Published in final edited form as:

Methods Mol Biol. 2014 ; 1091: 3–16. doi:10.1007/978-1-62703-691-7_1.

DisMeta – a Meta Server for Construct Design and Optimization

Yuanpeng Janet Huang^{*}, Thomas B. Acton, and Gaetano T. Montelione^{*}

Center for Advanced Biotechnology and Medication, Northeast Structural Genomics Consortium, Rutgers University

Summary

Intrinsically disordered or unstructured regions in proteins are both common and biologically important, particularly in regulation, signaling and modulating intermolecular recognition processes. From a practical point of view, however, such disordered regions often can pose significant challenges for crystallization. Disordered regions are also detrimental to NMR spectral quality, complicating the analysis of resonance assignments and three-dimensional protein structures by NMR methods. The DisMeta Server has been used by Northeastern Structural Genomics Consortium (NESG) as a primary tool for construct design and optimization in preparing samples for both NMR and crystallization studies. It is a meta-server that generates a consensus analysis of eight different sequence-based disorder predictors to identify regions that are likely to be disordered. DisMeta also identifies predicted secretion signal peptides, trans-membrane segments, and low complexity regions. Identification of disordered regions, by either experimental or computational methods, is an important step in the NESG structure production pipeline, allowing the rational design of protein constructs that have improved expression and solubility, improved crystallization, and better quality NMR spectra.

Keywords

intrinsically disorder protein prediction; construct design; construct optimization; hydrogen-deuterium exchange with mass spectrometry (HDX-MS)

1. Introduction

Intrinsically disordered or unstructured regions in proteins are both common and biologically important, particularly in modulating intermolecular recognition processes in cellular regulation and signaling. Intrinsically disordered proteins also have broad associations with human diseases (1-4). Intrinsic disorder, and or disorder-to-order structural transitions, is an important feature of many transcription factors (5,6), signaling scaffold proteins (7,8), stress-related proteins including protein chaperones (9) and hub proteins involved in protein-protein interactions (10,11).

From a practical point of view, highly homogeneous protein samples with minimal amounts of intrinsic disorder are generally more amenable for successful protein crystallization and

^{*}Corresponding Authors: Gaetano T. Montelione and Yuanpeng Janet Huang, 679 Hoes Lane, Piscataway NJ 08854, guy@cabm.rutgers.edu, yphuang@cabm.rutgers.edu.

structure determination by X-ray crystallography (12-14). While NMR can often be used successfully to study even fully disordered proteins, disordered segments of proteins can promote aggregation and deleteriously affect NMR spectral quality. In addition, a large percentage of targets for NMR studies, particularly human and other eukaryotic proteins, are multidomain proteins, which often misfold in prokaryotic systems (15). Such multi-domain proteins also often exceed the size limitations for NMR structure determination techniques. Domain parsing can be used to circumvent these significant issues.

It can be extremely challenging to predict the protein subsequence that will produce a soluble well-behaved protein, particularly in studies of domains for which the three-dimensional structure is not yet known. In particular, it can be challenging to accurately predict domain boundaries and locations of disordered residues. This information is critical for identifying open reading frames that can be expressed in soluble form in bacterial or even in eukaryotic expression systems.

The DisMeta Server has been developed by the Northeast Structural Genomics Consortium (NESG) as our primary tool for design and optimization of protein constructs expressed for both NMR and crystallization studies (<http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/>, Fig. 1). It is a meta-server that generates a consensus analysis of eight sequence-based disorder predictors to identify regions of the protein that are likely to be disordered (Fig 2B). The DisMeta server uses a consensus approach for disorder prediction. The consensus analysis is conservative in identifying disordered regions, minimizing the possibility of cutting into an ordered region of the protein due to an inaccurate prediction that it is disorder; i.e. low false positive rates. The server also identifies predicted secretion signal peptides using SignalP (16), transmembrane segments by TMHMM (17), low complexity SEG regions (18), secondary structure by PROFsec (19) and PSIPred (20), and ANCHOR (21) (Fig. 2A).

