Self-association of a highly charged arginine-rich cell-penetrating peptide

Giulio Tesei,a,*, Mario Vazdarb, Malene Ringkjöbing Jensenb, Carolina Cagnellb, Phil E. Masonc, Jan Heydab, Marie Skepöa, Pavel Jungwirthb,c, and Mikael Lundd,a

aDivision of Theoretical Chemistry, Lund University, POB 124, SE-22100 Lund, Sweden; bDivision of Organic Chemistry and Biochemistry, Rudjer Bošković Institute, POB 180, HR-10002 Zagreb, Croatia; cUniv. Grenoble Alpes, CNRS, CEA, CNRS, IBIS, F-38000 Grenoble; dInstitute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Flemingovo nám. 2, 16610 Prague 6, Czech Republic; cDepartment of Physical Chemistry, University of Chemistry and Technology, Technická 5, 16628 Prague 6, Czech Republic

This manuscript was compiled on July 7, 2017

Small angle X-ray scattering (SAXS) measurements reveal a striking difference in intermolecular interactions between two short highly charged peptides - deca-arginine (R10) and deca-lysine (K10). Comparison of SAXS curves at high and low salt concentration shows that R10 self-associates, while interactions between K10 chains are purely repulsive. The self-association of R10 is stronger at lower ionic strengths indicating that the attraction between R10 molecules has an important electrostatic component. SAXS data is complemented by NMR measurements and potentials of mean force between the peptides, calculated by means of umbrella sampling molecular dynamics (MD) simulations. All-atom MD simulations elucidate the origin of the R10-R10 attraction by providing structural information on the dimeric state. The last two C-terminal residues of R10 constitute an adhesive patch formed by stacking of the side chains of two arginine residues, and by salt bridges formed between the like-charge ion pair and the C-terminal carboxyl groups. A statistical analysis of the protein data bank reveals that this mode of interaction is a common feature in proteins.

Investigation of deca-arginine (R10) and deca-lysine (K10), in high and low ionic strength solutions conducted using small angle X-ray scattering (SAXS) measurements, all-atom molecular dynamics (MD) simulations, and 1H-13C heteronuclear single quantum coherence (HSQC) NMR measurements. The choice of the peptides is motivated by the biological interest in the high cellular uptake of arginine-rich peptides (RRPs), which has been the subject of several comparative studies. Oligo-arginine chains of six to fifteen amino acids readily translocate across cell membranes, while the translocation efficiency of oligo-lysines of equal length is considerably lower. If a certain threshold peptide concentration is exceeded, cell penetration of RRPs can occur in a non-endocytotic mode. Although the molecular details of this transduction mechanism still need to be clarified, an important feature has been identified as the Gdm+ moiety of the arginine side chain, which can form bidentate hydrogen bonds with phosphate and glycerol groups of the lipid molecules in the cell membrane. These interactions may promote the adsorption of RRPs on the membrane surface and perturb the packing of lipids in the bilayer, as well as significantly increase the lifetime of transient membrane pores.

Recent studies focusing on interactions of charged proteins in electrolyte solutions have highlighted the interplay of two counteracting electrostatic forces. The first originates from the presence of a localized distribution of charges defining an electrostatic patch on the protein surface. Depending on relative orientations, the charge distributions in the patches on the protein molecules can become complementary, thereby leading to an attractive electrostatic force. This anisotropic force is short-ranged, and is hereafter referred to as the electrostatic adhesive force. The other force is the double layer force arising from the Coulombic repulsion between like-charged molecules in the electrolyte medium. Both electrostatic adhesive and double layer forces are weakened by the presence of salt in the solution. As a non-trivial consequence, the propensity of the proteins to aggregate is heightened at low-to-intermediate ionic strengths. This owes to lowering of Coulombic repulsion due to salt screening of the net charge of the protein; in conjunction with the presence of the adhesive force, which operates at shorter distances, and is therefore less efficiently screened.

In this work, the competition between electrostatic adhesive and double layer forces, together with a chemically specific like-charge attraction between guanidinium (Gdm+) side chains, is reported for solutions of a small highly charged peptide. Observation of the complex aggregation behavior for a relatively simple molecule is substantiated by a comparative investigation of deca-arginine (R10) and deca-lysine (K10), in high and low ionic strength solutions conducted using small angle X-ray scattering (SAXS) measurements, all-atom molecular dynamics (MD) simulations, and 1H-13C heteronuclear single quantum coherence (HSQC) NMR measurements.

In the following sections, we first present experimental findings based on SAXS measurements on solutions of R10 and K10, as well as of R8KR and K8RK molecules. The latter two molecules are mutants of R10 and K10, respectively, resulting from a ninth residue interchange. Based on the SAXS results, we infer that an attraction, which is mainly of electrostatic origin, is present between R10 molecules, while the other molecules readily repel each other. Moreover, SAXS data for R10 at low ionic strength is consistent with the presence of dimers or oligomeric forms in solution. Subsequently, we report HSQC NMR data and potentials of mean force (PMFs) obtained from all-atom MD simulations, which support the interpretation of the scattering data and provide an atomistic description of R10 dimers. Finally, we analyze the occurrence of the inferred mode of interaction in cell-penetrating peptides | self-association | SAXS | NMR | MD simulations.

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

PNAS | July 7, 2017 | vol. XXX | no. XX | 1–11
biological systems by inspection of protein crystal structures in the protein data bank (PDB).

Results

Scattering Intensity Curves. Fig. 1 shows SAXS measurements on R10 and K10 concentration series at increasing ionic strength, $c_s$, at 293 K and pH 7.8. The scattering intensities $I(q)$, normalized by the peptide concentration, $c_p$, are reported as a function of the scattering vector, $q$.

