

NMR Structure Improvement: A Structural Bioinformatics & Visualization Approach.

by

Jeremy N. Block

Department of Biochemistry
Duke University

Date: _____

Approved:

David C. Richardson, Co-Supervisor

Jane S. Richardson, Co-Supervisor

John D. York

Pei Zhou

Leonard Spicer

Brian Kuhlman

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2010

ABSTRACT

NMR Structure Improvement: A Structural Bioinformatics & Visualization Approach.

by

Jeremy N. Block

Department of Biochemistry
Duke University

Date: _____

Approved:

David C. Richardson, Co-Supervisor

Jane S. Richardson, Co-Supervisor

John D. York

Pei Zhou

Leonard Spicer

Brian Kuhlman

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor in the Department of
Biochemistry in the Graduate School
of Duke University

2010

Copyright by
Jeremy N. Block
2010

Abstract

The overall goal of this project is to enhance the physical accuracy of individual models in macromolecular NMR (Nuclear Magnetic Resonance) structures and the realism of variation within NMR ensembles of models, while improving agreement with the experimental data. A secondary overall goal is to combine synergistically the best aspects of NMR and crystallographic methodologies to better illuminate the underlying joint molecular reality. This is accomplished by using the powerful method of all-atom contact analysis (describing detailed sterics between atoms, including hydrogens); new graphical representations and interactive tools in 3D and virtual reality; and structural bioinformatics approaches to the expanded and enhanced data now available.

The resulting better descriptions of macromolecular structure and its dynamic variation enhances the effectiveness of the many biomedical applications that depend on detailed molecular structure, such as mutational analysis, homology modeling, molecular simulations, protein design, and drug design.

Dedication

For Ian & Cindy Block, my parents.

For Rebecca Block, my sister.

For Henry & Betty Block, my paternal grandparents.

For Aaron & Rose Boskin, my maternal grandparents.

For all who confront significant challenges, add this work to the list of evidence demonstrating that hurtful stereotypes and misconceptions can be overcome.

Contents

Abstract	iv
List of Tables	xii
List of Figures	xiii
Acknowledgements	xix
1. Introduction	1
1.1 Field Development: Biochemistry, Structural Biology, and Bioinformatics	1
1.2 Relevance to the basic biomedical sciences' contribution to human health	4
1.3 Visualizing, Analyzing, & Validating Macromolecular Structures	6
1.4 Macromolecular Structure Quality	7
1.5 Tools for Structure Analysis, Visualization, & Validation	14
1.5.1 Mage & KiNG	15
1.5.2 PROBE All-Atom Contact Analysis	16
1.5.3 REDUCE	17
1.5.4 BioGeometry Local Density	19
1.5.5 MolProbity	20
1.5.6 WHATIF / WHAT_CHECK / PROCHECK / PROCHECK_NMR	21
1.5.7 QUEEN	24
1.5.8 Secondary Structure Assignment	24
1.6 Structure Calculation & Refinement Tools	25
1.6.1 AutoStructure / AutoQF	26
1.6.2 CYANA / DYANA	26
1.6.3 CNS / XPLOR / XPLOR-NIH	27
1.6.4 ARIA	28

1.7 Conclusions	28
2. The Inositol Signaling Pathway: Enzyme Structures and Specificity	29
2.1 MolProbity Diagnosis, Model Building, and Structure Refinement of Phosphatidylinositol 1-phosphatase (1Ptase)	31
2.2 Structure-Determination of a plant AtIpmk at 2.8Å, using tools in KiNG and MolProbity.....	44
2.3 Defining the Inositol Substrate-Specificity Motif.....	46
2.4 Conclusions	51
3. New Tools for Analysis & Error Diagnosis in NMR Structures.....	52
3.1 NMR Diagnostic Visualizations and related Error Analysis.....	52
3.1.1 Multi-Model Multi-Criterion Kinemages	52
3.1.2 Multi-Model Multi-Criterion Chart	55
3.1.3 NOEDisplay	57
3.1.4 All-Atom Contact Analysis of Under-Packing in Protein Models	58
3.2 NMR Database Analyses, and Specific Examples	66
3.2.1 NMR Structure Database Creation	66
3.2.2 Use of NOE Restraints per ordered residue as a data quality measure.....	71
3.2.3 Ramachandran Analysis for NMR Ensembles.....	72
3.2.4 Sidechain Rotamer Analysis for NMR Ensembles	74
3.2.5 Clashscore combined with H-bond Scores as an NMR structure quality factor	77
3.2.6 NMR Structure Chimera: A Test Case for a Best Parts Approach.....	80
3.2.7 Z-Domain Structures and Other Examples.....	82
3.3 NOE patterns used to identify RNA backbone conformations	89
3.3.1 Structural Overview of RNA	91
3.3.2 NOE's by CRMA.....	97

3.3.3	Semi-Quantitative Distances	98
3.3.4	Lookup Tables & Parallel Coordinate Plots	99
3.3.5	Test Structures	103
3.3.6	Results.....	105
3.3.7	Discussion.....	109
3.3.8	Conclusions.....	117
3.4	Concluding remarks on the new tools and analyses	117
4.	Visualizing & Analyzing NMR Structure Ensembles in Virtual Reality.....	119
4.1	KinImmerse: Macromolecular VR for NMR Ensembles	120
4.1.1	Background.....	121
4.1.1.1	The Tools: Molecular graphics and virtual reality.....	122
4.1.1.2	The Application Area: Macromolecular structures and NMR ensembles	127
4.1.2	Methods.....	128
4.1.2.1	VR hardware	128
4.1.2.2	The Syzygy toolkit	131
4.1.2.3	Callback-style API.....	132
4.1.2.4	Data structure design.....	133
4.1.2.5	Hierarchical bounding boxes.....	134
4.1.2.6	File format versus interactive display features	135
4.1.2.7	Kinemage construction for the test applications.....	135
4.1.2.8	Free Open Source Software.....	136
4.1.3	Results.....	137
4.1.3.1	User Interface	138
4.1.3.2	The Local Perspective on Models and Data in an NMR Ensemble	142

4.1.3.3 Visualization of RDC data and relationships.....	144
4.1.4 Discussion.....	147
4.1.4.1. Demo mode	147
4.1.4.2 Differences from single-screen molecular graphics or from most VR.....	148
4.1.4.3 Future Directions	149
4.2 Demonstrations and Use of KinImmerse in the DiVE.....	151
4.2.1 Orientation In Phases	151
4.2.1.1 Foyer.....	151
4.2.1.2 Facility	152
4.2.1.3 System.....	153
4.2.2 The Demonstrator	155
4.2.2.1 Demonstrator As Authority Figure	156
4.2.3 Orienting Attention.....	159
4.2.3.1 Voluntary vs. Involuntary Attention.....	160
4.2.3.2 Location Cueing.....	161
4.2.3.3 Crossmodal Cueing and Attention Shifting.....	163
4.2.3.4 Location Cueing and Motion Sickness.....	164
4.2.3.5 Expertise Effects on Location Cueing.....	165
4.2.3.6 Location Cueing & Co-Centering	166
4.2.4 Domain Experts' Experience	167
4.2.4.1 Biochemists and Structural Biologists	168
4.2.4.2 Engineers	169
4.2.4.3 Computer Scientists & Virtual Reality Experts	169
4.2.4.4 Visual Artists	170
4.3 Conclusions	171

5.	RDCvis – Visual Representation of RDC curves on an NMR Structural Model.....	173
5.1	Residual Dipolar Coupling use in Structural Biology.....	173
5.2	Visual Representation of an RDC	174
5.3	Displaying & Co-Centering RDCs on a structure model in KiNG.....	177
5.3.1	A walkthrough of RDCvis in KiNG.....	179
5.4	Results	185
5.4.1	The One Curve Rule	185
5.4.2	Curve Intersection.....	188
5.4.3	Orientation Dependent Variability	188
5.4.4	Planarity Problems.....	189
5.4.5	Error Model Issues	191
5.4.6	Other Examples	193
5.5	Discussion	195
6.	Application of tools to NMR structure improvement.....	197
6.1	Comparing Structure Determination Packages Incorporating RDC's	197
6.1.1	Cytochrome c maturation protein - CcmE.....	198
6.1.2	dvCcmE' Structure Determination Protocol Incorporating RDC's.....	200
6.2	Materials & Methods.....	200
6.2.1	Sample Preparation	200
6.2.2	Methods.....	202
6.2.2.1	NMR Spectroscopy	202
6.2.2.2	CNS Structure Determination and Validation.....	203
6.2.2.3	Xplor-NIH Structure Determination and Validation.....	204
6.3	Results	205
6.3.1	Curve Intersection Examples	206

6.3.2 One Curve Rule Examples.....	212
6.3.3 Hydrogen Bonding Examples.....	218
6.3.4 Error Model Examples.....	223
6.3.5 PEG vs. Gel vs. PEG & Gel Examples.....	227
6.3.6 CNS vs. Xplor Examples.....	236
6.3.7 Orientation Dependent Variability Examples.....	242
6.3.8 Other Examples.....	247
6.4 Discussion.....	254
6.4.1 Conclusions.....	254
6.4.2 Recommendations.....	255
6.4.3 Deposition.....	256
7. Conclusions & Future Directions.....	257
7.1 MolProbity NMR.....	257
7.2 KinImmerse.....	257
7.3 RDCvis.....	258
7.4 NMR Structure Improvement.....	259
Appendix.....	260
References.....	268
Biography.....	288

List of Tables

Table 1 - 1Ptase Re-Refinement Structure Information.....	41
Table 2 - MolProbity Statistics.....	43
Table 3 - The 10 NMR, RNA test structures and their number of residues demonstrating diversity in size.....	104
Table 4 - Conformer statistics: intersection conformer subsets, union conformer subsets, and each of these subsets extended to include conformers in the 10% error region for NOEs.....	107
Table 5 - Comparing Structure Determination Packages Incorporating RDCs	198

List of Figures

Figure 1-1 Resolution vs. Clashscore for the PDB.....	9
Figure 1-2 - PROBE All-Atom Contact Analysis.....	17
Figure 1-3 - ProCheck Ramachandran Data.....	23
Figure 1-4 - MolProbity Ramachandran Data.....	23
Figure 2-1 - Ribbon Schematic of 1Ptase	32
Figure 2-2 - Multicriterion Kinemage of 1INP.....	33
Figure 2-3 - First turn of a helix fit as Loop, in 1INP	35
Figure 2-4 - Clash repair in a turn: 1INP at left, rebuilt at right with register shift	37
Figure 2-5 - 1INP turn with density at left, rebuilt and re-refined structure at right.....	37
Figure 2-6 - 1Ptase register-shift error exit strand	38
Figure 2-7 - R Value vs. refinement round for re-refinement of 1INP	39
Figure 2-8 - MolProbity scores vs. refinement round for re-refinement of 1INP	40
Figure 2-9 - Validation Statistics for 1INP vs. Rebuilt 1INP	42
Figure 2-10 - 1Ptase Clashscores vs. PDB Sample	44
Figure 2-11 - Substrate and non-substrate inositol polyphosphate species	48
Figure 2-12 - Inositol polyphosphate species: proposed binding motif.....	49
Figure 3-1 - A Multi-Model Multi-Criterion Kinemage for an NMR Ensemble.....	53
Figure 3-2 - Multi-Model Multi-Criterion Chart for NMR Structures.....	56
Figure 3-3 - NOEDisplay Example on LpxC Structure.....	57
Figure 3-4 - Dihydrofolate Reductase Structures by NMR & Xray.....	62
Figure 3-5 - Hbonding dot counts vs. Percentage Loop	64
Figure 3-6 - DHFR Structure 1AO8 with Hbond dots & BioGeometry Analysis.....	65
Figure 3-7 - Ramachandran Outliers vs. Restraints Per Residue in NMR Structures	73

Figure 3-8 - Rotamer Outliers vs. Restraints Per Residue in NMR Structures.....	75
Figure 3-9 - Clashscore vs. Restraints Per Residue in NMR Structures.....	77
Figure 3-10 - Mainchain Hbonding vs. Restraints Per Residue in NMR Structures	78
Figure 3-11 - Favorable Dot Score vs. Restraints Per Residue in NMR Structures	79
Figure 3-12 - Chimera Plot of 1EGF Structure Ensemble.....	81
Figure 3-13 - Z-domain: Two Validation Criteria Compared Across Structure Ensembles	83
Figure 3-14 - 1J1H Structure with a Clashscore of 242.9.....	86
Figure 3-15 - Error Cluster in Proline Turn of Z-domain structure 1Q2N.....	87
Figure 3-16 - Surface Sidechain Conformations of Lys50 in 2SPZ.....	88
Figure 3-17 – RNA backbone suite.	92
Figure 3-18 - Heminucleotide designations and associated conformations	93
Figure 3-19 - RNA Backbone Nomenclature for GNRA tetraloop	95
Figure 3-20 - RNA Backbone Nomenclature for S-motif	95
Figure 3-21 - RNA backbone suite with potential H2'(n-1) NOEs shown	96
Figure 3-22 - RNA backbone NOE distance Lookup Table for H2'(n-1).....	100
Figure 3-23 - Parallel Coordinate Plot of RNA Backbone Rotamer NOE distances H1'(n-1)	102
Figure 3-24 - Combining the conformation subsets of multiple NOEs to get the intersection and union conformer subsets.....	105
Figure 3-25 - 1F9L residue 5: differences in NOE restraint distances (in text) and final model (as measured), distances, showing the problem with scaling.....	112
Figure 3-26 - 1YMO, has clashes in almost every residue. The clashscore for this model was 129.59 (Model 1). Only 38% of model suites match those predicted by the NOE constraints.....	113
Figure 3-27 - RNA Backbone NOE restraints that can conformationally restrict the suite to A-form.....	116
Figure 4-1 - The DiVE	129

Figure 4-2 - User with InterSense head tracker and 3D controller	130
Figure 4-3 - Flow chart of the KinImmerse logic.....	133
Figure 4-4 - InterSense Ultrasonic 3D Controller.....	139
Figure 4-5 - User toggling a subunit off using the menu in KinImmerse (left image on, right image off).....	140
Figure 4-6 - Annotation diagnosing a steric clash in KinImmerse.....	141
Figure 4-7 - NOEs for Ile3 H β of 1D3Z, in KiNG vs KinImmerse user session.....	142
Figure 4-8 - NH RDC curves, co-centered on the N atom.....	144
Figure 4-9 - Evaluating two clusters of loop models by RDC geometry.....	145
Figure 4-10 - High school students experiencing a molecule in the DiVE	148
Figure 4-11 - Interaction Control in the DiVE. Demonstrator in front with students behind.....	158
Figure 4-12 - Transition of interaction control from the demonstrator to a student in the DiVE.....	159
Figure 4-13 - Demonstrator employing multimodal location cueing verbally and interactively	162
Figure 4-14 - Demonstrator describing motion sickness effect while facing the audience.	165
Figure 4-15 - Kinemage drawing from KinImmerse of George W. Bush by Political Cartoonist Kevin "KAL" Kallaugh.....	171
Figure 5-1 - RDC curve examples from hyperboloid surfaces (dotted) intersecting with spheres (blue)	176
Figure 5-2 - RDCvis in KiNG entering PDB and RDC files.....	179
Figure 5-3 - RDCvis in KiNG choosing RDCs to display	181
Figure 5-4 - RDCvis in KiNG with RDCs shown.....	182
Figure 5-5 - Co-Centering tool in KiNG: Lys 100 as deposited	183
Figure 5-6 - Co-centering tool in KiNG: Lys 100 co-centered on the NH N atom.....	184
Figure 5-7 - Loop 35-41 in 1Q2N with RDC curves colored by agreement.....	186

Figure 5-8 - NH Bond Vector Pulled Out of Plane.....	190
Figure 5-9 - Excessively tight clustering of NH RDCs on Leu 77 in CCME	192
Figure 5-10 - Glutamine 36 in 2JNG loop	194
Figure 5-11 – Aspartic Acid 91 in 2JXX loop	195
Figure 6-1 dvCcmc' Ribbon of Model 1 from PDB: 2KCT	199
Figure 6-2 Ala 61 CNS with PEG + Gel RDCs	206
Figure 6-3 Asp 97 CNS with PEG + Gel RDCs	207
Figure 6-4 Asp 97 CNS Gel RDCs	208
Figure 6-5 Leu 67 CNS PEG + Gel RDCs	209
Figure 6-6 Phe 56 CNS with Peg + Gel RDCs	210
Figure 6-7 Phe 99 Xplor with Peg + GEL RDCs.....	210
Figure 6-8 Thr 122 CNS with PEG + Gel RDCs.....	211
Figure 6-9 Asp 116 CNS with PEG + Gel RDCs	213
Figure 6-10 Gly 72 Xplor with PEG + Gel RDCs. a. 7 models correct b. 13 models incorrect	214
Figure 6-11 Ser 84 CNS with PEG RDCs	215
Figure 6-12 Thr 52 CNS.....	216
Figure 6-13 Thr 52 Xplor with Gel	217
Figure 6-14 Ala 120 CNS with PEG RDCs.....	218
Figure 6-15 Gly 57 CNS with PEG RDCs.....	219
Figure 6-16 Thr 59 CNS with PEG + Gel RDCs.....	220
Figure 6-17 Ala 60 CNS with PEG + Gel RDCs	221
Figure 6-18 Ala 60 CNS with Gel RDCs	222
Figure 6-19 Ala 80 Xplor with Peg + GEL RDCs	223
Figure 6-20 Asp 62 CNS with PEG RDCs	224

Figure 6-21 Leu 77 CNS with PEG + Gel RDCs	225
Figure 6-22 Phe 118 CNS with PEG + Gel RDCs	226
Figure 6-23 Glu 108 CNS with PEG + Gel RDCs	227
Figure 6-24 Glu 108 Xplor with Gel RDCs	228
Figure 6-25 Gly 110 Xplor with Gel RDCs	229
Figure 6-26 Gly 110 Xplor with PEG RDCs.....	230
Figure 6-27 Gly 110 Xplor with PEG + Gel RDCs.....	231
Figure 6-28 Gly 110 CNS with PEG + Gel RDCs	231
Figure 6-29 Leu 111 CNS with PEG + Gel RDCs.....	232
Figure 6-30 Leu 111 Xplor with PEG RDCs	233
Figure 6-31 - Leu 11 Xplor with PEG + Gel RDCs	233
Figure 6-32 Thr 65 CNS with PEG + Gel RDCs.....	234
Figure 6-33 Thr 117 CNS with PEG + Gel RDCs	235
Figure 6-34 Asp 68 CNS	236
Figure 6-35 Asp 68 Xplor.....	237
Figure 6-36 Gly 109 CNS with PEG + Gel RDCs	238
Figure 6-37 Gly 109 Xplor with Gel RDCs	239
Figure 6-38 Val 53 Xplor with PEG RDCs	240
Figure 6-39 Val 53 CNS.....	241
Figure 6-40 Ala 81 Xplor with PEG + Gel RDCs.....	242
Figure 6-41 Ala 94 CNS with PEG + Gel RDCs	243
Figure 6-42 Arg 76 CNS with Gel RDCs	244
Figure 6-43 Lys 85 CNS with Gel RDCs	245
Figure 6-44 Lys 85 CNS with PEG + Gel RDCs	245
Figure 6-45 Val 103 CNS with PEG + Gel RDCs	246

Figure 6-46 Ala 112 CNS with PEG + Gel RDCs	247
Figure 6-47 Gly 63 CNS with PEG + Gel RDCs.....	248
Figure 6-48 Gly 63 Xplor with PEG RDCs	249
Figure 6-49 Gly 102 CNS with PEG RDCs	250
Figure 6-50 Gly 102 CNS with PEG+ Gel RDCs	251
Figure 6-51 Gly 102 Xplor with PEG RDCs.....	252
Figure 6-52 Gly 102 Xplor with PEG + Gel RDCs cluster of 17 models.....	253

Acknowledgements

It has been an honor and pleasure to work with some of the finest people in the world over the years, chief among them David & Jane Richardson. Besides my family, whom this work is dedicated to, their contributions to my life are more significant than anyone. Their contributions to science are exceptional and deserve international recognition; this type of recognition, in truth, would mean little to them which makes them even more deserving.

I'd like to acknowledge Laura Murray, Gary Kapral, Jeff Headd, Daniel Keedy, Christopher Williams, Michael Prisant, Swati Jain, Claire Vinson, Mike Word, Bradley Hintze, and Lizbeth Videau for everything they have done as friends, labmates, collaborators, and members of the laboratory family. I'd also like to acknowledge the members of my thesis committee, John York, Len Spicer, Pei Zhou, and Brian Kuhlman for their ongoing support of my work and guidance over the years.

I'd like to thank Bryan Arendall whose investment in others and his breadth of knowledge combined with a keen sense of self and commitment to quality inspires and challenges me. Without his counsel and mentorship, I'm not sure I'd have accomplished anything in science (yes, Bryan, Ace accomplished something).

I'd like to thank Ian Davis for his indulgence of my many absurd research ideas, including a few that actually worked! The years we spent in the lab together were amongst the most productive and interesting. The work Ian did in his thesis created the

sandbox in which all the rest of us play. His contributions to the productivity of others is surpassed only by his ability to add to the happiness in the lives of his friends and family.

I'd like to thank Vincent Chen for his bottomless patience. The serendipity that caused our thesis work to intersect has been a marvelous journey and I count him as one of my most cherished collaborators. His subtle humor and calm demeanor put everyone at ease, providing good balance for those of us who get rather boisterous. Without his developments in the KiNG software, much of this work would not be possible.

I'd like to acknowledge David Stein, Steve Feller, and Rachael Brady for their contributions to the virtual reality work. Many considered that project a pipedream at the outset; they did not (at least that I know of...). A special thanks to my close friend and collaborator David J. Zielinski, whose ability to teach me about bicycles, create virtual environments, and talk about all topics over a beer on a sunny afternoon continue to be a great source of joy for me.

I'd like to acknowledge the NorthEast Structural Genomics Consortium. Without their support, enthusiasm, continuing friendship, and keen interest in producing the highest quality NMR structures for use by the biomedical community, much of this work could never have happened. Specifically Gaetano Montelione, for his ongoing support in the creation and testing of quality measures for NMR structures as well as friendship and willingness to help in the adoption of the methods by the NMR community.

Special thanks to Aneerban Bhattacharya, a longstanding collaborator and good friend during his time at Rutgers and while he was at Duke. Aneerban was instrumental

in obtaining much of the data needed for analyzing NESG structures and was always up for heading to the farmers market, going for a hike, or trying a new restaurant in the Raleigh-Durham area.

Equally, Jim Aramini at the NESG was crucial in the fruitful work calculating structures using RDC's and enlisting Theresa Ramelot at Miami of Ohio to run calculations. Jim opened up his home outside of Philadelphia to me on multiple occasions for marathon work sessions, inevitably leading to trips into Philadelphia during the summer to see a Phillies baseball games or museum visits and stops at restaurants.

I'd like to acknowledge Stuart Endo-Streeter for the collaboration on his structure re-refinement and his continuing friendship. I learned a great deal about inositol modifying enzymes from Stuart and together we learned a great deal about how IP species bind in the active sites of these structures. There are few people I enjoy talking to about science more than Stuart, and even fewer of my friends who know more about martial arts than Stuart does (making him a worthy adversary on many fronts).

I'd like to acknowledge the fellowship support provided by an NIH Ruth L. Kirschstein predoctoral fellowship and the Structural Biology & Biophysics Training grant. Implied here are all those who make that happen in the Department of Biochemistry (Amy Norfleet, Esther Self, Marsha Brooks); the School of Medicine; and the Graduate School (Betty Jones, Susan Williford, Tomalei Vess, Jacqueline Looney, David Bell, Jo Rae Wright). I'm eternally grateful for the assistance offered by scores of other people along the way, a complete listing of which would be shockingly long.

1. Introduction

1.1 Field Development: Biochemistry, Structural Biology, and Bioinformatics

In the first edition of *Principles of Biochemistry*, the authors characterize the field of biochemistry as "the application of the principles and methods of chemistry to the field of physiology and biology" (Handler, 1954). Even then, leaders in the field found the growth of the field tough to wrap their heads around stating that, "biochemistry has grown increasingly broad in its scope, it has become more difficult to achieve success in providing a single course or textbook" (Handler, 1954).

Notably, x-ray crystallography is described only in passing in the 1950's biochemistry text and the structure of proteins is described as being highly regular and very complex (pg 179 of Handler, 1954). The latter is certainly still the case, but the former is now unsupported.

Not long after the publication of this text, Christian Anfinsen showed a near complete sequence of Ribonuclease at a symposium in Washington, DC in a presentation entitled *On the Structural Basis of Ribonuclease Activity* (Anfinsen, 1957). The hunt was already on for solving protein structures by x-ray diffraction and in 1960 Max Perutz published the paper *Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis*, closing the first stage of his twenty year chase for the 3D structure of hemoglobin and following up on his work with Kendrew determining the structure of myoglobin (Perutz, 1960).

One of the earliest textbooks referencing the importance of X-ray analysis in biochemistry does so in the context of saying that it is useful for mapping the active site of an enzyme and references Perutz' work (Kosower, 1962). Kosowers book, *Molecular Biochemistry*, represents one of the earliest books – now nearly 50 years ago – on what we now define as structural biology.

Interestingly, science writers in the 1960's knew about the importance of structural biology. Hans Neurath wrote in a 1964 piece for science writers that "Ever since 1927, when Sumner first isolated an enzyme in crystalline form and showed that it was protein, biochemists have known that the solution to the mystery of enzyme structure lies in unraveling the complicated chemical structure of proteins. Towards this end enormous progress has been made in recent years through development of procedures for the isolation and crystallization of other enzymes." (Neurath, 1964).

In the world of spectroscopy, the study of the molecular structures of proteins trailed the developments in x-ray diffraction methods. In 1962, Gerald King published a textbook entitled *Spectroscopy and Molecular Structure*. Nuclear magnetic resonance spectroscopy, described in only two instances, is characterized as "providing valuable information concerning the neighborhood forces and interactions to which a given atom in a molecule or crystal is subjected" (King, 1962). Indeed, this statement holds today. However, no mention is made of gaining any structural insights or the potential to use NMR for solving the 3D structure of molecules. This textbook counts as one of the ~30 references to NMR before 1965 that Wuthrich points out in the first chapter of his well known book *NMR of Protein and Nucleic Acids* (Wuthrich, 1986).

By the mid-1960's, the two methodologies – Xray and NMR – were likely on a crash course towards one another. The area of biochemistry called structural biology was emerging and the role of 3D structure determination of macromolecules, specifically proteins, touched a driving problem of biochemistry and broader biology; the relationship between biological structure and function.

Along another related line of development, early crystallographers at Cambridge, Oxford, and MIT were all using the few computers in the world (under development at those institutions, such as the 7094 and the IBM 360 at MIT) to run the Fourier transform calculations necessary to create the contoured maps of protein structures. This is the first known instance of large (in terms of pressing the limits of the then-current system) biological data sets being analyzed using computer technology.

Fast forward to the current day, where what we call the field of bioinformatics is defined as "the branch of science that deals with computer-based analysis of large biological data sets." (Westhead, 2003). Oddly enough, the modern conceptualization of bioinformatics – largely depicted as coming from the analysis of gene sequences - seems to ignore the structural biologists' work in the 1960's using computers to assist in the determination of 3D models of protein structures.

Early protein crystallographers using computers to analyze diffraction data and the development of molecular graphics to visualize these structures predates sequence bioinformatics by a few decades. The first visualization of a helix occurred in the 1960's on the same machines being developed at MIT used to calculate the Fourier transforms for creating the contour maps.

Structural biology - more specifically the determination of 3D structures of proteins and nucleic acids by crystallography - constituted the first use of computer-based analysis of large biological data sets. Moreover, what we now call structural bioinformatics (that part of bioinformatics primarily concerned with 3D structures of macromolecules) is the predecessor of modern day bioinformatics and a major partner in the early development of computer graphics.

In the most complete treatise on the subject of structural bioinformatics to date, Russ Altman & Jonathan Dugan definition states that it “focuses on the representation, storage, retrieval, analysis, and display of structural information at the atomic and sub-cellular spatial scales.” (Bourne, 2009). This is a functional definition that follows from the goals of structural bioinformatics which are described as seeking a high resolution understanding of biology (Bourne , 2009).

1.2 Relevance to the basic biomedical sciences' contribution to human health

With some lag from bench to clinic, the history of medicine and human health in the last century has steadily followed our ability to observe finer and finer biomolecular detail. Whether it was the early Nobel-Prize-winning work of determining amino acid sequences of proteins (Frederick Sanger) and using x-ray diffraction to elucidate their 3D structures (Dorothy Hodgkin, Max Perutz etc.), or the completion of 3x coverage of the human genome 50 years later (interestingly using a method for which Sanger received his 2nd Nobel Prize); physical and life scientists have been premiere innovators of techniques that reach both deeper and finer.

For the case of macromolecular 3D structures we have seen an unprecedented explosion in number, size, accuracy, and significance. This is in large part due to the crystallographers being among the first scientists to use computers to get insight into biological processes. Half a century later, we now have the field of structural bioinformatics, created by extremely powerful computers, an enormous and accessible databank of structural information (Protein Data Bank, Berman, 2006) and two different experimental techniques for determining 3D macromolecular structures.

Early x-ray methods combined sequence information, experimental diffraction data, and assumptive information derived from foundational physical and chemical principles & data (including small molecule structures). More recent x-ray and NMR methods have, in addition, a large body of empirical information from the collective knowledgebase developed over the last four decades of 3D structure determinations. There are currently ~65,000 structures in the PDB (Protein Data Bank). Each structure record can contain hundreds of thousands of pieces of information (atomic coordinates, structure factors for x-ray, restraints and assignments for NMR, secondary structure assignments, sequence information etc.). This bewildering amount of data is a challenging problem and in need of better organization. In cases where amount and type of data are sufficient, they should be mined for relationships that illuminate our understanding of biology and human health. However, reaching a critical mass of data is only part of the battle; knowing the quality of each piece of data is also essential to advancing this body of knowledge.

1.3 Visualizing, Analyzing, & Validating Macromolecular Structures

Just as 3D structure is central to biomolecular function, 3D visualization is central to understanding those structures and functions. Macromolecular structure is inherently complex, cooperative, handed, irregular, and mobile. Even for communicating specific structural concepts, static 2D images are second-best, while the discovery of new relationships is enormously enhanced in interactive systems that fully explore the third dimension. The Richardson lab has contributed significantly to how we visualize macromolecular structures devising at different levels of detail and abstraction various visualizations that have included ribbon diagrams (Richardson, 1981), PROBE contact dot surfaces (Word, 1999), multi-criterion visualizations of validation criteria on structures (Davis, 07), the kinemage graphics language (Richardson, 1992, 1994), the Mage and KiNG software packages (Richardson, 1992; Chen, 2009),

The Richardson lab has also developed a number of novel validation tools for 3D models of macromolecular structures. Much of the development began with studying the detailed local geometry of protein structures determined by x-ray crystallography. Integral to this work is the addition and optimization of all the hydrogen atoms to x-ray structure models, performed by the REDUCE software package (Word, 1999a). Historically, the explicit modeling of hydrogen atoms is done infrequently, and without explicitly modeling nearly half of the atoms a detailed understanding of local geometry is near impossible.

Many of the Richardson lab validation tools have been incorporated into the MolProbity webserver (Davis, 2007; Chen, 2010). This free, online service allows users

from all over the world to analyze, visualize, and correct macromolecular structures. MolProbity has become integral to day-to-day research tasks, giving convenient web access to core validation tools as well as including powerful command line functions for running tools in batch. While originally MolProbity only handled X-ray crystallographic protein structures, recent updates improved its capabilities for validating nucleic acid structures, and also NMR ensembles as I'll describe later (Davis, 2007). Since its release, MolProbity has become widely used for structure evaluation, and recent results show that MolProbity has improved the overall quality of structures being deposited into the PDB (Arendall, 2005; Chen, 2010).

1.4 Macromolecular Structure Quality

For data to be of high quality, it must be both accurate and precise. We define accuracy as how close a measurement is to its true value (estimated by the agreement between the data values a model predicts and the actual experimental data). Precision, however, relates to how reproducible a measurement is; of great concern but generally not the limiting factor for macromolecular structures. Instead, there is much interest in improving accuracy, since that is critical for such uses as understanding an enzymatic mechanism or designing a drug. For this project, our definition of model is a set of structure coordinates. A target for model refinement is defined as a mathematical score judging the fit of the values derived from the model to the experimental data on one hand, and to energetic or empirical expectations on the other. The choice of which terms to

include and how to set appropriate weights between them is important to the accuracy of the resulting model.

The methodology of protein crystallography is mature, powerful, and effective, and it has transformed our understanding of biology at the molecular level. However, independent determinations of the same structure show coordinate differences much larger than theoretical estimates (Kleywegt 1999; Mowbray, 1999). Structure-validation methods have been developed to provide assessments of overall reliability, based either on model-to-data agreement (Brunger 1992; Vaguine , 1999; van den Akker and Hol 1999; Kleywegt, 2004) or on geometrical criteria measurable from the model alone (Vriend 1990; Jones, 1991; Laskowski, 1993; Lovell, 2000 & 2003). There is still, however, much room for improvement.

In structures determined by crystallography, half the atoms (the hydrogens) are essentially invisible to current methodology. When they are added and optimized, their ‘all-atom’ steric contacts (hydrogen bonds, attractive van der Waals contacts, and steric clashes) provide an independent, sensitive, and powerful validation criterion (Word, 1999a; Lovell, 2003) now implemented in MolProbity (Davis, 2007; Chen, 2010). This all-atom contact evaluation at the local level is a crucial advantage to biomedical users of protein structures, since no level of global quality guarantees protection against a large local error in the region of interest (such as an active site where a drug is designed to bind).

Most importantly, when all-atom contact analysis is combined with improved rotamer, Ramachandran, and bond angle criteria, it can identify almost all fitting errors and can usually suggest how the model should be changed.

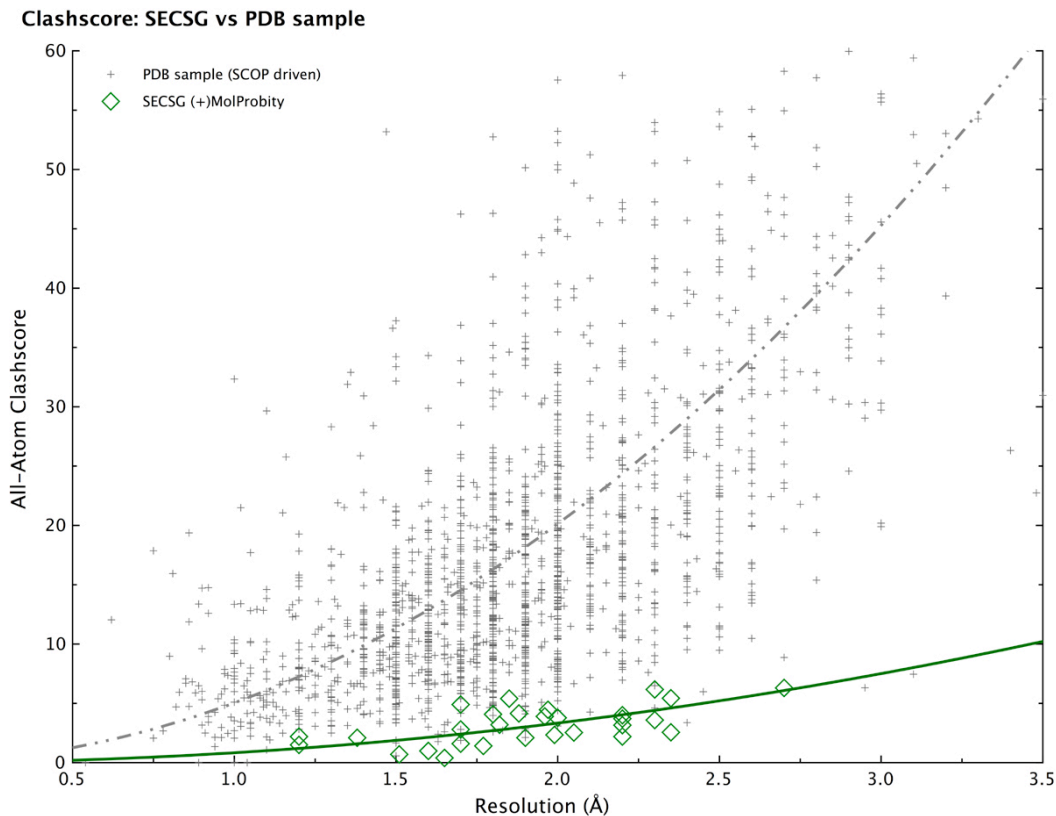


Figure 1-1 Resolution vs. Clashscore for the PDB

Using a combination of automated corrections (Word, 1999b) and considerable manual rebuilding (Richardson, 2003) the Richardson lab demonstrated that the accuracy of protein crystal structures can be greatly improved, either in the high-throughput context of structural genomics (Figure 1-1; Arendall, 2005) or for traditional crystallography. Most changes move atoms by several Å from one local minimum to another, such as for a sidechain or ligand originally fit backwards into ambiguous electron density. Structures using MolProbity-based protocols achieved order-of-

magnitude better clash, rotamer, and Ramachandran scores, with somewhat lowered R and free R, as exemplified by the plot of clashscores, for corrected (large diamonds) vs PDB-sample (small crosses) structures as shown in Figure 1-1. Before-and-after comparison of the local map, geometry, and contacts leaves no doubt that the changes are correct.

These methods are successful and are gaining increased acceptance for crystal structures, but they have not been applied yet to NMR structures. The original all-atom contact paper (Word, 1999a) showed that interiors of the best NMR structures of that era obeyed the all-atom constraints quite well, and a cross-check with Ad Bax's lab showed that his ubiquitin NMR structure (1D3Z) met our criteria better than the 1.8Å x-ray structure (1UBQ) while our corrected ubiquitin matched his RDC data better than 1UBQ did. This seemed promising, but no other NMR applications were done before I joined the lab. A large part of this work describes tailoring the MolProbity-based tools and methodology to the characteristics and needs of NMR protein structure determination, exploring the extent to which they can succeed in improving the accuracy of NMR models and developing new visualizations for NMR structure ensembles and their corresponding experimental data in the local context.

NMR structural biology takes advantage of a diverse and very rapidly developing set of experimental techniques, based on spectroscopic measurement of various interactions between the spins of atomic nuclei that are close together either through covalent bonds or through space. A subset of these NMR experiments are most relevant for this project: NOEs (Nuclear Overhauser Effect) measure through-space inter-proton

distances $<5\text{\AA}$, and are the central data used to determine 3D structure; ^3J couplings measure interactions between atoms 3 bonds apart, and are a function of the dihedral angle; and RDCs (Residual Dipolar Couplings) measure the angle of an internuclear bond vector with respect to the external magnetic field, when the molecule has a statistical weak alignment produced by an anisotropic medium (e.g., dilute liquid crystal).

The computational procedures used to derive structures from these NMR data are more rapidly changing and diverse than crystallographic methodology. There are three logically distinct stages, now increasingly combined into fairly automated systems. The first stage is assignment of each unique chemical shift in the spectrum to the resonance of a particular atom in the molecule; this puzzle is solved from an inventive panel of through-bond coupling experiments, and mostly precedes the stages I am concerned with. The second stage starts with the interlocking network of many NOE distance estimates between pairs of assigned atoms and uses simulated annealing or distance-geometry algorithms to come up with possible models of the molecule that are consistent with all those distances. The third stage is refinement of those proposed models; the ones that converge satisfactorily become the ensemble of models that constitutes the NMR structure.

The second and third stages are often done either in Cartesian space with CNS (Brunger, 1992) or NIH-XPLOR (Schwieters, 2003), or in dihedral-angle space with DYANA/CYANA (Guntert, 1997). Many methodological variables affect the results and are not standardized, such as how the NOE intensities are converted into distance

constraints, to what extent geometrical target values are utilized, or what error models to use in RDCs.

Before an NMR structure ensemble is deposited in the PDB, hopefully along with the experimental restraint data as well (now mandatory), its quality is checked with structure-validation software. In analogy to the R and R_{free} values (Brunger, 1992) for crystallography, agreement with the experimental NMR data should be assessed by the number of NOE constraint violations (when the distance between two atoms is inconsistent with their measured NOE), as tabulated by programs such as AQUA (Laskowski, 1996; Doreleijers, 1998). Usually the model with the fewest violations is made model 1 in the file.

Although there is not yet an agreed-upon standard, several attempts have been made to define a measure of agreement between the observed and the back-calculated spectra or data values; one example is the RPF score, using information retrieval statistics, developed by the Montelione lab (Huang, 2005a). The commonly reported rmsd between the models in the ensemble can be an indication of overall structural accuracy, but it is not quantitatively reliable because it can be made artificially low by tight scaling of allowed distance ranges.

A second aspect of structure validation is geometrical criteria such as ideality of bond lengths and angles, chirality, planarity, dihedral angles, and steric overlaps. Programs commonly used for NMR ensembles include WHAT IF, PROCHECK_NMR, QUEEN, and others as reviewed by Spronk (Spronk, 2004). Average values give a useful global evaluation, while individual outliers flag local problems and in principle

could be used to try making corrections. However, local listing is only sometimes for crystal structures and almost never for NMR structures. The list is long and unwieldy, and does not suggest how to make the needed changes. Validation is typically done only after the work is considered complete, and for NMR a change would in many cases mean redoing the entire coupled structure-determination and refinement process.

Most insidious, many NMR spectroscopists feel that dihedral and steric criteria should not be imported from outside sources and are not necessarily applicable to the more dynamic NMR structures in solution (Bertini, 2003). Similar ideas have even been proposed for surface sidechains in crystal structures (Carugo, 1997). However, a large body of work from the Richardson lab shows that Ramachandran, sidechain rotamer, and steric-clash outliers are almost entirely accounted for by poor data (Lovell, 2000; Lovell, 2003; Richardson, 2003; Butterfoss, 2005). In the well-ordered parts of high-resolution crystal structures there are 0.5% to 1% real cases of rotamer outliers, but they all have strong H-bond or packing interactions holding them in those strained conformations. For dynamic surface residues, there are no interactions that could hold them away from the favorable rotamer conformations. NMR structures typically contain 10% to 50% rotamer outliers, which are highly unlikely to be correct. Fortunately, there is a growing body of evidence, from theoretical analyses of experimental order parameter and RDC data, that the best NMR measurements show sidechains as moving between multiple closely rotameric states (Chou, 2001; Chou, 2003; Hu, 2005; Lindorff-Larsen, 2005).

It is biomedically important both to achieve more accurate macromolecular NMR structures and to have better structure-validation methodologies to assess that accuracy.

It is my hope that the tools and knowledge developed in this work will help both the practicing structural biologist and also the biomedical or bioinformatics researcher who makes use of the structures. NMR structures are in general less accurate than crystal structures and sometimes fail to work for purposes such as molecular replacement or homology modeling. The best NMR structures are very accurate, however, and it would be very valuable to bring more of them up to that level. More experimental data is the most effective way to increase accuracy (analogous to higher resolution in crystallography), but the new methods described in this work hold promise for extracting better accuracy from a given body of data. Structural accuracy is definitely critical for deriving functional/mechanistic insights, for theoretical simulations, for protein design, and especially for drug design.

Perhaps the most significant advantage of NMR over crystallography is that it can provide information about conformational dynamics. In order to fully utilize that potential, however, the signals of true variability must not be swamped out by noise from variability unsupported by experimental data and assigned only because it is in principle possible. I believe this work contributes insights that support the Richardson lab quest to bring NMR and x-ray structure methodologies closer to each other and closer to the underlying molecular reality.

1.5 Tools for Structure Analysis, Visualization, & Validation

1.5.1 Mage & KiNG

These are two molecular viewing software packages, developed in the Richardson lab. Both programs display kinemage files. Kinemages (Richardson 1992; Richardson 2001) are commented, hierarchical, 3D display lists in readable text form and are readily modified on-the-fly for research purposes. They are often used in combination with other interoperable software to evaluate or change properties of protein models, or of other graphical objects such as 3D data distributions.

KiNG ("Kinemage Next Generation") is a Java kinemage viewer used on MolProbity for clients and as a stand-alone application for 'Backrub' backbone movement (Davis, 2006; Chen, 2009) kinemage editing, and NMR functionality. Both Mage and KiNG have the important capability of interactively updating the all-atom contact display while refitting an idealized sidechain with the help of a hypertext rotamer distribution and torsion angle dials or sliders. KiNG includes most features of Mage plus an increasing number of novel capabilities, and it is the major test-bed for this project. An important recent enhancement in KiNG is the updatable display of NOE data by command-line calls to the NOEDisplay software (Coggins), shown on the 3D structure as dashed lines colored by agreement of the NOE distance in the data with the model. KiNG displays contoured electron density maps in O, CNS, or CCP4 formats with automatic update on recentering.

1.5.2 PROBE All-Atom Contact Analysis

The PROBE software (Word, 1999a) rolls a small sphere over the van der Waals surface of each atom in a model. At each point on an atom's surface that is within 0.5 Å of non-covalent contact with another atom PROBE produces a color-coded dot or spike for a visual display and a numerical term that contributes to a calculated score. For attractive van der Waals contacts the dots are blue or green and the score is $e^{-(\text{gap}/\text{err})^2}$ where gap is the distance to the other atom's surface and err is the probe radius of 0.25 Å. The favorable atomic overlap between an H-bond donor and acceptor produces pale green dots and is scored by overlap volume. An unfavorable steric clash overlap produces spikes rather than dots, of increasingly violent colors from yellow to hot pink, and again is scored by overlap volume (shown in Figure 1-2). Weighting factors among the 3 terms gives a score profile similar in shape to the van der Waals function for an isolated pair-wise interaction. However, the contact penalty for overlap increases much less steeply, because large clashes do not actually signify a problem in the molecule but just mean that there is an error in the model.

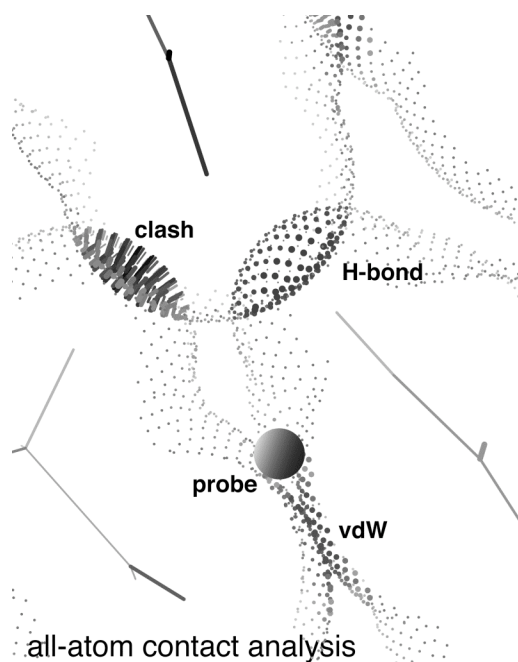


Figure 1-2 - PROBE All-Atom Contact Analysis

This sort of contact behavior is more complex, and we think more realistic, than a pair-wise potential in the common situation where multiple atoms intersect or especially where they completely shield one another. This project uses the global clashscore, which is the number of serious ($>0.4\text{\AA}$) clashes per 1,000 atoms in a structure. Similar scores based on actual dot counts are also explored in this project and are described later.

1.5.3 REDUCE

Addition and optimization of all H atoms is absolutely necessary for the detailed analysis of atomic contacts, 75% of which involve H on at least one side. The addition and optimization of H atoms is done by the program REDUCE (Word, 1999b) adding most H from local geometry and keeping all methyls staggered. REDUCE optimizes rotatable OH, SH, NH_3 and His protonation by a complete combinatorial analysis of local

H-bond networks. The first layer of water is used, but allowed to be donor or acceptor as needed. Since N vs. O of sidechain amides and N vs. C of His rings are hard to distinguish in electron density and are often misassigned, the H-bond network analysis in REDUCE optimizes their 180° flip orientation in crystal structures, considering both H-bonding and steric clashes. A commented PDB file is output, with the new H atoms and possible flips. For anyone who wants to assess the results of the automatic procedure in MolProbity, a Perl script called Flipkin produces a kinemage with views and comparison animation for each relevant sidechain.

For NMR structures the protons are already present, although we find that OH orientations are often poorly optimized. The flip analysis for Asn, Gln, and His is not performed for NMR structures, since the two sides of these groups are readily distinguished by NMR data. The REDUCE step could in principle be omitted for NMR models, but we prefer to use it for improved OH geometries and better comparisons (PROBE is very sensitive to small differences in assumed values for ideal bond lengths and angles). Coordinates that come directly from refinement (e.g. CNS format) rather than from the PDB need an additional conversion step, because the proton naming conventions unfortunately differ (e.g. opposite numbering order for handed methylene and methyl protons). This same conversion must be done consistently for the restraint files as well.

1.5.4 BioGeometry Local Density

Work by the BioGeometry group at Duke produced useful and clearly-defined interface surfaces with a non-arbitrary boundary definition (Ban, 2005). The software in this work was extended to quantify the packing density of proteins by measuring the local density of atoms (Ban, 2006). This method focuses on backbone amide and interior protons that best reflect the core packing of a protein, differing from other volume studies in that it investigates protons, rather than the heavier atoms. Local density Z-scores, based on the statistical deviation from distributions of proton local densities derived from a quality-filtered dataset of high-resolution protein structures, increase as the resolution of crystal structures gets worse and correlate with the Clashscore determined by PROBE. The work showed that NMR structures, both database-wide and individually, have problematic inflation and compression across a wide range of values for the mean pairwise RMSD between models. It also demonstrated that better refined NMR structures fare better, with significant improvements in packing quality seen for the final DrESS structures – a set of 100 re-refined NMR structures (Nabours, 2004; Ban, 2006). The new measures are especially designed to provide a previously missing validation criterion for loose packing of NMR structures, but are expected also to be useful in evaluating low-resolution x-ray structures or homology models.

The packing density of a protein had previously been investigated using volumetric measures based on the Voronoi diagram, a partitioning of the space in which the protein sits (Voronoi 1907, 1908; Richards 1974). The Voronoi diagram is used for both the computation of standard volumes in protein models (Gerstein, 1995; Lo Conte,

1999; Tsai, 1999; Tsai, 2002) and as a quality measure for protein models determined by X-ray crystallography (Pontius, 1996). The work in Ban et al. 2006, uses a generalization of the Voronoi diagram, known as the power diagram, in order to handle atoms of different radii. A power cell for a specific atom is the space closest to that atom versus all other atoms as measured by the power distance evaluated in three-dimensions. The local density for a proton is the fraction of the individual power cell occupied by the proton and is a number between zero and one.

These local density packing evaluations are publicly available as a stand-alone web service. Their outputs include PDF's containing graphs evaluating the density distributions for structural models, and kinemages that can be displayed in MAGE or KiNG.

1.5.5 MolProbity

MolProbity is a suite of software tools for assessing and improving the accuracy of macromolecular structures, consisting of a web service for analysis that is tightly integrated with 3D graphics programs for rebuilding (MAGE and KiNG). The MolProbity web service at <http://kinemage.biochem.duke.edu> (Davis, 2004; Davis, 2007; Chen, 2010) provides a convenient interface to hydrogen addition and all-atom contact analysis, as well as geometrical evaluations such as C β deviation and updated Ramachandran and sidechain rotamer analyses (Lovell, 2000; Lovell, 2003). Results are presented online both as tables and visually as a Java-based web-page embedded version of KiNG. Offline, the programs MAGE and KiNG provide interactive model rebuilding

tools, guided by dynamically updated displays of all-atom contacts and geometrical criteria as well as electron density maps or NOE constraints (Word 2000; Richardson 2003; Davis, 2005). MolProbity is written in PHP and runs under Apache on a dedicated Mac web-server machine.

Performing the automatic corrections in MolProbity, together with rebuilding based on the analyses, has been shown to significantly improve final re-refined models of x-ray structures. Crystallographic R and R_{free} typically drop by 1 – 2%, indicating a better overall fit to the diffraction data, as has been demonstrated through collaborations with the SouthEast Collaboratory for Structural Genomics (Arendall, 2005). More importantly, corrected regions show superior fit to the electron density and all local measures of steric and geometric quality improve at the same time, often by an order of magnitude. Thus, use of the MolProbity suite effectively increases the amount of reliable structural information that can be extracted from a given data set.

1.5.6 WHATIF / WHAT_CHECK / PROCHECK / PROCHECK_NMR

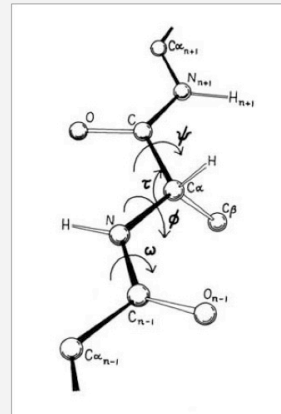
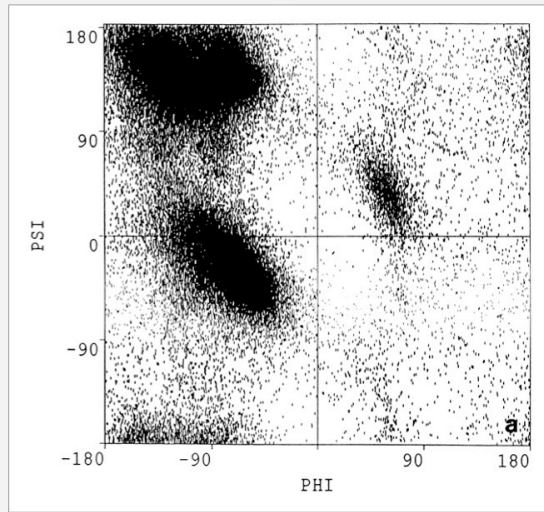
The software subset WHAT_CHECK (Hooft, 1996) is contained within and originated from WHAT IF (Vriend, 1990). WHAT_CHECK is commonly used for structure validation and verification checks during structure determinations and before submission of structures to the PDB (Hooft, 1997). It uses a number of different calculated structural parameters from the input PDB file and makes comparisons to standard values. The broad categories under which different analyses are performed in WHAT_CHECK are nomenclature, symmetry, geometry, and ‘structure’ (Hooft, 1997).

These categories include both local and global measures. Things such as bond lengths and bond angles (Engh and Huber 1991), Ramachandran values, ring planarity and proline puckering (Cremer, 1975) are routinely checked, among other things.

WHAT_CHECK uses empirical reference distributions for each of the measures evaluated, and based on a periodically updated but not listed set of the '300 best structures from the PDB' under 1.2Å resolution (Hooft, 1997).

PROCHECK (Laskowski, 1993), which performs both overall and residue-by-residue geometry evaluations of protein models, is an alternative to the WHAT IF / WHAT_CHECK suite of tools. It performs checks on bond angles and bond lengths, planarity checks on aromatic and amide sidechains, Ramachandran checks (Morris, 1992), sidechain chi angle checks, distance cutoff non-bonded interaction checks, calculated backbone hydrogen bonding energy checks, and disulfide bond checks. Many of these are based on distributions or calculations published by others, such as the DSSP energy calculation for backbone hydrogen bonding energy, or the Engh & Huber ideal values for bond lengths and angles. Notable here is the difference between the Ramachandran distribution used for PROCHECK in Figure 1-3, containing the entire unfiltered contents of the PDB from 1991, and the more recent MolProbity distribution which uses only high-resolution, quality-filtered data in Figure 1-4 (Lovell, 2003).

Ramachandran plots: then

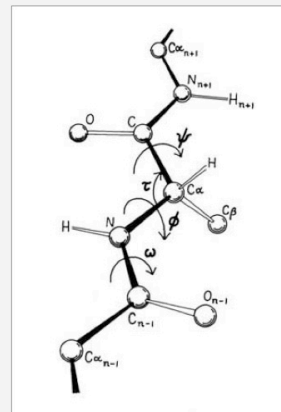
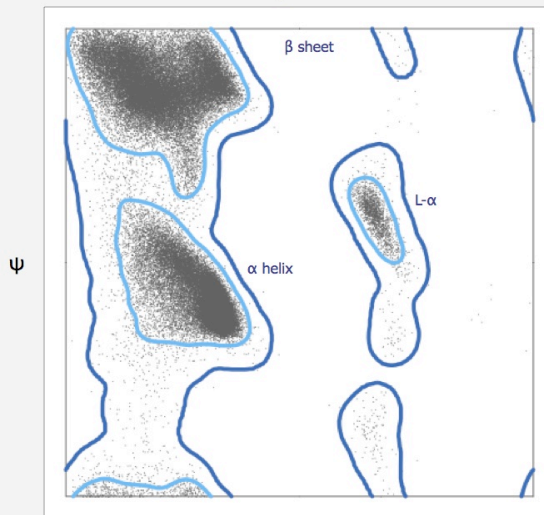


Drawing by Jane Richardson

ProCheck, 1991

Figure 1-3 - ProCheck Ramachandran Data

Ramachandran plots: now



Drawing by Jane Richardson

MolProbity, 2003

φ

Figure 1-4 - MolProbity Ramachandran Data

PROCHECK-NMR (Laskowski, 1996b) has similar functionalities as PROCHECK (and WHAT_CHECK) but was specifically developed to do both geometric and restraint analyses on NMR ensembles. It calculates geometrical values such as the circular variance (or spread) of the backbone dihedral angles of an ensemble; phi,psi distributions plotted against the old PROCHECK standard (Morris, 1992); coordinate RMSD from mean coordinates in an ensemble; and chi angle distributions for sidechains. Restraint analysis, done with the AQUA software (Laskowski, 1996b), plots restraint distances, number of restraints, and various restraint violation numbers both in a residue-by-residue and a model-by-model mode (Laskowski, 1996a)

1.5.7 QUEEN

QUantitative Evaluation of Experimental NMR restraints (QUEEN) uses restraint files in X-PLOR / CNS format and PDB files of structure coordinates as inputs (Nabuurs, 2003). QUEEN analyzes five types of distance restraints: intra-residue, sequential, medium range, long range, and inter-chain. It provides information about the relative contribution of a given restraint to the structure determination, identifying which distance restraints are important, which ones are unique, and which ones are redundant (Spronk, 2004).

1.5.8 Secondary Structure Assignment

The secondary structure assignment software DSSP takes PDB files as its input and classifies the secondary structure of each amino acid residue into one of nine

categories based on the hydrogen bonding pattern of the backbone (Kabsch and Sander 1983). DSSP is the standard used in PDB file headers, but other alternative systems have been developed since. DSSPcont (Andersen, 2002) uses an entire NMR ensemble to create weighted averages and assign a secondary structure classification at each amino acid position based on all the models. The software STRIDE (Frishman, 1995) likewise uses hydrogen bonding, but also includes Ramachandran values for the backbone as well. DEFINE (Richards, 1988) and P-SEA (Labesse, 1997) make their secondary structure assignments based on $C\alpha$ distances. The software PCURVE (Sklenar, 1989) assigns a smoothed axis for local backbone direction, and uses that for its assignment. Of particular note is that various secondary structure assignment methods assigned 20% of the residues in the same files to different states (Fourrier, 2004), thus making it important to note which package one is using, and to consider different ones if making critical judgments dependent on their strengths and weaknesses.

1.6 Structure Calculation & Refinement Tools

The vast number of algorithms, operations, inputs, outputs, options that can be turned on or off, and parameters that are adjustable within the following software packages make for a very large toolkit for macromolecular structure calculation and refinement. Notably, many of the packages contain format conversions allowing restraint lists and other file types to be used in other software packages. Therefore, each package may include novel algorithms and/or a novel sequence of operations, while still using

many other functions that are similar or the same implementations from other software packages.

1.6.1 AutoStructure / AutoQF

This software implements an algorithm that creates topology-constrained distance networks using graph theory (Huang, 2005b) taking as inputs the sequence, resonance assignments, and NOESY cross-peak lists. It then uses XPLOR/CNS or DYANA to generate structures iteratively. AutoQF, which determines RPF scores (described below), gives a quality check during each iteration (Huang, 2005a)

The RPF scores in AutoQF use information retrieval calculations to produce R-factor type measurements for NMR structure models that are sensitive to the vast amounts of true negatives inherent in NMR data – that is, cross peaks not observed (Snyder, 2005). The recall (R) is the percentage of assigned peaks in the NOESY spectra that match their distances in the structure model; whereas precision (P) measures the percentage of expected close proton pairs that are actually present in the NMR data when NOE interactions are back-calculated from the structure model (Baran, 2004). The F-measure score then combines R and P, returning a global measure of the structure's fit to the NOESY spectra.

1.6.2 CYANA / DYANA

DYnamics Algorithm for NMR Applications (DYANA) is a NMR structure calculation software package that uses molecular dynamics simulations in torsion angle

space. This use of torsion angle (internal) coordinates rather than Cartesian coordinates is its major unusual feature (Güntert, 1997; Güntert, 2002). DYANA is no longer actively supported, and is now a part of the Combined assignment and dYnamics Algorithm for NMR Applications (CYANA, Güntert, 2003). CYANA combines an improved torsion angle molecular dynamics simulation algorithm with an automated NOESY cross-peak assignment algorithm called CANDID (Herrmann, 2002).

1.6.3 CNS / XPLOR / XPLOR-NIH

Crystallography & NMR System (CNS), which is a newer version of XPLOR (Brunger, 1992b), is used for macromolecular structure determinations for both x-ray crystallographic and NMR experiments. It has both an HTML and a command-line interface. The software allows for operations on data structures such as crystallographic structure factors, electron density maps, and atomic properties. It contains extensive functionalities for improvement of crystallographic phases, including structure refinement in both real and reciprocal space (Brunger, 1998). It performs simulated annealing as well as energy minimization against either x-ray or NMR data.

XPLOR-NIH is a structure calculation package for NMR based on minimization protocols derived from molecular dynamics and simulated annealing. Its core is from XPLOR/CNS, including the use of distance geometry for creating a starting model, followed by rounds of simulated annealing. Its target function includes experimental terms based on agreement to the common types of NMR data, including NOE, J-coupling, RDC, and chemical shift anisotropy experiments. The target function also

includes the standard molecular mechanics energy terms for covalent and non-covalent geometry. XPLOR-NIH is somewhat unusual in that it includes empirical terms such as Ramachandran and rotamer potentials (Schwieters, 2003).

1.6.4 ARIA

Ambiguous Restraints for Iterative Assignment (ARIA) can deal with keeping alternatives for ambiguous distance restraints; it uses automated NOESY cross-peak assignments along with torsion angle dynamics, simulated annealing, explicit water refinement, and other force fields to accomplish a structure calculation (Nilges, 1995; Nilges, 1997; Linge, 2003). It has outputs and conversions such that it can be used in conjunction with CNS/XPLOR.

1.7 Conclusions

I've described a large variety of tools, metrics, methodologies, and visualizations. It is against this backdrop that I began to work towards understanding protein structure determination and validation with the goal of improving NMR structures. What came first, however, was an opportunity where I delved deeper into improving a protein structure determined by crystallography.

2. The Inositol Signaling Pathway: Enzyme Structures and Specificity

As previously described, the Richardson lab has developed new methods for validation and improvement of crystal structures and has demonstrated their effectiveness in high-throughput use (Arendall, 2005; Chen, 2010). A local opportunity arose for further test of these structure improvement methods and for me to gain direct experience using the MolProbity tools in crystallographic rebuilding and refinement of inositol phosphatase and kinase enzymes studied in the York laboratory.

D-myo-Inositol, a six-carbon cyclic alcohol, and its derivatives are involved in a highly specific and crucial signaling system unique to eukaryotes (Irvine, 2005; Irvine, 2001; Majerus, 1999; Xia, 2005; York, 2006; York, 2001) and provide regulation through a wide variety of phosphorylated species. Because of steric crowding, inositol phosphates cannot adopt a boat conformation and seem to be always in chair form. With the exception of the axial 2' position, all other substituents (OH or PO₄) in inositol are equatorial. The unique axial 2' position confers stereospecificity on the inositol species, where the addition of phosphate moieties in different combinations leads to a large number of distinct inositol polyphosphate (IP) variants. The inositol signaling code is thus a binary choice (OH or PO₄) in each of the six numbered and geometrically distinct positions around the ring. Recently recognized pyrophosphate substituents add another logical layer to these variants. Each unique IP variant has the potential for quite specific interactions with different molecules.

