Self-assembly of functionalized colloids and short amyloidogenic peptides: Modelling, theory and simulations
Luiken, J.A.

Citation for published version (APA):
Chapter 6

The primary nucleation kinetics of short fibril-forming amyloidogenic peptides

The primary nucleation step in amyloid fibril formation can, depending on the nature of peptide sequence, occur in one step, straight from a dilute solution, or in multiple steps, via oligomers or disordered aggregates. The precise kinetic pathways of these processes are poorly understood. Employing forward flux sampling and a mid-resolution coarse-grained force field, we analyzed the reactive pathways from the solvated state to the fibril nucleus for a system of 12 amyloidogenic peptides. We found that, in line with previous work, increasing the overall side-chain hydrophobicity switches the fibrillization mechanism from one- to two-step nucleation. Overall, in this mechanism, peptides first form dimers and trimers which then grow into a $\beta$-sheet. This sheet serves as a template for nucleation of additional $\beta$-sheets until the fibril nucleus is fully formed. The presence of the recently predicted polymerized phase in the nucleation pathway of intermediately hydrophobic peptides slows down the dynamics of fibril formation considerably, which may influence the timescale on which toxic early oligomers exist. The structure of the amyloid fibrils was found to be strongly dependent on the relative hydrophobic strength of side-chains along the sequence: $\beta$-sheets in the fibril are oriented such that a core of the relative strongest hydrophobic residues is formed along the fibril axis.

The contents of this chapter have been published:
6.1 Introduction

The self-assembly of amyloidogenic peptides into amyloid fibrils is interesting from a biological, medical and material science viewpoint. Fibril formation occurs readily \textit{in vivo} and \textit{in vitro} \cite{20} and is traditionally linked to neurodegenerative diseases such as type II diabetes, Parkinson’s, Alzheimer’s and Huntington’s disease \cite{21–24}. In the early stages of amyloid formation, proteins are known to aggregate into disordered oligomers. These early oligomers are considered toxic as they disrupt neuronal function \cite{23, 198}. Amyloid fibrils, however, also play functional roles in some organisms, such as bacteria \cite{153, 154}, and there has been increasing interest in the utilization of amyloids as strong biomaterials \cite{26, 27} such as micrometer long fibrils formed from silk-based block-copolymers \cite{29}, or strong fiber networks formed by proteins containing the IKVAV sequence \cite{28}.

For the development of treatments against deposition diseases, or for tailoring materials with strong physical features, control of the aggregation is crucial. Such control requires a good understanding of the early steps of amyloid formation. Molecular simulation can in principle provide such knowledge. Straightforward molecular dynamics (MD) using all atom force fields has been very useful in revealing the properties on the single peptide level, such as the dock-lock mechanism of amyloidogenic peptide association \cite{32, 199}. However, as the timescales on which the various steps of amyloid nucleation itself take place range from seconds to years \cite{200}, the computational expense associated with straightforward molecular dynamics (MD) simulations using all-atom force fields is prohibitively large. These long time scales are related to high free energy barriers that are present in the nucleation mechanism. There are two major solutions to this problem available in the literature. The first is to use enhanced sampling methods, such as replica exchange MD (REMD) \cite{84} or metadynamics \cite{201} to overcome the free energy barriers involved in fibril nucleation. As an example, Baftizadeh et al. \cite{202, 203} have investigated the aggregation of polyvaline and Amyloid-β using bias exchange metadynamics, and found non-classical nucleation behavior. REMD can be used for faster convergence towards equilibrium, but it is not easy to extract kinetics from REMD simulations. While in principle it would be possible to obtain insight in kinetic nucleation pathways from REMD simulations \cite{92}, the transitions often occur at unnaturally high temperatures and may thus not correctly represent the process at ambient or physiological conditions. The use of transition path sampling \cite{93, 97} or forward flux methodology \cite{101} circumvents this problem, and has \textit{e.g.} enabled the study of the docking mechanism of peptides \cite{33, 188, 204–206}, but has not yet been applied to larger amyloid systems.
The second solution to the timescale problem is to employ coarse-grained force fields, which contain less atomic and molecular detail, but which allow much longer simulation timescales. Coarse-grained molecular simulations have provided insight into the early steps of amyloid formation [29, 33, 64, 160, 166, 177, 207]. The mid-resolution model of Bereau and Deserno [41] provides the necessary tools to simulate amyloid formation, such as directional hydrogen bond interactions, dipole interactions and full sequence specificity. However, accessing the long timescales of amyloid fibril formation requires a combination with enhanced sampling techniques. Recently, we studied the aggregation thermodynamics of three short, amyloidogenic peptides (GNNQQNY, VEALYL and KLVFFAE) with REMD simulations using the model by Bereau and Deserno [207]. These REMD simulations yielded insight in the structural properties of the liquid phase and the fibril phase, and lead to the prediction of a new intermediate in the nucleation process: the polymerized (or associated) peptide liquid phase. In this phase, small clusters of aligned peptides coexist which possibly prolong the existence of the toxic early oligomers. We tested the robustness of this prediction with statistical mechanical theory [207].

In this work we aim to gain insight in the kinetic pathways of the primary nucleation steps of the three amyloidogenic peptides GNNQQNY, VEALYL and KLVFFAE. In addition we are interested in the dynamical role of the polymerized phase in the nucleation pathway. We utilize the forward flux sampling method [100–102] to investigate the nucleation dynamics of 12 amyloidogenic peptides in a dilute solution. For each of the three peptide sequences we compute the kinetics, the free energy and analyze the mechanism.

The remainder of this chapter is organized as follows. In the Methods section we introduce the simulation details and the analysis techniques. For a detailed description of the forward flux sampling method we refer to Section 2.4.2. In the Results and discussion section we present and discuss the results of the simulations for each peptide separately. We summarize the findings in the Conclusion section.