At $c_s = 0.060$ M and $c_s = 0.150$ M, the crowding effect on the scattering profiles of the two peptides is strikingly different. At low $q$ values, $I(q)/c_p$ decreases with increasing $c_p$ for K10, while it increases for R10. Scattering intensity curves obtained for K10 are as expected for polyelectrolyte solutions of low ionic strength [16, 17], displaying an increasingly pronounced maximum that shifts to higher $q$ with increasing $c_p$. Maxima are also observed in $I(q)/c_p$ for R10, however, with increasing $c_p$ the shift of the maxima to higher $q$ is less pronounced than for K10. At $c_s = 0.300$ M, $I(q)/c_p$ at low $q$ decreases with increasing $c_p$ for K10, while an oscillating trend is observed for R10. Fig. 1d shows $I(q)/c_p$ values extrapolated to $q = 0$ (SI Materials and Methods) for solutions of R10, K10, R8KR, and K8RK for increasing $c_p$ at $c_s = 0.150$ M. While for K10 and K8RK $I(0)/c_p$ values are similar to each other (as expected), the discrepancy in $I(0)/c_p$ values between R10 and R8KR is large compared to the small difference in molecular weights. Finally, $I(0)/c_p$ vs. $c_p$ has the same decreasing trend for K10, R8KR, and K8RK, while it has an increasing trend for R10 (SI Results, Fig. S6).

Potentials of Mean Force. Fig. 2 shows the PMFs obtained from umbrella sampling MD simulations for pairs of R10 and K10, as a function of the intermolecular separation. The intermolecular separation for R10 corresponds to the distance between the guanidino-C atoms of the ninth residues. This reaction coordinate was chosen based on inspection of the R10 dimer observed in previous MD simulations [8]. Analogously, the intermolecular separation for K10 is calculated as the distance between the ε-C atoms of the ninth residues. At $c_s = 0.01$ M, the free energy of interaction between pairs of K10 molecules is repulsive, and decays with increasing interparticle separation, as expected for two like-charged molecules in solution. In contrast, the PMF for R10 chains shows at low ionic strength a minimum at 0.4 nm, followed by two maxima at separations of 0.58 nm and 0.93 nm. As evidenced by the red and blue lines in Fig. 2, the positions of the minima is preserved at higher $c_s$. At $c_s = 0.07$ M, PMFs of both R10 and K10 show lower free energy values than at $c_s = 0$ M, corresponding to a decreased repulsion between the like-charged molecules due to electrostatic screening. At $c_s = 0.32$ M, the decay of the repulsive interaction with increasing intermolecular separation is considerably steeper than for the PMFs at lower $c_s$, nonetheless the free energy of the minimum at 0.4 nm separation between R10 molecules is higher compared to $c_s = 0.07$ M. At large separations, the PMFs agree with the Debye-Hückel approximation, as shown by the black points in Fig. 2. The vertical error bars reflect the fluctuation of separations between the center of mass of the peptides in the umbrella sampling simulation windows. While K10 molecules repel each other at all length scales and salt concentrations, R10 displays salt-dependent attraction at short separation. The free energy of the minimum at 0.4 nm varies non-monotonically with $c_s$, and it is the lowest at $c_s = 0.07$ M. On the contrary, the difference in free energy between the maximum at 0.58 nm and the minimum at 0.4 nm decreases with increasing $c_p$. To facilitate further discussion of the non-monotonic trend for the free energy minimum, Fig. 2 displays two-dimensional schematic representations of the R10 dimeric structures.

Hydrogen Bonding in Dimeric Structures. Fig. 3 displays the probability to form an H-bond between the C-terminal carboxyl group (COO⁻) of the first peptide and the last five

Fig. 1. Concentration normalized scattering intensities for deca-arginine (R10, full lines) and deca-lysine (K10, dashed lines) at various peptide concentrations, $c_p$, in 0.020 M Tris buffer solutions of 0.060 M (a), 0.150 M (b) and 0.300 M (c) ionic strength, $c_s$. Extrapolated $I(0)/c_p$ values for samples of R10, K10, R8KR and K8RK of increasing $c_p$ at $c_s = 0.150$ M (d).

Significance Statement

Arginine-rich cell-penetrating peptides are promising candidates for intracellular drug delivery. These cationic peptides spontaneously traverse biological membranes via a direct mode of entry which is not yet fully understood. In this study, we report the complex solution behavior of the cell-penetrating peptide deca-arginine. In spite of its large net positive charge, deca-arginine self-associates at low-to-intermediate ionic strengths, displaying an increasingly pronounced maximum that shifts to higher $q$ with increasing $c_p$. Maxima are also observed in $I(q)/c_p$ for R10, however, with increasing $c_p$ the shift of the maxima to higher $q$ is less pronounced than for K10. At $c_s = 0.300$ M, $I(q)/c_p$ at low $q$ decreases with increasing $c_p$ for K10, while an oscillating trend is observed for R10. Fig. 1d shows $I(q)/c_p$ values extrapolated to $q = 0$ (SI Materials and Methods) for solutions of R10, K10, R8KR, and K8RK for increasing $c_p$ at $c_s = 0.150$ M. While for K10 and K8RK $I(0)/c_p$ values are similar to each other (as expected), the discrepancy in $I(0)/c_p$ values between R10 and R8KR is large compared to the small difference in molecular weights. Finally, $I(0)/c_p$ vs. $c_p$ has the same decreasing trend for K10, R8KR, and K8RK, while it has an increasing trend for R10 (SI Results, Fig. S6).

