GENOME-WIDE PREDICTION OF INTRINSIC DISORDER; SEQUENCE ALIGNMENT OF INTRINSICALLY DISORDERED PROTEINS

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Uros Midic
January, 2012

Examining Committee Members:

Zoran Obradovic, Advisory Chair, Computer and Information Sciences
Alexander Yates, Computer and Information Sciences
Keith E. Latham, External Member, Biochemistry
Slobodan Vucetic, Computer and Information Sciences

ABSTRACT

GENOME-WIDE PREDICTION OF INTRINSIC DISORDER; SEQUENCE ALIGNMENT OF INTRINSICALLY DISORDERED PROTEINS

Uros Midic

Doctor of Philosophy

Temple University, 2012

Doctoral Advisory Committee Chair: Dr. Zoran Obradovic

Intrinsic disorder (ID) is defined as a lack of stable tertiary and/or secondary structure under physiological conditions in vitro. Intrinsically disordered proteins (IDPs) are highly abundant in nature. IDPs possess a number of crucial biological functions, being involved in regulation, recognition, signaling and control, e.g. their functional repertoire complements the functions of ordered proteins. Intrinsically disordered regions (IDRs) of IDPs have a different amino-acid composition than structured regions and proteins. This fact has been exploited for development of predictors of ID; the best predictors currently achieve around 80% per-residue accuracy.

Earlier studies revealed that some IDPs are associated with various human diseases, including cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, diabetes and others. We developed a methodology for prediction and analysis of abundance of intrinsic disorder on the genome scale, which combines data from various gene and protein databases, and utilizes several ID prediction tools. We used this methodology to perform a large-scale computational analysis of the abundance of

(predicted) ID in transcripts of various classes of disease-related genes. We further analyzed the relationships between ID and the occurrence of alternative splicing and Molecular Recognition Features (MoRFs) in human disease classes.

An important, never before addressed issue with such genome-wide applications of ID predictors is that – for less-studied organisms – in addition to the experimentally confirmed protein sequences, there is a large number of putative sequences, which have been predicted with automated annotation procedures and lack experimental confirmation. In the human genome, these predicted sequences have significantly higher predicted disorder content. I investigated a hypothesis that this discrepancy is not correct, and that it is due to incorrectly annotated parts of the putative protein sequences that exhibit some similarities to confirmed IDRs, which lead to high predicted ID content. I developed a procedure to create synthetic nonsense peptide sequences by translation of non-coding regions of genomic sequences and translation of coding regions with incorrect codon alignment. I further trained several classifiers to discriminate between confirmed sequences and synthetic nonsense sequences, and used these predictors to estimate the abundance of incorrectly annotated regions in putative sequences, as well as to explore the link between such regions and intrinsic disorder.

Sequence alignment is an essential tool in modern bioinformatics. Substitution matrices – such as the BLOSUM family – contain 20x20 parameters which are related to the evolutionary rates of amino acid substitutions. I explored various strategies for extension of sequence alignment to utilize the (predicted) disorder/structure information about the sequences being aligned. These strategies employ an extended 40 symbol alphabet which contains 20 symbols for amino acids in ordered regions and 20 symbols

for amino acids in IDRs, as well as expanded 40x40 and 40x20 matrices. The new matrices exhibit significant and substantial differences in the substitution scores for IDRs and structured regions. Tests on a reference dataset show that 40x40 matrices perform worse than the standard 20x20 matrices, while 40x20 matrices – used in a scenario where ID is predicted for a query sequence but not for the target sequences – have at least comparable performance. However, I also demonstrate that the variations in performance between 20x20 and 20x40 matrices are insignificant compared to the variation in obtained matrices that occurs when the underlying algorithm for calculation of substitution matrices is changed.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Zoran Obradovic for the chance to study in his lab at Temple University, and all his guidance and support. Dr. Slobodan Vucetic taught an excellent Bioinformatics course during my first semester at Temple University, and subsequently provided valuable advice on countless occasions. To Dr. Alexander Yates and Dr. Keith Latham I thank for their service on the advisory committee.

I could never have completed this dissertation without the help of Dr. Keith Dunker and Dr. Vladimir Uversky, from whom I learned a lot about protein intrinsic disorder, proteomics and bioinformatics in general, and with whom I collaborated on a number of projects and publications.

I deeply appreciate the help of fellow students, the ones who came before me (Bo Han, Kang Peng, and Hongbo Michael Xie), and those who joined the lab later (Alexey Uversky, Debasish Das, Dusan Ramljak, Joseph Jupin, Kosta Ristovski, Mohamed Ghalwash, Ping Zhang, Qiang Lou, Siyuan Ren, and Vladan Radosavljevic), as well as many others in the CIS department and beyond.

My deepest thanks go to my family and all the friends who supported and encouraged me throughout my time at Temple University. To my parents I am forever in debt for teaching me the value of knowledge and education.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

CHAPTER 1

INTRODUCTION

Significant experimental and computational data show that many biologically active proteins lack rigid 3-D structure, remaining unstructured, or incompletely structured, under physiological conditions, and, thus, these proteins exist as dynamic ensembles of interconverting structures. These proteins are known by different names, including intrinsically disordered (A K Dunker et al. 2001), natively denatured (Schweers et al. 1994), natively unfolded (Weinreb et al. 1996), intrinsically unstructured (Wright and Dyson 1999), and natively disordered (Daughdrill et al. 2005) among others. The terms intrinsic disorder (ID), intrinsically disordered protein (IDP), and intrinsically disordered region (IDR) will be used here.

The manifestation of ID is manifold, and functional disordered segments can be as short as only a few amino acid residues or can occupy rather long loop regions and/or protein ends. Proteins, even large ones, can be partially or even wholly disordered. Some IDPs and IDRs exhibit collapsed disordered conformations with pronounced residual structure (thus, resembling a molten globule), others can stay in extended highly disordered states (such as the random coil), while still others form collapsed random coils or semi-collapsed premolten globules (A K Dunker et al. 2001; Daughdrill et al. 2005; V N Uversky 2003; Crick et al. 2006; Tran, Mao, and Pappu 2008). The relationship among the different ID forms needs further study.

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, sequence complexity, hydrophobicity, aromaticity, charge, flexibility

index value, and type and rate of amino acid substitutions over evolutionary time. For example, IDPs are significantly depleted in bulky hydrophobic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe), which form and stabilize the hydrophobic cores of folded globular proteins. IDPs also possess a low content of Asn and of the cross-linking Cys residues. The residues that are less abundant in IDPs, and that are more abundant in structured proteins, have been called order-promoting amino acids. On the other hand, IDPs/IDRs are substantially enriched in polar and charged amino acids: Arg, Gln, Ser, Glu, and Lys and in structure-breaking Gly and Pro residues, collectively called disorder-promoting amino acid residues (A K Dunker et al. 2001; P Romero et al. 2001; P Radivojac et al. 2007). 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 (Creighton 1988), "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 (R M Williams et al. 2001; V N Uversky 2002).

Amino acid differences between IDPs and ordered proteins have been utilized to develop numerous disorder predictors, including PONDR® (Predictor of Naturally Disordered Regions) (P Romero et al. 2001), charge-hydropathy plots (CH-plots) (V N Uversky, Gillespie, and Fink 2000) and IUPred (Dosztanyi et al. 2005) to name a few. Intrinsic disorder predictors fall into two general groups. Per-residue predictors (such as the PONDR® group of predictors) output a score for each residue in a protein and are especially useful when applied to proteins having both structured and disordered regions. Per-protein (or per-sequences) type of algorithm gives a single prediction value for the entire protein sequence. This type is useful when the objective is to identify mostly or

wholly disordered or structured proteins. The charge-hydrophobicity (CH)-plot and the cumulative distribution function (CDF) are the two main predictors of this type (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005). The current state of the art in the field of IDP predictions, including advantages and drawbacks, has been summarized recently (He et al. 2009). Links to many of the servers for these predictors, can be found in the Disordered Protein Database, DisProt (www.disprot.org) (Sickmeier et al. 2007).

Although experimentally characterized IDPs have been discussed in the literature over at least four decades, these proteins have not been viewed as a group, but rather as a collection of unusual protein outliers. Bioinformatics is playing a major role in transforming this collection of examples into a sub-field of protein science. For example, soon after the first disorder predictor was developed (P Romero et al. 1997), it was shown that 25% of proteins in Swiss-Prot contained predicted ID regions longer than 40 consecutive residues and that about 11% of residues in Swiss-Prot were likely to be disordered (P Romero et al. 1998). Subsequent analyses confirmed these trends and revealed that eukaryotic proteomes are significantly more enriched in IDPs in comparison to bacterial and archaeal proteomes (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005; Ward et al. 2004). This increased utilization of IDPs in higher organisms was attributed to the greater need for signaling and coordination among the various organelles in the more complex eukaryotic domain (L M Iakoucheva et al. 2002).

In Section 2.2 I propose a methodology framework for prediction and analysis of protein ID on a genome-wide scale. I compare two indicators of abundance of predicted ID, and address the question of treatment of multiple isoforms encoded by the same gene. In Section 2.3, this methodology is then applied to the whole human genome, to analyze

abundance of ID in proteins related to various disease classes. The analysis revealed that (i) Intrinsic disorder is common in proteins associated with many human genetic diseases; (ii) Different disease classes vary in the IDP contents of their associated proteins; (iii) Molecular recognition features, which are relatively short loosely structured protein regions within mostly disordered sequences and which gain structure upon binding to partners, are common in the diseasome, and their abundance correlates with the intrinsic disorder level; (iv) Some disease classes have a significant fraction of genes affected by alternative splicing, and the alternatively spliced regions in the corresponding proteins are predicted to be highly disordered; and (v) Correlations were found among the various diseasome graph-related properties and intrinsic disorder.

In Chapter 3 I demonstrate that large-scale ID prediction on putative protein sequences in NCBI database, obtained with automated annotation procedures, can be unreliable. I explore the idea of using a predictor to distinguish between the confirmed protein sequences and synthetic nonsense amino acid sequences obtained by translating non-coding regions of the genomic sequences and translating coding regions in the wrong codon alignment. The application of these predictors to the unconfirmed portion of human proteins in NCBI database shows that the unconfirmed proteins contain such nonsense regions that partially bias the estimates of abundance of ID. However, the discrepancy in disorder prediction between confirmed and putative human protein sequences has not been fully explained.

Sequence alignment is an essential tool in modern bioinformatics. The goal of sequence alignment is to arrange two or more (nucleotide or amino acid) sequences in rows of equal length in an attempt to identify similar and evolutionary related sequences. The

alignment process allows mismatching and gaps where mismatches correspond to point mutations while gaps correspond to insertions and deletions. Most alignment algorithms, including BLAST (Altschul et al. 1990) and ClustalW (Chenna et al. 2003), use a matrix of parameters known as *substitution* or *scoring matrix* to assign scores to possible alignments and then look for an alignment with maximal score. Additionally, penalties for gaps can also be controlled with parameters, such as the gap opening penalty and the gap extension penalty.

Substitution matrices are derived from a set of "ground-truth" alignments; PAM matrices (Dayhoff and Schwartz 1978) were developed from a set of manually curated alignments, while BLOSUM matrices (S Henikoff and J G Henikoff 1992) were developed from alignments in the BLOCKS database (J G Henikoff and S Henikoff 1996). There is no natural golden standard for the choice of set of "ground-truth" alignments, and this choice is one of the main sources of variation between various substitution matrices. The score for matching of amino acids a_i and a_j is calculated as $score(a_i , a_j) = C \cdot \log_2(p_{ij} / q_i q_j)$, where p_{ij} is the observed frequency of a_i and a_j being aligned in the "ground-truth" alignments, while q_i and q_j are the observed frequencies of a_i and a_j, and the constant C is selected so that the error introduced by rounding all scores to the nearest integer is minimized. The score is positive if amino acids a_i and a_j are observed aligned as a pair more frequently than would be expected based on their individual frequencies, and negative if they are observed aligned less frequently than would be expected.

The difference in amino acid compositions between IDPs/IDRs and structured proteins casts doubt on the suitability of BLOSUM and PAM matrices for alignment of IDP

sequences, since q_i frequencies are different. The rates of sequence evolution in disordered versus ordered proteins were examined in (Celeste J Brown et al. 2002), where it was found that for 19 out of 26 families of proteins with confirmed intrinsic disorder, the disordered regions evolved significantly more rapidly than the ordered regions, while for only 2 families the opposite was true. A different rate of evolution in IDPs/IDRs means that the frequencies p_{ij} are also inappropriate, and a different substitution matrix is needed for alignment of IDP sequences.

To overcome the lack of "ground-truth" alignments for IDPs, an iterative approach has previously been used (Predrag Radivojac et al. 2002) to obtain a set of alignments of families of proteins with confirmed IDRs and the corresponding substitution matrix. The iterative procedure starts with the BLOSUM62 matrix, aligns all families of proteins and calculates the substitution matrix from obtained alignments. The two steps of alignment and calculation of the substitution matrix are then repeated until no significant changes are observed. The obtained matrix DISORDER is significantly different than the initial BLOSUM62 matrix. However, no clear-cut criterion was established for when this matrix should be used instead of the BLOSUM62 matrix. Furthermore, this matrix always assigns the same score to a pair of amino acids, regardless of whether they belong to IDRs or ordered regions of proteins.

In Chapter 4, I propose and test a radically new approach to protein sequence representation for the purpose of sequence alignment that takes into account the concept of intrinsic disorder and the differences in amino acid compositions and evolutionary rates. The proposed approach includes an extended amino acid alphabet with 40 symbols, which assigns two different symbols to the same amino acid depending on whether it belongs to

an intrinsically disordered region or a structured region. An iterative procedure, similar to (Predrag Radivojac et al. 2002) is used to obtain a 40x40 substitution matrix. This matrix has four 20x20 submatrices that correspond to aligning: 1) ordered to ordered regions, 2) and 2') ordered to disordered regions, and 3) disordered to disordered regions. The most important advantage of this approach is that the alignment algorithms such as Needleman-Wunsch (Needleman and Wunsch 1970), Smith-Waterman (Smith and Waterman 1981) and ClustalW (Chenna et al. 2003) can be modified to use the expanded substitution matrix and utilize the knowledge (experimentally determined or predicted) of intrinsically disordered regions in the sequences being aligned. I found significant and substantial differences between the four submatrices: the scores for alignment of disordered regions to disordered regions are higher than the scores for alignment of ordered regions to ordered regions, which in turn are higher than the scores for alignment of ordered regions to disordered regions. which is further empirical evidence of higher evolutionary rate in disordered regions. However, while this confirmed previously proposed difference in evolutionary rates between ID and ordered regions of proteins, tests with a reference dataset have shown that this approach cannot be successfully exploited to improve the retrieval of homologous sequences for ID query sequences. The extended amino acid alphabet idea was then applied to the original BLOSUM algorithm to construct 40x40 and 40x20 matrices from the BLOCKS database. These matrices were then compared to the standard 20x20 matrices, using the same reference dataset testing methodology. While the results were different, no type of matrix had a clear advantage over the others. Furthermore, simple changes to the underlying BLOSUM algorithm create regular matrices with significantly different performance than the regular matrix produced by the

original algorithm. The effects of all the intricate details of the BLOSUM and other matrix producing algorithms, as well as of the sequence alignment algorithms, have not yet been sufficiently understood. It is necessary to explore them before further attempts to incorporate disorder/order information into the sequence alignment algorithms.

CHAPTER 2

LARGE-SCALE PREDICTION OF INTRINSIC DISORDER

2.1 Related Work and Open Problems

Large-scale genome-wide prediction of ID has been used previously to confirm the ubiquity of ID (P Romero et al. 1998), and to compare the abundance of ID in various genomes and groups of genomes (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005; Ward et al. 2004). Two general types of predictors were used: per-residue predictors that output a score for each residue in a protein, and per-protein predictors that give a single prediction value for the entire protein. Per-residue predictors are especially useful when applied to proteins having both structured and disordered regions. Per-protein predictors are better suited when the objective is to identify mostly or wholly disordered or structured proteins.

The analysis of results of per-protein predictors is straightforward, as they give a single classification output for a whole protein. Predictions for sequences from two sets/genomes can be compared using simple statistical techniques.

Per-residue predictors output a vector of variable length with scores for each individual residue in a sequence. Before comparing predictions for sequences from two sets/groups, the vector of predictions for each sequence has to be summarized into a fixed number of observations. Two approaches have been used to achieve this. In the first approach, the predicted IDRs are identified in a sequence and the longest IDR in the sequence is identified. The length of the longest predicted IDR is then compared to a fixed threshold, and for a set of sequences the fraction of sequences with predicted IDR longer than the

threshold is reported. In the second approach, the fraction of disordered residues is reported. Both indicators (length of longest predicted IDR, fraction of disorder) have their advantages and potential problems. Fraction of disorder as an indicator has an issue with long sequences, as a long sequence with a relatively long predicted IDR can still have a small fraction of disordered residues. On the other hand, analysis shows that the length of longest predicted IDR is biased towards the long sequences, which leads to problems when comparing sets of sequences with significant difference in sequence length distributions. Furthermore, it is less stable with respect to the choice of prediction threshold.

Another important issue is that many genes produce multiple isoforms due to alternative splicing (AS). These differ in length, and due to the inclusion/exclusion of AS regions can have different predictions for regions near the AS splicing sites. This is usually addressed by using general methods for removal of redundant similar sequences. However, redundancy removal procedures select only one of the isoforms coded by the same gene. Additionally, they can remove one of several similar protein sequences even if they are encoded by different genes (orthologs and paralogs). To address this, I propose a method for compilation of prediction vectors for all isoforms of a gene into one long prediction vector that corresponds to the whole coding region of the gene.

2.2 Prediction of Disorder, Treatment of Multiple Isoforms, Comparison of ID Indicators

2.2.1 Methodology

Three predictors of intrinsic disorder were used on the protein sequences: PONDR® VSL2B, CH and CDF. **VSL2B** is a variant of VSL2 predictor described in (K Peng, P Radivojac, et al. 2006). For an amino acid sequence, VSL2B outputs ID prediction in [0,1] range per residue. These outputs are then compared to a threshold (the default threshold is 0.5) and residues with prediction value greater than the threshold are predicted to be ID.

CH and CDF give predictions on the level of whole proteins. The **CH** (Charge-Hydrophobicity) predictor is based on the finding (V N Uversky, Gillespie, and Fink 2000) that two sets of proteins – a set of natively unfolded proteins and a set of small globular folded proteins – occupy two distinct regions in the charge-hydrophobicity phase space, and can be almost perfectly separated with a straight line. The CH predictor calculates the mean hydrophobicity and the mean net charge for a protein sequence, identifies the part of the charge-hydrophobicity plane that the corresponding point belongs to, and calculates its distance from the separating line. The **CDF** predictor (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005; A K Dunker et al. 2000) compiles the predictions of a per-residue predictor to a single binary predictor per protein, by observing the *cumulative distribution function* (CDF) of per-residue predictions, and comparing it to a set of 7 boundary CDF points obtained from a training set (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005). In the case of multiple sequences for one gene, a weighted voting scheme is used to determine a single prediction for the gene. For the CH predictor, the mean of signed distances is calculated (distance is multiplied by –1 if

prediction is negative, i.e. protein is predicted to be ordered). The prediction for the gene depends on the sign of the weighted mean (disorder if the weighted mean is positive, order otherwise). Similarly to the CH predictor, CDF predictor has a parameter (*CDF count*), the mean of which over all proteins sequences for a gene is compared to the threshold to determine a single prediction for the gene.

In the case of multiple isoforms for one gene, the VSL2B predictor was applied to all isoform sequences. To compile all the sequences into one, they were aligned using a special multiple alignment algorithm, which is aimed at rediscovering identical exons in multiple sequences by only matching identical amino acids and optimizing the alignment for long contiguous matched subsequences (Figure 2.1). This was achieved by using a large gap-opening penalty (30 or more instead of a usual 11) and a small gap-extension penalty (0.5 or less). A refinement procedure was then applied to identify possible ambiguous situations where a residue from one sequence can be aligned to two identical amino acids (on two sides of a gap) from another sequence, which can lead to incorrect identification of exons. All multiple alignments were inspected to ensure that the alignment and refinement procedure performs correctly. The sequence that is compiled through this multiple alignment procedure includes all exons from individual sequences, and is considered to represent the whole gene sequence. The unique predictions for the sequence compiled from multiple isoforms are obtained by averaging. For each position in the compiled sequence, the compiled prediction is the mean of predictions for all residues from isoform sequences that are aligned at that position (Figure 2.2).

```
MSTPARRRLMRDFKRLQEDPPAGVSGAPSENNIMVWNAVIFGPEGTPFEDGTFKLTIEFTEEYPNKPPTVRFVSKMF...
MSTPARRRLMRDFKRLQEDPPAGVSGAPSENNIMVWNAVIFGPEGTPFED--------------------------...
------------------------------------------------------------------------MF...

 ...HPNVYADGSICLDILQNRWSPTYDVSSILTSIQSLLDEPNPNSPANSQAAQLYQENKREYEKRVSAIVEQSWRDC
 ...---VYADGSICLDILQNRWSPTYDVSSILTSIQSLLDEPNPNSPANSQAAQLYQENKREYEKRVSAIVEQSWRDC
```

Figure 2.1. Example of alignment of multiple gene isoforms (UBE2A).

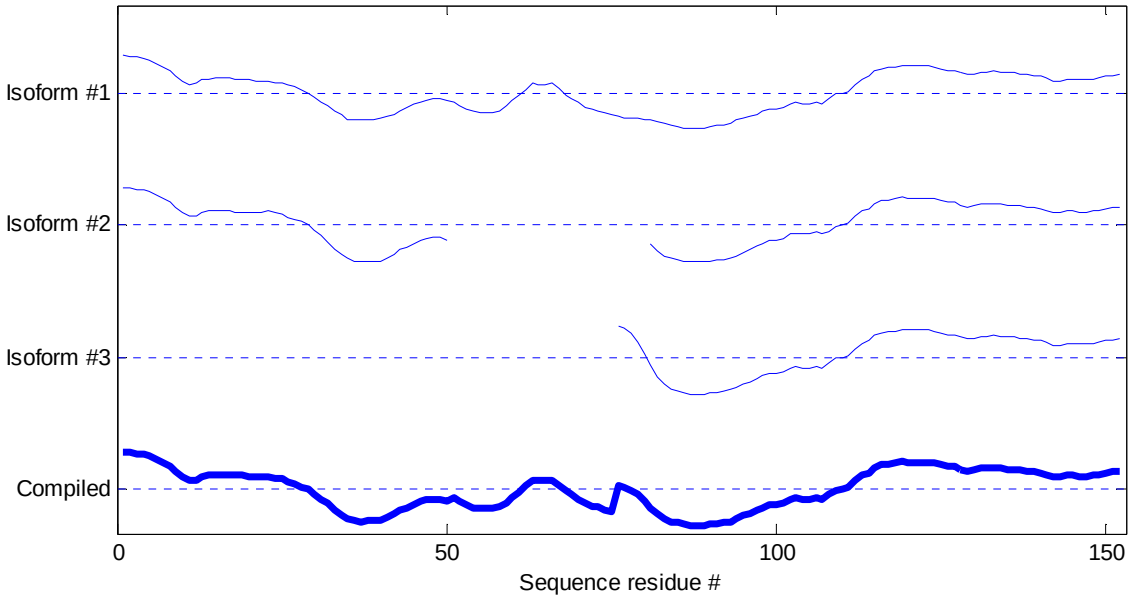

Figure 2.2. Example of VSL2B predictions for multiple isoforms (UBE2A).

Top three curves represent VSL2B predictions for three isoforms of UBE2A. Bottom curve represent the compiled predictions for the compiled sequence. Dashed horizontal lines correspond to the default threshold (0.5).

After obtaining predictions for all residues in a (compiled) sequence, and mapping them to binary predictions by comparing them with the threshold (0.5), we calculate the following values for the sequence. *Disorder total* is the total number of residues that are predicted to be ID. *Disorder content* (DC) is defined as the fraction of residues in the sequence that is predicted to be ID (disorder total/length of sequence). *Longest DR* (LDR) is the length of the longest predicted IDR in the sequence. For the purpose of comparison of DC and LDR, we also compute the ratio *longest DR*/*disorder total*.

2.2.2 Results and Discussion

The comparison of DC (disorder content) and LDR (length of longest disordered region) was performed on the set of 18109 human genes described in detail in Section 2.3 below. VSL2B predictions were obtained and compiled using the methodology described in Section 2.2.1 above.

Figure 2.3 illustrates how the choice of threshold affects the different indicators. As the threshold value increases from .35 to .4, the *total disorder* (which is linearly correlated to *disorder content*) decreases only slightly, while the longest disorder region is broken into two shorter regions and the LDR decreases abruptly. Furthermore, as the threshold decreases from .7 to 0, *total disorder* is close to being linear, LDR at first increases in small steps, and has a rapid increase at the lower values of the threshold.

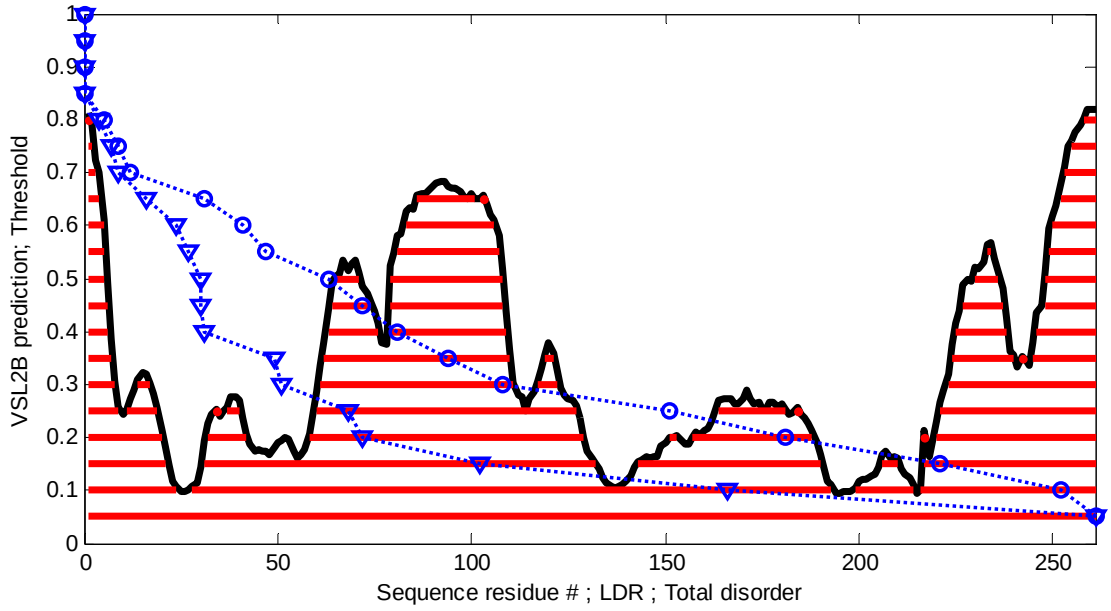

Figure 2.3. Comparison between the length of longest disordered region (LDR) and the total disorder, and effects of threshold selection.

Thick curve represents VSL2B prediction for C1orf38. Horizontal lines represent the regions that are predicted to be ID for various thresholds. Two overlaid dashed curves show the values of LDR (triangles) and total disorder (circles) for various thresholds.

Disorder content (DC) and lengths of longest disordered regions (LDR) were calculated for all 18109 sequences, for 13 threshold values between .35 and .65, with step .05. This produced matrices of size 18109x13. The values of DC obtained for different thresholds were then compared by calculating pair-wise correlation coefficients for all pairs of columns. Similarly, the matrix of all pairwise correlation coefficients was calculated for LDR. Table 2.1 shows the differences between corresponding correlation coefficients for DC and LDR. We can observe that the coefficients are larger for DC, which means that the values of DC obtained with different thresholds are more correlated than the corresponding values of LDR. This means that DC is more stable with respect to the small changes of threshold.

LDR can also be inappropriate as an indicator when a sequence has multiple predicted IDRs, and the longest IDR is only a small fraction of the total disorder in the sequence. Figure 2.4 shows the distributions of *LDR*/*total disorder* ratio for various thresholds. For the default threshold 0.5, this ratio is below .5 in more than half of sequences. In these sequences the longest predicted IDRs represent less than half of total disorder. This problem is even more emphasized for some lower threshold values.

LDR is biased with respect to the length of sequences, as can be observed in Figure 2.5. The LDR is correlated with the sequence length (corr. coef. = .53). More than three quarters of sequences in the six groups with longest sequences (longer than 600 residues) have at least one predicted IDR of length 50 or more. This bias can introduce further difficulties when comparing predicted ID in two sets of sequences with significant difference in length distributions.

While the state-of-the-art predictors like VSL2B achieve around 80% of per-residue accuracy, they still have a false-positive rate of around 20%. For any given IDR length threshold (e.g. L=30), the expected probability that a sequence will have a falsely predicted IDR of length L or longer grows with the length of the sequence. On the other hand, the expected contribution of false positives to DC is constant for all sequence lengths.

In summary, considering the stability issues of LDR and its bias towards long sequences, I decided to use DC as an indicator of ID abundance in a sequence.

Table 2.1. Differences between correlation coefficients obtained for DC and LDR with various thresholds in [.35, .65] interval.

