Intrinsic disorder in putative protein sequences

Uros Midic, Zoran Obradovic
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, PA, USA
uros@temple.edu, zoran.obradovic@temple.edu

***Abstract*— Intrinsically disordered proteins perform a variety of crucial biological functions despite lacking stable tertiary structure under physiological conditions in vitro. State-of-the-art sequence-based predictors of intrinsic disorder are achieving per-residue accuracies over 80%. In a genome-wide study we observed big difference in predicted disorder content between confirmed and putative human proteins, and suspected that this is due to large errors introduced by gene-finding algorithms for putative sequence annotation. To test this hypothesis we trained a predictor to discriminate sequences of real proteins from synthetic sequences that mimic errors of gene finding algorithms. Its application to putative human protein sequences shows that they contain a substantial fraction of incorrectly assigned regions. These regions are predicted to have higher levels of disorder content than correctly assigned regions. Our finding provides first evidence that current practice of predicting disorder content in putative sequences should be reconsidered, as such estimates are biased.**

Keywords: Protein intrinsic disorder; disorder prediction; gene finding.

I. INTRODUCTION

Intrinsically disordered proteins (IDPs) are proteins that lack stable tertiary structure under physiological conditions in vitro [1]. They are also known by other names, including natively denatured [2], natively unfolded [3], intrinsically unstructured [4], and natively disordered [5]. IDPs can be wholly disordered or partially disordered, where we can identify intrinsically disordered regions (IDRs) and ordered regions. Although they lack stable tertiary structure, the functional repertoire of IDPs complements the functions of ordered proteins. IDPs are involved in a number of crucial biological functions including regulation, recognition, signaling and control.

There are several crucial differences between amino acid sequences of IDPs/IDRs and structured globular proteins and domains. These differences include divergence in amino acid composition and sequence complexity, and consequently in physico-chemical properties like hydrophobicity, aromaticity, charge, and flexibility index value [6]. For example, IDPs possess a low content of N and of the cross-linking C residues and are significantly depleted in bulky hydrophobic (I, L, and V) and aromatic amino acid residues (W, Y, and F), which form and stabilize the hydrophobic cores of folded globular proteins. These amino acids have been called *order-promoting amino acids*. On the other hand, IDPs/IDRs are substantially enriched in polar and charged amino acids (R, Q, S, E, and K) and in structure-breaking G and P residues, collectively called *disorder-promoting amino acids* [1], [7], [8]. The difference in amino acid composition between predicted ordered and predicted disordered regions in human proteins is illustrated in Fig. 1.

Thus, in addition to the well-known "protein folding code" stating that all the information necessary for a given protein to fold is encoded in its amino acid sequence [9], "protein non-folding code" has been proposed, according to which the propensity of a protein to stay intrinsically disordered is likewise encoded in its amino acid sequence [10], [11]. This has been utilized to develop numerous predictors of intrinsic disorder (ID), which achieve over 80% of per-residue accuracy [12].

Large-scale genome-wide prediction of ID has been used previously to confirm the ubiquity of ID [13], and to compare the abundance of ID in various genomes and groups of genomes [14], [15]. In a previous study [16] we applied a per-residue predictor of ID to human proteins available in NCBI database, to compare disorder content in various classes of human proteins. One intriguing result was the vast discrepancy in predicted disorder content between the protein sequences with IDs starting with NP (protein records in advanced stage of the curation process, further referred as *NP sequences* and *NP class*) and protein sequences with IDs starting with XP (protein records in early stages of the curation process, further referred as *XP sequences* and *XP class*). The difference in distributions of predicted disorder content for NP class and XP class of sequences is shown in Fig. 2; we define *disorder content* (DC) as the fraction of residues in a sequence that are predicted to be disordered. Note that XP class has a low fraction of proteins

