ARTICLE IN PRESS

JOURNAL OF THE MECHANICAL BEHAVIOR OF BIOMEDICAL MATERIALS XX (XXXX) XXX–XXX

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jmbbm

Research paper

Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale

Graham Bratzel[a,b], *Markus J. Buehler*[a,*]

[a] *Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave. Room 1-235A&B, Cambridge, MA, USA*

[b] *Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA, USA*

ARTICLE INFO

Keywords:
Biological material
Molecular structure
Genetic sequence
Critical length scale
Size effect
Nanostructure
Molecular modeling
Materiomics
Spider silk

ABSTRACT

Spider silk is a self-assembling biopolymer that outperforms many known materials in terms of its mechanical performance despite being constructed from simple and inferior building blocks. While experimental studies have shown that the molecular structure of silk has a direct influence on the stiffness, toughness, and failure strength of silk, few molecular-level analyses of the nanostructure of silk assemblies in particular under variations of genetic sequences have been reported. Here we report atomistic-level structures of the MaSp1 protein from the *Nephila Clavipes* spider dragline silk sequence, obtained using an *in silico* approach based on replica exchange molecular dynamics (REMD) and explicit water molecular dynamics. We apply this method to study the effects of a systematic variation of the poly-alanine repeat lengths, a parameter controlled by the genetic makeup of silk, on the resulting molecular structure of silk at the nanoscale. Confirming earlier experimental and computational work, a structural analysis reveals that poly-alanine regions in silk predominantly form distinct and orderly β-sheet crystal domains while disorderly regions are formed by glycine-rich repeats that consist of 3_{10}-helix type structures and β-turns. Our predictions are directly validated against experimental data based on dihedral angle pair calculations presented in Ramachandran plots combined with an analysis of the secondary structure content. The key result of our study is our finding of a strong dependence of the resulting silk nanostructure depending on the poly-alanine length. We observe that the wildtype poly-alanine repeat length of six residues defines a critical minimum length that consistently results in clearly defined β-sheet nanocrystals. For poly-alanine lengths below six, the β-sheet nanocrystals are not well-defined or not visible at all, while for poly-alanine lengths at and above six, the characteristic nanocomposite structure of silk emerges with no significant improvement of the quality of the β-sheet nanocrystal geometry. We present a simple biophysical model that explains these computational observations based on the mechanistic insight gained from the molecular simulations. Our findings set the stage for understanding how

* *Corresponding author.* Tel.: +1 617 452 2750; fax: +1 617 324 4014.
E-mail address: mbuehler@MIT.EDU (M.J. Buehler).
URL: http://web.mit.edu/mbuehler/www/ (M.J. Buehler).

1751-6161/$ - see front matter
doi:10.1016/j.jmbbm.2011.07.012

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

variations in the spidroin sequence can be used to engineer the structure and thereby functional properties of this biological superfiber, and present a design strategy for the genetic optimization of spidroins for enhanced mechanical properties. The approach used here may also find application in the design of other self-assembled molecular structures and fibers and in particular biologically inspired or completely synthetic systems.

1. Introduction

Spider silk is an extraordinary biomaterial that surpasses most synthetic fibers in terms of toughness through a balance of ultimate strength and extensibility (Termonia, 1994; Simmons et al., 1996; Vollrath and Knight, 2001; Shao and Vollrath, 2002; Becker et al., 2003). The source of spider silk's remarkable properties has been attributed to the specific secondary structures of proteins found in the repeating units of spider silk (Hayashi et al., 1999), which self-assemble into a hierarchical structure. Experimental studies have primarily focused on mapping the repeating sequence units of spider silk and the basic structural building blocks and crystallinity of fibrils. The webs of higher spiders, including the Golden Orb-Weaver *Nephila clavipes*, are composed of different kinds of silk, each with distinct mechanical properties that are adapted for the purpose of that part of the web. Dragline silk, the strongest kind of silk, is used for the spokes and outer frame (Gosline et al., 1999). Two distinct proteins are typically found in dragline silks with similar sequences across species (Gatesy et al., 2001). The dragline silk of *N. clavipes*, one of the most studied spider silks, contains major ampullate spidroins MaSp1 and MaSp2 proteins with different repeat units and distinct mechanical functions (Hayashi and Lewis, 1998; Hayashi et al., 1999; Brooks et al., 2005; Holland et al., 2008). MaSp1 contains poly-alanine or poly-Ala $(A)_n$ and (GA) domains within glycine-rich $(GGX)_n$ repeats, where X typically stands for alanine (A), tyrosine (Y), leucine (L), or glutamine (Q). Studies have suggested that MaSp1 is more prevalent in the spider dragline silk than MaSp2, with a ratio of approximately 3:2 or higher, depending on the species (Hinman and Lewis, 1992; Guerette et al., 1996; Brooks et al., 2005; Sponner et al., 2005).

Recent investigations revealed that antiparallel β-sheet crystals play a key role in defining the mechanical properties of silk by providing stiff cross-linking domains embedded in a semi-amorphous Gly-rich matrix with extensible hidden-length (Thiel et al., 1997; van Beek et al., 2002; Lefevre et al., 2007; Keten and Buehler, 2010; Keten et al., 2010). Studies have also shown that the hydration level and solvent conditions (e.g. ion content and pH) play a large role in the structure and mechanical properties of silk proteins (Dicko et al., 2004; Rammensee et al., 2008) and even the transition from concentrated dope to final silk in the spinning duct. Variations in crystallinity and alignment within the silk thread, for example due to the reeling speed of the collected sample, have also been mapped to macroscopic mechanical properties (Du et al., 2006; Wu et al., 2009). The cross-linking β-sheet crystals employ a dense network of hydrogen bonds (Keten and Buehler, 2010; Keten et al., 2010), have dimensions of a few nanometers, and constitute at least 10%–15% of the silk volume. The existence of 3_{10}-helices and β-turn or β-spiral conformations has been suggested for the amorphous domains (Philip et al., 1994; Thiel et al., 1997; van Beek et al., 2002; Lefevre et al., 2007). However, no robust atomistic-level structural model with explicit solvent and a systematic analysis of sequence-structure correlations has yet been reported. It is anticipated that novel statistical mechanics approaches (Porter et al., 2005), experimental methods, such as X-ray diffraction and scattering (Riekel and Vollrath, 2001; Trancik et al., 2006), solid-state nuclear magnetic resonance (NMR) (Simmons et al., 1996; van Beek et al., 1999; Holland et al., 2008; Jenkins et al., 2010) and Raman spectroscopy (Rousseau et al., 2004; Lefevre et al., 2007; Rousseau et al., 2009), combined with multiscale atomistic-modeling methods such as those based on density functional theory (DFT) (Porter and Vollrath, 2008; Keten and Buehler, 2010) or molecular dynamics (MD) (Brooks, 1995; Ma and Nussinov, 2006; Buehler et al., 2008; Keten and Buehler, 2010) will provide more insight into the atomic resolution structure for spider silk and similarly complex materials. An earlier study of silk using replica exchange molecular dynamics (REMD) (Keten and Buehler, 2010) has yielded first results in comparison with experimental structure identification methods (Kummerlen et al., 1996a,b; van Beek et al., 2002; Holland et al., 2008). Other recent computational studies characterized the mechanics of a MaSp1-like protein. Cetinkaya et al. (2011), but the authors used a constructed structure that is potentially far from the native state. Owing to the lack of current large-scale atomistic models, the links between peptide sequence of actual silk protein remains poorly understood. The availability of powerful new methods to synthesize varied protein materials from the bottom up and will full control over the genetic sequence opens exciting opportunities to engineer materials such as spider silk for specific mechanical and other functional purposes.

