of large hydrophobic and charged amino acids into the protean segments, as well as (ii) the indecisiveness associated with their predicted secondary structures should serve constructively in unraveling a hidden consensus in promoting disorder-to-order transition.

Training a classifier to predict protean residues

To be able to predict protean residues from sequence, a random forest classifier was trained on the features described above. Most features have been calculated using a sliding window of 15 residues, optimized by trying different window sizes in the range of 9-21 and maximizing the performance (**Supplementary Fig.S5**). The optimal window size is in the center of the distribution of the length of protean segments (**Fig. 1**) and similar to what is used for predicting disorder residues[26]. Note that for all feature groups except Feature Group 1: Sequence Profiles, the number of features will remain the same even with a different window size. Among all features, some features might be non-informative, other might be redundant. Indeed, some features are similar in their physiochemical descriptions and therefore might be excluded without loss in performance. But sometimes it might be an advantage for the classifier to learn from explicit rather than implicit features. To find the best combination of the 7 feature groups, all 127 possible combinations were exhaustively examined by measuring the final cross-validated performance using MCC and F1-scores for each feature group combination.

The twenty best feature group combinations according to the MCC and F1-scores have been shown in **Supplementary Fig.S6** & **Fig.S7** respectively. The difference is small between the top feature group combinations. Also, the top-combinations as evaluated by MCC and F1 are not identical, whereas, using all features result in good scores being attained in both evaluations. Therefore, the combination of all feature groups was chosen judiciously. The absolute MCC and F1 score values are relatively small ~0.13, owing to a large number of false positives and negatives. However, the magnitude of the scores are comparable to other studies [26–29], and reflect the difficulty of predicting residues that will be ordered upon binding from information in one of the binding partners only. Further illustrated in the receiver operating characteristic (ROC) curves of the precision (PPV) vs. recall for the best combination (**Fig.7**). The ROC curves were constructed by varying the cutoff (P_{cut}) and calculating precision and recall for each cutoff. The random base line precision is 1.9% and the curve for the best combination is clearly above that and it can also be seen that 500 trees is slightly better than 50. But the question remains, if the rather modest 10% precision at 23% recall (P_{cut}>0.5) is useful at all? Considering that it is still five times better than expected by chance we would argue that it is useful given the alternative. But there is of course plenty of room for improvement, by incorporating additional information not directly obtained from the sequence, such as structure prediction to filter out residues that actually are ordered by themselves, and to predict the surrounding residues, and to be used as starting points in molecular simulations or docking studies.

Relative Importance of Features

In an effort to learn what features contributed to the overall prediction, the relative importance of each feature group was outputted from the classifier. To take care of the inherent randomness associated with the classifications, this relative importance was averaged over predictions of 500 decision trees. As we can see, there are three features that stand out above the rest (**Fig.8**): Feature 342 ('topographic length': Feature Group 6) is by-far the most important feature which describes the length of the topographic region where the current residue is located. Interestingly, the second most important feature (feature 340: Feature Group 5) is also a length descriptor, namely the more coarse-grained length of the ordered region corresponding to the current residue. Note that this feature will be 'zero' for all residues predicted to be disordered. The third most important feature (feature 334: Feature Group 5) is the predicted disorder score averaged over the current window size.

The other seven features in top ten were (4) the relative position of the ending residue with respect to the current one, which is detected to reside in an ordered region (feature: 339, group: 6, set to 'zero' if residue is predicted disordered), (5) length of the disordered region the current residue resides in (feature: 335, group: 6, set to 'zero' if the fragment is ordered), (6) the topography score (feature: 341, group: 7), (7) probability of the current residue to form a coil (feature: 333, group: 5), (8) probability of the current residue to form a helix (feature: 335, group: 5), (9) the relative position of the starting residue with respect to the current one, detected to reside in an ordered region (feature: 339, group: 6, set to 'zero' if the residue is predicted disordered), and (10) charge-neutrality of the current amino acid (set to 'zero' if charged).

True Positive Enrichment by analyzing the Proteus Score

A common test of machine learning predictors is to analyze the true positive enrichment by constructing score plots, which is more detailed compared to the ROC curves. Score plots are conventionally defined as the overlay of two independent evaluation measures, Positive Predicted Value (PPV) and recall as two distinct functions of the predicted score (the Proteus score in this case). Ideally, both the PPV and recall should be high but there is a conflict in finding as many true positive as possible (high recall) and at the same time have a high PPV (few false positives). In reality there will always be at a trade-off between the two, which is also the main reason to use the combined measure F1. In the current case (**Fig. 9A**), F1 peaks at around the score of 0.5, which is also the cutoff chosen for positive prediction in the final predictor (P_{cut}=0.5); corresponding to 10% PPV and 23% recall as discussed above. It can be noted that after that point the PPV increases quite rapidly, and scores >0.7 have PPV > 40%. Unfortunately there are rather few examples that obtain this high score resulting in a rather modest recall overall. Still, if the score is high we can certainly trust it to be a relatively accurate prediction. This is also reflected by analyzing the distribution of scores for protean and non-protean residues (**Fig. 9B**), where the score was found to be much higher for predicted protean residues than non-proteans with median values of 0.4 and 0.24 respectively with roughly equivalent median absolute deviations. It can also be seen that there are quite many high scoring outliers in the non-protean residues. These might of course be completely wrong, but there is also a possibility that these predictions are actually sites for yet unknown interactions. Since the study of transient interaction is difficult, and the focus of the structural biology community so far has been on stable interactions that can even form crystals, there is still a lot more to be discovered if the dynamics is also taken into account.

Benchmark on Independent Data Set

In any machine learning scheme it is an advantage if the final classifier can benchmarked on independent data, and against other classifiers. In the recent DISOPRED3 paper [26] the following methods were benchmarked ANCHOR [27] MoRFpred [28], MFSPSSMpred [29], and DISOPRED3 [26] using a set of 2,209 residues out of which 163 were protean (i.e., positive examples) from 9 proteins (see Material and Methods). None of the examples in the independent set were similar to any example used in training Proteus, thus before classifying, Proteus was retrained on the full non-cross validated training set. The predictions for the other methods were generously made available by the authors of DISOPRED3 through the following link: http://bioinfadmin.cs.ucl.ac.uk/downloads/DISOPRED/suppl_data/. The evaluation measures precision, recall, F1, and MCC were calculated for all methods using the binary classification of each method (**Fig.10**) or as ROC curves using the raw scores from each method (**Supplementary Fig.S8**), and overall Proteus is better in all measures. Proteus has the highest precision (0.26 compared to 0.22 for DISOPRED3, the second best), for a much larger recall (0.56 compared to 0.28 by ANCHOR, the next best). This combined improvement in both precision and recall is also naturally reflected in a concomitant increase in their trade-off, the F1-score (0.35 compared to 0.18 by DISOPRED3, the next best). It also attained a higher MCC value than the other methods (0.30 compared to 0.13 by DISOPRED3). Even though the independent set is small, the high recall is

particular encouraging if Proteus should be used as initial step before more elaborate approaches as discussed above, it is crucial not to lose too many true positives at an early stage.