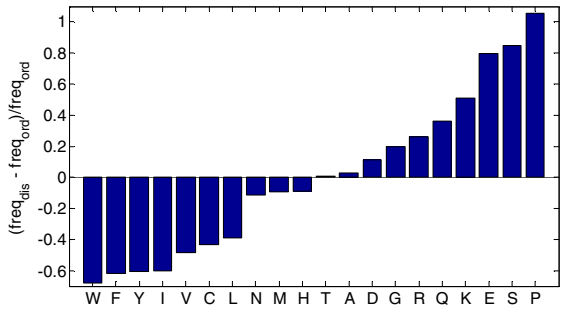

Figure 1. Comparison of amino-acid frequencies in predicted disordered ($freq_{dis}$) and ordered ($freq_{ord}$) regions of human proteins. Amino acids are sorted by the relative difference $(freq_{dis} - freq_{ord})/freq_{ord}$.

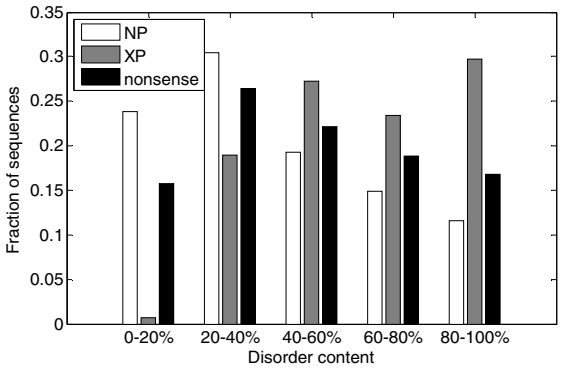

Figure 2. Distribution of predicted disorder content (DC) in confirmed human proteins (NP), putative human proteins (XP), and synthetic nonsense sequences. Histograms show fractions of sequences with various levels of DC.

predicted as fully and mostly ordered (DC < 20%), and that there is more than two-fold difference in fractions of proteins predicted as fully and mostly disordered (DC > 80%) in XP class compared to NP class.

The XP class contains putative protein sequences, predicted by gene finding algorithms, which have not yet been curated and confirmed. Gene finding is the problem of predicting the positions of genes, and the positions of exons and introns inside the genes, for a given genomic sequence. Most gene finding predictors use Bayesian networks, such as Interpolated Markov Models [17], Generalized Hidden Markov Models [18], and Generalized Pair HMMs [19]. These predictors exploit the following findings: 1) many signals involved in gene expression (e.g. promoters, splice junctions) exert specific motifs, and can be predicted from sequence, 2) protein-coding DNA have statistical properties (such as amino acid composition, length) that distinguish them from non-coding DNA, 3) signals and statistical properties are often conserved across related sequences (intra- and inter-species).

From the domain experts' point of view, the gene finding algorithms are useful, as they provide important guidelines for experimental research, where predicted putative sequences are confirmed or refined. However, inclusion of these putative sequences in large-scale analysis of ID is questionable. Even when predicted exons of a predicted protein sequence overlap with true exons, the overlap can be partial and non-coding DNA may be included in the predicted exons. Another possibility is that in predicted protein sequence, true exons are translated in wrong reading frame. Therefore, predicted protein sequences contain regions that come from non-coding DNA or incorrectly translated coding DNA, and are not present in true protein sequences. In further text we will refer to them as nonsense regions/sequences. Nonsense regions do not exist in real proteins, and the hypothetical structure they would conform to if they were synthesized is uncertain. Therefore, any prediction of structure – including prediction of intrinsic disorder – for nonsense regions and sequences is not valid. Inclusion of such sequences in genome-wide analysis of intrinsic disorder can possibly substantially bias the estimate of ID content in a genome. In our previous study [16] we decided to exclude XP sequences from analysis of ID in human genome. On the other hand, their exclusion from genome-wide analysis can also give an unrealistic estimate of ID content, especially if the proportion of unconfirmed sequences is high. If the higher predicted disorder content in XP sequences is realistic, then their exclusion can negatively bias the estimate of ID content in the genome.