Here, atomistic simulations are performed in order to identify structural models of spider silk proteins with the goal of developing a link between the poly-Ala repeat length and the resulting structure of β-sheet nanocrystals. The challenges of reaching native (that is, equilibrium) structures within the time-scales accessible to conventional molecular dynamics simulations require enhanced sampling methods such as replica exchange molecular dynamics (REMD) (Sugita and Okamoto, 1999; Earl and Deem, 2005). We employ REMD to investigate the structures formed by assemblies of segments of MaSp1 polypeptide chains and follow the REMD structure prediction step with a careful molecular dynamics equilibration in explicit water solvent to refine the geometry of the intermediate predicted structures. Along with other protein structure prediction approaches (Bradley et al., 2005; Zhang, 2008), REMD is considered to be a quite effective tool

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

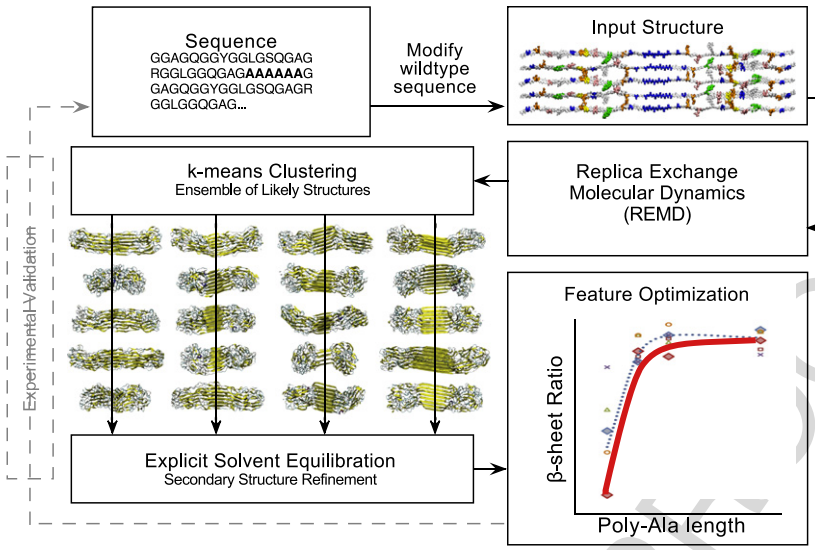

Fig. 1 – Design scheme of feature optimization through the creation and analysis of a test model ensemble. Replica exchange simulations (Keten and Buehler, 2010) are performed on a lattice of aligned strands of the *N. clavipes* MaSp1 peptide sequence with multiple cases of poly-Ala repeat length, including wildtype. The method of *k*-means clustering of the 300 K replica timeline determines the most probable native structures for each case. The secondary structure is then refined by equilibration with explicit solvent. Analysis of the secondary structure and dihedral angles determines the cases of poly-Ala length that result in the most defined β-sheet nanocrystals. While not included in this study, these predictions may be validated against experiment. Once the structure of nanocrystals is optimized, further modifications to the sequence may be explored to optimize other features using this general approach.

for investigating folding and aggregation of proteins, as it reduces the likelihood of kinetic trapping at non-native states (Sanbonmatsu and Garcia, 2002). Through a fast search of the conformation space at high temperatures and more detailed investigations at low temperatures, REMD allows the system to overcome energy barriers and local minima corresponding to non-native structures of proteins (Feig et al., 2003; Rao and Caflisch, 2003; Rhee and Pande, 2003; Miyashita et al., 2009) and facilitates identification of native protein structures from the amino acid sequence with atomistic resolution.

The plan of this paper is as follows. In Section 2, the simulation and structural analysis methods are described. Section 2.1 focuses on the creation of the test model ensemble, while Section 2.2 describes the secondary structure analysis. Section 3 reports the results of the computational experiments. A discussion of these results and conclusion follow in Section 4.

2. Materials and methods

In this section we describe the *in silico* setup and analysis of the model systems considered here. We then describe the methods of structure analysis. A summary of the methods used in this paper is presented in Fig. 1.

2.1. *Initial structure*

Previous computational studies have shown that the poly-Ala unit in the wildtype MaSp1 repeat unit forms a β-sheet nanocrystal (Keten and Buehler, 2010). The model systems treated here focus on a partial sequence with a single poly-Ala repeat with one instance of the Gly-rich semi-amorphous repeat unit on each side. The wildtype *N. clavipes* MaSp1 partial peptide sequence used in this study, in one-letter amino acid code, is GGAGQGGYGGLGSQGAGRGGLGGQGAG**AAAAAA**GGAGQGGYGGLGSQGAGRGGLGGQGAG. The wildtype poly-Ala length of six residues, bolded above (and shown as the letter "**A**"), is systematically varied among our test cases in order to study the effect of the unit's length on crystal formation.

We use REMD to create an ensemble of near-native, energy-minimized test models from an initial aligned lattice structure of partial MaSp1 strands. Experimental studies of recombinant silk (Rammensee et al., 2008) suggest that mechanical shear within a narrowing elongational flow extends and aligns MaSp chains during spinning, and that this encourages alanine "amyloidization" into extended nanoscale β-sheet crystals that cross-link a semi-amorphous filament network (Fig. 2(a)). Mimicking this process, an extended conformation is used when creating a single strand of the MaSp1 partial sequence unit using the TINKER Molecular Modeling Package with CHARMM-19 topology. A rectangular lattice structure (Fig. 2(b)) is created by arranging three parallel layers, where each layer is made of five MaSp1 strands in an anti-parallel arrangement in the side-chain direction. In this initial lattice structure, each strand is at a minimum distance of 10 Å to avoid side-chain interference (this distance is much larger than molecular interactions). While this simulation protocol is somewhat limited, as it does not account for all processes and reactions occurring during silk spinning (Eisoldt et al., 0000), the extended starting configuration mimics the elongational flow of the concentrated dope in the spinning duct and

