Universidade Federal de Juiz de Fora
Programa de Pós-graduação em Modelagem Computacional
219011 - Introdução à Mecânica Molecular (2019)

Estudo de agregação do surfactante dodecilfosfocolina (DPC) por dinâmica molecular: desenvolvimento de um novo protocolo para analisar computacionalmente a formação de agregados.

Autores: *Nícolas Glanzmann, nicolasglanz@gmail.com; Ruan Medina Carvalho, ruan.medina@engenharia.ufjf.br;*
Orientação*: Diego Enry Barreto Gomes, diego.enry@gmail.com; Priscila Vanessa Zabala Capriles Goliatt, priscila.capriles@ufjf.edu.br.*

Juiz de Fora, Novembro de 2019

Sumário do Trabalho

1. Introdução

Os compostos químicos capazes de diminuir a tensão superficial e/ou de interface do líquido no qual estão dissolvidos são classificados como surfactantes ou tensoativos, segundo definição da IUPAC, e são de grande importância pois em geral apresentam um comportamento em solução que permite amplas aplicações na área farmacêutica, industrial e ambiental (STEPHENSON et al., 2006). A principal aplicação industrial é o uso em produtos de limpeza e higiene pessoal. Contudo, a propriedade detergente destes compostos também confere aos mesmos aplicações mais tecnológicas como por exemplo seu uso para permitir a solubilização de macromoléculas ou moléculas hidrofóbicas. Neste aspecto, a dodecilfosfocolina (DPC, Figura 1) é um detergente que foi descrito como um bom surfactante para a análise em solução de proteínas de membrana por ressonância magnética nuclear (TIAN et al., 2000).

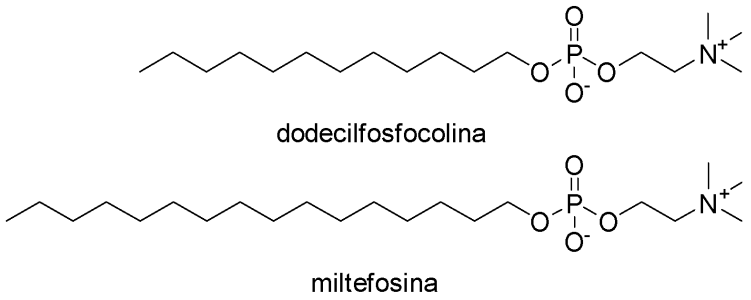

Figura 1. Estrutura dos surfactantes dodecilfosfocolina e miltefosina.

Sais orgânicos com cadeia lateral longa como o DPC tendem a se organizar quando dispostos em soluções aquosas, protegendo a parte hidrofóbica da molécula por meio de uma camada formada com a parte hidrofílica das moléculas e, nesse contexto, podem apresentar diversas formas de organização molecular como monocamadas ou bicamadas de compostos, vesículas ou lipossomos e micelas (DEAMER et al., 2002). Se o ambiente estiver concentrado o bastante, espera-se uma conformação micelar (em nanoesferas) para agrupamentos de moléculas de DPC por comparação com resultados observados para moléculas com propriedades similares na literatura (MARRINK; TIELEMAN; MARK, 2000), (PHILLIPS et al., 2005), (YOSHII; OKAZAKI, 2006), (YOSHII; IWAHASHI; OKAZAKI, 2006) e (LEBECQUE et al., 2017).

Marrink e colaboradores, no ano de 2000, foram capazes de simular por dinâmica molecular a auto agregação de moléculas de DPC formando uma única micela. (MARRINK; TIELEMAN; MARK, 2000). Neste trabalho, os átomos foram representados de forma reduzida, criando superátomos do tipo CH3, CH2 e etc. A simulação foi realizada utilizando o campo de forças do GROMACS, moléculas de água do tipo SPC e, no que diz respeito à concentração de surfactante, os valores foram retirados de um trabalho anterior no qual foi estudada a estabilidade de *clusters* de diferentes tamanhos (TIELEMAN; VAN DER SPOEL; BERENDSEN, 2000).

Alguns anos depois, em 2006, Stephenson e colaboradores (STEPHENSON et al., 2006) realizaram uma série de simulações envolvendo vários surfactantes, dentre eles, o DPC. Em seu trabalho, para os cálculos envolvendo DPC, eles utilizaram os mesmos parâmetros utilizados por Tieleman e Marrink (TIELEMAN; BERENDSEN, 1996; MARRINK;

TIELEMAN; MARK, 2000) e utilizaram uma estrutura termodinâmica para descrever a solução micelar. Desta forma, obtiveram valores de concentração micelar crítica (CMC) e número de agregação ($<N>_w$) tanto baseando-se em cargas atômicas do tipo ChelpG, para identificar a cabeça e cauda, quanto em cargas atômicas do tipo OPLS-AA, sendo que o método OPLS-AA forneceu valores muito próximos aos dados experimentais (Tabela 1).

	CMC teórico	**$<N>_w$ teórico**
ChelpG	0,24 mM	56
OPLS-AA	0.95 mM	48
Experimental	1,0 mM	44 ± 5

Tabela 1. Resultados Obtidos por Stephenson e colaboradores (STEPHENSON et al., 2006) ao replicar as condições de Marrink e colaboradores (MARRINK; TIELEMAN; MARK, 2000).

Para calcular experimentalmente a CMC, é necessário plotar um gráfico de uma propriedade física adequada em função da concentração do surfactante, sendo que uma mudança brusca na curva (ponto de inflexão) representa a CMC. Esta propriedade física pode ser tensão superficial, condutividade elétrica ou solubilidade de alguma substância marcadora que apresente uma banda particular no espectro de infravermelho ou ultravioleta-visível, por exemplo (DOMINGUÉS et al., 1997). Por Sua vez, o $<N>_w$ pode ser determinado experimentalmente aplicando diversas técnicas ao surfactante na CMC, como a calorimetria de titulação isotérmica (OLESEN; HOLM; WESTH, 2014), extinção de fluorescência (TEHRANI-BAGHA et al., 2012) e espalhamento de luz (THÉVENOT et al., 2005).

Tomando como base os dados experimentais, é esperado que, acima da CMC (1,0 mM), moléculas de DPC em solução aquosa deveriam assumir a forma de uma micela contendo 44 ± 5 moléculas com estrutura semelhante à micela representada na Figura 2. Outras formas que moléculas anfifílicas como a DPC tendem a se aglomerar são vesículas ou lamelas, como também pode ser observado na Figura 2. A estabilidade destas estruturas advém do fato de que são capazes de proteger a parte apolar da molécula de interagir com o solvente polar, favorecendo a interação de Van der Waals entre as cadeias, enquanto a cabeça polar fica em contato com o solvente. No trabalho de Marrink e colaboradores (MARRINK; TIELEMAN; MARK, 2000) a concentração utilizada (120 mM) foi muito superior à CMC e foi observada a formação de uma única micela contendo 54 moléculas de DPC. Este resultado é coerente pois, apesar de apresentar algumas moléculas a mais que o esperado compondo a micela, estas moléculas a mais não teriam condições de formar uma nova micela e, por isto, seria mais estável se incorporar em uma micela um pouco maior.

Estudos de agregação molecular como estes são de interesse devido a semelhança entre o DPC e a miltefosina (Figura 1), que são surfactantes análogos com variação no comprimento da cadeia carbônica lateral, sendo que a miltefosina é, também, um fármaco utilizado no combate à leishmaniose e diversas novas drogas são sintetizadas baseadas na estrutura deste composto. Nesse âmbito, o comportamento da miltefosina e derivados em água está diretamente relacionado à atividade biológica e o estudo da agregação de moléculas de DPC é um ponto de

partida para desenvolver métodos de simulação e análise de dados aplicáveis à miltefosina e compostos análogos.

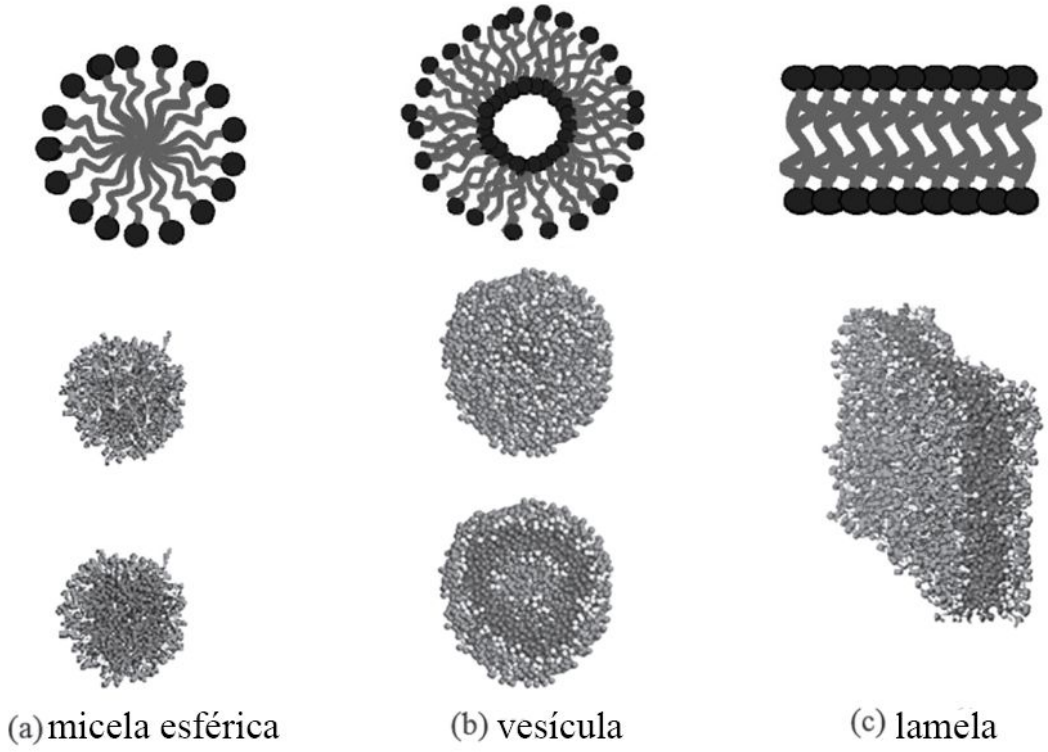

Figura 2. Estrutura de diferentes agregados moleculares (CHEN *et al.*, 2016; NISTICÒ; SCALARONE; MAGNACCA, 2017).

2. Objetivos

Recriar os sistemas de simulação de Marrink e colaboradores (2000) no intuito de desenvolver um protocolo similar, porém *all-atom*, para simular e analisar a auto agregação de moléculas de DPC por dinâmica molecular.

- Testar o uso de programas como Amber, PackMol, Antechamber, entre outros, em uma proposta de protocolo de estudo de agregação molecular.
- Descrever protocolo geral de forma a ser utilizado em análises de simulações de moléculas semelhantes - análogos não clássicos de Miltefosina - que apresentam menos dados experimentais.
- Comparar algoritmos de caracterização da formação de *clusters* por métodos de reconhecimento de padrões não supervisionados e comparar os métodos desenvolvidos com os resultados obtidos pela ferramenta de agrupamento molecular do GROMACS.

3. Material e Métodos

Nessa seção, apresenta-se o processo de desenho de molécula, parametrização molecular e simulação do sistema de moléculas de DPC em água. Apresenta-se também as estratégias desenvolvidas para o acompanhamento dos agrupamentos moleculares. Todos os arquivos dos passos descritos na metodologia e dos algoritmos desenvolvidos estão disponíveis para *download* no repositório online github.com/ruanmedina/DM-Self-Assembly-DPC.git.

3.1. Preparação do sistema

O primeiro passo do processo foi criar a molécula de DPC utilizando o programa `MSketch`. Para isso, deve-se considerar o estado de protonação desejado para a simulação e dessa forma, deve-se garantir que todos os átomos de hidrogênio foram inseridos corretamente. A estrutura tridimensional foi gerada e otimizada no programa, considerando apenas a disposição espacial dos átomos. O resultado do processo foi salvo em um arquivo `.SDF`.

A seguir, no segundo passo, com o arquivo de coordenadas otimizado, partiu-se para a parametrização da molécula. Utilizou-se o programa Antechamber para designar os tipos de átomos apropriados para o campo de forças `GAFF2`. Uma vez parametrizados, a estrutura foi otimizada considerando o método semi-empírico `AM1-BCC`. A seguir, as carga de cada átomo foi calculada com o modelo *Bond-Correction Charges* (`BCC`). O resultado do processo foi salvo em um arquivo `.mol2`.

O terceiro passo do processo consistiu em gerar o arquivo de parâmetros do campo de força `GAFF2` para a simulação. Para isso utilizou-se o programa `parmchk2`. O programa consegue designar os parâmetros necessários a partir do arquivo `.mol2`. Os parâmetros do modelo de simulação são salvos em um arquivo `.frcmod`.

As etapas posteriores demandam como entrada um arquivo de coordenadas em formato `.pdb`. Assim, o quarto passo do processo é utilizar o programa `Antechamber` novamente para gerar um arquivo nesse formato a partir do arquivo `.mol2`.