In this paper we explore the relationship between nonsense regions in XP sequences – introduced through errors made by gene finding procedures – and intrinsic disorder. In addition to the difference in amino acid composition between NP and XP sequences (further elaborated in the Experimental results section), we assumed that nonsense regions follow a different amino acid composition than true protein sequences. Therefore, instead of testing and improving the gene finding algorithms, we investigate whether nonsense regions can be detected from amino acid sequence, similarly to prediction of intrinsic disorder.

We developed a two-class predictor that aims at distinguishing true protein sequences from nonsense regions in putative sequences. Since no data is easily available about which regions of XP sequences are nonsense, we constructed synthetic nonsense sequences from mRNAs of the true protein sequences that form the other class.

Per-residue predictor achieved accuracy (~82%) and AUC (~.90) that is comparable to accuracy of state-of-the-art ID predictors, with good balance between sensitivity and specificity, and equal performance in disordered and ordered parts of the dataset. Per-protein was close to perfect (accuracy ~ 95%, AUC ~ .98).

Comparison of nonsense region prediction results for NP and XP sequences has shown that a substantial number of XP sequences contain nonsense regions. Furthermore, in XP sequences these regions have higher predicted disorder content, which was not observed in false positive regions in NP sequences. This indicates that detected errors in protein finding procedures that introduce nonsense regions into XP sequences are only partially responsible for higher disorder content in compared sets. However, the level of change in disorder prediction does not fully explain the discrepancy in predicted disorder content between XP and NP sequences.

In Section 2 we describe the methodology that we used to create the synthetic nonsense sequences, train and evaluate the nonsense predictor, and analyze results of the predictor for XP sequences. In Section 3 we present more details on the comparison of amino acid sequence composition, results of predictor evaluation, comparison of nonsense prediction in different classes of sequences, and the analysis of relationship between nonsense prediction and disorder prediction. In Section 4 we give a discussion of the results, open questions and directions for future research.

II. METHODOLOGY

A. Dataset and creation of synthetic nonsense sequences

The dataset was based on the set of human proteins obtained from NCBI that we used in our previous study [16]. For the control/negative class of true proteins we selected all human NP proteins that are listed as single isoforms of respective genes, i.e. the genes are not known to be involved in alternative splicing. Nonsense protein sequences for the

positive class were synthesized from mRNA sequences of confirmed proteins from the negative class. The exact locations of exons in these mRNA sequences are known, and the exons can only be translated correctly if they are read in one of three possible reading frames. For a given mRNA sequence and the protein it is translated to (top sequence in Fig. 3, where exons are shown in black), the procedure to synthesize nonsense sequences was the following:

For each of the three possible reading frames: 1) Translate the codons into amino acids, ignore/discard any stop codons (this amino acid sequence is further referred to as the *candidate sequence*). 2) Align the candidate sequence to the true protein sequence. 3) Identify any parts of the candidate sequence that are perfectly matched to the true protein sequence, and are at least 10 amino acids long (shown in dark gray at Fig. 3). These come from true exons that are correctly translated and are therefore removed from the candidate sequence. 4) The remaining parts of the candidate sequence (light gray in Fig. 3) are either coming from non-coding parts of mRNA or from incorrectly translated exons; therefore they can be considered nonsense sequences.

This procedure produces three nonsense sequences for each true protein sequence. We discarded short sequences for which construction of input features for prediction is not viable. The dataset contained 15,124 NP sequences and 45,038 synthetic nonsense sequences. We also assembled an additional set of 5,243 XP sequences, which is the focus of this study. Similarly to the training dataset, this set includes only proteins that are listed as single isoforms of their respective genes.

Sequences from both parts of the dataset and the additional set were preprocessed to construct predictive features, similarly to how features are constructed for PONDR family of ID predictor [7], [12], [20], [21]. For each fixed residue, a window of size 41 was positioned centered at the fixed residue. Amino acids in the window were counted and their frequencies were calculated; this produced 20 features that correspond to amino acid composition. Entropy was calculated from 20 amino acid frequencies; this feature measures local complexity of amino acid sequence. Local flexibility was approximated as the scalar product of 20 amino acid frequencies and 20 flexibility parameters, which were estimated empirically. Net charge and average hydrophobicity were calculated similarly to flexibility, and their ratio is used as an additional feature. Predictions of ID were obtained with the VSL2B predictor [12], which outputs a real value in the [0,1] range for each residue; these predictions are mapped to binary classification by applying the .5 threshold. To summarize the predicted ID in a protein sequence, we define *disorder content* (DC) for a sequence as the fraction of residues that are predicted to be in disordered regions.