6.2 Methods

6.2.1 Simulation details

We employed the coarse-grained (CG) model by Bereau and Deserno [41]. For a detailed description of this model we refer to Ref. 41 or Section 2.1.3. The integration was carried out by the ESPResSo [70] MD package using Langevin dynamics with a 1 fs timestep. The FRESHS [100, 208, 209]
package\(^1\) was modified to interface with the peptide builder\(^2\), which feeds coarse-grained peptides into the MD integrator. We solvated 12 peptides in a simulation box with sides \(L = 63.3\ \text{Å}\), corresponding to a concentration of approximately 0.07 M. Periodic boundary conditions apply in all directions. We performed five independent forward \((A \rightarrow B)\) and reverse \((B \rightarrow A)\) FFS simulations for each peptide, and calculate the order parameter and save an output structure every 1 ps.

The proper choice of order parameters and the set of interfaces that separates the solvated and aggregated state requires substantial knowledge of the system, which is not always available \textit{a priori}. Here we will make use of the results of REMD simulations of Ref. 207 using the same CG model, which indicate that the initial and final state are clearly separated in terms of the in-register contacts \(N_{C_{\alpha}}\), defined as when two \(C_{\alpha}\) atoms from the same residue on different strands are within 6.5 Å of each other. The 12-peptide GNNQQNY system forms below the ordering temperature \(T_o \approx 291\ \text{K}\), defined by a sharp peak in the heat capacity as obtained from the REMD simulations, a one-layered fibril consisting of around 9 peptides with the remaining peptides arranged in a second layer parallel to the first. The second layer stabilizes the fibril structure, as is shown by the 20 peptide system in Ref. 207. The order parameter \(\lambda\) for GNNQQNY is defined as the number of in-register contacts in the largest cluster. For KLVFFAE the fibril state consists of two parallel aligned layers \((T_o \approx 249\ \text{K})\), while for VEALYL the fibril consists of three parallel \(\beta\)-sheets \((T_o \approx 307\ \text{K})\). For these peptides we chose the order parameter \(\lambda\) to be the number of in-register contacts in the two largest clusters. For KLVFFAE and GNNQQNY we set boundary \(B\) to \(\lambda_B = 44\) in-register contacts, and for VEALYL to \(\lambda_B = 38\).

As mentioned above, the FFS algorithm is initiated by performing a straightforward MD run which terminates when 150 initial configurations are generated on boundary \(A\). Table 6.1 lists the threshold number of configurations \(C_i\) for each interface \(i\) required to initiate trials runs from \(i\), and the value of the order parameter at each interface. The FRESHS package imposes the condition \(\lambda_{i+1} > \lambda_i\), therefore for FFS simulations in the reversed direction \((B \rightarrow A)\) we define an additional order parameter \(\lambda'\):

\[
\lambda' = \lambda_A + \lambda_B - \lambda.
\]  

(6.1)

Random output structures on boundary \(B\) from the forward simulations are used as starting configurations for the initial MD runs in region \(B\).

\(^1\)The FRESHS package is publicly available at \url{http://www.freshs.org/dw/doku.php}
\(^2\)The peptide builder is publicly available on github: \url{https://github.com/tbereau/peptideB}
<table>
<thead>
<tr>
<th>i</th>
<th>$C_i$</th>
<th>$\lambda_i$</th>
<th>$\lambda_i'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>150</td>
<td>5 (A)</td>
<td>5 (B)</td>
</tr>
<tr>
<td>1</td>
<td>120</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>56</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>54</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>52</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>51</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>13</td>
<td>50</td>
<td>44 (B)</td>
<td>44 (A)</td>
</tr>
</tbody>
</table>

Table 6.1: Order parameter values $\lambda_i$ and $\lambda_i'$ and number of configurations ($C_i$) required per interface ($i$) for GNNQQNY and KLVFFAE. For VEALYL the boundary of $B$ is set to 38 in-register contacts (interface 11).

6.2.2 Cluster analysis

We obtain further insight into the aggregation process by following the formation and breaking of correctly formed peptide clusters of size $n$, by performing a cluster analysis on the frames along each reactive trajectory. Two peptides belong to the same cluster if they form at least 4 in-register contacts with each other. The algorithm is initiated by looping over all clusters in the first frame of a trajectory, and computing the size $n$ of each cluster. In the next frame we determine whether each cluster has grown, remained constant or shrunk, and we increment the corresponding histogram $h_{\text{grow}}(n)$, $h_{\text{same}}(n)$ or $h_{\text{shrink}}(n)$. This procedure is repeated until the end of the trajectory is reached. The probability that a cluster of size $n$ grows, $P_{\text{grow}}(n)$ is then calculated using

$$P_{\text{grow}}(n) = \frac{h_{\text{grow}}(n)}{h_{\text{grow}}(n) + h_{\text{same}}(n) + h_{\text{shrink}}(n)}.$$  \hspace{1cm} (6.2)

$P_{\text{same}}(n)$ and $P_{\text{shrink}}(n)$ are calculated in a similar manner. In some cases it is more insightful to calculate the growth probability relative to the shrinking probability:

$$P_{\text{grow}}(n) = \frac{h_{\text{grow}}(n)}{h_{\text{grow}}(n) + h_{\text{shrink}}(n)}.$$  \hspace{1cm} (6.3)
Although in principle it is possible to extract information on the critical nucleus size $n^*$ from this analysis, as $P_{\text{shrink}} > P_{\text{grow}}$ for $n < n^*$ while $P_{\text{shrink}} < P_{\text{grow}}$ for $n > n^*$ [210], we analyze only the reactive pathways and therefore only obtain insight into the nucleation mechanism rather than the size of the critical nucleus.