Para determinar caixas de simulação é necessário definir as quantidades de moléculas do solvente e do soluto a serem inseridas. Em alguns casos, pode ser necessário calcular grandezas como a massa molar de suas moléculas. O quinto passo consistiu em definir o modelo de solvente (*i.e.* água) a ser utilizado e calcular a massa molar das moléculas envolvidas. Para isso, utilizou-se do programa `OpenBabel` que recebe os arquivos `.pdb` do solvente e do soluto e retorna a massa molar de cada um deles. Com isso, pode-se efetuar cálculos para definir as configurações para a concentração desejada. Pode-se utilizar ainda, se necessário, o programa `Volume Guesser` para definir o tamanho necessário de caixa de simulação para obter concentrações esperadas**.**

O sexto passo consistiu na criação da caixa de simulação que garanta a concentração soluto/solvente desejada. Para este fim, foi utilizado o programa `PACKMOL`. O arquivo de configuração do programa exige dados do número de moléculas de solvente e soluto e as dimensões da caixa de simulação. Esses dados podem ser adquiridos pelo passo anterior. Além disso, o programa também exige um parâmetro de tolerância informando a distância mínima

entre moléculas inseridas na caixa de simulação (definiu-se `tolerance=2.0`). Na ocasião da simulação deste trabalho, esses dados puderam ser obtidos diretamente do trabalho de Marrink e colaboradores (2000). Dessa forma, definiu-se 22496 moléculas de água e 54 moléculas de DPC em uma caixa de 90 Å³. O arquivo `.inp` de entrada deve conter as seguintes instruções:

```
# A mixture of water and DPC

tolerance 2.0

# The file type of input and output files is PDB
filetype pdb
# The name of the output file
output mixture.pdb

structure water.pdb
  number 22496
  inside box 0. 0. 0. 90. 90. 90.
end structure

structure DPC_gaff.pdb
  number 54
  inside box 0. 0. 0. 90. 90. 90.
end structure
```

O sétimo passo consistiu em criar um arquivo de biblioteca para a posterior geração do arquivo de topologia da molécula estudada (i.e. DPC) compatível com o AMBER. Dessa forma, cria-se e executa-se um arquivo "`tleap_lib.in`" contendo as seguintes configurações:

```
# Carregar o GAFF2
source leaprc.gaff2
# Carregar os parâmetros GAFF da molécula DPC
loadAmberParams DPC_gaff.frcmod
# Carregar as coordenadas e ligações
MOL = loadMol2 DPC_gaff.mol2
# Salvar a topologia de referência para a molécula
saveOff MOL DPC_gaff.lib
```

Finalmente, o oitavo passo consistiu na preparação dos arquivos de topologia e coordenadas a serem utilizados pelo AMBER. Para isso, cria-se e executa-se um arquivo "`tleap.in`" contendo as seguintes configurações:

```
# GAFF2
source leaprc.gaff2
# Água TIP3P
source leaprc.water.tip3p
# Íons
```

```
loadAmberParams frcmod.ionsjc_tip3p

# Carregar os parâmetros GAFF a molécula
loadAmberParams DPC_gaff.frcmod

# Carregar a biblioteca referência para a molécula
loadOff DPC_gaff.lib

# Carregar o sistema
caixa = loadpdb mixture.pdb
saveAmberParm caixa caixa.prmtop caixa.inpcrd
savePDB caixa caixan.pdb

charge caixa

# Inserir o número suficiente de íons para
neutralizar
AddIons2 caixa Cl- 0
AddIons2 caixa Na+ 0

# Salvar sistema com íons: Topologia (prmtop) e
Coordenadas (rst7)
saveamberparm caixa ionized.prmtop ionized.rst7

quit
```

Utilizando o comando **`tleap -f tleap.in`** é possivel executar os comandos dos arquivos `.in` criados e assim obtemos os arquivos de coordenadas `(.rst7)` e topologia `(.prmtop)` necessários para a execução do programa de simulação `AMBER`.

3.2. Preparação dos arquivos de simulação

A simulação de dinâmica molecular contou com diversas etapas: Minimização de energia, aquecimento, densidade, equilibração e produção. Cada etapa da simulação será discutida a seguir.

A minimização de energia foi realizada em duas fases idênticas, sendo a única diferença que a primeira manteve as moléculas de DPC fixas e a segunda não. As minimizações contaram com 500 passos de minimização pelo método *steepest descent* e em seguida, mais 500 passos pelo método *conjugate gradient*. É interessante começar utilizando o método *steepest descent* , pois o mesmo é capaz de eliminar rapidamente possíveis regiões de tensão no sistema, todavia, este método converge lentamente próximo a um mínimo de energia e, por isso, é conveniente alternar para o método *conjugate gradient*, que é capaz de convergir mais rapidamente. A distância de *cutoff* utilizada foi de 8,0 angstroms para a minimização e todas as demais etapas.

O aquecimento consistiu em um aumento da temperatura de 290 K até 300 K, que ocorreu em 25000 passos de 0,002 ps. O termostato de Langevin foi utilizado com frequência de colisão 2,0 e condições periódicas de fronteira foram utilizadas para manter o volume constante,

porém, sem controle de pressão. As ligações envolvendo hidrogênios foram restringidas pelo SHAKE e não tiveram forças calculadas.

A etapa de densidade partiu da temperatura de 300 K atingida na etapa anterior e apresentou as mesmas configurações da etapa anterior, diferindo apenas na utilização de condições periódicas de fronteira com pressão constante por meio de um barostato de Berendsen com tratamento isotrópico e tempo de relaxação de 1,0 ps. O mesmo termostato e barostato utilizados nos dois últimos passos foram mantidos até o final da dinâmica molecular.

Assim como a minimização, a equilibração também ocorreu em duas fases idênticas, a primeira com moléculas de DPC fixas e a segunda, livres. Cada fase foi executada em 250000 passos de 0,002 ps e, neste caso, foi permitido um tempo de relaxação de 2,0 ps para o barostato.

Por fim, para a produção, foram mantidas as mesmas configurações da equilibração livre e o desenvolvimento espontâneo do sistema foi simulado em 100 partes, cada uma contendo 500000 passos de 0,002 ps. No total, a produção atingiu 100 ns.

3.3. Descrição de diferenças com Marrink (2000)

O trabalho realizado no ano de 2000 por Marrink e colaboradores apresentou algumas condições muito semelhantes ao que foi utilizado neste trabalho, por exemplo, o tamanho da caixa de simulação, o número de moléculas de água e de DPC, a temperatura e a pressão. Entretanto, a metodologias se diferem em questão de parâmetros de campo de força para as moléculas e nos métodos utilizados para a integração numérica. Além disso, no trabalho de Marrink, foi utilizado o campo de forças `GROMACS`, utilizando certo nível de *coarse grained* com representação implícita dos átomos de hidrogênio, enquanto o campo de forças `Amber`, todos os átomos do sistema foram considerados separadamente. Ademais, o `Amber` apresenta alto nível de otimização computacional que proporciona aumento no número de operações matemáticas efetuadas por unidade de tempo com relação ao `GROMACS` (ABRAHAM et al., 2019; CASE et al., 2019).

Em relação ao tipo de água utilizado, a propriedade diretamente relacionada ao comportamento de surfactantes é a tensão superficial, cujo valor experimental a 300 K é de 71,73 mJ m^{-2}. O modelo utilizado por Marrink (SPC) fornece um valor de tensão superficial de 54,6 mJ m^{-2}, enquanto o modelo utilizado nesta simulação (TIP3P) fornece um valor de 52,3 mJ m^{-2}. É possível notar que a representação desta propriedade é comparável nestes dois modelos, sendo o modelo SPC ligeiramente mais próximo do valor experimental, mas ainda assim, é possível notar que ambos apresentam valores muito baixos. Uma opção que melhor representaria esta propriedade seriam os modelos SPC/E, TIP4P/Ew e TIP4P/2005, que apresentam tensão superficial entre 63 e 69 mJ m^{-2}, se aproximando mais do resultado experimental (VEGA; MIGUEL, 2007).

Ainda, no trabalho publicado em 2000, foram utilizados os algoritmos LINCS para resolver os sistemas de equação não linear através de matrizes e SETTLE para resolver de forma analítica os sistemas de equação não linear de moléculas de água rígida. Em contraponto, neste

estudo foi utilizado o algoritmo SHAKE, que resolve os sistemas de equação não linear de forma iterativa pelo método de Gauss–Seidel, que apesar de ser mais custoso computacionalmente que o LINCS, proporciona resolver sistemas com maiores graus de liberdade em simulações com parametrização explícita de todos os átomos do sistema (ABRAHAM et al., 2019; CASE et al., 2019).

3.4. Análise de agrupamentos moleculares

Descreve-se nesta subseção a estratégia proposta para analisar os agrupamentos moleculares formados durante a dinâmica molecular de DPC. Uma vez que os agrupamentos moleculares observados variam os seus tamanhos e conformações durante a simulação, cabe-se acompanhar o número de moléculas que participam de cada estrutura estável na trajetória calculada. Neste trabalho, propõe-se acompanhar a evolução dessas organizações durante a dinâmica com o uso de algoritmos não supervisionados de agrupamento (clusterização). Como discutido por Johnston, M. A. et al. (JOHNSTON, M. A. et al, 2016), estratégias de estudo de formação de grupos moleculares deve levar em consideração o tamanho das moléculas em questão e informações de proximidade entre elas.

A primeira questão levantada é sobre a representação molecular escolhida. Usualmente, tende-se a simplificar a estrutura da molécula e a não trabalhar com todos os átomos da molécula durante a etapa de análise de agrupamentos. Neste cenário, poderia-se propor que cada molécula de DPC fosse representada pela posição espacial de um de seus átomos (*e.g.* um átomo de fósforo - P - por estar localizado na parte hidrofílica da molécula). Porém, o comprimento da molécula de DPC permite que haja agrupamentos de moléculas sem que suas cabeças hidrofílicas estejam extremamente próximas uma das outras.

Assim, para melhorar os resultados dos agrupamentos, foi proposto um modelo de representação computacional baseado no centro de massa de cada molécula de DPC. Essa estratégia aproxima os pontos de representação das moléculas em uma agrupamento molecular (*e.g.* micelas) para o centro de massa do próprio grupo. Para facilitar o processo computacional envolvido, os valores das coordenadas no espaço 3D foram normalizadas em valores de 0 a 1. Neste cenário, o número de grupos formados teria relação direta com o número de agregados observados e o tamanho dos grupos teria relação com o número de moléculas em cada agrupamento, valor que aproxima o número de agregação do composto em questão.

Partindo desse modelo de representação molecular, é necessário escolher um algoritmo de agrupamento. Existem diferentes abordagens computacionais para cálculos de agrupamentos de dados. Neste trabalho, foram aplicadas duas (2) estratégias distintas por meio dos algoritmos de agrupamento DBSCAN e Affinity Propagation. Os resultados obtidos por essas propostas serão comparados com os resultados obtidos pela ferramenta de agrupamento molecular do GROMACS.

Faz-se necessário testar diversas abordagens para acompanhamento de agregação molecular uma vez que cada algoritmo apresenta características distintas nos grupos formados. A Figura 3 apresenta os resultados obtidos no agrupamento de diferentes características de dados com 3 estratégias de agrupamento: 2.(a) os resultados obtidos pelo `DBSCAN`, 2.(b) os resultados obtidos pelo `Affinity Propagation`, e 2.(c) os resultados obtidos pelo

algoritmo `Ward` (que apresenta estratégia similar ao algoritmo utilizado pelo `GROMACS`). Percebe-se que os grupos formados possuem diferentes características que vezes facilitam o processo de identificação de agregados moleculares, e vezes podem contribuir para uma agregação errônea dos dados.

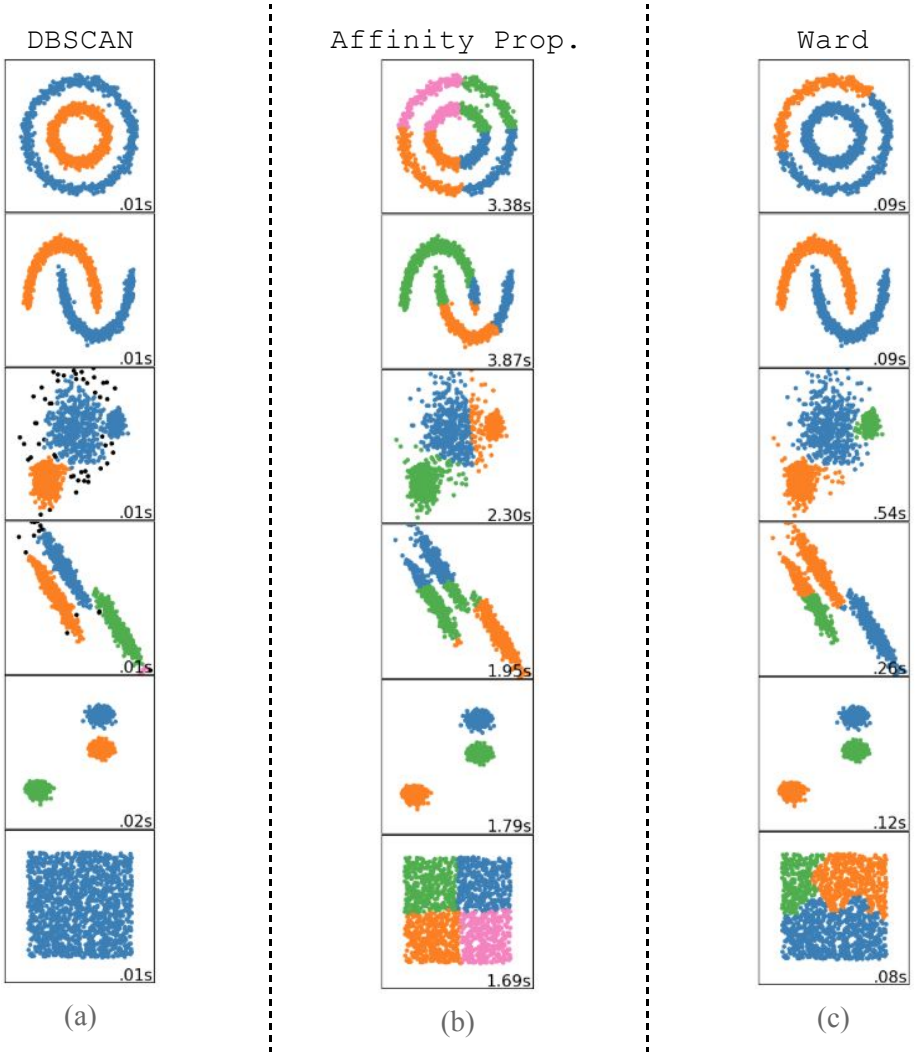

Figura 3. Resultados obtidos no agrupamento de diferentes características de dados com 3 estratégias de agrupamento: em (a) `DBSCAN`, em (b) `Affinity Propagation` e em (c) algoritmo hierárquico aglomerativo `Ward` (PEDREGOSA et al, 2011).

