Disorder in Proteins: Functional Lack of Structure

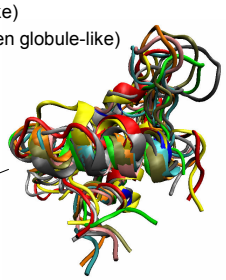
Lukasz Kurgan

Outline

- Disorder
  - definition
  - abundance
  - functional role and localization
- Prediction of disorder
  - benchmarking of disorder predictors
  - prediction of disorder-to-order transitioning protein-protein binding regions (MoRFpred)
- Applications of disorder predictions
  - disorder in nucleosome

Definition

- Intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs) exist as dynamic conformational ensembles which can be
  - collapsed-disordered (molten globule-like)
  - partially collapsed-disordered (pre-molten globule-like)
  - extended disordered (coil-like)



NMR solution structures (10 conformations) of the globular domain (residues 41–113) of the *G. gallus* (chicken) histon H1

Peng Z, Mizianty M, Xue B, Kurgan L, Uversky V. *Molecular BioSystems* 2012, 8:1886-1901

Abundance of disorder

- DisProt
  - curated database of experimentally verified disorder
  - current release 5.9: 653 proteins and 1421 disordered regions
- Disorder is abundant and (relatively) little explored
  - functions of disorder were studied in human and yeast proteomes
    - involved in regulation of transcription, cell signalling, kinase activity, and nucleic acid binding

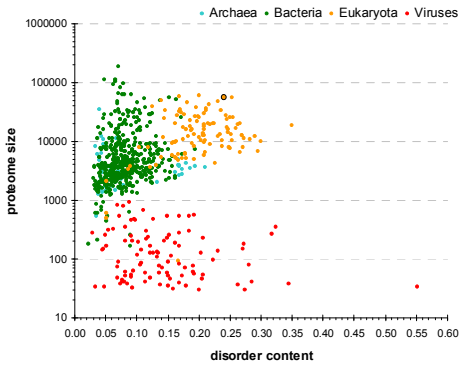
Kingdom	# sequences	% disorder
Archaea (6 species)	11,742	3.8
Bacteria (13 species)	35,389	5.7
Eukaryota (5 species)	88,531	18.9
PDB (non-redundant at 95% sequence identity)	7,169	3.2

Sickmeier M, et al. *Nucleic Acids Res.* 2007; 35:D786-93
Ward JJ, et al. *J Mol Biol.* 2004; 337(3):635-45
Lobley A, et al. *PLoS Comput Biol.* 2007; 3(8):e162
Pentony MM and Jones DT. *Proteins* 2010; 78(1):212-21

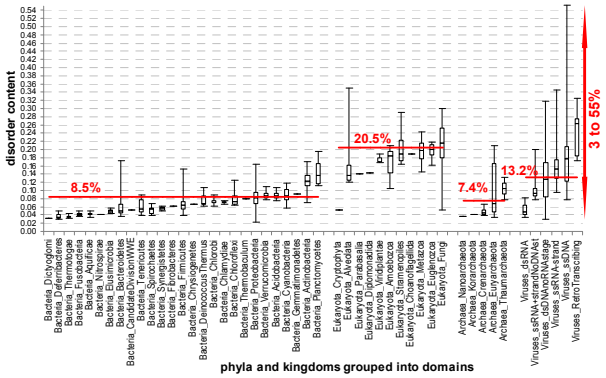
Abundance of disorder

- The setup
  - 965 complete proteomes and 6.4+ million proteins
    - 3.6% proteins from 59 species in archaea
    - 66.6% proteins from 471 species in bacteria
    - 29.5% proteins from 110 species in eukaryota
    - 0.3% proteins from 325 viral proteomes
  - consensus of 5 disorder predictors
    - compared to one predictor used in prior studies
  - analysis of
    - generic disorder characteristics
      - abundance and distribution
      - relation to evolution
    - functional analysis of disorder
      - disorder and post-translational modifications (PTMs) based on UniProt
      - disorder and protein functions based on Gene Ontology (GO)
      - disorder in cellular components based on GO