B. Predictor of nonsense regions in protein sequences

This prediction problem is novel, and therefore we could not utilize any of the existing protein-related prediction tools. Furthermore, we could not compare our results to any previously published results. Our goal was not to develop an optimal predictor, but rather to construct a simple predictor with reasonable accuracy and good balance between sensitivity and specificity. We briefly tested logistic regression and neural networks as the predictive model, with various sets of parameters. Here we present only the parameters that led to the best results that we have obtained. We used neural networks with 20 hidden nodes in a single hidden layer. We always trained ensembles of 10 neural networks, with randomly sampled training and validation sets. The training and validation sets (8% and 2% of the available data respectively) were sampled from the dataset. Only 10% of the available data was used per iteration to speed up the training and evaluation process. Because windows used to construct features for neighboring amino-acids are overlapping, the obtained features are similar, and therefore the redundancy allows for subsampling without significant loss of accuracy. Both training and validation sets were balanced (i.e. contained equal number of residues from positive and negative class), and samples from both classes were further balanced in terms of disorder to include equal number of residues predicted to be ordered and disordered. Targets for residues from two classes were encoded as .1 and .9. In the evaluation phase, the residues were classified by comparing their real-valued predictions with the .5 threshold.

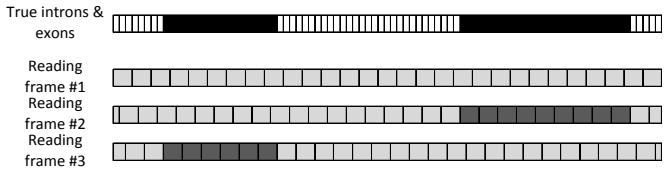

Figure 3. An mRNA sequence with known exon positions (black, dark grey) is translated using three different reading frames to synthesize nonsense protein sequences (light grey).

C. Evaluation

We performed both per-residue and per-sequence evaluation. In per-residue evaluation residues are observed separately, while in per-sequence evaluation predictions for all residues in a sequence are aggregated into one prediction (mean of per-residue predictions) and compared to a threshold. We used 10-fold cross-validation to evaluate the predictor, and the dataset was partitioned into 10 subsets so that residues from the same sequence were always members of the same subset. This partitioning both enables per-protein prediction and ensures fair testing in per-residue prediction, since neighboring residues in a sequence have similar input features and in most cases equal target values, and should therefore always be in the same subset.

We used two indicators of nonsense prediction level in a sequence. We define *nonsense content* as the fraction of predicted nonsense residues in a sequence; this indicator is analogous to disorder content. Another indicator is the mean of (real-valued) per-residue nonsense predictions in the sequence. Both indicators were used to compare results of prediction for NP and XP sequences.

III. Experimental Results

The motivation for this study was the discrepancy in predicted disorder content between NP and XP sequences (Fig. 2). This is consistent with the significant difference in amino acid composition between these two sets of sequences (Fig. 4). Majority of amino acids enriched in NP sequences are order promoting, while majority of amino acids enriched in XP

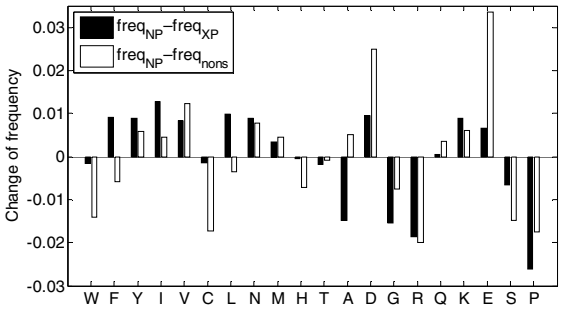

Figure 4. Changes in frequencies between XP (and synthetic nonsense sequences), and NP sequences. Same amino-acid order as in Fig. 1 is used.

