

# **Dynamika molekularna białek strukturalnie nieuporządkowanych oraz ich agregatów w ramach modeli gruboziarnistych**

Łukasz Mioduszewski

**Rozprawa doktorska**

Promotor:

Prof. dr hab. Marek Cieplak

Środowiskowe Laboratorium Fizyki Biologicznej

Instytut Fizyki, Polska Akademia Nauk

Warszawa 2020

# Podziękowania

Dziękuję

Rodzinie, w szczególności Mamie;

Przyjaciołom, w szczególności Tomkowi i Asi;

Agnieszce;

Współpracownikom, w szczególności:

dr hab. B. Różyckiemu za dyskusje i pomoc w rozwijaniu modelu PID,

dr B. R. H de Aquino za dyskusje i pomoc w algorytmie klastrowania,

mgr K. Wołkowi, dr M. Wojciechowskiemu i dr M. Chwastykowi za pomoc w zrozumieniu kodu,

mgr G. Matyszczakowi za pomoc w parametryzacji modeli,

mgr M. Raczkowskiemu za pomoc w tworzeniu potencjału dla mostków dwusiarczkowych;

prof. dr hab. A. L. Sobolewskiemu za symulacje DFT;

Promotorowi za to, że mogę liczyć na pomoc we wszystkim.

Wszystkie obliczenia wykonałem przy użyciu infrastruktury PL-Grid oraz klastra SL4 w IFPAN.

# Lista publikacji

Doktorat omawia i rozszerza wyniki opublikowane w sześciu artykułach [I -VI] i dwóch rozdziałach w książkach [VII -VIII]. Rozprawa integruje je w jedną spójną całość i zawiera wszystkie wyniki otrzymane w ramach pracy nad doktoratem, w tym niektóre dotychczas niepublikowane.

[I] Ł. Mioduszewski i M. Cieplak, Disordered peptide chains in an  $C_\alpha$ -based coarse-grained model. *Phys. Chem. Chem. Phys.* 2018, **20**, 19057-19070. <https://doi.org/10.1039/C8CP03309A>

[II] Ł. Mioduszewski, B. Różyczyki, i M. Cieplak. Pseudo-improper-dihedral model for intrinsically disordered proteins. *J. Chem. Theory Comput.* 2020, **16**(7), 4726–4733. <https://doi.org/10.1021/acs.jctc.0c00338>

[III] B. R. H. de Aquino, M. Chwastyk, Ł. Mioduszewski i M. Cieplak, The networks of the inter-basin traffic in intrinsically disordered proteins. *Phys. Rev. Res.* 2020, **2**, 013242. <https://doi.org/10.1103/PhysRevResearch.2.013242>

[IV] Ł. Mioduszewski i M. Cieplak. Protein droplets in systems of disordered homopeptides and the amyloid glass phase. *Phys. Chem. Chem. Phys.* 2020, **22**, 15592–15599. <https://doi.org/10.1039/DQCP01635G>

[V] Ł. Mioduszewski i M. Cieplak, Viscoelastic properties of wheat gluten in a molecular dynamics study, (w trakcie recenzji), 2020. <https://doi.org/10.1101/2020.07.29.226928>

[VI] Ł. Mioduszewski, M. Chwastyk i M. Cieplak, Contact-based molecular dynamics of structured and disordered proteins in a coarse-grained model: fixed contacts, switchable contacts and those described by pseudo-improper-dihedral angles, (w trakcie recenzji), 2020. Zawiera szczegóły techniczne dotyczące używania programu do symulacji.

[VII] M. Cieplak, M. Chwastyk, Ł. Mioduszewski i B. R. H. de Aquino, Transient knots in intrinsically disordered proteins and neurodegeneration, Rozdział w *Progress in Mol. Biol. and Translational Sci.* **174**, 79-103, ed. V. N. Uversky, Elsevier, 2020. <https://doi.org/10.1016/bs.pmbts.2020.03.003>

[VIII] M. Cieplak, Ł. Mioduszewski i M. Chwastyk, Contact-based analysis of aggregation of intrinsically disordered proteins, Rozdział w *Computer Simulations of Aggregation of Proteins and peptides*, ed. M. S. Li, Springer (przyjęte do publikacji).

Inne (artykuł [IX] powstał jeszcze przed rozpoczęciem studiów doktoranckich):

[IX] B. Różyczyki, Ł. Mioduszewski i M. Cieplak, Unbinding and unfolding of adhesion protein complexes through stretching: Interplay between shear and tensile mechanical clamps. *Proteins*, 2014, **82**, 3144-3153. <https://doi.org/10.1002/prot.24674>

# Prawa autorskie

Rozprawa zawiera rysunki będące kopią bądź modyfikacją rysunków z publikacji od [I] do [V],[VII] oraz [VIII]. Każdy taki rysunek zawiera odpowiednią informację w podpisie.

Rysunki z artykułów [I] i [IV] oraz z rozdziałów w książkach [VII] i [VIII] są udostępnione w tej rozprawie jedynie do celów edukacyjnych i naukowych, nie mogą być w żaden sposób publicznie rozpowszechniane ani modyfikowane.

Rysunki z artykułów [II], [V] oraz fragment rysunku z artykułu [IV] są udostępnione na licencji CC-BY, a zatem można je udostępniać i modyfikować pod warunkiem zacytowania oryginalnej pracy ([II], [III] lub [V]). Wszystkie pozostałe rysunki oraz tekst rozprawy mogą być publicznie rozpowszechniane oraz modyfikowane.

Wszystkie artykuły i rozdziały z Listy publikacji znajdują się na płycie CD dołączonej do rozprawy na potrzeby recenzji. Pozostają one do wyłącznej dyspozycji Recenzentów i są chronione prawami autorskimi przysługującymi wydawcom.

# Spis treści

<b>1 Wstęp</b>	<b>6</b>
1.1 Wprowadzenie . . . . .	6
1.2 Plan pracy . . . . .	12
<b>2 Budowa modelu gruboziarnistego</b>	<b>13</b>
2.1 Wprowadzenie . . . . .	13
2.2 Symulacje dynamiki molekularnej . . . . .	14
2.3 Sztywność łańcucha . . . . .	16
2.3.1 Potencjał kąta płaskiego . . . . .	17
2.3.2 Potencjał kąta dwuściennego . . . . .	17
2.4 Temperatura pokojowa $T_r$ . . . . .	20
2.5 Model quasi-adiabatyczny . . . . .	22
2.5.1 Postać potencjału . . . . .	22
2.5.2 Rodzaje kontaktów i ich charakterystyczne odległości . . . . .	22
2.5.3 Kryteria odległościowe . . . . .	29
2.5.4 Kryteria kierunkowe . . . . .	29
2.5.5 Kryteria związane z liczbą koordynacyjną . . . . .	33
2.5.6 Elektrostatyka . . . . .	37
2.5.7 Mostki dwusiarczkowe . . . . .	38
2.5.8 Zgodność z doświadczeniem i symulacjami pełnoatomowymi . . . . .	38
2.6 Model z Hamiltonianem niezależnym od czasu . . . . .	48
2.6.1 Dane z bazy PDB . . . . .	50
2.6.2 Prawoskrętność . . . . .	52
2.6.3 Kontakty typu bs . . . . .	54
2.6.4 Dane bez podziału na kontakty typu bb, bs i ss . . . . .	56
2.6.5 Implementacja potencjału PID . . . . .	57
2.6.6 Forma potencjału Lenard-Jonesa . . . . .	58
2.6.7 Oddziaływanie elektrostatyczne . . . . .	59
2.6.8 Porównanie wariantów modelu z doświadczeniem . . . . .	59
2.7 Podsumowanie . . . . .	70

<b>3 Wyniki symulacji pojedynczych łańcuchów</b>	<b>73</b>
3.1 Wprowadzenie . . . . .	73
3.2 Badanie dynamiki konformacyjnej homopeptydów . . . . .	73
3.3 Tworzenie węzłów . . . . .	78
3.4 Wyniki dla pojedynczych homopeptydów . . . . .	81
3.5 Próba zastosowania modeli dla białek uporządkowanych . . . . .	83
3.6 Podsumowanie . . . . .	84
<b>4 Symulacje wielu łańcuchów poliglutaminy i polialaniny</b>	<b>85</b>
4.1 Wprowadzenie . . . . .	85
4.2 Protokół symulacji . . . . .	86
4.2.1 Problemy z symulacją wielu łańcuchów . . . . .	87
4.2.2 Algorytm grupowania łańcuchów w agregaty . . . . .	88
4.3 Wyniki . . . . .	88
4.3.1 Agregacja dla różnych temperatur i gęstości . . . . .	88
4.3.2 Właściwości pojedynczych łańcuchów i ich par . . . . .	100
4.3.3 Dynamika łączenia się i rozpadu klastrów . . . . .	103
4.4 Podsumowanie . . . . .	111
<b>5 Symulacje białek glutenu, kukurydzy i ryżu</b>	<b>112</b>
5.1 Wprowadzenie . . . . .	112
5.2 Badane układy . . . . .	114
5.2.1 Fragmenty uporządkowane . . . . .	117
5.3 Protokół symulacji . . . . .	117
5.3.1 Wybór oddziaływanie ze ścianami . . . . .	121
5.3.2 Wybór gęstości . . . . .	123
5.4 Wyniki . . . . .	125
5.4.1 Właściwości białek glutenu, kukurydzy i ryżu . . . . .	125
5.4.2 Elastyczność glutenu na poziomie molekularnym . . . . .	130
5.4.3 Próba odtworzenia krzywych SAXS . . . . .	133
5.4.4 Moduł ścinania glutenu . . . . .	135
5.5 Podsumowanie . . . . .	137
<b>Bibliografia</b>	<b>139</b>

# Rozdział 1

## Wstęp

### 1.1 Wprowadzenie

Ponad jedna trzecia białek w organizmach eukariotycznych jest mniej lub bardziej nieuporządkowana [1], co oznacza, że nie mają one określonej struktury trzeciorzędowej. Takie białka nie zwijają się do struktury natywnej, lecz przyjmują różne kształty (konformacje), tak więc pojedyncze łańcuchy białek mogą wyglądać całkiem inaczej, nawet jeśli znajdują się w identycznych warunkach [2, 3]. Ta plastyczność kształtu pozwala białkom nieuporządkowanym reagować na zmiany otoczenia oraz wiązać się z wieloma innymi białkami. Z tego powodu pełnią one kluczową rolę w regulacji wielu procesów oraz przekazywaniu sygnałów w komórce [2, 4, 5, 6, 7, 8, 9, 10].