Species such as inositol-1,4,5-trisphosphate (IP₃) are well known for their direct involvement in important biological processes, such as triggering release of Ca²⁺ from intracellular stores (Cui, 2004; Schulz, 2004; Tisi, 2004; Wagner, 2004; Wagner, 2003; Zhu, 1999). In other systems, the amount of different IP species present is of primary importance in the signaling cascade; examples include immune cell development (Jayaraman and Marks, 1997), endoplasmic reticulum membrane-localized signaling pathways (Jesch, 2006), metabolism (Bechet, 1970; Delforge, 1975; Messenguy, 1976), stress response (Dubois, 2002; Xiong, 2001), and transcription and translation control (Odom, 2000; Saiardi, 2000; York, 1999; York, 2005)

Many of the details about inositol signaling remain unknown, yet new IP species are discovered *in vivo*, for which roles have not been proposed. In eukaryotic investigations (done in yeast), a major inositol polyphosphate population is inositol-1,2,3,4,5,6-hexakisphosphate (IP₆) (York, 1999). Other IP species exist with population increases under specific circumstances (Fujii, 2005; Seeds, 2005; Stevenson-Paulik, 2002), along with recent developments investigating the synthesis of pyrophosphorylated inositols (PP-IP species; Fridy, 2007).

The wide variety of IP species has a large number of modifying enzymes that go with them: inositol kinases and phosphatases. Multiple different IP kinases and phosphatases have been discovered and linked to various systems (Majerus, 1999; Shears, 2004; York, 2006). These modifying enzymes represent one component in a large network of interactions through which cellular signals are transmitted.

Phosphatidylinositol 1-phosphatase, and inositol polyphosphate multi-kinase are both part of this inositol modifying enzyme network.

2.1 MolProbity Diagnosis, Model Building, and Structure Refinement of Phosphatidyl Inositol 1-phosphatase (1Ptase)

Inositol polyphosphate 1-phosphatase (1Ptase) is a lithium-inhibited protein that requires Mg^{2+} as an enzymatic cofactor and removes the 1'-phosphate from $I(1,4)P_2$ and $I(1,3,4)P_3$. Open questions about the molecular target and side reactions of lithium therapy, commonly used in the treatment of bipolar disorder, add to interest in the 1Ptase molecule. The original wildtype crystal structure at 2.3Å resolution was published in 1994 (1INP; York, 1994). The overall fold is an alternatively layered $\alpha/\beta/\alpha/\beta$ sandwich (see ribbon schematic Figure 2-1), with a large, flat, six-stranded β -sheet near the center and a kink at the active-site DPIDST sequence motif.



Figure 2-1 - Ribbon Schematic of 1Ptase

Crystallographic data were collected for a D54A (phosphatase-dead) active-site mutant of 1Ptase by J.P. Xiong in 1997 in the York lab. Two D54A structures were investigated, apo at 2.5Å and with inositol 1,4 bis-phosphate (IP2) at 2.8Å resolution, but these were never deposited or published because they failed to refine satisfactorily.

The clashscore, Ramachandran and rotamer criteria, and the high *B*-factors and low electron density in many of the loops of these previous structures (see Figure 2-2 below, refer to chapter 3 for explanation of outlier glyphs) suggested the presence of fixable local errors. As of SCOP 1.75 (6/09), many other related enzyme structures have been solved, but none in the same protein family as 1Ptase. Therefore, in collaboration with Stuart Endo-Streeter, then a graduate student in the York lab, I used MolProbity-

based tools to diagnose and correct enough problems in the 1Ptase mutant model to allow successful refinement.

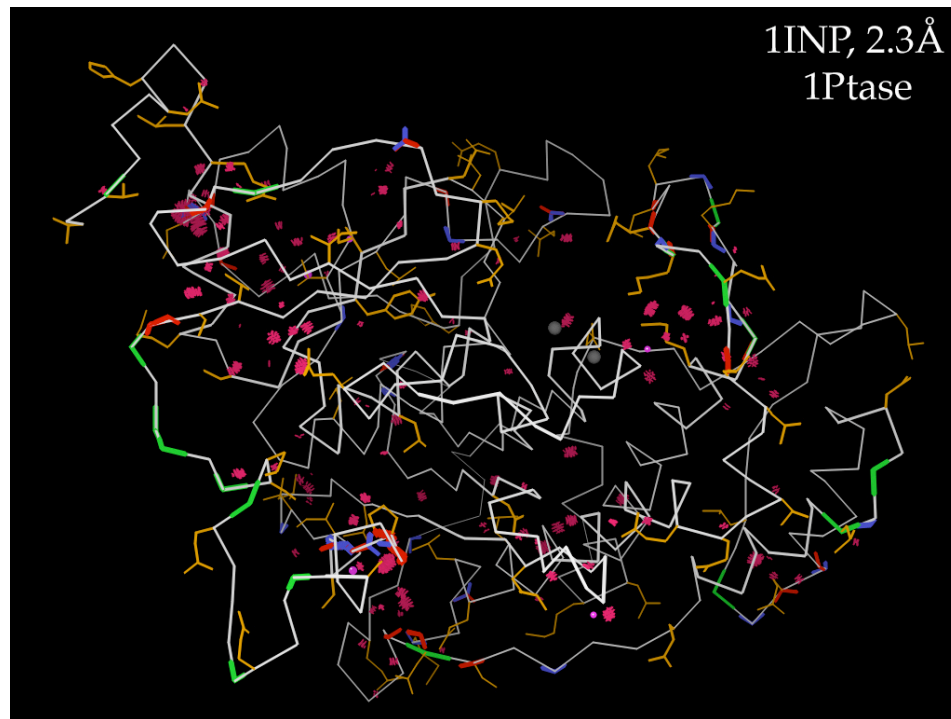


Figure 2-2 - Multicriterion Kinemage of 1INP

The deposited 1INP structure was used as the starting model, and corrections were also made in that structure. The three refinements were carried out in parallel. Most figures use 1INP, since the higher resolution shows the benefit of changes more clearly.

The initial 1Ptase structure was old enough not to have defined an R_{free} set (5-10% of the diffraction data held out of refinement to provide an unbiased cross-check), so that a pseudo- R_{free} set for our further use could be defined only after perturbing the model (and thus at least partially unbiasing the phases) with a high-temperature round of simulated annealing in CNS (Brunger, 1998). Unfortunately, this degrades the starting

structure significantly, but the benefits of having an R_{free} measure were judged to outweigh that degradation.

We utilized the MolProbity "multi-criterion" 3D kinemage graphics (generalized for multiple NMR models in chapter 3) to collect and visualize all of the local steric and geometric quality evaluations together on the 1Ptase 3D models. Correction was then attempted for local clusters of problems or large individual outliers, using either the traditional crystallographic rebuilding program O (Jones, 1991) or Ian Davis's KING graphics and modeling software (Davis, 2004; Chen, 2009) which displays electron density and interactively updated all-atom contact, rotamer, and Ramachandran criteria along with the changing molecular model. In addition to easy and interactively validated sidechain rebuilding, KING can make realistic small backbone movements with the BACKRUB tool (Davis, 2006) which often enables correction of otherwise recalcitrant problems because of its leverage on shifting sidechain position in a direction not accessed by the χ angle variables.

The first set of corrections made were 180° "flips" of Asn, Gln, and His sidechain groups, which are evaluated and performed automatically as part of the H-bond network optimization in REDUCE (Word, 1999b) and were rechecked every round as their neighbors refined. There is no cost in terms of fit to the experimental data, but it is valuable in terms of improved hydrogen bonding. In 1INP, there were four clear NQH flips (Asn17, Asn112, Gln208, Asn369). These flips correct a systematic error that occurs often in x-ray structures because the electron density does not discriminate between O and N or N and C atoms except at extremely high resolution. (This particular

problem is not relevant to NMR structures, however, where the NH or NH₂ would be clearly identified once its resonance(s) are assigned.)

A second type of correction is for tetrahedrally branched sidechains (such as Thr, Val, Leu, Ile, or Arg) which have been misfit backwards into ambiguous density (Lovell, 2000; Headd, 2009). For 1Ptase these were mainly leucines. (In NMR models, similar backward-fit sidechains can occur for Val or Leu because of incorrect stereospecific assignments for methyl groups.)

The structure of 1Ptase has many long loops with considerable disorder, which is not unusual for a large structure at this resolution. Several places had errors such as modeling backbone into what should be sidechain density and vice versa, resulting in diagnostic clashes and Ramachandran outliers. Especially notable were three separate places where an additional turn of α -helix could confidently be built into one end of a loop (residues 96-103, 132-139, and 387-392). The first of those cases is shown in the Figure 2-3 for both original and rebuilt versions.

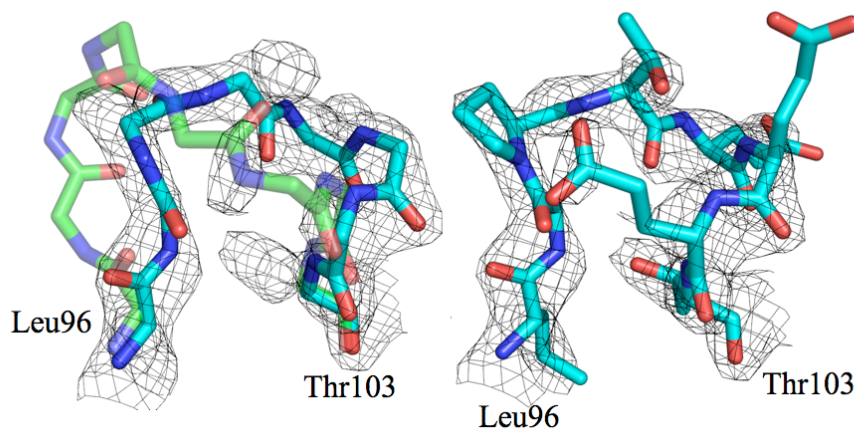


Figure 2-3 - First turn of a helix fit as Loop, in 1INP

Figure 2-3 shows the main chain of the original (green) and final (teal) models on the left, and the final model with sidechains at right. Shown are the 2Fo-Fc electron density maps at 2.3Å resolution, contoured at 1.5 σ .

The most dramatic change was correction of a sequence register-shift error by one residue in a beta strand (done both for the wildtype 1INP as well as the mutant 1Ptase), identified by intractable problems in fitting the neighboring turn and confirmed by clear electron density for a His ring misplaced by one position. The turn preceding the register-shifted strand was originally modeled as three residues (as shown on the left in the image pairs in Figure 2-4 and 2-5). The bad turn has a single severe backbone clash and a number of severe sidechain clashes; the backbone is pulled tight against the inner edge of its density; the Leu226 sidechain doesn't fit its density, and Gly227 has unfilled additional density. Three different people failed to fix these problems by local rebuilding, which suggested the possibility of a register-shift. An investigation of the strand following the turn showed good backbone density, plausible fit of mid-sized sidechains into mid-sized density, and complete disorder at the far end. However, a shift by one residue produced an excellent fit of the sidechains to the density. Most definitive were the improved turn (compare Leu226 in Figure 2-5), Gly227 now in a position without sidechain density, and the good fit of the ring-shaped density for His231 after the shift.

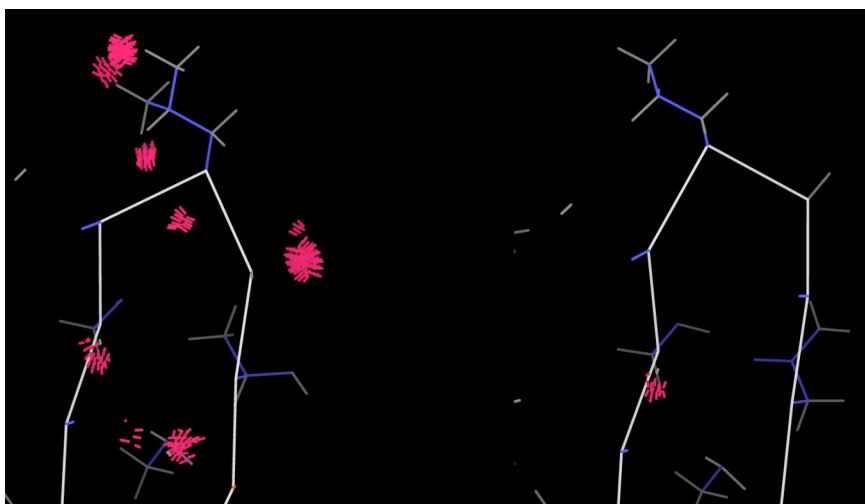


Figure 2-4 - Clash repair in a turn: 1INP at left, rebuilt at right with register shift

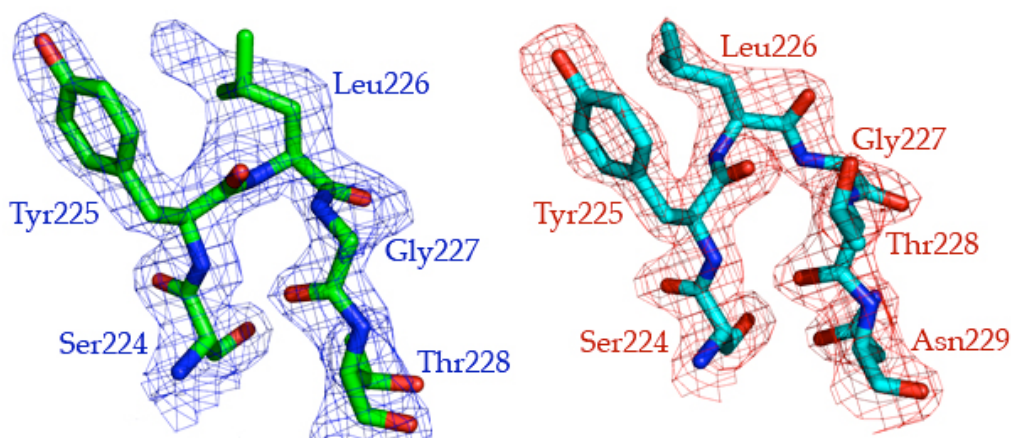


Figure 2-5 - 1INP turn with density at left, rebuilt and re-refined structure at right

The original fit (left in both Figure 2-4 and Figure 2-5) had a 3-residue turn with convincing density for the Tyr sidechain but not the Leu or Gly. At right in both figures, the following strand has been shifted up by one residue (note the Thr shift) to give a 4-residue turn and better density fit throughout.

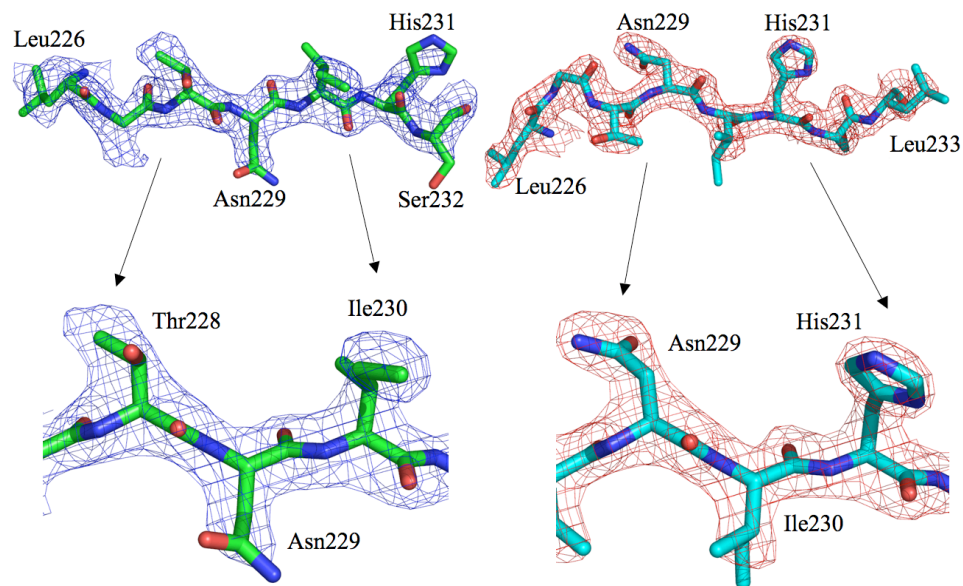


Figure 2-6 - 1Ptase register-shift error exit strand

Shown in Figure 2-6 is the β -strand electron density before and after turn change and register shift. Original (green) and final (teal) models of the β -strand, shifted by one residue after 224-228 turn rebuild are shown. Note relatively minor changes to area occupied by main-chain and significant changes in side-chain density, especially for Gly227 and for residues 228-231 in the expanded images.

As the structure refinement proceeded and model corrections were made, the new maps calculated after each round improved locally, enough to show additional residues previously invisible or unclear, which could now be modeled confidently. Also, these local corrections improved the phases in general, making the density more interpretable even in distant places such as the active site. The aforementioned register-shift correction was fixed and refined as the only change in round six of refinement. The resulting improvement was 1% in the crystallographic R and 1.6% in R_{free} for that single fix.

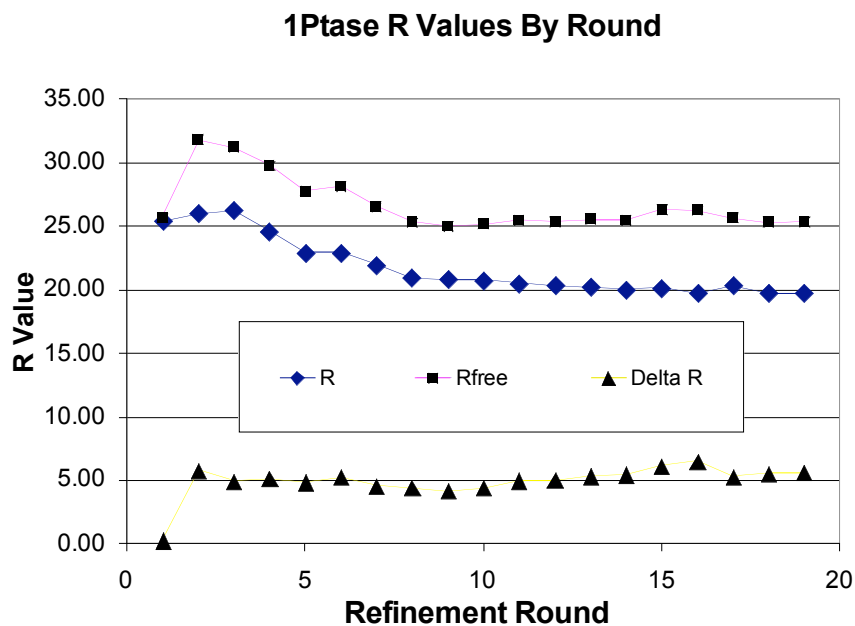


Figure 2-7 - R Value vs. refinement round for re-refinement of 1INP

The R and R_{free} were tracked through rounds of refinement (shown in Figure 2-7). The first round was the rigid-body fit of the original native model to the D54A dataset. R and R_{free} were judged to be too close together, and with no previously assigned R_{free} test set, no difference in the two would be expected. Rounds two through four were simulated annealing runs to “shake out” the influence of test-set reflections on the model, as can be seen by the increasing R_{free} values in the graph. Simulated annealing - carried out over a range of temperatures in CNS, from 1500K to 7000K - was halted after round four as the R_{free} score improved from the previous round, likely indicating that bias had been removed and that changes to the model were improving its fit to the data. Positional minimization was performed in CNS for round five. Rounds six through twenty were done including model fixups, positional minimization, and followed by B-factor

minimization. Rounds twenty-one to twenty-three, the final round, were performed in REFMAC and are not shown.

Of special note are changes in R/R_{free} after round six where the first group of changes to the model were made based on fixes identified in KiNG and MolProbity. The R/R_{free} scores dropped dramatically from 22.95% to 21.97% and 28.19% to 26.55% while the divergence between R and R_{free} fell from 5.25% to 4.58%.

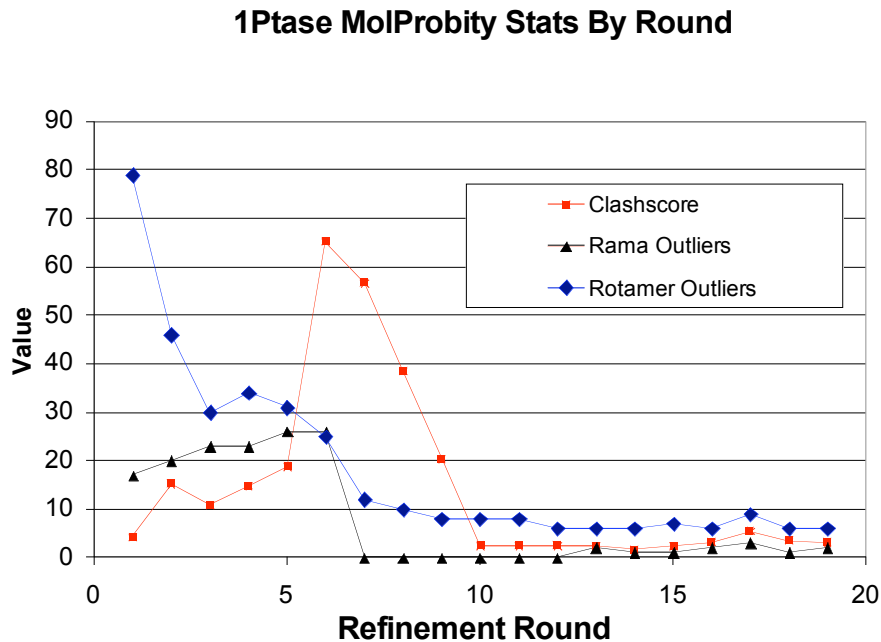


Figure 2-8 - MolProbity scores vs. refinement round for re-refinement of 1INP

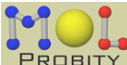
In general, the corrections more reflect the advances in available diagnostic tools since the original determination than any carelessness on the part of the previous crystallographer. The plot of R and R_{free} versus refinement round for the 1INP 1Ptase generally shows improved agreement between model and data in the same rounds where

the plotted MolProbity statistics (clashscore, rotamers <1%, Ramachandran outliers) substantially improve, as shown in Figure 2-8.

Shown in Table 1 below are the various crystallographic statistics for the original structure, the D54A mutant, and the D54A mutant with IP bound.

Table 1 - 1Ptase Re-Refinement Structure Information

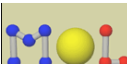
Data Collection			
	Native	D54A	D54A + I(1,4)P ₂
Space Group	P4 ₁	P4 ₁	P4 ₁
Unit cell (Å)	a=b=51.640 c=143.330 α=β=γ=90	a=b=51.547 c=143.105 α=β=γ=90	a=b=51.486 c=142.880 α=β=γ=90
Wavelength (Å)	1.53	1.53	1.53
Resolution (Å)	2.23-23.1	2.5-35.0	2.80-34.96
Unique reflections	15787	11831	8116
Completeness (%)	100 (100)	100 (100)	100 (100)
R _{sym} ^b		7.7 (37.2)	11.1 (48.2)
Refinement			
Resolution range (Å)	2.23-23.1 (2.23-2.29)	2.50-35.0 (2.50-2.57)	2.80-34.96 (2.80-2.87)
No. of reflections	15787	43072 (3225)	24476 (2115)
R (%)	19.44	18.64	17.21
R _{free} (%)	23.57	24.20	25.22
RMS deviations			
Bond length (Å)	0.009	0.011	0.014
Bond angle (°)	0.974	1.133	1.377
B factor (Å ²) mc	1.562	1.572	1.195
B factor (Å ²) sc	3.329	3.623	2.959
Protein atoms	2564	2561	2553
Ligand atoms	9	8	20
Water atoms	60	41	29
^a Data in parentheses are for highest resolution shells			
^b $R_{sym} = \frac{\sum_h \sum_i [(I_i(hkl) - \langle I(hkl) \rangle)^2]}{\sum_h \sum_i [I_i(hkl)]^2}$			



Analysis output: all-atom contacts and geometry for 1inpH.pdb

Summary statistics

All-Atom Contacts	Clashscore, all atoms:	23.46	43 rd percentile* (N=355, 2.30Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	20.06%	Goal: <1%
	Ramachandran outliers	4.27%	Goal: <0.2%
	Ramachandran favored	82.91%	Goal: >98%
	Cβ deviations >0.25Å	5	Goal: 0
	MolProbity score [^]	3.54	6 th percentile* (N=8909, 2.30Å ± 0.25Å)
	Residues with bad bonds:	1.75%	Goal: 0%
	Residues with bad angles:	10.00%	Goal: <0.1%



Analysis output: all-atom contacts and geometry for rnd01f_0.07_a1.5_d1.25_t1.75_5c_edH.pdb

Summary statistics

All-Atom Contacts	Clashscore, all atoms:	9.92	96 th percentile* (N=271, 2.50Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	1.48%	Goal: <1%
	Ramachandran outliers	0.32%	Goal: <0.2%
	Ramachandran favored	97.48%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score [^]	1.75	99 th percentile* (N=6960, 2.50Å ± 0.25Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.30%	Goal: <0.1%

Figure 2-9 - Validation Statistics for 1INP vs. Rebuilt 1INP

Figure 2-9 compares the summary statistics from MolProbity of the 1INP deposited in 1994 and the rebuilt structure. The new structure shows much improvement in the geometry and sterics.

Table 2 - MolProbity Statistics

	Native	D54A	D54A + I(1,4)P ₂
Resolution	2.23	2.5	2.8
Rotamer Out (%)	1.47	1.48	1.45
Rama Out (%)	0	0.32	0.32
Rama Fav (%)	96.53	97.48	97.46
Cbeta Dev	0	0	1
Bad Bonds (%)	0	0	0
Bad Angles (%)	0.3	0.3	0
Clashscore / Percentile	8.51 (93 rd)	9.92 (96 th)	14.08 (97 th)
MolProbity Score / Percentile	1.81 (93 rd)	1.75 (99 th)	1.89 (99 th)

The final MolProbity statistics for the three structures are shown in Table 2 (above). The clashscores and the MolProbity scores are very good for these resolutions, particularly the two structures at lower resolutions. When plotting the clashscores of these rebuilt structures vs. a PDB sample (solid diamonds in Figure 2-10 below), the improvement is that much more satisfying.

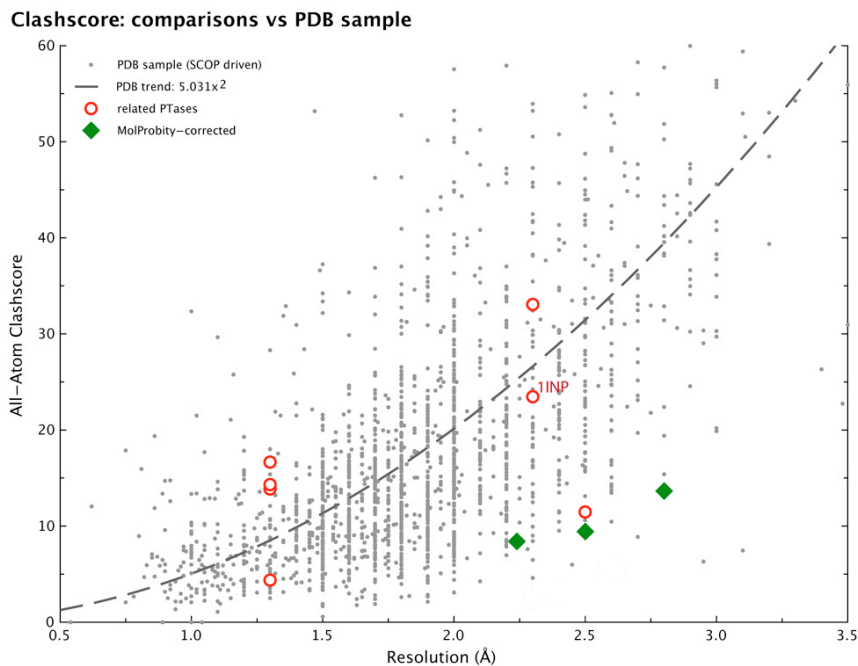


Figure 2-10 - 1Ptase Clashscores vs. PDB Sample

2.2 Structure-Determination of a plant *AtIpmk* at 2.8Å, using tools in *KiNG* and *MolProbity*

Building and refining the structure of *AtIpmk* (*Arabidopsis thaliana* Inositol phosphate multi-kinase) proved to be a challenging task for my collaborator Stuart Endo-Streeter in the York lab, due to the low resolution of the data. The level of detail visible in the electron density maps is limited, with density for side-chains difficult to interpret or unresolved. Similarly, many of the loops are without observable electron density, indicating disorder in the region.

Electron density in the core of *AtIpmk* is similar to that of the yeast homolog (*ScIpmk*, 2IEW and 2IF8, Holmes, 2006) and of the more specialized evolutionary descendent Inositol-(1,4,5)-trisphosphate 3-Kinase (IP3K, 1W2F, Gonzalez, 2004). The

core of the *ScIpmk* model provided a quick route to build the core of the *AtIpmk* model by molecular replacement. A poly-alanine model was generated from *ScIpmk* and fit into the *AtIpmk* experimental density map using the Molrep program from the CCP4 suite (CCP4, 1994) followed by a simulated annealing protocol from the Crystallographic and NMR System (CNS) program (Brunger, 1998) to refine the poly-alanine model coordinates. Next, Stuart used seleno-methionine anomalous difference density maps and the O program (Jones, 1991) to build model into all electron density interpretable at this stage. A second round of simulated annealing was done, followed by positional minimization in CNS.

Throughout the model building and refining process, experimental (F_o), simulated annealing composite-omit, $2F_o - F_c$, and difference ($F_o - F_c$) maps were used. Further model rebuilding also used KiNG and MolProbity (Davis, 2006; Davis, 2004; Lovell, 2003; Chen, 2009) to correct sidechain rotamers, improve backbone geometry using the Ramachandran analysis, and alleviate impossible steric overlaps using all-atom contacts. During this process, I consulted with Stuart on use of the tools in KiNG and MolProbity to improve geometric and steric criteria for the model, which was done in a similar way to the 1Ptase re-refinement previously described. Due to the lower resolution, this was more difficult than for 1Ptase, but also the help of these tools was more essential.

Finally, translation-libration-screw (TLS) refinement (Winn, 2001) and the REFMAC5 program (CCP4, 1994) from the CCP4 suite were used to finish refinement of the *AtIpmk* model, as determined by convergence to its final R/R_{free} values of 23.64/24.61 and modeling of all interpretable density.

The 2.8Å *AtIpmk* is the lowest-resolution independently-phased structure our lab has so far been involved with, and it showed the usefulness of the various tools in KiNG and the validation analyses in MolProbity even for a fairly large structure at relatively low resolution. The puzzle of how the *AtIpmk* active site manages multiple substrate specificity motivated the work described in the following section.

2.3 Defining the Inositol Substrate-Specificity Motif

The 6-position binary code represented by the molecular species in the inositol signaling pathway is elegant and clear. However, the basis for substrate-specificity in the multiple kinase reactions catalyzed by the Ipmk enzyme (whose crystal structure was the subject of the previous section) does not follow the same logic as the inositol ring nomenclature and had not previously been adequately understood. Therefore Stuart Endo-Streeter and I embarked on a search for an alternative logical scheme that would explain the complex pattern of specificity for IPMK substrates.

To identify a possible IP-substrate recognition motif, we compared the structures of all known substrate and non-substrate IP species. Our key hypothesis was that all inositol polyphosphate substrate species would bind with a conserved ring position and target hydroxyl orientation. The basis of this assumption is the identical nature of all the inositol phosphorylation reactions regardless of identity of the target hydroxyl: each Ipmk reaction is the addition of an ATP γ -phosphate to an unoccupied, equatorial inositol hydroxyl.

Within the single active site of Ipmk, the target inositol hydroxyl must always occupy the same position and orientation to be properly positioned for catalysis. Similarly, the positions of at least some subset of the inositol phosphate groups could be expected to bind in conserved positions and orientations within the specificity portion of the common active site, and should therefore have a fixed geometrical relationship to the kinase-target OH at which the chemistry occurs. Since the numbering of the target OH can differ from 3' to 5' to 6', the numbering of the key binding positions must also be variable. For this binding-centric analysis we designate the kinase-target OH position as κ and labeled the other five potential phosphorylation and binding positions as α through ε moving clockwise around the inositol ring from the κ -hydroxyl.

We used the canonical substrate of Ipmk, I(1,4,5)P₃, as the reference template. Additional IP species were overlaid, with their kinase-target hydroxyls modeled in the same position (κ) as the kinase-target 6'-hydroxyl of the reference template and with equivalent spatial positions for the six carbons of the chair-form inositol ring. The IP species with multiple free hydroxyls were modeled several times, such as both 3'- and 6'-target positions for I(1,4,5)P₃, and this process was performed for all known substrate and non-substrate IP species (see Figure 2-11).

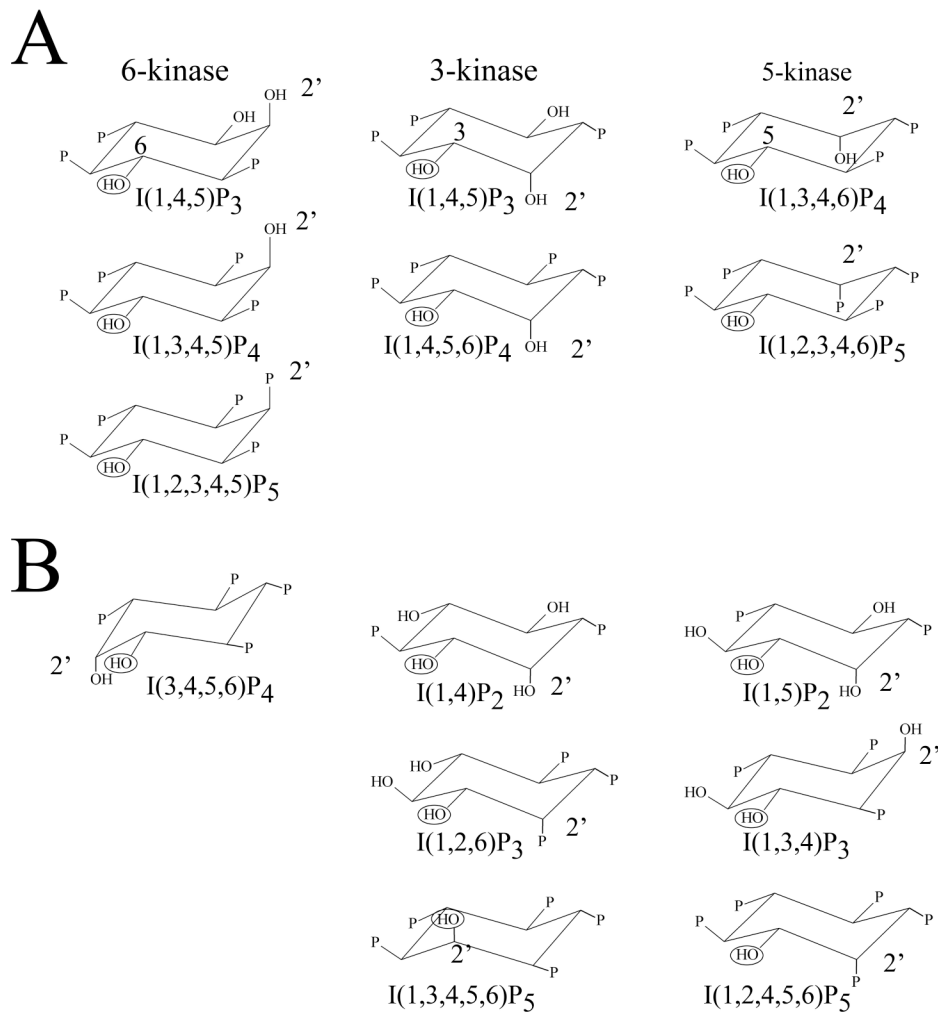


Figure 2-11 - Substrate and non-substrate inositol polyphosphate species

Figure 2-11, slightly modified from Stuart's thesis (Endo-Streeter, 2009), shows in part A the IP species known to be Ipmk substrates. As a frame of reference, inositol is drawn with the kinase-target (κ) hydroxyl circled and pointed down and toward the viewer; phosphates are labeled as P. Each species is in the column matching the kinase-target hydroxyl identity. All species in each column are shown in the same orientation and are for *AtIpmk*. In B, the IP species known not to be Ipmk substrates are shown. Non-substrate IP species are depicted with various of their hydroxyl groups aligned in the

target position. The third column includes test cases for non-substrate IP species in three different orientations with at least the 2' ring carbon numbered and the potential target hydroxyl circled and numbered.

Using this representation of the data, we soon identified some consistent characteristics of all substrate IP species. The first two positions clockwise of the kinase-target κ -hydroxyl (the α and β positions) are always phosphorylated in the known substrates, and they are always equatorial. They are predicted to have close interactions with the surrounding residues, in α and β binding-site pockets that control this specificity.

In contrast, one of those conserved α or β phosphates is often missing in non-substrates, such as for I(1,3,4)P₃ when either the 5' or 6'-hydroxyl is placed as potential target. The next clockwise position (γ), opposite to the κ -hydroxyl, appears to have no role in specificity as it can be either phosphorylated or not phosphorylated, and either equatorial or axial, in substrate and non-substrate species. It is probably solvent exposed, or at least open enough not to constitute a specific binding pocket.

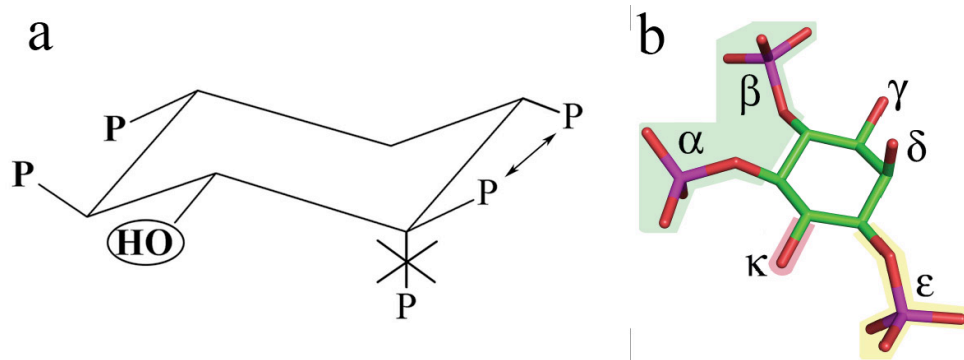


Figure 2-12 - Inositol polyphosphate species: proposed binding motif

Figure 2-12 is a diagram of the proposed inositol polyphosphate binding motif for IP species. Required phosphate groups are shown in bold in panel a, partially required ones indicated by an arrow, and exclusionary positions crossed out. The target hydroxyl is circled and phosphates are labeled as P. An equivalent schematic is shown in panel b, with the different sites labeled.

Later, Stuart refined this model further and proposed that at least one of the δ and ϵ -position pockets must also be occupied by a phosphate, probably having longer-range interactions with the surrounding residues. He noticed that no non-substrate species match both the α/β and the δ/ϵ patterns except I(1,2,4,5,6)P₅; however, that IP₅ species is unusual in having an axial 2'-phosphate group in the ϵ position (first counter-clockwise of the κ -hydroxyl in panel b above). The inference is that an axial phosphate group is prohibited in the ϵ -pocket as well as in the α and β pockets, where an axial PO₄ would probably result either in the loss of necessary interactions (α and β) or in steric conflicts (ϵ). An axial phosphate does appear to be permissible in the δ pocket, although perhaps not capable of interacting well since it occurs in the presence of a normal equatorial ϵ phosphate for I(1,2,3,4,5)P₅, where due to the ring geometry the axial δ PO₄ would point up from the plane of the inositol ring rather than down as for an axial ϵ group.

We worked on this concept at about the same time that Chang and Majerus proposed a substrate motif for *HsIpmk* (Chang, 2006). The *HsIpmk* motif proposed by Chang and Majerus works for *HsIpmk*, but not for the plant or yeast forms because it has a limited 6-kinase activity, none at all for I(1,3,4,5)P₄. They have not published how extensively they have tested the IP species for activity with *HsIpmk*, so it is unknown if

there are other IP species that may support or disprove their proposed motif. It is certainly not as specific as the final motif proposed for *AtIpmk*.

The above set of patterns taken together make up the *Ipmk* binding motif shown schematically in Figure 2-12 above. In this motif, a substrate IP species must bind in such a manner that both α and β pockets and at least one of the δ and ϵ pockets on IPMK are occupied by equatorial phosphate groups. Axial phosphates are prohibited except in the γ or δ positions. We propose that this motif corresponds to the requirements of the phosphate-binding pockets and surrounding residues in the active site of IPMK. It differs from standard ring numbering in classifying the six ring positions based on relationship to the active site (that is, to the target OH), and thereby succeeds in accounting for all the experimental data identifying *Ipmk* substrates and non-substrates.

2.4 Conclusions

From this practical experience of digging down into the details of crystallographic structure determination and improving the model accuracy using KiNG and MolProbity, I have not only learned those tools but have also developed insights about which types of analyses to test out on NMR structures. Along the way, a new biological understanding emerged to explain the structural basis behind the diverse but specific substrate recognition by *Ipmk* inositol kinase.

3. New Tools for Analysis & Error Diagnosis in NMR Structures

I implemented, deployed, and evaluated the use of all local criteria (all-atom clashes, underpacking, Ramachandran and rotamer outliers, bond angle distortions, constraint violations) across the models in an ensemble to diagnose groups in the wrong local-minimum conformation and to propose corrections.

I implemented in MolProbity the ability to create a multiple model, multiple criterion, kinemage display. NMR structures, commonly deposited as an ensemble of models present a unique challenge in analyzing and evaluating the structure models. Crystallographic structures are mainly deposited as a single structure model. The MolProbity system, initially created to analyze x-ray structures, thus needed to be extended.

3.1 NMR Diagnostic Visualizations and related Error Analysis

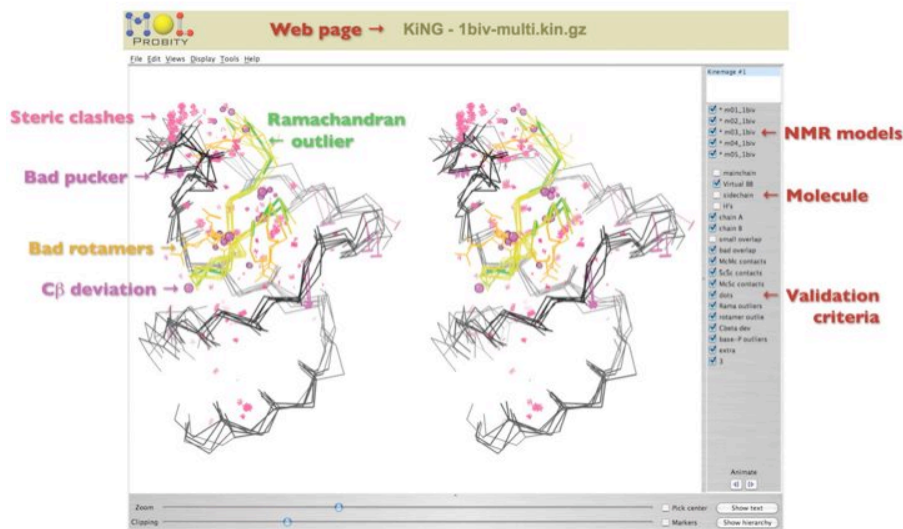
3.1.1 Multi-Model Multi-Criterion Kinemages

First, I wrote a utility for parsing out the many models of an NMR structure ensemble within MolProbity. The command line utility, written in PHP, is called by the MolProbity system and creates a separate file for each model of the structure ensemble (including the header information). Important for the system, the model record is modified such that all the other software will never ‘see’ it. This is necessary because a

number of structure analysis software tools only work on model 1, requiring this modification to analyze all models of the ensemble.

Next, I started with the multicriterion kinemage code Ian Davis wrote for the analysis of x-ray structures and devised a kinemage layout where each model within an NMR ensemble is assigned to a “group” (top level of the kinemage display-object hierarchy) that is animatable. For each individual group to be written, MolProbity calls the appropriate analysis functions (Ramachandran, rotamer, clashes, etc.) and performs the analyses on the appropriate single-model-file separated out from the multi-model PDB. The return is appended through a PHP script I wrote that iterates through the array of single-model-files. The resulting kinemage is a multi-model multi-criterion kinemage.

Multi-Model Multi-Criterion Kinemage



Ian W. Davis, Andrew Leaver-Fay, Vincent B. Chen, Jeremy N. Block, Gary J. Kapral, Xueyi Wang, Laura W. Murray, W. Bryan Arendall III, Jack Snoeyink, Jane S. Richardson, and David C. Richardson
MolProbity: all-atom contacts and structure validation for proteins and nucleic acids
Nucleic Acids Research Advance Access published online on April 22, 2007

Figure 3-1 - A Multi-Model Multi-Criterion Kinemage for an NMR Ensemble

Shown is a screenshot of a Multi-Model Multi-Criterion Kinemage, for the 1BIV RNA/protein complex displayed in stereo in KiNG (Davis, 2007). All the models of the ensemble appear as animatable groups. The various validation criteria are mapped onto the structure. Except for Ramachandran outliers, the color scheme shows warmer colors for outliers (such as hot pink steric clashes, gold sidechain rotamer outliers, and magenta ribose pucker or C- β deviations). Options are included in the MolProbity system to display ribbons, as well as vdW contacts and Hydrogen bonds. Subsets of the PROBE output also have their own master controls, allowing the user to display all or any subset of the mainchain-mainchain, mainchain-sidechain, sidechain-sidechain, or het-group contacts.

Critically different from X-ray structures, NMR spectroscopy derives data primarily from, and explicitly models, the hydrogens. Therefore, the NMR analyses require a different usage of REDUCE for adding and optimizing hydrogens. In all cases the NMR models are evaluated by only adding those Hydrogens the spectroscopist did not place themselves, and optimizing the ones modeled in order to standardize their bond lengths. Another critically important component is how the PDB format is used for NMR structure ensembles. One field of the format is reserved for occupancy values in the case of x-ray structures, and most of the Richardson lab software ignores atoms with occupancy <0.02. For NMR structure ensembles, that field can be used for anything the spectroscopist chooses. In order to effectively run PROBE on an NMR structure model, the appropriate flag must be set to assume an occupancy of 1.0 rather than read the value.

If this is not done, a structure model may give an erroneous output, such as failing to show steric clashes.

3.1.2 Multi-Model Multi-Criterion Chart

I built a chart companion to the multi-model multi-criterion kinemage that gives a visual representation in sequential order, across all models within an ensemble, of different local criteria. The three criteria chosen for representation are the presence of a clash, presence of a Ramachandran outlier, and presence of a rotamer outlier for a given residue. They are displayed as a 2D kinemage. This dense data visualization was very difficult to make satisfactorily user-friendly.

Working with Ian Davis, the original MolProbity developer, I decided to encode each criterion with the same color used in the multi-crit kinemage, for consistency. In order to accommodate an unknown number of models within a structure ensemble, the models are plotted as columns, with the sequence identity on the right. The visual representation selected made the density of the criteria higher to save space, and to make the visual effect more substantial when multiple local errors appear at the same residue within the same model or across multiple models. In the current implementation (shown below), the presence of errors at the same residue in multiple models of the structure ensemble appears as a smear of hot colors across a row, with blotches of greater intensity where multiple types of errors exist.

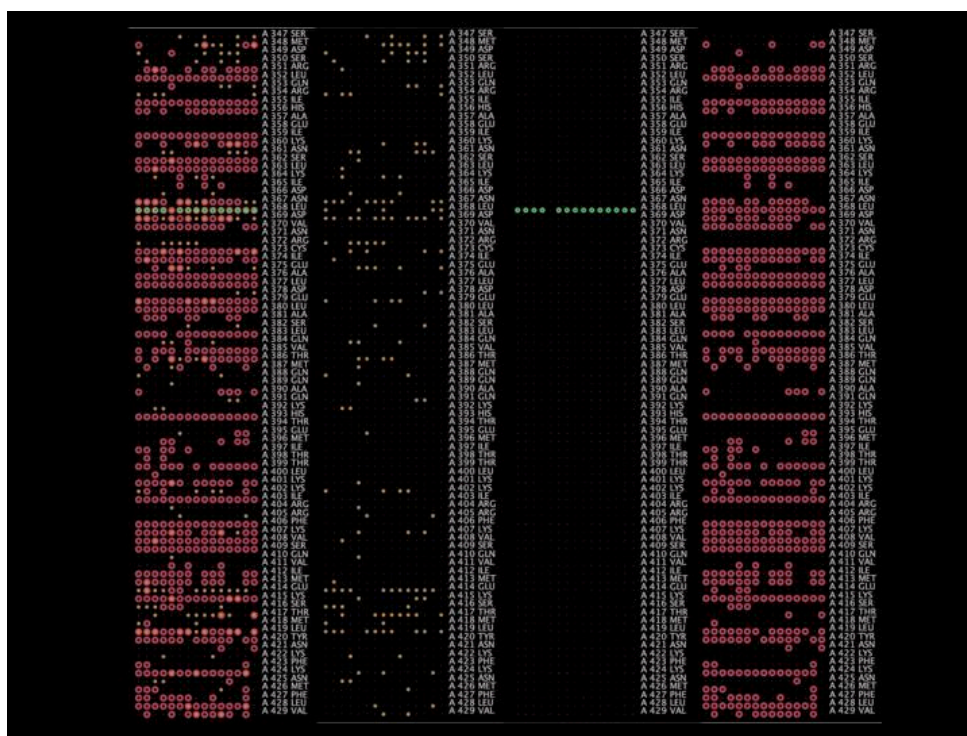


Figure 3-2 - Multi-Model Multi-Criterion Chart for NMR Structures

In the figure, the far right panel shows just the steric clashes (red circles) found in each residue of the models in the ensemble, where the models are arrayed as columns. Next to the clash representation are the Ramachandran outliers (in green). As expected, there are far fewer Ramachandran outliers (here there is only one residue where one occurs, but it happens to be modeled the same way on most of the models within the ensemble). To the left of this is the mapping of the rotamer outliers (as gold dots), and at the far left is a panel with all three indicators turned on at once. Each of these criteria appears as a clickable master button in the kinemage and can be toggled on or off. This visualization enables more detailed and close-up interrogation for pattern identification of local areas in the sequence where clusters of errors of different types appear.

3.1.3 NOEDisplay

In generalizing the MolProbity-based methods to improve accuracy for NMR ensembles, most of the validation measures and refitting tools are directly applicable, with NOE data and other experimental restraints replacing the match to electron density. However, types of systematic errors are very different for NMR (for example, there are no effects which produce 180° amide flips), and strategies for building corrected information into structure-determination processes are different because of experimental data content and structure-solution strategy for NMR vs. X-ray methodology.

My work toward defining procedures to propose corrections in NMR structures started with analysis of individual NMR structures from the PDB and from my collaborators, and with modifications to our software needed for these purposes.

Early work by Brian Coggins in the Zhou laboratory at Duke to his development of the NOEDisplay software as a plug-in to the existing visualization framework of KiNG (Chen, 2009). This gives continuously updated monitoring of NOE model-to-data agreement, along with the steric and geometric measures that can be updated interactively in KiNG, during model rebuilding.

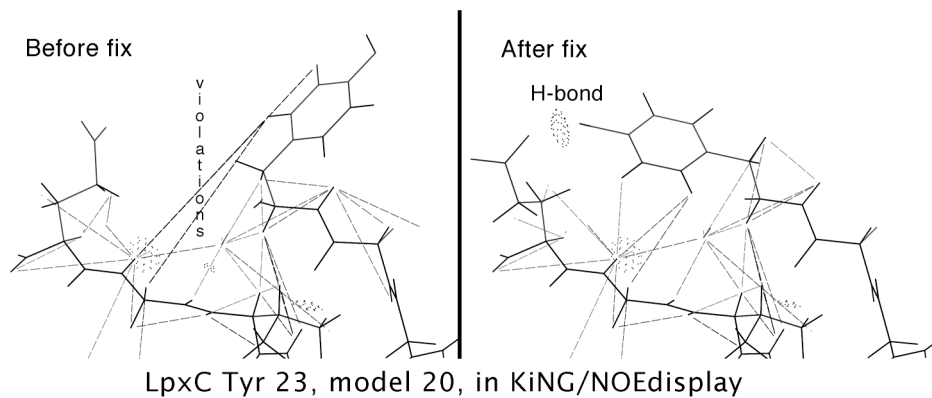


Figure 3-3 - NOEDisplay Example on LpxC Structure

The figure pair shows NOEDisplay being used in KiNG to aid in correcting a misplaced Tyr sidechain in an earlier version of the LpxC structure (1XXE, Coggins, 2005). The dashed lines are the NOE's connecting pairs of assigned hydrogens. The points where an NOE terminates away from any atom are "pseudoatom" positions often defined by structure determination packages (for instance in the center of a methyl). In the image on the left, there are two NOE outliers that do not satisfy the observed distance restraints in the experimental data. Using the sidechain rotation tool in KiNG, a different Tyr rotamer is selected that eliminates this mismatch of model to data, brings the NOE's into agreement, and creates a hydrogen bond with a neighboring residue (shown by the dotted lens shape of PROBE dots).

3.1.4 All-Atom Contact Analysis of Under-Packing in Protein Models

In x-ray crystallography, the Fourier transform of the diffraction pattern places the electron density in a correctly scaled 3D space (Blundell, 1976), providing a molecular envelope that cannot expand or contract significantly. Since NOE distance measurements reach no further than about 5Å and their quantitative scaling can be distorted by experimental artifacts, the dense network of observed NOE data determines local 3D structure very effectively in NMR but is not accurate for overall shape or scale (Cavanagh, 2006).

In contrast to a specific NOE, a residual dipolar coupling (RDC) measurement gives general orientation information that goes a long way toward resolving the overall shape problem, but does not address the scale problem (Blackledge, 2005). A radius-of-

gyration target is sometimes used in NMR to provide better long-range compactness and correct for the overwhelming repulsive terms of the force fields (Huang and Powers 2001). However, it is difficult to tune the overall effect correctly.

The fact that NMR measurements do not provide an envelope or overall scaling makes it unsurprising to find some NMR structures that appear to be expanded in relation to x-ray structures of the same protein. For the global case of expansion vs. contraction across entire NMR structures, evaluation and potential correction is definitely of interest. However, the reason underpacking analysis is crucial to my study of NMR structure improvement is that local expansion in NMR models can destroy the sensitivity, and even the ability, of all-atom clashes to identify incorrect conformations.

I approached the issue of under-packing using all-atom contact analysis because it is very sensitive to local structure as well as global properties, because its formalisms are the ones used for the best treatment of over-packing, and because the programs are available and modifiable within the Richardson lab.

Briefly, the general methodology of all-atom contact analysis has the hydrogens added and optimized by REDUCE (Word, 1999b) and the contacts calculated by PROBE (Word, 1999a). Hot pink spikes show physically unrealistic steric overlaps $>0.4\text{\AA}$. The clashscore is the number of such serious clashes per 1000 atoms. Favorable overlap of suitable H-bond donor and acceptor atoms is represented as lenses of pale green dots. PROBE calculates several numerical scores suitable for x-ray model validation, but did not previously evaluate under-packing.

From the NMR structures with coordinates and data in the PDB, I chose several small comparison sets for initial study, covering a wide range of structure types, qualities, and experimental and calculation methods. A particularly interesting set of three examples were *Lactobacillus casei* dihydrofolate reductases (DHFR's), both the NMR structures 1AO8 (Gargaro, 1998) and 1LUD (Polshakov, 2002) and the 1.7 Å resolution crystal structure 3DFR (Bolin, 1982). These are interesting because they are determined by three different groups, and two of them are NMR structures and one is an x-ray structure. Also, DHFR is a long-studied and important cancer target. The chemotherapeutic methotrexate is modeled in one of the NMR structures and the antibiotic trimethoprim is modeled in the other.

A set of ubiquitin structures include 1D3Z from the Bax lab (Cornilescu, 1998), which scores very well on existing quality measures and has unusually complete data. The other extreme of quality is represented by 1J1H (Someya, 2003) and 1JRM (Yee, 2002), which are unusually poor-scoring examples of different molecules from two different structural genomics centers. Another set were from our collaborators in the Montelione lab, including a series done on the Z domain of Staphylococcal protein A: 1SPZ, 2SPZ (Tashiro, 1997), and 1Q2N (Zheng, 2004), with increasingly more data and refinement. I scored these test structures in MOLPROBITY and analyzed the results by hand, tabulating the raw numbers of PROBE dots for each of five overlap ranges for every model in the ensemble and seeking correlations with data quality, model quality, and visual evaluation of the 3D ensembles.

I found the PROBE numerical clashscore and the hot-pink spike visualization of bad clashes behaved suitably for these NMR structures and in some instances are indicators by themselves that flag real problems. However, a measure of under-packing is necessary in order to tell if the clashscore is reasonable. Extreme cases, such as 1J1H, are very obvious visually and can have overall clashscores >200; these are structures with relatively few NMR measurements per residue and/or only partial refinement.

For comparison, a typical crystal structure has a clashscore of 45 at 3Å resolution, 20 at 2Å, and only 5 at 1Å (Arendall, 2005). Some NMR structures with excellent data quality (e.g., 1D3Z) have clashscores <10 and very few rotamer or Ramachandran outliers. However, not all NMR structures with low clashscores show high enough quality by other NMR and geometrical measures to inspire confidence in their accuracy; they may instead escape clashes by being underpacked.

Wide contacts (up to 0.5Å between surfaces) and close contacts (0 to 0.2 Å gap) together represent attractive van der Waals interactions. However, what I concluded from these preliminary studies is that neither their raw dot count as an area (dot density is typically $16/\text{Å}^2$) nor the exponentially-weighted value of their term in the PROBE score (Word, 1999a) seems to reflect looseness of packing in an appropriate way for NMR structures.

In contrast, the count per residue of backbone H-bond dots (which reflects both number and strength of H-bonds) was very promising and my focus shifted to investigating them. For instance, among the three DHFR structures, the backbone H-bond dot count correlates inversely with data quality measures (such as constraints per

residue or additional data types) and with small but perceptible differences in structure expansion. The figure below compares beta-sheet regions from DHFR structures. The H-bond dot lenses in the NMR model are thinner and fewer than those in the x-ray structure, and their size distribution is much broader. The backbone H-bond dot counts are 395 in the 1AO8 NMR structure shown at left vs. 1150 in the 3DFR x-ray model for the same local region of sheet. The other NMR model, 1LUD (not illustrated), with more data and better refined than 1AO8 has a count of 871 H-bond dots, intermediate between the 395 in 1AO8 and the 1150 in the 1.7Å x-ray structure. For these and other examples, the count of mainchain all-atom H-bond dots appears to be a suitable and sensitive measure of structure expansion.

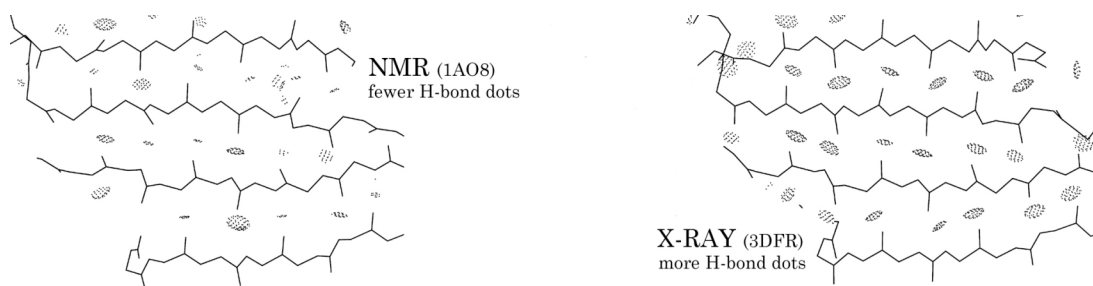


Figure 3-4 - Dihydrofolate Reductase Structures by NMR & Xray

The small subset of structures I analyzed by hand demonstrated a promising direction; the challenge then became re-framing the question and implementing a more robust approach to analyze larger, more representative datasets. To facilitate this, my advisor modified the PROBE software to allow for routine separation of dot types into additional classifications (based on backbone-backbone, sidechain-sidechain, and

backbone-sidechain contact types). This enables quantification of the H-bonding shown above as well as other analyses.

I wrote scripts for quantifying packing scores based on PROBE output data, taking advantage of the additional selections from newer versions of PROBE. These measures are implemented in MOLPROBITY (Davis, 2004; Davis, 2007), though they must be accessed through the command line and are not served up as analyses on the public web site.

In order to have a baseline to compare, I performed a first-pass analysis on the Top500 dataset from the Richardson lab (Lovell, 2003). This dataset contains non-homologous crystal structures from the PDB at $\leq 1.8\text{\AA}$ resolution that have acceptable clashscore and R-factor, and few extreme geometric outliers. Priority for inclusion was given to the highest resolution example available and to wild-type structures over mutants. Specifically disallowed structures included: unrefined structures, free-atom refined structures, structures with no B's or unrefined B's, and structures without specified sequences. All MOLPROBITY statistics (clash, Ramachandran, and rotamer), as well as dihedral angles, B factors, and secondary structure from DSSP for every residue of every structure in the Top500 are stored in a MySQL database.

For this Top500 data, I made nine plots to investigate overall hydrogen bonding trends. For each interaction type (mc-mc, mc-sc, sc-sc) and each type of secondary structure (alpha, beta, and loop) I plotted the average number of hydrogen bonding dots per residue in each protein as a function of the percentage of that secondary structure type in the model. Despite the coarseness of this level of data mining, a few strong trends

appeared. The figure below shows that protein structures containing increasing amounts of loop also have less backbone-backbone hydrogen bonding and more backbone-sidechain hydrogen bonding. This trend has been noted anecdotally (Richardson, 1981; Tainer, 1982), but is here shown clearly and in quantitative form.

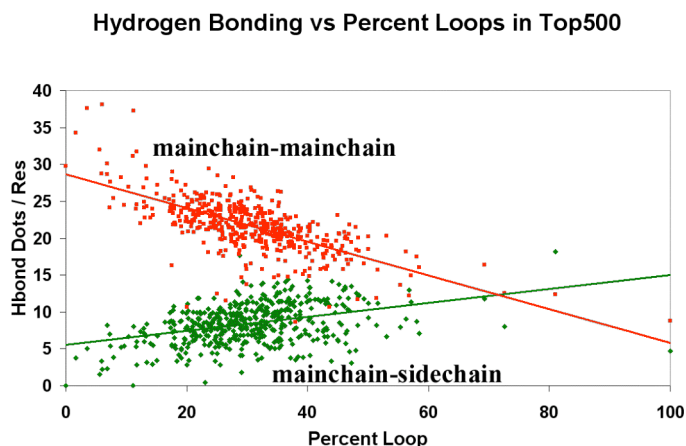


Figure 3-5 - Hbonding dot counts vs. Percentage Loop

An equivalent investigation for NMR structures lagged behind the x-ray investigation because the database with all evaluations performed needed first to be created for the NMR structures. Software existed to calculate most of the quantities needed, but much parsing and scripting was required to generate and add database information for all models of an ensemble.

In parallel with the work described above, a complementary analysis of underpacking was developed by Andrew Ban during his graduate work using the BioGeometry-based methods (Ban, 2006). We worked together to decide on selection criteria for the datasets (such as subsets from DrESS – Spronk, 2004 – and Top500 from the Richardson lab) used in his local density study. This was done such that there would

be a jointly informed dataset that would allow for meaningful comparisons, evaluation of the strengths and weakness of the two techniques, and a later decision on how best to combine them. His study confirmed that NMR structures with more complete data and/or refinement better match high-resolution x-ray structures in overall packing density.

The BioGeometry tools analyze packing in terms of atom-associated volumes, while the all-atom tools analyze packing along a sequence and across 3D space. The all-atom underpacking only scores polar atoms, while the BioGeometry version only scores buried atoms.