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

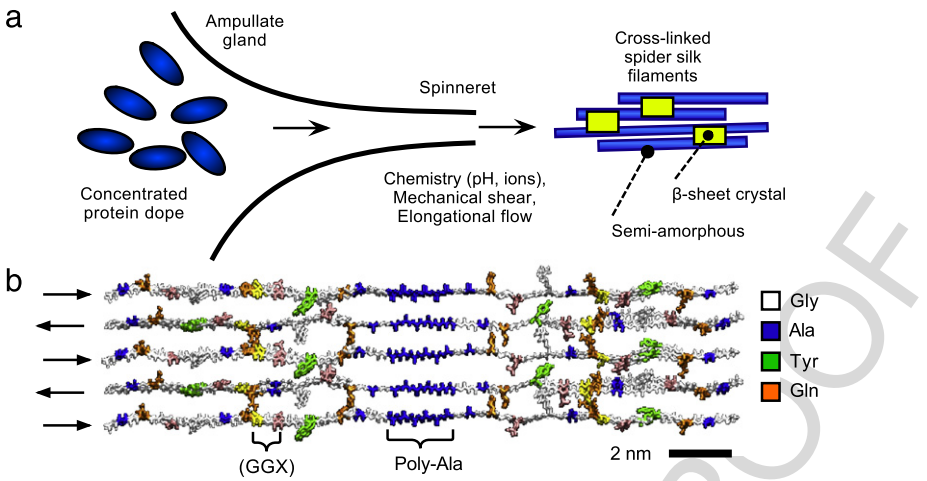

Fig. 2 – The REMD input structure is inspired by the natural silk spinning process. (a) A combination of chemistry and shear flow transform the concentrated protein dope into a filament network cross-linked by β-sheet crystals; based on insight gained from experimental work (Rammensee et al., 2008). (b) The extended lattice of repeat unit strands input into the REMD simulations mimic the elongational flow conditions within the spinneret.

encourages alanine aggregation leading to crystal formation rather than folding of each strand (Rammensee et al., 2008).

2.2. *Replica exchange molecular dynamics*

To create intermediate test structures from the initial lattice structure, we simulate the lattice with Langevin dynamics using CHARMM with the EEF1.1 implicit solvent force field. The implicit solvent model allows a simulation timestep of 2 fs by employing the SHAKE algorithm for hydrogen atoms. Solvent friction is added via a Langevin friction term that allows for high mobility and conformational sampling. While the EEF1.1 model has particular modifications and simplifications of solvent, side-chain, and hydrogen bond interactions, it is orders of magnitude faster than other implicit or explicit solvent models, making it ideal for preliminary simulations of the large-scale silk assembly processes. Since force fields are generally parameterized for room temperature calculations, we only pick final ensemble structures from the lowest temperature replica (*i.e.* 300 K) and use higher temperatures for fast conformational search and overcoming kinetic trapping in the REMD scheme. The REMD protocol is set up and performed using the MMTSB Toolset. We carry out long initialization runs of the extended lattice structure to obtain multiple distinct starting configurations to enhance better sampling in the production run. This is followed by a production run starting from the final configurations of the replicas from the initialization run using an exchange timestep of 2 ps to allow for relaxation of the system. In the production run, we simulate 64 replicas distributed evenly over a temperature range of 300–650 K. Each replica is simulated for a total of 10 ns, corresponding to a total simulation time of 640 ns for the peptide sequence. The EEF1.1 implicit solvent model becomes necessary for simulating each of 64 replicas with over 900 residues for 10 ns within a convenient user timeframe.

2.3. *Explicit solvent equilibration*

The *k*-means clustering algorithm from the MMTSB Toolset is performed on the last 1 ns of the 300 K replica timeline. The centers of the five largest clusters are exported for analysis as the intermediate test structures for the wildtype case of *N. clavipes* with 6 alanine residues in the poly-alanine region of the MaSp1 repeat unit. The creation of the initial lattice, REMD, and *k*-means clustering is performed again for cases of 2, 4, and 12 alanine residues in the poly-alanine region. The twenty resulting intermediate structures constitute the test model ensemble (see Fig. 3). The relative cluster sizes (*i.e.* time representation in the 300 K replica timeline) is used to weigh property averages. To obtain more realistic molecular conformation and tertiary protein structure, the principal structure, *i.e.* that from the largest cluster, for each test case is equilibrated for 20 ns in a wrapping periodic waterbox of TIP3P using NAMD. To prevent image interactions, the water box pads the protein by at least 10 Å. Equilibration is performed with Langevin dynamics at 300 K and with Particle Mesh Ewald (PME) electrostatics to more accurately capture solvent interactions.

2.4. *Analysis methods of structure predictions*

The secondary structure content of each equilibrated structure is determined using the STRIDE algorithm built into the VMD Molecular Graphics Viewer (Humphrey et al., 1996) and customized.tcl scripts. The STRIDE algorithm holds an advantage over DSSP and other secondary structural algorithms by employing pattern recognition of statistically-derived backbone dihedral angle information (Frishman and Argos, 1995). Using the secondary structure results, Ramachandran density plots are created for the principal (that is, the most represented) cluster center using a 10° bin size. Propensity for certain secondary structures along each strand are predicted by analyzing the peptide sequence

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

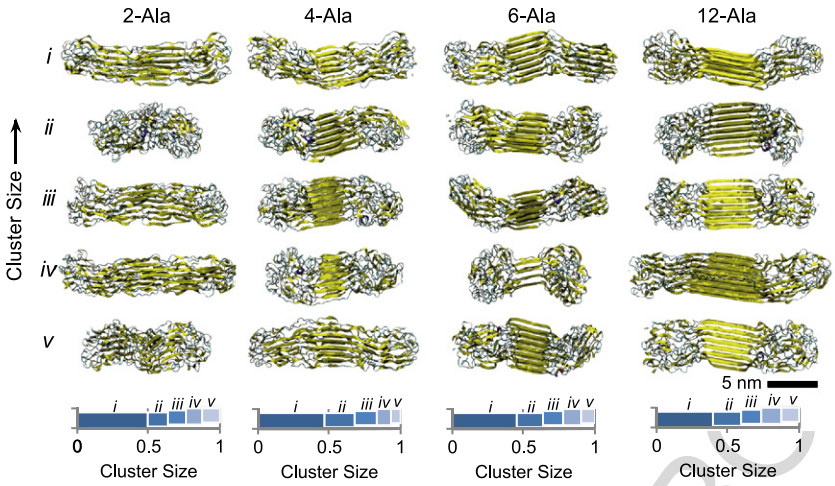

Fig. 3 – The test model ensemble after REMD with implicit solvent. The test models, in a cartoon representation colored by secondary structure, show the five largest *k*-means cluster centers after REMD. Relative *k*-means cluster sizes are shown underneath each test case for comparison. The largest clusters, *i.e.* the principal *i* structures, are most likely to represent native structures for each test case of poly-Ala length.

for each test case in the Protein Plot window in MATLAB for local hydrophobicity and Total β-strand properties. The definition of hydrophobicity as defined by Kyte and Doolittle (1982) is based on an index of relative hydropathicity ranging from −4.5 to +4.5. The value for an individual amino acid is a weighted average of the normalized transfer free energy from water to vapor, the fraction of side-chains found 100% buried in a sample of nearly 1300 experimentally studied proteins, and the fraction of side-chains found 95% buried in the same sample. Amino acids with a negative hydropathicity are considered hydrophilic, those with a positive hydropathicity are hydrophobic, and those near zero are ambivalent. For an entered peptide sequence, the hydropathicity of an individual residue is a weighted average of the hydropathicities of adjacent residues within a sliding window. Thus, a hydrophobic segment reinforces its own hydrophobicity.

