A Pseudopotential for Improving the Packing of Ellipsoidal Protein Structures Determined from NMR Data[†]

Charles D. Schwieters*,[‡] and G. Marius Clore*,[§]

Imaging Sciences Laboratory, Center for Information Technology, National Institutes of Health, Building 12A, Bethesda, Maryland 20892-5624, and Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, Bethesda, Maryland 20892-0510

Received: August 3, 2007; In Final Form: October 16, 2007

A pseudopotential has been introduced into NMR protein structure determination that effectively restrains molecular volume based on the observation that rigid, well-packed protein structures are approximately ellipsoidal. Allowing an ellipsoidal shape is more general than the single approximately spherical shape imposed by the radius of gyration pseudopotential introduced previously. We demonstrate that this new gyration volume term improves structures both during the calculation of an initial unknown fold and during final refinement.

Introduction

It is well-known that biomolecular structures determined using NMR distance restraints are poorly packed in comparison with those determined by X-ray diffraction.[1,2] X-ray diffraction is a direct imaging technique, while NMR relies on approximate distance restraints for translational atomic position information, and thus the resulting packing quality of NMR structures is quite dependent on the description of nonbonded interactions. Two general approaches to improving packing in NMR structure determination have involved the use of database packing pseudopotentials[3,4] and the practice of performing an additional refinement step in explicit solvent.[5] We favor the use of pseudopotential terms because they are less computationally demanding so that they can be enabled at all stages of structure determination. In such a context we have found that they can actually aid in convergence to the crystal structure. Here we present a pseudopotential that is an improvement on the radius of gyration restraint introduced previously.[3] The previous pseudopotential required the identification of approximately globular regions in the structure, while the new term allows much more general ellipsoidal shapes.

Theory

A simple, approximate relationship has been noted between radius of gyration (R_g) and the number of residues in well-packed single-domain proteins:

$$R_g \approx A N_{res}^{\ b} \tag{1}$$

with $A = 2.2$ and $b = 0.38$.[6] The basis for the relationship is the fact that proteins pack to a constant density of 1.43 ± 0.03 g cm^{-3},[7] implying that protein volume V is directly proportional to N_{res}, the number of constituent residues, assuming a uniform mix of residue types. If proteins were perfectly spherical, b would be 1/3. The observed deviation is due to the fact that larger proteins are less likely to be perfect spheres. The fact that the observed radius of gyration should be larger than that of a sphere informs us that proteins are more often closer to prolate spheroids (pencil-shaped) than oblate spheroids (saucer-shaped). The relationship in eq 1 has been exploited to improve the packing of NMR structures by employing a pseudopotential to bias structures toward satisfying this equation.[3]

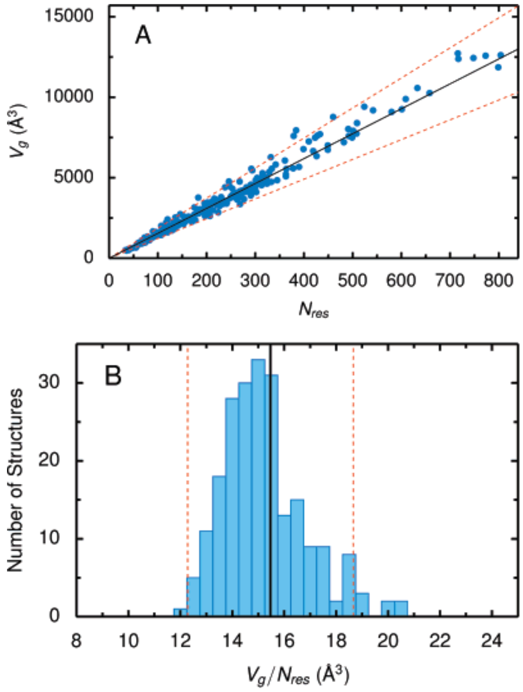

Figure 1. Gyration volume of proteins. (A) Gyration volume (V_g) as a function of number of residues for 220 well-packed crystal protein structures solved at a resolution of ≤1.2 Å with intimately interacting domains and/or subunits. The corresponding Pearson correlation coefficient is 0.989. (B) Distribution of V_g/N_{res} for these same structures. The mean value of the distribution is 15.47 ± 1.59 Å^3, and is denoted by the vertical black line. The dashed red lines in each plot denote the 5% and 95% confidence limits and are the bounds of the stiff quadratic potential in eq 5.