Badanie dynamiki zmian konformacji jest wyzwaniem nawet dla pojedynczego łańcucha: białka nieuporządkowane nie krystalizują ze względu na różnorodność kształtu, a widma NMR trudno zinterpretować [11]. W związku z tym istnieją duże problemy z uzyskaniem informacji na temat konformacji najczęściej przyjmowanych przez takie białko (choć ciągły rozwój metody cryo-EM zapowiada, że problemy te będą coraz mniejsze [12], a w bazie PDB pojawia się coraz więcej białek, w których znajdują się także położenia atomów z regionów nieuporządkowanych [13]). Obecnie do badania zespołu statystycznego konformacji białka nieuporządkowanego zwykle używa się metod o niższej rozdzielczości, takich jak pomiar fluorescencji z wykorzystaniem rezonansowego transferu energii (FRET [8, 14, 15]), niskokątowe rozpraszanie promieni rentgenowskich (SAXS [16, 17, 18]), czy wspomniany wcześniej magnetyczny rezonans jądrowy (NMR [19]). Metody te sprzężone są z symulacjami komputerowymi wykonywanymi przy użyciu więzów pochodzących z doświadczenia [8, 15, 16, 19, 20, 21, 22, 23]. Jednak naprawdę dobra symulacja białek nieuporządkowanych powinna odtwarzać ich właściwości bez użycia więzów z doświadczenia. Właśnie takie symulacje (bez użycia więzów) są tematem tej pracy.

Wiele białek nieuporządkowanych było symulowanych przy użyciu pełnoatomowych pól siłowych (takich potencjałów, w których jednemu rzeczywistemu atomowi odpowiada jedna symulowana kulka, czasem z wyjątkiem atomów wodoru [24, 25, 26, 27, 28, 29]). Przykłady tak symulowanych białek

to  $\alpha$ -synukleina [30], poliwalina (polyV [20]) i poliglutaminy (polyQ) [31, 32, 33, 34]. Jednak wierne odwzorowywanie każdego atomu znacznie ogranicza skale czasowe symulacji, a przez to możliwość próbkowania wielu konformacji. Ten problem można obejść rezygnując z odwzorowywania cząsteczek wody (metoda ukrytego rozpuszczalnika, ang. *implicit solvent*), używając obliczeń rozproszonych na superkomputerach i kartach graficznych, oraz rezygnując z “rzeczywistego” czasu na rzecz metod typu wymiana replik czy zmienne w czasie pole siłowe [20, 35, 36].

Jeśli rozważymy bardziej skomplikowany problem symulacji wielu oddziałujących ze sobą białek nieuporządkowanych, modele pełnoatomowe przestaną być wystarczające nawet przy zastosowaniu powyższych metod [37]. Gdyby nawet poświęcić wiele czasu i środków na zastosowanie symulacji pełnoatomowych dla bardzo dużych układów, detale na poziomie pojedynczych atomów nie byłyby po prostu interesujące w takim przypadku. Dlatego istnieje potrzeba rozwoju modeli gruboziarnistych, w których wiele rzeczywistych atomów jest odwzorowanych w symulacji w postaci jednego pseudoatomu. Można używać krótkich symulacji pełnoatomowych do parametryzacji modeli gruboziarnistych, jednak wówczas są one zbudowane tylko aby symulować jeden konkretny układ [38, 39, 40]. Naszym celem była konstrukcja modelu, który będzie działał dla jak najszerzej klasy białek wewnętrznie nieuporządkowanych.

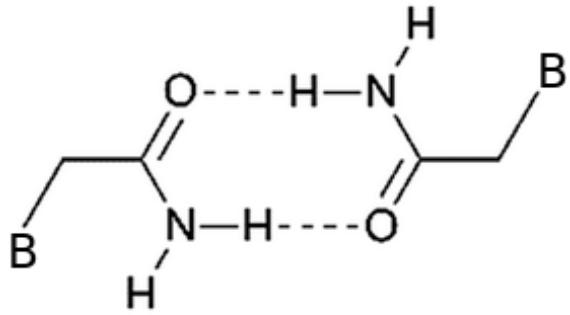
Istnieją modele białek, w których każdy aminokwas jest reprezentowany przez więcej niż jeden pseudoatom [41, 42, 43, 44, 45, 46, 47, 48], oraz takie, w których liczba atomów odpowiada liczbie aminokwasów [49, 50, 51, 52, 53]. Te drugie ze względu na szybkość symulacji są szczególnie przydatne przy analizie dużych układów w dużych skalach czasowych, na przykład przy analizie deformacji i składania kapsydów wirusów [54, 55, 56] czy przy odwzorowaniu rozciągania białek mikroskopem sił atomowych (AFM) z prędkościami odpowiadającymi eksperymentalnym **[IX]**,[57, 58].

Celem tej pracy jest właśnie symulacja jak największych układów w jak najdłuższych skalach czasowych, przy zachowaniu najistotniejszych cech każdego aminokwasu. W komórkach istnieją całe pozbawione membran organelle złożone z białek nieuporządkowanych (oraz RNA) [59, 60]. Są one nietrwałe i uczestniczą w takich procesach jak podział komórki [61], tworzenie jąderka [62], czy regulacja mitozy [63]. Takie organelle zachowują się jak krople cieczy: są bardzo lepkie, mogą się łączyć, wymieniać składem z cytoplazmą czy zmieniać kształt pod wpływem ruchu [64, 65, 66]. W otoczeniu takiej kropli tworzące ją białka występują w dużo mniejszym stężeniu. Dlatego tworzenie kropel przez białka nieuporządkowane można opisać jako przemianę fazową typu ciecz-ciecz [64]. Krople są utrzymywane razem przez przyciągające i anizotropowe oddziaływanie takie jak wiązania wodorowe. Istnienie takich kropel wykazano także w układach nieorganicznych [67, 68].

W przypadku kropel białkowych mają one rozmiar rzędu mikrona [64] więc pozostają na razie poza możliwościami modeli gruboziarnistych, jednak sam proces agregacji białek nieuporządkowanych w większe struktury jest już w zasięgu tych modeli (np. proces tworzenia się granicy faz został już zasymulowany [69]). Jednak w użytym do tego celu prostym modelu gruboziarnistym [49, 70] oddziaływanie przyciągające między aminokwasami polarnymi są bardzo słabe [71], dlatego nie nadaje się on do symulacji agregacji białek nieuporządkowanych bogatych w aminokwasy takie jak glutamina (Q). Białka z fragmentami składającymi się wyłącznie z glutaminy tworzą w komórce długo żyjące agregaty [72, 73, 74] (powstają one także dla białek polyQ, składających się w całości z glutamin, np. Q<sub>75</sub> [75]).

Agregacja białek bogatych w glutaminę jest ważnym tematem badań, ponieważ jest ona obserwowana w chorobach neurozwyrodnieniowych wywoływanych przez takie białka [78]. Nie oznacza to, że sama skłonność do agregacji odpowiada za ich toksyczność: długie łańcuchy polyQ tworzą czasem węzły [34], które mogą zatrzymywać ich degradację w proteasomie [79, 80]. Niedegradowane białka mogą się gromadzić w nadmiarze i powodować agregację, która jeszcze bardziej utrudnia degradację [81]. Najbardziej znaną z tych chorób neurozwyrodnieniowych jest choroba Huntingtona. Białko, które ją wywołuje (huntingtyna), zawiera przy N-końcu fragment, kodowany przez ekson 1 genu huntingtyny. Fragment ten jest odcinany przez kaspazy [82, 83]. Zawiera on trakt Q<sub>n</sub> glutamin, a toksyczność pojawia się gdy  $n$  przekracza próg około 40 – co odpowiada progowi tworzenia się węzłów [34]. Przedstawiony w tej pracy model potrafi symulować tworzenie się węzłów w łańcuchach poliglutaminy **[I]**. Dopiero niedawno agregaty tworzone przez białka kodowane przez ekson 1 huntingtyny zaczęto uznawać za krople - zaobserwowano w nich przejście ciecz-ciało stałe [84, 85, 86].