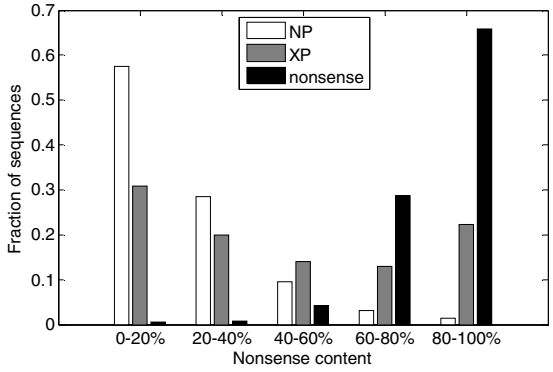

Figure 5. Distributions of predicted nonsense content in NP, XP, and synthetic nonsense sequences. Histograms show fractions of sequences with various levels of predicted nonsense content.

sequences are disorder promoting. However, we did not observe similar results when comparing amino acid composition in NP and synthetic nonsense sequences (Fig. 4). Two disorder promoting amino acids (D, E) are highly enriched in NP sequences, and several order promoting amino acids are enriched in nonsense sequences (e.g. W, C). While predicted disorder content in synthetic nonsense sequences is significantly higher than in NP sequences, it is significantly lower than in XP sequences (Fig. 2).

A. Evaluation of nonsense sequence predictor

The results of 10-fold cross-validation evaluation are summarized in Table 1. Since the positive class is much larger than the negative class, we measured specificity (true negative rate, accuracy on the negative class) and sensitivity (true positive rate, accuracy on the positive class) separately and used the average value of sensitivity and specificity as the adjusted measure of accuracy. For per-residue prediction we also evaluated the predictor separately on disordered and ordered regions. For per-protein evaluation, we list both the results for default threshold (.5), and the threshold that produces the best result (.53). The accuracy was only slightly higher in the disordered part of the dataset. While sensitivity was increased on disordered regions and specificity was increased on ordered regions, these differences were fairly small. The *area under curve* (AUC) for the *receiver operating characteristic* curve (ROC) is .9024 for per-residue prediction and .9846 per-protein prediction.

B. Comparison of predicted nonsense in NP and XP classes

As a part of the 10-fold cross-validation process, we obtained predictions for all NP and synthetic nonsense sequences. We could then use all 10 predictors as an ensemble for prediction on XP sequences, since they were not used in training; the ensemble predictor is expected to perform at least as well as its component predictors [22].

TABLE I. 10-FOLD CV EVALUATION OF PER-RESIDUE AND PER-PROTEIN NONSENSE SEQUENCE PREDICTORS

	Specificity	***Sensitivity.***	***Accuracy = (spec+sens)/2***
Per-residue			
Overall	**82.27%**	**82.67%**	**82.47%**
Ordered regions	81.15%	80.61%	80.88%
Disordered regions	82.44%	84.40%	83.42%
Per-protein			
Default thr. = 0.5	**93.29%**	**97.56%**	**95.42%**
Optimal thr. = 0.53	95.21%	96.14%	95.67%

We calculated nonsense content for all NP, XP and synthetic nonsense sequences. The distributions of nonsense content in the three groups of sequences are compared in Fig. 5. Difference between NP and synthetic nonsense sequences is expected in accordance with predictor evaluation results. However, the significant increase in nonsense content for XP sequences, compared to NP sequences, cannot be explained by the design of the predictor or attributed to noise. With respect to the input features, derived from amino acid sequence, significant portion of XP sequence regions are more similar to synthetic noise sequences than to NP sequences.