The data from these disorder consensus of prediction servers, along with multiple sequence alignments of homologous proteins and hidden Markov models characteristic of the targeted protein domain families (22,23), are used to predict possible structural domain boundaries. Based on this information, the user can generate nested sets of alternative constructs for full-length proteins, multi-domain constructs, and single-domain constructs. These alternative constructs often possess significantly better expression, solubility, and biophysical behavior than their full-length parent sequences, increasing the likelihood of success in crystallization and the efficiency of structure production. The success of this multiple construct strategy has been reported by the NESG and others (24-27). NESG has developed an automated construct design software, which generates nested sets of constructs based on the DisMeta report, to assist this construct design process for large-scale protein sample production (unpublished results) (27). In addition, the server identifies domain-sized regions of the protein that are amenable to high-throughput NMR studies, allowing structural studies of proteins that would otherwise be too large to study by NMR.

The results from the DisMeta Server have been compared with biophysical data on protein disorder, including HSQC NMR, ^{15}N nuclear relaxation rate, and hydrogen-deuterium exchange mass spectroscopy (HDX-MS) data (14) for many NESG protein targets. We have

found that consensus disordered region, regions predicted to be signal peptide, and/or trans-membrane regions and low complexity regions identified by DisMeta are often disordered in these experimental studies of protein samples. As shown in some examples in the Method Section, truncated constructs lacking residues from these disordered regions have been successfully generated and used to provide diffraction quality crystals and/or good NMR spectra suitable for determining the 3D protein structure (28,29,14,26,27).

2. Materials

2.1 Software used by the DisMeta Server

The DisMeta Server runs several different disorder prediction software, including DISEMBL(30), DISOPRED2 (31), DISPro (32), FoldIndex (33), GlobPlot2 (34), IUPred (35), RONN (36), VL2 (37). For more detailed reviews on disorder prediction methods, see reference (3).

The DisMeta Server also provides sequence-based structural prediction results from other bioinformatics software, including PROF (19), PSIPred (20), SignalP (16), TMHMM (17), Coils (38), SEG (18) and ANCHOR (21). These information are used together with the disorder predictions for construct design and optimization for both NMR and crystallization studies.

2.2 Amide hydrogen deuterium exchange with mass spectrometry (HDX-MS) detection

HDX-MS measurements are based on the concept that backbone amide protons in disordered regions are solvent accessible and therefore exchange with solvent deuterium ($^2\text{H}_2\text{O}$) at a faster rate than backbone amide protons in less solvent-accessible ordered regions where they are generally involved in hydrogen bonds. The degree of exchange over various time intervals is assessed by quenching the exchange kinetics by lowering the pH and temperature, fragmenting the protein by pepsin proteolysis, and measuring the mass of the resulting fragments by mass spectrometry. Peptides with greater mass (higher deuterium exchange) compared to the fully protonated control are identified. The protocols used in this work for HDX-MS studies have been described in detail elsewhere (14). The results are depicted graphically as a heat map (an example is shown in Fig. 2.4B); residues in peptides with the greatest amount of mass increase (disordered regions) are represented with red boxes. Residues in peptides showing little or no mass increase (ordered regions) are represented by blue boxes.

3. Methods

Identification of disordered regions of the protein by the DisMeta server, and elimination of these residues from the protein construct is one of the keys to the NESG construct design and optimization process. Over the past decade, the use of DisMeta for construct design and optimization has greatly increased the efficiency of protein sample and structure production for this large-scale structural genomics project.

3.1 Excluding consensus disorder regions from construct design

Figure 2 shows the DisMeta report for the *Escherichia coli* Spr lipoprotein (NESG target ER541), which originally provided NMR data of marginal quality, and no crystals in HPT crystal screens. The DisMeta report contains two parts: (A) sequence based bioinformatics prediction for construct design, and (B) disorder predictions from eight different servers. In this case, the disorder prediction programs provide a clear consensus result, namely strong evidence for disorder in the N-terminal region of the protein (red double-head arrow). On the basis of the disorder consensus results and secondary structure prediction, several truncated constructs lacking residues from this region were generated, ultimately leading to the production of Spr(37-162) (green double-headed arrow) whose solution NMR structure was subsequently determined by the NESG consortium (PDB ID, 2K1G) (28).