Warto zrobić tu dygresję na temat tego, dlaczego białka bogate w hydrofilową glutaminę tak chętnie łączą się ze sobą, mimo że jest ona hydrofilowa, więc woda powinna być dla tych białek dobrym rozpuszczalnikiem. Jest tak dlatego, że w teorii łańcuchy boczne glutaminy mogą tworzyć ze sobą po dwa wiązania wodorowe (Rys. 1.1), co może dawać równie korzystną energię swobodną co w przypadku wysycenia tych wiązań wodorowych przez wodę. Trzeba tu odróżnić amorficzne krople poliglutaminy, w których takie oddziaływanie łańcuchów bocznych mogą odgrywać znaczącą rolę, od złogów amyloidowych, które tworzą się przy użyciu oddziaływań łańcucha głównego [87].



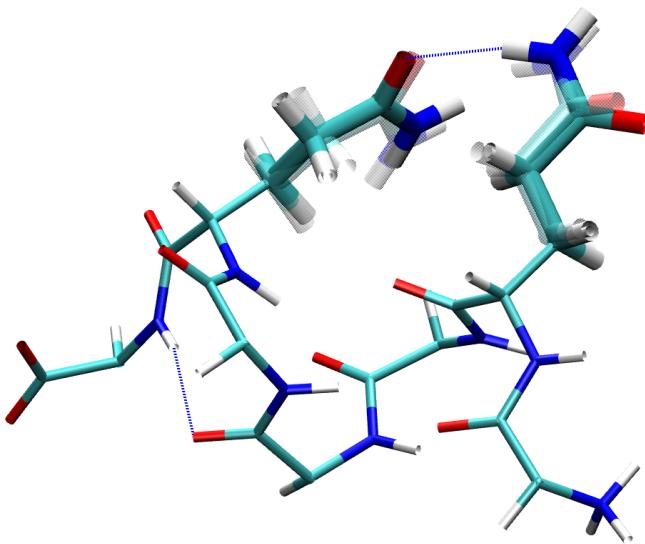
Rysunek 1.1: Schemat tworzenia wiązań wodorowych między łańcuchami bocznymi glutamin. Litera B oznacza resztę białka.

Prof. Sobolewski z IFPAN wykonał nieopublikowane symulacje bardzo dokładną metodą teorii funkcjonalów gęstości (DFT) dla 9 par glutamin<sup>1</sup>. Jeden z 9 przypadków jest pokazany na Rys. 1.2. Tym niemniej w żadnej symulacji przed lub po relaksacji DFT nie utworzyły się dwa wiązania wodorowe, co sugeruje, że między dwoma łańcuchami bocznymi glutamin nie zawsze powstają oba wiązania.

Agregacja białek bogatych w glutaminę była już symulowana przez modele gruboziarniste [87, 88, 89], jednak w pierwszej z cytowanych prac symulowano tylko dwa łańcuchy. W drugiej symulowano co najwyżej 1152 aminokwasy naraz, co daje niewielką statystykę dotyczącą rozmiarów klastrów tworzących przez agregujące łańcuchy (symulacje agregacji w tej rozprawie obejmują co najmniej 1800 aminokwasów). W trzeciej z cytowanych prac jeden pseudoatom reprezentował cały łańcuch, co uniemożliwiało badanie takich aspektów agregacji jak anizotropowość oddziaływań. Jednak przede wszystkim w żadnej z powyższych prac nie wyznaczono dla poliglutaminy diagramu fazowego ani nie wyznaczono prawa potęgowego dotyczącego czasu życia kontaktów między łańcuchami [IV].

Istnieje natomiast zestaw białek nieuporządkowanych bogatych w glutaminę, których agregaty jeszcze nigdy wcześniej (wg wiedzy autora) nie były symulowane: jest to gluten (Rys. 1.3). Jego niezwykłe właściwości wiskoelastyczne powodują, że posiada on zarówno cechy ciała stałego jak i cieczy. Te właściwości można scharakteryzować przez użycie dynamicznego modułu ścinania  $G^*$ , który jest jedną z niewielu wielkości niezależnych (w pewnym zakresie) od kształtu i wielkości próbki. Właśnie chęć

<sup>1</sup>struktury początkowe pochodziły z bazy PDB, oprócz glutamin symulowane były także aminokwasy sąsiadujące razem z nimi w sekwencji (razem 6 lub mniej aminokwasów). łańcuchy były zakończone grupą acetylową od N-końca i grupą N-Me od C-końca. Symulacje polegały na relaksacji na poziomie DFT/TPSS w bazie SVP w dielektrycznej wńęce przy założeniu, że rozpuszczalnikiem jest woda.



Rysunek 1.2: Peptyd z 6 aminokwasów (łańcuchy boczne glutamin pogrubione) pokolorowany wg schematu CPK (azot - niebieski, tlen - czerwony, węgiel - morski, wodór - biały). Pozycje łańcuchów bocznych glutamin przed relaksacją DFT są półprzezroczyste. Wiązania wodorowe są zaznaczone na niebiesko.

wyznaczenia  $G^*$  glutenu w symulacji była pierwotnym powodem powstania modeli gruboziarnistych opisanych w publikacjach [I, II]. Mimo że ostatecznie uzyskana wielkość różni się od tej uzyskanej w doświadczeniu [V], należy pamiętać, że dynamiczny moduł ścinania poznaje się poddając próbce periodycznym deformacjom, które w doświadczeniu są rzędu 1 Hz [90], a w symulacji rzędu 10 kHz [V]. Z powodu tak dużej różnicy skal czasowych nie spodziewamy się uzyskać zgodności ilościowej, zwłaszcza że  $G^*$  zależy od częstości. Jednak nasz model umożliwia nam poznanie molekularnych mechanizmów wpływających na elastyczność glutenu oraz pozwala wyznaczyć, które czynniki są dla tej elastyczności najważniejsze. Jest to szczególnie ważne, ponieważ białka glutenu są nierozpuszczalne w wodzie, co znaczaco utrudnia badanie ich konformacji [91]. Jedyne symulacje tych białek dotyczą pojedynczych łańcuchów, a nie agregatów [92, 93], tak więc symulacja jak wygląda odpowiedź glutenu na deformację na poziomie pojedynczych białek jest ważnym celem badawczym, zwłaszcza uwzględniając znaczenie przemysłowe elastyczności glutenu przy wypieku pieczywa [94].

Innym układem, którego tak jak gluten nie da się w praktyce symulować pełnoatomowo, jest cellulosem: kompleks białek trawiących celulozę, połączonych nieuporządkowanymi łącznikami [95, 96, 97]. Nie będzie on omawiany w tej rozprawie, może jednak być symulowany przedstawionym tu modelem.

Niezależnie od korzyści, jakie płyną z lepszego poznania mechanizmów agregacji poliglutaminy i elastyczności glutenu, model umożliwiający symulacje tysięcy aminokwasów w skalach czasowych rzędu



Rysunek 1.3: Rozciąganie makroskopowej próbki glutenu. Klatka z filmu [98].

milisekund (przy zachowaniu kluczowych cech tych aminokwasów) jest ciekawy sam w sobie.

Ze względu na naturę problemu (periodyczna deformacja glutenu i śledzenie kinetyki agregacji poliglutaminy) zdecydowaliśmy się na prostą dynamikę molekularną, bez użycia specjalnych technik próbkowania takich jak wymiana replik [35] czy metoda Monte Carlo [99]. Z tego samego powodu nie użyliśmy dynamiki Brownowskiej [100] ani nieciągłej dynamiki molekularnej [101, 102] (nie dałoby się w tym przypadku obliczyć naprężeń wywoływanych przez deformację glutenu). Nasz model jest na tyle uproszczony, że trudno byłoby wyprowadzić potencjał oddziaływań między aminokwasami bezpośrednio z rządzących nimi kwantowych praw fizyki (choć jest to możliwe [42]), dlatego zdecydowaliśmy się użyć potencjału empirycznego, który korzysta z wiedzy na temat istniejących struktur białek [103, 104, 105]. Ze względu na duże rozmiary badanych układów zrezygnowaliśmy z pseudoatomów reprezentujących wodę [46], a także ograniczyliśmy się do jednego pseudoatoma na aminokwas. Poza tym nasz model to jeden z niewielu, gdzie białko może być częściowo nieuporządkowane, a częściowo uporządkowane (inne modele które to umożliwiają [46, 106, 107, 108] nie mogły być zastosowane z któregoś z podanych wyżej powodów). Warto zauważyć, że modele gruboziarniste stosowano od dawna z dużym powodzeniem do badań dużych zmian konformacyjnych białek uporządkowanych [25, 99, 109, 110, 111, 112]. Sama konstrukcja modelu gruboziarnistego pozwala na lepsze zrozumienie badanego układu poprzez wyrażenie skomplikowanych oddziaływań w prostszy sposób: mimo że jeden aminokwas jest reprezentowany przez jeden pseudoatom, przedstawione tu modele potrafią odróżnić kiedy oddziałuje on przy pomocy łańcucha bocznego, a kiedy łańcucha głównego. Metoda ta może być wykorzystana także dla innych polimerów z grupami bocznymi. Podobne podejście zostało (niezależnie) zastosowane dla oddziaływania białek z łańcuchami kwasów nukleinowych RNA i DNA [113] (jeden nukleotyd był tam reprezentowany przez 3 pseudoatomy). Jednak w tej rozprawie prezentowane są (wg wiedzy autora) pierwsze dwa modele z jedną kulką na aminokwas, które uwzględniają podczas dynamiki molekularnej kierunkowość oddziaływań łańcucha bocznego i głównego. Kierunkowość oddziaływań nie jest brana pod uwagę w konkurencyjnym modelu [49, 70], użytym do symulacji wspomnianej separacji faz [69, 114].