The Total β-strand Preference (TBP) (Lifson and Sander, 1979) is a relative index ranging between 0 and 2 based on observed residue contacts in strand–strand interactions instead of the dihedral angles of a single residue. This index adds clarification to the traditional four-state distinction in secondary structure (α-helix, β-strand, reverse turn, and random coil) by focusing on tertiary structure environments such as the assembly of β-strands into a β-sheet. Index values for each naturally-occurring amino acid were determined for both antiparallel and parallel β-sheets in an experimental sampling of 30 proteins. In order to emphasize the predictions of both the hydropathicity index and the TBP index on β-sheet distribution within the MaSp1 repeat sequence, we multiply the values of both indices at each residue along the sequence.

3. Results

We first present our structural predictions as a function of poly-Ala repeat length variations as well as implicit and explicit water solvation during equilibration for a direct comparison between these two approaches. We focus primarily on dihedral angles and secondary structure of the predicted structures. Since the majority of MaSp1 (and silk in general) consist of glycine and alanine amino acids, we directly compare the dihedral φ–ψ angles of the glycine and alanine groups separately for our test cases with experimental data on spider silk proteins. Our studies of structural changes during equilibration with explicit solvent indicate that 20 ns provides sufficient convergence of three metrics: the root-mean-square deviation (RMSD), solvent-accessible surface area (SASA), and total β-sheet content (see Fig. 4). The relatively large change in each metric from the REMD-predicted structure suggests that equilibration in explicit solvent is a necessary step after REMD structure prediction with implicit solvent.

3.1. *Secondary structure*

Glycine residues (Fig. 5(a)) in implicit solvent show symmetry about the origin and a wide distribution about a peak at (−75°, +75°) for all cases of poly-Ala length. Glycine in explicit solvent shows very different peaks at (±90, 0) and (±90, 180). The peaks with explicit solvent are in better agreement with experimental findings (van Beek et al., 2002) as well as allowed Ramachandran regions. Alanine residues (Fig. 5(b)) in implicit solvent show a single peak at (−90°, +75°) for the 2-Ala principal structure. As the poly-Ala length is increased, the peak shifts to around (−140°, +140°). Alanine in explicit solvent show a similar progression toward (−140, +150°), which corresponds to a predominantly anti-parallel β-sheet conformation for the 6-Ala and 12-Ala principal structures and is in excellent agreement with experimental findings with peaks at (−135°, +150°) (van Beek et al., 2002). Note that as the number of simulated amino acids increases, the peaks soften, seen as a greater number of contour lines on the density plots. While the implicit and explicit water solvent models make simplifications for side-chain and hydrogen bond interactions in favor of computational efficiency, our secondary structure and dihedral angle analysis agree well with experimental findings.

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

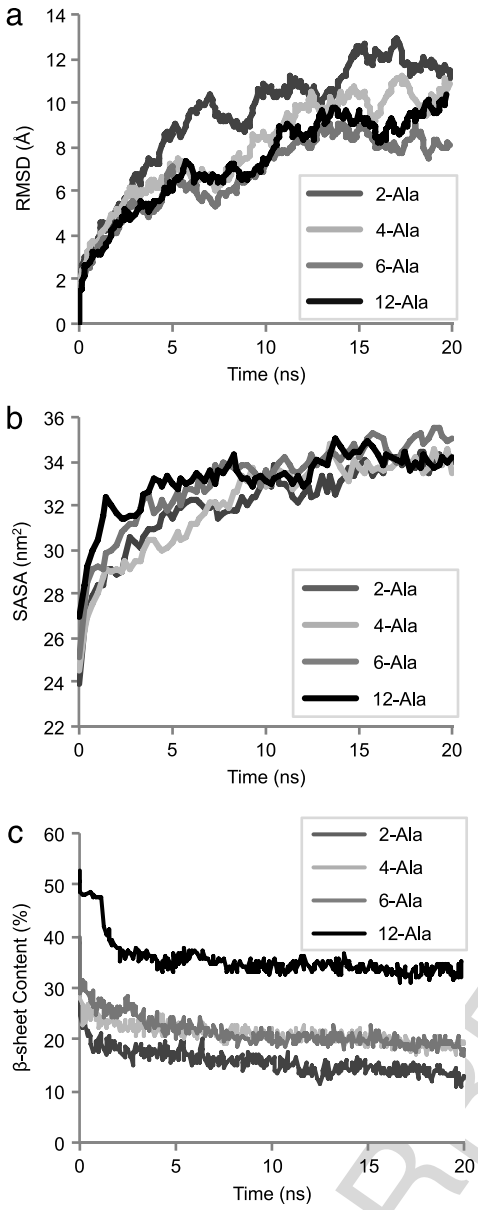

Fig. 4 – Convergence during equilibration of principal structures with explicit solvent. The decreasing rate of change in (a) root-mean-square deviation (RMSD), (b) solvent-accessible surface area (SASA), and (c) total β-sheet content indicate sufficient convergence of the principal structures for each test case after 20 ns. Further equilibration may improve accuracy but is not considered essential.

3.2. *Distribution of β-sheets*

We now focus on relative secondary structure ratios and the spatial distribution of β-sheet conformations for each test case considered here, visualized with cartoon representation in Fig. 6(a). With implicit solvent, we find an average, weighted according to relative cluster size, of approximately 50% β-sheet, 30% turn, and 20% random coil conformation for each test case of poly-Ala length. No helix conformation is reported by the STRIDE algorithm definition. Although the weighted average ratios do not vary more than 5% among the test cases, the distribution of discrete ratios decreases as the poly-Ala repeat length increases. The most consistent (*i.e.* tightest distribution) of β-sheet and turn conformations are observed for the wildtype 6-Ala test case. Consistency among conformational structures for each test case signals an implicit stability of the respective poly-Ala length.

After equilibration of the principal structures in explicit solvent, the semi-amorphous region shows an average of only 10% β-sheet for all test cases. The spatial distribution of residues with β-sheet conformation differs according to the poly-Ala repeat length. By plotting the occurrence of β-sheet for each residue ID among the fifteen chains of the test lattice (Fig. 6(b)), we directly observe β-sheet grouping for each test case. The 2-Ala test case in explicit solvent shows a soft peak in β-sheet occurrence centered at the poly-Ala region. The 46% peak occurrence implies inconsistent and weak crystal formation. For the 4-, 6-, and 12-Ala test cases, 90%–100% of chains have β-sheet conformation in the poly-Ala region, in sharp contrast to an average 10% occurrence outside this region. This shows very strong crystal definition for all cases with poly-Ala repeat length of 4 alanines or longer.

