JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

pubs.acs.org/JACS

Structure of a Functional Amyloid Protein Subunit Computed Using Sequence Variation

Pengfei Tian,[†] Wouter Boomsma,[‡] Yong Wang,[‡] Daniel E. Otzen,*[,§] Mogens H. Jensen,*[,†] and Kresten Lindorff-Larsen*[,‡]

[†]Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

[‡]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5 DK-2200 Copenhagen N, Denmark

[§]Interdisciplinary Nanoscience Center (iNANO), Centre for Insoluble Protein Structures (inSPIN), Department of Molecular Biology and Genetics, Aarhus University, Gustav Wieds Vej 14, 8000 Aarhus C, Denmark

Supporting Information

ABSTRACT: Functional amyloid fibers, called curli, play a critical role in adhesion and invasion of many bacteria. Unlike pathological amyloids, curli structures are formed by polypeptide sequences whose amyloid structure has been selected for during evolution. This important distinction provides us with an opportunity to obtain structural insights from an unexpected source: the covariation of amino acids in sequences of different curli proteins. We used recently developed methods to extract amino acid contacts from a multiple sequence alignment of homologues of the curli subunit protein, CsgA. Together with an efficient force field, these contacts allow us to determine structural models of CsgA. We find that CsgA forms a β-helical structure, where each turn corresponds to previously identified repeat sequences in CsgA. The proposed structure is validated by previously measured solid-state NMR, electron microscopy, and X-ray diffraction data and agrees with an earlier proposed model derived by complementary means.

Amyloid fibers are formed due to protein aggregation and are commonly associated with a variety of human diseases.[1] Knowing the molecular structures of amyloids provides a framework to understand the propensity to form such aggregates and to design potential inhibitors or regulators. In contrast to disease-related amyloid that is caused by the aggregation of misfolded or transiently unfolded proteins, functional amyloid is formed through a highly regulated protein assembly process, with the purpose to fulfill a specific biological function.[2] For instance, in humans, functional amyloid plays a vital role in physiological processes such as hemostasis and melanin synthesis.[3]

A class of functional amyloid fibers called curli assembles on the cell surface of Enterobacteriaceae such as *Escherichia coli* and *Salmonella* spp, where they are essential for binding to and internalization into the host cell and might also activate the host immune system.[2,4] Recent studies have shown that curli fibrils, like those formed from Aβ, can induce inflammatory responses.[5] A small organic molecule appears to promote oligomer assembly of both curli and α-synuclein,[6] and hence structural studies of curli fibrils might help understand the molecular origins of amyloid diseases.

Curli fibrils of *E. coli* are primarily aggregates of a subunit protein, CsgA, which is secreted as a soluble, unstructured protein and then aggregates in a manner controlled by the protein CsgB to form an amyloid fibril on the cell surface.[7] Due to the insoluble and noncrystalline nature of amyloid fibrils, techniques like solution NMR and X-ray crystallography are not easily applicable to determine amyloid structures. Solid-state NMR (ssNMR) and electron spin resonance spectroscopy can in certain cases provide structural constraints on amyloid fibrils and have been sufficient to rule out particular structural arrangements of curli.[8] Nevertheless a detailed structural characterization of CsgA fibrils remains an unsolved problem.

In contrast to disease-related amyloids caused by protein misfolding, functional amyloids are 'beneficial' aggregation systems that are evolved by nature and evolutionarily conserved across species.[2] This suggests that there is an evolutionary pressure to maintain the amyloid structure of these proteins and that only mutations that preserve the stability of the amyloid state will be allowed. Because mutations in amino acids that are in close spatial proximity are expected to be correlated, so as to maintain the stability and function of the protein, the pattern of covariation among residues in orthologs can provide information about tertiary contacts. We therefore hypothesized that by analyzing a multiple sequence alignment (MSA) of CsgA we might find an 'experimental' signal to determine an atomic-resolution structural model of the protein in its amyloid state.