Do analizy wyników prezentowanych tu symulacji wykorzystane zostały takie wielkości jak promień bezwładności, odległość między końcami łańcucha białkowego, liczba i rodzaj kontaktów między aminokwasami (definicje kontaktów podane będą w następnych rozdziałach), liczba i rozmiar wnęk tworzonych podczas agregacji (wyznaczonych algorytmem Spaceball [115, 116]) czy liczba splatań (wyznaczonych algorytmem Z1 [117, 118, 119, 120]).

## 1.2 Plan pracy

Rozprawa ta została podzielona na cztery główne rozdziały:

“Budowa modelu gruboziarnistego” przedstawia konstrukcję modelu gruboziarnistego do symulacji dynamiki molekularnej białek nieuporządkowanych bądź częściowo uporządkowanych **[I]**. Model ten był dalej rozwijany **[II]**, co także jest opisane, jednak do symulacji w pozostałych rozdziałach użyty został pierwszy model **[I]**. Szczegółowe techniczne potrzebne do użycia modelu na własnym komputerze nie są opisane w tej rozprawie, są jednak dostępne w artykule **[VI]**.

“Wyniki symulacji pojedynczych łańcuchów” opisuje symulacje pojedynczych łańcuchów białek nieuporządkowanych, w szczególności metody badań ich zmian konformacyjnych czy tworzenie się węzłów w poliglutaminie. Rozdział ten obejmuje publikacje **[I,III]**, jednak przedstawiona jest tylko niewielka część artykułu **[III]**, natomiast część wyników (klastrowanie konformacji i ich czasy życia) jest opisana szerzej niż w tych dwóch publikacjach.

“Symulacje wielu łańcuchów poliglutaminy i polialaniny” omawia wyniki uzyskane w artykule **[IV]**, zawiera także te informacje z rozdziałów w książkach **[VII,VIII]**, które powstały podczas prac nad doktoratem. Niektóre dane dotyczące innych metod wyznaczania diagramu fazowego nie były jednak jeszcze publikowane.

“Symulacje białek glutenu, kukurydzy i ryżu” to rozszerzona wersja artykułu **[V]**.

# Rozdział 2

## Budowa modelu gruboziarnistego

### 2.1 Wprowadzenie

Praca nad doktoratem doprowadziła do powstania dwóch odrębnych modeli gruboziarnistych opartych na dynamice atomów węgla  $C_\alpha$ .

Pierwszy opiera się na idei kontaktu między aminokwasami: para aminokwasów będąca blisko siebie (ewentualnie spełniająca jeszcze inne kryteria) tworzy kontakt.

Drugi model zawiera nowy, niestosowany wcześniej potencjał wykorzystujący niewłaściwy kąt dwuścienny. Jest on interesującą alternatywą dla pierwszego.

Podpunkty “Symulacje dynamiki molekularnej”, “Sztywność łańcucha” i “Wybór temperatury” są wspólne dla obu modeli. Jednak ponieważ większość symulacji w tej rozprawie wykonana została pierwszym modelem, to jemu poświęcone zostanie więcej uwagi.

Lista par aminokwasów będących w kontakcie wyznacza mapę kontaktów (dla danej konformacji białka). Mapa kontaktów jest pojęciem kluczowym dla modeli gruboziarnistych opartych na strukturze natywnej [50, 51, 121, 122, 123], w których natywna konformacja odpowiada minimum potencjału w modelu (lub jest bardzo blisko tego minimum). Takie modele używają metody ukrytego rozpuszczalnika (brak jest pseudoatomów reprezentujących wodę). Taka konstrukcja modelu jest zgodna z zasadą minimalnej frustracji i prowadzi do optymalnego lejka zwijania w przypadku białek uporządkowanych [124, 125]. Mapa kontaktów jest w takim wypadku tworzona na podstawie struktury natywnej [122, 126, 127, 128], a potencjał jest skonstruowany tak, aby tylko aminokwasy będące w tej natywnej mapie kontaktów przyciągały się podczas symulacji. Nie ma jednak jednego uniwersalnie przyjętego kryterium jak konstruować mapę kontaktów. W dotychczas używanym modelu gruboziarnistym dla białek uporządkowanych (opisywany tu model stanowi jego rozwinięcie i uzupełnienie dla białek nie-uporządkowanych) używamy mapy kontaktów opartej na przekrywaniu się kul reprezentujących ciężkie atomy w strukturze natywnej [129, 130, 131], tzn. dwa aminokwasy są w kontakcie jeśli ich ciężkie atomy się przekrywają. Sparametryzowany na podstawie eksperymentów rozciągania białek model [53] pokazuje, że potencjał Lenard-Jonesa (LJ) zastosowany dla aminokwasów z tak utworzonej mapy

kontaktów poprawnie odtwarza dynamikę rozciągania i zwijania białek. Aminokwasy spoza mapy kontaktów oddziałują ze sobą odpychającą częścią potencjału LJ.

Przedstawione powyżej podejście oparte na strukturze natywnej nie może być zastosowane do białek nieuporządkowanych, które nie mają jednej takiej struktury [1], lecz dynamicznie przyjmują wiele różnych konformacji. Nie mają też tak głębokich minimów w krajobrazie energetycznym, lecz przeskakują z jednego płytkego minimum w drugie [III]. Nie ma w związku z tym jednej stałej mapy kontaktów. Można jednak w każdej chwili uznać pewne pary aminokwasów za przyciągające się, tworząc chwilową mapę kontaktów. Tak jak w modelu dla białek uporządkowanych, pseudoatomy reprezentujące aminokwasy mają współrzędne odpowiadające atomom węgla  $C_\alpha$  tych aminokwasów.

O tym, która para aminokwasów tworzy chwilowy kontakt, decydują trzy kryteria:

1. odległość między aminokwasami
2. odpowiedni kierunek jaki miałyby grupy boczne i wiązania wodorowe łańcucha głównego (wyznaczony tylko na podstawie współrzędnych pseudoatomów, odpowiadających atomom  $C_\alpha$ )
3. liczba kontaktów jaką dany aminokwas może utworzyć

Jeśli wszystkie trzy kryteria są spełnione, kontakt zostaje utworzony. Zanim kryteria te zostaną omówione dokładniej, model zostanie omówiony od podstaw.

## 2.2 Symulacje dynamiki molekularnej

W modelu gruboziarnistym nie są symulowane żadne atomy wody ani rozpuszczonych w niej jonów (używamy metody ukrytego rozpuszczalnika). Układ ewoluje w czasie zgodnie z zasadami dynamiki molekularnej przy uwzględnieniu tłumienia zależnego od prędkości i szumu termicznego (dynamika Langevina). Rozpuszczalnik jest zatem reprezentowany przez tłumienie i szum termiczny, odpowiadający temperaturze  $T$ . Charakterystyczna skala czasowa  $\tau$  jest rzędu 1 ns [132], a dynamika jest nadtlumiona (ruch pseudoatomów jest bardziej dyfuzyjny niż balistyczny). Współczynnik tłumienia  $\gamma$  wynosi w tym modelu  $2m/\tau$ , gdzie  $m$  to średnia masa aminokwasu. Bardziej realistyczne wartości  $\gamma$  są około 25 razy większe [133], jednak przyjęcie ich zmniejszyłoby skalę czasową (symulacje są zatem ekstrapolacją dla długich skal czasowych).

Aminokwasy są połączone w łańcuch przy pomocy potencjału harmonicznego ze stałą sprężystością  $k = 100 \text{ \AA}^{-2} \cdot \epsilon$  z minimum odpowiadającym odległości 3.8 Å. Jednostka energii to  $\epsilon = 1.58 \text{ kcal/mol}$ ,

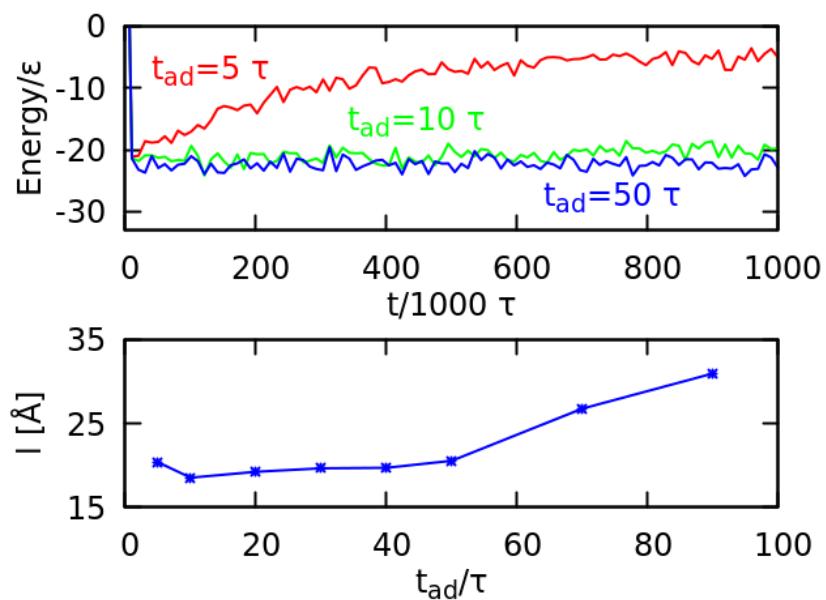
odpowiadająca (w dużym przybliżeniu) energii z jaką przyciągają się aminokwasy tworzące między sobą jedno wiązanie wodorowe [134].