Potentials of Mean Force. Fig. 2 shows the PMFs obtained from umbrella sampling MD simulations for pairs of R10 and K10, as a function of the intermolecular separation. The intermolecular separation for R10 corresponds to the distance between the guanidino-C atoms of the ninth residues. This reaction coordinate was chosen based on inspection of the R10 dimer observed in previous MD simulations [8]. Analogously, the intermolecular separation for K10 is calculated as the distance between the ε-C atoms of the ninth residues. At $c_s = 0.01$ M, the free energy of interaction between pairs of K10 molecules is repulsive, and decays with increasing interparticle separation, as expected for two like-charged molecules in solution. In contrast, the PMF for R10 chains shows at low ionic strength a minimum at 0.4 nm, followed by two maxima at separations of 0.58 nm and 0.93 nm. As evidenced by the red and blue lines in Fig. 2, the positions of the minima is preserved at higher $c_s$. At $c_s = 0.07$ M, PMFs of both R10 and K10 show lower free energy values than at $c_s = 0$ M, corresponding to a decreased repulsion between the like-charged molecules due to electrostatic screening. At $c_s = 0.32$ M, the decay of the repulsive interaction with increasing intermolecular separation is considerably steeper than for the PMFs at lower $c_s$, nonetheless the free energy of the minimum at 0.4 nm separation between R10 molecules is higher compared to $c_s = 0.07$ M. At large separations, the PMFs agree with the Debye-Hückel approximation, as shown by the black points in Fig. 2. The vertical error bars reflect the fluctuation of separations between the center of mass of the peptides in the umbrella sampling simulation windows. While K10 molecules repel each other at all length scales and salt concentrations, R10 displays salt-dependent attraction at short separation. The free energy of the minimum at 0.4 nm varies non-monotonically with $c_s$, and it is the lowest at $c_s = 0.07$ M. On the contrary, the difference in free energy between the maximum at 0.58 nm and the minimum at 0.4 nm decreases with increasing $c_p$. To facilitate further discussion of the non-monotonic trend for the free energy minimum, Fig. 2 displays two-dimensional schematic representations of the R10 dimeric structures.

Hydrogen Bonding in Dimeric Structures. Fig. 3 displays the probability to form an H-bond between the C-terminal carboxyl group (COO⁻) of the first peptide and the last five
C-terminal residues of the second peptide. The H-bond probabilities are calculated from umbrella sampling MD simulation windows, where the model peptides are at the closest separation along the reaction coordinate. It is evident for both R10 and K10 that the C-terminal COO\(^-\) of the first peptide is most likely to form H-bonds with the ninth residue of the second peptide. A non-monotonic dependence on \(c_p\) is observed for the H-bond probability between the C-terminal COO\(^-\) and the ninth residue. Specifically for R10 it increases approximately twofold from \(c_p = 0.01\) M to \(c_p = 0.07\) M, while at \(c_p = 0.32\) M it drops back to an intermediate value. The H-bond probability between the COO\(^-\) group and the ninth residue is significantly larger for R10 than for K10. Representative snapshots from the analyzed trajectories are included in Fig. 3 to show H-bonding pattern in the C-terminal residues of R10 and K10. In the R10 dimer, H-bonds are formed between the C-terminal COO\(^-\) and the Gdm\(^+\) moiety of the arginine side chains and are likely to be bidentate; whereas, in the K10 dimer, monodentate H-bonds are formed between the C-terminal COO\(^-\) and the amino group of the lysine side chains.

**1H-13C HSQC NMR Spectra.** Fig. 4 shows the regions of 2D 1H-13C HSQC NMR spectra of R10 and R8KR corresponding to correlations between Cα and H\(\alpha\) atoms. Peptide solutions have \(c_p = 0.025\) M, pH 5, and are measured at 293 K (SI Results, Fig. S7 for corresponding SAXS curves). With increasing \(c_p\), R8KR spectra superpose almost perfectly, indicating the absence of significant peptide-peptide attractions for R8KR. For R10 small changes in chemical shifts are observed for both N-terminal and C-terminal residues. As \(cp/I(0)\) is proportional to the derivative of the osmotic pressure, \(\Pi\), with respect to \(c_p\), Fig. 4c,d help to interpret the NMR data. At the lowest \(c_p\), R10 solutions have significantly lower \((\partial\Pi/\partial c_p)T\) at \(c_p = 0.025\) M than at \(c_p = 0.300\) M, while dilute R8KR solutions display a slighter and opposite trend for \((\partial\Pi/\partial c_p)T\) vs. \(c_p\). This suggests that at \(c_p = 0.025\) M dimers may be present in the dilute \(c_p = 1\) mM R10 solution, which also displays a different chemical shift for the N-terminal residue compared to the R8KR solution of \(c_p = 5\) mM. The weak dependence of \(cp/I(0)\) on \(c_p\) observed for R10 at \(c_p = 0.025\) M might owe to the repulsive nature of monomer-dimer and dimer-dimer interactions, as well as to a slight increase in dimer population. The latter is also reflected in the small changes in chemical shifts observed for R10 in the explored \(c_p\)-range (Fig. 4a).

**Occurrence of the Observed Mode of Interaction in Proteins.** We searched a selection of 10,388 entries of the PDB (SI Materials and Methods) for the observed mode of interaction in R10 dimers, consisting of H-bonding of two C-terminal COO\(^-\) groups with a Gdm\(^+\)-Gdm\(^+\) ion-pair. It was found to be present in 231 of the 1,697 protein crystal structures featuring at least one pair of stacked arginine residues (complete list in SI Results). The analysis focused on the COO\(^-\) groups of aspartate and glutamate side chains. However, in two proteins (PDB entries 5INJ and 4AZS), the Gdm\(^+\)-Gdm\(^+\) pair is in the active site and interacts with the COO\(^-\) of the substrate. These enzymes are prenyltransferase PrlB [18] and bifunctional methyl- transferase/kinase WbdD [19], while the substrates are tryptophan and S-adenosylmethionine, respectively. Fig. 5 shows three characteristic structures which exemplify the geometries of the interacting Gdm\(^+\) and COO\(^-\) moieties, as found in the crystal structures. In the most frequent arrangement I, each COO\(^-\) is H-bonded with a different Gdm\(^+\) moiety. In arrangement II, only one Gdm\(^+\) moiety is involved in the interaction with two COO\(^-\) groups. Finally, in the least occurring arrangement III, one of the COO\(^-\) groups is H-bonded with one Gdm\(^+\), whereas the second COO\(^-\) is approximately perpendicular to the molecular planes of the two Gdm\(^+\) and forms H-bonds with both Gdm\(^+\) moieties in the ion pair. The residues involved in the interactions belong to different protein chains in 45.4% of the occurrences, while in the remaining 54.6% the interaction is between residues of the same chain. The analysis further showed that geometrical arrangements involving like-charge ion pairing of lysine side chains, stabilized by salt bridges with COO\(^-\) groups of
aspartate and glutamate, are considerably less frequent than the mode of interaction characterized by arginine stacking (SI Results).

