

**NEW METHODS TO STUDY PROLINE-RICH
DISORDERED REGIONS AND THEIR STRUCTURAL
ENSEMBLES IN PROTEIN SIGNALING PATHWAYS**

LIU CHENGCHENG
(B.Sci. (Hons), NUS)

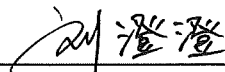
**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTATION AND SYSTEMS BIOLOGY
(CSB)
SINGAPORE-MIT ALLIANCE
NATIONAL UNIVERSITY OF SINGAPORE**

2012

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



Liu Chengcheng

24 August 2012

Acknowledgments

I would like to particularly thank my parents, who have given their full support in my entire undergraduate and graduate studies.

I am very grateful to my two thesis supervisors, Christopher Hogue and Michael Yaffe, both of whom gave me great inspiration and motivation in my research topic. I am impressed with Chris' novel and interesting insights in research. I deeply thank Chris for all the kind guidance, suggestions, effort and help throughout my entire PhD candidature, without which I could not have learnt and achieved so many meaningful things in this significant phase of my life. I truly give my thanks to Mike for his dedicated supervision and encouragement especially during my exchange at MIT. I would like to thank my qualifying examination committee members, Boon Chuan Low, Steve Rosen, Jianzhu Chen, who gave me great suggestions and advice in my thesis project. I also want to thank other SMA faculty, including Zhiyuan Gong, Chwee Teck Lim, Jie Yan and Sourav Saha Bhowmick, for their help and support. I thank for the encouragement from Lisa Tucker-Kellogg, Hanry Yu, and Yuzong Chen when I felt depressed in my study.

The work about molecular simulation of LRP6 intracellular domain in Chapter 1 of this thesis received inspiration about the simulation study of protein ActA, which was conducted by Mingxi Yao, a member of Hogue Lab and a graduate student in Mechanobiology Institute, Singapore. I thank Mingxi for all the helpful discussions and suggestions.

In Chapter 2 of this thesis, the work received from the help of Sihan Liu, a former SMA PhD candidate, and I thank him for the technical support.

The work in Chapter 3 was conducted in collaboration with Brian Joughin, a member of Yaffe Lab and a Research Scientist in David H. Koch Institute for Integrative Cancer Research at MIT. I am very grateful to Brian for his unique insights in the study of kinase-substrate specificity and other topics in computational biology.

Additionally, I would like to sincerely thank Narendra Suhas Jagannathan, Arun Chandramohan, Chen Zhao, Wenwei Xiang as well as other members in the Hogue Lab for their useful discussions. Furthermore, I extend my gratitude to the members in Yaffe lab, Dan Lim, Kylie Huang, Erik Wilker and so on, for their kind help when I was at MIT.

I had all the fun and joy with my fellow SMA-CSB classmates, Yingting Wu, Yujing Liu, Lu Huang, Huipeng Li, Lingbo Zhang and others.

Finally, I thank the financial support from Singapore-MIT Alliance and Mechanobiology Institute, Singapore.

Table of Contents

1	Introduction.....	1
2	The Effect of Spatial Constraints on An Ensemble of Proline-Rich Disordered Structures.....	41
2.1	Background	42
2.2	Results	47
2.2.1	LRP6 intracellular domain is predicted to be unfolded.	47
2.2.2	Radius of gyration distribution	47
2.2.3	End-to-end distance distribution.....	54
2.3	Discussion	58
2.3.1	LRP6 intracellular domain structure ensemble favors an elongated form when the Wnt/ β -catenin canonical pathway initiates.	58
2.3.2	Effects of the two spatial constraints	61
2.3.3	Elongation makes the phosphorylation of unfolded protein regions easier.....	64
2.4	Conclusions	69
2.5	Methods.....	71
2.5.1	Generation of conformers of LRP6 intracellular domain	71
2.5.2	Filtration of structural ensemble of LRP6 intracellular domain.....	72
2.5.3	Measurement	75
2.5.4	The Rgyr distribution and end-to-end distance distribution.....	76
2.5.5	Control experiment.....	76
2.5.6	Program development.....	77
2.5.7	Simulation procedure using structure [PDB:1CMK]	78
2.6	Acknowledgements	80
2.7	Author's Contributions.....	80
3	Sequence Detection of Proline/Serine-Rich Disordered Regions.....	81
3.1	Background	82
3.2	Implementation.....	85
3.2.1	Pro/Ser-rich disorder dataset	85
3.2.2	Third party datasets	86
3.2.3	The PSR index	87
3.2.4	Pro/Ser-rich disorder prediction	89
3.2.5	Prediction performance measures.....	89

3.2.6 Armadillo (2.0)	90
3.3 Results and Discussion.....	90
3.3.1 Amino acid composition in the datasets	90
3.3.2 Evaluation of Pro/Ser-rich disorder predictions	96
3.3.3 Server prediction examples	99
3.4 Conclusions	102
3.5 Author's Contributions.....	102
4 Sequence Analysis of Interpositional Dependence in Phosphorylation Motifs	103
4.1 Background	104
4.2 Results	108
4.2.1 Statistical significance of interpositional dependencies among kinase phosphorylation motifs.....	108
4.2.2 Incorporation of interpositional dependencies in predicting novel kinase phosphorylation sites	112
4.3 Discussion	120
4.4 Conclusion.....	125
4.5 Methods.....	126
4.5.1 Data sources.....	126
4.5.2 Data preparation	126
4.5.3 Simplified amino acid alphabet	128
4.5.4 Statistical analysis of enriched and reduced amino acid pairs.....	128
4.5.5 Statistical significance cutoff determination	131
4.5.6 First and second-order model prediction	132
4.5.7 Evaluation of first-and second-order models.....	133
4.6 Acknowledgement.....	136
4.7 Author's Contributions.....	136
5 Conclusions and Future Directions	137

Summary

In signaling and mechano-related pathways, a type of protein domain is critical for transducing signals. Such protein domains are located in the termini or flanked by folded domains, compositionally biased with prolines preventing folding into a single stable conformation. They are referred to as proline-rich disordered protein regions. This thesis presents a couple of new methods using molecular simulation, bioinformatics and statistical analysis to study the structural ensemble and sequences of proline-rich disordered regions. A new approach, involving simulating the membrane or nearby molecular assembly in the cellular context as simple planes in the conformational space of disordered protein regions, is described in the sampling structural ensembles of proline-rich disordered LRP6 intracellular domain in the initiation of Wnt/ β -catenin pathway. The new simulation approach shows that an elongated form dominates the conformational space of such proline-rich disordered regions when assembled with membranes or neighbor molecules that impose excluded volume constraints. A new amino acid propensity index called PSR is derived from a set of folded domains and a set of proline/serine-rich disordered regions. This index is used to predict long proline-rich disordered regions containing multiple serines, which could serve as phosphoacceptors in signaling pathways. New statistical analysis was done to further study the kinase-substrate specificity for kinases ATM/ATR, CDK1 and CK2, by including the second-order interpositional sequence dependence in the substrate phosphorylation peptides. The findings show that sequence alone is not sufficient to improve the accuracy of phosphorylation sites prediction for the kinases studied; instead, other parameters, especially co-localization,

surface accessibility etc, are required to be considered. This study can be extended to other kinases.

List of Tables

Table 1.1: Experimental methods for characterizing intrinsically disordered proteins.....	6
Table 1.2: A list of current disorder predictors with available URL and brief description.....	8
Table 1.3: Modular domains, phosphopeptide-binding domains and their specificities.	28
Table 1.4: Proline-rich regions with repeated proline-rich motifs.....	29
Table 1.5: Proline-rich regions without repeated proline-rich motifs.....	30
Table 2.1: Rgyr simulation results for LRP6 intracellular domain.....	52
Table 2.2: Rgyr simulation results for control sequence.	52
Table 2.3: End-to-end distance simulation results for LRP6 intracellular domain.	55
Table 2.4: T-test results on the constructed 100mer peptide.	69
Table 3.1: Calculated frequencies of amino acid residues in Pro/Ser-rich disorder dataset and MMDB-I domain dataset as well as the negative and normalized log ratios for PSR index.	88
Table 3.2: Amino acid composition difference in percentage between MMDB-I domain dataset and disordered protein segments in DisProt (v5.8).	92
Table 3.3: Amino acid composition difference in percentage between MMDB-I domain dataset and the curated Pro/Ser-rich disorder dataset from literature.	93
Table 3.4: Amino acid composition difference in percentage between MMDB-I linker dataset and disordered protein segments in DisProt (v5.8).	94
Table 3.5: Pro/Ser-rich disorder predictions.....	98
Table 4.1: A list of current phosphorylation site predictors.	107
Table 4.2: Substrate sequence position pairs demonstrating significant deviations from independence.	111

List of Figures

Figure 1.1: The protein sequence-structure-function paradigm.....	3
Figure 2.1: Two proposed initiation models of canonical Wnt/ β -catenin signalling pathways.....	44
Figure 2.2: Analysis of the human LRP6 protein [Swiss-Prot:O75581] using different predictors.....	49
Figure 2.3: Rgyr distribution of the initial conformational ensemble before filtration.....	50
Figure 2.4: Rgyr distributions of LRP6 ICD and control sequence.....	53
Figure 2.5: End-to-end distance distributions of D1 for LRP6 ICD and control sequence.....	56
Figure 2.6: End-to-end distance distributions of D2, D3, D4 and D5 for LRP6 ICD and control sequence.....	57
Figure 2.7: Simulation results from the study on structure [PDB:1CMK].....	67
Figure 2.8: Rgyr and end-to-end distance distributions of D1-40, D31-70 and D61-100 for the constructed 100mer alternating Pro/Ser peptide with substrate phosphorylation motif in the centre.....	68
Figure 2.9: Flow chart of the simulation process on LRP6 intracellular domain.....	78
Figure 3.1: Amino acid compositions of the datasets.....	95
Figure 3.2: Armadillo (2.0) Pro/Ser-rich disorder predictions for human proteins LRP6, WASP and MAP tau isoform 2.....	101
Figure 4.1: Comparison of ability of first- and second- order models to identify kinase substrates.....	118
Figure 4.2: Comparison of ability of first- and second- order models to correctly identify true positives, correcting for occurrence of amino acid pairs not present among training data.....	119
Figure 4.3: Model evolutionary fitness landscapes for substrates of kinases and phosphopeptide-binding domains.....	124
Figure 4.4: Data source and data preparation.....	127
Figure 4.5: Motif logos for substrates analyzed.....	129
Figure 4.6: ROC curves detail variation of true and false positive rates with probability score.....	135

List of Illustrations

Illustration 1.1: An illustration of energy landscape models for globular/folded proteins and intrinsically disordered/unfolded proteins.....	15
Illustration 2.1: Illustration of the spatial constraints.	74
Illustration 4.1: An illustration of statistical hypothesis testing as applied in this analysis.....	130

List of Symbols

R_g	Radius of Gyration	(Å)
r_k	Position of individual atoms of the structure	(--)
r_{mean}	Mean position of all atoms of the structure	(--)
$C_{aa,s}$	Frequency of amino acid aa in dataset s	(--)
$n_{aa,s}$	Occurrence of amino acid aa in dataset s	(--)
PSR_{aa}	PSR index of amino acid aa	(--)
$P_{Enrichment}$	Probability for Enrichment of an amino acid pair	(--)
$P_{Reduction}$	Probability for Reduction of an amino acid pair	(--)
$p^{(1)}(x_1, \dots, x_m)$	Probability of an m length of sequence in first-order model	(--)
$p^{(2)}(x_1, \dots, x_m)$	Probability of an m length of sequence in second-order model	(--)

Chapter 1

Introduction

Defining Protein Disorder

More than a century ago, the discovery about the structural fitness between an enzyme and a substrate led to the formation of the famous “lock and key” hypothesis, in which, the substrate (key) must possess a specific conformation to dock into the catalytic site (key-hole) of an enzyme (lock) [1]. The associated sequence-structure-function paradigm of protein folding states that the sequence of a protein determines its native three-dimensional structure in an aqueous environment, and a protein folds into a defined, stable and rigid three-dimensional structure to fulfill its functional purpose [2, 3]. The folding hypothesis has been demonstrated by a tremendous number of identified X-ray crystal structures and nuclear magnetic resonance conformers deposited in the Protein Data Bank (PDB) [4-9]. While many early scientists were aware that some protein sequences may not fold into such definite structures, the protein folding paradigm dominated our understanding of structure-function relationships. Now we are more aware of the significant fraction of proteins with native biological functions but that lack folded structure, either in their entirety or in portions. The evidence arises from proteins that either do not crystallize under any conditions, or whose determined structures have missing electron densities in X-ray diffraction, or that do not have stable defined structure in solution in nuclear magnetic resonance (NMR) spectrometry [10-27]. These flexible and disordered proteins or regions simply lack a unique folded conformation. They are frequently referred as flexible, mobile, partially

folded, natively denatured [28], natively unfolded [29, 30], intrinsically unstructured [31, 32], and recently a more common term, intrinsically disordered [33] (Figure 1.1). The definition of intrinsic disorder is clarified as regions in the protein structure where the equilibrium position of the backbone along with the dihedral angles, has no specific values and vary significantly over time [33, 34]. How can we describe such proteins? For the purpose of clarification, the conformational states that are available to proteins are defined here. First, the native state is a protein's observable conformation related to its biological functions [35]. A native state is often a folded state, which is structured and ordered [36] typically with common elements of protein folds such as secondary structure and a hydrophobic core. Yet, the native state of a protein sequence is not necessarily folded [37]; sometimes, it is rather an unfolded state, which is unstructured or disordered, not restricted to be a random coil, but possibly also consisting of extended disorder (pre-molten globule) and collapsed disorder (molten globule) [33, 38] components. If a protein's unfolded state is obtained through chemical denaturation, for example in high concentrations of urea, or at high temperature, such a state is normally referred as the denatured state, which is itself a non-native state [35]. Denatured states have common unstructured properties with intrinsically disordered proteins (IDPs), but the details of the types of conformations observed may differ. For over five decades Intrinsically Disordered Proteins by any name, have been considered to be mysterious as their structural features have remained evasive. Recent improvements in both experimental techniques and computational approaches are starting to improve our understanding of all forms of protein disorder.

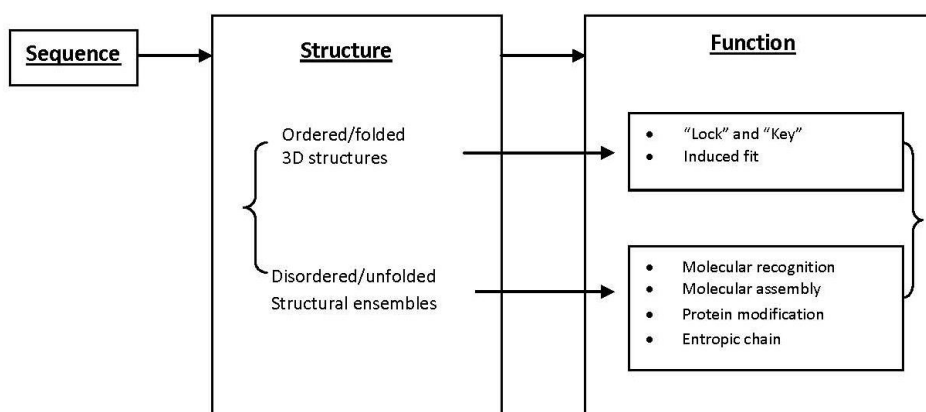


Figure 1.1: The protein sequence-structure-function paradigm.

Experimental Characterization

Most traditional experimental methods have limited abilities to characterize the 3-dimensional structure of intrinsically disordered proteins. NMR spectroscopy and circular dichroism (CD) spectropolarimetry [33] are the most useful of these. There are no examples of full-length disordered proteins that can be crystallized from solution thus their structures cannot be detected by X-ray crystallography, however small regions of disorder can be detected by the absence of data. For proteins having both ordered and disordered regions, they are able to crystallize on account of the ordered regions' crystallization. Disordered regions give incoherent X-ray scattering resulting in missing electron density [17, 39-44].

NMR is able to characterize protein disordered regions, transient secondary and tertiary structures as well. It can also be used to study the structure in a dynamic way [45-55]. A set of biophysical terms can be measured from NMR experiments including chemical shifts [56-58], scalar

couplings [59], residual dipolar couplings (RDCs) [60-64], and paramagnetic relaxation enhancement (PRE) effects [65]. These biophysical terms can often be expressed in terms of bond angle or atom distance information, and then used as restraints for fitting coordinate models of disordered proteins. Molecular simulations are necessary to generate examples of the conformational space that may be explored by disordered proteins, after which the associated NMR restraints can be applied to refine the models that fit the experimental data. Chemical shifts are the atoms' unique frequencies specified in the resonance spectrum. The deviation from random coil to helix and beta strand conformations can be determined by tables of chemical shift, and these inform us of evidence of local secondary structures [66-69]. Scalar couplings can inform us of the observed backbone dihedral angles in a protein structure. RDCs report the information about the bond angles and vectors relative to the core structure. PRE effects can provide long-range distance restraints.

CD identifies disordered proteins by measurement of low intensity near-UV backbone optical polarization information, which can be compared to standard protein folds. Deviation from folded backbone conformations can show a protein is intrinsically disordered [70, 71]. Other important techniques include small angle X-ray scattering (SAXS), hydrodynamic measurements such as size exclusion chromatography, infrared spectroscopy, fluorescence resonance energy transfer (FRET), conformational stability with effects of temperature and pH, mass spectrometry-based high resolution hydrogen-deuterium exchange, protease sensitivity and optical rotary dispersion (ORD). Table 1.1 provides a list of current experimental techniques for intrinsic disorder characterization.

SAXS can be applied to evaluate the size of protein structure in solution, which is then compared to its globular form with features like the signal changes at higher scattering angles, radius of gyration (R_{gyr}) and maximum dimension [72-75]. FRET captures the structural state by measuring the distance distribution between the donor and acceptor chromophores [76-79]. Taken together, these experimental measurements, especially from NMR [56-65], SAXS [80-82] and FRET [83-85], can often be used as sources for constructing ensembles for disordered proteins as fill in structural information missing from disordered regions. However the structures that result from these are often represented as an ensemble of 3-dimensional disordered structures, with some number of static structures that demonstrate the range of conformational variants that may fit the experimental data. The ensemble is implied to represent “snapshots” of the protein as it may dynamically meander and explore its native disordered states.

A combination of multiple experimental techniques will give more information about the identification and conformational states of intrinsic disorder over a single technique. Many experimentally identified disordered protein regions arising from conventional structures have been deposited into a database called DisProt [86]. However, difficulties exist in identifying sequence with intrinsic disorder, by a myriad of effects for example structural experimentation nuance of structure definition, protein expression, and reagents. A number of computational tools have been applied to the problem of identifying the specific regions that exhibit intrinsic disorder, which are becoming more helpful in working with intrinsic disordered proteins.

Table 1.1: Experimental methods for characterizing intrinsically disordered proteins

Major Experimental Methods For Study of Intrinsic Disordered Proteins

X-Ray Crystallography
Nuclear Magnetic Resonance (NMR) Spectroscopy
Small Angle X-ray Diffraction (SAXS)
Circular Dichroism (CD) Spectropolarimetry
Infrared Spectroscopy
Fluorescence Resonance Energy Transfer (FRET)
Size Exclusion Chromatography
Native Acrylamide Gel Electrophoresis
Conformational Stability (through Temperature or PH)
Mass Spectrometry-Based High Resolution
Hydrogen-Deuterium Exchange
Protease Sensitivity

Computational Prediction

R.J.P Williams proposed the first disorder predictor in 1979 based on the extremely high ratio between the number of charged residues and the number of hydrophobic residues [87, 88]. Secondary structure prediction algorithms starting with the GOR (Garnier, Osguthorpe, Robson) [89-92] indicated a fractional prediction of percent “coil” which may be interpreted as a lack of secondary structure and therefore a disordered region, however these tools were never widely used or tested with modern disordered datasets. The first well defined disorder predictors PONDR®s using artificial neural network algorithms were developed by the research group of Dunker, Obradovic and Uversky [30, 42, 93-102]. To date more than 50 computational approaches have been designed to discover disordered regions along protein sequences. Many of these predictors have online servers. Table 1.2 provides a series of current disorder predictors in details. These methods are discussed thoroughly in many review articles [103-106]. Disorder prediction was included in the

biennial Critical Assessment of Structure Prediction (CASP) since 2004 [107-111] which focuses on identification of structurally characterized small regions of disorder. This assessment brings further advancement in the development of disorder predictor design. At the same time, disorder predictors can give feedback to experimental protocols for accurate identification of intrinsic disorder. Among the published disorder predictors, such as , PONDR®s [93, 96, 98, 101, 102, 112, 113], DISOPRED [114, 115], RONN [116] and POODLE [117-120], machine learning algorithms including neural networks (NN) and support vector machines (SVMs) are used as the basic methods. The input features used in training these algorithms are largely different from each other, including amino acid composition, net charge, predicted secondary structure, and hydrophathy. Some predictors, such as GlobPlot [121] and IUPred [122, 123], use rather simple algorithms, yet they are able to effectively predict disordered regions. Some of the predictors have improved their efficiency through modifications. A number of metaprediction servers have also been developed, integrating different disorder predictors into a consensus prediction. Examples of metaprediction servers include DisPSSMP2 [124], PrDOS [125], MD [126], MFDp [127], GSmetaDisorder [128], which are generally able to produce better prediction results. Fundamentally, disorder predictors all rely on the properties of disordered regions that can be understood as amino acid compositional and contextual differences between ordered and disordered proteins.

Table 1.2: A list of current disorder predictors with available URL and brief description. Table adapted by author from [103-105].

Predictor	Publication year	Brief description
SEG[129] http://mendel.imp.ac.at/METHODS/seg.server.html	1994	SEG predicts low-complexity or compositional biased segments as well as non-globular domains. For predicting long and short non-globular domains, different parameters must be used. SEG is not trained as a disorder predictor, but as there is a correspondence between low-complexity sequence and disorder, often finds disordered regions.
HCA (Hydrophobic Cluster Analysis)[130] http://smi.snv.jussieu.fr/hca/hcaseq.html	1997	HCA predicts hydrophobic clusters, which tend to form secondary structure elements. This method is based on a helical visualization of amino acid sequence. The prediction output can display coiled coils, compositional biased regions and boundaries of disordered proteins.
PONDR® (XL1, VL1, XL-XT, VL2, VL3, VSL1, VSL2) [93, 96, 98, 101, 112, 113] http://www.pondr.com	1997-2006	PONDR®s includes a series of predictors which can predict disordered regions. The types of disordered regions predicted by PONDR® predictors include random coils, partially unstructured regions, and molten globules. It is trained with local amino acid composition, flexibility, hydrophobicity etc, using feed-forward neural network. These predictors perform well in disorder prediction as shown in many applications.
Charge/hydrophobicity method[30] http://www.pondr.com	2000	Charge/hydrophobicity method predicts fully unstructured domains (random coils) based on global sequence composition (hydrophobicity versus net charge). This method is expected to identify disordered regions that are not present in DisProt. Prior knowledge of modular organization of protein is required. It is only applicable to domains without disulfide bonds and without metal-binding regions.
GlobPlot [121] http://globplot.embl.de	2003	GlobPlot predicts regions with high propensity for globularity based on the Russell/Linding scale [121], which describes the relative propensity of an amino acid residue to be in an ordered (secondary structure) or disordered (random coil) state. The output provides an overview of modular organization of large proteins and shows changes of slope corresponding to domain boundaries. GlobPlot is user-friendly with built-in SMART, PFAM and low-complexity predictions.
DisEMBL[131] http://dis.embl.de	2003	DisEMBL is able to predict three kinds of disordered structure, including loops/coils (regions devoid of regular secondary structures), hot loops (highly mobile loops), and those that are missing from the PDB X-ray structures (REMARK465). The neural networks were trained with X-ray structure data. DisEMBL also displays the low-complexity regions and propensity of aggregation. Prediction using loops/coils predictor is most trusted.

NORSp[132] http://cubic.bioc.columbia.edu/services/NORSp	2003	NORSp predicts regions with No Ordered Regular Secondary (NORS) structure, most of which are highly flexible. It is based on secondary structure and solvent accessibility. NORSp generates and uses multiple sequence alignment. Some highly flexible regions are yet predicted to contain secondary structures.
DISOPRED [114] DISOPRED2 [115] http://bioinf.cs.ucl.ac.uk/disopred	2003	DISOPRED trains the whole sequence information using neural networks.
	2004	DISOPRED2 is trained with PSI-BLAST profiles using cascaded support vector machine (SVM) classifiers and generates and uses multiple sequence alignment. It predicts regions lack of ordered regular secondary structure. However, when there are few homologues, the prediction accuracy is lower.
Weather's method [133]	2004	Weather's method uses SVM analysis of a linear combination of composition vectors.
DRIPPRED [134] http://www.sbc.su.se/~maccallr/disorder/	2004	DRIPPRED is based on Kohonen's self-organizing map and received a good evaluation at CASP6.
FoldUnfold [135-137] http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi	2004	FoldUnfold is based on the idea that the structure of proteins is governed by the balance between the interaction energy of residues and their conformational entropy.
IUPred[122, 123] http://iupred.enzim.hu	2005	IUPred predicts regions that lack a well-defined 3D structure under native conditions. It is based on the idea that the energy resulting from inter-residue interactions is responsible for determining whether a protein forms structure or not. This method is expected to identify disordered proteins that are not present in DisProt and only applicable to proteins without disulfide bonds and without metal-binding regions.
RONN [116] http://www.strubi.ox.ac.uk/RONN	2005	RONN predicts regions that are lack of a well-defined 3D structure under native conditions. It trains on disordered proteins using bio-basis function neural network. RONN is restricted to search for short regions of disorder.
DISpro[138] http://scratch.proteomics.ics.uci.edu/	2005	DISpro is based on a one dimensional recursive neural network (1D-RNN) model, the flexibility of Bayesian model and a fast, convenient, parameterization of an artificial neural network (ANN).
FoldIndex [139] http://bip.weizmann.ac.il/fldbin/findex	2005	FoldIndex is used to analyze the ratio of net charge with hydrophathy locally using a sliding window. It predicts regions that have a low hydrophobicity and high net charge (loops or unstructured regions). FoldIndex provides prediction on probable short loops but no prediction on N- and C-termini.
PreLink[140] http://genomics.eu.org	2005	PreLink predicts regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner. It is based on compositional bias and low hydrophobic cluster content.
Spritz [141] http://distill.ucd.ie/spritz/	2006	Spritz consists of two specialized binary classifiers, one for short disordered regions and the other for long disordered fragments.
IUP[142]	2006	IUP is based on a Recursive Maximum Contrast Tree (RMCT) to recognize intrinsically disordered regions.

DisPSSMP[143] DisPSSMP2[124] http://biominer.bime.ntu.edu.tw/ipda/	2006	DisPSSMP is based on Radial Basis Function Networks with inputs from position-specific scoring matrices and other sequence properties.
	2007	DisPSSMP2 uses a two-level prediction scheme and a condensed position-specific scoring matrix.
NORSnet [144] http://cubic.bioc.columbia.edu/services/NORSp	2007	NORSnet uses feed-forward neural networks.
POODLE-S[118] http://mbs.cbrc.jp/poodle/poodle-s.html	2007	POODEL-S is a group of seven SVM predictors with each responsible for a specific region of the whole sequence.
POODLE-L [117] http://mbs.cbrc.jp/poodle/poodle-l.html		POODLE-L is composed of ten two-level SVM predictors.
POODLE-W [119] http://mbs.cbrc.jp/poodle/poodle-w.html		POODLE-W predicts disordered structures by using a Spectral Graph Transducer (SGT) and by training with a huge amount of structure-unknown sequences.
PrDOS[125] http://prdos.hgc.jp/cgi-bin/top.cgi	2007	PrDOS consists of two predictors, one of which uses the alignment of homologs.
metaPrDOS[145] http://prdos.hgc.jp/cgi-bin/meta/top.cgi	2008	MetaPrDOS is composed of seven individual predictors which areas follow: PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-S.
Bayes[146]	2008	Bayesian method computes the conditional probability of a sequence from a certain class and then infers the posterior probability of the class.
OnD-CRFs[147] http://babel.ucmp.umu.se/ond-crf/	2008	Conditional Random Fields (CRFs) method predicts the intrinsic disorder in proteins. CRF is a discriminatively supervised machine-learning method.
DISOclust[148] http://www.reading.ac.uk/bioinf/DISOclust/DISOclust_form.html	2008	DISOclust applies the principle that ordered residues within a protein target should be conserved in three-dimensional space within multiple models, whereas the residues that vary or are consistently missing may be correlated with the disordered structure.
MD [126] http://cubic.bioc.columbia.edu/newwebsite/services/md/index.php	2009	MD is a meta predictor composed of NORSnet, Ucon, PROFBval, DISOPRED2, IUPred, and FoldIndex.
CDF-ALL[149]	2009	CDF-ALL is a protein-level disorder meta predictor composed of CDFs from VLXT, VSL2, VL3, TopIDP, IUPred, and FoldIndex.
PreDisorder[150] http://casp.met.missouri.edu/predisorder.html	2009	PreDisorder uses a 1D recursive neural network with the input of a profile generated from PSI-BLAST, the predicted secondary structure and solvent accessibility.
POODLE-I[120] http://mbs.cbrc.jp/poodle/poodle-i.html	2010	POODLE-I is a meta predictor integrating POODLE-S, POODLE-L and POODLE-W.
PONDR-FIT[102] www.disprot.org	2010	PONDR-FIT is a meta predictor that is trained using ANN with the results of PONDR®VLXT, VL3, VSL2, IUPred, FoldIndex and TopIDP.
MFDp[127] http://biomine-ws.ece.ualberta.ca/MFDp.html	2010	MFDp is a meta predictor consisting of DISOPRED2, DISOclust, and IUPred. Other information, for example, PSSM, residue flexibility and back-bone dihedral torsion angles, etc are taken as input.

IsUnstruct[151]	2011	IsUnstruct is developed using Ising model which involves an estimation of the energy of the border between ordered and disordered regions.
DisCon[152] http://biomine.ece.ualberta.ca/DisCon/	2011	DisCon is based on a ridge regression model with the input of information on sequence, evolutionary profiles, and so forth.
DICHOT[153, 154] http://spock.genes.nig.ac.jp/~genome/DICHOT	2011	DICHOT system combines structural domain identification, DISOPRED2 disorder prediction and CLADIST classification program to predict structural domains and intrinsically disordered regions.
GSmetaDisorder[128] http://iimcb.genesilico.pl/metadisorder/	2012	GSmetaDisorder is a meta predictor that combines 12 disorder predictors: DisEMbL, DISOPRED2, DISpro, GlobPlot, iPDA, IUPred, Pdisorder, POODLE-S, PrDOS, Spritz, DisPSSMP and RONN.
CH-CDF plot[155]	2012	CH-CDF plot method is a combination of two methods: Charge/hydrophathy and CDF-ALL. It is able to predict proteins into four categories: structured, mixed, disordered and rare.
SPINE-D[156] http://sparks.informatics.iupui.edu/	2012	SPINE-D is based on a single neural network to predict if the residues are ordered or disordered and if they are in short or long disordered regions. Its evaluation was among the top servers in CASP9.

Studies have been carried out to learn about the difference in the amino acid compositions between ordered and disordered proteins using the sequences in DisProt. According to variation compared to DisProt, disordered regions contain higher percentages of disorder-promoting amino acids (A, G, R, Q, K, S, E and P) and lower percentages of order-promoting amino acids (W, F, Y, I, L, V, N and C) compared to the ordered regions [33, 96, 157-159]. This peculiarity in amino acid composition explains that disorder regions have overall low hydrophobicity and high net charge [30]. The sequence composition and order influence other biophysical properties of disordered regions, for example, flexibility index, helix propensities and strand propensities [157]. These biophysical properties together with amino acid sequence are treated as input features in the development of various sequence-based disorder predictors as discussed above and in Table 1.2. An amino acid scale was derived for better discrimination of order and disorder. The twenty residues are ranked according to their tendencies of promoting order to disorder as the following: W,F,Y,I,M,L,V,N,C,T,A,G,R,D,H,Q,K,S,E,P [160]. Note however that this ranking can be counter-intuitive. For example, glycine has the largest conformational space variation and would be expected to be on the extreme end of disorder promotion. Proline has the smallest conformational space and would be expected to be order promoting on that basis. However there is no simple correspondence between individual amino acid properties and structure disorder, simply because it is dependent on the context of neighboring residues and whether the sequence evolved some folded structure. Depending on the properties of the R-group in each residue, the twenty standard amino acids can be classified into several groups: non-

polar aliphatic (G, A, V, L, M and I), non-polar aromatic (F, Y and W), polar acidic (L, R and H), polar basic (D and E) and polar uncharged (S, T, C, P, N and Q). The aromatic residues (W, F and Y) as well as the bulky hydrophobic residues (I, L and V) are preferred in the hydrophobic core of folded globular domains. Thus, these residues are grouped into the order-promoting residues. Earlier studies show that low-complexity in amino acid composition infers the non-globular domains of proteins [161, 162]. A sequence is said to be of low-complexity if it is biased in local composition to one or more amino acids beyond what is expected in a normal sequence distribution. While low-complexity regions are often also intrinsically disordered, some are not, and some disordered regions fail to be detected by low-complexity locating software such as SEG [129]. It has been reported that amino acid composition alone cannot predict short-disordered regions (≤ 30 residues) effectively, but it is adequate to predict long-disordered regions (> 30 residues) accurately. Rauscher and Pomes [163] argued that for a protein polypeptide, when its sequence length increases, the amino acid composition is a sufficient criterion to predict long disorder regions, and at the same time, the sequence context become less important [163, 164].

Molecular Simulation

In order to understand how the conformations of intrinsically disordered proteins behave, ensembles are created by various means computational simulation together with restraint fitting as previously mentioned. The tools for molecular simulation are largely biased by a focus on structured proteins,

so exploration into the ensembles of disordered protein regions is limited by methods that have been more broadly used for the topics of protein folding and unfolding. Disordered proteins are anticipated to have a flat energy landscape (Illustration 1.1) and therefore adopt a large number of diverse conformational states at room temperature in solution. It is intriguing to compute the energy landscape of disordered proteins; however, the topic is beyond the scope of this thesis. In order to study the disordered protein conformations, an enormous conformational space needs to be sampled followed by some statistical analysis to understand the biophysical properties of the ensemble. To date, a few research groups have attempted to model disordered regions through an ensemble-based interpretation.

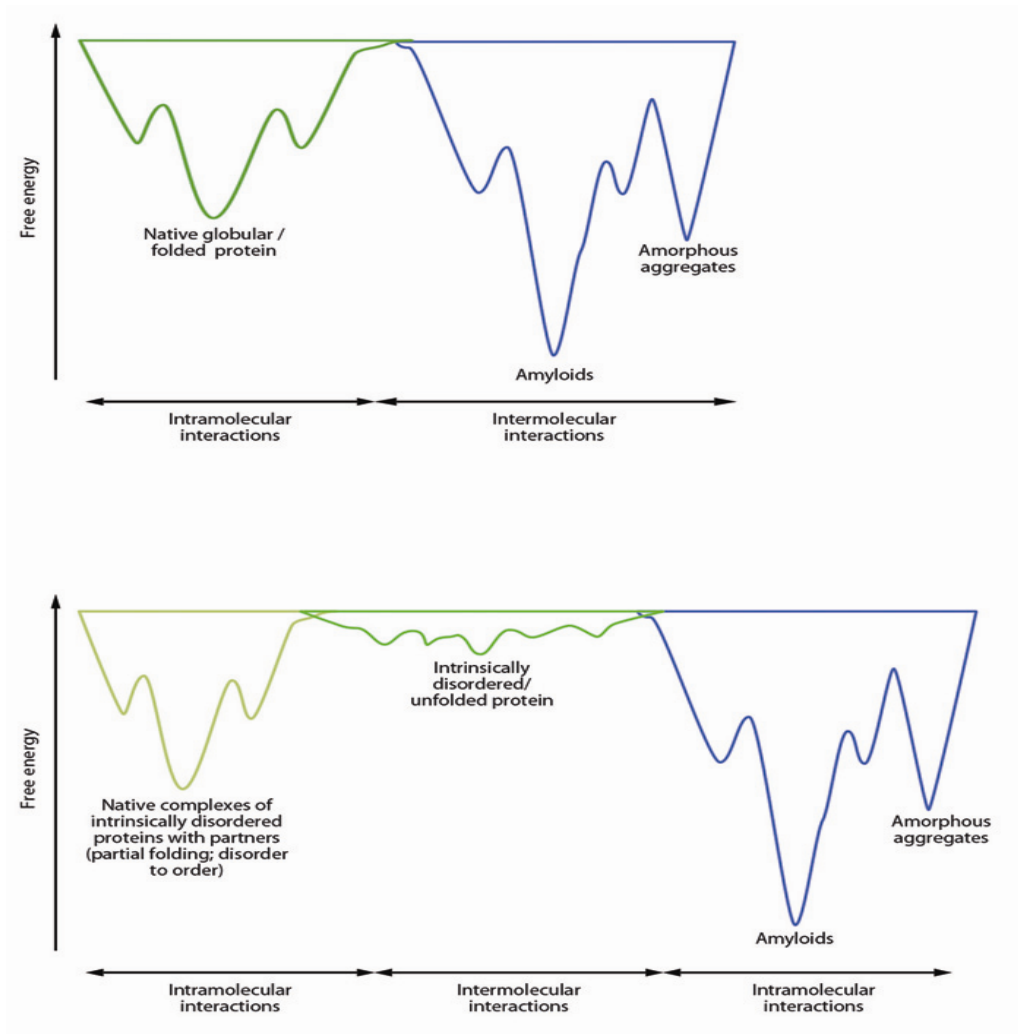


Illustration 1.1: An illustration of energy landscape models for globular/folded proteins and intrinsically disordered/unfolded proteins. This figure is designed based on the earlier energy funnel model proposed for globular/folded proteins and adapted by author from [9, 27, 165].