Równanie ruchu dla  $i$ -tego aminokwasu to:

$$m \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i - \gamma \frac{d\vec{r}_i}{dt} + \vec{\Gamma}_i \quad (2.1)$$

gdzie  $\vec{r}_i$  to pozycja aminokwasu,  $\vec{F}_i$  to siła wynikająca z potencjału,  $\gamma = 2m/\tau$  to współczynnik tłumienia, a  $\vec{\Gamma}_i$  to biały szum termiczny z wariancją  $\sigma^2 = 2\gamma k_B T$ .

Równania ruchu są rozwiązywane przez algorytm predyktor-korektor 5. rzędu [135]. Jednostka  $\tau$  odpowiada 200 krokom symulacji. Jak wspomniano wcześniej, kontakty tworzą się dynamicznie. Kontakt oznacza przyciągający potencjał LJ. Nagłe włączenie takiego potencjału mogłoby doprowadzić do niestabilności numerycznych, dlatego po utworzeniu kontaktu odpowiedni potencjał jest włączany quasi-adiabatycznie: głębokość studni potencjału rośnie liniowo od 0 do  $-\epsilon$  w ciągu  $10\tau$ . Taka skala czasowa (2000 kroków symulacji) jest dostatecznie dłuża, aby układ uległ termalizacji. Wyłączanie kontaktu zachodzi analogicznie. Krótsze czasy przełączania  $t_{ad}$  prowadzą do wzrostu całkowitej energii w układzie (górną część Rys. 2.1). Dłuższe czasy (do  $40\tau$ ) nie wpływają na wyniki symulacji, ale czasy powyżej  $40\tau$  prowadzą do zwiększonej mobilności łańcucha białkowego, która prowadzi do nadmiernego zwiększenia rozmiarów układu (dolna część Rys. 2.1).



Rysunek 2.1: Dane dla symulacji poliglutaminy o długości 30 aminokwasów ( $Q_{30}$ ) dla różnych czasów quasi-adiabatycznego przełączania kontaktów  $t_{ad}$ . Górnny panel pokazuje wygładzoną energię całkowitą  $E$  w funkcji czasu. Dolny panel pokazuje odległość między końcami  $l$  w funkcji  $t_{ad}$ , uśrednioną po 100 symulacjach. Błąd średniej jest rzędu rozmiaru punktu na wykresie. Wariancja  $\sigma^2$  także rośnie gdy  $t_{ad}$  przekracza  $50\tau$ . Oparte na rys. S7 z artykułu [I].

## 2.3 Sztywność łańcucha

Sztywność łańcucha jest zwykle utrzymywana dzięki potencjałom działającym na kąty płaskie i dwuścienne utworzone przez sąsiadujące ze sobą aminokwasy. W modelach opartych na strukturze natywnej minima tych potencjałów odpowiadają kątom ze struktury natywnej. W przypadku białek nieustrukturyzowanych tak jednoznacznych minimów nie ma, dlatego zamiast tego używamy potencjałów zaproponowanych dla białek nieuporządkowanych przez grupę Ghavani i in. [136]

Potencjały zostały uzyskane poprzez zastosowanie metody inwersji boltzmannowskiej do zestawu wyznaczonych eksperymentalnie kątów charakteryzujących wiązania chemiczne między aminokwasami dla nieuporządkowanych fragmentów białek (posiadających strukturę kłębka statystycznego, ang. *random coil*) [136]. Potencjały te są dość szerokie: w zasięgu fluktuacji termicznych  $kT$  znajduje się około  $60^\circ$  (potencjały oparte na strukturze natywnej są zwykle dużo węższe).

Użyty potencjał zależy od tego, jakie aminokwasy tworzą dany kąt. Aminokwas może należeć do jednego z trzech typów: glicyna, prolina albo reszta (X). Dla kąta płaskiego związanego ze środkowym z trzech aminokwasów jest 27 możliwych kombinacji. Zastosowany został tu potencjał poziomu 2 [136], co oznacza, że potencjał kąta płaskiego zależy tylko od środkowego aminokwasu i od tego czy poprzedzający go sąsiad jest proliną (co daje 6 kombinacji). Kolejność aminokwasów ma znaczenie, ponieważ odpowiada za chiralność łańcucha. Rozróżnianie glicyny i proliny ma szczególne znaczenie dla glutenu, który ma wysoką zawartość tych aminokwasów [137].

W przypadku kąta dwuściennego trzeba rozpatrzyć cztery sąsiednie aminokwasy. Na poziomie 2 potencjał dihedralny zależy od 2 środkowych, co daje 9 kombinacji.

Potencjały dla kątów płaskich i dwuściennych zostały uzyskane na podstawie biblioteki fragmentów o strukturze kłębka statystycznego dla temperatury pokojowej. Pierwotnie miały one odwzorowywać właściwości zdenaturowanych białek, w których aminokwasy się nie przyciągają [136]. W związku z tym potencjał ten nie wykazuje preferencji do żadnej struktury drugorzędowej (chyba że uznamy kłębek statystyczny za rodzaj struktury drugorzędowej). Jednak dla białek uporządkowanych w zwykłych warunkach aminokwasy mogą się przyciągać i tworzyć przejściowe bądź trwałe  $\alpha$ -helisy lub  $\beta$ -kartki. Aby takie struktury mogły powstać, dopuszczane jest tworzenie kontaktów między  $i$ -tym a  $i + 3$ im aminokwasem w łańcuchu ponieważ takie kontakty odpowiadają wiązaniom wodorowym stabilizującym  $\alpha$ -helisy w reprezentacji pełnoatomowej łańcucha głównego [138]. Natura kontaktów  $i, i + 4$  zostanie omówiona później.

Potencjały dla kątów płaskich są pokazane na Rys. 2.2, a dla kątów dwuściennych na Rys. 2.3. W obu przypadkach do potencjałów statystycznych została dopasowana funkcja analityczna: wielomian 6. rzędu dla kątów płaskich, a dla dwuściennych funkcja zawierająca funkcje sinus i cosinus kąta dwuściennego.

Wszystkie współczynniki tych funkcji, podane tutaj w kJ/mol, zostały przekształcone do jednostek energii używanych w programie,  $\epsilon = 6.6 \text{ kJ/mol} = 1.58 \text{ kcal/mol}$ . Wartość  $\epsilon$  odpowiada energii kontaktu wyprowadzonej z symulacji pełnoatomowych [139].

### 2.3.1 Potencjał kąta płaskiego

Potencjał statystyczny kąta płaskiego został dopasowany do funkcji:

$$ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g = 0 \quad (2.2)$$

gdzie  $x$  to kąt płaski w radianach. Współczynniki znajdują się w tabeli 2.1.

Kombinacja aminokwasów	g	f	e	d	c	b	a
OGY	137767.79	-417519.49	523500.78	-347689.12	129057.84	-25394.62	2070.23
OGP	54278.92	-166180.67	210155.26	-140514.29	52413.91	-10347.96	845.30
OPY	228674.80	-725717.73	953197.76	-663471.51	258240.30	-53322.81	4566.15
OPP	70917.09	-225383.01	295803.17	-205330.47	79600.56	-16366.17	1396.80
OXY	104836.85	-322892.77	411580.60	-277931.71	104885.76	-20978.03	1737.72
OXP	111628.30	-353562.64	462991.27	-320775.91	124020.92	-25374.95	2147.03