Discussion

In the SAXS curves of K10 at low $c_s$ (dashed lines in Fig. 1) we observe a typical polyelectrolyte behavior - the intermolecular interactions are repulsive, causing $I(q)/c_p$ in the low-$q$ range to decrease with increasing $c_p$, owing to the drop in osmotic compressibility [16]. The high-$q$ region is dominated by intra-molecular scattering and is thus $c_p$-independent. The maxima of $I(q)/c_p$ observed at $q = q_{max}$ result from intermolecular distance correlations [17]; the inverse of $q_{max}$ is related to the mean distance between nearest neighbors in solution. SAXS curves collected for R10 samples at $c_s = 0.060$ M and $c_s = 0.150$ M (solid lines in Fig. 1a,b) reveal a striking difference in intermolecular interactions compared to the case of K10. The increase of $I(q)/c_p$ in the low-$q$ range with increasing $c_p$ corresponds to an increase in osmotic compressibility, indicating a net attractive interaction between the R10 peptides. That is, in R10 solutions at intermediate $c_s$ an attraction dominates over the repulsive double layer force.

For solutions of $c_s = 0.300$ M (Fig. 1c), where electrostatic interactions are effectively screened, scattering intensities for K10 at increasing $c_p$ suggest an excluded-volume effect, i.e. an effective repulsion caused by the loss of entropy in the more crowded solutions. For R10 at $c_s = 0.300$ M, the effect of $c_p$ on the scattering intensities shows a nonmonotonic trend in $I(q)/c_p$ at small angles, indicating that the attractive interaction competes with the repulsive excluded-volume effect. Comparing SAXS curves for R10 at low and high $c_s$, it can be inferred that the attractive interaction has an important electrostatic component. The electrostatic adhesive force is strong at low $c_s$, where it outpowers the electrostatic repulsion between the highly charged peptides. In contrast, a much weaker attraction is required to compensate for the excluded-volume effect. In Fig. 1d, the decreasing trends of $I(0)/c_p$ values for K10, K8RK, and R8KR solutions show that the net interaction between these molecules is repulsive, whereas the large and increasing $I(0)/c_p$ values for R10 solutions of increasing $c_p$ indicate that at $c_s = 0.150$ M a fraction of R10 molecules self-associates into dimers. SAXS data provides evidence that the ninth arginine residue is essential for the attractive interaction, nonetheless, as shown by the scattering curves for K8RK molecules (SI Results, Fig. S6), the remaining arginine residues have a role in lowering the electrostatic repulsion.