Another way to compare per-residue nonsense predictions in NP and XP sequences is to calculate total predicted nonsense content (total fraction of residues predicted to be nonsense) for various levels of threshold between 0 and 1 and monitor the difference of these values (separation margin) between XP and NP sequences and between synthetic nonsense and NP sequences (figure not shown due to space constraint). The margin between NP and synthetic nonsense sequences, which equals 1–(*sensitivity*+*specificity*) for the predictor, has its peak of 64.95% at *thr* ~ .5. At the same threshold the margin between predicted nonsense content for XP (37.77%) and NP (17.73%) is 20.05%. If we assume that the predictor has the same false positive rate in XP as in NP (17.73%), then $1-\frac{1-.3777}{1-.1773}=24.36\%$ is a simple pessimistic estimate for the true fraction of nonsense region residues in XP sequences. Note that under the above stated assumption, predicted nonsense content for XP (37.77%) is split to the true nonsense content in XP sequences (24.36%) and the false positives in the remaining residues (1–.2436)*.1773, i.e. $.3777=.2436+(1-.2436).1773$, which is equivalent to $1-\frac{1-.3777}{1-.1773}=24.36\%$. With an additional assumption that the false negative rate in XP sequences is the same as in synthetic nonsense sequences (.1733), we can adjust the formula (omitted due to space constraint) for the fraction of nonsense region residues in XP sequences, which then gives an estimate of 31.48%.

We also applied the ensemble of per-protein predictors to XP sequences. 40.42% of XP sequences are predicted to be positive (i.e. nonsense), compared to only 6.71% of NP sequences. Per-protein predictors were developed to

discriminate between two extremes: real protein sequences and synthetic nonsense sequences. Therefore, the results of its application to XP sequences, with only partial nonsense regions, are not completely realistic. While we may not simply conclude that 40.24% of XP sequences are fully nonsense sequences (like the synthetic sequences), the per-protein predictor of nonsense sequences sees them as more similar to the synthetic nonsense sequences than to real NP protein sequences.

C. *Relationship between prediction of nonsense in XP sequences and prediction of intrinsic disorder*

After the computational experiments have indicated that XP sequences contain substantial fraction of nonsense regions, the important question is how these regions affect the prediction of disorder content in XP sequences. In XP sequences, 64.50% of all residues are predicted to be disordered. In regions predicted to be nonsense the fraction of predicted ID residues is increased to 75.10%, while in regions predicted to stem from coding DNA, the fraction of predicted ID residues is only 58.96%.

A new question arises, whether the positive correlation between prediction of nonsense and prediction of ID for XP sequences is specific for XP sequences, or whether it can also be observed in synthetic nonsense sequences, or even in the false positive regions in NP sequences predicted to be nonsense. To answer this question, in each of the three groups of sequences we calculate the Pearson correlation coefficient ρ between predicted disorder content and predicted nonsense content for all sequences, and calculate R^2 statistic and p-value for linear regression. These parameters indicate that the correlation is the strongest for XP sequences (ρ=.442, R^2=.196, and p~5E–250), and that there is a significant positive trend for synthetic noise sequences (ρ=.222, R^2=.049, and p~0), while there is no significant correlation for NP sequences (ρ=–.014, R^2~0, and p=.084).

IV. Discussion

In our previous study [16] we have observed a big increase in predicted disorder content for human protein sequences from NCBI with XP identifiers, as compared to human protein sequences with NP identifiers (Fig. 2). This difference was consistent with divergence in amino acid composition for NP and XP sequences (Fig. 4), since several order-promoting amino acids were highly enriched in NP sequences, and several disorder-promoting amino acids were highly enriched in XP sequences.

Sequences have XP identifiers when they are in early stages of curation, and many of them are just putative sequences submitted by the automated genome annotation procedure that utilizes gene finding algorithms. Since gene finding algorithms are not perfect, they introduce nonsense regions into XP sequences. We suspected that these nonsense regions may be one of the causes for the discrepancy in predicted disorder.