As características dos grupos formados é fator importante no problema tratado, uma vez que acompanhar o processo de formação de agregados moleculares apresentam grande variabilidade na distribuição das moléculas no espaço de simulação e diversos cenários que podem dificultar a identificação de grupos por estratégias computacionais. Com isso, busca-se investigar qual estratégia computacional, entre as testadas, apresenta melhores resultados e fácil adequação ao problema de acompanhamento de aglomeração molecular a priori. Mais especificamente, no caso da formação de micelas, que são estruturas de formato predominantemente esférico, é interessante encontrar um método capaz de reconhecer grupos que apresentam esse perfil, assim levando em conta não somente a proximidade entre os membros do grupo, bem como o formato do mesmo. Desta forma, um algoritmo ideal seria

capaz de observar duas micelas diferentes mesmo quando uma molécula de uma se encontra muito próxima a uma molécula da outra.

3.4.1. Abordagem baseada em densidade - DBSCAN

O DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um método baseado em densidade que trabalha com o conceito de " Regiões densas ", ou seja, concentrações elevadas de pontos. O algoritmo caracteriza um cluster como uma região densa cercada por uma região não densa. O algoritmo DBSCAN consegue encontrar agrupamentos de formas arbitrárias e é eficiente para grandes bancos de dados espaciais. O algoritmo busca *clusters* pesquisando a vizinhança de cada objeto e verifica se ele alcança, diretamente ou indiretamente, mais do que o número mínimo de objetos para formar um cluster (número mínimo de objetos e vizinhança de busca informadas pelo usuário - descritos a seguir); caso contrário, os pontos de dados são considerados *outliers* (ESTER, M. et al., 1996).

Como não se conhece nem o tamanhos das agrupamentos moleculares nem a quantidade de grupos a priori, o DBSCAN se mostra aplicável por não demandar previamente esses parâmetros. Além disso, o algoritmo permite a existência de *outliers*, não forçando a criação de grupos desnecessários e permitindo a identificação de moléculas que se desprendem de seus agrupamentos durante a simulação.

Dado o seu funcionamento, o algoritmo DBSCAN requer dois (2) parâmetros de entrada: `min_samples` que determina o mínimo de pontos para definir uma região densa (um cluster) e `eps` que determina o raio de vizinhança de busca com relação a cada ponto do sistema. O esquema do funcionamento do algoritmo é descrito na Figura 4.

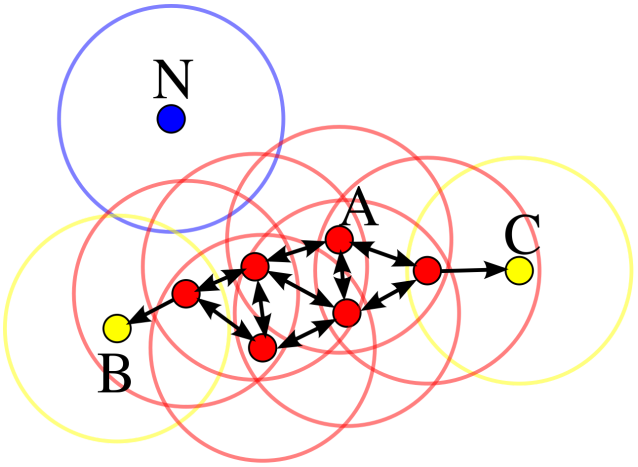

Figura 4. Esquema de funcionamento do algoritmo de agrupamento DBSCAN. Neste exemplo, `min_samples` = 4. O ponto **A** e os outros pontos vermelhos são pontos centrais, porque a área que circunda esses pontos em um raio `eps` contém pelo menos 4 pontos (incluindo o próprio ponto). Já que todos os pontos vermelhos são alcançáveis entre si, eles formam um único *cluster*. Os pontos **B** e **C** não são pontos centrais, mas são alcançáveis a partir de A (alcançáveis através de outros pontos centrais) e portanto, também pertencem ao *cluster*. O ponto **N** é um ponto de ruído que não é nem um ponto central nem diretamente acessível.

Os parâmetros necessários ao algoritmo permitem a adequação da estratégia de agrupamento ao problema de análise de agrupamentos moleculares. O primeiro parâmetro, de fato, ajuda no controle da contagem dos agrupamentos, pois é possível definir o tamanho mínimo dos agrupamentos de interesse. O segundo parâmetro apresenta maiores dificuldades na definição, pois exige um *cutoff* para a consideração de um ponto (molécula) em um grupo. Os valores utilizados partem de adaptações dos valores descritos por Marrink S. et al. (MARRINK; TIELEMAN; MARK, 2000) fixando `min_samples` = 2 e `eps` = 1.4.

3.4.2. Abordagem baseada em *Message Passing* - `Affinity Propagation`

O algoritmo Affinity Propagation é baseado no conceito de "Passagem de mensagem" entre pontos de dados. Cada um dos pontos de dados é um possível representante para um novo cluster. O método não necessita receber como entrada o número de agrupamentos como parâmetro de entrada. Para descobrir o número de clusters e formar os grupos de dados, uma função de similaridade é usada para concluir e atualizar uma matriz de similaridade em um processo iterativo (THAVIKULWAT, P., 2014), (NADIPALLI S., 2014).

O processo de cálculo inicia-se com o cálculo de uma matriz quadrada de similaridade S preenchida com o negativo da soma dos quadrados das diferenças de cada coordenada dos pontos. Onde X_i e X_j são as coordenadas dos pontos i e j.

$$s(i,j) = -\sum [|X_i - X_j|]^2$$

Uma matriz de avaliabilidade A é iniciada com zeros. O processo iterativo inicia-se calculando uma matriz de responsabilidade R pela seguinte fórmula:

$$R(i,k) = S(i,k) - max\{A(i,k') + S(i,k')\} \quad ; k = k'$$

A seguir, a matriz de avaliabilidade A é atualizada. Para a diagonal principal faz-se:

$$A(k,k) = \sum max\{0, r(i',k)\} \quad ; i' \neq k$$

Para o restante da matriz calcula-se:

$$A(i,k) = min\left\{0, r(k,k) + \sum max\{0, r(i',k)\}\right\} \quad ; i' \notin \{i,k\}$$

O processo acima continua sendo executado de forma iterativa até que o algoritmo chegue em um consenso para os valores das matrizes em questão. Por fim, uma matriz critério é calculada da seguinte forma:

$$C(i,k) = R(i,k) + A(i,k)$$

O valor mais alto do critério em cada linha é definido como valor exemplar do dado. Os dados que possuem os mesmos valores exemplares estão em um mesmo cluster.

3.4.3. Abordagem hierárquica do GROMACS - gmx-clustsize

Esta ferramenta computa a distribuição de clusters atômicos ou moleculares na fase gasosa. (ABRAHAM et al., 2019) O funcionamento é baseado em um estratégia de agrupamento hierárquico que trabalha para agrupar as moléculas próximas geometricamente até que a condição estabelecida pelo parâmetro de *cutoff* fornecida pelo usuário não possa ser mais atendida atendida ou parâmetros de controle sejam atingidos.

O método implementado pelo GROMACS se trata de uma estratégia hierárquica aglomerativa. Ou seja, o algoritmo inicia colocando cada uma das moléculas em um grupo unitário e itera comparando distância entre átomos ou o centro de massa das moléculas buscando juntar grupos que obedeçam um critério de conexão (*Linkage Criteria*). Esses critérios podem seguir diferentes regras. Por exemplo, pode-se computar distância entre grupos considerando as partes mais similares dos grupos (*single-linkage*), considerando as partes menos similares dos grupos (*complete-linkage*) ou o centróide (região central) dos grupos (*average-linkage*). O método implementado na função do GROMACS utiliza um critério diferente que compara distâncias entre pontos aleatórios de grupos diferentes, seguindo uma ordem predefinida dos pontos, e verifica se a distância entre esses pontos obedece *cutoff* informado pelo usuário.

O processo de cálculo de algoritmo gera um dendograma de união de grupos. É comum encontrar diversas estratégias de 'podagem' deste dendograma que buscam responder a pergunta: "Quando parar de unir os grupos em grupos maiores?". A estratégia utilizada pela ferramenta utiliza-se do parâmetro de cutoff para definir esse limite de aglomeração. Quando nenhuma das moléculas apresenta possibilidade de união, dado o critério de distância, com nenhuma outra molécula dos diferentes grupos formados até então, o algoritmo converge e retorna a resposta para o usuário.

Para a aplicação desta ferramenta, foi necessário converter a topologia e coordenadas geradas pelo AMBER para um formato compatível com o GROMACS utilizando o Python:

```
import parmed as pmd
mol = pmd.load_file('imaged.prmtop','imaged.rst7')
mol.save('gromacs.top',format='gromacs')
mol.save('gromacs.gro')
quit()
```

Em seguida, utilizando o programa GROMACS, gerou-se uma trajetória livre de moléculas de água com os seguintes comandos e o resultado foi salvo nos formatos do AMBER (.nc) e do GROMACS (.xtc):

```
cat <<EOF | cpptraj
parm ionized.prmtop
$(for i in $(seq 100) ; do echo "trajin
prod_${i}.nc" ;done)
strip :WAT
trajout imaged.nc
trajout imaged.xtc
go
quit
EOF
```

Feito isto, uma trajetória auxiliar para a análise foi gerada com saltos de 100 passos entre cada frame salvo na trajetória original para diminuir os cálculos. O arquivo .tpr de parâmetros e topologia do GROMACS foi preparado a partir dos arquivos .gro, .top e .mdp para enfim, realizar a análise com a ferramenta *clustsize* utilizando os parâmetros de condição periódica e o *cutoff* adequado. Os comandos desta etapa se encontram a seguir.

```
gmx trjconv -f imaged.xtc -o imaged_skip.xtc -skip 100
-timestep 2 -tu ps

gmx grompp -f grompp.mdp -c gromacs.gro -p gromacs.top -o
gromacs.tpr

gmx clustsize -s gromacs.tpr -f imaged_skip.xtc -mol -cut
1.8 -pbc
```

3.5. Detalhes de Implementação e condições periódicas

Exceto pela análise efetuada por meio do algoritmo do GROMACS, as análises de agrupamento propostas são executadas por um *script* em *Python3*. Foram utilizadas as bibliotecas *sklearn* para o pré-processamento e agrupamento dos dados e *mdtraj* para o auxílio na manipulação dos arquivos binários de trajetória (*.dcd*) com resultados da dinâmica. A estratégia de análise aplica o agrupamento em intervalos de 10 *frames*. Dessa forma, a amostragem de configurações do sistema é formada pelo conjunto de configurações do sistema em intervalos de 0,2 nanosegundos de dinâmica. Para isso, foi necessário ajustar os arquivos de trajetória do Amber para arquivos de trajetória binários na extensão .dcd. Utilizou-se o programa VMD (*Visual Molecular Dynamics*) para converter a trajetória de saída do Amber (.nc) para o formato desejado.

O algoritmo desenvolvido recebe um arquivo com os nomes dos arquivos de trajetória para auxiliar na ordem na qual os agrupamentos e gráficos de dispersão são gerados. O código inclui um dicionário com as massas dos átomos para cálculo do centro de massa de cada uma das moléculas em cada *frame* estudado. Os grupos observados são divididos em 5 classes, em uma adaptação da classificação proposta em por Marrink S. et al. (MARRINK; TIELEMAN; MARK, 2000): `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas). O *script* em `Python` em questão é apresentado no Apêndice deste trabalho.

Uma problemática na aplicação de algoritmos de agrupamento tradicionais para o problema de detecção de agrupamentos moleculares durante uma dinâmica molecular é o fato das simulações de DM utilizarem um espaço 3D de condições periódicas, como mostrado na Figura 5. Com isso, é possível que agrupamentos moleculares bem definidos se situem nas fronteiras da caixa de simulação e se partam na representação espacial. Como o algoritmo original não leva em consideração espaços periódicos, foi necessário propor uma alteração no método para que isso fosse levado em consideração.