Conclusion

With the realization that protein disorder is involved in a range of human diseases, including cancer, cardiovascular and neurodegenerative diseases, it is important to compile more and more structural information for these proteins to under their modus operandi. A first step in this direction is classification and prediction of protean segments. The literature shows that there is indeed much room for improvement for the existing predictors [26]. Proteus seems to perform better than the existing predictors on the available independent dataset. Of course this has to be re-evaluated when more data becomes available. It is also possible that combining different individual methods to build hybrid methods could be one way to increase the performance even further. It is also important to conceptualize the multiple sequence driven factors and their coordination holding the key 'consensus' in promoting the 'disorder-to-order' transitions. The 'consensus' is yet untangled and needs other exclusive studies to eventually be resolved, however, the current work explores certain empirically observed trend which appears to be instrumental in the transition from disorder to order. These factors include the reappearance of large hydrophobic and charged amino acids in the protean segments, which are significantly under-represented in the originally 'disordered' regions. The study also shows that there is an inherent indecisiveness in predicted secondary structure assignments associated with the protean segments, where protean regions seem to alters its path along the main-chain trajectory so frequently that it ends up being flexible. This is consistent with the notion of sustaining enough 'disorder' even in the bound form [4] which potentially help the proteins to sustain their binding promiscuity. To conclude, the study has both a basic and an applied content and should serve the IDP as well as the broad biological community in both ways.

The software package is available at https://github.com/bjornwallner/proteus

Tables

Table 1. Description of the datasets

Dataset	Proteins	Protean Residues	Non-Protean Residues	Total Residues
ProS	557	6,245	356,053	362,298
MoRF	840	10,549	494,264	504,813
ProS + MoRF (PnM)	1,397	16,794	850,317	867,111
Validation	9	163	2,046	2,209

Table 2. A Summary of Feature groups

Feature Group	Name	Feature Number	Count
1	Sequence Profile	1-300	20×15=300
2	Amino Acid Conservation	301	1
3	Amino Acid Concentration	302-321	20x1=20
4	Amino Acid Properties	322-330	4+3+1+1=9

5	Predicted Secondary Structure	331-333	3×1=3
6	Predicted Disorder	334-340	3+3+1=7
7	Disorder Topography	341-342	1+1=2

Figure Legends:

Fig.1. Distribution of size of the 'originally classified' Protean segments. The distribution is obtained from the combined 'PnM' training dataset.

Fig.2. Amino Acid Propensities in the 'predicted' disordered vs. ordered regions. The Black Horizontal Line (Propensity = 1.0) serves as the baseline; meaning no preferential occurrence of the said amino acid in the said class. A propensity greater and lesser than 1.0 represents over and under representations respectively.

Fig.3. Amino Acid Propensities in the 'originally classified' protean vs. non-protean segments. The Black Horizontal Line (Propensity = 1.0) serves as the baseline; meaning no preferential occurrence of the said amino acid in the said class. A propensity greater and lesser than 1.0 represents over and under representations respectively.

Fig.4. Secondary Structural probabilities in the 'predicted' disordered vs. ordered regions. H, E and C stands for α-Helix, β-Strand and Random Coil (non-helix, non-strand) respectively.

Fig.5. Secondary Structural probabilities in the 'originally classified' protean vs. non-protean segments. H, E and C stands for α-Helix, β-Strand and Random Coil (non-helix, non-strand) respectively.

Fig.6. Indecisiveness in adapting a particular secondary structure for the 'originally classified' protean vs. non-protean segments. Probability Distribution of the Altscore (see Text) have been drawn for both sets. Segments assigned as purely 'Coil' were excluded from both sets.

Fig.7. Receiver Operating Characteristic (ROC) curves to analyze the cross-validated performance of Proteus. All five separate training / test folds as well as the final five-fold cross-validated 'Proteus' predictions (mean) are tabulated. The dashed line (- -) with a slope of 1.0 represents the random baseline.

Fig.8. Relative feature importance. Feature 342 describes the topographic length , Feature 340 describes the 'length of the ordered region' and feature 334 is the predicted disorder score.

Fig.9. Analysis of Proteus Score for the cross-validated predictions. (A) Proteus score vs PPV (solid, blue), recall (dashed, red), and F1 (dotted, orange) for the cross-validated predictions. **(B)** Box plots showing the distribution of predicted Proteus scores for protean and non-protean residues. The median of the two distributions are shown by the horizontal red line in the middle of the two boxes.

Fig.10. Comparison of Proteus with other classifiers using ROC curves. All methods were tested on the same validation set of 9 proteins containing 2209 residues (total number of examples) with 163 protean (positive examples). AUC stands for Area Under the Curve which were calculated using the Trapezoidal numerical integration (trapz) function of MATLAB. The random baseline (dashed black line) corresponds to a purely random classifier.

References

1. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293: 321–331. doi:10.1006/jmbi.1999.3110

2. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, et al. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac Symp Biocomput Pac Symp Biocomput. 1998; 473–484.

3. Kulkarni P, Rajagopalan K, Yeater D, Getzenberg RH. Protein Folding and the Order/Disorder Paradox. J Cell Biochem. 2011;112: 1949–1952. doi:10.1002/jcb.23115

4. Uversky VN. Unusual biophysics of intrinsically disordered proteins. Biochim Biophys Acta. 2013;1834: 932–951. doi:10.1016/j.bbapap.2012.12.008

5. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181: 223–230.

6. Amadei A, Linssen AB, Berendsen HJ. Essential dynamics of proteins. Proteins. 1993;17: 412–425. doi:10.1002/prot.340170408

7. Harding HP, Zhang Y, Ron D. Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. Nature. 1999;397: 271–274. doi:10.1038/16729

8. Pestova TV, Hellen CUT. Coupled Folding during Translation Initiation. Cell. 2003;115: 650–652. doi:10.1016/S0092-8674(03)00981-4

9. Lau AY, Chasman DI. Functional classification of proteins and protein variants. Proc Natl Acad Sci U S A. 2004;101: 6576–6581. doi:10.1073/pnas.0305043101

10. Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol. 2004;5: R6.