The PMFs (Fig. 2) show that K10 molecules repel each other at all length scales and salt concentrations, while R10 molecules, despite bearing the same net charge, display a salt concentration-dependent attraction at short separation. This result is qualitatively insensitive to the choice of the force field (SI Results, Fig. S8). A minimum, at 0.4 nm separation, and an adjacent maximum, at 0.58 nm separation, are present in the PMFs of R10 at all $c_s$, as indicated by the blue and red vertical lines in Fig. 2. The free energy difference between the maximum and the minimum serves as an estimate of the strength of the adhesive force. The fact that the free energy difference diminishes with increasing $c_s$ confirms that the adhesive force has an important electrostatic component. The stability of the dimeric species is indicated by the free energy values of the minima in the PMFs of R10 and depends on the interplay between the double layer and the electrostatic adhesive force. From our simulation results, it can be inferred that the balance between the two forces, which is modulated by the ionic strength, favors the attraction at the intermediate $c_s = 0.07$ M, where the R10 dimer corresponds to the lowest free energy value. This result can be explained by a simple electrostatic argument, schematically illustrated in Fig. 2. The color scheme exemplifies that with increasing $c_s$ the repulsion between the outer positively charged sites is screened more efficiently than the attraction of oppositely charged sites in the binding region between the two peptides. As a result of the different efficiency by which the short-range attraction and the long-range repulsion are screened with increasing $c_s$, the net Debye-Hückel free energy of the dimer shows a nonmonotonic trend - the screening majorly affects like-charged side chains that are farther apart until $c_s$ is so large that even
the short-range attractions in the binding region are effectively screened. Competition between double layer and electrostatic adhesive forces has been observed for other biomolecules e.g. a monoclonal antibody mAb1 [1], and a globular milk protein lactoferrin [2]. In both cases, the attraction stems from a charged patch on the protein surface and becomes dominant at a critical salt concentration. MD simulations provide a detailed picture of the origin of the attraction between R10 peptides. The last two terminal residues of R10 can be identified as an adhesive patch displaying two oppositely charge sites. The positive site is the Gdm$^+$ moiety of the ninth arginine residue while the negative site is the C-terminal carboxyl group (COO$^-\$). We suggest that the observed attraction between R10 molecules occurs through complementarity of the charged groups in the adhesive patch of two interacting peptides. When two patches are at close separation, the Gdm$^+$ moieties of the ninth residues of the peptides form an ion pair. The divalent charge site, generated by the stacking of the arginine side chains, is stabilized by two intermolecular hydrogen bonds. These are salt bridges formed between the negatively charged C-terminal COO$^-$ and the positively charged Gdm$^+$ of the ninth residues. Theoretical and experimental studies suggest that Gdm$^+$ ions form weakly stable like-charge ion pairs in aqueous medium [20–27]. The free energy of interaction between two Gdm$^+$ ions is minimized when stacked parallel to each other in staggered geometry with carbon atoms separation between 0.35 nm and 0.46 nm [27]. The weak Gdm$^-$-Gdm$^+$ attraction in water is estimated to be around -2 kJ mol$^{-1}$ and is due to a combination of quadrupole-quadrupole interaction, dispersive forces, and the hydrophobic effect [27]. The hydrophilicity of the surfaces of Gdm$^+$, in conjunction with the ability of Gdm$^+$ to form H-bonds with functional groups in its molecular plane, has recently been confirmed from the analysis of interactions of arginine side chains in proteins in the PDB [28]. Our analysis of H-bonding involving C-termini of R10 or K10 (Fig. 3) shows that the probability of forming H-bonds between the ninth residues at close separation and COO$^-$ is remarkably higher in R10 than in K10. Since salt bridges involving COO$^-$ and ammonium (NH$_2^+$) or Gdm$^+$ moieties have comparable strengths [29], the higher probability observed for R10 highlights the important role of the arginine side chain in the R10-R10 patchy attraction. Due to the approximately tetrahedral geometry of NH$_2^+$, restraining the position of two lysine side chains at close separation hinders the formation of H-bonds with COO$^-\$. In contrast, the H-bonding of Gdm$^+$ occurs in the molecular plane and is unaffected by stacking. The probability for the formation of salt bridges involving the ninth residues of R10 and K10 has a non-monotonic dependence on $c_p$. The formation of salt bridges is the consequence of short-range attractive electrostatic interactions. Therefore, the trend can be rationalized by the same argument employed to explain the $c_p$-dependence of the free energy minima in the PMFs of R10. Our results highlight the fact that an important component of the described patchy attraction is of electrostatic origin and is imparted by two salt bridges. The main difference in the interaction between R10 e.g. K10 lies in the favorable Gdm$^+$-Gdm$^+$ stacking. As shown by the H-bond probabilities calculated for K10, in the absence of pairing of the positively charged side chains, C-terminal salt bridges are substantially less energetically favorable. The NMR data confirms that R8KR solution are monodisperse in the explored $c_p$ range, while R10 molecules self-associate. The observed changes in chemical shifts in Fig. 4a are consistent with a non-specific peptide-peptide interaction characterized by the stacking of side chains of several arginine residues of the R10 peptide. Accordingly, MD simulations indicate that, besides the ninth residue, the eighth and tenth residues also contribute to the attractive interaction (SI Results, Fig. S10), and that the stacking of Gdm$^+$ moieties occurring between the remaining residues can be stabilized by chloride ions [30] (SI Results, Fig. S13). For the N-terminal residues, the changes in chemical shift with increasing $c_p$ may be explainable by the favorable interaction of the N-terminus of one peptide with the binding region of the R10 dimer. Indeed, when the side chains of the tenth and eighth residues are constrained at stacking separations, the PMF as a function of distance between guanidino-C atoms of tenth and second residues of R10 shows two local minima at close separations. (SI Results, Fig. S11). The mode of interaction responsible for the adhesive force between R10 molecules is well represented in the PDB (Fig. 5). 231 X-ray crystal structures out of 10,388 entries present at least one ion pair formed by the stacking of two arginine side chains, which interacts with two carboxyl groups of aspartate or glutamate residues. Arginine residues are also found at protein-protein interfaces more frequently than lysine residues, as reflected in the ‘stickiness’ scale proposed by Levy and coworkers [31]. The concentration effect observed in the SAXS data from the dilution series of R10 at low-to-intermediate $c_p$ (Fig. 1, and SI Results, Fig. S6a,c) is consistent with a monomer-oligomer equilibrium where the population of oligomeric forms in solution increases with increasing $c_p$. This is further supported by concentration-dependent NMR $^1$H and $^{13}$C chemical shifts (Fig. 4), assuming fast exchange on the NMR chemical shift time scale between monomeric and oligomeric forms. Self-aggregation has been recently shown to be important for the effective translocation of a 9-amino acid-long peptide across the plasma membrane [32]. As a consequence, the propensity of R10 molecules to aggregate may contribute to the explanation of the high cellular uptake of RRPs. The transduction efficiency of oligo-arginines depends on the number of residues. It is maximal for chains of six to fifteen amino acids, but considerably lower for shorter as well as longer peptides [6, 11]. As the number of positively charged residues increases, the force balance governing the self-association of oligo-arginines gradually moves towards the repulsive double layer force. The presence of an upper limit in the range of optimal chain lengths may reflect the importance of self-association in the transduction of oligo-arginines. Conclusions We have employed a combination of MD simulations with SAXS and NMR experiments with the aim of elucidating the molecular mechanism of self-association of arginine-rich oligopeptides. Concentration studies point to an important electrostatic component of the attraction between R10 molecules, while single-point mutations underline a binding motif involving the C-terminus and the adjacent arginine.
residues. The present results not only support the notion of the important role of self-aggregation in the transduction of cell-penetrating peptides, but also open path to future studies of the potential biological roles of the newly discovered binding motif, as exemplified by its abundant occurrence in the PDB.

ACKNOWLEDGMENTS. The study was supported by the Science Faculty project grant program for research with neutrons and synchrotron light. The authors thank the Swedish Research Council; the Swedish Foundation for Strategic Research; LUNARC in Lund for computational resources; the European Synchrotron Radiation Facility (ESRF) for providing beamtime. PJ acknowledges the Czech Science Foundation for support via grant no. 16-01074S. We are grateful to Dr. Petra Pernot and Dr. Martha Brennich at ESRF, Grenoble for providing assistance in using beamline BM29 and to Ellen Rieffel for experimental assistance. We provide a Jupiter Notebook for reproducing the analyses of SAXS data and MD simulations, as well as all presented plots.

SI Results.

**Scattering Intensity curves.** Fig. S6a,b present SAXS measurements on R10, R8KR and K8RK concentration series for \( c_s = 0.025 \) M and \( c_s = 0.150 \) M at 293 K and pH 7.8. At low \( q \) values, with increasing \( c_p \), \( I(q)/c_p \) decreases for both R8KR and K8KR, while it increases for R10. Fig. S6c shows the effect of \( c_s \) on the scattering curves of R10 at \( c_p \approx 7 \) g/l: for \( q < 2 \) nm\(^{-1}\), \( I(q)/c_p \) increases with increasing \( c_s \) indicating that self-association is heightened at low-to-intermediate ionic strength. Fig. S1d shows \( I(q)/c_p \) values extrapolated to \( q = 0 \) for solutions of R10, K10, R8KR and K8RK of increasing \( c_p \) at \( c_s = 0.025 \) M. As observed at \( c_s = 0.150 \), \( I(0)/c_p \) values for R10 are much larger than expected based on the molecular weight, indicating oligomerization. Fig. S7a,b show SAXS curves for R10 and R8KR molecules in 0.035 M acetate buffer pH 5, i.e. at the same solution conditions of the \( ^{1}H-^{13}C \) HSQC NMR measurements. Comparison with Fig. S6a shows that scattering curves for R10 and R8KR are very similar in 0.020 M Tris buffer pH 7.8 and in 0.035 M acetate buffer pH 5. Interestingly, R10 solutions in acetate buffer yield higher \( c_p/I(0) \) values than in Tris buffer, suggesting that the acetate co-solute may compete with the C-terminal carboxyl groups and hamper the dimerization of R10.