6.2.3 The committor

The committor $P_B(x)$ is the probability that a trajectory initiated from configuration $x$ reaches boundary $B$ before it reaches boundary $A$. This probability increases from 0 to 1 as the reaction progresses from the initial to final state, therefore it is in principle the ideal reaction coordinate and, as such, is useful to measure the correlation between the chosen order parameter and the true reaction coordinate [102, 211, 212]. A committor value of around 0.5 has special significance, as here the configuration has an equal probability of reaching the initial or final state. These configurations comprise the transition state ensemble (TSE), and can be analyzed to obtain information on the reaction mechanism.

The procedure to calculate the committor for direct FFS is similar to that for branched growth FFS [213], with the main difference that the number of trials fired varies per configuration rather than per interface. For each configuration $j$ on interface $\lambda_i$ we estimate the corresponding committor $P_{Bj}^i$ from the sum of the committors of the configurations it connects to on the next interface, $P_{Bj}^{i+1}$, divided by the number of trials fired from this configuration, $k_j$:

$$P_{Bj}^i = \frac{\sum_{m=1}^{C_{i+1}} P_{Bm}^{i+1}}{k_j}. \quad (6.4)$$

We consider all configurations for which $0.35 < P_B(x) < 0.65$ to be part of the TSE. A good indication for a strong correlation of the order parameter with the true reaction coordinate is a highly peaked distribution of the order parameter in the TSE [96, 97].

6.2.4 Free energy analysis

The free energy $F$ is directly related to the distribution function $\rho(N_{C_{\alpha}})$ through

$$F(N_{C_{\alpha}}) = -k_BT \ln \rho(N_{C_{\alpha}}). \quad (6.5)$$

In case of a stable state $B$, $\rho(N_{C_{\alpha}})$ is extracted from FFS simulations by weighting the average time all trial runs from the forward ($A \rightarrow B$) and reverse simulations ($B \rightarrow A$) spent at order parameter $N_{C_{\alpha}}$ [102, 214]:

$$\rho(N_{C_{\alpha}}) = \rho_A \Phi_{A,0}\tau_+(N_{C_{\alpha}};\lambda_A) + \rho_B \Phi_{B,n}\tau_-(N_{C_{\alpha}};\lambda_B), \quad (6.6)$$
where $\rho_A$ is the probability of finding the system in region $A$ at steady state, $\Phi_{A,0}$ is the flux of trajectories from $A$ that reach interface $\lambda_0$ and $\tau_+(N_{C,\alpha};\lambda_A)$ is the average time a trajectory originating from $A$ has spent at order parameter $N_{C,\alpha}$. Similarly, $\rho_B$ is the steady-state probability of finding the system in region $B$, $\Phi_{B,n}$ is the flux of trajectories from $B$ that reach interface $\lambda_n$ and $\tau_-(N_{C,\alpha};\lambda_B)$ is the average time a trajectory originating from $B$ has spent at order parameter $N_{C,\alpha}$.

We extract the probabilities $\rho_A$ and $\rho_B$ from the rate constants using the relation

$$\rho_A k_{AB} = \rho_B k_{BA} \tag{6.7}$$

and the fact that $\rho_A + \rho_B \approx 1$ at steady state. The fluxes $\Phi_{A,0} = N_f/\tau$ and $\Phi_{B,0} = N_f/\tau$ are calculated during the initial straightforward MD simulation in the corresponding regions, where $\tau$ is the total simulation time of the initial run. The function $\tau_+(N_{C,\alpha};\lambda_A)$ is calculated using data from the forward simulations [102]:

$$\tau_+(N_{C,\alpha};\lambda_A) = \pi_+(N_{C,\alpha};\lambda_A) + \sum_{i=1}^{n-1} \pi_+(N_{C,\alpha};\lambda_i) \prod_{j=0}^{i-1} P(\lambda_{j+1}|\lambda_j), \tag{6.8}$$

where $\pi_+(N_{C,\alpha};\lambda_i)$ is the average time a run originating from interface $\lambda_i$ spent with order parameter $N_{C,\alpha}$. The function $\tau_-(N_{C,\alpha};\lambda_A)$ is calculated in a similar fashion with data from the reverse simulations.

### 6.3 Results and discussion

#### 6.3.1 The GNNQQNY peptide

In Ref. 207 we inferred the nucleation behavior of the weakly hydrophobic (hydrophilic) GNNQQNY peptide on the basis of the aggregation thermodynamics. Using Wertheim’s first-order perturbation theory and statistical associating fluid theory we showed that polymerization in the gas phase is possible for weakly hydrophobic peptides. However, in the REMD data we did not find evidence for a full polymerization transition for the 12-peptide GNNQQNY system.

The polymerization transition is signified by an inflection point in the aggregation fraction $\Theta$ as function of temperature (see Fig. 6.1). Although clustering is possible in the gas phase ($\Theta > 0$ for $T > T_o$), there is no inflection point for $T > T_o$. We therefore hypothesize that a critical fibril nucleus forms during the polymerization process, thus allowing the system to crystallize before it can fully polymerize. Based on the thermodynamic
behavior of GNNQQNY we argued that slightly below the ordering temperature peptides nucleate in one step directly from the dilute solution.

![Figure 6.1: Plot of the aggregation fraction versus temperature.](image)

Here we study the mechanism of fibril formation of GNNQQNY by performing FFS simulations. We set the simulation temperature to 285 K, slightly below $T_o$, and define boundaries $\lambda_A = 5$ and $\lambda_B = 44$ with $\Delta\lambda = 3$. In the first phase of FFS, a straightforward MD simulation, initiated from a random solvated state, is run until it crosses boundary $A$ in the positive direction 150 times (see Table 6.1). In the second phase, paths are generated from these crossing points, for each interface, saving each crossing point of the next interface. As $C_B = 50$, we generate eventually 50 paths in the transition path ensemble (TPE). As some of these paths are likely to be correlated, we perform five independent FFS simulations and combine the data. To calculate the steady-state probabilities $\rho_A$ and $\rho_B$, the rate constant $k_{BA}$ and the free energy curve we initiate five reverse FFS simulations from boundary $B$ using random output structures on boundary $B$ from the forward simulations. Below we will discuss the results of the FFS simulations.