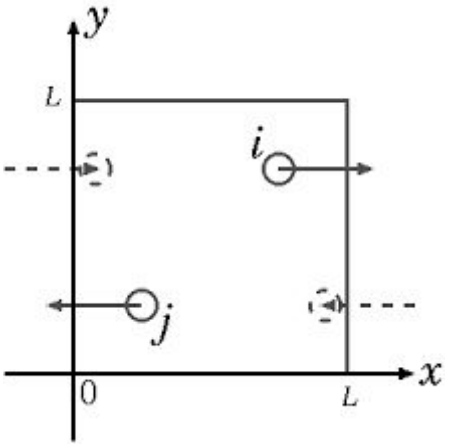

Figura 5. Esquema bidimensional de um sistema com condições periódicas de contorno. Nesse cenário, as partículas `i` e `j` conseguem transitar pela caixa de simulação delimitada atravessando as fronteiras e sendo re-inseridas na posição espelhada da caixa. Logo, partículas posicionadas em extremos opostos da caixa de simulação estão próximas uma das outras de tal forma que sofrem com a interação mútua.

A alteração proposta se baseia na definição de uma nova função de distância utilizada nos cálculos de vizinhança de cada molécula. A função define uma distância máxima que pode ocorrer entre duas moléculas na caixa de simulação normalizada e permite a identificação de distâncias que permeiam pela fronteira periódica. Para isso, utiliza-se da forma algébrica da função mínimo descrita abaixo. A função retorna o mínimo entre dois valores a e b informados.

$$min\{a,b\} = \frac{|a+b| - |a-b|}{2}$$

Para cada cálculo de distância entre dois pontos i e j, deve-se calcular a distância entre os centros de massa em cada uma das coordenadas p. Esse valor de distância já considera medidas que transpassam a caixa periódica por escolher o mínimo entre a distância entre as

moléculas dentro da caixa ou o valor complementar considerando um tamanho D de caixa periódica.

$$d_p(i,j) = min\left\{\left[|x_p(i) - x_p(i)|\right], \left[D - |x_p(i) - x_p(i)|\right]\right.$$

Substituindo a função mínimo definida anteriormente temos para cada coordenada p:

$$d_p(i,j) = \frac{|D| - |D - 2 * |x_p(i) - x_p(i)||}{2}$$

Assim, para cálculo geral temos que a função distância entre duas moléculas `i` e `j` proposta está disposta a seguir, onde $X_i(x_x(i), x_y(i), x_z(i))$ e $X_i(x_x(j), x_y(j), x_z(j))$ são os vetores de coordenadas dos centros de massa das moléculas.

$$d(i,j) = \sqrt{d_x^2(i,j) + d_y^2(i,j) + d_z^2(i,j)} = \sqrt{\sum_p^{\{x,y,z\}} d_p^2(i,j)}$$

3.5.1. DBSCAN

O agrupamento foi implementado em um espaço tridimensional por meio do DBSCAN com parâmetros `eps=1.4` e `min_samples=2`. A modificação no método tradicional implica na utilização da função de distância $d(i,j)$ no cálculo de vizinhos mais próximos. Os demais parâmetros são utilizados como descrito anteriormente.

3.5.2. Affinity Propagation

A alteração consiste de uma mudança na inicialização da matriz de similaridade. Ao invés de usar a diferença entre os valores de cada coordenada dos pontos na fórmula de inicialização de S, utiliza-se a expressão $d_p(i,j)$ que considera as distâncias periódicas das caixas de simulação. Dessa forma, temos a inicialização de S da seguinte forma:

$$s(i,j) = -\sum_p^{\{x,y,z\}} d_p^2(i,j)$$

O restante do processo iterativo segue como descrito anteriormente.

3.6. Extração de Energias Potenciais

O cálculo das energias do sistema foi realizado utilizando a ferramenta `cpptraj` do AMBER. Para isso, foi utilizado o arquivo `imaged.nc` que foi gerado anteriormente e contém

a trajetória de todas as partes da produção e não apresenta moléculas de água. O arquivo de parâmetros utilizado foi o `imaged.prmtop`. Foi escolhido o método *Particle Mesh Ewald* por ser um método eficiente e levar em conta as condições periódicas de fronteira. A sequência de comandos utilizada se encontra a seguir:

```
cat <<EOF | cpptraj
parm imaged.prmtop
trajin imaged.nc
energy imaged out ene.agr etype pme
go
quit
EOF
```

4. Resultados e Discussão

Nesta seção apresenta-se os resultados obtidos com as análises da dinâmica molecular realizada. As discussões estão divididas em análises qualitativas e quantitativas. As comparações entre os métodos de agrupamento são apresentadas, juntamente com uma breve demonstração de sensibilidade de parâmetros de entrada quando necessário.

Todos os arquivos de resultado estão disponíveis para *download* no repositório *online* github.com/ruanmedina/DM-Self-Assembly-DPC.git.

4.1. Análise Qualitativa

Nesta seção apresenta-se os resultados das análises qualitativas realizadas pela inspeção visual das trajetórias. Para isso, utilizou-se do programa VMD (*Visual Molecular Dynamics*) que possui estratégias de visualização científica para a bioinformática e biologia computacional.

Na Figura 6 o sistema pode ser observado em diferentes momentos da simulação. É possível observar que, na fase inicial, o sistema começa a se organizar com uma tendência de aproximação entre as moléculas de DPC e, rapidamente, a partir de 8 ns, já é perceptível a organização das moléculas em pequenos aglomerados, ainda com muitas moléculas livres ou em grupos menores. Além disso, analisando a Figura 7a (11 ns), foi constatado que a formação dos pequenos aglomerados se dá a partir da formação de pares (em vermelho), seguida da aproximação de conjuntos com 3 ou mais moléculas de DPC que alinham suas carbônicas formando pequenas lamelas (em azul). A partir de 20 ns, os 3 *clusters* majoritários já contém a maioria das moléculas do sistema e, em 43ns (Figura 7b) todas as moléculas de DPC já se encontram em um dos 3 *clusters*, que se mantém predominantemente estáveis na fase final da simulação estes *clusters* se mantém predominantemente estáveis até o final da simulação com pequenos desprendimentos e reincorporações.

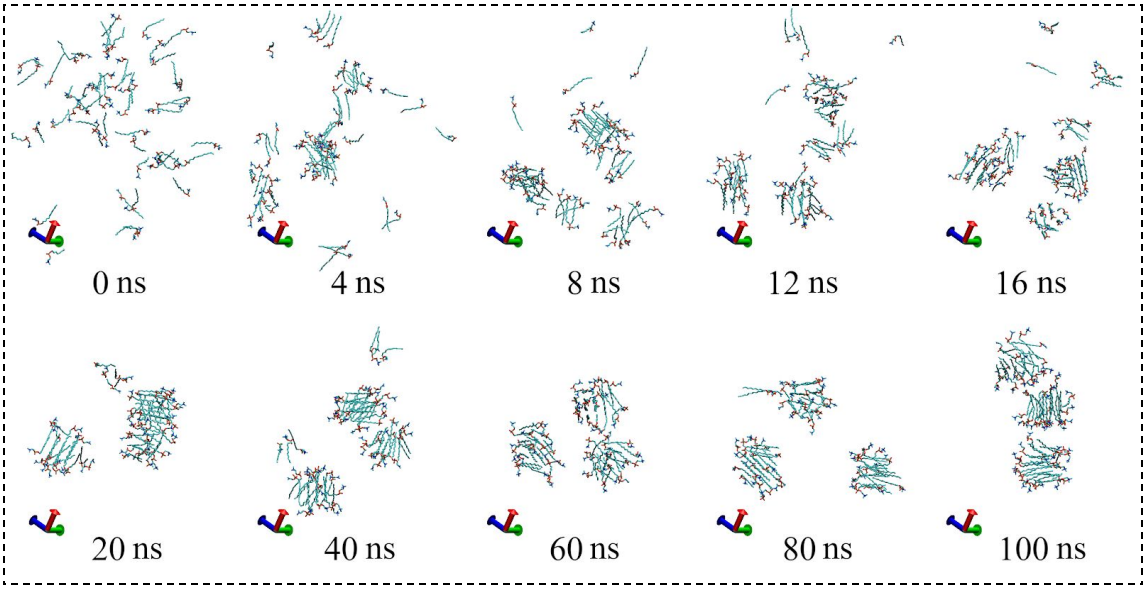

Figura 6. Moléculas de DPC em diferentes estágios da simulação.

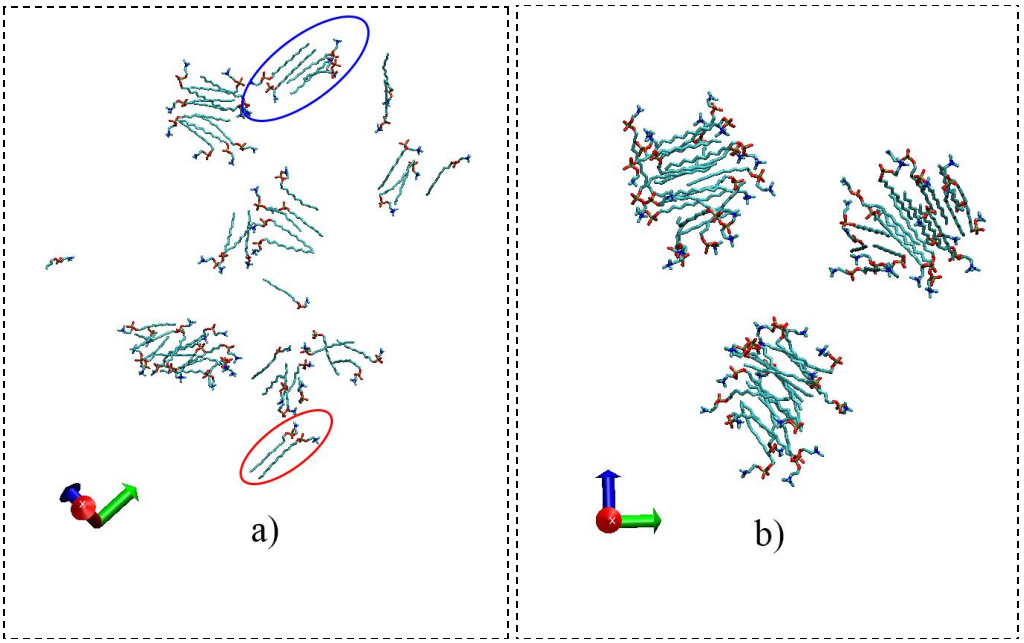

Figura 7. a) Moléculas de DPC em 11 ns de simulação evidenciando os pequenos aglomerados. b) Formação dos 3 *clusters* majoritários em 43 ns de simulação.

4.2. Análise Quantitativa

Nesta seção apresenta-se os resultados das análises quantitativas realizadas pelos algoritmos de agrupamento propostos com as modificações de distância periódica apresentadas em contraponto com os resultados obtidos por suas versões padrões e com a ferramenta de análise de agrupamentos do `GROMACS`.

4.2.1. DBSCAN

A Figura 8 mostra o gráfico de agrupamentos moleculares durante os 100ns da dinâmica efetuada calculados pelo algoritmo `DBSCAN`. O gráfico mostra a separação de classes definida na metodologia. É possível observar uma grande quantidade de grupos no início da dinâmica, principalmente de grupos compostos por poucas moléculas. Esse comportamento era esperado já que a simulação é iniciada com uma distribuição aleatória das moléculas. Nos primeiros 10ns de moléculas o comportamento do sistema é de diminuição do número de agrupamentos `small_clst` e aumento do número de agrupamentos `intermediate_clst` e `large_clst`. Esse comportamento indica uma aproximação entre os menores grupos para formar agrupamentos com maiores números de moléculas.

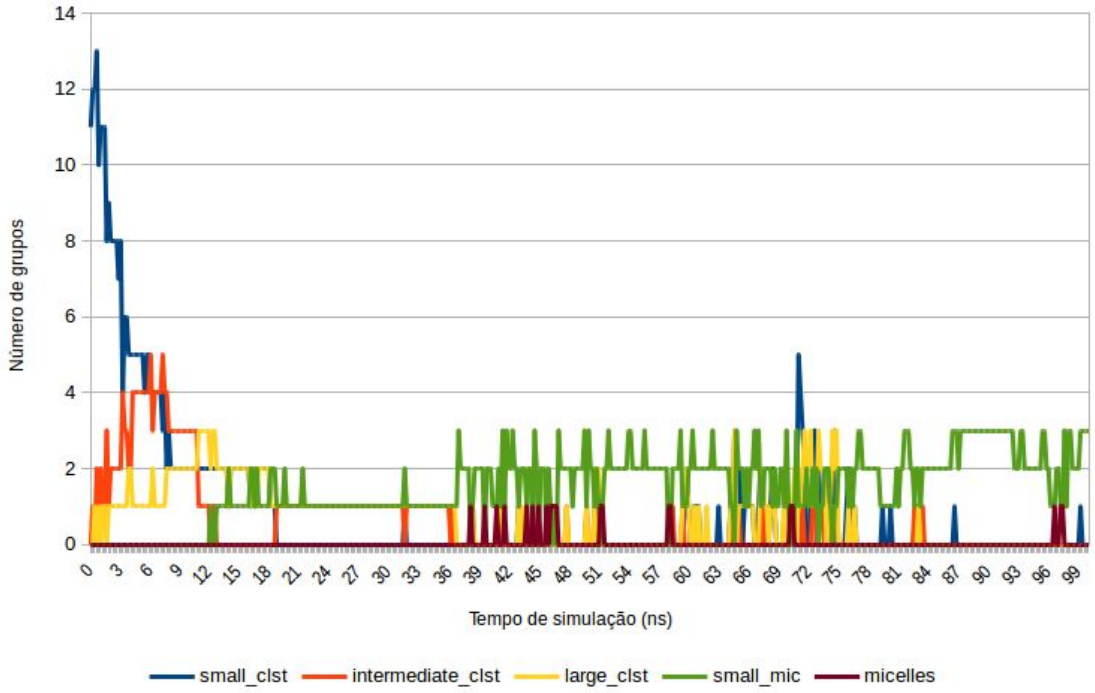

Figura 8. Gráfico da dinâmica de agrupamento de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `DBSCAN` modificado para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

A partir dos 10ns de dinâmica é possível observar a diminuição no número de grupos `intermediate_clst` e `large_clst` para o surgimento de agrupamentos `small_mic`. Essas micelas formadas são mantidas durante toda a dinâmica, porém os cálculos de identificação de agrupamentos indica diversas oscilações na composição desses grupos. A partir dos 35ns de dinâmica parece haver uma mudança no patamar do número de grupos `small_mic`. Neste ponto as oscilações nas identificações indicam uma valores entre 1 e 3 micelas. Uma análise mais detalhada desse processo permite identificar que a queda no número de agrupamentos `small_mic` ocorre ao mesmo tempo que o aparecimento de agrupamentos `micelles`. Isso ocorre pela proximidade geométrica de dois (2) dos grupos

observados, o que acaba dificultando e confundindo o cálculo do algoritmos em situações nas quais os grupos se aproximam momentaneamente durante a dinâmica.