There is a long history for the idea of using coevolution for molecular structure prediction.[9−14] Recent growth in sequence databases and new, efficient algorithms to disentangle indirect couplings in a network has dramatically improved our ability to predict residue−residue contacts, thereby greatly enhancing the practical applicability of the method.[10,15−19] Using the contact information as structural restraints in molecular simulations, high-quality de novo models have been obtained for several protein families and for both soluble and membrane proteins.[15−19] To our knowledge, the approach has not yet been applied to functional amyloid structures, and the current

Received: September 11, 2014
Published: November 21, 2014

dx.doi.org/10.1021/ja5093634 | J. Am. Chem. Soc. 2015, 137, 22−25

study therefore serves as a probe into the potential applicability of the approach in this domain.

We initiated our analysis by conducting a CsgA homology search using HHblits[20] and the uniprot20 database, leading to a MSA of 390 entries (SI). We note that none of the homologue sequences has an associated experimentally solved structure, which rules out template-based modeling as a feasible path toward a structural model (see SI: text). Instead, we use the MSA to infer tertiary contacts using a consensus prediction from two different methods: the maximum entropy direct coupling analysis approach (EVCOUPLING),[15] and a method based on sparse inverse covariance estimation (PSICOV).[18] The cutoff of the *E*-value (10^{-3}) and the number of high ranking residue−residue pairs (top 50) predicted by each method were chosen as a trade-off between a maximum number of sequences in the alignment and adequate coverage of the whole CsgA sequence by MSA. The results from the covariance analysis are remarkably clear (Figure 1): there is a striking pattern of contacts parallel to the diagonal, suggesting that the internally homologous segments R1−R5 are arranged in a parallel fashion, compatible, e.g., with a helical-like structure with a period of ∼23 residues. The predicted contacts display significant overlap (Figure 1), and we reduce noise further by considering only contacts occurring in both predictions.

Given CsgA's role as a functional aggregate, a central question is whether the observed contacts are intermolecular or whether they describe interactions between monomers in the fibrillar assembly. To address this question, we used an approach that has been proposed in the context of a similar issue in NMR spectroscopy:[21] simulations using either two or three CsgA molecules are conducted allowing each contact to be satisfied either internally or between monomers (Figure S3). The results suggest strongly that the predicted contacts are dominated by internal interactions within the individual monomer subunits.

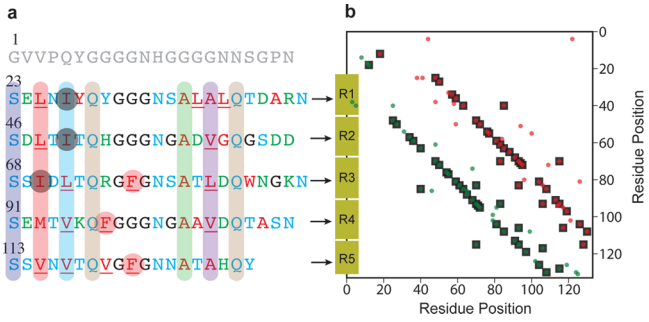

Figure 1. CsgA sequence and contact prediction from a MSA. The N-terminal signal sequence of CsgA (not shown) is cleaved after secretion, leaving the primary sequence of the amyloid fibril subunit, CsgA, as shown in (a). The sequence from residue 23 to 131 contains five imperfect repeats (R1−R5) that are aligned vertically to highlight the repeat structure. The individual amino acids are color-coded in red (hydrophobic), blue (polar), and green (acidic). The background colors of seven amino acid columns correspond to the colors of the hydrophobic cores in Figure 2e,f. (b) We used residue coevolution information in a MSA with 390 entries to predict spatially close residues. The plot shows the 50 most likely contacts calculated either by the EVCOUPLING (lower triangle, green) or PSICOV (upper triangle, red) method, in both cases restricting to pairs with a minimum sequence separation of six residues. The symmetry of the plot (larger version is shown in Figure S1) is evidence of the robustness of the contact predictions, and a consensus prediction is shown in black (also listed in Table S1).

To investigate the subunit structure in greater detail, we constructed a hybrid pseudoenergy function, $E_{tot} = E_{cov} + E_{ProFASi}$, where E_{cov} is a contact potential that drives covarying amino acids close together in space (see SI), and $E_{ProFASi}$ is an implicit solvent, all-atom energy function, which has been successfully used in simulations of aggregation and reversible protein folding.[22] The hybrid model allows us to complement the nonlocal information extracted from the derived contacts with features captured by standard physical force fields such as hydrogen bonding, hydrophobicity, and excluded volume. Simulations were carried out using enhanced Monte Carlo sampling[23] in the ProFASi simulation framework.[22] All simulations were started from a fully extended structure.