The spatial distribution of residues with β-sheet conformation in explicit solvent (Fig. 7(a)) agrees well with the hydropathicity and Total β-sheet Preference (TBP) predictions in Fig. 7(b). Crystal definition and consistency is made more readily apparent through direct observation of β-sheet occurrence among only the poly-Ala residues (Fig. 7(c)), again with averages weighted by relative cluster size for implicit solvent test cases. Explicit solvent cases refer only to principal structures. For both solvation models, the 2-Ala test group averages 45%–65% β-sheet conformation in the poly-Ala region, with a very wide distribution of discrete points. On the other hand, the 4-, 6-, and 12-Ala test groups average more than 90% β-sheet conformation for both solvent models. The wildtype 6-Ala test case shows the highest average β-sheet content and smallest distribution. The 12-Ala test case is almost as defined, but shows that larger crystals may be degraded by detrimental effects at this length scale, such as hydrogen bond saturation.

4. Discussion and conclusion

We presented results from atomistic REMD simulations on MaSp1 protein segments of the dragline spider silk of *N. clavipes*. We have illustrated that critical conditions for particular secondary structure formation can be found through systematic variation of the peptide sequence alone, and that the length of the poly-Ala repeat unit is critical in defining identifiable stable β-sheet nanocrystals. Specifically, there exists a strong scaling effect where a minimum length of poly-Ala repeats is found. Averaging over all probable structures of each test case shows a minimum poly-Ala length of at least 4 alanine residues for consistent crystal formation. However, this assumes perfect alignment of alanine side-chains before agglomeration. During natural spinning, a minimum of 6 alanines is more realistic for the formation of robust β-sheet nanocrystals, and this is indeed the wildtype poly-Ala length for *N. clavipes*. Non-araneoid and

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

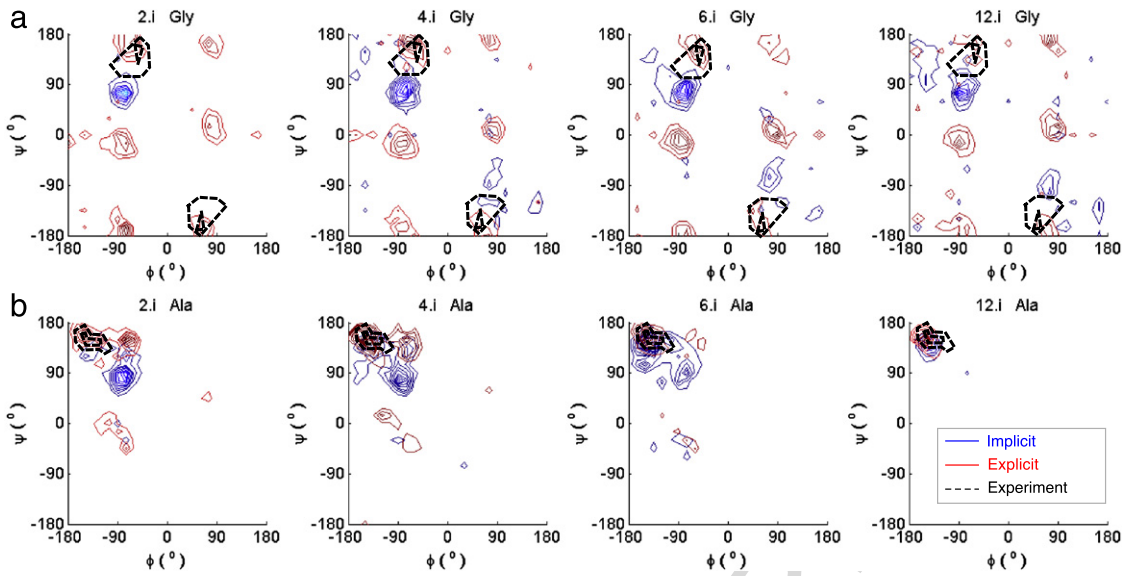

Fig. 5 – Experimental validation of the in silico silk structure predictions based on implicit and explicit water models: Ramachandran density plots of the (a) glycine and (b) alanine in the principal *i* structures. Blue shows conformation with implicit solvent, red with explicit solvent, and dotted black shows experimental peaks (van Beek et al., 2002). For both glycine and alanine, explicit solvent results in peaks closer to both "allowed" Ramachandran regions and to experimental peaks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

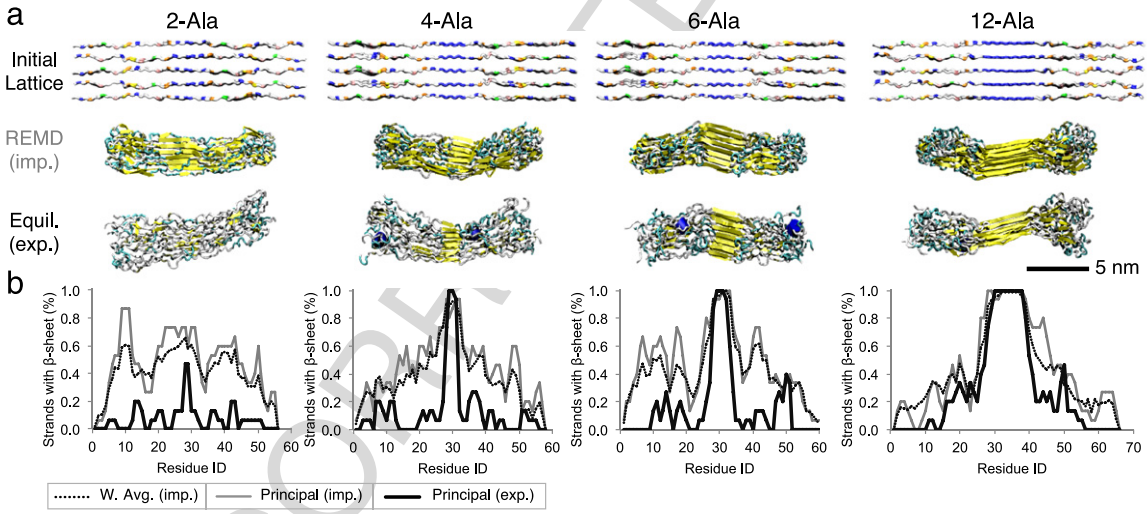

Fig. 6 – Analysis of the β-sheet content, which depends heavily on solvent choice, for implicit and explicit water models. (a) The initial lattice structure before REMD shows the location of the poly-Ala in blue. This also illustrates the coiled hidden length afforded by the semi-amorphous regions after REMD and equilibration. (b) Crystal definition is inferred by the percent of strands with β-strand conformation at each residue for the principal structure for each test case. Solvent effects are clearly observable; after diffusion of explicit solvent, most β-sheets in the semi-amorphous regions are disrupted, while the hydrophobic poly-Ala crystal remains defined. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Minor ampullate (e.g. capture) silk feature short repeats 2–4 Ala in length, seen in Fig. 8. Major ampullate (e.g. dragline) silk, the stiffest and strongest silk, features longer 6-Ala or 8-Ala repeats, depending on species, allowing larger cross-linking β-sheet crystals. While a longer poly-Ala region results in consistently defined crystals, the synthesis of poly-Ala regions longer than 8 alanines may be too metabolically expensive and thus prohibitive during the evolution of this species (Craig et al., 1999), and may be without any further mechanical payoff as suggested in earlier work (Keten et al., 2010).