Thr.	.350	.375	.400	.425	.450	.475	.500	.525	.550	.575	.600	.625	.650
.350		.0065	.0073	.0057	.0066	.0023	.0231	.0166	.0132	.0199	.0138	.0063	.0163
.375	.0065		.0045	.0055	.0100	.0077	.0319	.0266	.0240	.0309	.0258	.0183	.0294
.400	.0073	.0045		.0042	.0112	.0112	.0364	.0316	.0305	.0382	.0339	.0271	.0373
.425	.0057	.0055	.0042		.0107	.0122	.0398	.0365	.0359	.0449	.0417	.0355	.0469
.450	.0066	.0100	.0112	.0107		.0057	.0189	.0182	.0194	.0256	.0239	.0197	.0279
.475	.0023	.0077	.0112	.0122	.0057		.0166	.0178	.0201	.0281	.0271	.0240	.0334
.500	.0231	.0319	.0364	.0398	.0189	.0166		.0043	.0086	.0128	.0133	.0122	.0153
.525	.0166	.0266	.0316	.0365	.0182	.0178	.0043		.0065	.0130	.0146	.0145	.0182
.550	.0132	.0240	.0305	.0359	.0194	.0201	.0086	.0065		.0088	.0117	.0130	.0182
.575	.0199	.0309	.0382	.0449	.0256	.0281	.0128	.0130	.0088		.0048	.0081	.0138
.600	.0138	.0258	.0339	.0417	.0239	.0271	.0133	.0146	.0117	.0048		.0051	.0113
.625	.0063	.0183	.0271	.0355	.0197	.0240	.0122	.0145	.0130	.0081	.0051		.0086
.650	.0163	.0294	.0373	.0469	.0279	.0334	.0153	.0182	.0182	.0138	.0113	.0086	

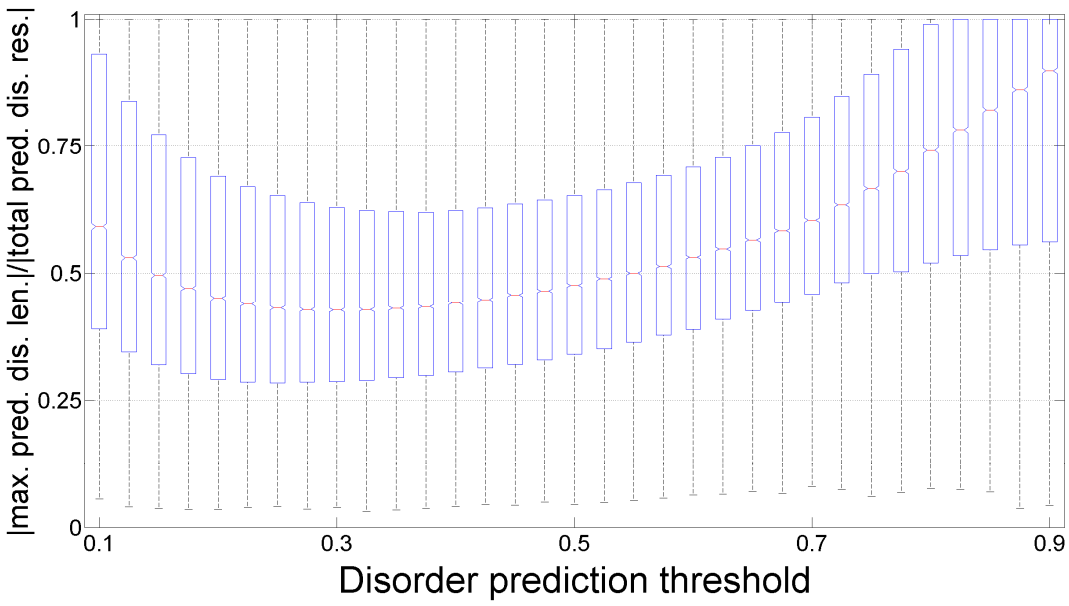

Figure 2.4. Distributions of *LDR/total disorder* ratio for various values of threshold. For each value of threshold, the distribution of *LDR/total disorder* ratio is shown as a boxplot. In each boxplot, the boxes show the range between 25-th and 75-th percentile, while red line represent the median.

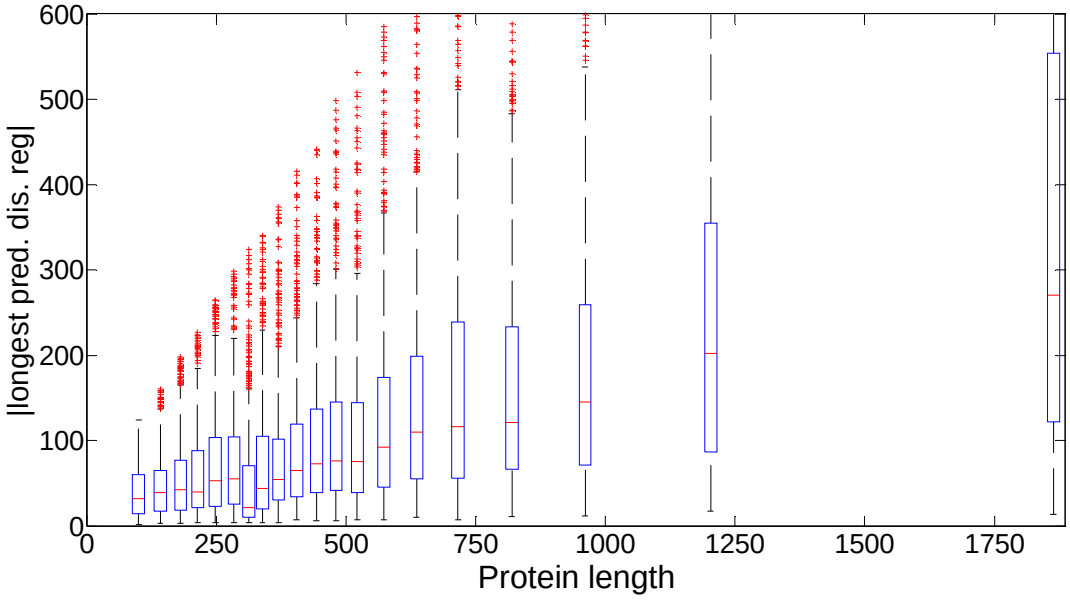

Figure 2.5. Comparison of LDR distributions for various sequence lenghts. **Human genes are grouped by sequence length into 20 equally sized groups. For each group, the boxplot shows the distribution of LDR, and the horizontal position of the boxplot corresponds to the median length of sequences in that group. Note that the graph is truncated at the top.**

2.3 Large-Scale Prediction of ID in Human Genome With Focus on Human Genetic Diseases

2.3.1 Introduction

IDPs carry out numerous biological functions, many of which obviously rely on high flexibility and lack of stable structure. These functions are diverse and complement those of ordered proteins and protein regions. While structured proteins are mainly involved in molecular recognition leading to catalysis or transport, disordered proteins and regions are typically involved in signaling, recognition, regulation, and control by a diversity of mechanisms (Xie, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Vladimir N Uversky, et al. 2007; Slobodan Vucetic et al. 2007; Xie, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Zoran Obradovic, et al. 2007).

IDPs play crucial roles in protein-protein interaction networks, which generally involve a few proteins binding to many partners (called hub proteins or hubs) and many proteins interacting with just a few partners. Consideration of structure data revealed that several hub proteins are entirely disordered, from one end to the other, and to be capable of binding large numbers of partners, other hubs contain both ordered and disordered regions, and some hubs are structured throughout (A K Dunker et al. 2005). Fully disordered hubs can serve as scaffolds for organizing the components of multi-step pathways (Cortese, V N Uversky, and Keith Dunker 2008). For the mixed-structure hubs, many, but not all, of the interactions map to the regions of disorder. For the highly structured hubs (such as 14-3-3 (C J Oldfield et al. 2008) and calmodulin (P Radivojac et al. 2006)), the binding regions of their partner proteins are intrinsically disordered (V N

Uversky, C J Oldfield, and A K Dunker 2005). Overall, these observations support two previously proposed mechanisms by which ID is utilized in protein-protein interactions: namely, one disordered region binding to many partners and many disordered regions binding to one partner (V N Uversky, C J Oldfield, and A K Dunker 2005; A K Dunker et al. 1998).

The binding diversity of IDPs plays important roles in the establishment, regulation and control of various signaling networks. Such disorder-based signaling is further modulated in multi-cellular eukaryotes by posttranslational modification and by alternative splicing, both of which very likely occur much more often in IDRs compared to structured regions of proteins (Xie, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Zoran Obradovic, et al. 2007; P R Romero et al. 2006). Locating alternative splicing in disordered regions avoids the folding problems that arise upon removal of segments from structured domains. The flexibility of IDRs facilitates the binding of the enzymes that bring about the disorder-associated posttranslational modifications. It has been suggested that the intersection of binding sites, posttranslational modifications, and alternative splicing variants within IDRs provide a powerful combination to bring about signaling diversity in different cell types (Xie, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Zoran Obradovic, et al. 2007; C J Oldfield et al. 2008; P R Romero et al. 2006).

Many IDPs and IDRs fold upon binding with their specific partners. Said partners include other proteins, nucleic acids, membranes or small molecules (A K Dunker et al. 2002). The concept of the “molecular recognition feature,” abbreviated as MoRF, was introduced to describe short, intrinsically disordered regions that “morph” from disorder-

to-order upon partner recognition (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Mohan et al. 2006; Vacic et al. 2007). Based on several specific features in the disorder prediction scores, a predictor of helix-forming MoRFs was elaborated (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Yugong Cheng et al. 2007). The application of this predictor to several proteomes revealed that such foldable recognition features are especially abundant among eukaryotic proteins (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Yugong Cheng et al. 2007). MoRFs that form sheet or irregular structure also exist (Mohan et al. 2006; Vacic et al. 2007). Predictors of these non-helical MoRFs have not yet been developed, so the predictions of helix-forming MoRFs should be regarded as providing lower-bound estimates of binding sites in disordered regions.

Proteins are involved in virtually all cellular and in many extracellular processes. Protein dysfunction can therefore cause development of various pathological conditions and a broad range of human diseases known as protein-conformation or protein-misfolding diseases. Such diseases arise from the failure of a specific peptide or protein to adopt its functional conformational state; i.e., from protein misfolding and malfunctioning.

Misfolding diseases can affect a single organ or be spread through multiple tissues. Consequences of misfolding include protein aggregation, loss of normal function, and gain of toxic function. Misfolding and misfunction can originate from point mutation(s) or result from an exposure to internal or external toxins, from impaired posttranslational modification (phosphorylation, advanced glycation, deamidation, racemization, etc.), from an increased probability of degradation, from impaired trafficking, from lost binding partners or from oxidative damage among other causes. These factors can act independently or in complex associations with one another (V N Uversky, C J Oldfield,

and A K Dunker 2008). Furthermore, numerous IDPs are associated with human diseases such as cancer (L M Iakoucheva et al. 2002), cardiovascular disease (Y Cheng et al. 2006), amyloidoses (V N Uversky 2008a), neurodegenerative diseases (V N Uversky 2008b), diabetes and others (V N Uversky, C J Oldfield, and A K Dunker 2008). Based on these intriguing links among intrinsic disorder, cell signaling and human diseases, suggesting that protein conformational diseases may result not only from protein misfolding, but also from misidentification and missignaling (V N Uversky, C J Oldfield, and A K Dunker 2005), the "disorder in disorders" or D^2 concept was recently introduced (V N Uversky, C J Oldfield, and A K Dunker 2008).

Recently, to estimate whether human genetic diseases and the corresponding disease genes are related to each other at a higher level of cellular and organism organization, a bipartite graph was utilized in a dual way: to represent a network of genetic diseases, the "human disease network", HDN, where two diseases are directly linked if there is a gene that is directly related to both of them, and a network of disease genes, the "disease gene network", DGN, where two genes are directly linked if there is a disease to which they are both directly related (Goh et al. 2007). This framework, called the human diseasome, systematically linked the human disease phenome (which includes all the human genetic diseases) with the human disease genome (which contains all the disease-related genes). This diseaseome opens a new avenue for the analysis and understanding of human genetic diseases, moving from single gene-single disease viewpoint to a framework-based approach (Goh et al. 2007).

The analysis of the HDN and DGN properties revealed that these networks are significantly different in many aspects from randomly generated networks of the same

size. By these analyses the various diseases became classified into 20 types, some diseases were unclassified, and several diseases were annotated as belonging to multiple classes. Similarly, genes were clustered into classes via their associations with specific diseases (Goh et al. 2007). Analysis of this network of genetic diseases and disease genes linked by known disease-gene associations revealed the common genetic origin of many diseases. The vast majority of these disease genes was non-essential and showed no tendency to encode hub proteins. Overall, the expression pattern of these disease-related genes indicated that they are localized in the functional periphery of the network (Goh et al. 2007).

In collaboration with domain experts, we started from the disease-related classification of genes from (Goh et al. 2007) and then performed a large-scale analysis of the abundance of intrinsic disorder in transcripts of the various disease-related genes. Since structural information was available only for a limited number of these proteins, we used intrinsic disorder predictions. We also analyzed the correlation between various HDN/DGN graph-related properties of genes and intrinsic disorder. We compared the occurrence of alternative splicing in various disease classes and analyzed the relationship between alternative splicing and intrinsic disorder. In essence, the aim of our study was to build an unfoldome, which we define as the IDP-containing subset of a given genome, associated with human genetic diseases.

2.3.2 Methodology

The basis for our experimental dataset is the dual Human Disease Network/Disease Gene Network (HDN/DGN) (Goh et al. 2007). It consists of two types of nodes that represent human genes (1,777) and diseases (1,284), and links that connect diseases with

related genes. A disease and a gene were connected by a link if mutation(s) in the corresponding gene were implicated in the given disease (Goh et al. 2007). The network is dual, because it can be observed as both a Human Disease Network (two diseases are linked if they are both related to the same gene), or as a Disease Gene Network (two disease genes are linked if they are both related to the same disease).

We augmented the set of disease genes from DGN with human genes with known protein sequences. Protein sequences for all human genes were collected from NCBI Gene database; we excluded all model proteins obtained solely with automated genome annotation processing (proteins with IDs that start with "XP_"). After this exclusion, our dataset consists of 1,751 human disease related genes and 16,358 other human genes with known protein sequences. If several protein sequences were collected for a single gene; i.e., for genes with multiple alternatively spliced isoforms, then any duplicate sequences were discarded.

The diseases in DGN were grouped into twenty classes. In addition to these twenty classes we introduced sets of unclassified diseases and diseases belonging to multiple classes as two separate disease classes. We used this approach to classify genes as well. In our model, a gene belongs to all classes to which its related diseases also belong. Furthermore, since a gene can be related to multiple diseases that belong to various classes, we defined an additional *multiple class gene* group. Thus, overall, this approach defined 22 gene classes: the twenty original classes, as well as classes of *unclassified genes* (related to unclassified diseases) and *multi-class disease genes* (genes related to diseases that belong to multiple classes). Note that the 22 gene classes were not necessarily disjoint, and that all genes from *multiple class gene* class also belonged to at

least two more classes. Two more sets were used for comparison: *disease genes* (this set included all genes from DGN with known protein sequences; i.e., genes from all 22 previously defined classes), and *human genes* (this was the whole dataset that included the *disease genes* set). Table 2.2 contains preliminary statistics for 22 disease/gene classes and 3 additional classes of genes, namely *multiple class genes*, *disease genes*, and *human genes*.

2.3.2.1 Intrinsic Disorder Prediction

We used the methodology described in Section 2.2 to obtain disorder content (DC) from VSL2B predictions, as well as CH and CDF predictions, for proteins (or groups of proteins) encoded by all 18109 collected human genes.

2.3.2.2 A-Morf Predictions

The predictor of an α-helix forming Molecular Recognition Feature (α-MoRF) is based on observations that predictions of order in otherwise highly disordered proteins corresponds to protein regions that mediate interaction with other proteins or DNA. This predictor focuses on short binding regions within long regions of disorder that are likely to form helical structure upon binding (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005). It uses a stacked architecture, where PONDR® VLXT is used to identify short predictions of order within long predictions of disorder and then a second level predictor determines whether the order prediction is likely to be a binding site based on attributes of both the predicted ordered region and the predicted surrounding disordered region. An α-MoRF prediction indicates the presence of a relatively short (~ 20 residues), loosely structured helical region within a largely disordered sequence (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005). Such regions gain functionality upon a disorder-to-helix

transition induced by binding to partner sequences (Mohan et al. 2006; Vacic et al. 2007). Recently it has been indicated that the α-MoRF predictor has a poor sensitivity, i.e., misses many α-MoRF regions , due to the small set of α-MoRF regions used in its development. In this study, the modified α-MoRF predictor, α-MoRF-PredII, was used (Yugong Cheng et al. 2007). This algorithm was improved by including additional α-MoRF examples and their cross species homologues in the positive training set, carefully extracting monomer structure chains from PDB as the negative training set and including attributes from recently developed disorder predictors, secondary structure predictions, and amino acid indices as attributes (Yugong Cheng et al. 2007).

2.3.2.3 Alternative Splicing Analysis

For genes with multiple isoforms, the alignments of multiple isoforms described in Section 2.2 provide the information on the alternative splicing regions. We define the alternative splicing regions (AS regions) as exons that are expressed in some, but not all protein sequences for a given gene. Similarly as for a whole gene, we define disorder content for an AS region as the fraction of its residues that are predicted to be disordered.

Table 2.2. Disease class names and acronyms, number of diseases and number of genes related to disease classes.

The first 22 classes are sorted in descending order with respect to the median of disorder content.

Class name	Acronym	Number of diseases	% (of 1284)	Number of genes	% (of 1751)
Skeletal	SKEL	64	4.98%	56	3.20%
Bone	BONE	30	2.34%	44	2.51%
Dermatological	DERM	48	3.74%	80	4.57%
Cancer	CANC	113	8.80%	207	11.82%
Developmental	DEVE	37	2.88%	53	3.03%
Multi-class disease	MCD	155	12.07%	209	11.94%
Cardiovascular	CARD	41	3.19%	96	5.48%
Muscular	MUSC	31	2.41%	68	3.88%
Immunological	IMMU	69	5.37%	115	6.57%
Ophthamological	OPHT	62	4.83%	120	6.85%
Connective tissue dis.	CTD	28	2.18%	51	2.91%
Endocrine	ENDO	56	4.36%	96	5.48%
Neurological	NEUR	117	9.11%	254	14.51%
Psychiatric	PSYC	17	1.32%	30	1.71%
Ear, Nose, Throat	ENT	6	0.47%	44	2.51%
Respiratory	RESP	13	1.01%	34	1.94%
Renal	RENA	36	2.80%	58	3.31%
Hematological	HEMA	88	6.85%	146	8.34%
Nutritional	NUTR	4	0.31%	22	1.26%
Gastrointestinal	GI	23	1.79%	34	1.94%
Unclassified	UNCL	31	2.41%	29	1.66%
Metabolic	META	215	16.74%	289	16.50%
Multiple class genes	MULT			295	16.85%
Disease genes	DIS			1751	100.00%
Human genes	HUM			18109	

2.3.2.4 Statistical Analysis of the Data

When disorder content measurements – as predicted by VSL2B predictor – for all genes in a disease class were observed as a sample, we used statistical tests to compare the samples arising from different disease classes. Since we cannot make any assumptions on the distributions for disorder content in disease classes, we used the nonparametric Mann-Whitney U test (Wilcoxon rank-sum test) (Mann and Whitney 1947; Wilcoxon 1945) to test whether two samples of observations (i.e. disorder content for two classes) came from the same distribution. The Mann-Whitney U test was not appropriate for similar comparison in the case of CH and CDF predictors, as their predictions were binary. For these two predictors, we counted the number of positive (disordered) and negative (ordered) observation in two samples (classes) and then used the χ^2 test to estimate the likelihood of whether the two samples come from the same distribution.

We dealt with the possible problems of multiple hypotheses testing by controlling *false discovery rate* (FDR) with the Benjamini-Hochberg (for independent tests) (Benjamini and Hochberg 1995) or with the Benjamini-Yekutieli method (Benjamini and Yekutieli 2001).

Several of our hypotheses dealt with the dependency between graph-related numeric properties of nodes representing genes and their disorder content. For a fixed gene, the numeric properties were defined as:

- number of related diseases: number of diseases the gene is directly related to,
- number of related disease classes: number of distinct disease classes that diseases related to the gene belong to, or

- degree: number of other genes that are related to the diseases the gene is related to; or defined in the terms of DGN graph: the number of other genes that are directly linked to the gene (through some disease node).

For such hypotheses we used (first-order) linear regression to model the relationship, and then we used the corresponding F-statistic to assess the validity of the linear model.

The HDN/DGN graph was not completely connected. Using the usual definition of connectivity in graphs, we identified the connected components. One of the components was large and included 516 disease nodes and 903 gene nodes. All of the remaining components contained 15 genes or less; for example, 399 components contain only one gene each. We split the set of disease genes (DIS) into the set of 896 disease genes that belong to the large component (LARGECOMP) and the set of 855 disease genes that belong to one of the smaller components (SMALLCOMPS). Note that although the 16 disease genes with no available protein sequences were not included in the DIS set, and therefore neither in LARGECOMP nor the SMALLCOMPS set, these 16 genes were still included in the HDN/DGN graph for the purpose of identification of connected components.

2.3.3 Results

2.3.3.1 Analysis of ID in Human Diseasome

Prediction of intrinsic disorder using PONDR® VSL2B predictor on all 30053 initially collected protein sequences showed significant differences in predicted ID content for the 7525 (25.04%) model protein sequences obtained with automated genome annotation processing, and the 22528 (74.96%) protein sequences with additional experimental support. The medians of disorder content for model protein sequences was much higher

(68.6% vs. 37.5%), as well as the first quartile (37.9% vs. 21.4%) and the third quartile (96.5% vs. 61.7%). Furthermore, 40.6% of model protein sequences were predicted to have disorder content above 80%, compared to only 11.3% for remaining sequences.

The boxplot in Figure 2.6 depicts the distributions of disorder content for genes in 25 classes. The 22 disease classes are sorted according to their medians of disorder content. The distributions for the majority of classes appear to be positively skewed. The ranges of disorder content between the first and the third quartile differ greatly between classes. For example, *connective tissue disorder* (CTD) class is ranked eleventh in disorder content median among the 22 disease classes, but has the highest third quartile.

The distributions of disorder content in disease classes are further compared in histograms in Figure 2.7. The various classes have irregular disorder content distributions that can hardly be fit by any of the standard distributions. Furthermore, the distributions associated with the different disease classes are dissimilar both in shape and size. For these reasons we use a nonparametric test, Wilcoxon rank-sum test (Mann and Whitney 1947; Wilcoxon 1945), to compare the distributions by comparing their medians.

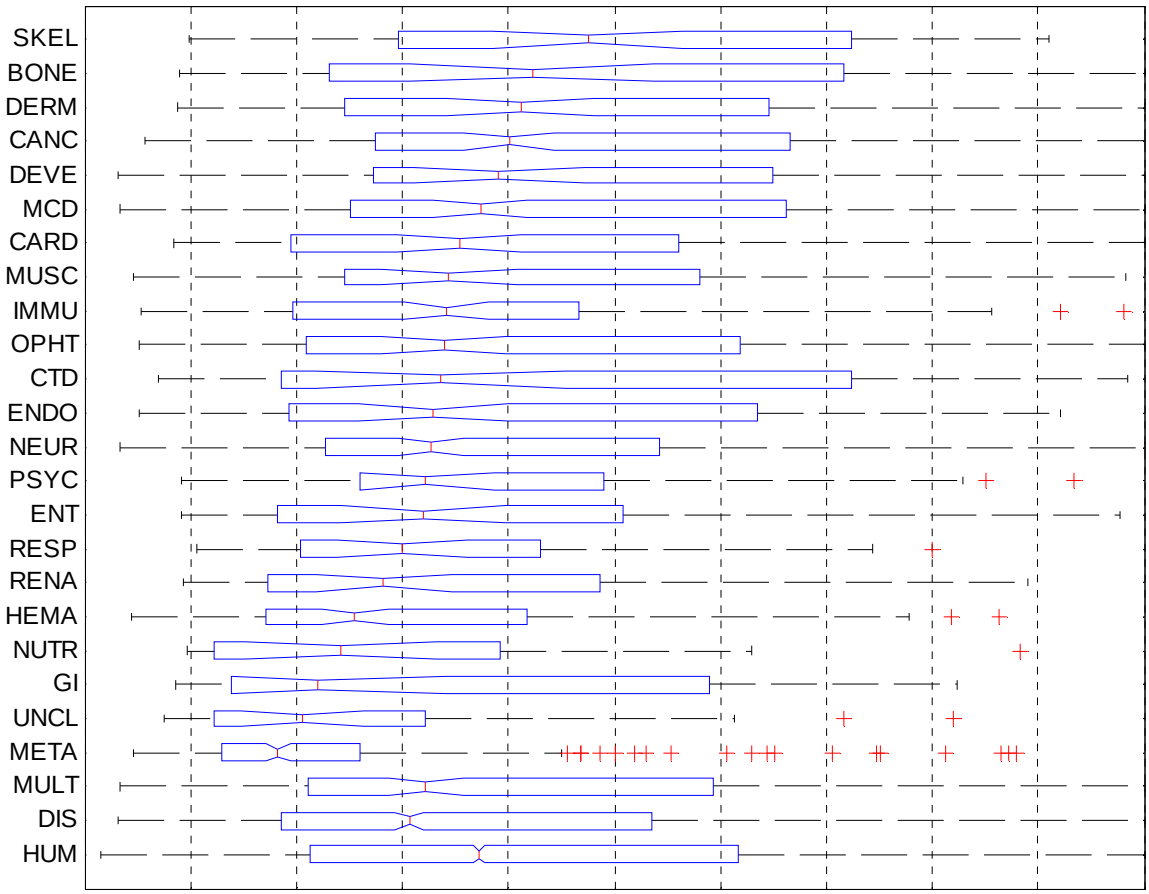

Figure 2.6. Comparison of disorder content distributions in disease classes and human gene class (boxplots).

The 22 disease classes are sorted according to their disorder content medians. The boxes in the boxplot represent the first quartile (left edge), median (line in the middle), and third quartile (right edge); the whiskers extend to the lowest/highest values within the 1.5IQR interval from the box (IQR is the range between the first and the third quartile), while the + signs represent the outliers. Medians for two classes can be compared by looking at the notches at their median lines; if the notches do not overlap, the medians are different at the significance level $\alpha = 0.05$.

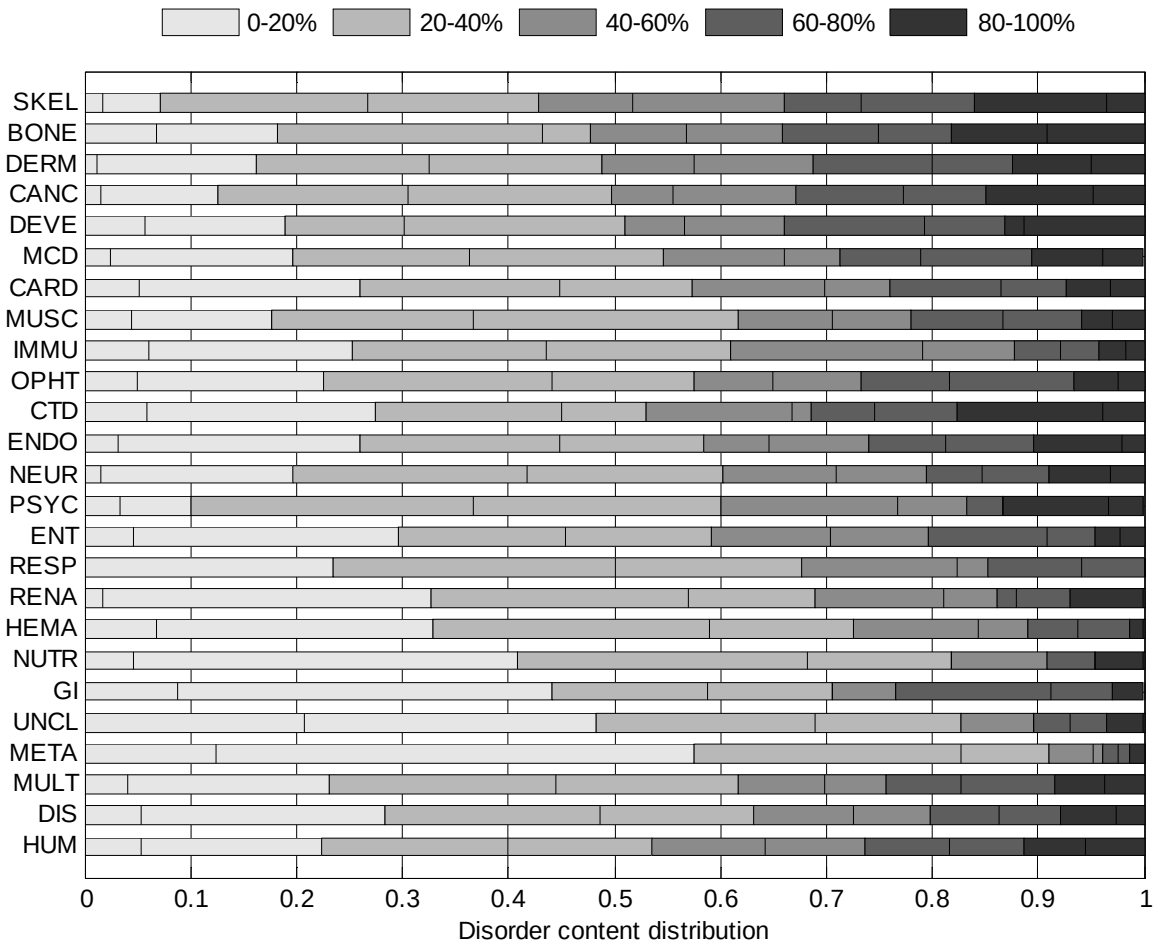

Figure 2.7. Comparison of disorder content distributions in disease classes and human gene class (stacked histograms).

The histograms are stacked horizontally to save space. They show what fraction of genes in each class has disorder content within various ranges. Each of the five major ranges, that cover 20% each, is further split into two smaller 10% ranges (they use the same color, but are divided with a line). Distributions can be visually compared by observing the balance between darker and lighter shades of gray; the class with a darker histogram has on average more disorder content.

Figure 2.8 shows an overview of pair-wise comparisons of disorder content medians. We used Benjamini-Yekutieli (BY) method of *false discovery rate* (FDR) control (Benjamini and Yekutieli 2001), as the *family-wise error rate* multiple comparisons methods, such as the Tukey-Kramer method (Tukey 1953; Kramer 1956), are much more conservative. With an FDR rate of 0.05, it is expected that 2.8 of 56 class pairs reported to have significantly different disorder content medians were false discoveries. The BY method is still quite conservative as it does not make any assumption on the independence of the pair-wise comparisons. Therefore we included Table 2.3 which shows the top 15 *p*-values and BY adjusted *p*-values for comparison of disease classes with *disease gene* (DIS) set, as well as for comparison of disease classes with *human gene* (HUM) set. Several other classes, besides the one indicated in Figure 2.2, can be considered to have disorder content medians significantly different from the DIS and HUM classes, depending on how strict the comparisons are to be. For example, *cancer gene* class has (borderline) significantly different disorder content median than the *human gene* set with a BY false discovery rate of 0.05. Several other classes have low p-values in comparison with *human gene* set, but the adjustment for the BY method pushes them above the 0.05 limit. Note that adjusted *p*-values would be ~3.7 times smaller if we used the Benjamini-Hochberg false discovery method (Benjamini and Hochberg 1995), which makes an assumption that the tests are independent.

Table 2.3. Comparison of disorder content medians in disease classes to disease gene set (DIS) and human gene set (HUM).