We computed the committor values for the forward transition $A \rightarrow B$ and plotted the distribution of the in-register contacts evaluated over the transition state ensemble (TSE) in Fig. 6.2a. Note that the probability of finding a transition state is peaked around 38 contacts, near boundary $\lambda_B = 44$. This suggests that the addition of peptides to large aggregates, consisting of around 38 contacts or more, is the rate-limiting step, and that state $B$, in fact, might be metastable. Indeed, the computed rate constants (see Table 6.2) are $k_{AB} = 4.79 \cdot 10^2 \text{ s}^{-1}$ and $k_{BA} = 7.61 \cdot 10^4 \text{ s}^{-1}$, i.e. the re-
verse reaction proceeds significantly faster than the forward reaction. The ratio of the rate constants $k_{AB}/k_{BA} = 0.006$ is equal to the equilibrium constant for the reaction. Using the well known relation for the free energy difference $\Delta F_{AB} = F_B - F_A = -k_B T \ln \frac{k_{AB}}{k_{BA}} = 5.07 k_B T$. The crossing probabilities, flux and rate constants are listed in Table 6.2. The steady state probability of finding the system in state $A$, given by

$$\rho_A = \frac{k_{BA}}{k_{AB} + k_{BA}} \approx 0.994,$$  \hspace{1cm} (6.9)

and the probability of finding the system in state $B$, $\rho_B = 1 - \rho_A \approx 0.006$, reflect these results; at 285 K large $\beta$-sheets of GNNQQNY peptides are at best metastable with respect to the dissolved A state.

The results of the free energy analysis are shown in Fig. 6.2b. We can identify several peaks belonging to the subsequent addition of peptides to the largest cluster. Clusters with size $n = 3$ contain at least 8 contacts, and those with size $n = 4$ at least 12 contacts, therefore the first and second peak are likely due to the formation of trimers and tetramers (see insets in Fig. 6.2b for simulation snapshots). The top of the barrier is around 27 contacts, which corresponds to clusters consisting of 4 to 6 peptides. Such clusters are likely to act as a critical nucleus. The right-most metastable minimum around $\sim 37$ contacts is due to clusters of size $n = 7, 8$ and 9. Beyond this point, $\beta$-sheets become increasingly less stable. The metastable minimum is around $4 - 5 k_B T$ above the minimum for state A, roughly in agreement with the free energy difference based on the rate constants. The fact that this metastable minimum is not corresponding to the definition of state B ($\lambda \geq 44$), reveals the limitation of $N_{C\alpha}$ to describe this state.
Moreover, the location of the maximum in the free energy ($\lambda \approx 25$) does not agree with the distribution of the order parameter in the TSE in Fig. 6.2a. Also the height of the free energy barrier (about $5k_B T$ and $2k_B T$ for the forward and backward transition, respectively) is much lower than the rate constants suggest. This is indicative of an overlap of state B with a large part of the barrier in terms of the order parameter $N_{Ca}$. 

**Table 6.2:** Crossing probabilities per interface, the flux ($s^{-1}$) and the rate constants ($s^{-1}$) for the forward and reverse FFS simulations of all peptides.

<table>
<thead>
<tr>
<th></th>
<th>GNNQQNY</th>
<th>VEALYL</th>
<th>KLVFFAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$P(\lambda_{i+1}</td>
<td>\lambda_i)$</td>
<td>$P(\lambda_{i+1}'</td>
</tr>
<tr>
<td>1</td>
<td>0.1182 0.1872</td>
<td>0.2722 0.7057</td>
<td>0.1118 0.3414</td>
</tr>
<tr>
<td>2</td>
<td>0.4050 0.2376</td>
<td>0.4119 0.3058</td>
<td>0.2774 0.7531</td>
</tr>
<tr>
<td>3</td>
<td>0.3405 0.3550</td>
<td>0.3694 0.3220</td>
<td>0.2635 0.8603</td>
</tr>
<tr>
<td>4</td>
<td>0.2913 0.6084</td>
<td>0.2722 0.1831</td>
<td>0.3558 0.9558</td>
</tr>
<tr>
<td>5</td>
<td>0.2996 0.6805</td>
<td>0.4428 0.3696</td>
<td>0.2943 0.8712</td>
</tr>
<tr>
<td>6</td>
<td>0.2357 0.7501</td>
<td>0.4048 0.3844</td>
<td>0.4438 0.7073</td>
</tr>
<tr>
<td>7</td>
<td>0.3776 0.8401</td>
<td>0.5606 0.3483</td>
<td>0.3430 0.6093</td>
</tr>
<tr>
<td>8</td>
<td>0.3031 0.8374</td>
<td>0.5142 0.3549</td>
<td>0.2677 0.2092</td>
</tr>
<tr>
<td>9</td>
<td>0.6474 0.8282</td>
<td>0.5456 0.5779</td>
<td>0.2831 0.1179</td>
</tr>
<tr>
<td>10</td>
<td>0.6506 0.8930</td>
<td>0.5864 0.4793</td>
<td>0.4604 0.2043</td>
</tr>
<tr>
<td>11</td>
<td>0.7702 0.8734</td>
<td>0.5664 0.5128</td>
<td>0.4107 0.3527</td>
</tr>
<tr>
<td>12</td>
<td>0.8748 0.8151</td>
<td></td>
<td>0.4494 0.3150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\Phi$</th>
<th>$k$</th>
<th>$\Phi$</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$2.89 \cdot 10^7$</td>
<td>$6.31 \cdot 10^7$</td>
<td>$1.55 \cdot 10^7$</td>
<td>$2.59 \cdot 10^7$</td>
</tr>
<tr>
<td></td>
<td>$4.79 \cdot 10^2$</td>
<td>$7.61 \cdot 10^4$</td>
<td>$1.39 \cdot 10^3$</td>
<td>$90.1$</td>
</tr>
</tbody>
</table>