[†] Part of the "Attila Szabo Festschrift".

[*] Corresponding authors. E-mail: Charles.Schwieters@nih.gov (C.D.S.); mariusc@intra.niddk.nih.gov (G.M.C.).

[‡] Center for Information Technology.

[§] National Institute of Diabetes and Digestive and Kidney Diseases.

10.1021/jp076244o This article not subject to U.S. Copyright. Published 2008 by the American Chemical Society
Published on Web 12/19/2007

TABLE 1: Effect of V_g and R_g Restraints on Refinement Accuracy

	accuracy to X-ray structure (Å)		
protein[a]	no term	V_g	R_g
BAF	1.18 ± 0.12	1.13 ± 0.10	0.95 ± 0.09
p53	0.53 ± 0.06	0.52 ± 0.06	0.51 ± 0.06
e-gp41	1.82 ± 0.13	1.24 ± 0.07	3.26 ± 0.11
EIN	1.75 ± 0.22	1.93 ± 0.19	4.12 ± 0.57

[a] The number of experimental NMR restraints used in calculating the structures is as follows: BAF[9] (89 residues per subunit), 1655 restraints per monomer comprising 864 distance restraints including 48 intermolecular restraints, 257 torsion angle restraints, 66 ^{3}J coupling restraints, 165 13 C_α/β secondary chemical shift restraints and 259 dipolar coupling restraints; tetramerization domain of p53[11] (42 residues per subunit; residues 319−360), 1118 restraints per monomer comprising 938 distance restraints including 309 intersubunit restraints, 71 torsion angle restraints, 36 ^{3} J coupling restraints and 73 ^{13} C_α/β chemical shift restraints; e-gp41[13] (123 residues per subunit; residues 27−149), 2160 restraints per monomer comprising 1500 distance restraints, including 232 intermolecular restraints, 360 torsion angle restraints, 35 ^{3}J coupling restraints, 26 $^3DC_\alpha$ (ND) isotope shift restraints, and 239 $^{13}C_\alpha/\beta$ chemical shift restraints; EIN[15] (259 residues), 4089 restraints comprising 3120 distance restraints, 549 torsion angle restraints, 163 ^{3}J coupling restraints, and 257 $^{13}C_\alpha/\beta$ secondary chemical shift restraints. The conformational torsion angle database potential[20] was employed in all calculations. The target values of R_g given by $2.2N_{res}^{0.38}$ are as follows: BAF, 15.63 Å for residues 3−89 of both subunits taken together; p53, 13.57 Å for residues 325−355 of the four subunits taken together; e-gp41, 20.60 Å for residues 29−148 for the three subunits taken together; EIN, 17.93 Å for residues 1−250. The precision of the determined backbone coordinates for no term, V_g and R_g refinement, respectively, is as follows: BAF dimer: 0.35, 0.36, 0.34 Å; p53 tetramer (residues 326−354): 0.22, 0.26, 0.29 Å; e-gp41 trimer: 0.70, 0.56, 2.82 Å; EIN: 0.96, 0.93, 1.56 Å.

In fact, the relationship in eq 1 is known to break down for elongated structures, such that it is necessary to define multiple overlapping globular regions when using the R_g pseudopotential on such structures.[3] Defining such regions is possible only after a “rough” structure is determined, thus limiting the applicability of that approach when determining *de novo* structures.

If, instead of a sphere, we make the base approximation that a protein can be described as an ellipsoid, it is convenient to describe the shape using the gyration tensor:

$$G = \frac{1}{N}\sum_{i=1}^{N} \Delta q_i \otimes \Delta q_i \qquad (2)$$

where N is the number of atoms, and $\Delta q_i = q_i - q_c$ is the difference between atom i’s position q_i and the centroid position q_c (average atom position). Here, ⊗ denotes outer vector product.