Comparison with DIS			Comparison with HUM		
	p-value	FDR *p*-value		*p*-value	FDR *p*-value
META	$9.10 \cdot 10^{-31}$	$7.81 \cdot 10^{-29}$	META	$1.38 \cdot 10^{-50}$	$1.25 \cdot 10^{-48}$
CANC	$9.76 \cdot 10^{-09}$	$4.19 \cdot 10^{-07}$	DIS	$6.16 \cdot 10^{-15}$	$2.79 \cdot 10^{-13}$
SKEL	$3.92 \cdot 10^{-05}$	0.001123	HEMA	$7.13 \cdot 10^{-08}$	$2.15 \cdot 10^{-06}$
MCD	0.000548	0.011771	UNCL	0.000192	0.004349
DERM	0.001852	0.031810	CANC	0.002397	0.043445
HEMA	0.003684	0.052740	NUTR	0.007141	0.107855
UNCL	0.004152	0.050941	SKEL	0.011080	0.143441
DEVE	0.008386	0.090036	GI	0.015816	0.179167
NEUR	0.033455	0.319267	IMMU	0.016768	0.168843
BONE	0.042742	0.367102	RENA	0.026136	0.236856
NUTR	0.063282	0.494113	RESP	0.093824	0.772967
MULT	0.090375	0.646849	MULT	0.105644	0.797813
MUSC	0.122463	0.809091	DERM	0.178919	1.247247
GI	0.130811	0.802516	ENT	0.195823	1.267578
ENDO	0.164391	0.941288	DEVE	0.208293	1.258409

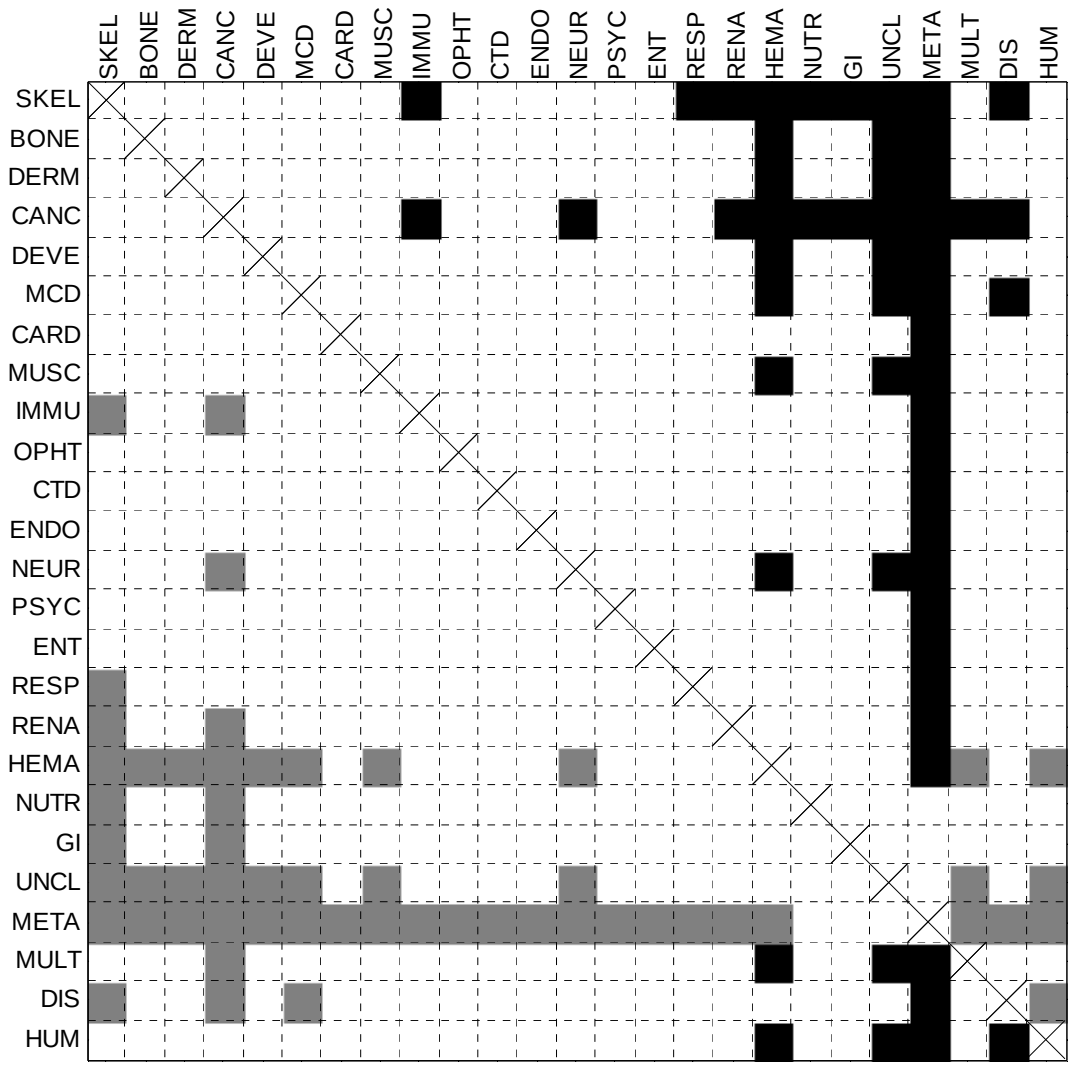

Figure 2.8. Pairwise comparison of disorder content medians for disease classes and human gene class.

Filled squares represent pairs for which adjusted Wilcoxon rank sum test *p*-values are smaller than α=0.05 (*p*-values are adjusted for false discovery rate control with Benjamini-Yekutieli method). Squares are filled black if the median for the row class is greater than the median for the column class, or gray if the median for the row class is smaller than the median for the column class.

We continued with the investigation of the relationship between disorder content and several HDN/DGN graph-related properties. We used linear regression to model disorder content as a linear function of *number of related diseases for a gene* (Figure 2.9), *number of related disease classes for a gene* (Figure 2.10), and *gene degree in DGN* (Figure 2.11). For all three cases, the F-test gave *p*-values that were smaller than 0.05; for the *number of related diseases* and *gene degree* the *p*-values were smaller than 0.01. Although it is not likely that the observed linear trends were obtained by pure chance, they explained only a very small amount of variation in the disorder content; the respective R^2 values were $6.12 \cdot 10^{-3}$, $3.51 \cdot 10^{-3}$, and $6.10 \cdot 10^{-3}$.

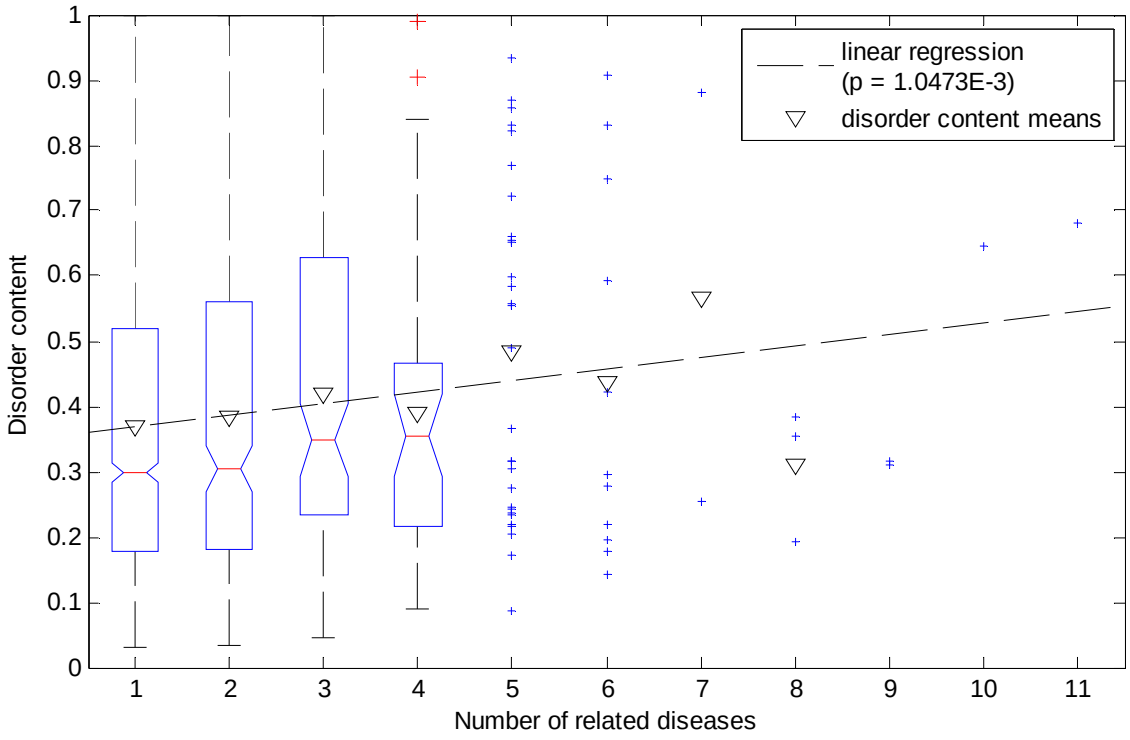

Figure 2.9. Linear regression of disorder content with respect to number of related diseases (for genes).

The genes with number of related diseases up to 4 are represented as a boxplot, while the remaining genes are represented as points. Note that the disorder content means (inverted triangles) for subsets are greater than the respective medians, because the disorder content distributions in these subsets are positively skewed.

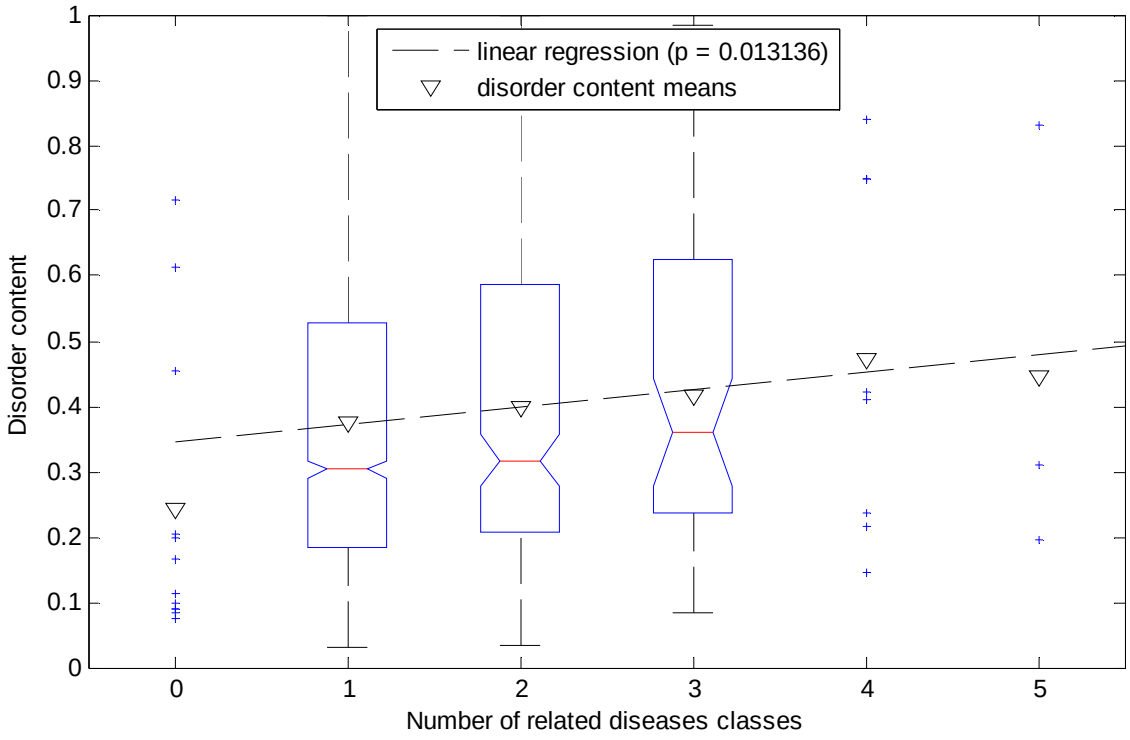

Figure 2.10. Linear regression of disorder content with respect to number of related disease classes (for genes).

The genes with number of related disease classes between 1 and 3 are represented as a boxplot, while the remaining genes are represented as points. Note that the disorder content means (inverted triangles) for subsets are greater than the respective medians, because the disorder content distributions in these subsets are positively skewed.

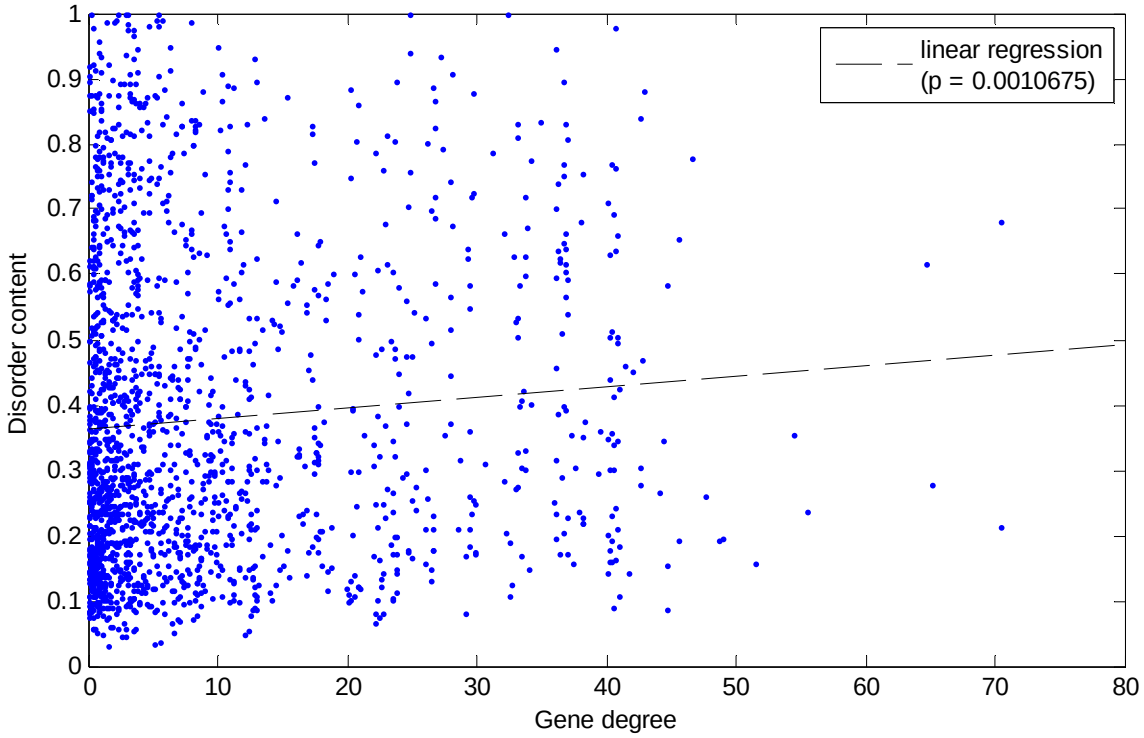

Figure 2.11. Linear regression of disorder content with respect to gene degree in Disease Gene Network.

The disease genes set DIS is split almost evenly between LARGECOMP, the 896 (51.17%) disease genes in the large DGN component, and SMALLCOMPS, the 855 (48.83%) disease genes in the remaining small DGN components. This split can be further observed in individual disease classes. The histogram in Figure 2.12 shows the split between LARGECOMP and SMALLCOMPS for all disease gene classes. Using the χ^2 test to compare the split in each class to the overall split in the disease gene set, we identified classes of disease genes that were significantly overrepresented or underrepresented in LARGECOMP. For example, 85.99% of genes related to cancer diseases belong to the large component, while only 19.03% of genes related to metabolic diseases belong to the large component. We then compared the medians of disorder content for genes from LARGECOMP and SMALLCOMPS for each class individually, as well as for the whole disease genes set. The median of disorder content for LARGECOMP genes was significantly greater than for SMALLCOMPS genes, with an adjusted *p*-value of $7.56 \cdot 10^{-7}$ on the rank sum test. Similarly, the median of disorder content for LARGECOMP genes related to metabolic diseases was significantly greater than for the SMALLCOMPS genes related to metabolic diseases, with an adjusted *p*-value of 0.0112. These comparisons are illustrated in Figure 2.13. Substantial differences between disorder content medians for genes in LARGECOMP and genes in SMALLCOMPS can also be observed for several other classes; in the majority of cases, the median for the LARGECOMP genes is greater than the median for the SMALLCOMPS genes. However, none of these differences were statistically significant; which was partially due to the small numbers of genes in subsets compared.

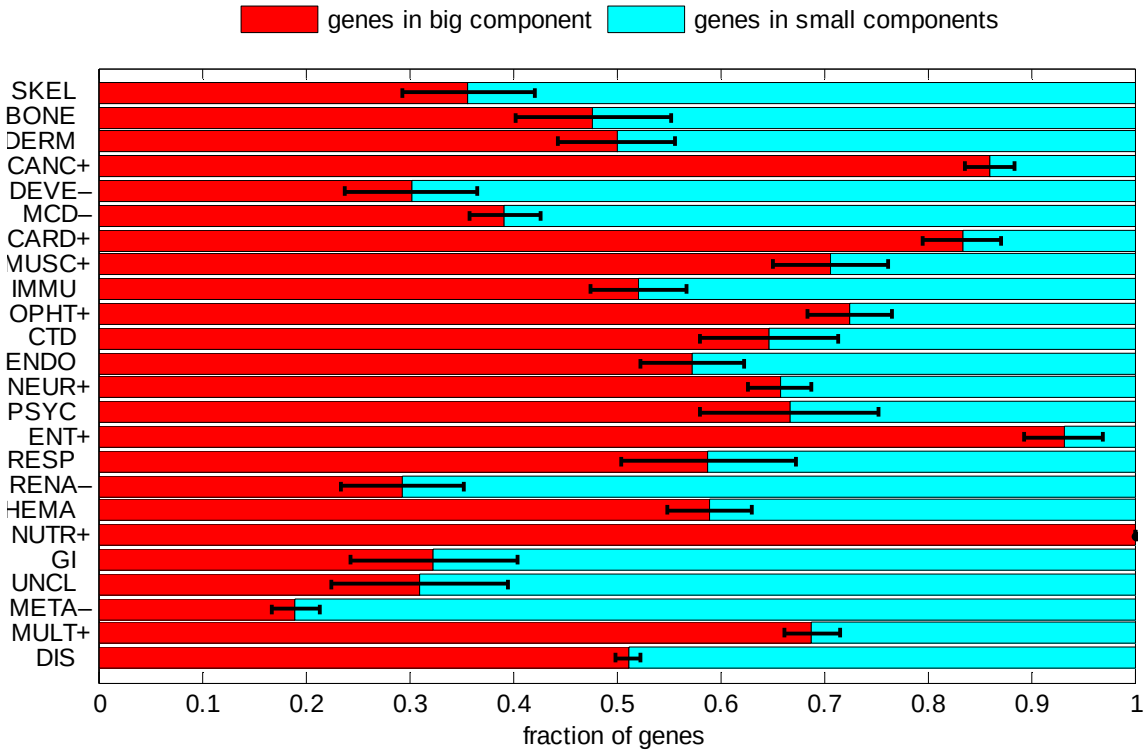

Figure 2.12. Comparison of fractions of disease genes in the large component and the small components of the Disease Gene Network.

The classes with the + signs after their acronyms are significantly overrepresented in the big component; the classes with the – signs after their acronyms are significantly underrepresented in the big component. The error bars represent one standard deviation or 68.2% confidence interval.

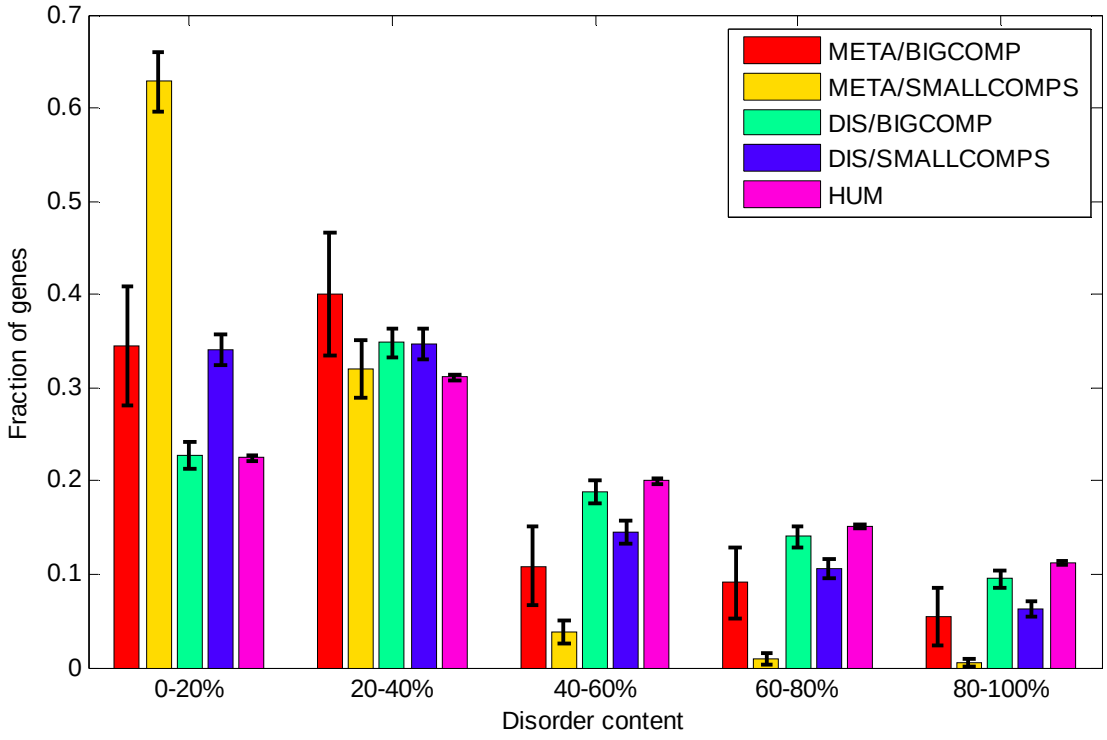

Figure 2.13. Comparison of distributions of disorder content in the large component and the small components of the Disease Gene Network for genes related to metabolic diseases, and for the whole disease gene set.

Distribution of disorder content for human gene set is included for comparison. The error bars represent one standard deviation or 68.2% confidence interval.

2.3.3.2 Alternative Splicing and ID in Human Diseasome

We applied similar methodology to analyze alternative splicing. We divided the set of all genes (HUM) into the set of genes with multiple isoforms and the set of genes with a single isoform. The same division can also be applied to all disease classes, and the disease gene set. The comparison of fractions of genes with multiple isoforms is shown in Figure 2.14.

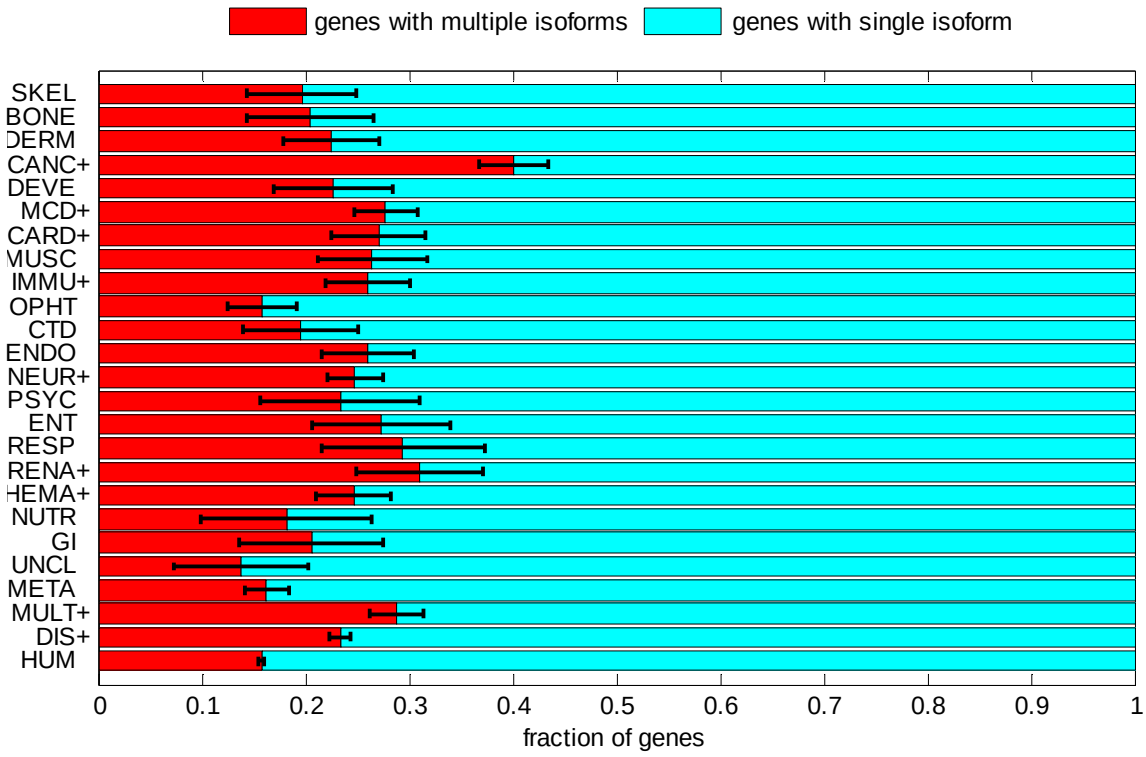

Figure 2.14. Comparison of fractions of disease genes with multiple isoforms (i.e. with alternative splicing) and with single known isoform.

The classes with the + signs after their acronyms have significantly higher fraction of genes with multiple isoforms than the human gene set. The error bars represent one standard deviation or 68.2% confidence interval.

The disease gene set DIS had significantly higher fraction of genes with multiple isoforms than the human gene set HUM. Out of 1751 disease related genes, 410 genes (23.4%) had multiple isoforms (average of 2.77 for disease related genes with multiple isoforms), and they included 991 alternatively spliced regions (2.41 AS regions per disease-related gene with multiple isoforms). Out of 16358 non-disease genes, 2445 (14.95%) had multiple isoforms (average of 2.51 for non-disease genes with multiple isoforms), and they included 4954 AS regions (2.02 AS regions per non-disease gene with multiple isoforms).

Furthermore, all the disease classes but one (unclassified diseases) had higher fraction of genes with multiple isoforms than the HUM set, and for several classes this difference in fractions was statistically significant. The highest fraction of genes with multiple isoforms was 40.10% for the cancer disease gene class.

The comparisons of distributions of disorder content for genes with multiple isoforms with genes with single isoform showed that for three sets the medians of disorder content for genes with multiple isoforms were significantly greater than for genes with single isoform: human genes set HUM (BY adjusted $p = 1.50 \cdot 10^{-7}$), disease genes set DIS (BY adjusted $p = 5.08 \cdot 10^{-7}$) and multiple class genes set MULT (adjusted $p = 0.0176$). Individual tests for three disease classes also returned low p-values (hematological, $p = 0.0196$; renal, $p = 0.0283$; bone, $p = 0.0291$), but the corresponding BY adjusted p-values were above $\alpha = 0.05$.

Figure 2.15 shows the distributions of disorder content for genes with multiple isoforms (disease, non-disease, and all genes) and for all human genes. Although there are significant differences in medians, the distributions have similar shapes; the peaks are in

the 20-40% range, and the fractions decrease with the increase in the disorder content. Figure 2.15 also shows the disorder content distributions for AS regions in disease related and non-disease genes. These two distributions have different shape than shape of the disorder content distributions for whole proteins; the fractions decrease with the increase in the disorder content, but then suddenly increase in the 80-100% range. The rank sum test for medians shows that the distribution of disorder content in AS regions is significantly different from distributions of disorder content in whole proteins for all genes ($p \sim 10^{-142}$), as well as for subset of genes with multiple isoforms ($p \sim 10^{-48}$). However, as is clearly seen in Figure 2.15, the distributions of disorder content for AS regions in disease genes and non-disease genes were not significantly different ($p = 0.5278$). We compared the disorder content distributions for AS regions for genes from individual classes to the overall distribution for AS regions from all human genes. The distributions for classes with significant statistical results are shown in Figure 2.16. For developmental and neurological disease classes, the fraction of AS regions in 80-100% range is significantly increased. Similarly, there is an increase in 0-20% range for hematological disease class. Metabolic disease class is an extreme case, as there is both a big increase in 0-20% range and decrease in 80-100% range; the AS regions in metabolic disease genes have significantly less disorder when compared to whole sequences in human genes.

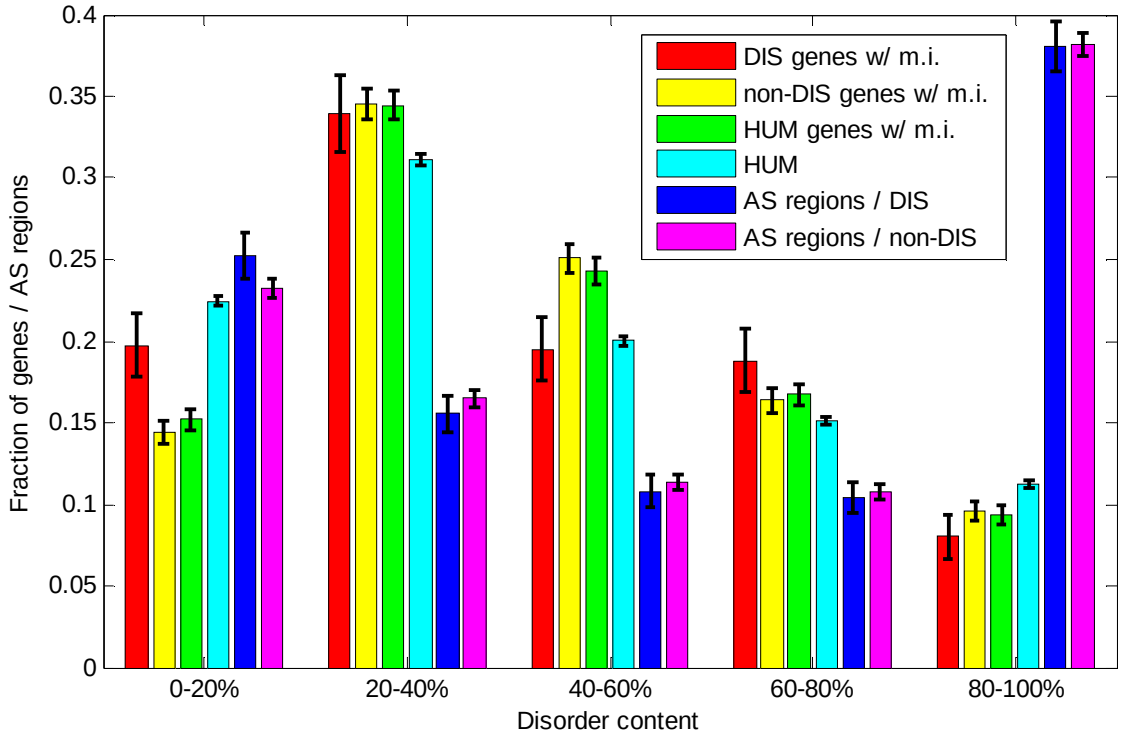

Figure 2.15. Comparison of disorder content distributions for the whole proteins and for the alternative splicing (AS) regions.

Series #1–4 represent disorder content distributions in 1) disease genes with multiple isoforms, 2) non-disease genes with multiple isoforms, 3) human genes with multiple isoforms, 4) all human genes. Series #5 and #6 represent disorder content distributions for AS regions in disease genes, and AS regions in non-disease genes. The error bars represent one standard deviation or 68.2% confidence interval.