Molecular dynamics (MD) and Monte Carlo (MC) algorithms are commonly applied to simulate the conformational space of disordered regions. While widely used for folded protein conformational studies and docking, MD has some shortcomings in addressing broad conformational sampling of IDP. MD employs Newton's formula of motions $F = \frac{dp}{dt}$ (F is the force, p is the momentum and t is the time.) in a small time frame when sampling conformers of disordered regions. MD is a helpful method to model conformers of disordered regions; nonetheless, it has constrained usage in modeling long disordered regions because the time frame required would be incredibly small, *i.e.* nanoseconds. The basic algorithm calculates the energy associated with covalent bonds, dihedral angles, torsions, van der Waals interaction (Leonard-Jones potential) as well as an electrostatic potential (Coulomb potential). Every term requires parameterization which is mutually mentioned as the force field [166]. Force fields have been refined over the past two decades but were initially slow to accurately represent the observed distribution of backbone angles found in the PDB database. More recently, MD has been improved to do conformational sampling with replica exchange [167], accelerated [168] or quenched MD [57]. Lei and Duan gave more details in their review article on MD sampling approaches [169].

MC sampling is a stochastic process that favors or disfavors a protein conformation by determining if the calculated values agree with the experimental measurements or not, and a free energy potential is calculated in the meanwhile. MD and MC are often coupled together or integrated into other techniques to sample a conformational pool and search for a subset

ensemble, of which structural properties most resemble those from experimental measurements. A few methods are discussed below.

The Hilser group developed a statistical thermodynamic model called COREX [170, 171], which is able to calculate free energy and entropy by partitioning the protein sequence into a number of folding units with a window length. Each folding unit is computationally rotated out from the folded structure and calculations are made to determine the energy and whether the structure should be folded or unfolded. Each partitioning will generate $2^N - 2$ partially folded conformational states, where N is the number of folding units. The partitioning process is iterated through the entire sequence. Eventually, the total number of conformational states equals to the summation of the number of partially folded conformational states generated in each partitioning in addition to the fully folded state and fully unfolded state, ie. $(\sum 2^{N_i} - 2) + 2$ where N_i is the number of folding units in each partition. This algorithm calculates the entropy and hence can report the Gibbs free energy of each conformational state. COREX requires a crystal structure of the studied protein as a template. It has therefore been demonstrated useful in investigating the cooperative [172-175] and allosteric behaviors [176] of protein conformations. However, this method cannot be applied to intrinsically disordered proteins with no starting structure. Without identification of any partially folded regions, this approach is of limited value for IDP analysis.

TraDES (Trajectory Directed Ensemble Sampling) [177, 178] is an unbiased all-atom conformational sampling software which can generate both native and non-native conformational states. The software uses dictionaries of backbone conformations from a high quality nonredundant set of PDB

structures for selecting backbone angle conformations, and the backbone-dependent rotamer library of Dunbrack [179] for placement of amino acid side chains. Originally designed for sampling conformational space to find folded proteins by brute force, it was the first such program to be adapted by NMR researchers for generating ensembles of unrestricted IDP structures prior to restraint fitting. As it is a validated $O(N \log N)$ algorithm, it is much faster than other methods at sampling conformational space. Validated backbone atom and side chain placement accuracy have made it a system of great utility for IDP studies.

The TRADES software is divided into two phases. The first phase reads in a protein sequence, and provides a trajectory distribution file, which stores the chemical graph of the structure with any post-translationally modified amino acids, together with the distributions of Ramachandran dihedral angles for each residue. It is called a trajectory distribution because it contains the information for sampling the conformational space of the protein as modeled as an N-to-C terminal build up process. Each possible 3D protein structure is considered a single trajectory through the distribution. Trajectory distributions can be created using combinations of Ramachandran space gathered from specific secondary structure, for example TraDES can create all-coil or all-beta structure samples, or it can use a 3-state secondary structure prediction such as the GOR method to bias the trajectory distribution of each residue to more frequently sample its most preferred secondary structure.

The TraDES trajectory distribution file serves as an input to the *ab initio* conformer generator, which is the second phase of the TraDES system. This samples the space encoded by the trajectory distribution to rapidly make

a large sample of plausible unfolded protein conformers. It works by adding residues one by one from N-to C- terminus based on probabilistic geometry sampling. TraDES sampling does not apply any explicit potential functions, and creates structures with a combination of statistics and sterics. Philosophically, the TraDES structure sampling method avoids energy computations while building protein conformers, in order to avoid any bias arising from any particular force field. Thus, any energy scoring function can be estimated on the sampled conformers which are all-atom models. TraDES outputs potential terms including Zhang potential (an atom-based statistical potential showing the amount of favorable contacts) [180], Bryant-Lawrence potential (a residue-based threading potential) [181], and the VSCORE potential (an atom-contact based scoring function) [182]. TraDES is able to reconstruct folded proteins matching high quality PDB structures to very low RMSD (Root Mean Square Deviation) tolerances, which is a form of validation to demonstrate that native structures embedded within the trajectory distribution can indeed appear in the sample, if it is sufficiently large. TraDES is also used as the initial step for conformational sampling in other Monte Carlo methods, for example the NMR package ENSEMBLE [59, 183-186]. ENSEMBLE allocates weights to each conformer in a TraDES-generated 3D structure ensemble to optimize the mapping between the ensemble-averaged properties and experimental data. The experimental restraints used in ENSEMBLE are chemical shifts, NOEs, PREs, RDCs, hydrogen exchange protection factors, solvent-accessible surface area, and hydrodynamic radius. ENSEMBLE was originally applied to calculate the native and non-native states of drk SH3 domains [183, 184], but now it has become more widespread

in the NMR community. Sample and Select (SAS) [84, 187, 188] is another Monte Carlo approach that assign equal weight to each conformer in the ensembles and select a subset of conformations that minimize the difference between predicted and experimental data.

Other NMR research groups have built systems similar to TraDES, but have implemented models that are more restrictive to predict the disorder conformational space by assuming the disordered regions most likely adopt random coil structures. Jha *et al.* made a statistical coil model which can produce an equilibrium ensemble of polypeptides from Monte Carlo simulations [189, 190]. Firstly, they constructed a coil library consisting of residues that lie outside of helices, sheets and turns from an X-ray structure dataset of 2020 peptide chains. Then, the conformational state is generated by assigning each residue specific ϕ , ψ angles of a type of Ramachandran basin (α_R , β , PPII, α_L and γ) according to the basin's frequency in the coil library. Note that this set of basins is much coarser than the 400x400 divisions used in TraDES. A statistical potential is calculated as the simulation process carries on. The modeling results agree with the experimental RDC values of denatured proteins, whereas it does not explain if these conformations are native states.

Another example of a TraDES inspired package is the Flexible-Meccano (FM) packaged developed by the Blackledge group [60, 62-64, 191, 192]. Like TraDES, their algorithm generates backbone structures with an N-to-C terminal build up sampling from specific coil regions obtained from high-quality nonredundant crystal structures. FM has demonstrated great utility in matching observed RDC data alone or together with a RDC-restrained

molecular dynamics refinement. For disordered proteins, the conformational state is formed by constructing consecutive peptide planes and tetrahedral junctions from the selected ϕ, ψ angles which are randomly retrieved from a loop library, which is similarly to the coil library built in the study by Jha et al., but with less X-ray structures, *i.e.* 500, and different resolution thresholds. FM was applied to study the disordered regions in the nucleocapsid-binding domain of Sendai virus phosphoprotein [62], and the ensembles were used to demonstrate that the experimental RDC and SAXS results are dominated by coil behavior. This approach is further integrated into an ensemble optimization method to quantitatively search the subset ensemble that matches SAXS data in a Monte Carlo way [63].

In choosing between MC and MD techniques, it is useful to note that there is an ongoing debate as to whether disordered regions can be modeled simply with random coils or they actually contain a certain amount of local or long-range contacts [193]. However this debate may be missing the point of context, in that there may be instances of disordered proteins that have no local or long range contacts, while there may be others that do. From the standpoint of evolution, either outcome may have a specific fitness or capability. Given this, a genetic algorithm has been coupled to FM by the Blackledge group, which has produced the program ASTEROIDS [56, 60, 192], stochastically searches for conformations whose predicted conformational variants are in an agreement with the experimental values.

MD approaches tend to produce limited samples of conformational space owing to the energy function's propensity to drive towards local minima. The MC methods used by TraDES and similar approaches do not suffer this

limitation. Additional approaches, for instance, Rosetta [194], CNS [195] and Xplor-NIH [196], apply simulated annealing in their mechanisms. Rosetta creates ensembles of structures by swapping nine-residue long fragments, which takes upon possible local structures that are found in a known similar protein sequence [194]. This approach can be described as a simulated annealing process that considers Bayesian scoring functions, but it lacks an ability to broadly sample the conformational space of disordered proteins owing again to its tendency to optimize the energy of folded regions. The original CNS used simulated annealing to generate conformers by starting with an all-beta strand extended configuration with plausible geometry [195], however this method is inefficient at producing large samples of conformational space. Xplor-NIH, an improved version of Xplor [197], and the updated CNS software can sample structural ensembles via NMR experimental restricted simulated annealing and energy minimization. Energy-minima Mapping and Weighting (EMW) algorithm [57, 198] assigns a statistical weight from 0 to 1 to each conformer and optimize the conformational ensemble at the same time according to a simulated annealing protocol.

The above methods and other molecular simulation techniques not covered in the discussion all attempt to search for an ensemble out of the conformational space that corresponds to the experimental data with or without some energy scoring function. None of these methods, however, take spatial or steric boundaries such as membranes or close-packing into consideration, which disordered regions may actually encounter in a cellular context. In this thesis, spatial constraints comprising of membranes and nearby

molecules or assemblies, are examined to determine whether they may alter the conformational space available to the disordered region of a protein. The restriction of conformational space sampling caused by neighboring structures or membranes may alter the ensemble structure of the disordered region, thereby modifying its statistical structural properties, and alter its functional role.

Prevalence, Function, and Disease Impact of Intrinsically Disordered

Intrinsically disordered proteins are prevalent in the three kingdoms of life [115, 199]. Bioinformatics technique has predicted that 33% of eukaryotic proteins contain disordered regions. The content of protein disorder is predicted to be 4.2% and 2% in bacteria and archaea [115]. Researchers argue that the prevalent existence of protein disorder in higher organisms may stem from the much more complicated signaling and regulation systems, in which they play important roles. The functions of intrinsically disordered proteins are summarized as four categories: molecular recognition, in which they act as effectors and scavengers displaying sites for post-translational modifications; molecular assembly; protein modification; and entropic chain activities [158, 200]. They are involved in a multitude of cellular processes, for example, transcription, translation, cell cycle control and signal transduction. Moreover, protein disorder is often associated with Alzheimer's disease, Parkinson's disease, and others which are collectively known as neurodegenerative conformational diseases [201]. It is reported that $57\pm 4\%$ of cardiovascular

disease associated proteins and $79\pm 5\%$ of cancer associated proteins are predicted to contain disorder regions with a length of more than 30 consecutive residues [202, 203].

Proline-Rich Disordered Regions

In the regulation of signal and mechano-transduction, a collection of “hub” proteins, such as, α -synuclein, p53, 14-3-3, AXIN, are indispensable proteins which bind to a number of other proteins via protein-protein interactions [204-207]. When they are removed via knock-out or knock-down experiments, the missing “hub” proteins will disrupt the necessary interactions with their partner proteins resulting in unsuccessful binding and signal transduction. Studies have been done to find that these “hub” proteins and their interacting partners interact with each other via the disordered regions within both [208-214]. Many of which carry short binding motifs within proline-rich regions such as the yeast protein Las17 [215]. Another example is the tumor suppressor p53, which is the central hub protein in a complex signaling network. The N-terminal domain (NTD; residues 1-94) of p53 containing a proline-rich region (PRR; residues 61-93) is intrinsically disordered and interacts with Tfb1 (PDB:2GS0), Mdm2 (PDB:1YCR) and Rpa70 (PDB:2B3G) [216, 217]. AXIN is a scaffold protein in Wnt [218], TGF- β [219], c-Jun N terminal/stress-activated protein kinase (JNK) [220] and p53 pathways. The highly disordered fragment of residues 383-480 in AXIN is compositionally biased with proline and is able to bind GSK3 β (PDB:1O9U) and β -catenin(PDB:1QZ7) [221].

Proline in Intrinsically Disordered Regions

A great deal of study has been done on proline-rich regions, whose properties and behaviors originate from the special amino acid proline. Proline is ranked in the first place in the amino acid scale of promoting disorder [160]. This is due to the peculiar amino acid configuration of proline compared to the rest of its peers. The proline side chain is cyclized back onto the backbone amide position, a unique configuration that grants proline the following distinct properties.

First of all, proline has a very restricted backbone conformation. The ϕ dihedral angles are limited to take a value around -65° [222, 223]. The value of ψ dihedral angle is not as constrained and is free to be in the α -helical region ($\psi \approx -40^\circ$) or the β -sheet region ($\psi \approx +150^\circ$). Studies of prolines in crystal structures show that approximately 44% of prolines are in the α region and 56% are in the β region [224, 225]. The preceding residue of proline in Xaa-Pro dipeptide, greatly affects the conformation of proline. A hydrophobic preceding residue or cis bound in Xaa-Pro creates a higher tendency for proline to be in β region. When the preceding residue Xaa is a tyrosine residue, the fraction of Xaa-Pro cis conformation was observed to increase from 5-6% up to 19% [224, 226].

Second, for a given Xaa-Pro dipeptide, proline also affects the conformation of its preceding residue via the bulky N-CH₂ group, disfavoring the α -helix conformation of the preceding residue [224, 225, 227]. The preceding residue Xaa tends to be in the β conformation when the Pro ϕ angle

is constrained to be around -65° . Hence, the Xaa-Pro dipeptide is likely to be fairly rigid and extended.

Third, with the amide proton replaced by a CH_2 group and its bulky side-chain, proline fails to act as a hydrogen bond donor and disrupts the structure of both helix and β -sheet. Within a helix, proline causes a 'kink' to form, bending the helix at the point where the hydrogen bonding network is disrupted. In addition, proline substitution disrupts the normal pattern of β -sheet hydrogen-bonding conformations [228] in both the parallel and antiparallel forms. This implies that proline-containing regions are incapable of binding to proteins that form strand-edge protein interactions, such as the crystallin family of chaperones. Proline is often found at the beginning of a helix. The reason is mainly because the ϕ dihedral angle of proline is constrained to an angle normally found in a helix [229].

Proline-Rich Motif, Proline-Rich Regions, and Polyproline II Helix

Many short sequence segments with identified interaction and function contain at least one conserved and functionally required proline. These short sequences are referred as proline-rich motifs, which can be recognized by several modular domains and phosphopeptide-binding domains. Table 1.3 lists current known modular domains and phosphopeptide-binding domains and their binding specificities related to proline-rich motifs. Proline-rich motifs often appear in cluster in a much longer proline-rich region with up to hundreds of residues (See Table 1.4 for some examples). Some proteins whose proline-rich regions do not contain repeated proline-rich motifs are listed in

Table 1.5. More examples were discussed in an earlier review by Williamson [229]. These proline-rich disordered regions are involved in various biological processes, including endocytosis [230], cell protrusion and mobility [231], transcription, immune response, and signal transduction as listed in Table 1.4 and Table 1.5. Table 1.3, Table 1.4 and Table 1.5 are compiled from a survey of literature.

Table 1.3: Modular domains, phosphopeptide-binding domains and their specificities. Table adapted by author [232].

Domain	Example proteins	Specificity	Reference
Modular domains			
SH3			
Class I	Src, Yes, Lyn, Abl, Grb2 A, PI3K, Fyn	(R/K)P ϕ P χ ϕ P	[233-237]
Class II	Src, Cortactin, p53BP2, PLC γ , Crk A, Amphiphysin, Nck SH3-B, CAP SH3-C	ψ PP ϕ P ϕ (R/K)	[232, 238-243]
WW			
Class I	YAP65, Nedd-4, Dystrophin, BAG3	(L/P)PP(Y/pY)	[244-251]
Class II	FBP-11	PPLPP	[252]
Class III (a)	FE65	(p/ ϕ)P(p/g)PPpR	[253]
Class III (b)	FBP21	(p/ ϕ)PP(R/K)gpPp	[254]
Class IV	Ess1/Pin1, Nedd-4	(pS/pT)P	[255-257]
Class V	PRP40-2	(p/ ϕ)PPPPP	[258]
EVH1			
Class I	Ena/VASP, Mena, Evl	(D/E)FP χ ϕ P	[259, 260]
Class II	Homer/Vesl	PP χ χ (F/Y)	[261, 262]
Class III	WASP/N-WASP	LPPPEP	[263]
Class IV	SPRED	Not Defined	[264, 265]
GYF			
	CD2-binding protein 2(CD2BP2)	(R) χ χ PP χ R	[266, 267]
UEV			
	Tsg101	P(T/S)AP	[268, 269]
Profilin			
	Profilin (single-domain protein)	Poly-L-proline	[270, 271]
Phosphopeptide-binding domains			
SH2	Src	pYEEI	[272]
PTB	SHC	NPpY	[273]
14-3-3	14-3-3 ζ	RSXpSXP or RXY/FXpSXP	[274]
WW Class IV	Pin 1	(pS/pT)P	[256]
FHA	Rad53 FHA 1	pTXXD	[275]
WD40	B-TrCP	DpSGXXpS	[276]
MH2	Smad2	SpSMpS-COOH	[277]
Polo-box	Plk1	S(pS/pT)P	[278]
BRCT	BRCA1	(pS/pT)XX(F/Y)	[279]

Note: The aliphatic, hydrophobic and any amino acid are represented by symbols ψ , ϕ and χ respectively. Phosphoserine and phosphothreonine, phosphotyrosine are represented by pS, pT and pY. Residues that are favored but not highly conserved are represented by lowercase letters. Among the multi copies of a modular domain in a protein, the one nearest to the N-terminus is represented as domain A.

Table 1.4: Proline-rich regions with repeated proline-rich motifs.

Name	Organism	Proline-rich motifs	Function	Position	References
Fc receptor	Streptococcus agalaciae	(T/S/A/I/L/VP)30	Binds peptidoglycan	Membrane anchor	[280]
LRP6	Human	PPP(S/T)PX(S/T)	Activation of Wnt/ β -catenine pathway	Intracellular domain	[216]
Acta (Actin assembly-inducing protein)	Listeria	FPPPP	Actin polymerization	N-terminus	[281]

Table 1.5: Proline-rich regions without repeated proline-rich motifs.

Name	Organism	Function	Position	Reference
SH3-binding protein 1	Mouse	Binds SH3 domains		[282]
Ig alpha-1	Human	Immunoglobulin	Linker	[283]
FAK (Focal adhesion kinase 1)	Human	Cell migration, adhesion and protrusion, etc.	Near C-terminus	[284]
BAG3 (BAG family molecular chaperone regulator 3)	Human	Anti-apoptotic		[285]
Wasp/N-Wasp	Human	Actin polymerization		[286, 287]
Ena/VASP	Human	Associated with actin		[288]
p53	Human	Tumor suppressor	N-terminus	[216]

A short sequence of two or more proline residues in a row in these proline-rich motifs frequently tend to adopt a left-handed poly-L-proline II helix (PPII) with $\phi=-78^\circ$ and $\psi=+149^\circ$ [289]. PPII helix is an extended structure with three residues per turn and all Xaa-Pro bonds in trans conformation. Another conformation called right-handed poly-L-proline I helix (PPI), is observed in solvents such as propanol and butanol [290]. Different from PPII, PPI has all Xaa-Pro bonds in cis conformation.

The PP II helix is a special conformation, which has continuous prolines constructing a hydrophobic strip around the helix. At the same time, the carbonyls on the backbone present hydrogen bonding sites. As a result, PPII helix is able to provide both an accessible hydrophobic surface and a hydrogen bonding site. This much more rigid and extended structure described as ‘sticky arm’ has been observed near the N and C termini of proteins [229]. In addition, they are commonly observed in the exposed surface of globular proteins. A discontinuous stretch of prolines can also form a PPII helix. As a matter of fact, PPII type structure can form in solution in peptides without the

proline [291]. The PPII conformation has been observed to be the dominant conformation of disordered protein regions, including both proline-rich and proline-free disordered regions [292, 293]. Continuous PPII helices do not appear as stretches in disordered regions as observed in experimental and simulation studies [294-296], however the majority of the amino acid backbone conformations are in a PPII conformation. PPII helix tends to exist in long extended proline-rich regions. Such regions are highly flexible, resistant to interactions with chaperones, thermo tolerant, and are difficult to characterize in the experiments of X-ray crystallography or NMR spectrometry [232, 297]. Crystallographers have found that it is best to remove these sequence regions from proteins in order to obtain crystal structures of folded protein domains connected to proline-rich disordered regions. Thus, a comprehensive understanding about the structural role of these long proline-rich disordered regions is needed.

Several studies claim that proline-rich disordered region have auto-inhibition effects in the entire protein structure. Kim *et al.* 2000 reported the inactivated structure of WASP (Wiskott - Aldrich syndrome protein), which is induced by the autoinhibitory contact between the GTPase-binding domain (GBD) and the proline-rich C-terminal region. Cdc42 can bind to the GBD and disrupt the hydrophobic core, resulting in the release of the C terminus, which is able to interact with actins. ALIX (apoptosis-linked gene 2 (ALG-2)-interacting protein X) is important in apoptosis, endocytosis and is also associated in ESCRT pathway [298]. The C-terminal of ALIX is a proline-rich region with a length of 150 residues and a percentage of 33% proline. ALIX proline-rich region can form a contact with another domain in ALIX called

Bro1 domain, leading to the autoinhibition of the binding activities of ALIX [299-301]. Interestingly, ALIX can also dimerize through its proline-rich region to mediate HIV-1 budding [302]. As a result, it seems that binding to other molecules depending on context (e.g. Cdc42 binds to GBD in WASP and ALIX dimerization) could relieve the autoinhibition effect induced by proline-rich regions.

Another structural role of high amount of proline locating in disordered regions may be to disrupt soluble oligomers [303], amorphous aggregates and amyloid fibrils which are harmful and associated in diseases, including Alzheimer's disease (amyloid β -protein (A β)), Parkinson's disease (α -synuclein aggregation), type II diabetes (amylin) and Huntington's disease (polyglutamine repeats) [201]. Almost all amyloid fibrils have some beta-strand type protein-protein interaction, which is disrupted by the presence of proline.

Unlike folded domains, the intrinsically disordered have stronger tendency to mis-fold, but they are protected by their natural sequence features comprising of high net charge, low hydrophobicity and the presence of proline [304]. Rauscher et al have shown that a high content of proline and glycine can help the elastomeric proteins escape the amyloid formation. Elastomeric proteins are a group of entropic chains with characteristics of intrinsic disorder, including elastin, spider silk, abductin, wheat gluten, and resilin [305].

The extraordinary roles of these proline-rich disordered regions in signaling pathways is in part due to the fact that they often contain phosphorylation sites and phosphor-amino-acid recognition motifs. Phosphorylation by kinases is of extreme importance in regulation and

signaling networks, as it changes the chemical nature of a short peptide motif to an altered state of structure and charge, facilitating binding to some recognition domain and forming a protein-phosphopeptide-motif interaction.

Many aspects of cellular biology, including DNA damage response, cell cycle control, differentiation and apoptosis [306-308], are regulated by the complex interplay of protein kinases and their substrates. It is estimated that about one third in the eukaryotic proteome is phosphorylated [309]. A conclusion has been arrived that these phosphorylation sites are likely to occur in disordered regions, because a significant enrichment of disorder-promoting residues is observed surrounding the phosphorylation sites (Ser, Thr or Tyr) and in particular proline is highly populated [310]. For example, the N-terminal region of p53 is proline-rich unfolded and two phosphorylation sites of ATM/ATR reside in this region [216, 217]; LRP6 intracellular domain is enriched in proline and serine (15-20%) containing five iterated phosphorylation motifs of kinase GSK3 β and CK1, and this region has no structure identified mainly because it refuses to crystallize [311]. Until now, numerous methods have been developed to identify the 3-10 residue long phosphorylation motifs or recognition motifs, many of which are in fact proline-directed, specifically identified by a kinase or a signaling modular domain (for example, SH2, SH3, PDZ and 14-3-3, and so forth) [312-314].

There are two major bioinformatic tools available online for searching phosphorylation sites or phosphopeptide recognition motifs. One is called NetPhos which is based on neural network algorithm [315]. The other one is Scansite which represents the motifs with the position-specific scoring matrix (PSSM) derived from oriented peptide library experiments [316, 317]. Such

scoring matrix is capable of indicating the preference for each amino acid residue occurring at that position relative to the phosphorylation site within the recognition motif. PSSM assumes independence between positions in the motif, as it calculates scores at each position independently from the residue type at other positions. Because the kinome involves a large number of members and the experimental results still contain redundant phosphorylation or recognition sites, the prediction accuracy still remains to be improved [318]. An alternative research area to study is the evolutionary rates of substrate sequences for kinases and for phosphopeptide-binding domains.

The Present Work

This thesis focuses on the proline-rich disordered protein regions in signaling and mechano-related pathways. We propose a hypothesis that the structural ensembles of proline-rich disordered protein regions will adopt an altered form that dominates their conformational space with facing the spatial constraints imposed by membrane, nearby molecules or molecular complexes in real cellular context. A structural simulation approach was developed to examine the consequences of spatial constraints, which may explain the biological relevance of conformational change occurring in the proline-rich disordered regions in the initiation of signaling pathway. The abundance of protein-rich disordered regions is high in the signaling pathways. This may largely be due to the functional sites they have in their sequence. The proline-rich regions usually contain multiple phosphorylation sites, with higher frequencies of serines and threonines (less tyrosines), whose activation is pivotal in the signal

transduction. Therefore, it is vital to locate these proline/serine-rich disordered regions in the proteomes associating with signaling and mechano-related pathways. We developed an amino acid propensity index as a tool to search for such regions. Finally, we take a deeper look into the interpositional sequence dependencies around the phosphorylation sites which have a preference to appear in the disordered regions. A fast and easy evolutionary rate for the substrate motif is observed. This agrees with the earlier report about the study of phosphorylation sites in the disordered regions which tend to have a fast evolutionary rate compared to the ordered folded domains [310]. The entire thesis involves various research fields, including structural simulation, biophysics, computer science, sequence analysis, mathematics and biostatistics.

In Chapter 2 of the thesis, the molecular simulation on the intracellular domain of a protein called LRP6 is described, which is a single-pass type I transmembrane protein in the Wnt signaling pathway [319]. The intracellular domain of LRP6 is intrinsically disordered and enriched in proline and serine [311]. Mutation of this segment has been reported to lead to inactivate the Wnt pathway [320]. A molecular simulation on this proline-rich region was carried out by constructing an initial large scale sampling of conformational space with TraDES followed by filtration with two spatial constraints, which are modeled as the excluded volume effects imposed by the plasma membrane and nearby molecules or molecular assemblies. We find that an elongated conformation described by an increased ensemble-averaged radius of gyration, dominates the structural ensemble of LRP6 intracellular domain. In particular, this elongation happens on the near-membrane domain, described by an

increased ensemble averaged end-to-end distance. In the Wnt signaling context, such a conformational change could be the reason why the phosphorylation motif closest to the membrane gets activated first and propagate the reactions further down to the phosphorylation sites behind, as discovered in reported experiments [320-322].

In Chapter 3, an algorithm was developed to search for proline/serine-rich disordered regions. A previous method designed to search for domain linkers was based on a domain-linker amino acid propensity index [323]. Originating from these simple log-odds and z-score approach, a new amino acid index called “PSR” was derived from the amino acid compositional bias between a set of Pro/Ser-rich disordered regions and a folded domain dataset. The amino acid index “PSR” is able to detect protein regions of interest with a high sensitivity and reasonably good specificity. The current disorder predictors only focus on finding the disordered regions without giving a clue about the particular compositional bias within their predictions. On the other hand, protein regions predicted by software (mainly SEG [129]) to contain composition bias or low sequence complexity, are not always disordered. Therefore, the “PSR” index provides an alternative insight into the sequence composition of disordered regions. This new method considers the previous finding that amino acid composition alone is sufficient to predict long disordered regions (>30aa) [163, 164]. The training dataset was constructed by manually collecting Pro/Ser-rich disordered regions with amino acid length of more than 40 residues. The prediction method searches for Pro/Ser-rich regions with a length of 50 residues in the training dataset. A web server called Armadillo (2.0) was implemented for the PSR index. It provides a

platform for researchers who are interested to find out more about the proline-rich disordered regions with multiple phosphorylation sites, most of which are indispensable in the signaling pathways.

Finally, in Chapter 4, the sequence information of the substrate motifs of several kinases was studied to learn more about the kinases' specificities towards their substrates, most of which have phosphoacceptors in proline-rich regions. The interpositional sequence dependencies surrounding the phosphorylation site were examined in large sets of sequences and found to be rare. Two models were designed using the first-order information and second-order information to predict novel substrates for three types of kinases. The first-order is based on the assumption that every position is independent of each other. The second-order model incorporated the interpositional sequence dependencies discovered from the substrate dataset we collected from experimental results. The second-order model did not provide much improvement in prediction power than the first-order model. The results suggest the position-independent model does well enough to predict novel phosphorylation sites. In order to achieve higher kinase specificity toward their substrates, other biological and cellular contexts should be included rather than considering the sequence information only.

The results also imply a fast evolutionary rate in the substrate motifs, which coincides with the evidence that phosphorylation sites are more often located in disordered regions that also evolve at a faster rate [310]. The substrate motifs indicated that kinase activity is confined to a small and smooth fitness energy landscape. This is different from the rough evolutionary landscape of fitness between phosphopeptide-binding domains and their

substrates, within which interpositional sequence dependence does exist. This difference between kinases and phosphopeptide-binding domains indicates that the kinases recognize their phosphorylation sites with less specificity than do phosphopeptide-binding domains binding to their substrates. Evolutionarily speaking, phosphorylation is a more random process than phospho-amino-acid recognition. A cellular context, for example, the structural accessibility, could regulate the specificity of a kinase on its substrates. In the case of auto-inhibition of a disordered segment, the inhibition may be caused by steric blockage of the phosphor-amino-acid, in one dominant ensemble configuration. Modification of the overall ensemble shape may relieve the auto-inhibition of the disordered regions by exposing the blocked or buried serines or threonines in the disordered ensemble for kinases to access easily.

Proline dipeptides have a natural tendency to adopt an extended PPII form [324, 325]. Autoinhibition may be induced by evolutionary changes such that this naturally extended form becomes more compact, either by long or short-range inter-molecular interactions. By definition, autoinhibition precludes intra-molecular interactions as cause and are inhibiting effects demonstrated in the absence of other molecules. Quite often autoinhibition is relieved by removing larger portions of the disordered sequence and revealing only a smaller portion of peptide with the phosphorylation motif. It follows, logically, that autoinhibition in a disordered protein region is caused by either a stable single structure, or by bulk properties of the disordered ensemble. In this thesis the latter is the focus. As disordered regions will have generally weak long or short-range inter-molecular interactions, they may be sensitive to the conformational space they are free to sample. While this may be difficult

to imagine, any spatial constraint on conformational shape, through proximity to nearby membrane or complexes, may alter the freedom of the disordered ensemble to stop it from adopting structural trajectories that lead to the autoinhibiting conformations, thereby preventing the autoinhibiting conformations from dominating the ensemble.

This concept of ensemble autoinhibition is novel. To better understand it, consider the analogy of a swordsman, defending himself from an attacker. A spatial constraint in this case would be like backing the swordsman into a corner. While the corner would not directly pin down the arms of the swordsman, it would restrict his dynamic freedom of movement and make him more vulnerable to a frontal attack. The historical example of the battle of Cannae offers another such analogy, where sword-bearing Roman soldiers were so tightly packed together that they could not lift or swing their swords to defend themselves, and were defeated by the smaller army of Hannibal.

There are several explanations for the enrichment of proline in disordered regions, including resistance to amyloid formation, resistance to chaperone binding, resistance to proteolysis, thermo tolerance, and the natural tendency of proline to adopt an elongated PPII configuration. Elongated configurations may provide the most accessible opportunity for kinase binding and phosphorylation, a concept tested in this thesis by examining the conformational space preferences of docked phosphorylation motif binding regions. The compositional bias towards proline may therefore be functional in the sense that the polymer properties of the disordered region need to be conserved within a certain compositional threshold to remain functional. In addition, this compositional bias limits the evolutionary rate of phosphopeptide-

binding substrates, many of which require a proline-directed position or even proline-rich relative to the phosphorylated residues.

In this thesis, a variety of computational methods have been applied to obtain a better understanding of the disordered regions enriched in prolines, in terms of analyzing the structural ensembles, sequence compositional bias, as well as phosphorylation motifs. Our hypothesis that proline-rich disordered regions tend to adopt an elongated or extended conformation when they encounter spatial constraints in a cellular context is demonstrated by conformational sampling. The simulation results provide information that so far cannot be measured by structure experiments and may help interpret the phenomena that remain unexplained from experiments. With computational tools, we were able to design an amino acid index “PSR” with the implementation of Armadillo (2.0) web interface for searching proline-rich disordered regions with multiple phosphorylation sites. The research provides a systematic view of these particular protein regions. Finally, with sequence analysis algorithms, the sequence preferences of short phosphorylation motifs were investigated, a great number of which are found within disordered proline-rich regions. From the perspectives of evolution, we showed that the interpositional sequence dependencies around the phosphorylation sites motifs are actually rare, which agrees with other findings about the fast evolutionary rate of the disordered regions where the functional sites are likely to reside in. Collectively, the results present new knowledge to help explain some of the outstanding mysteries underlying the sequence, structure and function of phosphorylatable, proline-rich disordered regions.

Chapter 2

The Effect of Spatial Constraints on An Ensemble of Proline-Rich Disordered Structures¹

Abstract

LRP6 is a membrane protein crucial in the initiation of canonical Wnt/ β -catenin signaling. Its function is dependent on its proline-serine rich intracellular domain. LRP6 has five PPP(S/T)P motifs that are phosphorylated during activation, starting with the site closest to the membrane. Like all long proline-rich regions, there is no stable 3D structure for this isolated, contiguous region. In our study, we use a computational simulation tool to sample the conformational space of the LRP6 intracellular domain, under the spatial constraints imposed by (a) the membrane and (b) the close approach of the neighboring intracellular molecular complex, which is assembled on Frizzled when Wnt binds to both LRP6 and Frizzled on the opposite side of the membrane. We observe that an elongated form dominates in the LRP6 intracellular domain structure ensemble. By docking simulation we show that kinases prefer elongated substrates, and that this elongation could relieve conformational auto-inhibition of the PPP(S/T)PX(S/T) motif binding sites and allow GSK3 and CK1 to approach their phosphorylation sites, thereby activating LRP6 and the downstream pathway. We propose a model in which

¹ Portions of the work written in this chapter have been previously published as:

Chengcheng Liu, Mingxi Yao and Christopher WV Hogue: Near-membrane ensemble elongation in the proline-rich LRP6 intracellular domain may explain the mysterious initiation of the Wnt signalling pathway. *BMC Bioinformatics*. 2011, **12**(Suppl 13): S13

the conformation of the LRP6 intracellular domain is elongated before activation. This is based on the intrusion of the Frizzled complex into the ensemble space of the proline-rich region of LRP6, which alters the shape of its available conformational space. To test whether this observed ensemble conformational change is sequence dependent, we did a control simulation with a hypothetical sequence with 50% proline and 50% serine in alternating residues. We confirm that this ensemble neighborhood-based conformational change is independent of sequence and conclude that it is likely found in all proline-rich sequences. These observations help us understand the nature of proline-rich regions which are both unstructured and which seem to evolve at a higher rate of mutation, while maintaining sequence composition.

2.1 Background

Wnt induced signaling pathways play essential roles in development and disease [326-328]. Currently, two initiation models of the canonical Wnt/ β -catenin signaling pathway have been proposed as illustrated in Figure 2.1 [329-331]. One could be referred to as the sequential recruitment/amplification model, in which Wnt stimulation is proposed to recruit the scaffold protein AXIN to approach the membrane through the bridging interactions between frizzled (FZD) and dishevelled (DVL), as well as between DVL and AXIN. GSK3 (glycogen synthase kinase 3) in association with AXIN thereafter is able to phosphorylate the LRP5/6 PPP(S/T)P motif near the membrane. Initial phosphorylation creates a docking site for AXIN and thereby recruits additional AXIN-GSK3 to promote further LRP6 phosphorylation [332]. The

second model is the signalosome/aggregation model. Recent results showed that a signalosome is formed by aggregated LRP6 and AXIN when Wnt is present. Clustering of LRP6 leads to the phosphorylation of T1479 by CK1 γ (casein kinase 1 γ) and subsequent phosphorylation of the PPP(S/T)P motif by GSK3 [321]. Phosphorylated LRP6 recruits AXIN resulting in the disruption of the “ β -catenin destruction complex”, which comprises of AXIN, APC (the tumor suppressor *Adenomatous polyposis coli*), GSK3 and CK1 α [333, 334]. This results in the stabilization of a cytoplasmic pool of β -catenin. Free β -catenin enters the nucleus and activates gene transcription by binding to the TCF/LEF (T cell factor/Lymphoid enhancer factor) family of transcription factors [335-337]. Thereafter, the activation of LRP5/6 is indispensable to initiate the downstream intracellular Wnt signaling cascade, in order to stabilize β -catenin.