3.2 Excluding the TMHMM and SignalP regions from construct design

The TMHMM and SignalP regions are often disordered for proteins expressed and purified in bacterial expression systems, although they are ordered based on disorder consensus prediction. For example, NESG target SyR11 (full length 155, PDB ID: 2K3A) (29) is a bacteria putative secretory antigen. Both SignalP_nn and SignalP_hmm methods (16) predict polypeptide segment 1-29 to be a signal peptide. Disorder consensus predicts that the region of residues 1-29 is ordered, and the region around 25-49 is disordered (Fig. 3A). The NMR ^1H - ^{15}N HSQC spectra are shown for the full-length 1-155 protein construct (Fig. 3B) and truncated (residues 50-155) protein construct (Fig. 3C). The truncated (residues 50-155) construct was selected by removing both the signal peptide and the disorder consensus regions. Figure 3D shows the differences of two HSQC spectra (Fig 3B and C). Many overlapping peaks in the full-length protein with chemical shift values typical of disordered residues are absent in the truncated protein construct (Fig 3D). The amide ^{15}N and ^1H resonance frequencies for the remaining ordered residues are identical in both spectra, confirming that the SignalP region 1-29 are disordered and deletion of the disordered region 1-49 does not disturb the structure of the remaining protein.

E. coli inner membrane lipoprotein YiaD (NESG target ER553) has two TMHMM regions: 13-32 and 42-63. The DisMeta disorder consensus predicts that the region around polypeptide segment 1-13 is disordered, and the TMHMM regions 13-32 and 42-63 are ordered (Fig. 4A). HDX-MS analysis (14) shows that region ~1-59 is in fact disordered, which is consistent with the DisMeta prediction by combining both the TMHMM regions and the disorder consensus regions.

3.3 Low complexity SEG regions when used together with the disorder consensus prediction, are good disorder indicators

Studies have shown that sequence regions with low complexity nearly always correspond to polypeptide segments that do not fold into ordered structures, or to regions of proteins that form fibrous or extended structures (39), whereas intrinsically-disordered regions do not always possess low sequence complexity (40,39). Both SEG analysis (18) for complexity and order-disorder prediction are useful and complementary in the analysis of protein sequences (41) and in construct design and optimization.

In our experience, many low complexity SEG regions are next to or overlapped with disorder consensus regions, and those SEG regions next to the disorder consensus regions in the sequence are often indeed disordered. Examples are shown in Figure 3, 4 and 5. In Figure 3A, the low complexity regions are in polypeptide segments 4-21, 32-48 and 73-87. Except for polypeptide segment 73-87, the two other low complexity regions are overlapped with either the SignalP region or the disorder consensus regions. Figure 4A shows that the low complexity region polypeptide segment 40-58 overlaps with the TMHMM region, which is disordered in protein samples produced in our *E. coli* expression system, as shown by the HDX-MS (14). The low complexity region polypeptide segment 202-207 also overlaps with disorder consensus region at the C-terminal.

C. elegans TPPP family protein CE32E8.3 (NESG target WR33) has been solved by the NESG consortium. It consists of five helices with an intrinsically disordered region in the C-terminal one-third of the protein sequence. HDX-MS data (14) (Fig 5B) shows that regions ~1-12 (6xHis tag) and ~121-190 are disordered. Figure 5A shows that polypeptide segments 3-25 and 125-136 are low complexity regions. The region polypeptide segment 3-15 is overlapped with the disorder consensus. The region polypeptide segment 125-136 is very close to the C-terminal disorder consensus. Combining the SEG prediction with the disorder consensus agrees well with the HDX-MS data, which shows that the region of polypeptide segment 121-190 is disordered (Fig 5B) (14). Figure 5C compares the NMR ^1H - ^{15}N HSQC spectra for the full length CE32E8.3 protein 1-190 and the truncated protein construct residues 1-125, that was used for the solution structure determination. This comparison shows the presence of many overlapping peaks in the full-length protein with chemical shift values typical of disordered residues. These peaks are absent in the truncated protein construct. The amide ^{15}N and ^1H resonance frequencies for the remaining ordered residues are identical in both spectra, confirming that deletion of the disordered sequence does not disturb the structure of the remaining protein (14).