**Force Field Dependence of the PMFs.** The preference of the AMBER ff03 force field over the OPLS/AA or CHARMM22 force fields, used in previous comparative simulation studies on homopolymers of arginine and lysine [8, 9], is motivated by two reasons. First, the correction proposed by Best and coworkers [33] has been shown to be able to reproduce the conformational properties of small intrinsically disordered proteins, whereas widely used force fields developed for globular proteins generate significantly less extended structures [33, 34]. Second, the OPLS/AA and CHARMM22 force fields, as opposed to the AMBER ff03 force field, have been shown to overestimate the probability of arginine/aspartate association [29].

**Additional interaction modes between R10 molecules.** Side chain stacking of the ninth arginine residues is necessary for the attractive interaction. On top of it, as shown by the scattering curves for K8RK molecules, the other arginine residues have a fundamental role in lowering the electrostatic repulsion. Moreover, the lower propensity to self-associate of R4, compared to R10, might be related to the lower bioavailability observed experimentally.

**PMF between tetra-arginine molecules.** Figure S8 shows the PMF between two tetra-arginine (R4) molecules at \( c_s = 0.07 \) M as a function of separation between guanidino-C atoms of the third residue. The free energy profile has a considerably shallower local minimum compared to the PMF for R10 at similar \( c_s \). This result suggests that the arginine side chains contribute in an additive fashion to lower the electrostatic repulsion. Moreover, the lower propensity to self-associate of R4, compared to R10, might be related to the lower bioavailability observed experimentally.
Fig. 8. Potentials of mean force (PMFs) for pairs of R10 molecules as a function of the separation between the guanidino-C atoms of the ninth residues. PMFs are from umbrella sampling MD simulations, using the Amber ff99SB-ILDN (black solid line) and AMBER f03WS (red line) force fields, at \( c_s = 0.01 \) M (a), \( c_s = 0.07 \) M (b) and \( c_s = 0.32 \) M (c). PMFs for pairs of R10 (black solid line) and R4 (blue line) from simulations with the Amber ff99SB-ILDN force field at \( c_s = 0.07 \) M (d).

Fig. 9. Snapshots from molecular dynamics simulations illustrating the interaction between C-terminal residues of two R10 molecules at close separation. Dashed lines represent the hydrogen bonds between the negatively charged carboxyl groups and the positively charged guanidinium moieties of the ninth residues. Colored circles represent the probability of H-bonds between the carbonyl groups of one peptide and the last five residues of the other peptide. Snapshots and probabilities are obtained from MD simulations performed using the Amber ff99SB-ILDN force field and constraining the separations between the ninth residues’ guanidino-C atoms to 0.42 ± 0.04 nm.

Fig. 10. Potentials of mean force (PMFs) for pairs of R10 (red line) and R8KR (blue line) molecules as a function of the separation between the guanidino-C atoms of the tenth and eighth residues (\( CZ_{10} - CZ_8 \)) (a). PMFs for pairs of R10 (black line) molecules as a function of the separation between the guanidino-C atoms of the ninth residues (\( CZ_9 - CZ_9 \) (b)). PMFs are from umbrella sampling MD simulations, using the Amber ff99SB-ILDN force field, for systems at \( c_s = 0.07 \) M. Average \( CZ_9 - CZ_9 \) calculated from simulations of R8KR (blue points) and R10 (red points) molecules at constrained \( CZ_{10} - CZ_8 \) (c). Average \( CZ_{10} - CZ_8 \) calculated from simulations of R10 (black points) molecules at constrained \( CZ_9 - CZ_9 \) (c).

Fig. 11. Potentials of mean force (PMFs) for pairs of R10 (red lines) and R8KR (blue lines) molecules as a function of the separation between the guanidino-C atoms of the tenth and second residues (\( CZ_{10} - CZ_2 \)) (a), and of the tenth and eighth residues (\( CZ_{10} - CZ_8 \)) (b). PMFs are from umbrella sampling MD simulations, using the Amber ff99SB-ILDN force field, for systems at \( c_s = 0.07 \) M. PMFs in (a) are shifted to make the free energy at large separation coincide with the free energy of the local minimum at \( CZ_{10} - CZ_8 = 0.45 \) nm in the corresponding PMFs in (b).
CZ10 – CZ8 or CZ9 – CZ9. For the derivation of the Jacobian correction, we consider the probability distributions of c.o.m. separations for constant values of the constrained coordinate. In the case of umbrella sampling simulations along CZ9 – CZ9, we calculate the average distances between c.o.m. and CZ9 in each peptide. The Jacobian correction is obtained from the sum, over the explored C9 – C9 values, of the distributions of distances between points on the surfaces of two spheres of radius equal to the c.o.m. – CZ9 and centers at CZ9 – CZ9 separation. Figure 12a shows PMFs as a function of c.o.m. – c.o.m. for R10 molecules at c9 = 0.07 M obtained using the AMBER f99SB-ILDN (A99) and AMBER f03WS (A03) force fields. Figure 12b shows the comparison between PMFs as a function of C9 – C9 as obtained by applying the 2D WHAM method on simulations biased along CZ10 – CZ8, or directly by biasing the C9 – C9 reaction coordinate. The good agreement between these PMFs indicates that the simulations are converged. Assuming that the PMFs as a function of c.o.m. – c.o.m. separation can be considered to be pair-potentials at infinite dilution, we estimate the dissociation constant of the dimerization process according to [36]

\[ K_D = \frac{3}{4\pi R^3} \int_0^\infty 2r^2 - \frac{PMF(r)}{RT} \, dr \]

where r is the c.o.m. – c.o.m. separation, r0 delimits the binding region and it is set to 2 nm and 2.2 nm for the PMF obtained with the A03 and A99 force fields, respectively. Using the A99 force field for R10 at c9 = 0.07 M we estimate a dissociation constant \( K_D = 0.069 \text{ M} \), whereas the corresponding value obtained using the A03 force field is \( K_P = 0.338 \text{ M} \).