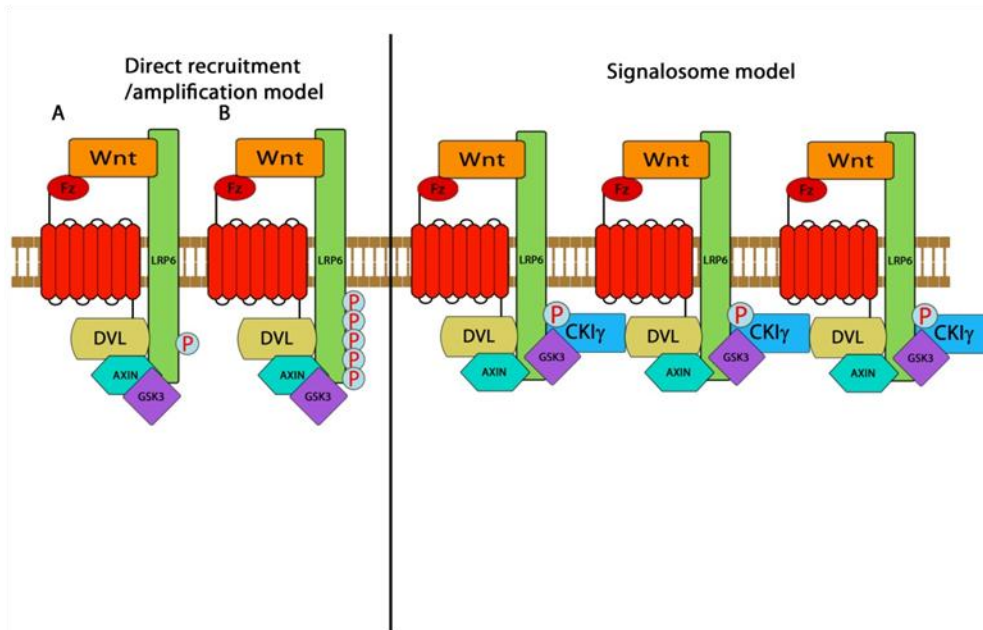


Figure 2.1: Two proposed initiation models of canonical Wnt/ β -catenin signalling pathways. In the sequential recruitment/amplification model (left), Wnt-induced FZD-LRP6 complex formation promotes initial LRP6 phosphorylation via DVL recruitment of the AXIN-GSK3 complex. Initial LRP6 phosphorylation provides docking sites and thereby recruits additional AXIN-GSK3 complex to promote further LRP6 phosphorylation if LRP6 multimerizes. In the signalosome/aggregation model (right), Wnt induces clustering of LRP6, leading to its phosphorylation by CK1 and subsequently by GSK3 and recruitment of AXIN.

LRP6/LRP5/Arrow belongs to a subfamily of LDL receptors (LDLR) [319]. Human LRP6 is a type I single-pass transmembrane protein. Its modular extracellular domain has three basic domains; the YWTD (tyrosine, tryptophan, threonine and aspartic acid)-type β -propeller domain, the EGF (epidermal growth factor)-like domain, and the LDLR type A (LA) domain. This region has crystal structures present in PDB database [338-340]. During signaling pathway initiation, Wnt binds the cysteine-rich domain of FZD proteins and exhibits a Wnt1-dependent association with LRP6 extracellular domains *in vitro* [341, 342]. However, the interaction between Wnt and LRP6

is weaker compared to the interaction between Wnt and FZD [341]. It is therefore more likely that a Wnt-FZD complex binds to the LRP6 extracellular domain. After deletion of its extracellular domain, the LRP6 protein is still capable activating the Wnt/ β -catenin signaling pathway [343].

The LRP6 intracellular domain is rich in proline (~15%) and serine (~20%). Sequence alignment shows that it includes a S/T cluster and downstream five reiterated PPP(S/T)PX(S/T) motifs, each of which contains a PPP(S/T)P motif phosphorylated by GSK3 and juxtaposed to a CK1 phosphorylation site [311]. Such dual phosphorylation is essential to stabilize the pool of β -catenin in the cytoplasm [344]. The phosphorylation of the S/T cluster has also been characterized, particularly the phosphorylation of T1479 by CK1 γ [321, 322]. It is believed that the phosphorylated S/T cluster promotes downward PPP(S/T)PX(S/T) phosphorylation [320]. He's group previously showed that a LRP6 mutant lacking most of the intracellular domain is a loss-of-function [341]. In addition, a truncated LRP6 comprising its transmembrane and intracellular domain is constitutively active in Wnt signaling transduction [345-347]. A single PPP(S/T)P motif transferred to a LRP6 variant lacking the extracellular domain activates the Wnt/ β -catenin signaling pathway. Phosphorylated PPP(S/T)PX(S/T) motifs provide docking sites for AXIN, which associates with GSK3 to promote proximity to LRP6 [348].

So far, no stable structure has been obtained from this isolated and contiguous LRP6 intracellular domain in current structure databases. The LRP6 intracellular domain is expected to be natively unstructured (unfolded or

disordered) since its composition is enriched with proline whose conformation is limited [311, 329, 338, 349]. There has been little study on the conformational behavior of LRP5/6 before activation, when Wnt induces signal transduction.

No matter which initiation model applies to the canonical Wnt/ β -catenin signaling pathway, the conformation of LRP6 has to face spatial constraints imposed by (i) the plasma membrane and (ii) a nearby molecule or molecular assemblies, which could be neighboring LRP6 molecules or a Wnt-FZD-DVL-AXIN-GSK3 assembly. We hypothesize that these two spatial constraints would restrict LRP6 intracellular domain conformational space to cause its conformation to adopt a more extended or elongated form before it is activated and docked by AXIN.

The TraDES software package was used [177] to sample the conformational space of the LRP6 intracellular domain with an initial ensemble of over 390,000 structures, then spatial constraints were applied to determine whether a change in ensemble structure may be observed as a result of neighboring molecular complexes moving closer due to Wnt/ β -catenin signaling pathway initiation. The same system of constraints was also studied to determine whether it applies to observable structural change in an ensemble of unfolded states in a sequence independent manner by using a sequence of uniform composition of alternating proline/serine residues of the same length as the intracellular region of LRP6.

2.2 Results

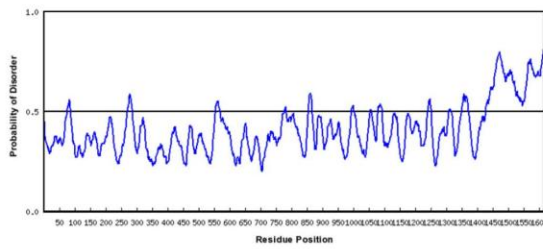
2.2.1 LRP6 intracellular domain is predicted to be unfolded.

No stable structure has been documented for the LRP6 intracellular domain in current structure databases. This region is expected to be natively unstructured because it is enriched with proline and serine. Several protein disorder prediction tools predict that this region is disordered or unfolded. Figure 2.2 gives the prediction results from disorder predictors; RONN [116], IUPred [122], Globplot [121], PONDR-FIT [102] and FoldIndex [139]. This unfolded intracellular protein region most probably tends to have random coiled conformations, which auto-inhibits the structure itself to avoid interactions with other molecules [350]. Like most other disordered protein regions, it exists as an ensemble of structures which can be generated by TraDES in this simulation study.

2.2.2 Radius of gyration distribution

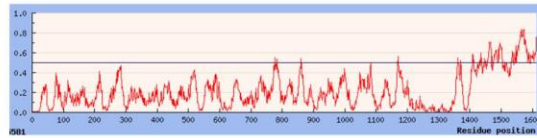
Radius of gyration (Rgyr) measures the openness of the whole structure. A structure with a larger Rgyr has more sparse atoms within it. Figure 2.3 displays the Rgyr distribution of the initial conformational ensemble (before filtration) in the LRP6 intracellular domain simulation experiment. The number of generated conformers and average Rgyr are provided in the second column in Table 2.1. Conformers with different values of Rgyr were checked. It was observed that conformers with smaller Rgyr have more compact structures; while, conformers with larger Rgyr tend to adopt more open or extended conformations. In this distribution, the conformation of the generated

conformers changes from compact, to more open, to more extended as their R_{gyr} increases.



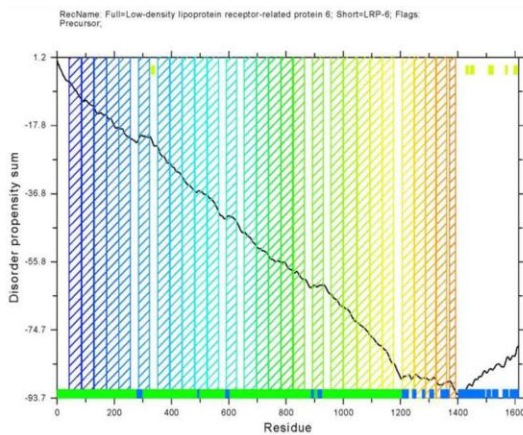
RONN of LRP6

Region 1427-1613 is predicted as disordered.



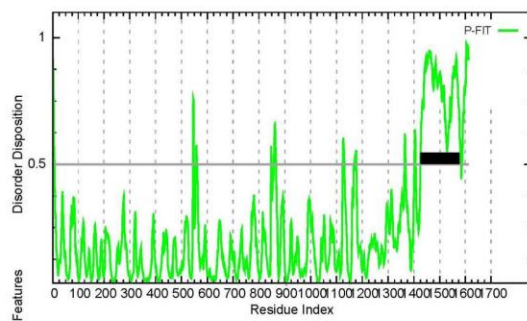
IUPred of LRP6

C terminus (LRP6 intracellular domain) is predicted as disordered.



Globplot of LRP6

Regions 1402-1496, 1501-1516, 1521-1541, 1558-1577, and 1583-1613 are not reliably predicted as globular domains.



PONDR-FIT of LRP6

Region 1425-1613 is predicted as disordered (thick black line).

FoldIndex of LRP6

Region 1486-1613 is predicted as unfolded (red label).



Figure 2.2: Analysis of the human LRP6 protein [Swiss-Prot:O75581] using different predictors. The graphical output of each method and the corresponding interpretation is shown. The precise boundaries of ordered and disordered regions were derived from the corresponding text output (not shown). The intracellular domain is unfolded, whereas the extracellular domain is folded/ structured.

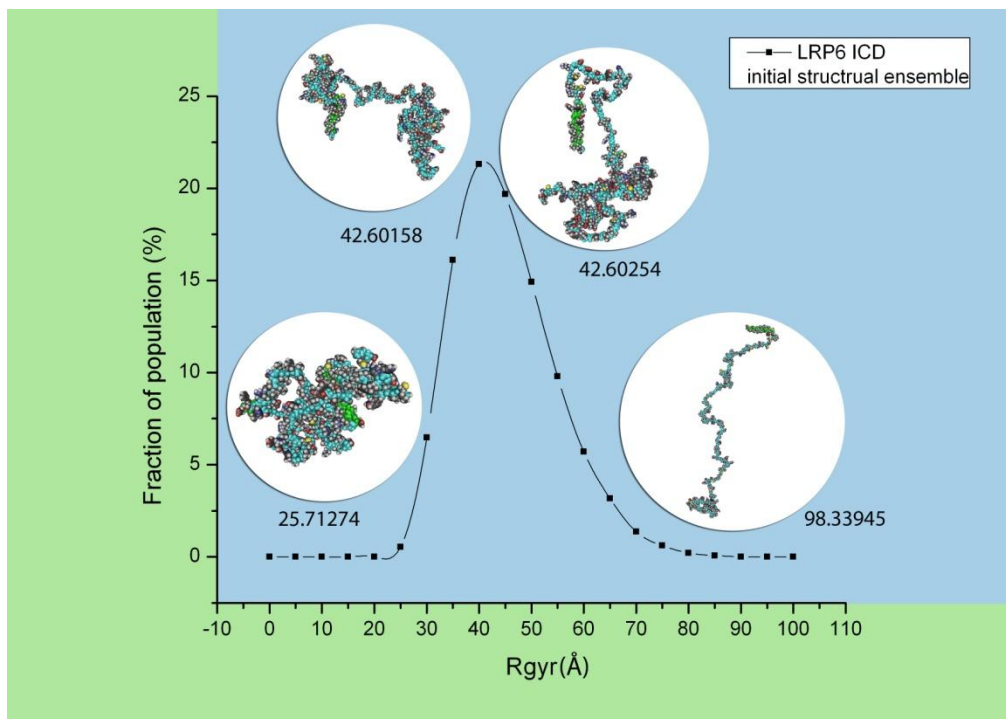


Figure 2.3: Rgyr distribution of the initial conformational ensemble before filtration. Conformers shown in the graph are some examples in the initial conformational ensemble of LRP6 intracellular domain. The number below each conformer is the value of its radius of gyration. A conformer with a smaller value of radius of gyration has a compact conformation (the structure on the left). A conformer with a larger value of radius of gyration has an extended conformation (the structure on the right). Two conformers with a mean value of radius of gyration are shown in the middle.

Simulation experiments were carried out both for the LRP6 intracellular domain (ICD) and the control sequence. See results in Table 2.1 and Table 2.2. A Unix script was written to obtain 10000 conformers that pass Constraint 1 and Constraint 2 ($\delta = 20.0 \text{ \AA}$) out of the initial structural ensemble. These 10000 conformers were then filtered by Constraint 2 with parameter δ set to 5.0 \AA and 10.0 \AA . This parameter represents the distance from the vertical plane to the plane defined by the transmembrane helix and origin point (0,0,0). The average Rgyr of the structural ensemble gets larger after each constraint is applied, as shown in Table 2.1, Table 2.2 and Figure 2.4, while the ensemble size decreases. In both the LRP6 intracellular domain and control sequence simulation experiments, after each filtration, conformers in the structural ensemble surviving from the spatial constraints tend to possess more open or extended conformations. This was also indicated by the observation that after each constraint more fractions of structural ensemble appear to have Rgyr larger than the average (i.e. the Rgyr distribution curves of structural ensembles after Constraint 1 and Constraint 2 shift to the right of the Rgyr distribution curve of the initial structural ensemble). In addition, when the distance δ gets smaller, the average Rgyr in the structural ensemble that survived Constraint 2 gets bigger, while the number of structures that passes the constraint decreases.

Table 2.1: Rgyr simulation results for LRP6 intracellular domain

LRP6 ICD simulation	Initial structural ensemble	Structural ensemble after Constraint 1	Structural ensemble after Constraint 2		
			$\delta=20.0$	$\delta=10.0$	$\delta=5.0$
No. Structures	396339	36025	10000	4939	2192
Average. Rgyr (Å)	42.57	46.43	47.43	48.36	48.73
Minimum. Rgyr (Å)	20.73	23.12	23.12	26.18	26.18
Maximum. Rgyr (Å)	98.60	95.02	95.02	95.02	95.02

Control sequence simulation	Initial structural ensemble	Structural ensemble after Constraint 1	Structural ensemble after Constraint 2		
			$\delta=20.0$	$\delta=10.0$	$\delta=5.0$
No. Structures	181833	33556	10000	5632	3276
Average. Rgyr (Å)	44.15	47.94	48.73	49.63	59.78
Minimum. Rgyr (Å)	20.43	22.20	23.89	24.74	27.15
Maximum. Rgyr (Å)	101.04	101.04	93.77	93.77	92.39

Table 2.2: Rgyr simulation results for control sequence. The control sequence has LRP6 transmembrane region and poly(Pro-Ser)₅₀ polypeptide substituting LRP6 intracellular domain. The theoretical maximum Rgyr for a completely extended chain (a polypeptide containing 100 Gly residues, i.e. poly(Gly)₁₀₀ polypeptide) is calculated from TraDES to be around 90Å. The poly(Gly)₁₀₀ polypeptide is constrained to take a β -strand conformation in its trajectory distribution file (Phi = -119; Psi = 113; Peak Magnitude = 100).

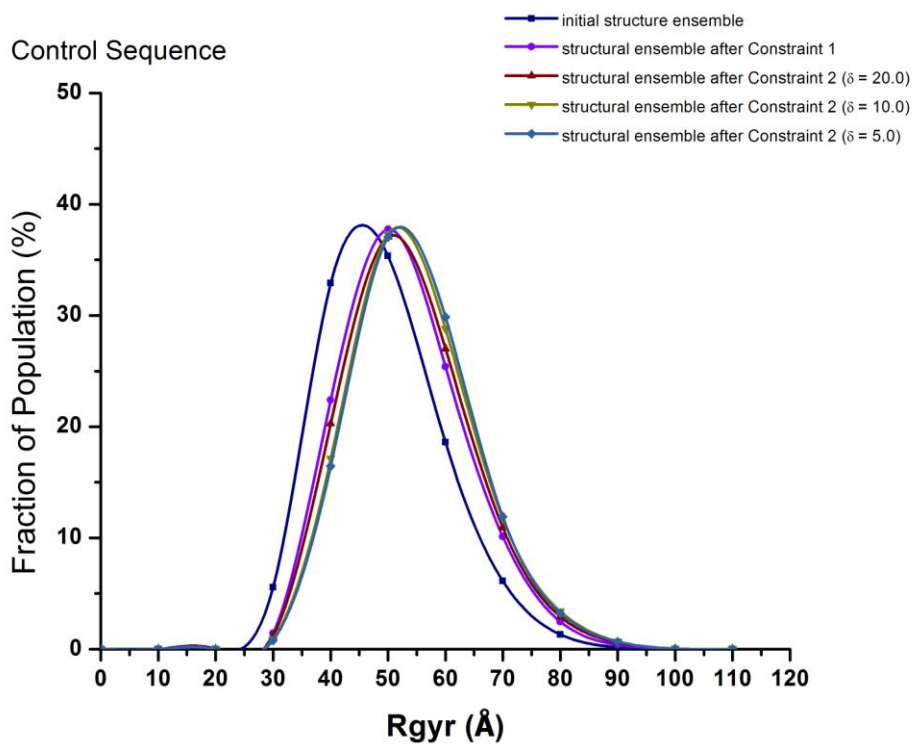
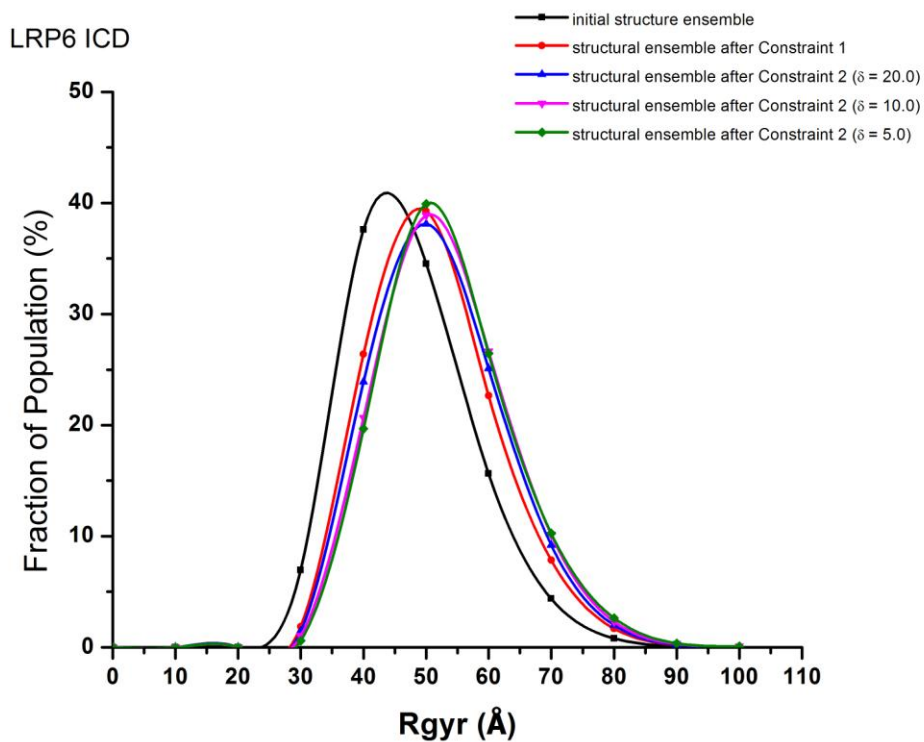


Figure 2.4: Rgyr distributions of LRP6 ICD and control sequence. The distance δ is set to 20.0Å, 10.0Å, and 5.0Å.

2.2.3 End-to-end distance distribution

To examine local effects in the ensemble, five different end-to-end distances with equal length were calculated in the LRP6 intracellular domain simulation experiment. Each distance contains at least one conserved PPP(S/T)PX(S/T) motif. Table 2.3 lists the exact description of the distance endpoints. For each end-to-end distance, the difference between the average Rgyr of the initial structural ensemble and that of the structural ensemble after Constraint 1 is provided in the column titled as “ $\Delta_{\text{mean}}(\text{Constraint1})$ ”. The difference between the average Rgyr of the initial structural ensemble and that of the structure ensemble after Constraint 2 under different values of distance δ are shown in the columns under the title of “ $\Delta_{\text{mean}}(\text{Constraint2})$ ”. For both the LRP6 intracellular domain and control sequence, out of the five end-to-end distances, the distribution of D1 displays largely increased mean values of the structural ensembles after each constraint. This was indicated by the positive differences in Table 2.3. It also shows that after each constraint is applied, the average value of D1 gets larger. In Figure 2.5, the D1 distribution curves of structural ensembles after Constraint 1 and after Constraint 2 move to the right of the D1 distribution curve of the initial structural ensemble. Meanwhile, more fractions of structural ensemble after each constraint are found at a larger value of D1 on the distribution curves. This indicates that the region corresponding to D1, within the LRP6 intracellular domain conformers in the surviving structural ensemble, occupies the constrained conformational space by adopting a preferred elongated or extended statistical conformation. This region starts right from the beginning of the LRP6 intracellular domain and

extends to the end of the first PPP(S/T)PX(S/T) motif. It is the closest membrane region inside the LRP6 intracellular domain. Additionally, as the distance δ gets smaller, the mean of D1 of the structural ensemble after Constraint 2 also gets bigger. However, for both the LRP6 intracellular domain and the control sequence, the rest of the four end-to-end distances' (D2-D5) distributions show no prominent changes compared to the change in D1 distribution after applying each constraint and furthermore they overlap with each other as seen in Figure 2.6. Thus the spatial constraint only changes the local ensemble statistical conformation at the near-membrane region spanned by D1.

	Length	Motifs contained	start-end	Δmean (Constraint1) (\AA)	Δmean (Constraint2) (\AA)		
					$\delta=20$	$\delta=10$	$\delta=5$
D1	103aa	S/T cluster; Motif 1	24-126	12.7220	17.5067	21.9135	24.2415
D2	103aa	Motif 1&2	64-166	0.8405	1.9484	2.4047	2.7648
D3	103aa	Motif 1,2&3	106-208	-1.3586	-2.7764	-2.8862	-3.0856
D4	103aa	Motif 2,3&4	124-226	-1.6533	-3.5066	-3.4477	-3.5721
D5	103aa	Motif 2,3,4 &5	141-243	-1.4545	-3.1360	-3.1388	-3.3186

Table 2.3: End-to-end distance simulation results for LRP6 intracellular domain. The table shows the length, motif contained, starting and ending positions in LRP6 simulation sequence, and the differences between the average Rgyr of structural ensembles after each constraint and the initial structural ensemble.

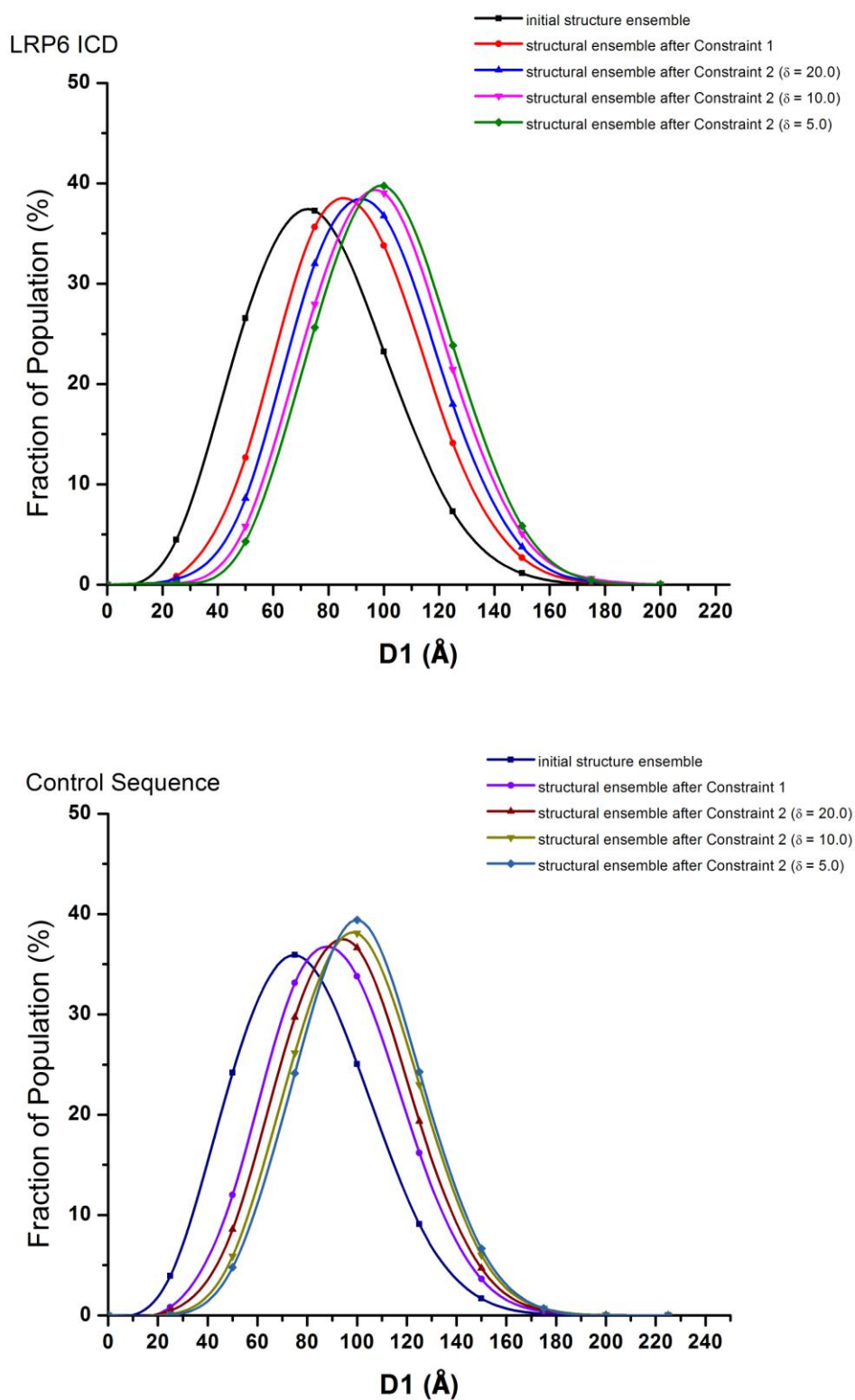


Figure 2.5: End-to-end distance distributions of D1 for LRP6 ICD and control sequence. The distance δ was set to 20.0Å, 10.0Å, and 5.0Å.

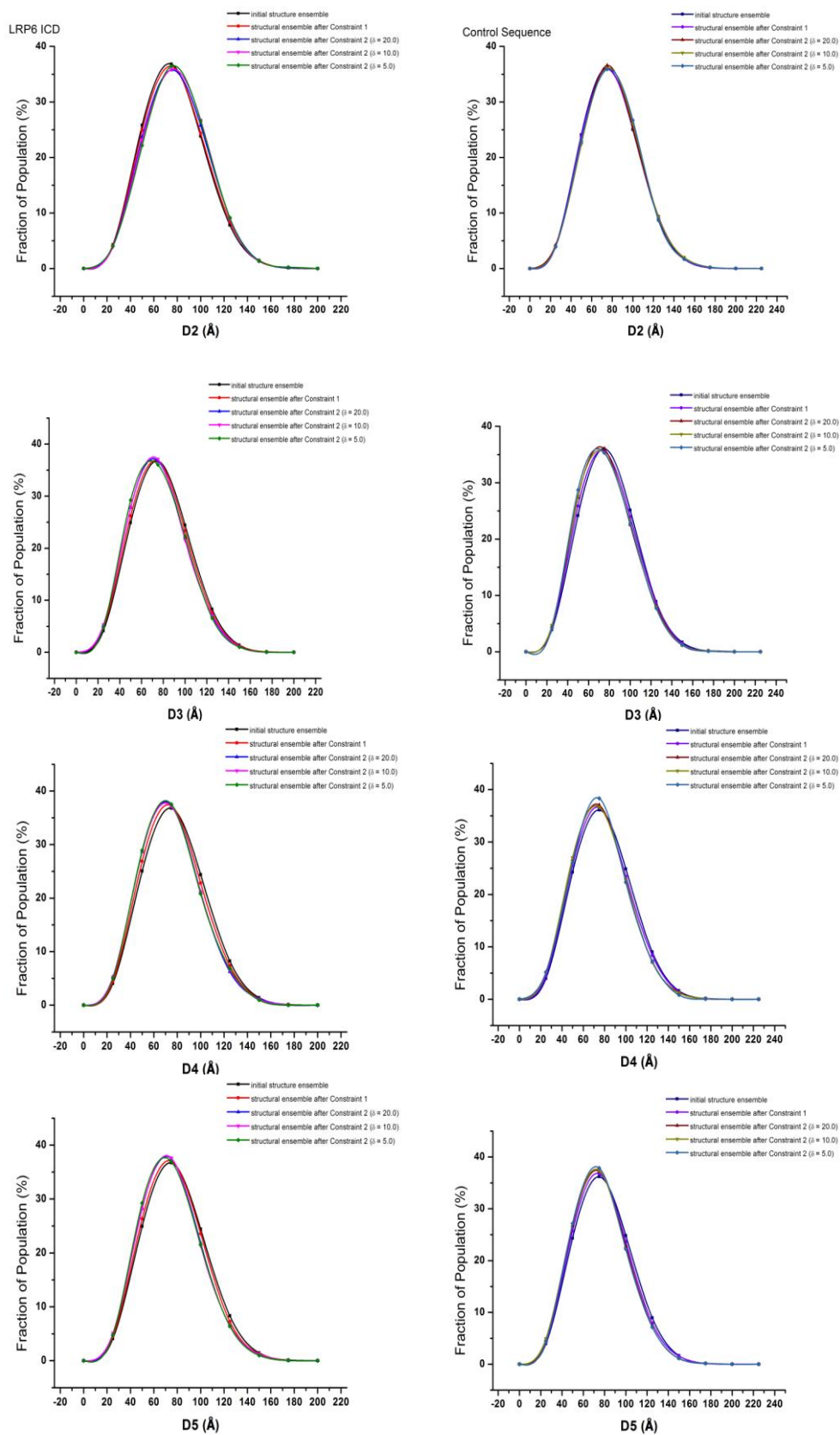


Figure 2.6: End-to-end distance distributions of D2, D3, D4 and D5 for LRP6 ICD and control sequence. The distance δ was set to 20.0Å, 10.0Å, and 5.0Å

2.3 Discussion

2.3.1 LRP6 intracellular domain structure ensemble favors an elongated form when the Wnt/ β -catenin canonical pathway initiates.

In the LRP6 intracellular domain simulation experiment, greater proportions of the structural ensemble are observed to have R_{gyr} of a larger value than the average after each spatial constraint is applied, in comparison with the initial structural ensemble (Figure 2.4). It shows that the two spatial constraints make the LRP6 intracellular domain likely to adopt a more open or elongated global conformation. The plasma membrane and neighboring assemblies formed by Wnt-FZD-DVL-AXIN-GSK3 or neighboring LRP6 aggregation could limit the LRP6 intracellular domain to form fewer numbers of random coiled structures. Instead, the LRP6 intracellular domain tends to form more elongated conformations as the spatial constraints exclude its volume near the membrane in the cell.

The implications of this novel observation are that *in vivo*, plasma membrane and nearby assemblies or molecules could result in a reduction in conformational space forcing an environment that forces the ensemble LRP6 disordered states into a more elongated population when a Wnt signal triggers the pathway. Such elongation behavior might grant kinases CK1 and GSK3 open access to the phosphorylation sites within the LRP6 intracellular domain, and subsequently LRP6 could be activated through these phosphorylation events. We propose that when the Wnt pathway initiates, the LRP6 intracellular domain is elongated to reduce the auto-inhibition before it is activated.

A statistically observable conformational change occurs to the LRP6 intracellular domain structural ensemble after applying spatial constraints. It is intriguing to investigate if the overall conformational change is localized within the LRP6 intracellular domain during Wnt canonical pathway initiation. The distributions of the five calculated end-to-end distances could reflect the openness of the subsequences in the LRP6 intracellular domain. The first end-to-end distance D1, which measures the openness of the region that is closest to the membrane on the LRP6 intracellular domain. The distribution curves for D1 show that this region gets longer in more conformers out of the structural ensemble after constraints are applied (Figure 2.5). This shows that the near-membrane region in the LRP6 intracellular domain elongates or extends when its conformational space is limited by the plasma membrane and nearby assemblies or molecules. Such an extended conformation could allow CK1 to more easily reach the S/T cluster and initiate phosphorylation. This may also explain the experimental finding that S/T cluster phosphorylation by CK1 promotes the downward activation of the PPP(S/T)PX(S/T) motif [8, 23, 24]. On the contrary, the end-to-end distances D2, D3, D4, and D5 hardly show any changes in their distribution curves between original and filtered structural ensembles (See Figure 2.6). The means of these distributions of structural ensemble after filtration are in fact smaller than that of the initial structural ensemble. This suggests the regions corresponding to these distances are on average less extended in the conformers surviving from filtration. The protein regions corresponding to the five end-to-end distances are gradually further away from the transmembrane helix, which determines the location of the plasma membrane. The region corresponding to D1 is the closest to the

plasma membrane followed by D2. The observations on the distribution curves of these distances suggest that the spatial constraints exclude to a great extent the volume of the LRP6 intracellular domain at the near-membrane location in the cell.

Additionally, the same behaviors are observed in Rgyr distributions and end-to-end distributions in the simulation experiment as for the control sequence with simple alternating proline/serine composition. The structural changes can be demonstrated in a hypothetical sequence with as much as 50% proline and 50% serine. This indicates that the observed elongation property is not strictly a feature of the LRP6 sequence itself, but may be a general feature of proline-serine rich disordered regions. This may represent the biophysical basis by which such regions can tolerate high levels of mutation whilst conserving approximate composition without affecting the underlying function of the disordered region. Hence, it can be concluded that such an elongation process induced by membrane and neighboring assembly/aggregation is sequence independent but it is likely compositionally dependent. Further studies with more amino acid control sequences would be required to determine if this is a property restricted to proline/serine-rich disordered regions, or common to all disordered sequence compositional variants.

2.3.2 Effects of the two spatial constraints

The observation of the near-membrane effect serves as the key finding in this simulation study. The membrane-anchor issue has been discussed in several published papers that claimed the LRP6 intracellular domain needs anchoring to the membrane to process signaling [338]. Arrow/LRP5/LRP6 mutants without the extracellular domain with which to anchor to the membrane constitutively activate the β -catenin pathway in mammalian cells. The LRP5 intracellular domain is unable to activate the signaling pathway unless it is anchored to the membrane [345-347]. In the simulation, the horizontal plane mimics the constraint imposed by the membrane plane. The vertical plane mimics the constraint imposed by nearby assemblies or molecules. Experiments show that the components in the assembly, for example, DVL, AXIN and GSK3 accumulate near the membrane when Wnt interacts with FZD and initiates the pathway [332]. Furthermore, CK1 that is responsible for the S/T cluster phosphorylation events is a near-membrane kinase [322, 351].

The second constraint also occurs near the membrane. If the second constraint is more stringent and the vertical plane gets closer to the conformer (i.e. δ gets smaller), the spatial volume of the conformer is excluded more. Such an excluded volume effect forces the conformer of the LRP6 intracellular domain to go through an elongation process. We propose this elongation might be necessary for the phosphorylation of the LRP6 intracellular domain.