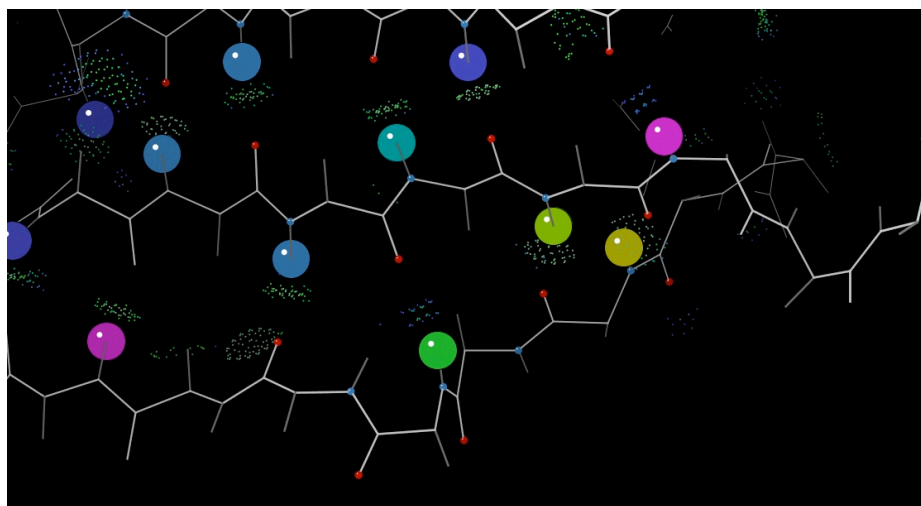


Figure 3-6 - DHFR Structure 1AO8 with Hbond dots & BioGeometry Analysis

The figure shows both the local density visualization from BioGeometry (the balls colored by local density Z-score) and the all-atom contact analysis of mainchain-mainchain hydrogen bonding for a section of the previously described DHFR structure. Interestingly, the PROBE analysis shows a bifurcated hydrogen bond (on both sides of the green ball, below center), which can only be seen by the wide contacts in this clearly underpacked portion of sheet (and only one dot is observed). The BioGeometry analysis

did not pick this up at all. In places where the hydrogen bonding on the mainchain seems reasonable, the PROBE and BioGeometry analyses seem to be in agreement. At the surface (not shown), the PROBE analysis behaves much more sensibly, as it does not run into a significant limitation of the Voronoi partition; namely that BioGeometry only applies to closed partitions of space, making the scores unavailable at the surface.

In summary, from these two studies of protein underpacking, it seems likely that the new all-atom H-bond measures and the BioGeometry volume measures, taken together, provide a good first-step assessment of local underpacking.

3.2 NMR Database Analyses, and Specific Examples

3.2.1 NMR Structure Database Creation

Some of the data for an in depth-study of local criteria to evaluate NMR structure models were in the pre-existing PDB coordinate and restraint files, but needed to be parsed and organized for this purpose. The rest I generated using many different pieces of software, resulting in thousands of diverse pieces of information on each of the thousands of individual structure models. For example, it took nearly three days to analyze all of the roughly 6,600 individual structure models by running a combined analysis script I wrote in MolProbity for outputting the various Ramachandran, rotamer, and all atom contact analysis results.

For earlier studies in the lab, preliminary explorations have usually been done in spreadsheet form, but to support the main research a very capable centralized

organization for data management and analysis is essential. This project reached that point once I decided to analyze thousands of models. My strategy was to create an extensive MySQL database, in order to manage and mine the data needed.

MySQL databases are powerful, open-source relational databases available for Linux, Mac OSX, and PC platforms and were already in use within the lab. Relational databases hold all of their data in tables. All the operations on the database are done on the tables themselves, and the results are output into tables. The ability to create linkages between tables in the database allows important flexibility in data organization and a robust method of mining that data.

A central strength of MySQL is the use of SQL (Structured Query Language) to create tables and linkages and then to investigate the data relationships. MySQL is extensively used world-wide, and its development has advanced substantially since the founders at the MySQL AB company redid the original mSQL. The developers of MySQL provide it as an open-source distribution, with frequently-updated online documentation and manuals.

The most common implementations of MySQL databases are those interfaced by creating pages that are accessible on web-browsers. The MySQL infrastructure here is interfaced with the programming language PHP for web use in what is normally termed the 'Dynamic Duo.' There are a number of open-source implementations of PHP interfaces to MySQL.

The primary worldwide database of macromolecular structure coordinates is the Protein Data Bank (Berman, 2000), containing over 65,000 total entries and about 5000

protein NMR structures. Additional restraint files for NMR structures are also available through the Biological Magnetic Resonance Bank (BMRB; (Seavey, 1991)). Using the advanced search tools available at the new PDB website (<http://www.rcsb.org>) a great deal of filtering based on many different criteria can be accomplished online.

Other published datasets of NMR structures also influenced this project. DrESS (Nabuurs, 2004)) and RECOORD (Nederveen, 2005)) contain before-and-after datasets of re-refined NMR structures. An unusually high-quality ensemble for ubiquitin (1XQQ) obtained by refining several entire 16-member ensembles, rather than each model, against the data including order parameters was also looked at (Lindorff-Larsen, 2005). The Richardson lab also has several later ensembles produced by Lindorff-Larsen with different experimental data and variants of the refinement methods. An analysis of the structures, while very interesting, did not yield very much in terms of insight beyond the conclusion that even within this quite dynamic ensemble the geometry seemed relatively well-behaved when compared to the original 1D3Z structure determined by Bax, and both ensembles are very accurate. Analyzing 128 model ensembles made a very useful driving problem for the development of user-friendly multi-criterion kinemages.

Choosing MySQL as the database and PHP as the script-language interface to a web-browser sets the framework. Populating the data tables in the database and querying the database provides the specific details. I used standard tools in PHP, for creating and extracting data about the NMR structures and populating the tables of the database. This was accomplished through the open source front-end for MySQL called PHP MyAdmin.

Three levels of database tables were intended: File Level Tables, Residue Level Tables, and Atom Level Tables. These are described below, as well as some of the data sources needed to populate them. Ultimately, I only created the file level tables.

The file level tables denote the presence or absence of different types of experimental data and global validation scores from MOLPROBITY. Methodological information (such as experimental conditions or refinement software) is available in the headers of '.pdb' format coordinate files. However, the tools available from the PDB to extract this type of information are designed for the mmCIF format files, which have the additional advantages of an organization deliberately similar to that of relational databases and of having the most complete and updated header information. Therefore, using the pre-existing PDB tools, I worked with Ian Davis in developing scripts to populate some of the file level tables with data, mostly from the PDB format files.

All existing Richardson lab software operates on coordinate information from PDB-format files and makes very little use of header information; only KiNG has been modified to read mmCIF input. Therefore, the all-atom contact and geometrical information for the database came from PDB-format files. Our source directories of suitably selected and cleaned coordinate files (for example, with duplicate chains or nucleic acids deleted) is in PDB format. Later additions at the file level could include biological data (e.g., source organism or disease relevance), solution or crystallization conditions (increasingly now being gathered in the TargetDB and PepCDB set up by the Protein Structure Initiative), or fold classification from SCOP (Murzin, 1995).

While I did not build the residue or atom level tables, it is useful to describe how they could be done and some important components for them. The centrally important residue level data tables would be similar to the ones already implemented in our Top500 MySQL database. DANGLE can provide all dihedral-angle information at the residue level for all models and can flag bond length and bond angle outliers. MOLPROBITY output can fill tables for clashes, rotamer scores, H-bond scores, and Ramachandran scores at the residue level. DSSP can provide the secondary structure classification for each residue of each model. QUEEN and the BMRB can be used to fill in data about number and type of NMR restraints for each residue collected by the spectroscopist and assigned.

The new tools I created as part of this project can serve to populate many of the atom level tables. PROBE, and Gary Kapral's scripts for parsing its 'unformatted' output, can be used to extract input data for atom level tables with counts of different all-atom contact dot types, specifically including clashes and the H-bonds used for all-atom packing analysis. Atomic packing-density scores from the BioGeometry project can populate tables for the relevant atom types. A combination of QUEEN and the BMRB can be used to create tables of NMR restraints assigned to each atom (including experiment type and a flag for ambiguity). Many atom-level properties are pairwise interactions, in which case the identity of the other atom can also be stored.

The final database I created, with just the file level tables, was used to do a number of coarse analyses on a dataset of 339 NMR structure ensembles – the same as used in Ban, 2006 –where NOE data is present and deposited in the PDB.

3.2.2 Use of NOE Restraints per ordered residue as a data quality measure

In crystallographic structures, resolution has been used for over 50 years as an indicator of data quality. All structure quality measures, such as Ramachandran outliers, rotamer outliers, and clashscore are plotted against crystallographic resolution. A single data quality measure for NMR structures is not quite so simple to identify.

In my initial study of structures I spent a great deal of time looking at ‘.mr’ files of deposited restraints for NMR structures. Restraints for NOE’s, J-couplings, RDC’s, H-bonds, and various other lesser-used restraints are included in this file. There is no formal structure for a ‘.mr’ file, which makes parsing them incredibly difficult. However, my experience looking at NOE’s modeled onto a structure using NOEDisplay convinced me that both at a local level, and across the whole structure model, the number of experimentally observed restraints per ordered residue is a useful and meaningful measure of data quality for NMR structures.

In no database or format is there a requirement to deposit the number of restraints per residue for an NMR structure. In order to determine this, I used the BMRB to obtain the NOE restraint files for the NMR structures. The files chosen were specifically selected for structures with deposited NOE’s. I only used NOE’s of roughly $<6.0\text{\AA}$ and not longer range NOE’s such as those observed in paramagnetic relaxation experiments (Clare, 2002; Kosen, 1989).

These files were populated into a MySQL database, and associated with their appropriate PDB structure. Next, I extracted the number of residues from the PDB files

through the PHP MyAdmin interface. Then, I used SQL queries to tabulate the number of NOE restraints in each file and computed the number of NOE restraints per residue. It is these data and files that the large analyses of local criteria were performed on, allowing the plots that are described in the next sections to be created.

While I did not perform a more sophisticated study that included tabulating other types of restraints (RDC's, j-couplings, etc.), attempting to interpret the relative contributions of the restraints to a data quality measure; I conclude that even without those analyses, the number of NOE restraints per ordered residue seems specific enough and a compelling indicator representing a critical first step in defining a data quality measure for NMR structures. Further work in this area could include looking at structures determined by both NMR and X-ray and comparing the behavior of restraints-per-residue versus crystallographic resolution. Evaluative criteria of geometric and steric outliers along with other structure validation criteria would serve as intermediaries to help estimate how commensurate the cross-validated structures are with one another.

3.2.3 Ramachandran Analysis for NMR Ensembles

Using the data in the NMR structure database, I analyzed all the Ramachandran outliers for each model in each ensemble of structures. I performed the analysis as previously described, using MolProbity, then plotted the percentage of residues with Ramachandran outliers in a given structure model vs. the number of NOE restraints per residue for the structure determinations based on 2D ϕ, ψ plots for general case, Gly, Pro, and prePro residues where the "Outlier" contour encloses 99.95% of high-quality data for

general, 99.8% for specific cases (Lovell, 2003). I chose percentage of Ramachandran outliers in order to make the structures of differing lengths comparable to one another. It is also a useful metric for comparing against X-ray structures, where the ideal is at the 0.01% level for high-quality data.

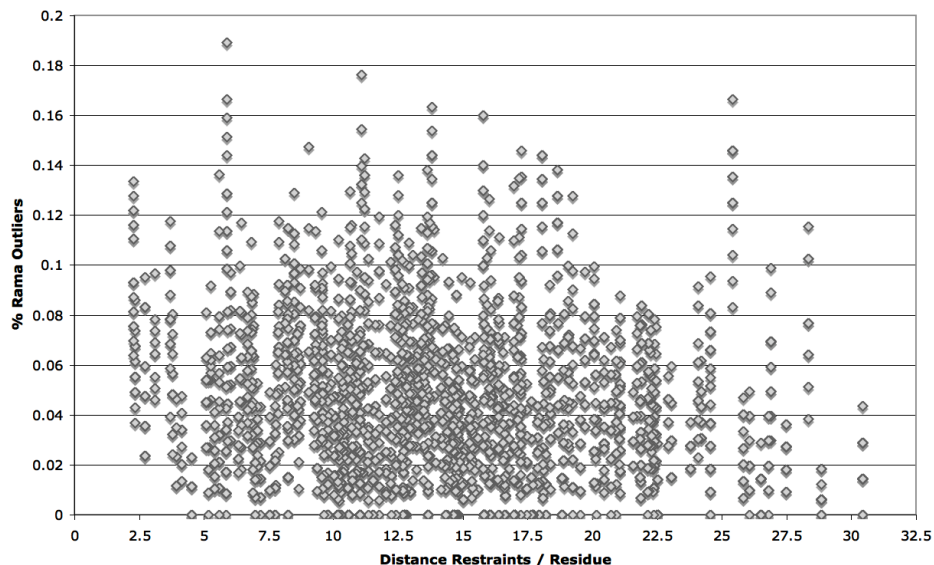


Figure 3-7 - Ramachandran Outliers vs. Restraints Per Residue in NMR Structures

This figure shows that most of the NMR structures analyzed have percentage of outliers significantly higher than X-ray structures. There are limitations to these data. Since there are multiple models in a given ensemble, I decided not to take an average of the numbers within the ensemble or just use one of the models. Each model in the ensemble of models a spectroscopist deposits to the PDB is supposed to represent a plausible model based on the data. Therefore each model should be evaluated, as done here.

Another limitation of this plot is that it does not identify model 1 of the NMR structure ensemble. It would be useful to know whether most model 1's selected are better than the rest of their ensemble counterparts. The reason this is useful is to better understand the realism of variation within an ensemble of structure models determined by the software packages. The percentage of rotamer outliers does decrease with more restraints-per-residue, although this has a rather low slope and correlation coefficient.

Because outliers in backbone Ramachandran are useful, the MolProbity site now has an output in the NMR section that creates a multi-model Ramachandran analysis that is served to the user as a PDF. Like its sister on the x-ray side of MolProbity, it has different plots for pre-Pro, Pro, Gly, and all of the rest combined.

3.2.4 Sidechain Rotamer Analysis for NMR Ensembles

Similar to the analysis of Ramachandran outliers for each model in a NMR structure ensemble, I plotted the percentage of rotamer outliers in a structure model versus the number of NOE restraints per residue in the model. Data from the MolProbity analysis are based on updated multidimensional χ angle distributions (Lovell, 2000; Chen, 2010). A rotamer “outlier” is outside the contour that encloses 99% of the high-quality data.

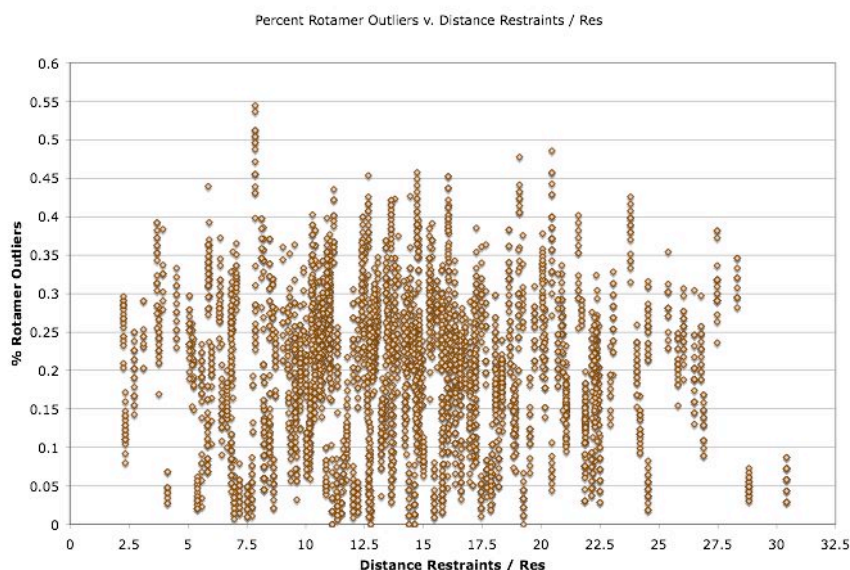


Figure 3-8 - Rotamer Outliers vs. Restraints Per Residue in NMR Structures

The figure shows no observable trend. However, these data support a hypothesis that sidechain rotamers are not being adequately addressed in NMR structure determinations. A majority of the structures have greater than 15% of their sidechains modeled as rotamer outliers.

In order to better understand and make a rotamer analysis useful for NMR, a few things need to be developed in the field. First, the community needs to perform stereospecific assignment of protein sidechains in order to increase the amount of assignable data on sidechains. Second, the structure determination packages need to incorporate sidechain rotamers routinely and in a sensible manner. With the current data, it is not feasible to identify which PDB structures have stereospecific assignments to sidechains, and which structures used rotamer terms in the determination software.

I also developed the NMR rotamer analysis, which was used by the laboratory as

part of the assessment of CASP (Critical Assessment of Structure Prediction) models (Keedy, 2009). Rotamer correctness (*corRot*) was defined as the match of valid rotamer names between model and target (Keedy, 2009). For X-ray targets, the target rotamer set consisted of all residues for which a valid rotamer name could be assigned (i.e. not < 1% rotamer score and not undefined because of missing atoms). For NMR targets, a target rotamer was defined only at those residues for which one named rotamer comprised a specified percentage (85, 70, 55, and 40% for sidechains with one, two, three, and four chi angles respectively) of the ensemble. Consideration was given to requiring a sufficient number NOE restraints in a residue for it to be included, but a conclusion was reached that in practice this would be largely redundant with the simpler consensus criterion. Considering the uncorrelated plot in Figure 3-8, the analysis of sidechain rotamers for NMR targets in CASP unfortunately compared predicted models to target NMR structures using one of the weakest feature in these targets.

Using rotamer names based on multidimensional distributions rather than simple agreement of individual chi1, or chi1 and chi2, values has the advantage of favoring predictions in real local-minimum conformations and with good placement of the functional sidechain ends. However, a disadvantage is that matching is all-or-none; for example, model rotamers tttm and mmmm would be equally “wrong” matches to a target rotamer tttt in our formulation, meaning the *corRot* score is more stringent for long sidechains. This motivated the relaxation of the previously listed % criteria as a function of sidechain length.

3.2.5 Clashscore combined with H-bond Scores as an NMR structure quality factor

In X-ray structures, the PROBE clashscore is routinely used as a structure quality measure (Arendall, 2005). I plotted this for the NMR structure models in the database against the number of NOE restraints per residue. For comparison purposes, in Xray structures the average clashscore at 2Å is nearly 20, around 35 at 3Å and 70 at 4Å.

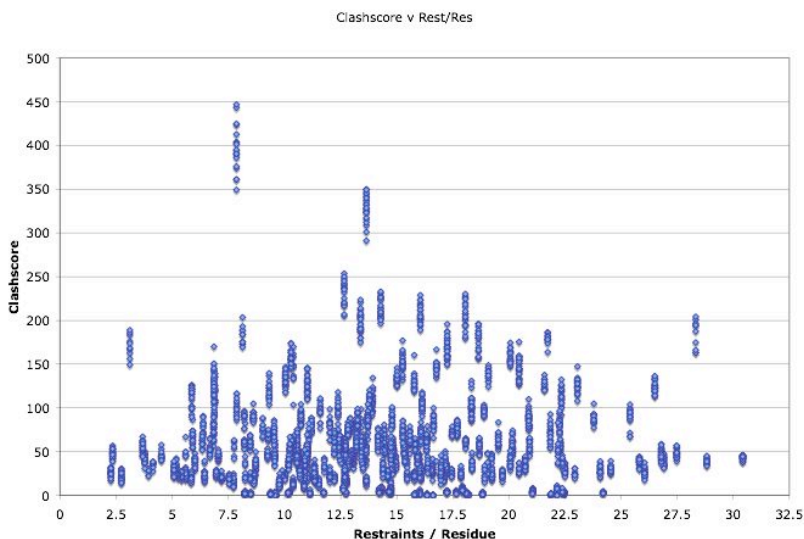


Figure 3-9 - Clashscore vs. Restraints Per Residue in NMR Structures

These data provide evidence that NMR structures have more steric clashes modeled into them than X-ray structures in spite of their explicitly modeled H's.

Based on the plot, I hypothesize that when a low number of restraints per residue is present, a low Clashscore might be warranted, but the structure is probably underpacked. I also hypothesize that at a high number of restraints per residue, a low Clashscore is more likely warranted, but might still be underpacked. In order to test this

hypothesis, I calculated the number of mainchain-mainchain hydrogen bond dots per 1,000 atoms in the structure: a 'Mainchain H-bond Score' similar to the Clashscore, which is the number of clashes per 1,000 atoms. I then plotted this against the number of NOE restraints per residue.

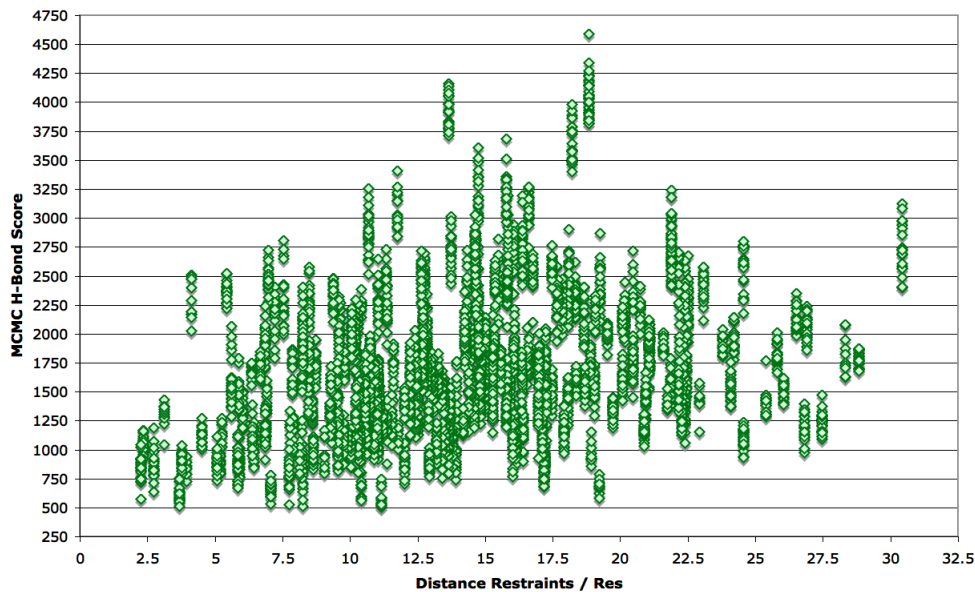


Figure 3-10 - Mainchain Hbonding vs. Restraints Per Residue in NMR Structures

The figure shows that mainchain H-bond Score increases as the number of NOE restraints per residue increases. These data, in combination with the Clashscore data suggest that at a low number of NOE restraints per residue, a low Clashscore, and a low Mainchain H-bond Score flag a structure model for potentially being underpacked. Similarly, at a high number of NOE restraints per residue; a low Clashscore and a low Mainchain H-bond Score flag a structure model for potentially being underpacked. Unlike the situation where little data is present, a high Mainchain H-bond Score and a

low Clashscore when a large number of NOE's are present indicates the structure is higher quality.

Recognizing that for underpacked structures it may be necessary to count interactions at somewhat larger separation distances, I hypothesized that the sum of the mainchain-mainchain H-bonds and the wide and close mainchain-mainchain vdW dots together would increase as the number of NOE restraints per residue increased. I calculated these values as a score per 1,000 atoms in the structure model as a Favorable Dot Score, and plotted them.

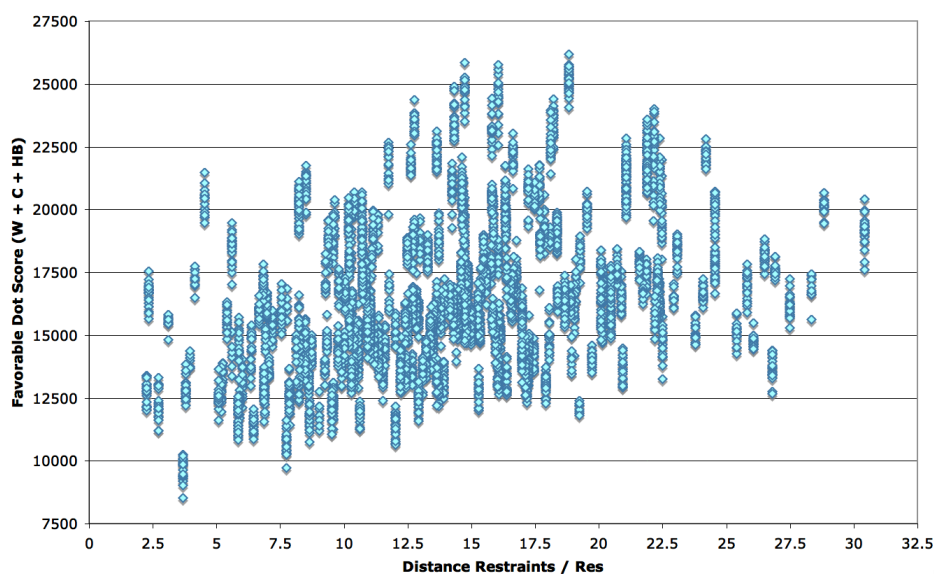


Figure 3-11 - Favorable Dot Score vs. Restraints Per Residue in NMR Structures

The figure shows the Favorable Dot Score plotted against NOE restraints per residue. As predicted, the score increases as the number of restraints per residue increases. The behavior of this value is similar to that of the Mainchain H-bond Score, though the trend is more pronounced. This is not surprising because counting the vdW contributions is the same as observing more contact dots at longer distances.

This work can be extended and refined by the implementation of residue-level and atom-level tables in the database. This would enable a more detailed study of the impacts by specific secondary structural elements on these scores.

3.2.6 NMR Structure Chimera: A Test Case for a Best Parts Approach

In order to evaluate the extensibility of the MolProbity geometric criteria for use in NMR structure improvement, a comparison of models within structure ensembles as well as between whole ensembles is needed. Comparing models within an ensemble provides a way to identify clusters of errors in a structure model. Comparing separate structure ensembles is important globally for identifying the best experimental or refinement methodologies, and locally for identifying specific types of systematic errors. For initial trials, multicriterion kinemages with all models of an NMR ensemble were something I created by hand, by editing together separate runs on the single models and modifying the colors and master buttons for intelligibility. They then became useful for investigating 3D clustering of geometric errors mapped on the structure, and I implemented them in MolProbity where they are now served up on the web (as described in section 3.1.1).

I also created two-dimensional plots of validation outliers by hand, with each member of the NMR ensemble separated vertically and the sequence running horizontally. Clustering of problem areas is immediately evident, especially in the highly mobile chain ends, which can be “trimmed” away for clarity especially in the 3D graphics. It is also clear from these plots that model 1 is often not the highest-quality

model in an ensemble by these criteria. It is for this reason that I analyze all models in an ensemble and plot them without averaging the number of outliers in each across the ensemble. This became so useful that I decided to build this type of plot into MolProbity, where it is now served up to the community (as described in section 3.1.2).

One use for which such 2D plots were very important was in the construction of a best-parts “chimera” for the 1EGF ensemble (Montelione , 1992) done jointly by Jessica Allison, Jane Richardson, and myself. Concentrating on the backbone and disulfides, we divided the structure into local segments with low rmsd at their ends and identified the best model or models for each segment. Originally we expected primarily to be avoiding problems (clashes and Ramachandran outliers), but we soon discovered that positive factors were just as important (H-bonding, compactness vs underpacking, and even “reasonable-looking” familiar conformations).

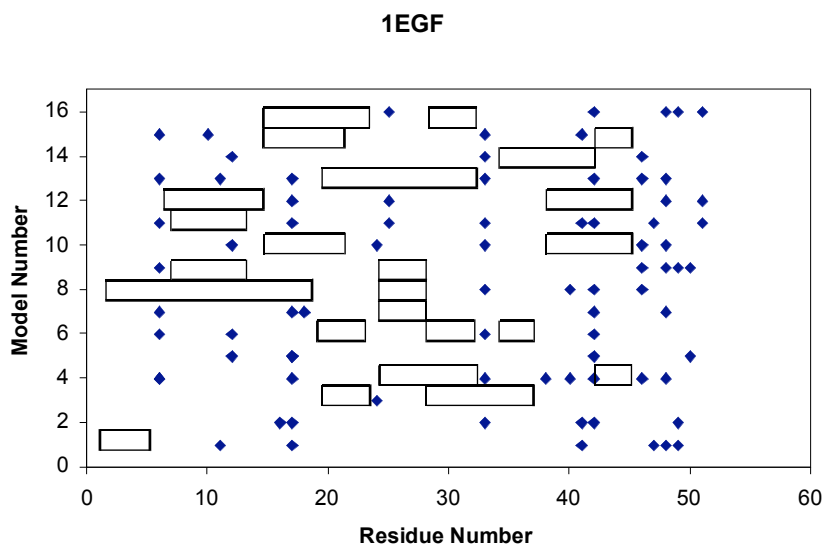


Figure 3-12 - Chimera Plot of 1EGF Structure Ensemble

The plot in Figure 3-12 shows boxes for the accepted models in each region, and also diamonds for the Ramachandran outliers which were avoided. At the C-terminal end, there were no reasonable conformations and the last seven residues would likely be trimmed off the structure model due to a lack of observed data. Although this modeling attempt was not remarkably successful, construction of a chimera (with some further refinement) still seems useful for an accurate, final NMR ensemble, providing a single model more globally accurate than a minimized average structure or than just using model 1. However, for a low-data ensemble like 1EGF this sort of examination would be better utilized to define additional restraints, then used for an entire new cycle of structure determination. For example, in 1EGF the cross-model comparison defined correct vs. incorrect conformations for the two central disulfides, which could profitably be restrained.

3.2.7 Z-Domain Structures and Other Examples

I created analogous 2D plots to the ones for the Chimera example for comparisons between different structures as well. The example in the figure below plots Clashscore and % Ramachandran outliers for each model in the three Z-domain structure ensembles done by the Montelione group at the NorthEast Structural Genomics Consortium (NESG). In order of determination, 1SPZ is the oldest, followed by 2SPZ and then 1Q2N (Tashiro, 1997; Zheng, 2004). All scores (including rotamers, not shown) improve from 1SPZ to 2SPZ, which was refined with a high degree of care and attention because of a controversy about the helix orientations. Clashscore goes up again somewhat in the later

1Q2N ensemble, but Ramachandran outliers achieve <1% because of peptide orientation information from NH RDC measurements. The 1Q2N ensemble also has more α -helical residues and more well-ordered residues.

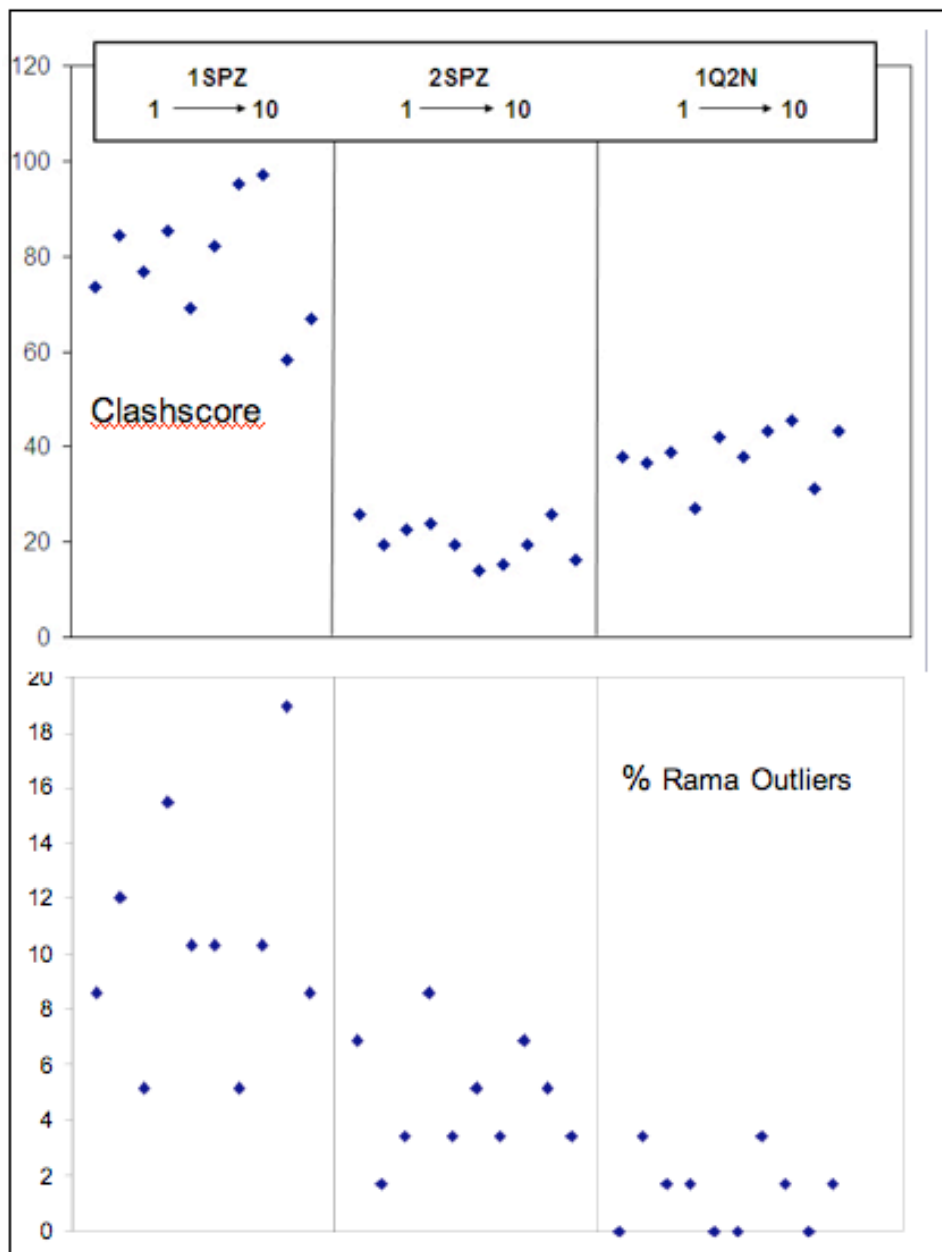


Figure 3-13 - Z-domain: Two Validation Criteria Compared Across Structure Ensembles

The plots and kinemages made by hand were promising, but extremely time intensive. This put priority on the development of more automated and flexible procedures. In collaboration with Ian Davis, I implemented methods for handling multiple models in the new version of MOLPROBITY, with a separate page tailored for NMR-specific questions, as has been described earlier in this chapter. Included to date are features that create either a multi-model multicriterion kinemage and a multi-model chart of summary statistics, or else single-model evaluations done one at a time by the old system. The new MOLPROBITY NMR facility also can accept an input restraint list in CNS format and use NOEDisplay to return a multicriterion kinemage that also maps onto the 3D structure any NOE restraint violations (shown as dashed red lines). However, this functionality is available only in the command line.

The fundamental task of a multicriterion kinemage is the same for NMR as for x-ray structures: to show clearly the clustering of all types of errors mapped onto the 3D structure. Unfortunately, this is much harder for NMR because an order of magnitude more information is being displayed per residue. The extreme case that broke all our earlier tools is the dynamic, 128-model ubiquitin ensemble of 1XQQ (Lindorff-Larsen et al. 2005); its 3D variation can now be successfully studied.

I used the infrastructure from the new tools I created to study a variety of examples. The initial structure survey documented a large number of structures for which sidechains with no experimental data at all are given rotamer-outlier conformations for many, or even for half of the ensemble members (eg 1KKG (Huang, 2003) with 55% poor rotamers total). Favorable rotamer conformations, or even low-

energy staggered values for individual χ angles, are not taken into account at all in many NMR structure determinations, probably because of historical reluctance to use any x-ray-derived information.

Structures from Gerhard Wagner's lab, analyzed when he visited here, provide informative examples because his lab uses a wide variety of software and strategies. The 1Z9E integrase-binding domain (Cherepanov, 2005), which Wagner identified as especially carefully done, has Clashscores of ~ 28 and Ramachandran outliers of $\sim 1\%$, which correspond to scores for a typical x-ray structure at 2.3\AA resolution. However, 27% of the rotamers are poor, which is exceptionally bad for an x-ray structure but could easily have been avoided, because most of those rotamer outliers are on the surface where there is room to choose favorable alternatives.

In contrast, the Wagner lab's 2AIV structure of a nucleoporin domain (Robinson, 2005) is more mobile and was refined with DYANA rather than XPLOR; it achieves just 7% bad rotamers per model, but its clashscore >100 and Ramachandran outliers of 10% are very bad. Overall, 2AIV seems much less accurate than the integrase-binding structure from the Wagner group, but shows that sidechain rotamers (as well as covalent geometry) can be given largely canonical values without contradicting the NMR data.

Similarly, in the Montelione-lab Z-domain structures, Lys 50 has data only to show that Cd is within NOE distance of its own Ha and NH. In 1Q2N only 3 of the 10 models have acceptable rotamers for Lys 50, and none make H-bonds. In 2SPZ all Lys 50 conformations H-bond either to Glu 47 or Asp 53, but half the rotamers are good and half are not.

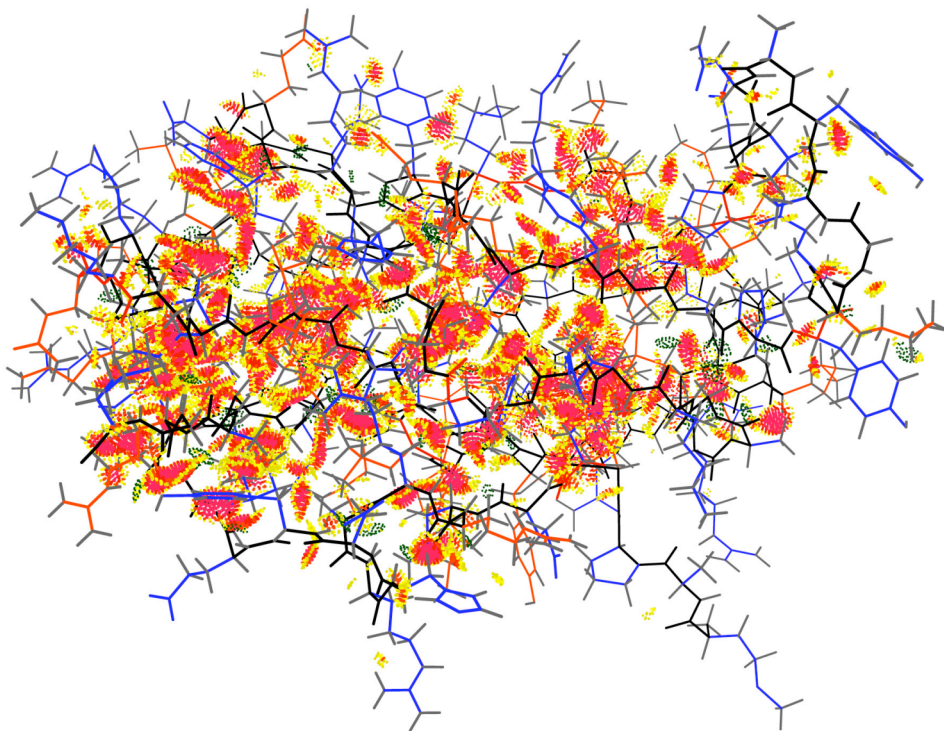


Figure 3-14 - 1J1H Structure with a Clashscore of 242.9

The most egregious example was the 1J1H structure with a Clashscore of 242.9 (model 1 shown above in Figure 3-14). A closer examination of the clashes using the multi-model multi-criterion kinemage revealed that most of the clashes include sidechain atoms, had 30.29% bad rotamers, 8.26% Ramachandran outliers, and a MolProbity score of 4.68 (0th percentile).

The Montelione lab structures showed a case for 2SPZ where all 4 ensemble members with a frayed helix had bad geometry for residues 22-25 (clash, Ramachandran, and rotamer outliers) while the other 6 models use Asn 23 as a classic helix N-cap and have no validation problems; they also match the conformation later assigned for all models in 1Q2N, based on RDC data.

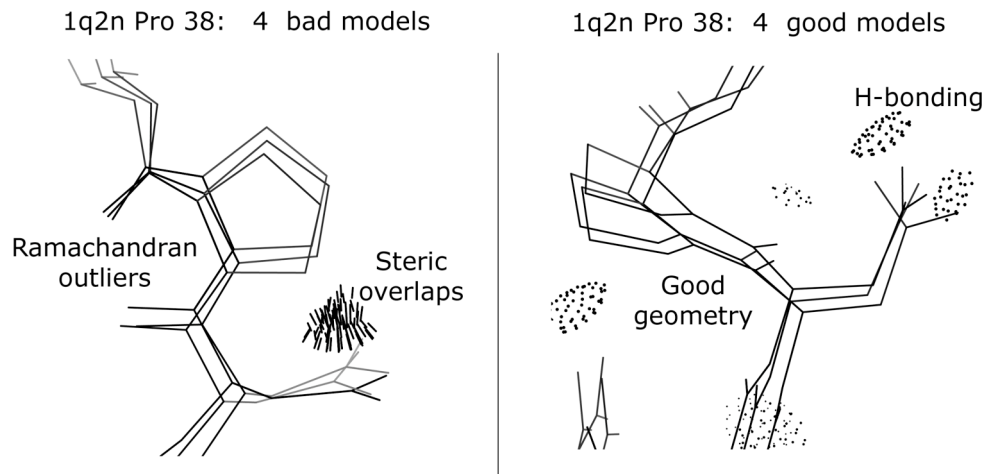


Figure 3-15 - Error Cluster in Proline Turn of Z-domain structure 1Q2N

In 1Q2N, even with RDC measurements, one group of models (at left in figure) showed multiple clashes and outliers while a second group (at right) showed excellent geometry and additional H-bonding for a tight turn. Presumably both peptide orientations were consistent with the RDC value, but this analysis clearly shows that only one of those alternatives is correct. In 2SPZ, this same 37-40 loop shows many different conformations, most with outliers, but no cases of the conformation known to be correct. (Note: this example was very significant in other parts of my work and is revisited in greater detail in chapter 4 and chapter 5).

The 2SPZ structure also demonstrated the common problem with surface sidechains where little or no experimental distance restraints are present.

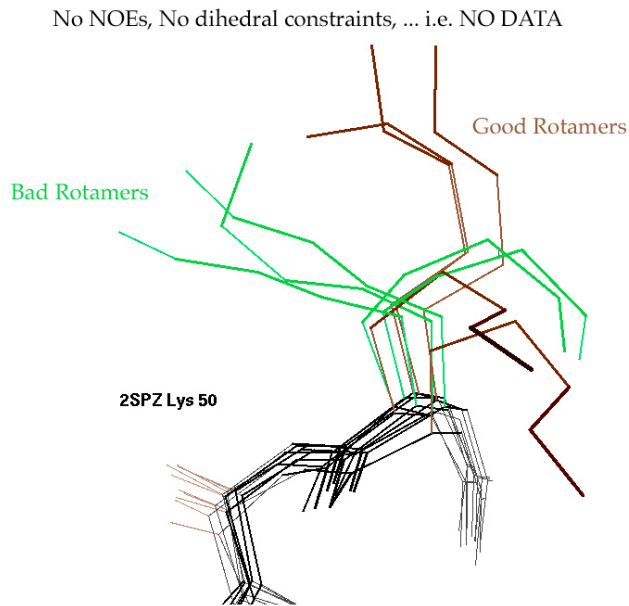


Figure 3-16 - Surface Sidechain Conformations of Lys50 in 2SPZ

The figure shows Lys50, with no distance restraints present in the experimental data, modeled using rotameric conformations in some models in the ensemble and outlier conformations in others. This surface exposed sidechain, I conclude, should only be modeled with rotamers where there is little or no experimental evidence to the contrary.

A similar case occurs in the unpublished Montelione-lab FGF structure, where the 10 models for loop 135-145 differ but all have both backbone clashes and Ramachandran outliers. Such patterns mean that the sampling for the loop is inadequate, which could perhaps be remedied if it was recognized that no acceptable alternatives had been found. These and related analyses provided additional restraints for the Z-domain structure that were sent to the Montelione group for their consideration in further refinements.

3.3 NOE patterns used to identify RNA backbone conformations

RNA has been implicated in a rapidly growing number of biological processes, from transcription and translation to catalysis and gene regulation. These diverse functions owe much to RNA's ability to form many types of stable 3D structures. Base-pairing and base-stacking are the dominant factors in these structures, but backbone structure is known to have key roles in catalysis and binding. Unfortunately, RNA backbone has been a thorn in the side of crystallographers and NMR spectroscopists alike.

In the past decade, advances in technical methodology and biological motivation have led to the watershed ribosome structures in the year 2000 (Ban, 2000; Schluenzen, 2000; Wimberly, 2000), and to many other RNA crystal structures since then. This enormous increase in high quality information on RNA 3D structure has revolutionized RNA structural bioinformatics, leading to, among other things, the discovery that the RNA backbone is rotameric when divided into sugar-to-sugar units called suites (Murray, 2003). Each of the conformers is assigned a unique name, which can be combined with the names of neighboring suites to form suitestrings—a 1D descriptor of 3D conformations (Richardson, 2008). NMR structures of RNA were quite dominant before 2000, but crystallography of RNA has been surging forward. There are over five times as many solved RNA structures presently in the PDB as there were at the end of 2000, with the increased rate mainly due to x-ray crystallography. It would aid overall progress and provide a complementary, more dynamic perspective if NMR methodology could catch up.

The usual way to determine nucleic acid NMR structures relies heavily on NOE (Nuclear Overhauser Effect) through-space distance constraints, allowing the placement of hydrogen atoms ~ 5 Å or less apart. But determining the detailed conformation of RNA backbone using NMR is quite tricky, since the density of observable and useful proton-proton distances is much lower than for proteins, and the interesting RNA structures tend to be the most difficult to analyze (Varani, 1996). In practice, it is easy to identify the regions of A-form RNA structure vs the rest. In addition to their role in refinement, the NOE constraints are used to guide early model building, such as determining helices and hairpin loops in RNA, and in some cases structural features as small as single-base bulges. One place where NMR has an advantage over crystallography is that sugar pucker can be determined by J-coupling measurements that reflect individual torsion angles in the ring. Late in the refinement process, RDC (residual dipolar coupling) orientation measurements are sometimes added, especially to determine long-range shape; RDCs will be discussed in later chapters.

My work, in collaboration with Gary Kapral, maps out all the expected interatomic distances between hydrogen atoms (and thus the potentially observable NOEs) along RNA backbone in each of the suite conformers, allowing the user to display them both in tabular form and in parallel coordinates (Inselberg, 2009; Chen, 2010). Viewing the possible NOE distances in parallel coordinates is an especially revealing way to identify patterns corresponding to particular backbone conformations. As with suitestrings, series of such conformations have their own identifiable multi-residue patterns; thus, an S-motif and a GNRA tetraloop imply distinct, repeatable NOE patterns.

Even if the base is facing the solvent, and thus has little NOE data to identify the base position, the observed NOEs can be used to systematically pare down the possible backbone conformers at that suite; these can then be combined to identify common suitestrings and thus the local RNA 3D structure. Taken together, patterns of distance constraints and the system of suite conformers should provide a powerful tool to help elucidate the 3D conformations of RNA backbone in NMR structures.

3.3.1 Structural Overview of RNA

RNA can be divided into two major components per nucleotide residue: the backbone, consisting of the sugar and phosphate, and the base, which can be a two-ring purine (G,A) or a one-ring pyrimidine (C,U). The structure of RNA has been traditionally described solely by the base-pairing and base-stacking interactions, which are most readily identified. More recently, it has been discovered that the backbone, which has 6 variable dihedral angles per residue, adopts distinct, favored 3D conformations. These backbone conformations have been described in a number of ways—through binning the alpha, gamma, delta, and zeta torsion angles (Hershkovitz, 2003), through identifying backbone rotamers in a new “suite” division from sugar to sugar, spanning 7 dihedrals from δ_{n-1} to δ_n , (Murray, 2003), and through smoothing and peak-picking of the 12 backbone dihedrals of a dinucleotide along with the two χ angles of the bases (Schneider, 2004); these three alternative divisions can be seen in Figure 3-17 below.

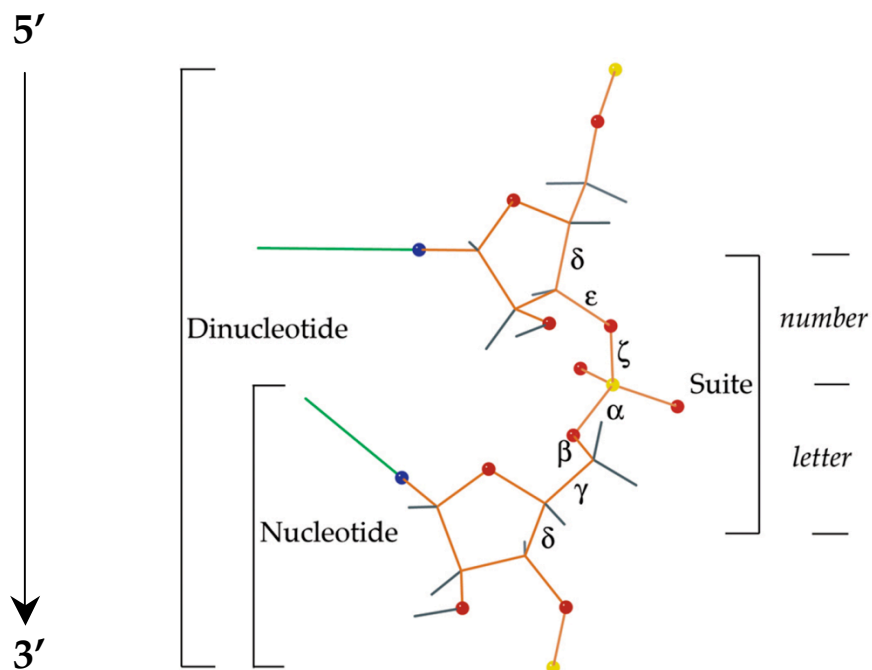


Figure 3-17 – RNA backbone suite.

The above figure shows the suite dihedral angles labeled. Also marked are the divisions into suite, nucleotide (residue) and dinucleotide. The modular heminucleotide units and appropriate nomenclature are shown along the right edge. The groups that developed these separate backbone classification schemes came together under the auspices of the RNA Ontology Consortium (Leontis, 2006) to define a unified nomenclature, and identified 54 well-conserved backbone conformations throughout RNA structures (Richardson, 2008).

Modular consensus nomenclature

- for $\delta\epsilon\zeta$ heminuclotides:
- C3'endo puckers – odd numbers:
 - 1 = 3' -e m
 - 3 = 3' -e t
 - 5 = 3' -e p
 - 7 = 3' -e -e
 - 9 = 3' -e e
 - & = 3' t -e
- for $\alpha\beta\gamma\delta$ heminuclotides:
- For C3'endo pucker:
 - a = m t p 3' 1a is A-form
 - c = t t t 3' 1c is variant of A-form
 - d = p t p 3' inverted “p”; see below
 - e = -e p t 3' 1e is stack-shift dent;
 - f = t e t 3'
 - g = t t p 3' 1g is suite 1-2 of GNRA
 - h = m t t 3'
 - i = p e t 3'
 - j = p e t 3'
 - l = m e p 3'
 - m = m -e p 3' minor 1a shoulder
 - n = p t t 3' 6n is 2'3' Z-form
- C2'endo puckers – even numbers:
 - 2 = 2' -e m
 - 4 = 2' -e t
 - 6 = 2' -e p
 - 8 = 2' -e -e
 - 0 = 2' -e e
 - # = 2' t e
- For C2'endo pucker:
 - b = m t p 2' 2b would be B-form DNA
 - o = m t m 2' 1o and 2o both put bases opposite
 - p = p t p 2' most p angles in 2' set
 - q = p e t 2'
 - r = p t m 2' rare reverse order of m t p
 - s = m p t 2' 4s is suite 2-3 of S-motif
 - t = t t t 2' all-trans
 - z = t t p 2' 5z is 3'2' Z-form
 - [= m -e p 2' 1[is commonest intercalation

Figure 3-18 - Heminucleotide designations and associated conformations

In the figure above, the mean dihedral angles are to the right of each conformation. Notation of 3' and 2' indicate the pucker (δ nearly 84° or 147°, respectively); m signifies near -60° (minus); t, 180° (*trans*); p, +60° (plus); e, 120° +/- 25°; and -e, -120° +/- 25°.

The modular nomenclature for RNA backbone conformation uses letter-number combinations to give a 1D representation of the 3D dihedrals by describing the $\delta\epsilon\zeta$ dihedrals of the preceding residue by a number (or number-like single character), and the $\alpha\beta\gamma\delta$ torsions of the current residue by a letter. Odd numbers, and letters “a-n”, represent conformations with a C3'-endo sugar pucker; even numbers, and letters “o-z”, represent conformations with C2'-endo sugar puckers (and a δ near 147°). The notable exception to this rule is the letter b, where 2b is reserved for B-form DNA (although it cannot occur in RNA); “b” is therefore a 2'-pucker heminuclotide.

Only some combinations of a number and letter designate actually observed, favorable suite conformations. The 54 officially recognized backbone conformers include 3'3', 3'2', 2'3, and 2'2' suites (Richardson, 2008). A complete table with the angle values, roles, and representative examples can be found on the Richardson laboratory website. Using this modular nomenclature, the conformation designated as "1a" refers to the backbone dihedrals needed to construct an A-form RNA helix, with the "1" describing the $\delta\epsilon\zeta$ torsions, and the "a" describing the $\alpha\beta\gamma\delta$ torsions; from the table above, it can be inferred that 1a must have 3'3' sugar pucker. "[[" was chosen for the commonest intercalation conformation, since the "[[" character resembles the shape of a suite with bases spread apart for intercalation. A pair of underscores, "__", is used to indicate regions where dihedral data is undefined, such as chain ends and internal chain breaks (disordered loops, etc.). Finally, the "!!" designation is used to indicate that the suite does not fall into any of the 54 recognized conformers; this can either be due to errors in the model or to interesting regions of strained geometry, such as is often found in active sites.

RNA structural motifs, such as the tetraloop or the S-motif, can be specified by the corresponding strings of suites that describe their backbone dihedrals, as shown in the following two figures.

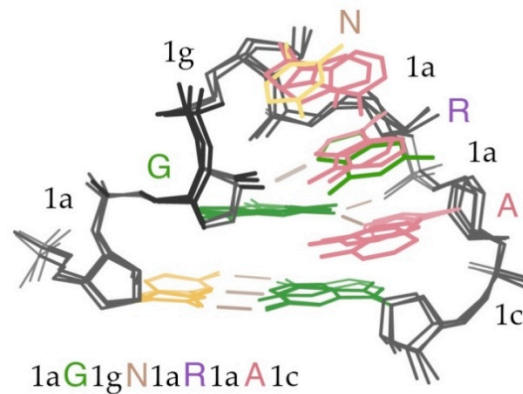


Figure 3-19 - RNA Backbone Nomenclature for GNRA tetraloop

The GNRA tetraloop has suitestring 1a1g1a1a1c; bases can be included to fully describe the RNA structure: 1aG1gN1aR1aA1c. Strings of non-A-form RNA indicate large deviations from helices, usually in the form of stem-loops, internal loops, or junctions.

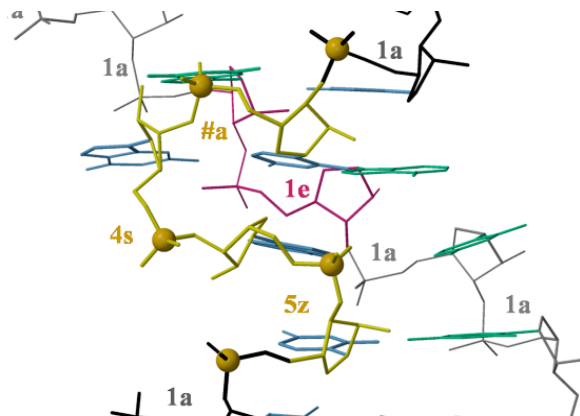


Figure 3-20 - RNA Backbone Nomenclature for S-motif

The distinctive S-shape of the primary strand of the aptly named S-motif, for example, has a suitestring of 5z4s#a. The corresponding back strand of the S-motif has suitestring 1a1e1a, the 1e conformation being necessary to make the stack switch that accommodates the primary strand's distinctive shape.

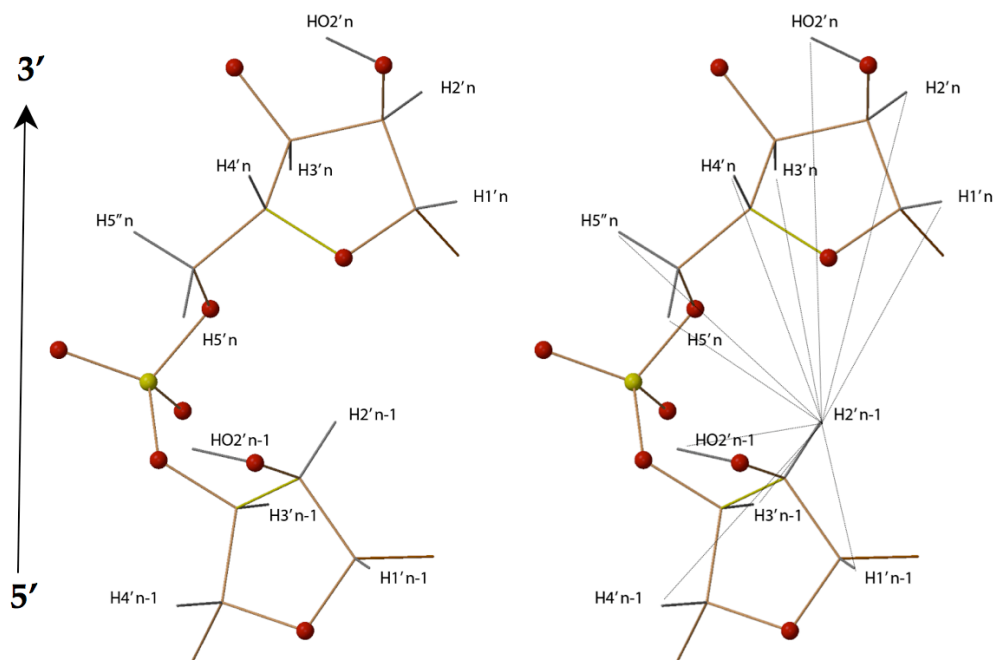


Figure 3-21 - RNA backbone suite with potential H2'(n-1) NOEs shown

The RNA backbone within a suite has 12 available hydrogens whose 3D arrangement is determined by the suite conformation. Shown above is an RNA backbone suite with all the hydrogen atoms named (on the left of the image pair), and lines for the eleven potential backbone NOE's involving the H2'(n-1) atom (on the right of the image pair). Since NOEs are based on the distance between hydrogens, one can calculate the expected NOE's for each RNA backbone conformation based on their relative hydrogen positions. The resulting calculations can be used as a lookup table for identification of RNA backbone structure during an NMR 3D structure determination and building process.

3.3.2 NOE's by CRMA

NOE constraints are often determined by calculating distance between atoms via the relaxation matrix. The relaxation matrix itself is related to the intensity of the NOE peaks (Cavanagh, 2006). This method is much more complete and exact than the semi-quantitative distance approach. In practice, not all the intensities are known, resulting in an incomplete intensity matrix and thus interatomic distances cannot be computed exactly. A general solution for this problem (called the Constraint Relaxation Matrix Approach - CRMA) is to build a model, often idealized A-form helix for RNA (Boelens, 1989; Schmitz, 1995), from which a model relaxation matrix is calculated. Observed NOEs are substituted for the theoretical values where possible, and a hybrid matrix is constructed from which distances are calculated. This process of substitution and back-calculation is iterated several times until the calculated and experimental distances are within an acceptable range (Wijmenga, 1998). Many alternative methods have been proposed for building a complete relaxation matrix (Borgias, 1990; Kaluarachchi, 1991; van de Ven, 1991).

A severe disadvantage of using CRMA is overdependence on starting models (Borgias, 1989; Borgias, 1990). Using A-form RNA in the starting model yields an A-form-based model intensity matrix. A-form RNA accounts for >70% of the residues in RNA structures (Murray, 2003) and thus will give a good overall match to the data. However, this method will severely hamper the ability to model non-A-form structure and muddy the waters when looking at interesting areas, particularly binding and active sites.

In contrast, using NOE lookup tables could enable non-A-form structure to be determined for particular stretches of residues, allowing for better starting models with less A-form bias. Combining these lookup tables with known suitestrings, whole motifs could be handled in CRMA without resorting to A-form placeholders in the initial structure.

3.3.3 Semi-Quantitative Distances

The semi-quantitative distance approach relies on a large number of NOE constraints rather than precision of a given NOE. The NOE distances are estimated based on their relative intensities with respect to generally observed reference NOEs as measured on the same sample, and are classified only into bins as strong, medium, or weak. These reference atoms are often the H5-H6 distance (2.43Å) for strong, H1'-H3' (~3.5Å) for medium, and H6/H8 to H1' for weak (>5Å) (Varani, 1996). Using a relative scale based on reference atoms is important for offsetting the influence of mixing time on the intensity of the peaks when spin diffusion skews the distance measurements over longer mixing times.

In the semi-quantitative approach, each bin corresponds to a rough distance estimate, less than the upper bound of the constraint for that bin. Due to the difficulty in determining the lower bound accurately, most constraints have no lower bound. Looking at these upper bounds provides an invaluable resource for comparing observed NOEs to NOE constraints calculated from ideal RNA conformers because more NOEs in each bin

means more cross-references to RNA backbone conformations; thus one can attempt to create a more robust estimate of which conformation is present.

3.3.4 Lookup Tables & Parallel Coordinate Plots

Using the 54 defined RNA backbone conformations, Gary Kapral and I created a series of twelve tables, one for each of the twelve hydrogens in the suite considered as one end of the potential NOE pair. Names of the 54 rotamers are listed vertically and the twelve hydrogens for the other end of the pair are along the horizontal direction. Each table contains shaded-out cells (in purple) where the interatomic NOE distances are long enough ($>5.2 \text{ \AA}$) that they are very unlikely to be observed experimentally.