4.1. *Geometric interference of side-chains*

We explain the observation of a minimum poly-Ala length based on a simple biophysical model (Fig. 9(a)). Alanine

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

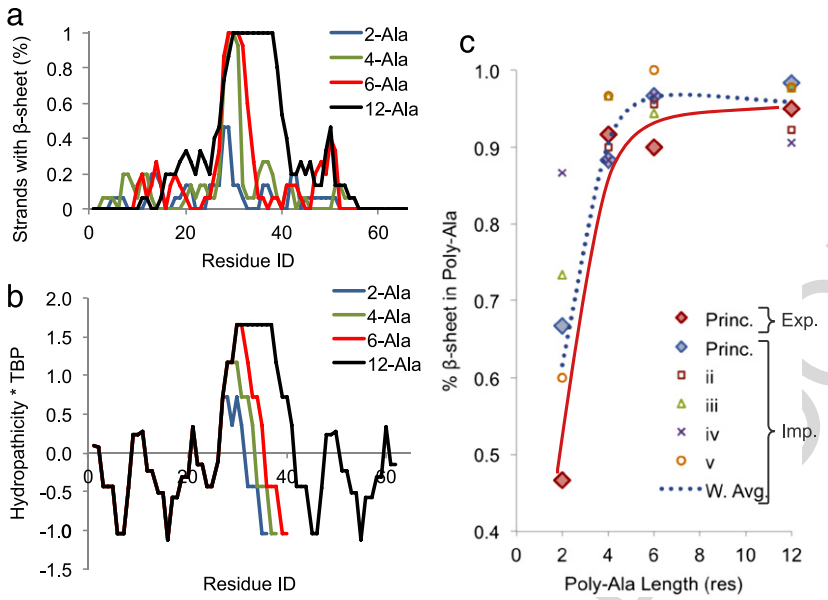

Fig. 7 – β-sheet distribution and crystal definition after structure prediction. (a) β-sheet distribution, averaged over the 15 strands of each principal structure, after equilibration with explicit solvent. (b) Multiplying the hydropathicity and Total β-sheet Preference indices closely predicts the observed β-sheet distribution and crystal definition. Only the values for the 12-Ala case are shown past the poly-Ala region for clarity; the values of the other cases are identical beyond their cutoffs. (c) The β-sheet content within only the poly-Ala region for both implicit and explicit solvents indicates a critical length scale for crystal formation.

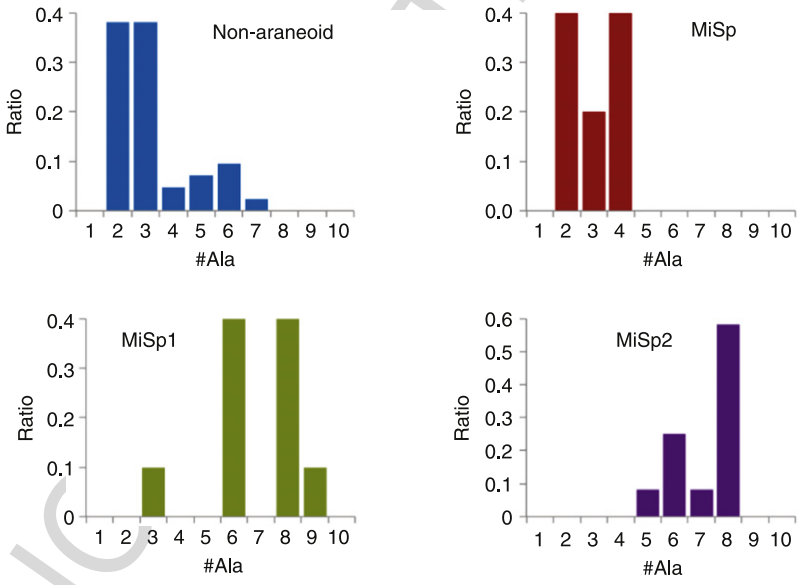

Fig. 8 – Poly-Ala length distribution among spider silk proteins across several species. Ratios are normalized to unity for each category. Values taken from the consensus sequences listed in Gatesy et al. (2001).

side-chains (*i.e.* nonpolar methyl groups) alternate sides of the backbone along a β-strand. This alignment and the small size of the alanine side-chain allow the side-chains of adjacent β-sheets to zip together, in turn allowing aligned poly-Ala β-sheets to closely stack out-of-plane. Geometric interference of the side-chains after stacking provides additional bending and torsional rigidity to the multi-layer crystal. In addition, the hydrophobic and nonpolar nature of alanine reinforces crystal stability by preventing water diffusion between the stacked β-sheets. Indeed, after equilibration of the principal structures in explicit solvent, water was found to have diffused within the semi-amorphous region, but no water was observed within the β-sheet crystal. Side-chain zipping also offers resistance to peeling, as deformed β-strand backbones force side-chains to mechanically clamp onto other adjacent side-chains. The 2-Ala β-strands are not long enough to zip together and are easily cleaved by water or peeled by boundary conditions

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

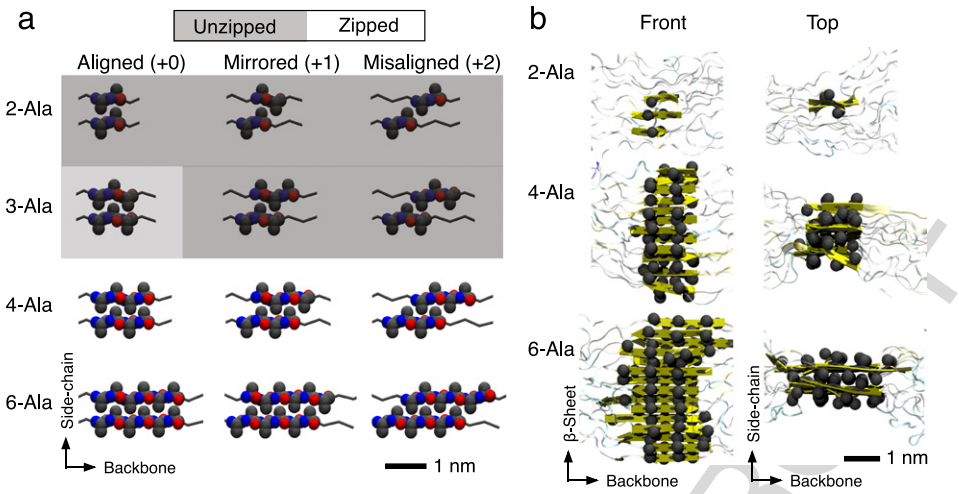

Fig. 9 – Efficient stacking of alanine side-chains determines a critical poly-Ala length for crystal stability. (a) Ala side-chains alternate direction along a β-strand, allowing poly-Ala β-sheets to stack out-of-plane to reinforce hydrophobicity and provide rigidity through geometric interference. (b) Molecular simulation results directly illustrate these trends in crystal stability. Ala side-chain carbon atoms are shown in gray to illustrate alignment in the β-sheet direction and stacking in the side-chain direction.

of the less-dense semi-amorphous region. This explains the low β-sheet content of the poly-Ala region of the 2-Ala test case. Also, 3-Ala or 4-Ala β-strands must be highly aligned to allow side-chain zipping, but can resist cleavage in this case. The 6-Ala (wildtype) case may be misaligned during amyloidization and still result in some side-chains zipping. This margin of error may be structurally worth the energetic cost of additional poly-Ala synthesis.