A Figura 8 também nos permite observar uma oscilação considerável no sistema durante 60ns a 80ns. Nesse intervalo da dinâmica os resultados mostram o aumento repentino do número de grupos de tamanho intermediário e um pico de `small_clst`. Isso ocorre, pois durante esse intervalo na dinâmica um número considerável de moléculas se desprendem de seus agrupamentos. Esse comportamento pode indicar que o sistema não se apresenta completamente equilibrado. Esse comportamento também expõe uma das dificuldades da simulação computacional do *self-assembly* de moléculas em micelas.

De forma geral, o número de `small_mic` parece guiar a informação dos agrupamentos estáveis na dinâmica. As oscilações no número de grupos medido é perceptível por conta das características do sistema e do algoritmo. O número máximo de grupos `small_mic` apontam para a formação de 3 grupos estáveis. Tais tendências apresentadas anteriormente podem ser analisadas pelas linhas de tendência por média móvel (período 15 *frames*) apresentadas pela Figura 9.

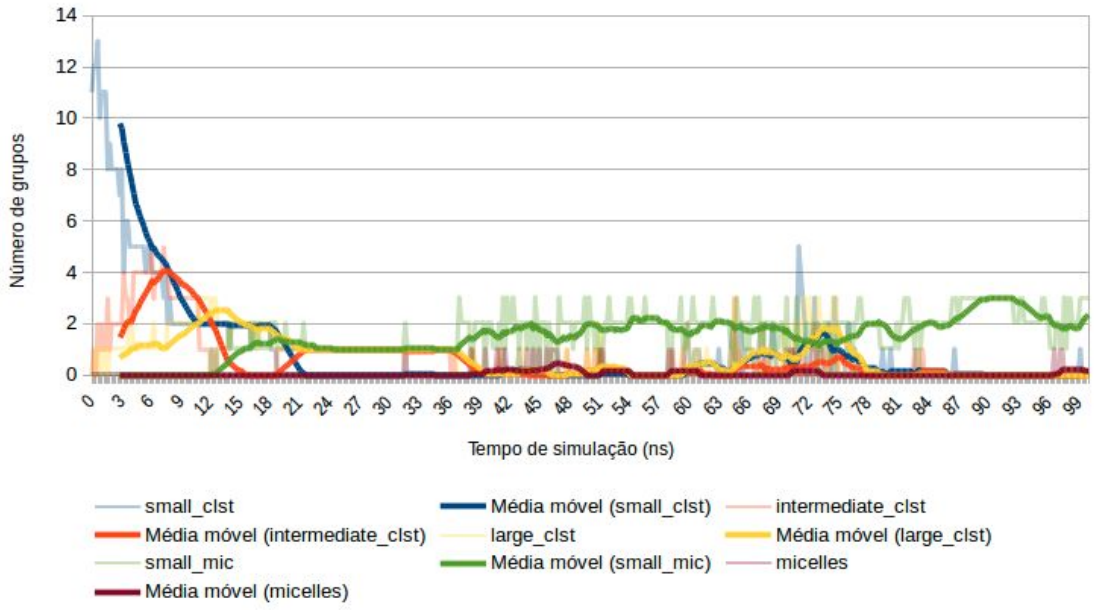

Figura 9. Gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 15 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método DBSCAN modificado para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

A Figura 10 apresenta os resultados dos agrupamentos moleculares em certos momentos da dinâmica molecular. Percebe-se que os grupos se formam de forma estável de forma gradativa, podem a proximidade entre alguns deles dificulta a identificação dos agregados.

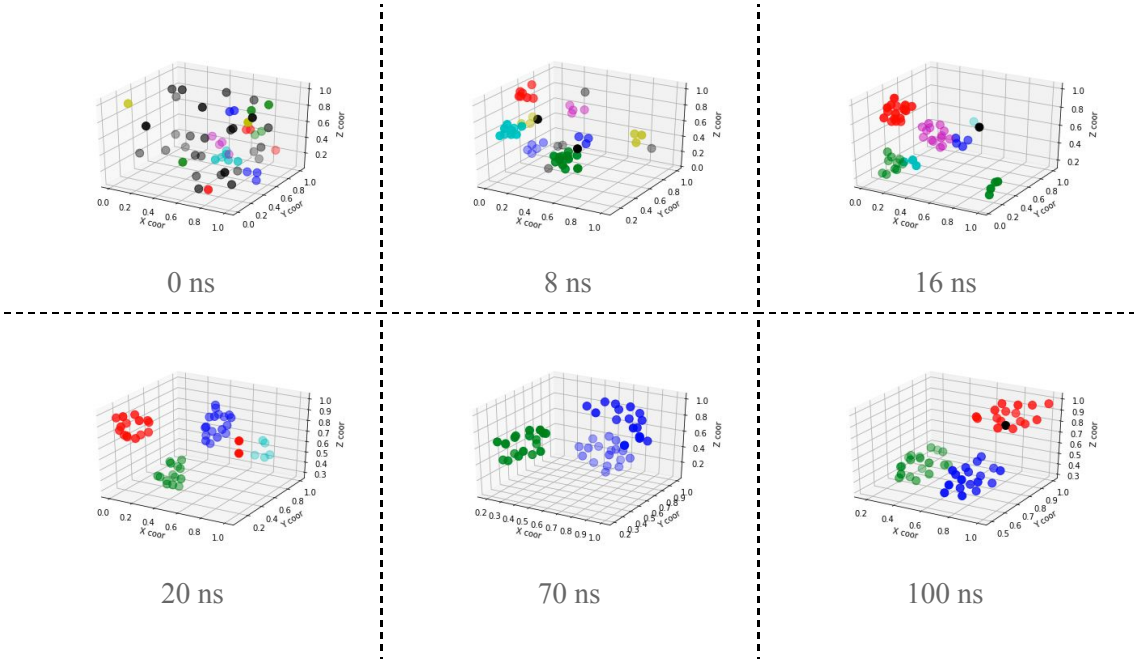

Figura 10. Agrupamentos identificados pelo algoritmo DBSCAN em diversos estágios da dinâmica molecular realizada. Cada cor representa um agregado molecular distinto segundo o algoritmo.

A Tabela 2 apresenta um resumo do número de grupos nas classes `large_clst`, `small_mic` e `micelles` durante diferentes faixas da dinâmica. Já a Tabela 3 considera somente as classes `small_mic` e `micelles` nos cálculos dos valores resumo. Com esses valores podemos entender melhor como se dá a formação desses agrupamentos de tamanhos consideráveis. Ambas situações demonstram aumento no número médios de cluster calculados. As tabelas apresentam valores consideravelmente diferentes para os valores médios até os 80ns de dinâmica, primeiramente pelo alto número de grupos na classe `large_clst` no início da dinâmica e posteriormente pelo aumento de grupos nessa classe na região de oscilação identificada entre 60-80ns.

	0-20ns	**20-40ns**	**40-60ns**	**60-80ns**	**80-100ns**
Mínimo	0	1	1	1	1
Máximo	3	3	3	4	3
Mediana	2	2	2	2	2
Média ± Desvio Padrão	2 ± 0,953	2 ± 0,246	2,07 ± 0,517	2,4 ± 0,725	2,37 ± 0,646

Tabela 2. Valores resumo das medidas do número de *clusters* nas classes **`large_clst`**, **`small_mic`** e **`micelles`** em diferentes faixas da simulação.

	0-20ns	20-40ns	40-60ns	60-80ns	80-100ns
Mínimo	0	1	1	0	1
Máximo	2	3	3	3	3
Mediana	0	1	2	2	2
Média ± Desvio Padrão	0,46 ± 0,642	1,16 ± 0,395	1,93 ± 0,555	1,70 ± 0,674	2,36 ± 0,644

Tabela 3. Valores resumo das medidas do número de *clusters* nas classes **`small_mic`** e **`micelles`** em diferentes faixas da simulação.

Nos resultados anteriores percebe-se oscilação nas medidas efetuadas, porém é importante ressaltar que o nível de oscilação nos resultados calculados é maior se não tivessem seus resultados atenuados por conta da utilização da proposta do uso de uma medida de distância periódica. A Figura 11 e Figura 12 mostram os mesmos resultados, porém obtidos pelo algoritmo padrão do `DBSCAN`. É visível que as oscilações presentes são consideravelmente maiores, principalmente nos primeiros 30ns da dinâmica.

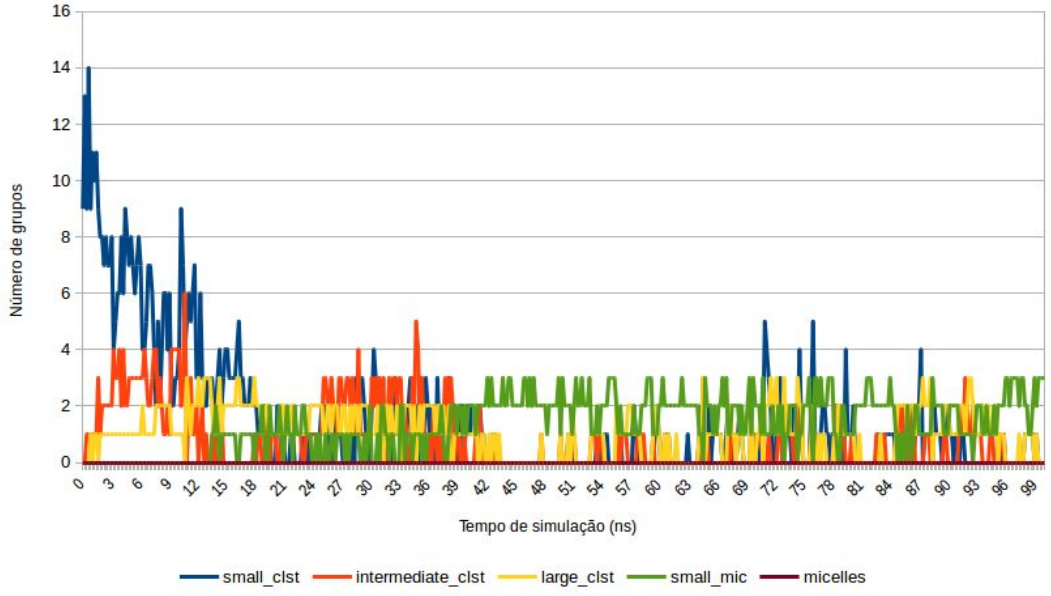

Figura 11. Gráfico da dinâmica de agrupamento de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `DBSCAN` **sem** modificação para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

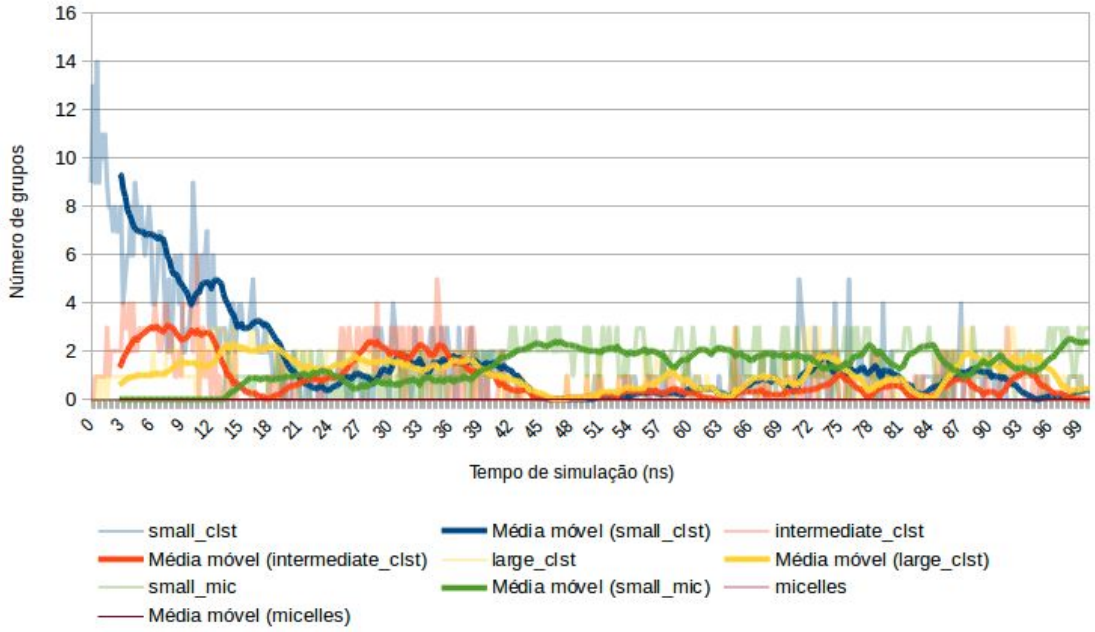

Figura 12. Gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 15 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `DBSCAN` **sem** modificação para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

A Figura 13 representa o número de grupos durante os 100ns de dinâmica sem considerar a divisão de classes de agrupamentos. No gráfico podemos observar a diminuição gradual do número de grupos. A partir dos 35ns é possível observar um número médio de 3 grupos. É importante ressaltar a presença da oscilação entre os 65ns a 75ns já mencionado.