11. E. B. Starikov BN. Entropy-enthalpy compensation as a fundamental concept and analysis tool for systematical experimental data. Chem Phys Lett. 2012;538: 118–120. doi:10.1016/j.cplett.2012.04.028

12. Fisher CK, Stultz CM. Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol. 2011;21: 426–431. doi:10.1016/j.sbi.2011.04.001

13. Chebaro Y, Ballard AJ, Chakraborty D, Wales DJ. Intrinsically Disordered Energy Landscapes. Sci Rep. 2015;5: 10386. doi:10.1038/srep10386

14. chaos_complexity_entropy.pdf [Internet]. Available: http://necsi.edu/projects/baranger/cce.pdf

15. Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes - Chemical Society Reviews (RSC Publishing) [Internet]. [cited 7 May 2016]. Available: http://pubs.rsc.org/en/content/articlelanding/2011/cs/c0cs00057d#!divAbstract

16. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16: 18–29. doi:10.1038/nrm3920

17. Uversky VN. Intrinsically disordered proteins may escape unwanted interactions via functional misfolding. Biochim Biophys Acta. 2011;1814: 693–712. doi:10.1016/j.bbapap.2011.03.010

18. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys. 2008;37: 215–246. doi:10.1146/annurev.biophys.37.032807.125924

19. Fukuchi S, Sakamoto S, Nobe Y, Murakami SD, Amemiya T, Hosoda K, et al. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res. 2012;40: D507-511. doi:10.1093/nar/gkr884

20. Cheng S, Cetinkaya M, Gräter F. How Sequence Determines Elasticity of Disordered Proteins. Biophys J. 2010;99: 3863–3869. doi:10.1016/j.bpj.2010.10.011

21. Theillet F-X, Kalmar L, Tompa P, Han K-H, Selenko P, Dunker AK, et al. The alphabet of intrinsic disorder. Intrinsically Disord Proteins. 2013;1: e24360. doi:10.4161/idp.24360

22. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein Disorder Prediction: Implications for Structural Proteomics. Structure. 2003;11: 1453–1459. doi:10.1016/j.str.2003.10.002

23. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005;21: 3433–3434. doi:10.1093/bioinformatics/bti541

24. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7: 208. doi:10.1186/1471-2105-7-208

25. Shimizu K, Hirose S, Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics. 2007;23: 2337–2338. doi:10.1093/bioinformatics/btm330

26. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinforma Oxf Engl. 2015;31: 857–863. doi:10.1093/bioinformatics/btu744

27. Mészáros B, Simon I, Dosztányi Z. Prediction of Protein Binding Regions in Disordered Proteins. PLOS Comput Biol. 2009;5: e1000376. doi:10.1371/journal.pcbi.1000376

28. Disfani FM, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012;28: i75–i83. doi:10.1093/bioinformatics/bts209

29. Fang C, Noguchi T, Tominaga D, Yamana H. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics. 2013;14: 300. doi:10.1186/1471-2105-14-300

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

31. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292: 195–202. doi:10.1006/jmbi.1999.3091

32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped Blast and PsiBlast: a new generation of protein database search programs. NUCLEIC ACIDS Res. 1997;25: 3389–3402.

33. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23: 1282–1288. doi:10.1093/bioinformatics/btm098

34. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27: 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

35. Cooper GM. The Cell. 2nd ed. Sinauer Associates; 2000.

36. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157: 105–132. doi:10.1016/0022-2836(82)90515-0

37. Munson M, Balasubramanian S, Fleming KG, Nagi AD, O'Brien R, Sturtevant JM, et al. What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. Protein Sci Publ Protein Soc. 1996;5: 1584–1593.

38. Betz SF. Disulfide bonds and the stability of globular proteins. Protein Sci Publ Protein Soc. 1993;2: 1551–1558.

39. P Y Chou, Fasman and GD. Empirical Predictions of Protein Conformation. Annu Rev Biochem. 1978;47: 251–276. doi:10.1146/annurev.bi.47.070178.001343

40. Visiers I, Braunheim BB, Weinstein H. Prokink: a protocol for numerical evaluation of helix distortions by proline. Protein Eng. 2000;13: 603–606. doi:10.1093/protein/13.9.603

41. Das M, Basu G. Glycine Rescue of β-Sheets from cis-Proline. J Am Chem Soc. 2012;134: 16536–16539. doi:10.1021/ja308110t

42. Rauscher S, Baud S, Miao M, Keeley FW, Pomès R. Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. Struct Lond Engl 1993. 2006;14: 1667–1676. doi:10.1016/j.str.2006.09.008

43. Singh G. Association between intrinsic disorder and serine/threonine phosphorylation in Mycobacterium tuberculosis. FASEB J. 2015;29: 563.4.

44. Haynes C, Iakoucheva LM. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. Nucleic Acids Res. 2006;34: 305–312. doi:10.1093/nar/gkj424

45. Basu S, Bhattacharyya D, Wallner B. SARAMAint: The Complementarity Plot for Protein–Protein Interface. J Bioinforma Intell Control. 2014;3: 309–314. doi:10.1166/jbic.2014.1103

46. Basu S, Wallner B. Finding correct protein–protein docking models using ProQDock. Bioinformatics. 2016;32: i262–i270. doi:10.1093/bioinformatics/btw257