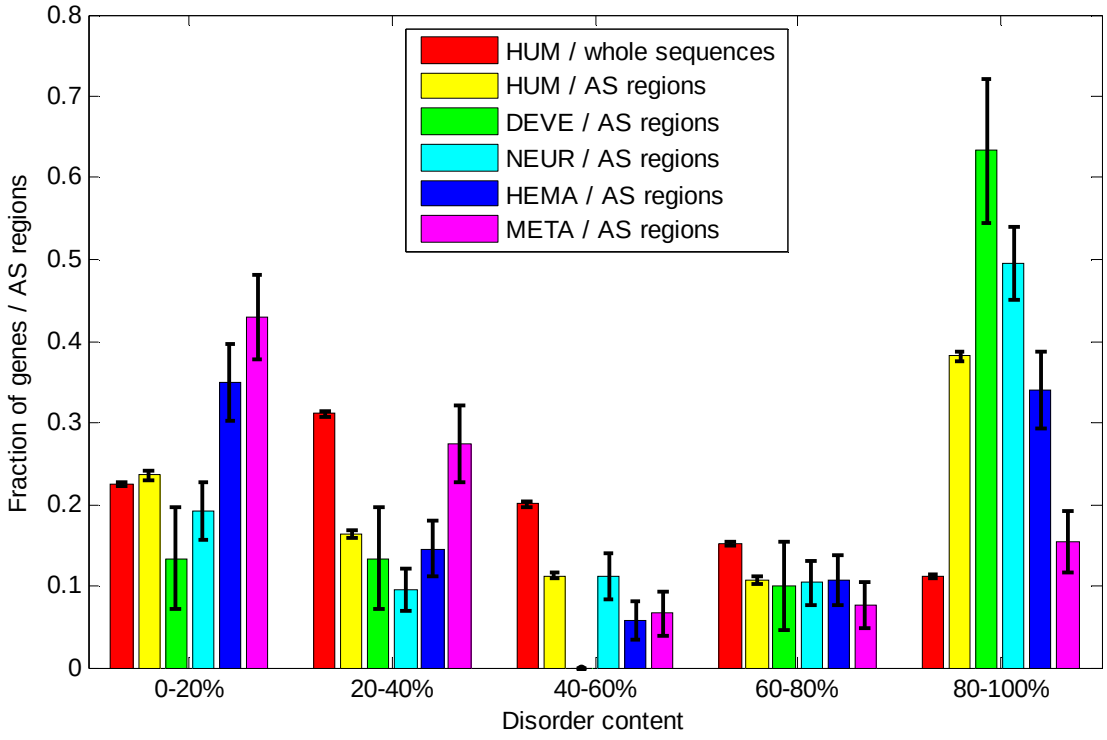

Figure 2.16. Comparison of disorder content distributions for AS regions in various classes of human genes.

Series #1 and #2 represent disorder content distribution for whole human gene sequences and AS regions in human genes. Series #3–6 represent disorder content distributions for AS regions in: 3) developmental, 4) neurological, 5) hematological, and 6) metabolic disease classes. The error bars represent one standard deviation or 68.2% confidence interval.

2.3.3.3 α-MoRFs in the Human Diseasome

Figure 2.17 compiles the α-MoRF prediction data and shows the fractions of genes with predicted α-MoRFs and the densities of α-MoRFs (number of α-MoRFs per residue) for all disease classes, as well as for sets of all disease genes and all human genes. The overall fractions of disordered residues are included for comparison. The fractions of genes with predicted α-MoRFs are highly correlated to fractions of disordered residues (corr. coefficient ~ 0.89).

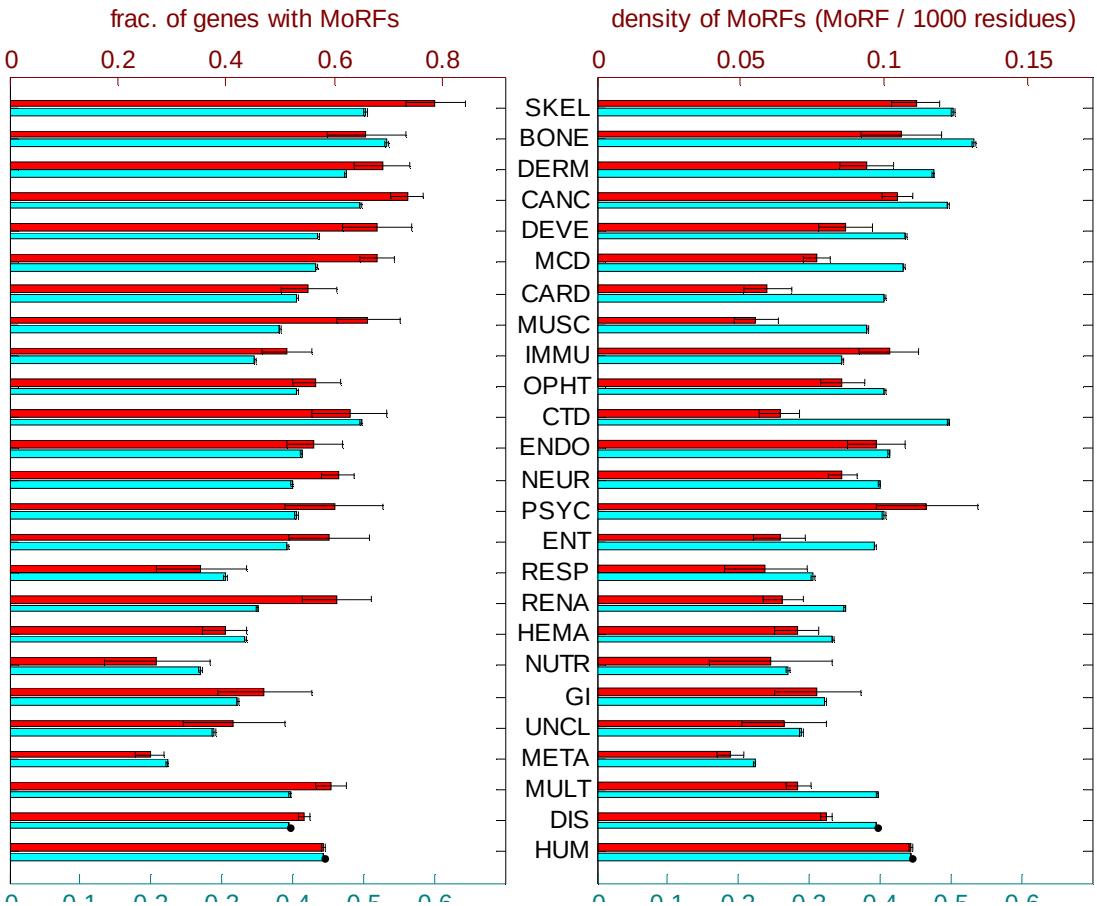

Figure 2.17. Comparison of fractions of genes with predicted α-MoRFs and densities of α-MoRFs with fractions of disordered residues.

The plot on the left compares fractions of genes with predicted α-MoRFs (top/first series) with fractions of disordered residues (bottom/second series). The plot on the right compares densities of α-MoRFs (top/first series) with fractions of disordered residues (bottom/second series). In both plots the series are shown with different scales, such that the values for HUM set are aligned. The error bars represent one standard deviation or 68.2% confidence interval.

2.3.3.4 α-MoRFs and Alternative Splicing in the Human Diseasome

Figure 2.18 compares the overall density of predicted α-MoRFs versus density of predicted α-MoRFs in AS regions for the 25 classes. The differences between densities of MoRFs (overall vs. AS regions) are significant for the majority of classes (listed by increasing *p*-values: HUM, NEUR, META, CANC, DIS, GI, DEVE, IMMU, ENDO, RESP, BONE, DERM, MUSC, CARD, HEMA, ENT), borderline significant for NUTR and OPHT, and not significant for the remaining classes (RENA, MCD, UNCL, SKEL, MULT). Two classes (PSYC, UNCL) have no α-MoRFs predicted in AS regions (while genes in both classes have very small number of residues in AS regions, for PSYC class this difference in densities is still statistically significant). Finally, Table 2.4 lists the quotients of the MoRF density in AS regions over the overall MoRF density, as well as corresponding *p*-value for comparison of these densities.

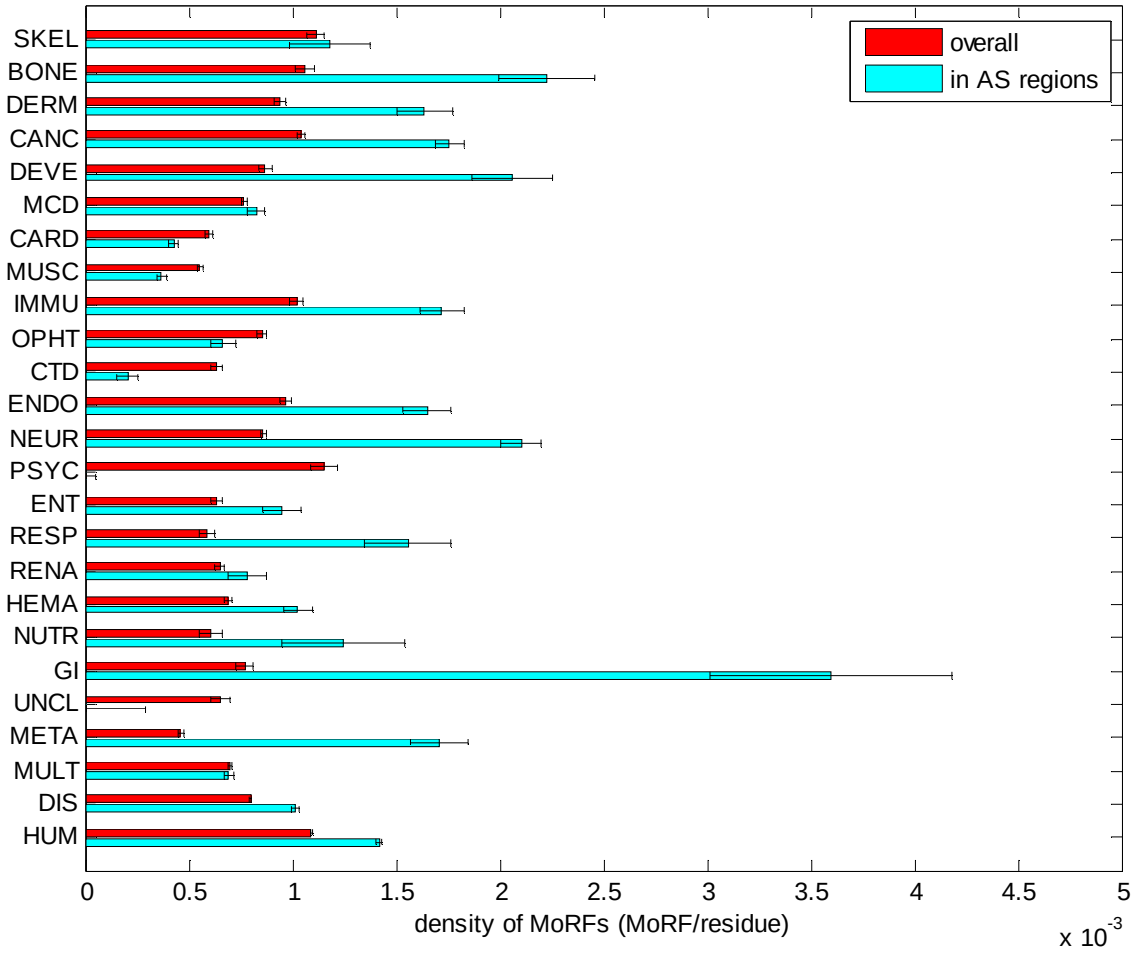

Figure 2.18. Comparison of overall density of predicted MoRFs vs density of predicted MoRFs in AS regions for 25 classes/sets.

The error bars represent one standard deviation or 68.2% confidence interval.

Table 2.4. Comparison of densities of predicted α-MoRFs in AS regions and complete genes (ratios of densities of predicted α-MoRFs in AS regions and overall densities of predicted α-MoRFs; *p*-values for comparison of densities).

Class acronym	Density of MoRFs in AS / Overall density of MoRFs	p-value for comparison of densities of MoRFs
GI	4.68	$3.84 \cdot 10^{-021}$
META	3.71	$1.16 \cdot 10^{-061}$
RESP	2.65	$1.19 \cdot 10^{-010}$
NEUR	2.45	$2.54 \cdot 10^{-076}$
DEVE	2.37	$4.76 \cdot 10^{-017}$
BONE	2.10	$1.28 \cdot 10^{-010}$
NUTR	2.06	0.017694
DERM	1.75	$5.11 \cdot 10^{-010}$
ENDO	1.71	$2.93 \cdot 10^{-011}$
IMMU	1.69	$1.92 \cdot 10^{-013}$
CANC	1.68	$1.07 \cdot 10^{-031}$
ENT	1.50	0.00079422
HEMA	1.48	$4.65 \cdot 10^{-007}$
HUM	1.30	$6.12 \cdot 10^{-233}$
DIS	1.27	$1.10 \cdot 10^{-030}$
RENA	1.21	0.55174
MCD	1.08	0.55638
SKEL	1.06	2.8735
MULT	0.99	3.0621
OPHT	0.78	0.049566
CARD	0.72	$3.30 \cdot 10^{-007}$
MUSC	0.66	$8.89 \cdot 10^{-009}$
CTD	0.32	$3.78 \cdot 10^{-006}$
PSYC	0	$2.07 \cdot 10^{-005}$
UNCL	0	0.56375

2.3.3.5 Evaluation of ID Prediction by Binary Classifiers

We compared the fractions of genes predicted to be disordered by per-protein predictors CDF and CH in Figure 2.19. Overall, the CDF predictor identified more genes to be disordered than the CH predictor. The ratio was 2.79 for human gene set, and 4.64 for the disease genes set. For disease classes it ranged from 2.63 for hematological disorder genes to 15.50 for immunological disorder genes; additionally, for three disease classes – respiratory, renal and unclassified – the CH predictor predicted all genes to be ordered. The correlation coefficient for two vectors of fractions was 0.66.

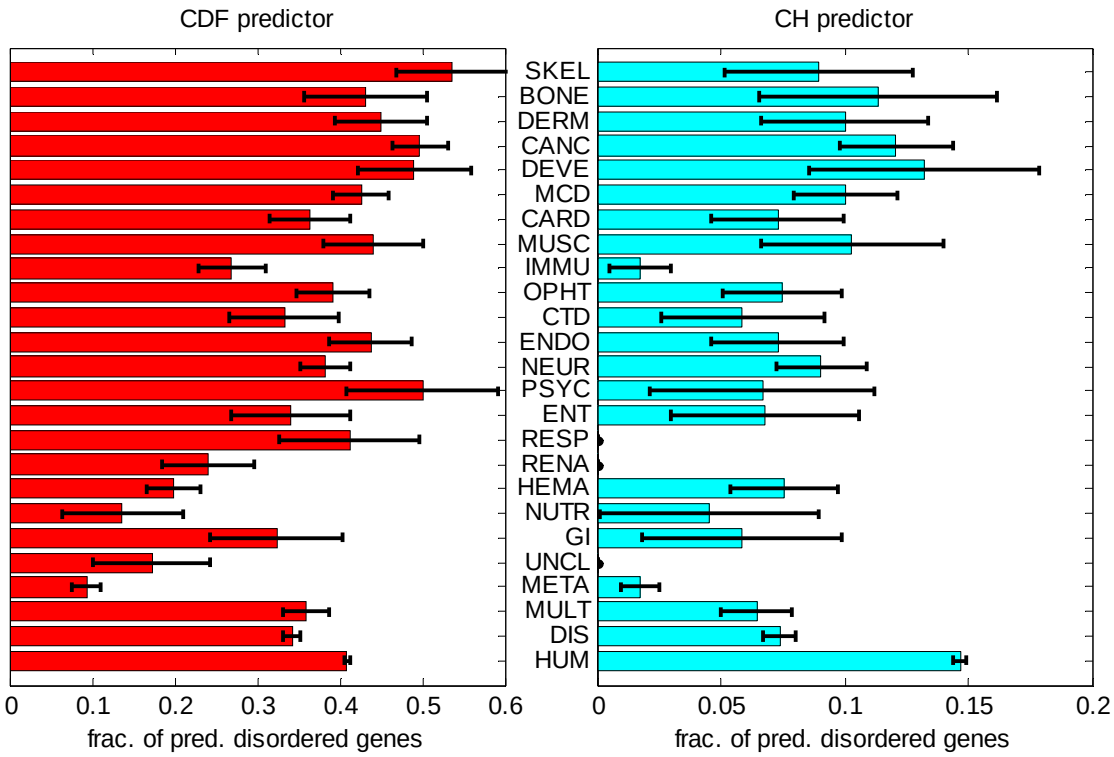

Figure 2.19. Fractions of genes predicted to be disordered by CDF and CH predictors.

The error bars represent one standard deviation or 68.2% confidence interval.

When compared to HUM gene set, the CDF predictor identified significantly different fractions of disorder for the classes META, DIS, HEMA, and (borderline significance) IMMU, while the CH predictor predicted significantly different fractions of disorder for the classes DIS, META, MULT, IMMU, and RENA (in all these classes, fractions of predicted disordered genes were significantly smaller than the fraction for HUM set). When compared to DIS gene set, the CDF predictor identified significantly different fractions of disorder for classes META, CANC, HEMA, and SKEL, while the CH predictor predicted significantly different fractions of disorder for class META. Classes META and HEMA had smaller fractions of predicted disordered genes than DIS set, and classes CANC and SKEL had greater fractions of predicted disordered genes than DIS set.

The relationship between alternative splicing and intrinsic disorder, as predicted by CDF and CH predictors, can only be observed at the level of whole proteins. For the CDF predictor, classes with significantly different fractions of predicted disordered genes in genes with single isoform and in genes with multiple isoforms were: DIS, HUM, RENA, MULT, and (borderline significance) HEMA; in all cases, fraction of predicted disordered genes for genes with multiple isoforms was greater than for genes with a single isoform. For the CH predictor, significant difference of fractions of predicted disordered genes in genes with a single isoform and in genes with multiple isoforms was only observed in the HUM set.

In general, CDF predicts a much higher fraction of genes to be disordered than CH. Vectors of fractions of predicted disordered genes in various classes for CDF and CH predictors are fairly correlated, though there are several classes with substantial differences. For example, IMMU, RESP, RENA, and UNCL have very low (or even zero)

fractions of disordered genes for CH predictor. The relative difference between HUM and DIS sets is much larger for the CH predictor (approximately two-fold) than for the CDF predictor. For the CDF predictor, the fractions of predicted disordered genes for several classes are higher than (although not strictly significantly higher) or similar to the same fraction in the HUM set, while this is not the case for CH predictor.

Overall, the fractions of predicted disordered genes for both binary predictors are correlated to medians of disorder content for VSL2B predictors, but there are some striking differences. For example, low fractions of disordered genes in IMMU class for both predictors, or relatively high fraction of PSYC and RESP classes for CDF predictor. Looking only at the medians without at least comparing whole distributions is not a good way to compare prevalence of intrinsic disorder in two classes/sets of genes.

The difference between these two methods in the magnitude of predicted disorder is generally similar to previously published data (C J Oldfield, Y Cheng, Cortese, C J Brown, et al. 2005; Y Cheng et al. 2006; Mohan et al. 2008). This difference was explained by the fact that the CH-plot is a linear classifier that takes into account only two parameters of the particular sequence – charge and hydropathy, whereas CDF analysis is dependent upon the output of the PONDR® VLXT predictor, a nonlinear neural network classifier, which was trained to distinguish order and disorder based on a significantly larger feature space that explicitly includes net charge and hydropathy. According to these methodological differences, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, PONDR-based CDF analysis may discriminate all disordered

conformations including molten globules from rigid well-folded proteins. Therefore, this discrepancy in the disorder prediction by CDF and CH-plot might provide a computational tool to discriminate proteins with extended disorder from native molten globules, which might be predicted to be disordered by CDF, but compact by CH-plot. This model is consistent with the behavior of several intrinsically disordered proteins (e.g., (Lavery and McEwan 2008)).

Figure 2.20 compares the results of the CH-plot and CDF analyses by showing the distributions of proteins in each disease within the CH-CDF phase space. In these plots, each spot corresponds to a single protein and its coordinates are calculated as a distance of this protein from the boundary in the corresponding CH-plot (Y-coordinate) and an averaged distance of the corresponding CDF curve from the boundary (X-coordinate). Positive and negative Y values correspond to proteins which, according to CH-plot analysis, are predicted to be natively unfolded or compact, respectively. Whereas positive and negative X values are attributed to proteins that, by the CDF analysis, are predicted to be ordered or intrinsically disordered, respectively. Therefore, each plot contains four quadrants: (-, -) contains proteins predicted to be disordered by CDF, but compact by CH-plot (i.e., potential native molten globules); (-, +) includes proteins predicted to be disordered by both methods (i.e., proteins with extended disorder); (+, -) contains ordered proteins; (+, +) includes proteins predicted to be disordered by CH-plot, but ordered by the CDF analysis. A sharp cut-off at the right side of each plot is due to the upper limit of a difference between the CDF curve (which might have a maximum value of 1.0) and a boundary separating IDPs and ordered proteins in CDF plots. Figure 2.20 suggests that the majority of the wholly disordered proteins could possibly be native molten globules.

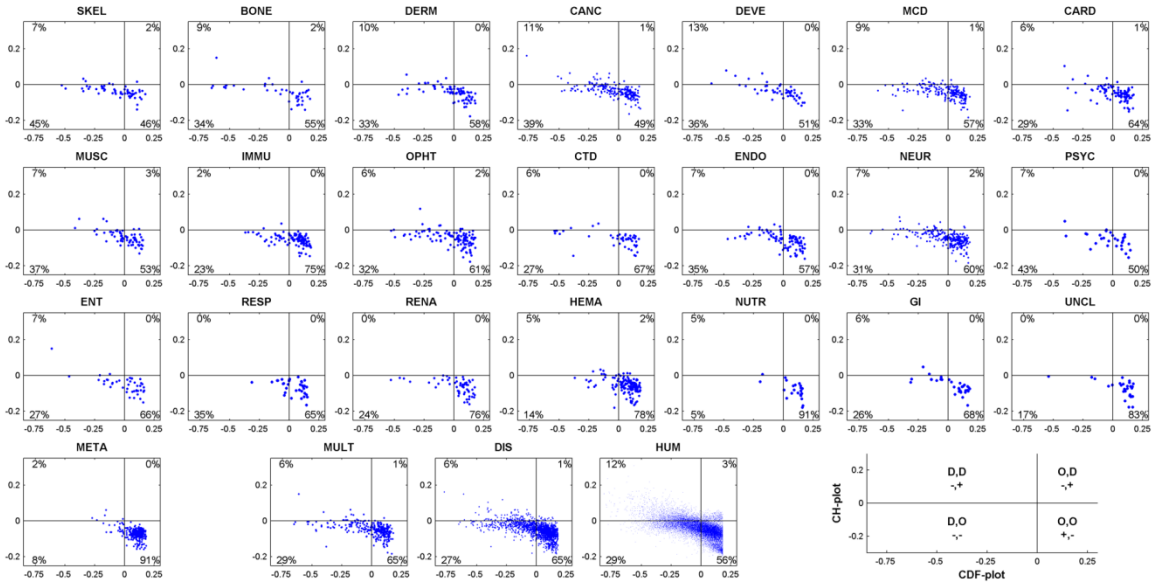

Figure 2.20. Comparison of CDF and CH predictions in various disease gene classes and gene sets.

Each spot represents a gene whose coordinates were calculated as the distance of the corresponding point in the CH-plot from the boundary (x-coordinate) and the averaged distance of the corresponding CDF-curve from the CDF boundary (y-coordinate). Four quadrants in each plot correspond to the following predictions: (–,–) proteins predicted to be disordered by CDF, but compact by CH, (–,+) proteins predicted to be disordered by both methods, (+,–) ordered proteins, (+,+) proteins predicted to be disordered by CH, but ordered by CDF. This is further illustrated by an explanatory plot at the bottom right corner. Percentages represent the fractions of genes in the corresponding quadrants.

2.3.4 Discussion

An important assumption that we made was that ID predictors have no bias towards any class of genes. Although errors are unavoidable in prediction of disorder, we assumed that both false positive and false negative errors occur equally likely in all gene classes. Under this assumption we can expect that any observed variations in predicted disorder content between disease classes are due to real variations in disorder content and not due to bias introduced by prediction. Although we have not found any obvious reason for questioning these assumptions, more structural data are needed to test such biases.

2.3.4.1 Intrinsic Disorder in Human Genetic Diseases

Contrary to our initial expectations based on known abundance of ID in such diseases as cancer (L M Iakoucheva et al. 2002), cardiovascular disease (Y Cheng et al. 2006), amyloidoses (V N Uversky 2008a), neurodegenerative diseases (V N Uversky 2008b), diabetes and others (V N Uversky, C J Oldfield, and A K Dunker 2008), the disease genes have in general slightly lower disorder content than the non-disease genes. This can be explained by the fact that the human disease network (HDN) and disease gene network (DGN) are based on the genetic diseases and genes, mutations in which were associated with disease development, respectively. Based on the expression pattern analyses of the DGN genes it has been concluded that they are mostly localized in the functional periphery of the protein-protein interaction network (Goh et al. 2007). This peripheral localization of most disease genes was explained assuming that mutations in topologically central, highly connected, and widely expressed genes were more likely to result in severe impairment of normal development, leading to early lethality and therefore to deletion from the population, whereas mutations compatible with survival into the reproductive

years were more likely to be maintained in a population (Goh et al. 2007). Overall, the vast majority of disease genes in DGN was non-essential and showed no tendency to encode hub proteins (Goh et al. 2007). On the other hand, the above mentioned studies on various individual diseases (L M Iakoucheva et al. 2002; V N Uversky, C J Oldfield, and A K Dunker 2008; Y Cheng et al. 2006; V N Uversky 2008a; V N Uversky 2008b) dealt with all proteins known to be associated with a given disease and not just those proteins bearing the disease-promoting mutations. Therefore, the various datasets of proteins associated with individual diseases contained wider variety of proteins, including hubs. It is important to remember that hub proteins were shown to be highly enriched in intrinsic disorder (A K Dunker et al. 2005; Cortese, V N Uversky, and Keith Dunker 2008; C J Oldfield et al. 2008; Haynes et al. 2006; Patil and Nakamura 2006; Dosztanyi et al. 2006; Ekman et al. 2006; Singh et al. 2006). In fact, hubs were shown to have multiple interactions, either being intrinsically disordered and serving as an anchor, or acting as a stable globular scaffold that interacts with intrinsically disordered regions of its targets (A K Dunker et al. 2005; Cortese, V N Uversky, and Keith Dunker 2008; C J Oldfield et al. 2008; Haynes et al. 2006; Patil and Nakamura 2006; Dosztanyi et al. 2006; Ekman et al. 2006; Singh et al. 2006). Therefore a systematic depletion of hub proteins in HDN and DGN can in part explain their slightly lower disorder contents.

2.3.4.2 Hubness and Intrinsic Disorder in Human Diseasome

Linear regression of disorder content with respect to number of related diseases, number of related disease classes, and gene degree, shows that the correlation between disorder content and these graph-related gene features are positive and significant. The very low R^2 coefficient tells us that disorder content cannot be predicted from these

features (which was never our intention), but that the positive correlations should be observed as trends.

A number of related disease classes and gene degree are features related to whether a gene/protein is a hub. The observed trends in predicted disorder content provide additional support for the hypothesis that hub proteins are more likely to be disordered, to accommodate the various interactions and functions they are involved with (A K Dunker et al. 2005). All three graph-related gene features are related to the partition of the HDN/DGN graph into one large connected component and a series of small connected components. Genes for which any of the three graph-related features is a high number belong to the large component. Since such genes are more likely to be disordered, they contribute to the difference in disorder content between large component and small components. This difference is particularly significant for genes related to metabolic diseases. More than 60% of metabolic disease genes that belong to the small components have disorder content in the 0-20% range, and further 30+% have disorder content in the 20-30% range. On the other hand, 25% metabolic disease proteins that belong to the large component have disorder content higher than 40%, which is lower when compared to other disease proteins, but substantially higher than the level of ID in metabolic disease proteins in the small component. Of note, most of metabolic disease genes in the large component are also related to disease from other classes.

The difference in disorder content between one large connected component of HDN/DGN and remaining small connected components has to be observed with caution. The connectivity of HDN/DGN is influenced heavily by small components. Only one link between a gene/disease in the large component and a disease/gene in some small

components that has not yet been established, but is discovered in the future can change the partition completely, by leading to inclusion of that whole small component into the large component.

2.3.4.3 Alternative Splicing, Intrinsic Disorder and Human Genetic Diseases

Prediction of intrinsic disorder in proteins encoded by genes with alternative splicing shows that AS regions have much higher predicted disorder content than the whole protein sequences. This is in agreement with previous observations (P R Romero et al. 2006). No difference was observed in disorder content for AS regions in disease and non-disease genes/proteins; the distributions were almost identical. However, alternative splicing can be observed as an important link between diseases and intrinsic disorder, as several disease classes have significantly higher fraction of genes with multiple isoforms; i.e., with AS regions. The presence of AS regions in such genes is associated with increased disorder content. Distributions of disorder content in AS regions were fairly similar across various genes, except for three classes. DEVE and NEUR have a very high fraction of highly disordered AS regions (disorder content 80-100%). This fact might be related to the functionality of proteins involved in these diseases (see above). AS regions in META genes are much less disordered than AS regions in other disease classes, just like whole META gene sequences are much less disordered than other disease genes.

2.3.4.4 Abundance of α-MoRFs in Proteins Associated with Human Genetic Diseases

IDRs frequently participate in protein-protein interactions and molecular recognitions (A K Dunker et al. 2001; Daughdrill et al. 2005; P Radivojac et al. 2007; L M Iakoucheva et al. 2002; V N Uversky, C J Oldfield, and A K Dunker 2005; C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Tompa 2002). Many IDPs and IDRs undergo disorder-to-

order transitions upon binding, which is crucial for recognition, regulation, and signalling (A K Dunker et al. 2001; Wright and Dyson 1999; V N Uversky, Gillespie, and Fink 2000; C J Oldfield et al. 2008; C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Mohan et al. 2006; Vacic et al. 2007; Yugong Cheng et al. 2007; A K Dunker and Z Obradovic 2001; Dyson and Wright 2002; Dyson and Wright 2005). A recent confounding observation is that not all specific interactions between intrinsically disordered proteins and their partners are necessarily accompanied by the disorder-to-order transitions, but may somehow remain unstructured even after binding (A Sigalov, D Aivazian, and L Stern 2004; A B Sigalov 2004; A B Sigalov 2006; A B Sigalov et al. 2006; A B Sigalov, Zhuravleva, and Orekhov 2007). Nevertheless, a correlation has been established between the specific pattern in the PONDR® VL-XT curve and the ability of a given short disordered regions to undergo a disorder-to-helix transition upon binding (E Garner et al. 1999). Based on these specific features, a predictor helix-forming MoRFs was recently developed (C J Oldfield, Y Cheng, Cortese, P Romero, et al. 2005; Yugong Cheng et al. 2007). Not all helix forming MoRF regions share these same features, and some MoRFs form β- or irregular structure rather than the α-helix (Mohan et al. 2006; Vacic et al. 2007). A further complication is that MoRFs can exhibit partner-dependent structures, with at least one example morphing into helix, sheet, or irregular structure, depending on the partner (C J Oldfield et al. 2008). Overall, therefore, these predicted MoRFs represent only fractions of the total numbers of MoRFs for each organism.