Simulations designed to find low-energy conformations consistently resulted in β-helical CsgA structures (Figure 2). The β-helix motif provides an extremely stable architecture owing to the hydrogen bonds formed between the β-strands (Figure 2c,d) and the 'rectangular' hydrophobic core between layers of β-sheets (Figure 2e,f). Surprisingly we find in our simulations that CsgA can attain both left-handed (Figure 2a,c,e) and right-handed (Figure 2b,d,f) β-helices and that the two are found about equally often in our simulations.

Previously, a right-handed β-helix model of CsgA was proposed based on the assembly of short fragments with matching secondary structures.[24] It appears that the right-handedness of CsgA in this previous study was dictated by the choice of a right-handed β-helix template. To our knowledge, the only β-helix

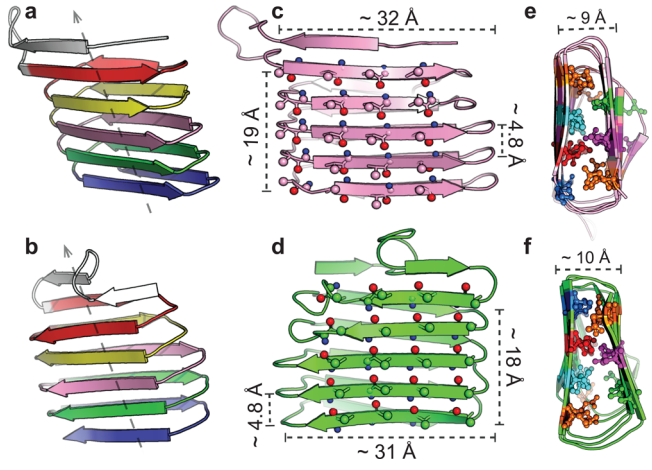

Figure 2. CsgA forms a β-helix. The predicted structures obtained from the simulation of CsgA result in both (a) left-handed and (b) right-handed β-helical structures. The structures shown are the lowest energy structures found in 64 simulations, 11 and 16 of which converged to the same left- and right-handed β-helix fold, respectively, with the remainder trapped in conformations of high energies. From top to bottom, the R1−R5 repeats are shown as red, yellow, pink, green, and blue, respectively. The arrow indicates a likely fibril axis. Panels (c) and (d) illustrate the location of the side chains on the surface of both forms (only Cβ atoms are shown), highlighting polar and acidic amino acids such as Ser, Glu, Asn, Thr, Asp, and Lys. Hydrogen-bonded backbone oxygens and protons are shown in red and blue, respectively. The distance between the parallel β-strands along the axis is ∼4.8 Å. The length of the R1−R5 domains along the fibril axis is ∼19 Å. The width of the β-helical core is around ∼31 Å. (e,f) The 'rectangular' hydrophobic cores are primarily formed by side chains of Ala, Ile, Leu, Met, and Val. The colors of the residues in (e) and (f) correspond to those used in Figure 1b, so that, e.g., the Ser ladder is blue and the Gln is orange. The distance between the two opposite β-strands is ∼10 Å.

dx.doi.org/10.1021/ja5093634 | J. Am. Chem. Soc. 2015, 137, 22−25

fibrillar state that has previously been solved to atomic resolution by experiment is formed by the HET-s (218−289) prion, whose monomer has a left-handed orientation.[25]

In an attempt to determine whether CsgA in its natural form is dominated by the left- or right-handed topology, we calculated the total energy for the two cases. We found, however, that the energies of the structures are very similar (Figure 3a), in part due to the similar helical structures with opposite handedness (Figure S4), the similar packing of the two different cores (Figure 2) and the achiral information provided by the predicted contacts.