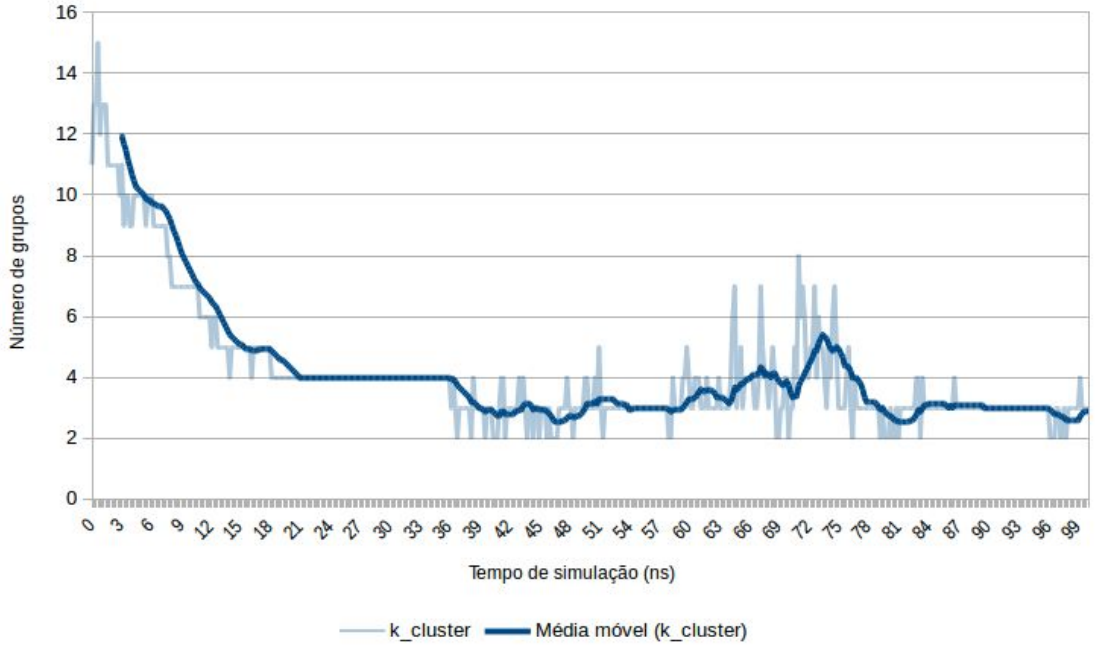

Figura 13. Gráfico da dinâmica de agrupamento de moléculas de DPC de 0-100 ns da dinâmica

molecular de produção pelo método DBSCAN modificado para atributos periódicos. O gráfico mostra a soma do número de grupos com mais de duas (2) moléculas.

Uma das dificuldades presentes na utilização do método de agrupamento DBSCAN é a definição de *cutoff* para o parâmetro `eps` (raio de busca). A Figura 14 mostra os resultados obtidos com a variação do parâmetro `eps`. O método se mostra sensível a variações no parâmetro de entrada. Valores baixos parecem favorecer o número de grupos encontrados e ressaltam a presença de oscilações nos sistemas. Valores altos para o parâmetro tendem a perdas na observação das nuances do sistema na dinâmica.

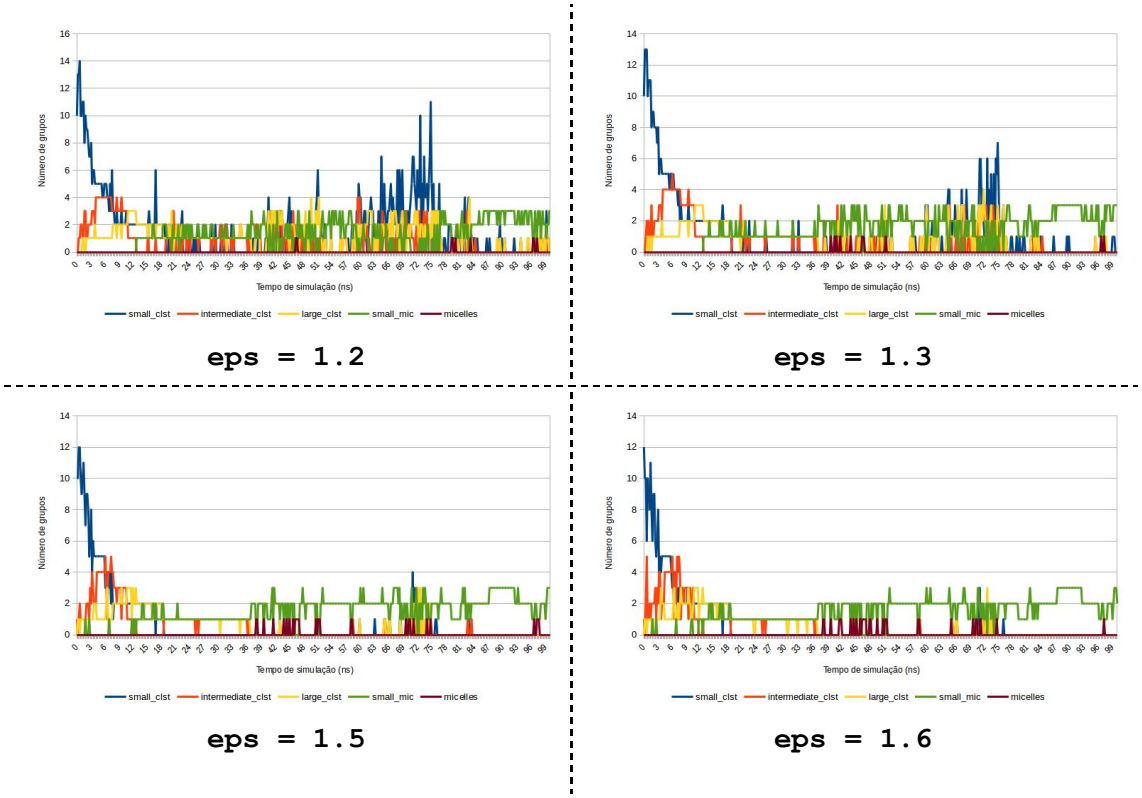

Figura 14. Comparação entre gráfico da dinâmica de agrupamento de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método DBSCAN modificado para atributos periódicos com a variação do parâmetro de raio de busca. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

4.2.2. Affinity Propagation

A Figura 15 mostra o gráfico de agrupamentos moleculares durante os 100ns da dinâmica efetuada calculados pelo algoritmo Affinity Propagation. Uma grande vantagem do método em questão é a menor sensibilidade a variações de seus parâmetros de entrada. O algoritmo não funciona a base de definições de valores de *cutoff*. Contudo, os resultados do método não foram promissores. Os resultados apresentados na Figura 15 mostram a baixa capacidade do método de acompanhar as mudanças ocorridas no sistema durante sua evolução.

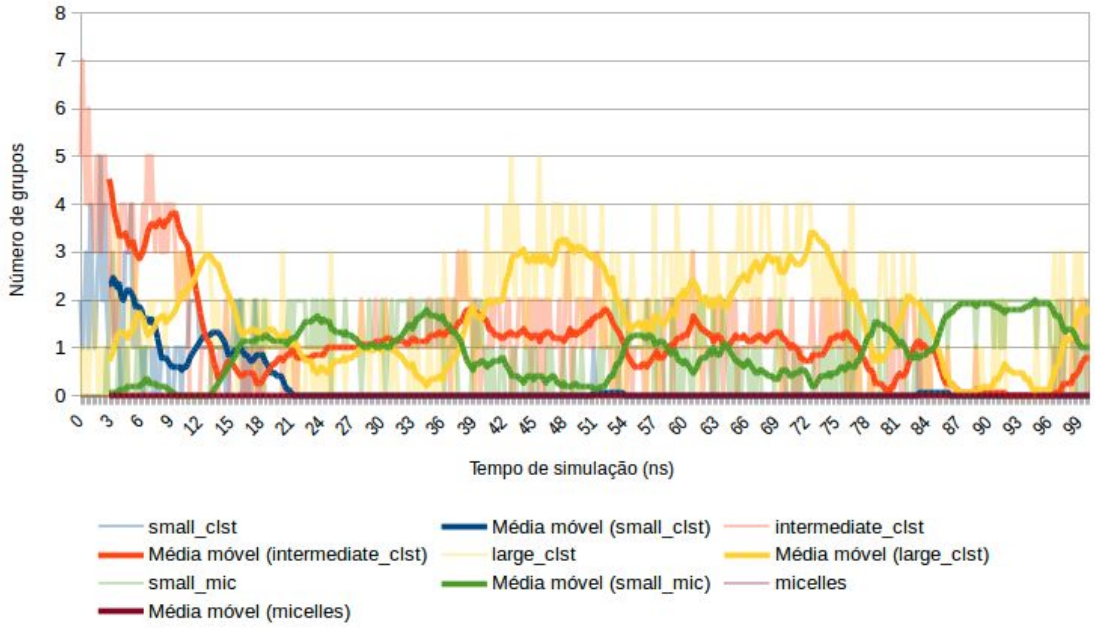

Figura 15. Gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 15 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `Affinity Propagation` modificado para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

É importante ressaltar que a mudança proposta ao método foi capaz de trazer maior detalhamento da evolução do sistema por parte do método, porém, como visto, não o necessário. A critério de comparação, a Figura 16 mostra os mesmos resultados obtidos pela implementação clássica do método. Percebe-se grande oscilação e ainda menos detalhamento das nuances da evolução do sistema.

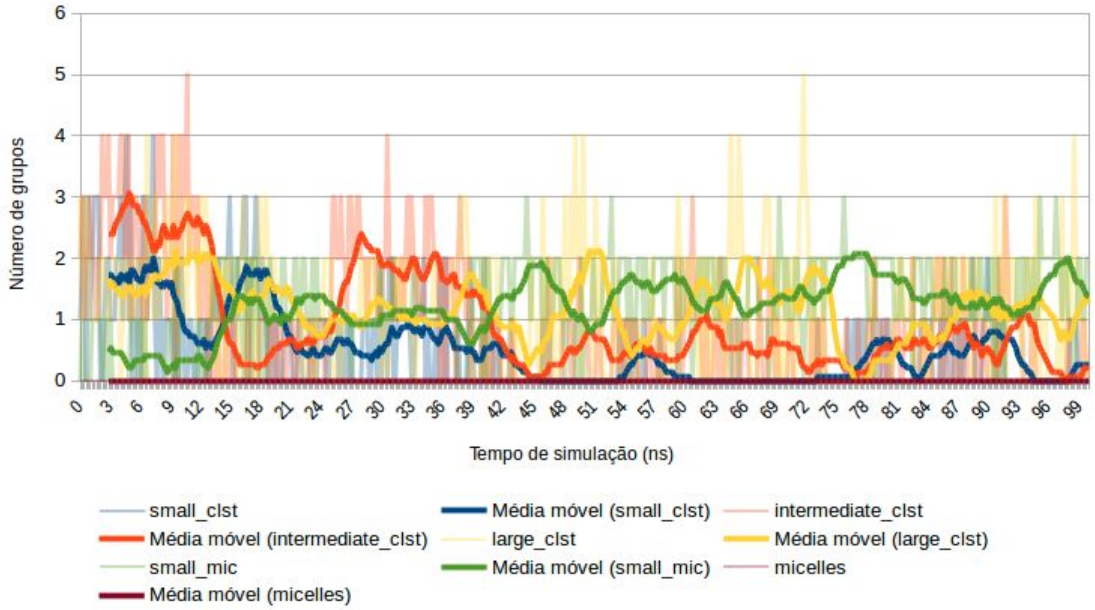

Figura 16. Gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 15 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `Affinity Propagation` **sem** modificacao para atributos periódicos. O gráfico mostra os dados para 5 tamanhos de grupos: `small_clst` (de 2 a 4 moléculas), `intermediate_clst` (5-8), `large_clst` (9-15), `small_mic` (16-31) e `micelles` (32 ou mais moléculas).

4.2.3. gmx-clustsize

Utilizando a ferramenta *clustsize* do `GROMACS`, também foi possível obter uma distribuição com o número de *clusters* durante cada etapa da simulação (Figura 17).