The application of the α-MoRF predictor to various datasets reveals that helix forming molecular recognition features are highly abundant in proteins associated with all human genetic diseases as well as in proteins encoded by disease genes and by all human genes,

suggesting the existence of extensive interaction networks. In the HUM set, 57.9% of human genes contain α-MoRFs. In the DIS set, 54.4% of all disease-associated genes contain α-MoRFs, with significant variation between various disease classes, ranging from 26.0% in metabolic diseases and 27.3% in nutritional disorders to 73.4% in cancer and 78.6% in skeletal diseases. In most disease classes some long, highly disordered proteins have multiple predicted α-MoRF regions (Table 2.4) that may potentially serve as binding sites for multiple proteins. For example, DMD from CARD/MUSC (7 predicted α-MoRFs, 3771 amino acids, 54.4% disorder content); MITF from MCD (5, 598, 75.6%); DTNA from CARD (4, 767, 61.0%); EDA from DERM (4, 460, 63.9%); PLEC1 from DERM/MUSC (3, 4904, 56.6%); BRCA1 from CANC (3, 1864, 80.6%); GNAS from BONE/CANC/MCD/ENDO (3, 1323, 74.9%); OPA1 from OPHT (3, 1015, 36.1%); CD44 from HEMA (3, 807, 76.0%); COLQ from NEUR (3, 622, 76.0%); PITX2 from MCD/OPHT (3, 385, 81.0%); FAS from CANC/IMMU (3, 376, 59.3%); MXI1 from CANC (3, 320, 88.8%).

Interestingly, fractions of proteins with predicted α-MoRF regions were highly correlated with the content of predicted disorder in a given dataset (correlation coefficient is ~ 0.89). This suggests that the major function of IDRs in the proteins from analyzed datasets is protein-protein interaction. α-MoRFs, being disordered in the unbound state and gaining α-helical structure upon interaction with binding partners, suit ideally this function. In fact, it has been proposed that that the involvement of IDRs in protein-protein interactions have several advantages (Cortese, V N Uversky, and Keith Dunker 2008), including: (i) Decoupled specificity and strength of binding (high-specificity-low-affinity interactions); (ii) Increased speed of interaction due to greater capture radius and the

ability to spatially search interaction space; (iii) Efficient regulation via rapid degradation; (iv) Increased interaction (surface) area per residue; (v) Strengthened encounter complex (less stringent spatial orientation requirements); (vi) A single disordered region may bind to several structurally diverse partners; (vii) Many (structured) proteins may bind a single disordered region; (viii) Less sterically restricted to allow elongation of binding area; (ix) Efficient regulation via posttranslational modification; (x) Ease of regulation/redirection by alternative splicing; (xi) Overlapping binding sites due to extended linear conformation; (xii) High evolutionary rate; (xiii) Flexibility that allows masking (or not) of interaction sites or allow interaction between bound partners. Many of these features are specific properties of α-MoRFs.

2.3.4.5 Abundance of α-MoRFs in Alternatively Spliced Regions of Proteins From Human Diseasome

Interestingly, our analysis revealed that α-MoRFs are abundantly present in alternatively spliced regions of proteins from some human genetic diseases. This observation is very important as it sheds some light on the potential functional repertoire of alternatively spliced regions. In several diseases, these regions play a crucial role in protein-protein interaction, as they are enriched in molecular recognition features.

CHAPTER 3

PREDICTION OF INTRINSIC DISORDER IN PUTATIVE SEQUENCES

3.1 Motivation and Related Work

At the time of collection of protein sequences for large-scale ID analysis in human genome (Chapter 2), we collected 30,053 human RefSeq protein sequences in NCBI database. 22,528 of these were sequences whose identifiers start with NP. This code denotes that there is substantial experimental confirmation for these sequences. Other 7,525 were sequences whose identifiers start with XP, which is a code for putative model sequences, obtained with NCBIs automated annotation procedure, that lack sufficient experimental confirmation. We compared the distributions of disorder content in these two sets of sequences (Figure 3.1). The difference in predicted disorder content between sets of XP sequences (putative/unconfirmed protein sequences) and NP sequences (experimentally confirmed protein sequences) is greater in both magnitude and statistical significance than difference between any two classes in our final data set for large-scale ID analysis in human genome. The simplest explanation for this is that the automated annotation procedure has a high error rate that introduces a large number of incorrect amino acid sequences. Alternatively, this dramatic difference in the level of predicted ID between the experimentally and automatically identified proteins could be due to the bias of the existing identification techniques toward the ordered proteins. To some extent this resembles a problem the Structural Genomics Initiative Centers are facing, where the use of the traditional target search criteria (mostly based on the sequence identity) and protein purification and isolation methods generated mostly ordered targets, whereas alternatively

identified and purified proteins awaiting structure determination were richer in disorder than an average protein in PDB (C J Oldfield, Ulrich, et al. 2005; Balasubramanian et al. 2000). It has been pointed out that this bottleneck was determined by the strategy chosen where in efforts to identify proteins with novel folds researchers started with proteins having amino acid sequences unlike those of proteins with known 3D structures (C J Oldfield, Ulrich, et al. 2005; Balasubramanian et al. 2000). In a similar manner, traditional experimental approaches developed for protein identification could be biased toward order (as ordered well-folded proteins were at the research focus for many years), whereas predictive tools are mostly dealing with the remaining part of the proteomes and therefore are inevitably identifying more disordered proteins. Two groups of sequences also have significantly different amino acid distributions (Figure 3.2). The enrichment of several amino acids in confirmed sequences (F, I, L, N, V, Y) and in putative sequences (G, P, R, S) is consistent with the order promoting vs. disorder promoting classification of amino acids.

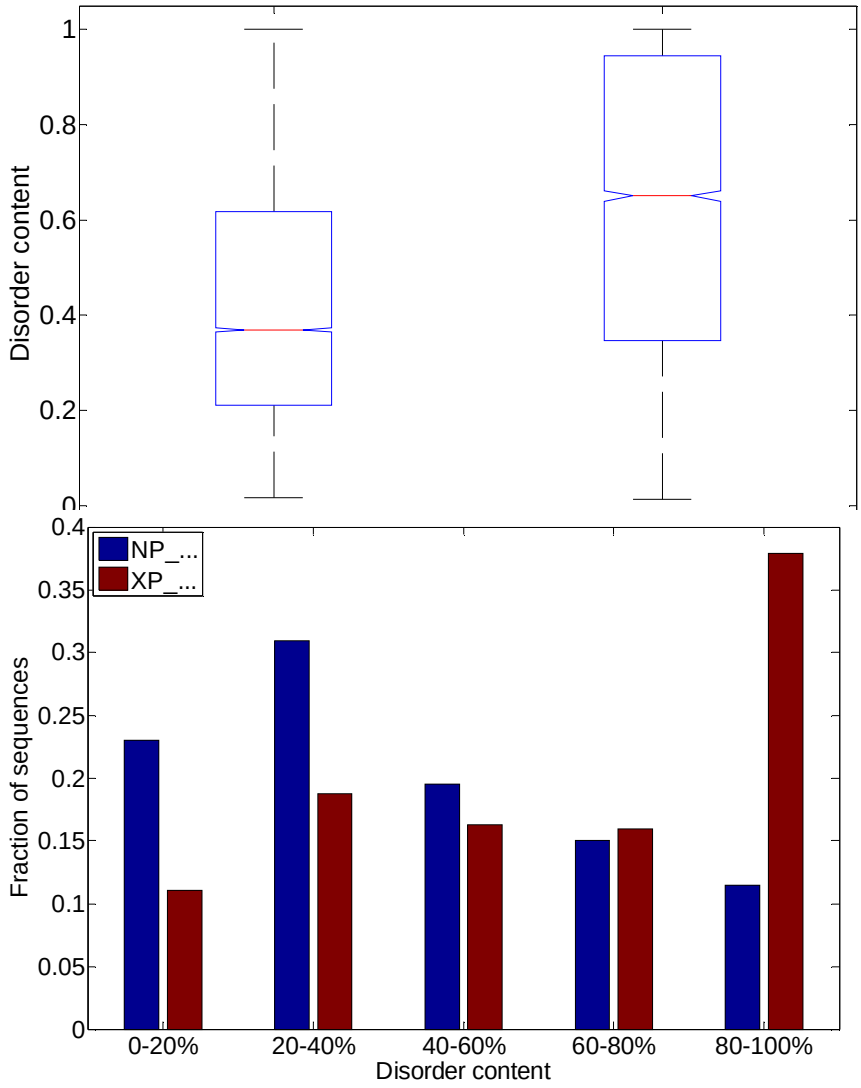

Figure 3.1. Comparison of predicted disorder content distributions in the confirmed human protein sequences (NP_...) and the putative human protein sequences (XP_...) from the dataset described in Chapter 2.

Top: Boxplot comparison. Bottom: Comparison of histograms with respect to the disorder content.

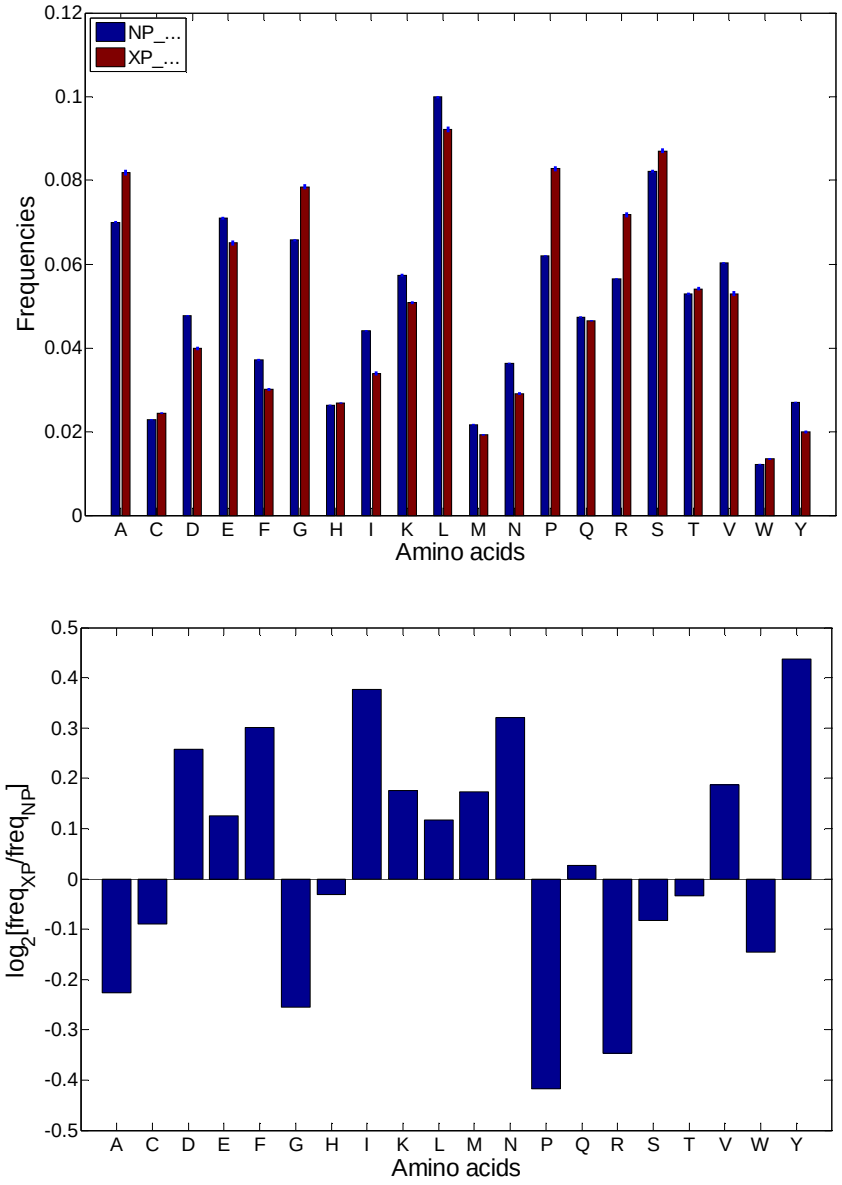

Figure 3.2. Comparison of amino acid compositions in the confirmed human protein sequences (NP_...) and the putative human protein sequences (XP_...) from the dataset described in Chapter 2.

Top: Direct comparison of frequencies (error bars are too small to be visible). Bottom: Log_2-ratios of frequencies; amino acids with positive values are enriched in confirmed sequences, amino acids with positive values are enriched in putative sequences.

The predicted sequences were unevenly distributed between disease-related and disease-unrelated proteins. In fact, the majority of the putative sequences were products of the non-disease genes. Therefore, including such sequences into the data set would introduce significant bias for disorder in the non-disease gene part of the data set. Based on these observations, we decided to exclude such sequences from the final datasets.

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 predictors use Bayesian networks, such as Interpolated Markov Models (Salzberg et al. 1998), Generalized Hidden Markov Models (Burge and Karlin 1997), and Generalized Pair HMMs (Pachter, Alexandersson, and Cawley 2002). These predictors exploit the following findings: 1) many signals involved in gene expression (e.g. promoters, splice junctions) exert specific patterns, known as 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, these prediction models perform well, 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 genomic regions or incorrectly translated coding regions, 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 Chapter 2 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 incorrectly annotated 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 Chapter I 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, I assumed that nonsense regions follow a different amino acid composition than the true protein sequences. Therefore, instead of testing and improving the gene finding algorithms, I investigate whether nonsense regions can be detected from amino acid sequence, similarly to prediction of intrinsic disorder.

I 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, I constructed synthetic nonsense sequences from genomic regions of the true protein sequences that form the other class.

The methodology that was used to create the synthetic nonsense sequences, train and evaluate the nonsense predictor, and analyze results of the predictor for XP sequences is described in Section 3.2. Section 3.3 presents 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. Section 3.4 provides a discussion of the results and conclusion.

The dataset that was used in the initial attempt at performing this analysis was based on the dataset used in Chapter 2, which was retrieved from the NCBI database in 2007, and included only the human genome. Since additional information about genes and proteins was required to answer open questions and improve several shortcomings of the setup for the initial study, I downloaded all the necessary information from the NCBI database again in 2011 and performed analysis with improved methodology and, in addition to the updated human dataset, also three new datasets: mouse, fruitfly and zebrafish. This chapter presents the methodology and the results of the newer, extended study. However, the old methodology and results are also mentioned where appropriate, since the comparison of the results gives an important insight into the trends of the Gene and Protein sections of the NCBI database, that are relevant for the topic of this Chapter.

3.2 Methodology

3.2.1 Dataset and Creation of Synthetic Nonsense Sequences

We created four datasets, one for each of the following species: Homo sapiens (human), Mus musculus (mouse), Drosophila melanogaster (fruitfly), Danio rerio

(zebrafish). For each of the organisms, we downloaded genomic records with sequences and annotation about all genes with RefSeq protein records. These records contain the genes' nucleotide sequences, as well as position of all parts of mRNA sequences: 3' and 5' UTRs (untranslated regions) and coding regions (exons). From this information we could also easily identify intronic regions.

For the control/negative class of true proteins we selected either NP protein sequences that are listed as single isoforms of respective genes (i.e. the genes are not known to be involved in alternative splicing), or representative sequences compiled from multiple NP sequences for genes with multiple isoforms (i.e. alternatively spliced). A representative sequence was compiled by translating all exon regions in a genes sequence; this is similar to the methodology used in Chapter 2 to obtain one representative sequence for alternatively spliced genes. The only exceptions were the alternatively spliced genes for which an exon was translated in different codon alignment in different isoforms; such genes were not used in the study.

Nonsense protein sequences for the positive class were synthesized from coding and noncoding regions of the genomic sequences of genes whose representatives form the negative class. The exact locations of exons in these genomic sequences are known, and the exons can only be translated correctly if they are read in one of the three possible reading frames. For a given annotated genomic sequence and the protein it is translated to (top sequence in Figure 3.3, where exons are shown in black), the procedure to synthesize nonsense sequences was the following:

Crop the gene's nucleotide sequence by removing all nucleotides from noncoding regions that are further than 120 nucleotides away from the closest exon. The obtained

nucleotide sequence can be translated in three different reading frames, for each of these three 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 Figure 3.3). 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 Figure 3.3) are either coming from non-coding regions or from incorrectly translated exons; therefore they can be considered to be nonsense sequences.

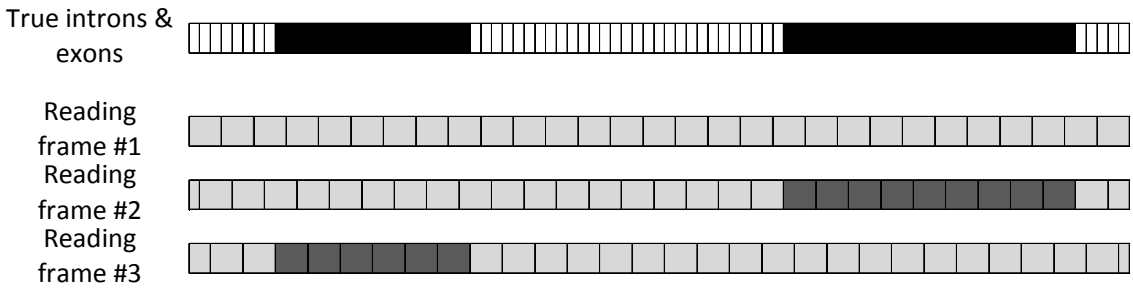

Figure 3.3. Illustration of the procedure to synthesize nonsense protein sequence from genomic sequences with confirmed exon positions.

A genomic sequence with known exon positions (black, dark grey) is read and translated in three different ways, with different starting codon positions. Parts of an obtained amino acid sequence that align perfectly to parts of the confirmed protein are discarded (dark grey), while the remainder is kept as a synthesized nonsense sequence (light grey). The schema is simplified, true sequences and exons are longer than the ones depicted here.

This procedure produces three nonsense sequences for each true protein sequence. As an additional set, we selected representative sequences for genes with XP protein sequences the same way as for genes with NP protein sequences. We discarded short sequences for which construction of input features for prediction is not viable. The overview of the number of sequences in the three groups (NP, XP, synthetic nonsense) for four genomes is given in Table 3.1.

Table 3.1. Overview of numbers of sequences in datasets for nonsense prediction

Organism	NP sequences	XP sequences	Synthetic nonsense sequences
Homo sapiens	14353	307	42923
Mus musculus	14661	799	43808
Drosophila melanogaster	12190	0	36240
Danio rerio	9331	7867	27897

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 (P Romero et al. 2001; Zoran Obradovic et al. 2003; Kang Peng et al. 2005; Kang Peng, Predrag Radivojac, et al. 2006). 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 (Kang Peng, Predrag Radivojac, et al. 2006); these predictions are mapped to binary classification by applying the .5 threshold. To summarize the predicted ID in a protein sequence, we used *disorder content* (DC) as defined in Chapter 2. We labeled amino acids in synthetic nonsense sequences with information about their origin, i.e. whether the central nucleotide of the corresponding codon was a part of coding region or non-coding region. For amino acids in all sequences we calculated the distance of the codon from the nearest border between exon and a non-coding region. Both of these labels were later used in balancing of the training set.

The main difference in the above described datasets and the dataset in the initial study is that in the initial study only the mRNA sequences of the human proteins were used as the source for synthesis of nonsense sequences. The translation of non-coding regions was therefore limited only to upstream and downstream untranslated regions (3'UTR and 5'UTR) if they were included in the mRNA sequence at all. We also excluded all genes that were known to be alternatively spliced. The dataset contained 15,124 NP sequences and 45,038 synthetic nonsense sequences, as well as the additional set of 5,243 XP sequences.

3.2.2 Prediction 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 were overlapping, the obtained features were similar, and therefore the redundancy allowed 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 balanced in terms of disorder to include equal number of residues predicted to be ordered and disordered. We further balanced the nonsense class by sampling equal number of residues obtained by translating non-coding regions and residues obtained by translating coding regions. We also balanced both nonsense and true protein class by sampling equal numbers of residues obtained from regions in vicinity of an exon/non-coding region border (50nt or less) and of residues obtained from regions far from such borders (more

than 50nt). 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.

In the initial study, we balanced the training dataset only with respect to the class and the predicted disorder, but not with respect to the origin of the amino acids.

3.2.3 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.

To analyze the significance of input features for prediction of nonsense, we used approximation of partial derivatives of prediction function. Partial derivative of prediction function *pred* with respect to *i*-th feature f_i at point x was approximated as

$\partial pred_{f_i}(x) \approx \big(pred(x + \epsilon_i) - pred(x)\big)/\epsilon$, where $\epsilon_i = \epsilon(0, \dots, 1, \dots, 0)$ is the vector with value ϵ at *i*-th elementh and value 0 at all other elements. The mean of such estimates for feature f_i over all points in the dataset $\sum_{j=1}^{n} \partial pred_{f_i}(x_j)/n$ was then used to estimate both significance (absolute value) and direction (sign) of contribution of feature f_i to prediction function.

3.3 Experimental Results

The motivation for this study was the discrepancy in predicted disorder content between NP and XP sequences. The same difference is preserved in the dataset for Homo sapiens (Figure 3.4), although there is a change in the distribution of disorder content for XP sequences. There are also differences in distributions of disorder content between NP and XP sequences for Mus musculus and Danio rerio (Figure 3.4), but they are not as large as for Homo sapiens. The distribution curve for disorder content in synthetic nonsense sequences in Danio rerio is strongly skewed to the right.

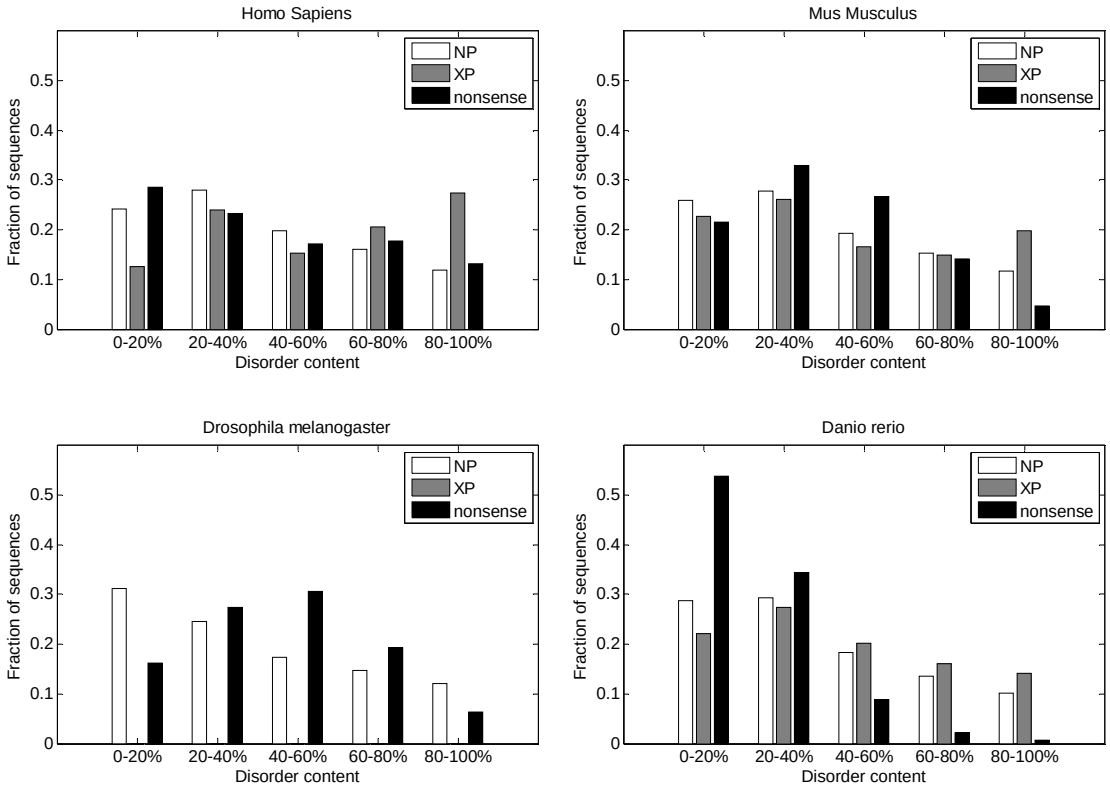

Figure 3.4. Distributions of predicted disorder content (DC) in confirmed proteins (NP), putative proteins (XP), and synthetic nonsense sequences.

Histograms show fractions of sequences with various levels of DC.

In the synthetic nonsense sequences for Homo sapiens, we can observe a large difference in distributions for residues originating from non-coding regions and exons (Figure 3.5, top). Distribution for the residues originating from non-coding regions is strongly biased towards order. Distribution of disorder content for residues originating from coding regions further from exons' borders (i.e. in the middle of exons) is fairly uniform. Contrary to the residues originating from non-coding regions, distribution of disorder content for residues originating from coding regions near the exons' borders is strongly biased towards disorder. This is also preserved for NP and XP sequences (Figure 3.5, bottom), although XP sequences have higher fraction of residues with high levels of disorder prediction. These differences in distributions of disorder content were the reason for the additional balancing of the dataset introduced after the initial study.

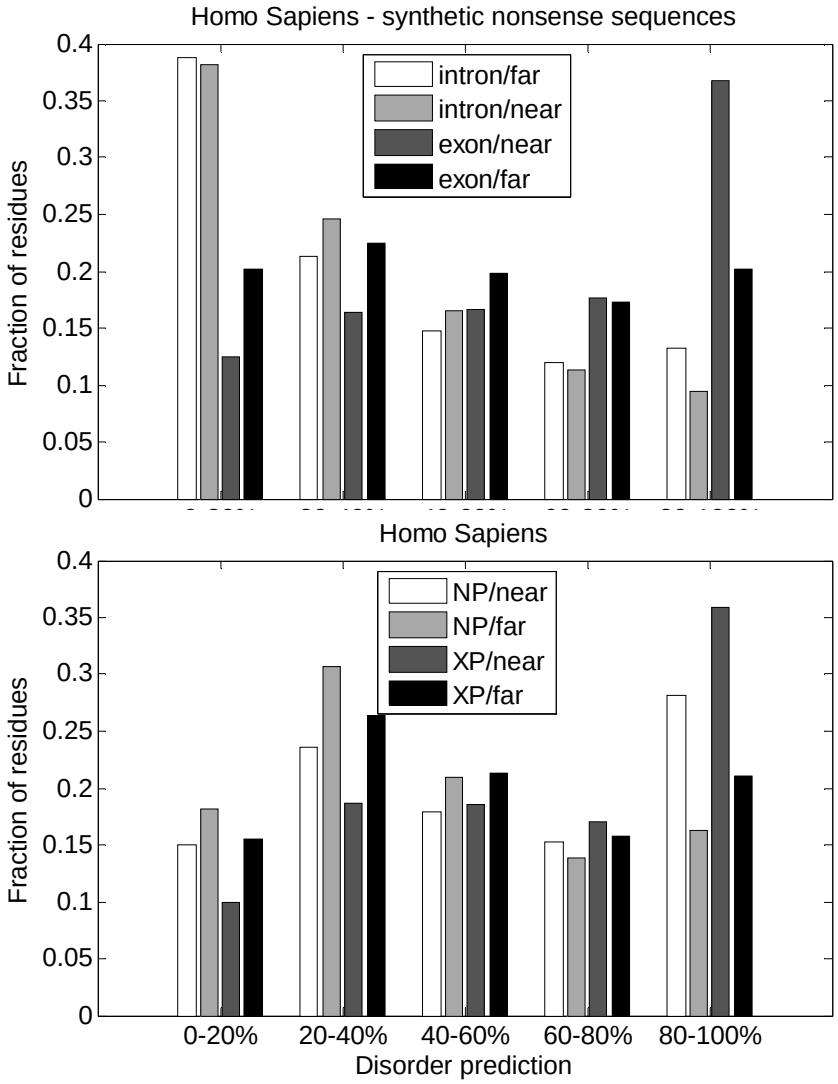

Figure 3.5. Distributions of predicted disorder in human synthetic nonsense, NP and XP sequences – comparison by position of codons in genomic sequences.

Residues were grouped by the positions of their codons in genomic sequences (non-coding or exons, near or far from non-coding/exon border). Histograms show fractions of residues with various levels of disorder prediction.

3.3.1 Evaluation of Nonsense Sequence Predictor

The results of 10-fold cross-validation evaluation of nonsense predictors are summarized in Table 3.2. 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. We also report area under ROC curve (AUC). For per-residue prediction we also perform separate evaluation of predictors for disordered and ordered regions, separate evaluation for nonsense regions originating from exons and nonsense regions originating from non-coding regions (in both cases the negative class remains the same, i.e. contains all NP sequences).

All indicators of predictor's performance on Homo sapiens dataset showed a small improvement compared to the results of the initial study. To test the reason behind that improvement we trained a predictor on the Homo sapiens dataset for which the training dataset was only balanced with respect to true/nonsense and order/disorder criteria, but not with respect to the non-coding/exonic origin and the near/far from non-coding–exon border criteria.

The evaluation of these predictors (results not shown) showed similar improvement compared to results for the initial study. Therefore we can conclude that additional balancing did not directly affect the performance of the predictor. Instead, the improvement in performance can be attributed to one of the following: 1) changes in the NCBI dataset that occurred over last three years (refinement of NP sequences and upgrading of XP sequences to NP status), 2) inclusion of more intronic regions into the synthetic nonsense part of the dataset, 3) inclusion of sequences with alternative splicing.

Table 3.2. 10-fold cross-validation evaluation of per-residue and per-protein nonsense sequence predictors.

	Specificity	Sensitivity.	Accuracy = (spec+sens)/2	Area under curve
Homo sapiens				
Per-residue				
Overall	**83.47%**	**84.42%**	**83.94%**	**0.9189**
Ordered regions	82.35%	84.03%	83.19%	0.9119
Disordered regions	84.82%	85.04%	84.93%	0.9276
Nonsense ~ introns	83.47%	84.15%	83.81%	0.9174
Nonsense ~ exons	83.47%	84.94%	84.20%	0.9217
Per-protein	**94.43%**	**98.81%**	**96.62%**	**0.9933**
Mus musculus				
Per-residue				
Overall	**82.28%**	**83.28%**	**82.78%**	**0.9087**
Ordered regions	81.15%	82.94%	82.05%	0.9016
Disordered regions	83.69%	83.88%	83.79%	0.9182
Nonsense ~ introns	82.28%	81.92%	82.10%	0.9022
Nonsense ~ exons	82.28%	85.91%	84.09%	0.9213
Per-protein	**94.02%**	**98.85%**	**96.43%**	**0.9938**
Drosophila melanogaster				
Per-residue				
Overall	**84.57%**	**87.14%**	**85.86%**	**0.9360**
Ordered regions	82.05%	85.59%	83.82%	0.9187
Disordered regions	87.70%	88.86%	88.28%	0.9538
Nonsense ~ introns	84.57%	81.18%	82.88%	0.9105
Nonsense ~ exons	84.57%	90.04%	87.31%	0.9485
Per-protein	**96.97%**	**97.54%**	**97.26%**	**0.9938**
Danio rerio				
Per-residue				
Overall	**83.29%**	**87.12%**	**85.20%**	**0.9297**
Ordered regions	80.80%	88.41%	84.61%	0.9262
Disordered regions	86.85%	82.38%	84.61%	0.9266
Nonsense ~ introns	83.29%	88.00%	85.64%	0.9338
Nonsense ~ exons	83.29%	85.47%	84.38%	0.9220
Per-protein	**95.98%**	**99.60%**	**97.79%**	**0.9980**

3.3.2 Comparison of predicted nonsense in NP and XP sequences

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 (Breiman 1996).