Control simulations of CsgA using only the predicted contacts but not the ProFASi force field resulted in structures with roughly the right topology, but no regular secondary structure, and simulations with ProFASi but no contacts resulted in an extended β-sheet structure that is incompatible with the predicted contacts (Figure S5). We also predicted contacts for a protein with a known β-helical structure and found these to be in excellent agreement with the experimentally determined structure (Figure S6).

To validate our models, we compared the structures to a range of previously measured experimental data. X-ray fiber-diffraction experiments on CsgA fibrils in vitro resulted in atomic spacings of ∼4.7 and ∼9 Å, as is commonly found in amyloid fibrils.[8] These values coincide with those in our structures (Figure 2) where we find the distance between β-strands within each β-sheet to be ∼4.8 Å, and the distance between the two β-sheets is ∼9−10 Å.

Electron microscopy studies of CsgA reveal highly narrow fibrils with a diameter of only ∼30 Å,[8] in good agreement with the width of the β-sheet in our structures (∼30−31 Å; Figure 2c,d). In experiments, such narrow fibrils were often found to associate laterally to form wider bundles. Comparing our CsgA structures to the known structure of HETs prion fibrils,[25] one would expect the growing direction of the fibril to be approximately perpendicular to the β-strands and parallel to the interstrand hydrogen bonds (Figure 2a,b). This would be compatible with an N- to C-terminal aggregation scheme. The large exposed lateral areas are covered by polar and acidic amino acids side-chains (Figure 2c,d), which we speculate might form sites for lateral fibril association.

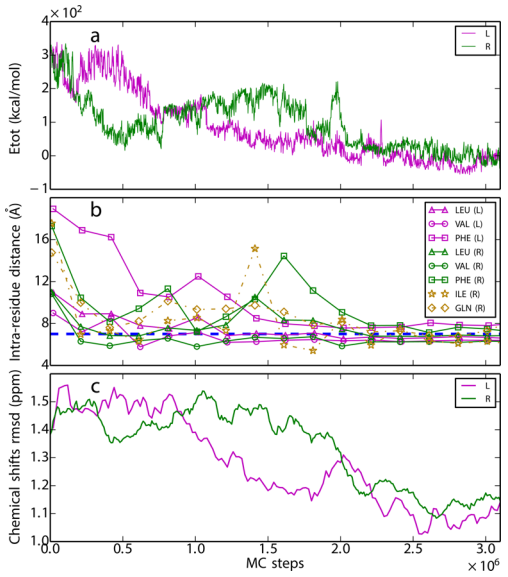

Figure 3. Validation of CsgA structures. The plots show how (a) the energy (E_{tot}), (b) inter-residue distances, and (c) calculated chemical shifts (rmsd) progress during two representative simulations of CsgA, that lead either to a (purple) left-handed or (green) right-handed β-helical conformation. The lowest energy structures of these two simulations are shown in Figure 2. The energies shown in (a) report on E_{tot}, and thus both include the force field energy and the distance restraints extracted from the MSA. The blue dashed line in (b) represents the experimental value. Data in (b) and (c) were smoothed[26] to reduce noise and ease visualization. In all three panels, L and R refer to the simulations leading to a left- and right-handed structure, respectively. In (b) the distances for ILE(L) and GLN(L) were left out for clarity and can be found in Figure S7.

Dark-field transmission electron microscopy was previously used to estimate the mass-per-length of individual CsgA fibrils, which was found to be ≤1.5 kDa/Å.[8] Based on the molecular mass of CsgA (13.9 kDa), this suggests a lower limit for the length of an individual unit along the fibrils axis to be ∼9 Å, which is also compatible with our model.

More detailed structural information on the CsgA fibril has been provided by ssNMR experiments.[8] Inter-residue distances were previously measured between ^{13}C-labeled carbonyls of Val, Leu, or Phe residues separately. In practice, the experiments probe the distance to the nearest neighbor of the same amino acid type which was found experimentally to be ∼7 Å for all three types of amino acid studied. We calculated the corresponding distances throughout our simulations (Figure 3b). As the simulations progress to form either left- or right-handed structures, the distances converge to values that are in good agreement with the experiments (6−8 Å). For all simulations that converged to the left-handed β-helical conformations we observe Boltzmann-averaged inter-residue distances of Leu, Val, and Phe to be 7.2, 6.3, 7.9 Å, respectively. For the simulations that converged to the right-handed β-helical conformations, the corresponding values are 7.0, 6.1, and 7.6 Å.