Figure 6.3: (a) Growth and shrinking of GNNQQNY clusters along the TPE. (b) Growing, shrinking and stability of clusters.
By performing a time series cluster analysis of the transition path ensemble (see Methods section), we illustrate the decline in growth probability for these clusters. Fig. 6.3a shows the growth probability relative to the shrinking probability. In the reactive nucleation paths it is likely that small peptide clusters \( n < 4 \) shrink on average \( i.e. \ P_{\text{shrink}} > P_{\text{grow}} \). Clusters consisting of 4 or 5 peptides are most likely to grow, but the growth probability steeply declines for \( n > 5 \). This suggests that nuclei of 4 to 5 peptides are characteristic for a barrier crossing, but as was pointed out in the Methods section, they do not necessarily correspond to the critical nucleus since we only take reactive pathways into account. Indeed, this characteristic nucleus size does not correspond to the maximum in the free energy barrier in Fig. 6.2b. We included the fraction of unchanged clusters between snapshots in the time series cluster analysis in Fig. 6.3b. Here we indeed observe a steep decline in the unchanged configurations for \( n > 7 \); larger clusters are thus more likely to shrink than to remain stable in the absence of free monomers\(^1\). We can attribute the decrease in stability to one or more of the following reasons: 1) the REMD data in Fig. 6.1 indicates that at \( T = 285 \) K some peptides have begun assembling into the fibril, however the system is not yet fully aggregated (\( \Theta \approx 0.55 \)), thus large \( \beta \)-sheets (\( n > 5 \)) may not be thermodynamically stable at this temperature; 2) the FFS algorithm drives the formation of a single large cluster rather than the formation of two layers, which was shown to be the thermodynamically most stable state at temperatures \( T < T_o \) in Ref. 207. Therefore state \( B \) as we defined it may not correctly represent the equilibrium structure of the fibril. Finally, 3) it may be due to a finite-size effect: the free monomer concentration declines as the cluster size increases. The timescale \( \tau_{\text{grow}} \) on which a monomer docks to the fibril is inversely proportional to the monomer concentration: \( \tau_{\text{grow}} \propto L^3/(N - n) \), and thus increases with increasing \( n \). The timescale \( \tau_{\text{shrink}} \) on which peptides at the fibril ends detach, however, is independent of the monomer concentration, therefore for large \( n \) the rate of detachment may exceed the rate of attachment (\( \tau_{\text{grow}} > \tau_{\text{shrink}} \)), and large \( \beta \)-sheets may become more likely to shrink than grow in a given time frame.

By analyzing the reactive pathways of GNNQQNY we reconstructed the early aggregation mechanism, and visualized it in Fig. 6.4a-c. The initial state consists of random-coil peptides in solution (a). Then, the peptides form dimers and trimers via a docking mechanism [33] (b) until the critical nucleus size is reached (c). The local environment of aligned peptides acts as a template that helps peptides lock firmly into place, and the resulting \( \beta \)-sheet grows through peptide addition at the ends of the sheet.

\(^1\)We note that increasing the time between frames would decrease the unchanged fraction for all \( n \).
This self-templated nucleation effect has been previously reported in Ref. 177 for nucleation inside oligomer droplets. The formation of the critical nucleus thus proceeds directly from solution, the hallmark of a one-step nucleation mechanism (1SN). For large clusters peptides detach more frequently due to the instability of large sheets or due to monomer depletion (e). Nevertheless clusters are unlikely to become smaller than the critical nucleus size. The remaining free monomers form a second β-sheet that interacts laterally and aligns in-plane with the first sheet (f). In-plane aligned β-sheets of GNNQQNY peptides were observed in earlier work [41], where the driving force was found to be the attraction between C-terminal tyrosines. From Fig. 6.4f we can conclude that rather than growing a cluster larger than seven peptides, the additional peptides prefer to form a second layer. This second layer provides additional stability through lateral hydrophobic interactions between tyrosine residues, thereby forming a hydrophobic core along the fibril axis. This structure may well be a transition state in the formation of the two-layered fibril (g). To test this hypothesis we performed additional FFS simulations using the number of in-register contacts in the two largest clusters (data not shown). However, here two independent β-sheets formed in solution, rather than a two-layered fibril.

The GNNQQNY results are in good agreement with previous experimental work [215] and simulations using atomistic and other CG models [158, 165, 216], where amyloid formation was found to proceed in a single step from solution (1SN) with a critical nucleus of four to six peptides.
The VEALYL peptide is the most hydrophobic of the three peptides, and has the strongest overall side-chain interactions\cite{41} compared to GN-NQQNY and KLVFFAE. These interactions drive the formation of a dense liquid phase prior to forming fibrillar structures. We set the simulation temperature to 300 K, slightly below the ordering temperature of $T_o \approx 307$ K determined from the REMD simulations. The boundaries of the disordered droplet state $A$ and the fibrillar state $B$ are set to $\lambda_A = 5$ and $\lambda_B = 38$ respectively, with a spacing of $\Delta \lambda = 3$ between interfaces.

During the initial run in region $A$ we found that peptides quickly coalesce into a dense disordered liquid of peptides, driven by the strong hydrophobic interactions between the side-chains, which suggests a two-step nucleation (2SN) mechanism. The distribution of the order parameter in the TSE of the $A \rightarrow B$ transition is plotted in Fig. 6.5a, and is peaked around 32 contacts. The fibril state of the 12-peptide VEALYL system is more stable than the disordered state. The forward and reverse reaction rate constants are $k_{AB} = 1.39 \cdot 10^3$ s$^{-1}$ and $k_{BA} = 90.1$ s$^{-1}$, respectively. Table 6.2 lists the crossing probabilities, flux and rate constants. These rate constants translate into steady-state probabilities of $\rho_A = 0.0608$ and $\rho_B = 0.939$. The free energy of fibril formation $\Delta F_{AB} = -k_B T \ln \frac{k_{AB}}{k_{BA}} = -2.73 k_B T$ is indeed negative.