Liu and colleagues [347] demonstrated that a truncated LRP6 comprising of its transmembrane and cytoplasmic domains is expressed as a constitutively active monomer whose signaling ability is inhibited by forced dimerization. Also, Wnt is shown to activate canonical signaling through LRP6 by inducing an intracellular conformational switch which relieves allosteric inhibition imposed on the intracellular domains. This paper published in 2003 is the only one until now on the conformational behavior of the LRP6 intracellular domain through experiments. There is however no evidence to prove such a conformational switch in terms of indicating the changes in the LRP6 intracellular domain structural ensemble. In the paper published by Yasui *et al.* [352], the authors conclude that the LRP6 extracellular domain does not form homodimers in solution and speculate that weak dimerization may occur only at the cell surface where the receptors are confined in the 2D plane. In our current simulation study, we focus on the conformational change of LRP6 intracellular domain under spatial constraints in the initiation of the canonical Wnt signaling pathway. Our results show that the spatial constraints cause the structural ensemble of the intracellular domain to adopt an extended or elongated form which relieves the allosteric inhibition. This provides another explanation for why wild-type LRP6 and LRP6 mutant without an extracellular domain behave differently with or without the presence of Wnt. The LRP6 mutant without an extracellular domain is free from the auto-inhibitory effect imposed by its extracellular domain. The LRP6 intracellular domain anchored to the plasma membrane only faces the spatial constraint caused by the plasma membrane. It can adopt a more open or elongated conformation to relieve the auto-inhibition caused by this unfolded

region itself, to allow CK1 and GSK3 access. For wild-type LRP6, without the presence of Wnt, membrane constraint is not enough to relieve the auto-inhibition caused by its extracellular and intracellular domains. It requires another constraint to relieve the auto-inhibitory effect caused by the extracellular domain. When Wnt is present, it forms a complex with FZD and interacts with the LRP6 extracellular domain. Though this interaction may be weak, the conformational space of the LRP6 intracellular domain is excluded. The domain is therefore forced to adopt a more open or extended structure for it to reduce the tension. Wnt-FZD hence imposes another spatial constraint to LRP6. In the initiation complex, Wnt is not the only component; FZD, DVL, AXIN and GSK3 also participate in the process. Hence, they together could form the second spatial constraint on LRP6 to amplify the effect of auto-inhibition. Such amplification would be helpful to the activation of the LRP6 intracellular domain and the stabilization of β -catenin. Our model and results can help explain the results obtained by Liang *et al* who recently discovered that the previously functional unknown protein TMEM198 is able to promote LRP6 phosphorylation in the Wnt signaling pathway. TMEM198 functions as a membrane scaffold protein, assembling kinases and substrates into a higher-molecular-weight complex prior to phosphorylation, but it promotes LRP6 phosphorylation through a mechanism independent of FZD and DVL [353]. Like FZD, TMEM198 could recruit CK1 as well as other molecules to form a nearby molecular assembly close to LRP6. Any nearby assembly, together with the membrane, can impose the spatial constraints to the conformational space of the LRP6 intracellular domain so that this region will be elongated for kinases CK1 and GSK3 to gain easy access for phosphorylation. Liang *et*

al. observe TMEM198 to associate with LRP6, however, unlike FZD, the interaction is likely mediated by the transmembrane domains between LRP6 and TMEM198 which can bring the TMEM198-CK1 complex more close to LRP6. The findings in Liang *et al.* also demonstrate that near-membrane is the key point in the simulation model. The interaction between TMEM198 and LRP6 at transmembrane domains takes place at the membrane plane. It amplifies the vertical spatial constraint by recruiting CK1 which is near-membrane localized. The spatial constraints can come from any nearby molecules or molecular assemblies. These include neighboring LRP6 molecules in the signalosome/aggregation model, Wnt-FZD-DVL-AXIN-GSK3 assembly in the sequential/amplification model or other discovered molecular assemblies such as the TMEM198-CK1 complex reported in the paper by Liang *et al.*

2.3.3 Elongation makes the phosphorylation of unfolded protein regions easier.

In this chapter, it is proposed that the elongation form may be required for the phosphorylation of the LRP6 intracellular domain. This can be further examined by docking sampled ensemble conformations of the region around the phosphorylation motif to a Kinase structure to determine whether elongated forms are preferred. Since there is no structure for the LRP6 intracellular domain present in the structural databases and it is very hard to get the kinase-substrate crystal structures experimentally, we used a homologous structure complex [PDB:1CMK] to demonstrate that in a general case, elongation is required for phosphorylation taking place in the

conformation of an unfolded protein region. [PDB:1CMK] contains a cAMP dependent protein kinase catalytic subunit and its inhibitor, a 22aa long peptide binding to the kinase catalytic site. The sampling procedure can be referred in the method section of this chapter. A 100mer sequence was constructed that includes the binding peptide in the middle and a repeated proline-serine extending to both termini. We used TraDES to generate a structural ensemble with the constructed sequence. Out of the 347426 conformers generated, 10000 passed the aligning, merging and crashe-checking requirements. These stages systematically remove ensemble structures that have steric clashes with the kinase. These surviving conformers of the phosphorylation motif region are available for docking and are not sterically autoinhibited. It is noted that there is an order of magnitude reduction in the number of sampled conformers after removal of steric clashes. This indicates that Kinase binding is only sterically compatible with a reduced subset of the substrate protein conformational space, compared to that of the unbound substrate. Figure 2.7 shows some examples of the sampled substrate conformers that are available and unavailable for docking. We calculated the Rgyr and end-to-end distances of Region 1-40, Region 31-70, Region 61-100 along the sequence and compared the distribution curves between 10000 conformers that are available for docking and 337426 conformers that are sterically unavailable for docking (Figure 2.8). We used a t-test to see whether there is a difference between the means of the datasets (Table 2.4). For Rgyr and end-to-end distance of the three regions, the p-values are significantly small ($p < 2.2e-16$), indicating that the structures available for the kinase to access and bind have, on average significantly larger Rgyr and longer end-to-

end distances. We can therefore say that the structural ensemble of a disordered proline-serine-rich sequence containing a kinase motif is required to transition to a more open and elongated form for binding to a generic kinases without steric autoinhibition. This proposition fits as elongation can reduce the auto-inhibition of the unfolded protein's random coiled structure. As we employed a generic alternating proline/serine sequence in this computer experiment, we can generalize that kinase homologue structures require an elongated form of the phosphorylation site region in order to carry out binding and phosphorylation. As we will see in Chapter 4 the substrate specificity of the kinase does not go very far beyond the phosphorylated residue, and there is little or no specific contact between a kinase and residues upstream or downstream of the substrate amino acid. Therefore it is unlikely that the kinase itself will induce a conformational change by specifically binding to the length of the substrate protein observed to span the elongated form. Instead it seems probable that kinase may have to wait until the elongated form presents itself.

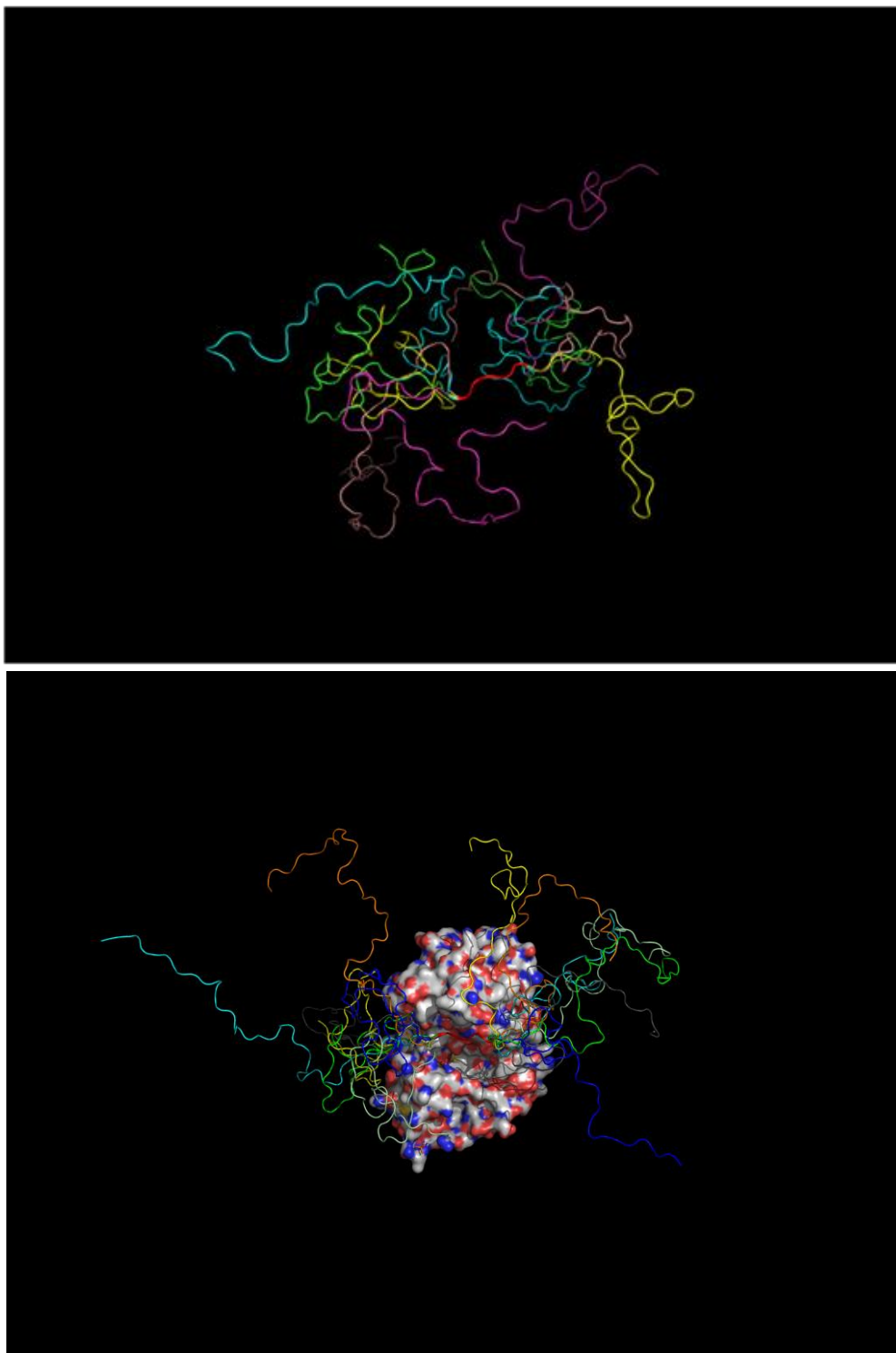


Figure 2.7: Simulation results from the study on structure [PDB:1CMK]. Examples of the generated conformers of the constructed 100mer peptide are shown. The conformers are aligned to the peptide in [PDB:1CMK] complex. (Upper): Conformers not available for docking by steric interference with the kinase. The regions near binding site appear to be more random coiled (Lower): Conformers available for docking that do not exhibit steric interference with the kinase.

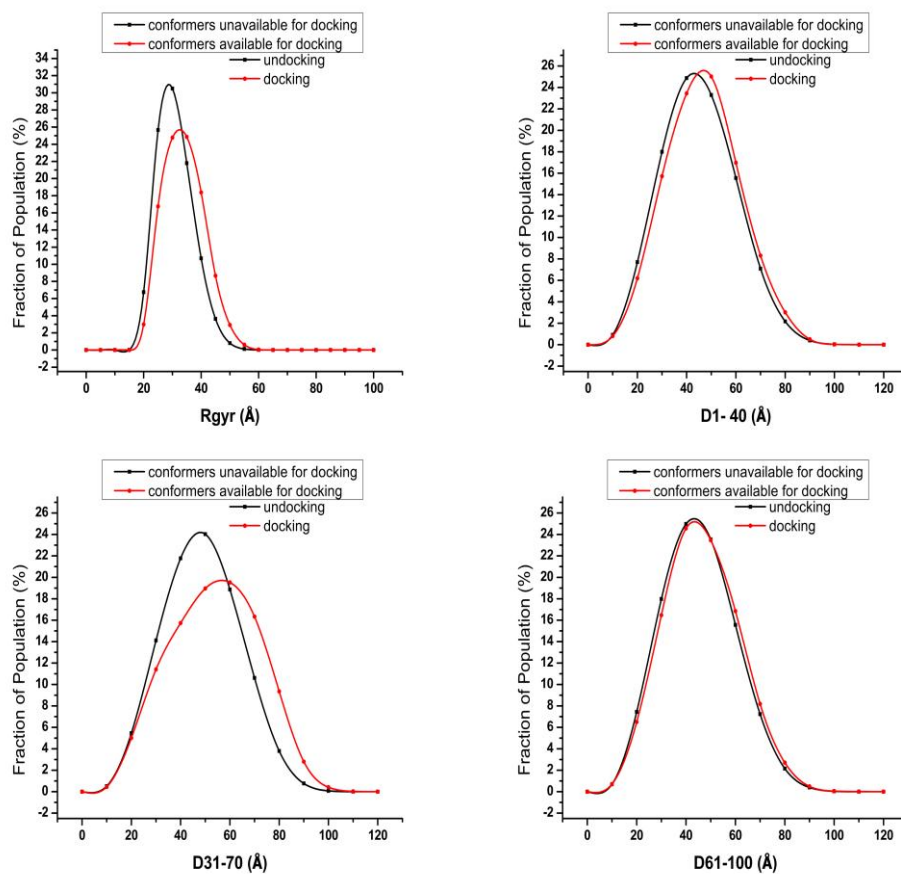


Figure 2.8: Rgyr and end-to-end distance distributions of D1-40, D31-70 and D61-100 for the constructed 100mer alternating Pro/Ser peptide with substrate phosphorylation motif in the centre.

t test			
Alternative hypothesis: True difference in means is less than 0, i.e. the mean of undocking conformers' Rgyr or end-to-end distance is less than that of docking conformers			
	Mean of undocking conformers (Å)	Mean of docking conformers (Å)	p-value
Rgyr	28.4383	31.4704	<2.2e-16
D1-40	40.0255	41.8216	<2.2e-16
D31-70	43.6516	48.8396	<2.2e-16
D61-100	40.2149	41.3728	1.032e-14

Table 2.4: T-test results on the constructed 100mer peptide. The table shows t-test results on the Rgyr distributions and end-to-end distance distributions of docking and undocking structural ensembles of the constructed 100mer peptide. The difference in distance is greatest over the span encompassing the kinase binding phosphorylation motif.

2.4 Conclusions

The Rgyr distributions of structure ensembles of the LRP6 intracellular domain were compared before and after applying spatial constraints. The whole structure was observed to require a more open or extended conformation to pass the spatial constraints and it was found that the near-membrane region is in fact elongated with the applied horizontal and vertical spatial constraints. During the initiation, the spatial constraints caused by the plasma membrane and nearby assemblies or molecules force an elongation form to dominate the conformational space of the LRP6 intracellular domain. We demonstrated that such an elongation process is required for unfolded protein structures because it can relieve the auto-inhibitory effect and grant kinases easy access. The near-membrane LRP6 intracellular domain extension could expose the S/T cluster phosphorylation site for CK1, which subsequently promotes downward PPP(S/T)PX(S/T) phosphorylation events.

This study elaborates details on the activation of LRP6 through its conformational change in the current Wnt/ β -catenin pathway initiation models.

TraDES provides a new way to investigate signal transduction mechanisms through computational structure sampling. More importantly, it demonstrates a way to study the conformational behavior of other proline-rich disordered protein regions including those in signaling pathways and mechanobiological systems. The Wnt/ β -catenin signaling pathway plays important roles in cancer and diseases. The LRP5/6 mutation is responsible for bone density disorders, ocular disorders and disorders of cholesterol and glucose metabolism. The findings in this study could pave the way to the development of new therapeutics through structure based drug design with the consideration of spatial constraints imposed by cellular components.

Experiments proposed to validate the LRP6 elongation model are single-molecule fluorescence resonance energy transfer (SM-FRET) and time-resolved fluorescence resonance energy transfer (TR-FRET). TraDES was originally validated with successful comparison to TR-FRET distribution [177]. SM-FRET and TR-FRET have been applied to study the conformations of full-length p53, which has both folded and intrinsically disordered domains [85]. SM-FRET can measure the radius of gyration of the LRP6 intracellular domain. TR-FRET can measure the end-to-end distance distribution within the LRP6 intracellular domain. Collectively these experiments may provide significant validation of the findings presented in this study.

2.5 Methods

2.5.1 Generation of conformers of LRP6 intracellular domain

The conformers of the LRP6 intracellular domain were generated using programs VISTRAJ and FOLDTRAJ from TraDES package [177], by providing the corresponding segment sequence. The sequence used was the 1613-residue LRP6 precursor retrieved from Uniprot database with definition line:

```
>sp|O75581|LRP6_HUMAN Low-density lipoprotein receptor-related protein  
6 OS=Homo sapiens GN=LRP6 PE=1 SV=1
```

The 19-residue signal peptide region was deleted from the N-terminal. As the extracellular domain from residue 20 to residue 1370 has its structure derived from X-ray diffraction in PDB database with accession ID “1N7D” [PDB:1N7D], this region was also deleted. The regions containing transmembrane and intracellular domain were taken as the segment sequence (residue 1371 to residue 1613) to generate conformers. The segment sequence was used in VISTRAJ to generate a trajectory distribution file for LRP6 intracellular domain, which contains the probabilistic distribution of ϕ / ψ angles in Ramachandran space for each residue in the segment sequence. This segment is predicted to be unfolded and has no apparent secondary structure. The “standard” method was used with no secondary structure predictions added. In this way the trajectory generated for each amino acid residue was based on its observed distribution of ϕ / ψ angles in a non-redundant subset of the PDB database. The trajectory distribution file was then manually edited to constrain the 23 residues from N-terminal side (T1371-I1393) to take an α -

helix conformation by replacing their random trajectory distribution with a fixed helical backbone conformation ($\Phi = -57^\circ$; $\Psi = -47^\circ$; Peak Magnitude =100). This was used for the first plane constraint anchor point for each member of the structural ensemble. The modified trajectory distribution file was next used by FOLDTRAJ to generate all-atom structure models of the LRP6 segment sequence. FOLDTRAJ generates off-lattice unbounded all-atom protein structures by amino acid residue random walks. The ϕ / ψ angles of the residues are obtained by sampling the Ramachandran space based on the trajectory distribution, and side chain rotamers are sampled from backbone dependent rotamer frequencies. Illustration 2.1 displays examples of generated conformers of LRP6 intracellular domain.

2.5.2 Filtration of structural ensemble of LRP6 intracellular domain

The conformers generated have no geometrical boundaries other than the steric hindrance of the sequence itself. However, *in vivo*, LRP6 is a transmembrane receptor and its intracellular domain would have its available conformational space limited by the cell membrane. If a structure generated has part of the peptide penetrating the plasma membrane, the conformation is not feasible *in vivo*, so we remove it from the structure ensemble through a defined horizontal plane mimicking the plasma membrane plane. During Wnt signaling pathway initiation, the assembly Wnt-FZD-DVL-AXIN-GSK3 or neighboring LRP6 molecule gets to the proximity of a LRP6 molecule. This will cause a second steric constraint to LRP6 intracellular domain. We defined a vertical plane perpendicular to the plasma membrane plane in order to

further filter the structure ensemble, and varied its distance to the membrane anchor point to represent possible close approach conditions. Illustration 2.1 shows some conformers of LRP6 intracellular domain that pass only Constraint 1 (Illustration 2.1B) or both Constraint 1 and Constraint 2 (Illustration 2.1C).

2.5.2.1 Constraint 1: Horizontal plane

In order to filter out the conformations that having part outside cell membrane, a program was developed to test whether a generated structure is in a conformation that can be bounded by a membrane. This check is done by constructing a virtual plane at the transmembrane site (Residue I1613), which is perpendicular to the inner membrane helix region. The rest of residues that should be inside the membrane are checked for whether they lie on the opposite site of the plane of the inner membrane helix. A Unix shell script was written to filter ensembles of structures that pass this constraint test in batches.

2.5.2.2 Constraint 2: Vertical plane

To further filter the structural ensemble and simulate the situation when an assembly or a neighboring molecule gets close to LRP6, a program was developed to test if a conformation is bounded by a plane that is between the intracellular domain of LRP6 and a neighboring object. We built another virtual plane at distances δ of 5.0Å, 10.0Å and 20.0Å to the transmembrane helix. All the residues should reside on one side of this plane. Another Unix shell script was written to complete this constraint based filtering.

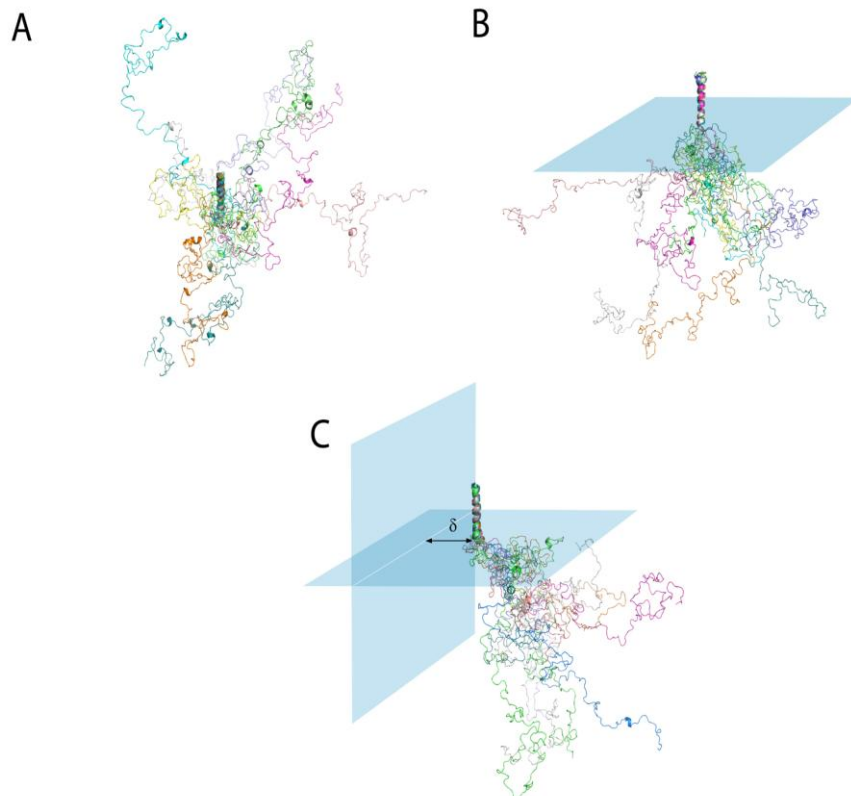


Illustration 2.1: Illustration of the spatial constraints. A. Ten aligned conformers of LRP6 intracellular domain which are not filtered by any spatial constraints. They are free to explore a spherical shaped conformational space. B. Constraint 1: Horizontal Plane Ten aligned conformers of LRP6 intracellular domain which pass the horizontal spatial constraint, i.e. membrane plane. They can explore a hemispherical conformational space. Signal peptide and extracellular domain are deleted. The transmembrane region (23 residues) is constrained as an α -helix. C. Constraint 2: Vertical Plane Ten aligned conformers of LRP6 intracellular domain which pass both horizontal and vertical spatial constraints. They are free to explore a smaller transected hemispherical shaped conformational space. The vertical constraint is imposed by the Wnt-FZD-DVL-AXIN-GSK3 assembly. δ is the distance between the vertical plane and transmembrane helix.

2.5.3 Measurement

Under each constraint, we measured the Radius of Gyration (R_{gyr}) to see the openness of the whole structure. In the meantime, the end-to-end distances with equal residues were measured to see the openness of LRP6 intracellular subsequences containing conserved PPP(S/T)PX(S/T) motifs. We can determine whether there is a conformational change by comparing the distributions of R_{gyr} and end-to-end distances of the structural ensembles under each constraint.

2.5.3.1 Measurement of radius of gyration

The generated structures by TraDES package are stored in NCBI ASN.1 format. It contains the locations of all the atoms inside the structure including hydrogen. Thus the distances between atoms and the radius of gyration could be calculated directly from the locations of the atoms. Radius of gyration is a measure of the dimensions of the peptide chains in polymer physics. It is defined as follows:

$$R_g^2 = \frac{1}{N} \sum_{k=1}^N (r_k - r_{mean})^2 \quad (2.1)$$

Where r_k is the position of individual atoms of the structure and r_{mean} is the mean position of the atoms (defined as the center of gravity of the structure). Radius of gyration is a root mean square distance of individual atom to the center of the structure. The higher the radius of gyration is, the sparser the atoms are in the structure. Therefore, this term can measure the openness of the whole structure.

2.5.3.2 Measurement of end-to-end distance

The end-to-end distance is defined by the distance between the α -carbon of the residues with a number of residues in between. The exact distances are those between C24 and S126, G64 and V166, T106 and L208, E124 and S226, T141 and S243. The averages of the end-to-end distances over large population indicate the openness of the subsequence within LRP6 intracellular domain.

2.5.4 The Rgyr distribution and end-to-end distance distribution

The Rgyr distributions (fraction of population in structural ensemble vs Rgyr) after Constraint 1 and Constraint 2 are plotted to compare the openness of structure ensemble. If mean of Rgyr has a shift to a higher value, it would indicate that the structure prefers an open and extended conformation based on the application of physical constraint.

The end-to-end distance distributions after Constraint 1 and Constraint 2 are also plotted to compare the openness of the LRP6 intracellular domain subsequence. If average end-to-end distance turns larger, it indicates the subsequence within LRP6 intracellular domain favors an open and elongated form.

2.5.5 Control experiment

A control sequence with the same length of LRP6 segment containing a transmembrane and intracellular domain was constructed. The control sequence has LRP6 transmembrane region and repeated proline-serine peptides substituting LRP6 intracellular domain. Conformers were generated using TraDES constraining transmembrane portion to adopt an alpha helix.

This initial structural ensemble was then filtered by Constraint 1 and Constraint 2. Rgyr and end-to-end distances were calculated and plotted into distribution curves. The control sequence is the following.

```
TNTVGSVIGVIVTIFVSGTVYFIPSPSPSPSPSPSPSPSPSPSPSPSPSSPSPSPS  
PSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPS  
PSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPS  
PSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPS  
PSPSPSPSPSPSPSPSPSPS
```

2.5.6 Program development

All the programs developed for Constraint 1 and Constraint 2 are based on NCBI C toolkit, MMDBAPI and libraries in TraDES package (<http://trades.blueprint.org/>). MMDBAPI implements data structures for describing biological sequence and 3D structure data and tools for easy access and manipulation of data, either in file system or in memory, based on ASN.1 standard. The TraDES package contains powerful function to manipulate and analyze the 3D structure of proteins such as aligning proteins by SVD (Singular Value Decomposition) method and to calculate Rgyr of structures. Shell scripts are created for processing and analyzing the structure ensemble in batch. The file format for storing 3D structures is .val file that can be visualized by Cn3D program from NCBI, or converted into PDB format. The whole simulation process can be viewed in the flow chart (Figure 2.9).

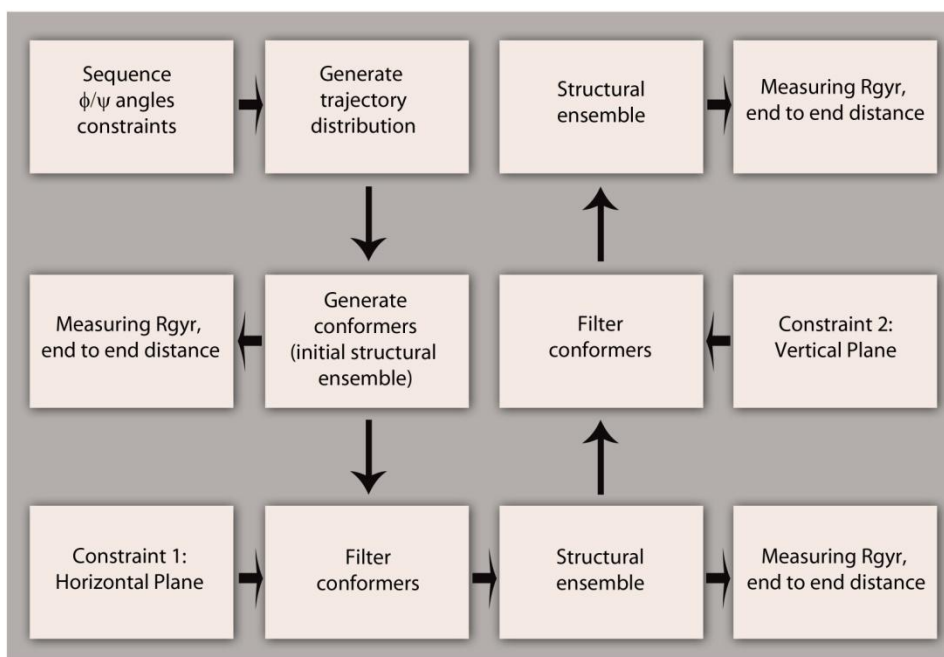


Figure 2.9: Flow chart of the simulation process on LRP6 intracellular domain.

2.5.7 Simulation procedure using structure [PDB:1CMK]

Alignment (SALIGN)

The structure [PDB:1CMK] chain I was aligned to the constructed 100mer peptide binding site by program SALIGN. SALIGN aligned two peptides (or part of the peptides) of the same length by superimposing them using a singular value decomposition (SVD) method. It takes in two structures and rotates and translates the second structure to align with the first one, then outputs the transferred structure to a new structure file. The exact parts of the sequence of the two sequences to be aligned can be specified respectively. There are three alignment methods in SALIGN-all atom, backbone and α -carbon. Each of the methods only takes their respective group of atoms into consideration. The backbone atoms were used for the 5 residues of the binding

site in the constructed 100mer peptide, which was superimposed by SVD onto the backbone atoms of chain E of 1CMK. As a result of each alignment, the whole structure record [PDB:1CMK] is transformed to the position and orientation where its chain E is superimposed with the binding site of the constructed 100mer peptide. The RMSD (Root Mean Square Deviation) of the alignment is calculated as a measure of quality of the alignment. A shell script was created to run this process in batch. The mean RMSD for conformers available for docking is 0.3463 and the best RMSD is 0.1660.

Merging (VALMERGE) and Crashes-checking (CRASHCHK)

After [PDB:1CMK] was aligned to the binding site of the constructed peptide, chain E of the [PDB:1CMK] structure was merged with the constructed 100mer peptide structure to form a single *.val file containing the bound protein complex. Next, a check was done to determine whether the kinase catalytic subunit had significant steric clashes with the atoms of the sampled 100mer peptide. If no crashes were detected, the binding site would be determined to be available for binding. Two programs VALMERGE and CRASHCHK were developed previously in our lab to do these tasks and modified slightly for creating protein complex files as described. VALMERGE merges the structure in multiple *.val files to form a complex after they are properly aligned by SALIGN. Individual chains could be specified so that only these are copied and merged into the output structure. CRASHCHK is a program that iteratively scans through each chain of the structure and output the number and identity of all the pairs of atoms that are too close or superimposed on one another. The atom-atom distance is

compared to the Van der Waal distances of the two atoms as thresholds. The developed programs in this study are included in TraDES package, which are available in the TraDES 2.0 package at <ftp://ftp.blueprint.org/pub/TraDES>. TraDES 2.0 was released on 6 June 2012 in open-source under a BSD license and renames FOLDTRAJ to TRADES, and VALMERGE to STRMERGE. Refer to <ftp://ftp.blueprint.org/pub/spatialConstraints/> for detailed account of the software, scripts and parameters used in this study.

2.6 Acknowledgements

This work was supported by the Singapore-MIT Alliance Fellowship and Mechanobiology Institute of Singapore.

2.7 Author's Contributions

Chengcheng Liu carried out the simulation studies, developed the software, performed the statistical analysis and wrote the manuscript. Mingxi Yao participated in the program development. Christopher W.V. Hogue participated in the design of the study, the program development, and coordination.

Chapter 3

Sequence Detection of Proline/Serine-Rich Disordered Regions²

Abstract

Most compositionally biased proline-rich regions are usually disordered and contain multiple serines or threonines as phosphoacceptors. We refer to these regions as Pro/Ser-rich disordered regions. They play important roles in signaling pathways, but sequence annotation of these regions is not complete in most protein databases. In addition, there is a lack of consensus on how to define these regions. In this study, we define Pro/Ser-rich disordered regions as those long disordered regions (>40aa) and enriched in prolines and serines. Many machine learning algorithms have been proposed to predict disordered regions. However, their predictions do not give a clue about the compositional bias within the sequence. They are fundamentally based on amino acid properties of protein sequence, which determines the structure and in turn the function. The amino acid composition of the protein should identify many protein structural properties. Previous study on predicting the linkers between domains used a simple amino acid composition propensity as a discriminating linker prediction index. Referring to this method, we generated an amino acid index called Pro/Ser-Rich (PSR) based on the compositional bias between a set of curated Pro/Ser-rich disordered regions and a set of folded domains. PSR index is a simple and effective approach to predict Pro/Ser-rich disordered regions. A web server called Armadillo (2.0) has been updated to

² Portions of the work written in this chapter are being prepared in a manuscript.

perform queries and prediction tasks. Armadillo (2.0) incorporates PSR index into its previous domain-linker propensity indices. It can predict linker regions as well as Pro/Ser-rich disordered regions within a protein sequence.

3.1 Background

Disordered protein regions exist as structural ensembles where the equilibrium position of the backbone atoms and their dihedral angles have no set positions or values and vary significantly over time [33, 34]. These regions are discovered and studied by different techniques. For example, X-ray crystallography defines missing electron density in many protein structures as disordered protein [43, 44, 354]; nuclear magnetic resonance (NMR) is able to assign resonances for the disordered protein fragments [52, 61, 193, 355-357]; circular dichroism (CD) can detect disordered regions by the near-UV CD spectrum with low intensity. Other spectroscopic techniques including far-UV CD [70, 358-363], Fourier transform infrared (FTIR) [30], electron paramagnetic resonance (EPR), and optical rotary dispersion (ORD) are also employed to identify disordered regions [364]. Additionally, fluorescence resonance energy transfer (FRET) can provide more information of the conformations of disordered protein regions [115, 365].

The fraction of protein disorder in several genomes was predicted in high content in eukarya (33%) compared to 4.2% in bacteria and 2% in archaea [115]. This may result from the more complex regulation and signaling systems in higher organisms. The number of disorder predictors has increased rapidly. The algorithms that these predictors based upon include

simple sequence complexity, complicated machine learning methods such as neural network and support vector machine, and other novel proposed biophysical models. Many reviews have been published to discuss the details about these disorder predictors [104, 105, 366, 367]. Many research groups have participated in the biennial Critical Assessment of Techniques for Protein Structure Prediction (CASP) since 2004 to demonstrate the performance of their disorder predictors [368].

Disordered and ordered regions have different amino acid compositions. Higher frequencies of R, K, E, P and S amino acids and lower frequencies of C, W, Y, I and V amino acids have been found in disordered regions compared to ordered regions [96]. By comparing amino acid composition and various biophysical attributes between ordered and disordered fragments, a novel amino acid scale was constructed to discriminate order and disorder. This provides a new ranking for the residues for their tendencies to promote disorder (from order promoting to disorder promoting): W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P [160].

Disorder proteins have four main functions: (1) molecular recognition: effectors, scavengers and displaying sites for post-translational modifications, (2) molecular assembly, (3) protein modification, and (4) entropic chain activities [158, 200]. Many disordered proteins, for example, α -synuclein, p53, 14-3-3, AXIN, breast cancer type 1 susceptibility protein (BRCA1), microtubule-associated protein 2 (MAP2), titin etc, serve as hubs in signaling pathways to interact with multiple partners [204-207, 369]. The D2 concept or “disorder in disorders” has been proposed to illustrate that disordered regions

are highly involved in neurodegenerative diseases/conformational diseases (for example, Alzheimer's disease, Parkinson's disease, Down's syndrome, etc), cardiovascular diseases, and cancer [370].

Romero *et al.* demonstrated that low sequence complexity cannot determine the protein disorder since the prediction only involves direct sequence analysis without using any structural information [96]. However, a large number of disordered proteins have been identified by low sequence complexity, and many of these proteins contain compositional sequence biases. For example, glycine-rich sequences are disordered regions; regions enriched in proline, glutamic acid, serine, and threonine (PEST) are mostly disordered. Our study focuses on compositionally biased proline-rich regions, which have been associated with protein disorder and contain multiple serines as phosphorylation sites. Proline and serine are two amino acids that are highly ranked to promote disorder [160]. Proline has very limited conformational space as its unique side-chain is cyclised onto the backbone nitrogen atom [224]. Polyprolines are likely to adopt the PPII (Polyproline II) helix, which is an extended structure with three residues per turn. Such conformations tend to exist in extended disordered regions that are hard to characterize using X-ray cryptography or NMR [371]. Phosphoacceptor serines or threonines are often located immediately preceding proline in proline-rich regions. Like most disordered proteins, Pro/Ser-rich disordered proteins are biologically important. Many identified disordered regions are enriched in proline and serine. For example, casein, tau protein, etc [232]. There is no such sequence annotation as "Pro/Ser-rich Disordered" in protein databases. In UniProt database, a type of compositional bias feature is indicated as "Pro/Ser-rich".

However, it does not give its structural information as order or disorder. In addition, there is no consensus on this term during curation. To determine whether a region is disordered and highly enriched in proline and serine requires a threshold frequency and other attribute, for example, protein region length.

Current disorder predictors can predict disordered and ordered regions; however, their predictions do not indicate the compositional bias within the sequence. Earlier study developed a discriminating linker prediction index (DLI) to predict the linkers between domains by using a simple amino acid composition propensity [323]. Following this method, an amino acid propensity index called Pro/Ser-Rich (PSR) was computed based on the compositional bias between a set of curated Pro/Ser-rich disordered regions collected by the author, and a set of protein folded domains. The PSR index is a simple and effective approach to predict Pro/Ser-rich disordered regions. A web server called Armadillo (2.0) (<http://web2.mbi.nus.edu.sg>) has been updated to perform queries and prediction tasks. Armadillo (2.0) incorporates the PSR index into its previous domain-linker propensity indices. This tool gives more information about protein disorder by predicting linker regions as well as Pro/Ser-rich disordered regions within a protein sequence.

3.2 Implementation

3.2.1 Pro/Ser-rich disorder dataset

For this study, 357 protein regions were collected from 388 published papers on PubMed identified with the query “proline serine rich region”. These protein regions were classified into four main categories: PRR (Pro-rich

region), PSR (Pro/Ser-rich region), PST (Pro/Ser/Thr-rich region) and PEST (Pro/Glu/Ser/Thr-rich regions). These terms are used by the authors in the found publications. The same protein region may be mentioned differently using above terms according to the authors' opinions in different papers. A manual filtering ensured that candidate proteins had the reported Pro/Ser-rich region that was greater than 40 residues in length, and had random coiled / disordered / unfolded structure or had no reported structure in the PDB database. This was done because amino acid composition has been reported to be sufficient for predicting long disordered regions [163, 164]. The filtering leaves a subset containing 125 Pro/Ser-rich disordered regions with start residue and end residue annotated as in the publication. The proteins in this dataset had a mean length of 148.7 (± 149.8) residues and a median length of 102 residues. The dataset can be downloaded from <http://web2.mbi.nus.edu.sg/>.

