Supplementary Material for

The Expanded FindCore Method for Identification of a Core Atom Set for Assessment of Protein Structure Predictions

David A. Snyder[1*], Jennifer Grullon[1], Yuanpeng J. Huang[2], Roberto Tejero[2,3], Gaetano T. Montelione [2]*

---

[1]Department of Chemistry, William Paterson University, 300 Pompton Road
Wayne, New Jersey 07470, USA

[2] Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, and Northeast Structural Genomics Consortium, 679 Hoes Lane, Piscataway, New Jersey, 08854, USA

[3]Departamento de Química Física, Universidad de Valencia, Avenida Dr. Moliner 50 46100 Burjassot, Valencia, SPAIN

Structural Validation of Expanded FindCore Atom Sets

For high-quality structures, the Expanded FindCore core atom set should have good validation statistics. Two common means of assessing the validity and coherency of NMR-derived structural ensembles are an analysis of dihedral angles and an analysis of H-bond completeness. Comparing dihedral angle quality with a definition (e.g. the expanded core or the original FindCore core) of a core **atom** set requires a rule as to how to label dihedral angles as "core" or "non-core". For our purposes, the backbone (ϕ and ψ) dihedral angles for the i^{th} residue are associated with the core if all the heavy atoms involved in their calculation (N, Cα and C' from residue i as well as C' from residue i-1 and N from residue i+1) are in the core. The sidechain dihedral angles for the i^{th} residue are in the (either expanded or FindCore) core if all the (heavy) atoms in that residue are in the core.

One method to quantify the quality of dihedral angles is to calculate G scores for both the backbone (ϕ and ψ) and side-chain dihedral angles using the PROCHECK software package.[1, 2] A G score of less than -1 corresponds to dihedral angles having a less than 25% probability of occurring together in a valid structure. Supplementary Table S2 reports the percentage of residues with backbone (ϕ and ψ) dihedral angles and with side chain dihedral angles with G scores of less than -1 as grouped by whether those angles are in the core or not for both the original and Expanded FindCore definitions of the core atom set. Goodness of fit analysis of the associated contingency tables cross-classifying residues into core and non-core sets vs. high and low G-score sets indicates that (for backbone dihedral angles), non-core residues (with either definition of core) are significantly more likely to have G scores less than -1 than are core residues. However, the Expanded FindCore method, as measured by summing the goodness of fit χ^2 statistics associated with

each model's contingency table, better associates residues having poor PROCHECK G scores with the non-core set than does the original FindCore method. Similar trends were observed for both backbone and sidechain dihedrals, but when considering all dihedral angles in the PROCHECK analysis, the G scores for residues with all (heavy) atoms in the FindCore core are not significantly more likely to be less than -1 than are G scores for residues with all (heavy) atoms in the expanded core and at least one atom *not* in the original FindCore core.

We also analyzed the consistency of hydrogen bonded interactions across the ensemble of conformations (Supplementary Table S3), calculated using the PDBStat software package.[3] Hydrogen bonds occur more consistently for pairs of core atoms for both the FindCore and Expanded FindCore definitions of core. However, the FindCore definition best distinguishes core atom pairs from atom pairs including non-core atoms in terms of hydrogen bond consistency. This is not surprising as FindCore defines core atoms on the basis of an order parameter that sums over pairs of well-defined atoms, and hydrogen bonding constrains pairs of atoms. This suggests that the decision of whether to validate all core atoms as defined by the Expanded FindCore or to restrict such validations to the original FindCore atom set may depend on the nature of the validation under consideration and that the Expanded FindCore method complements rather than replaces the original FindCore definition.

Supplementary Table S1.

GDT-TS scores for predictions assessed against 18 experimental NMR structures from each of 5 top ranked groups, using in each case three representations of the NMR structure: (i) the untrimmed (Untr.) coordinates, (ii) coordinates trimmed using the original FindCore method (FC), and (iii) coordinates trimmed using the expanded FindCore method (ExpFC). In each case we report the mean value and standard deviation of the GDT-TS score computed for each of the multiple structures in the NMR ensemble.