3.4 Domain parsing based on the disorder consensus prediction

Multi-domain proteins often misfold in prokaryotic system (15) and also often exceed the size limitations for high-throughput NMR structure determination techniques. Domain parsing can be used to circumvent these significant issues. The data from these disorder consensus of prediction servers, along with multiple sequence alignments of homologous proteins and hidden Markov models characteristic of the targeted protein domain families (22,23) are used to predict possible structural domain boundaries.

Disorder consensus prediction can be used alone to predict the domain boundary in some cases. An example is shown in Fig 6. The C-terminal part of the DisMeta report of Gamma-interferon-inducible protein 16 (NESG target HR4626) from *Homo sapiens* is shown in Fig. 6. There are two ordered regions, corresponding approximately to polypeptide segments 206-385 and 577-768 (green double-headed arrows), linked by a flexible disordered linker. The DisMeta consensus ordered regions are consistent with the domain constructs that provided these 3D structures (i.e. HR4626A, PDB 2OQ0, 200-390 and HR4626B, PDB 3B6Y, 576-761). Indeed, the DisMeta predictions are in better agreement with the

experimentally determined domain parsing than the domain regions polypeptide segments of residues 201-370 and 575-740 predicted by PFAM (42).

3.5 Multiple constructs as a strategy for obtaining samples suitable for structural analysis

It is extremely challenging to predict the protein subsequence that will produce a soluble well-behaved protein, particularly for protein containing domains for which the three-dimensional structure is not yet known. This arises from problems with accurately predicting the domain boundaries and locations of disordered residues, as this information is critical for designing an open reading frame that results in high level expression and solubility in bacterial expression system. Currently, our approach is to take advantage of our high-throughput cloning and expression platform (26,27) and produce several alternative constructs, varying the termini of a targeted domain based on the DisMeta predictions, followed by small scale expression and solubility screening, as well as NMR and crystallization screening, to identify the protein subsequence with the best behavior (24-27).

An example of our domain parsing and multiple construct approach is shown in Fig. 7A. Consensus analysis of several disorder prediction algorithms (see Disorder Consensus panel) suggests that the C-terminal half of the 434-residue protein from *Porphyromonas gingivalis* Q7MX54 (NESG ID: PgR37) contains disordered regions. Cloning and expression analysis of the full-length protein in our bacterial expression system results in no detectable expression, supporting the disorder prediction based on the fact that proteins with significant disorder are often degraded in the *E. coli* cell. DisMeta analysis was then used for construct optimization, and alternative constructs were designed using these data together with information for the database of protein families that includes their annotations and multiple sequence alignments (DUF477). These alternative constructs are depicted schematically in Fig. 7B. The two expression constructs comprised of residues 54–187 and 59–182 also did not express at detectable levels. However, a slightly longer construct (residues 35–182) was highly expressed and soluble and ultimately allowed the structure of this targeted domain to be solved by NMR (Protein Data Bank ID:2KW7). Interestingly, the three-dimensional structure reveals the presence of two short helical regions between residues 38–48 and a β -strand for residues 54–56. These helices and the β -strand are tightly packed against each other and to other regions of the protein. The loss of these interactions in the shorter constructs likely destabilizes the protein, leading to degradation in the expression host (27).

4. Conclusion

The DisMeta server provides a consensus analysis of eight disorder predictors, as well as predictions of secondary structure, signal peptides, trans-membrane helical regions, and low complexity regions of the protein sequence using publicly available servers. These data are combined into a single simple to read report. DisMeta analyses have been used for protein construct design and optimization in the large-scale sample and structure production pipeline of the Northeast Structural Genomics Consortium of the Protein Structure Initiative. These disorder predictions have allowed production of many protein samples that have been used in many successful NMR and X-ray crystallography studies, some examples of which are illustrated in this paper. The server is freely available on line to the scientific community, and should be useful both to small laboratories focused on specific biological problems and

to large-scale protein sample production efforts, including antigen sample production projects.

Acknowledgments

We thank H. Zheng, S. Sharma, A. Ertekin, and R. Xiao for providing the HDX-MS data illustrated in this paper, and J. Aramini for providing the NMR spectrum shown in Fig. 5. The NMR data shown in Fig 3 were recorded by P. Rossi. This work was supported by a grant from the National Institute of General Medical Sciences Protein Structure Initiative U54-GM074958 (to G.T.M.).