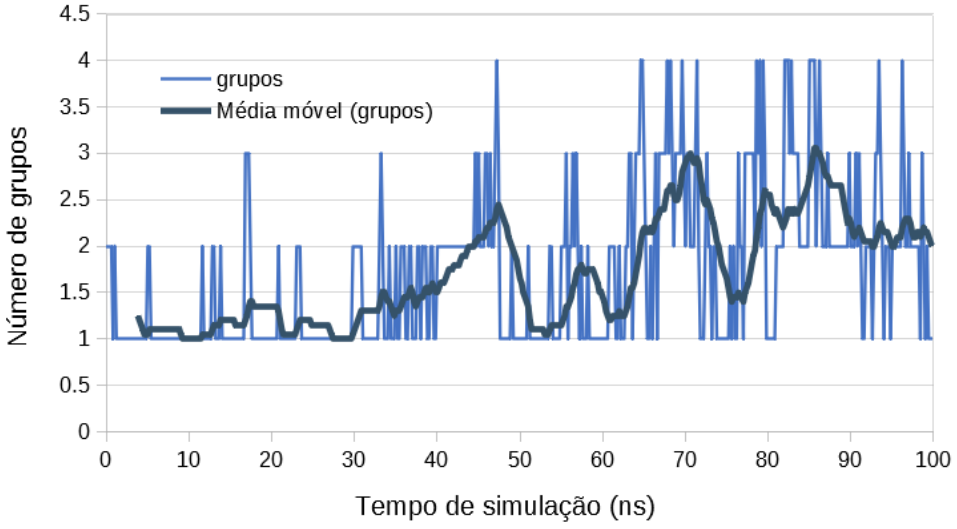

Figura 17. Gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 20 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `gmx-clustersize` considerando condições periódicas. O gráfico mostra a soma do número de grupos com mais de duas (2) moléculas dentro de um *cutoff* de 1,8 nm.

Para gerar os resultados da Figura 17 foi utilizado um *cutoff* de 1,8 nm, pois foi o valor que melhor se adequou às condições da ferramenta para fornecer um resultado satisfatório em relação à simulação realizada. Conforme esperado, na fase inicial da simulação as moléculas se encontram mais distribuídas e apenas um ou dois *clusters* são computados. Enquanto isso, a partir da metade da simulação, as moléculas se encontram agrupadas nos 3 *clusters* majoritários e, por isso, o número de clusters gira em torno de 3. Com certa frequência um número de *clusters* igual a 2 é computado, provavelmente devido à aproximação entre dois dos *clusters* principais, fato que também é observado visualmente no resultado da simulação.

Outros valores de *cutoff* foram utilizados, fornecendo resultados menos coerentes com o que foi observado na simulação. Valores de *cutoff* menores que 1,8 nm encontraram mais de 3 grupos na fase final na qual a simulação revelou 3 grupos razoavelmente estáveis. Enquanto isso, valores maiores que 1,8 nm encontraram menos de 3 grupos na fase final da simulação. O

melhor resultado foi obtido com *cutoff* de 1,8 nm, pois o número de grupos está em torno de 3, ocasionalmente variando devido à aproximação afastamento dos grupos ou conjuntos de moléculas dentro de um grupo (Figura 18).

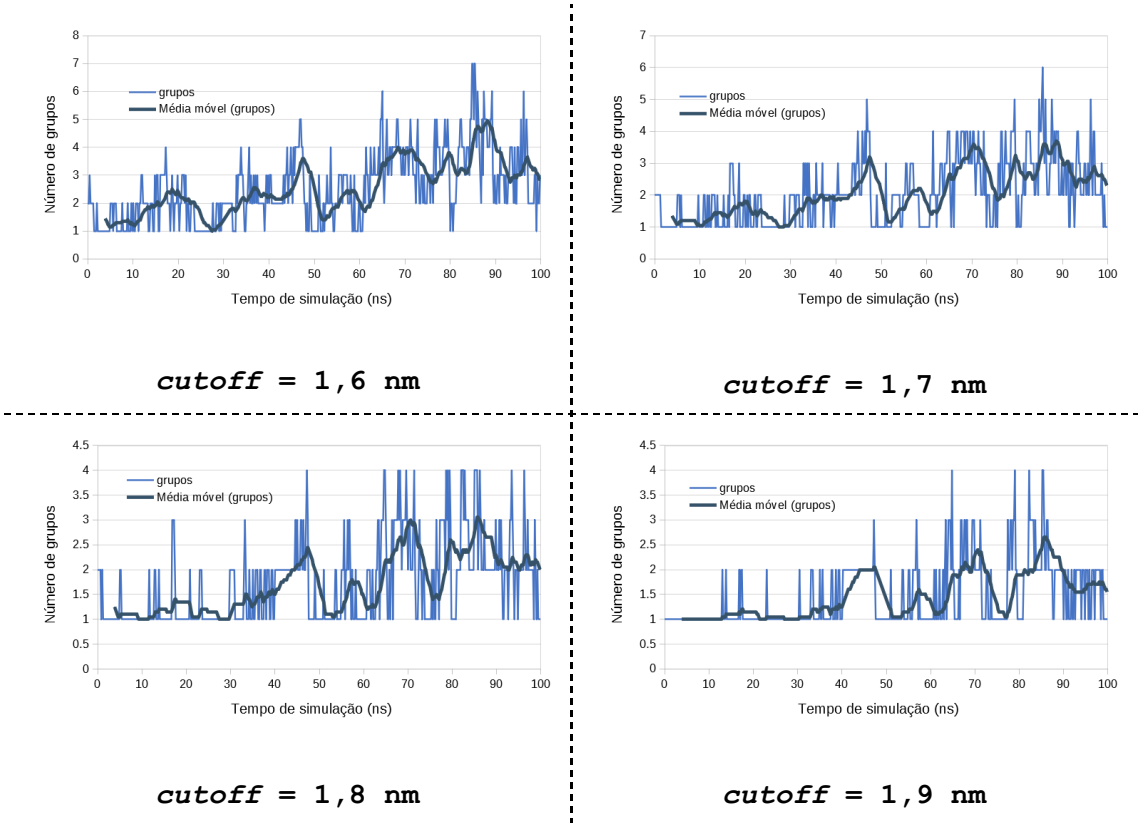

Figura 18. Comparação entre gráfico da dinâmica de agrupamento contrastado pelos valores de média móvel (período 20 *frames*) de moléculas de DPC de 0-100 ns da dinâmica molecular de produção pelo método `gmx-clustsize` com a variação do parâmetro de *cutoff*. O gráfico mostra a soma do número de grupos com mais de duas (2) moléculas.

4.3. Análise de Energias Potenciais

Para verificar a convergência da dinâmica, a energia potencial do sistema foi analisada e o gráfico que representa a energia total durante o tempo de simulação pode ser encontrado na Figura 19. É interessante notar que o momento no qual foi observada a formação dos 3 agregados principais (~43 ns) coincide com a faixa de tempo na qual a energia do sistema começa a atingir um patamar que parece ser próximo ao patamar energético de equilibração. Ainda, pode ser visualizada uma tendência da evolução do sistema para um ponto de mínimo de energia a medida que ocorre a agregação das moléculas do sistema. Para atingir uma configuração globalmente equilibrada seria necessário que um maior tempo de simulação fosse executado. Uma permanência no estado atual mostraria uma convergência do sistema.

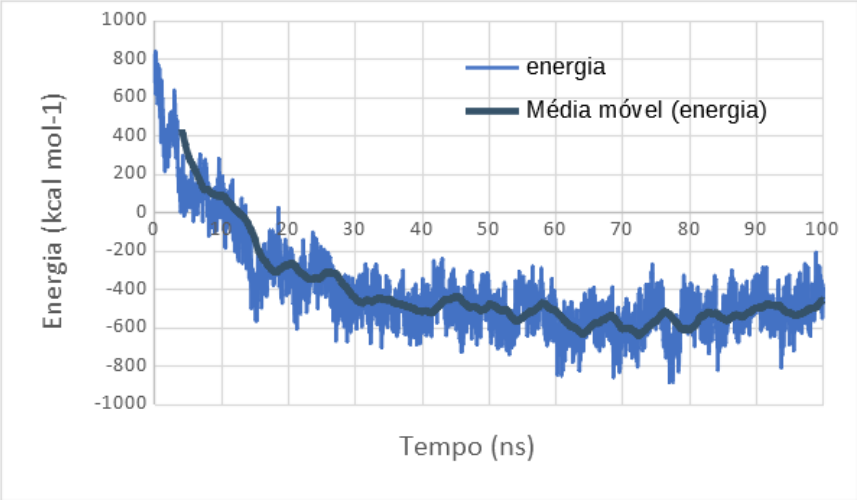

Figura 19. Energia potencial total associada às moléculas de DPC no sistema em função do tempo de simulação contrastada pelos valores de média móvel (período 200 *frames*).

As energias potenciais específicas também foram calculadas individualmente e podem ser encontradas na Figura 20. Pode ser inferido que as energias de ligação e angular não sofreram muita alteração, permanecendo estáveis durante toda a simulação. Isto indica que estas energias estavam bem equilibradas antes da etapa de produção e durante a dinâmica e que a movimentação das moléculas não acarretou mudanças significativas nas ligações e ângulos das moléculas.

A Figura 20 também nos mostra que a energia de diedro por sua vez decaiu no início da simulação e depois permaneceu razoavelmente constante, evidenciando que a dinâmica provocou uma alteração nos diedros para minimizar a energia, que pode estar associado a torções de ligação que ocorreram para gerar uma conformação mais favorável para a formação dos agregados. A energia de Van der Waals foi a que mais sofreu alteração, o que é coerente visto que à medida que os aglomerados foram formados e os grupos apolares foram protegidos da interação com o solvente, a energia referente à interação de Van der Waals das caudas hidrofóbicas com moléculas de água foi substituída por interações cauda/cauda, causando uma diminuição de energia. A energia eletrostática, associada às cargas presentes na cabeça polar dos surfactantes, também diminuiu com a formação dos aglomerados e isto pode estar relacionado à presença de mais interações elétricas intermoleculares.

Sendo assim, é possível concluir que a estabilização do sistema foi regida principalmente pelas energias de Van der Waals e elétrica, pois o decaimento das mesmas é muito favorecido pela formação dos agregados.

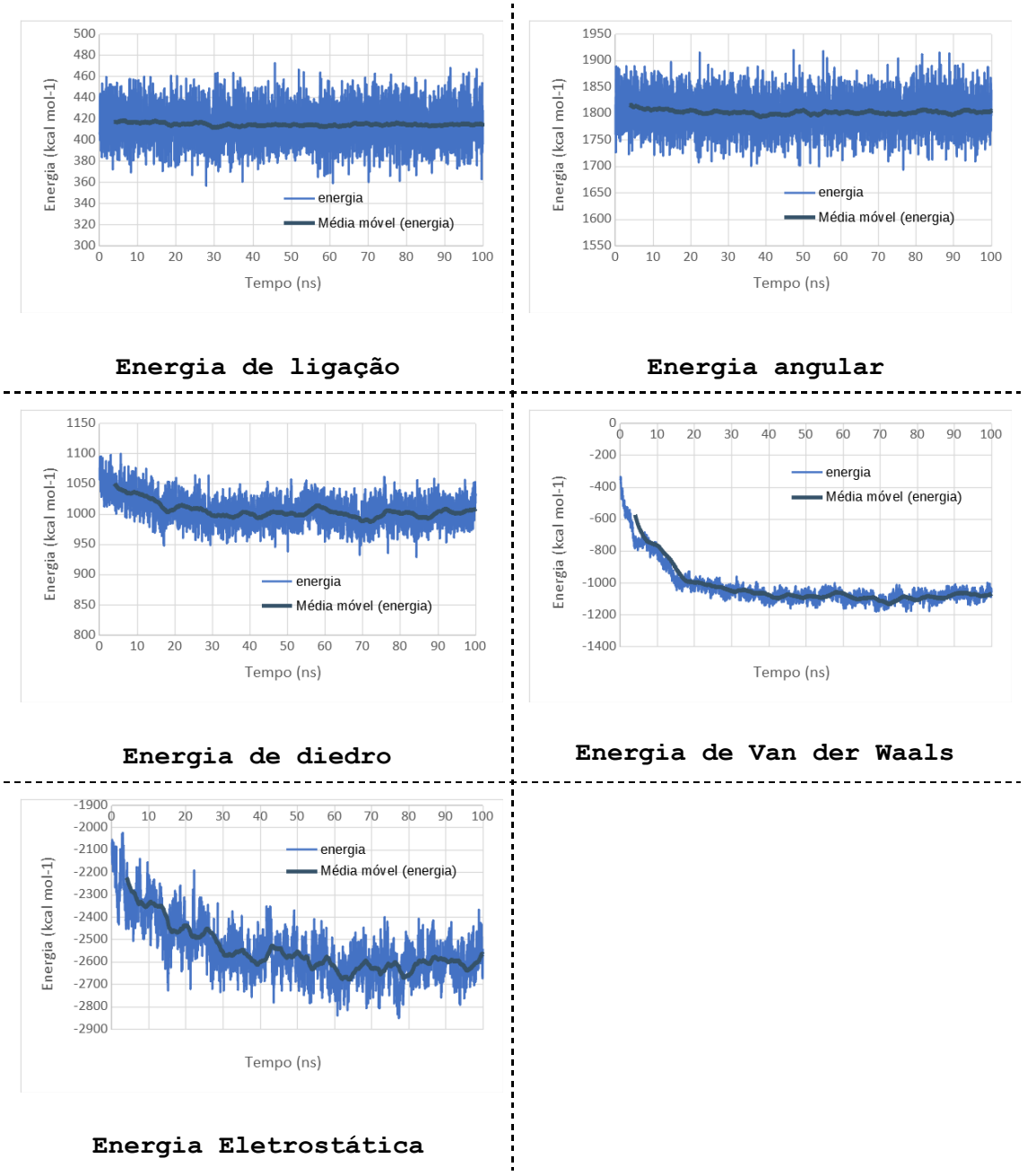

Figura 20. Energias potenciais associadas às moléculas de DPC em função do tempo de simulação contrastadas pelos valores de média móvel (período 200 *frames*).