	HHpred-thread			PMS			QUARK			RaptorX-ZY			Zhang-server		
Target	**Untr.**	**FC**	**ExpFC**	**Untr.**	**FC**	**ExpFC**	**Untr.**	**FC**	**ExpFC**	**Untr.**	**FC**	**ExpFC**	**Untr.**	**FC**	**ExpFC**
T0655	0.6255 0.0095	0.8007 0.0146	0.7103 0.0111	0.6527 0.0117	0.8413 0.0118	0.7419 0.0129	0.6475 0.0155	0.8000 0.0181	0.7161 0.0176	0.5442 0.0134	0.7133 0.0098	0.6256 0.0149	0.6546 0.0143	0.8075 0.0169	0.7262 0.0169
T0657	0.7658 0.0132	0.9844 0.0065	0.8476 0.0127	0.7518 0.0133	0.9681 0.0096	0.8276 0.0118	0.7860 0.0110	0.9904 0.0059	0.8637 0.0112	0.7661 0.0172	0.9848 0.0069	0.8461 0.0126	0.7809 0.0098	0.9883 0.0063	0.8604 0.0088
T0662	0.7113 0.0194	0.8106 0.0217	0.7533 0.0184	0.7247 0.0230	0.8112 0.0213	0.7645 0.0224	0.7590 0.0213	0.8359 0.0239	0.8013 0.0214	0.7304 0.0223	0.8201 0.0228	0.7712 0.0220	0.7648 0.0210	0.8462 0.0219	0.8074 0.0203
T0665	0.7453 0.0268	0.9570 0.0171	0.9070 0.0236	0.7343 0.0201	0.9657 0.0161	0.9041 0.0233	0.7666 0.0252	0.9772 0.0175	0.9368 0.0253	0.7457 0.0257	0.9679 0.0177	0.9123 0.0223	0.7750 0.0261	0.9846 0.0143	0.9476 0.0220
T0668	0.2247 0.0051	0.3263 0.0057	0.2434 0.0048	0.2657 0.0063	0.3747 0.0099	0.2931 0.0079	0.3430 0.0085	0.5428 0.0088	0.3763 0.0093	0.2391 0.0056	0.3447 0.0071	0.2646 0.0064	0.3365 0.0101	0.4952 0.0055	0.3707 0.0102
T0669	0.5226 0.0089	0.7062 0.0122	0.5827 0.0107	0.5794 0.0124	0.7836 0.0203	0.6401 0.0136	0.5583 0.0093	0.7173 0.0114	0.6198 0.0117	0.5151 0.0058	0.6872 0.0114	0.5732 0.0074	0.5743 0.0079	0.7292 0.0118	0.6370 0.0088
T0675	0.3575 0.0166	0.4354 0.0078	0.4138 0.0208	0.3882 0.0133	0.5156 0.0099	0.4686 0.0318	0.3735 0.0078	0.5148 0.0118	0.4658 0.0330	0.3618 0.0069	0.5047 0.0079	0.4579 0.0274	0.3765 0.0076	0.5159 0.0126	0.4726 0.0330
T0677_A	0.7563 0.0206	0.9532 0.0146	0.8997 0.0176	0.7759 0.0235	0.9439 0.0120	0.9009 0.0129	0.7129 0.0180	0.9264 0.0153	0.8656 0.0173	0.7513 0.0189	0.9518 0.0139	0.8906 0.0150	0.6979 0.0184	0.8921 0.0167	0.8475 0.0175
T0677_B	0.3322 0.0061	0.4334 0.0080	0.3911 0.0086	0.3299 0.0077	0.4436 0.078	0.3931 0.0098	0.3325 0.0056	0.4480 0.0077	0.4004 0.0072	0.3150 0.0082	0.4282 0.0064	0.3757 0.0076	0.3135 0.0060	0.4259 0.0064	0.3738 0.0071
T0709	0.7000 0.0256	0.8619 0.0179	0.8489 0.0259	0.6962 0.0245	0.8368 0.0206	0.8341 0.0239	0.7326 0.0267	0.8737 0.0222	0.8545 0.0202	0.7144 0.0217	0.8566 0.0224	0.8534 0.0179	0.7280 0.0259	0.8694 0.0220	0.8511 0.0200