H2' n-1 to:	H3' n-1	H4' n-1	H1' n-1	H2' n-1	HO2' n-1	H5' n	H5'' n	H4' n	H3' n	H2' n	HO2' n	H1' n
1a	2.459	3.773	2.770	0.000	2.597	2.951	4.295	4.232	4.335	5.389	6.030	4.039
1c	2.458	3.765	2.773	0.000	2.615	4.358	3.356	3.733	4.956	5.735	5.910	3.840
1e	2.459	3.773	2.770	0.000	2.252	5.125	4.498	4.822	6.127	6.958	6.959	4.964
1f	2.459	3.773	2.771	0.000	2.621	4.307	3.280	4.717	5.516	6.891	6.969	5.466
1g	2.459	3.774	2.771	0.000	2.650	4.865	3.804	5.206	2.504	4.547	5.883	6.037
1L	2.468	3.813	2.756	0.000	2.546	3.801	5.041	4.215	4.679	5.100	5.693	3.153
1m	2.463	3.796	2.761	0.000	2.604	2.469	3.802	4.738	5.220	6.882	7.306	5.835
3a	2.466	3.804	2.759	0.000	2.650	5.851	6.692	7.972	6.780	8.463	9.819	8.146
3d	2.466	3.803	2.758	0.000	2.638	4.682	4.543	6.759	7.009	9.332	9.109	9.235
3g	2.466	3.804	2.759	0.000	2.864	6.077	5.891	8.097	6.905	9.216	9.854	9.562
5d	2.458	3.765	2.773	0.000	2.614	5.277	4.759	6.954	5.495	7.879	8.647	8.577
5j	2.468	3.820	2.754	0.000	2.570	5.566	5.615	6.813	8.065	9.938	9.407	8.888
5n	2.468	3.813	2.756	0.000	2.620	6.810	5.531	7.555	7.903	9.313	9.879	7.887
7a	2.463	3.789	2.764	0.000	2.137	5.060	6.019	6.560	5.819	7.004	8.276	6.084
7d	2.464	3.797	2.762	0.000	2.597	3.288	2.145	4.123	4.913	6.982	6.086	7.232
9a	2.463	3.789	2.765	0.000	2.528	5.561	6.633	7.824	7.446	9.163	9.985	8.294
&a	2.460	3.781	2.768	0.000	2.290	4.020	5.008	5.635	4.784	6.168	7.365	5.477
1b	2.464	3.797	2.762	0.000	2.627	2.598	4.142	4.147	5.318	4.586	6.504	4.746
1o	2.464	3.796	2.762	0.000	2.644	2.699	3.726	5.484	4.501	4.589	7.064	6.951
1t	2.459	3.773	2.770	0.000	2.287	4.557	3.657	4.034	5.693	5.064	6.580	4.576
1l	2.463	3.790	2.765	0.000	2.627	2.545	3.866	4.875	5.612	5.304	7.536	6.168
1z	2.462	3.789	2.765	0.000	2.683	5.078	4.301	5.939	3.873	3.461	5.904	6.413
3b	2.466	3.804	2.759	0.000	2.620	5.968	6.465	8.215	7.429	6.548	9.381	8.932
5p	2.464	3.797	2.762	0.000	2.573	5.757	5.260	7.276	5.630	5.301	7.911	8.142
5q	2.461	3.780	2.768	0.000	2.591	5.236	5.286	6.583	7.610	8.067	10.047	8.782
5r	2.463	3.788	2.764	0.000	2.626	5.876	5.862	7.181	6.693	8.001	9.881	9.827
5z	2.463	3.788	2.764	0.000	2.584	5.562	6.156	6.997	6.140	4.193	6.896	6.430
7p	2.464	3.797	2.762	0.000	2.609	3.478	3.257	5.214	5.588	6.892	8.542	7.950
7r	2.466	3.804	2.759	0.000	2.585	3.555	2.336	4.477	5.956	6.070	7.883	6.371
0a	2.524	3.887	3.084	0.000	2.490	6.087	6.812	8.516	7.988	10.006	10.738	9.516
0i	2.523	3.888	3.083	0.000	2.782	7.480	6.508	6.330	7.980	8.157	7.984	5.587
0k	2.524	3.887	3.084	0.000	2.846	5.905	5.236	7.520	8.232	10.357	10.155	9.709
2a	2.506	3.894	3.085	0.000	2.834	4.171	5.663	6.354	6.891	8.329	8.618	6.903
2g	2.490	3.896	3.086	0.000	2.769	6.884	6.535	7.743	5.010	6.562	8.386	7.940
2h	2.520	3.888	3.086	0.000	2.412	3.519	5.127	5.266	5.750	7.814	6.808	8.457
4a	2.506	3.897	3.079	0.000	2.842	5.746	7.248	7.464	7.607	8.543	9.456	6.896
4d	2.527	3.886	3.084	0.000	2.802	6.598	5.023	7.352	5.433	7.631	8.246	9.228
4g	2.521	3.888	3.086	0.000	2.818	7.099	7.609	9.063	7.556	9.317	10.705	9.293
4n	2.501	3.894	3.087	0.000	2.864	6.354	4.733	7.129	6.938	8.316	9.142	7.114
6d	2.514	3.891	3.085	0.000	2.831	6.411	5.741	8.187	7.195	9.640	10.009	10.265
6g	2.505	3.894	3.086	0.000	2.791	6.106	7.220	7.807	7.401	8.515	9.639	7.297
6j	2.491	3.897	3.086	0.000	2.862	5.799	6.031	7.546	8.397	10.523	10.113	9.888
6n	2.528	3.886	3.084	0.000	2.802	6.819	5.418	7.677	7.354	8.724	9.642	7.614
8d	2.523	3.887	3.084	0.000	2.807	5.522	4.278	6.596	6.233	8.568	8.207	9.432
#a	2.521	3.887	3.086	0.000	2.864	5.682	6.710	8.154	7.785	9.621	10.323	8.878
0b	2.520	3.888	3.086	0.000	2.843	5.549	6.638	8.020	8.413	7.791	10.354	9.082
2o	2.513	3.890	3.085	0.000	2.845	4.501	6.031	6.079	3.488	4.274	5.627	6.848
2u	2.502	3.895	3.086	0.000	2.819	4.863	6.235	5.690	7.246	8.762	9.697	8.656
2l	2.506	3.898	3.079	0.000	2.808	3.880	4.874	6.347	7.091	7.076	9.366	8.126
2z	2.495	3.896	3.086	0.000	2.864	6.699	5.777	7.509	5.164	4.988	7.164	7.989
4b	2.505	3.895	3.086	0.000	2.798	5.869	7.033	8.291	8.326	7.237	9.939	8.790
4p	2.528	3.886	3.083	0.000	2.863	6.488	4.810	7.042	4.600	5.156	6.687	7.919
4s	2.528	3.886	3.083	0.000	2.821	6.792	7.523	5.225	7.904	8.914	8.772	7.088
6p	2.507	3.897	3.079	0.000	2.841	6.774	5.569	7.965	6.002	6.110	8.393	8.994

Figure 3-22 - RNA backbone NOE distance Lookup Table for H2'(n-1)

Each suite conformer is identified by its appropriate number/letter combination describing the dihedral-angle values. Cross-referencing the observed NOE with the

calculated NOEs in the table allows the spectroscopist to determine what subset of the 54 backbone conformations are possible given the data. For example, a spectroscopist observing a strong NOE of 3.2 Å between H2' and H1' of the following residue uses the H2'_{n-1} table, and will see that the only possible calculated NOE in the H1'_n column that fits is 3.153 Å, belonging to the 1L conformation. The 1L conformation differs from A-form in the β and ε dihedrals resulting in a more twisted base positioning.

If, on the other hand, a medium strength NOE of 4.0 Å is observed, the 1a and 1c (as well as 1L) conformations constitute the reasonable subset of the 54 rotamer choices. For a longer-distance NOE of 4.7 Å or so, the best match would now include 1b or 1t conformations (both 3'2' puckers). Importantly, the observed NOEs in the semi-quantitative method are only representing upper bounds; conformations with lower NOE constraints (1a, 1c, and 1L in this case) may still be candidates.

Not all observed NOEs will be deterministic—many will result in ten or so conformations that match the data. This can be refined further by referencing multiple sets of observed NOEs and paring down the subset of backbone conformations plausible; using the 4.7 Å NOE example above, a second observed NOE between H2'_{n-1} and H5''_n of 3.6 Å further pares the choices down to 1c or 1t, which can easily be distinguished based on sugar pucker determination. Pucker combinations can be identified on the table by color: 3'3' is highlighted in blue, 3'2' in green, 2'3' in yellow, and 2'2' in red. Sugar pucker can be easily determined independently with J-coupling or ³¹P chemical shifts (Varani, 1996). [It is interesting that distinct ³¹P chemical shifts for C2'-endo vs. C3'-endo ribose puckers are probably looking at the same geometrical relationships that

permit pucker determination in X-ray models by measuring the perpendicular distance from the 3' phosphate to the line of the glycosidic (sugar-to-base) bond (Richardson, 2009; Chen, 2010).]

Ultimately, the strength of the approach is that even though multiple RNA backbone conformations are possible, the number of plausible ones that match the data will be reduced significantly compared to the 54 known RNA backbone rotamers and allows the spectroscopist to begin with more reasonable starting models for relaxation matrix generation.

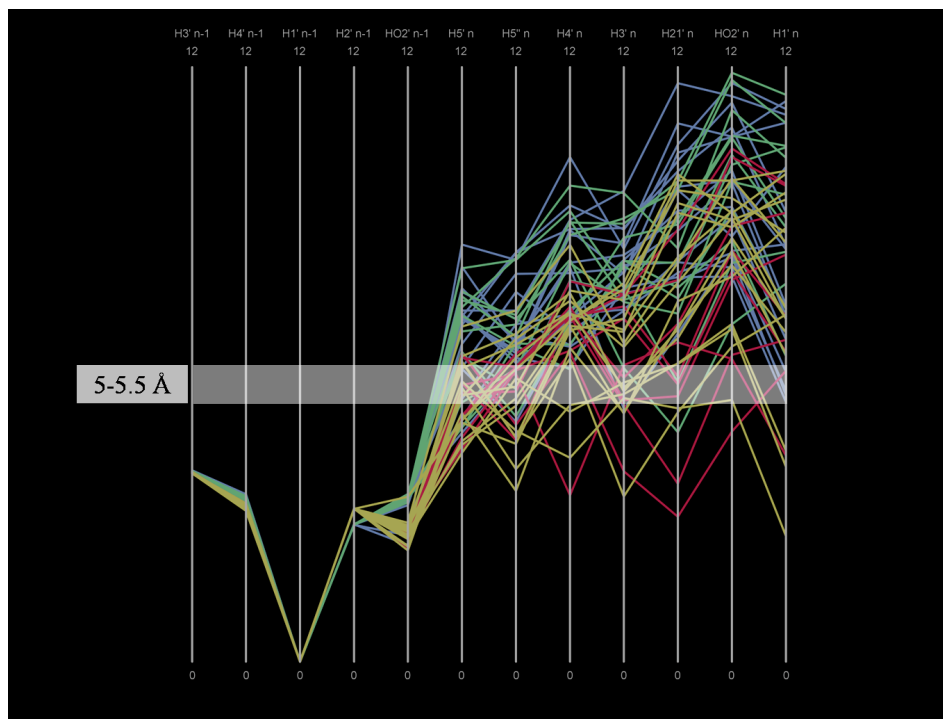


Figure 3-23 - Parallel Coordinate Plot of RNA Backbone Rotamer NOE distances H1'(n-1)

Complementary to the tables are parallel coordinate plots where each axis in the plot represents a NOE distance between two hydrogens within the suite and is populated

by the distances for each one of the 54 RNA backbone conformations. Twelve parallel coordinate plots were made, one for each hydrogen in the suite. The twelve plots provide a visual representation of the NOEs found in the tables. A dividing bar is shown at 5.0-5.5 Å, indicating where it becomes difficult to experimentally observe an NOE. The color scheme for pucker combinations is maintained (blue for 3'3', green for 3'2', yellow for 2'3', red for 2'2'). Each conformation has a unique polyline representing the ideal NOE constraints; the parallel coordinate view can act similar to the lookup table, but also give the spectroscopist insight into what other NOEs might be observed for each conformation, aiding in the interpretation of the experimental data.

3.3.5 Test Structures

A series of test structures culled from the Protein Data Bank were used to assess the limits of the method. A search performed at www.pdb.org used the Advanced Search, a selection of Experimental Method being solution NMR, with data present followed by a second query for molecule type with the restriction of RNA. A total of 271 structure hits returned from the late 1990's until today. Each structure was run through the program Suitename (Richardson, 2008) to get their suite conformer assignments. Ten structures were selected for an initial test of the method as seen in Table 2; these examples were selected for their variety of length, structural motifs, and function and for containing relatively high numbers of non-A-form suites.

Table 3 - The 10 NMR, RNA test structures and their number of residues demonstrating diversity in size.

PDB code	Author	Structure description	Residues
1F9L	Rudisser	PSABC hairpin of GI ribozyme	22
1I3Y	Blanchard	A loop of 23S rRNA	19
1IKD	Ramos	tRNA acceptor stem	22
1LDZ	Hoogstraten	Lead-dependent Ribozyme	30
1T4X	Popenda	Z-RNA	12
1UUU	Sich	Hairpin loop with CGUUUCG motif	19
1YMO	Theimer	P2b-P3 pseudoknot from telomerase RNA	47
2B7G	Johnson	Smaug recognition element	19
2HGH	Lee	TF IIIA with 5S rRNA	55
2JYM	Schwalbe	Stem-loop of hepatitis B virus PTRE	22

The .mr data files containing the NMR restraints used to calculate the structures were downloaded and non-NOE data spliced out. Additionally, any NOES that were not between backbone atoms of adjacent nucleotides were removed; only combinations of the 12 atoms we know to constrain the suite were analyzed. We also did not use suites labeled as “!!” by Suitename. Each NOE distance restraint observed within a suite was

matched with its possible backbone conformations, by finding the appropriate lookup table and then indexing down to the value of the NOE restraint and assigning the appropriate subset of 3D backbone conformers. Where more than one NOE was present for the same suite, each NOE was used to generate a subset of possible conformers. The resulting conformer subsets were combined to form two new subsets, one with the intersection of all NOE-derived conformer subsets for that suite, and one with the union of these subsets. An example of this can be seen in Figure 3-24.

				Conformer Subsets
NOE1:	17 RGUA H2'	18 RCYT H5'	2.70	1a,1c,1e,1f,1L,&a,1b,1t,
NOE2:	17 RGUA H2'	18 RCYT H1'	5.00	1a,1b,1m,1o,1[,
Intersection Conformer Subset:				1a, 1b
Union Conformer Subset:				1a,1b,1c,1e,1f,1L,1m,1o,1t,1[,&a

Figure 3-24 - Combining the conformation subsets of multiple NOEs to get the intersection and union conformer subsets.

3.3.6 Results

Our study looked at 10 NMR structures of varied size and shape, giving us a potential set of 267 suites to investigate. Unfortunately, it became clear that the 1F9L structure contained unreasonably short NOE constraints and was removed. Of the 245 residues remaining, 152 had observed NOEs between suite-constraining backbone atoms; 34 of these were assigned !! by Suitename, leaving us 118 suites, for detailed study.

The intersection and union conformer subsets were generated for each suite and compared to the suite present in the deposited structure (the suite assigned by Suitename).

If the actual suite is present in the conformer subset, it means the suite in the structure model at that location which was submitted to the PDB is consistent with our method of predicting RNA backbone conformations based on NOEs. If the modeled suite is not contained within the conformer subset, then the prediction has failed at that point.

Throughout the 9 test structures, the model suite matches the intersection conformer subsets 61% of the time, while it matches the union conformer subsets 79% of the time. This is intuitive in that the intersection will have smaller subsets of conformers and therefore less chance of matching the modeled suite. On the other hand, these numbers seemed rather low, likely because of error in NOE distance constraints which often have error of roughly 10%. Adjusting the lookup tables to separately report subsets within this 10% margin of error, by adding 10% to the value, improved the resulting match with the model suite by 15%. Overall, the model matched our new intersection conformer subsets 76% of the time, while it matched the union conformer subset 92%. A summary of these statistics can be found below in Table 4 (NB: table values are only for central values of each conformer, so they also have a substantial uncertainty range).

Table 4 - Conformer statistics: intersection conformer subsets, union conformer subsets, and each of these subsets extended to include conformers in the 10% error region for NOEs.

	Intersection Conformer Subsets	Union Conformer Subsets	Intersection Conformer Subsets +10% NOE error	Union Conformer Subsets +10% NOE error
Total suites (non!!)	118	118	118	118
Total matching suites	72	93	90	108
Total suites nonAform	42	42	42	42
Total matching suites (non-Aform)	17	34	24	39
Overall % matching suites	61%	79%	76%	92%
Overall % matching suites (non-Aform)	40%	81%	57%	93%
Average % matching suites per structure	51%	73%	67%	88%
Average % matching suites per struc (non- Aform)	46%	81%	61%	90%
Average reduction in conf space per structure	84%	63%	80%	61%
Average reduction in conf space:non-A- form	80%	56%	76%	48%
Conformer space saved if puckers are known			91%	86%
Conf. space saved if puckers are known (nonAform)			88%	82%

The second test of the method involved determining how effectively it restricts the RNA backbone conformer space to a limited set of suites. To evaluate this, the percentage of the 54 recognized conformers present in each conformer subset was determined both for intersection and union subsets. Similarly, the percentage of

conformers discarded by using this prediction method was calculated. This represents the percentage of suites not considered for model building, thus saving time in calculations and in choosing a model to fit the data. Overall, for intersection subsets there is an 80% reduction in the considered conformational space, with a corresponding 61% reduction for the union subsets.

While these numbers are very good for a first look, the truly interesting part of this method is its treatment of non A-form RNA. Throughout the test structures, there are 42 non A-form suites with useful backbone NOEs. Of these non A-form suites, 57% match with their corresponding intersection subsets and 93% with their union subsets. Conformation space reduction is as follows: 76% for intersect subsets and 48% for union subsets. Astonishingly, in these test structures one can use the combination of NOE constraints to restrict the possible non A-form suites with 93% certainty while still cutting the number of possible conformations in half.

Of course, NOEs are not the only piece of structural information a spectroscopist has when determining an RNA structure; J-coupling and ^{31}P chemical shifts can give a good indication of sugar pucker. If the sugar pucker is known, one can include this information in the lookup table, and narrow down the conformational space further. When puckers of both residues in the suite are known, only 5 of the 54 conformers will, on average, be valid choices for intersection subsets, and a slightly larger 8 for union subsets. With most of the conformation space pared down, building the structure using known RNA backbone conformers can be greatly simplified, particularly with regards to non A-form suites that are often difficult to determine by other means.

3.3.7 Discussion

This proposed method of using theoretical NOE constraints from ideal RNA backbone rotamers as an aid in structure determination is very promising. The number of matches between the model suite and the NOE-derived theoretical conformation subsets shows that even with only one or two NOEs, the number of candidate suites can be pruned from 54 down to a much more manageable number. This method requires that the spectroscopist needs only to know which two atoms are assigned for a given RNA backbone NOE and the NOE distance. Using this information, the spectroscopist chooses the appropriate lookup table based on the atoms involved, and inputs the distance constraint. The lookup table outputs the resulting conformer subset for that NOE, as well as a second conformer subset containing suites that are in the 10% NOE error region. These RNA backbone conformer subsets can be used as a guide for building the molecule or as a check when reviewing the results from a round of refinement, or validating a completed structure. Overall, the lookup tables could save time and lead to more accurate structures.

There are two overarching features in the lookup tables. The first is suite coverage—how many of the 54 suites are suggested by any given NOE constraint. Shorter NOE distance constraints, 3-4Å, lead to small numbers of suites in any given conformer subset; conversely, NOE distance constraints beyond 6Å often yield conformer subsets containing 30 to 40 possible suites. The specific RNA backbone atoms assigned to a given NOE have a significant impact on how conformationally

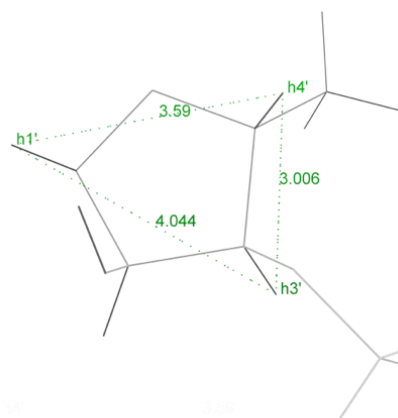
restrictive the assigned RNA backbone NOE is. For example, an NOE distance constraint of 4.5Å between H2'_{n-1} and H5'_n yields a conformation subset of 22 (29 including the 10% error region suites); on the other hand a constraint of 4.5Å between H1'_{n-1} and H1'_n yields a conformation subset of only 4 suites. This observation was implied by Varani (Varani 1996), but is made explicitly clear with these lookup tables.

The second feature of the lookup tables is that in ideal conditions, the modeled suite should match one of the suites from the NOE-derived conformer subset. When multiple RNA backbone NOEs are observed for the same suite, then their respective conformer subsets should be reconciled. This method does so by taking the intersection and the union of all conformer subsets for a given suite. In theory, the modeled suite would always be present in the intersection subset, but in practice the modeled suite matches the intersection subset 76% of the time for all structures, and 57% for non A-form structures. The reasons for this are many, as any systematic errors in the model building or in the NOE measurement could result in the modeled suite not matching any suites in the conformer subset. To alleviate this problem, the union of multiple NOE-derived conformer subsets is determined, lessening the impact of error in an individual NOE measurement. This brings the number of matches to 92% in all suites, and 93% for non A-form suites. Unfortunately, the union of multiple conformer subsets increases the total number of suites in the final subset, making the conformer reduction less effective.

When building a structure, the lookup tables serve to inform the spectroscopist which suites to consider first, namely, the intersection conformer subset. With only an average of 5 suites for a given residue, the intersection subset provides a selection of

RNA backbone conformations consistent with all available NOEs, and is small enough that each suite can be tried to find the best fit for the data. Only after the possibilities in the intersection set have been tried and dismissed should the union conformer subset be used. In this way, the intersection subset acts as a kind of upper bound, and the union subset a lower bound.

The removal of the 1F9L structure from the study was due to unreasonably short NOE constraints. The restraints placed on adjacent residues in 1F9L were as tight or tighter than restraints usually found within a residue. The differences between the recorded NOE (included in the .mr data file) and the actual model can be seen in Figure 3-25 below. 1F9L not only uses unreasonably tight NOEs, it also uses standard values such that an observed NOE always has the same distance restraint. This is true also of some other structures in the late 1990's. In some cases, there are large discrepancies between the deposited NOE distance restraint value and the value measured on the structure itself (an example is also found in a protein described in the next chapter).



```

{constraints for residue 5 from expt}.
assign (residue 5 and name h3')(residue 5 and name h1') 2.0 0.0 2.0.
assign (residue 5 and name h4')(residue 5 and name h1') 2.0 0.0 2.0.
assign (residue 5 and name h3')(residue 5 and name h4') 2.0 0.0 2.0.

```

Figure 3-25 - 1F9L residue 5: differences in NOE restraint distances (in text) and final model (as measured), distances, showing the problem with scaling.

Notably, the higher the clashscore, the fewer the suite matches in the test structures. In 2HGH, Model 1 has a clashscore of 0.95, and 100% of the model suites matched those predicted by the NOE constraints. In the case of 1F9L (clashscore = 47.35), it appears the excessively short NOE constraints made the overall structure too tightly packed and contributed to the many steric overlaps. An especially egregious structure, 1YMO (38% match), also had a very high clashscore at 129.59 (below, Figure 3-26).



Figure 3-26 - 1YMO, has clashes in almost every residue. The clashscore for this model was 129.59 (Model 1). Only 38% of model suites match those predicted by the NOE constraints.

Unlike 1F9L, some of the NOEs for 1YMO were reasonable and the model suites matched with their conformer subsets, despite the high clashscore. While most of the steric overlaps in 1YMO occurred between adjacent residues, the suites that matched the NOE-derived conformer subset eschewed clashes. These data suggest that structures with large amounts of clashes will have a large number of suites in the model that do not match those suggested by the lookup table method. This discrepancy disappears almost completely in areas with few clashes, where one can observe 80% to 90% matches between the modeled suite and the NOE suggested conformer subset from the lookup tables. This suggests that we can use discrepancies between the model suite and the conformer subset as an early indicator of potential error regions during the model building process and lends credence to the idea that using the suite conformer subsets to guide model building will in fact lead to better models.

The NOE lookup tables can also be used to diagnose areas where the structure should be refit. 1T4X is structure of Z-form RNA, and contains 48 backbone NOEs for its 12 residues, with no A-form suites present in the structure. The second suite in this structure is labeled as a !! by SuiteName and is defined by 7 NOEs. When looking at the NOE suggested conformer subset from the lookup tables, the union conformer subset for this suite contains 49 conformer possibilities. This is unlike the other suites in the structure, which have roughly 12 conformers in each union conformer subset. Part of the problem with the second suite of 1T4X is that no single conformation will fit all the NOE constraints, but there are many conformers that fit at least one of the NOEs. The most overlap is from conformation 5z, shared by 5 of NOE subsets. What may be occurring is the structure determination software is trying to compensate by fitting something in between, and the result is not within the set of acceptable conformations. Substituting the ideal 5z suite conformer into the structure shows that the 5z fits quite well into the structure, and small rotations in α and ζ are all that are needed to move the original suite to the 5z conformation. This choice is further bolstered by the fact that ideal Z-form DNA also contains the 5z suite conformer at this position.

When the intersection of 3D backbone conformer subsets determined by multiple observed NOEs does not match with the suite-string analysis of the model, there is either a problem with the structure model (or with the data), or with variability in the conformer not reported in the tables. In one sense, the rotamer could be an outlier if we believe that the data are correct and the structure model is problematic. In reverse, this could be an indication of an error in an NOE assignment, or a demonstration of the limitations of

accuracy in the data. Understanding what factors impact the range of values of the observed NOE is therefore a critical factor for this analysis.

There are two primary effects impacting the range of values around the observed NOE value that should be included when defining a subset based on an observed NOE distance restraint. First, the ranges of the 54 backbone rotamers around each dihedral will impact the NOE distances in our tables (how much ‘give’ on the value is acceptable—especially how short a given distance can get for each conformer); this is the rotamer contribution. Second, effects such as NOE scaling and how constraint bounds are set impact NOE distance restraint values used to define subsets of conformations for each observed NOE, plus the possibility of an incorrect assignment; this is the NMR-side contribution. A systematic evaluation of these effects will be important for spectroscopists intending to use this system.

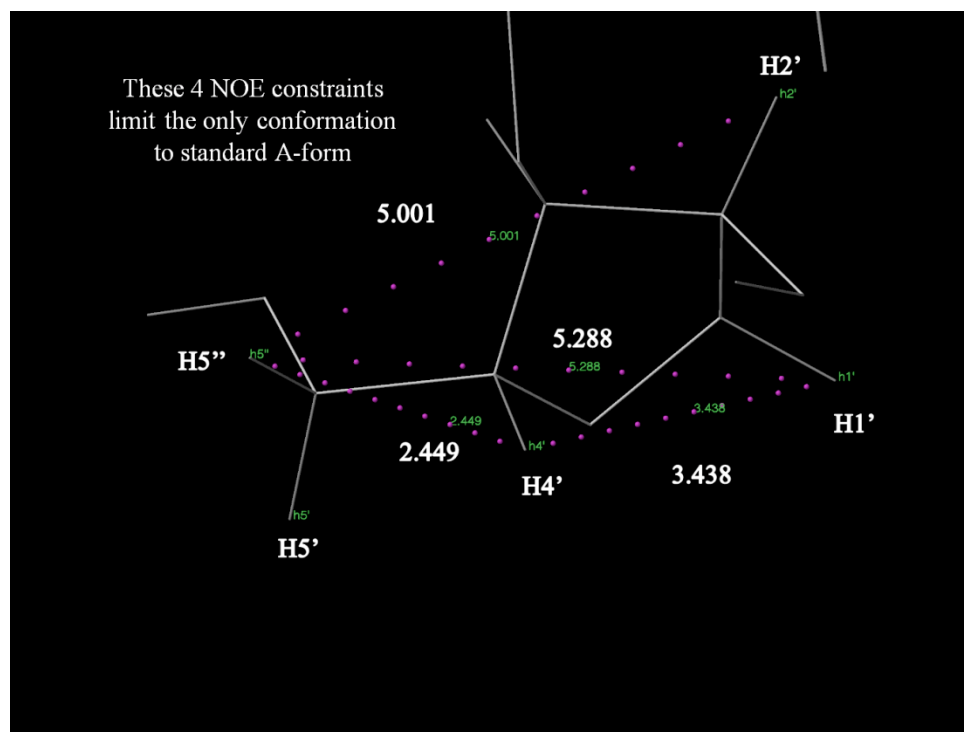


Figure 3-27 - RNA Backbone NOE restraints that can conformationally restrict the suite to A-form

While some sets of restraints may conformationally restrict a suite to A-form (above, Figure 3-27), the structure determination packages make this observation difficult for non A-form RNA. This is to be expected, since no current structure determination package uses RNA backbone rotamers at all. Additionally, NOEs are seldom observed for some atom pairs in the lookup tables, for very technical experimental reasons, and are at the limits of the experimental methods currently available. In short, it is a challenge to determine an RNA structure by NMR and create a realistic and high quality model.

3.3.8 Conclusions

Despite having contributions to an error model from rotamericities and scaling, the backbone in-suite NOE distance restraints conformationally restrain the backbone rotamer options for a given suite to a smaller subset of the 54 rotamers.

All these data show that RNA structures done by NMR are not particularly well characterized or standardly created. Future development of this work will include addition of the more common base-base and base-backbone NOEs which, while not as directly conformationally restraining to the backbone, are useful in coarse measures like determining sugar pucker. This information would then be correlated with the rest of the observed NOEs to constrain the structure further.

It is known that many suite clusters contain few examples of the “ideal” suite, and in fact hover more distant from the cluster center from which ideal dihedral values are derived. To improve the accuracy of suite assignments, it will be necessary to calculate the ranges of NOE constraints for each suite conformation. Additionally, bringing this information together into a single data quality metric for ease of use should be evaluated. One proposal would be to base it on the percentage of conformations fitting the particular residue’s NOE constraints.

3.4 Concluding remarks on the new tools and analyses

In the first and second section of this chapter, I describe new tools to visualize NMR structure ensembles of Protein and RNA. I investigated a large number of NMR structure ensembles in order to understand how the structure models behaved when

analyzing local criteria such as measures of underpacking (using steric clashes and mainchain Hydrogen bonding combined), backbone Ramachandran, and sidechain rotamer analysis. This produced a number of new tools and local criteria for identifying errors in NMR structures. These tools and visualizations were implemented in MolProbity, and made available to the biomedical community.

In the third section of this chapter I, describe new tools to visualize and analyze patterns of NOE data for building structure models of RNA. I investigated a small set of RNA structures with NOE's available in order to understand how structure determination might be improved by using a series of lookup tables to conformationally restrict the possible RNA backbone suites to conformations that are consistent with the observed data. These sets of possible RNA backbone configurations could used by spectroscopists to evaluate whether or not to include more restraints in the structure determination process..

Finally, the experimental data – NOE's – used to propose a measure of data quality for NMR structures (number of NOE restraints per residue in a structure) is just one type of NMR data used by structural biologists to build 3D models of macromolecules. Another type of data, Residual Dipolar Couplings (RDC's) was used in a few of the examples I investigated, and the importance of this type of data and what it could mean for analyzing and improving NMR structures will be described in greater detail in the next few chapters. The 1Q2N structure from the Montelione group plays a prominent role in the next two chapters.

4. Visualizing & Analyzing NMR Structure Ensembles in Virtual Reality

The collaborative nature of the work described in this chapter became so critical to the success of the project that it merits explanation. The level of complication involved in creating a piece of software implemented in a 6-sided CAVE environment requires the efforts of individuals from a variety of backgrounds each bringing expertise to the work. Put simply, this is impossible for one person to do alone.

I proposed this project to my thesis advisors David and Jane Richardson in Biochemistry and to the head of the Visualization Technology Group (VTG) – Rachael Brady – in the Pratt School of Engineering, both at Duke. Rachael provided unrestricted access to the DiVE (Duke immersive Virtual Environment) for the project, and welcomed me as a de facto member of her group, even sending me to the IEEE Virtual Reality conference in Charlotte.

I wrote the original VirTools prototype with help from Claire Vinson. This prototype, critical in the decision to create a full software platform, taught us all about the challenges of building software for virtual environments. Ian Davis, Claire Vinson, and I programmed related utilities needed for the early prototype. David J. Zielinski – a virtual reality programmer and engineer in the VTG – wrote the program KinImmerse and I worked closely with him on all details of implementation. Vincent Chen wrote RDCvis (described in a later chapter) and provided kinemages for a number of important examples in this chapter.

In collaboration with Dave Zielinski, and with assistance from Vincent Chen, David Richardson and Jane Richardson; I devised, performed, and evaluated the research usage tests over a period of two years. I conducted most of the teaching sessions and demonstrations with assistance from Dave Zielinski. As time went on, Dave Richardson began to conduct teaching sessions and demonstrations, though all along advised me and helped shape my thinking about how to do demonstrations.

Steve Feller in the VTG provided technical assistance, troubleshooting mechanical, electrical, and other problems. Steve was helpful in discussions Dave Zielinski and I had concerning technical implementation of features and interesting new directions.

David Stein, who is part of the Duke-Durham Neighborhood Partnership, provided perhaps the best set of testers of KinImmerse: hundreds of Durham Public School students of various ages who visited Duke through a laundry list of programs David developed and I assisted with.

4.1 KinImmerse: Macromolecular VR for NMR Ensembles

In molecular applications, virtual reality (VR) and immersive virtual environments have generally been used and valued for the visual and interactive experience -- to enhance intuition and communicate excitement -- rather than as part of the actual research process. This work develops and presents the software infrastructure needed to see whether that limitation can be overcome.

The Syzygy open-source toolkit for VR software was used to write the KinImmerse program, which translates the molecular capabilities of the kinemage graphics format into software for display and manipulation in the DiVE (Duke immersive Virtual Environment) or other VR system. KinImmerse is supported by the flexible display construction and editing features in the KiNG kinemage viewer and it implements new forms of user interaction in the DiVE.

In addition to varied and powerful molecular visualizations and navigation, the KinImmerse software provides an initial set of research tools for manipulation, identification, co-centering of multi-model ensembles, free-form 3D annotation, and output of results. The molecular research test case shown here is analysis of the local neighborhood around an individual atom within an ensemble of NMR models, as defined by the NMR experimental data, including target curves for residual dipolar couplings (RDCs).

The potential for production-level molecular research in the DiVE is very promising, and KinImmerse allows us and others to test the effectiveness of such a system.

4.1.1 Background

3D molecular structures and their visualizations matter because the progress of biology and medicine in the last century has steadily followed our ability to observe finer and finer biomolecular detail and then integrate it to higher levels of complexity. Although the information-storage capability of DNA arises from its sequence and the specificity of its base pairing, almost all other biological functions (catalysis, gene

expression, specific binding, cellular structure, growth, signaling, mobility, etc.) are due to the detailed 3D structural relationships in and between protein and nucleic acid molecules. Knowing those structural relationships is essential to understanding our own, and all the rest of, biology.

4.1.1.1 The Tools: Molecular graphics and virtual reality

Just as 3D structure is central to biomolecular function, 3D visualization is central to understanding those structures and functions. Macromolecular structure is complex, cooperative, handed, irregular, and mobile. For communicating specific structural concepts, static 2D images are only second-best, while the discovery of new relationships is enormously enhanced in interactive systems that fully explore the third dimension.

In the beginning days of structural biology, interactive computer graphics was not possible for macromolecules, and comprehension of the structures came from physical models that were labor-intensive, expensive, and susceptible to time and gravity. One of the most widely used types were the Kendrew-Watson skeletal models where the bonds are brass rods and the atoms are their joints. This shows connectivity and identity well, and is open enough to see and even reach into the center of a molecule, but does not convey volume and surface. Another widely used physical model was the Cory-Pauling-Koltun (CPK) space-filling model where each atom is a plastic sphere of slightly under van der Waals radius. The CPK models are colored by atom type and show surface shape and atom interactions directly, but obscure all interior information and connectivity.

Simplification was later available with C α backbone models bent from steel wire (e.g. Rubin, 1972).

The earliest molecular graphics on the computer used C α or all-atom "stick" representations, with a single drawn vector per bond allowing rotation on high-end machines of the time (Katz, 1972). With the advent of bitmap displays, the advantages and disadvantages of CPK models were reproduced in static gray-scale computer images (Porter, 1978), with real-time performance possible only much later. Such vector and CPK systems enabled analysis of active sites and recognition of structural motifs. Interactive fitting of crystallographic models into electron density contours was first achieved with the Grip-75 system at UNC Chapel Hill (Britton, 1978), then adopted in FRODO and later "O" (Jones, 1978; Jones, 1986; Jones, 1991) and others (e.g. Xtalview, McRee, 1999; Coot, Emsley, 2004); becoming an essential part of determining protein crystal structures.

With the advent of color, dot surfaces (Connolly, 1993) became the first widely successful innovation in molecular representation to originate on the computer side; in combination with stick models they allow the dual 1D and 3D nature of macromolecules to be visualized at the same time, bonded atomic interactions by the sticks and surface interactions by the dots. Ribbon schematics developed as 2D hand drawings for publication (Richardson, 1981) were adapted for 3D computer graphics in the 1980's (Carson, 1986), based on earlier ribbon-like drawings.

In the early 90's molecular graphics migrated to personal desktops and into classrooms, with kinemages and Mage (Richardson, 1992) or with RasMol (Sayle, 1995)

as well as others. Some awkward but temporarily necessary conventions have fortunately been superseded: half-bond coloring mostly replaced by ball & stick, and copious labels by identification when picking. David Goodsell's hand drawings had the feeling of depth with shaded edge outlines with a simple software version available (Chimera, Peterson, 2004) and a more elaborate one in QuteMol (Tarini, 2006). Several advanced techniques such as transparency and volume-rendering have been tried but have not yet found their ideal interactive uses or ideal implementation. The highest-quality rendering with ray-tracing, even now, is usually restricted to static images or frame-by-frame movies.

Many excellent software systems are currently available for interactive molecular graphics, such as PyMol (DeLano, 2002), DeepView (Guex, 1997), KiNG (Lovell, 2003; Chen, 2009), Chimera (Pettersen, 2004), MolMol with specific NMR features (Koradi, 1996), or Coot with specific crystallographic features (Emsley, 2004). The current state of the art on fast desktops or laptops -- with flexible representations, high resolution, stereo, manipulation and calculation features, and smooth rotation for full models as big as the ribosome -- is an enormously effective tool in routine use by all structural biologists and biomedical researchers.

The most interactive and immersive of computer graphics are those that use virtual reality techniques, such as head-mount displays and trackers, force-feedback, tracked gloves or wands, multi-sense modalities, wall-size displays, and surround-projection CAVEs (Sherman, 2002). Such immersive virtual environments have made a great impact in many fields from gaming to surgery, but have so far seen only limited use for macromolecular structures.

The first multi-wall CAVE virtual environment was developed 17 years ago in Chicago at the Beckman Institute (Cruz-Neira, 1993). It was very early made to display molecules, using a plug-in version of the VMD molecular-dynamics and display software (Humphrey, 1996). Molecular displays are part of the demonstration repertoire, but production-level work in CAVE systems has been almost entirely in other subject areas.

A group at University of California Irvine modified MolScript on an SGI Onyx to display molecules on the 4-screen CAVE at Mississippi State (Moritz, 2004). That system can show, move, and compare ball & stick or ribbon representations of several proteins at once, floating above a gridded floor.

A modification to VMD was made to put it into a CAVE for the display of protein structures by using Sherman's FreeVR library (Sherman, <http://www.freevr.org>). The VMD software package is sometimes used by the general biomedical community for simple viewing and animations, but most of its users are researchers who run molecular dynamics simulations, and the more sophisticated functions in VMD are geared towards those users. The VR VMD could not be used for our study, because it does not support cluster-based projection systems such as the DiVE and is neither updated, nor freely available.

Over the last thirty years the computer science department at UNC Chapel Hill has contributed to virtual reality systems for visualizing and investigating molecules (Britton, 1981; Brooks, 1985; Bergman, 1993; Surles, 1994), including headmount displays (HMD's; Chung, 1989), a two-wall display for joint molecular graphics work by two people (Arthur, 1998), and especially focusing on force-feedback devices (Brooks,

1990), which they have developed into a highly effective system for actual physical manipulation of molecules in conjunction with atomic-force microscopes (Taylor, 1993; Fisher, 2005).

A report in the Protein Data Bank newsletter (PDB, 2006) describes a system called PDB in a CAVE, built on the COVISE platform and providing virtual-reality displays of proteins and nucleic acids in ribbon and other representations from web-downloaded PDB files, including animation capability. So far it has been used interactively in demonstration mode with large wall displays, but claims to be capable of display in CAVEs. Amira, a commercial software package, also provides support for molecular viewing of PDB files in CAVE-type systems, though the plugin is not distributed standardly with Amira.

All of the above systems provide effective molecular visualization in immersive VR. However, they all use standard representations (usually ribbons and CPKs), many are not open source and hence not modifiable for new uses, only Amira supports measurements, and as far as we can tell none of them have any tools for model manipulation, annotation, output, or other research interactions. In recent years, most of the early limitations to VR and especially CAVE systems (low resolution, slow rendering, tracking latency) have been overcome, creating the opportunity to try for production-level molecular research applications. This project develops open-source software with very flexible display creation and suitable interface tools for molecular-structure research within the 6-surface surround of the DiVE or other VR system.

4.1.1.2 The Application Area: Macromolecular structures and NMR ensembles

The 3D coordinates and experimental data for macromolecular structures are made publicly available in the international Protein Data Bank, or PDB (Berman, 2000), which currently contains over 60,000 entries. There are two principal experimental techniques for determining these structures in atomic detail: x-ray crystallography and NMR (Nuclear Magnetic Resonance spectroscopy). Both depend on constructing a molecular model consistent with many thousands of individual measurements and also with bond lengths and angles known from chemistry and with the amino acid or base sequence. The logic of the two methods differs, however -- crystallographic data directly give position in 3D space but atom identities must be inferred, while NMR data show local distance or angle relationships between identified atoms but their positions must be inferred. The two types of NMR data central to 3D structure determination are: 1) the NOE (Nuclear Overhauser Effect) that measures through-space distance between two atoms closer than about 5Å, and 2) the RDC (Residual Dipolar Coupling) that measures the angular relationship between a specific interatomic bond vector and the magnetic field direction in partially-ordered experimental samples (Cavanagh, 2006). The RDC value measured for a specific pair of atoms places that bond direction somewhere along a symmetrical pair of ellipse-like target curves (see results). Crystal structures typically consist of a single model, with a “B-factor” estimate of positional uncertainty. NMR structures are reported as ensembles of multiple models each of which is consistent with the data; differences between the models can result either from incomplete data or from real motion in the molecule (Cavanagh, 2006).

Both crystal and NMR structures are very reliable, especially those with the most experimental data (high resolution for x-ray, many restraints per residue for NMR), but both are susceptible to local errors that can hurt their uses for other biomedical research. We have contributed new methods for the diagnosis and correction of local misfittings in protein and RNA structures (Lovell, 2003; Word, 1999a; Davis, 2006; Davis, 2007), which have proven highly effective in routine crystallographic use (Arendall, 2005; Chen 2010). We would like to develop related methods suitable for improving the accuracy of NMR structures. That is a harder task because of the less direct relationship of the experimental data to 3D space, the complication of multiple models, and a tradition of determining and analyzing the models computationally rather than visually, and globally rather than locally. Good molecular graphics for NMR ensembles are especially challenging, and neither display of RDC target curves on the model nor the explicitly local perspective suggested by the local nature of the experimental data are currently used. We believe immersive VR display should be particularly effective for achieving and utilizing that local perspective.

4.1.2 Methods

4.1.2.1 VR hardware

These displays were shown in the Duke Immersive Virtual Environment (DiVE), a 6-sided, fully immersive VR system approximately 2.9m x 2.9m x 2.9m. The walls are flexible screens with wooden and acrylic frame. Images on each of the six sides currently have 1056 x 1056 resolution, with stereo-switched between eyes at 110 Hz. The ceiling

and floor are rigid acrylic, 20mm and 50mm thick, respectively. The door opens manually by sliding and the walls are removable for screen replacement.

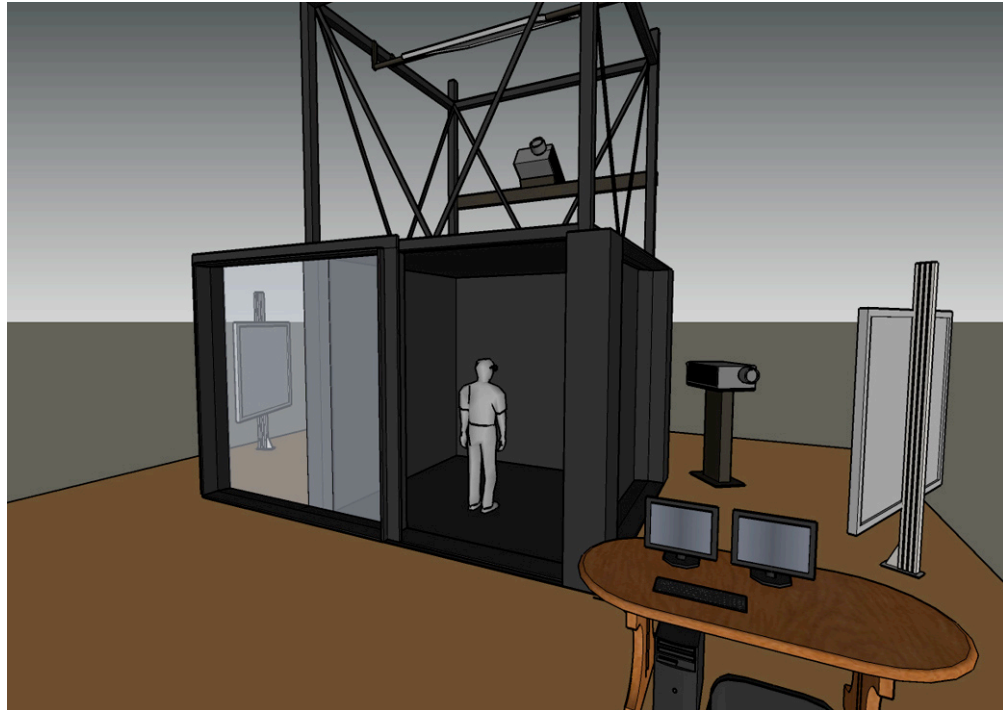


Figure 4-1 - The DiVE

Active stereoscopic vision is enabled through the CrystalEyes system from StereoGraphics Corp., including a master pair of stereo goggles that is head-tracked to set the viewing point and direction. A hand-held 3D mouse or 'wand' from InterSense Technologies includes a joystick and four button controls on the face and one button control on the underside of the wand. The InterSense IS-900 inertial/ultrasound tracking system determines the position and orientation of the stereo glasses and the hand-held 3D controller.

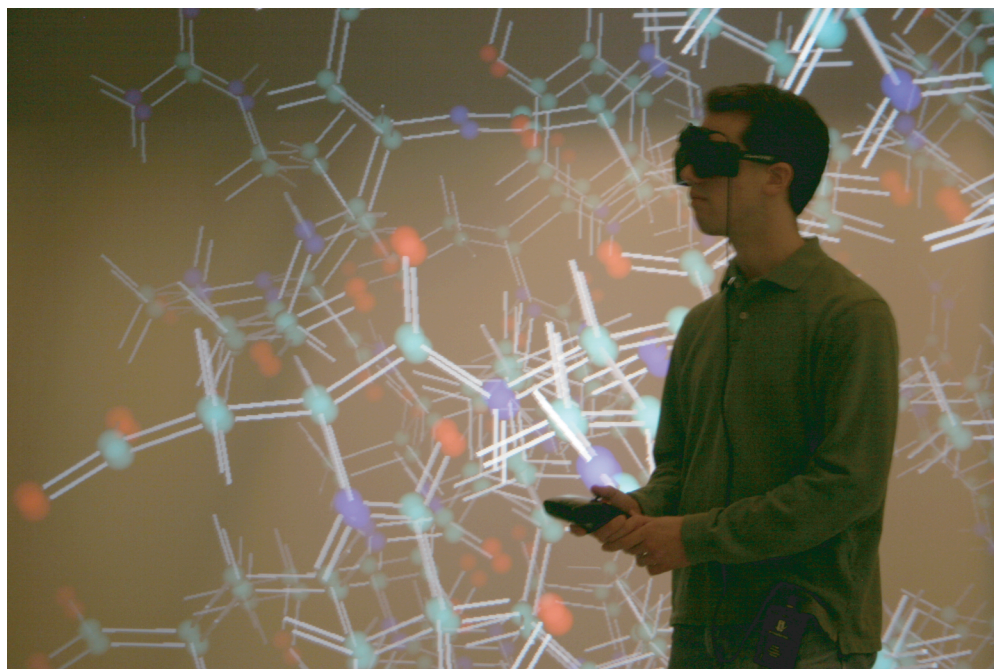


Figure 4-2 - User with InterSense head tracker and 3D controller

The molecular coordinate data for VR display in KinImmerse typically originates in files downloaded from the PDB (Berman, 2000). The PDB-format file is parsed either on the MolProbity web site (Davis, 2007; Chen 2010), or in the Prekin-Mage system (Richardson, 1992; Richardson 2001), or in the KiNG kinemage display program (Lovell, 2003; Chen, 2009), where molecular representations can be produced, modified, and then output in the kinemage graphics format (Richardson, 1992), a hierarchical, human- or machine-readable text file format for the display of various graphics primitives. KiNG (in Java) and Prekin-Mage (in C) are open-source, multi-platform, and available from <http://kinemage.biochem.duke.edu>.

The first proof-of-concept system for displaying molecules in the DiVE was implemented using a modified output from KiNG and a VirTools (Dev 3.5; <http://www.virttools.com>) application that constructed VR displays from the KiNG

output. That system showed the feasibility and promise of the kinemage-to-DiVE route, but required a commercial software system (VirTools) and had the limitation, traditional in VR, of using only surface-graphics primitives and not points or lines. For example, a covalent bond vector, which would be represented by a line in normal kinemage format, became a narrow 4-sided cylinder in the VirTools application. Those limitations are overcome in the present KinImmerse program, using the open-source Syzygy toolkit and implements both surface and line graphics objects. KinImmerse reads kinemage files directly, and the current version has implemented recognition of a large portion of the kinemage format (such as the group/subgroup/list heirarchy, pointID information, colors and line widths, ribbons, surfaces, etc.). Input to KinImmerse is therefore a kinemage-format file, either a pre-existing one or one created to suit the current VR objectives.

4.1.2.2 The Syzygy toolkit

In order to support a variety of immersive virtual reality systems, the software programming toolkit "Syzygy" (Schaeffer, 2003) was utilized for developing KinImmerse. It provides an abstracted interface to the programmer so that regardless of the particular display system, tracking system, operating system (support for Linux, MSWindows, MacOSX), or number of networked render nodes used, the application itself does not have to be modified. By instead using XML configuration files, Syzygy provides display support for head-mounted displays, cave-type systems, or tiled display walls, and also a desktop simulator mode useful for development. In order to facilitate the best immersive experience, a head and hand tracking system is often utilized. Syzygy

directly supports a number of tracking systems, as well as many more through its interfaces to VRPN and Trackd. VRPN is a public domain library which provides a device-independent and network-transparent interface to virtual-reality peripherals (Taylor, 2001). Trackd is a commercial library which "takes information from a variety of tracking and input devices and makes that information available for other applications to use". [<http://www.vrco.com/trackd>]

4.1.2.3 Callback-style API

The Syzygy toolkit is easy to use, since its interface is somewhat similar to the GLUT API application programming interface (Kilgard, 1996), which is familiar to most programmers in the OpenGL community. In order to convey what work was necessary to achieve the final KinImmerse program, we will discuss this interface in more detail. A self-contained program (not using GLUT or Syzygy) could have a linear flow of custom-written steps as follows: (1) read tracking/sensors, (2) update-world (based on sensor data), (3) set viewing transform, (4) draw-world, (5) repeat back to (1).

The Syzygy callback-style API can provide step 1 of reading the sensor data, step 3 of setting the viewing transform based on head position and screen geometry (as specified in a text configuration file), as well as the looping functionality of step 5. To complete this call-back system, the programmer need only write the registered update-world function that decides what objects to modify in the scene (step 2) and the registered draw-world function using OpenGL calls (step 4), for the API to call as needed. Thus we need not be concerned with the specifics of the tracking system used or the screen

geometry. Syzygy takes this a step farther by allowing us to hide the fact that we are running the application over a cluster of computers. The complete Syzygy version functions roughly as follows (with functions in thick-lined boxes custom coded, though distributed across N clients, as shown in the next figure):

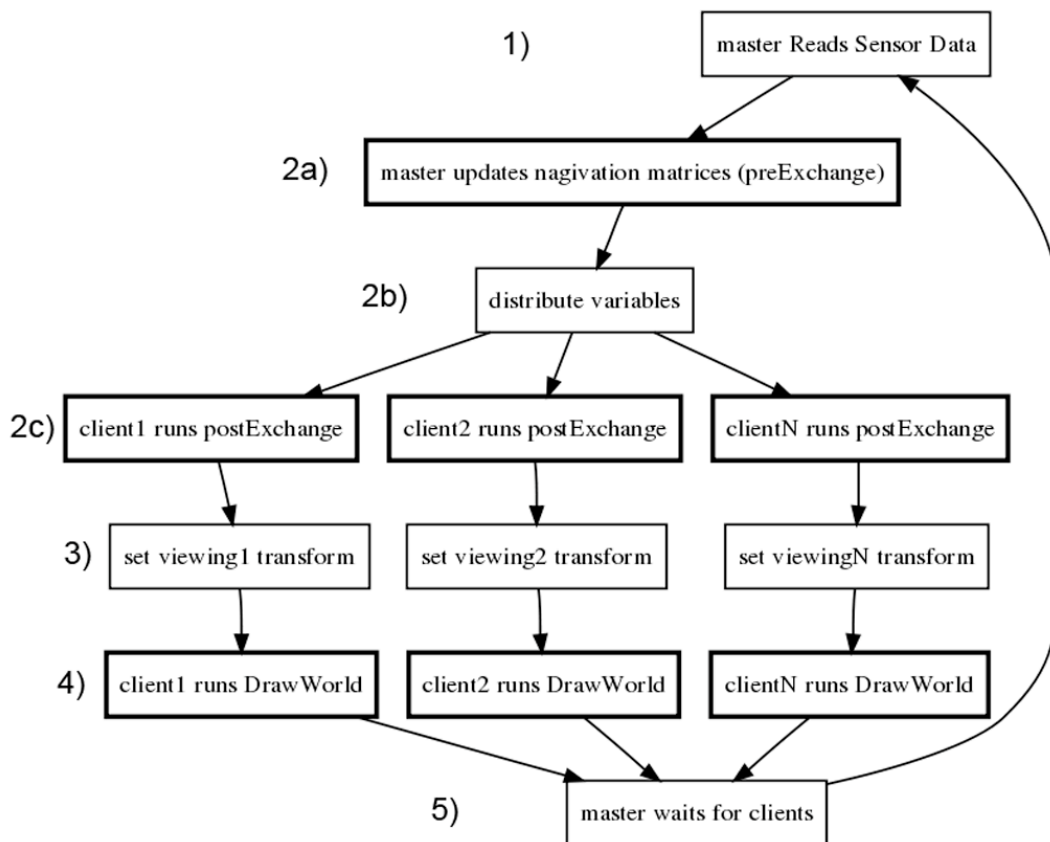


Figure 4-3 - Flow chart of the KinImmerse logic

4.1.2.4 Data structure design

The internal data design of KinImmerse mimics the heirarchical nature of the kinemage format. As the kinemage file is loaded, a series of containers (groups, subgroups, and lists in kinemage terminology) along with objects (points, lines, spheres,

etc.) are created and added to their proper parent objects. There are some differences between the levels in kinemage format (e.g., animation is done on groups, and individual display objects are in lists). These properties are implemented as extra restrictions on the KinImmerse containers, to ensure that output files produced from the DiVE are compatible with later display on standard single-screen systems using KiNG or Mage.

The kinemage format also has a cross-cutting system of "masters" that can control display-object visibility with flags that can occur on any object at any of the container levels. Masters will need to be implemented as a special case in KinImmerse, since they are orthogonal to the major kinemage hierarchy. This will be part of a more general future rework of the menuing system.

4.1.2.5 Hierarchical bounding boxes

Another advantage of using the hierarchical internal data representation is ease of integration of axis-aligned bounding boxes. At each item (whether container or object) we compute a bounding box for that item. This is quick to recompute, and in practice is only recomputed for the whole structure after a rotation/translation/scale operation is completed. For wand detection of the closest object to the wand tip, we have the defined bounding boxes for all objects and use a collision detection algorithm which checks to see if the current wand point is inside the container's bounding box, before descending to check if the wand point is inside any of its children. This potentially speeds up the collision detection process by eliminating checks on objects that, as a group, are outside of the wand point.

4.1.2.6 File format versus interactive display features

The single-screen kinemage display programs Mage and KiNG both can read, interpret, and write out essentially all aspects defined in the kinemage format. Their basic on-screen functions are very similar, but they implement a somewhat different set of functionalities for user editing and manipulation, such as the separate dockable, internally rotatable molecular fragments in Mage (Murray, 2005) or the "backrub" tool for local shifts of protein backbone in KiNG (Davis, 2006). Analogously, the new KinImmerse application needs to read, interpret, and write out all basic aspects of the kinemage format, but its interactive display and interaction tools can be quite different, as either required or enabled by the VR environment.

4.1.2.7 Kinemage construction for the test applications

The kinemage graphics files for demonstration use (ribbons, ball&stick, and space-filling) were produced by the Molikin feature of KiNG, with some minimal on-screen or in-file editing to optimize display in the DiVE. The dot surfaces of all-atom contact analysis were calculated in MolProbity (Word, 199a; Davis, 2007). The primary NMR example is the 10-model 1D3Z structure of ubiquitin (Cornilescu, 1998), solved from unusually complete NOE and RDC data. The 1Q2N Z-domain ensemble (Zheng, 2004) and the 2I5O polymerase γ -domain ensemble (Bomar, 2007) were also used. The multi-model ensemble kinemages were produced by the MolProbity site's standard procedures for analyzing NMR structures (Davis, 2007). Approximate starting global

superposition of the ensemble models was done with the docking tools in KiNG, since the models diverge significantly in the deposited 1D3Z PDB file. Color-coded dotted lines to represent NOE data are produced by the NOEDisplay plug-in to KiNG (Coggins).

The representation of target curves for RDC data was produced by the RDCDisplay plug-in to KiNG implemented specifically for this project. In order to generate target curves, RDCDisplay requires a set of experimental RDC values and PDB-format coordinates of a model. From this information, it calculates a Saupe alignment tensor (Cavanagh, 2006) and uses the tensor to determine the quartic equations of the RDC curves for each internuclear vector. Each pair of target curves is drawn as polygonal curves on a sphere centered on one of the atoms of the internuclear vector, and they represent the possible orientations of the vector that are compatible with its experimentally-measured RDC value.

Using the co-centering feature built into KinImmerse to translationally superimpose the different models of an NMR structure (with RDC curve visualizations) on a single atom allows users to see instantly which models have internuclear vectors that do not line up on the correct RDC curves, indicating a region in which those models do not fit the experimental data and possibly have errors.

4.1.2.8 Free Open Source Software

Finally, as we desire creating a tool that can benefit and be accessible to as many as possible, we have made the KinImmerse system dependent only on libraries/toolkits

that are open-source (e.g. Syzygy, VRPN, KiNG), as well as releasing the application itself under a BSD-style open source license.

4.1.3 Results

The KinImmerse system can directly parse nearly all kinemage graphics files and display them on the 6 sides of the DiVE or in other VR systems. Representations include ribbon schematics, stick figures, ball-and-stick, space-filling spheres, electron density contours, dot surfaces, and various abstract notations that use simple graphics primitives (such as symmetry or helix axes, 3D scatterplots of data, or the NMR data representations described in this paper). After user interaction, the modified kinemage file can be written out again, for later use in KinImmerse VR or in Mage or KiNG single-screen or web-based graphics. Loading and saving of files is controlled through a separate Java GUI on the command console.

For general molecular displays in the DiVE we have shown ribbon, ball&stick, and space-filling representations of proteins and nucleic acids. Protein/protein or protein/nucleic acid interactions have been shown as polygonal Voronoi surfaces (Ban, 2004) along with all-atom contact dot surfaces (Word, 1999a). Correction of individual residue conformations in protein crystal structures has been shown by before-and-after kinemages with electron density contours, all-atom contact dots, and local stick figures in a ribbon context. Our major test application -- NMR ensembles and data -- is described below. For all of these subjects and for both novice and expert users, it was found anecdotally that the KinImmerse display provides significantly better perception of the 3-

dimensional relationships than motion plus stereo in single-screen systems. This is especially true for the master user, but also holds for the other viewers.

4.1.3.1 User Interface

The interface implements a mix of metaphors from virtual reality and molecular graphics, optimized for domain-specific interactions that enable scientific insight into the 3D structure of biological macromolecules. The dominant schema is that of person-centered control. All users have stereo goggles and can move about freely. The master user wears the head-tracked goggles that control the position and direction of view, navigates and controls mode with the hand-held physical controller, and points its virtual wand to select or grab objects. The InterSense handheld controller acts as a 3D mouse, topped by a central joystick and a crescent of four buttons (red, yellow, green, and blue) in easy thumb reach. The virtual wand appears as a white pointer stick projecting forward from the controller.

As well as moving about in the room, the user navigates by a flexible, gaming-style point-to-fly navigation. Pushing on the joystick flies the user through the scene in the direction it is pushed, relative to the direction the handheld controller is pointing. Typically the joystick is pushed forward, to fly toward where the wand is pointing. More push moves faster. For extended work sessions, the user can navigate virtually from a “command chair” with padded feet to protect the DiVE floor, which also allows the use of notes, laptops, or multiple VR interface devices.



Figure 4-4 - InterSense Ultrasonic 3D Controller

The display objects (that is, the molecules) can be manipulated in several different manners. Holding down the trigger button on the underside of the wand is a "grab" function that locks the graphics to the wand. This enables full 6-degree-of-freedom orientation of the graphics image, since the ultrasonic tracker follows the translational and rotational position and motion of the handheld controller. This feels like grabbing the molecule with your hand and turning it about. While the green button is pressed, pushing the joystick forward or back scales the display larger or smaller. This allows the user to zoom in or out of the display. If an animation pair or sequence is defined in the kinemage (usually to show conformational changes), pressing the red button advances by one step in that sequence.

A KinImmerse menu to show or hide elements of the kinemage is presently provided as a fixed menu list along one edge of the left side wall. It uses a direct hit of the virtual wand tip on the menu item to change its state. On/Off control of graphics

objects is essential for serious production work such as the NMR analyses, and is also very convenient for demonstration or teaching mode.

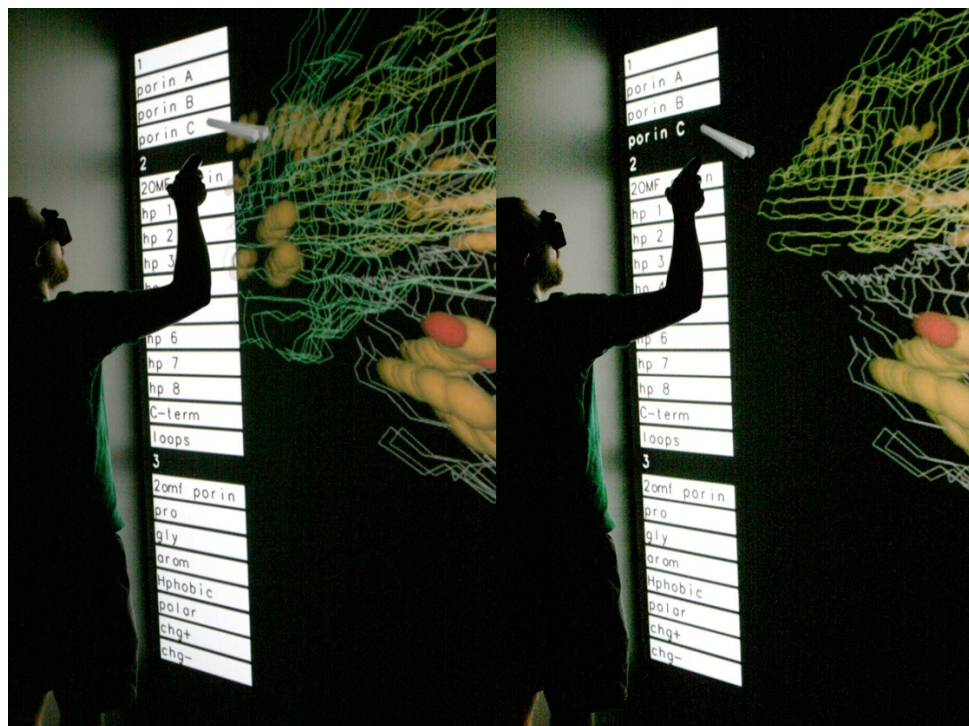


Figure 4-5 - User toggling a subunit off using the menu in KinImmerse (left image on, right image off)

While the user touches the tip of the virtual wand to a point of interest (an atom in our examples), a bounding box appears around the object and the object's identifying information is shown in one corner of each screen. The content of such information for an atom usually includes molecule, residue type, sequence number, and atom type (plus model number if NMR), but is freely specifiable in the kinemage.

Holding down the blue button enables user 3D annotation: drawing a freehand line in 3 dimensions with the tip of the virtual wand as it is moved. The skillful user can write text as well as draw 3D glyphs. These annotation marks become a part of the

kinemage, so standard desktop kinemage programs can be used later to view annotations made while immersed in VR space (see below).