Figure 1

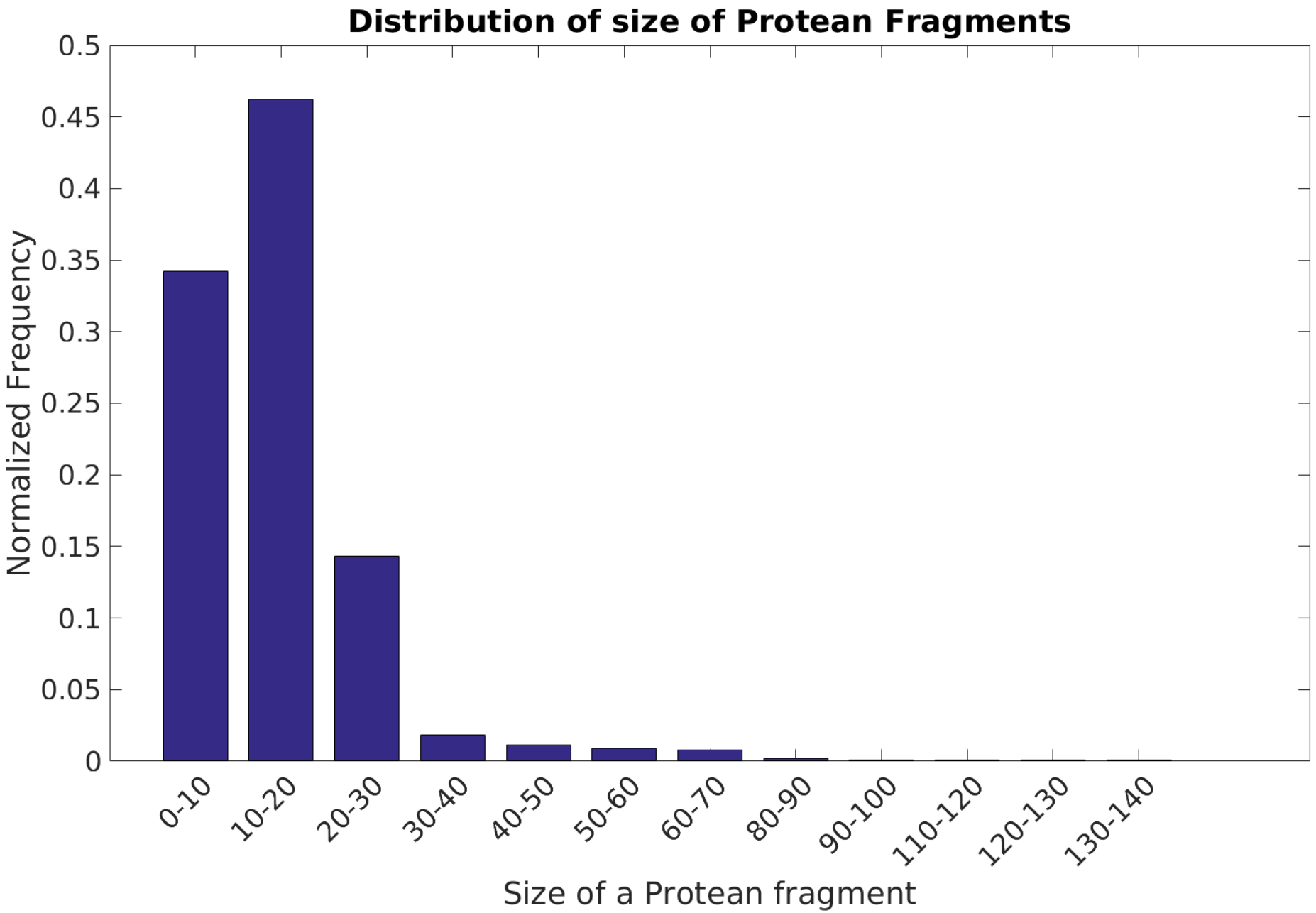

Distribution of size of Protean Fragments
Normalized Frequency
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
80-90
90-100
110-120
120-130
130-140
Size of a Protean fragment

Figure 2

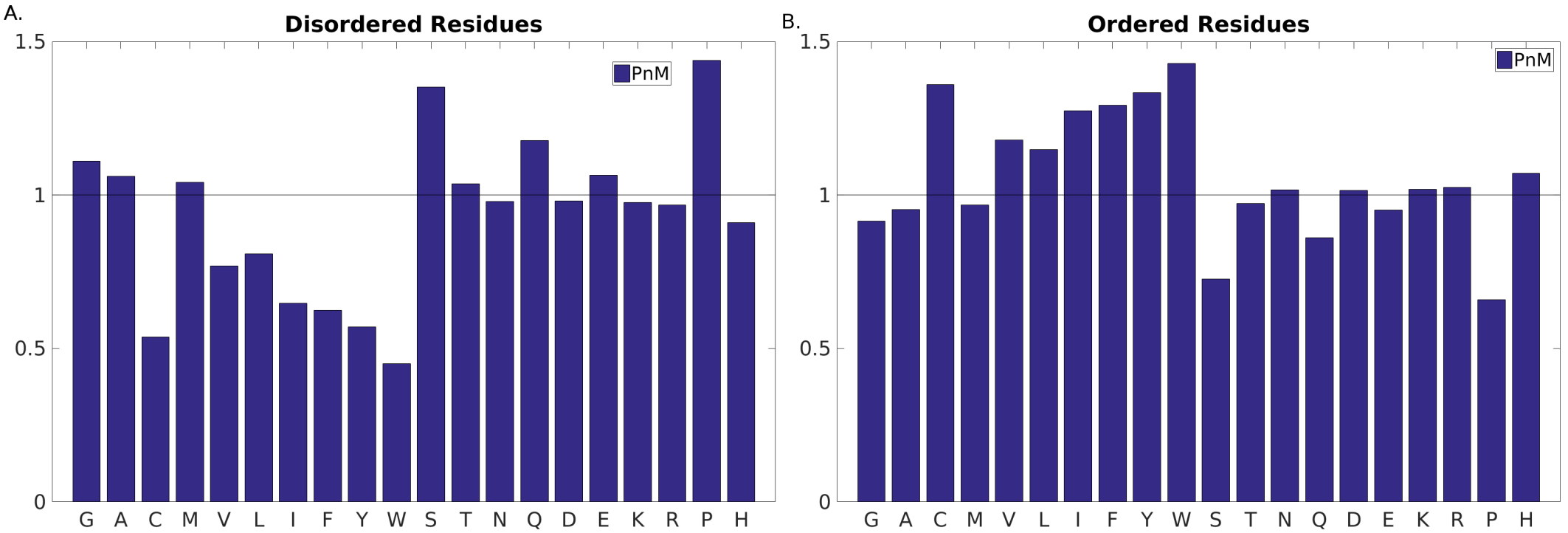

A.
Disordered Residues
PnM
1.5
1
0.5
0
G A C M V L I F Y W S T N Q D E K R P H
B.
Ordered Residues
PnM
1.5
1
0.5
0
G A C M V L I F Y W S T N Q D E K R P H