**Fig. 12.** Potentials of mean force (PMFs) for pairs of R10 molecules as a function of the separation between the centers of mass (c.o.m. – c.o.m.) at c9 = 0.07 M. PMFs are calculated using the AMBER f99SB-ILDN (blue line) and AMBER f03WS (red line) force fields (a). PMFs for pairs of R10 molecules as a function of the separation between the guanidino-C atoms of the ninth residues (CZ9 – C9) calculated from umbrella sampling MD simulations where the biased coordinate is CZ9 – CZ9 (blue line) and CZ10 – CZ8 (red line).

PMFs for pairs of R10 molecules as a function of the separation between the centers of mass (c.o.m. – c.o.m.) at c9 = 0.07 M. PMFs are calculated using the AMBER f99SB-ILDN (blue line) and AMBER f03WS (red line) force fields (a). PMFs for pairs of R10 molecules as a function of the separation between the guanidino-C atoms of the ninth residues (CZ9 – C9) calculated from umbrella sampling MD simulations where the biased coordinate is CZ9 – CZ9 (blue line) and CZ10 – CZ8 (red line).

**Fig. 13.** Average separation between guanidino-C atoms ((CZ – CZ)) and coordination numbers with chloride anions of closely spaced residues on two interacting R10 molecules simulated with the AMBER f03WS force field (a). Average separation between \( \epsilon \) atoms ((C – C)) and coordination numbers with chloride anions of closely spaced residues on two interacting K10 molecules simulated with the AMBER f03WS force field (b).

**Coordination with Chloride Ions.** We calculated radial distribution functions, RDFS, between chloride ions and pairs of residues of two interacting peptides. The analysis is performed on umbrella sampling simulation windows at c9 = 0.07 M where the ninth residues of two R10 or K10 peptides are restrained at close separation. RDFS are calculated as a function of the distance between the ions and the guanidino-C or \( \epsilon \) atoms of two arginine and lysine side chains, respectively. The first peak of the RDF is integrated to obtain the number of chloride ions coordinated with the pair of residues. Average residue-residue separations are measured as the distance between guanidino-C ((CZ – CZ)) and \( \epsilon \) atoms ((C – C)) of arginine and lysine side chains, respectively. The high average number of ions coordinated with the stacked residues in R10 suggests that the non-specific interaction is counterion-mediated [30]. However, coordination numbers are low for the 9-9 and 9-10 pairs of residues in R10, indicating that the mode of interaction identified in this work does not involve chloride ions.
solutions contained 0.02 M Tris which, besides keeping a constant pH, acts as hydroxyl radical scavenger [37] lowering the risk of radiation damage [38]. NaCl was added to obtain ionic strength, $c_s$, of 0.025, 0.06, 0.15, 0.3 M. The pH set was 7.8 to add a mixture of HCl 1 M in volumes of 3.5 ml and 3.9 ml to low and high $c_p$ solutions, respectively. For NMR experiments, we used 0.035 M acetic buffer solutions pH 5, prepared by mixing glacial acetic acid (Scharlau SA) and 3 NaOH (Merck KGaA). The buffer has $c_s = 0.025$ M, and NaCl (Scharlau) was added to obtain $c_p = 0.060$ M.

**Sample Purification Procedure.** For each dilution series in peptide concentration, $c_p$, around 40 mg of peptide powder was dissolved in 2 ml MilliQ water. In order to remove the buffer and salts in the peptide powder, the aqueous solution was ultrafiltrated with Vivaspin 2 Hydrocol Columns (molecular weight cut-off of 2 kDa). Ultrafiltration was performed at 277 K and 3,500 RCF adding aliquots of buffer solution until a volume of 5 ml had flowed through the column. The 150 µl concentrate was dialyzed on a rocking plate with Slide-A-Lyze MINI Dialysis Devices (Hydrocol membrane with 2 kDa MWC) for at least 20 hours at 277 K against 3.4 ml of the same buffer solution used for the ultrafiltration. Serial dilutions of R10 and K10 in the concentration range 2.5-45 g/l were made from the concentrated dialed solutions and the buffer solution in contact with the semipermeable membrane. This protocol provides peptide solutions of well defined composition as well as buffer solutions that exactly match the background solution.

**Peptide Concentration Measurements.** The peptide concentration was measured from the absorbance at 214 nm using NanoDrop 2000c spectrophotometer. The extinction coefficients were determined by amino acid analysis (Aminosyraanalyscentralen, Uppsala, Sweden). Before SAXS measurement, each sample was centrifuged (14,000 rcf for 45 minutes at 277 K). The supernatant was collected, stirred and quantified for peptide concentration immediately before the SAXS measurement. Peptide absorption at 214 nm is mainly due to the $\pi \to \pi^*$ transition in the peptide bond, with minor contributions from the arginine- and lysine-rich peptides, respectively. Our measured extinction coefficients differ by less than 14% from those estimated according to Kuipers and Gruppi [39].

**Small-angle Scattering Measurements.** Scattering measurements were performed in beamline BM29 [40] at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. The distance between the Pilatus 1M detector and the sample was set to 2.872 m, while the wavelength of the incident beam was 0.099 nm. The range of the scattering vector was 0.0326 $\leq q \leq 4.946$ nm. The sample holder was thermostated at 293 K. For each measurement, ten frames separated in time were collected and compared to assess the absence of radiation damage. Each measurement required 50 µl of sample and, with the aim of improving the signal to noise ratio, measurements on 3 mM solutions were repeated three times. The buffer was measured before and after each sample measurement, and the average intensities of the buffer were subtracted from the intensities of the sample using PRIMUS (ATAS package 2.7.1) [41]. The reproducibility of the results was verified by measuring SAXS on samples obtained from independent preparation procedures, from three beam times in September 2015, April 2016, and November 2016. The SAXS curves reported in Fig. 1, S6 and S7 were smoothed using the LOWESS method [42]. $I(q)$ values reported in Fig. 1, Fig. 4, S6, and S7 are obtained from averaging $I(q)$ in the range 0.819 $\leq q \leq 0.2986$ nm.