References

1. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol.* 2002; 12(1):54–60. [PubMed: 11839490]
2. Iakoucheva LM, Brown CJ, Lawson JD, et al. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology.* 2002; 323(3):573–584. [PubMed: 12381310]
3. Radivojac P, Iakoucheva LM, Oldfield CJ, et al. Intrinsic disorder and functional proteomics. *Biophys J.* 2007; 92(5):1439–1456. [PubMed: 17158572]
4. Kovacs D, Szabo B, Pancsa R, et al. Intrinsically disordered proteins undergo and assist folding transitions in the proteome. *Arch Biochem Biophys.* 2012
5. Liu J, Perumal NB, Oldfield CJ, et al. Intrinsic disorder in transcription factors. *Biochemistry.* 2006; 45(22):6873–6888. [PubMed: 16734424]
6. Minezaki Y, Homma K, Kinjo AR, et al. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *Journal of molecular biology.* 2006; 359(4):1137–1149. [PubMed: 16697407]
7. Balazs A, Csizmek V, Buday L, et al. High levels of structural disorder in scaffold proteins as exemplified by a novel neuronal protein, CASK-interactive protein1. *Febs J.* 2009; 276(14):3744–3756. [PubMed: 19523119]
8. Buday L, Tompa P. Functional classification of scaffold proteins and related molecules. *Febs J.* 2010; 277(21):4348–4355. [PubMed: 20883491]
9. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *Faseb J.* 2004; 18(11):1169–1175. [PubMed: 15284216]
10. Dosztanyi Z, Chen J, Dunker AK, et al. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res.* 2006; 5(11):2985–2995. [PubMed: 17081050]
11. Haynes C, Oldfield CJ, Ji F, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS computational biology.* 2006; 2(8):e100. [PubMed: 16884331]
12. Pantazatos D, Kim JS, Klock HE, et al. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc Natl Acad Sci U S A.* 2004; 101(3):751–756. [PubMed: 14715906]
13. Spraggon G, Pantazatos D, Klock HE, et al. On the use of DXMS to produce more crystallizable proteins: structures of the *T. maritima* proteins TM0160 and TM1171. *Protein Sci.* 2004; 13(12):3187–3199. [PubMed: 15557262]
14. Sharma S, Zheng H, Huang YJ, et al. Construct optimization for protein NMR structure analysis using amide hydrogen/deuterium exchange mass spectrometry. *Proteins.* 2009; 76(4):882–894. [PubMed: 19306341]
15. Netzer WJ, Hartl FU. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature.* 1997; 388(6640):343–349. [PubMed: 9237751]
16. Emanuelsson O, Brunak S, von Heijne G, et al. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2007; 2(4):953–971. [PubMed: 17446895]
17. Krogh A, Larsson B, von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology.* 2001; 305(3):567–580. [PubMed: 11152613]

18. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 1996; 266:554–571. [PubMed: 8743706]
19. Rost B, Yachdav G, Liu J. The PredictProtein server. *Nucleic Acids Res.* 2004; 32(Web Server issue):W321–326. [PubMed: 15215403]
20. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics.* 2000; 16(4):404–405. [PubMed: 10869041]
21. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics.* 2009; 25(20):2745–2746. [PubMed: 19717576]
22. Dessailly BH, Nair R, Jaroszewski L, et al. PSI-2: structural genomics to cover protein domain family space. *Structure.* 2009; 17(6):869–881. [PubMed: 19523904]
23. Hunter S, Jones P, Mitchell A, et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research.* 2012; 40(Database issue):D306–312. [PubMed: 22096229]
24. Graslund S, Sagemark J, Berglund H, et al. The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr Purif.* 2008; 58(2):210–221. [PubMed: 18171622]
25. Chikayama E, Kurotani A, Tanaka T, et al. Mathematical model for empirically optimizing large scale production of soluble protein domains. *BMC Bioinformatics.* 2010; 11:113. [PubMed: 20193068]
26. Xiao R, Anderson S, Aramini J, et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J Struct Biol.* 2010; 172(1):21–33. [PubMed: 20688167]
27. Acton TB, Xiao R, Anderson S, et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods in enzymology.* 2011; 493:21–60. [PubMed: 21371586]
28. Aramini JM, Rossi P, Huang YJ, et al. Solution NMR structure of the NlpC/P60 domain of lipoprotein Spr from *Escherichia coli*: structural evidence for a novel cysteine peptidase catalytic triad. *Biochemistry.* 2008; 47(37):9715–9717. [PubMed: 18715016]
29. Rossi P, Aramini JM, Xiao R, et al. Structural elucidation of the Cys-His-Glu-Asn proteolytic relay in the secreted CHAP domain enzyme from the human pathogen *Staphylococcus saprophyticus*. *Proteins.* 2009; 74(2):515–519. [PubMed: 18951393]
30. Linding R, Jensen LJ, Diella F, et al. Protein disorder prediction: implications for structural proteomics. *Structure.* 2003; 11(11):1453–1459. [PubMed: 14604535]
31. Ward JJ, Sodhi JS, McGuffin LJ, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol.* 2004; 337(3):635–645. [PubMed: 15019783]
32. Cheng J, S M, Baldi P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery.* 2005; 11:213–222.
33. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics.* 2005; 21(16):3435–3438. [PubMed: 15955783]
34. Linding R, Russell RB, Neduva V, et al. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 2003; 31(13):3701–3708. [PubMed: 12824398]
35. Dosztanyi Z, Csizmok V, Tompa P, et al. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol.* 2005; 347(4):827–839. [PubMed: 15769473]
36. Yang ZR, Thomson R, McNeil P, et al. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics.* 2005; 21(16):3369–3376. [PubMed: 15947016]
37. Vucetic S, Brown CJ, Dunker AK, et al. Flavors of protein disorder. *Proteins.* 2003; 52(4):573–584. [PubMed: 12910457]
38. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. *Science.* 1991; 252(5009):1162–1164. [PubMed: 2031185]
39. Romero P, Obradovic Z, Li X, et al. Sequence complexity of disordered protein. *Proteins.* 2001; 42(1):38–48. [PubMed: 11093259]

40. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* 1999; 462(3):363–367. [PubMed: 10622726]
41. Weathers EA, Paulaitis ME, Woolf TB, et al. Insights into protein structure and function from disorder-complexity space. *Proteins.* 2007; 66(1):16–28. [PubMed: 17044059]
42. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic acids research.* 2010; 38(Database issue):D211–222. [PubMed: 19920124]

Dismeta

Disorder Prediction Meta-Server

[Main](#)

The Dismeta server shows consensus results of these protein disorder predictors:

DISEMBL	DISOPRED2	DISpro	DRIPPRED	FoldIndex
FoldUnfold	GlobPlot2	IUPred	RONN	VSL2

It also reports predictions from these sequence analysis tools:

coils	ANCHOR	SignalP	TMHMM	SEG	PROFphd	PSIPred
-----------------------	------------------------	-------------------------	-----------------------	---------------------	-------------------------	-------------------------

All tools use default setup. Tools in [green](#) run on DisMeta server site.

Email address:

Protein name or NESG targetID (no longer than 10 characters):

Sequence for non-NESG target(letters only):

SignalP option - Organism:
☒ gram+
☐ gram-
☐ euk

Figure 1.

The DisMeta webserver interface. The interface is designed to be simple and easy to use. All tools use default setup parameters. The programs DISOPRED2, DISpro, VSL2, coils, SignalP, TMHMM, SEG, PROFphd, PSIPred are installed locally, and run on DisMeta server site. Others results are retrieved over the Internet from the corresponding web servers.

A. Disorder Prediction

B. Disorder Prediction

Figure 2.