Figure 3

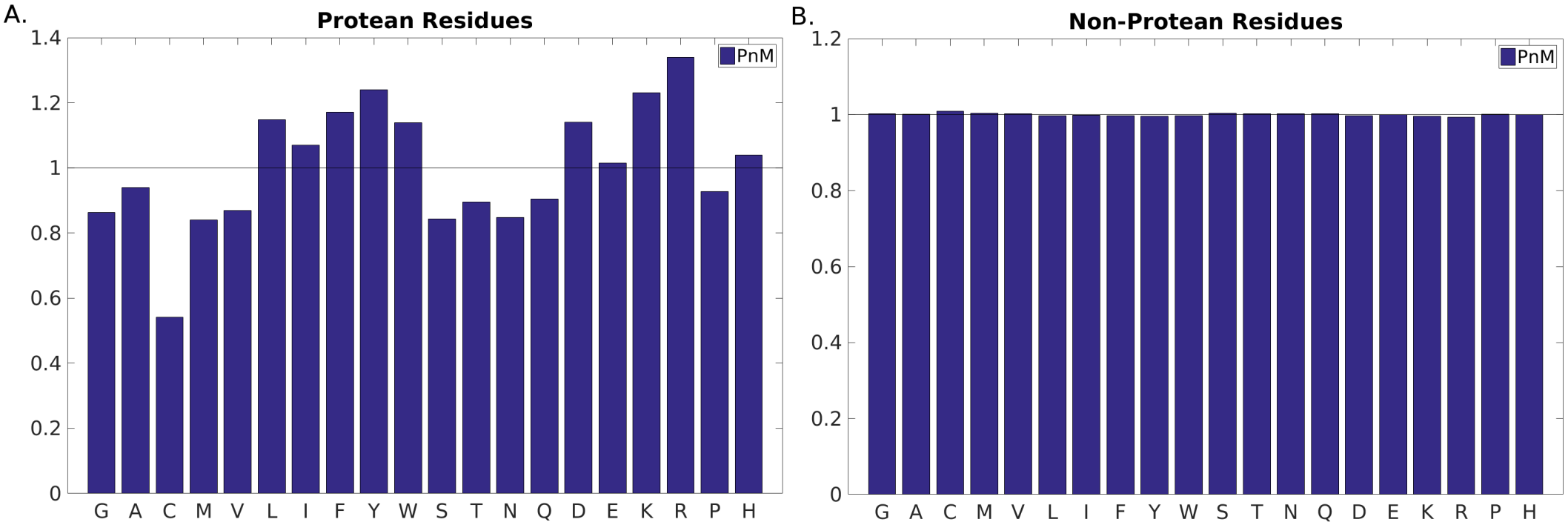

A.
Protean Residues
PnM
1.4
1.2
1
0.8
0.6
0.4
0.2
0
G A C M V L I F Y W S T N Q D E K R P H
B.
Non-Protean Residues
PnM
1.2
1
0.8
0.6
0.4
0.2
0
G A C M V L I F Y W S T N Q D E K R P H

Figure 4

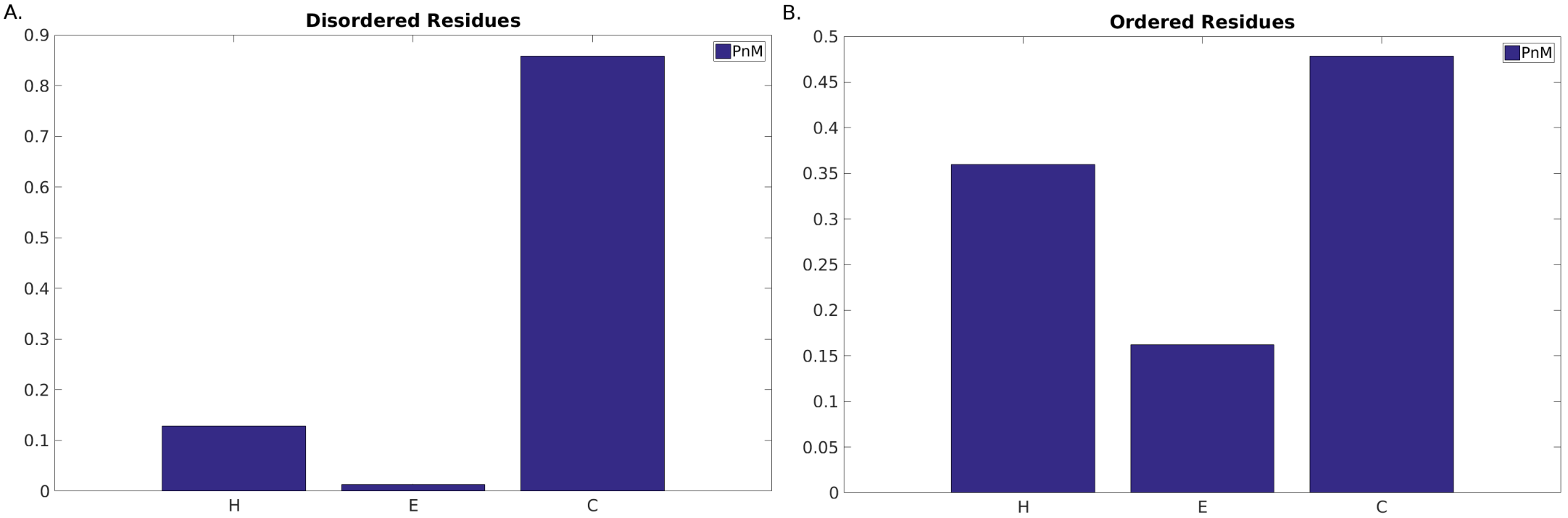

A.
Disordered Residues
PnM
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
H
E
C
B.
Ordered Residues
PnM
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
H
E
C

Figure 5

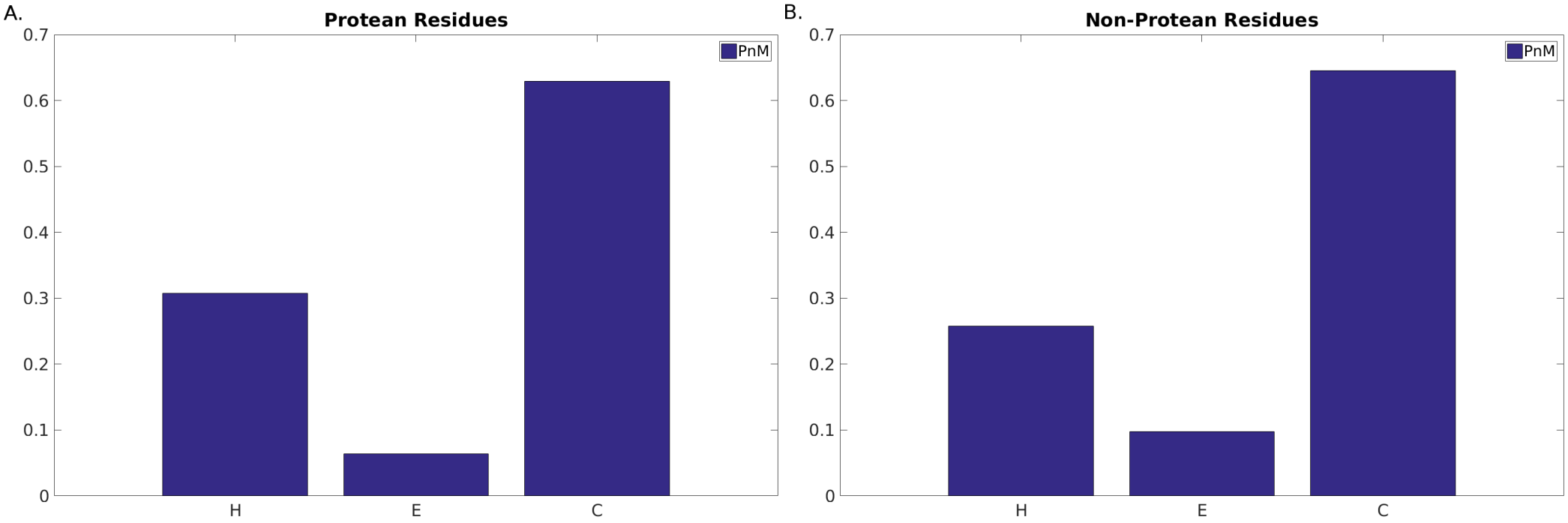

A.
Protean Residues
PnM
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
H
E
C
B.
Non-Protean Residues
PnM
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
H
E
C