$^{1}H_{-^{13}}C$ HSQC NMR Measurements. NMR measurements were performed at 293 K on Agilent or Bruker spectrometers operating at 600 MHz (room temperature probe). Solutions of R10 and R8KR molecules are prepared in 0.035 M acetic buffer solutions pH 5. The ionic strength of the buffer is $c_s = 0.025$ M. Samples of $c_p = 0.06$ M are obtained by addition of NaCl.

**SAXS-derived Molecular Weights.** The monodispersity of our samples at high $c_s$ and low $c_p$ was probed by evaluating the molecular weights of the peptides from the scattering intensity extrapolated at $q = 0$ via Guinier plots. We used $M_W = N_A I(0)/ [c_p \sigma^2 (\rho_m - \rho_{sol})^2]$, where $M_W$ is the molecular weight of the peptide, $I(0)$ is the absolute scattering intensity at zero angle calculated using the Guinier approximation, $\sigma$ is the partial specific volume (K10 and K8RK: 0.84 cm$^3$/g, R10 and R8KR: 0.80 cm$^3$/g [44], $\rho_m$ is the scattering density of the peptide (K10: 1.13 $\times 10^4$ cm$^{-2}$, R10: 12.1 $\times 10^4$ cm$^{-2}$) while $\rho_{sol}$ is the scattering density of a 300 mM NaCl solution (9.51 $\times 10^4$ cm$^{-2}$ [45]). The Guinier plots (Fig. S14a) for R10 and K10 samples with the lowest measured $c_p$ and $c_s = 0.300$ M were linearly fitted to $I(q) \approx \ln I(0) - (q/R_g^2)/5$, where $R_g$ is the peptide radius of gyration. The small discrepancies between SAXS-derived $M_W$s and theoretical values can be attributed to the uncertainty on the partial specific volume or to the presence of intermolecular interactions [43]. Fig. S14b shows the log-log plot of the scattering intensities used for the Guinier plots. The plateaus in the low-$q$ range confirm that the samples were monodisperse.

**Molecular Dynamics Simulation Protocol.** R10 and K10 are modeled with the all-atom force field AMBER03WS with the correction of protein-water interaction proposed by Best et al. and the TIP4P/2005 water model [33]. The correction consists of increasing the Leonard-Jones $\epsilon$ parameter by 10%, between oxygen atoms of water molecules and all peptide atoms. R10 and K10 have a net charge number of +10. The carbonyl group is negatively charged, while the amino acid side chains and the N-terminal amine groups are protonated. Umbrella sampling [46] was used to calculate the PMF as a function of separation between guanidino-C and $\pi$-C separation by 2 nm. The PMF was constructed using the weighted histogram analysis method (WHAM).

**Fig. 14. Guinier plots of R10 (red line), R8KR (blue line), K10 (green line), K8RK (black line) from SAXS curves at the lowest measured $c_p$ at $c_s = 0.300$ M (a).** The PMF was constructed using the weighted histogram analysis method (WHAM).
method (WHAM) [52], estimating the uncertainty by performing 10 bootstraps of the free energy profile. The convergence of the PMFs was evaluated by inspecting the effect of increasing the simulation time of the umbrella sampling windows by 20 ns. A simulation time of at least 70 ns was sampled for each umbrella sampling window.

Molecular structures were rendered using VMD [53].

Hydrogen Bond Probabilities. H-bonds between C-terminal carboxyl groups and amine groups on the residues’ side chains are identified based on the criterion outlined previously [54], using the routine implemented in MDTraj [55]. The analysis is performed on umbrella sampling simulations where the peptides are at the closest distance along the reaction coordinate. A maximum of one H-bond is counted for each residue in each frame, so that monodentate and bidentate H-bonds are weighted equally.

Analysis of the Protein Data Bank. The PISCES web server [56] was used to obtain a list of PDB X-ray crystal structures with maximum 25% sequence identity, minimum resolution of 0.25 nm and R factor lower than 0.25. The list was generated on July 5 2017 and included 10,388 entries. The corresponding PDB files were downloaded from the RCSB PDB [57] and parsed using a Python script. MDTraj v1.7.2 [55] was used to search arginine guanidino-C atoms or lysine ε-C atoms closer than 0.4 nm, as well as to identify H-bonds between like-charged ion pairs and carboxyl groups of glutamate or aspartate residues. In the case of arginine and lysine pairing, 1,594 out of 1,697 and 395 out of 423 protein crystal structures, respectively, were missing hydrogen atoms. Hydrogen atoms were added to the protein structures using PDBFixer v1.4, a routine of the OpenMM molecular simulation toolkit [58]. Proteins presenting at least one pair of stacked arginine/lysine residues forming H-bonds with two carboxyl groups were selected.

Debye-Hückel Predictions. The Debye-Hückel approximation is used to estimate free energies of interaction between peptides at large separations. Each umbrella sampling window corresponds to a restricted position $r$ along the reaction coordinate and to a more extensively fluctuating mass-center separation, $r_{cm-cm}$. The Debye-Hückel free energy was calculated, for each umbrella sampling window, using the average mass-center separation, $\langle r_{cm-cm} \rangle$, and $w(r) = RT \lambda_B Z^2 \exp[-\langle r_{cm-cm} \rangle / \lambda_D] / \langle r_{cm-cm} \rangle$, where $Z=10$ is the net charge number of the peptides, $\lambda_B = 0.713$ nm is the Bjerrum length for water at $T = 293$ K, $\lambda_D = 0.304 / \sqrt{c_s}$ is the Debye screening length of a monovalent salt solution of concentration $c_s$ and $R$ is the gas constant. Electrostatic free energies at the $i$-th charge site, in the schematic illustration of the R10 dimer shown in Fig. 3, were calculated with $w_i = RT \sum_j \lambda_B z_i z_j \exp[-r_{ij} / \lambda_D] / r_{ij}$, where $r_{ij}$ is the distance between the $i$-th and $j$-th sites, of charges $z_i$ and $z_j$, respectively.