In Fig. 6.5b we plotted the free energy as function of the number of in-register contacts in the two largest clusters. The minimum at 6-9 contacts corresponds to a disordered oligomeric state containing mostly monomers.

![Figure 6.5](image)

Figure 6.5: (a) Distribution of the order parameter of the TSE configurations for the $A \rightarrow B$ transition in the 12-peptide VEALYL system. (b) Free energy versus in-register contacts in the two largest clusters. Simulation snapshots are shown as insets.
and either one dimer, two dimers or one trimer. The two peaks following this minimum correspond to the formation of a small \( \beta \)-sheet consisting of 4 peptides. This \( \beta \)-sheet forms a hydrophobic surface in the oligomer which serves as a template for monomers to align and form the second \( \beta \)-sheet flat on the first sheet. This process is known as self-templated nucleation [177]. Subsequently, the remaining monomers align and a stable three-layered fibril is formed. The minimum at \( N_{C_{\alpha}} = 35 \) is roughly 2 \( k_B T \) lower in free energy than the dissolved state, in agreement with the rate constant estimates. The structure of the fibril corresponds well with the thermodynamically most stable structure found from REMD simulations, and consists of three sheets hooked into each other through their strongly hydrophobic C-terminal leucine and tyrosine residues. Several representative simulation snapshots are shown in Fig. 6.5b. As in the case of GNNQQNY, the location of the maximum in the free energy (\( \lambda \approx 25 \)) does not agree with the distribution of the order parameter in the TSE in Fig. 6.5a, and free energy barrier (about \( 4 k_B T \) and \( 6 k_B T \) for the forward and backward transition, respectively) is much lower than the rate constants indicate. Again, this suggests a limitation of order parameter \( N_{C_{\alpha}} \).

A time series cluster analysis of all reactive pathways reveals that the growth probability of dimers and trimers exceeds the shrink probability (Fig. 6.6a). The growth probability abruptly decays to zero for \( n \geq 4 \), which confirms that in the reactive pathways the peptides form a second sheet rather than attaching to the first sheet. The fraction of clusters that remain of constant size (green line in Fig. 6.6b) also indicates that the stability of larger clusters steeply declines with \( n > 4 \). We can conclude that 1) the overall hydrophobicity of the chains drives a hydrophobic collapse into a disordered oligomer prior to forming the fibril (2SN), and 2) rather than forming a flat three-layered fibril, the \( \beta \)-sheets are oriented such that a hydrophobic core of leucine and tyrosine residues forms along the fibril axis, therefore the structural properties of the fibril are determined by the relative side-chain hydrophobicity along the peptide chain.

Based on inspection of the reactive pathways we visualized the aggregation mechanism in Fig. 6.7. First, freely soluble peptides (a) rapidly coalesce into a disordered oligomer (b). Inside the oligomer, peptides form persisting bonds through their strongly hydrophobic C-terminal residues and form a critical nucleus. The nucleus grows until a single \( \beta \)-sheet consisting of 4 peptides is formed (c). The formation of a second sheet proceeds more rapidly as the first sheet serves as a template for nearby peptides [177]. Due to this self-templated nucleation effect the fibril forms in a highly cooperative manner. The second sheet lies on top of the first sheet and is aligned parallel to maximize the hydrophobic interactions between the C-terminal residues (d). In the last step, the remaining peptides align to form a third
Figure 6.6: (a) Growth and shrinking of VEALYL clusters along the TPE. (b) Growing, shrinking and stability of clusters.

Figure 6.7: Schematic of the nucleation pathway of VEALYL peptides.

Sheet (d) and the resulting fibril undergoes a structural change to maximize the hydrophobic interactions. The fibrillar structure consists of three sheets hooked into each other to form a hydrophobic core along the fibril axis. Subsequent growth would take place along the fibril axis. Indeed, the REMD simulations in Ref. 207 for 20 peptides show more elongated structures.

Solid-state NMR spectroscopy in combination with atomistic MD simulations in recent work [217] reveal that VEALYL peptides arrange in anti-parallel $\beta$-sheets stacked parallel in a steric zipper structure. These authors have identified several aggregated states: disordered oligomeric, microcrystalline and fibrillar. We have identified the oligomeric intermediate in our simulations of the VEALYL peptide, however due to the lack of side-chain detail and electrostatic interactions on glutamic acid in the CG model we did not observe anti-parallel $\beta$-sheets or steric-zipper interfaces.
6.3.3 The KLVFFAE peptide

KLVFFAE peptides have an overall hydrophobic strength in between that of VEALYL and GNNQQNY. We initially performed FFS simulations of KLVFFAE peptides at 240 K, slightly below the ordering temperature $T_o \approx 249$ K determined from REMD simulations, using various order parameters (in-register contacts, total contacts, in-register contacts in largest cluster(s) and the nematic order parameter $P_2$ [181, 191]) to drive the formation of the thermodynamically most stable structure at this temperature, the two-layered fibril. We found that, similar to VEALYL, the peptides quickly collapse into a liquid disordered phase, driven by strong side-chain interactions. Subsequently, the peptides polymerize to form a liquid of randomly oriented dimers, trimers, etc.: a polymerized or associated liquid [207]. This polymerized liquid appears almost frozen as the small clusters are kinetically arrested. In the thermodynamically preferred pathway these clusters must either dissolve and reassemble in the correct way, or diffuse and align to form the fibril. However, the extremely slow intrinsic dynamics of the polymerized liquid at 240 K would render such a pathway too long for path simulations. At this temperature, the polymerized phase is essentially a kinetic trap and the system is not able to escape within a reasonable (simulation) time. In order to force the system to follow an accessible kinetic pathway of fibril formation, we therefore raised the simulation temperature to 285 K. This reduced the degree of polymerization and led to faster dynamics in the oligomer droplet (Fig. 6.8). As the fibrillar aggregate consists of two parallel $\beta$-sheets, we set the in-register contacts in the two largest clusters as the order parameter. The initial and final state are defined by boundaries $\lambda_A = 5$ and $\lambda_B = 44$ respectively, with a spacing of $\Delta \lambda = 3$. Using the number of in-register contacts in the two largest clusters as the order parameter, the system was able to reach the final interface consistently.