The trace of the gyration tensor is simply related to the radius of gyration by $G_{Tr} = R_g^2$, but, as noted above, this latter quantity cannot be predicted without the assumption of a fixed, approximately spherical protein shape. On the other hand, as long as a protein’s shape does not strongly deviate from that of an ellipsoid, the volume associated with G should obey the relationship

$$V_g \approx V_g^{res} N_{res} \qquad (3)$$

where the gyration volume is given by

$$V_g \equiv \frac{4}{3}\pi\sqrt{|G|} \qquad (4)$$

V_g^{res} is an average volume per residue and $|\cdots|$ denotes determinant.

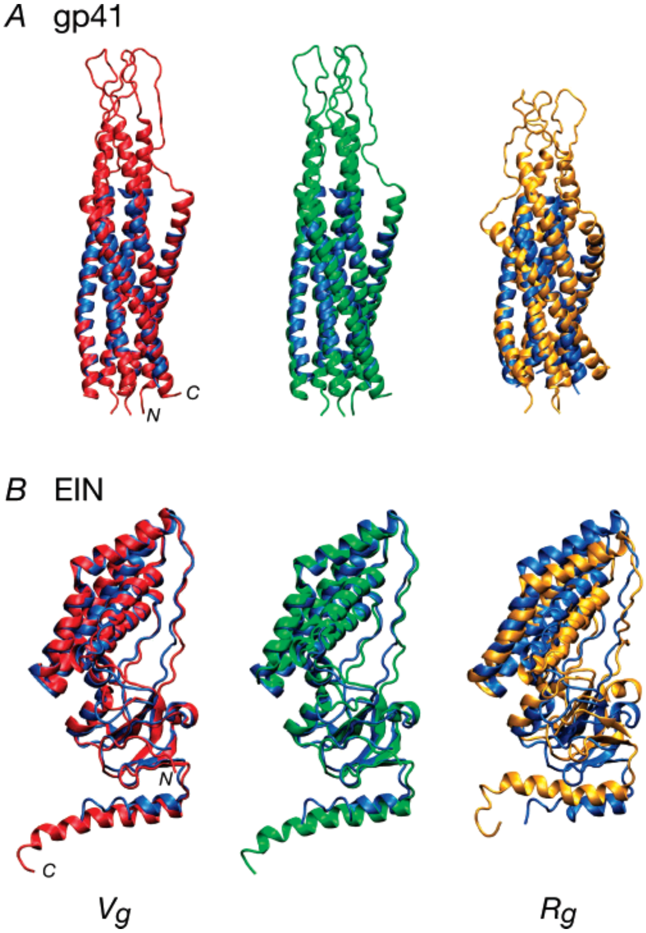

Figure 2. NMR structures computed by simulated annealing refinement of a good starting model generated from the NMR data. Reference X-ray structures are shown in blue. The red structures on the left were calculated using the V_g pseudopotential, the green structures in the center were calculated with no packing pseudopotential, and the yellow structures on the right were calculated including the R_g-based pseudopotential with a single selected region. (A) The structure of e-gp41; (B) the N-terminal domain of enzyme I. For both structures, the R_g-calculated structures are severely compacted along the long axis. For e-gp41, inclusion of the V_g term prevents expansion perpendicular to the long axis, while the V_g term has little effect on the calculated structure of EIN.

We manually scanned the 281 protein X-ray structures in the Protein Data Bank solved at a resolution of 1.2 Å or better, and removed extended structures and structures with multiple domains that do not interact strongly with one another to arrive at a set of 220 well-packed structures for which we calculated R_g and V_g. A plot of V_g versus N_{res} for these structures is shown in Figure 1A. For these structures we found that V_g correlates better with N_{res} than R_g does with $N_{res}{}^b$ with Pearson correlation coefficients of 0.989 and 0.941, respectively. A very weak dependence of V_g/N_{res} on N_{res} was found, with a correlation coefficient of 0.207, consistent with constant protein density. A histogram of V_g/N_{res} is shown in Figure 1B.