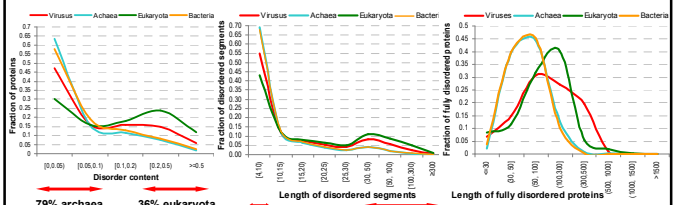
Abundance of disorder



Abundance of disorder



Abundance of disorder



79% archaea
77% bacteria
63% viruses
46% eukaryota

36% eukaryota
21% viruses

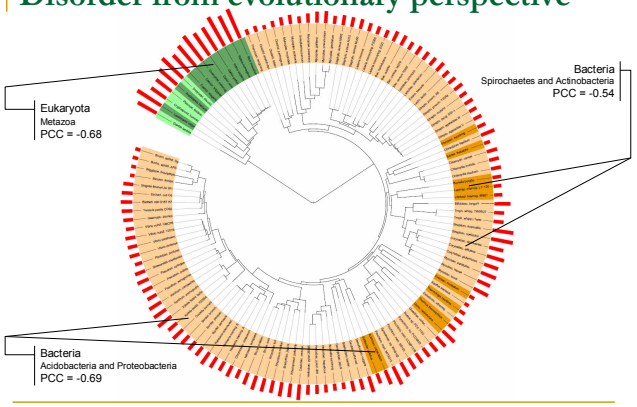
69% archaea
67% bacteria
55% viruses
43% eukaryota

25% eukaryota
16% viruses

89% bacteria
86% archaea
53% viruses
52% eukaryota

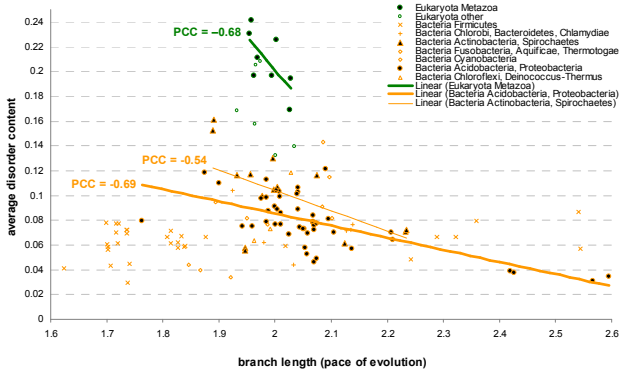
20% viruses
8% eukaryota

Disorder from evolutionary perspective



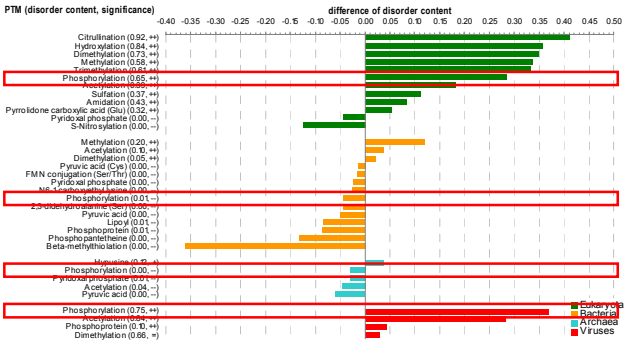
Ciccarelli FD, et al. *Science* 2006, 311(5765):1283-7

Disorder from evolutionary perspective



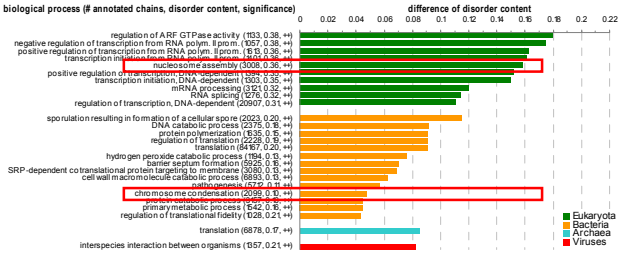
Functional role and localization

- Disorder in PTM sites



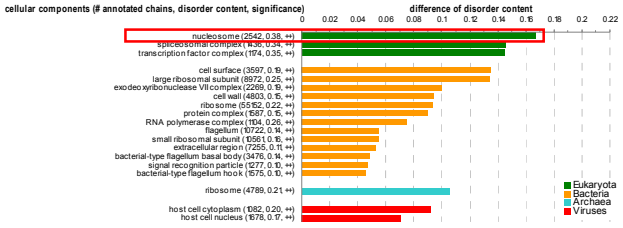
Functional role and localization

		Archaea	Bacteria	Eukaryota	Viruses
biological processes	total # of processes	12	318	104	2
	# of processes enriched in disorder	1	14	10	1