From an analysis of the committor for the $A \rightarrow B$ transition we again obtain the distribution of the order parameter in the TSE, shown in Fig. 6.9a. As expected, the TSE structures are located on the interfaces near boundary $B$. Table 6.2 lists the computed crossing probabilities, flux and rate constants. The reverse reaction rate $k_{BA} = 68.4 \, s^{-1}$ is now approximately 6 times faster than the forward reaction rate $k_{AB} = 11.7 \, s^{-1}$, and the corresponding steady-state probabilities are $\rho_A = 0.854$ and $\rho_B = 0.146$. The free energy for fibril formation is $\Delta F_{AB} = -k_B T \ln \frac{k_{BA}}{k_{BA}} = +1.77 \, k_B T$, indicating that the two-layered fibril is indeed thermodynamically metastable at 285 K.

Fig. 6.9b shows the calculated free energy as function of the number of in-register contacts in the two largest clusters. We can identify various free
Figure 6.8: Cluster size distribution at various temperatures obtained from the REMD simulations in Ref. 207. The occurrence of small clusters decreases with increasing temperature.

Figure 6.9: (a) Distribution of the order parameter of the TSE configurations for $A \rightarrow B$ transition of the 12-peptide KLVFFAE system. (b) Free energy versus in-register contacts in the two largest clusters.

energy minima. The left-most minimum at $N_C = 4$ contacts corresponds to a disordered oligomer droplet consisting of one dimer and 10 monomers. A large peak separating the first and second minimum is likely to correspond to the formation of the critical nucleus of three aligned peptides, in line with results from higher resolution MD simulations in earlier work [162]. Two possible oligomer compositions contribute to the minimum at $N_C = 8$: a droplet comprising of a trimer and nine monomers, or containing at least two dimers. The minimum at $N_C = 12$ is due to oligomers containing either one tetramer and eight monomers, or a trimer and at least one dimer. Oligomer structures corresponding to the minimum at 16 contacts...
consist of either 1) one pentamer and 7 monomers, 2) a tetramer and at least one dimer, or 3) at least two trimers. The free energy minima at $N_C = 8, 12$ and 16 contacts therefore correspond to different configurations in the polymerized phase. The right-most minimum contains contributions from oligomer configurations with $29 \lesssim N_{C_{\alpha}} \lesssim 37$ contacts in the two largest clusters. These highly ordered configurations generally consist of at least two aligned tetramers (see Fig. 6.9b, inset) and are therefore close in structure to the fibril state. The minimum is about $2 k_B T$ above the free energy of the disordered state, consistent with the rate constant estimates. The free energy maximum at $\lambda \approx 20$ disagrees with the distribution of the order parameter in the TSE in Fig. 6.5a, and the free energy barrier (about $6 k_B T$ and $4 k_B T$ for the forward and backward transition, respectively) is again much lower than the rate constants indicate, which is deemed a limitation of the choice of the order parameter $N_{C_{\alpha}}$.

Fig. 6.10 shows the fraction of growing and shrinking clusters as function of the cluster size in the transition path ensemble. The growth probability of the trimer is (only slightly) lower than that of the dimer (Fig. 6.10a), suggesting that the trimer indeed acts as the critical nucleus ($n^* = 3$). Again, we stress that only the reactive paths are taken into account here. We can attribute the fact that the shrinking probability is always greater than the growth probability to the elevated simulation temperature. The stability of clusters (Fig. 6.10b, green line) further decays for $n > 4$, and plateaus around $n = 8$, indicating that these cluster sizes are not very stable.

Inspection of the reactive pathways in the transition path ensemble led to a visualization of the various steps in the aggregation mechanism in Fig. 6.11. First, hydrophobic side-chain interactions drive the formation of a disordered liquid droplet (b) from solvated peptides (a). Polymerization of peptides in the droplet leads to the formation of a polymerized peptide liquid (c). The small clusters in this polymerized phase must either dissolve and reassemble, or diffuse and align to form a single large $\beta$-sheet consisting of 4 or 5 peptides (d). Due to the slow intrinsic dynamics of the polymerized liquid both these processes are slow. Subsequently, peptides align with the first sheet to form a second sheet, with the remaining peptides still dissolved in the droplet (e). This conformation maximizes the contact surface between the strongly hydrophobic phenylalanine residues at the centers of the chains, thereby forming a hydrophobic core along the fibril axis. The remaining peptides align to form a flat two- or three-layered fibril (f), with layers sliding and rotating relative to each other (g), which is likely due to the elevated simulation temperature.