Based on the difference in amino acid composition (Fig. 4), we assumed that nonsense regions can be predicted from sequence. Since no data on nonsense regions was available we developed a simple procedure to construct synthetic nonsense sequences from real protein sequences and their mRNAs (Fig. 3). These sequences have different amino acid composition than their real counterparts (Fig. 4.), although they also differ greatly from XP sequences, as they have higher fractions of some order-promoting amino acids and lower fractions of some disorder-promoting amino acids. We suspect that the procedure to construct synthetic nonsense sequences can be adjusted to better mimic the errors made by gene finding algorithms, however at this point it is not clear how that can be done.

Using a simple prediction model, we have successfully trained a predictor that discriminates true NP sequences from synthetic nonsense sequences. All input features were based only on local sequence information, and were constructed using methodology similar to many predictors of intrinsic disorder. This result confirmed the assumption that nonsense regions can be predicted from sequence alone. The trained predictor has very good per-residue accuracy (~82.5%), comparable to predictors of intrinsic disorder (Table 1). More importantly, it is very well balanced (i.e. has similar sensitivity and specificity) and performs equally well on predicted disordered regions and predicted ordered regions.

We have also developed a simple method to aggregate per-residue predictions and output per-protein predictions. Performance of per-protein predictor is very close to optimal, with accuracy ~95% and AUC ~ .98. However, it is only feasible to use per-protein prediction when a sequence is either a true protein sequence or the whole sequence is nonsense. We applied both per-residue and per-protein predictors to XP sequences. We used various methods to compare results of nonsense prediction for NP and XP sequences. Per-protein predictor classified ~40% of XP sequences as fully nonsense sequences, compared to only ~6% of NP sequences. While this estimate is not realistic, it is indicative of how many XP sequences are more similar, in terms of input features, to synthetic nonsense sequences than to real NP sequences. Per-residue predictor also gave very different predictions for NP and XP sequences. The difference in distributions of nonsense content (fraction of residues in a sequence predicted to be in nonsense regions) is significant (Fig. 5). For example, 48.33% of XP sequences have predicted nonsense content greater than 40%, compared to only 12.84% of NP sequences.

We analyzed the total nonsense content (total fraction of residues in predicted nonsense regions) for NP, XP and synthetic sequences at various values of threshold. The separation margin between predicted nonsense contents for NP and synthetic nonsense sequences peaks around the default threshold .5, and the margin between predicted nonsense contents for NP and XP is close to its maximum (~20%) at that threshold. Under the assumption that in XP sequences the false positive rate is the same as in NP sequences, we provided a pessimistic estimate for disorder content of ~24%. After adding the assumption that in XP sequences the false negative rate is the same as in the synthetic nonsense sequences, we estimated that ~31% of residues in XP sequences belong to nonsense regions. Note that both of these estimates were calculated using simple formulas, under very strong assumptions.

Predicted nonsense regions in XP sequences have higher disorder content (75.1%) than the remaining regions (59.0%). More importantly, there is a significant positive linear dependency between predicted nonsense content and predicted

disorder content in XP sequences, as indicated by fairly high Pearson correlation coefficient, as well as the R^2 statistic and low p-value for the corresponding linear regression model. While similar positive linear dependency (albeit with lower correlation coefficient) is observed in synthetic nonsense sequences, it is completely absent from NP sequences.

All these experimental results support the hypothesis that the presence of nonsense regions in XP sequences – introduced by errors of gene finding procedures – significantly increases the predicted disorder content, and therefore introduce bias to genome-wide estimate of disorder content. Yet, we were only able to partially explain the discrepancy in disorder content estimates for NP and XP sequences. It is still possible that the proteins, which are currently covered with XP records, in fact have higher average disorder content than NP sequences. However, even if that is the case we cannot be sure what portion of the difference in predicted disorder content is due to the real difference, and what portion is due to errors in XP sequences that are to be eventually corrected. More importantly, reliability of estimates of disorder content by ID predictors is decreasing with the ratio between the number of curated protein sequences and the number of putative predicted protein sequences in a genome.