T0711	0.7318 0.0192	0.9444 0.0088	0.8095 0.0141	0.6992 0.0122	0.9430 0.0097	0.773 0.0110	0.7364 0.0143	0.9375 0.0155	0.7922 0.0125	0.7803 0.0192	0.9514 0.0093	0.8259 0.0138	0.7205 0.0109	0.9306 0.0152	0.7793 0.0123
T0714	0.8563 0.0220	0.9434 0.0201	0.8563 0.0220	0.8112 0.0256	0.8965 0.0173	0.8112 0.0256	0.8804 0.0224	0.9495 0.0206	0.8804 0.0224	0.8584 0.0150	0.9232 0.0168	0.8584 0.0150	0.8993 0.0206	0.9554 0.0205	0.8993 0.0206
T0716	0.7153 0.0330	0.9538 0.0211	0.9191 0.0366	0.7282 0.0326	0.9638 0.0137	0.9361 0.0335	0.7127 0.0366	0.9532 0.0190	0.9160 0.0341	0.5072 0.0210	0.6702 0.0119	0.6745 0.0254	0.7014 0.0353	0.9426 0.0235	0.9012 0.0365
T0727	0.5835 0.0071	0.7080 0.0118	0.5930 0.0079	0.6413 0.0091	0.7479 0.0096	0.6532 0.0099	0.9854 0.0056	1.0000 0.0000	0.9900 0.0047	0.5723 0.0099	0.7077 0.0085	0.5815 0.0095	0.9859 0.0066	1.0000 0.0000	0.9908 0.0052
T0729	0.8081 0.0151	0.9201 0.0161	0.8290 0.0150	0.8182 0.0114	0.9332 0.0156	0.8376 0.0124	0.9354 0.0237	0.9936 0.0076	0.9441 0.0195	0.8111 0.0158	0.9198 0.0184	0.8308 0.0166	0.9389 0.0248	0.9951 0.0066	0.9490 0.0200
T0731	0.6218 0.0187	0.8777 0.0249	0.8402 0.0277	0.5943 0.0164	0.8625 0.0187	0.8209 0.0198	0.6293 0.0263	0.8888 0.0285	0.8523 0.0315	0.6136 0.0211	0.8691 0.0275	0.8291 0.0256	0.6279 0.0275	0.8843 0.0292	0.8493 0.0313
T0751	0.3831 0.0093	0.9006 0.0109	0.9011 0.0118	0.3759 0.0094	0.8883 0.0120	0.8880 0.0129	0.3746 0.0075	0.8886 0.0110	0.8910 0.0095	0.3872 0.0128	0.9048 0.0102	0.9060 0.0111	0.3923 0.0112	0.9111 0.0128	0.9133 0.0122
T0754	0.5291 0.0104	0.7167 0.0144	0.6300 0.0136	0.5342 0.0103	0.7175 0.0115	0.6299 0.0126	0.5305 0.0099	0.7058 0.0109	0.6289 0.0125	0.5224 0.0091	0.7097 0.0099	0.6166 0.0117	0.6727 0.0156	0.8519 0.0231	0.7703 0.0162

Supplementary Table S2: PROCHECK Validation of Core vs. Non-Core Residues for both Original and Expanded FindCore

		Low G-Score (Backbone) Residues (%)	χ^2	Low G-Score (All) Residues (%)	χ^2
FindCore					
	Core	15.7%		11.1%	
	Non-Core	60.7%	450.2	23.9%	50.4
Expanded FindCore					
	Core	22.0%		15.4%	
	Non-Core	77.9%	482.6	26.8%	67.3
	EFC Only	40.3%	86.8	19.7%	18.4*

Top row of χ^2 statistics compares core and non-core residues under the original FindCore definition. Next row of χ^2 statistics compares core and non-core residues under the Expanded FindCore definition. Last row compares residues only in the core under the Expanded FindCore definition to those in the core under the original definition. The Supplementary Materials text under the heading "Structural Validation of Expanded FindCore Atom Sets" describes the standard of -1 used for low PROCHECK [1, 2] G-scores. *This value is **not** significant ($p = 0.24$). All other χ^2 statistics in this table have $p < 0.01$

Supplementary Table S3: Consistency of Hydrogen Bonding Across Models for Core vs. Non-Core Atoms for both Original and Expanded FindCore

	Hydrogen Bonding Consistency (%)	χ^2
FindCore		
Core	56.7%	
Core/Non-Core	15.8%	
Non-Core	10.0%	14498.7
Expanded FindCore		
Core	41.3%	
Core/Non-Core	9.5%	
Non-Core	8.8%	10119.9

Top row of χ^2 statistics compares core atom pairs with atom pairs containing a non-core atom under the original FindCore definition. Next row of χ^2 statistics compares core atom pairs with atom pairs containing a non-core atom under the Expanded FindCore definition.

Supplementary References

1. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283-291.

2. Laskowski R, Rullmann JA, MacArthur M, Kaptein R, Thornton J. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477-486.

3. Tejero R, Snyder D, Ma B, Aramini JM, Montelione GT. PDBStat: A Universal Restraint Converter and Restraint Analysis Software Package for Protein NMR. J Biomol NMR 2013;56:337-351.