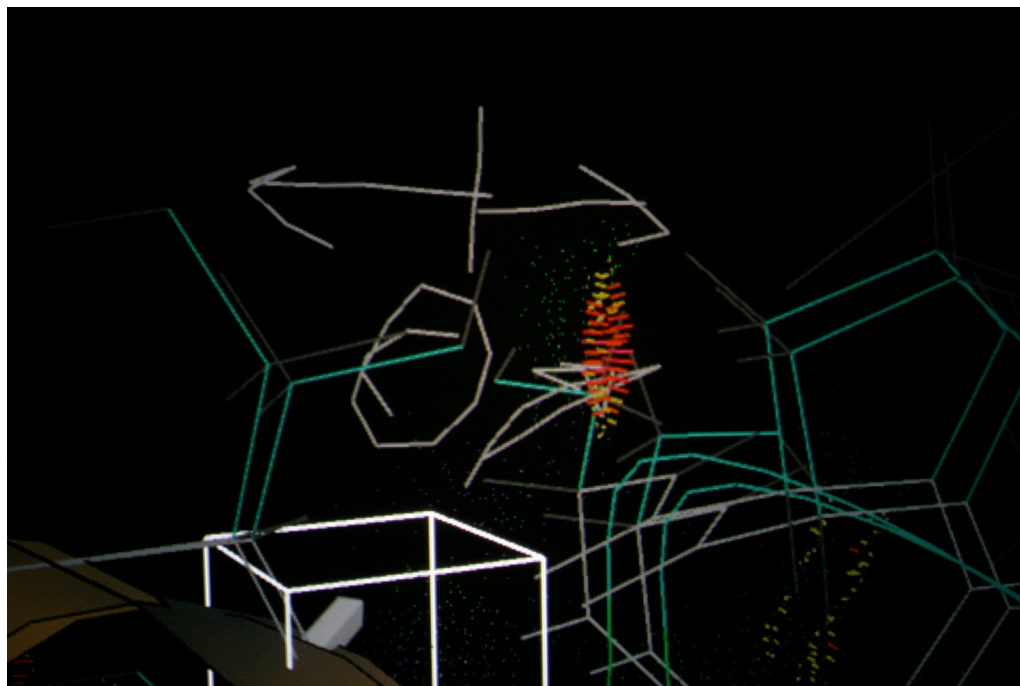


Figure 4-6 - Annotation diagnosing a steric clash in KinImmerse

Hitting the yellow button instructs the system to translationally co-center preselected mobile groups (such as NMR models) onto the picked point. The program identifies which points to co-center by their sharing a common string with the name of the selected point. This requires a specially constructed kinemage and illustrates the power of using an authored display descriptor such as the kinemage format. For the NMR examples under study the co-centered points are atoms of the same name in different models of the ensemble, with a set number of characters kept identical in the point names. However, this function could work on any kinemage that has points with well-behaved names.

4.1.3.2 The Local Perspective on Models and Data in an NMR Ensemble

The central test of KinImmerse as part of the research process is visualization of NMR structural ensembles and local analysis of the relationships between the NMR experimental data and the models derived from those data. Currently KinImmerse supports representation of NOE distance data as dashed lines and of RDC orientation data as pairs of target curves (see Background for their meaning and Methods for their production).

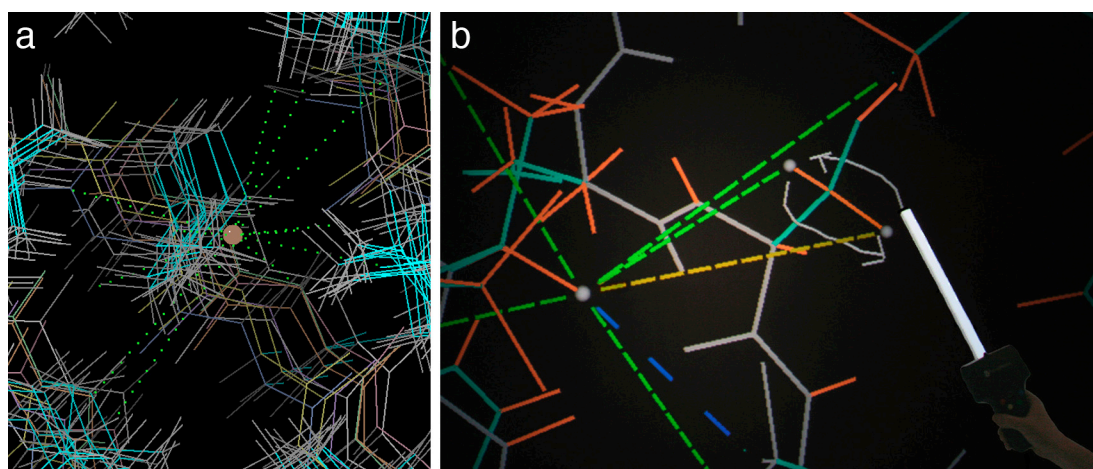


Figure 4-7 - NOEs for Ile3 H β of 1D3Z, in KiNG vs KinImmerse user session

Shown in panel a of the figure is a standard visualization (in KiNG) of the set of NOEs observed for the C β H atom of Ile 3 in the 10-model ensemble of the high-accuracy 1D3Z structure of ubiquitin, as deposited. In the DiVE, this system was initially viewed in the prototype VirTools display, where both orientational and translational superposition of models was done in the DiVE by interactive 6-degree-of-freedom docking onto a chosen reference model. We learned that the 1D3Z ensemble is extremely tight once globally superimposed, certainly understandable for models optimized against many NOEs and sets of RDCs in two different media, but not evident

in the deposited ensemble. Interactive translational centering on an atom of interest was found to be extremely valuable in assessing the model-to-data relationships within a local region. Therefore in the current KinImmerse system, global superposition of models is done as a pre-processing step if needed, and a new translational “co-centering” operation was implemented as part of the user interface (see Methods). For the working session illustrated in panel b, the 10 models were all co-centered on the H β of Ile 3, but only one model is shown in the figure for clarity on the page.

For studying NOE data in the DiVE, one can locate any restraint violations, color-coded in red (or near-violations in yellow), in the whole ensemble or across one model. In an area of interest, one co-centers the ensemble on a particular H atom for detailed local analysis. 1D3Z has no restraint violations, but we noticed that the relative NOE intensities observed between the central Ile 3 C β H and the two C β protons of Ser 65 are reversed relative to the respective interatomic distances in the model. In panel b, the user has drawn two arrows with the 3D annotation tool to record that observation. This type of minor discrepancy could have three quite different origins: a reversed resonance assignment, a different sidechain χ_1 conformer for Ser 65, or a quite plausible 1Å measurement uncertainty. Co-centering on Ser 65 C β to check its other NOE data, we found that the relative distances for sensitive pairs (from 1H β vs 2H β , to 65 NH, to 66 NH, and to Phe 45 1H ϵ /2H ϵ) are all neatly reversed by about 1Å as well, which would be unlikely from independent measurement errors. Modeling the other two Ser 65 rotamers in KiNG (not yet implemented in KinImmerse) shows them to have less optimal but not impossible sterics and H-bonding, and to only approximately reverse the NOE distances.

The consistent pattern of relative distance reversal is therefore most compatible with reversed assignments of 65 $1H\beta$ and $2H\beta$, and could prompt re-examination of the evidence for those assignments.

4.1.3.3 Visualization of RDC data and relationships

The below figure shows the pair of target curves for the RDC of a specific NH bond vector in one model of an NMR ensemble, color-coded by model-to-data agreement (see Methods). They lie on a sphere centered on the N atom; if the H atom lies at any point on one of the curves, then its back-calculated RDC would exactly match the observed value – which is very nearly the case for all 10 models in this example.

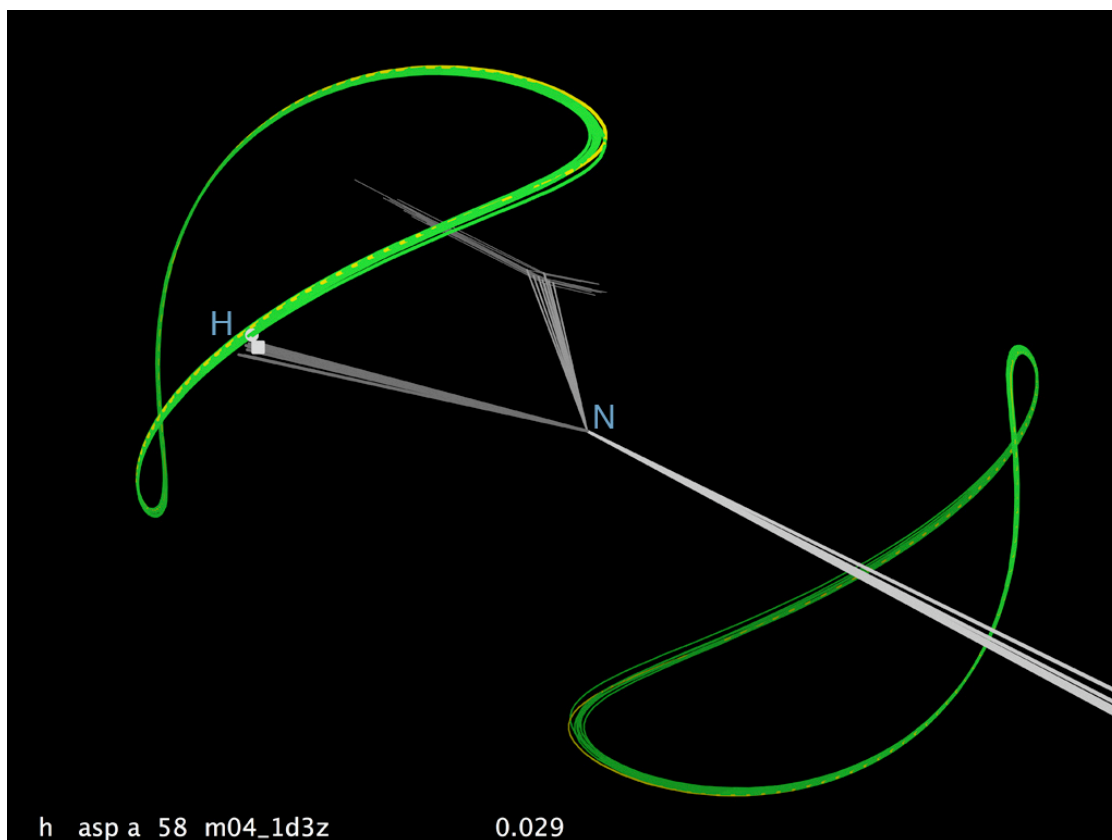


Figure 4-8 - NH RDC curves, co-centered on the N atom

Other possible RDC information can be calculated by RDCDisplay and viewed in KinImmerse. A curve can be back-calculated from the NH vector orientation in each model, with two outer curve sets showing the results of an RDC either +1Hz or -1Hz from the measured value. RDC measurements can be very precise, and in many cases the strips of probable orientation on the sphere are quite narrow. However, the RDC equations are highly non-linear, and in some cases a small change in RDC values can encompass rather large changes in model bond-vector orientation.

In analyzing the 1Q2N ensemble for the B domain of Staphylococcal protein A (Zheng, 2004), one sees that including RDC data improved model-to-data agreement and other quality criteria (Huang, 2005; Davis, 2007) for the backbone compared with the earlier 1SPZ and 2SPZ structures. But the helix 2-3 loop in 1Q2N shows two quite different conformations in nearly equal numbers for residues 37-39, with no restraint violations. We would like to understand how both conformations can fit the RDC data, and if possible to decide whether one of the conformational groupings is wrong or whether the two groups together represent a valid ensemble.

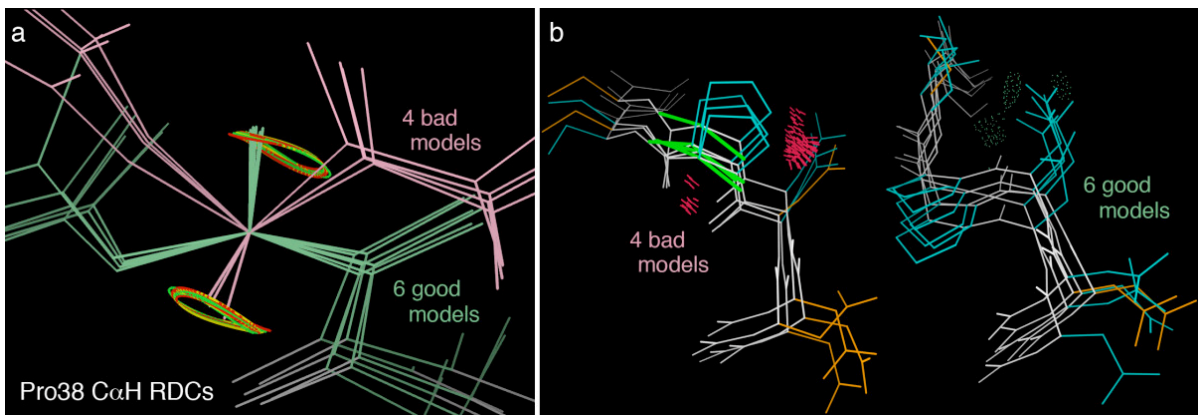


Figure 4-9 - Evaluating two clusters of loop models by RDC geometry

Panel a of the figure above shows this loop region of the 1Q2N ensemble backbone in KinImmerse co-centered on the C α atom of Pro 38, with the C α H RDC curves on each model. The two clusters of different loop conformations are colored pink and green (done in KiNG). Residue 38 is a proline which therefore has no NH, and the 37 and 39 C α H RDCs were not observed, a level of data incompleteness not uncommon in loops. The ensemble sampling identified unique conformational clusters for residues 36 and 40, anchoring the loop ends at the C-cap and N-cap of the two helices it joins. The largest conformational difference is at Pro 38, and it can be seen the figure that the C α -C α H bond vectors point in nearly opposite directions for the two model clusters and that their H atoms lie on opposite branches of the RDC target curves. The fact that the two groups of model conformations both match points on different target curves is problematic, because such a match would result in a different experimental observation. Other local NMR data is relatively sparse and has already been taken into account in the structure-solution process, but we can obtain an independent assessment from the MolProbity structure validation site. Panel b of the figure shows the multi-criterion MolProbity kinemage (Davis, 2007) for this region of the 1Q2N ensemble, with Ramachandran outliers in magenta, sidechain rotamer outliers in gold, all-atom steric clashes as clusters of hot pink spikes, and hydrogen bonds (H-bonds) as lenses of pale green dots. The conformational cluster of models 2, 3, 7, and 8 (with Rama outlier Pro 38) has many clashes and rotamer outliers and no H-bonds, while models in the other conformational cluster (with good 37-8 Rama values) are nearly free of outliers and make

several H-bonds. The conclusion, therefore, is that only the second model cluster represents a valid conformation for this loop.

In addition to obtaining specific new information about the model-to-data relationships in NMR structural ensembles for two different proteins, we were convinced of the value of the co-centering operation pioneered in KinImmerse, which has now also been implemented as a tool in the traditional KiNG display system. Such propagation of new visualization and interface ideas out into non-VR software is another kind of benefit from exploratory visualization research in systems such as the DiVE.

4.1.4 Discussion

4.1.4.1. Demo mode

KinImmerse (and the earlier prototype VirTools system) was immediately successful in demonstration (“demo”) mode, engaging and exciting viewers from all levels of sophistication, who enjoyed flying through into the major and minor groove of DNA or having the β -barrel ribbon of a membrane-pore protein pulled down over their heads. Pharmacology and engineering classes were motivated by having their projects end up as kinemages in the DiVE. In a Howard Hughes "Phage Hunting" summer program for at-risk High School students, the DiVE session was their highest-rated section of the program. Such demo and educational uses of VR are certainly not new to this application and are not its major purpose, but they are a very worthwhile side benefit.



Figure 4-10 - High school students experiencing a molecule in the DiVE

4.1.4.2 Differences from single-screen molecular graphics or from most VR

KinImmerse is a combination between the conventions of single-screen molecular graphics and usual VR applications, and thus differs in some ways from either.

Navigation and viewpoint are quite unlike the window-into-a-box of traditional molecular graphics, where one centers on a point by name, by preset views, or by picking and zooming, and where stereo and display-object motion by dragging give a good perception of depth, but the viewer is always on the outside. In VR the viewer is inside the world of the display, and walking or 3D flying to a point of interest is the dominant metaphor. In KinImmerse the user can move about freely in the room/molecule or can push on the joystick of the handheld controller and fly (virtually) in that direction. Up to 5 or 6 users

can be inside the DiVE together, all with stereo goggles but only one with the head-tracked master goggles and the hand-held controller. The controller has a virtual wand extension whose tip can directly select points in 3D or grab and manipulate display objects, giving the user very active and tactile control.

On the other hand, KinImmerse supports line and dot graphics, in addition to the triangulated surfaces that are the assumed primitives in nearly all VR displays. The information text shown when a point is selected goes beyond purely visual presentation into more specific and quantitative interactions with the displayed molecules, while the command-chair mode makes extended work sessions feasible. As far as we know, freehand 3D annotation, co-centering, and local display of RDC curves are all novel functionalities in either single-screen or VR applications. We believe this unusual combination of immersive context and detailed control should enable enhanced types of molecular research.

4.1.4.3 Future Directions

Display of information text in the DiVE corners is very effective, and distance and other measurements can be added to that mode. The current menu system is workable but not graceful, and other alternatives should be tested. Similarly, there should be further optimization of the lighting model for different molecular representations. There will probably need to be either a way to pre-specify a view/location or a way to move to specified features.

For the NMR work, developing ways to represent other data types, such as dihedral-angle restraints, is a logical next step. On the hardware side, resolution and fill rate will increase with installation of a new computer cluster. Most importantly, KinImmerse enables further exploration of using experimental data display and local perspective for research analysis of NMR structural ensembles.

Another line of work that could reasonably be followed up on is that of integrating the structure model fixup capabilities of KiNG into KinImmerse. To facilitate this, incorporating a second wand for two-handed manipulation and control is ideal. This could be implemented such that two separate 'grab' actions of the user (one with each wand) when a refitting tool is active could enable the structure-quality-monitored rotation of different dihedral angles of a structure. Simply stated, grab the bonds and move them while monitoring to ensure that the changes are consistent with our knowledge of how protein structures behave.

4.2 Demonstrations and Use of KinImmerse in the DiVE

4.2.1 Orientation In Phases

In a virtual reality demonstration, the audience must become oriented to the system. A demonstration makes the task an accelerated teaching session. I approach this in three stages, each in different physical locations; (1) in the foyer (2) in the facility room but not inside the system (3) inside the system itself.

4.2.1.1 Foyer

Orientation begins in the foyer where the audience cannot directly see the facility. This is done because the sense of excitement people have when they view the facility reduces their ability to pay attention to information and remember directions or instructions. The foyer is an ideal place to cover ground rules.

Typically, I explain the basics of protein structures to a lay audience in the foyer. For school groups, this might be very simple and include asking individuals to recall what DNA, RNA, and proteins are in very general terms. For undergraduates or graduate students with training in the basic biomedical sciences or chemical sciences, this is mostly unnecessary, besides possibly discussing some details of the structure model used in the demonstration.

Next, I make connections to real-world examples that the group might know well. The most readily accessible parallels are to video game systems. The Nintendo Wii (Nintendo, 2009), which uses a motion-tracked controller, is the most common example

used. It is described as a very simplified version of the wand controller used in the facility. In a similar fashion, the software packages used to develop the demonstrations are described as being the same or similar to ones used by video game developers, such as VirTools (Virtools, 2009). VirTools is a rapid prototyping package used extensively by developers of Xbox (Microsoft, 2009), Playstation (Sony, 2009), and the Wii (Nintendo, 2009) games. At this point, a parallel has been made to the controller, and the software that creates the environment. The last part of the contextual placement is the interaction. I often use the example of the movie *Minority Report* (Spielberg, 2002) to draw a parallel to the types of interactions between the person and the data, as depicted in the film when Tom Cruise manipulates information floating in front of him at a futuristic looking terminal.

4.2.1.2 Facility

Standing in the facility is an important transition step. This is where the demonstrator can sense the level of reaction to the system that might occur when the audience is inside using it. It is a good time to reinforce what was said in the foyer and allows the audience to get over some of their initial excitement.

The room containing the virtual reality chamber is high, and the virtual reality cube is a dominating looking object with sharp lines, dark colors, cables and wires, mirrors, and often perceived as an actualization of something most people believe only exists in science fiction novels or films.

Next, I describe the breadth of expertise necessary to create a virtual reality environment and experience. I discuss the role of scientists who have expertise in a specific domain, the role of engineers who build and maintain the system along with building associated devices, the role of computer scientists and programmers who transform the concept into software, and the role of artists who often provide the critically important touches to the designed environments that confer realism.

I follow this with an explanation of the basics of virtual reality. I focus on the three elements: creating a virtual world, immersion into that world, and sensory feedback to the virtual world (Sherman, 2003). I stress the importance of interacting with the system through using the head-tracked stereo goggles and the tracked wand. I explain that virtual environments are intended to fool various human senses into believing you are somewhere else experiencing something else.

4.2.1.3 System

Inside the system itself, I start out wearing the head-tracker and holding the wand. I reiterate the importance of creating a virtual world, immersing the user into the world, and the various ways we have designed feedback to the system. I put this in terms of the human computer interaction as I describe and demonstrate each function of KinImmerse.

The first thing I demonstrate is that the system tracks quite precisely with the movements of the individual wearing the head tracker and holding the wand. I do so by asking others to stand directly behind me and look in the same direction as I am. I then pan my head back and forth from left to right slowly. While panning, I ask whether

anyone notices the movement of the objects in front of me. This, I tell them, is an example of the system interacting with the user and a good reason to always look in the general direction of the person using the tracking system.

Next, I hold my arm out fully extended and describe the wand and explain that there are buttons on the wand, a joystick in the center, and a trigger button underneath. From this pose, I explain that the joystick is a fly navigation, and that this is most similar to how Superman flies. I stress that it *flies in the direction you point it when you press forward on the joystick*. I perform this task slowly and tell those in the facility that they should be respectful of others around them and not fly without announcing the direction they are going and recommend that they fly slowly.

I use the fly navigation to fly towards the molecule being displayed. As the molecule approaches the boundary of the front screen to cross into a virtually recreated space in the actual room, I tell others that we can ‘grab’ the molecule as if it were a real object. As the molecular model crosses into the perceived real space, others experience the realism conferred for the first time. Then, I show them which button performs the 6DoF grab and demonstrate it slowly.

From here, the other functionalities can be explained while someone besides the demonstrator is using the system. I reiterate that the most important part of the demonstration is for everyone to use the system and that it is time for someone else to use the tracking system and have control. I recommend that each person use the system for a minute or two and then pass it along to someone else. Ideally, five to ten minutes for each person is desired, but the lower bound is approximately one to two minutes.

As I hand off the head tracker and wand to another person, I tell others in the room that it is important to watch how one another use the system and tell the first user that it is their responsibility to remind the next person what the functions are. This encourages peer learning and as I've found, it is more effective than the demonstrator looking over the shoulder of each user.

4.2.2 The Demonstrator

The demonstrator has to overcome many obstacles in order to keep the audience in check and to orient them properly to receive the educational and interactive products the system delivers. This is made more difficult when there are multiple people, when they are of a younger age, when they are inexperienced with this type of system, and when they are overcome with a sense of excitement and wonderment produced by the system.

It is critical to know the audience in order to determine which elements must be explained in more detail and which instructions to stress more than others. For younger people, I stress the importance of following my directions and learning from one another. For older audiences I stress the importance of taking it slow, and remind them that younger people who are more experienced with video games often have an easier time learning how to use the system. In almost all cases, six or seven people are in the room for the demonstration. This requires asking people to be mindful of the walls and to do their best to announce what they are doing as they interact with the system. A larger

number of people also necessitates that everyone scrunch closer together and do their best to set their gaze in the same direction as the person controlling the system.

When a user is struggling with a function, as demonstrator, I verbally ask if they would like me to help them by putting my hand over theirs and help guide their arm and fingers in performing the operation. This is almost always a graciously accepted offer. Next, I use one of my arms to overlay my forearm on theirs slowly and wrap my hand gently around the top part of the wand while physically guiding their arm in the direction needed and moving their thumb to the correct button and press their thumb to enable the action. It is my experience that this is incredibly effective in helping a confused user become reoriented. However, it requires a great deal of finesse and explicit and clear communication between the demonstrator and the user in order to avoid startling someone or making them uncomfortable. It took me over a year of doing demonstrations before I felt I could do this type of corrective demonstration behavior. I do not recommend an inexperienced demonstrator do this.

4.2.2.1 Demonstrator As Authority Figure

The facility is large, expensive, can be disorienting, and has equipment all over the place. The demonstrator needs to be perceived as an absolute authority figure so the audience will immediately turn to the demonstrator for information, to report something they learned or found problematic, and to look for directions and approval before or after taking an action. This level of situational control over the facility is essential to running a smooth demonstration.

Situational control by the demonstrator is accomplished primarily through nonverbal expressions of power/dominance over the audience. A number of classic signals are used to achieve this. They can be categorized as physical potency, resource control, and interaction control (Manusov, 2006).

Physical Potency: Typically, this is parsed into three types of concepts, threat, size or strength, and expressivity. In the demonstration, the demonstrator uses size or strength and expressivity the most in establishing dominance over the audience. Standing erect with shoulders squared, speaking with a loud and firm tone, and dressing in a clean cut and professional manner all are nonverbal cues of dominance that aid the demonstrator in controlling the demonstration and guiding the audience (Hall, 2006).

Resource Control: This is usually separated into four sub-categories; (1) command of space, (2) precedence, (3) prerogative, (4) and possession of other valued commodities (Manusov, 2006). The command of space is one of the easiest ones when giving a tour in the virtual reality facility. It's a large and impressive facility, it is scarce as a resource, and access is controlled. The demonstrator has total command of the space as perceived by the audience. This translates into a perception that the demonstrator is a clear leader and in control of the resource (Remland, 1981). The precedence component is most notable in demonstrating the controls of the system. Since the demonstrator needs to stand in front of the audience and show the controls, this shows a very clear precedence and sets the demonstrator up as leader and authority. The presence of prerogative is initially strong, but becomes much weaker during the sessions as the point of virtual reality is to create immersion and allow interaction. This means the specific

allowance of others to have prerogative and interact at will with the system. Finally, the possession of other valued commodities is extremely high. Since there are only about a dozen of these facilities in the world, this is a high status symbol conferring authoritativeness on the demonstrator.

Interaction Control: This is sub-categorized into five effects; (1) centrality, (2) elevation, (3) initiation, (4) nonreciprocation, and (5) task performance cues (Manusov, 2006). As demonstrator, I begin by showing the different interactions in the system to the audience; this starts off as a very high power asymmetry.

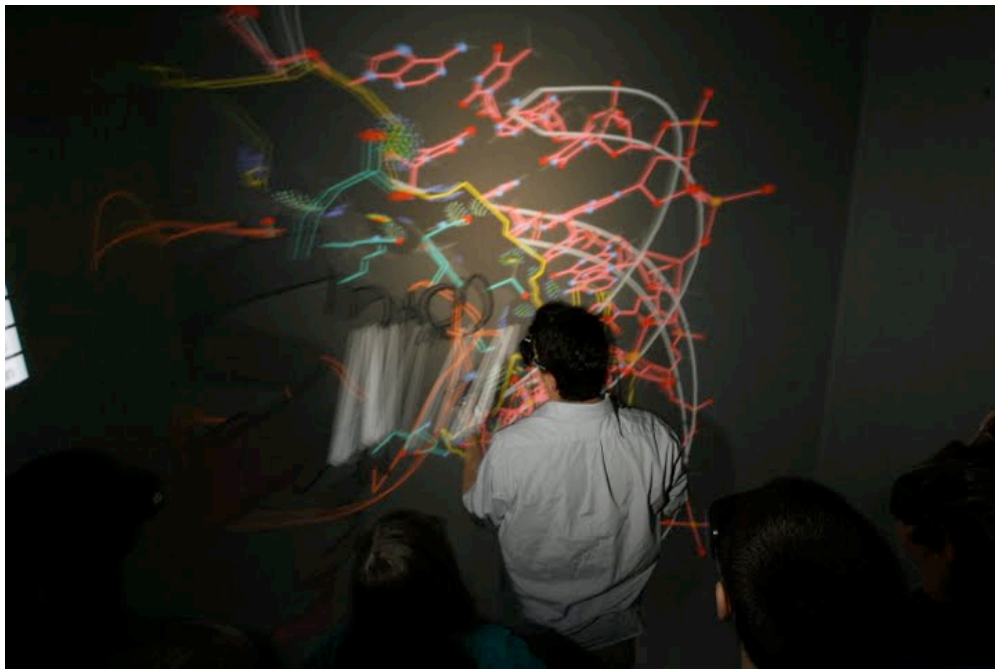


Figure 4-11 - Interaction Control in the DiVE. Demonstrator in front with students behind.

Interestingly, this power effect *diminishes by design* as the demonstrator allows each user to interact with the system. This redistributes the power of interaction control to the users and creates a sense of learning and personal ownership by others as they become empowered users of the system rather than passive viewers.



Figure 4-12 - Transition of interaction control from the demonstrator to a student in the DiVE

4.2.3 Orienting Attention

The scientist Hermann von Helmholtz was known for innovation and discovery in a vast number of fields. His ophthalmoscope revolutionized the entire field of ophthalmology and variants of it are used today. Likewise, his contributions in physics and thermodynamics are widely known. Sometimes forgotten is that Helmholtz made key discoveries in the theory of attention. In 1871, Helmholtz concluded that one could

shift their focus of attention without ocular movement and made the claim that visual analysis was based more on where we focus our attention than the point in which our eyes gaze (Helmholtz, 1871; Wright, 2008). Since that initial discovery, the field has grown and moved forward. Today, much of the interest for researchers is over exactly how this attention shifting is accomplished.

4.2.3.1 Voluntary vs. Involuntary Attention

Important to this work is the distinction between voluntary and involuntary attention. In the 1940's, Kohler elucidated that voluntary attention of an individual originates from self and goes to the object, while involuntary attention originates from the object grabbing the attention of the individual (Kohler, 1947). In the DiVE, we take advantage of immersion into the molecule as a dominant force capturing the user's attention involuntarily.

In the immersive system, demonstrations are done using a mixture of techniques that seek both to orient people physically – head and body motions, ocular movement and gaze all done by instruction – and to orient their attention either jointly or independently of physical orientation through the virtual environment of KinImmerse. The *physical* orientation is overt orienting, as described by the famous conditioning studies of Pavlov (Pavlov, 1927), and the *attention* orienting is covert orienting as described by Posner (Posner, 1978).

Intriguing for the study of macromolecular structure is how covert attention orienting is often described in cognitive terms, and overt orienting in physical terms.

Immersive molecular graphics representations are abstract, complex, irregularly shaped, and the importance of handedness and myriad other fine details are critical to allow comprehension. Because of all these difficulties, increasing the degree of overt, physical orienting, so that the system grabs the user's attention effectively, reduces the cognitive load placed on the user by the covert orienting through instruction. Together, this can reduce user fatigue and increase the length of time attention can remain on one object or area of interest for a user investigating a molecule.

4.2.3.2 Location Cueing

Shifting focus to one portion of the visual field increases efficiency in analysis of the items located in that region. Location cueing – using signaling means to shift setup and shift focus in a particular way – is an important part of any virtual reality system and even more so for demonstrations. With the distractions present and a virtual environment emulating multiple modalities of interaction, the demonstrator must use location cueing effectively.

Early research on location cueing demonstrated that shifting attention to an area prior to onset increased target identification (Eriksen, 1972). This was later refined and the elements in it are: (a) a central fixation point subjects are instructed to look at, (b) a target item to be identified, and (c) a location cue given immediately before target appearance (Wright, 2008). I use this in demonstrations by having users look at the front screen of the DiVE, verbally describing the item the user should bring into view, and

using the virtual wand displayed as an easily visible location cue for bringing the target object to the location.



Figure 4-13 - Demonstrator employing multimodal location cueing verbally and interactively

Interestingly, and by design, KinImmerse uses multiple modalities for location cueing to increase effectiveness. The vocal mode of describing to the user what will be coming and why, combined with the visual modality of the virtual wand in the virtual environment are simultaneous location cues. Additionally, the bounding box around the object touched by the wand gives a proximity cue for the area being brought forward for investigation and discussion. Figure 4-14 shows the demonstrator giving verbal instructions that an object is about to be ‘grabbed’ in the virtual environment by using the physical wand to reach out and perform the task.

This hybrid approach describes using symbolic and direct location cues. The meanings for symbolic and direct cues differ. Symbolic cues are considered goal-driven

and more focused, whereas direct cues acquire meaning by being near an area and are therefore considered more sensory. Symbolic cues are usually pointed very specifically at a target, whereas direct cues are present in close proximity to a target location (Wright, 2008). In the immersive environment, direct cues are much less clear. The wand used as a pointer might be called symbolic and the bounding boxes that appear around a selected or 'touched' atom might be considered direct. However, because of the interactivity of the system and the multiple modalities employed, the distinction is much less useful in this context.

4.2.3.3 Crossmodal Cueing and Attention Shifting

The combination of cues implemented in KinImmerse for the DiVE provides multiple modes which heavily capture, sustain and orient attention, thus achieving the goal of immersion in the virtual environment. This cueing combination is also called a crossmodal model for cueing and shifting attention. In the literature, this is described as being one of the hardest things to study. One significant reason posed for the difficulty in studying crossmodal models is a potential fundamental difference in the mechanism of different cueing modalities where localization of the stimuli of the cue in space is necessary (Wright, 2008).

The spatial localization is important for discriminating between targets, not the mere detection of a target. Therefore, the virtual environment and the multimodal cueing approach is anecdotal evidence supporting the usefulness of crossmodal models. It is important to note that details of the crossmodal model are still argued, especially over

whether the effects of each added modality are additive or greater than the sum of the modes. It is my experience with students and novice subjects in the DiVE using KinImmerse that a multimodal cueing paradigm has more than an additive effect on orienting attention. I do not, however, have a good sense for which mechanism is dominant (which mode contributes more and in what way) and that is the central debate still underway in the field.

My experience in the DiVE using KinImmerse is that virtual environments have potential for the study of attention, and to gain a deeper understanding of the interplay of modalities in a crossmodal cueing system with the advantage of a more controlled environment.

4.2.3.4 Location Cueing and Motion Sickness

One of the common concerns people have about virtual reality systems is motion sickness and disorientation. Users complain about distractions due to abrupt changes in the visual system when another user is controlling it. There is an interesting reason for this that is noted in the literature. Abrupt onset stimuli are known to immediately capture attention. This is a guiding principle advanced by Titchener in 1908. However, in the 1990's this view was challenged by the observation that goal-driven stimuli can override an abrupt attention capture (Wright, 2008). In the virtual reality system, the multiple modes of cueing act in a goal-driven way to override any abrupt – and usually inadvertent or distracting – attention capture outside of where the demonstrator wants to focus the demonstration.



Figure 4-14 - Demonstrator describing motion sickness effect while facing the audience.

Once the transition of control of the system has gone from the demonstrator to the other users, they are noticeably less aware of the potential for motion sickness, and increasingly do not use verbal cues to tell each other what they are about to do. This produces the observed result of people making others dizzy in the system by not giving them sufficient location cues.

4.2.3.5 Expertise Effects on Location Cueing

Subject or domain experts in a field often have expected location cues based on the history and development of their field. For example, a structural biologist would expect Carbon to be drawn white or black, Nitrogen to be blue, and Oxygen to be red. They might also have expectations as to which visual representations go with which type of message being presented. A ribbon schematic is the default style for an overall feeling

of the shape and topology of a protein structure, while a space filling model with atom colors (rather than dot surface or translucent surface) is expected for showing surface shape and interactions on the macromolecule. What the literature suggests happens is that these expectations by domain experts yields mixed results in terms of the subject's sense of effectiveness of the location cues and makes for conflicting claims concerning immersiveness when compared to novices with little domain expertise (Wright, 2008).

This supports my observations of showing the KinImmerse software package to different structural biologists. Some experienced scientists find it to be exceptional in orienting their attention. Others, mainly those who work with half-bond coloring schemes or those who especially value rendered and drop-shadowed graphics are less inclined to believe that the immersive system can be cueing their attention effectively. Thankfully, the body language of these critics quietly signals to me otherwise: their head movements are automatic and clearly following the wand position in the system as I move it, while they are verbally stating that they are not affected by or convinced of the immersive ability of the system. Mixed claims of effectiveness indeed come from domain experts.

4.2.3.6 Location Cueing & Co-Centering

As described earlier, the co-centering command was developed in KinImmerse for aiding the visual comprehension of local relationships between NMR models and NOE or RDC experimental data. It co-locates the same atom across all models within a structural ensemble while maintaining orientation, and demonstrates some interesting

cueing effects. Simpler location cueing effects occur when the demonstrator discusses a local area while moving it around using the 6-degree-of-freedom (6DoF) control, the bounding box, and the wand extending to the point of interest. Simultaneously, the demonstrator describes in a multimodal fashion the movements being made.

When the co-centering operation is activated, the user experiences an abrupt location cue when the many individual structure models in the ensemble snap to co-locate the atom of interest in the same position in the virtual environment. In this way, a strong goal-driven location cueing comes first, followed by an abrupt capturing of the subject's attention and a sudden increase in comprehensibility. Usually the many models are poorly overlaid prior to co-centering, looking messy and difficult to understand. When the co-centering happens, there is suddenly a satisfying disorder-to-order transition making the data in the local area clear and accessible. This observation was critical in the decision to implement the visualization of RDC's in the desktop software KiNG by Vincent Chen and the programming of the co-centering function into KiNG (Chen, 2009).

4.2.4 Domain Experts' Experience

Each demonstration has a different set of goals. Most school groups are being brought here for the same reason; to encourage young students to focus their energies on excelling in math and science. Domain experts in engineering, computer science, or biochemistry have a different set of goals. Getting domain experts from different disciplines to understand the contribution of their work to an interdisciplinary project,

recognize the usefulness and importance of the work, and to acknowledge there are difficulties in evaluating it because it does not fall primarily within any single domain represents a grand challenge in interdisciplinary research.

4.2.4.1 Biochemists and Structural Biologists

Biochemists are primarily interested in how the virtual reality system can assist their research. Within biochemistry, structural biologists who probe the fine details of macromolecular structures are particularly interested in new techniques that can help them achieve a better understanding of these structures. The development of KinImmerse allows a biochemist to quickly create a visualization to interact with in a 6-sided CAVE by taking a PDB file (protein data bank formatted coordinate file, Berman 2006), making a kinemage file (Richardson, 1992), and loading the kinemage directly into KinImmerse. The entire process takes under five minutes for a fast lash up.

Most importantly, KinImmerse facilitates quicker conceptualization of the complicated macromolecular structures in the mind of the user, and a better feeling for the local relationships of data and model. This was not systematically tested, but every user who was asked whether it was easier to comprehend the shape and relationships of the molecule in the DiVE or on a desktop responded that it was easier in the DiVE. This could be further tested in a true human subject study with proper controls.

4.2.4.2 Engineers

Engineers are primarily interested in some of the technical challenges of the DiVE. The use of the tracking system and the fine control over the structure being experienced in the virtual environment is a feat of modern technology. Performance of the tracking system is currently fast enough that latency in system response to a user action is rarely noticed by the user. The main limitation is that latency of the system degrades near the seams and walls of the room, due to bodily interference with the signal and to the unfavorable angle at which the head tracker and wand tracker are oriented with respect to the ceiling-level sensors.

The ease of using the handheld wand is a testament to how well engineered it is. It is relatively easy for users to get used to manipulating the four buttons and joystick on the face of the controller. The button underneath, in a logical place to be pressed by the index finger, is also well designed and could be called a ‘trigger’ button. In a newer research version of the KinImmerse software, this is mapped as the 6DoF ‘grab’ button. The virtual aspects of the wand are also intuitive, such as the white “light saber” for touching objects and the “fly” navigation familiar from computer games. Unfortunately, the wand does not stand up well to being dropped, but that has nothing to do with how ergonomic it is.

4.2.4.3 Computer Scientists & Virtual Reality Experts

One strength of KinImmerse for domain experts in computer science and virtual reality is the selection of Syzygy as the back-end: they appreciate that it is widely used,

open-source, and works with a variety of devices. Virtual reality practitioners may criticize KinImmerse because many of the individual functions are not novel. This is a failure to recognize the synergy and novelty of suitably combining these features into a useful platform for research.

It is difficult to engage and impress these domain experts, in part because of the way publications and systems become known by the community. In many cases, developments are not published in journal articles or other standard modes that create the literature of a field. Conferences, symposia, and even white papers from companies or university groups can carry the same weight as formal publications. This hinders investigating and understanding the different capabilities previously described by others in the field.

4.2.4.4 Visual Artists

Perhaps least acknowledged by basic scientists and visualization experts are visual artists. I was lucky enough to have met Kevin ‘KAL’ Kallaugher at a lecture in the Sanford School of Public Policy at Duke. KAL is an internationally recognized political cartoonist whose work appears weekly in *The Economist*, where he serves as their primary cartoonist.

I invited KAL to use KinImmerse after hearing his lecture. He agreed and we went into the DiVE. KAL is not primarily trained in the physical or life sciences. I gave him an introduction to the DiVE and asked him to draw using the annotation feature. Within two minutes he had produced the figure below.

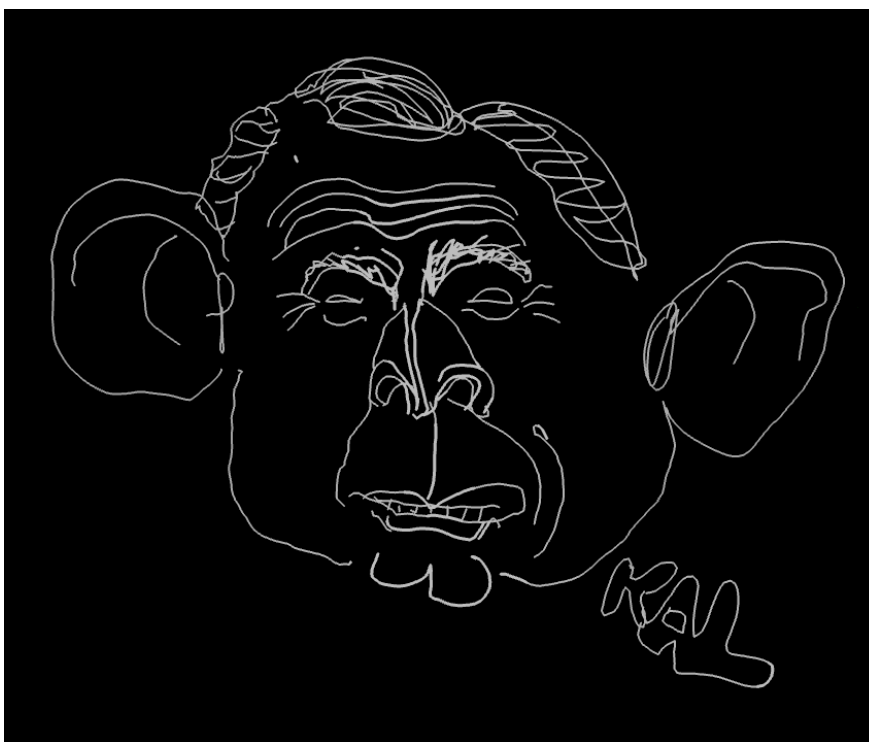


Figure 4-15 - Kinemage drawing from KinImmerse of George W. Bush by Political Cartoonist Kevin "KAL" Kallaugher

The figure shows a caricature of then President George W. Bush. Not captured in this figure is the depth to the drawing, which resembles the useful outlines of a sculptural bust. I concluded from my interactions with KAL that KinImmerse is easy to train a visual artist to use, and is valuable for transforming sculptural concepts from the artist's hand into a computer through the annotation feature.

4.3 Conclusions

Combining the DiVE immersive VR hardware with the kinemage graphics format, the KiNG display software, and the Syzygy VR toolkit enabled development of the open-source KinImmerse software system for immersive molecular graphics in the

service of production research uses. Possible input builds on the broad, flexible base of representations from years of using kinemages for molecular and other graphics, while KinImmerse provides both immersive and local perspectives and adds new interactive features.

The initial research application explored the study of NMR experimental data in context of interactive local superposition of models in the NMR structural ensemble, and its ability to spark new scientific insights is clear. Long-term success would constitute having NMR structural biologists using these tools routinely to help improve the quality of their structures or their understanding of them.

5. RDCvis – Visual Representation of RDC curves on an NMR Structural Model

The overall message of this chapter is that visualizing the RDC curves in their structural context, especially when combined with other structure quality visualizations allows users to easily identify and study areas of their models which need improvement.

5.1 Residual Dipolar Coupling use in Structural Biology

Significant advancements in using Nuclear Magnetic Resonance (NMR) spectroscopy to investigate the structure and dynamics of macromolecules have continued steadily over the last few decades. With the advent of measuring Residual Dipolar Couplings (RDC's) in the last ten years, the suite of investigative tools has added a new and powerful member to its ranks. Both structure determination and dynamics investigations now routinely use RDC experiments alongside more traditional cross-relaxation techniques that give assignable through-space and through-bond information such as Nuclear Overhauser Effect (NOE) distance restraints and angular J-couplings.

RDCs came from the discovery that diluting protein in liquid crystal solutions allows for easy measurements of couplings without degrading the high quality spectra needed for high resolution protein NMR (such as what is needed to collect NOE data). These partially aligned media allow for the measurement of RDCs observed over a large range of values (tens of Hertz). The RDC itself includes a term from the alignment tensor – a dimensionless vector denoting the preferential orientational averaging of the molecule – and can be described as an orientation of the internuclear vector (Blackledge,

2005). Examples of the internuclear vectors where RDCs could be measured include the NH or C α H bonds (and others). However, much work has been done investigating NH RDCs because they were noticed to be quite sensitive, useful in the study of dynamic motions, and are now the most common RDCs measured (Blackledge, 2005).

In structure determination of macromolecules, RDC's are used extensively for constructing protein backbones from RDC measurements in two alignment media using NH RDC's or two different RDCs (Cornilescu, 1998), or alternatively RDC data used in one medium to improve the accuracy of models calculated from NOE data. At a larger scale, RDC measurements are used to understand the relative orientation of macromolecules and the quaternary geometry of proteins with multiple domains (Blackledge, 2005).

5.2 Visual Representation of an RDC

Software for generating RDC visualizations, dubbed RDCvis and built into KiNG (Chen, 2009), was implemented by Vincent Chen using the Java programming language. As inputs, it requires a PDB format coordinate file and an NMR restraints file (in CNS format) with RDC data. RDCvis outputs the RDC visualizations in kinemage format (Richardson, 1992), as a standalone file that is routinely appended onto an existing multi-model kinemage for viewing in KiNG.

One note is that a significant barrier to using RDCvis is the lack of consistency in the deposited NMR restraints files. A more strictly defined standard data-file format

would make RDCvis more straightforward to use and thus routinely useful to a wider community.

The approach used by RDCvis to draw the RDC curves is based on a singular value decomposition (SVD) method to calculate a Saupe alignment tensor from the RDCs (Losonczi, 1999). The basic RDC equation between two spin 1/2 nuclei can be expressed in the following form:

$$D^{nm} = D_{\max}^{nm} \left\langle \frac{3 \cos^2 \theta - 1}{2} \right\rangle$$

where D_{\max}^{nm} is the maximal possible dipolar coupling value, which is related to the gyromagnetic ratios and the distance between the two atoms, and θ is the angle between the internuclear vector and the external magnetic field (Losonczi, 1999). This equation can be converted to the following quadratic form:

$$D^{nm} = D_{\max}^{nm} v_{nm}^T S v_{nm}$$

where v_{nm} and v_{nm}^T represent the internuclear vector and internuclear vector transposed, and S is the Saupe matrix, a 3 x 3 matrix with five degrees of freedom, whose elements describe the direction and asymmetry of the partial molecular orientation in the experimental system (Yan, 2005).

In order to draw RDC curves, RDCvis uses the input set of experimental RDC values and the input PDB file to generate a series of linear equations that can be solved using SVD to give the best-fit elements of the Saupe matrix. A minimum of five individual RDC measurements are required to solve for the five degrees of freedom of the Saupe matrix; more are of course needed for reliable accuracy.

RDCvis then uses the resulting tensor to determine the quartic equation of the RDC curves for each internuclear vector. These curves arise from the intersection of a hyperboloid conic surface (calculated from the RDC value and the equation) with the sphere of all the possible orientations of an internuclear vector (Figure 5-1). Each pair of target curves is drawn as polygonal curves that lie on a sphere centered on one of the atoms of the internuclear vector (for instance, the N atom for an NH RDC). The paired curves represent the locus of possible orientations of the vector (and thus possible positions for the H) that match the experimentally-measured RDC value. In order to remain consistent with standard practice in NMR structure determination, RDCvis calculates a Saupe tensor separately for each model of the ensemble.

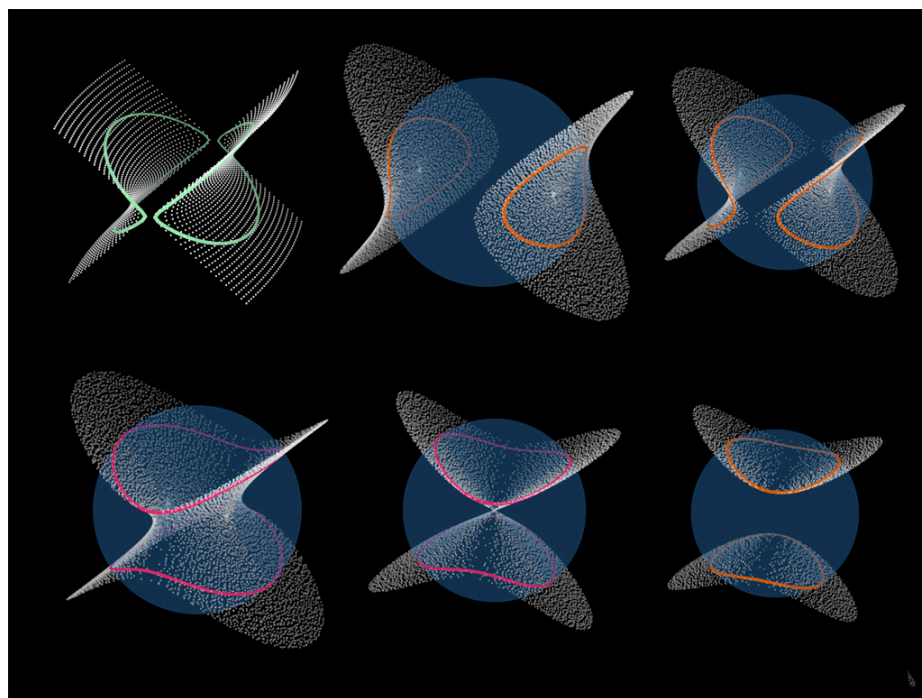


Figure 5-1 - RDC curve examples from hyperboloid surfaces (dotted) intersecting with spheres (blue)

Throughout the rest of this chapter (and others to follow), a Saupe curve, RDC curve, and target curve are all referring to curves plotted using the kinemage format. These curves plotted using the kinemage graphics format, take advantage of the powerful and extensive infrastructure that already exists for manipulating and viewing kinemages in Mage, KiNG, and KinImmerse (Richardson, 1992; Chen, 2009; Block, 2009).

5.3 Displaying & Co-Centering RDCs on a structure model in KiNG

Presented in the previous chapter, the software KinImmerse (Block, 2009) was developed for production as well as demo use of kinemage molecular graphics in a fully immersive environment. It provides varied and powerful molecular visualizations and navigation, and an initial set of research tools for manipulation, identification, co-centering of multi-model ensembles, free-form 3D annotation, and output of results.

In general, even the most well defined NMR ensembles will have enough deviation from model to model that a close-up comparison of the behavior of residues is difficult with an overall superposition. When all the models are visible, the visual clutter from all of the models is too overwhelming for reasonable analysis. Viewing models one at a time resolves the issue of clutter, but it is still difficult to compare one model to the others. On-demand local superimposition of the models is one possible solution, however for visualizing RDC data, which is directly related to the global orientation of the model, any rotation of the models would alter the relationship of the model to the RDCs. Therefore, the solution performed by the co-centering tool is to translate all the

models onto a single point with no rotational aspect, to maintain the global orientation of the models.

In the majority of cases co-centering reduces the visual clutter dramatically, to a degree that users can make a meaningful observation about the model-by-model agreement of the internuclear bond vector to the RDC curves and visually assess the match of the model to the data in the local context of the structure models in the ensemble. There are some situations where the co-centering may not be enough help. Particularly, in regions where there is a limited amount of experimentally observed data, the different models of the ensemble may have wildly different conformations, which makes co-centering less effective.

The co-centering function in the DiVE was so effective at revealing the relationship of structures to their RDC data that it was subsequently implemented in KiNG. Similar to the DiVE implementation, the aim was to make the KiNG version as simple to use as possible. Users merely need to activate the tool, at which point a simple button click on a point in a multi-group kinemage will translate the corresponding point in each group onto the selected point. Options are included for resetting the groups to their original relationships.

In addition to the single-button click co-center, the co-center tool in KiNG includes a number of extra options. During the course of testing the co-centering tool, we realized that a useful feature would be the ability to scan through the structure being studied. To enable this, Vincent added a “slide” function, which re-centers the view and simultaneously co-centers on the next or previous residue in the sequence. This feature

allows users to easily scan along the polypeptide chain, analyzing the fit to the RDC curves as they go.

5.3.1 A walkthrough of RDCvis in KiNG

Presented here is a walkthrough in KiNG of visualizing RDC's on an NMR structure ensemble. The example shown is the CCME structure determined by the NorthEast Structural Genomics consortium, described in greater detail in the next chapter.

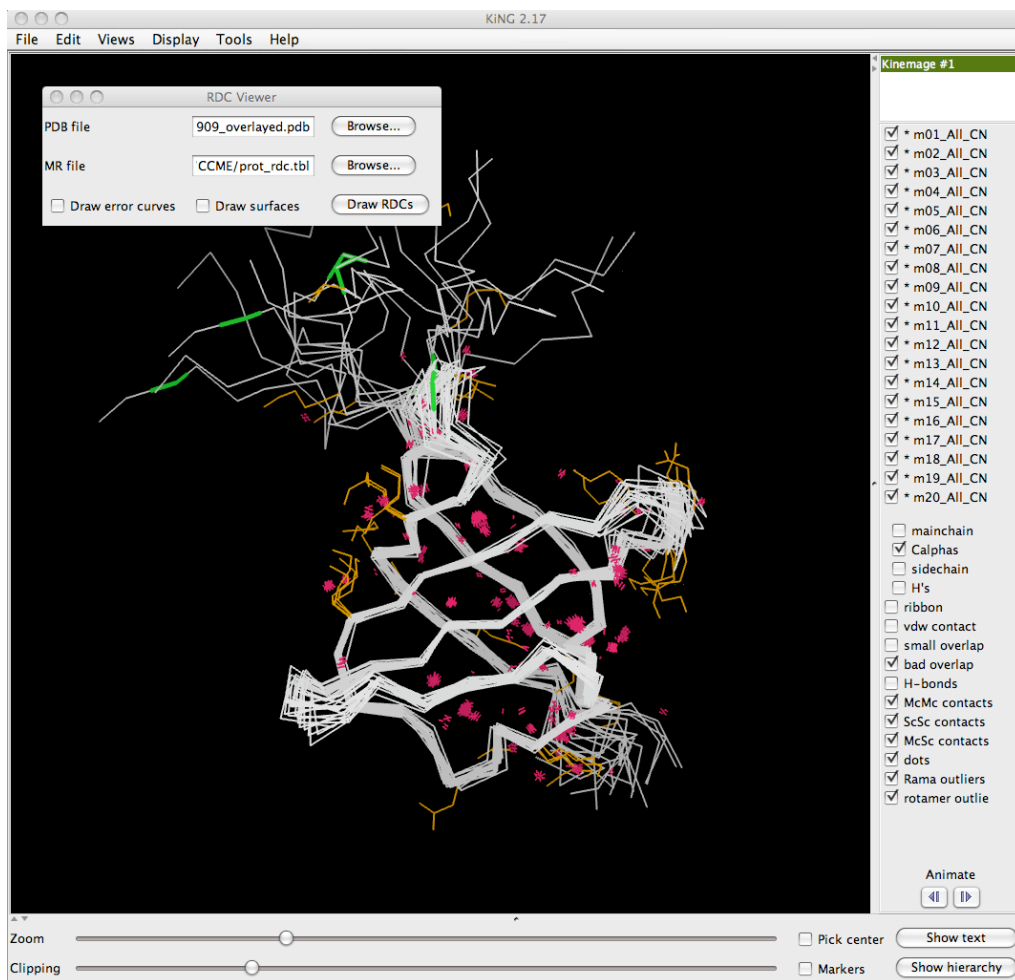


Figure 5-2 - RDCvis in KiNG entering PDB and RDC files

This shows a multi-model multi-criterion kinemage created in MolProbity (as described in chapter three). It is displayed in KiNG showing the alpha carbon backbone trace of all models of the ensemble as well as the geometric and steric outliers mapped onto the structure. The dialog box, at upper left, is accessed by going into the ‘Tools’ drop down menu and the ‘Specialty’ submenu where the user clicks on ‘RDCvis tool.’ This dialog box allows the user to specify the two necessary inputs for RDCvis, a PDB coordinate file (Berman, 2006), and an NMR restraint file containing CNS-formatted RDC restraints (Brunger, 1998).

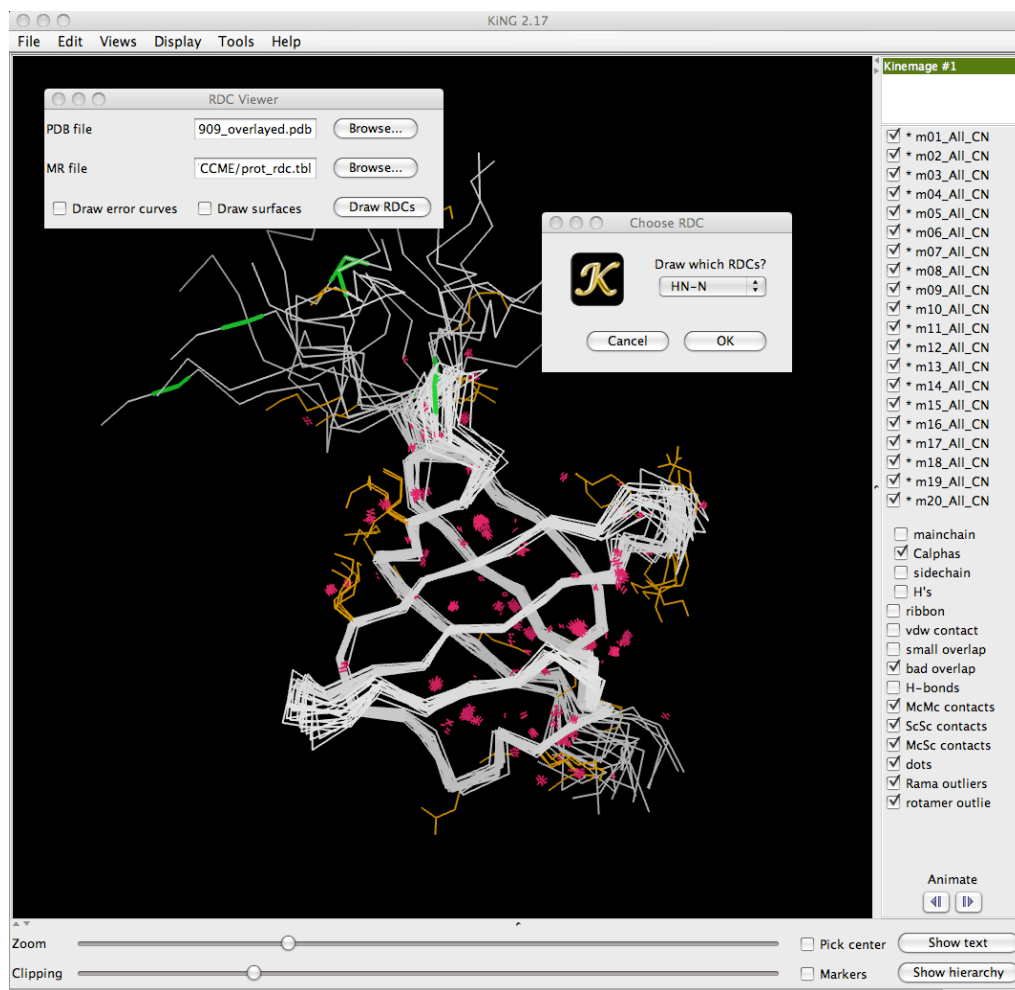


Figure 5-3 - RDCvis in KiNG choosing RDCs to display

The figure shows the dialog box that appears when the PDB and RDC restraint files are selected. It prompts the user to select the internuclear vector for which to create RDC visualizations. In Figure 5-3, NH RDCs are being selected for visualizing on the structure.

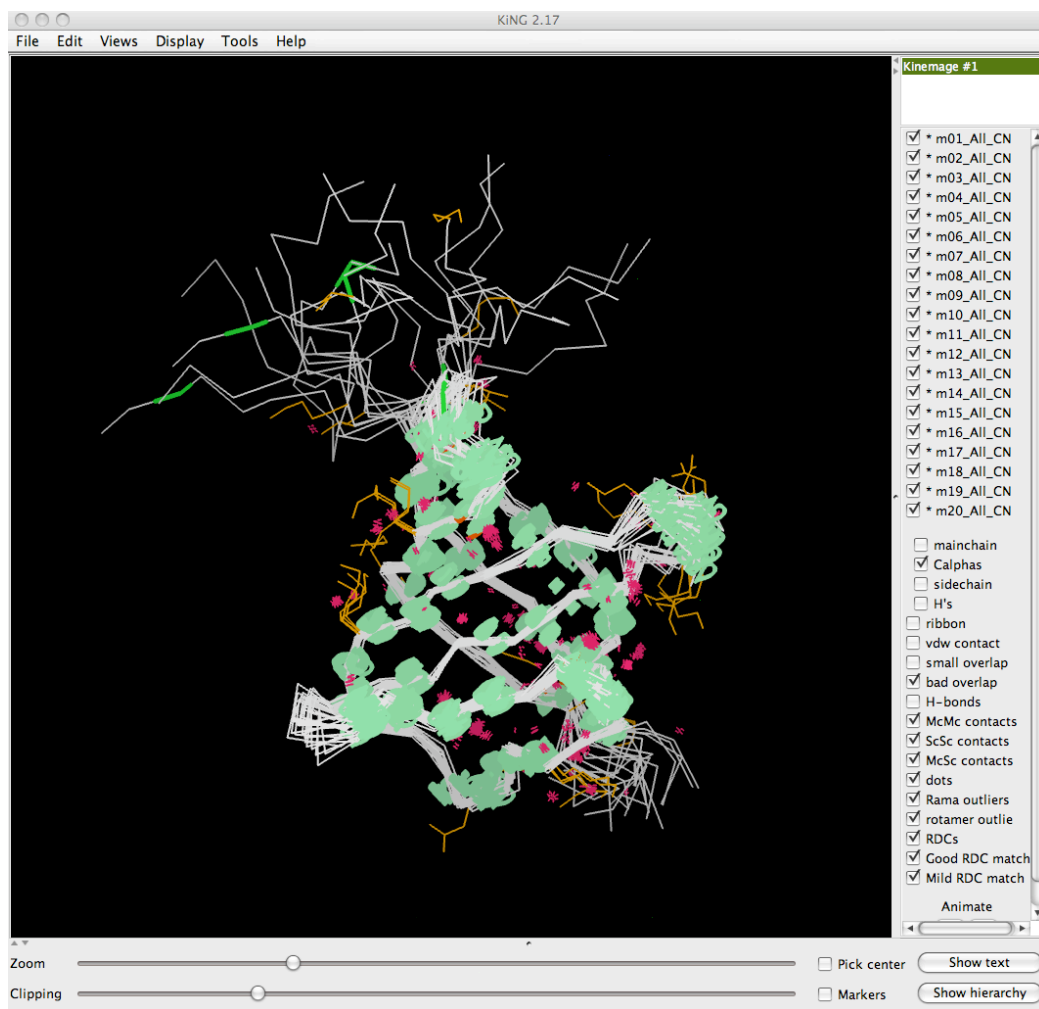


Figure 5-4 - RDCvis in KiNG with RDCs shown

Figure 5-4 shows the NH RDC's mapped onto all models of the ensemble. They appear in this zoomed out figure as blobs at various points along the alpha carbon trace.