Thus, V_g was deemed a good candidate on which to base a pseudopotential. We chose a potential energy form with two components:

$$E_{gyr} = w_{gyr}\,(w_{gyr}^{(1)}\,E_p(V_g - V_g^{res}; 0) + w_{gyr}^{(2)}\,E_p(V_g - V_g^{res}; DV_g)) \qquad (5)$$

where

TABLE 2: Effect of Incorporation of the V_g and R_g Restraints on Accuracy for Structure Determination from an Extended Initial Structure

	no term[a]			V_g			R_g		
protein	accuracy[b]	viols[c]	N_c[d]	accuracy	viols	N_c	accuracy	viols	N_c
BAF	3.13 ± 1.92	64	9	2.92 ± 1.43	72	7	3.07 ± 1.32	85	7
p53	0.52 ± 0.05	0.3	29	0.55 ± 0.06	0.3	29	0.54 ± 0.09	0.0	30
e-gp41	3.62 ± 4.02	153	11	1.55 ± 0.39	94	24	3.92 ± 0.28	89	0
EIN	1.60 ± 0.16	0.1	23	1.72 ± 0.43	0.1	24	3.57 ± 0.26	0.0	0

[a] This annealing protocol started with a random extended structure with correct covalent geometry for each structure. Molecular dynamics were then run at 5000 K for 80 ps or 10 000 steps using the internal variable module[21] with a reduced force constant on the NOE energy term. Simulated annealing in torsion angle coordinates was then performed by lowering the temperature from 5000 to 25 K in 25 K increments, with 0.4 ps or 200 molecular dynamics steps taken at each temperature. During annealing, the force constants of various potential terms were scaled as in the standard protocol.[18] Finally, all-atom minimization in torsion angle and Cartesian coordinates was performed as a final step. [b] Accuracy in Å relative to the X-ray structure averaged over the top 50% of structures sorted by the sum of bond, angle, improper, dihedral, NOE, and dipolar coupling energy terms. [c] Average number of NOE violations > 0:5 Å reported for the top 50% of structures [d] Number of converged structures out of a total of 30 calculated. For these purposes, a structure is defined as converged if its backbone accuracy is within 2.5 Å of the X-ray coordinates.

$$E_p(x,\Delta x) = \begin{cases} (x - \Delta x)^2 & \text{for } x > \Delta x \\ (x + \Delta x)^2 & \text{for } x < -\Delta x \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

We used values $V_g^{res} = 15.47$ Å^3, $\Delta V_g = 3.19$ Å^3, $w_{gyr}^{(1)} = 0.0001$ kcal/Å^3 and $w_{gyr}^{(2)} = 1$ kcal/Å^3, so that the first component is a soft bias toward the mean of the distribution, while the second is a strong term restraining V_g to the range between the vertical red lines in Figure 1. This potential term was implemented within the Xplor-NIH package.[8]

Results and Discussion

We compared the use of the V_g and R_g restraints on the structure calculations of four proteins for which both NMR and X-ray structures are available: the dimeric barrier-to-autointegration factor (BAF) (∼21 kDa),[9,10] the tetramerization domain of p53 (∼20 kDa),[11,12] the trimeric ectodomain of SIV gp41 (e-gp41) (∼44 kDa),[13,14] and the N-terminal domain of enzyme I of the *Escherichia coli* phosphoenolpyruvate/sugar phosphotransferase system (EIN) (∼30 kDa).[15,16] The first two multimers comprise roughly globular structures, while the latter two structures are rather elongated (i.e., nonspherical).[17] Disordered tail regions were omitted in the calculation of V_g and R_g for all structures.