Functional role and localization

		Archaea	Bacteria	Eukaryota	Viruses
cellular components	total # of components	6	61	50	5
	# of components enriched in disorder	1	13	3	2



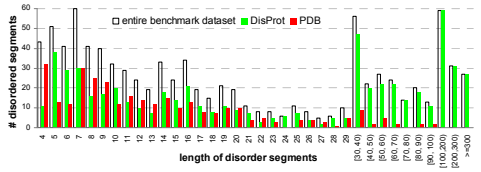
Prediction of disorder

- disorder annotations find a wide range of applications
  - structural and functional studies of certain protein families
  - proteome-scale analysis
- disorder prediction enjoys a relatively strong interest
  - 40+ (non-incidental) methods
  - included in CASP since 2002
    - 32 groups in CASP9 (22 servers and 10 human-expert groups)
    - relatively large room for further improvements
    - lack of improvement since CASP8
    - new breakthroughs are needed
    - lack of data with functionally relevant long disordered segments
    - run currently at CASP10

Monastyrskyy et al. *Proteins* 2011
He et al. *Cell Res.* 2009

Benchmarking prediction of disorder

- Large(r) scale assessment of (some/most of) end-user accessible disorder predictors
  - new dimensions: disordered segment, prediction of chains with long disordered segments, prediction vs. protein size, disorder content
- Benchmark dataset with 494 proteins
  - 289 from DisProt using DisProt annotations and alignment into PDB structures to transfer structure-based annotations + 205 collected from PDB and annotated analogously to CASP



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

- 18 predictors (21 including sub-versions)
  - must be available as a web server or/and standalone implementation

Prediction method			in	URL	
Name (year published)		Algorithm used	CASP9	standalone program (SP); web server (WS)	
MFDp	2010	Support vector machine	X	WS	http://biomine-ws.ece.ualberta.ca/MFDp.html
PONDR-FIT	2010	Neural network		WS	http://www.disprot.org/predictors.php
MULTICOM/PreDisorder	2009	Neural network	X	SP+WS	http://casp.rnet.missouri.edu/predisorder.html
MD	2009	Neural network		SP+WS	https://rostlab.org/owiki/index.php/Metadisorder
DISOCLUST	2008	Scoring function	X	SP+WS	http://www.reading.ac.uk/bioinf/DISOclust/
PrDOS	2007	SVM + templates	X	WS	http://prdos.hgc.jp/cgi-bin/top.cgi
Norsnet	2007	Neural network		SP+WS	https://rostlab.org/owiki/index.php/Norsnet
Ucon	2007	Scoring function		WS	https://rostlab.org/owiki/index.php/UCON
ProfBval	2006	Neural network		SP+WS	https://rostlab.org/owiki/index.php/Profbval
Spritz	2006	Support vector machine	X	WS	http://distill.ucd.ie/spritz/
VSL2B	2006	Support vector machine		SP+WS	http://www.ist.temple.edu/disprot/Predictors.html
DISpro	2005	Neural network		SP+WS	http://scratch.proteomics.ics.uci.edu/
FoldIndex	2005	Scoring function		WS	http://bioportal.weizmann.ac.il/fldbin/findex
IUPred (2 versions)	2005	Scoring function		SP+WS	http://iupred.enzim.hu/
RONN	2005	Neural network		SP+WS	http://www.strubi.ox.ac.uk/RONN
DISOPRED2	2004	Support vector machine	X	SP+WS	http://bioinf.cs.ucl.ac.uk/disopred/
DisEMBL (3 versions)	2003	Neural network		SP+WS	http://dis.embl.de/
GlobPlot	2003	Scoring function		SP+WS	http://globplot.embl.de/