We calculated nonsense content for all NP, XP and synthetic nonsense sequences. The distributions of nonsense content in the three groups of sequences (NP, XP, synthetic nonsense) for four datasets are compared in Figure 3.6. Difference between NP and synthetic nonsense sequences is expected in accordance with predictor evaluation results. However, the significant increase in nonsense content for human 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 human XP sequence regions are more similar to synthetic noise sequences than to NP sequences. There is also a (much smaller) difference in nonsense content between mouse NP and XP seuqences. However, distributions of nonsense content for NP and XP sequences in Danio rerio are almost the same.

Comparison between nonsense content prediction for NP and XP sequences and the effects of the choice of threshold is further elaborated in Table 3.3, which lists fractions of sequences that are predicted to be ”mostly nonsense” (i.e. nonsense content is greater than some threshold). Here we can again observe the drop of the fraction for XP sequences in Mus musculus and especially in Danio rerio.

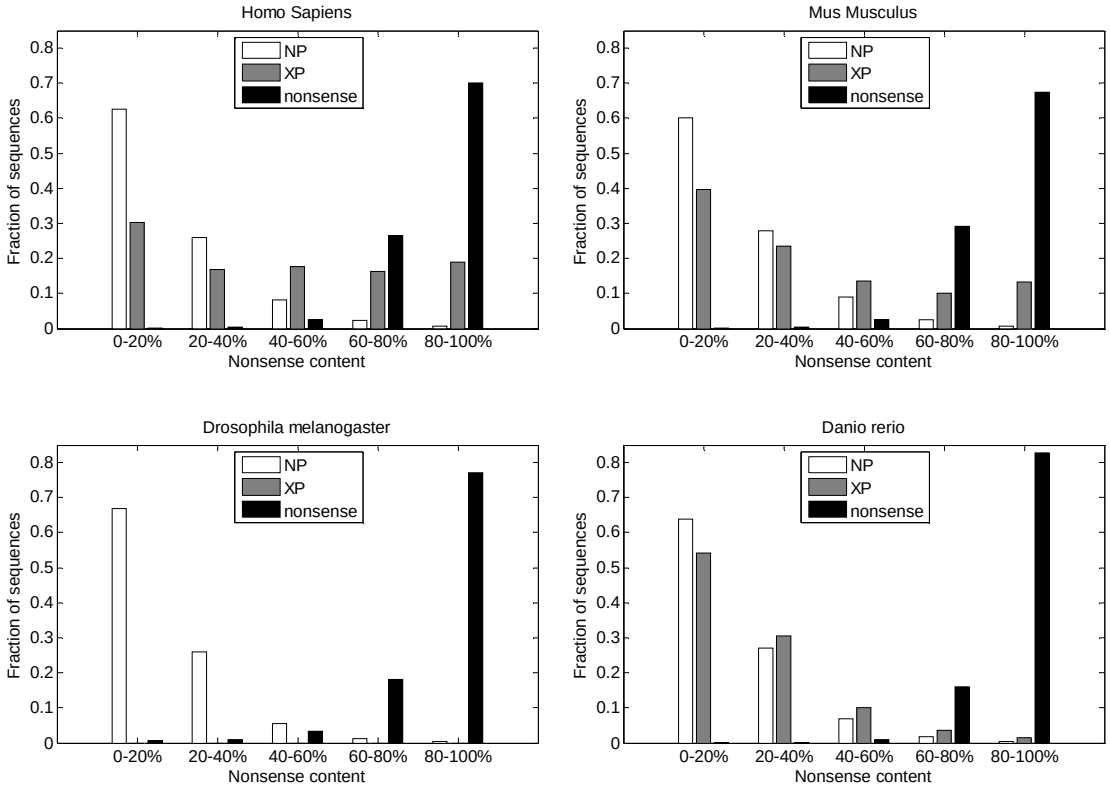

Figure 3.6. Distributions of predicted nonsense content (NC) in confirmed proteins (NP), putative proteins (XP), and synthetic nonsense sequences.

Histograms show fractions of sequences with various levels of NC.

Table 3.3. Comparison of fractions of NP, XP and synthetic nonsense sequences with nonsense content greater than threshold = .5 (, .4, .6).

Organism	Threshold	NP	XP	Synth. nons.
Homo sapiens	0.4	11.34%	52.77%	99.38%
Homo sapiens	**0.5**	**5.98%**	**44.30%**	**98.64%**
Homo sapiens	0.6	3.07%	35.18%	96.71%
Mus musculus	0.4	12.22%	36.92%	99.37%
Mus musculus	**0.5**	**6.55%**	**29.41%**	**98.70%**
Mus musculus	0.6	3.20%	23.28%	96.77%
Drosophila melanogaster	0.4	7.37%		98.48%
Drosophila melanogaster	**0.5**	**3.71%**		**97.40%**
Drosophila melanogaster	0.6	1.72%		95.12%
Danio rerio	0.4	8.95%	15.32%	99.78%
Danio rerio	**0.5**	**4.39%**	**9.01%**	**99.53%**
Danio rerio	0.6	2.10%	5.17%	98.76%

We also compared the total fractions of residues predicted to be in nonsense regions (Table 3.4). While the margin between total nonsense content between NP and synthetic nonsense sequences, which equals (*sensitivity*+*specificity*)–1, is increased for Homo sapiens compared to the dataset from the initial study, the margin between total nonsense content in XP and NP sequences is decreased (from 20.05% to 18.09%). The same margin is further decreased for Mus musculus, and almost non-existent for Danio rerio.

Table 3.4. Total (per-residue) predicted nonsense content in NP, XP and nons sequences, and the margin of nonsense content between NP and XP, and between NP and synthetic nonsense sequnces.

Organism	NC_NP	NC_XP	NC_XP – NC_NP	NC_nons	NC_nons – NC_NP
Homo sapiens	16.57%	34.65%	18.09%	84.42%	67.86%
Mus musculus	17.75%	27.21%	9.46%	83.29%	65.54%
Drosophila melanogaster	15.41%			87.15%	71.74%
Danio rerio	16.69%	18.81%	2.12%	87.13%	70.44%

3.3.3 Relationship between prediction of nonsense in XP sequences and prediction of intrinsic disorder

After the computational experiments have indicated that human (and to some extent mouse) 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 human XP sequences, 55.53% of all residues are predicted to be disordered. In regions of human XP sequences that are predicted to be nonsense the fraction of predicted ID

residues is increased to 64.87%, while in regions predicted not to be nonsense, the fraction of predicted ID residues is only 50.58%. It is interesting to note here that in the mouse dataset, predicted fraction of ID residues is very similar in predicted nonsense regions of XP sequences (48.79%), regions of XP sequences that are predicted not to be nonsense (49.09%) and overall XP sequences (49.01%). Furthermore, in the zebrafish dataset, the difference is inverted compared to the human dataset: 46.69% overall, 38.13% in predicted nonsense regions, and 48.67% in remaining regions.

A new question arises whether the positive difference between prediction of nonsense and prediction of ID for human 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 indicators of correlation between prediction of nonsense and prediction of disorder for NP, XP and synthetic nonsense sequences are listed in Table 3.5. There is a significant positive correlation between predicted nonsense and predicted disorder in XP sequences in Homo sapiens. Surprisingly all correlation coefficients (including for XP sequences) in Danio rerio, as well as all correlation coefficients for NP sequences, are negative. However, the corresponding R^2 values are low.

Table 3.5. Correlation of disorder content (DC) and nonsense content (NC) for NP, XP and synthetic nonsense sequences

	NP			**XP**			***Synt. nonsense***		
Organism	Corr. coeff.	R^2	*p*	Corr. coeff.	R^2	*P*	Corr. coeff.	R^2	*p*
Homo sapiens	-0.085	0.007	~0	0.354	0.125	~0	0.252	0.063	~0
Mus musculus	-0.123	0.015	~0	0.019	0.000	0.59	0.227	0.051	~0
Drosophila melanogaster	-0.120	0.014	~0	0.000	0.000	~0	0.098	0.010	~0
Danio rerio	-0.173	0.030	~0	-0.157	0.025	~0	-0.203	0.041	~0

Table contains Pearson correlation coefficients, and the R^2 statistics and *p*-values for linear regression of disorder content and nonsense content.

It is interesting to note here that the correlation indicators for human XP sequences were much stronger in the initial study (ρ=.442, R^2=.196, and p~5E–250). In the initial study we also produced scatterplots of predicted nonsense content against predicted disorder content (Figure 3.7). While the points representing NP sequences were clustered at the bottom (low level of prediction for nonsense) and the points representing synthetic nonsense sequences were clustered at the top (high level of prediction for nonsense), the points representing XP sequences form two clusters – in the upper-right corner (high disorder prediction, high nonsense prediction) and the lower right corner (low disorder prediction, low nonsense prediction). The decreased strength of the correlation may be attributed to the improved curation of the XP part of the human RefSeq sequences.

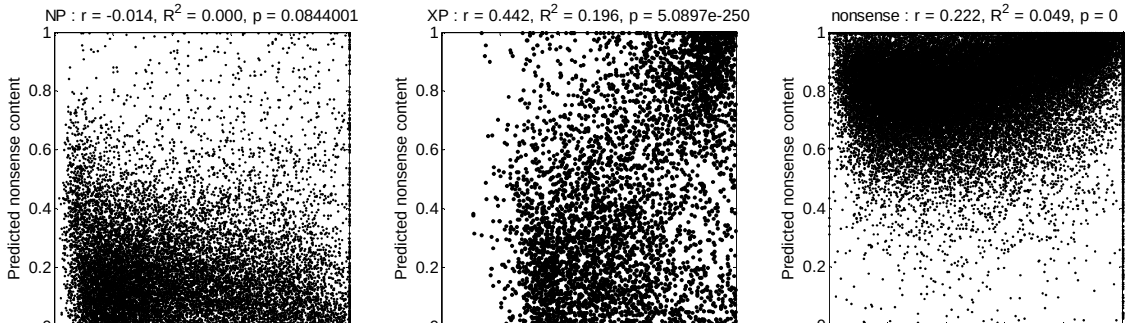

Figure 3.7. Scatter-plots for predicted disorder content (x axis) vs. predicted nonsense content (y axis) for NP, XP, and synthetic nonsense sequences (from the initial study).

Pearson-correlation coefficients *r*, and R^2 statistics and p-values for linear regression models are shown above the plots.

3.3.4 Analysis of input features for nonsense prediction

The approximate means of partial derivatives of prediction function with respect to 23 input features over all points in the balanced training human dataset are shown in Figure 3.8. Figure 3.8 also shows the frequencies of the amino acids corresponding to the first 20 input features, which are sorted according to their order (left) or disorder (right) promoting tendency. There is no obvious link between the sign and/or direction of the mean partial derivatives on one side and the amino acid frequencies and/or their order/disorder promoting property. Both positive and negative values are present among both the disorder promoting and order promoting amino acids. There are several pairs of amino acids with very similar frequencies and very different values of mean partial derivatives.

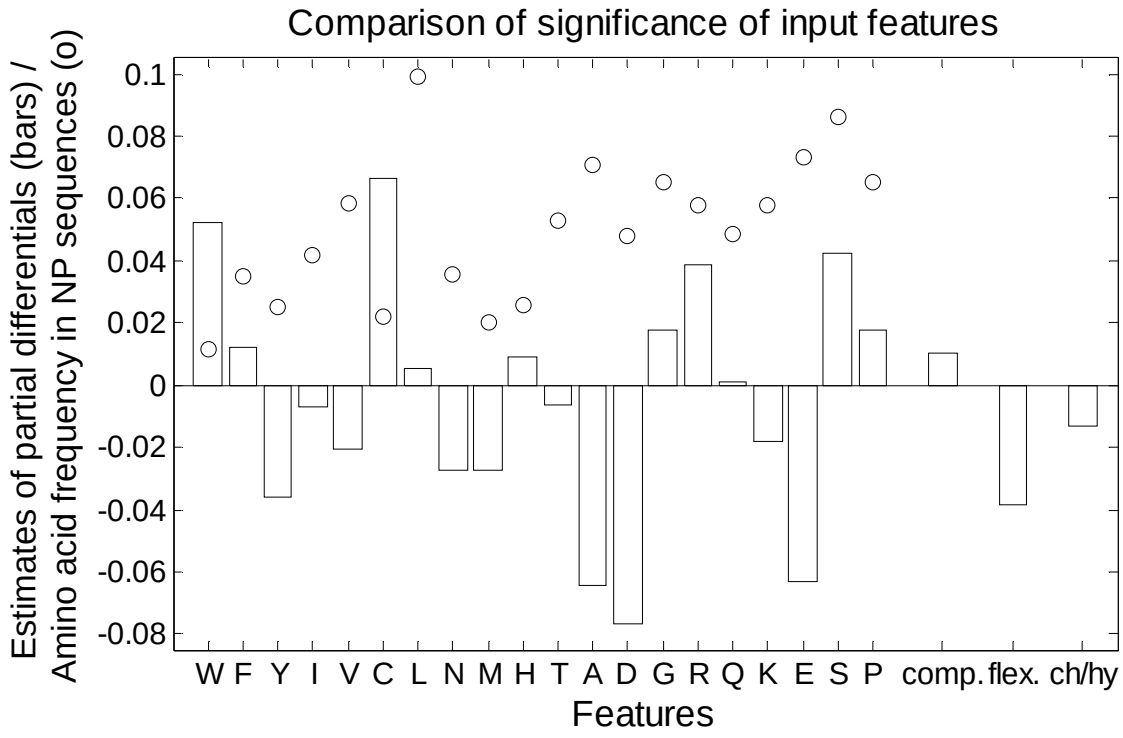

Figure 3.8. Comparison of significance of input features.

Bars represent approximated 23 means of partial derivatives (with respect to 23 input features) over all points in the dataset. Circles represent the frequencies of the amino acids that the first 20 features are based on. The amino acids are ordered from most order promoting (left) to most disorder promoting (right).

3.4 Discussion

In the study described in Chapter 2 (Midic et al. 2009) 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 (Figure 3.1). This difference was consistent with the divergence in amino acid composition for NP and XP sequences (Figure 3.2), 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 (Figure 3.2), 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 genomic sequences (Figure 3.3). These sequences have different amino acid composition than their real counterparts (Figure 3.4), although in human and mouse genome 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.

Using a simple prediction model, we have successfully trained predictors that discriminate 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. The predictors have very good per-residue accuracies (82%–86%) and AUCs (> .9), comparable to predictors of intrinsic disorder (Table 3.2). More importantly, they are very well balanced (i.e. has similar sensitivity and specificity) and perform equally well on predicted disordered regions and predicted ordered regions, as well as on synthetic nonsense sequence regions originating from coding and non-coding genomic regions. These results confirm the assumption that nonsense regions can be predicted from sequence alone.

We have also used a simple method to aggregate per-residue predictions and obtain per-protein predictions. The performance of per-protein predictors is very close to optimal, with accuracies 96%–98% and AUC ~ .99. However, it is only feasible to use per-protein predictors 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 ~44% of human 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. Similar large discrepancy was observed for Mus musculus (~30% vs 7%), but not for Danio rerio (~9% vs 4%).

Per-residue predictor also gave very different predictions for human NP and XP sequences. The differences in distributions of nonsense content (fraction of residues in a

sequence predicted to be in nonsense regions) are substantial for Homo sapiens and Mus musculus, but not for Danio rerio (Figure 3.6, Table 3.4).

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 human 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% in mRNAnons, ~18% in GNMCnons dataset) at that threshold.

Predicted nonsense regions in human XP sequences have higher total disorder content (64.9%) than the remaining regions of human XP sequences (50.6%). 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 a similar positive linear dependency (albeit with lower correlation coefficient) is observed in synthetic nonsense sequences, it is completely absent from NP sequences. However, similar significant correlation is absent in Mus musculus, while in Danio rerio the correlation is significant and negative. In Danio rerio, predicted nonsense regions in XP sequences have lower total disorder content (38.1%) than the remaining regions of human XP sequences (48.67%).

These experimental results support the hypothesis that the presence of nonsense regions in human XP sequences – introduced by errors of gene finding procedures – significantly increases the predicted disorder content, and therefore introduces bias to genome-wide estimate of disorder content.

However, the same conclusion cannot be reached for Mus musculus and Danio rerio. Danio rerio has very similar distributions for predicted disorder content in NP and XP sequences, as well as very similar distributions for predicted nonsense content in NP and XP sequences. Furthermore, it has the lowest levels of predicted nonsense in XP sequences of all three compared organisms. Most importantly, the contribution of nonsense regions in XP sequences to predicted disorder content is at most minimal.

We were only able to partially explain the discrepancy in disorder content estimates for human 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. Differences in datasets and results for Homo sapiens, between the initial study and the new study presented here, clearly show the trend that more and more XP sequences are being curated and eventually have they status upgraded, which leads to decrease in discrepancy between predicted disorder contents, as well as to lower predicted nonsense content.

CHAPTER 4

SEQUENCE ALIGNMENT OF INTRINSICALLY DISORDERED PROTEINS

4.1 Introduction

The first substitution matrices, Point Accepted Mutation (PAM) family of matrices (Dayhoff and Schwartz 1978), were obtained from a set of manually curated alignments of 71 families of proteins, containing 1572 observed mutations. PAM1 matrix is normalized so that it roughly corresponds to sequences that allow 1 mutation for every 100 residues. Matrices that correspond to higher mutation rates are obtained by implying a Markov chain model of protein mutation. For virtually any mutation rate, the PAM matrix can be extrapolated by applying an appropriate exponent to the underlying normalized [p_{ij}] matrix of PAM matrices for lower mutation rates.

BLOSUM matrices (S Henikoff and J G Henikoff 1992) were developed from alignments in the BLOCKS database (J G Henikoff and S Henikoff 1996). Unlike the PAM matrices, which utilize extrapolation, the BLOSUM matrices are based on observed alignments with required mutation rates. The number in the matrix name corresponds to the threshold imposed on the similarity of clusters which are used to compute the matrices. Higher numbered matrices should be used for alignments of closely related sequences, while lower numbered matrices should be used for more divergent sequences.

For both matrices, the score $score(a_i,a_j)$ for matching amino acids a_i and a_j is calculated as $C \cdot \log_2(p_{ij} / q_i q_j)$, where p_{ij} is the observed frequency of a_i and a_j being aligned in the "ground-truth" alignments, while q_i and q_j are the observed frequencies of a_i and a_j, and the constant C is selected so that the error introduced by rounding all scores to the nearest

integer is minimized. The score is positive if amino acids a_i and a_j are observed aligned as a pair more frequently than would be expected based on their individual frequencies, and negative if they are observed aligned less frequently than would be expected.

The difference in amino acid compositions between IDPs/IDRs and structured proteins casts doubt on suitability of BLOSUM and PAM matrices for alignment of IDP sequences (since frequencies q_i are different). Rates of sequence evolution in disordered versus ordered proteins were examined in (Celeste J Brown et al. 2002), where it was found that for 19 out of 26 families of proteins with confirmed intrinsic disorder, the disordered regions evolved significantly more rapidly than the ordered regions, while for only 2 families the opposite was true. A different rate of evolution in disordered proteins means that the frequencies p_{ij} are also inappropriate, and a different substitution matrix is needed for alignment of IDP sequences.

The question of differences in amino acid frequencies q_i was addressed for the general case in (Schäffer et al. 2001), which applies matrix operations on the $[p_{ij}]$ which adapt the resulting substitution matrix to the change in amino acid distributions.

The first attempt to address the question of relationship between ID and sequence alignment was made in (Predrag Radivojac et al. 2002). An iterative approach was used to obtain a set of alignments of families of proteins with confirmed IDRs and the corresponding substitution matrix. The iterative procedure starts with the BLOSUM62 matrix, aligns all families of proteins and calculates the substitution matrix from obtained alignments. The two steps of alignment and calculation of the substitution matrix are then repeated until no significant changes are observed. The obtained matrix DISORDER is significantly different than the initial BLOSUM62 matrix. However, no clear-cut criterion

was established for when this matrix should be used instead of the BLOSUM62 matrix. Furthermore, this matrix always assigns the same score to a pair of amino acids, regardless of whether they belong to IDRs or ordered regions of proteins.

In Section 4.2 I describe a new approach that is an extension of the iterative approach used in (Predrag Radivojac et al. 2002), yet attempts to address its shortcomings. The main idea for improvement was to use an extended 40-symbol alphabet (with 20 symbols for amino acids in ordered regions and 20 symbols for amino acids in disordered regions) and an expanded 40x40 substitution matrix. The 40x40 matrix obtained through the iterative estimation procedure has several important characteristics related to intrinsic disorder. However, tests have shown that this matrix is inferior to BLOSUM62 and other standard 20x20 matrices, as it suffers greatly from the problem of false positives. Furthermore, there is no simple way to incorporate a parameter, similar to the ones used in BLOSUM and PAM families of matrices, into the sequence selection and preprocessing procedure. This makes the comparison with BLOSUM matrices unfair.

In Section 4.3 I describe a different approach to allow for a fair comparison with BLOSUM matrices. We downloaded the BLOCKS5.0 database (J G Henikoff and S Henikoff 1996) and followed the procedure used for creation of BLOSUM matrices (S Henikoff and J G Henikoff 1992), except for the correction of the normalization procedure (Styczynski et al. 2008). Similarly to the iterative procedure described in Section 4.2, we represented all sequences whose parts form BLOCKS with the 40-symbol alphabet. This allowed for the creation of expanded matrices using BLOSUM methodology. In addition to 40x40 matrices, we also created 40x20 matrices and 20x20 matrices. 40x20 matrices can be used in a scenario where intrinsic disorder is predicted for a query sequence, but it

is too expensive to perform prediction for the whole target database of sequences. The 20x20 matrices, similar to the original BLOSUM matrices, but created with the revised algorithm, were created for comparison purposes. We did not compare expanded matrices with widely used original BLOSUM matrices, because it was shown that only by pure chance some of them (e.g. BLOSUM62) have significantly better performance than their counterparts obtained with revised algorithm.

4.2 Iterative estimation procedure

4.2.1 Methodology

4.2.1.1 Dataset

To overcome the limitation on the size of dataset from (Predrag Radivojac et al. 2002), where only proteins with confirmed IDRs were used, prediction of ID was used to label the IDRs in protein sequences, which in turn allowed the selection of arbitrary families of protein sequences for the dataset. The selection procedure began by randomly selecting 1000 protein sequences from the UNIREF database as “anchors” for families. BLAST queries were performed for these sequences against the UNIREF database to obtain families of similar sequences. From the BLAST results only those sequences were kept that satisfied the following criteria: 1) the difference in sequence length compared to the anchor sequence was less than 10%, and 2) the global sequence identity with the anchor sequence was at least 90% (note that significance of BLAST results is estimated based on local identity and/or similarity). The families with less than 10 sequences were discarded. To limit the computational requirements a threshold of 900 was imposed on the length of sequences and reduced the large families to only 50 sequences by random sampling. The

resulting dataset contains 600 families with between 10 and 50 sequences (436 families, or 72%, contain 50 sequences). The average length of sequences in 600 families ranges between 27 and 811, while the median is 312.

VSL2B predictor (K Peng, P Radivojac, et al. 2006) was used to predict IDR in all protein sequences, since this was the most accurate disorder predictor at two consecutive community-wide protein structure prediction assessment experiments (CASP 6-7). VSL2B predicted that 18% of residues in the constructed dataset belong to IDRs.

4.2.1.2 An Iterative Procedure for Estimation of a 40x40 Substitution Matrix

Modifications of Needleman-Wunsch and Smith-Waterman algorithms (global and local pairwise sequence alignment) for use with extended alphabet and an expanded 40x40 substitution matrix are fairly straightforward. I implemented a multiple-sequence alignment algorithm based on ClustalW, as described in (Chenna et al. 2003), with necessary modifications. To save computation time, the all-to-all pairwise sequence identities were precomputed using the Smith-Waterman algorithm and BLOSUM62 matrix (ClustalW uses a heuristic to estimate pairwise identities) the same guiding tree and weights were used for multiple-sequence alignment in all iterations.

The following iterative procedure was used for sequence alignment and estimation of the 40x40 substitution matrix:

1. Initialize the 40x40 matrix (as explained below).
2. Obtain multiple-sequence alignment for each family using the current matrix.
3. Calculate a new matrix from the alignments obtained in step 2.
4. Go back to step 2, unless the changes between iterations are negligible.

The first step of the iterative procedure initializes the matrix to a 40x40 matrix made up of four copies of the BLOSUM62 substitution matrix (Figure 4.1). This means that in the first iteration of alignment, the disorder prediction information is ignored.

After the alignments are obtained in step 2, the new substitution matrix is calculated using the following procedure:

1. Initialize an array for matrix M to zeros.
2. For each family of sequences:

   For each pair of sequences seq_i, seq_j, with weights w_i, w_j, for which $i < j$,

   For each pair of matched amino-acids from seq_i and seq_j, (excluding "matches" to gaps):

   increase the cell in the array corresponding to the two matched amino-acid by $w_1 w_2$.
3. Calculate matrix of amino acid pair frequencies $P=|p_{ij}|$ as $P=(M+M')/2\sum_{i,j} m_{ij}$.
4. Calculate frequencies for amino acids $q_i=\sum_j p_{ij}$
5. Calculate all scores using the formula: $score(a_i,a_j)=2\log_2(p_{ij}/q_i q_j)$

The value of constant $C = 2$ is the same as in the calculation of the BLOSUM62 matrix, so the same gap penalty values could be used.

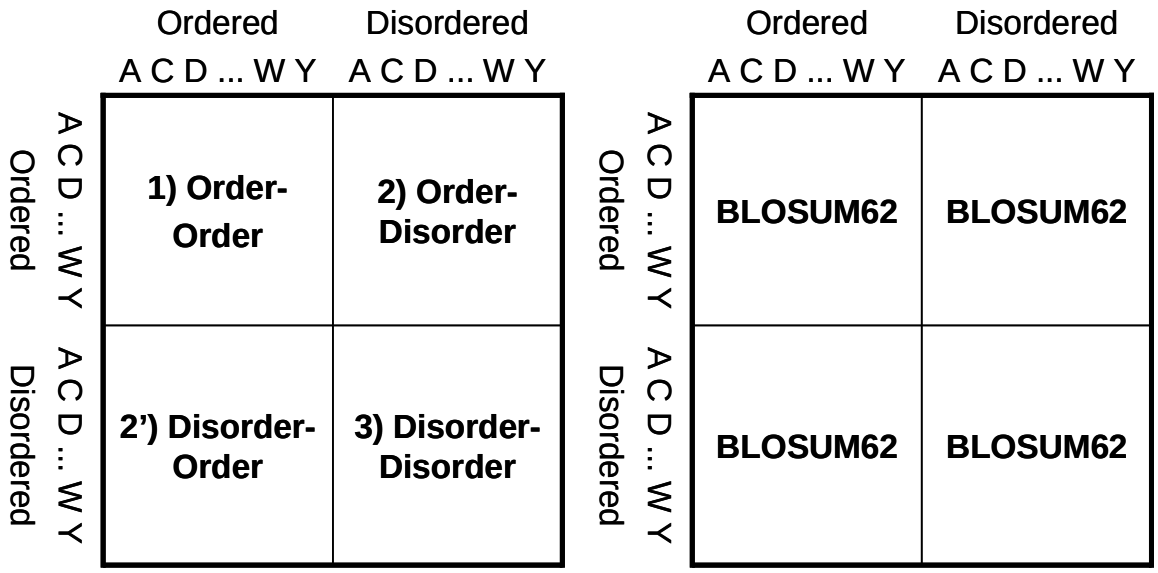

Figure 4.1. A 40x40 substitution matrix and its submatrices.

A 40x40 substitution matrix (left) consists of four 20x20 sub-matrices. The initial matrix for the iterative procedure (right) is made up of four copies of the BLOSUM62 matrix.

4.2.1.3 Experiments

All experiments with the described iterative procedure were performed with the default values for gap penalties in BLAST algorithm: 11 for gap opening and 1 for gap extension. In the main experiment I used the whole dataset to obtain a 40x40 substitution matrix.

To test the stability of the iterative procedure with respect to the choice of the dataset, it was also ran six times with six different subsets of the dataset, each time randomly selecting only half of the sequence families. If the procedure was stable, it was expected that six obtained matrices would be similar.

As a control experiment, the dataset was modified by assigning randomly generated numbers instead of disorder predictions (I draw random numbers from a similar distribution as the values of disorder predictions). By comparing the matrices obtained in the main and control experiment, it was possible to identify which properties of the matrix obtained in the main experiment were specific to ID and were not obtained by pure chance.

4.2.1.4 Evaluation

The proposed alignment approach with expanded substitution matrix was compared to standard local alignment with BLOSUM62 matrix, using an alignment evaluation protocol that is similar as in (Schäffer et al. 2001).

The protocol includes a set of 103 query sequences, a set of 6341 target sequences, and a manually curated set of correct <query, target> pairs. For each query sequence, the sequence is aligned with a pairwise alignment algorithm with all 6341 target sequences. Alignment scores are mapped into E-scores, and the target sequences are sorted according to obtained E-scores in the ascending order. Information-retrieval evaluation techniques

are then used to evaluate the alignments. Target sequences that were labeled as correct hits for that query (by domain experts) are then located in the sorted list, and their ranks in the list are recorded. These ranks are then compared across alignment algorithms. In the ideal case, all true hit targets should be in the top positions in the list.

Another way to compare results of two alignment algorithms is by using ROC curves. ROC curves are constructed by sorting all E-scores in ascending order and adding aligned pairs to the set of hits, which is equivalent to increasing the value of the threshold imposed on E-scores. At each iteration, we calculate the Coverage as *number of correct hits / number of all true hits* and Errors Per Query (EPQ) as *number of errors / number of query sequences*. ROC curve is then obtained by plotting EPQ on the x-axis against Coverage on the y-axis. While area under curve can be calculated similarly as for the 2-class classification problems, a potential difficulty is that for this problem the early retrieval (corresponding to the leftmost part of the ROC curve) is the most important, and AUC does not necessarily reflect that.

4.2.2 Experimental results

The convergence criterion for the iterative procedure used to estimate the 40x40 substitution matrix is that the absolute values of updates for all parameters in the matrix fall below 0.5. This relaxed criterion is due to the fact that in applications the values in the matrix are usually rounded to the nearest integers to allow usage of integer arithmetic. The substitution matrix estimation procedure converged in five iterations as illustrated in Figure 4.2.

The 40x40 matrix obtained in the main experiment with the whole dataset is displayed in Table 4.1. The values in the obtained matrix were compared with the values in the initial BLOSUM62 matrix in Figure 4.3.