Structure Refinement. A conventional gentle simulated annealing protocol[18] was employed to calculate 50 structures starting from good models with the correct overall fold. Of these structures, those with no structural violations[19] were used for generating statistics. The force constants for the V_g and R_g terms were held constant at the values $w_{gyr} = 1$ and $w_{rgyr} = 100$ kcal mol^{-1}, respectively. For the four structures, the V_g and R_g restraints were applied to all residues, excluding disordered tail regions. The definitions of these regions include the following residues ranges: 3−89 (BAF), 325−355 (p53), 29−148 (e-gp41), and 1−250 (EIN). Table 1 shows that, within the accuracy of the calculations, the E_{gyr} refinement term generally improves the agreement with the corresponding crystal structure coordinates. The most marked improvement in structure accuracy is for e-gp41. In Figure 2A it can be seen that the e-gp41 structure computed without the V_g term is expanded perpendicular to the long axis relative to the X-ray structure, while the V_g-calculated structure is significantly less so. For the other structures, the effect of the V_g term is smaller: the agreement with the corresponding X-ray structures is within the spread of structures. While the accuracies of the EIN structures computed with and without the V_g term lie within the respective computational uncertainties, the fact that the V_g structures have a slightly worse accuracy may indicate that the ellipsoidal approximation is slightly violated by this structure.

For comparison, we also calculated structures including the R_g term with a single selection consisting of all of the well-ordered residues. This is in contrast to the method of breaking the structure into overlapping globular regions employed in previous work.[3] Table 1 and Figure 2 show that the two elongated structures are extremely distorted relative to those calculated with the E_{gyr} term or without any packing potential, as expected. In contrast, the two multimers that pack to globular moieties show higher accuracy when the R_g term is included. So much so, that the R_g-refined BAF structures are actually slightly more accurate than those determined without that term.

Structure Determination from Extended Initial Structure. As the real value of the V_g restraint comes from being able to apply the potential without already knowing the structure, we again calculated the same four structures using the same NMR data, but this time the initial structures were random extended chains and a more vigorous simulated annealing protocol was employed using Xplor-NIH's internal variable module[21] to perform dynamics and minimization in torsion-angle space. In this modified protocol, initial molecular dynamics were carried out at 5000 K, and the force constant w_{gyr} on E_{gyr} was scaled up from 0.002 to 1 over the course of annealing. Table 2 shows that use of the V_g term produces more accurate structures or (at worst) has no effect. Analogous calculations were carried out using the R_g term, and again greatly distorted structures were produced for the elongated proteins e-gp41 and EIN. In fact, the R_g term is shown to be particularly dangerous in this context, as demonstrated by the EIN results in Table 2, as all the nuclear Overhauser effect (NOE) restraints could be satisfied when the R_g term was used, but the resulting structure is quite distorted.

One would expect that the accuracy of the computed structures would be less when determining structures from an extended chain relative to that computed by a refinement protocol, such as that of the previous section. A comparison of Tables 1 and 2 shows that refinement does indeed produce more accurate structures for 3 of 4 proteins. It is interesting, however, that the annealed structure of EIN is of slightly higher accuracy than the refined structure.

Use of the V_g target is also promising as a means to improve convergence, as shown in the number of converged structures (N_c) column in Table 2. For e-gp41, use of the V_g term results in a factor of 2 improvement in convergence. The results for the globular structures show little change in convergence with the addition of the V_g or R_g pseudopotentials.

Concluding Remarks

We have introduced an empirical pseudopotential restraining the gyration volume, V_g, that can improve the packing of protein structures generated from NMR data and is safe to use for well-structured protein domains of various shapes. The previously introduced restraint on the radius of gyration, R_g,[3] can be used for final refinement of regions that are already known to be globular and, in some cases, produces better results, but it is not safe for early stages of refinement. The V_g potential should be particularly useful for automatic structure determination procedures such as PASD,[22] ATNOS/CANDID,[23] or Autostructure.[24] In the context of such protocols, the V_g packing term can be safely used to improve the produced structures and to aid in convergence of the respective algorithms.

Acknowledgment. We thank Attila Szabo for useful suggestions. This work was supported by the CIT and NIDDK Intramural Research programs of the NIH.