Figure 6

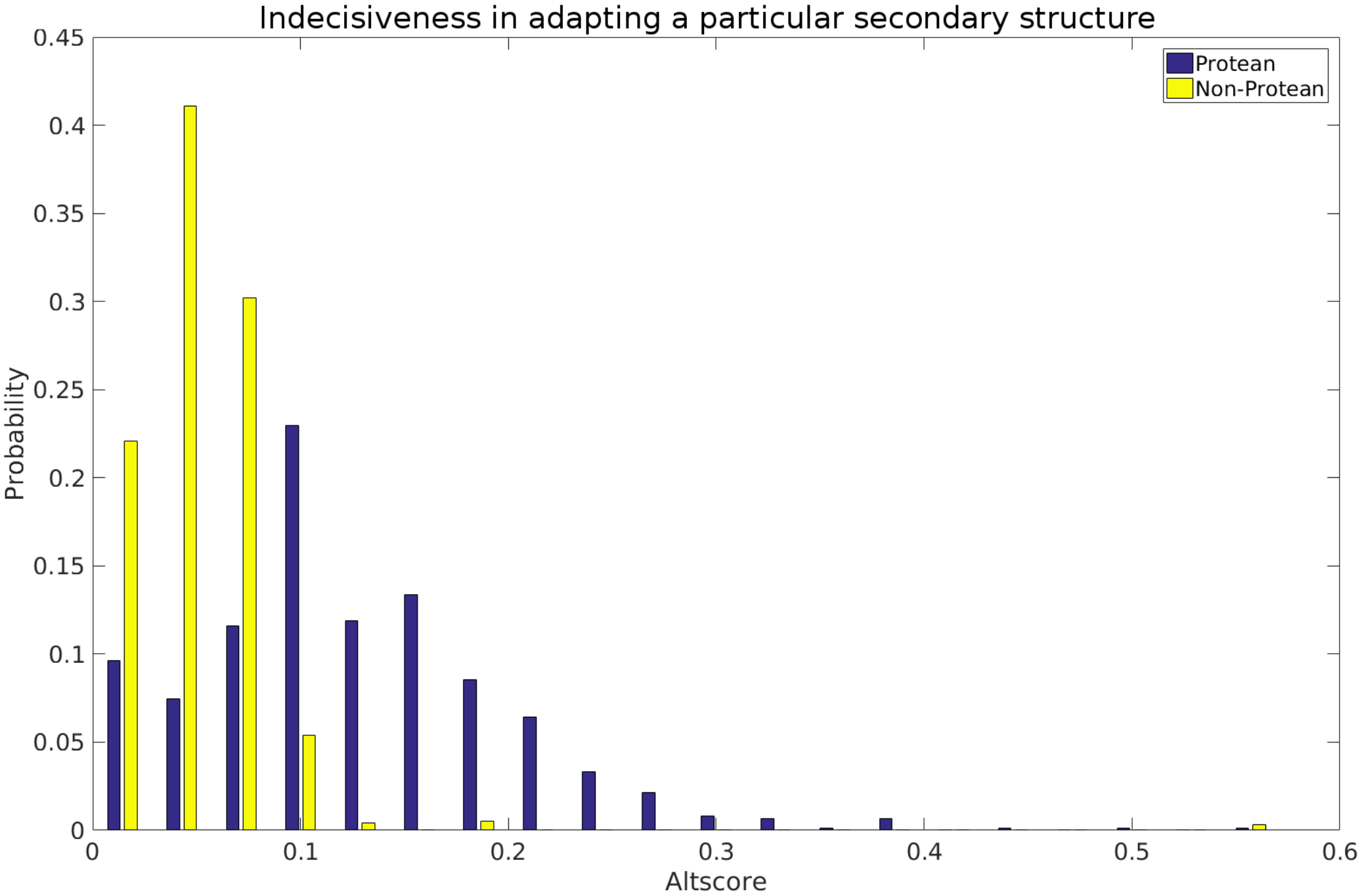

Indecisiveness in adapting a particular secondary structure
Protean
Non-Protean
Probability
Altscore
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 7