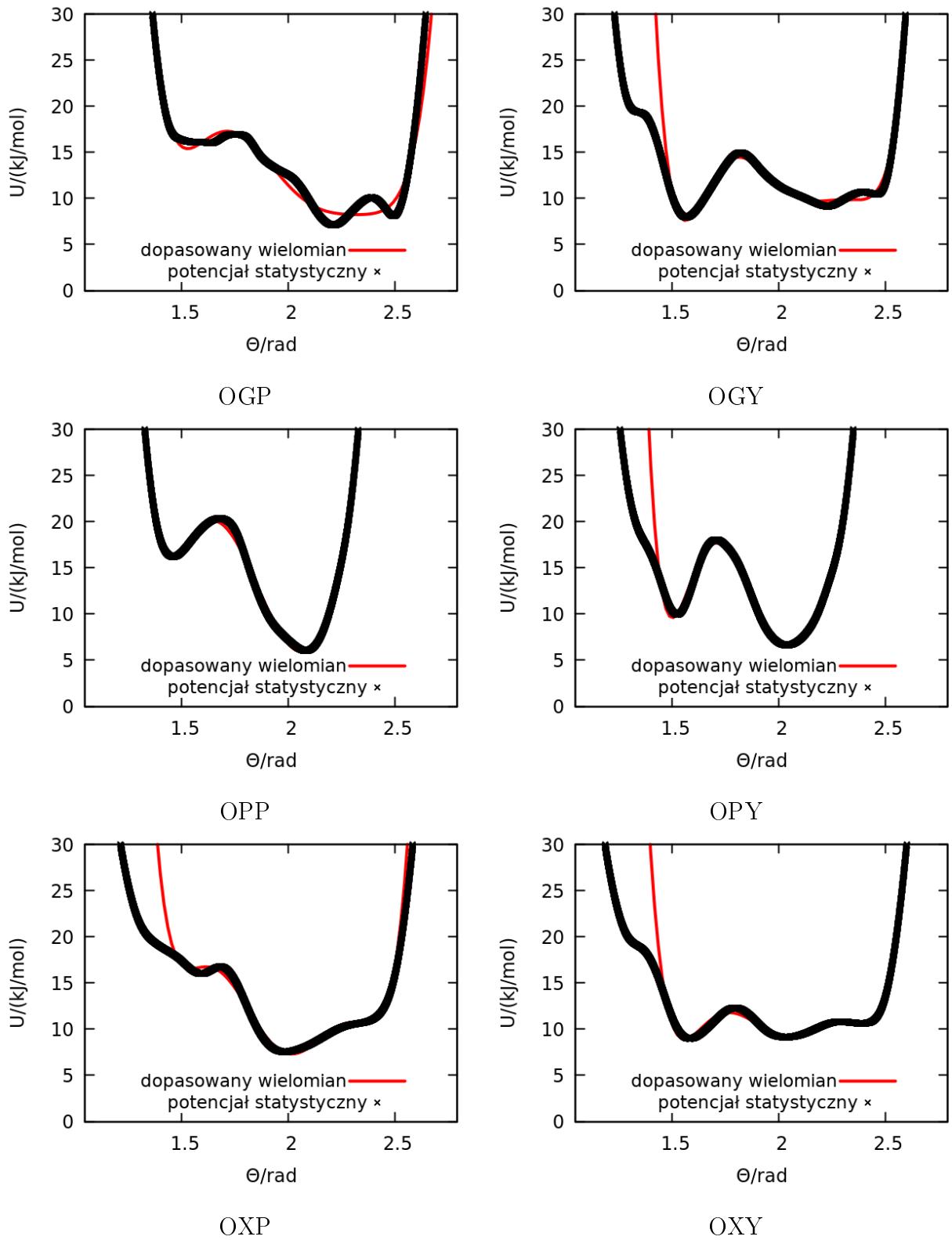
Tabela 2.1: Współczynniki użyte w dopasowaniu wielomianu 6. stopnia (w jednostkach kJ/mol) dla 6 kombinacji: O to dowolny aminokwas, Y dowolny poza proliną, X dowolny poza glicyną (G) i proliną (P).

### 2.3.2 Potencjał kąta dwuściennego

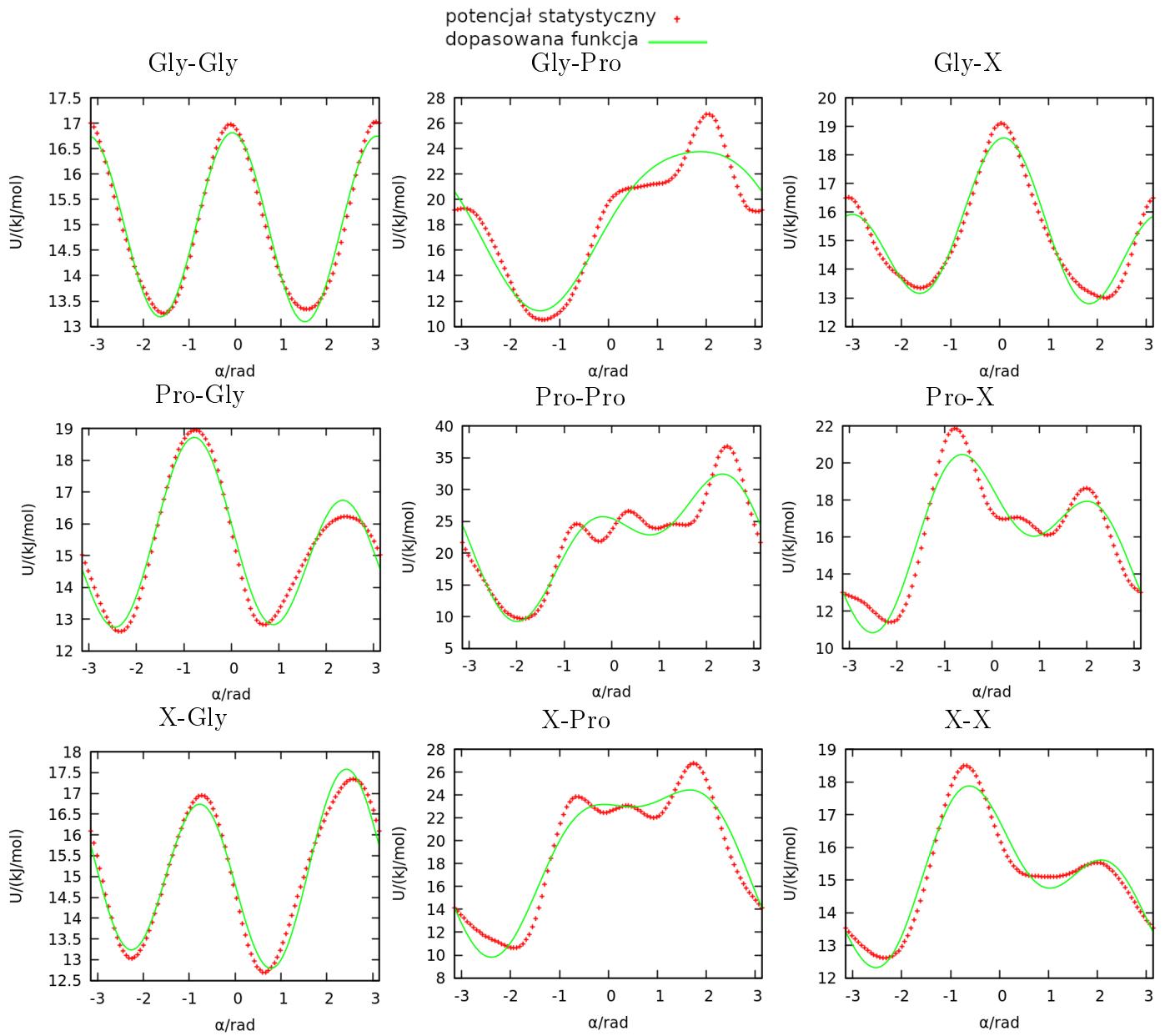
Potencjał statystyczny kąta dwuściennego został dopasowany do funkcji:

$$a \sin(x) + b \cos(x) + c \sin^2(x) + d \cos^2(x) + e \sin(x) \cos(x) + f \quad (2.3)$$

gdzie współczynniki  $a, b, c, d, e$  są podane w tabeli 2.2.



Rysunek 2.2: Statystyczny potencjał kąta płaskiego [136] (czarny) i dopasowany wielomian (czerwony) dla kombinacji opisanych w tabeli 2.1. Oparte na rys. S5 z artykułu [I].



Rysunek 2.3: Statystyczny potencjał kąta dwuściennego [136] (punkty) i dopasowana funkcja (linie) dla kombinacji opisanych w tabeli 2.2. Oparte na rys. S6 z artykułu [I].

Kombinacja aminokwasów	f	a	b	c	d	e
GG	2.117	-0.008	0.004	-0.125	0.425	-0.061
GP	2.639	0.929	-0.185	0.016	0.286	0.073
GX	2.149	-0.006	0.203	-0.161	0.461	0.133
PG	2.165	-0.102	0.109	0.149	0.152	-0.742
PP	3.205	1.171	0.091	-0.254	0.558	-1.570
PX	2.304	0.115	0.429	0.201	0.100	-0.803
XG	2.136	0.018	-0.071	0.122	0.179	-0.624
XP	2.740	0.739	0.686	0.219	0.083	-0.791
XX	2.142	0.006	0.257	0.155	0.146	-0.448

Tabela 2.2: Współczynniki dopasowania potencjału kąta dwuściennego (w jednostkach kJ/mol) dla 9 kombinacji 2 środkowych aminokwasów. X oznacza dowolny aminokwas poza glicyną (G) i proliną (P).