Our simulation results of poly-Ala length test cases in explicit solvent illustrate these trends in the β-sheet nanocrystal stability (Fig. 9(b)). For the 2-Ala test case, a single β-sheet of only three strands is found in the interior of the protein. Another β-sheet is not present for stacking in the side-chain direction. In contrast, the 4-Ala test case shows a very aligned two-layer crystal that is open to exterior water. The 6-Ala case shows a similar crystal and also illustrates the tolerance of misalignment. In the center of the crystal, several β-strands are misaligned by two residues. However, the core of the crystal remains 4-Ala in width and thus remains stable in the presence of water. The 12-Ala test case (omitted from Fig. 9(b)) shows highly ordered stacking and is in some places three layers thick. Therefore, MD simulation of MaSp1 poly-Ala amyloidization demonstrates that 4-Ala repeats are sufficient for the formation of β-sheet nanocrystals in the final silk. However, repeats of 6-Ala (wildtype) or longer allow misalignment of the poly-Ala during amyloidization.

The most important findings of this study is that the critical conditions for particular secondary, tertiary, and quaternary structure formation, in particular of stable β-sheet nanocrystals, can be found through systematic variation of the peptide sequence alone, and that the length of the poly-Ala repeat unit is critical in defining identifiable β-sheet nanocrystals. Specifically, there exists a strong scaling effect where a minimum length of poly-Ala repeats is required. Our results also confirm that the glycine-rich regions form semi-extended 3_{10}-helix type structures and not alpha-helix or beta-helix structures, in agreement with experimental NMR studies (Kummerlen et al., 1996a,b). We showed that the poly-Ala region agglomerates under elongation during the spinning process into dominantly anti-parallel β-sheet nanocrystals.

The efficacy of the REMD simulation method can most readily be seen in the final shapes of the β-sheet nanocrystals in our ensemble of cases considered. While the initial lattice structure arranges three layers of strands in the anti-parallel direction, the majority of the final nanocrystals are only two layers in depth. This illustrates the ability of the higher temperature replicas in disrupting initial hydrogen bond networks in order to find structures with lower potential energy. These lower-energy structures are favored in replica exchanges and in turn form the basis of the final lower-temperature structures. While our choice of a three-by-five initial lattice structure is based on intuition of physical conditions, it is limited by current computational constraints. However, we assume that the high-temperature replicas are able to explore conformations beyond those in the neighborhood of a user-defined, unnatural initial structure. While REMD with implicit solvent is useful in approaching native structures, structure predictions after equilibration with explicit solvent are treated as more realistic situations than those with simplified implicit solvent models due to the nature of the derivations of the CHARMM force-fields used here.

4.2. *Conclusion*

The test model ensemble (Fig. 3) shows that the clearly defined crystals are 2–4 nm in length, depending on poly-Ala length, and consistently 3.1–3.4 nm in width (*i.e.* in the side-chain direction), no matter the poly-Ala length. With identical simulation conditions, each test case produces a nanocrystal that self-assembles into a critical width at which hydrogen bonds within the β-sheet gain a strong character through cooperativity (Keten et al., 2010). In addition to being more metabolically expensive to synthesize (Craig et al., 1999), longer nanocrystals may also prohibit certain mechanisms

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

at a higher hierarchical level (Nova et al., 2010). To test the macroscale effects of crystal size and connectivity, the atomistic structure predictions of REMD simulations may be used to train a coarse-grain model of the protein network within the core of a spider silk strand. Such a network would be too large to simulate with atomistic resolution, but the deformation and failure of the network would depend heavily on the shear behavior of the relatively small nanocrystals and the extensible hidden length of the amorphous regions.

Studying the characteristic hierarchies, the coordination of disparate secondary structure features, and the incorporated failure mechanisms in spider silk can have a considerable impact on the design strategy of recombinant silk composites for medical applications. By understanding how the repeat unit sequences dictate the mechanical properties of the final silk, genetic mutation of these sequences may offer customization of recombinant silk behavior. An understanding of the interplay between β-sheet nanocrystals and amorphous networks during failure can also offer insights into the design of silk-like synthetic fiber composites that also employ hydrogen-bond networks at their base hierarchical level (Naraghi et al., 2010).

Acknowledgments

GB acknowledges support from a NDSEG fellowship. Support provided by DOD-MURI and DOD-PECASE is acknowledged. We acknowledge helpful discussions with M. Feig and S. Keten on the general topic of replica exchange molecular dynamics and silk molecular structures.

REFERENCES

Becker, N., Oroudjev, E., et al., 2003. Molecular nanosprings in spider capture-silk threads. Nat. Mater. 2 (4), 278–283.

Bradley, P., Misura, K.M.S., et al., 2005. Toward high-resolution de novo structure prediction for small proteins. Science 309 (5742), 1868–1871.

Brooks, C.L., 1995. Methodological advances in molecular-dynamics simulations of biological-systems. Curr. Opin. Struct. Biol. 5 (2), 211–215.

Brooks, A.E., Steinkraus, H.B., et al., 2005. An investigation of the divergence of major ampullate silk fibers from Nephila clavipes and Argiope aurantia. Biomacromolecules 6 (6), 3095–3099.

Buehler, M.J., Keten, S., et al., 2008. Theoretical and computational hierarchical nanomechanics of protein materials: deformation and fracture. Progr. Mater. Sci. 53 (8), 1101–1241.

Cetinkaya, M., Xiao, S., et al., 2011. Silk fiber mechanics from multiscale force distribution analysis. Biophys. J. 100 (5), 1298–1305.

Craig, C.L., Hsu, M., et al., 1999. A comparison of the composition of silk proteins produced by spiders and insects. Int. J. Biol. Macromol. 24 (2–3), 109–118.

Dicko, C., Vollrath, F., et al., 2004. Spider silk protein refolding is controlled by changing pH. Biomacromolecules 5 (3), 704–710.

Du, N., Liu, X.Y., et al., 2006. Design of superior spider silk: from nanostructure to mechanical properties. Biophys. J. 91 (12), 4528–4535.

Earl, D.J., Deem, M.W., 2005. Parallel tempering: theory, applications, and new perspectives. Phys. Chem. Chem. Phys. 7 (23), 3910–3916.

Eisoldt, L., Hardy, J.G., et al. The role of salt and shear on the storage and assembly of spider silk proteins. J. Struct. Biol., 170 (2) 413–419.

Feig, M., MacKerell, A.D., et al., 2003. Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations. J. Phys. Chem. B 107 (12), 2831–2836.

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins: Struct., Funct., Bioinf. 23 (4), 566–579.

Gatesy, J., Hayashi, C., et al., 2001. Extreme diversity, conservation, and convergence of spider silk fibroin sequences. Science 291 (5513), 2603–2605.