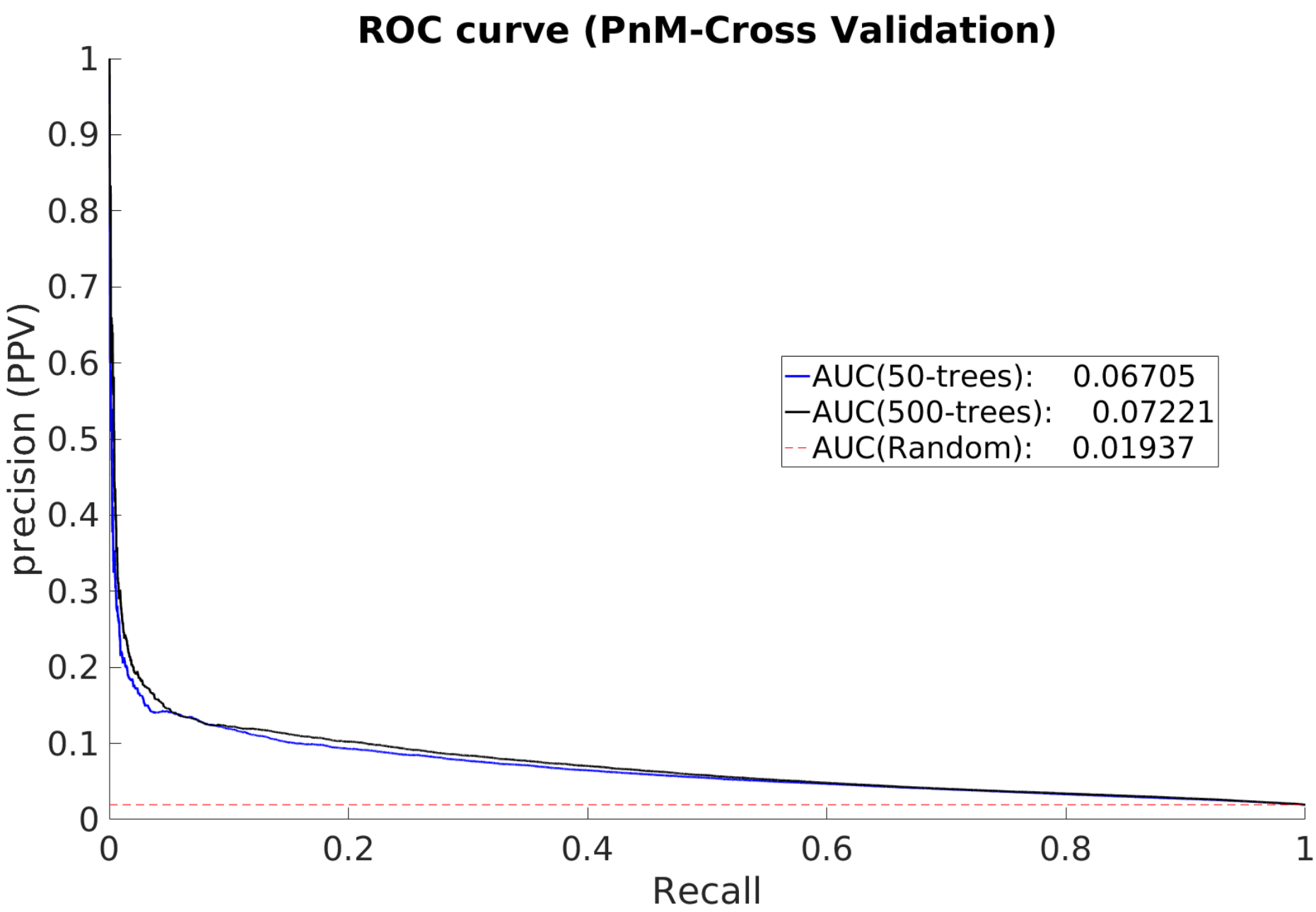

ROC curve (PnM-Cross Validation)
precision (PPV)
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
AUC(50-trees): 0.06705
AUC(500-trees): 0.07221
AUC(Random): 0.01937

Figure 8

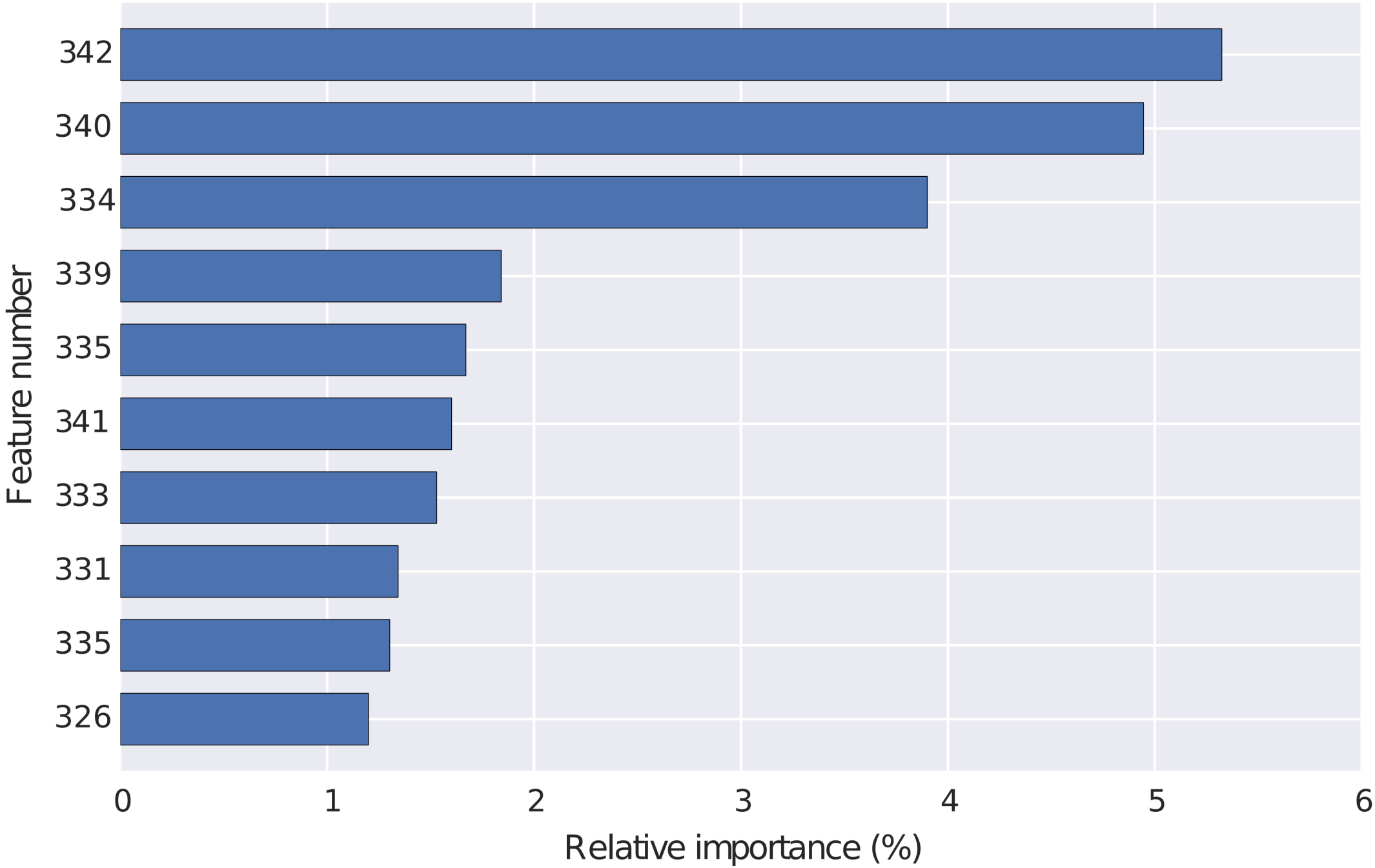

342
340
334
339
335
341
333
331
335
326
Feature number
0
1
2
3
4
5
6
Relative importance (%)