References and Notes

(1) (a) Clore, G. M.; Gronenborn, A. M. *Ann. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 29−63. (b) Gronenborn, A. M.; Clore, G. M. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 351−385.

(2) Abagayan, R. A.; Totrov, M. M. *J. Mol. Biol.* **1997**, *268*, 678−685.

(3) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

(4) (a) Kuszewski, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2001**, *123*, 3903−3918. (b) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2003**, *125*, 1518−1525.

(5) Nederveen, A. J.; Doreleijers, J. F.; Vranken, W.; Miller, Z.; Spronk, C. A. E. M.; Nabuurs, S. B.; Guentert, P.; Livny, M.; Markley, J. L.; Nilges, M.; Ulrich, E. L.; Kaptein, R.; Bonvin, A. M. J. J. *Proteins* **2005**, *59*, 662−672.

(6) Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 338−352.

(7) Quillin, M. L.; Matthews, B. W. *Acta Cryst. D.* **2000**, *56*, 791−794.

(8) (a) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 66−74. (b) Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. NMR Spectrosc.* **2006**, *48*, 47−62.

(9) Cai, M.; Huang, Y.; Zheng, R.; Wei, S.-Q.; Ghirlando, R.; Lee, M. S.; Craigie, R.; Gronenborn, A. M.; Clore, G. M. *Nat. Struct. Biol.* **1998**, *5*, 903−909.

(10) Umland, T. C.; Wei, S. Q.; Craigie, R.; Davies, D. R. *Biochemistry* **2000**, *39*, 9130−9138.

(11) Clore, G. M.; Ernst, J.; Clubb, R. T.; Omichinski, J. G.; Kennedy, W. M. P.; Sakaguchi, K.; Appella, E.; Gronenborn, A. M. *Nat. Struct. Biol.* **1995**, *2*, 321−332.

(12) Jeffrey, P. D.; Gorina, S.; Pavletich, N. P. *Science* **1995**, *267*, 1498−1502.

(13) Caffrey, M.; Kaufman, J.; Stahl, S. J.; Wingfield, P. T.; Covell, D. G.; Gronenborn, A. M.; Clore, G. M. *EMBO J.* **1998**, *17*, 4572−4584.

(14) Malashkevich, V. N.; Chan, D. C.; Chutkowski, C. T.; Kim, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9134−9139.

(15) Garrett, D. S.; Seok, Y.-J.; Liao, D. T.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. *Biochemistry* **1997**, *36*, 2517−2530.

(16) Liao, D. I.; Silverton, E.; Seok, Y.-J.; Lee, B. R.; Peterkofsky, A.; Davies, D. R. *Structure* **1996**, *4*, 861−872.

(17) Clore, G. M.; Gronenborn, A. M.; Szabo, A.; Tjandra, N. *J. Am. Chem. Soc.* **1998**, *120*, 4889−4890.

(18) Nilges, M.; Gronenborn, A. M.; Brünger, A. T.; Clore, G. M. *Protein Eng.* **1988**, *2*, 27−38.

(19) A geometrical restraint is counted as violated if it differs from the expected value by more than a specified threshold. The thresholds for the various restraints are as follows: NOE distance restraints, 0.5 Å; torsion angle restraints, 5°; bond lengths, 0.05 Å; bond angles, 5°; improper dihedral angles, 5°.

(20) (a) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *Protein Sci.* **1996**, *5*, 1067−1080. (b) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1997**, *125*, 171−177. (c) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2002**, *124*, 2866−2867.

(21) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *152*, 288−302.

(22) Kuszewski, J.; Schwieters, C. D.; Garrett, D. S.; Byrd, R. A.; Tjandra, N.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 6258−6273.

(23) (a) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **2002**, *24*, 171−189. (b) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.

(24) (a) Huang, Y. J.; Tejero, R.; Powers, R.; Montelione, G. T. *Proteins* **2006**, *15*, 587−603. (b) Huang, Y. J.; Powers, R.; Montelione, G. T. *J. Am. Chem. Soc.* **2005**, *127*, 1665−1674.