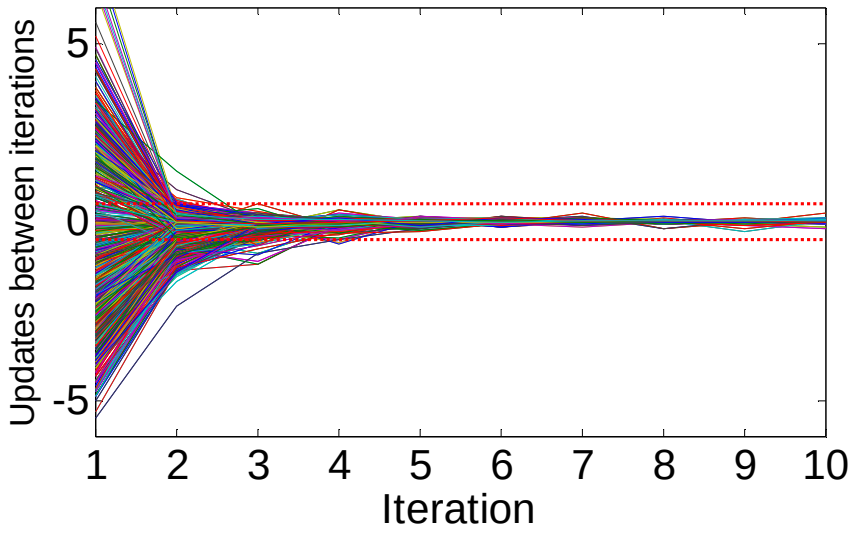

Figure 4.2. Convergence of the iterative substitution matrix estimation procedure. The updates between iterations for all 400 matrix elements are shown for the first 10 iterations. Horizontal lines are at $y = \pm 0.5$.

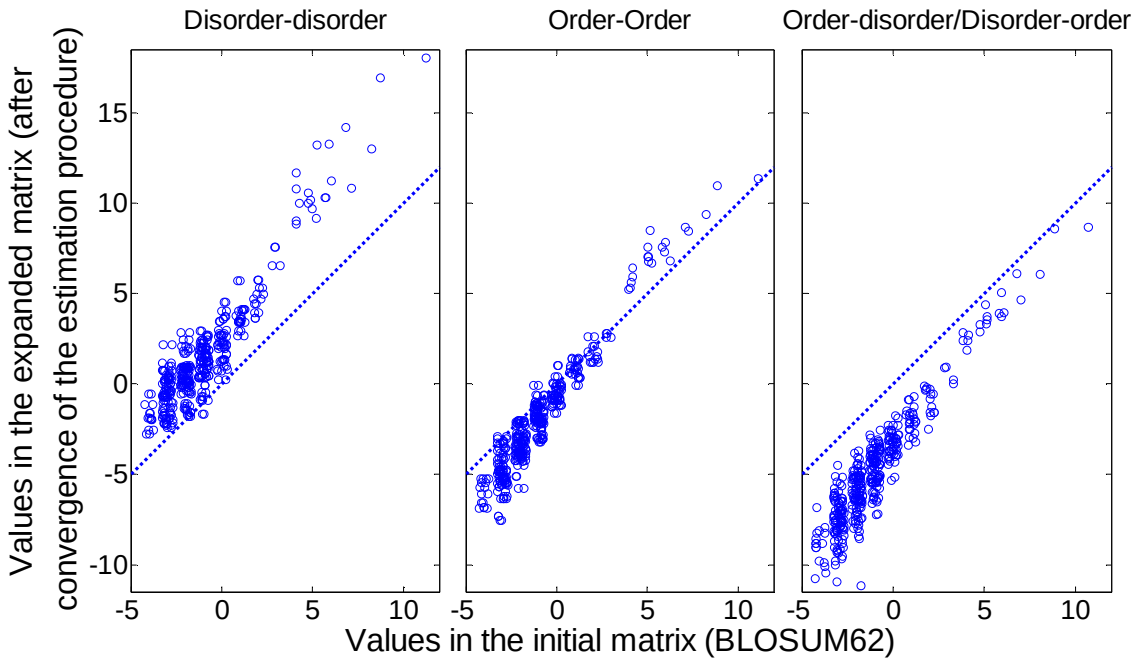

Figure 4.3. Comparison of the obtained 40x40 substitution matrix and the initial matrix.

The scatterplots compare the values in the three 20x20 submatrices of the 40x40 substitution matrix (obtained with iterative procedure) and the corresponding values in the initial matrix (BLOSUM62).

The stability of the iterative procedure was checked by examining the distribution of standard deviations of six values obtained for each matrix element in the experiments repeated with six random subsets of the dataset (each subset contains 300 randomly selected sequence families, i.e. half of the dataset). Substitution matrices obtained in these six experiments were fairly similar, with standard deviation for 85% of matrix elements being smaller than 0.5 (histogram omitted for lack of space). The greatest instability is observed for scores related to the least frequent amino acid types in disordered regions. This is expected, since $\log_2(p_{ij}/q_iq_j)$ is least stable for small values of p_{ij}, q_i and q_j.

For the 40x40 matrix obtained in the control experiment with the randomized dataset, its four 20x20 submatrices were compared among themselves. These four submatrices were practically identical, with 0.305 being the highest standard deviation for four related elements in these submatrices, and with standard deviation smaller than 0.1 for 85.5% of 400 submatrix positions. The four submatrices of the matrix obtained in the control experiment were also compared with the order-order submatrix of the substitution matrix obtained in the main experiment. The differences follow a distribution similar to a normal distribution with $\mu = .24$ and $\sigma = .16$, meaning that although the submatrices are very close, the scores are slightly higher in the order-order submatrix of the expanded substitution matrix from the main experiment.

Table 4.1. The 40x40 substitution matrix obtained with the iterative procedure.

	Order																				Disorder																			
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	11	0	-1	-5	0	-3	-3	-5	-5	-4	-3	-4	-5	-2	-2	-2	-1	-2	-2	-3	9	-3	-4	-7	-3	-6	-6	-7	-7	-6	-5	-5	-7	-7	-6	-5	-4	-5	-5	-5
S	0	6	1	-1	1	-1	0	-1	-1	-1	-2	-1	-1	-3	-4	-4	-3	-4	-3	-5	-6	2	-2	-4	-3	-4	-3	-3	-4	-4	-5	-4	-4	-7	-8	-8	-7	-8	-8	-9
T	-1	1	7	-2	0	-3	-1	-2	-2	-1	-2	-2	-1	-1	-2	-3	-1	-4	-4	-4	-6	-1	3	-5	-3	-6	-3	-4	-4	-4	-4	-4	-4	-4	-5	-5	-6	-4	-7	-7	-9
P	-5	-1	-2	8	-1	-4	-3	-2	-2	-2	-3	-3	-2	-5	-5	-5	-4	-6	-5	-5	-9	-3	-5	5	-4	-6	-6	-4	-4	-5	-5	-6	-5	-8	-8	-8	-7	-9	-9	-10
A	0	1	0	-1	5	-1	-2	-2	-1	-1	-3	-2	-2	-2	-2	-3	-1	-3	-4	-4	-5	-2	-3	-3	2	-3	-4	-4	-3	-4	-5	-4	-4	-5	-6	-5	-4	-7	-7	-8
G	-3	-1	-3	-4	-1	7	-1	-2	-3	-3	-4	-3	-3	-5	-7	-6	-5	-6	-6	-6	-9	-4	-7	-7	-4	4	-4	-4	-5	-6	-6	-5	-6	-9	-11	-10	-9	-10	-9	-11
N	-3	0	-1	-3	-2	-1	8	1	-1	0	1	-1	0	-4	-6	-5	-5	-5	-3	-5	-8	-2	-3	-6	-4	-3	4	-1	-3	-3	-2	-4	-3	-7	-8	-8	-8	-8	-6	-9
O D	-5	-1	-2	-2	-2	-2	1	7	2	0	-1	-2	-1	-6	-8	-7	-6	-7	-5	-7	-11	-3	-4	-5	-4	-4	-1	4	0	-3	-4	-5	-4	-9	-10	-9	-8	-9	-8	-10
r E	-5	-1	-2	-2	-1	-3	-1	2	7	2	-1	-1	1	-4	-6	-5	-4	-6	-4	-5	-10	-3	-4	-4	-3	-5	-3	0	3	-2	-4	-3	-2	-7	-8	-7	-6	-9	-8	-10
d Q	-4	-1	-1	-2	-1	-3	0	0	2	8	1	1	1	-2	-4	-3	-4	-5	-3	-4	-8	-3	-4	-4	-3	-5	-3	-2	-1	4	-2	-2	-1	-5	-7	-5	-6	-7	-6	-8
e H	-3	-2	-2	-3	-3	-4	1	-1	-1	1	9	0	-1	-3	-5	-4	-4	-2	1	-2	-7	-4	-5	-6	-5	-6	-3	-3	-4	-2	6	-3	-4	-7	-8	-6	-8	-5	-3	-6
r R	-4	-1	-2	-3	-2	-3	-1	-2	-1	1	0	7	3	-3	-5	-4	-4	-5	-3	-3	-8	-4	-4	-5	-4	-5	-4	-4	-3	-2	-3	4	0	-6	-7	-6	-6	-8	-6	-7
K	-5	-1	-1	-2	-2	-3	0	-1	1	1	-1	3	7	-3	-5	-4	-4	-6	-4	-5	-8	-3	-4	-5	-4	-5	-3	-3	-2	-2	-4	0	3	-7	-7	-6	-6	-8	-7	-10
M	-2	-3	-1	-5	-2	-5	-4	-6	-4	-2	-3	-3	-3	8	1	2	0	0	-2	-2	-7	-5	-4	-8	-5	-8	-6	-7	-6	-4	-6	-5	-5	4	-3	-1	-3	-4	-5	-6
I	-2	-4	-2	-5	-2	-7	-6	-8	-6	-4	-5	-5	-5	1	6	1	3	-1	-3	-3	-6	-6	-4	-7	-5	-9	-7	-8	-7	-6	-6	-6	-6	-2	3	-1	0	-3	-5	-7
L	-2	-4	-3	-5	-3	-6	-5	-7	-5	-3	-4	-4	-4	2	1	5	0	0	-2	-2	-6	-6	-5	-7	-5	-9	-8	-8	-7	-5	-5	-5	-6	-2	-2	2	-2	-2	-4	-5
V	-1	-3	-1	-4	-1	-5	-5	-6	-4	-4	-4	-4	-4	0	3	0	6	-1	-3	-4	-5	-5	-3	-6	-3	-7	-6	-7	-5	-6	-6	-6	-6	-3	0	-2	3	-4	-5	-7
F	-2	-4	-4	-6	-3	-6	-5	-7	-6	-5	-2	-5	-6	0	-1	0	-1	8	3	1	-7	-6	-7	-8	-6	-9	-8	-8	-8	-8	-3	-7	-8	-4	-4	-2	-4	5	1	-2
Y	-2	-3	-4	-5	-4	-6	-3	-5	-4	-3	1	-3	-4	-2	-3	-2	-3	3	9	1	-6	-6	-6	-8	-6	-8	-5	-6	-6	-5	-1	-5	-6	-5	-5	-4	-6	1	6	-2
W	-3	-5	-4	-5	-4	-6	-5	-7	-5	-4	-2	-3	-5	-2	-3	-2	-4	1	1	11	-6	-7	-7	-8	-7	-8	-8	-9	-8	-6	-4	-5	-7	-5	-6	-4	-6	-1	0	9
C	9	-6	-6	-9	-5	-9	-8	-11	-10	-8	-7	-8	-8	-7	-6	-6	-5	-7	-6	-6	17	2	1	-1	1	1	1	-2	-3	0	0	1	-1	-2	0	0	1	1	2	3
S	-3	2	-1	-3	-2	-4	-2	-3	-3	-3	-4	-4	-3	-5	-6	-6	-5	-6	-6	-7	2	9	4	2	3	2	4	2	1	2	1	1	1	-1	0	0	1	0	0	-1
T	-4	-2	3	-5	-3	-7	-3	-4	-4	-4	-5	-4	-4	-4	-4	-5	-3	-7	-6	-7	1	4	10	2	4	1	3	2	1	2	1	1	2	1	2	1	3	0	0	-2
P	-7	-4	-5	5	-3	-7	-6	-5	-4	-4	-6	-5	-5	-8	-7	-7	-6	-8	-8	-8	-1	2	2	11	3	0	1	1	0	2	1	0	0	-2	0	1	1	-1	-2	-1
A	-3	-3	-3	-4	2	-4	-4	-4	-3	-3	-5	-4	-4	-5	-5	-5	-3	-6	-6	-7	1	3	4	3	9	2	2	2	2	3	1	1	2	0	1	1	3	0	0	-1
G	-6	-4	-6	-6	-3	4	-3	-4	-5	-5	-6	-5	-5	-8	-9	-9	-7	-9	-8	-8	1	2	1	0	2	10	2	2	1	0	0	2	0	-2	-2	-2	0	-2	-2	0
D N	-6	-3	-3	-6	-4	-4	4	-1	-3	-3	-3	-4	-3	-6	-7	-8	-6	-8	-5	-8	1	4	3	1	2	2	11	4	2	3	3	1	3	-1	0	-1	0	-1	1	-2
i D	-7	-3	-4	-4	-4	-4	-1	4	0	-2	-3	-4	-3	-7	-8	-8	-7	-8	-6	-9	-2	2	2	1	2	2	4	10	5	2	2	0	1	-2	-1	-2	0	-2	-1	-3
s E	-7	-4	-4	-4	-3	-5	-3	0	3	-1	-4	-3	-2	-6	-7	-7	-5	-8	-6	-8	-3	1	1	0	2	1	2	5	9	4	1	1	3	-1	-1	-1	0	-2	-2	-2
o Q	-6	-4	-4	-5	-4	-6	-3	-3	-2	4	-2	-2	-2	-4	-6	-5	-6	-8	-5	-6	0	2	2	2	3	0	3	2	4	11	4	3	3	0	0	1	1	-1	0	1
r H	-5	-5	-4	-5	-5	-6	-2	-4	-4	-2	6	-3	-4	-6	-6	-5	-6	-3	-1	-4	0	1	1	1	1	0	3	2	1	4	13	3	1	-1	0	0	0	2	4	1
d R	-5	-4	-4	-6	-4	-5	-4	-5	-3	-2	-3	4	0	-5	-6	-5	-6	-7	-5	-5	1	1	1	0	1	2	1	0	1	3	3	10	5	-1	0	0	0	-2	-1	2
e K	-7	-4	-4	-5	-4	-6	-3	-4	-2	-1	-4	0	3	-5	-6	-6	-6	-8	-6	-7	-1	1	2	0	2	0	3	1	3	3	1	5	10	-1	0	-1	0	-2	-1	-2
r M	-7	-7	-5	-8	-5	-9	-7	-9	-7	-5	-7	-6	-7	4	-2	-2	-3	-4	-5	-5	-2	-1	1	-2	0	-2	-1	-2	-1	0	-1	-1	-1	13	4	4	3	2	0	0
I	-6	-8	-5	-8	-6	-11	-8	-10	-8	-7	-8	-7	-7	-3	3	-2	0	-4	-5	-6	0	0	2	0	1	-2	0	-1	-1	0	0	0	0	4	12	5	7	4	2	1
L	-5	-8	-6	-8	-5	-10	-8	-9	-7	-5	-6	-6	-6	-1	-1	2	-2	-2	-4	-4	0	0	1	1	1	-2	-1	-2	-1	1	0	0	-1	4	5	10	4	5	2	2
V	-4	-7	-4	-7	-4	-9	-8	-8	-6	-6	-8	-6	-6	-3	0	-2	3	-4	-6	-6	1	1	3	1	3	0	0	0	0	1	0	0	0	3	7	4	11	3	1	0
F	-5	-8	-7	-9	-7	-10	-8	-9	-9	-7	-5	-8	-8	-4	-3	-2	-4	5	1	-1	1	0	0	-1	0	-2	-1	-2	-2	-1	2	-2	-2	2	4	5	3	13	8	6
Y	-5	-8	-7	-9	-7	-9	-6	-8	-8	-6	-3	-6	-7	-5	-5	-4	-5	1	6	0	2	0	0	-2	0	-2	1	-1	-2	0	4	-1	-1	0	2	2	1	8	14	6
W	-5	-9	-9	-10	-8	-11	-9	-10	-10	-8	-6	-7	-10	-6	-7	-5	-7	-2	-2	9	3	-1	-2	-1	-1	0	-2	-3	-2	1	1	2	-2	0	1	2	0	6	6	18

4.2.3 Evaluation

The proposed alignment approach with the expanded substitution matrix performed much worse than the standard alignment algorithm with the BLOSUM62 matrix in the evaluation with the reference query/target dataset. The ranks of the true hits were consistently worse for the proposed approach than for the standard approach. While the proposed alignment approach correctly aligned the true <query,target> pairs, their rank in the E-value sorted list was often pushed down by irrelevant <query,target> matches that were assigned higher scores and lower E-values. Manual inspection suggested that the high scores assigned to such unrelated pairs of sequences are due to the fact that pairs of residues from ID regions in two sequences are more likely to get high scores than pairs of residues from other regions. High positive scores are far overwhelming over the negative scores in the disorder-disorder part of the expanded matrix. This leads to a higher probability that long stretches from two unrelated sequences may be aligned with a very high score just by chance. As the evaluation shows, this probability is too high and introduces a prohibitively high level of false positives.

4.2.4 Discussion

The iterative procedure for estimation of the 40x40 substitution matrix that is described in this section is an effective way of overcoming the lack of ground-truth alignments. The resulting substitution matrix is the fixed point of the mapping defined by steps 2 and 3 of the procedure. It also has the property that it both produces the alignments in step 2, and it is derived from the same alignments.

In the obtained expanded substitution matrix substantial differences were observed between the scores assigned to alignment of disordered-disordered, ordered-ordered and ordered-disordered pairs of amino acids. These differences provide further evidence that evolutionary rates in disordered and ordered regions of proteins are different and that BLOSUM62 and other matrices are not appropriate for alignment of IDPs. In contrast to BLOSUM62 matrix that tends to penalize matching of non-identical amino acids, the expanded matrix tends to assign higher scores (or at least smaller penalties) to the matching of non-identical amino acids in the disordered regions, where due to higher evolutionary rate such mismatches are more likely to occur in nature. The scores for alignment of ordered regions of two sequences in our expanded matrix are similar to scores assigned by the BLOSUM62 matrix. Finally, the expanded matrix assigns the lowest scores (or more precisely: highest penalties) for matching amino acids in ordered regions in one sequence to amino acids in disordered regions in another sequence. This is consistent with the conservation of position and extent of disordered regions in homologous sequences.

The experiments with the random subsets of the dataset showed that the procedure is stable with respect to the selection choice of the protein sequences in the dataset (as long as the selection is done randomly). The results also emphasize the importance of using a large dataset. Furthermore, the results of the experiment with the randomized dataset showed that the differences between four 20x20 submatrices observed in the main experiment were not obtained by chance and that they clearly come from the differences between evolutionary rate in ordered and disordered regions of proteins.

While the expanded substitution matrix with all its properties can be considered a further empirical evidence for previously proposed hypothesis of different evolutionary mechanism in ID, its utility for alignment of sequences is questionable. The evaluation results have shown that, in the general case, the proposed alignment approach is inferior to standard alignment algorithm.

4.3 BLOCKS and BLOSUM Revisited

4.3.1 BLOSUM algorithm and proposed adjustments

BLOSUM matrices (S Henikoff and J G Henikoff 1992) were derived from the BLOCKS database (J G Henikoff and S Henikoff 1996). Each "block" is a gapless multiple alignment of conserved segments from various sequences.

The number in the name refers to the parameter used in preprocessing of the data, which is related to evolutionary distance. For example, before calculation of BLOSUM62, sequences in each block are clustered such that no two sequences from two different clusters have 62% or greater identity. Lower numbers lead to smaller number of bigger clusters, while higher numbers lead to greater number of smaller clusters.

After clustering of sequences in a block the algorithm proceeds by, for each column in the block, counting all pairs of amino acids from different clusters. The contribution of each pair of amino acids is normalized by dividing it with the product of the two sizes of clusters. After normalization, for a block with sequence length l and k clusters, the total sum of contributions is $k(k-1)l/2$. Note that in the original implementation of BLOSUM algorithm this normalization was not implemented correctly, and contribution was only divided by the size of the first cluster (Styczynski et al. 2008). This meant that the

matrices produced by that implementation of the algorithm changed if the order of sequences in the input files changed.

The remaining steps are similar to other algorithms. Contributions of all amino acid pairs from all blocks are counted together and summed up in a matrix M_{ij}. This matrix is normalized to calculate the matrix of amino acid pair frequencies $P=|p_{ij}|$ as $P=(M+M')/2\sum_{i,j}m_{ij}$. From this matrix we calculate marginal frequencies for amino acids $q_i=\sum_j p_{ij}$. From p_{ij} and q_i to calculate all scores using the formula: $score(a_i,a_j)=C\log_2(p_{ij}/q_iq_j)$. We can use this algorithm to calculate both 20x20 and 40x40 matrices. To obtain a 40x20 matrix, the algorithm has to be adjusted to calculate separately marginal frequencies by rows and columns of matrix P.

To obtain disorder prediction for sequence segments and thus facilitate the use of a 40-symbol alphabet, disorder predictor was applied to the original sequences whose segments form the blocks, and not to the segments themselves; segments are too short to obtain reliable prediction, especially at the ends of segments. After prediction on the whole sequences, disorder prediction outputs for their segments were cut out and stored with the BLOCKS.

4.3.2 Testing

For each obtained matrix, we can calculate its relative entropy as $\sum_{i,j}q_iq_j\log_2(p_{ij}/q_iq_j)$, and its expected score as $\sum_{i,j}p_{ij}\log_2(p_{ij}/q_iq_j)$. Relative entropy is important in the testing stage, as comparison of results obtained with two matrices is only fair if they have the same relative entropy (Altschul 1991).

We calculated the 20x20 matrices for various (round integer) values of clustering parameter. We then calculated 40x20 and 40x40 matrices for various values of clustering

parameter on a finer grid, and found matrices whose relative entropies are closest to the relative entropies of the 20x20 matrices. This was to ensure that each 20x20 matrix has corresponding 40x20 and 40x40 matrices with which it can be fairly compared.

For comparison of matrices we used the same methodology as described in Section 4.2.1.4.

4.3.3 Proposed changes to the normalization performed in BLOSUM algorithm

As already mentioned in Section 4.3.1, the contribution of each pair of amino acids is normalized by dividing with the product of the two sizes of clusters. After normalization, for a block with sequence length l and k clusters, the total sum of contributions of amino acid pairs in that block is $k(k–1)l/2$. Therefore the contributions of blocks with the higher number of clusters tend to dominate over the contributions of blocks with the lower number of clusters. The number of clusters is correlated with the number of sequences in a block, and the number of sequences in a block depends on the progress of the curation of the block.

We decided to change the normalization formula, and use a further division by $k(k–1)/2$, and in another version by $(k–1)/2$. In the first case, the total sum of contributions of amino acid pairs in that block is l (the length of block), and in the other case it is kl. While in the second case we still have the factor k, and blocks with larger number of clusters can still dominate the calculation of matrix M, this factor is linear and not quadratic as in the original algorithm.

We tested two new approaches to normalization by calculating 20x20 matrices with equal relative-entropy as the original 20x20 matrix, and then comparing the three matrices.

4.3.4 Experimental results

In this section we compare matrices of size 20x20, 40x20 and 40x40 obtained with BLOSUM algorithm (corrected and adapted for extended alphabet). We add to the comparison the similar matrices produced by variations of the BLOSUM algorithm: we will refer to the algorithm with additional normalization factor $k(k–1)/2$ as Variant A, and to the algorithm with normalization factor $(k–1)/2$ as Variant B.

We compare the entropies of matrices obtained for commonly used values of clustering parameter (50, 62, 80) in Table 4.2. For comparison, the entropies of corresponding BLOSUM matrices are 0.4808 for BLOSUM50, 0.6979 for BLOSUM62, and 0.9868 for BLOSUM80. Entropies for matrices of three different sizes obtained with regular algorithm are fairly similar; same is true for matrices obtained with algorithms Variant A and Variant B. However, entropies of matrices obtained with various variants differ greatly. For matrices obtained with algorithm Variant B the entropy is higher than for regular matrices, and it is even higher for matrices obtained with algorithm Variant A. Variant B (normalization by $(k–1)/2$) lessens the dominance of blocks with high number of clusters, while Variant A (normalization by $k(k–1)/2$) completely alleviates the problem and allows for information from all blocks to be weighted equally.

ROC curves for evaluation of these 9 groups of matrices are compared in Figure 4.4–Figure 4.6. The 20x20 and 40x20 matrices in all cases have interlaced ROC curves, and it is hard to establish clearly which of the two matrices has better performance. In all cases the 40x40 matrix performs much worse than the other two matrices, which is consistent with the evaluation of the matrix obtained with iterative procedure described in Section 4.2.

Table 4.2. Comparison of entropies for matrices obtained with adjusted BLOSUM algorithm and algorithm variants A and B

Identity 50%

	Regular	Variant A	Variant B
20x20	0.4876	0.5922	0.5488
40x20	0.4887	0.5934	0.5498
40x40	0.4933	0.6002	0.5553

Identity 62%

	Regular	Variant A	Variant B
20x20	0.7096	0.8540	0.7972
40x20	0.7107	0.8550	0.7981
40x40	0.7158	0.8621	0.8036

Identity 80%

	Regular	Variant A	Variant B
20x20	0.9699	1.1847	1.0870
40x20	0.9714	1.1856	1.0879
40x40	0.9779	1.1944	1.0945

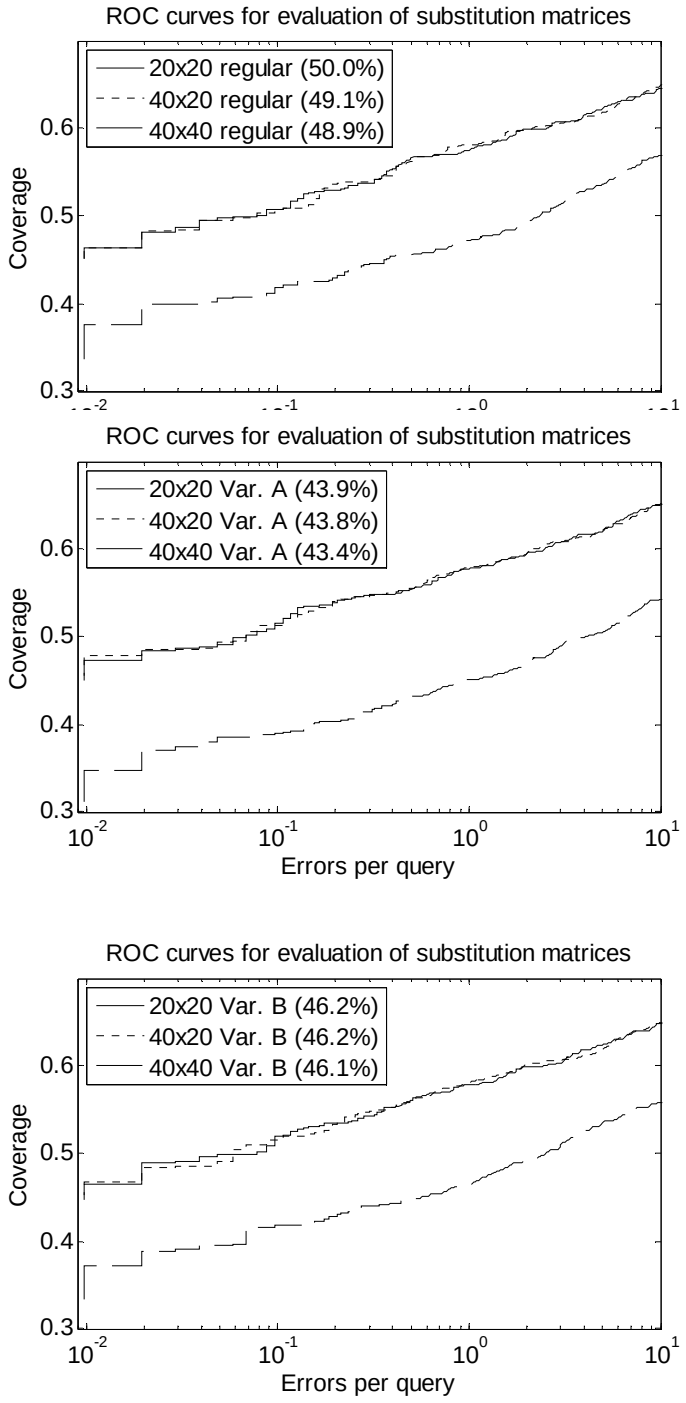

Figure 4.4. Comparison of ROC curves for evaluation of 20x20, 40x20 and 40x40 matrices (equivalent to BLOSUM50), obtained with various algorithm variants.

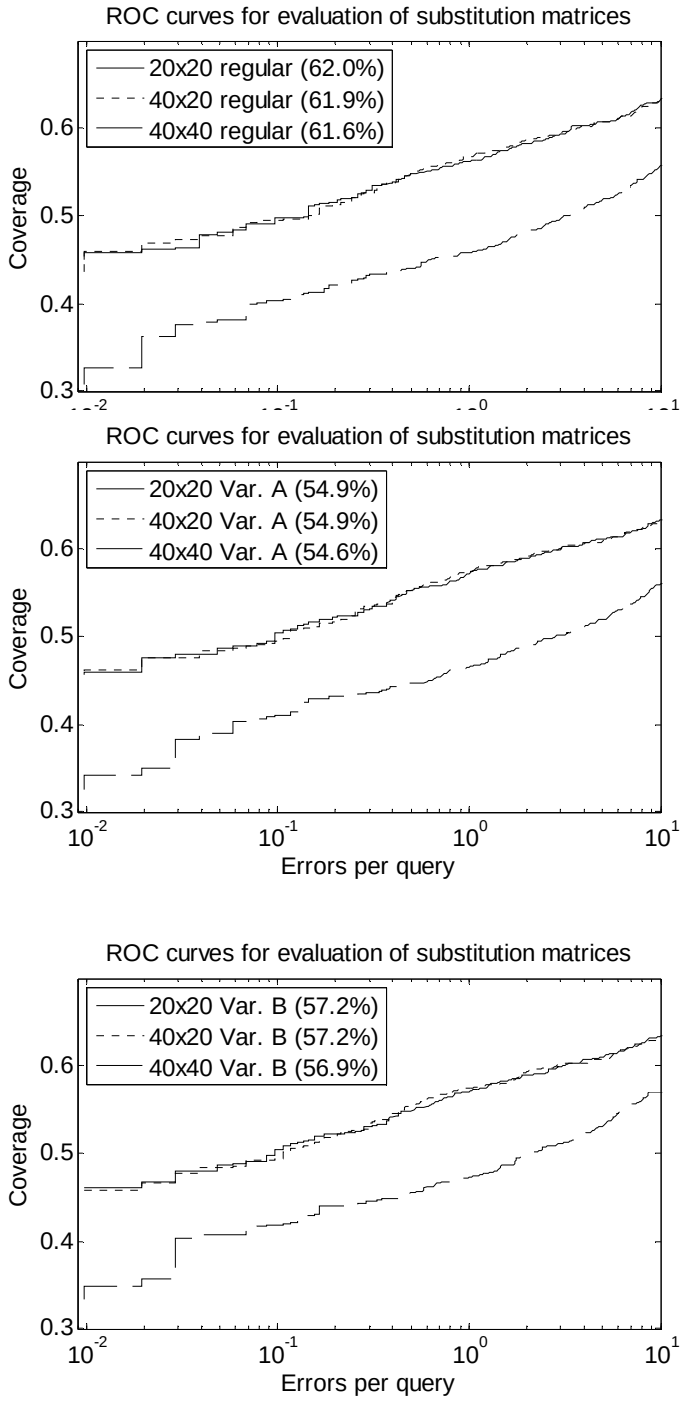

Figure 4.5. Comparison of ROC curves for evaluation of 20x20, 40x20 and 40x40 matrices (equivalent to BLOSUM62), obtained with various algorithm variants.