The fibrillar structure of the Amyloid-$\beta_{16-22}$ (KLVFFAE) segment has been studied using solid-state NMR spectroscopy [218] and MC simulations [168] and was found to be mostly in anti-parallel arrangement, whereas here
we exclusively find parallel $\beta$-sheets. We have attributed this to the lack of electrostatic interactions at the N- and C-terminal residues of KLVFFAE in the model by Bereau and Deserno [41] in earlier work [207]. The two-step nucleation mechanism of KLVFFAE was also clearly observed in experimental work employing fluorescence imaging [219], where a hydrophobic collapse drives the formation of densely packed aggregates: a state critical for amyloid nucleation [220]. In addition, the size of the critical nucleus corresponds with previous simulation results [162], however we must be careful to draw such a conclusion since here the nuclei growth analysis was performed on the reactive pathways only.
6.4 Conclusion

We performed forward flux sampling simulations of the fibril-forming peptide sequences GNNQQNY, KLVFFAE and VEALYL using the mid-resolution coarse-grained model of Bereau and Deserno [41]. From the REMD simulations in Ref. 207 we obtained definitions for the stable states, the solvated (disordered) state and the fibril nucleus state, in terms of the number of in-register contacts in the (two) largest cluster(s). In agreement with previous work, we found that the fibril nucleation behavior is indeed in large part determined by the overall side-chain hydrophobicity: the nucleation mechanism switches from one-step- to two-step nucleation with increasing hydrophobic strength. The details of the mechanism and the fibrillar structure depend on the relative side-chain hydrophobicity along the peptide sequence.

For the weakly hydrophobic GNNQQNY peptide, where fibril formation is dominated by hydrogen bond formation rather than hydrophobic interactions, strongly hydrophobic tyrosine residues drive the docking of peptides directly from solution (one-step nucleation) until a critical nucleus is formed, consisting of a four peptide β-sheet ($n^* = 4$). Beyond this critical nucleus, the sheet is more likely to grow than shrink. However, the stability steeply declines for $n > 7$, thereby indicating that large one-layered fibrils of GNNQQNY are not stable. Possible explanations for the low stability are 1) at the used simulation temperature only half of the system is aggregated, 2) monomer depletion effects and 3) the thermodynamically most stable structure is a two-layered fibril rather than a one-layered fibril. The formation of the two-layered fibril is often preceded by the formation of small β-sheets that align in-plane with the tyrosine residues, thus forming a hydrophobic core along the fibril axis.

The VEALYL peptides form fibrils via a two-step condensation and ordering mechanism (2SN). First, strong hydrophobic interactions drive the formation of a disordered oligomer droplet. Inside this oligomer peptides align to form the critical nucleus, a dimer or a trimer ($n^* = 2, 3$), which grows into a β-sheet consisting of four peptides. This sheet serves as a hydrophobic surface for the peptides still dissolved in the droplet, and a second β-sheet forms parallel to the first via self-templated nucleation. The formation of the third β-sheet proceeds in a similar manner, but at this stage the fibril undergoes a structural change to maximize hydrophobic interactions between the C-terminal leucine and tyrosine residues. The fully formed fibril consists of three β-sheets hooked into each other through the C-terminal residues, which now form a hydrophobic core along the fibril axis.

The kinetic pathway of KLVFFAE fibril formation was only accessible
for an elevated temperature (above the ordering temperature), due to the slow dynamics in the low temperature polymerized liquid oligomer. The reaction scheme of KLVFFAE suggests two possible pathways. In the kinetically accessible high temperature pathway, fibril formation follows a two-step condensation and ordering mechanism (2SN) where peptides first condense into a disordered liquid. Following the formation of the critical nucleus, a $\beta$-sheet consisting of three peptides ($n^* = 3$), fibril formation proceeds in a highly cooperative manner: the first sheet serves as a template for the formation of the second sheet. The fibril consists of two flat parallel $\beta$-sheets in which hydrophobic interactions between the phenylalanine residues are maximized, thereby forming a hydrophobic core along the fibril axis.

In the low temperature pathway (below $T_o$), peptides are likely to aggregate following a condensation, polymerization and ordering mechanism (2SN + polymerization). First, peptides coalesce into disordered oligomers driven by hydrophobic interactions. KLVFFAE peptides are of intermediate hydrophobicity, therefore hydrogen bond formation plays an important role in the aggregation processes inside the droplet [207]. This leads to the formation of the polymerized phase which consists of several slowly diffusing peptide clusters. The conversion of the polymerized droplet into the fibril is very slow as it requires the clusters to diffuse and align or dissolve and reassemble to form the first $\beta$-sheet.

Our results indicate that an increase in the relative hydrophobic- to hydrogen bond strength switches the nucleation pathway from 1SN to 2SN, and that for intermediate hydrophobicity an additional polymerization step in the 2SN mechanism is possible. The relative side-chain hydrophobicity along the peptide chain (i.e. the sequence) plays an important role in the details of the reaction mechanism and the structural properties of the fibril: there is a strong tendency of all peptides to form a hydrophobic core along the fibril axis. The side-chain hydrophobicity is what differentiates peptides, whereas the hydrogen bond contribution can be considered sequence independent. Hydrogen bond formation not only stabilizes the assemblies, but also induces polymerization in disordered oligomers, which slows down the nucleation process. Polymerization of disordered oligomers may thus lead to an increased lifetime of early oligomers, which are known to be involved in neurodegenerative disorders such as Alzheimer’s disease [23, 25].

The model in combination with FFS simulations has been successful in elucidating the nucleation mechanism as function of the side-chain hydrophobicity, We note that the reaction kinetics have been derived using an implicit-solvent four-beads-per-amino-acid CG model, therefore the free energy landscape is smoothed and the timescales presented will be faster than those obtained experimentally or from atomistic simulations. While a
simple scaling of the kinetics is probably not possible, the model is nevertheless able to give qualitative insight in the nucleation mechanism. However, molecular simulations using a higher resolution model are recommended to resolve the structural details of the polymerized phase and the various aggregated states of amyloid fibrils. Finally, in FFS the equations of motion are integrated forward in time only and might lead to unwanted dependence on the initial configurations. This is especially the case when the order parameters used to describe the interfaces are not optimal. The discrepancy between the location of the TSE of the forward transition and the maximum of the free energy reveals that this might indeed be the case. Time reversible methods such as TPS [93–97] or TIS [98, 99] may ameliorate this situation in a future study.