3.2.2 Third party datasets

Two third party datasets were also used in this study. A number of 1069 disordered protein segments were extracted from the dataset DisProt Release 5.8 (http://www.disprot.org/data/version_5.8/disprot_fasta_v5.8.txt). The MMDB-I dataset from Dumontier *et al*, 2005 [323] consisted of 585 proteins with two to ten non-redundant VAST domains and at least one linker. We used the pre-computed amino acid composition of domains in MMDB-I to develop our PSR index as a reference state for the log-odds scoring function. Thus the scoring function will remain near zero for the majority of folded domains in protein sequences, which in the MMDB-I set are distinctly separated from any

potentially linker regions. Thus the combination of PSR and DLI may offer discrimination between short linkers and Pro/Ser-rich disordered regions.

3.2.3 The PSR index

The frequencies of amino acid residues in the domains from MMDB-I dataset (d) and the Pro/Ser-rich disorder dataset (t) are shown in Table 3.1. Amino acid frequency as described in [323] is the ratio of number of an amino acid to that of all amino acids. The amino acid frequency $C_{aa,s}$ was determined for each amino acid aa from the ratio of the occurrence of each amino acid $n_{aa,s}$ in a set s compared to its total occurrence $\sum n_{aa,s}$ as shown in equation (3.1):

$$C_{aa,s} = \frac{n_{aa,s}}{\sum_{aa} n_{aa,s}} \quad (3.1)$$

The PSR index describes the Pro/Ser-rich disorder propensity in the form of a log-likelihood of any amino acid residue to be found in a Pro/Ser-rich disordered region. Amino acid propensity refers to the natural inclination or tendency for an amino acid to behave in a particular manner where all factors are not known [323]. The PSR index is calculated from the amino acid frequency $C_{aa,s}$ as the negative log ratio between the amino acid frequencies in Pro/Ser-rich disorder dataset t and in folded domains d , as shown in equation (3.2). The PSR is then normalized to a zero mean and unit standard deviation.

$$PSR_{aa} = -\log \frac{C_{aa,t}}{C_{aa,d}} \quad (3.2)$$

Table 3.1: Calculated frequencies of amino acid residues in Pro/Ser-rich disorder dataset and MMDB-I domain dataset as well as the negative and normalized log ratios for PSR index.

Amino Acid	<i>d</i>		<i>t</i>		Negative Log Ratio	Normalized PSR index
	$n_{aa,d}$	$C_{aa,d}$	$n_{aa,t}$	$C_{aa,t}$	PSR_{aa}	
A	19504	8.5	1547	8.3	0.010	-0.333
C	2869	1.3	140	0.8	0.221	0.515
D	13370	5.9	700	3.8	0.191	0.392
E	15177	6.6	920	4.9	0.127	0.136
F	9108	4.0	350	1.9	0.325	0.931
G	17201	7.5	1244	6.7	0.051	-0.172
H	5246	2.3	418	2.2	0.009	-0.341
I	13089	5.7	427	2.3	0.396	1.217
K	13290	5.8	781	4.2	0.141	0.190
L	20855	9.1	1062	5.7	0.203	0.440
M	4966	2.2	284	1.5	0.153	0.238
N	9644	4.2	471	2.5	0.221	0.513
P	10311	4.5	3658	19.7	-0.640	-2.945
Q	8483	3.7	776	4.2	-0.051	-0.581
R	11480	5.0	821	4.4	0.055	-0.152
S	13021	5.7	2397	12.9	-0.355	-1.801
T	12765	5.6	1318	7.1	-0.104	-0.793
V	16454	7.2	836	4.5	0.204	0.444
W	3256	1.4	103	0.6	0.410	1.271
Y	8172	3.6	333	1.8	0.300	0.829
Total	228261	100.0	18586	100.0		
Mean					0.093	0.000
STD					0.249	1.000

3.2.4 Pro/Ser-rich disorder prediction

After each residue of a protein sequence is assigned a value according to its normalized PSR index, a numeric profile can be generated. Next, a 15 residue moving window is used to smooth the numeric profile by giving an average value to the central residue. Then, a second smoothing is completed by an inverse discrete Fourier transform followed by a low-pass filter with a cutoff frequency of 1/25. The residue profile distribution is then transformed to a standard normalized distribution (zero mean, unit standard deviation).

$$Z = \frac{x - \mu}{\sigma}$$

Z-scores less than or equal to -3.43 were predicted as Pro/Ser-rich disordered regions. This threshold was trained on the training dataset and empirically tested giving the best prediction sensitivity. It was determined from the density distributions of those belong to Pro/Ser-rich disordered regions and those do not in the training dataset.

3.2.5 Prediction performance measures

For the determination of the performance of PSR, the first quantity measured is prediction sensitivity, which is $TP/(TP+FN)$, where TP is the number of residues that are in true positive Pro/Ser-rich disordered regions and FN is the number of false negative, i.e. residues that are in Pro/Ser-rich disordered regions and have not been predicted. The second quantity measured is specificity, which is $TN/(TN+FP)$ where FP is the number of false positives, i.e. predicted residues that do not fall on a real Pro/Ser-rich disordered region.

3.2.6 Armadillo (2.0)

The first Armadillo program has been updated to a new version which incorporates PSR index. Armadillo (2.0) can take inputs of FASTA sequences or NCBI GI numbers. The combined program generates predictions about linkers and Pro/Ser-rich disordered regions. A tab-delimited file including the normalized scores from each of the amino acid indices can be generated. A new web interface to Armadillo (2.0) (<http://web2.mbi.nus.edu.sg>) is available. It generates graphical output using Dojox Charting (<http://dojotoolkit.org>). The prediction program will be made available under a BSD style open source license.

3.3 Results and Discussion

3.3.1 Amino acid composition in the datasets

To develop a propensity index for searching Pro/Ser-rich disordered regions, we compared the differences in the amino acid composition profiles of the four datasets: domains and linkers in MMDB-I, disordered protein segments in DisProt Release 5.8, and Pro/Ser-rich disorder dataset from PubMed. The comparison is shown in Table 3.2, Table 3.3, Table 3.4 and Figure 3.1. The discriminating difference existing within the domain and linker compositions was applied to generate the linker propensity index as described by Dumontier *et.al*. The domains appear to have more hydrophobic residues Leu, Val, Ile, Ala and to a lesser extent Arg, Met, and Tyr. Linker regions have higher percentages of Pro, Gly, and less content of Asn, Asp, Ser and Thr. In contrast, disordered regions are enriched in Ser, Glu, Pro, Lys and to a lesser extent Gln,

Asp and Thr; while, the Pro/Ser-rich disordered regions are highly enriched in Pro, Ser and to a lesser extent Thr and Gln. These regions have a higher content of Thr comparing to disordered regions. This may be due to the fact that Pro-rich regions always contain multiple Ser and Thr, which fit exactly into our interest. These observations confirm that Pro/Ser-rich disordered regions are a subset of disordered regions with compositional bias.

Table 3.2: Amino acid composition difference in percentage between MMDB-I domain dataset and disordered protein segments in DisProt (v5.8).

Amino Acid	% Domain	% DisProt (v5.8)	% Difference
L	9.14	6.37	2.77
I	5.73	3.11	2.62
V	7.21	5.39	1.81
F	3.99	2.34	1.65
Y	3.58	2.02	1.56
W	1.43	0.64	0.79
N	4.22	3.61	0.61
C	1.26	0.79	0.47
A	8.54	8.09	0.46
R	5.03	4.61	0.42
M	2.18	1.80	0.38
H	2.30	1.95	0.35
G	7.54	7.42	0.11
T	5.59	5.77	-0.18
D	5.86	6.40	-0.55
Q	3.72	5.32	-1.60
K	5.82	8.13	-2.31
P	4.52	7.44	-2.93
E	6.65	9.81	-3.16
S	5.70	8.98	-3.27

Table 3.3: Amino acid composition difference in percentage between MMDB-I domain dataset and the curated Pro/Ser-rich disorder dataset from literature.

Amino Acid	% Domain	%PSR in PubMed	% Difference
I	5.73	2.3	3.43
L	9.14	5.71	3.43
V	7.21	4.5	2.71
F	3.99	1.88	2.11
D	5.86	3.77	2.09
Y	3.58	1.79	1.79
E	6.65	4.95	1.7
N	4.22	2.53	1.69
K	5.82	4.2	1.62
W	1.43	0.55	0.88
G	7.54	6.69	0.85
M	2.18	1.53	0.65
R	5.03	4.42	0.61
C	1.26	0.75	0.51
A	8.54	8.32	0.22
H	2.3	2.25	0.05
Q	3.72	4.16	-0.44
T	5.59	7.09	-1.5
S	5.7	12.9	-7.2
P	4.52	19.68	-15.16

Table 3.4: Amino acid composition difference in percentage between MMDB-I linker dataset and disordered protein segments in DisProt (v5.8).

Amino Acid	% Linker	%DisProt (v5.8)	% Difference
G	10.59	7.42	3.17
P	10.07	7.44	2.62
N	5.20	3.61	1.59
F	3.89	2.34	1.55
Y	2.88	2.02	0.86
D	7.10	6.40	0.70
H	2.48	1.95	0.52
W	1.16	0.64	0.52
C	1.27	0.79	0.48
I	3.58	3.11	0.46
T	6.18	5.77	0.41
L	6.72	6.37	0.35
M	1.46	1.80	-0.34
R	4.15	4.61	-0.46
V	4.83	5.39	-0.56
A	6.45	8.09	-1.64
Q	3.52	5.32	-1.80
S	6.85	8.98	-2.13
K	5.48	8.13	-2.65
E	6.15	9.81	-3.66

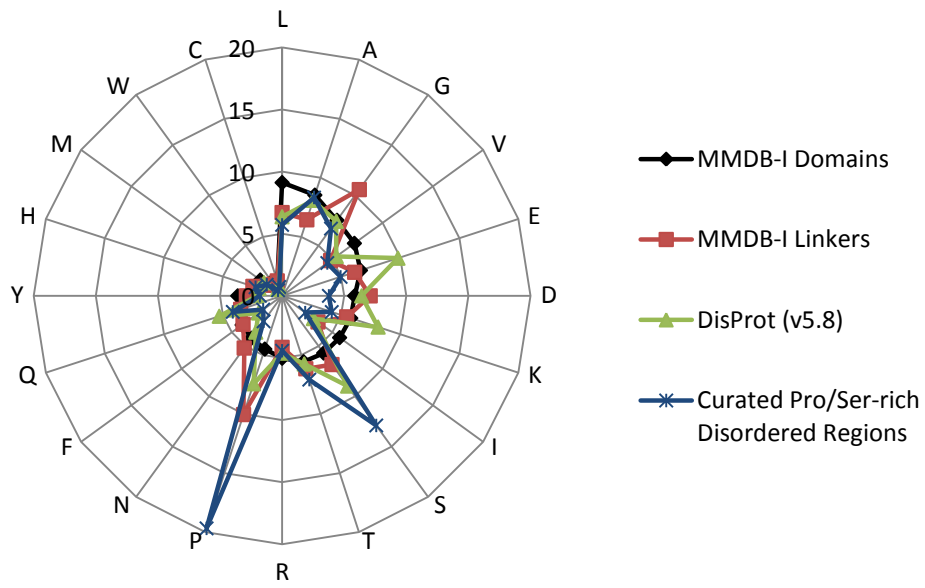


Figure 3.1: Amino acid compositions of the datasets. The amino acids were put around the radar clockwise. The domain curve spirals in and goes around the clock in decreasing amino acid composition frequencies.

3.3.2 Evaluation of Pro/Ser-rich disorder predictions

The smoothed numeric profile of the sequence generated from PSR index was used to make predictions (See Implementation). Cross-validation and dataset scoring were used to determine the prediction sensitivity and specificity.

3.3.2.1 Cross-validation

The first approach was a five-fold cross-validation procedure in which the training dataset was randomly divided into five groups. The members from four groups were used to train the PSR index and then used to predict the excluded fifth group. This process was repeated over 100 times to calculate the sensitivity and specificity of prediction. PSR index achieved 64 (± 6)% in sensitivity and 66 (± 5)% in specificity for predictions from the curated Pro/Ser-rich disordered regions.

3.3.2.2 Dataset scoring

The second approach was to use PSR index to predict the training dataset and get the training error. Prediction on the training dataset averaged 64% in sensitivity and 73% in specificity for PSR index. For comparison, we built three tandem pipeline approaches each of which adopted one of the three disorder predictors: DisEMBL [131], GlobPlot [121] and PrDOS [125]. These disorder predictors predict disordered regions without information about the compositional bias. Thus, we applied another software package “ps_scan” [372] as a secondary step to further predict proline-rich regions. The overlapping prediction between each disorder predictor and ps_scan will reflect the Pro/Ser-rich disordered regions that we are interested in (Table 3.5).

PSR predictions appear to be much more sensitive (64%) but less specific (73%) than the tandem pipeline approaches which contained a disorder predictor and the composition scanner ps_scan (see Table 3.5). The majority of residues in Pro/Ser-rich disordered regions are predicted correctly, and the greatest error appears to be the prediction of residues in domains and in disordered regions which are not Pro/Ser-rich. The sensitivity is also limited by the fact that the dataset has inherent multi-domain proteins with discontinuous segments. The high specificities of the three tandem approaches may be attributed to the Pro-rich predictions of ps_scan (79%), which could predict 62 Pro-rich regions (49.6%) that are in the curated Pro/Ser-rich disorder dataset. The software ps_scan is a PROSITE scanning program which uses a “profile library” to search for specified pattern. In the study, a “Proline-rich region profile” was included in order to search for regions that are significantly enriched in proline. Moreover, PROSITE contains a profile library for sequence regions enriched in a particular amino acid; however, this only focuses on finding compositional biased regions with low-complexity that may not be disordered. Additionally, the software ps_scan prediction does not involve end-effect averaging [372]. PSR is a unique dual predictive program and there is no other existing software to directly compare with it. Here, we adopted a tandem pipeline approach using existing tools to do a “fabricated comparison”.

		Tandem pipeline (Disorder predictor combined with ps_scan)			
		PSR index	DisEMBL & ps_scan	GlobPlot & ps_scan	PrDOS & ps_scan
training dataset	%s	64	21	37	33
	%p	73	99	98	98

Table 3.5: Pro/Ser-rich disorder predictions. A summary of the results obtained for Pro/Ser-rich disorder prediction in the dataset of Pro/Ser-rich disordered regions from literature. Percentage sensitivity (%s) and specificity (%p) for PSR index, and three disorder predictors (DisEMBL, GlobPlot and PrDOS) combined with ps_scan.

3.3.3 Server prediction examples

Armadillo (2.0) web server predictions for the LRP6, WASP and MAP tau proteins are shown in Figure 3.2. LRP6 is a member in the subfamily of LDL receptors (LDLR) [319]. It is a type I single-pass transmembrane protein, which contains an extracellular domain, a helical transmembrane domain and a cytoplasmic domain. The extracellular domain consists of three basic domains: the YWTD (tyrosine, tryptophan, threonines and aspartic acid)-type β -propeller domain, the EGF (epidermal growth factor)-like domain, and the LDLR type A (LA) domain [338]. The cytoplasmic domain of LRP6 (residue 1317 to residue 1613) is enriched in prolines and serines including a S/T cluster and downstream five reiterated PPP(S/T)PX(S/T) motifs, whose phosphorylation is crucial in the activation of the Wnt/ β -catenin signaling pathway [311]. Until now, no stable structure has been reported for LRP6 cytoplasmic domain in current structure databases. This region is expected to be natively unfolded due to its Pro/Ser-rich compositional bias [311, 338]. Armadillo (2.0) predicts a Pro/Ser-rich disordered region from residue 1380 to residue 1613, which corresponds to the LRP6 cytoplasmic domain (Figure 3.2(a)). Figure 3.2(b) shows LRP6 conserved domains that are obtained by searching against the NCBI's conserved domain database (CDD). The second example is the WASP protein, which regulates actin filament reorganization and polymerization. Human WASP has a Pro-rich region from residue 160 to residue 404, which also contains several phosphorylation sites [286]. The PSR index predicts four Pro/Ser-rich disordered regions (residue 138-residue 242, residue 292-residue 342, residue 374-residue 424 and residue 431-residue 483),

three of which cover its Pro-rich region (Figure 3.2(c)). The last example is the microtubule-associated protein tau, which promotes microtubule assembly and stability. Abnormally hyperphosphorylated tau is found in Alzheimer's disease [373]. In the human MAP tau isoform 2 protein sequence, PSR predicts two Pro/Ser-rich disordered regions: residue 32-residue 82 and residue 158-residue 218 (Figure 3.2(e)), the latter of which covers the reported Pro-rich region (172-251). However, it also identifies the N terminus as Pro/Ser-rich region (Figure 3.2(f)).

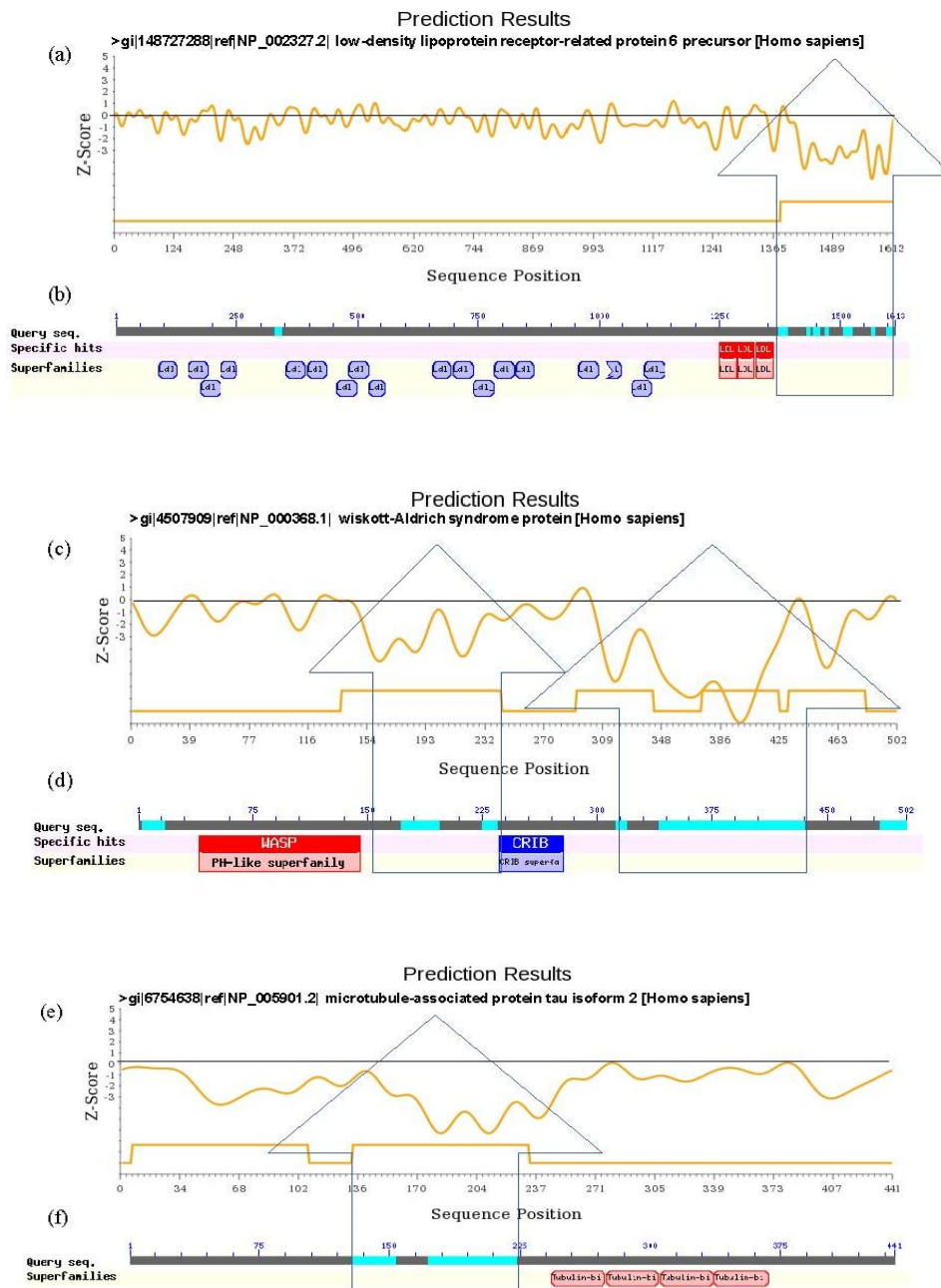


Figure 3.2: Armadillo (2.0) Pro/Ser-rich disorder predictions for human proteins LRP6, WASP and MAP tau isoform 2. By default, the Armadillo 2.0 web server shows all the linker predictions and Pro/Ser-rich disorder predictions. The graphs here show only Pro/Ser-rich disorder predictions for LRP6 (a), WASP (c) and MAP tau isoform 2 (e). Each graph displays the smoothed Z-score profile along the sequence (x-axis). The domain architectures are shown from querying the NCBI's conserved domain database.

3.4 Conclusions

In this study, we have developed a simple approach to predict disordered regions whose composition is biased with proline and serine. PSR index achieves 64% prediction sensitivity for its training dataset. This method may not give higher specificity than the tandem pipeline approach that was discussed; however it is a more sensitive tool for finding Pro/Ser-rich disordered proteins. The concept of this study is to explore a subset of disordered proteins in terms of their biased proline/serine-rich composition. As our comparison set was to a combined set of prediction methods including the ps_scan tool, the results show that there are distinct areas of possible improvement. For example, specificity may be increased by incorporating other amino acid attributes or scores, or by combining different prediction methods. In general, the log-odds score performed well as a simple tool to detect Pro/Ser-rich disordered regions that may be highly present in human genome. Meanwhile, the newly updated web server Armadillo (2.0) is available for usage.

3.5 Author's Contributions

Chengcheng Liu collected the data, developed the software, modified the interface, performed the statistical analysis and wrote the manuscript. Christopher W.V. Hogue participated in the design of the algorithm and the study.

Chapter 4

Sequence Analysis of Interpositional Dependence in Phosphorylation Motifs³

Abstract

A significant number of phosphorylation sites in substrate sequences are characterized in vitro from oriented peptide library screening. The specificity of kinase for the substrate sequence mostly comes from the sequence information at positions in the proximity to the site of phosphorylation, but this excludes the interpositional dependence among the positions. It is now plausible to obtain large quantities of in vivo substrates using high-throughput techniques like mass spectrometry. We used data from experiments on the kinases ATM/ATR and CDK1/Cyclin B as well as curated CK2 substrates, to examine the abundance of interpositional dependencies between positions within a substrate motif. Incorporating these interpositional dependencies, we used probabilistic models to predict kinase phosphorylation sites. In the results, a scarcity of interpositional sequence dependencies is observed, and these dependencies in fact do little help to improve the prediction accuracy of the probabilistic models. Our results may imply that other components of biological and cellular context should be included to improve the ability of the models, rather than only considering the sequence alone. The results also

³ Portions of the work written in this chapter have been previously published as: Brian A. Joughin, Chengcheng Liu, Douglas A. Lauffenburger, Christopher W.V. Hogue and Michael B. Yaffe. Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signaling networks. *Philos. Trans. R. Soc. B.* **367**: 2574-2583

suggest that the kinase substrate fitness exists in a smooth energetic landscape evolutionarily. Other research results suggest the interpositional dependences do lie in the substrate motifs of phosphopeptide-binding domains, like SH2 etc. Our data indicates that the evolution of fitness of phosphopeptide-binding domains may limit the new functional substrate molecules in the phospho-signaling pathways.

4.1 Background

Phosphorylation is a post-translational modification that is crucial in various biological processes, including cell cycle control, signal transduction, cell motion, enzyme regulation and others [306-308]. The number of kinases that are responsible for the phosphorylation of serine, threonine and tyrosine on about one third of eukaryotic proteins, has been identified to be more than 500 [374]. The traditional ways used to characterize phosphorylation sites are mutational analysis and Edman degradation chemistry, but more recently advanced mass spectrometry techniques has shown numerous phosphorylation sites on a large variety of proteins which have not yet been found to be involved in biological processes. Nevertheless, it is difficult to assign the discovered phosphorylation sites to the kinases that are responsible for the modification, which means the kinase-substrate specificity is mostly unknown. Bioinformatic tools have developed to solve the kinase-substrate specificity, commonly with describing a 3-10 residue long sequence motif with the phosphorylation site specifically for a kinase [375]. The sequence motifs only give a simple glimpse of the specificities of kinases, because they do not serve

as a sufficient nor a necessary condition when determine if a sequence is a substrate of a specific kinase.

A list of current useful phosphorylation predictors are provided in Table 4.1. Some of these phosphorylation predictors like NetPhos and GPS2.0 are based on PROSITE pattern searches [312-314], which in fact will give many false-positive examples. NetPhosK, NetPhorest, KinasePho, PredPhospho and PPSP etc. use machine-learning methods, including neural networks, hidden Markov models, support vector machines and Bayesian decision theory, which are not quite optimal prediction methods if the training dataset is small. Besides, these algorithms normally retrieve the learning set from databases Phospho.ELM [376, 377] and Swiss-Prot/TrEMBL [378], which collects phosphorylation sites from different experimental methods and hypothetical sites which are derived from sequence similarity to known substrates instead of experimental results. This could cause high redundancy and large noise in the dataset. Scansite [317] applies a position-specific scoring matrix (PSSM), which is constructed from oriented peptide library screening in vitro. It only covers a short central motif (3-10 residues) to do the prediction task, which may increase the false positive rate if the evolved specificity covers a larger region of sequence. The scoring matrix PSSM calculates the prediction score based on the amino acid frequency at each position within the central motif assuming these positions are independent, which describes a first-order model of sequence specificity. As such, there is a lack of consideration about the interpositional dependencies within the phosphorylation motif, which could possibly exist around the phosphorylation

sites *in vivo* substrates and determine the kinase specificity. This describes a higher-order model of sequence specificity, i.e. the amino acids binding evolved co-dependencies in the neighboring positions relative to the phosphorylation site. If such higher-order information missing in PSSM does indeed exist, would it be useful to better describe the specificity of a kinase *in vivo*?

To approach this problem of determining higher-order specificity, a large number of substrates for a kinase from a single experiment are required for the analysis of interpositional dependence. The current database Phospho.ELM only curates a lower percentage (12%) of phosphorylation sites specific to a particular kinase [376]. For most kinases, they do not have a known specificity sequence motif according to a computational study on linear motif atlas for phosphorylation-dependent signaling conducted by Huang *et al* [379]. Moreover, a large dataset of phosphorylation sites for most kinases does not come from a single experiment. Fortunately, a few single experiments generating a large dataset of phosphorylation sites for a specific kinase have been described in the literature. These give us resources to study the interpositional dependence surrounding the phosphorylation site and to test if higher-order model would give a better prediction result than a PSSM does.

Table 4.1: A list of current phosphorylation site predictors.

Name	URL	Year	Reference
NetPhos	http://www.cbs.dtu.dk/services/NetPhos	1999	[315]
Scansite2.0	http://scansite.mit.edu	2003	[317]
DIPHOS	http://www.ist.temple.edu/DIPHOS	2004	[310]
NetPhosK	http://www.cbs.dtu.dk/services/NetPhosK	2004	[380]
PredPhospho	http://pred.ngri.re.kr/PredPhospho.htm	2004	[381]
PPSP	http://www.webcitation.org/query.php?url=http://bioinformatics.lcd-ustc.org/PPSP&refdoi=10.1186/1471-2105-7-163	2006	[382]
pKaPS	http://mendel.imp.ac.at/sat/pkaPS	2007	[383]
KinasePhos2.0	http://KinasePhos2.mbc.nctu.edu.tw	2007	[384, 385]
NetPhosYeast	http://www.cbs.dtu.dk/services/NetPhosYeast	2007	[386]
GPS2.0	http://bioinformatics.lcd-ustc.org/gps2/down.php	2008	[387]
Predikin	http://predikin.biosci.uq.edu.au	2008	[388]
NetPhorest	http://netphorest.info	2008	[389]
Phospho.ELM	http://phospho.elm.eu.org/	2008	[376, 377]
NetworKin	http://networkin.info	2008	[390]
PhosphoSitePlus	http://www.phosphosite.org	2012	[391]

In this study, we take a look at the second-order (pairwise positions) interpositional dependence in three large phosphorylation datasets of kinases ATM and ATR, CDK1 and CK2. ATM/ATR are the central components of DNA Damage Response (DDR) [392]. They are activated when the double stranded breaks (DSB) initiate which create deadly DNA lesions. ATM and ATR phosphorylate one or more key proteins in the DDR signaling network e.g. the activation of cell-cycle checkpoints proteins and they share common substrates. CDK1 (cyclin-dependent kinase 1) is activated indirectly by ATM/ATR. ATM/ATR phosphorylates CHK1/CHK2 (serine/threonine kinases) that inactivates CDC25C/CDC25A through phosphorylation. CDC25C/CDC25A in turn can activate CDK1/CDK2 by phosphorylation [392]. CDK1 drives G2 to M phase in cell-cycle control [393]. These two types of protein kinases are critical in DNA damage response signaling

network. CK2 is a ubiquitous, active serine/threonine kinase whose substrates include numerous signaling proteins [394, 395].

4.2 Results

4.2.1 Statistical significance of interpositional dependencies among kinase phosphorylation motifs

First, the substrate dataset of ATM/ATR [392] was analyzed to see if there are any second-order interpositional sequence dependencies, which means the preference of amino acid pair co-occurring at a pair of positions. At the positions in the substrate motif relative to the phosphorylation site, the number of occurrence of each amino acid was counted. At each pair of positions, the number of co-occurrences of each pair of amino acids was counted as well. These numbers are applied in the hypergeometric distribution to calculate the degree of enrichment and reduction for a pair of amino acids at pairwise positions, comparing to the distribution of co-occurrences that would appear randomly given the individual actual occurrences of each amino acid and actual co-occurrences of each amino acid pair (See Methods). From the structural perspective, the kinase would evolve so that binding preferred or avoided particular biophysical feature presented by amino acid pairs at pairwise positions relatively close to the site of phosphorylation. The substrate sequence may also undergo evolutionary selection induced by the functional constraint of binding and recognition. If there are significant interpositional dependencies, a large set of enriched or reduced amino acid pairs with significant deviations would be expected. In unexpected contrast, only small deviations in the frequencies of individual amino acid pairs were observed

from what might be most expected by chance. When using a rigorous criterion to control the false positive rate, these deviations were not statistically significant (see Table 4.2). This means, the data indicates that motif-neighboring sequence deviations arise randomly and not from the evolution of interpositional dependencies.

Though the ATM/ATR substrate dataset is large containing 861 peptide sequences, only a few statistically significant deviations from what would be expected by chance under a position-independent model were identified (Table 4.2). These observations consist of the site of phosphorylation (the phosphoserine or phosphothreonine). Phosphoserine co-occurs more frequently with proline or glycine at position -1, with glycine at position +2, and with serine at position +3. In contrast, phosphothreonine co-occurs less frequently with these amino acids at these positions. The contradicting results may originate from the oriented phosphorylated position, which creates a space of 40 amino acid pairs together with the substrate serine or threonine. A pair of arbitrary positions can create a space of 400 (20 by 20) amino acid pairs. The reduced dimensionality in the second position may enable the statistical boosting power for identifying lower degrees of interpositional dependence.

The same methodology was applied to the substrate datasets of Cdk1/Cyclin B [393] and CK2 [389]. Surprisingly, the set of 71 proline-directed phosphor peptides of Cdk1/Cyclin B has no pairs of individual amino acids observed with significant enrichment or reduction. In addition, the 432 substrate dataset for CK2 only has one amino acid pair at one pair of positions

observed to deviate significantly from what might be expected by chance (see Table 4.2).

In order to increase the statistical power, a decreased space of amino acid pairs at a pair of positions could be done by grouping the 20 amino acids into 6 functional groups: acidic, basic, hydrophobic, polar, aromatic, and a “structural” category including proline and glycine (see Methods). Thus, the space of amino acid pairs is reduced from 20 by 20 to 6 by 6. This way, however, only detects a small number of statistically significant deviations: one case for Cdk1/Cyclin B and four more cases for CK2 (see Table 4.2).

The phosphorylated substrate dataset used in this study are characterized *in vivo*, and is sufficiently large to show a signal should it be present. The results indicate that they do not present a statistically significant indication of the interpositional dependencies within the motif relative to the site of phosphorylation.

Table 4.2: Substrate sequence position pairs demonstrating significant deviations from independence.

Kinase	Position 1	Position 2	Motif	Type	P-value ^a
ATM/ATR	0	2	<u>pS</u> -Q-G	Enriched	4.93 x 10 ⁻⁴
	0	2	<u>pT</u> -Q-G	Reduced	4.93 x 10 ⁻⁴
	0	3	<u>pS</u> -Q-X-S	Enriched	1.16 x 10 ⁻³
	0	3	<u>pT</u> -Q-X-S	Reduced	1.16 x 10 ⁻³
	-1	0	<u>Struct.</u> - <u>pS</u> -Q	Enriched	9.23 x 10 ⁻⁵
	-1	0	<u>Struct.</u> - <u>pT</u> -Q	Reduced	9.23 x 10 ⁻⁵
Cdk1/ Cyclin B	3	4	<u>pS/pT</u> -P-X- <u>Basic</u> - <u>Polar</u>	Reduced	1.63 x 10 ⁻³
CK2	1	2	<u>pS/pT</u> -E-E	Enriched	2.25 x 10 ⁻⁵
	2	4	<u>pS/pT</u> -X- <u>Polar</u> -X- <u>Polar</u>	Enriched	1.39 x 10 ⁻³
	4	5	<u>pS/pT</u> -X-X-X- <u>Acidic</u> - <u>Acidic</u>	Enriched	2.82 x 10 ⁻⁴
	2	4	<u>pS/pT</u> -X- <u>Hyd</u> -X- <u>Struct.</u>	Enriched _b	3.14 x 10 ⁻³
	2	5	<u>pS/pT</u> -X- <u>Hyd</u> -X-X- <u>Hyd.</u>	Enriched _b	1.29 x 10 ⁻³

^aRaw, uncorrected p-value is reported when significance is indicated by comparison to empirical multiple hypothesis testing control (see Methods).

^bThese amino acid pairs were indicated only by the method of Benjamini and Hochberg [396] and not by the empirical heuristic.

Grouped amino acid definitions: Structural (G, P), Basic (K, R), Acidic (D, E), Aromatic (F, Y, W), Hydrophobic (A, I, L, M, V), Polar (C, H, N, Q, S, T).

4.2.2 Incorporation of interpositional dependencies in predicting novel kinase phosphorylation sites

The number of instances of statistically significant deviations from a position-independent description of substrate specificity for these kinases is few. But this does not rule out the possibility that the kinase recognizes the phosphorylation site dependent on the nearby positions. Other factors such as the transient nature of kinase binding, phosphorylation and release may suggest that kinase substrate recognition is different from phosphopeptide-binding modules, which are not as transient in their binding. But to identify a lower degree of interpositional dependence, a large set of substrate sequences for the kinase would be required. The observed small number of interpositional dependent instances may indicate that the kinases tend to recognize their substrate in a position-independent way, or that the cooperative or uncooperative effect of a pair of amino acid is too minute to be discovered provided the currently available sample size of substrate sequences. It may simply reflect evolutionary pressure of kinase binding and phosphorylation as a transient phenomenon, where an increased binding and recognition sequence dependence of the neighboring amino acids around the phosphorylation substrate amino acid may have made the binding less transient.

Despite the very few number of discovered amino acid pairs that are enriched or reduced at pairs of positions, some sub-significant interdependencies might still exist in the sequence motif to facilitate kinase specificity towards its phosphorylation site. The interpositional dependencies can be incorporated into a probabilistic model to predict novel

phosphorylation sites. Here, second-order interpositional sequence information is utilized without considering higher order information. A first-order model is also set up and compared with the second-order model to examine if the interpositional dependent information improves the performance of models in detecting phosphorylation sites (see Methods).

4.2.2.1 Prediction of true substrates from mock substrates made of shuffled controls

For each kinase, the substrate dataset is randomly divided into a 90% training set and a 10% test set. The 90% training set is used to train a first-order and a second-order models, whose ability was tested to identify the withheld true test set from among a background of mock substrates (see Figure 4.1a and Figure 4.2a). The mock substrates are shuffled negative controls generated by shuffling amino acids among substrate sequences, while maintaining their positions relative to the phosphorylation site. In this manner, the individual amino acid frequencies at a single position are preserved, but the interdependencies at pairs of positions are broken. The procedure repeats 1000 times leading to 1000 first-order and second-order models. From the results, it is observed that both first-order and second-order models do a particularly good job of identifying true substrates from among the mock substrates. The first-order models score potential substrates of the kinase based on a function of the probabilities of individual amino acids occurring at single positions among the training data, without considering all higher-order combinations. The second-order models score potential substrates accounting for both the individual probabilities at single positions and the probabilities of pairs of

amino acids at pairs of positions, without considering the triplet and higher-order combinations (see Methods). These models are used to score a set of 870 potential ATM/ATR substrates (87 true substrates, 783 shuffled negative control substrates). The top 10% (87/870) of scores are taken as predicted substrates from the models. First-order models captured a median value of 10.3% (9/87) true positives among putative hits, while the second order models capture a median value of 12.6% (11/87) true positives (See Figure 4.1a). Similar results are seen for CK2 (first-order, 13.6% (6/44); second-order, 18.2% (8/44)). Random selection of sequences would produce 10% true positives.