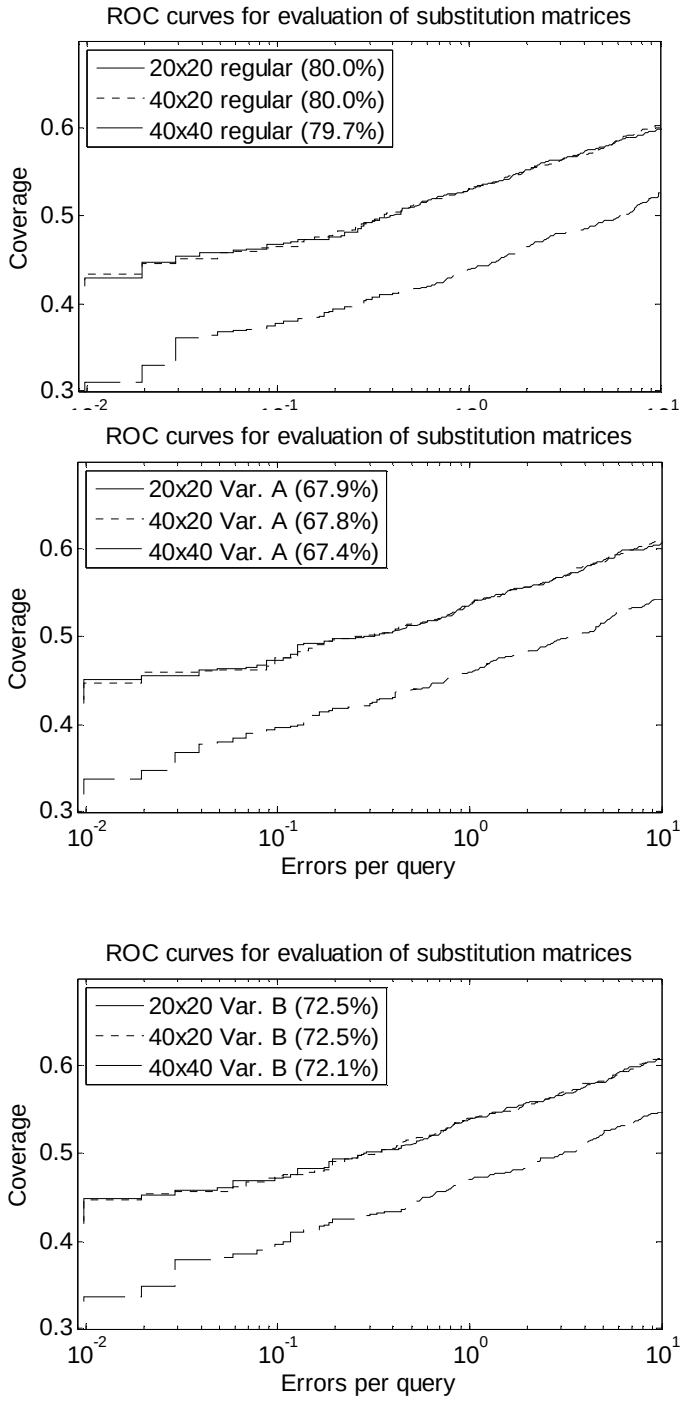

Figure 4.6. Comparison of ROC curves for evaluation of 20x20, 40x20 and 40x40 matrices (equivalent to BLOSUM80), obtained with various algorithm variants.

Finally, we compare the matrices of the same size obtained by various algorithm variants (regular, variant A, variant B) in Figure 4.7 (20x20) and Figure 4.8 (40x20).

For all 6 combinations of clustering thresholds and matrix sizes it is hard to compare the performance of two matrices obtained with modified variants of the algorithm, as the ROC curves are interlaced.

For lower clustering identity threshold (50%, upper plots in Figure 4.7 and Figure 4.8), the ROC curve for matrix obtained with the regular algorithm is close to the ROC curves of two variant matrices, although we can observe that in certain regions of EPQ, ROC curve for regular matrix is clearly below the ROC curves for two variant matrices. This is consistent with the fact that for the lower clustering identity threshold, the obtained clusters are larger, and the numbers of clusters for all blocks are generally lower than for higher clustering identity thresholds. Because the numbers of clusters are lower, the problem of dominance of blocks with high cluster numbers is less obvious.

For the two higher clustering identity thresholds (62% and 80%, middle and lower plots in Figure 4.7 and Figure 4.8), we can clearly observe that the ROC curve for regular matrix is below the ROC curves for two variant matrices. This means that the regular matrix performs worse than two variant matrices, when matrices are obtained with higher clustering identity thresholds. At these thresholds, the numbers of clusters for blocks are larger. The original BLOSUM algorithm loses parts of information available from the blocks with smaller numbers of clusters, because of the dominance of blocks with higher numbers of clusters.

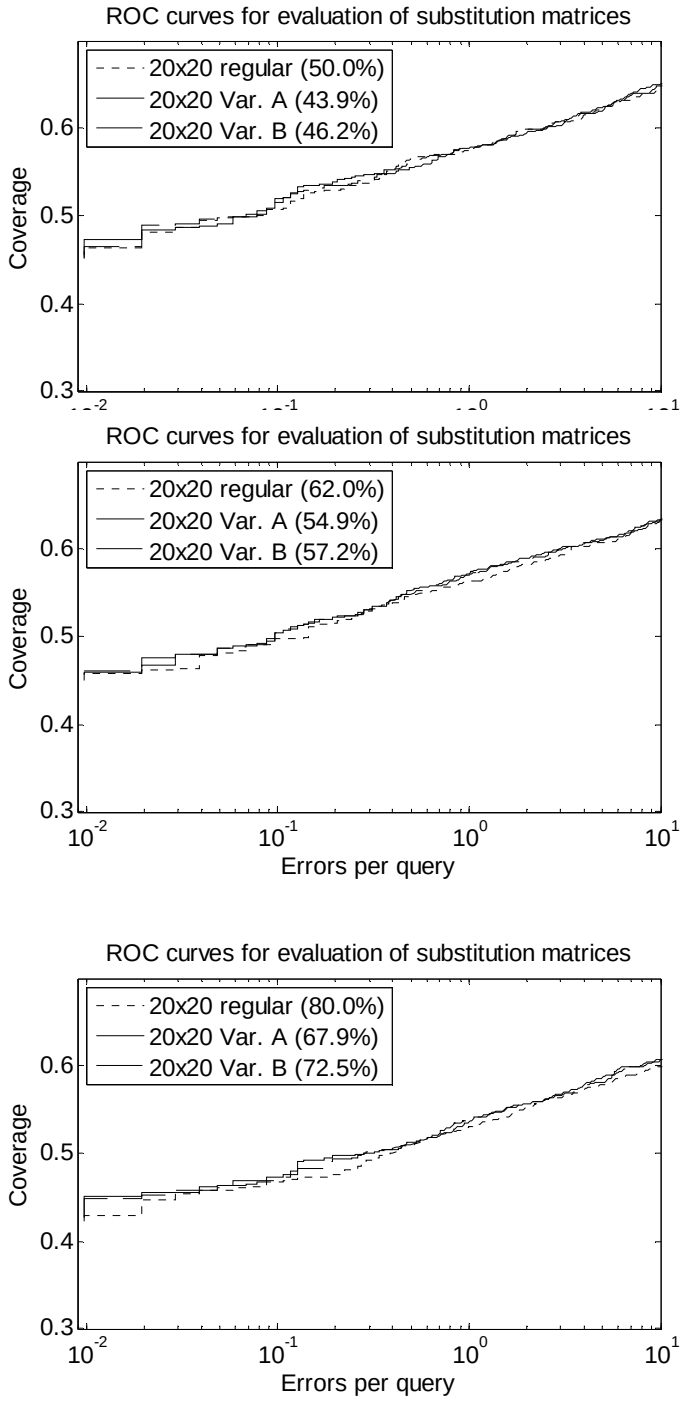

Figure 4.7. Comparison of ROC curves for evaluation of regular, variant A and variant B 20x20 matrices.

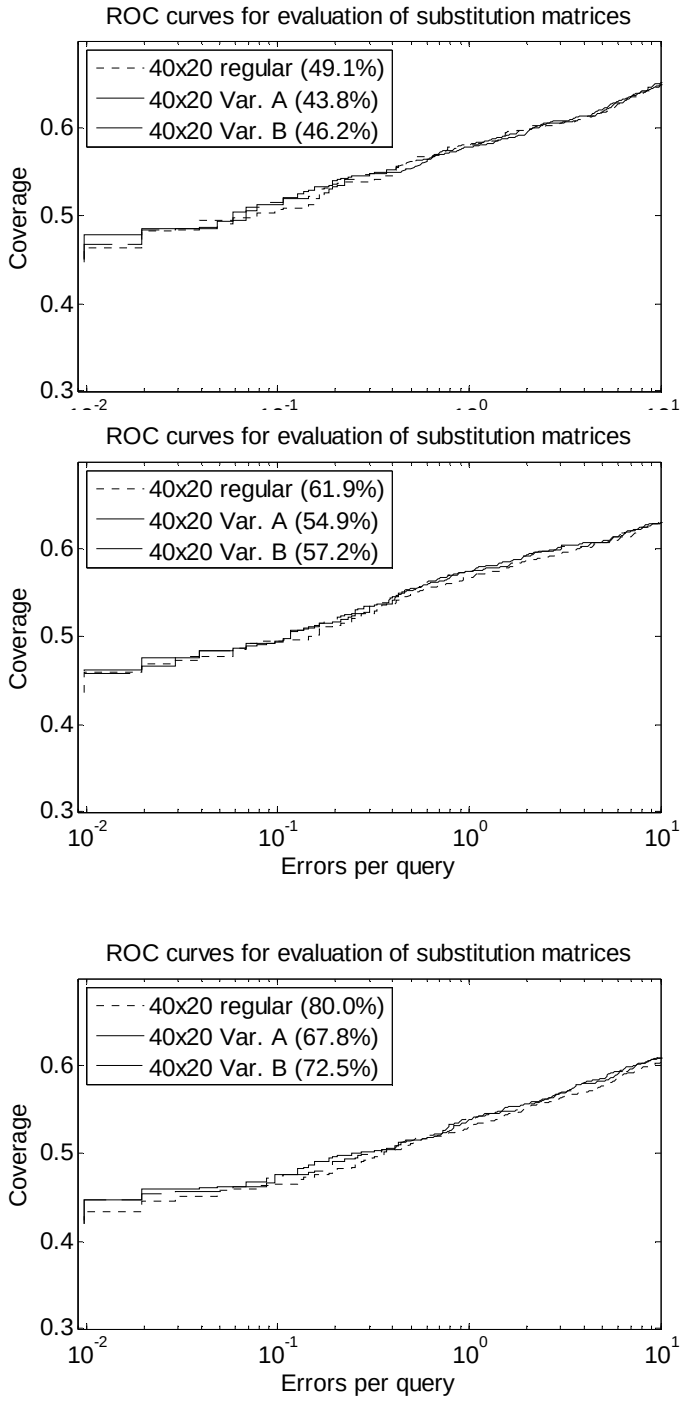

Figure 4.8. Comparison of ROC curves for evaluation of regular, variant A and variant B 40x20 matrices.

4.3.5 Conclusion

We altered the original BLOSUM algorithm to facilitate the use of disorder information and calculate matrices of sizes 40x20 and 40x40 from BLOCKS database.

Similarly to the situation with the matrix obtained by the iterative procedure described in Section 4.2, full matrices of size 40x40 perform much worse than the corresponding 40x20 and 20x20 matrices.

It is hard to compare the performance of 20x20 and corresponding 40x20 matrices, as their ROC curves are very close, they are interlaced and they change order in different regions of EPQ. It is therefore not clear if the inclusion of disorder information can help improve sequence alignment.

On the other hand, a very basic question related to the underlying BLOSUM algorithm – the question of normalization with respect to numbers of clusters in the blocks – has a great effect on the outcome of the algorithm, and on the performance of the produced matrix. For higher values of clustering identity threshold, the matrices produced with two proposed variants of normalization formula performed better than the corresponding matrices produced by the regular algorithm.

More importantly, the variation in performance that is introduced by the change in the algorithm itself is much larger than the variation introduced by the inclusion of the disorder information. This suggests that the inclusion of disorder into sequence alignment process may not be the top priority, and that instead there are still more open important questions about the basic sequence alignment and substitution matrix producing algorithms.

REFERENCES

Altschul, S F. 1991. “Amino acid substitution matrices from an information theoretic perspective.” *Journal of molecular biology* 219 (3) (June): 555-65. http://www.ncbi.nlm.nih.gov/pubmed/2051488.

Altschul, S F, W. Gish, W. Miller, E W Myers, and D J Lipman. 1990. “Basic local alignment search tool.” *Journal of molecular biology* 215 (3) (October): 403-10. doi:10.1006/jmbi.1990.9999. http://www.ncbi.nlm.nih.gov/pubmed/2231712.

Balasubramanian, S, T Schneider, M Gerstein, and L Regan. 2000. “Proteomics of Mycoplasma genitalium: identification and characterization of unannotated and atypical proteins in a small model genome.” *Nucleic Acids Res* 28 (16): 3075-3082.

Benjamini, Y, and D Yekutieli. 2001. “The control of the false discovery rate in multiple testing under dependency.” *Annals of Statistics* 29: 1165-1188.

Benjamini, Y., and Y. Hochberg. 1995. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Breiman, Leo. 1996. “Bagging predictors.” *Machine Learning* 24 (2) (August): 123-140. doi:10.1007/BF00058655. http://www.springerlink.com/index/10.1007/BF00058655.

Brown, Celeste J, Sachiko Takayama, Andrew M Campen, Pam Vise, Thomas W Marshall, Christopher J Oldfield, Christopher J Williams, and a Keith Dunker. 2002.

"Evolutionary rate heterogeneity in proteins with long disordered regions." *Journal of molecular evolution* 55 (1) (July): 104-10. doi:10.1007/s00239-001-2309-6. http://www.ncbi.nlm.nih.gov/pubmed/12165847.

Burge, C, and S Karlin. 1997. "Prediction of complete gene structures in human genomic DNA." *Journal of molecular biology* 268 (1) (April): 78-94. doi:10.1006/jmbi.1997.0951. http://www.ncbi.nlm.nih.gov/pubmed/9149143.

Cheng, Y, T LeGall, C J Oldfield, A K Dunker, and V N Uversky. 2006. "Abundance of intrinsic disorder in protein associated with cardiovascular disease." *Biochemistry* 45 (35): 10448-10460.

Cheng, Yugong, Christopher J Oldfield, Jingwei Meng, Pedro Romero, Vladimir N Uversky, and A Keith Dunker. 2007. "Mining alpha-helix-forming molecular recognition features with cross species sequence alignments." *Biochemistry* 46 (47) (November): 13468-77. doi:10.1021/bi7012273. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570644&tool=pmcentrez&rendertype=abstract.

Chenna, Ramu, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. 2003. "Multiple sequence alignment with the Clustal series of programs." *Nucleic acids research* 31 (13) (July 1): 3497-500. http://www.ncbi.nlm.nih.gov/pubmed/12824352.

Cortese, M S, V N Uversky, and A Keith Dunker. 2008. "Intrinsic disorder in scaffold proteins: Getting more from less." *Prog Biophys Mol Biol* 98 (1): 85-106.

Creighton, T E. 1988. “The protein folding problem.” *Science* 240 (4850): 267-344.

Crick, S L, M Jayaraman, C Frieden, R Wetzel, and R V Pappu. 2006. “Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions.” *Proc Natl Acad Sci USA* 103 (45): 16764-16769.

Daughdrill, G W, G J Pielak, V N Uversky, M S Cortese, and A K Dunker. 2005. “Natively disordered proteins.” *Handbook of Protein Folding*: 271-353.

Dayhoff, M.O., and R.M. Schwartz. 1978. A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*, 345-358. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315.

Dosztanyi, Z, J Chen, A K Dunker, I Simon, and P Tompa. 2006. “Disorder and sequence repeats in hub proteins and their implications for network evolution.” *J Proteome Res* 5 (11): 2985-2995.

Dosztanyi, Z, V Csizmok, P Tompa, and I Simon. 2005. “IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.” *Bioinformatics* 21 (16): 3433-3434.

Dunker, A K, C J Brown, J D Lawson, L M Iakoucheva, and Z Obradovic. 2002. “Intrinsic disorder and protein function.” *Biochemistry* 41 (21): 6573-6582.

Dunker, A K, M S Cortese, P Romero, L M Iakoucheva, and V N Uversky. 2005. "Flexible nets. The roles of intrinsic disorder in protein interaction networks." *Febs J* 272 (20): 5129-5148.

Dunker, A K, E Garner, S Guilliot, P Romero, K Albrecht, J Hart, Z Obradovic, C Kissinger, and J E Villafranca. 1998. "Protein disorder and the evolution of molecular recognition: theory, predictions and observations." *Pac Symp Biocomput*: 473-484.

Dunker, A K, J D Lawson, C J Brown, R M Williams, P Romero, J S Oh, C J Oldfield, et al. 2001. "Intrinsically disordered protein." *Journal of molecular graphics & modelling* 19 (1) (January): 26-59. http://www.ncbi.nlm.nih.gov/pubmed/11381529.

Dunker, A K, Z Obradovic, P Romero, E C Garner, and C J Brown. 2000. "Intrinsic protein disorder in complete genomes." *Genome informatics. Workshop on Genome Informatics* 11 (January): 161-71. http://www.ncbi.nlm.nih.gov/pubmed/11700597.

Dunker, A K, and Z Obradovic. 2001. "The protein trinity - linking function and disorder." *Nat Biotechnol* 19 (9): 805-806.

Dyson, H J, and P E Wright. 2002. "Coupling of folding and binding for unstructured proteins." *Curr Opin Struct Biol* 12 (1): 54-60.

———. 2005. "Intrinsically unstructured proteins and their functions." *Nat Rev Mol Cell Biol* 6 (3): 197-208.

Ekman, D, S Light, A K Bjorklund, and A Elofsson. 2006. “What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?” *Genome Biol* 7 (6): R45.

Garner, E, P Romero, A K Dunker, C Brown, and Z Obradovic. 1999. “Predicting Binding Regions within Disordered Proteins.” *Genome Inform Ser Workshop Genome Inform* 10: 41-50.

Goh, K I, M E Cusick, D Valle, B Childs, M Vidal, and A L Barabasi. 2007. “The human disease network.” *Proc Natl Acad Sci USA* 104 (21): 8685-8690.

Haynes, C, C J Oldfield, F Ji, N Klitgord, M E Cusick, P Radivojac, V N Uversky, M Vidal, and L M Iakoucheva. 2006. “Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.” *PLoS Comput Biol* 2 (8): e100.

He, Bo, Kejun Wang, Yunlong Liu, Bin Xue, Vladimir N Uversky, and A Keith Dunker. 2009. “Predicting intrinsic disorder in proteins: an overview.” *Cell research* 19 (8) (August): 929-49. doi:10.1038/cr.2009.87. http://www.ncbi.nlm.nih.gov/pubmed/19597536.

Henikoff, J G, and S Henikoff. 1996. “Blocks database and its applications.” *Methods in enzymology* 266 (January): 88-105. http://www.ncbi.nlm.nih.gov/pubmed/8743679.

Henikoff, S, and J G Henikoff. 1992. “Amino acid substitution matrices from protein blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22) (November): 10915-9.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=50453&tool=pmcentrez&rendertype=abstract.

Iakoucheva, L M, C J Brown, J D Lawson, Z Obradovic, and A K Dunker. 2002. "Intrinsic disorder in cell-signaling and cancer-associated proteins." *J Mol Biol* 323 (3): 573-584.

Kramer, C Y. 1956. "Extension of multiple range tests to group means with unequal numbers of replications." *Biometrics Bulletin* 12: 307-310.

Lavery, D N, and I J McEwan. 2008. "Structural characterization of the native NH2-terminal transactivation domain of the human androgen receptor: a collapsed disordered conformation underlies structural plasticity and protein-induced folding." *Biochemistry* 47 (11): 3360-3369.

Mann, H B, and D R Whitney. 1947. "On a test of whether one of two random variables is stochastically larger than the other." *Annals of Mathematical Statistics* 18: 50-60.

Midic, Uros, Christopher Oldfield, A Keith Dunker, Zoran Obradovic, and Vladimir Uversky. 2009. "Protein disorder in the human diseasome: unfoldomics of human genetic diseases." *BMC Genomics* 10 (Suppl 1): S12. doi:10.1186/1471-2164-10-S1-S12. http://www.biomedcentral.com/1471-2164/10/S1/S12.

Mohan, A, C J Oldfield, P Radivojac, V Vacic, M S Cortese, A K Dunker, and V N Uversky. 2006. "Analysis of molecular recognition features (MoRFs)." *J Mol Biol* 362 (5): 1043-1059.

Mohan, A, W J Sullivan, P Radivojac, A K Dunker, and V N Uversky. 2008. “Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes.” *Mol Biosyst* 4 (4): 328-340.

Needleman, S B, and C D Wunsch. 1970. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” *Journal of molecular biology* 48 (3) (March): 443-53. http://www.ncbi.nlm.nih.gov/pubmed/5420325.

Obradovic, Zoran, Kang Peng, Slobodan Vucetic, Predrag Radivojac, Celeste J Brown, and A Keith Dunker. 2003. “Predicting intrinsic disorder from amino acid sequence.” *Proteins* 53 Suppl 6 (January): 566-72. doi:10.1002/prot.10532. http://www.ncbi.nlm.nih.gov/pubmed/14579347.

Oldfield, C J, Y Cheng, M S Cortese, C J Brown, V N Uversky, and A K Dunker. 2005. “Comparing and combining predictors of mostly disordered proteins.” *Biochemistry* 44 (6): 1989-2000.

Oldfield, C J, Y Cheng, M S Cortese, P Romero, V N Uversky, and A K Dunker. 2005. “Coupled folding and binding with alpha-helix-forming molecular recognition elements.” *Biochemistry* 44 (37): 12454-12470.

Oldfield, C J, J Meng, J Y Yang, M Q Yang, V N Uversky, and A K Dunker. 2008. “Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners.” *BMC Genomics* 9 (Suppl 1): S1.

Oldfield, C J, E L Ulrich, Y Cheng, A K Dunker, and J L Markley. 2005. “Addressing the intrinsic disorder bottleneck in structural proteomics.” *Proteins* 59 (3): 444-453.

Pachter, Lior, Marina Alexandersson, and Simon Cawley. 2002. “Applications of generalized pair hidden Markov models to alignment and gene finding problems.” *Journal of computational biology□: a journal of computational molecular cell biology* 9 (2) (January): 389-99. doi:10.1089/10665270252935520. http://www.ncbi.nlm.nih.gov/pubmed/12015888.

Patil, A, and H Nakamura. 2006. “Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.” *FEBS Lett* 580 (8): 2041-2045.

Peng, K, P Radivojac, S Vucetic, A K Dunker, and Z Obradovic. 2006. “Length-dependent prediction of protein intrinsic disorder.” *BMC Bioinformatics* 7: 208.

Peng, Kang, Predrag Radivojac, Slobodan Vucetic, A Keith Dunker, and Zoran Obradovic. 2006. “Length-dependent prediction of protein intrinsic disorder.” *BMC bioinformatics* 7 (January): 208. doi:10.1186/1471-2105-7-208. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1479845&tool=pmcentrez&rendertype=abstract.

Peng, Kang, Slobodan Vucetic, Predrag Radivojac, Celeste J Brown, a Keith Dunker, and Zoran Obradovic. 2005. “Optimizing long intrinsic disorder predictors with protein evolutionary information.” *Journal of bioinformatics and computational biology* 3 (1) (February): 35-60. http://www.ncbi.nlm.nih.gov/pubmed/15751111.

Radivojac, P, L M Iakoucheva, C J Oldfield, Z Obradovic, V N Uversky, and A K Dunker. 2007. “Intrinsic disorder and functional proteomics.” *Biophys J* 92 (5): 1439-1456.

Radivojac, P, S Vucetic, T R O’Connor, V N Uversky, Z Obradovic, and A K Dunker. 2006. “Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition.” *Proteins* 63 (2): 398-410.

Radivojac, Predrag, Zoran Obradovic, C.J. Brown, and A.K. Dunker. 2002. Improving sequence alignments for intrinsically disordered proteins. In *Pac. Symp. Biocomput*, 7:589–600. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.1309&rep=rep1&type=pdf.

Romero, P R, S Zaidi, Y Y Fang, V N Uversky, P Radivojac, C J Oldfield, M S Cortese, M Sickmeier, T LeGall, and Z Obradovic. 2006. “Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.” *Proc Natl Acad Sci USA* 103 (22): 8390-8395.

Romero, P, Z Obradovic, C R Kissinger, J E Villafranca, E Garner, S Guilliot, and A K Dunker. 1998. “Thousands of proteins likely to have long disordered regions.” *Pac Symp Biocomput*: 437-448.

Romero, P, Z Obradovic, C Kissinger, J E Villafranca, and A K Dunker. 1997. “Identifying disordered regions in proteins from amino acid sequence.” *1997 Proceedings of International Conference on Neural Networks* 1: 90-95.

Romero, P, Z Obradovic, X Li, E C Garner, C J Brown, and A K Dunker. 2001. "Sequence complexity of disordered protein." *Proteins* 42 (1) (January 1): 38-48. http://www.ncbi.nlm.nih.gov/pubmed/11093259.

Salzberg, S L, A L Delcher, S Kasif, and O White. 1998. "Microbial gene identification using interpolated Markov models." *Nucleic acids research* 26 (2) (January 15): 544-8. http://www.ncbi.nlm.nih.gov/pubmed/9421513.

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

Schäffer, A A, L Aravind, T L Madden, S Shavirin, J L Spouge, Y I Wolf, E V Koonin, and S F Altschul. 2001. "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." *Nucleic acids research* 29 (14) (July): 2994-3005. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=55814&tool=pmcentrez&rendertype=abstract.

Sickmeier, M, J A Hamilton, T LeGall, V Vacic, M S Cortese, A Tantos, B Szabo, P Tompa, J Chen, and V N Uversky. 2007. "DisProt: the Database of Disordered Proteins." *Nucleic Acids Res* (35 Database): D786-793.

Sigalov, A B. 2004. "Multichain immune recognition receptor signaling: different players, same game?" *Trends Immunol* 25 (11): 583-589.

———. 2006. “Immune cell signaling: a novel mechanistic model reveals new therapeutic targets.” *Trends Pharmacol Sci* 27 (10): 518-524.

Sigalov, A B, D A Aivazian, V N Uversky, and L J Stern. 2006. “Lipid-binding activity of intrinsically unstructured cytoplasmic domains of multichain immune recognition receptor signaling subunits.” *Biochemistry* 45 (51): 15731-15739.

Sigalov, A B, A V Zhuravleva, and V Y Orekhov. 2007. “Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form.” *Biochimie* 89 (3): 419-421.

Sigalov, A, D Aivazian, and L Stern. 2004. “Homooligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif.” *Biochemistry* 43 (7): 2049-2061.

Singh, G P, M Ganapathi, K S Sandhu, and D Dash. 2006. “Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes.” *Proteins* 62 (2): 309-315.

Smith, TF, and MS Waterman. 1981. “Identification of common molecular subsequences.” *J. Mol. Bwl* 147: 195–197. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.2897&rep=rep1&type=pdf.

Styczynski, Mark P, Kyle L Jensen, Isidore Rigoutsos, and Gregory Stephanopoulos. 2008. “BLOSUM62 miscalculations improve search performance.” *Nature*

biotechnology 26 (3) (March): 274-5. doi:10.1038/nbt0308-274. http://www.ncbi.nlm.nih.gov/pubmed/18327232.

Tompa, P. 2002. “Intrinsically unstructured proteins.” *Trends Biochem Sci* 27 (10): 527-533.

Tran, H T, A Mao, and R V Pappu. 2008. “Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins.” *J Am Chem Soc* 130 (23): 7380-7392.

Tukey, J W. 1953. “The problem of multiple comparisons.”

Uversky, V N. 2002. “Natively unfolded proteins: a point where biology waits for physics.” *Protein Sci* 11 (4): 739-756.

———. 2003. “Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?” *Cell Mol Life Sci* 60 (9): 1852-1871.

———. 2008a. “Amyloidogenesis of natively unfolded proteins.” *Current Alzheimer Research* 5 (3): 260-287.

———. 2008b. “Intrinsic disorder in proteins associated with neurodegenerative diseases.” *Protein Folding and Misfolding: Neurodegenerative Diseases*: 21-75.

Uversky, V N, J R Gillespie, and A L Fink. 2000. “Why are ‘natively unfolded’ proteins unstructured under physiologic conditions?” *Proteins* 41 (3): 415-427.

Uversky, V N, C J Oldfield, and A K Dunker. 2005. “Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.” *J Mol Recognit* 18 (5): 343-384.

———. 2008. “Intrinsically disordered proteins in human diseases: Introducing the D2 concept.” *Ann Rev Biophys Biomol Structure* 37: 215-246.

Vacic, V, C J Oldfield, A Mohan, P Radivojac, M S Cortese, V N Uversky, and A K Dunker. 2007. “Characterization of molecular recognition features, MoRFs, and their binding partners.” *J Proteome Res* 6 (6): 2351-2366.

Vucetic, Slobodan, Hongbo Xie, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Zoran Obradovic, and Vladimir N Uversky. 2007. “Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions.” *Journal of proteome research* 6 (5) (May): 1899-916. doi:10.1021/pr060393m. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2588346&tool=pmcentrez&rendertype=abstract.

Ward, J J, J S Sodhi, L J McGuffin, B F Buxton, and D T Jones. 2004. “Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.” *J Mol Biol* 337 (3): 635-645.

Weinreb, P H, W Zhen, A W Poon, K A Conway, and P T Lansbury. 1996. “NACP, a protein implicated in Alzheimer’s disease and learning, is natively unfolded.” *Biochemistry* 35 (43): 13709-13715.

Wilcoxon, F. 1945. “Individual comparisons by ranking methods.” *Biometrics Bulletin* 1: 80-83.

Williams, R M, Z Obradovic, V Mathura, W Braun, E C Garner, J Young, S Takayama, C J Brown, and A K Dunker. 2001. “The protein non-folding problem: amino acid determinants of intrinsic order and disorder.” *Pac Symp Biocomput*: 89-100.

Wright, P E, and H J Dyson. 1999. “Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.” *Journal of molecular biology* 293 (2) (October 22): 321-31. doi:10.1006/jmbi.1999.3110. http://www.ncbi.nlm.nih.gov/pubmed/10550212.

Xie, Hongbo, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Zoran Obradovic, and Vladimir N Uversky. 2007. “Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.” *Journal of proteome research* 6 (5) (May): 1917-32. doi:10.1021/pr060394e. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2588348&tool=pmcentrez&rendertype=abstract.

Xie, Hongbo, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Vladimir N Uversky, and Zoran Obradovic. 2007. “Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.” *Journal of proteome research* 6 (5) (May): 1882-98. doi:10.1021/pr060392u.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2543138&tool=pmcentrez&rendertype=abstract.