There is space for improvement of our nonsense region predictor. We have only addressed the human protein sequences, and extension to other genomes (of both close and distant organisms) is the next natural step. From our experience with the development of predictors of ID, we believe that further optimization of prediction model and training methodology will soon lead to "saturation point", which is also the reason why we did not put much effort into elaborating that part of the study. Instead, we believe that much more can be gained by further refining the dataset. So far we have completely excluded from any consideration all proteins with alternative splicing (AS), because their inclusion requires more complicated synthetic nonsense sequence generation, and in part because AS regions have been shown to be highly disordered [16], [23]. We are working on improving the procedure for generation of synthetic nonsense sequences to also include AS sequences and to give greater focus to areas around splicing points and other genomic regions where errors by gene finding algorithms are expected.

ACKNOWLEDGMENT

This work was supported in part by a grant to Z. O. from the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

REFERENCES

[1] A. K. Dunker et al., "Intrinsically disordered protein.," *Journal of molecular graphics & modelling*, vol. 19, no. 1, pp. 26-59, Jan. 2001.

[2] O. Schweers, E. Schonbrunn-Hanebeck, A. Marx, and E. Mandelkow, "Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure," *J Biol Chem*, vol. 269, no. 39, pp. 24290-24297, 1994.

[3] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury, "NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded," *Biochemistry*, vol. 35, no. 43, pp. 13709-13715, 1996.

[4] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.," *Journal of molecular biology*, vol. 293, no. 2, pp. 321-31, Oct. 1999.

[5] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker, "Natively disordered proteins," *Handbook of Protein Folding*, pp. 271-353, 2005.

[6] P. Romero, Z. Obradovic, and A. K. Dunker, "Intelligent data analysis for protein disorder prediction," *Artificial Intelligence Review*, vol. 14, no. 6, pp. 447–484, 2000.

[7] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein.," *Proteins*, vol. 42, no. 1, pp. 38-48, Jan. 2001.

[8] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder and functional proteomics," *Biophys J*, vol. 92, no. 5, pp. 1439-1456, 2007.

[9] T. E. Creighton, "The protein folding problem," *Science*, vol. 240, no. 4850, pp. 267-344, 1988.

[10] R. M. Williams et al., "The protein non-folding problem: amino acid determinants of intrinsic order and disorder," *Pac Symp Biocomput*, pp. 89-100, 2001.

[11] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Sci*, vol. 11, no. 4, pp. 739-756, 2002.

[12] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder.," *BMC bioinformatics*, vol. 7, p. 208, Jan. 2006.

[13] P. Romero et al., "Thousands of proteins likely to have long disordered regions," *Pac Symp Biocomput*, pp. 437-448, 1998.

[14] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker, "Comparing and combining predictors of mostly disordered proteins.," *Biochemistry*, vol. 44, no. 6, pp. 1989-2000, Feb. 2005.

[15] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J Mol Biol*, vol. 337, no. 3, pp. 635-645, 2004.

[16] U. Midic, C. Oldfield, A. K. Dunker, Z. Obradovic, and V. Uversky, "Protein disorder in the human diseasome: unfoldomics of human genetic diseases," *BMC Genomics*, vol. 10, no. 1, p. S12, 2009.

[17] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models.," *Nucleic acids research*, vol. 26, no. 2, pp. 544-8, Jan. 1998.

[18] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA.," *Journal of molecular biology*, vol. 268, no. 1, pp. 78-94, Apr. 1997.

[19] L. Pachter, M. Alexandersson, and S. Cawley, "Applications of generalized pair hidden Markov models to alignment and gene finding problems.," *Journal of computational biology□: a journal of computational molecular cell biology*, vol. 9, no. 2, pp. 389-99, Jan. 2002.

[20] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker, "Predicting intrinsic disorder from amino acid sequence.," *Proteins*, vol. 53 Suppl 6, pp. 566-72, Jan. 2003.

[21] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, a K. Dunker, and Z. Obradovic, "Optimizing long intrinsic disorder predictors with protein evolutionary information.," *Journal of bioinformatics and computational biology*, vol. 3, no. 1, pp. 35-60, Feb. 2005.

[22] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, Aug. 1996.

[23] P. R. Romero et al., "Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 22, pp. 8390-5, May. 2006.