A different situation is observed in the results for CDK1/Cyclin B. There is less than 10% of the test data scored with a nonzero score of being a potential substrate under the second-order model. This means the top 10% of the test data is actually supplemented with the potential substrates with a probability score of 0, which are randomly selected from all the potential substrates with a probability score of 0 (see Figure 4.1a). To solve this problem, we set a rule to constrain the number of top-scoring predictions for each of 1000 random separations of training and test data to the least value of three numbers, including 10% of the dataset, or the number of predictions having positive scores under the first- or second-order model (see Figure 4.2a). Such approach leads to a mean of 1.32 sequences per test data set are assigned a non-zero score over 1000 tests. For both the first- and second-order models, the median fraction of true positive substrates among these predictions selected is 0%.

In either way, the second-order model does not show substantially better performance than the first-order model. The fact is that the qualities of both models are similar in detecting the substrate peptides from among the true and mock test data. Another similarity lies in the sequences of the true test data and shuffled negative control substrates, which still preserves the amino acid frequencies at individual positions.

4.2.2.2 Prediction of true substrates from mock substrates made of proteomic peptides

The shuffled negative control substrates can be a very stringent dataset for the probabilistic models to identify the true positive substrates from within. A less strict and more realistic test might be done by identifying true substrates from among a background of mock substrates, which are potential proteomic peptides conforming to the known consensus phosphorylation motif for the kinases (see Figure 4.1b and Figure 4.2b). For ATM/ATR, the motif “S/T-Q” was used to select mock substrates. For CDK1/Cyclin B, the motif “S/T-P” was used. For CK2, the majority of the substrate dataset contain the motif “S/T-X-X-D/E”, and this subset was extracted for analysis too. The CK2 phosphorylation motif was used to randomly select proteomic peptides, which were used to test CK2 probabilistic models built from either the full dataset or the subset that conform to the motif. Once again, the dataset was split into a training set (90% substrate dataset) which was used to train the first-and second-order models, and a test set which was combined with true positives (10% substrate dataset) and nine times as many proteomic peptides. The top 10% of high-scoring peptides were taken as potential substrates (see Figure

4.1b). The procedure repeats for 1000 times, and the median fraction of true positives among the top 10% high-scoring peptides was recorded. Because less than 10% of test data was scored non-zero for CDK1/Cyclin B and CK2, the procedure was repeated by making the number of top-scoring predictions in each test equal to the 10% of the substrate data, or the smaller number of predictions having non-zero scores under the first- or second-order model, whichever is the least (see Figure 4.2b).

With the proteomic mock substrates, all models present better predictive ability than that observed in the previous prediction of true substrates from shuffled negative controls. For ATM/ATR, the first-order models predicted a median of 36.8% (32/87) of the top 10% high-scoring sequences were true positives, versus 31.0% (32/87) predicted by second-order models (see Figure 4.1b). For CDK1/Cyclin B, the first model predicted 50.0% (4/8) of hits as true positives, though the second-order only obtained 12.5% (1/8) true positives. For CK2, the prediction results are similar in predicting the true positives from mock substrates with the entire dataset (first-order, 45.5% (20/44); second-order, 34.1% (15/44)) or the subset that conforms to the “S/T-X-X-D/E” motif (first-order, 54.663.6% (18/33); second-order, 33.3% (11/33)). The procedure was repeated limiting the number of top-scoring predictions in each test to 10% of the test data, or the fewer number of peptides with positive scores under the first- or second-order model (see Figure 4.2b). For CDK1/Cyclin B, this produced an average of 1.04 among the 117 trials that receive non-zero scores, and the rest 883 trials had not a single non-zero score prediction. In the 117 trials, the median

fraction of true positives among the scoring peptides is 100% under the first-order model and 0% for the second-order model. This is mostly due to the fact that most of these 117 trials have only one sequence with nonzero score.

With the mock substrates coming from proteomic peptides, the models do better jobs in detecting the true substrates from the mock data. The second-order model, however, is less accurate than the first-order model. This indicates that the frequencies of amino acid pairs among pairs of positions in the training data can not represent the frequencies of amino acid pairs in the test data. Therefore, the second-order models are prone to overfit to the training data.

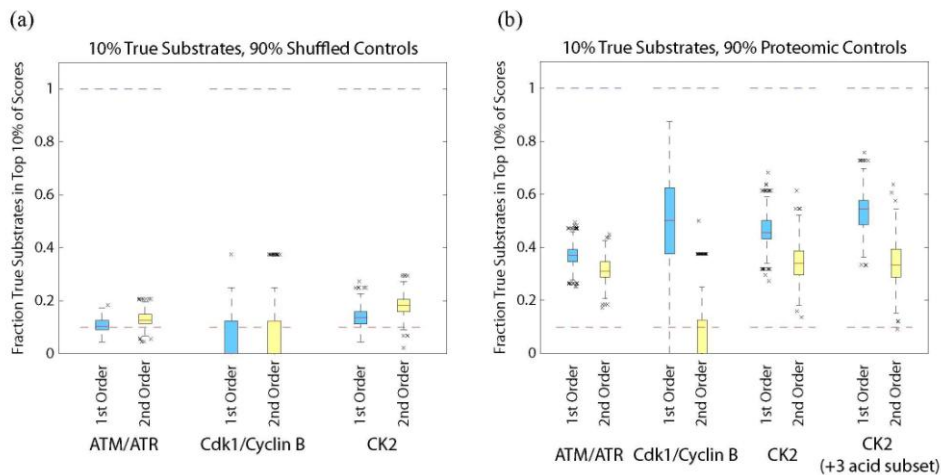


Figure 4.1: Comparison of ability of first- and second- order models to identify kinase substrates. A field of true kinase substrates withheld from training was hidden among mock substrates for each kinase. A field of 10% true and 90% mock substrates was scored using first- and second-order models, and the fraction of true substrates in the top 10% of highest-scoring sequences was counted. The procedure was repeated 1000 times. Plotted boxes span the 25th to 75th percentile of values, with the red line in the boxes marking the medians. Whiskers extend 1.5 times the distance between the 25th and 75th percentiles, and any points more distant from the median are explicitly plotted. Blue and red dashed lines at the value 10% and 100% represent the maximum possible fraction of true substrates in the top 10% of scores and the fraction expected if true and mock substrates were scored randomly, respectively. (A) Mock substrates generated by shuffling true substrates to maintain the probability of each amino acid at each position while breaking interpositional dependencies. (B) Mock substrates chosen by randomly selecting sequences from the human proteome conforming to basic known elements of kinase specificity: “pS/pT-P” for ATM/ATR, “pS/pT-P” for CDK1/Cyclin B, and “pS/pT-X-X-D/E” for CK2. Because CK2 phosphorylates a number of true substrates that do not have +3 D/E, the CK2 models were trained and tested both with all substrate sequences and with only +3 D/E sequences included.

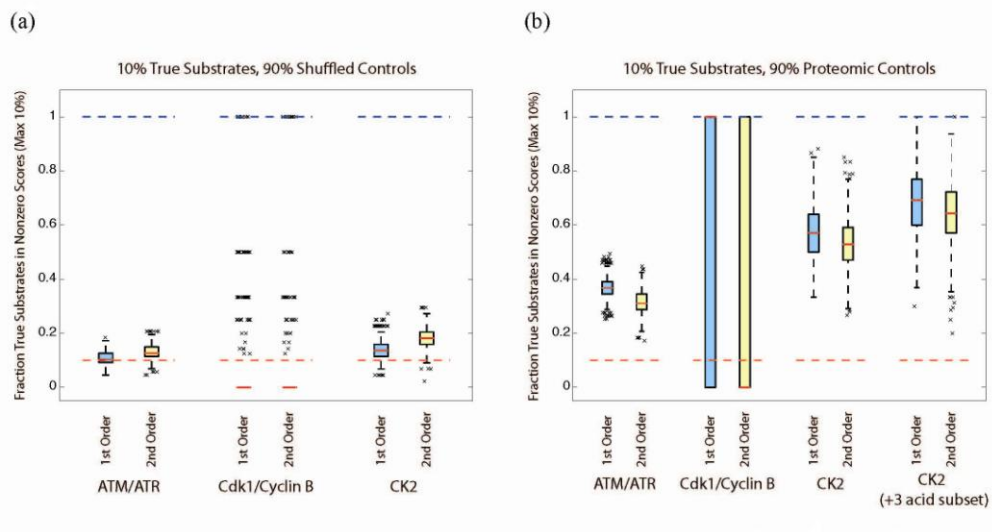


Figure 4.2: Comparison of ability of first- and second- order models to correctly identify true positives, correcting for occurrence of amino acid pairs not present among training data. As in Figure 4.1, but rather than examining the top 10% of scored test sequences, a number of sequences for each random splitting of test and training data was examined equal to the least of: 10% of the tested sequences, or the number of tested sequences given a nonzero score under the first- or second-order model. (A) Mock substrates generated by positionwise shuffling of true substrates. (B) Mock substrates chosen by randomly selecting sequences from the human proteome conforming to basic known elements of kinase specificity.

4.3 Discussion

It is found that the interpositional dependencies among the positions relative to the site of phosphorylation in the ATM/ATR, CDK1/Cyclin B and CK2 substrates sequences are strikingly rare, as studied here. Yet, still a relatively small number of pairs of amino acids that act cooperatively or uncooperatively were identified. For ATM/ATR, serine is favored with a statistical significance over threonine at the site of phosphorylation, which prefers proline or glycine at position -1, glycine at position +2, or serine at position +3. Threonine at the phosphorylation site coincidentally dislikes these instances. This can be explained by the biophysical difference between serine and threonine. Threonine has an additional methyl group attached to the beta carbon, which may render steric clash with the residues at positions in tandem or make the substrate backbone configuration difficult for a kinase to access in the presence of other residue. Other amino acid pairs statistically significantly enriched or reduced at pairs of positions observed in this study are more difficult to explain without structural data.

Notably the crystal structures of the kinases examined here in complex with their substrates do not exist, which is itself a reflection of the transient nature of kinase-substrate peptide binding. A crystal structure of kinase Cdk2/Cyclin A in complex with an optimized substrate peptide was identified [397]. Cdk2 has 66% sequence identity with Cdk1 and similar substrate specificity. In this structure complex, the amino acid side chains of the substrate peptide have no close contacts. This partially explains the scarcity of interpositional dependencies among CDK1/Cyclin B substrate peptide

sequences. Given the close structure similarity of the kinase core amongst all kinases, it may be a structural generalization that there are few close contacts in most kinases.

The datasets used in this study may not be perfect for characterizing the accurate kinase specificity toward its phosphorylation sites. The dataset containing ATM/ATR substrates is determined from a single experimental study. While, these two kinases should be treated individually as they may differ in their specificities in their substrates, that is, they prefer different amino acids at an individual position or different amino acid pairs at pairwise positions [392]. The substrate dataset for CK2 involves the curation of results from earlier experiments which are subject to study biases [394]. For the substrate dataset of CDK1/Cyclin B, the number of identified phosphorylation sites was small, which is perhaps not sufficient enough to tell the subtle frequency of amino acid pairs at pairwise positions [312]. The three datasets studied here span a wide range of sizes and associate with several collection of methodologies. Across all three cases, the same consistent pattern of rare interpositional dependencies is discovered.

It is surprising to find out that the enrichment or reduction of amino acid pairs at pairs of positions in the substrates is not statistically significant than what would be expected, given that each amino acid were independently recognized by a kinase. This could be subject to three possibilities. First, it could be true for a kinase to recognize amino acid of its substrate independently, but this seems not biophysically plausible. Second, it is possible that the second-order information does lie in the substrate sequences,

however, at a lower level requiring a much larger sample size of substrates necessary to detect, however this is unlikely given the large sample sizes used here. Third, when a large number of potential substrates are present *in-vitro*, the kinase will show its statistically significant preferences for certain amino acid combinations. The situation may be different *in vivo*, where the kinase specificity can be limited by the sequences encoded in the genome, the same subcellular localization, and structural accessibility.

The lower level of interpositional dependence in the kinase substrates brings new implications for the evolution of phosphorylation sites and of phosphorylation signaling pathways. If each substrate sequence position contributes independently to the ability of a kinase to recognize its phosphorylation sites, the evolution fitness landscape of substrate as a function of the amino acid at each position will be smooth with a single minimum, as shown in Figure 4.3. There are no non-global local minima as traps in kinase substrate fitness space. For any non-ideal substrate, one or more single mutations would improve the fitness of the substrate for the kinase, and other conjunct double or higher-multiple mutations are not necessary. Most research groups studying phosphopeptide-binding domains, e.g. 14-3-3, SH2, etc have reported that there is interpositional dependence within the phosphopeptide-binding motifs. The substrates of phosphopeptide-binding protein 14-3-3 in the study of Yaffe *et al.*, adhered to one of two mutually exclusive sequence modes [274]. Liu et al. demonstrated that SH2 domains recognize and bind to phosphotyrosine-containing peptides, in which amino acid at certain positions influences amino acid at other positions [398]. The SH2 domain of Crk binds

to phosphotyrosine-containing peptides at the presence of a leucine or a proline in the +3 position. Only a leucine at +3 position allows a proline at +2 position, while a proline at +3 does not. Other signaling domains like SH3, PDZ and WW domains have been shown to have similar properties in their specificities on the substrate sequences [399]. Though these domains usually do not bind phosphopeptides, a subset of WW domains do bind to peptides with phosphoserine and phosphothreonine peptides with proline at position +1.

The results show that the substrate sequence is recognized by the kinase in a position-independent way, but it is recognized by the phosphopeptide-binding domains and signal modular domains, for example, 14-3-3, SH2, WW, SH3 and PDZ etc, in a position-dependent way involving second- or higher-order cooperation. This difference may infer that the evolution of kinase substrate sequences is a fast or easy process in the evolution of phosphorylation signaling pathways. Single mutations are adequate to improve an imperfect substrate sequence to be favored by the kinase without the necessity of higher-order conjunct mutations. This is not the case for the substrates of phosphopeptide-binding domains or other signal modular domains. These domains bind the phosphorylated left behind after kinase activation. Their substrate sequences may require higher-order mutations to evolve so that they become more suited for the phosphopeptide-binding domains to bind (see Figure 4.3) in a non-transient manner. Such binding makes the phosphorylation site play a functional role in signaling pathways. The phenomenon is consistent with the earlier report that there is an enormous number of phosphorylation sites with no functional roles [389]. It is

also consistent with a simple evolutionary process of kinase substrate in the evolution of signaling pathways.

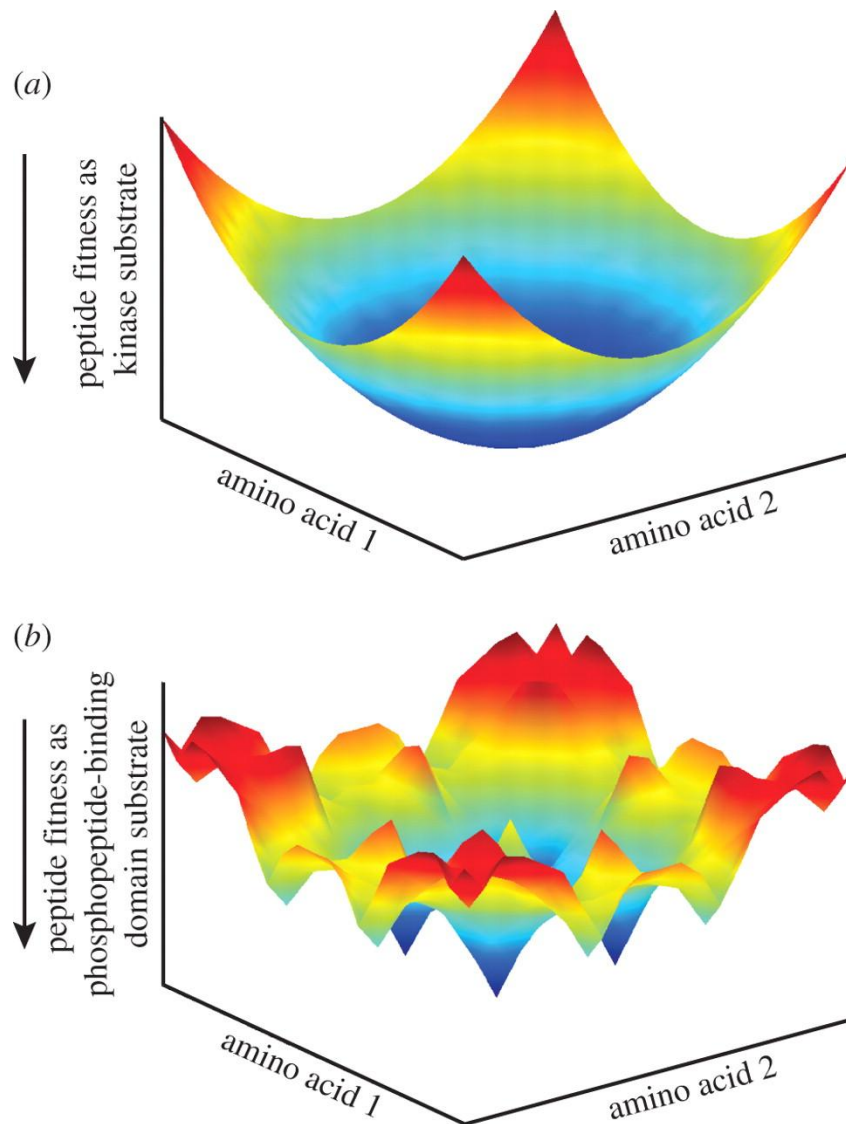


Figure 4.3: Model evolutionary fitness landscapes for substrates of kinases and phosphopeptide-binding domains. (A) Data presented in this paper indicates that kinase substrate fitness may be positionwise independent in the substrate amino acid sequence. (B) Data presented elsewhere [274, 399, 400] indicates that phosphopeptide-binding domains express significant interpositional dependencies. The smooth landscape is compatible with rapid evolution of neighboring sequence as is found in intrinsically disordered regions.

4.4 Conclusion

This study has shown that the first-order model is likely to be sufficient to describe the specificity of kinase for the substrate motif relative to the site of phosphorylation. The addition of second-order information to the first-order model does not seem to improve the ability of the kinase studied here to predict true substrates from among a background of mock substrates. Hence, in order to predict novel substrates, other contextual information may be required, for instance known interactions, subcellular localization, protein structure and distal site recognition may be considered as well, other than considering the sequence alone [390].

The previous finding indicates that intrinsic protein disorder is enriched in the regions in the proximity to the site of phosphorylation [310]. The disordered regions have a fast evolutionary rate, which is consistent with the proposition that neighboring residues around phosphorylation motifs may exhibit evolutionary variation without affecting kinase substrate fitness.

An assumption has been made that the first-order information embedded in the substrate motif reflects evolutionary freedom in the substrate sequences for the specificity of kinases. The energy landscape for substrate fitness is smooth without requiring the higher-order mutations to improve the fitness of any non-optimal potential substrate sequence. While this study has focused on three different kinases, it may not be a sufficient generalization without further examination of substrate fitness for other kinases. However the finding that kinases possess minimal motif information corresponds well to the

general structure and mechanisms known to be exhibited by kinases, namely the transient nature of substrate binding and release.

4.5 Methods

4.5.1 Data sources

Three long lists of substrates from literature were chosen. The first dataset contains 894 phosphorylation sites with the canonical “pS/pT-Q” motif on 700 human proteins identified as the target sites of ATM and ATR by the groups of Gygi and Elledge, who conduct the single experiment using antibody capture, peptide immunoprecipitation (IP) and SILAC followed by liquid chromatography-tandem MS (LC-MS/MS) [392]. The second dataset contains 77 phosphorylation sites for CDK1/Cyclin B identified by covalent capture in the report of Blethrow *et al*, 71 of which are with the canonical proline-directed “pS/pT-P” motif [393]. In addition, a third dataset comes from an expert curating collection of 432 phosphorylation sites of CK2, which are identified specifically by specific amino acid and directly used here [397]. The detailed experimental procedures in each study are provided in Figure 4.4.

4.5.2 Data preparation

For the datasets of ATM/ATR, CDK1/Cyclin B and CK2, the phosphopeptides, the phosphorylation sites and the corresponding gene IDs are collected from the source publication. Each protein sequence is obtained by using the phosphopeptides (without the phosphate ‘p’) to BLAST against the Swiss-Prot database. After preparation, 861 peptide sequences with motif “S/T-Q” for ATM/ATR, 71 peptide sequences with motif “S/T-P” for CDK1/Cyclin B and

432 peptide sequences for CK2 were generated. Each peptide sequence has a window length of 13 residues including the site of phosphorylation, as well as the 6 residues upstream and 6 downstream of the phosphorylation site. Figure 4.4 gives a summary of the steps of data sources and data preparation.

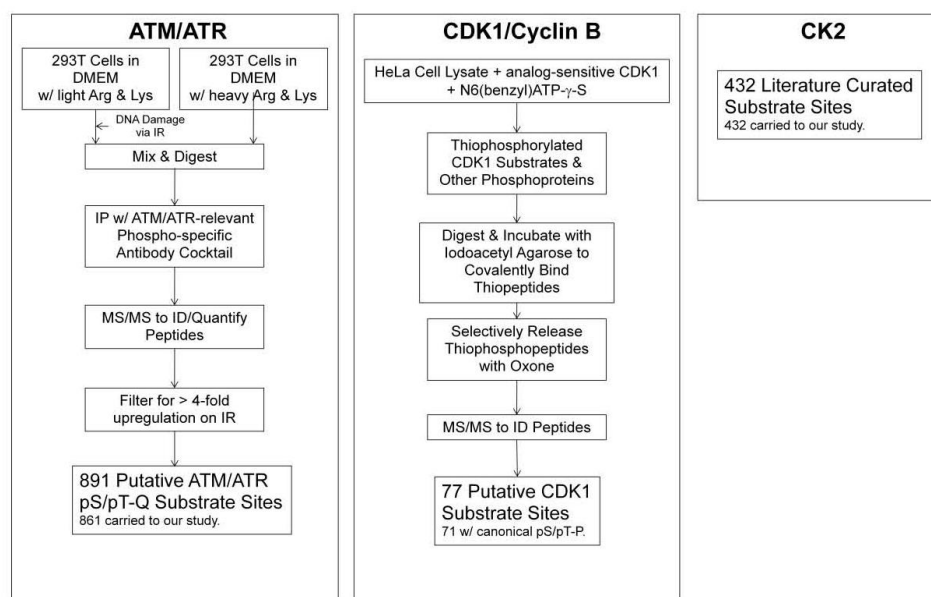


Figure 4.4: Data source and data preparation.

4.5.3 Simplified amino acid alphabet

We also conducted a simplified analysis using the datasets by grouping the 20 amino acids into six groups as follows: Structural (P, G) ->P, Basic (K, R) -> L, Acidic (D, E) -> E, Aromatic (F, Y, W) -> W, Hydrophobic (A, L, I, M, V) -> L, and Polar (C, H, N, Q, S, T) -> Q. Significant results that are generated using the simplified alphabet are reported only when no pair of individual amino acids from the pair of classes is itself capable of representing the significant result.

4.5.4 Statistical analysis of enriched and reduced amino acid pairs.

Exact hypergeometric tests were performed to find whether the type of amino acid at one position within the consensus sequence motif influences the type of amino acid at another position either positively (enrichment) or negatively (reduction). This means to identify those pairs of amino acids at pairs of positions relative to the phosphor-residue, which co-occurred among substrate sequence with a frequency not adequately explained by their individual prevalence and chance. The positions considered were selected by inspection of a sequence log made from substrate sequences as a sequential set of positions that contain more information than the background. The motif logos [401] made from the three datasets for the kinases are shown below (see Figure 4.5).

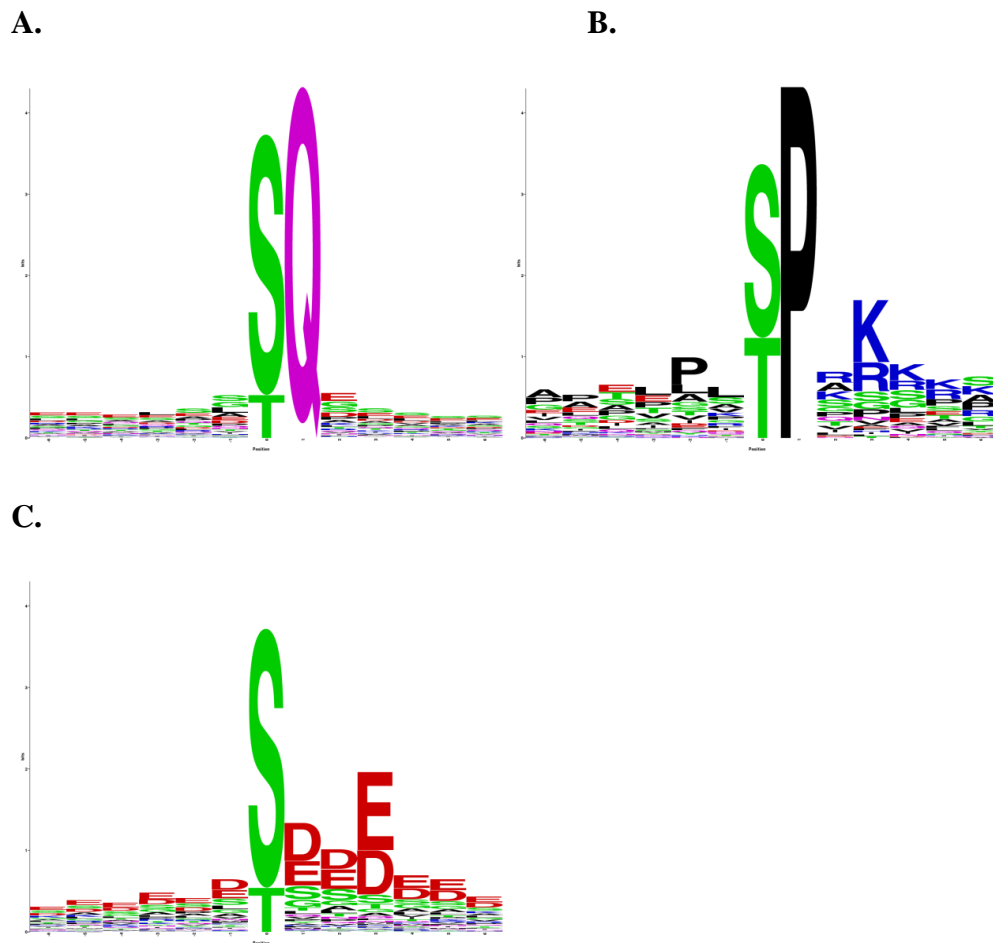


Figure 4.5: Motif logos for substrates analyzed. Sequence position within substrates is along the x-axis. Height of the stacks on the y-axis indicates information content, while the distribution of amino acids within the stack represents frequency. Logos generated with WebLogo 3[401]. (A) 861 substrates of ATM/ATR [392]. (B) 71 substrates of CDK1/Cyclin B [393]. (C) 432 substrates of CK2 [395].

Null hypothesis: amino acid A at position i and amino acid B at position j are independent

Alternative hypothesis: they are dependent, i.e. amino acid A at position i influences amino acid B at position j , either positively (enrichment) or negatively (reduction).

If we define N as the number of peptide sequences, n as the number of amino acid A at position i , m as the number of amino acid B at position j , k as the observed amino acid pair AB, and x as the number of selected AB, the probability for enrichment or reduction is equivalent to the probability of selecting $\geq k$ or $\leq k$ di-amino-acid pair AB (See Illustration 4.1).

$$P_{Enrichment} = \sum_{x=k}^{\min(n,m)} \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad (4.1)$$

$$P_{Reduction} = \sum_{x=0}^k \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad (4.2)$$

If $P_{Enrichment}$ or $P_{Reduction}$ is less than 0.05, the null hypothesis might be rejected, i.e. AB pair is enriched or reduced at position pair i and j .

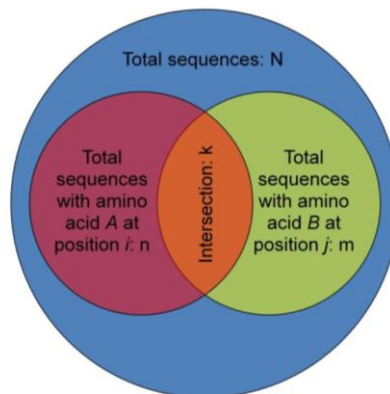


Illustration 4.1: An illustration of statistical hypothesis testing as applied in this analysis.

4.5.5 Statistical significance cutoff determination

At each pair of positions, for each kinase, up to 400 pairs of amino acids are tested for statistically significant enrichment or reduction. It is expected that by chance about 20 statistical significance values lower than 0.05 per 400 tests. In order to control the false positive discovery rate, two criteria were applied.

The first is an empirical method. For each kinase, 1000 negative control datasets were generated which shuffled the amino acids at each position among all substrate peptide sequence. This maintains an identical amino acid composition at each position but breaking any interpositional dependencies, because the amino acid pairs at a pairwise positions were scrambled. For each control dataset, the significances of enrichment or reduction for all pairs of amino acids at all pairs of positions were calculated, as well as the lowest significance value at that position pair. For the true dataset, the significance results which had a more significant p-value than 95% of the randomized controls were noted.

A more rigorous criteria described by Benjamini and Hochberg was used to control the false discovery rate [396]. A q^* value of 0.05 was used, and the set of hypotheses tested were only those that could not be trivially accepted: for enrichment, only those pairs of amino acids that occur at least once independently, and for reduction only those pairs of amino acids that co-occur once or more.

Though the false discovery rate control procedure is only well-suited to series of significance tests that are statistically independent, and the

interdependence of statistical significance values calculated by studying the frequency of pairs of amino acids at each pair of positions is difficult to accurately characterize, results were in strong agreement with those generated using the empirical procedure.

4.5.6 First and second-order model prediction

To predict kinase specific phosphorylation sites, two probabilistic models are built based on the Generalized Kirkwood Superposition Approximation (GKSA) as indicated by Killian et al [402]. The models approximate the probability across all amino acid positions considering only single positions, or single positions and pairs of positions.

The first-order approximation of the probability of a protein sequence (x_1, \dots, x_m) with a length of m residues at positions 1 to m is the product of the independent probability. This model does not consider any cooperative effect of pairwise or higher-order combinations of amino acids.

$$p^{(1)}(x_1, \dots, x_m) = \prod_{i=1}^m p_1(x_i) \quad (4.3)$$

The second-order approximation of the same probability takes the individual and pairwise influences into account, but excludes any higher-order effects. The product in the numerator is calculated from all the probabilities of pairs of amino acids at all pairs of positions. The product in the denominator is calculated from the probabilities of individual amino acids at single positions as in the first-order model.

$$p^{(2)}(x_1, \dots, x_m) = \frac{\prod_{c_2}^m p_2}{[\prod_{c_1}^m p_1]^{m-2}} \quad (4.4)$$

4.5.7 Evaluation of first-and second-order models

Each data set was split into 90% training and 10% test data. The training data were used to build the first-order and second-order models to predict substrate specific phosphorylation sites. The test data were supplemented with nine times as many mock substrates. The mock substrate peptide sequences are generated in a way by shuffling amino acids among the test sequences within each amino acid position. This way preserves exactly the frequencies of amino acids at each position and scrambles the amino acid pairs at pairwise positions. An alternative way to generate mock substrate sequences is by selecting appropriate sequences randomly from the human proteome (International Protein Index [403], v3.55). These sequences were selected from among all sites containing the amino acid sequence “S/T-Q” for ATM/ATR, the sequence “S/T-P” for CDK1/Cyclin B, and the sequence “S/T-X-X-D/E” for CK2. Because an acid at the +3 position relative to the site of phosphorylation is not necessarily required for CK2, additional models were trained and tested against proteomic mock substrates for CK2 using only the subset of substrates conforming to the “S/T-X-X-D/E” motif. The ability of each probabilistic model to identify true substrates among a background of mock substrates was measured by counting the number of true substrates among the best 10% of potential substrates scored. This procedure was repeated using 1000 random divisions of the substrate data into test and training sets, and the mean and

standard deviation are reported. In some cases, the best 10% of potential substrates under a second-order model included some substrates with a probability score of 0. In these cases, the potential substrates with a score of 0 included in the best 10% were selected randomly from all potential substrates with a score of 0.

Because a 0% probability score is not deemed as prediction of phosphorylation site, a maximum the best 10% of potential substrates are taken with limitation to the positive probability scores. For each of the 1000 repeated procedures, the same number of high-ranking sequence was examined for both the first- and second-order model including the least of 10% of the sequence, the number given a nonzero score by the first-order model, or the number given a nonzero score by the second-order model. If no sequences passed these criteria for one of the 1000 divisions, this division was not included in distributions to calculate median statistics. This procedure lead to see 87 of 870 sequences in all 1000 test sets for ATM/ATR, a mean of 2.01 of 80 sequences in 654 of 1000 test sets for CDK1/Cyclin B, and 44 of 440 sequences in all 1000 CK2 test sets, when comparing to positionwise-shuffled substrate sequences. When comparing to proteomic mock substrate sequences, 87 of 870 sequences in all 1000 ATM/ATR test sets, a mean of 1.04 of 80 sequences in 117 of 1000 CDK1/Cyclin B test sets, and a mean of 25.2 or 15.4 in all 1000 test sets for CK2 were used, when the training and test data did or did not include sequences not matching the “S/T-X-X-D/E” motif, respectively. See Figure 4.6 for the ROC curves for first- and second-order models.

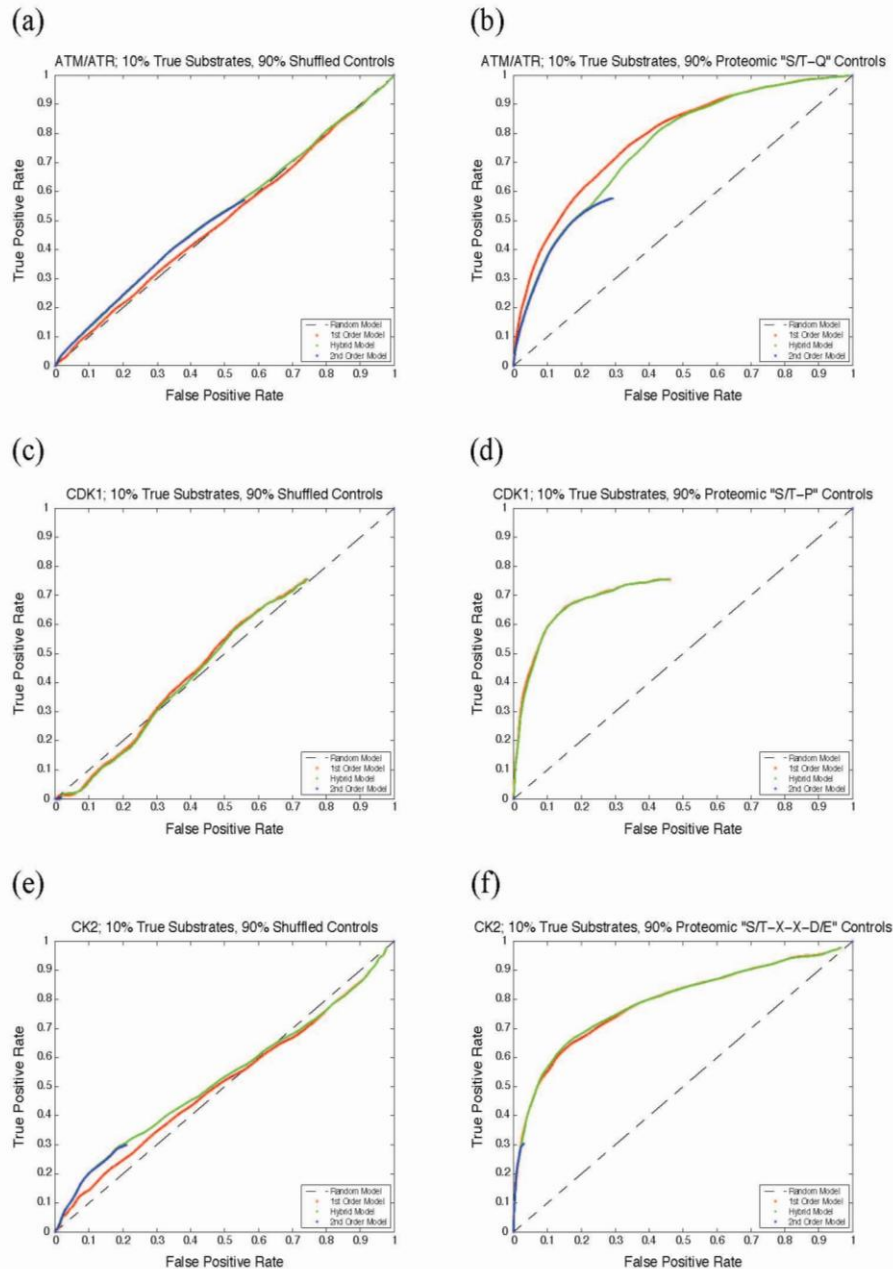


Figure 4.6: ROC curves detail variation of true and false positive rates with probability score. As probability score according to first-order (red lines), second-order (blue lines) and hybrid (green lines, taken as a linear combination of 5% of the second-order score plus 95% of the first-order score for a given sequence) models is varied, the fraction of test sequences (true positive rate) and mock sequences (false positive rate) with an equal or better probability score is plotted. (A, B) ATM/ATR, (C, D) CDK1/CyclinB, (E, F) CK2. (A, C, E) Mock sequences generated by positionwise shuffling of test sequences. (B, D, F) Mock sequences generated by selection of proteomic sequences fitting fundamental elements of substrate motif. Discontinuities represent the point at which all remaining probability scores are zero, at which point the true and false positive rates are both 1.

4.6 Acknowledgement

The authors thank particularly Dr. David Clarke, for helpful discussion. This work was supported by NIH grants CA112967, ES015339, and R24-DK090963, and by a Singapore-MIT Alliance Fellowship.