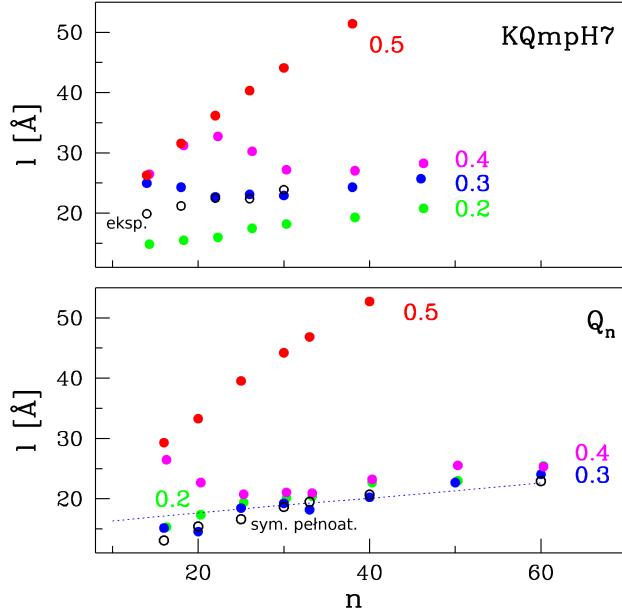
## 2.4 Temperatura pokojowa $T_r$

Optymalna temperatura do przeprowadzania symulacji w modelach gruboziarnistych zależy od przyjętej sztywności łańcucha [140]. Zakres temperatur, w którym białka ustrukturyzowane zwijają się najszybciej (w modelu opartym na strukturze natywnej) wynosi między 0.3 a 0.35  $\epsilon/k_B$  jeśli sztywność jest opisana potencjałem chiralnościowym [140], ale dla potencjału kątów płaskich i dwuściennych (z minimami dla struktury natywnej) ten zakres zwiększa się do około 0.7  $\epsilon/k_B$ . Z wartości jednostki energii  $\epsilon$  i stałej Boltzmanna wynika, że temperatura pokojowa  $T_r$  wynosi około 0.35  $\epsilon/k_B$  (dokładnie 0.38), jednak jak widać na przykładzie białek uporządkowanych optymalna temperatura symulacji może być inna. Dlatego trzeba było zbadać jakie  $T_r$  przyjąć dla modelu białek nieuporządkowanych.

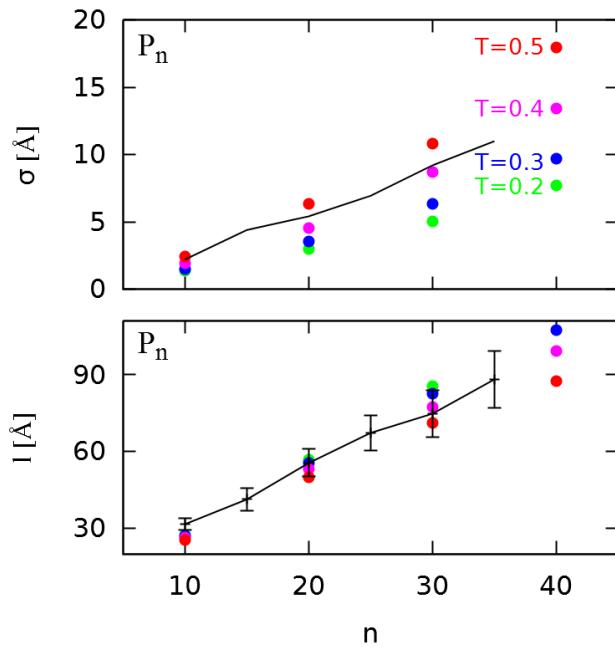
Można do modelu opartego na strukturze natywnej zastosować sztywność łańcucha jak dla białek nieuporządkowanych, wtedy optymalny zakres temperatur zwijania dla białek o kodach PDB 1GB1, 1TIT, 1UBQ, and 2M7D wynosi między 0.3 a 0.35  $\epsilon/k_B$ .

Temperatura  $T_r$  dla białek nieuporządkowanych została wyznaczona na podstawie symulacji poliglutaminy i poliproliny (polyQ i polyP). Średnią odległość między końcami łańcucha można wtedy porównać z wynikami doświadczalnymi [14, 141] (a w przypadku polyQ także z wynikami symulacji pełnoatomowych [34]). Rys. 2.4 pokazuje, że dla temperatury 0.3  $\epsilon/k_B$  średnia odległość między końcami w symulacji najlepiej zgadza się z wynikami symulacji pełnoatomowych (dolny panel) i doświadczenia (górny panel), które były prowadzone dla temperatury pokojowej. Zgodność dla  $T = 0.2 \epsilon/k_B$  jest porównywalnie dobra dla polyQ, ale wyniki dla polyP pokazują, że właściwa temperatura jest większa lub równa 0.3.

Poliprolina praktycznie nie tworzy kontaktów przyciągających między odległymi aminokwasami i posiada dużą sztywność [14], dlatego także jest dobrym układem do wyznaczenia temperatury symulacji. Rys. 2.5 wskazuje, że  $0.38\epsilon/k_B$  jest faktycznie najbardziej odpowiednią temperaturą. Jednak większość symulacji została wykonana w temperaturze  $0.35\epsilon/k_B$  (chyba że zaznaczono inaczej).



Rysunek 2.4: Średnia odległość między końcami łańcucha  $l$  w funkcji liczby aminokwasów w łańcuchu  $n$  w porównaniu do symulacji pełnoatomowych [34] polyQ (dół) oraz do  $l$  wyznaczonego doświadczalnie dla łańcuchów postaci KKWQ <sub>$m$</sub> AKK (góra). Cztery temperatury symulacji są podane w jednostkach  $\epsilon/k_B$ . Pełne punkty odpowiadają symulacjom, czarne puste kółka odpowiadają symulacjom pełnoatomowym [34] lub eksperymentom [141]. Błąd średniej jest mniejszy niż rozmiar punktów. Oparte na rys. S8 z artykułu [I].



Rysunek 2.5: Średnia odległość między końcami łańcucha  $l$  (dół) oraz odchylenie średniej  $\sigma$  (góra) w funkcji liczby aminokwasów w łańcuchu  $n$  w porównaniu do eksperymentu [14]. Cztery temperatury symulacji są podane w jednostkach  $\epsilon/k_B$ . Pełne punkty odpowiadają symulacjom, czarne krzywe eksperymentowi.

## 2.5 Model quasi-adiabatyczny

### 2.5.1 Postać potencjału

Oddziaływanie między aminokwasami bazują na modelach opartych o mapę kontaktów [53, 57, 142], ale mapa kontaktów nie pochodzi teraz ze struktury natywnej, lecz jest uaktualniana w każdym kroku symulacji. Warto pamiętać, że położenie  $i$ -tego pseudoatomu w symulacji ma odpowiadać położeniu atomu węgla  $C_\alpha$  w rzeczywistości.

Kiedy aminokwasy  $i$  oraz  $j$  (w odległości  $r_{i,j}$  od siebie) są w kontakcie, ich oddziaływanie opisuje potencjał Lennard-Jonesa (LJ):

$$V_{L-J}(r_{i,j}) = 4\epsilon \left[ \left( \frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left( \frac{\sigma_{i,j}}{r_{i,j}} \right)^6 \right] \quad (2.4)$$

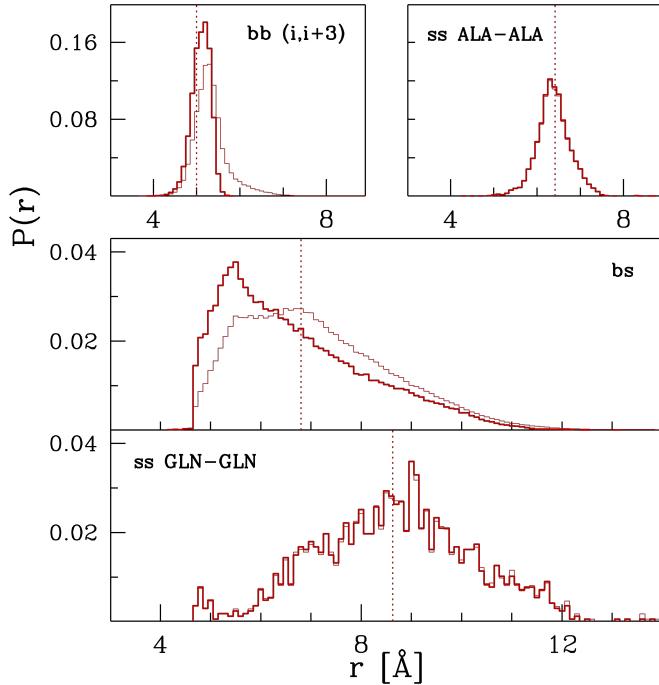
gdzie głębokość potencjału  $\epsilon$  jest taka sama jak w modelu dla białek uporządkowanych: wielkość około 110 pN Å została w nim wyznaczona przez dopasowanie do wyników eksperymentalnego rozciągania białek [57]. Wartość ta jest równa podanej wcześniej wartości 1.58 kcal/mol [139].

Drugi parametr potencjału LJ,  $\sigma_{i,j} = r_{min} \cdot (0.5)^{\frac{1}{6}}$ , (gdzie  $r_{min}$  oznacza położenie minimum potencjału), zależy od tego jakie aminokwasy są w kontakcie. Wybór  $\sigma_{i,j}$  i kryteria utworzenia kontaktu są omówione poniżej. Kontakt jest natomiast zrywany zawsze gdy  $r_{i,j} > f \sigma_{i,j}$  (gdzie  $f = 1.5$ , choć w podpunkcie 2.5.3 dyskutowana jest też wartość 1.3).