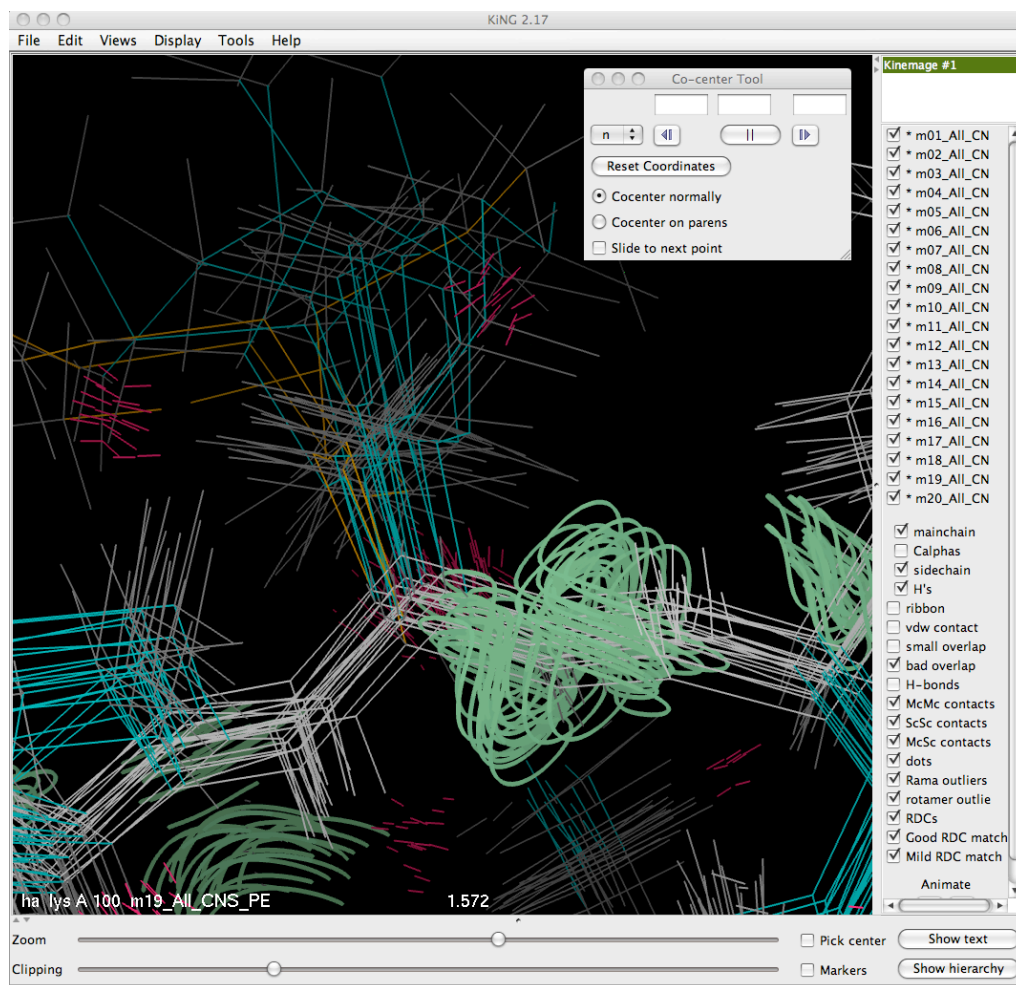


Figure 5-5 - Co-Centering tool in KiNG: Lys 100 as deposited

The figure above shows a zoomed-in view of Lysine 100 in the structure. The RDC curve pairs are mapped onto each of the models in the structure ensemble. The dialog box on the upper right of the image is the co-centering tool, turned on through the ‘Tools’ menu in KiNG and the ‘Kin editing’ submenu. Once the dialog box appears, the user can “co-center normally” by mouse-clicking the atom (in any model) around which to co-center. The user may also select the “slide” option and then the forward or back arrows in the box, to co-center on the next residue in sequence.

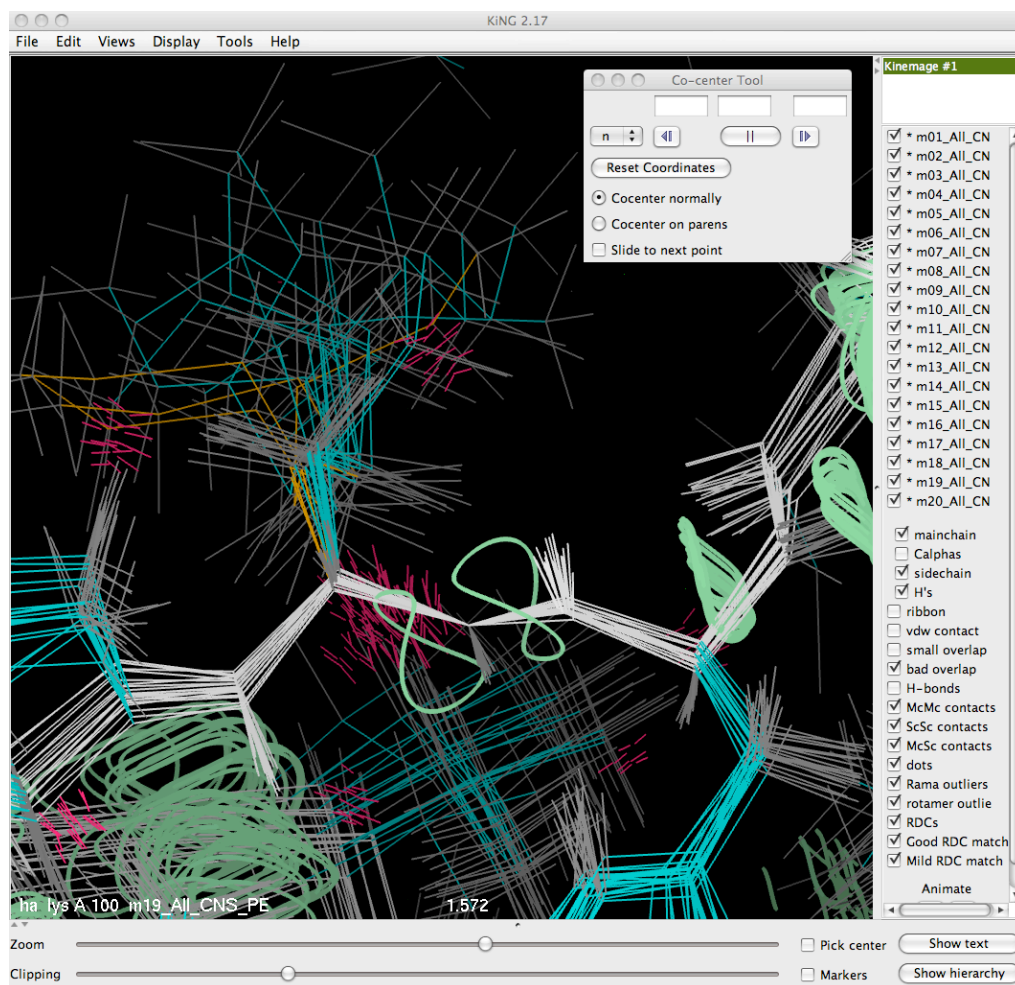


Figure 5-6 - Co-centering tool in KiNG: Lys 100 co-centered on the NH N atom

The figure shows the same Lysine 100, this time co-centered on the backbone Nitrogen. The ensemble of co-centered models allows easy interpretation of the match to the RDC curve by the NH bond vectors, with all the Hydrogens in this case pointing to and nearly touching the RDC curve in a narrowly defined bundle.

5.4 Results

A number of example structures containing RDC restraints were obtained from the NorthEast Structural Genomics Consortium (NESG) and were run through MolProbity to create multi-model multi-criterion kinemages. Then, their RDC target curves were modeled onto the structure using RDCvis. Each residue with a measured RDC was co-centered and viewed for all structure models within the ensemble. A number of interesting observations came from this work.

5.4.1 The One Curve Rule

As described in Chapter 3, the 1Q2N structure of Z-domain in Protein A determined by the NESG contained a loop region where two quite different conformations were modeled in the structure ensemble, despite the inclusion of RDC data. Originally, when looking at the region using only the multi-model multi-criterion kinemage, it was hypothesized that one of the conformations was a correct interpretation and the other was incorrect (based on the local geometry and sterics), and that somehow there was an ambiguity in match to the RDCs. However, at that stage I had no easy way to test that idea.

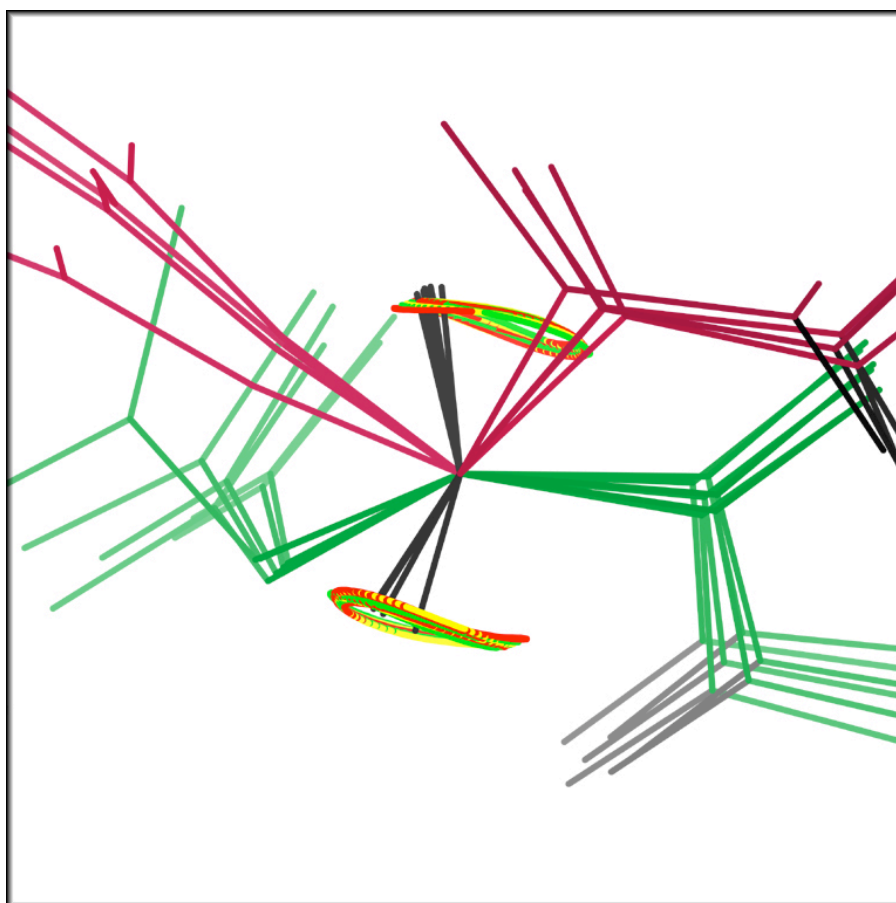


Figure 5-7 - Loop 35-41 in 1Q2N with RDC curves colored by agreement

Later, with the tools developed in Chapter 4, I investigated the 35-41 loop region of 1Q2N in the DiVE virtual reality facility, using KinImmerse and visualizing the RDC curves directly on the structure. That work showed (Figure 5-7) that the two conformations had their C α H bond vectors (in black) pointed to opposite limbs of the RDC target curves.

Specifically, the intersection of the hyperboloid conic surface with the internuclear vector sphere usually results in a pair of symmetric curves, both of which taken together delineate all possible internuclear vector directions that match a given

RDC value. The example in Figure 5-7 is shown with green RDC curves where the value calculated from a model is <1Hz different from the value measured, and red for >3Hz difference. I argue that having two distinct clusters does not mean that this bond vector in the real molecule moves to match both structural possibilities, but is probably due to an ambiguity in the joint implications of the RDC plus other data and geometry.

The usual methodology for use of RDCs in structure solution implicitly assumes one conformation, so that in general a given bond vector should only line up with one or the other curve. Motion or multiple conformations are of course possible, and for loops even probable, but this is not the way to identify such motion. It would require an extremely unlikely coincidence for a motion or conformational change to line up each of two orientational clusters for a given internuclear vector exactly on a different one of the two curves. Even if a residue were sampling conformations that could match both curves, this would result in averaging of the RDC and end up with a different, smaller RDC value. This averaging affect has been treated in the literature, where others have tried to develop a model of conformational sampling that stays in agreement with the observed RDC's (Clare, 2004; Brooks, 1997; Hess, 2003). I conclude that this behavior in a structure ensemble - of two distinctly different conformations pointing the internuclear vector towards opposite curves - is a potential systematic error allowed by the usual procedure of requiring each individual model to match the scalar RDC value, without considering the relationship of the models to one another or to the target curves.

5.4.2 Curve Intersection

When considering the range of possible RDC curve shapes, there is a unique transition point where the hyperboloid conic curves create two great circles that intersect (top left in Figure 5-1) and that each bisect the internuclear vector sphere. At RDC values near this transition point, it is possible that a given internuclear vector would match both curves if it pointed toward one of the curve intersections. As in the more general case, in this situation the internuclear vectors should still all point in nearly the same direction.

When RDC alignments are experimentally observed in two or more different media (such as the CcmE example in the next chapter), multiple RDC curve pairs could be present at a given position. Internuclear vectors pointing to an intersection between Saupe curves from RDCs obtained by different alignment media are structurally important; these signify a match of the internuclear vector to independently observed RDC measurements from different experiments.

5.4.3 Orientation Dependent Variability

There exists some variation in the internuclear vector match to the RDC data drawn as a curve. This flexibility can result in a fanning out of the internuclear vector along a target curve. The likely contributors to this variation are the “orientation dependent variability,” and the error model of observed RDC’s. I will later discuss modeling the error of the observed RDCs.

Generally, the orientation of the alignment tensor to the molecule (and its rhombicity) will determine the shape of the Saupe curves at each internuclear bond vector where they are experimentally observed. In addition, the orientation of the local structural features of the molecule in relation to the given Saupe curve shape will determine the amount and direction of variation allowable for structural interpretation.

Both orientation of the tensor to the molecule and orientation of the local structural features in relation to the Saupe curve interact with one another to impact the potential structural interpretations. For example, if a Saupe curve is relatively flat, a peptide rotation approximately around the $C\alpha$ - $C\alpha$ direction could swing the NH bond vector along the Saupe curve if the curve tangent has the right relationship to the $C\alpha$ - $C\alpha$ axis. The resulting fan of solutions match along an extended segment of the curve while staying consistent with the RDC data.

An orientation dependent variability should not be taken to imply dynamics. Rather, it demonstrates that for a given RDC in a local area, there is an arc along which multiple positions remain consistent with the data because of the orientation and shape of the RDC curve in the local environment of the structure model.

5.4.4 Planarity Problems

The figure below (Figure 5-8) shows that in the 2JNG structure from the NESG (Kaustov, 2007) at Histidine 69, the NH bond vector for five of the twenty submitted models in the ensemble is out of plane by greater than 0.28Å. An RDC observed at the

NH of this residue is important for understanding why the models contain out-of-plane peptide geometry.

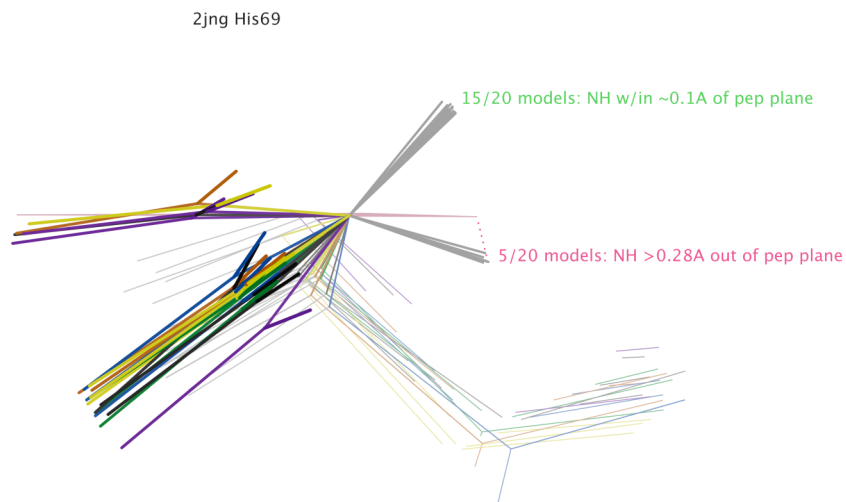


Figure 5-8 - NH Bond Vector Pulled Out of Plane

The ensemble has two distinct orientations for this peptide (clearly separate in the dark colors at front). For the 15-model cluster, the NH bond vectors are almost exactly in the peptide plane, while for the 5-model cluster the NHs are an implausible 15° out of plane, pulled by constrained match to the nearest RDC target curve.

This example is a cautionary one showing visually a known problem of mis-weighting terms in the target functions of structure determination packages that incorporate RDC data (Blackledge, 2005). In such cases, the resulting awkward peptide geometry almost certainly results from an over-weighting of the RDC restraints relative to the geometry terms.

5.4.5 Error Model Issues

The error model used for an NMR ensemble deposited to the PDB is rarely reported. This is not surprising since the full details of input values for structure determination and refinement are too numerous for regular deposition by most structural biologists. From informal discussion with spectroscopists, I do know that one common way of estimating error for an RDC is simply to use 10% of the observed total range (in Hertz) as the error specified in structure determination packages that refine against RDC restraints (like CNS).

What I observed, when investigating NMR structure ensembles with RDCs visualized on the models, were numerous instances where clustering of internuclear vectors on the RDC curves is extraordinarily tight - perhaps too tight, as strongly suggested by cases with two tight clusters widely separated. An example of this is shown below (Figure 5-9).

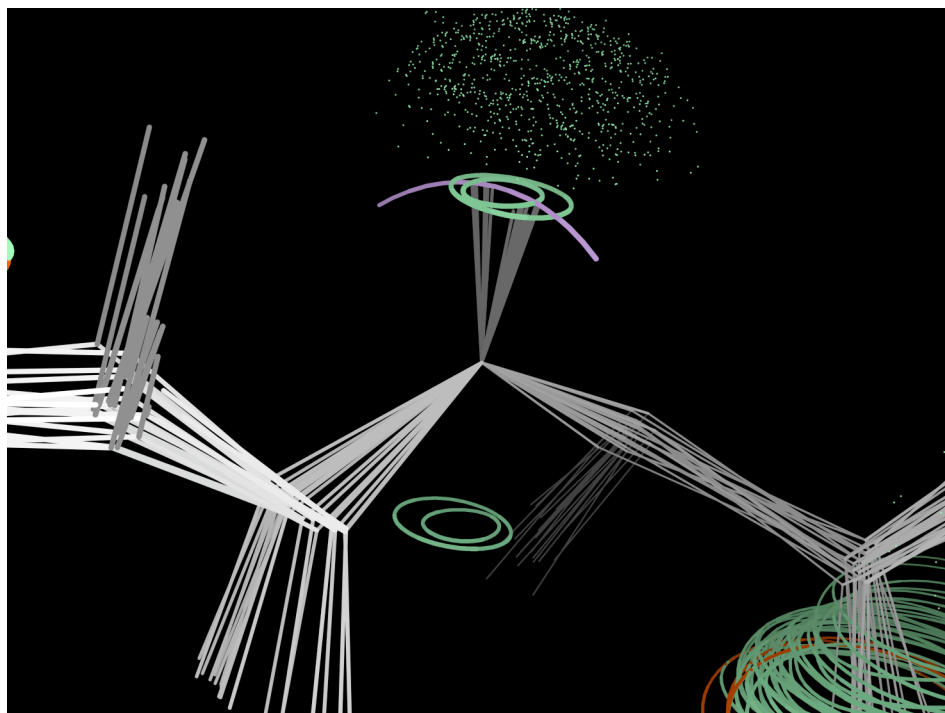


Figure 5-9 - Excessively tight clustering of NH RDCs on Leu 77 in CCME

The figure shows curves for two sets of NH RDCs on the CcmE ensemble at the Leucine 77 position. One RDC set was collected in PEG, and the other in Gel. Notice that one pair of curves is smaller and lies inside the other. At the top, the NH vector of the models in the ensemble are very tightly divided into two separate groups, each coupled to the curve. Instead, an arc of possible structural models bisecting the RDC curve and connecting these populations should be possible (shown in the figure as a drawn arc).

In the example, either the weight of matching the model to the RDC is too heavy, or the error model is too stringent. Some in the field have argued that a tight coupling to the RDC solution is valid (Clare, 2004), though these new visualizations imply that the error model used makes the distinction between the two observed conformations

artificial. On the other hand, the Bax group and others investigated the local geometry around NH RDCs, concluding that fluctuations exist in the NH bond orientation observed (Ulmer, 2003; Bremi, 1997; Fischer, 1997; Lienin, 1998).

I conclude that often the error estimate used for an RDC measurement does not realistically reflect error in the observation from the spectrometer. Additionally, if a rule such as 10% of the range is universally applied to all RDCs in the list of restraints, it may be inappropriate. In addition, CNS and Xplor-NIH are known informally to have had a bug in the code for some time (now fixed) that ignored the error estimates input by the user. Overall, distorted ensemble clustering (split, tight, or asymmetrical) is seen in many, but not all, RDC-based NMR structures.

5.4.6 Other Examples

I obtained a number of other example structures with RDC data from the NESG. These were analyzed in collaboration with Bart Bartkowiak who was rotating in the Richardson lab at the time. The NESG structures and data were provided by my collaborators with a consistently applied set of structure-determination protocols.

In the 2JNG structure, a loop problem not unlike the one previously shown for 1Q2N (Figure 5-7) appeared when co-centering on Glutamine 36 (depicted in the three-panel Figure 5-10).

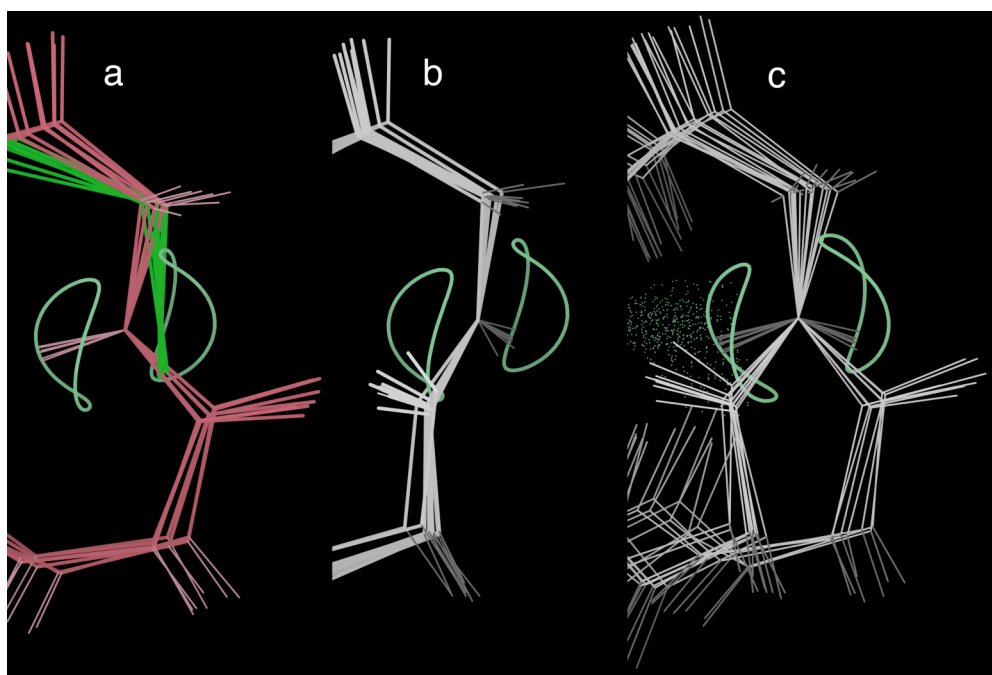


Figure 5-10 - Glutamine 36 in 2JNG loop

Panel a in the figure is a cluster of models in the ensemble adopting an unfavorable configuration, as indicated by the cluster of Ramachandran outliers. Panel b in the figure shows a cluster of plausible conformations in the ensemble with preferred backbone geometry. Finally, panel c (note the Hydrogen bond for the good cluster, visible here) in the figure depicts both clusters at the same time, with a clear use of opposite RDC target curves and a divergence of the protein backbone in opposite directions when co-centered on the NH of the Glu.

Another example, from 2JXX (Butler, 2007), has a similar situation in a loop with an Asp (as shown in the Figure 5-11).

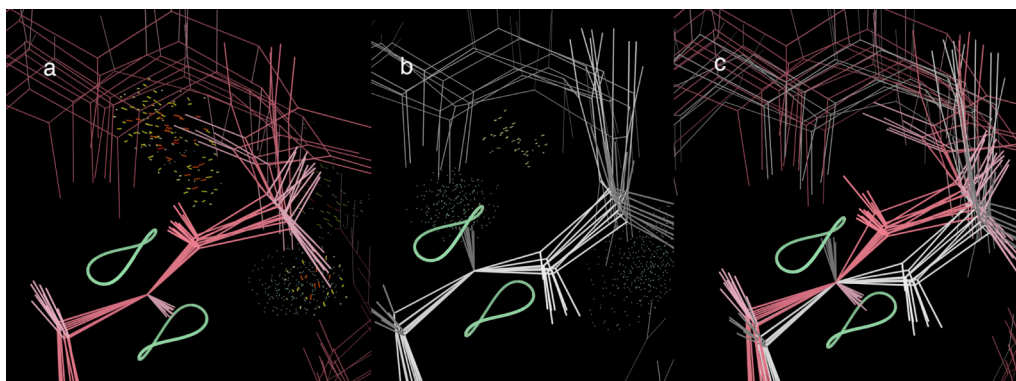


Figure 5-11 – Aspartic Acid 91 in 2JXX loop

In panel a, one cluster of models in the structure ensemble have the internuclear vector of the NH pointing outwards and match the lower RDC target curve. In panel b, a second cluster of models match the upper RDC target curve, pointing inward and making a Hydrogen bond (shown as pale green lenses of dots) with a nearby carbonyl. In panel c, both clusters are shown at once. The interpretation in this instance is that panel b is the preferred structural interpretation because of the presence of the Hydrogen bond.

5.5 Discussion

It is clear that co-centering, originally inspired by VR techniques and implemented in the DiVE, is effective when implemented in KiNG for routine use on the desktop. Further, by taking advantage of the new RDCvis code, this work identifies some patterns of frequent, systematic errors analogous to those previously found for crystal structures, but now arising from properties of RDC data that were detected by visualizing RDC target curves mapped onto models in a structure ensemble.

Orientation dependent variability gave insight on the limitations of RDC data under circumstances dictated by local geometry. This observation, as well as the non

planarity, the too-tight clustering, and especially the two-curve examples, indicate problems with the usual error models for RDCs. The one-curve rule, the simplest indicator of a potential systematic error in NMR structure ensembles, could be automated in a majority of cases.

6. Application of tools to NMR structure improvement

One way to approach improving NMR structures is by analyzing an output ensemble and feeding modified models or new restraints back into a new cycle of structure determination and refinement. As shown to work for x-ray structures (Arendall, 2005) using the methods available in the Richardson lab and through the MolProbity web service, this type of methodology is what I addressed in earlier chapters. Here it will be applied to a specific experimental NMR structure.

A second, more difficult, approach is to incorporate use of the new tools and analyses more directly into structure determination protocols. What follows is a description of the first steps aimed at enabling that process, using RDCvis, MolProbity analyses, and the specific example structure from the NESG.

6.1 Comparing Structure Determination Packages Incorporating RDC's

In order to test the use of RDCvis and MolProbity during macromolecular structure determination, I designed an experiment using the two major packages (CNS, and Xplor-NIH) that incorporate RDC data into the software, applied to a single test structure with multiple RDC datasets available. The table below summarizes the structure calculations performed for the comparison.

Table 5 - Comparing Structure Determination Packages Incorporating RDCs

Experimental Data Used	CNS	Xplor-NIH
NOE data only	X	X
NOE + RDC 1	X	X
NOE + RDC 2	X	X
NOE + RDC 1 + RDC 2	X	X

I worked with Jim Aramini at the NorthEast Structural Genomics Consortium (NESG) to identify a candidate structure for this investigation. The NESG data is readily available, and their structure determinations and refinements are performed in a standardized way. Additionally, NESG focuses on NMR structure determination and routinely collects RDC data, allowing for swift identification of a candidate structure for this study.

This experiment addresses two obvious initial questions – the relative merits of alternative software and the degree of benefit from an additional RDC dataset – and brings into focus several new considerations highlighted by the new visualizations.

6.1.1 Cytochrome c maturation protein - CcmE

CcmE from *Desulfovibrio vulgaris* (dvCcmE) is used here because RDC data are now available in a second medium and the protein is interesting biochemically.

In *E.coli*, CcmE plays an important role in Cytochrome C biogenesis by making a transient linkage to heme and transferring the heme to apocytochrome C in the presence of other factors (Kranz, 2009). It is proposed that H130 and Y134 are important parts of the heme-binding site in *E.coli* (Harvat, 2009).

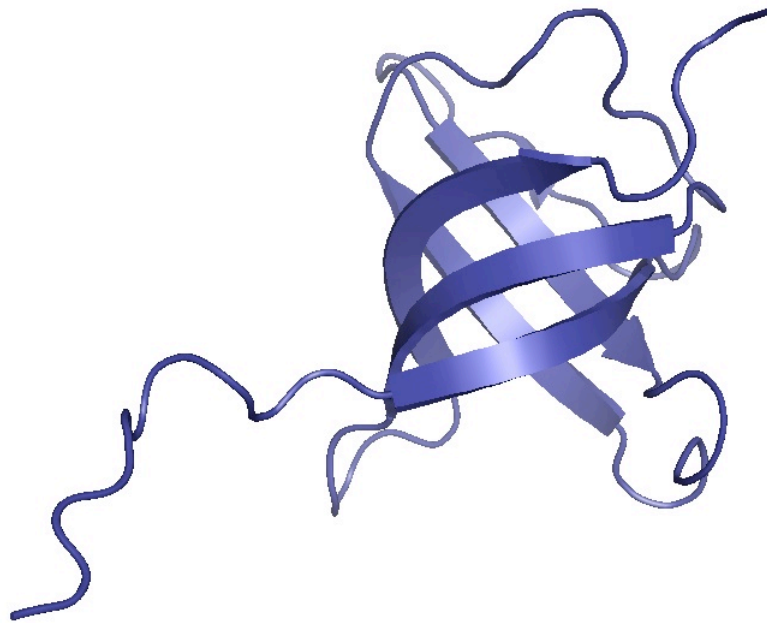


Figure 6-1 dvCcmE' Ribbon of Model 1 from PDB: 2KCT

A small handful of structures exist for CcmE in the PDB: an apo ensemble dvCcmE' by the NESG with one RDC dataset in 2009 (2KCT: Aramini, unpublished Figure 6-1), an apo ensemble in *E.coli* done in 2004 (1SR3: Enggist, 2002), and a more truncated apo minimized-average structure in *E.coli* done in 2002 (1J6Q) along with its ensemble counterpart (1LM0: Arnesano, 2002).

6.1.2 dvCcmE' Structure Determination Protocol Incorporating RDC's

In order to compare protocols, the dvCcmE' structure was determined using Peg RDC's alone, Gel RDC's alone, and both Peg & Gel RDC's together. These three determinations, performed by James Aramini at the NESG, were produced using a CNS protocol. After these determinations were done, the information was sent to Theresa Ramelot – another member of the NESG - at Miami University of Ohio, where she produced the structure ensembles for dvCcmE' using a Xplor-NIH protocol. This was performed in a blinded fashion, preventing the individuals performing the structure determinations from seeing the completed ensembles the other person produced.

6.2 Materials & Methods

The sample preparation, and NMR spectroscopy were performed by the NESG. The structure determination and validations were completed through numerous sessions where I went to Rutgers and worked directly with James Aramini while he computed the structures. The Xplor-NIH structures were performed at Miami University of Ohio and data shared via email.

6.2.1 Sample Preparation

The 85-residue construct from the CcmE gene of *Desulfovibrio vulgaris* (NESG ID, DvR115G; hereafter referred to as dvCcmE'), comprising residues 44-128 of the 137-residue full-length protein, was cloned into a pET21 expression vector (Novagen) containing an N-terminal methionine and C-terminal affinity tag (LEHHHHHH), yielding the plasmid DvR115G-21.1. This domain of dvCcmE' was selected by manual construct

optimization based on secondary structure and disorder prediction methods using the DisMeta server (Huang, 2009). The DvR115G-21.1 plasmid was transformed into codon enhanced BL21 (DE3) pMGK *Escherichia coli* cells, and cultured in MJ9 minimal medium (Jansson, 1996) containing ($^{15}\text{NH}_4$) $_2\text{SO}_4$ and U - ^{13}C -glucose as the sole nitrogen and carbon sources.

Initial cell growth was carried out at 37 °C and protein expression was induced at 17 °C by 1 mM isopropyl- β -D-thiogalactopyranoside (IPTG). Expressed proteins were purified using an ÄKTApurify™ (GE Healthcare) two-step protocol consisting of HisTrap HP affinity chromatography followed directly by HiLoad 26/60 Superdex 75 gel filtration chromatography. The final yield of purified isotopically-enriched dvCcmE' was \approx 28 mg/L of culture. Samples of U - ^{13}C , ^{15}N and U -5%- ^{13}C , 100%- ^{15}N dvCcmE' for NMR spectroscopy were concentrated by ultracentrifugation to 1.3 mM in 95% H_2O /5% D_2O solution containing 20 mM ammonium acetate, 200 mM NaCl, 10 mM DTT, 5 mM CaCl_2 at pH 4.5.

Sample purity and molecular mass were confirmed by SDS-PAGE and MALDI-TOF mass spectrometry (MALDI-TOF mass of U - ^{13}C , ^{15}N dvCcmE' (Da): experimental, 10,791.2; expected, 10,950; the difference is appropriate for cleavage of the N-terminal Met). Analytical gel filtration chromatography, static light scattering and ^{15}N T_1 and T_2 relaxation data demonstrate that the protein is monomeric in solution under the conditions used in the NMR studies.

6.2.2 Methods

6.2.2.1 NMR Spectroscopy

All NMR data were collected at 25°C on Bruker AVANCE 600MHz and 800MHz NMR spectrometers at the NESG equipped with 1.7-mm TCI and 5-mm TXI cryoprobes, respectively, processed with NMRPipe (Delaglio, 1995), and visualized using SPARKY (Goddard). Complete ^1H , ^{13}C , and ^{15}N resonance assignments for dvCcmE' were determined using conventional triple resonance NMR methods and deposited in the BioMagResDB (BMRB accession number 16096). All spectra were referenced to internal DSS. Backbone assignments were made using AutoAssign 2.4.0 (Moseley, 2001) and the PINE 1.0 server (Bahrami, 2009) using peak lists from 2D ^1H - ^{15}N HSQC and 3D HNCO, HN(CA)CO, HN(CO)CA, HNCA, CBCA(CO)NH and HNCACB spectra. Side chain assignment was completed manually using 3D HBHA(CACO)NH, HCCH-COSY, HCCH-TOCSY and (H)CCH-TOCSY experiments. Stereospecific isopropyl methyl assignments for all Val and Leu residues were deduced from characteristic cross-peak fine structures in high resolution 2D ^1H - ^{13}C HSQC spectra of the U -5%- ^{13}C , 100%- ^{15}N -enriched sample (Neri, 1989).

Resonance assignments were validated using the Assignment Validation Suite (AVS) software package (Moseley, 2004). Three-bond $^3J(\text{H}^{\text{N}}-\text{H}^{\text{a}})$ scalar couplings were determined using the 3D HNHA experiment (Vuister, 1993). ^1H - ^{15}N heteronuclear NOE and ^{15}N T_1 and T_2 relaxation measurements were made using gradient sensitivity-enhanced 2D heteronuclear NOE and ^{15}N T_1 and T_2 (CPMG) relaxation experiments, respectively (Farrow, 1994).

Two sets of ^{15}N - ^1H residual dipolar couplings (RDCs) were obtained for dvCcmE' by weakly aligning the protein in 4.2% C_{12}E_5 poly(ethylene glycol) (PEG)/hexanol or compressed polyacrylamide gel (Ruckert, 2000), in the same buffer as described above. Residual dipolar couplings for each unambiguous NH resonance were measured by the difference in ^1H , ^{15}N -HSQC-TROSY resonance frequencies between isotropic and aligned spectra.

6.2.2.2 CNS Structure Determination and Validation

The solution NMR structure of dvCcmE' was calculated using CYANA 2.1 (Guntert, 1997; Herrmann, 2002) supplied with peak intensities from 3D simultaneous CN NOESY (Pascal, 1994) ($t_m = 100$ ms) and 3D ^{13}C -edited aromatic NOESY ($t_m = 120$ ms) spectra, together with dihedral angle constraints computed by TALOS+ (Cornilescu, 1999; using only the constraints with the highest confidence and using TALOS+ uncertainties), and N-H residual dipolar couplings from one or both of the two different alignments (see below).

The 20 structures with lowest target function out of 100 in the final cycle calculated were further refined by restrained molecular dynamics in explicit water using CNS 1.1 (Brunger, 1998; Linge, 2003) and the PARAM19 force field, using the final NOE derived distance constraints, TALOS+ dihedral angle constraints and RDC values. Structural statistics and global structure quality factors, including Verify3D (Luth, 1993), ProsaII (Sippl, 1993), PROCHECK (Laskowski, 1993), and MolProbity (Lovell, 2003; Davis, 2007) raw and statistical Z-scores, were computed using the PSVS 1.3 software

package (Bhattacharya, 2007). The global goodness-of-fit of the final structure ensembles with the NOESY peak list data were determined using the RPF analysis program (Huang, 2005).

6.2.2.3 Xplor-NIH Structure Determination and Validation

Each of the 20 Cyana-3.0 structures calculated previously were separately refined with a restrained simulated annealing protocol that uses many of the updated features of the Xplor-NIH software (version 2.20.0; Legler, 2004; Cai, 2007). These include the IVM module for torsion angle and rigid body dynamics (Schwieters, 2001), a radius of gyration term to represent the weak packing potential (Kuszewski, 1999), and database potentials of mean force to refine against $C\alpha/C\beta$ chemical shifts (Kuszewski, 1995), multidimensional torsion angles (Kuszewski, 1997; Kuszewski, 2000), a backbone hydrogen bonding term (Grishaev, 2004), and RDC restraints (Clare, 1998).

The topology and parameter files used were protein.top and protein.par, which were designed to agree with bond lengths and angles from the CSDX force field (Engh, 1991). The radius of gyration was applied to residues 52-127 with the target value of $2.2N_{res}^{0.38} = 11.4 \text{ \AA}$, where N_{res} is 76 residues (Kuszewski, 1999). The backbone hydrogen bonding term was used in free mode so that identification of backbone hydrogen bonding was fully automated without user input (Grishaev, 2004). The D_a (the axial component of the alignment tensor, D) and R_h (rhombicity = $D_{rhombic}/D_a$) for each alignment medium were determined from the calcTensor script, which calculates initial values of the tensor using singular value decomposition based on the RDC

alignment tensor determined from the input starting structures and RDCs (Clare, 1998). The structures were calculated by simulated annealing in torsion angle space with cooling from 3000K to 25K with initial and final energy minimizations.

6.3 Results

I ran each of the resulting structure ensembles through the NMR side of MolProbity (Davis, 2007) to create multimodel multicriterion kinemages and charts (as described in Chapter 3). I then parsed out the NH RDC data from the NMR restraints files provided and loaded them into each of the corresponding kinemages using RDCvis (as discussed in Chapter 5). Next, I took the various kinemages into the DiVE and walked through the structures using the co-center feature (as discussed in Chapter 4) to assist in orienting myself with the data. After looking at the ensembles in the DiVE, I visualized each RDC using RDCvis on the desktop in KiNG, stepping residue-by-residue through each 20 model ensemble looking at the geometry and sterics in the local environment together with the RDC curves visualized on the structure.

The sections that follow contain examples with accompanying figures intended to survey what could become a typical set of findings using RDCvis during refinement and model building for NMR structure determinations where RDC measurements are taken. The RDC curves themselves are always across all models within an ensemble when co-centered at a given residue. However, because of the exact overlap, the colors of the curves in the examples correspond to just one of the models, while all (or many) of the NHs for a given residue are displayed.

6.3.1 Curve Intersection Examples

As discussed in the previous chapter (5.4.2), when RDC alignments are experimentally observed in two or more different media, multiple RDC curve pairs could be present at a given position. Internuclear vectors pointing to an intersection between RDCs obtained by different alignment media signify a match of the internuclear vector to independently observed RDC measurements from different experiments.

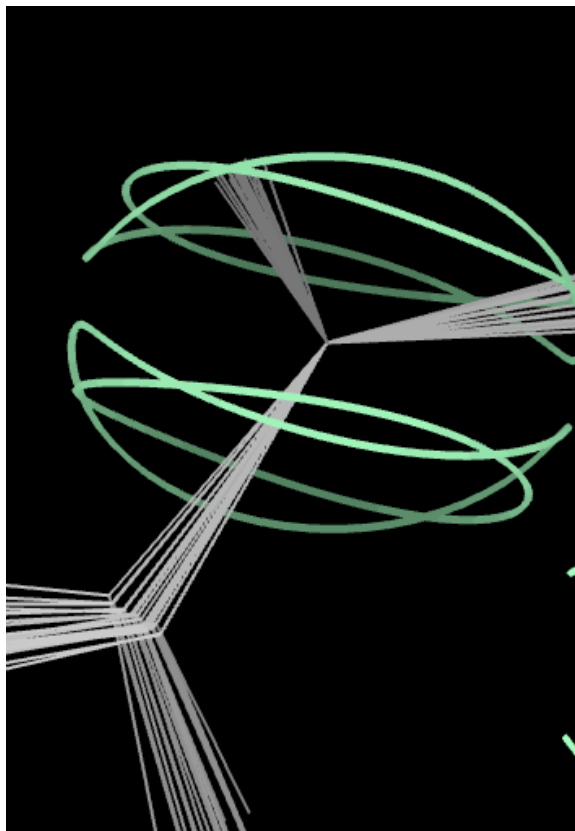


Figure 6-2 Ala 61 CNS with PEG + Gel RDCs

NH RDC data for two media are present at Ala 61 (Figure 6-2); the RDC curves intersect and all the NH's are pointing very near to an intersection point. A slight fanning

is observed, consistent with a peptide motional axis potentially from an orientation dependent variability (see Chapter 5).

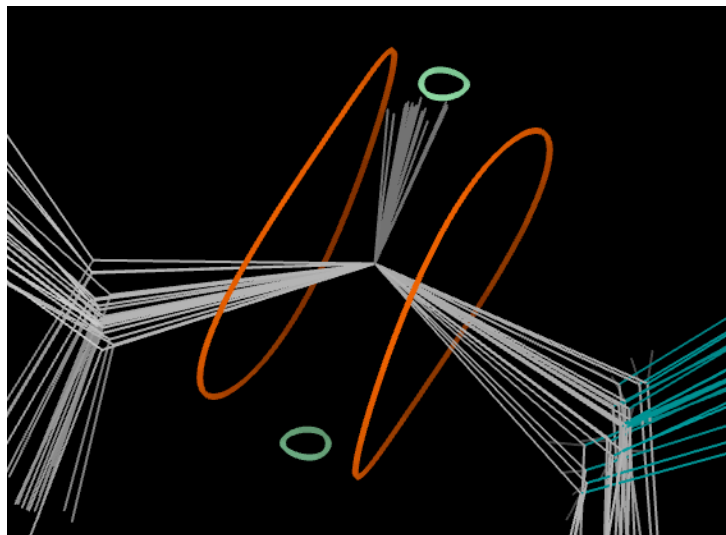


Figure 6-3 Asp 97 CNS with PEG + Gel RDCs

Asp 97 shows the Gel experimental data (wide flat circles), and the PEG experimental data (smaller circles in the center of Figure 6-3). Interestingly, the NH's for all the models in the ensemble cluster together, close to the small PEG RDC curve. This behavior is noticed in both CNS and Xplor (not shown) ensembles that used both PEG and Gel data together.

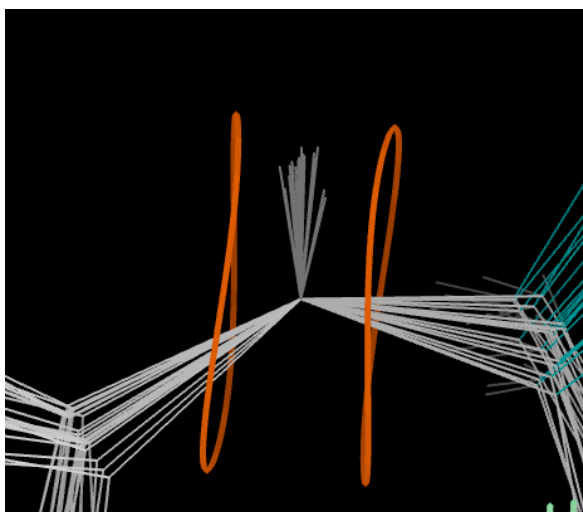


Figure 6-4 Asp 97 CNS Gel RDCs

In Asp 97, the match to the PEG RDC curves (not shown, but similar to Figure 6-3) is quite good in both the CNS and Xplor determinations that used only the PEG RDC experimental data. Likewise, in both CNS and Xplor, the determinations using Gel RDC experimental data show a clustering of NH's (Figure 6-4) in between the solution curves not touching either one, presumably because surrounding constraints on the structure prevented a better match.

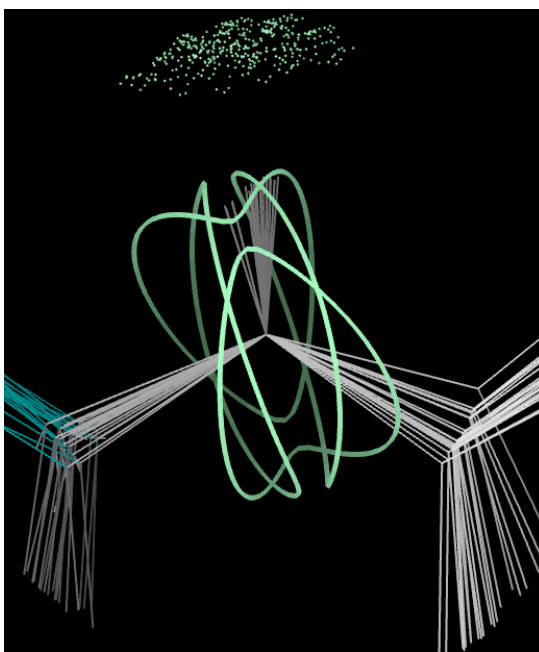


Figure 6-5 Leu 67 CNS PEG + Gel RDCs

Leu 67 (Figure 6-5) shows some very positive attributes. First, the NH's all point to (or very near) the intersection of the solution curves from the PEG and Gel data. Second, all the NH's have a strong hydrogen bond. Interestingly, most of the Leucines in this structure had NH RDC data observed in both PEG and Gel experiments.

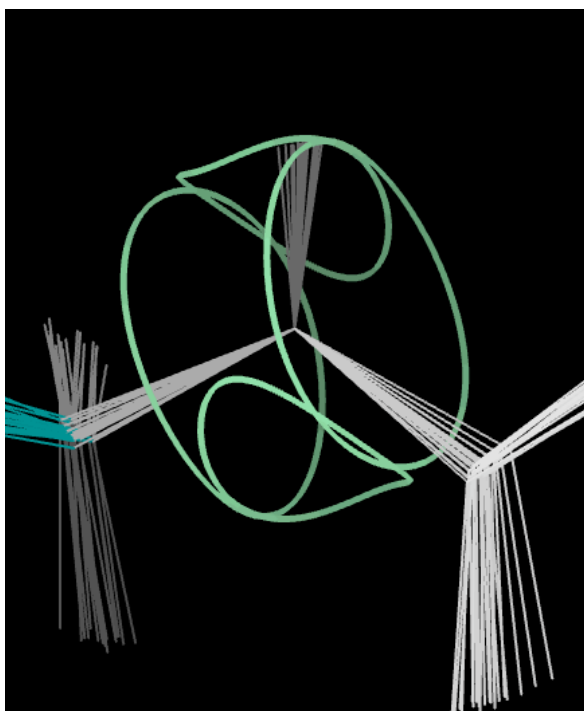


Figure 6-6 Phe 56 CNS with Peg + Gel RDCs

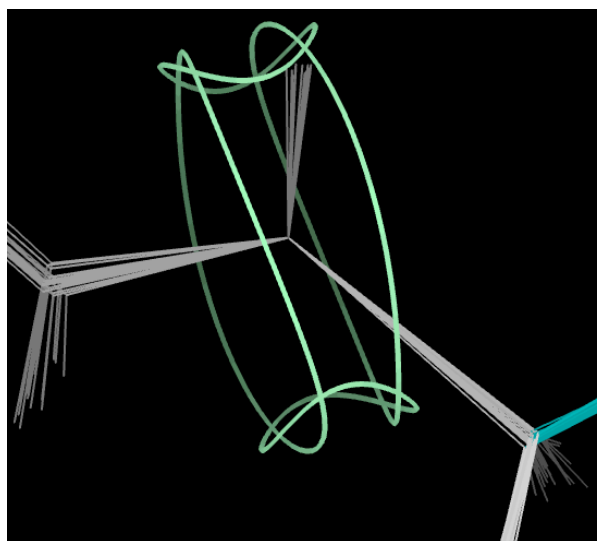


Figure 6-7 Phe 99 Xplor with Peg + GEL RDCs

The NH's for Phe where RDC data was present in the structures (which was most of them) looked very good, such as the two examples above (Figure 6-6, 6-7) showing

Phe 56 determined in CNS and Phe 99 determined in Xplor. Interestingly, since aromatics have a distinct part of the spectra where they are observed in an HSQC experiment, they are more readily observed in RDC measurements and here they were handled quite well by CNS and Xplor-NIH. Additionally, the interpretations of Phe residues may turn out cleaner than most because they are commonly found in the core of proteins and usually well packed.



Figure 6-8 Thr 122 CNS with PEG + Gel RDCs

Thr 122 has PEG and Gel RDC data present. In the CNS ensemble with both data (shown in Figure 6-8), all models in the ensemble have a nice hydrogen bond, and the NH's touch an RDC curve from the PEG data. Interestingly, the RDC curves from the PEG and Gel datasets never touch, and the other ensembles all look exactly the same at this position (not shown) with none of them pointing towards the RDC curves from Gel.

In a number of cases, curves of various relative shapes intersect well, and NHs in the ensemble touch very near the intersection point (Figure 6-2, 6-5, 6-6, & 6-7). However, this is not always true when using more than one set of RDC data (as shown in Figure 6-3, 6-4, & 6-8).

There is very little difference between CNS and Xplor determinations for the examples where the data matches quite nicely and the NH's point towards intersection points. However, for those examples where the NH's do not point towards an intersection (such as where the RDC curves do not intersect), it is more complicated and they don't always agree. These differences for the difficult cases probably arise from actual methodological advantages of one program, but sometimes merely from differences in relative weighting factors for which there is no clear scientific basis for the choice.

6.3.2 One Curve Rule Examples

As I described in chapter 5 (5.4.1), the intersection of the hyperboloid conic surface with the internuclear vector sphere results in a pair of symmetric curves, both of which taken together delineate all possible internuclear vector directions that match a

given RDC value. I argued that having two distinct clusters does not mean that this bond vector in the real molecule moves to match both structural possibilities, but is often due to an ambiguity in the joint implications of the RDC plus other data and geometry.

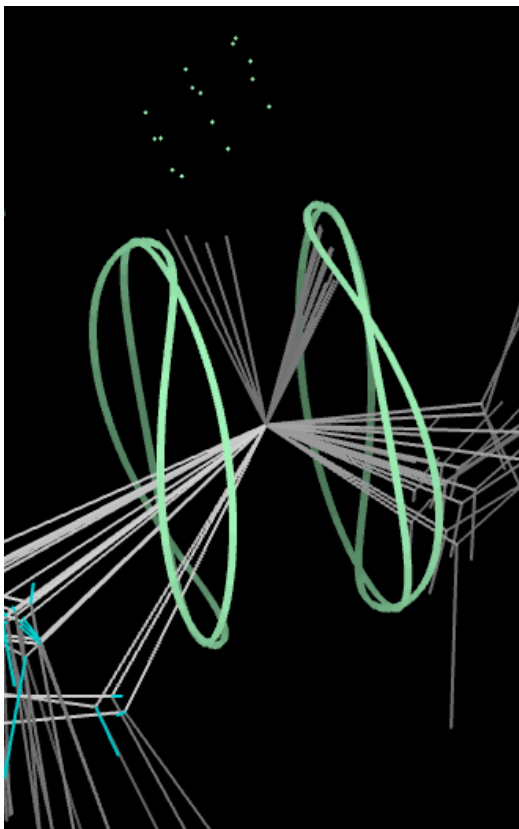


Figure 6-9 Asp 116 CNS with PEG + Gel RDCs

Asp 116 (Figure 6-9) shows an error where NH's in the ensemble are pointing to opposite curves, even with data from two media. It also shows a fanning of solutions (but with a gap) along a peptide motional axis, demonstrating orientation dependent variability. The behavior shown in Figure 6-9 is the same for both Xplor and CNS calculations, and remains the same when looking at determinations that use Gel or PEG RDC experimental data alone (not shown). Such cases imply that at least one of the

clusters is very likely in error, since there is no reason why these clusters should both match the same target curve pair.

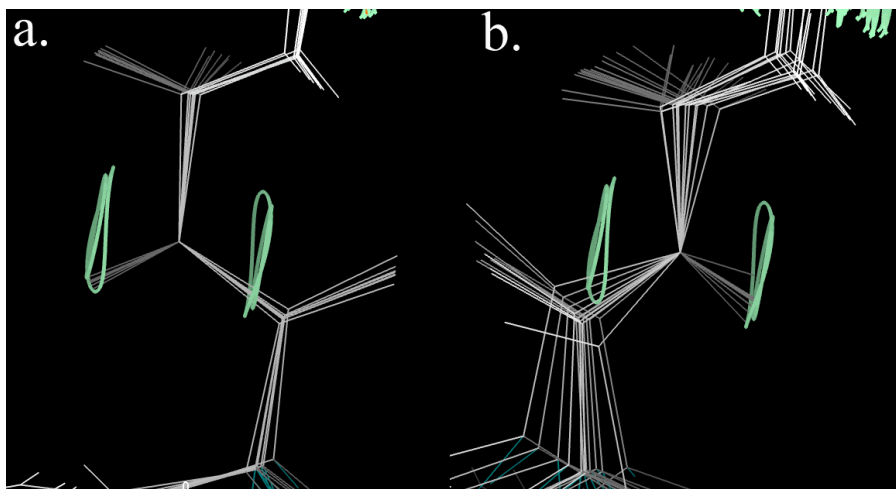


Figure 6-10 Gly 72 Xplor with PEG + Gel RDCs. a. 7 models correct b. 13 models incorrect

Gly 72 is interesting for multiple reasons. First, it shows the type of systematic error where the NH vectors are modeled to both curves. There is a population of 7 models pointing in one direction (Figure 6-10 a) and 13 models in the other direction (Figure 6-10 b). In the Xplor ensembles, this problem persists in all three of the determinations (not shown). This could be expected, since the RDC curves are almost identical for PEG & Gel as evident in Figure 6-10. Perhaps more interesting, the CNS structures (not shown) did not contain an error at this position in any of the three determinations and resemble the tight ensemble of 7 models shown in panel a of Figure 6-10.

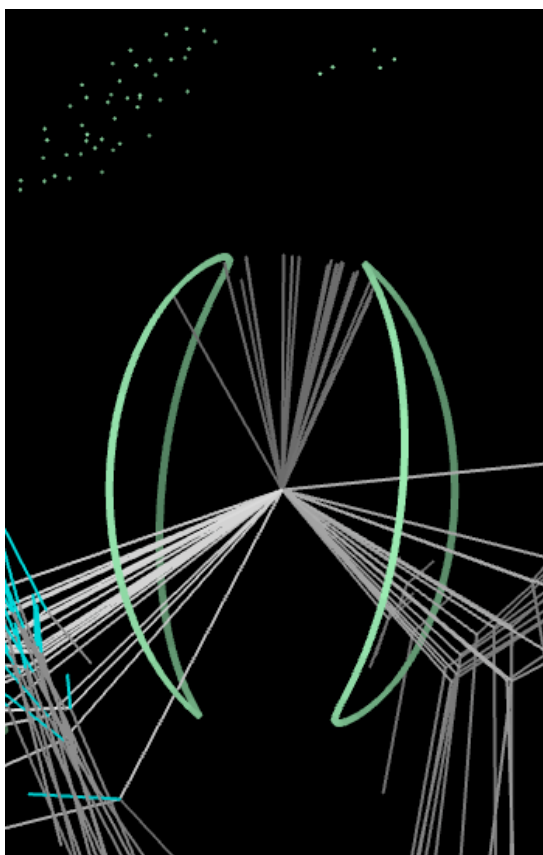


Figure 6-11 Ser 84 CNS with PEG RDCs

Ser 84 has PEG data only and shows the NH's pointing towards opposing curves and fanning between this position (Figure 6-11). However, the small population of models pointing towards the curve on the left makes a hydrogen bond and I conclude that this should be the preferred configuration. Uncertainty in the measurement is relatively high at this point because it is near to a X crossing at the undefined singular-point on the equator of the sphere.

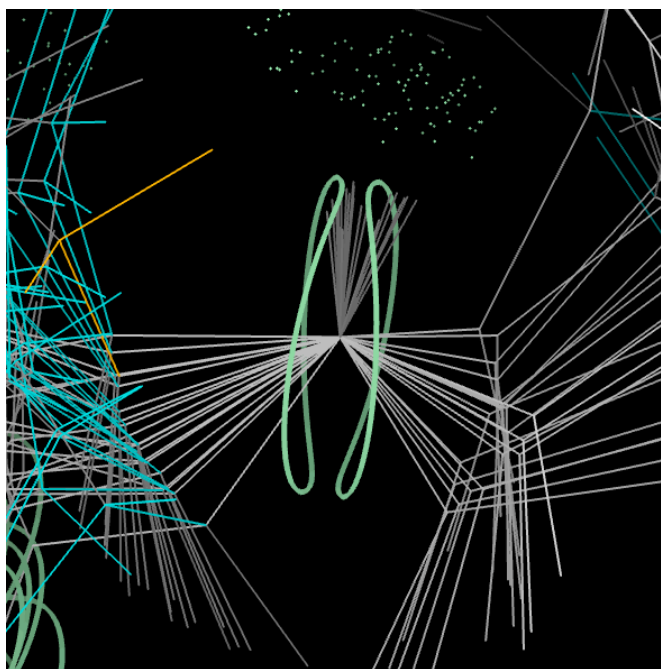


Figure 6-12 Thr 52 CNS

The NH's of Thr 52 in CNS (Figure 6-12, ensemble done with PEG + Gel data, only Gel data present at this position) are a good example of what might occur when the various solutions modeled in the ensemble are within the error of each other and pointing towards opposite RDC curves (see Chapter 5 discussion). Notably, there is little experimental data for adjacent residues (either NOE's nor RDC's). This means few observables are present to pin down this part of the model, which presumably contributes to the variability shown.

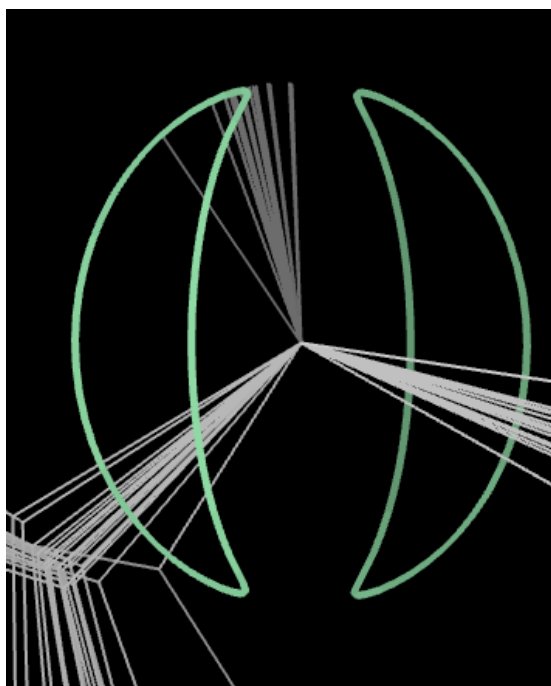


Figure 6-13 Thr 52 Xplor with Gel

However, the Xplor ensembles for Thr 52 do not exhibit the same problem. The target curves drawn by Xplor are shaped differently enough at this position and enough further apart that all NH's in the ensemble point towards only one of the curves (shown in Figure 6-13). Presumably, the differences in the shapes of the curves is in part attributable to the rhombicity term for the RDCs, causing the curves to become more arched in the Xplor structure at this position and flatter in the CNS structure as shown in Figure 6-12.

Unlike the examples in chapter 5, where the loop 35-41 in the 1Q2N structure had clear steric and geometric problems supporting a decision that one configuration was correct, no such clear examples are present in the CcmE structure. I expect that in practice overall, there will be a mixture of both types of cases.

6.3.3 Hydrogen Bonding Examples

In the CcmE structure, a number of striking examples included residues where the NH makes a Hydrogen bond. In practice, care should be taken when putting weights onto geometric vs. experimental terms in the structure determination packages. Due to the local nature of the experimental data for NMR (NOEs, or as is the case here, RDCs), weighting of the geometric terms vs. experimental terms could contribute differentially at a given residue.

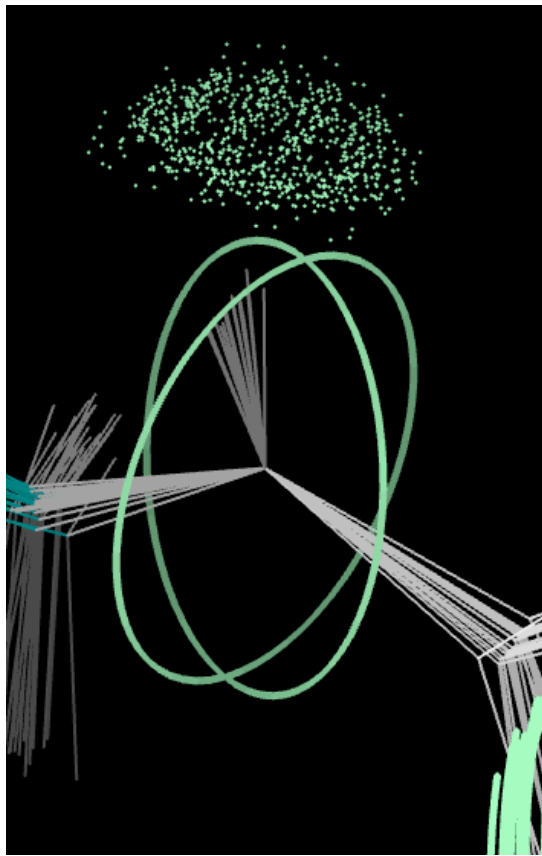


Figure 6-14 Ala 120 CNS with PEG RDCs

Ala 120 shows fanning out along a motional axis of the peptide as well as a strong Hydrogen bond for all the NH's in the ensemble (Figure 6-14). Interestingly, the RDC

target curves touch one another, but the intersection point is not automatically the solution (though it could be), because it does not represent the joint solution of two experimental observations. This is an example of the relatively rare case where the exact “X” shape occurs for a curve pair.

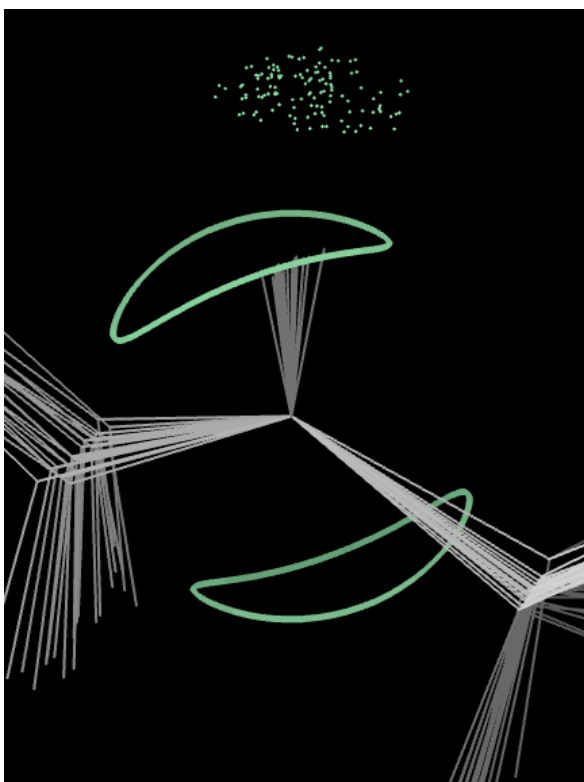


Figure 6-15 Gly 57 CNS with PEG RDCs

Gly 57 shows a fanning shape of NH's pointing along the RDC solution curve (Figure 6-15). Also present in each of the models is a Hydrogen bond, albeit a weak one. This could be weak for any number of reasons, some of which were discussed in Chapter 3 and are often observed in NMR structures.