5. Conclusão

Este trabalho apresentou a descrição de uma metodologia de simulação de moléculas partindo do desenho da geometria da molécula, parametrização de átomos e os modelos de arquivos de configuração da simulação compatíveis com campo de força do `Amber`. A validação do processo ocorreu pela simulação de 100ns de dinâmica de auto agregação de 54 moléculas de DPC em meio aquoso e comparação com os resultados obtidos por Marrink e colaboradores (2000).

As análises dos resultados ocorreram de forma qualitativa, por meio da visualização das trajetórias da simulação, e de forma quantitativa, por meio da aplicação de diferentes métodos de agrupamento. Alterações foram propostas aos métodos que não consideravam atributos periódicos em sua forma padrão na literatura e demonstraram melhora significativa nos resultados. Os resultados mostraram que o método `Affinity Propagation` não foi capaz de acompanhar a dinâmica de formação de agregados moleculares. O método `DBSCAN` acompanhou bem a formação dos grupos por pequenas oscilações causada pela proximidade geométrica dos grupos formados. Os resultados obtidos pelo `DBSCAN` foram, a princípio, mais razoáveis que os resultados obtidos pelo método de agrupamento do `GROMACS` (`gmx clustsize`), porém seus resultados são bem sensíveis à escolha do parâmetro de entrada do raio de busca (`eps`).

Foi possível verificar de forma quantitativa e qualitativa o processo de formação de três (3) agregados moleculares de tamanho considerável (17 a 19 moléculas) na dinâmica efetuada. Na dinâmica efetuada por Marrink e colaboradores (2000) o sistema convergiu para uma grande micela com todas as moléculas do sistema. Acredita-se que as diferenças entre os campos de força e métodos de integração dos programas `GROMACS` e `AMBER` sao os principais responsáveis por essa diferença. Segundo os dados experimentais, como o sistema se encontra acima da CMC, o mesmo deveria convergir para a formação de micelas. O número de moléculas utilizado por Marrink e colaboradores (MARRINK; TIELEMAN; MARK, 2000) e repetido neste trabalho apresenta mais moléculas que o número de agregação, porém menos moléculas do que seria o suficiente para formar duas micelas. No caso das condições utilizadas por Marrink e colaboradores (campo de forças `GROMACS` e água do tipo `SPC`) o sistema convergiu para a formação de uma única micela contendo mais moléculas de DPC do que seria esperado experimentalmente e, no caso do presente trabalho (campo de forças `AMBER` e água do tipo `TIP3P`), o sistema convergiu para a formação de três micelas menores, distribuindo a quantidade de moléculas de DPC. Neste aspecto, seria interessante repetir a simulação em condições mais próximas à CMC e utilizando em torno de 44 moléculas de DPC, para observar se uma única micela se formaria.

A análise da evolução das energias potenciais mostraram que o sistema se encontra ainda em um processo de equilibração, o que caracteriza uma das maiores dificuldades desse tipo de estudo de auto agregação molecular. Para atingir uma configuração globalmente equilibrada é necessário que um maior tempo de simulação seja executado. Uma permanência no estado atual mostraria uma convergência do sistema.

De forma geral, os métodos de análise de agrupamentos moleculares propostos se mostraram adequados para serem aplicados em estudos de auto agregação, fornecendo resultados mais representativos do que a ferramenta do `GROMACS gmx clustsize`. Sobre o protocolo seguido para realizar as simulações, seria importante permitir um tempo maior de simulação do que o realizado para confirmar a convergência do sistema. Caso, ainda assim, não fosse atingido um resultado satisfatório de equilibração do sistema, seria necessário otimizar o protocolo de forma a reproduzir o que é esperado segundo os dados experimentais. Essa validação permitiria efetuar de forma confiável simulações em moléculas semelhantes, como a Miltefosina e análogos não clássicos da mesma como sais orgânicos de 1,2,3-triazol, para fornecer dados sobre seus mecanismos de agregação molecular.

Referências

ABRAHAM, M. J.; VAN DER SPOEL, D.; LINDAHL, E.; HESS, B. e a equipe de desenvolvimento do GROMACS, GROMACS User Manual version 2019. http://www.gromacs.org.

CALIXTO, S. L.; GLANZMANN, N.; SILVEIRA, M. M. X.; GRANATO, J. T.; SCOPEL, K. K. G.; AGUIAR, T. T.; DAMATTA, R. A.; MACEDO, G. C.; SILVA, A. D.; COIMBRA E. S. Novel organic salts based on quinoline derivatives: The *in vitro* activity trigger apoptosis inhibiting autophagy in *Leishmania* spp. **Chemico-Biological Interactions**, v. 293, p. 141-151, 2018.

CASE, D. A.; BEN-SHALOM, I. Y.; BROZELL, S. R.; CERUTTI, D. S.; CHEATHAM, III, T. E.; CRUZEIRO, V. W. D.; DARDEN, T. A.; DUKE, R. E.; GHOREISHI, D.; GIAMBASU, G.; GIESE, T.; GILSON, M. K.; GOHLKE, H.; GOETZ, A. W.; GREENE, D.; HARRIS, R.; HOMEYER, N.; HUANG, Y.; IZADI, S.; KOVALENKO, A.; KRASNY, R.; KURTZMAN, T.; LEE, T. S.; LEGRAND, S.; LI, P.; LIN, C.; LIU, J.; LUCHKO, T.; LUO, R.; MAN, V.; MERMELSTEIN, D. J.; MERZ, K. M.; MIAO, Y.; MONARD, G.; NGUYEN, C.; NGUYEN, H.; ONUFRIEV, A.; PAN, F.; QI, R.; ROE, A.; ROITBERG, D. R.; SAGUI, C.; SCHOTT-VERDUGO, S.; SHEN, J.; SIMMERLING, C. L.; SMITH, J.; SWAILS, J.; WALKER, R. C.; WANG, J.; WEI, H.; WILSON, L.; WOLF, R. M.; WU, X.; XIAO, L.; XIONG, L.; YORK, D. M.; KOLLMAN, P. A. AMBER 2019, University of California, San Francisco, 2019. https://ambermd.org/doc12/Amber19.pdf.

CHEN, J.; YANG, C.; GUO, J.; ZHU, D.; FU, S.; YANG, Z.; ZHONG, X. Mesoscopic Simulations on the Aggregate Behavior of Oligomeric Adamantane Surfactants in Aqueous Solutions. **Tenside Surfactants Detergents**, v. 53, n. 2, p. 120-126, 2016.

DEAMER, D.; DWORKIN, J. P.; SANDFORD, S. A.; BERNSTEIN, N. P.; ALLAMANDOLA, L. J. The first cell membranes. **Astrobiology**, v. 2, p. 371-81, 2002.

DOMINGUÉS, A.; FERNÁNDES, A.; GONZÁLES, N.; IGLESIAS, E.; MONTENEGRO, L. Determination of Critical Micelle Concentration of Some Surfactants by Three Techniques. **Journal of Chemical Education**, v. 74, n. 10, p. 1227-1231, 1997.

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. 1996. p. 226-231.

GLANZMANN, N.; CARMO, A. M. L.; ANTINARELLI, L. M. R., COIMBRA, E. S.; COSTA, L. A. S.; SILVA, A. D. Synthesis, characterization, and NMR studies of 1,2,3-triazolium ionic liquids: a good perspective regarding cytotoxicity. **Journal of Molecular Modeling**, v. 24, n. 160, p. 1-7, 2018.

IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. **Blackwell Scientific Publications, Oxford** (1997). Online version (2019-) created by S. J. Chalk. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook.

JOHNSTON, Michael A. et al. Toward a standard protocol for micelle simulation. **The Journal of Physical Chemistry B**, v. 120, n. 26, p. 6337-6351, 2016.

LEBECQUE, S.; CROWET, J. M.; NASIR, M. N.; DELEU, M.; LINS, L. Molecular dynamics study of micelles properties according to their size. **Journal of Molecular Graphics and Modelling**, v. 72, p. 6-15, 2017.

MARRINK, S. J.; TIELEMAN, D. P.; MARK, A. E.; Molecular Dynamics Simulation of the Kinetics of Spontaneous Micelle Formation. **The Journal of Physical Chemistry B**, v. 104, p. 12165-12173, 2000.

MARTINS, R. C.; DORNELES, G. P.; TEIXEIRA, V. O. N.; ANTONELLO, A. M.; COUTO, J. L.; RODRIGUES JÚNIOR, L. C.; MONTEIRO, M. C.; PERES, A.; SCHREKKER, H. S.; ROMÃO, P. R. T.; Imidazolium salts as innovative agents against Leishmania amazonensis. **International Immunopharmacology**, v. 63, p. 101-109, 2018.

NISTICÒ, R.; SCALARONE, D.; MAGNACCA, G. Sol-gel chemistry, templating and spin-coating deposition: A combined approach to control in a simple way the porosity of inorganic thin films/coatings. **Microporous and Mesoporous Materials**, v. 248, p. 18-29, 2017.

NADIPALLI, S., VENKATA D. and KARTEEKA P.. Relative Performance of K-Means, Single Linkage and Affinity Propagation in Cluster Analysis, 2014.

OLESEN, N. E.; HOLM, R.; WESTH, P. Determination of the aggregation number for micelles by isothermal titration calorimetry. **Thermochimica Acta**, v. 588, p. 28-37, 2014.

PHILLIPS, J. C.; BRAUN, R.; WANG, W.; GUMBART, J.; TAJKHORSHID, E.; VILLA, E.; CHIPOT, C.; SKEEL, R. D.; KALÉ, L.; SCHULTEN, K. Scalable molecular dynamics with namd. **Journal of Computational chemistry**, v. 26, n. 16, p. 1781-1802, 2005.

PEDREGOSA et al.; Scikit-learn: Machine Learning in Python. **JMLR 12**, pp. 2825-2830, 2011.

STROPPA, P. H. F.; ANTINARELLI, L. M. R.; CARMO, A. M. L.; GAMEIRO, J.; COIMBRA, E. S.; SILVA, A. D. Effect of 1,2,3-triazole salts, non-classical bioisosteres of miltefosine, on Leishmania amazonensis. **Bioorganic & Medicinal Chemistry**, v. 25, p. 3034-3045, 2017.

TEHRANI-BAGHA, A. R.; KÄRNBRATT, J.; LÖFROTH, J.-E.; HOLMBERG K. Cationic ester-containing gemini surfactants: Determination of aggregation numbers by time-resolved fluorescence quenching. **Journal of Colloid and Interface Science**, v. 376, n. 1, p. 126-132, 2012.

THAVIKULWAT, P. Affinity propagation: a clustering algorithm for computer-assisted business simulations and experiential exercises. In: **Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference**. 2014.

THÉVENOT, C.; GRASSL, B.; BASTIAT, G.; BINANA, W. Aggregation number and critical micellar concentration of surfactant determined by time-dependent static light scattering

(TDSLS) and conductivity. **Colloids and Surfaces A: Physicochemical and Engineering Aspects**, v. 252, n. 2-3, p. 105-111, 2005.

TIAN, C.; KARRA, M. D.; ELLIS, C. D.; JACOB, J.; OXENOID, K.; SÖNNICHSEN, F.; SANDERS, C. R.; Membrane Protein Preparation for TROSY NMR Screening. **Methods In Enzymology**, v. 394, p. 321-334, 2005.

TIELEMAN, D. P.; BERENDSEN, H. J. C. Molecular dynamics simulations of a fully hydrated dipalmitoylphosphatidylcholine bilayer with different macroscopic boundary conditions and parameters. **The Journal of Chemical Physics**, v. 105, p. 4871-4880, 1996.

TIELEMAN, D. P.; VAN DER SPOEL, D.; BERENDSEN, H. J. C. Molecular Dynamics Simulations of Dodecylphosphocholine Micelles at Three Different Aggregate Sizes: Micellar Structure and Chain Relaxation. **The Journal of Physical Chemistry B**, v. 104, n. 27, p. 6380-6388, 2000.

VEGA, C.; DE MIGUEL, E. Surface tension of the most popular models of water by using the test-area simulation method. **The Journal Of Chemical Physics**, v. 126, p. 154707-154710, 2007.

YOSHII, N.; IWAHASHI, K.; OKAZAKI, S. A molecular dynamics study of free energy of micelle formation for sodium dodecyl sulfate in water and its size distribution. **The Journal of Chemical Physics**, v. 124, n. 18, p. 184901, 2006.

YOSHII, N.; OKAZAKI, S. A molecular dynamics study of structural stability of spherical sds micelle as a function of its size. **Chemical Physics Letters**, v. 425, n. 1-3, p. 58-61, 2006.