We find larger fluctuations for the distances between Phe residues than for Val and Leu during our simulations. This is presumably caused by the larger sequence separations between Phe residues in the primary structure (Figure 1a, underlined residues), causing these to be more sensitive reporters of the atomic structure than Val and Leu for which there are pairs of residues with little or no sequence separation. The experimentally measured distances between Phe thus more directly probe the compactness of repeats R3, R4, and R5 and are fully compatible with the β-helix motif we propose. We suggest that the validity of our model could be tested further by measuring the inter-residue distances of Ile residues, which are distributed evenly along the first half of the CsgA sequence (positions 27, 50, and 70), and whose distances are sensitive to the conformation of R1, R2, and R3 (Figures 3b and S7). Gln is another potential candidate for such validations (Figures 3b and S7).

As a final means to validate our structures, we turned to backbone chemical shifts measured by ssNMR, as these provide sensitive probes in particular of the secondary structure. Using SPARTA+[27] we calculated the chemical shifts of C_α, C_β, and C′ atoms in Ala, Val, Thr, Glu, Asn, Ser, Thr, and Leu residues and compared the results to experimental ssNMR data (Table S2).[8] Also in this case our structural models appear to be fully consistent with experiments, with all Glu, Ser, and Thr residues, and the majority of Ala and Val residues located in β-strands, while Asn residues are found in loops. We quantified the agreement with experiments by calculating the root-mean-square-deviation (rmsd) between experiment and simulations for all 21 chemical shifts (Figure 3c). The results show clearly that as the structures proceed toward the β-helical structure, the agreement with independent experimental data continues to

dx.doi.org/10.1021/ja5093634 | J. Am. Chem. Soc. 2015, 137, 22−25

improve. The final values obtained, with rmsd ∼1.1 ppm, are close to the inherent uncertainty of the chemical shift predictions of SPARTA+, suggesting that the structures are in very good agreement with experiments. We note that the experiments measure the chemical shifts of CsgA in the context of a full fibril, while our calculations includes only interactions within the monomer of the fibrils. Our finding that the agreement between experiment and simulation is close to the accuracy of SPARTA+ suggests that this is a valid approximation. As is the case for the molecular energies and inter-residue distances, the available NMR chemical shifts (averaged over the same type of residues) do not allow us to distinguish between left- and right-handed structures. Thus, although we expect that one of the two orientations dominates in nature, we cannot currently determine which. Indeed, it has previously been highlighted that even with high-resolution ssNMR data it is difficult to determine the correct handedness of the helix, though measurements of short intermolecular H^N-H^α distances made it possible to determine that HET-s(218−289) forms a left-handed β-helix in its fibrillar state.[28] Finally, we note that among the recently proposed methods for covariation-based contact prediction, the performance seems to be fairly similar. In addition to the similar results reported for EVCOUPLING and PSICOV (Figure 1b), we also repeated the analyses with a third method, GREMLIN,[19] obtaining very similar results (Figure S8).

To our knowledge, this is the first time the analysis of correlated mutations and computer simulations have been used together to study the structure of a functional amyloid. We find both left- and right-handed β-helical structures from our simulations, and suggest that one or the other or intriguingly perhaps both, represents the structural unit in curli fibrils. The structures have been extensively validated and found to be in good agreement with the available experimental data, and we suggest further experiments to increase the resolution of our structure. Our structure is also compatible with a previously proposed model of the related AgfA protein,[24] which was based on an analysis of the predicted pattern of secondary structure and loops. The clear structural signature of the amyloid state in the evolutionary record of CsgA also supports the crucial functional role of these amyloids. Finally, we note that although the amino acid sequences have provided a clear signal of the structure of individual subunits of CsgA, we have not been able to find any equally strong sequence signals to provide information about interactions between individual structural subunits. Since the function of the molecule is linked to its fibrillar state, the lack of such a signal is surprising, possibly indicating that contact specificity at the interface is reduced compared to that internally in the monomer. Since no experimental data is currently available on the interaction between monomers, the full fibrillar state remains difficult to investigate. However, we envisage that with the rapid growth of sequence databases, and increasingly sensitive methods to extract structural information from them, it will become possible to extend the scope of this approach to study how entire fibrils are organized.