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

Predictor	Disorder prediction per-residue					Disorder prediction per-segment
	Sw	AUC	SENS	SPEC	MCC	SOV
MFDp	0.504	**0.815**	0.732	0.772	**0.45**	61.0
MD	0.473	**0.814**	0.657	0.816	**0.44**	45.0
VSL2B	0.469	0.790	0.766	0.704	0.41	**62.9**
PreDisorder	0.458	0.789	0.724	0.733	0.40	58.4
PONDR-FIT	0.441	0.786	0.617	0.824	0.42	54.5
PrDOS	0.409	0.785	0.563	0.846	0.40	53.2
DISOPRED2	0.444	0.781	0.639	0.804	0.41	50.3
DISpro	0.239	0.776	0.297	**0.942**	0.32	31.7
IUPredL	0.409	0.775	0.566	0.843	0.40	32.9
DISOCLUST	0.411	0.774	0.774	0.637	0.35	59.5
IUPredS	0.378	0.774	0.510	0.868	0.39	45.8
RONN	0.413	0.762	0.655	0.758	0.37	50.3
Norsnet	0.354	0.737	0.522	0.832	0.35	20.7
Ucon	0.330	0.736	0.541	0.788	0.31	26.0
Profbval	0.223	0.695	**0.833**	0.39	0.20	48.4
FoldIndex	0.312	NA	0.592	0.719	0.28	36.7
SPRITZ	0.310	NA	0.493	0.817	0.30	36.9
DisEMBL_R465	0.244	NA	0.306	0.938	0.32	34.9
DisEMBL_HL	0.222	NA	0.428	0.793	0.22	44.4
GlobPlot	0.181	NA	0.352	0.829	0.19	33.3
DisEMBL_LR	0.177	NA	0.760	0.417	0.16	54.1

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

Predictor	Disorder prediction per-residue					Disorder prediction per-segment
	Sw	AUC	SENS	SPEC	MCC	SOV
MFDp	0.504	0.815	0.732	0.772	0.45	61.0
MD	0.473	0.814	0.657	0.816	0.44	45.0
VSL2B	0.469	0.790	0.766	0.704	0.41	62.9
PreDisorder	0.458	0.789	0.724	0.733	0.40	58.4
PONDR-FIT	0.441	0.786	0.617	0.824	0.42	54.5
PrDOS	0.409	0.785	0.563	0.846	0.40	53.2
DISOPRED2	0.444	**0.781**	0.639	0.804	**0.41**	**50.3**
DISpro	0.239	**0.776**	0.297	0.942	**0.32**	**31.7**
IUPredL	0.409	**0.775**	0.566	0.843	**0.40**	**32.9**
DISOCLUST	0.411	**0.774**	0.774	0.637	**0.35**	**59.5**
IUPredS	0.378	**0.774**	0.510	0.868	**0.39**	**45.8**
RONN	0.413	0.762	0.655	0.758	0.37	50.3
Norsnet	0.354	0.737	0.522	0.832	0.35	20.7
Ucon	0.330	0.736	0.541	0.788	0.31	26.0
Profbval	0.223	0.695	0.833	0.39	0.20	48.4
FoldIndex	0.312	NA	0.592	0.719	0.28	36.7
SPRITZ	0.310	NA	0.493	0.817	0.30	36.9
DisEMBL_R465	0.244	NA	0.306	0.938	0.32	34.9
DisEMBL_HL	0.222	NA	0.428	0.793	0.22	44.4
GlobPlot	0.181	NA	0.352	0.829	0.19	33.3
DisEMBL_LR	0.177	NA	0.76	0.417	0.16	54.1

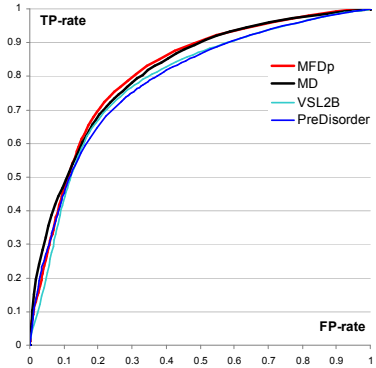
Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