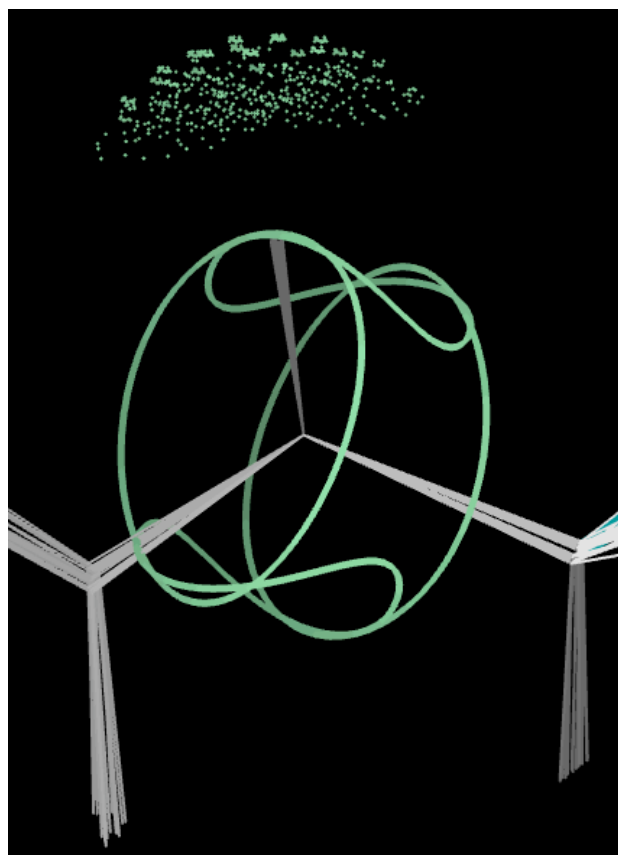


Figure 6-16 Thr 59 CNS with PEG + Gel RDCs

Thr 59 (Figure 6-16) from CNS with RDCs from PEG and Gel shows very nice hydrogen bonds from all the NH's in the ensemble, and a tight clustering of the NH's touching an intersection point between the two sets of RDC curves.

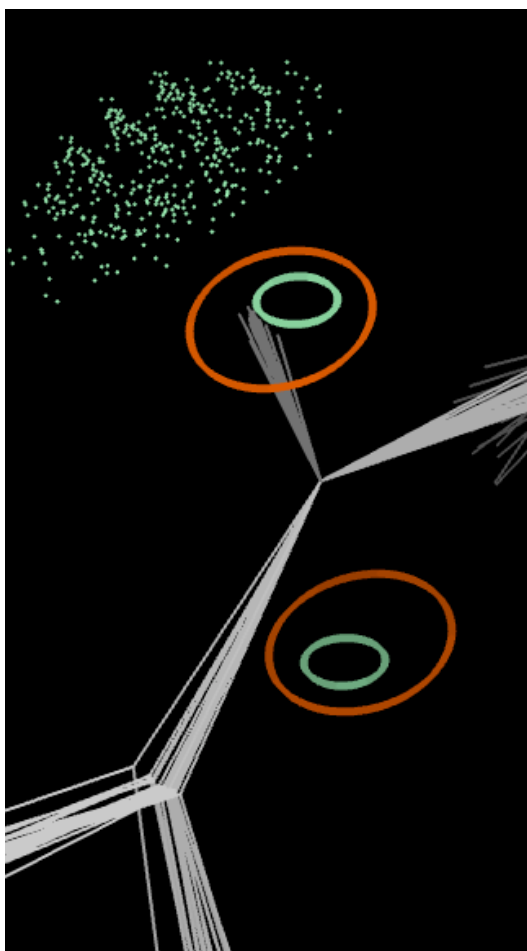


Figure 6-17 Ala 60 CNS with PEG + Gel RDCs

The RDC curves from the two different media do not intersect at Ala 60, although they are close in space (Figure 6-17). The models show the NH's clustering together much closer to one of the two curves. The pale green dots denote a strong hydrogen bond. The balancing of terms in the target function between the different RDCs and the hydrogen bond are one factor in this observed local conformation. Additionally, if the maximum value of the RDC datasets is similar for the two different media, the smaller (more towards the pole) RDC curve is indeed the more accurate of the two, perhaps account for the tight clustering near the green curve.

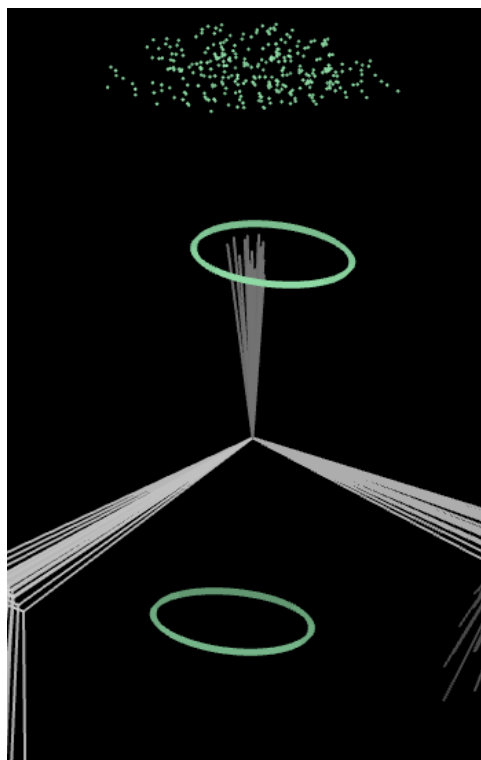


Figure 6-18 Ala 60 CNS with Gel RDCs

For the Gel-only RDC data (Figure 6-18), the NH's cluster together in all the models, pointing toward their hydrogen bond partner. When compared to the ensemble with data from two media (Figure 6-17), one interpretation is that the H-bond term in the target function 'won' over the match to the RDC terms. This same behavior is present in the Xplor determinations at Ala 60 (not shown), and Ala 80 (below, Figure 6-19).

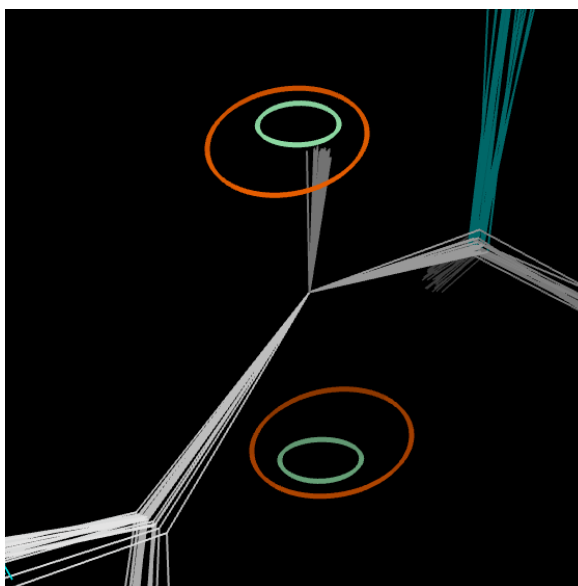


Figure 6-19 Ala 80 Xplor with Peg + GEL RDCs

These examples suggest that the weighting of the refinement terms can impact the local structural interpretation. RDCvis shows visually the results of how the practicing spectroscopist weighs the geometric and experimental terms in the refinement software. It also gives insight into how the two RDC datasets are weighted with respect to each other in the structure calculation.

6.3.4 Error Model Examples

As discussed in the previous chapter (5.4.5) I observed NMR structure ensembles with RDCs visualized on the models, where clustering of internuclear vectors on the RDC curves is extraordinarily tight - perhaps too tight.

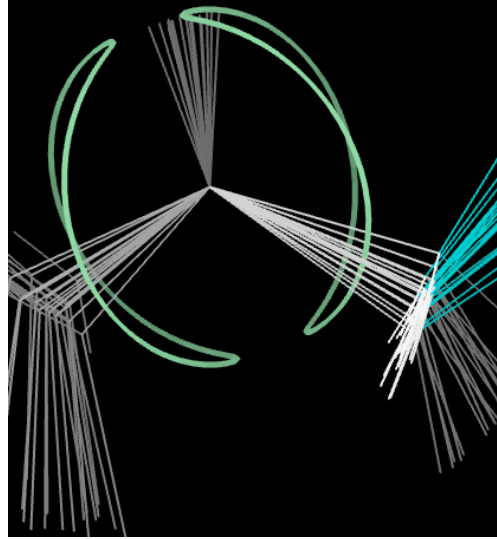


Figure 6-20 Asp 62 CNS with PEG RDCs

Asp 62 (Figure 6-20) shows a fanning of NH's along the solution curve and consistent with an orientation-dependent variability. However, some of the NH's creep towards the area between the two solution curves. It is likely that these are still within error since this position is near an X critical point.

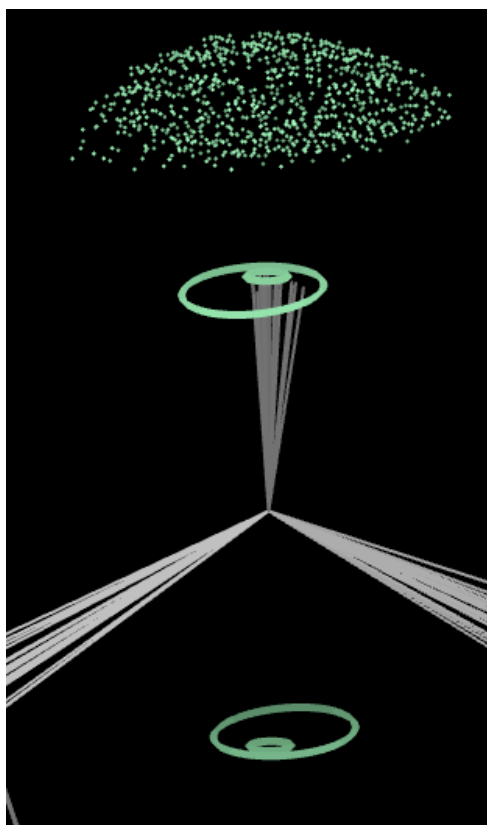


Figure 6-21 Leu 77 CNS with PEG + Gel RDCs

Leu 77 has very nice Hydrogen bond. Also, with two sets of RDC data available, all of the determinations are likely holding the NH's too tightly to the inner curve (as shown in Figure 6-21). The error model is probably too strict and contributes to this observation.

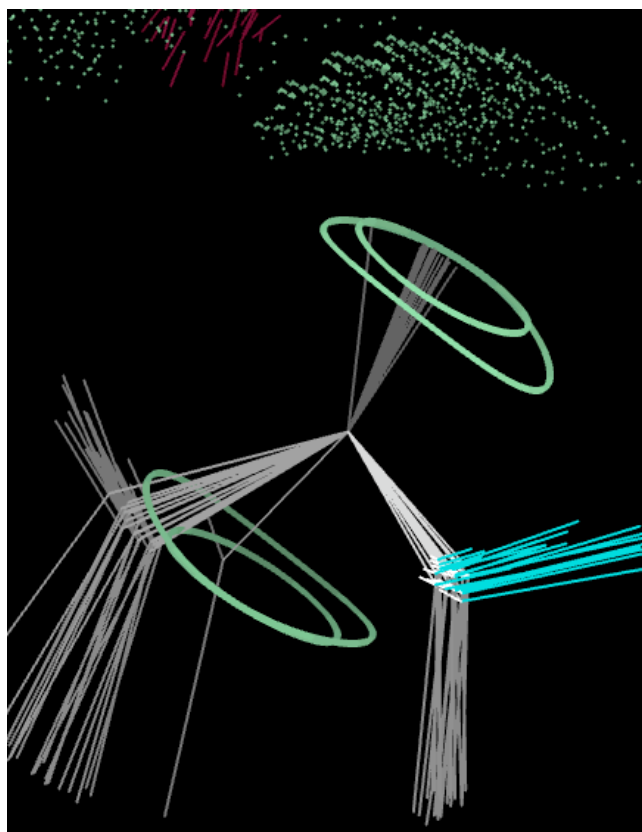


Figure 6-22 Phe 118 CNS with PEG + Gel RDCs

Phe 118 in the CNS ensemble using both sets of RDC data shows one model far away from the cluster of other models (shown in Figure 6-23). However, the outlier model still makes a hydrogen bond and does not have other problems with the local geometry. The model lies along the area where the two curves are ($\pm 2\text{Hz}$) likely close enough to be within error. I conclude that while this is far away from the rest of the cluster, there is insufficient data to suggest that this is an implausible configuration.

6.3.5 PEG vs. Gel vs. PEG & Gel Examples

Chief among the goals of this study was to use a real structural example with RDCvis in order to better understand where adding more data (or a different dataset at the same position) will produce local configurations that are better determined when put through the paces of structure determination. It is accepted that more data usually equates to more accurate structures and higher quality structures (discussed in greater detail in chapter 3), but such generalizations should be tested for the case of multiple RDCs.

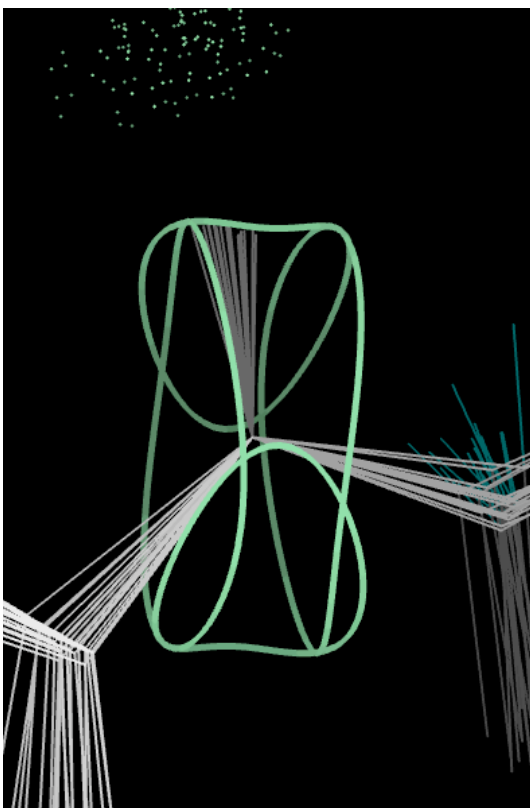


Figure 6-23 Glu 108 CNS with PEG + Gel RDCs

Glu 108 has experimental data from both Gel and PEG present. The intersection point of the two sets of RDC data shows a clustering of the NH's, and most of the models

have a Hydrogen bond present (Figure 6-23). There is some fanning of solutions along one of the curves, though this is probably within error for both curve pairs.

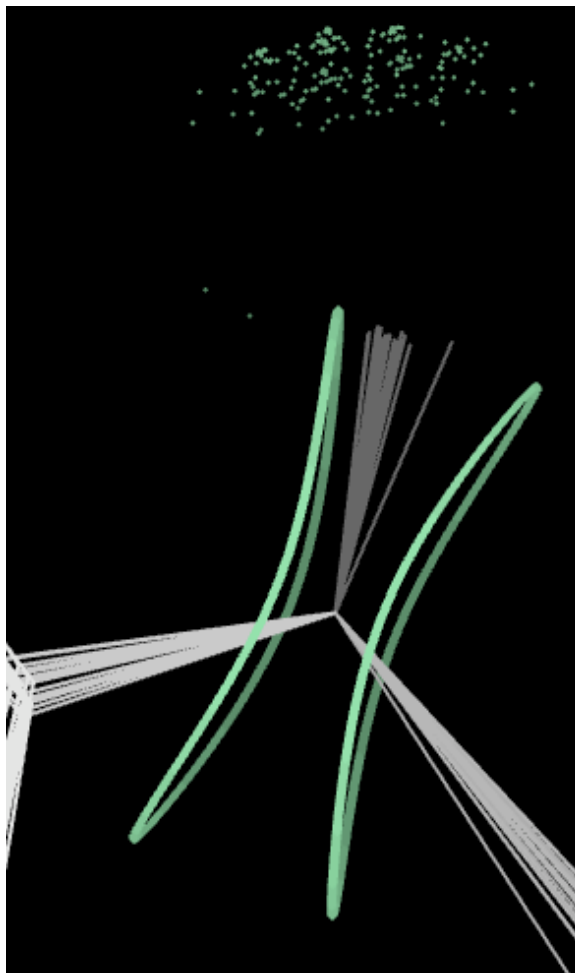


Figure 6-24 Glu 108 Xplor with Gel RDCs

Interestingly, Glu 108 in the Xplor determination using Gel alone shows the NH's clustering in between the two curves (Figure 6-24). Here, orientation-dependent variability does not go in the direction that would help satisfy this RDC. However, the curves are near the equator, where uncertainty is greater. Also, the addition of the PEG data in the determination in both CNS as shown before and Xplor (not shown) resolve

this problem while making a Hydrogen bond. At this position, the addition of more data adequately resolved the issue.

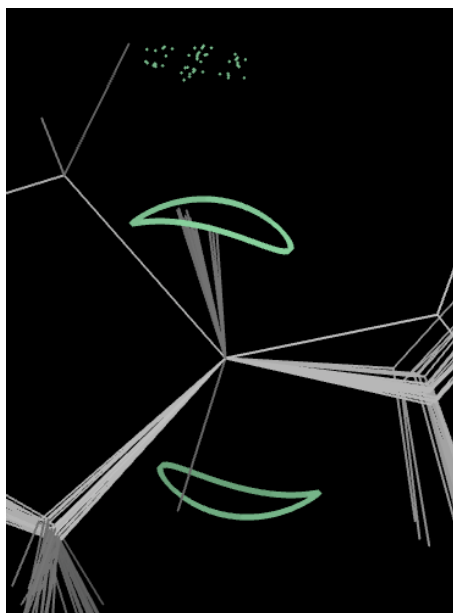


Figure 6-25 Gly 110 Xplor with Gel RDCs

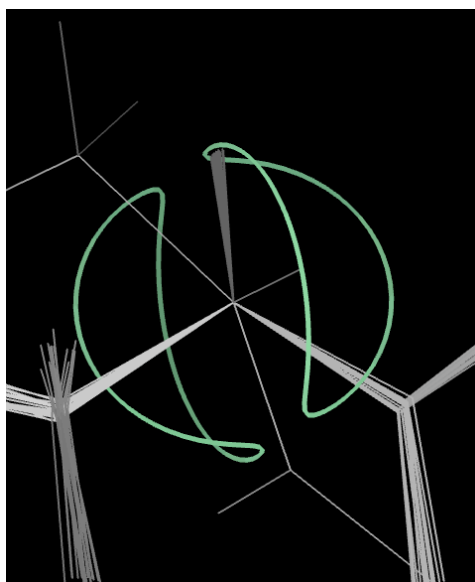


Figure 6-26 Gly 110 Xplor with PEG RDCs

At Gly 110 in the Xplor structures (shown above in Figure 6-25 and 6-26 respectively), a problem arises when using Gel RDC data where models are pointing at two opposite RDC curves (Figure 6-25), whereas in the PEG structure (Figure 6-26) the outlier is on the same curve. The Xplor ensemble using both data together still has one model off to the side, while all the rest point the NH's towards the intersection point (shown below in Figure 6-27).

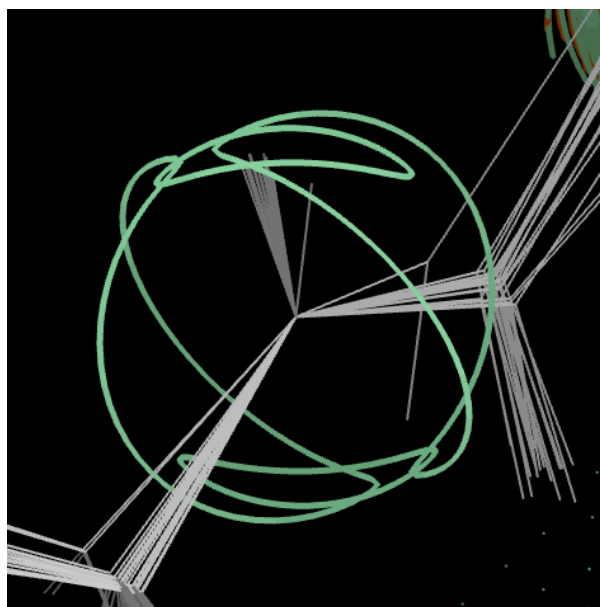


Figure 6-27 Gly 110 Xplor with PEG + Gel RDCs

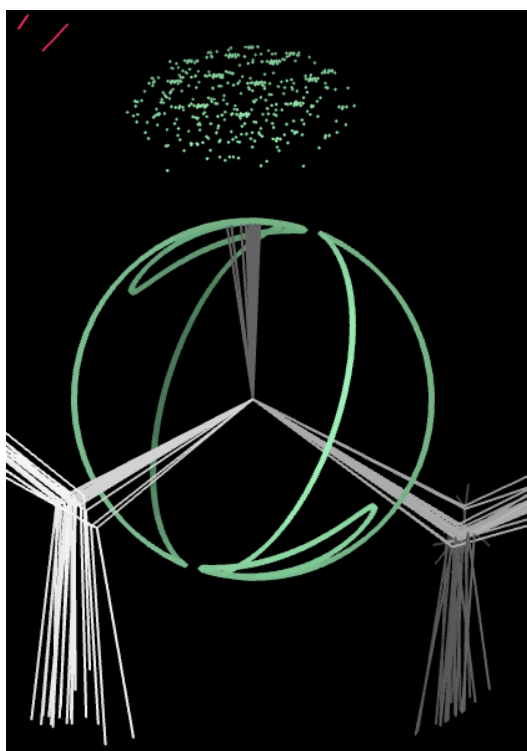


Figure 6-28 Gly 110 CNS with PEG + Gel RDCs

When using both data together in CNS, (shown above in Figure 6-28) all of the NH's point towards the intersection point of the RDC curves from the two media and make a Hydrogen bond. Thus, the single-medium problems were resolved in CNS but not in Xplor-NIH by adding a second dataset. The differences in target-curve and ensemble positioning that produce this better outcome are rather small and subtle.

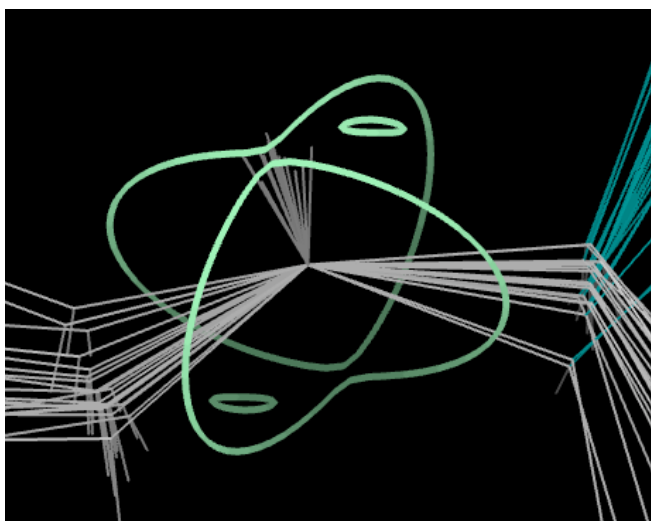


Figure 6-29 Leu 111 CNS with PEG + Gel RDCs

Leu 111 has RDC data available in both PEG and Gel. In both the CNS and Xplor ensembles using both sets of data (CNS shown in Figure 6-29), the NH's all point near the close approach of the PEG RDC curve pair (large curves).

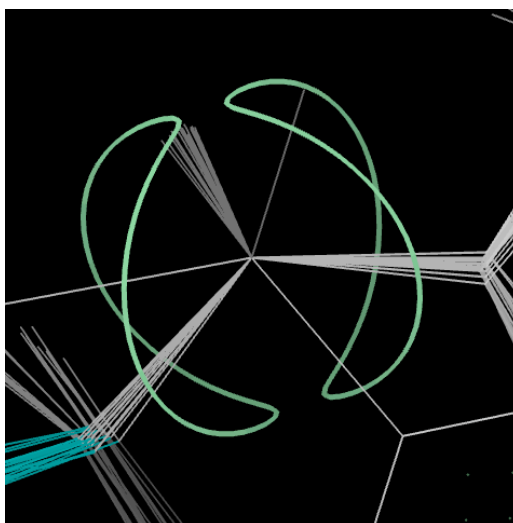


Figure 6-30 Leu 111 Xplor with PEG RDCs

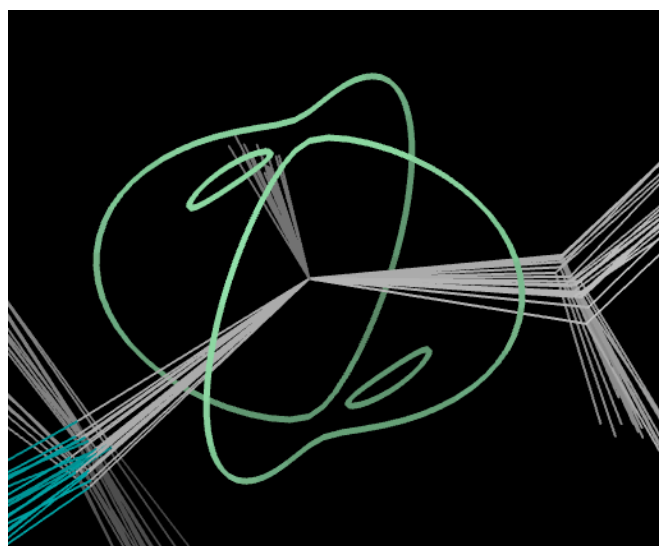


Figure 6-31 - Leu 11 Xplor with PEG + Gel RDCs

The PEG determination in Xplor has one outlier model where the NH points towards the opposite curve and is far out along that curve (shown in Figure 6-30 and seen from the other side than the previous figure). This problem was resolved in the Xplor-NIH ensemble that used both PEG and Gel RDC data and shown in Figure 6-31.

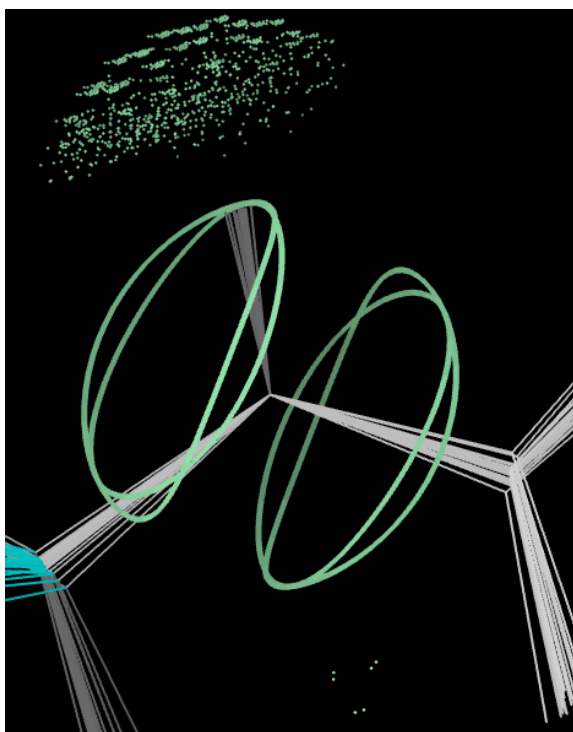


Figure 6-32 Thr 65 CNS with PEG + Gel RDCs

Comparing the Thr 59 (Figure 6-16 in section 6.3.3) with the Thr 65 above (Figure 6-32), one notices that while both have very nice hydrogen bonding, and tight clustering of the NH's modeled – the shape of the RDC curves from PEG and Gel are much more similar to one another at position 65 in the CNS ensemble. Here, the addition of the extra data adds little in terms of extending the local structural interpretation (but is not needed), whereas in the following example (Figure 6-33), extra information that is different would have been very helpful.

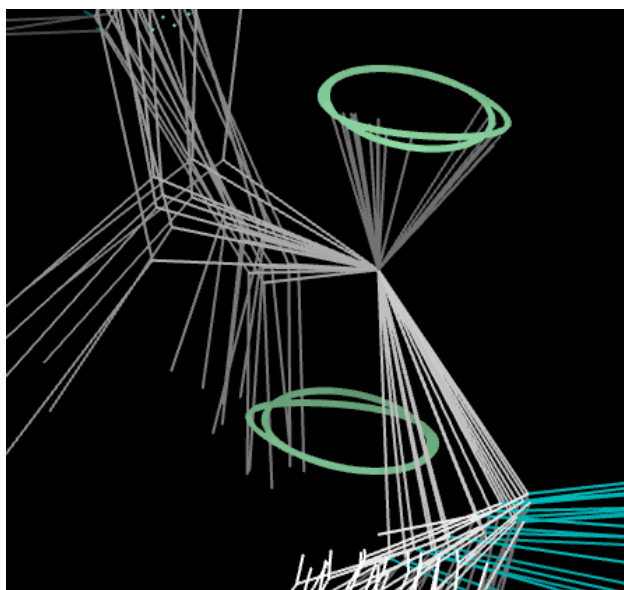


Figure 6-33 Thr 117 CNS with PEG + Gel RDCs

Thr 117 has data from both PEG and Gel available (Figure 6-33). While all the NH's modeled are touching both curves, the addition of extra data does not assist the structural interpretation here. More importantly, this example shows orientation-dependent variability resulting in configurations spread out widely around the RDC curve. This is consistent in the Xplor-NIH ensembles and in ensembles using only PEG or Gel data alone (not shown), making a single interpretation difficult without the addition of more data (such as NOE's or differences in validation criteria).

6.3.6 CNS vs. Xplor Examples

Comparing CNS vs. Xplor structure ensembles is difficult. The resulting ensembles could be slightly different from one another even if the same exact procedure were followed twice using the same determination package and the same inputs, since there are random steps involved. That said, comparing the two packages in this manner with varying amounts of data has not been reported in the literature (to my knowledge).

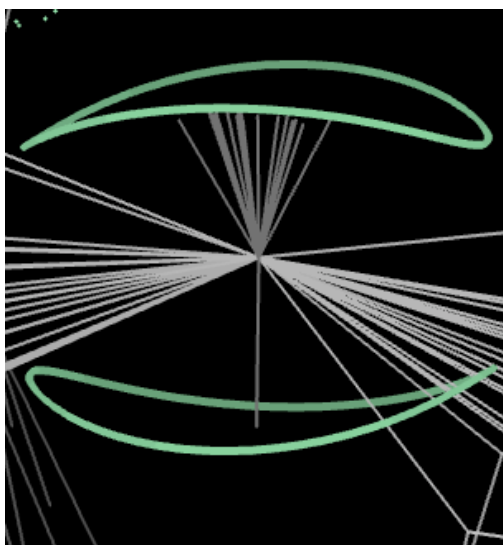


Figure 6-34 Asp 68 CNS

Asp 68 depicts in CNS (Figure 6-34) a systematic error where one of the NH's is pointing to the opposite RDC curve from the others (the One-Curve Rule discussed in Chapter 5 and more examples in section 6.3.2). Interestingly, Asp 68 has only PEG RDCs observed, but the CNS determination using both PEG & Gel RDC experiments to calculate the ensemble (shown in Figure 6-34) is the only one with an NH pointing to the opposite curve. Asp 68 has a fanning of NH's along the curve, satisfying orientation-

dependent variability, and the other CNS determinations (PEG or Gel alone, not shown) had the same fanning.

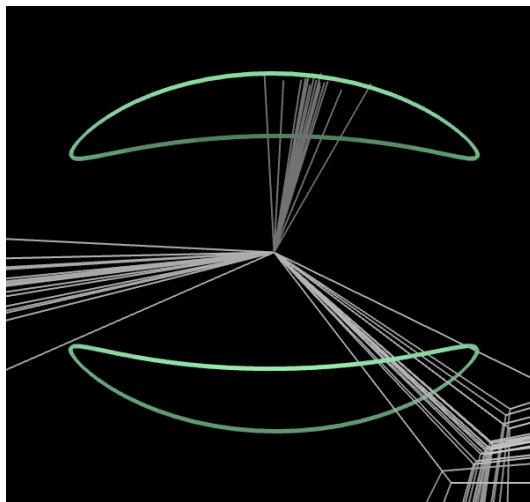


Figure 6-35 Asp 68 Xplor

To muddy the waters, Asp 68 in the Xplor ensembles (Figure 6-35) using both RDC experiments to calculate the ensemble did not have any NH's point to the opposite curve and the companion Xplor determination with only a single dataset used (PEG or Gel alone, not shown) looked similar. The Asp 68 was handled better in Xplor than in CNS. However, with few other experimental observations in this local area to restrain the backbone or sidechain geometry, this part of the structure is not well “understood.”

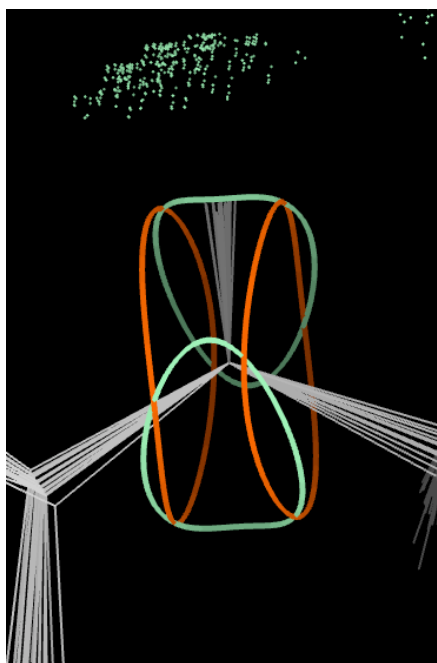


Figure 6-36 Gly 109 CNS with PEG + Gel RDCs

Gly 109 has RDC data available for both PEG and Gel (CNS ensemble shown in Figure 6-36). In both CNS and Xplor, when both RDC datasets are used in the structure determination, the NH's fan out slightly along the green PEG solution curve (though not in the direction of a motional axis of the peptide) while making a hydrogen bond.

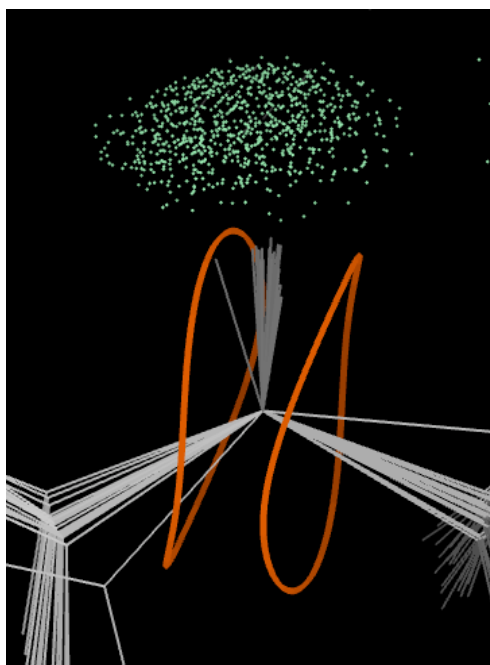


Figure 6-37 Gly 109 Xplor with Gel RDCs

When using Gel alone, in both CNS and Xplor structures, all of the NH's – except for one model in the Xplor structure shown here – do not match the Gel RDC curves (Xplor shown in Figure 6-37). Therefore, at Gly 109 in Xplor using only Gel RDCs, the hydrogen bond presumably trumps over a match to the Gel RDC data.

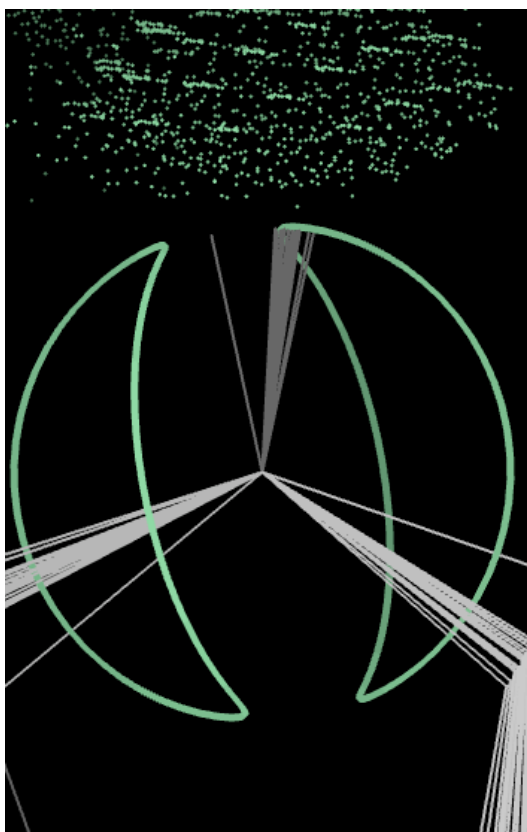


Figure 6-38 Val 53 Xplor with PEG RDCs

In the Xplor ensembles for Val 53, when no RDC data is used, the structures are all consistent and tightly clustered (not shown). When the PEG data is used in Xplor, one of the models points more towards the opposite RDC solution curve (Figure 6-38). This behavior persists in the Xplor ensemble (not shown, where both sets of data are used though only PEG data was observed at this NH).

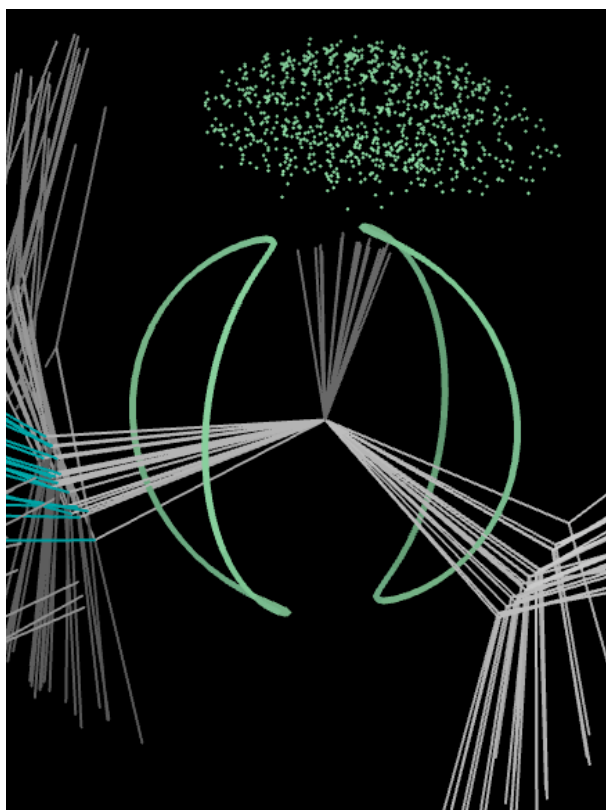


Figure 6-39 Val 53 CNS

In the CNS ensembles, more of the NH's spread into the space between the two curves (Figure 6-39, both RDC datasets used in the calculation, but only PEG data observed). In both the CNS and Xplor ensembles, none of the NH's cross all the way over to the opposite curve, and a tighter clustering is observed in the Xplor ensembles. The position on the sphere is near the X point, so that the RDC does not very strongly constrain the conformation at Val 53.

6.3.7 Orientation-Dependent Variability Examples

As described in chapter 5, there exists some variation in the internuclear vector match to the RDC data drawn as a curve, that can result in a fanning out of the internuclear vector along a target curve that I've termed "orientation-dependent variability," quite acceptable when it follows the path of allowable peptide rotation. The orientation of the alignment tensor to the molecule (and its rhombicity) will determine the shape of the RDC curves, and the orientation of the local structural features of the molecule in relation to the given RDC curve shape will determine the amount and direction of variation allowable for structural interpretation.

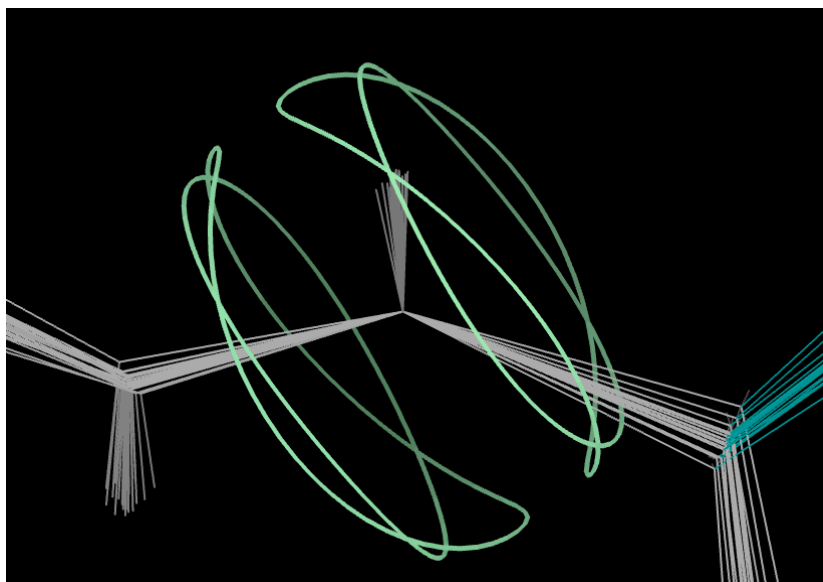


Figure 6-40 Ala 81 Xplor with PEG + Gel RDCs

The NH's of Ala 81 (Figure 6-40) all point towards the intersection point of the RDC curves for the two media. Unlike Ala 61 (section 6.3.1), a peptide motion here move the NH's perpendicular to rather than along the solution curve, and the resulting cluster of solutions observed could thus be interpreted as restrictively as shown.

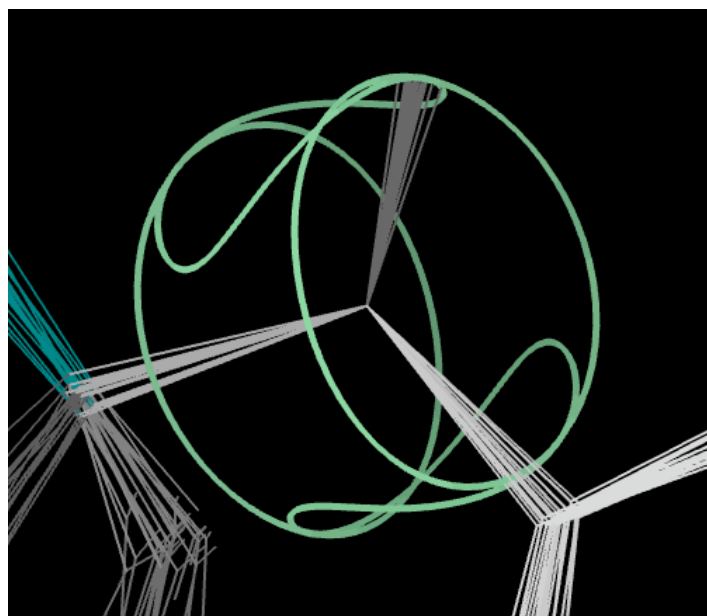


Figure 6-41 Ala 94 CNS with PEG + Gel RDCs

The two pairs of RDC target curves of Ala 94 for the two media form a striking cylinder shape around the NH's. Interestingly, the intersection of the two curves - where a fan of solutions appear - is wide and the tangent is along the motional axis of a peptide movement (Figure 6-41). It is possible that even more movement could be supported by the data.

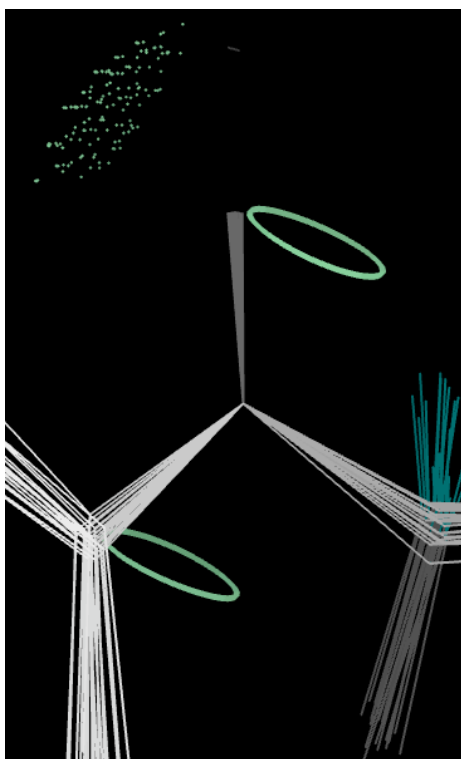


Figure 6-42 Arg 76 CNS with Gel RDCs

Similar to Ala 120 (section 6.3.3), Arg 76 matches the NH's near to the RDC solution curve while also making a Hydrogen bond (Figure 6-42 above). For a peptide rotation here, the fan of solutions would not follow the curve and therefore the tight clustering seems very reasonable.

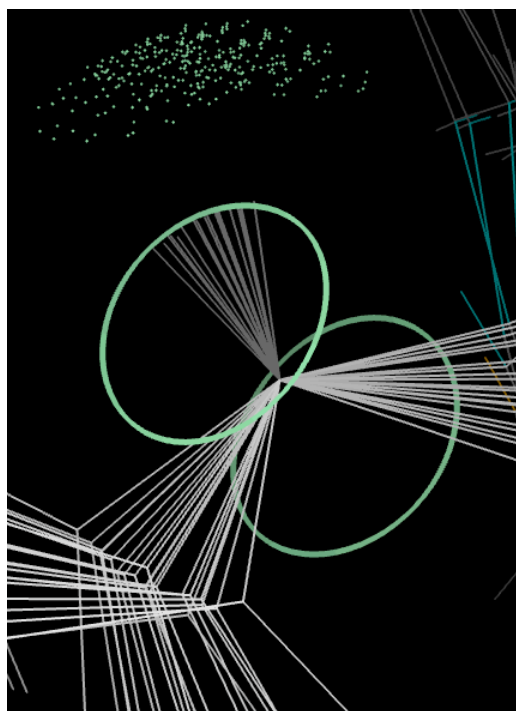


Figure 6-43 Lys 85 CNS with Gel RDCs

Lys 85 in the CNS ensemble with Gel RDC data (Figure 6-43 above) shows an orientation-dependent variability where the NH's fan out along the solution curve and most of the models make Hydrogen bonds.

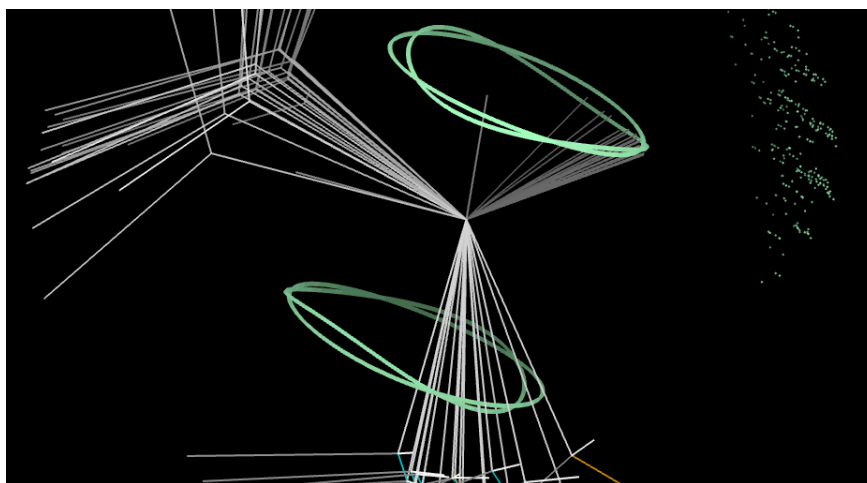


Figure 6-44 Lys 85 CNS with PEG + Gel RDCs

When RDC data is added from PEG and determined with CNS (shown in Figure 6-44) the result fans out even further and includes one model that remains on the same curve as the other solutions, but points in the opposite direction and does not make a Hydrogen bond. Unfortunately, the RDCs from Peg and GEL are similar enough that the addition of more experimental data does little to help the structural interpretation, and seems to actually make it a bit worse, since the single outlier is not well supported.

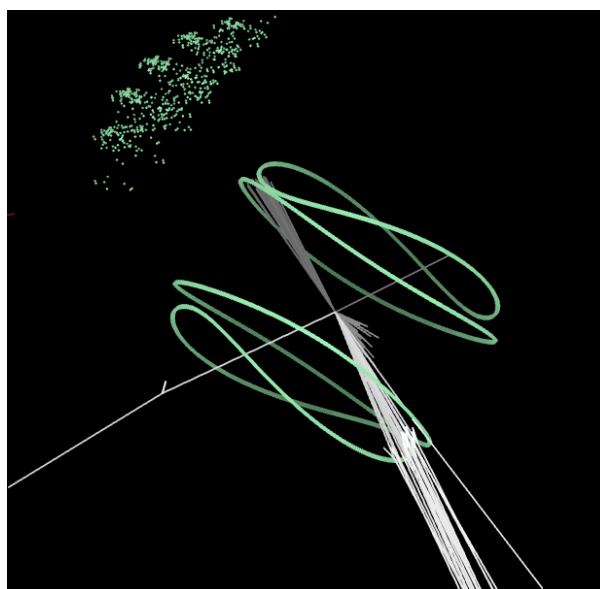


Figure 6-45 Val 103 CNS with PEG + Gel RDCs

Val 103 has data from both PEG and Gel present. In the CNS ensemble (shown in Figure 6-45), one of the models is significantly different from the rest. It does not have a hydrogen bond like the others. However, it is pointing to an area along the same RDC curve and clearly not matching both sets of RDC data. What I conclude from this example is that the configurations with a hydrogen bond are much more favorable and the one model pointing in the other direction is an outlier. Notably, this was the only

ensemble where this appeared. The other five looked very good at this position (not shown), even when using only a single set of RDC data.

6.3.8 Other Examples

In a few instances in CcmE, visualizing RDC data on the structures using RDCvis allowed me to identify especially thorny and complicated examples.

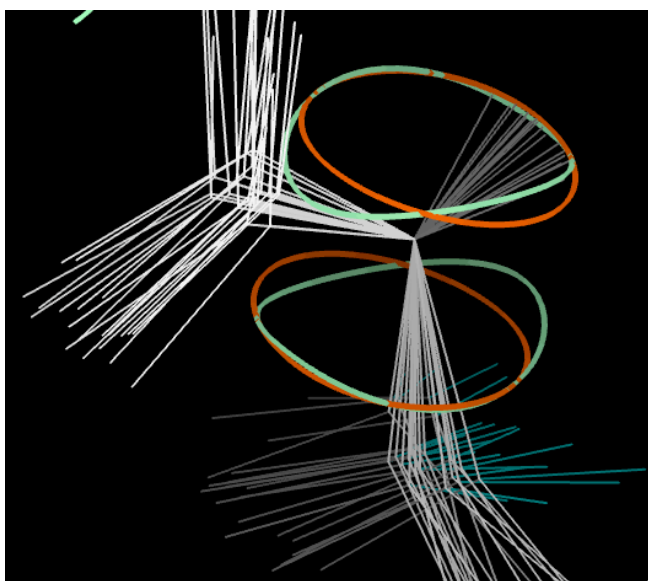


Figure 6-46 Ala 112 CNS with PEG + Gel RDCs

The NH's of Ala 112 (Figure 6-46) fan out along the intersection range between the sets of curves. If one were to view looking down on the curves from the top, the NH's spread out along about $\frac{1}{4}$ of the RDC curve. When I looked at whether adding more data (PEG vs. Gel vs. PEG and Gel combined) could tighten things up a bit, there is no ensemble that looks reasonable for all of the configurations, and no clearly preferred configuration.

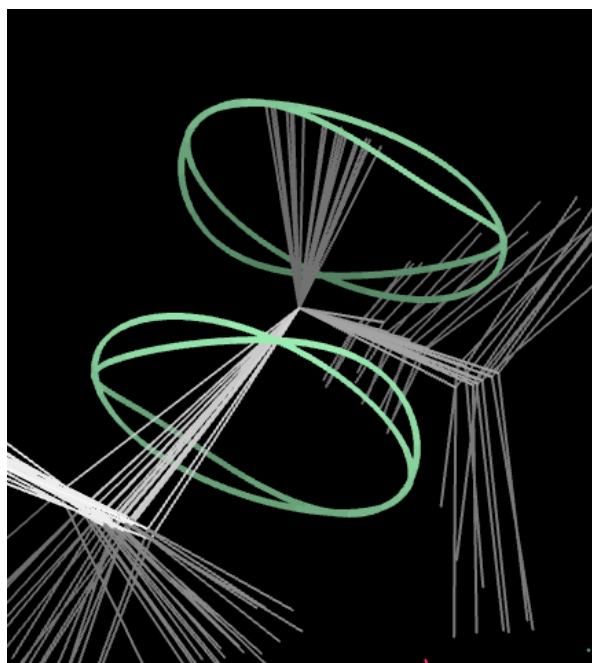


Figure 6-47 Gly 63 CNS with PEG + Gel RDCs

When RDCs from PEG and Gel are present, Gly 63 shows a broad fan of orientation-dependent variability (Figure 6-47 above), in the CNS-determined ensemble, and in most of the other structure ensembles.



Figure 6-48 Gly 63 Xplor with PEG RDCs

However, clustering is more pronounced (tighter, and clearly separated) in the Xplor ensemble with Peg RDCs (Figure 6-48 above) where there is a noticeable gap between the NH clusters and the resulting backbone geometry before and after the NH is significantly different for each cluster, pointing the carbonyl in very different directions. This region of the structure is not generally well restrained by other data, and it is unclear that this two-cluster ensemble is any more accurate than the more continuous fans.

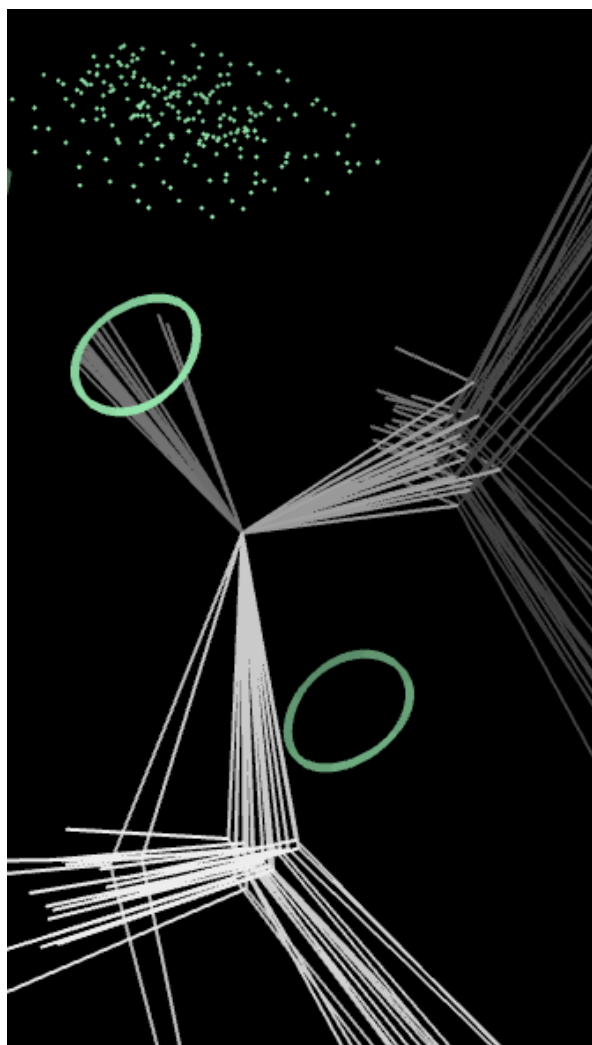


Figure 6-49 Gly 102 CNS with PEG RDCs

In the CNS structures, Gly 102 is sometimes artificially separated into distinct, tight clusters and more movement along a peptide motional axis could be allowable (more orientation-dependent variability). This happens with distinct clusters in the PEG-only and Gel-only ensembles (PEG shown in Figure 6-49, Gel not shown).

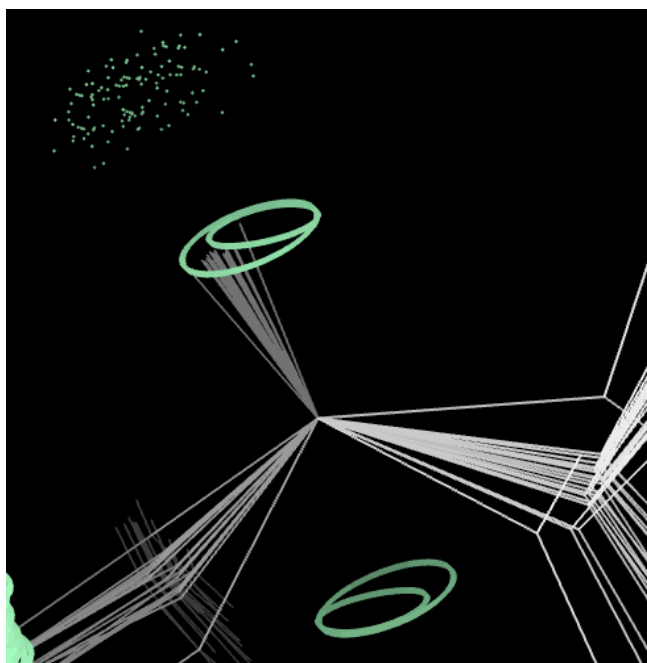


Figure 6-50 Gly 102 CNS with PEG+ Gel RDCs

When both PEG and Gel data are used in CNS, the distinct clusters mostly disappear and a tighter grouping is seen between the RDC curves (shown in Figure 6-50). Ultimately, the addition of more data helped here.

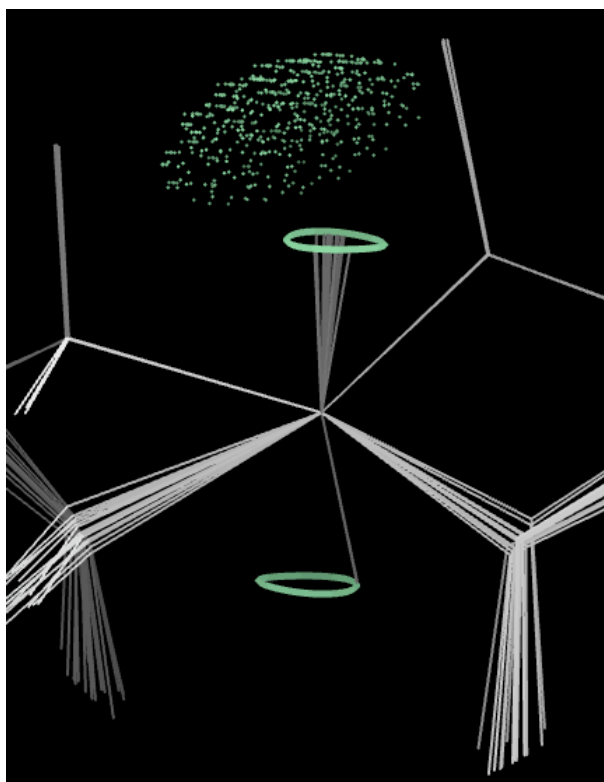


Figure 6-51 Gly 102 Xplor with PEG RDCs

The Xplor structures tell a slightly different story for Gly 102 than the CNS structures. In all three Xplor ensembles, a small number of models point to the opposite solution curve (PEG shown in Figure 6-51). However, the majority of models point the other way and make a Hydrogen bond. This strongly suggests that the rare models without the Hydrogen bond are likely incorrect.

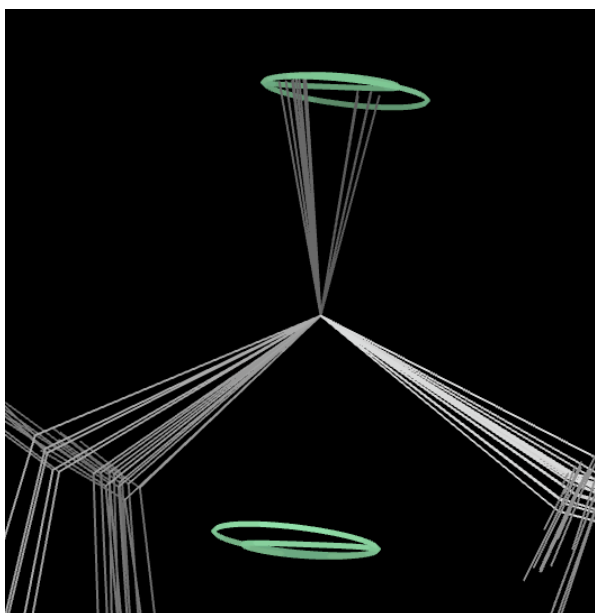


Figure 6-52 Gly 102 Xplor with PEG + Gel RDCs cluster of 17 models

Interestingly, 17 of the 20 models in the Xplor ensemble where both Peg and GEL RDCs are present make a hydrogen bond, but two distinct sub-clusters appear on opposite sides of the same RDC curve (shown in Figure 6-52). As described in Chapter 5, an arc of conformations across the center of the curves, connecting the two groups, might be reasonable with a less inflexible error model. This example shows that the use of RDCvis can guide the user in identifying more than one layer of problem at a given residue; here there may be both an artificially inflexible error model and also a problem with pointing NH's to opposite curves, where one of those issues (the opposite curve problem) is resolved by adding more data to the structure determination.

6.4 Discussion

The tools, visualization modes, and techniques developed in previous chapters proved critical in the analysis of the CcmE structure. The results showed many different types of examples in the various ensembles using RDC data. A wide variety of problems were illustrated, generating insights on diagnosing those problems and in a few instances suggesting proposed solutions.

6.4.1 Conclusions

Using more than one set of RDC data indeed produces unique solutions much more often. Typically this occurs where the two pairs of RDC curves are different enough to have unique intersection points (as in section 6.3.1). Subtle differences in the tensor combined with local structure can in some cases produce pairs of RDC curves that are nearly indistinguishable from one another (as in Figures 6-9, 6-10, 6-32, 6-33), in which case the additional data is unhelpful.

Most of the time, the results from CNS and Xplor-NIH were in very good agreement. There were some cases where results from the two determination packages differed (in section 6.3.6). The shape of the RDC curve for the same location and with the same data can be subtly different between CNS and Xplor-NIH structures because of how the tensors are determined and refined (resulting in slightly different RDC curves). Despite these differences, there is no overall superiority of CNS vs. Xplor-NIH, with some residues looking better in each. This is not surprising, though satisfying to have tested more thoroughly.

There is considerable non-uniformity of uncertainty in allowable directions on the hyperboloid sphere of the RDC. This issue arises from: a) non-uniform uncertainty in the RDC values which are tighter near the poles, looser near the equator of the sphere, and very loose near the intersection point, and b) non-uniform uncertainties in the model (especially for NHs), since a peptide can rotate easily only around $C\alpha-C\alpha$ direction. These non-uniformities imply that any constant error estimate is unrealistic, such as the 3Hz we used for the CcmE work, or the rule-of-thumb 10% of the total RDC range. In addition, both in the CcmE structures and others I have looked at, there is some evidence that the error model is overly restrictive; for instance, NHs fanning out along a curve but not at all perpendicular to it (e.g. Figure 6-46), or an interrupted fan of NHs across a small circle with nothing in the middle (e.g. Figure 5-9).

6.4.2 Recommendations

Several aspects of the procedures used here could suitably be incorporated into structure determination protocols.

This work attempts to move the NMR community towards visualizing and inspecting each RDC in its local context with geometric and steric criteria shown, in order to better understand the match of the models with the underlying RDC data. This is just as necessary as for a crystallographer to study the electron density.

The One Curve Rule would be very useful for the NMR community as an obvious red flag indicating that a residue needs closer scrutiny.

Loops are very likely to have multiple conformations. However, simple treatments of RDC data are only valid for a single tight ensemble. Therefore, a good practice (already done by some labs) would be to first determine the structure of just the core (not including loops). The core could be identified using secondary structure assignments combined with order parameter information. The loops could then be built as a later step onto the core scaffold.

6.4.3 Deposition

The six ensembles calculated for this work will be deposited along with the restraint data, and the 2KCT structure (PEG RDCs alone in CNS) will be obsolete. Additionally, a best-parts approach (similar to the chimera approach described in chapter 3) can now be performed on the ensembles to rule out poor individual models, using steric and geometric criteria, along with match of NOE and RDC data to the models (using NOEdisplay and RDCvis) to create a suitable ensemble for deposition representing a final, high quality, structure of CcmE.

7. Conclusions & Future Directions

This thesis develops productive new ways of representing, interacting with, and improving NMR ensembles.

7.1 MolProbity NMR

These new NMR validation methods are gaining wide use by the NMR community, in addition to their direct use in the MolProbity website. The early and continuous collaboration with the NESG led to inclusion of MolProbity analyses in the PSVS system they wrote and use. Recently, after I gave a talk at an NESG workshop in Buffalo, descriptions of MolProbity usage during NMR structure determination will now become part of the NESG Wiki. The structural genomics community, as a whole, widely adopted the use of MolProbity tools as standard methods in their work, first for X-ray and now for NMR structures. More formal inclusion of this work into standard practice will result from the Richardson lab involvement in both the wwPDB Xray Validation Task Force, and the newly constituted NMR Validation Task Force.