The DisMeta report for the *Escherichia coli* Spr lipoprotein, NESG target ER541. The full-length protein provided NMR data of marginal quality, and no crystals in HTP crystallization screens. The DisMeta report contains two parts: (A) sequence based bioinformatics prediction for construct design, and (B) disorder predictions from eight different servers. In this case, the disorder prediction programs provide a clear consensus result, with strong evidence for disorder in the N-terminal region of the protein (red double-headed arrow). On the basis of the disorder consensus results and secondary structure prediction, several truncated constructs lacking residues from this region were generated, ultimately leading to the production of construct Spr(37-162) (green double-headed arrow) whose solution NMR structure was solved in NESG consortium (PDB ID, 2K1G) (28). Sample preparation and NMR data collection for NESG target ER541 are described in reference (28).

Figure 3.

(A) The DisMeta report for NESG target SyR11 (full length 155), a bacterial secretory antigen. Both SignalP_nn and SignalP_hmm methods predict residues 1-29 to be a signal peptide. Disorder consensus predicts that the polypeptide segment region 1-29 is ordered, and the region around 25-49 is disordered. (B) The NMR ^1H - ^{15}N HSQC spectra are shown for the full-length 1-155 protein construct, and (C) truncated (50-155) protein construct. (D) The differences between the two HSQC spectra (B) and (C). Sample preparation and NMR data collection for NESG target SyR11 are described in reference (29).

Figure 4.

(A) The DisMeta report for *E. coli* inner membrane lipoprotein YiaD (NESG target ER553). It has two TMHMM regions: 13-32 and 42-63. The disorder consensus predicts that the polypeptide segment region 1-13 is disordered, and the TMHMM regions 13-32 and 42-63 are ordered. (B) HDX-MS analysis (14) shows that polypeptide segment region ~1-59 is in fact disordered. This is consistent with the DisMeta results combining predictions of the TMHMM regions and the disorder consensus region.

Figure 5.

(A) The DisMeta report for *C. elegans* TPPP family protein CE32E8.3 (NESG target WR33), whose structure has been solved by the NESG consortium (PDB ID 1PUL). It consists of five helices with an intrinsically disordered region in the C-terminal one-third of the protein sequence. The region 3-25 and 125-136 are identified as the SEG regions. (B) The HDX-MS data (14) shows that regions ~1-12 (6×His tag) and ~121-190 are disordered. The region 3-15 is overlapped with the disorder consensus. The region 125-136 is very close to the C-terminal disorder consensus. The combined result of the low complexity SEG regions and the disorder consensus, agrees well with the HDX-MS data, which shows that the polypeptide region 121-190 is disordered (14). (C) Comparison of the NMR ^1H - ^{15}N HSQC spectra for the full length CE32E8.3 protein, residues 1-190, and the truncated protein construct used for the solution structure determination, residues 1-125. This comparison shows the presence of many overlapping peaks in the full-length protein with chemical shift values typical of disordered residues. These peaks are absent in the truncated protein construct. The amide ^{15}N and ^1H resonance frequencies for the remaining ordered residues are identical in both spectra, confirming that deletion of the disordered sequence does not disturb the structure of the remaining protein. Sample preparation, NMR data collection, and HDX-MS data collection for NESG target WR33 are described in reference (14).

Figure 6. Domain parsing based on disorder consensus prediction

The C-terminal part of the DisMeta report of Gamma-interferon-inducible protein 16 (NESG target HR4626) from *Homo sapiens* is shown. There are two ordered regions, corresponding to residues ~206-385 and 577-768 (green double-headed arrows), linked by a flexible disordered linker. The DisMeta consensus ordered regions are consistent with the domain constructs for which structures were determined and deposited in the PDB (i.e. HR4626A, PDB 2OQ0, 200-390 and HR4626B, PDB 3B6Y, 576-761). Indeed, the ordered regions identified by DisMeta have better agreement with the domain boundaries revealed by these experimental structures than the domains predicted by PFAM (42), polypeptide segments 201-370 and 575-740.

A. Disorder Prediction

B. Alternative Constructs

Figure 7.
(A) The DisMeta report for the *Porphyromonas gingivalis* protein Q7MX54 (NESG ID: PgR37). (B) Schematic representation of construct optimization for NESG protein target PgR37 from PFAM domain family DUF477, including full-length, residues 54–187, residues 59–182, and residues 35–182. Only the 35–182 construct produced a soluble-expressed protein and ultimately, an NMR structure (Protein Data Bank ID: 2KW7).