Predictor	Disorder prediction per-residue					Disorder prediction per-segment	CASP9	
	Sw	AUC	SENS	SPEC	MCC	SOV	MCC	AUC
MFDp	0.504	**0.815**	0.732	0.772	**0.45**	61.0	**2 (0.46)**	**7 (0.82)**
MD	0.473	0.814	0.657	0.816	0.44	45.0		
VSL2B	0.469	0.790	0.766	0.704	0.41	62.9		
PreDisorder	0.458	**0.789**	0.724	0.733	**0.40**	58.4	**5 (0.41)**	**3 (0.85)**
PONDR-FIT	0.441	0.786	0.617	0.824	0.42	54.5		
PrDOS	0.409	**0.785**	0.563	0.846	**0.40**	53.2	**3 (0.42)**	**1 (0.85)**
DISOPRED2	0.444	**0.781**	0.639	0.804	**0.41**	50.3	**1 (0.51)**	**2 (0.85)**
DISpro	0.239	0.776	0.297	0.942	0.32	31.7		
IUPredL	0.409	0.775	0.566	0.843	0.40	32.9		
DISOCLUST	0.411	**0.774**	0.774	0.637	**0.35**	59.5	**12 (0.34)**	**9 (0.82)**
IUPredS	0.378	0.774	0.510	0.868	0.39	45.8		
RONN	0.413	0.762	0.655	0.758	0.37	50.3		
Norsnet	0.354	0.737	0.522	0.832	0.35	20.7		
Ucon	0.330	0.736	0.541	0.788	0.31	26.0		
Profbval	0.223	0.695	0.833	0.39	0.20	48.4		
FoldIndex	0.312	NA	0.592	0.719	0.28	36.7		
SPRITZ	0.310	NA	0.493	0.817	**0.30**	36.9	**16 (0.33)**	**21 (0.75)**
DisEMBL_R465	0.244	NA	0.306	0.938	0.32	34.9		
DisEMBL_HL	0.222	NA	0.428	0.793	0.22	44.4		
GlobPlot	0.181	NA	0.352	0.829	0.19	33.3		
DisEMBL_LR	0.177	NA	0.76	0.417	0.16	54.1		

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

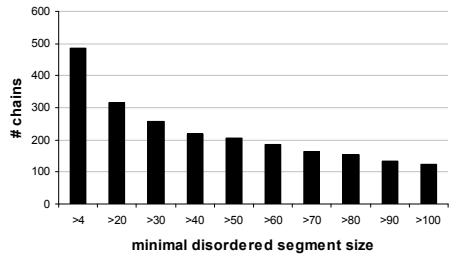
Benchmarking prediction of disorder



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

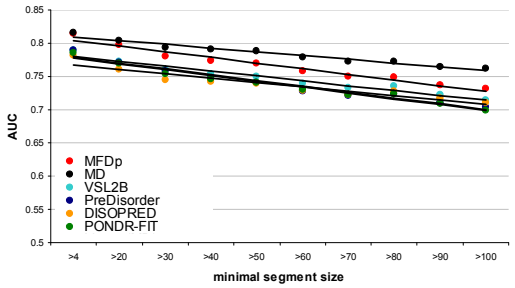
- proteins with disordered segments > threshold
  - similar to analysis in CAP9 but includes longer segments



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

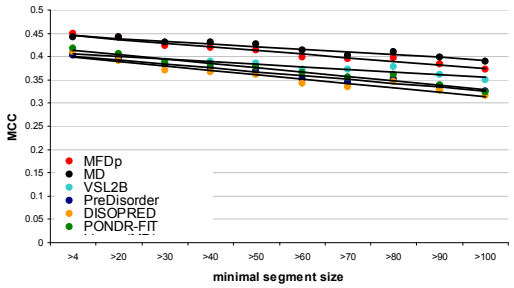
- proteins with disordered segments > threshold