Gosline, J., Guerette, P., et al., 1999. The mechanical design of spider silks: from fibroin sequence to mechanical function. J. Exp. Biol. 202 (23), 3295–3303.

Guerette, P.A., Ginzinger, D.G., et al., 1996. Silk properties determined by gland-specific expression of a spider fibroin gene family. Science 272 (5258), 112–115.

Hayashi, C.Y., Lewis, R.V., 1998. Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J. Mol. Biol. 275 (5), 773–784.

Hayashi, C.Y., Shipley, N.H., et al., 1999. Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. Int. J. Biol. Macromol. 24 (2–3), 271–275.

Hinman, M.B., Lewis, R.V., 1992. Isolation of a clone encoding a 2nd dragline silk fibroin-nephila-clavipes dragline silk is a 2-protein fiber. J. Biol. Chem. 267 (27), 19320–19324.

Holland, G.P., Creager, M.S., et al., 2008. Determining secondary structure in spider dragline silk by carbon–carbon correlation solid-state NMR spectroscopy. J. Am. Chem. Soc. 130 (30), 9871–9877.

Humphrey, W., Dalke, A., et al., 1996. VMD—visual molecular dynamics. J. Mol. Graphics 14, 33–38.

Jenkins, J.E., Creager, M.S., et al., 2010. Quantitative correlation between the protein primary sequences and secondary structures in spider dragline silks. Biomacromolecules 11 (1), 192–200.

Keten, S., Buehler, M.J., 2010. Atomistic model of the spider silk nanostructure. Appl. Phys. Lett.

Keten, S., Xu, Z., et al., 2010. Nanoconfinement controls stiffness, strength and mechanical toughness of beta-sheet crystals in silk. Nat. Mater. 9, 359–367.

Kummerlen, J., van Beek, J.D., et al., 1996a. Local structure in spider dragline silk investigated by two-dimensional spin-diffusion nuclear magnetic resonance. Macromolecules 29 (8), 2920–2928.

Kummerlen, J., van Beek, J.D., et al., 1996b. Local structure in spider dragline silk investigated by two-dimensional spin-diffusion nuclear magnetic resonance. Macromolecules 29 (8), 2920–2928.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1), 105–132.

Lefevre, T., Rousseau, M.E., et al., 2007. Protein secondary structure and orientation in silk as revealed by Raman spectromicroscopy. Biophys. J. 92 (8), 2885–2895.

Lifson, S., Sander, C., 1979. Antiparallel and parallel beta-strands differ in amino acid residue preferences. Nature 282 (5734), 109–111.

Ma, B.Y., Nussinov, R., 2006. Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. Curr. Opin. Chem. Biol. 10 (5), 445–452.

Miyashita, N., Straub, J.E., et al., 2009. Transmembrane structures of amyloid precursor protein dimer predicted by replica-exchange molecular dynamics simulations. J. Am. Chem. Soc. 131 (10), 3438–3439.

Naraghi, M., Filleter, T., et al., 2010. A multiscale study of high performance double-walled nanotube-polymer fibers. ACS Nano 4 (11), 6463–6476.

Nova, A., Keten, S., et al., 2010. Molecular and nanostructural

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012

mechanisms of deformation, strength and toughness of spider silk fibrils. Nano Lett. 10 (7), 2626–2634.

Philip, M.C., Stephen, A.F., et al., 1994. Mechanical and thermal properties of dragline silk from the spider *I*Nephila clavipes/*I*. Polym. Adv. Technol. 5 (8), 401–410.

Porter, D., Vollrath, F., 2008. The role of kinetics of water and amide bonding in protein stability. Soft Matter 4 (2), 328–336.

Porter, D., Vollrath, F., et al., 2005. Predicting the mechanical properties of spider silk as a model nanostructured polymer. Eur. Phys. J. E 16 (2), 199–206.

Rammensee, S., Slotta, U., et al., 2008. Assembly mechanism of recombinant spider silk proteins. Proc. Natl. Acad. Sci. USA 105 (18), 6590–6595.

Rao, F., Caflisch, A., 2003. Replica exchange molecular dynamics simulations of reversible folding. J. Chem. Phys. 119 (7), 4035–4042.

Rhee, Y.M., Pande, V.S., 2003. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. Biophys. J. 84 (2), 775–786.

Riekel, C., Vollrath, F., 2001. Spider silk fibre extrusion: combined wide- and small-angle X-ray microdiffraction experiments. Int. J. Biol. Macromol. 29 (3), 203–210.

Rousseau, M.E., Lefevre, T., et al., 2004. Study of protein conformation and orientation in silkworm and spider silk fibers using Raman microspectroscopy. Biomacromolecules 5 (6), 2247–2257.

Rousseau, M.E., Lefevre, T., et al., 2009. Conformation and orientation of proteins in various types of silk fibers produced by nephila clavipes spiders. Biomacromolecules 10 (10), 2945–2953.

Sanbonmatsu, K.Y., Garcia, A.E., 2002. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins: Struct., Funct., Genet. 46 (2), 225–234.

Shao, Z.Z., Vollrath, F., 2002. Materials: surprising strength of silkworm silk. Nature 418 (6899), 741–741.

Simmons, A.H., Michal, C.A., et al., 1996. Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. Science 271 (5245), 84–87.

Sponner, A., Schlott, B., et al., 2005. Characterization of the protein components of Nephila clavipes dragline silk. Biochemistry 44 (12), 4727–4736.

Sugita, Y., Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 314 (1–2), 141–151.

Termonia, Y., 1994. Molecular modeling of spider silk elasticity. Macromolecules 27 (25), 7378–7381.

Thiel, B.L., Guess, K.B., et al., 1997. Non-periodic lattice crystals in the hierarchical microstructure of spider (major ampullate) silk. Biopolymers 41 (7), 703–719.

Trancik, J.E., Czernuszka, J.T., et al., 2006. Nanostructural features of a spider dragline silk as revealed by electron and X-ray diffraction studies. Polymer 47 (15), 5633–5642.

van Beek, J.D., Hess, S., et al., 2002. The molecular structure of spider dragline silk: folding and orientation of the protein backbone. Proc. Natl. Acad. Sci. USA 99 (16), 10266–10271.

van Beek, J.D., Kummerlen, J., et al., 1999. Supercontracted spider dragline silk: a solid-state NMR study of the local structure. Int. J. Biol. Macromol. 24 (2–3), 173–178.

Vollrath, F., Knight, D.P., 2001. Liquid crystalline spinning of spider silk. Nature 410 (6828), 541–548.

Wu, X., Liu, X.-Y., et al., 2009. Unraveled mechanism in silk engineering: fast reeling induced silk toughening. Appl. Phys. Lett. 95 (9), 093703-3.

Zhang, Y., 2008. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18 (3), 342–348.

Please cite this article in press as: Bratzel, G., Buehler, M.J., Sequence-structure correlations in silk: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized at a critical length scale. Journal of the Mechanical Behavior of Biomedical Materials (2011), doi:10.1016/j.jmbbm.2011.07.012