4.7 Author's Contributions

Brian A. Joughin carried out statistical analysis and wrote the manuscript that was published online. Chengcheng Liu collected the data and participated in the statistical analysis. Douglas A. Lauffenburger, Christopher W.V. Hogue and Michael B. Yaffe participated in the design of the study.

Chapter 5

Conclusions and Future Directions

In this thesis project, new computational methods have been developed to study the structural ensemble and sequence information of disordered regions particularly enriched in prolines. Although these approaches are here only applied to a single case or a set of subjects, they can be extended and generalized to other disordered regions which are functional in signaling pathways.

In Chapter 2, a structural ensemble sampling approach was described to show that a structural ensemble elongation is observed in a specific near-membrane proline-rich disordered region of LRP6 intracellular domain, which has crucial roles in the activation of Wnt signaling pathways. This is achieved by applying simple planar spatial constraints to the LRP6 intracellular domain structural ensembles mimicking the excluded volume effects of the membrane and nearby molecule or molecular assembly. The statistical conformational change of the LRP6 intracellular domain ensemble is determined by comparing the distributions of radius of gyration before and after applying the spatial constraints, and the distributions of pairwise distances across phosphorylation motifs. With the spatial constraints imposed by the membrane and nearby molecules, we see a dramatic decrease in the available conformational space accessible to the disordered ensemble, and an increased fraction of structural ensemble on average has a larger value of radius of gyration. This indicates a new observation that intrinsic ensembles can be reorganized without binding by neighboring molecules, membranes or

structures. The pairwise distance distributions show a clear extension of the region of sequence in the near-membrane region of LRP6 intracellular domain, where the initial phosphorylation events are experimentally demonstrated to take place. We argue that the spatial constraints alone are responsible for the structural elongation in this region, and demonstrate further that a statistical elongation is seen in substrate peptides docked to a kinase protein. While this is not adequate proof of causation, the correspondence between the spatial constraints effect on the near-membrane region and its known experimental role in first site of activation leads us to a hypothesis that spatial constraints may play a role in activation. This would proceed by the constraining system extending the proline-rich disordered region, reduces the random auto-inhibition of the ensemble and making the phosphorylation sites more accessible. In general the mechanism of LRP6 activation has gone without a molecular explanation, and it seems surprising that the first site of phosphorylation would be nearest the membrane. The results of the simulation provided here, while not conclusive, provide the first mechanistic plausible explanation as to how the extracellular binding of Wnt leads to the activation of LRP6 phosphorylation, and points squarely to a non-binding phenomenon of a major rearrangement of the conformational space accessible to the disordered LRP6 region. The extended conformation may be a common requirement for proteins docking to motif binding sites on disordered regions. If these normally tend to be randomly coiled, an extension to reveal their functions may be the simple result of spatial constraints once the protein is sufficiently near to its functional site. This may have profound implications on how the cell is organized and help explain why specific binding does not

proliferate at the point of protein synthesis. If intrinsically disordered proteins of a certain length and composition can auto-inhibit their own motifs, they have a mechanism to deploy to their intended site. Such mechanisms are well known in the cases of folded protein, in particular the insulin gene has three forms, pre-proinsulin, pro-insulin and insulin, which represent stages in the extracellular deployment of the insulin gene. Intrinsic disordered regions that are sensitive to spatial constraints may serve a similar purpose in inhibiting molecular function on the way to cellular localization and deployment, to prevent sensitive signaling networks from becoming active prior to the proper assembly. After assembling the molecular signaling network, the disordered protein ensemble configuration may be elongated by mechanical reorganization of spatial constraints, causing the localized extension of the protein backbone and optimizing access to embedded protein binding or phosphorylation motifs that were otherwise autoinhibited.

The enrichment of proline can increase the potential of an extended backbone configuration due to the uniqueness of proline configuration, however many proline-rich regions are interrupted by other amino acids, so that it is difficult to deduce their properties from sequence alone. The sampling of protein conformational space by tools like TraDES allows the determination of the normal unbound state of the protein backbone. Consistent with early and recent findings of proline-rich disordered regions, it seems that it is important for them to transition to an elongated conformation to fulfill their functions [404]. The protein paradigm: sequence determines structure, which determines function may be modified by our findings. The new knowledge from this initial may well mean that the context of the available

conformational space is also important in determining function, as it seems that the neighborhood of an intrinsically disordered protein and its surrounding spatial constraints may well modulate function in a specific and potentially predictable manner.

The approach described in this Chapter 2 differs from the current molecular simulation approach of Molecular Dynamics, which, after a long period of development, has gained a well-regarded ability to match experimental results. The TraDES software adopted in our simulation approach has been demonstrated to be well suited for sampling conformational space of disordered proteins. The growing use of TraDES like structure sampling methods followed by fitting NMR restraints to find representative ensembles has helped validate the underlying sampling and constraint approach used here. There is a general lack of consideration of spatial constraints in simulation methods, however it is clear that close packing and assembly may present a protein molecule with many such constraints in different regions of the cell. The spatial constraints induced by subcellular localization and proximity of nearby molecules can greatly limit the conformational space of protein disordered regions and as shown in our results, the restriction of conformational space can force an ensemble into a predominant conformation such as an elongated state.

This approach can be applied to study many other single-pass transmembrane proteins, whose intracellular domain has been reported to be mostly disordered [405]. Additionally, the modeled planes may be replaced by other objects or spatial systems representing chromatin or specific three dimensional structures like the ribosome, and systematically create a limited

space for examining the fit of a structural ensemble of any given disordered region. Hence, the modeling method is not exclusive to near-membrane location, but can be applied to constructing planes or objects that represent various spatial constraints originating from different factors in the cellular context. This approach generates information that of course can only be validated by experimental results, however straightforward experiments of this kind are difficult to design without some precise control of the nanoenvironment of experimental molecules. Transmembrane proteins, in particular, have been challenging for structural biology. Methods using fluorescence energy transfer between labeled amino acids can help generate distance distributions to directly compare to the TraDES results.

As indicated in the introduction, proline/serine-rich intrinsically disordered regions are special cases of intrinsic disorder that warrant specific detection and separate annotation from non-proline disordered regions. In Chapter 3 we address the detection of proline-rich regions containing multiple serines or threonines as phosphoacceptors. The algorithm in the work involves developing a log-odd based amino acid propensity index calculated from the compositions of a set of Pro/Ser-rich disordered regions and a set of folded domains as a reference composition. From CASP, many disorder predictors have demonstrated their prediction accuracy, but they do not provide further compositional bias information in the sequence. Though the database UniProt also documents the compositional bias features of a protein, these are largely predicted using the PROSITE pattern predictors, which produce many false positives. A new dataset of Pro/Ser-rich sequences was curated for this study and used for the scoring function developed in Chapter 3. This effort helps

ensure that the PSR predictor is appropriately trained and can find proline-rich disordered regions with a range of different sizes. The evaluation result shows a modest improvement in accuracy produced from the algorithm but with slightly less specificity compared to a traditional pattern finder `ps_scan`, which can search for proline-rich regions using a hidden Markov chain profile. `Ps_scan` can only provide low-complexity information without inferring the disorder capacity in a protein region. The specificity is slightly lower owing to the various lengths of disordered Pro/Ser-rich regions, and the challenge of determining the ends of the scoring function signal. A version that is under development that uses TraDES conformational space information to help bound the termini of these regions shows promise, but was not fully implemented at the completion of this thesis. The study illustrates that for disordered predictors, the compositional bias should be incorporated into the prediction and annotation. This should not just be limited to proline-rich but also extended to other amino-acid-rich disordered regions. For example, glutamine-rich disordered regions, which are identified as important study subjects in conformational diseases, and asparagine-rich regions which are likely to form amyloid plaques.

Finally in Chapter 4, the focus of the study shifted to understand the specificity of the phosphorylation sites which are largely present in proline-rich disordered regions. Our initial intention was to develop a probabilistic model with second- or higher-order interdependencies within the substrate positions in the proximity to the site of phosphorylation. However, we discover that the interpositional dependencies are strikingly rare and incorporating the information is actually not adding boosting power to predict

true phosphorylation sites from a background of mock substrates. Though the dataset we used in the work was not separated into ordered and disordered regions, the results we obtained, in terms of fast evolutionary rate of substrate sequences for a kinase, agreed with the earlier report that phosphorylation sites or other functional motifs are surrounded with amino acids that are disorder-promoting. In order to improve the accuracy of prediction model for a specific kinase, other biological or cellular contexts must be considered too. The DISPHOS and other predictors have added in the properties of disordered protein residues, such as protein flexibility, but they still produce a big number of false positive rates. Hence, other elements, for example, subcellular localization and structural accessibility can affect the kinase specificity in the substrate for the phosphorylation sites.

This coincides nicely with our molecular structural simulation study. Near-membrane effects can limit the conformational space of disordered region to adopt a predominant elongation configuration for the kinase to access. This could explain why GSK3 is able to phosphorylate several sites in motifs present in LRP6 that are not its recognition motif identified *in vitro*. Since spatially constrained disordered regions presents an easily assessable extended conformation near membrane where GSK3 accumulates, the phosphorylation event can naturally take place. The lack of interpositional dependencies may seem contradicting with other research groups' reports about substrates of the phosphopeptide-binding domains or other signal modular domains, which do have a preference over certain amino acid pairs at certain positions, i.e. second- or higher-order interpositional dependencies. Thus, the residue positions in proximity to the phosphorylation site are readily

mutated, which is consistent with the fast evolutionary rate of disordered regions in which they are often found.

For the substrates of those phosphopeptide-binding domains or signaling modular protein domains, the evolutionary energy landscape as a function of amino acid at positions is not smooth with non-global minima, i.e. the binding sites are limited by the mutation rates at certain positions. This can explain why many phosphorylated sites are not recognized by these domains and become functionless. That means the phosphorylation sites activated by kinase may be, in many cases, nonspecifically phosphorylated, but functional domain binding requires a phosphorylated residue with local interpositional dependencies. Interestingly, many positions relative to the phosphorylation site in the substrates of phosphopeptide-binding domains require the presence of proline or glycine, for example, SH2, SH3, and they are located in a proline-rich motif. These motifs have been reported to adopt an extended structure for binding, which contribute to the importance of enrichment of prolines in disordered regions. The statistical analysis can extend to other large quantities of substrates for other kinases and explore more hidden information about interpositional dependencies.

There are additional considerations that arise from the juxtaposition of the three main themes of this thesis. Firstly, the general evolutionary transition to adapt a proline-rich disordered region into a signaling pathway becomes much clearer in light of these results. The results in Chapter 4 point to a nonspecific phosphorylation tendency of originating protein kinases, and that over evolutionary time this specificity is not greatly enhanced at the level of motif recognition. In the context of proline-rich regions, the tendency of

proline to form elongated regions free in solution offers a natural state that might be modified by mutations of amino acids or by domain insertions into disordered regions, or insertions of disordered regions into existing folded regions. Should mutations push the ensemble state into a more coiled or autoinhibited conformation that is sensitive to spatial constraints, this offers a simple evolutionary mechanism to allow the phosphorylation of a residue to act as a chemical marker of cellular position or neighboring state. This allows another small step in evolution to arise, that of a specific phosphoresidue-binding protein that recognizes this chemical marker as distinct from other phosphorylations that are not sensitive to their spatial neighborhood. These small steps in protein evolution may lead to the evolution of specific signaling pathways as follows: (a) non-specific (or less-specific) phosphorylation of many residues (Chapter 4), (b) disordered domain autoinhibition possibly evolving from proline-rich compositionally biased sequence elongation (Chapter 3), (c) relief of that autoinhibition by evolution towards some specific spatial environment (Chapter 2), (d) specific phosphoresidue recognition by duplication and mutation of a phosphor-residue binding domain, and finally (e) integration of the signaling pathway from the bound phosphopeptide – disordered region complex by modular sequence insertion.

Driven by an interest to understand the structural ensemble and sequence properties of disordered regions, particularly enriched in prolines, we have implemented a number of computational methods which may be applied to study other disordered regions. Our preliminary findings with these methods highlight that there is much still to be understood about the mechanism of these regions. The evolutionary and mechanistic implications of

this study are themselves predictive, self-consistent and generally applicable. It is anticipated that the effect of spatial constraints on disordered regions may have broad implications for understanding the evolution and mechanism of signal transduction.

Bibliography

1. Fischer, E., *Einfluss der configuration auf die wirkung der enzyme*. Ber Dt Chem Ges, 1894. **27** p. 2985-2993.
2. Mirsky, A.E. and L. Pauling, *On the Structure of Native, Denatured, and Coagulated Proteins*. Proceedings of the National Academy of Sciences of the United States of America, 1936. **22**(7): p. 439-47.
3. Anfinsen, C.B., *Principles that govern the folding of protein chains*. Science (New York, N.Y.), 1973. **181**(4096): p. 223-30.
4. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis*. Nature, 1958. **181**(4610): p. 662-6.
5. Kendrew, J.C., et al., *Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution*. Nature, 1960. **185**(4711): p. 422-7.
6. Perutz, M.F., et al., *Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution obtained by X-ray analysis* Nature, 1960: p. 416-422.
7. Blake, C.C., et al., *Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution*. Nature, 1965. **206**(4986): p. 757-61.
8. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures*. Journal of molecular biology, 1977. **112**(3): p. 535-42.
9. Schultz, C.P., *Illuminating folding intermediates*. Nature structural biology, 2000. **7**(1): p. 7-10.
10. Breslow, E., et al., *Relative Conformations of Sperm Whale METMYOGLOBIN AND APOMYOGLOBIN IN SOLUTION*. The Journal of biological chemistry, 1965. **240**: p. 304-9.
11. Boulik, M., et al., *An investigation of the conformational changes of histone F2b by high resolution nuclear magnetic resonance*. European journal of biochemistry / FEBS, 1970. **17**(1): p. 151-9.
12. Stellwagen, E., R. Rysavy, and G. Babul, *The conformation of horse heart apocytochrome c*. The Journal of biological chemistry, 1972. **247**(24): p. 8074-7.
13. Fisher, W.R., H. Taniuchi, and C.B. Anfinsen, *On the role of heme in the formation of the structure of cytochrome c*. The Journal of biological chemistry, 1973. **248**(9): p. 3188-95.

14. Bloomer, A.C., et al., *Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits*. Nature, 1978. **276**(5686): p. 362-8.
15. Bode, W., P. Schwager, and R. Huber, *The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution*. Journal of molecular biology, 1978. **118**(1): p. 99-112.
16. Venyaminov, A.T., et al., *Absorption and Circular Dichroism Spectra of Individual Proteins from Escherichia coli Ribosomes*, Pushchino, Russia. 1981.
17. Huber, R. and W.S. Bennett, *Functional significance of flexibility in proteins*. Biopolymers, 1983. **22**(1): p. 261-79.
18. Sigler, P.B., *Transcriptional activation. Acid blobs and negative noodles*. Nature, 1988. **333**(6170): p. 210-2.
19. Watts, J.D., et al., *Thymosins: both nuclear and cytoplasmic proteins*. European journal of biochemistry / FEBS, 1990. **192**(3): p. 643-51.
20. Holt, C. and L. Sawyer, *Caseins as rheomorphic proteins: interpretation of primary and secondary structures of the α 1-, β - and κ -caseins*. J. Chem. Soc., Faraday Trans., 1993. **89**(15): p. 2683-2692.
21. Isbell, D.T., et al., *Metal ion binding to dog osteocalcin studied by 1H NMR spectroscopy*. Biochemistry, 1993. **32**(42): p. 11352-62.
22. Pontius, B.W., *Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association*. Trends in biochemical sciences, 1993. **18**(5): p. 181-6.
23. Gast, K., et al., *Prothymosin alpha: a biologically active protein with random coil conformation*. Biochemistry, 1995. **34**(40): p. 13211-8.
24. Holt, C., N.M. Wahlgren, and T. Drakenberg, *Ability of a beta-casein phosphopeptide to modulate the precipitation of calcium phosphate by forming amorphous dicalcium phosphate nanoclusters*. The Biochemical journal, 1996. **314** (Pt 3): p. 1035-9.
25. Holt, C., et al., *A core-shell model of calcium phosphate nanoclusters stabilized by beta-casein phosphopeptides, derived from sedimentation equilibrium and small-angle X-ray and neutron-scattering measurements*. European journal of biochemistry / FEBS, 1998. **252**(1): p. 73-8.

26. Uversky, V.N., et al., *Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH*. *Biochemistry*, 1999. **38**(45): p. 15009-16.
27. Jahn, T.R. and S.E. Radford, *The Yin and Yang of protein folding*. *Febs Journal*, 2005. **272**(23): p. 5962-5970.
28. Schweers, O., et al., *Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure*. *The Journal of biological chemistry*, 1994. **269**(39): p. 24290-7.
29. Weinreb, P.H., et al., *NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded*. *Biochemistry*, 1996. **35**(43): p. 13709-15.
30. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* *Proteins*, 2000. **41**(3): p. 415-27.
31. Tompa, P., *Intrinsically unstructured proteins*. *Trends in biochemical sciences*, 2002. **27**(10): p. 527-33.
32. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions*. *Nature reviews Molecular cell biology*., 2005. **6**(3): p. 197-208.
33. Dunker, A.K., et al., *Intrinsically disordered protein*. *Journal of molecular graphics & modelling*, 2001. **19**(1): p. 26-59.
34. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics*. *Protein Sci*, 2002. **11**(4): p. 739-56.
35. Daura, X., *Molecular dynamics simulation of peptide folding*. *Theor. Chem. Acc.*, 2006. **116**: p. 297-306.
36. Daura, X., et al., *Unfolded state of peptides*. *Advances in protein chemistry*, 2002. **62**: p. 341-60.
37. Dill, K.A. and D. Shortle, *Denatured states of proteins*. *Annual review of biochemistry*, 1991. **60**: p. 795-825.
38. Uversky, V.N., *Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?* *Cellular and molecular life sciences : CMLS*, 2003. **60**(9): p. 1852-71.
39. Douzou, P.P., G. A., *Proteins at work: "stop-action" pictures at subzero temperatures*. *Adv. Protein Chem*, 1984(36): p. 245-361.
40. Hobohm, U. and C. Sander, *Enlarged representative set of protein structures*. *Protein Sci*, 1994. **3**(3): p. 522-4.

41. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000*. Nucleic Acids Res, 2000. **28**(1): p. 45-8.
42. Obradovic, Z., et al., *Predicting intrinsic disorder from amino acid sequence*. Proteins, 2003. **53 Suppl 6**: p. 566-72.
43. Le Gall, T., et al., *Intrinsic disorder in the Protein Data Bank*. J Biomol Struct Dyn, 2007. **24**(4): p. 325-42.
44. Mohan, A., V.N. Uversky, and P. Radivojac, *Influence of Sequence Changes and Environment on Intrinsically Disordered Proteins*. PLoS computational biology, 2009. **5**(9).
45. Weiss, M.A., et al., *Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA*. Nature, 1990. **347**(6293): p. 575-8.
46. Arcus, V.L., et al., *Toward solving the folding pathway of barnase: the complete backbone ¹³C, ¹⁵N, and ¹H NMR assignments of its pH-denatured state*. Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(20): p. 9412-6.
47. Arcus, V.L., et al., *A comparison of the pH, urea, and temperature-denatured states of barnase by heteronuclear NMR: implications for the initiation of protein folding*. Journal of molecular biology, 1995. **254**(2): p. 305-21.
48. Eliezer, D., et al., *Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding*. Nature structural biology, 1998. **5**(2): p. 148-55.
49. Dobson, C.M. and P.J. Hore, *Kinetic studies of protein folding using NMR spectroscopy*. Nature structural biology, 1998. **5 Suppl**: p. 504-7.
50. Dyson, H.J. and P.E. Wright, *Equilibrium NMR studies of unfolded and partially folded proteins*. Nature structural biology, 1998. **5 Suppl**: p. 499-503.
51. Dyson, H.J. and P.E. Wright, *Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states*. Methods in enzymology, 2001. **339**: p. 258-70.
52. Dyson, H.J. and P.E. Wright, *Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance*. Advances in protein chemistry, 2002. **62**: p. 311-40.
53. Barbar, E., *NMR characterization of partially folded and unfolded conformational ensembles of proteins*. Biopolymers, 1999. **51**(3): p. 191-207.

54. Bracken, C., *NMR spin relaxation methods for characterization of disorder and folding in proteins*. Journal of molecular graphics & modelling, 2001. **19**(1): p. 3-12.
55. Yao, J., et al., *NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding*. Biochemistry, 2001. **40**(12): p. 3561-71.
56. Jensen, M.R., et al., *Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts*. Journal of the American Chemical Society, 2010. **132**(4): p. 1270-2.
57. Huang, A. and C.M. Stultz, *The effect of a DeltaK280 mutation on the unfolded state of a microtubule-binding repeat in Tau*. PLoS Comput Biol, 2008. **4**(8): p. e1000155.
58. Fisher, C.K., A. Huang, and C.M. Stultz, *Modeling intrinsically disordered proteins with bayesian statistics*. Journal of the American Chemical Society, 2010. **132**(42): p. 14919-27.
59. Marsh, J.A. and J.D. Forman-Kay, *Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints*. Journal of molecular biology, 2009. **391**(2): p. 359-74.
60. Nodet, G., et al., *Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings*. Journal of the American Chemical Society, 2009. **131**(49): p. 17908-18.
61. Jensen, M.R., et al., *Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings*. Structure, 2009. **17**(9): p. 1169-85.
62. Bernadó P., et al., *A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(47): p. 17002-7.
63. Bernadó P., et al., *Structural characterization of flexible proteins using small-angle X-ray scattering*. Journal of the American Chemical Society, 2007. **129**(17): p. 5656-64.
64. Bernadó P., et al., *Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings*. Journal of the American Chemical Society, 2005. **127**(51): p. 17968-9.
65. Ganguly, D. and J. Chen, *Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states*. Journal of molecular biology, 2009. **390**(3): p. 467-77.

66. Wishart, D.S., B.D. Sykes, and F.M. Richards, *The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy*. *Biochemistry*, 1992. **31**(6): p. 1647-51.
67. Wishart, D.S. and B.D. Sykes, *Chemical shifts as a tool for structure determination*. *Methods in enzymology*, 1994. **239**: p. 363-92.
68. Spera, S. and A. Bax, *Empirical correlation between protein backbone conformation and C-alpha and C-beta C-13 nuclear-magnetic-resonance chemical-shifts*. *J Am Chem Soc*, 1991. **113**: p. 5490-5492.
69. Wishart, D.S., B.D. Sykes, and F.M. Richards, *Relationship between nuclear magnetic resonance chemical shift and protein secondary structure*. *Journal of molecular biology*, 1991. **222**(2): p. 311-33.
70. Adler, A.J., N.J. Greenfield, and G.D. Fasman, *Circular dichroism and optical rotatory dispersion of proteins and polypeptides*. *Methods Enzymol*, 1973. **27**: p. 675-735.
71. Fasman, G.D., *Circular Dichroism and the Conformational Analysis of Biomolecules*1996, New York: Plenum Press.
72. O.Glatter, O.K., *Small Angle X-ray Scattering*. England, London, 1982.
73. Doniach, S., et al., *Partially folded states of proteins: characterization by X-ray scattering*. *J Mol Biol*, 1995. **254**(5): p. 960-7.
74. Semisotnov, G.V., et al., *Protein globularization during folding. A study by synchrotron small-angle X-ray scattering*. *J Mol Biol*, 1996. **262**(4): p. 559-74.
75. Kataoka, M. and Y. Goto, *X-ray solution scattering studies of protein folding*. *Fold Des*, 1996. **1**(5): p. R107-14.
76. Lackowicz, J., *Principals of Fluorescence Spectroscopy*1999, New York: Kluwer Academic/Plenum Publishers.
77. Stryer, L., *Fluorescence spectroscopy of proteins*. *Science*, 1968. **162**(3853): p. 526-33.
78. Permyakov, E.A., *Luminescence spectroscopy of proteins*1993, London: CRC PRes.
79. G.V. Semisotnov, N.A.R., O.I. Razgulyaev, V.N. Uversky, A.F. Gripas, R.I. Gilmanshin, *Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe*. *Biopolymers*, 1991(31): p. 119-128.

80. Yang, S., et al., *Multidomain assembled states of Hck tyrosine kinase in solution*. Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(36): p. 15757-62.
81. Rozycki, B., Y.C. Kim, and G. Hummer, *SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions*. Structure, 2011. **19**: p. 109-116.
82. Petsko, G.A. and D. Ringe, *Fluctuations in protein structure from X-ray diffraction*. Annual review of biophysics and bioengineering, 1984. **13**: p. 331-71.
83. Ferreon, A.C.M., et al., *Single-molecule fluorescence studies of intrinsically disordered proteins*. Methods in enzymology, 2010. **472**: p. 179-204.
84. Chen, Y., S.L. Campbell, and N.V. Dokholyan, *Deciphering protein dynamics from NMR data using explicit structure sampling and selection*. Biophys J, 2007. **93**(7): p. 2300-6.
85. Huang, F., et al., *Multiple conformations of full-length p53 detected with single-molecule fluorescence resonance energy transfer*. Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(49): p. 20758-63.
86. Sickmeier, M., et al., *DisProt: the Database of Disordered Proteins*. Nucleic Acids Res, 2007. **35**(Database issue): p. D786-93.
87. Williams, R.J., *The conformational mobility of proteins and its functional significance*. Biochemical Society transactions, 1978. **6**(6): p. 1123-6.
88. Williams, R.J., *The conformation properties of proteins in solution*. Biological reviews of the Cambridge Philosophical Society, 1979. **54**(4): p. 389-437.
89. Garnier, J., D.J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins*. J Mol Biol, 1978. **120**(1): p. 97-120.
90. Gibrat, J.F., J. Garnier, and B. Robson, *Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs*. J Mol Biol, 1987. **198**(3): p. 425-43.
91. Garnier, J., J.F. Gibrat, and B. Robson, *GOR method for predicting protein secondary structure from amino acid sequence*. Methods Enzymol, 1996. **266**: p. 540-53.

92. Kloczkowski, A., et al., *Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence*. Proteins, 2002. **49**(2): p. 154-66.
93. Romero, P., et al., *Identifying disordered regions in proteins from amino acid sequence*. 1997. **1**: p. 90-95.
94. Romero, Obradovic, and Dunker, *Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family*. Genome informatics. Workshop on Genome Informatics, 1997. **8**: p. 110-124.
95. Romero, P., et al., *Thousands of proteins likely to have long disordered regions*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 1998: p. 437-48.
96. Romero, P., et al., *Sequence complexity of disordered protein*. Proteins, 2001. **42**(1): p. 38-48.
97. Garner, E., et al., *Predicting Binding Regions within Disordered Proteins*. Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 41-50.
98. Li, X., et al., *Predicting Protein Disorder for N-, C-, and Internal Regions*. Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 30-40.
99. Vucetic, S., et al., *Flavors of protein disorder*. Proteins, 2003. **52**(4): p. 573-84.
100. Radivojac, P., et al., *Prediction of boundaries between intrinsically ordered and disordered protein regions*. Pac Symp Biocomput, 2003: p. 216-27.
101. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder*. Proteins, 2005. **61 Suppl 7**: p. 176-82.
102. Xue, B., et al., *PONDR-FIT: a meta-predictor of intrinsically disordered amino acids*. Biochim Biophys Acta, 2010. **1804**(4): p. 996-1010.
103. Ferron, F., et al., *A practical overview of protein disorder prediction methods*. Proteins, 2006. **65**(1): p. 1-14.
104. He, B., et al., *Predicting intrinsic disorder in proteins: an overview*. Cell Research, 2009. **19**(8): p. 929-949.
105. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins*. Brief Bioinform, 2010. **11**(2): p. 225-43.

106. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods*. Molecular Biosystems, 2012. **8**(1): p. 114-21.
107. Melamud, E. and J. Moult, *Evaluation of disorder predictions in CASP5*. Proteins, 2003. **53 Suppl 6**: p. 561-5.
108. Jin, Y. and R.L. Dunbrack, Jr., *Assessment of disorder predictions in CASP6*. Proteins, 2005. **61 Suppl 7**: p. 167-75.
109. Bordoli, L., F. Kiefer, and T. Schwede, *Assessment of disorder predictions in CASP7*. Proteins, 2007. **69 Suppl 8**: p. 129-36.
110. Noivirt-Brik, O., J. Prilusky, and J.L. Sussman, *Assessment of disorder predictions in CASP8*. Proteins, 2009. **77 Suppl 9**: p. 210-6.
111. Monastyrskyy, B., et al., *Evaluation of disorder predictions in CASP9*. Proteins, 2011. **79 Suppl 10**: p. 107-18.
112. Peng, K., et al., *Optimizing long intrinsic disorder predictors with protein evolutionary information*. J Bioinform Comput Biol, 2005. **3**(1): p. 35-60.
113. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder*. BMC bioinformatics, 2006. **7**: p. 208.
114. Jones, D.T. and J.J. Ward, *Prediction of disordered regions in proteins from position specific score matrices*. Proteins, 2003. **53 Suppl 6**: p. 573-8.
115. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life*. J Mol Biol, 2004. **337**(3): p. 635-45.
116. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*. Bioinformatics (Oxford, England), 2005. **21**(16): p. 3369-76.
117. Hirose, S., et al., *POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions*. Bioinformatics (Oxford, England), 2007. **23**(16): p. 2046-53.
118. Shimizu, K., S. Hirose, and T. Noguchi, *POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix*. Bioinformatics (Oxford, England), 2007. **23**(17): p. 2337-8.
119. Shimizu, K., et al., *Predicting mostly disordered proteins by using structure-unknown protein data*. BMC bioinformatics, 2007. **8**: p. 78.

120. Hirose, S., K. Shimizu, and T. Noguchi, *POODLE-I: Disordered Region Prediction by Integrating POODLE Series and Structural Information Predictors Based on a Workflow Approach*. In *Silico Biol*, 2010. **10**(3): p. 185-91.
121. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder*. *Nucleic acids research*, 2003. **31**(13): p. 3701-8.
122. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content*. *Bioinformatics*, 2005. **21**(16): p. 3433-4.
123. Dosztányi, Z., et al., *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins*. *Journal of molecular biology*, 2005. **347**(4): p. 827-39.
124. Su, C.T., C.Y. Chen, and C.M. Hsu, *iPDA: integrated protein disorder analyzer*. *Nucleic Acids Res*, 2007. **35**(Web Server issue): p. W465-72.
125. Ishida, T. and K. Kinoshita, *PrDOS: prediction of disordered protein regions from amino acid sequence*. *Nucleic acids research*, 2007. **35**(Web Server issue): p. W460-4.
126. Schlessinger, A., et al., *Improved disorder prediction by combination of orthogonal approaches*. *PloS one*, 2009. **4**(2): p. e4433.
127. Mizianty, M.J., et al., *Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources*. *Bioinformatics*, 2010. **26**(18): p. i489-96.
128. Kozłowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins*. *BMC bioinformatics*, 2012. **13**(1): p. 111.
129. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures*. *Computers & chemistry*, 1994. **18**(3): p. 269-85.
130. Callebaut, I., et al., *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives*. *Cell Mol Life Sci*, 1997. **53**(8): p. 621-45.
131. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics*. *Structure (London, England : 1993)*, 2003. **11**(11): p. 1453-9.

132. Liu, J. and B. Rost, *NORSp: Predictions of long regions without regular secondary structure*. Nucleic acids research, 2003. **31**(13): p. 3833-5.
133. Weathers, E.A., et al., *Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein*. FEBS letters, 2004. **576**(3): p. 348-52.
134. MacCallum, R.M., *Order/disorder prediction with self organizing maps*.
135. Garbuzynskiy, S.O., M.Y. Lobanov, and O.V. Galzitskaya, *To be folded or to be unfolded?* Protein science : a publication of the Protein Society, 2004. **13**(11): p. 2871-7.
136. Galzitskaya, O.V., S.O. Garbuzynskiy, and M.Y. Lobanov, *FoldUnfold: web server for the prediction of disordered regions in protein chain*. Bioinformatics (Oxford, England), 2006. **22**(23): p. 2948-9.
137. Galzitskaya, O.V., S.O. Garbuzynskiy, and M.Y. Lobanov, *Expected packing density allows prediction of both amyloidogenic and disordered regions in protein chains*. Journal of Physics-Condensed Matter, 2007. **19**(28).
138. Cheng, J.S., M.J.; Baldi, P., *Accurate prediction of protein disordered regions by mining protein structure data*. Data Mining and Knowledge Discovery, 2005. **11**: p. 213-222.
139. Prilusky, J., et al., *FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded*. Bioinformatics (Oxford, England), 2005. **21**(16): p. 3435-8.
140. Coeytaux, K. and A. Poupon, *Prediction of unfolded segments in a protein sequence based on amino acid composition*. Bioinformatics (Oxford, England), 2005. **21**(9): p. 1891-900.
141. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines*. Nucleic acids research, 2006. **34**(Web Server issue): p. W164-8.
142. Yang, M.Q. and J.Y. Yang, *IUP: Intrinsically unstructured protein predictor - A software tool for analyzing polypeptide sequences*. BIBE 2006: Sixth IEEE Symposium on Bioinformatics and BioEngineering, Proceedings, 2006: p. 3-11.
143. Su, C.T., C.Y. Chen, and Y.Y. Ou, *Protein disorder prediction by condensed PSSM considering propensity for order or disorder*. BMC bioinformatics, 2006. **7**.

144. Schlessinger, A., M. Punta, and B. Rost, *Natively unstructured regions in proteins identified from contact predictions*. Bioinformatics (Oxford, England), 2007. **23**(18): p. 2376-84.
145. Ishida, T. and K. Kinoshita, *Prediction of disordered regions in proteins based on the meta approach*. Bioinformatics (Oxford, England), 2008. **24**(11): p. 1344-8.
146. Bulashevskaya, A. and R. Eils, *Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered*. Journal of theoretical biology, 2008. **254**(4): p. 799-803.
147. Wang, L. and U.H. Sauer, *OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields*. Bioinformatics (Oxford, England), 2008. **24**(11): p. 1401-2.
148. McGuffin, L.J., *Intrinsic disorder prediction from the analysis of multiple protein fold recognition models*. Bioinformatics (Oxford, England), 2008. **24**(16): p. 1798-804.
149. Xue, B., et al., *CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions*. FEBS letters, 2009. **583**(9): p. 1469-74.
150. Deng, X., J. Eickholt, and J. Cheng, *PreDisorder: ab initio sequence-based prediction of protein disordered regions*. BMC bioinformatics, 2009. **10**: p. 436.
151. Lobanov, M.Y. and O.V. Galzitskaya, *The Ising model for prediction of disordered residues from protein sequence alone*. Phys Biol, 2011. **8**(3): p. 035004.
152. Mizianty, M.J., et al., *In-silico prediction of disorder content using hybrid sequence representation*. BMC bioinformatics, 2011. **12**: p. 245.
153. Fukuchi, S., et al., *Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains: its application to human transcription factors*. BMC Struct Biol, 2009. **9**: p. 26.
154. Fukuchi, S., et al., *Binary classification of protein molecules into intrinsically disordered and ordered segments*. BMC Struct Biol, 2011. **11**: p. 29.
155. Huang, F., et al., *Subclassifying disordered proteins by the ch-cdf plot method*. Pac Symp Biocomput, 2012: p. 128-39.
156. Zhang, T., et al., *SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method*. J Biomol Struct Dyn, 2012. **29**(4): p. 799-813.

157. Williams, R.M., et al., *The protein non-folding problem: amino acid determinants of intrinsic order and disorder*. Pac Symp Biocomput, 2001: p. 89-100.
158. Radivojac, P., et al., *Intrinsic disorder and functional proteomics*. Biophys J, 2007. **92**(5): p. 1439-56.
159. Vacic, V., et al., *Composition Profiler: a tool for discovery and visualization of amino acid composition differences*. BMC bioinformatics, 2007. **8**: p. 211.
160. Campen, A., et al., *TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder*. Protein Pept Lett, 2008. **15**(9): p. 956-63.
161. Wootton, J.C. and S. Federhen, *Analysis of compositionally biased regions in sequence databases*. Methods Enzymol, 1996. **266**: p. 554-71.
162. Wootton JC, F.S., *Statistics of local complexity in aminoacidsequences and sequence databases*. Comp Chem, 1993. **17**: p. 149-163.
163. Rauscher, S. and R. Pomes, *Molecular simulations of protein disorder*. Biochem Cell Biol, 2010. **88**(2): p. 269-90.
164. Szilagyi, A., D. Gyorffy, and P. Zavodszky, *The twilight zone between protein order and disorder*. Biophys J, 2008. **95**(4): p. 1612-26.
165. Turoverov, K.K., I.M. Kuznetsova, and V.N. Uversky, *The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation*. Prog Biophys Mol Biol, 2010. **102**(2-3): p. 73-84.
166. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules*. Nat Struct Biol, 2002. **9**(9): p. 646-52.
167. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*. Chem Phys Lett, 1999. **314** p. 141-151.
168. Chen, L.Y. and N.J.M. Horing, *An exact formulation of hyperdynamics simulations*. The Journal of chemical physics, 2007. **126**(22): p. 224103.
169. Bezsonova, I., et al., *Oxygen as a paramagnetic probe of clustering and solvent exposure in folded and unfolded states of an SH3 domain*. J Am Chem Soc, 2007. **129**(6): p. 1826-35.
170. Hilser, V.J. and E. Freire, *Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen*

- exchange protection factors*. Journal of molecular biology, 1996. **262**(5): p. 756-72.
171. Hilser, V.J., et al., *A statistical thermodynamic model of the protein ensemble*. Chemical reviews, 2006. **106**(5): p. 1545-58.
 172. Hilser, V.J., et al., *The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble*. Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(17): p. 9903-8.
 173. Liu, T., S.T. Whitten, and V.J. Hilser, *Ensemble-based signatures of energy propagation in proteins: a new view of an old phenomenon*. Proteins, 2006. **62**(3): p. 728-38.
 174. Liu, T., S.T. Whitten, and V.J. Hilser, *Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(11): p. 4347-52.
 175. Pan, H., J.C. Lee, and V.J. Hilser, *Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble*. Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(22): p. 12020-5.
 176. Hilser, V.J. and E.B. Thompson, *Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(20): p. 8311-5.
 177. Feldman, H.J. and C.W. Hogue, *A fast method to sample real protein conformational space*. Proteins, 2000. **39**(2): p. 112-31.
 178. Feldman, H.J. and C.W.V. Hogue, *Probabilistic sampling of protein conformations new hope for brute force ?* Proteins, 2002. **46**(1): p. 8-23.
 179. Dunbrack, R.L., Jr. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction*. J Mol Biol, 1993. **230**(2): p. 543-74.
 180. Zhang, C., et al., *Determination of atomic desolvation energies from the structures of crystallized proteins*. J Mol Biol, 1997. **267**(3): p. 707-26.
 181. Bryant, S.H. and C.E. Lawrence, *An empirical energy function for threading protein sequence through the folding motif*. Proteins, 1993. **16**(1): p. 92-112.