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

- proteins with disordered segments > threshold



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

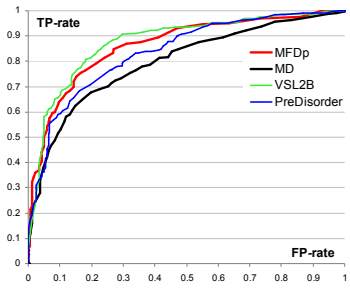
Benchmarking prediction of disorder

Predictor	Prediction of proteins with long (>30 residues) disordered segments			
	MCC	SENS	SPEC	AUC
VSL2B	**0.615**	0.884	0.720	**0.878**
MFDp	0.554	0.818	0.733	0.863
PONDR-FIT	0.580	0.702	0.873	0.857
Ucon	0.505	0.496	0.958	0.849
DISpro	0.411	0.368	**0.966**	0.846
PrDos	0.519	0.578	0.915	0.844
predisorder	0.505	0.764	0.742	0.837
RONN	0.545	0.810	0.733	0.837
IUPredL	0.497	0.570	0.903	0.827
DISOPRED2	0.536	0.694	0.839	0.826
IUPredS	0.511	0.609	0.886	0.826
DISOCLUST	0.488	0.833	0.644	0.824
Norsnet	0.488	0.496	0.945	0.804
MD	0.483	0.651	0.826	0.796
Profbval	0.399	0.721	0.678	0.764
FoldIndex	0.364	0.868	0.462	NA
SPRITZ	0.409	0.411	0.941	NA
DisEMBL_HL	0.381	0.601	0.775	NA
GlobPlot	0.365	0.457	0.877	NA
DisEMBL_R465	0.357	0.337	0.949	NA
DisEMBL_LR	0.149	**0.930**	0.165	NA

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

- predictions of proteins with disordered segments > 30

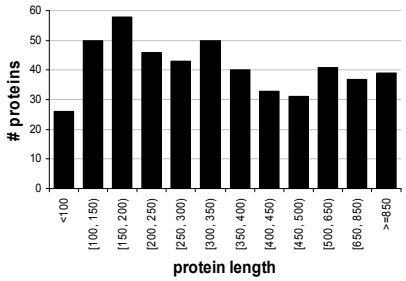


we can find that a given chain has long disordered segments, but exact location and size of these segments may not be very accurate

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

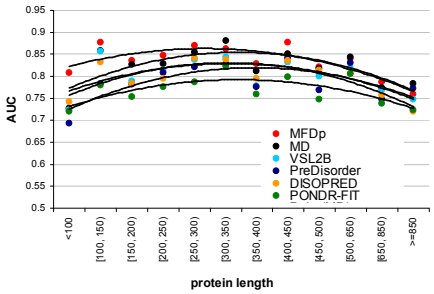
- predictive quality vs. protein size



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

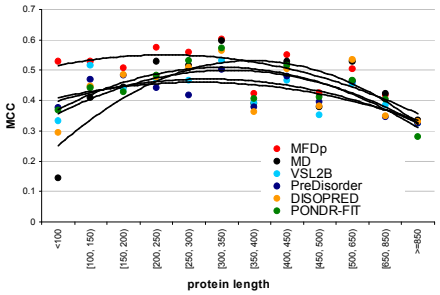
- predictive quality vs. protein size



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

- predictive quality vs. protein size



Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

Benchmarking prediction of disorder

Predictor	Prediction of disorder content		
	MSE	MAE	PCC
Ucon	**0.057**	0.163	**0.62**
PONDR-FIT	**0.058**	**0.156**	**0.62**
IUPredS	0.063	**0.153**	0.58
RONN	0.067	0.182	0.55
MFDp	0.068	0.165	**0.62**
predisorder	0.068	0.194	0.60
DISOPRED2	0.070	0.162	0.55
PrDos	0.072	0.159	0.51
VSL2B	0.073	0.201	0.61
IUPredL	0.074	0.168	0.55
DisEMBL_HL	0.074	0.197	0.45
MD	0.085	0.193	0.60
FoldIndex	0.087	0.222	0.50
DisEMBL_R465	0.090	0.179	0.51
GlobPlot	0.091	0.200	0.29
DISpro	0.095	0.186	0.50
SPRITZ	0.096	0.185	0.29
Norsnet	0.098	0.194	0.46
DISOCLUST	0.101	0.250	0.53
DisEMBL_LR	0.217	0.420	0.22
Profbval	0.220	0.428	0.35

Peng Z, Kurgan L, *Current Protein and Peptide Science* 2012, 13:6-18

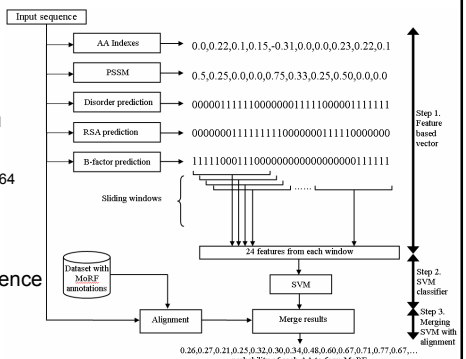
MoRFpred

- Molecular recognition features (MoRFs)
  - short (5-25 amino acids) regions located within longer intrinsically disordered regions that bind to protein partners via disorder-to-order transitions
  - relatively common, particularly in eukaryotes
  - enriched in proteins with regulatory and signal transduction functions
- Lack of high-quality predictive models
  - α-MoRF-PredII (Cheng et al., 2007)
  - ANCHOR (Mészáros et al., 2009)