Figure 9

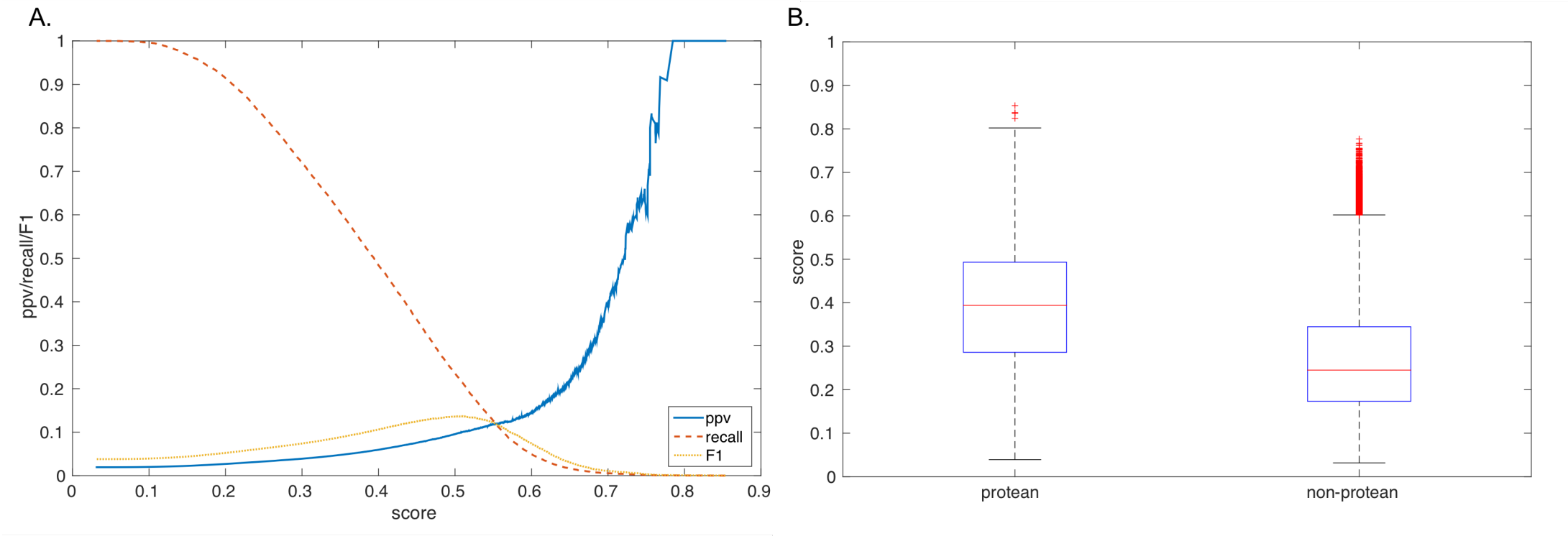
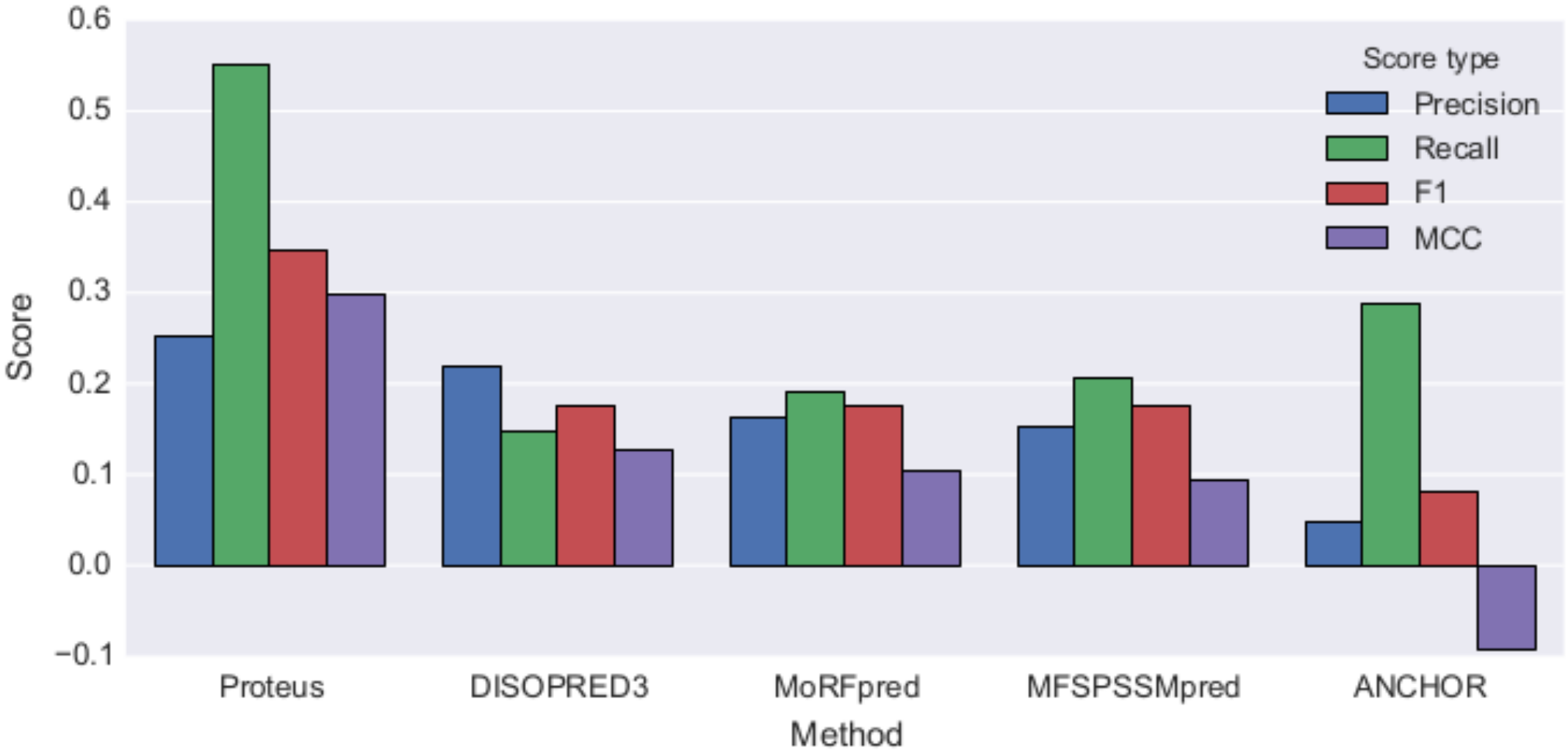

A.
ppv/recall/F1
score
ppv
recall
F1
B.
score
protean
non-protean

Figure 10


0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
Score
Score type
Precision
Recall
F1
MCC
Proteus
DISOPRED3
MoRFpred
MFSPSSMpred
ANCHOR
Method