182. McConkey, B.J., V. Sobolev, and M. Edelman, *Discrimination of native protein structures using atom-atom contact scoring*. Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3215-20.
183. Choy, W.Y. and J.D. Forman-Kay, *Calculation of ensembles of structures representing the unfolded state of an SH3 domain*. Journal of molecular biology, 2001. **308**(5): p. 1011-32.
184. Marsh, J.A., et al., *Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure*. Journal of molecular biology, 2007. **367**(5): p. 1494-510.
185. Marsh, J.A., et al., *Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators*. Structure (London, England : 1993), 2010. **18**(9): p. 1094-103.
186. Mittag, T., et al., *Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase*. Structure (London, England : 1993), 2010. **18**(4): p. 494-506.
187. Stelzer, A.C., et al., *Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs*. Methods (San Diego, Calif.), 2009. **49**(2): p. 167-73.
188. Frank, A.T., et al., *Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition*. Nucleic acids research, 2009. **37**(11): p. 3670-9.
189. Jha, A.K., et al., *Statistical coil model of the unfolded state: resolving the reconciliation problem*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(37): p. 13099-104.
190. Jha, A.K., et al., *Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library*. Biochemistry, 2005. **44**(28): p. 9691-702.
191. Hus, J.C., D. Marion, and M. Blackledge, *Determination of protein backbone structure using only residual dipolar couplings*. Journal of the American Chemical Society, 2001. **123**(7): p. 1541-1542.
192. Salmon, L., et al., *NMR characterization of long-range order in intrinsically disordered proteins*. Journal of the American Chemical Society, 2010. **132**(24): p. 8407-18.
193. Mittag, T. and J.D. Forman-Kay, *Atomic-level characterization of disordered protein ensembles*. Current opinion in structural biology, 2007. **17**(1): p. 3-14.

194. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions*. Journal of molecular biology, 1997. **268**(1): p. 209-25.
195. Brünger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination*. Acta crystallographica. Section D, Biological crystallography, 1998. **54**(Pt 5): p. 905-21.
196. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package*. Journal of magnetic resonance (San Diego, Calif. : 1997), 2003. **160**(1): p. 65-73.
197. Brünger, A.T., *XPLOR Manual Version 3.1*, 1993, Yale University Press: New Haven.
198. Yoon, M.-K., et al., *Residual structure within the disordered C-terminal segment of p21(Waf1/Cip1/Sdi1) and its implications for molecular recognition*. Protein science : a publication of the Protein Society, 2009. **18**(2): p. 337-47.
199. Oldfield, C.J., et al., *Comparing and combining predictors of mostly disordered proteins*. Biochemistry, 2005. **44**(6): p. 1989-2000.
200. Dunker, A.K., et al., *Intrinsic disorder and protein function*. Biochemistry, 2002. **41**(21): p. 6573-82.
201. Uversky, V.N. and A.L. Fink, *Conformational constraints for amyloid fibrillation: the importance of being unfolded*. Biochim Biophys Acta, 2004. **1698**(2): p. 131-53.
202. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins*. J Mol Biol, 2002. **323**(3): p. 573-84.
203. Cheng, Y., et al., *Abundance of intrinsic disorder in protein associated with cardiovascular disease*. Biochemistry, 2006. **45**(35): p. 10448-60.
204. Dunker, A.K., et al., *Flexible nets. The roles of intrinsic disorder in protein interaction networks*. Febs Journal, 2005. **272**(20): p. 5129-48.
205. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling*. J Mol Recognit, 2005. **18**(5): p. 343-84.
206. Bustos, D.M. and A.A. Iglesias, *Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins*. Proteins, 2006. **63**(1): p. 35-42.

207. Cortese, M.S., V.N. Uversky, and A.K. Dunker, *Intrinsic disorder in scaffold proteins: getting more from less*. Prog Biophys Mol Biol, 2008. **98**(1): p. 85-106.
208. Patil, A. and H. Nakamura, *Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks*. FEBS Lett, 2006. **580**(8): p. 2041-5.
209. Ekman, D., et al., *What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?* Genome Biol, 2006. **7**(6): p. R45.
210. Haynes, C., et al., *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes*. PLoS Comput Biol, 2006. **2**(8): p. e100.
211. Dosztanyi, Z., et al., *Disorder and sequence repeats in hub proteins and their implications for network evolution*. J Proteome Res, 2006. **5**(11): p. 2985-95.
212. Singh, G.P. and D. Dash, *Intrinsic disorder in yeast transcriptional regulatory network*. Proteins, 2007. **68**(3): p. 602-5.
213. Singh, G.P., M. Ganapathi, and D. Dash, *Role of intrinsic disorder in transient interactions of hub proteins*. Proteins, 2007. **66**(4): p. 761-5.
214. Dunker, A.K., et al., *The unfoldomics decade: an update on intrinsically disordered proteins*. BMC Genomics, 2008. **9** Suppl 2: p. S1.
215. Madania, A., et al., *The Saccharomyces cerevisiae homologue of human Wiskott-Aldrich syndrome protein Las17p interacts with the Arp2/3 complex*. Mol Biol Cell, 1999. **10**(10): p. 3521-38.
216. Dawson, R., et al., *The N-terminal domain of p53 is natively unfolded*. J Mol Biol, 2003. **332**(5): p. 1131-41.
217. Muller-Tiemann, B.F., T.D. Halazonetis, and J.J. Elting, *Identification of an additional negative regulatory region for p53 sequence-specific DNA binding*. Proc Natl Acad Sci U S A, 1998. **95**(11): p. 6079-84.
218. Hart, M.J., et al., *Downregulation of beta-catenin by human Axin and its association with the APC tumor suppressor, beta-catenin and GSK3 beta*. Curr Biol, 1998. **8**(10): p. 573-81.
219. Furuhashi, M., et al., *Axin facilitates Smad3 activation in the transforming growth factor beta signaling pathway*. Mol Cell Biol, 2001. **21**(15): p. 5132-41.

220. Zhang, Y., et al., *Axin forms a complex with MEKK1 and activates c-Jun NH(2)-terminal kinase/stress-activated protein kinase through domains distinct from Wnt signaling*. J Biol Chem, 1999. **274**(49): p. 35247-54.
221. Xue, B., A.K. Dunker, and V.N. Uversky, *The roles of intrinsic disorder in orchestrating the wnt-pathway*. J Biomol Struct Dyn, 2012. **29**(5): p. 843-61.
222. Balasubramanian, R., et al., *Studies on the conformation of amino acids. VI. Conformation of the proline ring as observed in crystal structures of amino acids and peptides*. Int J Protein Res, 1971. **3**(1): p. 25-33.
223. Morris, A.L., et al., *Stereochemical quality of protein structure coordinates*. Proteins, 1992. **12**(4): p. 345-64.
224. MacArthur, M.W. and J.M. Thornton, *Influence of proline residues on protein conformation*. J Mol Biol, 1991. **218**(2): p. 397-412.
225. Nicholson, H., et al., *Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme*. Biopolymers, 1992. **32**(11): p. 1431-41.
226. Reimer, U., et al., *Side-chain effects on peptidyl-prolyl cis/trans isomerisation*. J Mol Biol, 1998. **279**(2): p. 449-60.
227. Hurley, J.H., D.A. Mason, and B.W. Matthews, *Flexible-geometry conformational energy maps for the amino acid residue preceding a proline*. Biopolymers, 1992. **32**(11): p. 1443-6.
228. Wood, S.J., et al., *Prolines and amyloidogenicity in fragments of the Alzheimer's peptide beta/A4*. Biochemistry, 1995. **34**(3): p. 724-30.
229. Williamson, M.P., *The structure and function of proline-rich regions in proteins*. Biochem J, 1994. **297** (Pt 2): p. 249-60.
230. Dafforn, T.R. and C.J. Smith, *Natively unfolded domains in endocytosis: hooks, lines and linkers*. EMBO Rep, 2004. **5**(11): p. 1046-52.
231. Holt, M.R. and A. Koffer, *Cell motility: proline-rich proteins promote protrusions*. Trends Cell Biol, 2001. **11**(1): p. 38-46.
232. Kay, B.K., M.P. Williamson, and M. Sudol, *The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains*. FASEB J, 2000. **14**(2): p. 231-41.

233. Cheadle, C., et al., *Identification of a Src SH3 domain binding motif by screening a random phage display library*. J Biol Chem, 1994. **269**(39): p. 24034-9.
234. Rickles, R.J., et al., *Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries*. EMBO J, 1994. **13**(23): p. 5598-604.
235. Sparks, A.B., et al., *Identification and characterization of Src SH3 ligands from phage-displayed random peptide libraries*. J Biol Chem, 1994. **269**(39): p. 23853-6.
236. Feng, S., et al., *Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions*. Science, 1994. **266**(5188): p. 1241-7.
237. Aasland, R., et al., *Normalization of nomenclature for peptide motifs as ligands of modular protein domains*. FEBS Lett, 2002. **513**(1): p. 141-4.
238. Sparks, A.B., et al., *Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2*. Proc Natl Acad Sci U S A, 1996. **93**(4): p. 1540-4.
239. Knudsen, B.S., et al., *Affinity and specificity requirements for the first Src homology 3 domain of the Crk proteins*. EMBO J, 1995. **14**(10): p. 2191-8.
240. Grabs, D., et al., *The SH3 domain of amphiphysin binds the proline-rich domain of dynamin at a single site that defines a new SH3 binding consensus sequence*. J Biol Chem, 1997. **272**(20): p. 13419-25.
241. Quilliam, L.A., et al., *Isolation of a NCK-associated kinase, PRK2, an SH3-binding protein and potential effector of Rho protein signaling*. J Biol Chem, 1996. **271**(46): p. 28772-6.
242. Kurakin, A., N.G. Hoffman, and B.K. Kay, *Molecular recognition properties of the C-terminal Sh3 domain of the Cbl associated protein, Cap*. J Pept Res, 1998. **52**(5): p. 331-7.
243. Kang, H., et al., *SH3 domain recognition of a proline-independent tyrosine-based RKxxYxxY motif in immune cell adaptor SKAP55*. Embo Journal, 2000. **19**(12): p. 2889-2899.
244. Chen, H.I. and M. Sudol, *The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules*. Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7819-23.

245. Schild, L., et al., *Identification of a PY motif in the epithelial Na channel subunits as a target sequence for mutations causing channel activation found in Liddle syndrome*. EMBO J, 1996. **15**(10): p. 2381-7.
246. Rentschler, S., et al., *The WW domain of dystrophin requires EF-hands region to interact with beta-dystroglycan*. Biol Chem, 1999. **380**(4): p. 431-42.
247. Chen, H.I., et al., *Characterization of the WW domain of human yes-associated protein and its polyproline-containing ligands*. J Biol Chem, 1997. **272**(27): p. 17070-7.
248. Linn, H., et al., *Using molecular repertoires to identify high-affinity peptide ligands of the WW domain of human and mouse YAP*. Biol Chem, 1997. **378**(6): p. 531-7.
249. Ilsley, J.L., M. Sudol, and S.J. Winder, *The interaction of dystrophin with beta-dystroglycan is regulated by tyrosine phosphorylation*. Cell Signal, 2001. **13**(9): p. 625-32.
250. Doong, H., et al., *CAIR-1/BAG-3 abrogates heat shock protein-70 chaperone complex-mediated protein degradation: accumulation of poly-ubiquitinated Hsp90 client proteins*. J Biol Chem, 2003. **278**(31): p. 28490-500.
251. Beere, H.M., *Death versus survival: functional interaction between the apoptotic and stress-inducible heat shock protein pathways*. J Clin Invest, 2005. **115**(10): p. 2633-9.
252. Bedford, M.T., D.C. Chan, and P. Leder, *FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands*. EMBO J, 1997. **16**(9): p. 2376-83.
253. Ermekova, K.S., et al., *The WW domain of neural protein FE65 interacts with proline-rich motifs in Mena, the mammalian homolog of Drosophila enabled*. J Biol Chem, 1997. **272**(52): p. 32869-77.
254. Bedford, M.T., R. Reed, and P. Leder, *WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: the proline glycine and methionine-rich motif*. Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10602-7.
255. Ranganathan, R., et al., *Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent*. Cell, 1997. **89**(6): p. 875-86.
256. Lu, P.J., et al., *Function of WW domains as phosphoserine- or phosphothreonine-binding modules*. Science, 1999. **283**(5406): p. 1325-8.

257. Schutkowski, M., et al., *Role of phosphorylation in determining the backbone dynamics of the serine/threonine-proline motif and Pin1 substrate recognition*. *Biochemistry*, 1998. **37**(16): p. 5566-75.
258. Otte, L., et al., *WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains*. *Protein Sci*, 2003. **12**(3): p. 491-500.
259. Gertler, F.B., et al., *Mena, a relative of VASP and Drosophila Enabled, is implicated in the control of microfilament dynamics*. *Cell*, 1996. **87**(2): p. 227-39.
260. Niebuhr, K., et al., *A novel proline-rich motif present in ActA of Listeria monocytogenes and cytoskeletal proteins is the ligand for the EVH1 domain, a protein module present in the Ena/VASP family*. *EMBO J*, 1997. **16**(17): p. 5433-44.
261. Naisbitt, S., et al., *Shank, a novel family of postsynaptic density proteins that binds to the NMDA receptor/PSD-95/GKAP complex and cortactin*. *Neuron*, 1999. **23**(3): p. 569-82.
262. Tu, J.C., et al., *Homer binds a novel proline-rich motif and links group 1 metabotropic glutamate receptors with IP3 receptors*. *Neuron*, 1998. **21**(4): p. 717-26.
263. Volkman, B.F., et al., *Structure of the N-WASP EVH1 domain-WIP complex: insight into the molecular basis of Wiskott-Aldrich Syndrome*. *Cell*, 2002. **111**(4): p. 565-76.
264. Harmer, N.J., et al., *1.15 A crystal structure of the X. tropicalis Spred1 EVH1 domain suggests a fourth distinct peptide-binding mechanism within the EVH1 family*. *FEBS Lett*, 2005. **579**(5): p. 1161-6.
265. Holtzman, J.H., et al., *Miniature protein ligands for EVH1 domains: interplay between affinity, specificity, and cell motility*. *Biochemistry*, 2007. **46**(47): p. 13541-53.
266. Nishizawa, K., et al., *Identification of a proline-binding motif regulating CD2-triggered T lymphocyte activation*. *Proc Natl Acad Sci U S A*, 1998. **95**(25): p. 14897-902.
267. Dustin, M.L., et al., *A novel adaptor protein orchestrates receptor patterning and cytoskeletal polarity in T-cell contacts*. *Cell*, 1998. **94**(5): p. 667-77.
268. Pornillos, O., et al., *Structure and functional interactions of the Tsg101 UEV domain*. *EMBO J*, 2002. **21**(10): p. 2397-406.

269. Pornillos, O., et al., *Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein*. Nat Struct Biol, 2002. **9**(11): p. 812-7.
270. Schutt, C.E., et al., *The structure of crystalline profilin-beta-actin*. Nature, 1993. **365**(6449): p. 810-6.
271. Tanaka, M. and H. Shibata, *Poly(L-proline)-binding proteins from chick embryos are a profilin and a profilactin*. Eur J Biochem, 1985. **151**(2): p. 291-7.
272. Songyang, Z., et al., *SH2 domains recognize specific phosphopeptide sequences*. Cell, 1993. **72**(5): p. 767-78.
273. Songyang, Z., et al., *The phosphotyrosine interaction domain of SHC recognizes tyrosine-phosphorylated NPXY motif*. J Biol Chem, 1995. **270**(25): p. 14863-6.
274. Yaffe, M.B., et al., *The structural basis for 14-3-3:phosphopeptide binding specificity*. Cell, 1997. **91**(7): p. 961-71.
275. Durocher, D., et al., *The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms*. Mol Cell, 2000. **6**(5): p. 1169-82.
276. Winston, J.T., et al., *The SCFbeta-TRCP-ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in IkappaBalpha and beta-catenin and stimulates IkappaBalpha ubiquitination in vitro*. Genes Dev, 1999. **13**(3): p. 270-83.
277. Wu, J.W., et al., *Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling*. Mol Cell, 2001. **8**(6): p. 1277-89.
278. Elia, A.E., et al., *The molecular basis for phosphodependent substrate targeting and regulation of Plks by the Polo-box domain*. Cell, 2003. **115**(1): p. 83-95.
279. Manke, I.A., et al., *BRCT repeats as phosphopeptide-binding modules involved in protein targeting*. Science, 2003. **302**(5645): p. 636-9.
280. Shiung, Y.Y., et al., *An anti-IgE monoclonal antibody that binds to IgE on CD23 but not on high-affinity IgE.Fc receptors*. Immunobiology, 2012. **217**(7): p. 676-83.
281. Call, G.S., et al., *Zyxin phosphorylation at serine 142 modulates the zyxin head-tail interaction to alter cell-cell adhesion*. Biochem Biophys Res Commun, 2011. **404**(3): p. 780-4.

282. Shim, J.H., et al., *Epigallocatechin gallate suppresses lung cancer cell growth through Ras-GTPase-activating protein SH3 domain-binding protein 1*. *Cancer Prev Res (Phila)*, 2010. **3**(5): p. 670-9.
283. Halim, A., et al., *Human urinary glycoproteomics; attachment site specific analysis of N- and O-linked glycosylations by CID and ECD*. *Mol Cell Proteomics*, 2012. **11**(4): p. M111 013649.
284. Almeida, E.A., et al., *Matrix survival signaling: from fibronectin via focal adhesion kinase to c-Jun NH(2)-terminal kinase*. *Journal of Cell Biology*, 2000. **149**(3): p. 741-54.
285. Carra, S., S.J. Seguin, and J. Landry, *HspB8 and Bag3: a new chaperone complex targeting misfolded proteins to macroautophagy*. *Autophagy*, 2008. **4**(2): p. 237-9.
286. Chellaiah, M.A., et al., *Phosphorylation of a Wiscott-Aldrich syndrome protein-associated signal complex is critical in osteoclast bone resorption*. *J Biol Chem*, 2007. **282**(13): p. 10104-16.
287. Martinez-Quiles, N., et al., *Erk/Src phosphorylation of cortactin acts as a switch on-switch off mechanism that controls its ability to activate N-WASP*. *Mol Cell Biol*, 2004. **24**(12): p. 5269-80.
288. Krause, M., et al., *Ena/VASP proteins: regulators of the actin cytoskeleton and cell migration*. *Annu Rev Cell Dev Biol*, 2003. **19**: p. 541-64.
289. Cowan, P.M., S. McGavin, and A.C. North, *The polypeptide chain configuration of collagen*. *Nature*, 1955. **176**(4492): p. 1062-4.
290. Creighton, T., *Conformational properties of polypeptide chains*, in *Proteins. Structures and Molecular Properties* 1984, W.H. Freeman and Co.: New York. p. 159-197.
291. Adzhubei, A.A. and M.J. Sternberg, *Left-handed polyproline II helices commonly occur in globular proteins*. *J Mol Biol*, 1993. **229**(2): p. 472-93.
292. Shi, Z., et al., *Polyproline II propensities from GGXGG peptides reveal an anticorrelation with beta-sheet scales*. *Proc Natl Acad Sci U S A*, 2005. **102**(50): p. 17964-8.
293. Whittington, S.J., et al., *Urea promotes polyproline II helix formation: implications for protein denatured states*. *Biochemistry*, 2005. **44**(16): p. 6269-75.
294. Makowska, J., et al., *Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded*

- peptides and proteins*. Proc Natl Acad Sci U S A, 2006. **103**(6): p. 1744-9.
295. Zagrovic, B., et al., *Unusual compactness of a polyproline type II structure*. Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11698-703.
296. Shi, Z., et al., *Conformation of the backbone in unfolded proteins*. Chem Rev, 2006. **106**(5): p. 1877-97.
297. Ren, X. and J.H. Hurley, *Proline-rich regions and motifs in trafficking: from ESCRT interaction to viral exploitation*. Traffic, 2011. **12**(10): p. 1282-90.
298. Odorizzi, G., *The multiple personalities of Alix*. J Cell Sci, 2006. **119**(Pt 15): p. 3025-32.
299. Zhou, X., et al., *Decoding the intrinsic mechanism that prohibits ALIX interaction with ESCRT and viral proteins*. Biochem J, 2010. **432**(3): p. 525-34.
300. Zhou, X., et al., *The HIV-1 p6/EIAV p9 docking site in Alix is autoinhibited as revealed by a conformation-sensitive anti-Alix monoclonal antibody*. Biochem J, 2008. **414**(2): p. 215-20.
301. Pan, S., et al., *Involvement of the conserved adaptor protein Alix in actin cytoskeleton assembly*. J Biol Chem, 2006. **281**(45): p. 34640-50.
302. Carlton, J.G., M. Agromayor, and J. Martin-Serrano, *Differential requirements for Alix and ESCRT-III in cytokinesis and HIV-1 release*. Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10541-6.
303. Fowler, D.M., et al., *Functional amyloid--from bacteria to humans*. Trends Biochem Sci, 2007. **32**(5): p. 217-24.
304. Monsellier, E. and F. Chiti, *Prevention of amyloid-like aggregation as a driving force of protein evolution*. EMBO Rep, 2007. **8**(8): p. 737-42.
305. Rauscher, S., et al., *Proline and glycine control protein self-organization into elastomeric or amyloid fibrils*. Structure, 2006. **14**(11): p. 1667-76.
306. Harper, J.W. and S.J. Elledge, *The DNA damage response: ten years after*. Mol Cell, 2007. **28**(5): p. 739-45.
307. Zhou, B.B. and S.J. Elledge, *The DNA damage response: putting checkpoints in perspective*. Nature, 2000. **408**(6811): p. 433-9.
308. Shiloh, Y., *ATM and related protein kinases: safeguarding genome integrity*. Nat Rev Cancer, 2003. **3**(3): p. 155-68.

309. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
310. Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation*. Nucleic Acids Res, 2004. **32**(3): p. 1037-49.
311. Niehrs, C. and J. Shen, *Regulation of Lrp6 phosphorylation*. Cell Mol Life Sci, 2010. **67**(15): p. 2551-62.
312. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins*. Nucleic Acids Res, 2003. **31**(13): p. 3625-30.
313. Hulo, N., et al., *The PROSITE database*. Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.
314. Hulo, N., et al., *The 20 years of PROSITE*. Nucleic Acids Res, 2008. **36**(Database issue): p. D245-9.
315. Blom, N., S. Gammeltoft, and S. Brunak, *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites*. J Mol Biol, 1999. **294**(5): p. 1351-62.
316. Yaffe, M.B., et al., *A motif-based profile scanning approach for genome-wide prediction of signaling pathways*. Nat Biotechnol, 2001. **19**(4): p. 348-53.
317. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs*. Nucleic Acids Res, 2003. **31**(13): p. 3635-41.
318. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. **298**(5600): p. 1912-34.
319. Brown, S.D., et al., *Isolation and characterization of LRP6, a novel member of the low density lipoprotein receptor gene family*. Biochem Biophys Res Commun, 1998. **248**(3): p. 879-88.
320. Yum, S., et al., *The role of the Ser/Thr cluster in the phosphorylation of PPPSP motifs in Wnt coreceptors*. Biochem Biophys Res Commun, 2009. **381**(3): p. 345-9.
321. Bilic, J., et al., *Wnt induces LRP6 signalosomes and promotes dishevelled-dependent LRP6 phosphorylation*. Science, 2007. **316**(5831): p. 1619-22.
322. Davidson, G., et al., *Casein kinase 1 gamma couples Wnt receptor activation to cytoplasmic signal transduction*. Nature, 2005. **438**(7069): p. 867-72.

323. Dumontier, M., et al., *Armadillo: domain boundary prediction by amino acid composition*. J Mol Biol, 2005. **350**(5): p. 1061-73.
324. Deber, C.M., et al., *Nuclear magnetic resonance evidence for cis-peptide bonds in proline oligomers*. J Am Chem Soc, 1970. **92**(21): p. 6191-8.
325. Brown, A.M. and N.J. Zondlo, *A Propensity Scale for Type II Polyproline Helices (PPII): Aromatic Amino Acids in Proline-Rich Sequences Strongly Disfavor PPII Due to Proline-Aromatic Interactions*. Biochemistry, 2012. **51**(25): p. 5041-51.
326. Logan, C.Y. and R. Nusse, *The Wnt signaling pathway in development and disease*. Annu Rev Cell Dev Biol, 2004. **20**: p. 781-810.
327. Clevers, H., *Wnt/beta-catenin signaling in development and disease*. Cell, 2006. **127**(3): p. 469-80.
328. Klaus, A. and W. Birchmeier, *Wnt signalling and its impact on development and cancer*. Nat Rev Cancer, 2008. **8**(5): p. 387-98.
329. Wu, D. and W. Pan, *GSK3: a multifaceted kinase in Wnt signaling*. Trends Biochem Sci, 2010. **35**(3): p. 161-8.
330. Angers, S. and R.T. Moon, *Proximal events in Wnt signal transduction*. Nat Rev Mol Cell Biol, 2009. **10**(7): p. 468-77.
331. Verheyen, E.M. and C.J. Gottardi, *Regulation of Wnt/beta-catenin signaling by protein kinases*. Dev Dyn, 2010. **239**(1): p. 34-44.
332. Zeng, X., et al., *Initiation of Wnt signaling: control of Wnt coreceptor Lrp6 phosphorylation/activation via frizzled, dishevelled and axin functions*. Development, 2008. **135**(2): p. 367-75.
333. Kimelman, D. and W. Xu, *beta-catenin destruction complex: insights and questions from a structural perspective*. Oncogene, 2006. **25**(57): p. 7482-91.
334. Ha, N.C., et al., *Mechanism of phosphorylation-dependent binding of APC to beta-catenin and its role in beta-catenin degradation*. Mol Cell, 2004. **15**(4): p. 511-21.
335. Liu, C., et al., *Control of beta-catenin phosphorylation/degradation by a dual-kinase mechanism*. Cell, 2002. **108**(6): p. 837-47.
336. van Noort, M., et al., *Wnt signaling controls the phosphorylation status of beta-catenin*. J Biol Chem, 2002. **277**(20): p. 17901-5.
337. Molenaar, M., et al., *XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos*. Cell, 1996. **86**(3): p. 391-9.

338. He, X., et al., *LDL receptor-related proteins 5 and 6 in Wnt/beta-catenin signaling: arrows point the way*. Development, 2004. **131**(8): p. 1663-77.
339. Springer, T.A., *An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components*. J Mol Biol, 1998. **283**(4): p. 837-62.
340. Jeon, H., et al., *Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair*. Nat Struct Biol, 2001. **8**(6): p. 499-504.
341. Tamai, K., et al., *LDL-receptor-related proteins in Wnt signal transduction*. Nature, 2000. **407**(6803): p. 530-5.
342. Semenov, M.V., et al., *Head inducer Dickkopf-1 is a ligand for Wnt coreceptor LRP6*. Curr Biol, 2001. **11**(12): p. 951-61.
343. Brennan, K., et al., *Truncated mutants of the putative Wnt receptor LRP6/Arrow can stabilize beta-catenin independently of Frizzled proteins*. Oncogene, 2004. **23**(28): p. 4873-84.
344. Zeng, X., et al., *A dual-kinase mechanism for Wnt co-receptor phosphorylation and activation*. Nature, 2005. **438**(7069): p. 873-7.
345. Mao, B., et al., *LDL-receptor-related protein 6 is a receptor for Dickkopf proteins*. Nature, 2001. **411**(6835): p. 321-5.
346. Mao, J., et al., *Low-density lipoprotein receptor-related protein-5 binds to Axin and regulates the canonical Wnt signaling pathway*. Mol Cell, 2001. **7**(4): p. 801-9.
347. Liu, G., et al., *A novel mechanism for Wnt activation of canonical signaling through the LRP6 receptor*. Mol Cell Biol, 2003. **23**(16): p. 5825-35.
348. Tamai, K., et al., *A mechanism for Wnt coreceptor activation*. Mol Cell, 2004. **13**(1): p. 149-56.
349. Piao, S., et al., *Direct inhibition of GSK3beta by the phosphorylated cytoplasmic domain of LRP6 in Wnt/beta-catenin signaling*. PloS one, 2008. **3**(12): p. e4046.
350. Kim, A.S., et al., *Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein*. Nature, 2000. **404**(6774): p. 151-8.
351. Price, M.A., *CKI, there's more than one: casein kinase I family members in Wnt and Hedgehog signaling*. Genes Dev, 2006. **20**(4): p. 399-410.

352. Yasui, N., et al., *Detection of endogenous LRP6 expressed on human cells by monoclonal antibodies specific for the native conformation*. J Immunol Methods, 2010. **352**(1-2): p. 153-60.
353. Liang, J., et al., *Transmembrane protein 198 promotes LRP6 phosphorylation and Wnt signaling activation*. Mol Cell Biol, 2011. **31**(13): p. 2577-90.
354. Bhalla, J., et al., *Local flexibility in molecular function paradigm*. Mol Cell Proteomics, 2006. **5**(7): p. 1212-23.
355. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. Journal of molecular biology, 1999. **293**(2): p. 321-31.
356. Eliezer, D., *Biophysical characterization of intrinsically disordered proteins*. Curr Opin Struct Biol, 2009. **19**(1): p. 23-30.
357. Eliezer, D., *Characterizing residual structure in disordered protein States using nuclear magnetic resonance*. Methods Mol Biol, 2007. **350**: p. 49-67.
358. Wang, X., et al., *Characterizing the conformational ensemble of monomeric polyglutamine*. Proteins, 2006. **63**(2): p. 297-311.
359. Provencher, S.W. and J. Glockner, *Estimation of globular protein secondary structure from circular dichroism*. Biochemistry, 1981. **20**(1): p. 33-7.
360. Johnson, W.C., Jr., *Secondary structure of proteins through circular dichroism spectroscopy*. Annu Rev Biophys Biophys Chem, 1988. **17**: p. 145-66.
361. Woody, R.W., *Circular dichroism*. Methods Enzymol, 1995. **246**: p. 34-71.
362. Kelly, S.M. and N.C. Price, *The application of circular dichroism to studies of protein folding and unfolding*. Biochim Biophys Acta, 1997. **1338**(2): p. 161-85.
363. Vassilenko, K.S. and V.N. Uversky, *Native-like secondary structure of molten globules*. Biochim Biophys Acta, 2002. **1594**(1): p. 168-77.
364. Chen, E., et al., *The kinetics of helix unfolding of an azobenzene cross-linked peptide probed by nanosecond time-resolved optical rotatory dispersion*. J Am Chem Soc, 2003. **125**(41): p. 12443-9.
365. Semisotnov, G.V., et al., *Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe*. Biopolymers, 1991. **31**(1): p. 119-28.

366. Bourhis, J.M., B. Canard, and S. Longhi, *Predicting protein disorder and induced folding: from theoretical principles to practical applications*. *Curr Protein Pept Sci*, 2007. **8**(2): p. 135-49.
367. Dosztanyi, Z., et al., *Prediction of protein disorder at the domain level*. *Curr Protein Pept Sci*, 2007. **8**(2): p. 161-71.
368. Kryshtafovych, A., K. Fidelis, and J. Moult, *CASP9 results compared to those of previous CASP experiments*. *Proteins*, 2011. **79 Suppl 10**: p. 196-207.
369. Radivojac, P., et al., *Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition*. *Proteins*, 2006. **63**(2): p. 398-410.
370. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically disordered proteins in human diseases: introducing the D2 concept*. *Annu Rev Biophys*, 2008. **37**: p. 215-46.
371. Sreerama, N. and R.W. Woody, *Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(pro)II conformations*. *Proteins*, 1999. **36**(4): p. 400-6.
372. Gattiker, A., E. Gasteiger, and A. Bairoch, *ScanProsite: a reference implementation of a PROSITE scanning tool*. *Appl Bioinformatics*, 2002. **1**(2): p. 107-8.
373. Moore, C.L., et al., *Secondary nucleating sequences affect kinetics and thermodynamics of tau aggregation*. *Biochemistry*, 2011. **50**(50): p. 10876-86.
374. Johnson, S.A. and T. Hunter, *Kinomics: methods for deciphering the kinome*. *Nat Methods*, 2005. **2**(1): p. 17-25.
375. Amanchy, R., et al., *A curated compendium of phosphorylation motifs*. *Nature biotechnology*, 2007. **25**(3): p. 285-6.
376. Dinkel, H., et al., *Phospho.ELM: a database of phosphorylation sites--update 2011*. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D261-7.
377. Diella, F., et al., *Phospho.ELM: a database of phosphorylation sites--update 2008*. *Nucleic acids research*, 2008. **36**(Database issue): p. D240-4.
378. Farriol-Mathis, N., et al., *Annotation of post-translational modifications in the Swiss-Prot knowledge base*. *Proteomics*, 2004. **4**(6): p. 1537-50.
379. Miller, M.L., et al., *Linear motif atlas for phosphorylation-dependent signaling*. *Science signaling*, 2008. **1**(35): p. ra2.

380. Blom, N., et al., *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence*. Proteomics, 2004. **4**(6): p. 1633-49.
381. Kim, J.H., et al., *Prediction of phosphorylation sites using SVMs*. Bioinformatics, 2004. **20**(17): p. 3179-84.
382. Xue, Y., et al., *PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory*. BMC bioinformatics, 2006. **7**: p. 163.
383. Neuberger, G., G. Schneider, and F. Eisenhaber, *pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model*. Biol Direct, 2007. **2**: p. 1.
384. Wong, Y.H., et al., *KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns*. Nucleic Acids Res, 2007. **35**(Web Server issue): p. W588-94.
385. Huang, H.-D., et al., *KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites*. Nucleic acids research, 2005. **33**(Web Server issue): p. W226-9.
386. Ingrell, C.R., et al., *NetPhosYeast: prediction of protein phosphorylation sites in yeast*. Bioinformatics, 2007. **23**(7): p. 895-7.
387. Xue, Y., et al., *GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy*. Mol Cell Proteomics, 2008. **7**(9): p. 1598-608.
388. Saunders, N.F., et al., *Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites*. BMC bioinformatics, 2008. **9**: p. 245.
389. Landry, C.R., E.D. Levy, and S.W. Michnick, *Weak functional constraints on phosphoproteomes*. Trends in genetics : TIG, 2009. **25**(5): p. 193-7.
390. Linding, R., et al., *NetworkKIN: a resource for exploring cellular phosphorylation networks*. Nucleic acids research, 2008. **36**(Database issue): p. D695-9.
391. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse*. Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.

392. Matsuoka, S., et al., *ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage*. Science (New York, N.Y.), 2007. **316**(5828): p. 1160-6.
393. Blethrow, J.D., et al., *Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(5): p. 1442-7.
394. Meggio, F. and L.A. Pinna, *One-thousand-and-one substrates of protein kinase CK2?* FASEB J, 2003. **17**(3): p. 349-68.
395. Salvi, M., et al., *Extraordinary pleiotropy of protein kinase CK2 revealed by weblogo phosphoproteome analysis*. Biochim Biophys Acta, 2009. **1793**(5): p. 847-59.
396. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing*. Statistics in medicine, 1990. **9**(7): p. 811-8.
397. Brown, N.R., et al., *The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases*. Nature Cell Biology, 1999. **1**(7): p. 438-443.
398. Liu, B.A., et al., *SH2 domains recognize contextual peptide sequence information to determine selectivity*. Mol Cell Proteomics, 2010. **9**(11): p. 2391-404.
399. Gfeller, D., et al., *The multiple-specificity landscape of modular peptide recognition domains*. Mol Syst Biol, 2011. **7**: p. 484.
400. Inatsuka, C.S., et al., *Pertactin is required for Bordetella species to resist neutrophil-mediated clearance*. Infect Immun, 2010. **78**(7): p. 2901-9.
401. Crooks, G.E., et al., *WebLogo: a sequence logo generator*. Genome research, 2004. **14**(6): p. 1188-90.
402. Killian, B.J., J. Yundenfreund Kravitz, and M.K. Gilson, *Extraction of configurational entropy from molecular simulations via an expansion approximation*. The Journal of chemical physics, 2007. **127**(2): p. 024107.
403. Kersey, P.J., et al., *The International Protein Index: an integrated database for proteomics experiments*. Proteomics, 2004. **4**(7): p. 1985-8.
404. Boze, H., et al., *Proline-rich salivary proteins have extended conformations*. Biophys J, 2010. **99**(2): p. 656-65.

405. De Biasio, A., et al., *Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand Delta-4*. *J Proteome Res*, 2008. **7**(6): p. 2496-506.