■ ASSOCIATED CONTENT

Supporting Information

Details on methods and data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

dao@inano.au.dk
mhjensen@nbi.dk
lindorff@bio.ku.dk

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Lundbeck Foundation, Villum Foundation, and Novo Nordisk Foundation. We thank Wei Chen for help with TOC graphics.

■ REFERENCES

(1) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333−366.
(2) Otzen, D.; Nielsen, P. H. *Cell. Mol. Life Sci.* **2008**, *65*, 910−927.
(3) Fowler, D. M.; Koulov, A. V.; Balch, W. E.; Kelly, J. W. *Trends Biochem. Sci.* **2007**, *32*, 217−224.
(4) Barnhart, M. M.; Chapman, M. R. *Annu. Rev. Microbiol.* **2006**, *60*, 131.
(5) Tükel, Ç.; Wilson, R. P.; Nishimori, J. H.; Pezeshki, M.; Chromy, B. A.; Bäumler, A. J. *Cell Host Microbe* **2009**, *6*, 45−53.
(6) Horvath, I.; Weise, C. F.; Andersson, E. K.; Chorell, E.; Sellstedt, M.; Bengtsson, C.; Olofsson, A.; Hultgren, S. J.; Chapman, M.; Wolf-Watz, M.; Almqvist, F.; Wittung-Stafshede, P. *J. Am. Chem. Soc.* **2012**, *134*, 3439−3444.
(7) Wang, X.; Hammer, N. D.; Chapman, M. R. *J. Biol. Chem.* **2008**, *283*, 21530−21539.
(8) Shewmaker, F.; McGlinchey, R. P.; Thurber, K. R.; McPhie, P.; Dyda, F.; Tycko, R.; Wickner, R. B. *J. Biol. Chem.* **2009**, *284*, 25065−25076.
(9) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. *Proteins: Struct., Funct., Bioinf.* **1994**, *18*, 309−317.
(10) Lapedes, A. S.; Giraud, B. G.; Liu, L.; Stormo, G. D. *Lect. Notes-Monogr. Ser.* **1999**, *33*, 236−256.
(11) Fodor, A. A.; Aldrich, R. W. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 211−221.
(12) Altschuh, D.; Lesk, A.; Bloomer, A.; Klug, A. *J. Mol. Biol.* **1987**, *193*, 693−707.
(13) Neher, E. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 98−102.
(14) Taylor, W. R.; Hatrick, K. *Protein Eng.* **1994**, *7*, 341−348.
(15) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. *PLoS One* **2011**, *6*, e28766.
(16) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E1293−E1301.
(17) Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S. *Cell* **2012**, *149*, 1607−1621.
(18) Jones, D. T.; Buchan, D. W.; Cozzetto, D.; Pontil, M. *Bioinformatics* **2012**, *28*, 184−190.
(19) Kamisetty, H.; Ovchinnikov, S.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 15674−15679.
(20) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. *Nat. Methods.* **2011**, *9*, 173−175.
(21) Nilges, M. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 297−309.
(22) Irbäck, A.; Mohanty, S. *J. Comput. Chem.* **2006**, *27*, 1548−1555.
(23) Ferkinghoff-Borg, J. *Eur. Phys. J. B* **2002**, *29*, 481−484.
(24) Collinson, S.; Parker, J.; Hodges, R.; Kay, W. *J. Mol. Biol.* **1999**, *290*, 741−756.
(25) Wasmer, C.; Lange, A.; Van Melckebeke, H.; Siemer, A. B.; Riek, R.; Meier, B. H. *Science* **2008**, *319*, 1523−1526.
(26) Savitzky, A.; Golay, M. J. *Anal. Chem.* **1964**, *36*, 1627−1639.
(27) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13−22.
(28) Van Melckebeke, H.; Wasmer, C.; Lange, A.; AB, E.; Loquet, A.; Böckmann, A.; Meier, B. H. *J. Am. Chem. Soc.* **2010**, *132*, 13765−13775.