Mészáros B, et al. *PLoS Comput. Biol* 2009; 5:e1000376
Cheng Y, et al. *Biochemistry* 2007; 46(47):13468-77
Mohan A, et al. *J Mol Biol*. 2006; 362(5):1043-59

MoRFpred

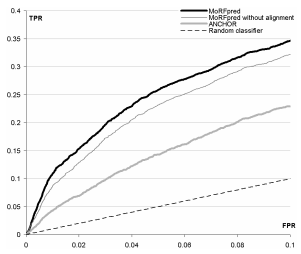
- a new and large MoRF dataset
- novel design:
  - custom-designed and new inputs
  - 24 features selected from 1764
- Support Vector Machine (SVM) combined with alignment/sequence similarity



Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MoRFpred

- tested on independent (low similarity) proteins



Predictor	TPR	FPR	Success rate		AUC	
α-MoRF-PredI	0.12	0.04	0.16	++	NA	NA
α-MoRF-PredII	0.26	0.10	0.30	++	NA	NA
ANCHOR	0.39	0.25	0.61	++	0.60	++
MoRFpred	0.25	0.05	0.72		0.67	
MoRFpred (to match the highest TPR)	0.39	0.14	0.72		0.67	
MoRFpred (to match the lowest FPR)	0.22	0.04	0.72		0.67	

Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MoRFpred

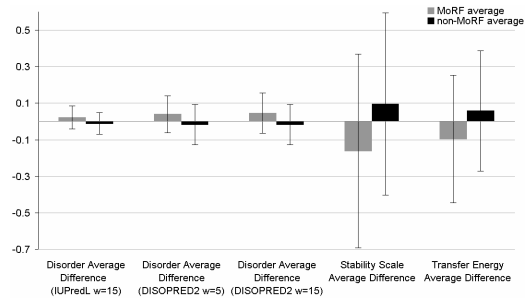
- more tests on new and new-experimental data

Dataset	Predictor	TPR	FPR	Success rate	AUC
TEST 2012	MoRFpred	0.236	0.045	0.756	0.697
	MD	0.613	0.436	0.578	0.679
	ANCHOR	0.433	0.236	0.578	0.638
	IUPredS	0.449	0.287	0.600	0.634
	IUPredL	0.572	0.382	0.600	0.620
	MFDp	0.754	0.556	0.556	0.620
	Spine-D	0.720	0.522	0.467	0.605
	DISOPRED2	0.534	0.455	0.244	0.548
	DISOclust	0.653	0.593	0.556	0.512
	α-MoRF-PredI	0.091	0.030	0.133	NA
	α-MoRF-PredII	0.291	0.096	0.311	NA
EXPER 2008-12	MoRFpred	0.210	0.077	0.750	0.636
	MD	0.690	0.702	0.500	0.616
	ANCHOR	0.548	0.481	0.500	0.556
	IUPredL	0.724	0.714	0.375	0.471
	IUPredS	0.486	0.554	0.250	0.451
	MFDp	0.919	0.839	0.500	0.337
	Spine-D	0.710	0.783	0.250	0.330
	DISOPRED2	0.481	0.720	0.125	0.310
	DISOclust	0.581	0.791	0.250	0.290
	α-MoRF-PredI	0.000	0.069	0.000	NA
	α-MoRF-PredII	0.238	0.161	0.250	NA

Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MoRFpred

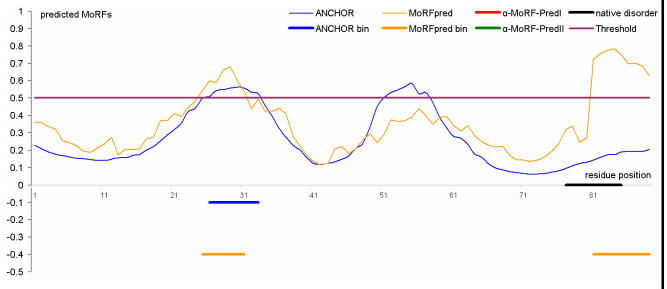
- "smart" use of ML enables development a strong predictor in spite of weak predictive inputs



Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MoRFpred

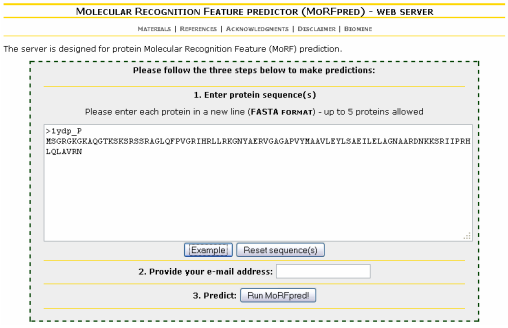
- H2A class histone protein (native MoRF region folds into coil)



Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MoRFpred

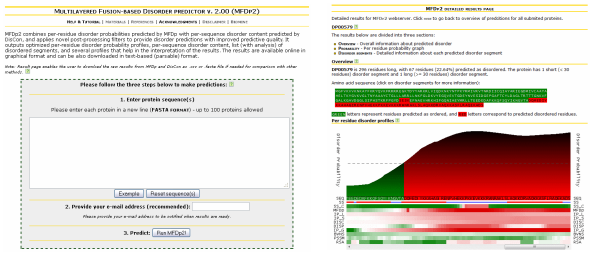
- http://biomine.ece.ualberta.ca/MoRFpred/



Miri Disfani F, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. *Bioinformatics* 2012 (ISMB'2012), 28(12):i75-i83

MFDp

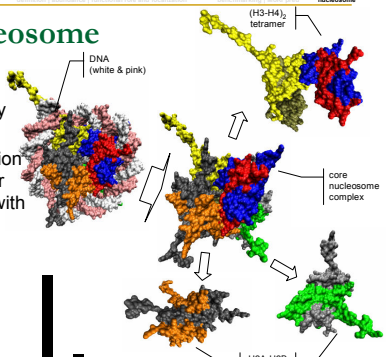
- http://biomine.ece.ualberta.ca/MFDp2/
  - visited (Jan 2011 to May 2012) from 43 countries (top 5: USA, Canada, France, China, and UK) and 165 cities
  - ranked 2nd (among 32 participants) in binary disorder prediction during the most recent CASP9 experiment

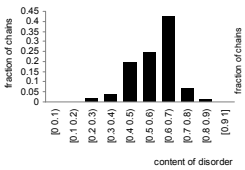


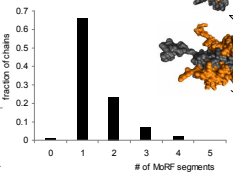
Monastyrskyy B, et al., *Proteins* 2011; 79(S10):107-118
Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Miri Disfani F, Kurgan L. *Bioinformatics* 2010(ECCB'2010), 26(18):i489-96

Disorder in nucleosome

- 2007 histones; 746 species
- all members of histone family are intrinsically disordered
- plays role in heterodimerization and formation of higher order oligomers, and interactions with DNA and other proteins
- is highly conserved



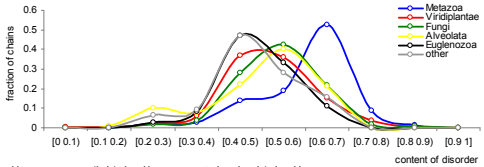




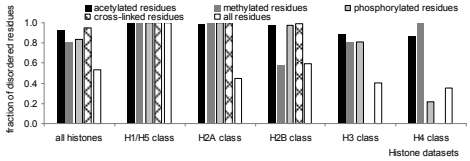
Peng Z, Mizianty M, Xue B, Kurgan L, Uversky V. *Molecular BioSystems* 2012, 8:1886-1901

Disorder in nucleosome

- is highly abundant across different phyla from Eukaryotes



- is enriched in PTM sites



Peng Z, Mizianty M, Xue B, Kurgan L, Uversky V. *Molecular BioSystems* 2012, 8:1886-1901

Acknowledgements

People

Zhenling Peng	Ph.D. student
Marcin Mizianty	Ph.D. student
Jing Yan	Ph.D. student
Dr. Keith A. Dunker	Indiana University
Dr. Vladimir Uversky	University of South Florida

Funding

NSERC CRSNG
Natural Sciences and Engineering Research Council of Canada

Alberta Innovates Technology Futures