7.2 KinImmerse

The development of KinImmerse serves as a driving problem for virtual reality. The complications associated with modeling and manipulating macromolecular structures can push the edge of the capabilities of a virtual reality system such as the 6-sided DiVE immersive CAVE system at Duke. Certainly, only some of the elements necessary for creating a system capable of production research use for the structural biologist are

themselves novel features of virtual reality systems. However, the useful combination of new interface features with the new hardware and software will lead to many more discoveries, and may spread to usage in other scientific applications.

What the DiVE work did for the progress of this thesis cannot be overstated. The development of the co-centering tool, a critical part of the RDCvis tool and the analysis of RDCs in the local context of a structure, was first devised and implemented in KinImmerse, and the first RDC examples were visualized in KinImmerse. One could argue that these developments could have occurred without it, but the fact remains that KinImmerse is where they came into being.

7.3 RDCvis

The development of RDCvis, and the co-centering tool inspired by and first tested in KinImmerse, allows NMR spectroscopists to interact with their RDCs, along with the geometric and steric criteria from MolProbity, within the local context characteristic of NMR. New patterns of model and data behavior, such as the One Curve Rule and Orientation Dependent Variability and the more complicated issues surrounding the non-uniformity of uncertainty in the RDCs, give insights into how to think about the match of the model to the RDC data beyond looking at numbers in a table.

7.4 NMR Structure Improvement

Overall, this work surveys a number of areas where the geometric and steric criteria developed in the Richardson lab can be usefully and appropriately applied to the determination, visualization, and improvement of NMR structures. As part of this thesis work, the CcmE structure has been improved. Rooted in the understanding of macromolecules developed by crystallographic work in the last 60 years, bolstered by the methods developed for NMR in the last 30 years; this work attempts to bring a more critical eye to NMR research in order to push forward towards an improved understanding of the underlying molecular reality.

Appendix

The table below contains the PDB identifier (four letter alpha-numeric code) and the PDB title for the NMR structures analyzed in chapter three.

1C7V	CALCIUM VECTOR PROTEIN
1DV9	BETA-LACTOGLOBULIN
1E17	AFX
1E41	FADD PROTEIN
1EGX	VASODILATOR-STIMULATED PHOSPHOPROTEIN
1EHX	SCAFFOLDIN PROTEIN
1EIW	HYPOTHETICAL PROTEIN MTH538
1EO1	HYPOTHETICAL PROTEIN MTH1175
	6-HYDROXYMETHYL-7,8-DIHYDROPTERIN
1EQ0	PYROPHOSPHOKINASE
1EZO	MALTOSE-BINDING PERIPLASMIC PROTEIN
1EZY	REGULATOR OF G-PROTEIN SIGNALING 4
1F0Z	THIS PROTEIN
	PROTEIN (APOPTOSIS REGULATOR BAX, MEMBRANE
1F16	ISOFORM ALPHA)
1F6V	DNA TRANSPOSITION PROTEIN
1FA4	PLASTOCYANIN
1FAF	LARGE T ANTIGEN
1FH3	LQH III ALPHA-LIKE TOXIN
1FHO	UNC-89
1FI6	EH DOMAIN PROTEIN REPS1
1FJD	PEPTIDYL PROLYL CIS/TRANS ISOMERASE (PPIASE)
1FPW	CALCIUM-BINDING PROTEIN NCS-1
1FR0	ARCB
1FZT	PHOSPHOGLYCERATE MUTASE
1G03	HTLV-I CAPSID PROTEIN
1G47	PINCH PROTEIN
1G4F	BETA2-GLYCOPROTEIN I
1G6E	ANTIFUNGAL PROTEIN
1G6J	UBIQUITIN
1G9L	POLYADENYLATE-BINDING PROTEIN 1
1GD5	NEUTROPHIL CYTOSOL FACTOR 1
1GE9	RIBOSOME RECYCLING FACTOR
1GGW	PROTEIN (CDC4P)

1GH1 NONSPECIFIC LIPID TRANSFER PROTEIN
1GH8 TRANSLATION ELONGATION FACTOR 1BETA
1GHH DNA-DAMAGE-INDUCIBLE PROTEIN I
1GJX PYRUVATE DEHYDROGENASE
1GO0 50S RIBOSOMAL PROTEIN L30E
1GXE MYOSIN BINDING PROTEIN C, CARDIAC-TYPE
1H2O MAJOR ALLERGEN PRU AV 1
1H3Z HYPOTHETICAL 62.8 KDA PROTEIN C215.07C
1H4B POLCALCIN BET V 4
1H5P NUCLEAR AUTOANTIGEN SP100-B
1H95 Y-BOX BINDING PROTEIN
1H9C PTS SYSTEM, CHITOBIOSE-SPECIFIC IIB COMPONENT
1HA6 MACROPHAGE INFLAMMATORY PROTEIN 3 ALPHA
1HPW FIMBRIAL PROTEIN
1HS7 SYNTAXIN VAM3
1HZK C-1027 APOPROTEIN
1E0Z FERREDOXIN
1I11 TRANSCRIPTION FACTOR SOX-5
1I42 P47
1IBI CYSTEINE-RICH PROTEIN 2
1IE5 NEURAL CELL ADHESION MOLECULE
1IEH BRUC.D4.4
1IEZ 3,4-Dihydroxy-2-Butanone 4-Phosphate Synthase
POLYADENYLATE-BINDING PROTEIN, CYTOPLASMIC
AND NUCLEAR
1IFW
1IG6 MODULATOR RECOGNITION FACTOR 2
1IIO conserved hypothetical protein MTH865
1IRY hMTH1
1IX5 FKBP
1IYY RIBONUCLEASE T1
1J0T MOLT-INHIBITING HORMONE
1J2O Fusion of Rhombotin-2 and LIM domain-binding protein 1
1J3C High mobility group protein 2
1J3X High mobility group protein 2
1J7Q Calcium Vector Protein
1J8C ubiquitin-like protein hPLIC-2
1J8K FIBRONECTIN
1JAS UBIQUITIN-CONJUGATING ENZYME E2-17 KDA
1JCU conserved protein MTH1692
1JDQ HYPOTHETICAL PROTEIN TM0983
HYPOTHETICAL 8.6 KDA PROTEIN IN AMYA-FLIE
1JE3 INTERGENIC REGION
1JFN APOLIPOPROTEIN A, KIV-T6
1JGK CANDOXIN

1JH3 TYROSYL-TRNA SYNTHETASE
1JI8 dissimilatory siroheme-sulfite reductase
1JJG M156R
1JNJ beta2-microglobulin
1JNS PEPTIDYL-PROLYL CIS-TRANS ISOMERASE C
1JQR DNA POLYMERASE BETA-LIKE
1JT8 PROBABLE TRANSLATION INITIATION FACTOR 1A
1JW2 HEMOLYSIN EXPRESSION MODULATING PROTEIN Hha
1JW3 Conserved Hypothetical Protein MTH1598
1JYT Olfactory Marker Protein
1JZU lipocalin Q83
1K0S CHEMOTAXIS PROTEIN CHEW
1K0X Melanoma Derived Growth Regulatory Protein
1K19 Chemosensory Protein CSP2
1K1C catabolite repression HPr-like protein
1K3J Protein Kinase SPK1
1K5W Synaptotagmin I
1K8H Small protein B
1KKG ribosome-binding factor A
1KMD Vacuolar morphogenesis protein VAM7
1KOT GABARAP
HEPATOCYTE NUCLEAR FACTOR 3 FORKHEAD
1KQ8 HOMOLOG 1
1KVI Copper-transporting ATPase 1
1L1P trigger factor
1L3G TRANSCRIPTION FACTOR Mbp1
1L5I Rep protein
1L7B DNA LIGASE
1L7Y HYPOTHETICAL PROTEIN ZK652.3
1LG4 Prion-like protein
1LKN hypothetical protein tm1112
1LS4 Apolipoprotein III
1M12 SAPOSIN C
1M39 Caltractin, isoform 1
fusion of the LIM interacting domain of ldb1 and the N-terminal
1M3V LIM domain of LMO4
1M5Z AMPA receptor interacting protein
1M7T Chimera of Human and E. coli thioredoxin
1M94 Protein YNR032c-a
1M9G Monellin chain B and Monellin chain A
1M9L Outer Arm Dynein Light Chain 1
1MG8 Parkin
1MJD DOUBLECORTIN
1MK3 Apoptosis regulator Bcl-W

Wiskott-Aldrich syndrome protein (WIP), GSGSG linker, and(N-
1MKE WASP)
1MM4 CrcA protein
1MVG Liver basic Fatty Acid Binding Protein
1MX7 CELLULAR RETINOL-BINDING PROTEIN I, APO
1MZK KINASE ASSOCIATED PROTEIN PHOSPHATASE
1N3G Protein yfiA
1N4C Auxilin
1N6U Interferon-alpha/beta receptor beta chain
1N88 Ribosomal protein L23
1N91 orf, hypothetical protein
1NEE Probable translation initiation factor 2 beta subunit
1NI7 Hypothetical protein ygdK
1NM7 Peroxisomal Membrane Protein PAS20
1NMW Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1
1NNV Hypothetical protein HI1450
1NO8 ALY
1NSO PROTEASE
1NWB Hypothetical protein AQ_1857
1NWV Complement decay-accelerating factor
1NXI conserved hypothetical protein VC0424
1NY8 Protein yrbA
1NY9 Transcriptional activator tipA-S
1NYA Calerythrin
1NYN Hypothetical 12.0 kDa protein in NAM8-GAR1 intergenic region
1NYO Immunogenic protein MPT70
1NYP PINCH protein
1NZP DNA polymerase lambda
1O1W RIBONUCLEASE H
1O6X PROCARBOXYPEPTIDASE A2
BIOTIN CARBOXYL CARRIER PROTEIN OF
1O78 METHYLMALONYL-COA CARBOXYL-TRANSFERASE
1O7B TUMOR NECROSIS FACTOR-INDUCIBLE PROTEIN TSG-6
1O8R GUANYLIN
1OJG SENSOR PROTEIN DCUS
1ON4 Sco1
1ONB helicase NS3
1OQA Breast cancer type 1 susceptibility protein
1OQK conserved protein MTH11
1OVQ Hypothetical protein yqgF
1OWA Spectrin alpha chain, erythrocyte
1OYI double-stranded RNA-binding protein
1OZI protein tyrosine phosphatase
1P1D Glutamate receptor interacting protein

1P1T Cleavage stimulation factor, 64 kDa subunit
1P4S Adenylate kinase
1P68 De novo designed protein S-824
1P6R Penicillinase repressor
1P6T Potential copper-transporting ATPase
1P88 3-phosphoshikimate 1-carboxyvinyltransferase
1P8A protein tyrosine phosphatase
1P9K orf, hypothetical protein
1PA4 Probable ribosome-binding factor A
1PBU Elongation factor 1-gamma
1PC0 Hypothetical protein AF1917
1PC2 mitochondria fission protein
1PFJ TFIIF basal transcription factor complex p62 subunit
1PJW envelope protein
1PJZ Thiopurine S-methyltransferase
1PLO TGF-beta receptor type II
1PN5 NACHT-, LRR- and PYD-containing protein 2
1POQ YPM
1POZ CD44 antigen
1PQN Spliceosomal U5 snRNP-specific 15 kDa protein
1PQX conserved hypothetical protein
1PSY 2S albumin
1PU1 Hypothetical protein MTH677
1PU3 P-30 protein
1PUN Mutator mutT protein
1PUX Sporulation initiation phosphotransferase F
1PUZ conserved hypothetical protein
1Q27 Putative Nudix hydrolase DR0079
1Q38 Fibronectin
1Q56 Agrin
1Q60 General transcription factor II-I
1Q80 Sarcoplasmic calcium-binding protein
1Q8X Cofilin, non-muscle isoform
1Q9P HIV-1 Protease
1QVP Diphtheria toxin repressor
1QVX Focal adhesion kinase 1
1QWV Pheromone-binding protein
1QZP dematin
1R36 C-ets-1 protein
1R4Y Ribonuclease alpha-sarcin
1R57 conserved hypothetical protein
1R73 50S ribosomal protein L29
1R9P NifU-like protein
1RDU conserved hypothetical protein

1RFL Probable tRNA modification GTPase trmE
1RG6 second splice variant p63
1RGW ZASP protein
1RH8 Piccolo protein
1RHW Dynein light chain 1, cytoplasmic
1RHX conserved hypothetical protein TM0979
1RI0 Hepatoma-derived growth factor
1RI9 FYN-binding protein
1RJA Tyrosine-protein kinase 6
1RJH Tetranectin
1RK7 Superoxide dismutase [Cu-Zn]
1RKN Thermonuclease
1RL1 Suppressor of G2 allele of SKP1 homolog
1RQ8 conserved hypothetical protein
1RQM Thioredoxin
1RQS 50S ribosomal protein L7/L12
1RSF Coxsackievirus and adenovirus receptor
1RW2 ATP-dependent DNA helicase II, 80 kDa subunit
1RWU Hypothetical UPF0250 protein ybeD
1RXL Afimbrial adhesin AFA-III
1RY4 CG5884-PA
1RYJ unknown
1RYK Protein yjbJ
1S04 hypothetical protein PF0455
1S3A NADH-ubiquinone oxidoreductase B8 subunit
1S6D Albumin 8
1S6I Calcium-dependent protein kinase SK5
1S6U Copper-transporting ATPase 1
1S7E Hepatocyte nuclear factor 6
1SA8 Fatty acid-binding protein, intestinal
1SB6 copper chaperone ScAtx1
1SCV Troponin C, slow skeletal and cardiac muscles
HOMOLOGUE OF THE THETA SUBUNIT OF DNA
1SE7 POLYMERASE III
1SE9 ubiquitin family
1SGO Protein C14orf129
1SJG Toluene-4-monooxygenase system protein C
1SJQ Polypyrimidine tract-binding protein 1
1SJR Polypyrimidine tract-binding protein 1
1SLJ Ribonuclease E
1SNL Nucleobindin 1
1SO9 Cytochrome C oxidase assembly protein ctaG
1SOU 5,10-methenyltetrahydrofolate synthetase
1SOY CyaY protein

1SPK RIKEN cDNA 1300006M19
 1SQ8 dh434
 1SQR 50S ribosomal protein L35Ae
 1SR3 APO-CCME
 1SRZ Gamma-aminobutyric acid type B receptor, subunit 1
 1SS6 NSFL1 cofactor p47
 1SSF Transformation related protein 53 binding protein 1
 1SW8 Calmodulin
 1SXD GA repeat binding protein, alpha
 1SXE Transcriptional regulator ERG
 1T0G cytochrome b5 domain-containing protein
 1T0Y tubulin folding cofactor B
 1T17 conserved hypothetical protein
 1T3K Dual-specificity tyrosine phosphatase
 1T3V conserved hypothetical protein
 1TDP carnobacteriocin B2 immunity protein
 1TE4 conserved protein MTH187
 1TIZ calmodulin-related protein, putative
 1TK7 CG4244-PB
 1TKN Amyloid beta A4 protein
 1TL4 Copper transport protein ATOX1
 1TM9 Hypothetical protein MG354
 1TQ1 senescence-associated family protein
 1TR4 26S proteasome non-ATPase regulatory subunit 10
 1TTX Oncomodulin
 1TVI Hypothetical UPF0054 protein TM1509
 1U5L prion protein
 1U6F RNA-binding protein UBP1
 1U81 ADP-ribosylation factor 1
 1UFM COP9 complex subunit 4
 1UFW Synaptojanin 2
 1UFZ Hypothetical protein BAB28515
 1UHC KIAA1010 protein
 1UHP hypothetical protein KIAA1095
 1UHU product of RIKEN cDNA 3110009E22
 1UHZ staufen (RNA binding protein) homolog 2
 1UIT HUMAN DISCS LARGE 5 PROTEIN
 1UJD KIAA0559 protein
 1UQV STE50 PROTEIN
 1UW0 DNA LIGASE III
 1UZC HYPOTHETICAL PROTEIN FLJ21157
 1V5J KIAA1355 protein
 1VEH NifU-like protein HIRIP5
 1W0A ALPHA-HEMOGLOBIN STABILIZING PROTEIN

1W4U UBIQUITIN-CONJUGATING ENZYME E2-17 KDA 2
1WI5 RRP5 protein homolog
1WI8 Eukaryotic translation initiation factor 4B
1WI9 Protein C20orf116 homolog
1WIA hypothetical ubiquitin-like protein (RIKEN cDNA 2010008E23)
1WIB 60S ribosomal protein L12
1WID DNA-binding protein RAV1
1WIF RIKEN cDNA 4930408O21
1WIH mitochondrial ribosome recycling factor
1WII Hypothetical UPF0222 protein MGC4549
1WIJ ETHYLENE-INSENSITIVE3-like 3 protein
1WIX Hook homolog 1
1WJ6 KIAA0049 protein
1WKI LSU ribosomal protein L16P
1WOT PUTATIVE MINIMAL NUCLEOTIDYLTRANSFERASE
1XFL Thioredoxin h1
1XHS Hypothetical UPF0131 protein ytfP
1XJS NifU-like protein
Chimeric CD3 mouse Epsilon and sheep Delta Ectodomain
1XMW Fragment Complex
1XN5 BH1534 unknown conserved protein
1XN8 Hypothetical protein yqbG
1XN9 30S ribosomal protein S24e
1XNE hypothetical protein PF0469
1XO3 RIKEN cDNA 2900073H19
1XO8 At1g01470
1XOY hypothetical protein At3g04780.1
1XPN hypothetical protein PA1324
1XPV hypothetical protein XCC2852
1XPW LOC51668 protein
1XSA Bis(5'-nucleosyl)-tetrphosphatase
1XU0 prion protein
1XWE Complement C5
1Y15 Major prion protein
1Y6D Phosphorelay protein luxU
1YEL At1g16640
1YHD UPF0269 protein yggX
1M4O Methionine Salvage Pathway Enzyme E-2/E-2'
1N4T 2',3'-cyclic nucleotide 3'-phosphodiesterase
1N9D Prolactin
1R83 DNA-binding protein 7a
1XO4 Proposed Acetyl Transferase
1XO9 hypothetical protein At3g03773
1Y41 Translationally controlled tumor protein

References

- Andersen CAF, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. *Structure* 10: 175-184.
- Anfinsen, C. (1957) On the Structural Basis of Ribonuclease Activity. *Molecular Structures and Biological Specificity*. Linus Pauling & Harvey Itan Ed. Waverly Press, Baltimore MD
- Arendall WB, III, Tempel W, Richardson JS, Zhou W, Wang S et al. (2005) A test of enhancing model accuracy in high-throughput crystallography. *Journal of Structural & Functional Genomics* 6: 1-11.
- Arnesano F, Bianci L, Barker PD, Bertini I, Rosato A, Su XC, Viezzoli MS (2002) Solution structure and characterization of the heme chaperone CcmE. *Biochemistry* 41: 13587-13594
- Arthur K, Preston T, Taylor RM II, Brooks FP Jr, Whitton MC, Wright WV: The PIT: Design, Implementation, and Next Steps. *Proc. 2nd Internat. Immersive Projection Tech. Workshop*, Ames, Iowa, 1998
- Arthur, K, Preston, T, Taylor II, RM, Brooks, FP, Whitton, MC, Wright, MV (1998) Designing and Bulding the PIT: a Head-Trackted Stereo Workspace for Two Users. *2nd International Immersive Projection Technology Workshop*. Ames Iowa, May 1998
- Ban, N., Nissen, P., Hansen, J. Moore, P.B., and Steitz, T.A. (2000). "The complete atomic structure of the large ribosomal subunit at 2.4Å resolution." *Science* 289: 905-920
- Ban Y-EA (2005) Applications of Computational Geometry and Topology to Structural Biology: Protein-Protein Interfaces and Protein Packing. Durham: Duke University.
- Ban Y-EA, Edelsbrunner H, Rudolph J (2005) Interface surfaces of protein-protein complexes. *Journal of the Association for Computing Machinery*
- Ban Y-EA, Rudolph J, Zhou P, Edelsbrunner H (2006) Evaluating the quality of NMR structures by local density of protons. *Proteins: Structure, Function, and Bioinformatics*
- Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput Biol* 5: e1000307.
- Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. *Chemical Reviews* 104: 3541-3556.

- Bechet, J., Greenson, M., and Wiame, J.M. (1970). Mutations affecting the repressibility of arginine biosynthetic enzymes in *Saccharomyces cerevisiae*. *European journal of biochemistry / FEBS* 12, 31-39.
- Bergman, L.D., J.S. Richardson, D.C. Richardson, F.P. Brooks, Jr. (1993): "VIEW—An Exploratory Molecular Visualization System with User-Definable Interaction Sequences," *Computer Graphics (Proceedings of SIGGRAPH 93)*, August 1993: 117-126.
- Berman (2006) Protein Data Bank. Available: <http://www.rcsb.org/pdb>.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al. (2000) The Protein Data Bank. *Nucleic Acids Research* 28: 235-242.
- Bernado, M. Blackledge (2004) Local Dynamic Amplitudes on the Protein Backbone from Dipolar Couplings: Toward the Elucidation of Slower Motions in Biomolecules *J. Am. Chem. Soc.*, 2004, 126 (25), pp 7760–7761
- Bertini I, Cavallaro G, Luchinat C, Poli I (2003) A use of Ramachandran potentials in protein solution structure determinations. *Journal of Biomolecular NMR* 26: 355-366.
- Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* 66: 778-795.
- Blanchard, S. C., Puglisi, J.D. (2001). "SOLUTION STRUCTURE OF THE A LOOP OF 23S RIBOSOMAL RNA." *Proc.Natl.Acad.Sci.* 98: 3720-3725.
- Blundell TLJ, L.N. (1976) *Protein Crystallography*: Academic Press, New York.
- Boelens, R., Koning, T.M.G., van der Marel, G.A., van Boom, J.H., and Kaptein, R. (1989). "Iterative procedure for structure determination from proton-proton NOEs using a full relaxation matrix approach. Application to a DNA octamer." *J. Magn. Reson.* 82: 290-308.
- Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J (1982) Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. *The Journal of Biological Chemistry* 257: 13650-13662.
- Bomar MG, Pai MT, Tzeng SR, Li SS, Zhou P (2007). Structure of the ubiquitin-binding zinc finger domain of human DNA Y-polymerase eta. *Embo Reports*, 8: 247-251.
- Bremi, T, Bruschiweiler, R (1997) Locally Anisotropic Internal Polypeptide Backbone Dynamics by NMR Relaxation *J. Am. Chem. Soc.*, 1997, 119 (28), pp 6672–6673
- Borgias, B. A., Gochim, M., Kerwood, D.J., and James, T.L. (1990). "Relaxation matrix analysis of 2D NMR data." *Prog. NMR Spectrosc.* 22(1): 83-100.
- Borgias, B. A., James, T.L. (1989). "Two-dimensional nuclear Overhauser effect: Complete relaxation matrix analysis." *Methods Enzymol.* 176: 169-183.

- Borgias, B. A., James, T.L. (1990). "MARDIGRAS-A procedure for matrix analysis of relaxation for discerning geometry of an aqueous structure." *J. Magn. Reson.* 87: 475-487.
- Brooks, CL, Karplus, M, Pettitt, BM, (1987) *Proteins, A Theoretical Perspective of Dynamics, Structures and Thermodynamics.* Wiley, New York.
- Brooks FPJ (1996) The Computer Scientist as Toolsmith II. *Communications of the ACM* 39: 61-68.
- Brooks, FP , Pique, M. (1984). *Computer Graphics for Molecular Studies in Molecular Dynamics and Protein Structure: Proceedings of a Workshop held 13-18 May 1984 at UNC* Jan Hermans ed., Published by Chapel Hill, University of North Carolina. Distributed by Polycrystal Book Service, 1985 pg 109
- Brooks, FP, Ou-Young, M, Batter, JJ, Kilpatrick, JP (1990) Project GROPE - Haptic Displays for Scientific Visualization. *Proceedings of SIGGRAPH '90 in Computer Graphics* 24 ACM SIGGRAPH, New York, 1990, pp 177-185
- Britton, E., J.S. Lipscomb, M. Pique, W.V. Wright, F.P. Brooks, Jr. (1981) The GRIP-75 Man-Machine Interface. Invited videotape presented at *1981 SIGGRAPH conference*, August 1981.
- Britton EG, Lipscomb JL, Pique ME (1978) Making nested rotations convenient for the user. *Computer Graphics* 1978, 12: 222-227
- Brunger AT (1992a) Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355(6359): 472-475.
- Brunger AT (1992b) *X-PLOR version 3.1: A System for X-ray Crystallography and NMR.* New Haven, CT: Yale University Press.
- Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Cryst D* 54: 905-921.
- Butterfoss G, Richardson JS, Hermans J (2005) Protein imperfections: separating intrinsic from extrinsic variation of torsion angles. *Acta Cryst D* 61: 88-98.
- Cai M, Huang Y, Suh JY, Louis JM, Ghirlando R, Craigie R, Clore GM. (2007) Solution NMR structure of the barrier-to-autointegration factor-emerin complex. *J Biol Chem* 2007;282:14525-14535.
- Carugo O, Argos P (1997) Correlation between side-chain mobility and conformation in protein structures. *Protein Engineering* 10(7): 777-787.
- Carson WM, Bugg CE (1986) Algorithm for ribbon models of proteins. *J. Molec. Graphics* 1986, 4: 121-122
- Cavanagh J, Fairbrother WJ, Palmer AG III, Rance M, Skelton NJ (2006) *Protein NMR Spectroscopy: Principles and Practice* (2nd Ed). 2006. Academic Press, San Diego

- CCP4: Collaborative Computational Project, Number 4. (1994). The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst.* D50, 760-763
- Chang, S.C., and Majerus, P.W. (2006). Inositol polyphosphate multikinase regulates inositol 1,4,5,6-tetrakisphosphate. *Biochemical and Biophysical Research Communications* 339, 209-216.
- Cherepanov P, Sun ZY, Rahman S, Maertens G, Wagner G et al. (2005) Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nat Struct Mol Biol* 12(6): 526-532.
- Chou JJ, Bax A (2001) Protein side-chain rotamers from dipolar couplings in a liquid crystalline phase. *J Am Chem Soc* 123(16): 3844-3845.
- Chou JJ, Case DA, Bax A (2003) Insights into the mobility of methyl-bearing side chains in proteins from (3)J(CC) and (3)J(CN) couplings. *J Am Chem Soc* 125(29): 8959-8966.
- Chung, J.C., M.R. Harris, F.P. Brooks, Jr., H. Fuchs, M.T. Kelley, J. Hughes, M. Ouh-young, C. Cheung, R.L. Halloway, Eriksen, C.W., Hoffman, J.E., (1972). Some characteristics of selective attention in visual perception determined by vocal reaction time. *Perceptions & Psychophysics* 11, 169-171.
- Clore GM, Garrett DS (1999) R-factor, Free R, and Complete Cross-Validation for Dipolar Coupling Refinement of NMR Structures. *Journal of the American Chemical Society* 121: 9008-9012.
- Clore GM, Schwieters CD (2004) How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J Am Chem Soc* 126(9): 2923-2938.
- Coggins, Brian: NOEDisplay plug-in to KiNG. personal communication.
- Coggins BE, McClerren AL, Jiang L, Li X, Rudolph J et al. (2005) Refined solution structure of the LpxC-TU-514 complex and pKa analysis of an active site histidine: insights into the mechanism and inhibitor design. *Biochemistry* 44: 1114-1126.
- Connolly ML (1993) Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1993, 221: 709-713
- Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J Am Chem Soc* 120: 6836-6837.
- Cornilescu G, Delaglio F, Bax A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13: 289-302.

- Clore GM, Gronenborn AM, Tjandra N. (1998) Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J Magn Reson* 1998;131:159-162.
- Clore, GM, Schwieters, CD. (2002) Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr Opin. Struct. Bio.* Vol 12 Issue 2. pg 146-153.
- Cremer D, Pople JA (1975) A general definition of ring puckering coordinates. *Journal of the American Chemical Society* 97: 1354-1358.
- Cruz-Neira C, Sandin D, DeFanti T (1993) Surround-screen projection-based virtual reality: The design and implementation of the CAVE. *ACM SIGGRAPH Proc.* 1993, 93: 135-142
- Cui, J., Matkovich, S.J., deSouza, N., Li, S., Roseblit, N., and Marks, A.R. (2004). Regulation of the type 1 inositol 1,4,5-trisphosphate receptor by phosphorylation at tyrosine 353. *The Journal of biological chemistry* 279, 16311-16316.
- Davis IW, Murray LW, Richardson JS, Richardson DC (2004) MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research* 32, Web Server Issue: W615-W619.
- Davis IW, Arendall WB, III, Richardson DC, Richardson JS (2006) The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure*.
- Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB III, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* 2007, 35: W375-W383
- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6: 277-293.
- DeLano WL: The PyMOL Molecular Graphics System. 2002. DeLano Scientific, Palo Alto, CA, USA
- Delforge, J., Messenguy, F., and Wiame, J.M. (1975). The regulation of arginine biosynthesis in *Saccharomyces cerevisiae*. The specificity of argR- mutations and the general control of amino-acid biosynthesis. *European journal of biochemistry / FEBS* 57, 231-239.
- Doreleijers JF, Rullmann JAC, Kaptein R (1998) Quality Assessment of NMR Structures: a Statistical Survey. *Journal of Molecular Biology* 281: 149-164.
- Dubois, E., Scherens, B., Vierendeels, F., Ho, M.M., Messenguy, F., and Shears, S.B. (2002). In *Saccharomyces cerevisiae*, the inositol polyphosphate kinase activity of Kcs1p is required for resistance to salt stress, cell wall integrity, and vacuolar morphogenesis. *The Journal of biological chemistry* 277, 23755-23763.

- Emsley P, Cowtan K (2004): Coot: model-building tools for molecular graphics. *Acta Crystallogr.* 2004, D 60: 2126-2132
- Endo-Streeter, S. (2009). Structural Studies of *Arabidopsis thaliana* Inositol Polyphosphate Multi-Kinase. Ph.D Dissertation. Duke University.
- Enggist E, Thony-Meyer L, Guntert P, Pervushin K. (2002) NMR Structure of Heme Chaperone Ccme Reveals a Novel Functional Motif. *Structure* 10:1551-1557
- Engl RA, Huber R (1991) Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Crystallographica, Section A* 47: 392-400.
- Farrow NA, Muhandiram R, Singer AU, Pascal SM, Kay CM, Gish G, Shoelson SE, Pawson T, Forman-Kay JD, Kay LE (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ¹⁵N NMR relaxation. *Biochemistry* 33: 5984-6003.
- Fischer, MW, Zeng, L, Pang, Y, Hu, W, Majumdar, A, Zuiderweg, ER (1997) Experimental characterization of models for backbone pico-second dynamics in proteins. Quantification of NMR auto- and cross correlation relaxation mechanisms involving different nuclei of the peptide plane. *J. Am. Chem. Soc.* 119 (1997), pp. 12629–12642P.
- Fisher J, Cummings J, Desai KV, Vicci L, Wilde B, Keller K, Weigle C, Bishop G, Taylor RM II, Davis CW, Boucher R, O'Brien ET, Superfine R (2005) Three-dimensional force microscope: A nanometric optical tracking and magnetic manipulation system for the biomedical sciences. *Rev. Scientific Instruments* 2005, 76: 53711-22
- Fourrier L, Benros C, de Brevern AG (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* 5: 58.
- Fridy, P.C., Otto, J.C., Dollins, D.E., and York, J.D. (2007). Cloning and characterization of two human VIP1-like inositol hexakisphosphate and diphosphoinositol pentakisphosphate kinases. *The Journal of biological chemistry* 282, 30754-30762.
- Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* 23: 566-579.
- Fujii, M., and York, J.D. (2005). A role for rat inositol polyphosphate kinases rIPK2 and rIPK1 in inositol pentakisphosphate and inositol hexakisphosphate production in rat-1 cells. *The Journal of biological chemistry* 280, 1156-1164.
- Gargaro AR, Soteriou A, Frenkiel TA, Bauer CJ, Birdsall B et al. (1998) The solution structure of the complex of *Lactobacillus casei* dihydrofolate reductase with methotrexate. *Journal of Molecular Biology* 277: 119-134.
- Gerstein M, Tsai J, Levitt M (1995) The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *Journal of Molecular Biology* 249: 955-966.
- Goddard TD, Kneller DG. SPARKY 3. University of California, San Francisco

- Gonzalez, B., Schell, M.J., Letcher, A.J., Veprintsev, D.B., Irvine, R.F., Williams, R.L. (2004) Structure of a Human Inositol 1,4,5-Trisphosphate 3-Kinase; Substrate Binding Reveals Why It is not a Phosphoinositide 3-Kinase. *Mol.Cell* 15: 689
- Grishaev A, Bax A. (2004) An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *J Am Chem Soc* 2004;126:7281-7292.
- Gueux N, Peitsch MC (1997): SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 1997, 18: 2714-2723
- Güntert P, Mumenthaler C, Wuthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA *Journal of Molecular Biology* 273: 283-298.
- Güntert P (2003) Automated NMR protein structure calculation. *Progress in NMR Spectroscopy* 43: 105-125.
- Güntert P, Mumenthaler C, Herrmann T (2002) DYANA Version 1.5: Users Manual.
- Hall, J.A. et.al. (2006). Attributing the sources of accuracy in unequal-power dyadic communication: Who is better and why? *Journal of Experimental Social Psychology*, 42, 18-27.
- Harvat EM, Redfield C, Stevens JM, Ferguson SJ. (2009) Probing the Heme-Binding Site of the Cytochrome c Maturation Protein CcmE (dagger). *Biochemistry* 48:8 1820-1828.
- Helmholtz, H von. (1871) Ubet die Zeit, welche nötig ist, damit ein Gesichtseindruck zum Bewusstsein kommy, *Berliner Monatsherichte, Juni.* 333-337
- Herrmann T, Güntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA *Journal of Molecular Biology* 319: 209-227.
- Herskovitz, E., Tannenbaum, E., Howerton, S.B., Sheth, A., Tannenbaum, A., and Williams, L.D. (2003). "Automated identification of RNA conformational motifs: Theory and application to Hm LSU 23S rRNA." *Nucleic Acids Res.* 31: 6249-6257.
- Hess B, Scheek RM. (2003) Orientation restraints in molecular dynamics simulations using time and ensemble averaging. *J Magn Reson.* 2003 Sep;164(1):19-27.
- Holmes, W., Jogl, G. (2006) Crystal structure of inositol phosphate multikinase 2 and implications for substrate specificity. *J.Biol.Chem.* 281: 38109-38116
- Hooft R (1997) A WHAT IF check report: what does it mean? Available: <http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/explain.html>. 2006.

- Hooft RW, Sander C, Vriend G (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. *Computer Applications in the Biosciences* 13: 425-430.
- Hooft RWW, Vriend G, Sander C, Abola EE (1996) Errors in Protein Structures. *Nature* 381: 272.
- Hoogstraten, C. G., Legault, P., Pardi, A. (1998). "NMR solution structure of hte lead-dependent ribozyme: evidence for dynamics in RNA catalysis." *J. Mol. Biol.* **284**(2): 337-350.
- Hu H, Hermans J, Lee AL (2005) Relating side-chain mobility in proteins to rotameric transitions: insights from molecular dynamics simulations and NMR. *J Biomol NMR* 32(2): 151-162.
- Huang X, Powers R (2001) Validity of using the radius of gyration as a restraint in NMR protein structure determination. *J Am Chem Soc* 123(16): 3834-3835.
- Huang YJ, Powers R, Montelione GT (2005a) Protein NMR Recall, Precision, & F-measure Scores (RPF Scores): Structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society* 127: 1665-1674.
- Huang YJ, Tejero R, Powers R, Montelione GT (2005b) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins: Structure, Function, and Bioinformatics*
- Huang YJ, Swapna GVT, Rajan PK, Ke H, Xia B et al. (2003) Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from *Escherichia coli*. *Journal of Molecular Biology* 327: 521-536.
- Huang, Y. J., and Montelione, G.T. (2009) DisMeta disorder prediction server. Rutgers University. <http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/>
- Humphrey W, Dalke A, Schulten K (1996): VMD – visual molecular dynamics. *J. Molec. Graphics* 1996, 14, 33-38
- Irvine, R.F. (2005). Inositide evolution - towards turtle domination? *The Journal of physiology* 566, 295-300.
- Irvine, R.F., and Schell, M.J. (2001). Back in the water: the return of the inositol phosphates. *Nature reviews* 2, 327-338.
- Jansson M, Li Y-C, Jendeberg L, Anderson S, Montelione GT, Nilsson B (1996) High level production of uniformly ¹⁵N- and ¹³C-enriched fusion proteins in *Escherichia coli*. *J Biomol NMR* 7: 131-141.
- Johnson, P. E., Donaldson, L.W. (2006). "RNA recognition by the Vts1p SAM domain." *Nat.Struct.Mol.Biol.* **13**: 177-178.

- Jones TA, Zou J-Y, Cowan SW, Kjeldgaard M (1991) Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models. *Acta Crystallographica*, Section A 47: 110-119.
- Jayaraman, T., and Marks, A.R. (1997). T cells deficient in inositol 1,4,5-trisphosphate receptor are resistant to apoptosis. *Molecular and cellular biology* 17, 3005-3012.
- Jesch, S.A., Liu, P., Zhao, X., Wells, M.T., and Henry, S.A. (2006). Multiple endoplasmic reticulum-to-nucleus signaling pathways coordinate phospholipid metabolism with gene expression by distinct mechanisms. *The Journal of biological chemistry* 281, 24070-24083.
- Jones, T.A. and Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J* 5, 819-822.
- Jones TA (1978): A graphics model building and refinement system for macromolecules. *J. Applied Crystallogr.* 1978, 11: 268-272
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577-2637.
- Kaluarachchi, K., Meadows, R.P., Gorenstein, D.G. (1991). "How accurately can oligonucleotide structures be determined from the hybrid relaxation rate matrix/NOESY distance restrained molecular dynamics approach?" *Biochemistry* 30(36): 8785-8797.
- Katz L, Levinthal C (1972): Interactive computer graphics and representation of complex biological structures. *Ann. Rev. Biophys. Bioengin.* 1972, 1: 465-504
- Kaustov L, Lukin J, Lemak A, Duan S, Ho M, Doherty R, Penn LZ, Arrowsmith CH. (2007) The conserved CPH domains of Cul7 and PARC are protein-protein interaction modules that bind the tetramerization domain of p53. *J Biol Chem.* Apr 13;282(15):11300-7. Epub 2007 Feb 12.
- Kieffer, B, Atkinson, A (2004) The role of protein motions in molecular recognition: insights from heteronuclear NMR relaxation measurements *Prog. NMR Spectrosc.* 44 (2004) 141.
- Kilgard MJ (1996): The OpenGL Utility Toolkit (GLUT) Programming Interface: API Version 3, Silicon Graphics Incorporated, 1996.
- King, G. (1962) Spectroscopy and Molecular Structure. Holt Inc. New York.
- Kleywegt GJ (1999) Experimental assessment of differences between related protein crystal structures. *Acta Cryst D* 55: 1878-1884.
- Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A et al. (2004) The Uppsala Electron-Density Server. *Acta Cryst D* 60: 2240-2249.

- Kohler, J (1947) *Gestalt Psychology: An introduction to new concepts in modern psychology*. New York: Liveright Publications.
- Koradi R, Billeter M, Wüthrich K (1996): MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* 1996, 14: 51-55.
- Kosen, PA. (1989) Spin labeling of proteins. *Methods Enzymol* 177, pp. 86–121.
- Kosower, E. (1962) *Molecular Biochemistry*. McGraw Hill, New York.
- Kranz RG, Richard-Fogal C, Taylor JS, Frawley ER (2009) Cytochrome c biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control. *Microbiol. Mol. Biol. Rev.* 2009 Sep; 73(3): 510-528.
- Kuszewski J, Gronenborn AM, Clore GM (1997) Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *Journal of Magnetic Resonance* 125: 171-177.
- Kuszewski J, Gronenborn AM, Clore GM (1999). Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. *J Am Chem Soc* 1999;121:2337-2338.
- Kuszewski J, Qin J, Gronenborn AM, Clore GM. (1995) The impact of direct refinement against $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ chemical shifts on protein structure determination by NMR. *J Magn Reson Ser B* 1995;106:92-96.
- Kuszewski J, Clore GM. (2000) Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. *J Magn Reson* 2000;146:249-254.
- Labesse G, Colloc'h N, Pothier J, Moron JP (1997) P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Computer Applications in the Biosciences* 13: 291-295.
- Laskowski R, MacArthur M, Rullmann T (1996a) AQUA and PROCHECK-NMR Operating manual.
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) ProCheck - A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26: 283-291.
- Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM (1996b) AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR* 8: 477-486.
- Lee, B. M., Xu, J., Clarkson, B.K., Martinez-Yamout, M.A., Dyson, J.H., Case, D.A., Gottesfeld, J.M. (2006). "Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA." *J.Mol.Biol.* 357: 275-291.

- Legler PM, Cai M, Peterkofsky A, Clore GM. (2004) Three-dimensional solution structure of the cytoplasmic B domain of the mannitol transporter IImannitol of the escherichia coli phosphotransferase system. *J Biol Chem* 2004;279:39115-39121.
- Leinin, SF, Bremi, T, Brutscher, B, Bruschweiler, R, Ernst, RR (1998) Anisotropic Intramolecular Backbone Dynamics of Ubiquitin Characterized by NMR Relaxation and MD Computer Simulation *J. Am. Chem. Soc.*, 1998, 120 (38), pp 9870–9879
- Leontis, N. B., Altman, R.B., Berman, H.M., Brenner, S.E., Brown, J.W., Engelke, D.R., Harvey, S.C., Holbrook, S.R., Jossinet, F., Lewis, S.E., et al. (2006). "The RNA Ontology Consortium: An open invitation to the RNA community." *RNA* 12: 533-541.
- Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433: 128-132.
- Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. *Proteins* 50: 496-506.
- Liu Z-J, Tempel W, Ng JD, Lin D, Shah AK et al. (2005) The high-throughput protein-to-structure pipeline at SECSG. *Acta Cryst D*61: 679-684.
- Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology* 285: 2177-2198.
- Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson.* 1999 Jun;138(2):334-42.
- Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. *Proteins: Structure, Function, and Genetics* 40: 389-408.
- Lovell SC, Davis IW, Arendall WB, III, de Bakker PIW, Word JM et al. (2003) Structure Validation by C α Geometry: ϕ, ψ and C β Deviation. *Proteins: Structure, Function and Genetics* 50: 437-450.
- Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**: 83-85.
- Majerus, P.W., Kisseleva, M.V., and Norris, F.A. (1999). The role of phosphatases in inositol signaling reactions. *The Journal of biological chemistry* 274, 10669-10672.
- Manusov, V., Patterson, M. (2006) *The SAGE Handbook of Nonverbal Communication*. SAGE Publications, Thousand Oaks, CA. pp 287-292.
- McRee, D.E. (1999) XtalView/Xfit - A Versatile Program for Manipulating Atomic Coordinates and Electron Density. *Journal Structural Biology*, vol. 125, pp. 156-165.

- Messenguy, F. (1976). Regulation of arginine biosynthesis in *Saccharomyces cerevisiae*: isolation of a cis-dominant, constitutive mutant for ornithine carbamoyltransferase synthesis. *Journal of bacteriology* 128, 49-55.
- Microsoft, (2009). Xbox – www.microsoft.com/xbox
- Montelione GT, Wuthrich K, Burgess AW, Nice EC, Wagner G et al. (1992) Solution structure of murine epidermal growth factor determined by NMR spectroscopy and refined by energy minimization with restraints [erratum appears in *Biochemistry* 1992 Oct 20;31(41):10138]. *Biochemistry* 31: 236-249
- Moritz E, Meyer J (2004): Interactive protein structure visualization using virtual reality. 4th Proc. IEEE Symp. Bioin. Bioeng. 2004.
- Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Genetics* 12: 345-364.
- Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods Enzymol* 339: 91-108.
- Moseley HNB, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J Biomol NMR* 28: 341-355.
- Mowbray SL, Helgstrand C, Sigrell JA, Cameron AD, Jones TA (1999) Errors and reproducibility in electron-density map interpretation. *Acta Cryst D* 55: 1309-1319.
- Murray, L. W., Arendall III, W.B., Richardson, D.C., and Richardson, J.S. (2003). "RNA backbone is rotameric." *Proc. Natl. Acad. Sci.* 100: 13904-13909.
- Murray LW, Richardson JS, Arendall WB III, Richardson DC (2005): RNA Backbone Rotamers - Finding your way in 7 dimensions. *Biochemical Society Transactions* (UK) 2005, 33: 485-487
- Murzin A, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247: 536-540.
- Nabuurs SB, Spronk CA, Krieger E, Maassen H, Vriend G et al. (2003) Quantitative evaluation of experimental NMR restraints. *J Am Chem Soc* 125(39): 12026-12034.
- Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AM et al. (2004) DRESS: a database of Refined solution NMR structures. *Proteins* 55: 483-486.

- Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CAEM et al. (2005) RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*
- Neri D, Szyperski T, Otting G, Senn H, Wüthrich K (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28**: 7510-7516.
- Neurath, H. (1964) *The Chemistry of Proteins, in Concepts in Biochemistry: Ten Essays Prepared for Science Writers at The Sixth International Congress of Biochemistry. July 26th to August 1st 1964, New York City. Publisher Unknown*
- Nilges M (1995) Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities *Journal of Molecular Biology* 245: 645-660.
- Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta- spectrin. *Journal of Molecular Biology* 269: 408-422.
- Nintendo, (2009). *The Nintendo Wii.*
- Odom, A.R., Stahlberg, A., Wenthe, S.R., and York, J.D. (2000). A role for nuclear inositol 1,4,5-trisphosphate kinase in transcriptional control. *Science* 287, 2026-2029.
- Pascal SM, Muhandiram DR, Yamazaki T, Forman-Kay JD, Kay LE (1994) Simultaneous acquisition of ^{15}N - and ^{13}C -edited NOE spectra of proteins dissolved in H_2O . *J Magn Reson Ser B* **103**: 197-201.
- Pavlov, I. P. (1927) *Conditioned reflexes.* London: Oxford University Press
- PDB and CallIT2: PDB in a CAVE: Virtual reality environment highlights PDB structures. *PDB Newsletter* spring 2006, 29: 1
- Pearlman, DA, Case, DA, Caldwell, JW, Ross, WR, Cheatham, TE, Debolt, S, Ferguson, D, Seibel, G, Kollman, P. , (1995) *Comput. Phys. Commun.* 91
- Perutz, MF, Rossmann, MG, Cullis, AF, Muirhead, H, Will, G, North AC, (1960) Structure of Haemoglobin: a three-dimensional Fourier synthesis at 5-5.5Å resolution, obtained by x-ray analysis. *Nature* 13;185(4711);416-22
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 2004, 25: 1605-1612
- Pique, M (1989) Exploring Virtual Worlds with Head-Mounted Displays, *Non-Holographic True 3-D Dimensional Display Technologies, SPIE Proceedings*, 1083, Los Angeles, CA, January 15-20.

- Polshakov VI, Smirnov EG, Birdsall B, Kelly G, Feeney J (2002) NMR-based solution structure of the complex of *Lactobacillus casei* dihydrofolate reductase with trimethoprim and NADPH. *Journal of Biomolecular NMR* 24: 67-70.
- Pontius J, Richelle J, Wodak SJ (1996) Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology* 264(1): 121-136.
- Popenda, M., Milecki, J., and Adamiak, R.W. (2004). "High salt solution structure of a left-handed RNA double helix." *Nucleic Acids Res.* 32: 4044-4054.
- Porter, TK (1978) Spherical shading. *Computer Graphics* 1978, 12: 282-285
- Ramos, A., Varani, G. (1997). "Structure of the acceptor stem of *Escherichia coli* tRNA Ala: role of the G3.U70 base pair in synthetase recognition." *Nucleic Acids Res.* 25(11): 2083-2090.
- Remland M. (1981). Developing leadership skills in nonverbal communication: A situational perspective. *Journal of Business Communications.* 18, 18-31.
- Richards FM (1974) The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *Journal of Molecular Biology* 82: 1-14.
- Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Bioinformatics* 3: 71-84.
- Richardson JS (1981) The Anatomy and Taxonomy of Protein Structure. *Adv. Prot. Chem.* 1981, 34: 167-339
- Richardson DC, Richardson JS (1992) The Kinemage: A Tool for Scientific Illustration. *Protein Science* 1: 3-9.
- Richardson DC, Richardson JS (1994) Kinemages - Simple Macromolecular Graphics for Interactive Teaching and Publication. *Trends in Biochemical Sciences* 19: 135-138.
- Richardson JS (2003) All-atom contacts: A new approach to structure validation. In: Bourne PE, Weissig H, editors. *Methods of Biochemical Analysis: Structural Bioinformatics*. New York: John Wiley & Sons, Inc. pp. 305-320.
- Richardson JS, Richardson DC (2001) "MAGE, PROBE, and Kinemages", chapter 25.2.8. In: Rossmann MG, Arnold E, editors. *International Tables for Crystallography*. Dordrecht: Kluwer Academic Publishers, The Netherlands. pp. 727-730.
- Richardson JS, Arendall WB, III, Richardson DC (2003) New Tools and Data for Improving Structures, Using All-atom Contacts. In: Carter CW, Jr., Sweet RM, editors. *Methods in Enzymology: Macromolecular Crystallography*, Pt D. New York: Academic Press. pp. 385-412.

- Richardson JS, Richardson DC, Tweedy NB, Gernert KM, Quinn TP et al. (1992) Looking at proteins: representations, folding, packing, and design. *Biophysical Journal* 63: 1186-1209.
- Richardson, J. S., Schneider, B., Murray, L.W., Kapral, G.J., Immormino, R.M., Headd, J.J., Richardson, D.C., Ham, D., HersHKovits, E., Williams, L.D., Keating, K.S., Pyle, A.M., Micallef, D., Westbrook, J., and Berman, H.M. (2008). "RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)." *RNA* 14: 465-481.
- Robinson MA, Park S, Sun ZY, Silver PA, Wagner G et al. (2005) Multiple conformations in the ligand-binding site of the yeast nuclear pore-targeting domain of Nup116p. *J Biol Chem* 280(42): 35723-35732.
- Rubin BH, Richardson JS (1972) The simple construction of protein α -carbon models. *Biopolymers* 1972, 11: 2381-2385
- Rudisser, S., Tinoco, Jr., I. (2000). "Solution Structure of a 22-Nucleotide Hairpin Similar to the P5ABC Region of a Group I Ribozyme with Cobalt(III)hexammine Complexed to the GAAA Tetraloop." *J. Mol. Biol.* **295**: 1211-1223.
- Rückert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J Am Chem Soc* **122**: 7793-7797.
- Rudisser, S., Tinoco, Jr., I. (2000). "Solution Structure of a 22-Nucleotide Hairpin Similar to the P5ABC Region of a Group I Ribozyme with Cobalt(III)hexammine Complexed to the GAAA Tetraloop." *J. Mol. Biol.* **295**: 1211-1223.
- Saiardi, A., Caffrey, J.J., Snyder, S.H., and Shears, S.B. (2000). Inositol polyphosphate multikinase (ArgR^{III}) determines nuclear mRNA export in *Saccharomyces cerevisiae*. *FEBS letters* 468, 28-32.
- Sayle R, Milner-White EJ (1995) RasMol: Biomolecular graphics for all. *Trends in Biochem. Sci.* 1995, 20: 374
- Schaeffer B, Goudeseune C: Syzygy (2003) Native PC cluster VR. *Technical report from the Integrated Systems Laboratory, Beckman Institute, U IL Urbana-Champaign, 2003, <http://www.isl.uiuc.edu/szg/>.*
- Scheek, RM, Torda, A, Kemmink, J, van Gunsteren, WF (1991) in: J. Hoch (Ed.), *Computational Aspects of the Study of Biological Macromolecules by NMR*, Plenum Press, New York.
- Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gleuhmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F., et al. (2000). "Structure of functionally activated small ribosomal subunit at 3.3 Å resolution." *Cell* 102: 615-623.
- Schmitz, U., and James, T.L. (1995). "How to generate accurate solution structures of double-helical nucleic acid fragments using nuclear magnetic resonance and restrained molecular dynamics." *Methods Enzymol.* 261: 3-44.

- Schneider, B., Moravek, Z., and Berman, H.M. (2004). "RNA conformational classes." *Nucleic Acids Res.* 32: 1666-1677.
- Schwalbe, M., Ohlenschlager, O., Marchanka, A., Ramachandran, R., Hafner, S., Heise, T., Grolach, M. (2008). "Solution structure of stem-loop alpha of the hepatitis B virus post-transcriptional regulatory element." *Nucleic Acids Res.* 36: 1681-1689.
- Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* 160: 65-73.
- Schwieters CD, Clore GM. (2001) Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *Journal of Magnetic Resonance* 2001;152:288-302.
- Schulz, I., and Krause, E. (2004). Inositol 1,4,5-trisphosphate and its co-players in the concert of Ca²⁺ signalling--new faces in the line up. *Current molecular medicine* 4, 313-322.
- Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. *J Biomol NMR* 1(3): 217-236.
- Seeds, A.M., Bastidas, R.J., and York, J.D. (2005). Molecular definition of a novel inositol polyphosphate metabolic pathway initiated by inositol 1,4,5-trisphosphate 3-kinase activity in *Saccharomyces cerevisiae*. *The Journal of biological chemistry* 280, 27654-27661.
- Shears, S.B. (2004). How versatile are inositol phosphate kinases? *The Biochemical journal* 377, 265-280.
- Sherman WR, Craig AB (2002): *Understanding Virtual Reality: Interface, Application, and Design*. 2002. Morgan Kauffmann, San Francisco
- Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* 17: 355-362.
- Sklenar H, Etchebest C, Lavery R (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins: Structure, Function, and Bioinformatics* 6: 46-60.
- Snyder DA, Bhattacharya A, Huang YJ, Montelione GT (2005) Assessing precision and accuracy of protein structures derived from NMR data. *Proteins: Structure, Function, and Bioinformatics* 59: 655-661.
- Someya T, Nameki N, Hosoi H, Suzuki S, Hatanaka H et al. (2003) Solution structure of a tmRNA-binding protein, SmpB, from *Thermus thermophilus*. *FEBS Letters* 535: 94-100.
- Sony, (2009). Playstation www.sony.com/playstation/

- Spielberg, S. (2002) *Minority Report*. Twentieth Century Fox. USA
- Spronk CA, Nabuurs SB, Krieger E, Vriend G, Vuister GW (2004) Validation of protein structures derived by NMR spectroscopy *Progress in Nuclear Magnetic Resonance Spectroscopy* 45: 315-337.
- Stevenson-Paulik, J., Odom, A.R., and York, J.D. (2002). Molecular and biochemical characterization of two plant inositol polyphosphate 6-/3-/5-kinases. *The Journal of biological chemistry* 277, 42711-42718.
- Surles, M.C., J.S. Richardson, D.C. Richardson, F.P. Brooks, Jr. (1994) Sculpting Proteins Interactively: Continual Energy Minimization Embedded in a Graphical Modeling System," *Protein Science*, 3: (February 1994), 198-210.
- Tarini, M., Cignoni, P., Montani, C. (2006). Ambient Occlusion and Edge Cueing to Enhance Real Time Molecular Visualization. *IEEE Trans. Vis & Comp Graphics*. Vol 12, No. 5.
- Taylor RM II, Robinett W, Chi VL, Brooks FP Jr, Wright WV, Williams RS, Snyder EJ (1993) The Nanomanipulator: A Virtual-Reality Interface for a Scanning Tunneling Microscope. *ACM SIGGRAPH Proc.* 1993, 93: 127-34.
- Taylor RM II, Hudson TC, Seeger A, Weber H, Juliano J, Helser AT (2001) VRPN: A Device-Independent, Network-Transparent VR Peripheral System. *Proc. ACM Symp. on Virtual Reality Software & Technology 2001, VRST 2001 Banff Centre, Canada.*
- Tashiro M, Tejero R, Zimmerman DE, Celda B, Nilsson B et al. (1997) High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *Journal of Molecular Biology* 272: 573-590.
- Theimer, C. A., Blois, C.A., Feigon, J. (2005). "Solution structure of the P2b-P3 pseudoknot from human telomerase RNA." *Mol.Cell* 17: 671-682.
- Tisi, R., Belotti, F., Wera, S., Winderickx, J., Thevelein, J.M., and Martegani, E. (2004). Evidence for inositol triphosphate as a second messenger for glucose-induced calcium signalling in budding yeast. *Current genetics* 45, 83-89.
- Torda AE, Scheek RM, van Gunsteren WF. (1990) Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J Mol Biol.* 1990 Jul 5;214(1):223-35.
- Tsai J, Gerstein M (2002) Calculations of protein volumes: sensitivity analysis and parameter database. *Bioinformatics* 18: 985-995.
- Tsai J, Taylor R, Chothia C, Gerstein M (1999) The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology* 290: 253-266.
- Tsui V, Zhu L, Huang TH, Wright PE, Case DA. (2000) Assessment of zinc finger orientations by residual dipolar coupling constants *J Biomol NMR.* 2000 Jan;16(1):9-21.

- Ulmer TS, Ramirez BE, Delaglio F, Bax A. (2003) Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J Am Chem Soc.* Jul 30;125(30):9179-91.
- Vaguine AA, Richelle J, Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Cryst D* 55: 191-205.
- van de Ven, F. J. M., Blommers, M.J.J., Schouten, R.E., Hilbers, C.W. (1991). "Calculation of interproton distances from NOE intensities. A relaxation matrix approach without requirement of a molecular model." *J. Magn. Reson.* 94: 140-151.
- van den Akker F, Hol WGJ (1999) Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structures. *Acta Cryst D* 55: 206-218.
- Varani, G., Aboul-ela, F., Allain, F. H.-T (1996). "NMR investigation of RNA structure." *Progress in Nuclear Magnetic Resonance Spectroscopy* 29(1-2): 51-127.
- VirTools, (2009). www.virttools.com
- Voronoi G (1907) Nouvelles applications des paramètres continus a la théorie des formes quadratiques, premier mémoire, sur quelques propriétés es formes quadratiques positives parfaites. *Journal für die Reine und Angewandte Mathematik* 133: 97-178.
- Voronoi G (1908) Nouvelles applications des paramètres continus a la théorie des formes quadratiques, deuxième mémoire, recherche sur les paralléloèdres primitifs. *Journal für die Reine und Angewandte Mathematik* 134: 198-287.
- Vriend G (1990) WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics* 8(1): 52-56.
- Vuister GW, Bax A (1993) Quantitative *J* correlation: a new approach for measuring homonuclear three-bond $J(\text{H}^{\text{N}}\text{H}^{\alpha})$ coupling constants in ^{15}N -enriched proteins. *J Am Chem Soc* 115: 7772-7777.
- Wagner, L.E., 2nd, Li, W.H., Joseph, S.K., and Yule, D.I. (2004). Functional consequences of phosphomimetic mutations at key cAMP-dependent protein kinase phosphorylation sites in the type 1 inositol 1,4,5-trisphosphate receptor. *The Journal of biological chemistry* 279, 46242-46252.
- Wagner, L.E., 2nd, Li, W.H., and Yule, D.I. (2003). Phosphorylation of type-1 inositol 1,4,5-trisphosphate receptors by cyclic nucleotide-dependent protein kinases: a mutational analysis of the functionally important sites in the S2+ and S2- splice variants. *The Journal of biological chemistry* 278, 45811-45817.
- Westhead, DR. (2003) Bioinformatics. BIOS Scientific Publishers Ltd. Oxford UK

- White, A, Handler, P, Smith, S, Stetten, D (1954) Principles of Biochemistry, 1st Ed. McGraw Hill, New York
- Wijmenga, S. S., van Buuren, Bernd N. M. (1998). "The use of NMR methods for conformational studies of nucleic acids." *Progress in NMR Spectroscopy* 32: 287-387.
- Wimberly, B. T., Broderson, D.E., Clemons, W.M., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000). "Structure of the 30S ribosomal subunit." *Nature* 407: 327-339.
- Winn, M.D., Isupov, M.N., and Murshudov, G.N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta crystallographica* 57, 122-133.
- Word JM (2000) All-Atom Small-Probe Contact Surface Analysis: an information-rich description of molecular goodness-of-fit [Ph.D. dissertation]. Durham, N.C.: Duke University. 274 p.
- Word JM, Lovell SC, Richardson JS, Richardson DC (1999b) Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *Journal of Molecular Biology* 285: 1735-1747.
- Word JM, Bateman RC, Jr., Presley BK, Lovell SC, Richardson DC (2000) Exploring Steric Constraints on Protein Mutations using Mage/Probe. *Protein Science* 9: 2251-2259.
- Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME et al. (1999a) Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogens. *Journal of Molecular Biology* 285(4): 1711-1733.
- Wright, R. (2009) Orienting of Attention. Oxford Univ. Press.
- Wuthrich, K (1986) NMR of Protein and Nucleic Acids. Wiley Interscience. New York.
- Xia, H.J., and Yang, G. (2005). Inositol 1,4,5-trisphosphate 3-kinases: functions and regulations. *Cell research* 15, 83-91.
- Xiong, L., Lee, B., Ishitani, M., Lee, H., Zhang, C., and Zhu, J.K. (2001). FIERY1 encoding an inositol polyphosphate 1-phosphatase is a negative regulator of abscisic acid and stress signaling in Arabidopsis. *Genes & development* 15, 1971-1984.
- Yan J, Delaglio F, Kaerner A, Kline AD, Mo H, Shapiro MJ, Smitka TA, Stephenson GA, Zartler ER. (2004) Complete relative stereochemistry of multiple stereocenters using only residual dipolar couplings. *J Am Chem Soc.* 2004 Apr 21;126(15):5008-17.
- Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM et al. (2003) Structural proteomics: Toward high-throughput structural biology as a tool in functional genomics. *Accounts of Chemical Research* 36: 183-189.

- Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A et al. (2002) An NMR approach to structural proteomics. *Proc Natl Acad Sci U S A* 99(4): 1825-1830.
- York JD, Ponder JW, Chen ZW, Mathews FS, Majerus PW (1994) Crystal structure of inositol polyphosphate 1-phosphatase at 2.3-Å resolution. *Biochemistry* 33(45): 13164-13171.
- York, J.D. (2006). Regulation of nuclear processes by inositol polyphosphates. *Biochimica et biophysica acta* 1761, 552-559.
- York, J.D., Guo, S., Odom, A.R., Spiegelberg, B.D., and Stolz, L.E. (2001). An expanded view of inositol signaling. *Advances in enzyme regulation* 41, 57-71.
- York, J.D., Odom, A.R., Murphy, R., Ives, E.B., and Wenthe, S.R. (1999). A phospholipase C-dependent inositol polyphosphate kinase pathway required for efficient messenger RNA export. *Science* 285, 96-100.
- York, S.J., Armbruster, B.N., Greenwell, P., Petes, T.D., and York, J.D. (2005). Inositol diphosphate signaling regulates telomere length. *The Journal of biological chemistry* 280, 4264-4269.
- Zheng D, Aramini JM, Montelione GT (2004) Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. *Protein Science* 13: 549-554.
- Zhu, C.C., Furuichi, T., Mikoshiba, K., and Wojcikiewicz, R.J. (1999). Inositol 1,4,5-trisphosphate receptor down-regulation is activated directly by inositol 1,4,5-trisphosphate binding. Studies with binding-defective mutant receptors. *The Journal of biological chemistry* 274, 3476-3484.

Biography

I grew up near Syracuse, NY in a small town named Manlius. Growing up disabled due to a degenerative eye disorder, I recognized hard work and academic achievement as being critical to overcoming my visual limitations. The son of a policeman and a social worker, my parents stressed the importance of school, encouraged my sister and I to get jobs during our youth, supported me running track & cross country in high school and college, instilled a strong sense of faith, and nurtured a firm commitment to public service.

I strive to connect disciplines both inside and outside the classroom by blending my technical/scientific background with public policy. Ultimately, I'm most interested in developing and implementing policy solutions addressing the most thorny and complicated problems in order to improve the lives of every American while making the United States a model citizen in the global community.

I am a proud mentor of students with visual impairments in the school district where I was a student growing up in New York. In Durham, my work with the Duke-Durham Neighborhood Partnership promoting science and mathematics education for adverse circumstance youth populations is a joy of mine and I feel is one of the many important pieces necessary for the revitalization of the education system in Durham.