### 2.5.2 Rodzaje kontaktów i ich charakterystyczne odległości

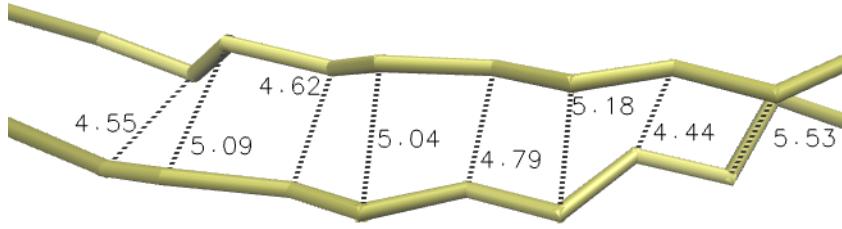
Zakładamy, że fizyka oddziaływania między dwoma bliskimi sobie aminokwasami jest uniwersalna, zatem parametryzację kontaktów do modelu białek nieuporządkowanych zaczeliśmy od stworzenia bazy kontaktów z 21 090 białek uporządkowanych z bazy CATH [143] (wybrane zostały białka, których podobieństwo sekwencji nie przekracza 40%: cath-dataset-nonredundant-S40.pdb). Lista kontaktów w tych białkach została ustalona na podstawie kryterium przekrywania ciężkich atomów [131]. Ciężkie atomy są w nim reprezentowane jako sfery o promieniu zależnym od pozycji w aminokwasie [129]. Promienie wg. Tsai i in. zostały pomnożone przez czynnik 1.24 aby przekrywające się sfery mogły faktycznie odpowiadać przyciąganiu między atomami (czynnik 1.24 odpowiada punktowi przegięcia potencjału LJ) [144]. Jeśli choć jedna sfera należąca do aminokwasu  $i$  przekrywa się z choć jedną sferą należącą do aminokwasu  $j$ , uznajemy istnienie kontaktu natywnego między aminokwasem  $i$  oraz  $j$ . Mimo że pseudoatomy w naszym modelu reprezentują tylko pozycje atomów węgla  $C_\alpha$ , już kryterium wyboru kontaktów z bazy białek uporządkowanych zawiera informacje na temat położenia grup bocznych i rozmiarów pojedynczych atomów.

Rozróżniamy trzy rodzaje kontaktów na podstawie tego, które atomy się przekrywają: kontakty między łańcuchami bocznymi (ss od ang. *sidechain-sidechain*), między fragmentami łańcucha głównego (bb od ang. *backbone-backbone*) oraz między łańcuchem głównym i bocznym (bs od ang. *backbone-sidechain*). Dla danej pary aminokwasów przekrywanie może prowadzić do więcej niż jednego rodzaju kontaktu naraz. Niezależnie od tego ilu rodzajów jest kontakt, zawsze jest on liczony jako jeden. Odległość przypisana danemu kontaktowi odpowiada odległości między atomami  $C_\alpha$ . Wybrane rozkłady tych odległości dla kontaktów bb, bs oraz ss są przedstawione na Rys. 2.6.



Rysunek 2.6: Przykłady rozkładów odległości  $C_\alpha - C_\alpha$  dla kontaktów z bazy CATH. Rozkłady dotyczą kontaktów, które są tylko jednego rodzaju (bb, bs albo ss). Cienkie linie odpowiadają kontaktom na podstawie kryterium przekrywania, natomiast grube linie uwzględniają tylko kontakty, które spełniają także podane niżej kryteria kierunkowe. Pionowe linie przerywane oznaczają odległości przyjęte w modelu. Odpowiadają wartościom średnim z rozkładów zaznaczonych grubą linią. Panel w lewym górnym rogu dotyczy tylko kontaktów bb między  $i$ -tym a  $i + 3$ -im aminokwasem i odpowiada średniej odległości 5.0 Å. Dla dalszych kontaktów średnia to 4.8 Å, patrz Rys. 2.7. Środkowy panel dotyczy kontaktów bs typu  $i, i + k$ , gdzie  $k > 4$ . Kolejne panele dotyczą kontaktów typu ss między alaninami (prawy górny róg) i między glutaminami (dolny panel). W przypadku kontaktów ss rozkład zawiera wszystkie kontakty postaci  $i, i + k$ , gdzie  $k > 2$ . Podobne rozkłady odległości w kontaktach ss dla innych kombinacji aminokwasów przedstawiają Rys. 2.8, 2.9, 2.10. Modyfikacja rys. 1 z artykułu [I].

Aby otrzymać wartościowe informacje z rozkładów odległości, z bazy kontaktów wybrane zostały te, które są tylko jednego rodzaju (ponieważ w modelu kontakt między aminokwasami także może być tylko



Rysunek 2.7: Przykład  $\beta$ -kartki z zaznaczonymi odległościami  $C_\alpha - C_\alpha$ .

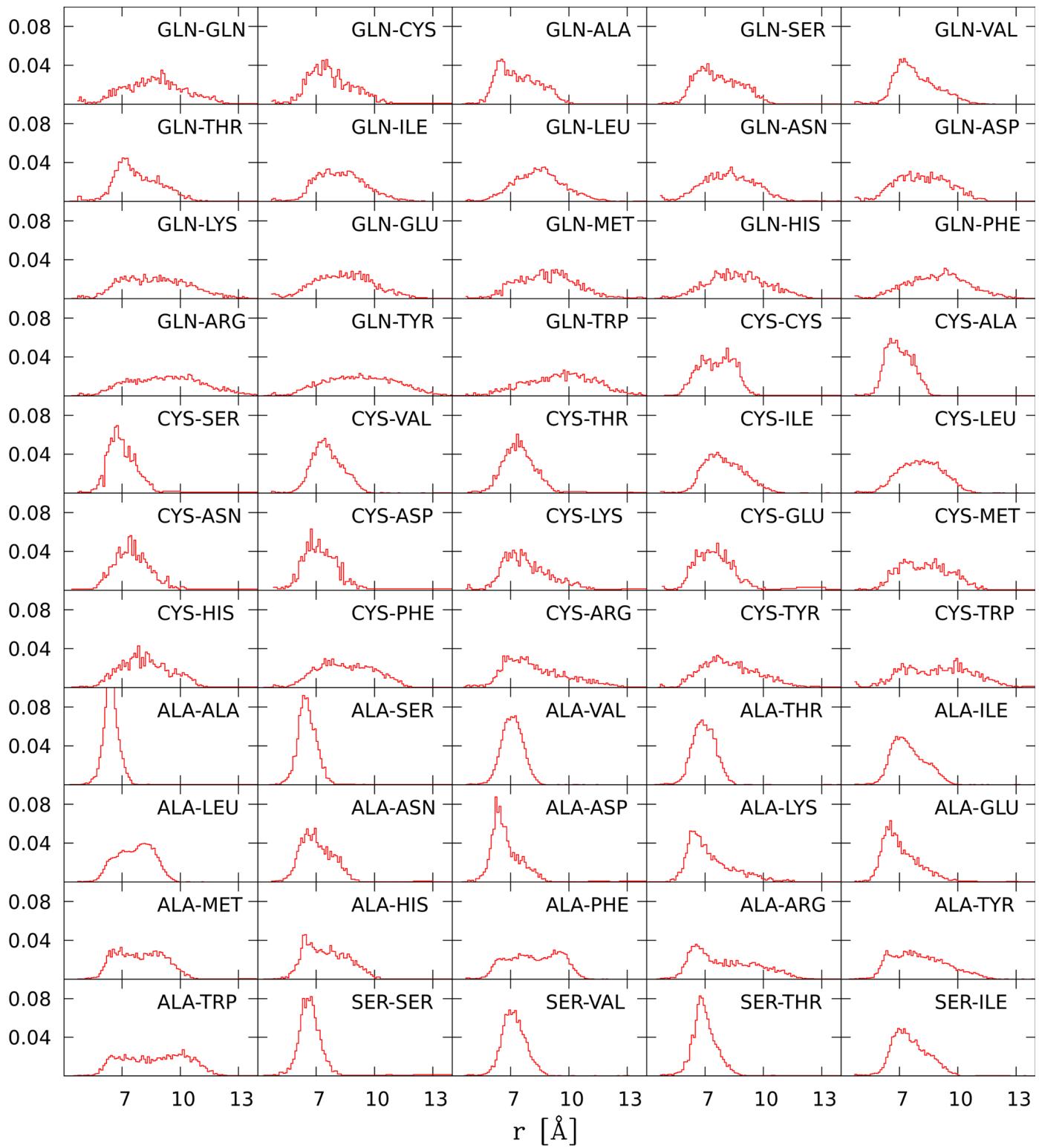
jednego rodzaju naraz - choć w trakcie symulacji rodzaj ten może się zmieniać). Ponieważ tożsamość aminokwasu jest określana tylko przez jego łańcuch boczny, rozkłady odległości zostały podzielone na możliwe pary aminokwasów tylko dla kontaktów ss. Średnie odległości uzyskane z rozkładów dla kontaktów ss są podane w tabeli 2.3. W naszym modelu jednoimiennie naładowane aminokwasy nie mogą tworzyć kontaktów ss, tak samo glicyna i prolina. Dlatego tabela 2.3 zawiera tylko 165 wartości (zamiast 210).

Poza tożsamością aminokwasów tworzących kontakt znaczenie ma też, ile aminokwasów znajduje się pomiędzy nimi w sekwencji białka. Np. odległości bb różnią się dla kontaktów  $i, i + 3$  oraz dla dalszych. Przyjęto minimum w 5 Å, aby dobrze oddać naturę kontaktów  $i, i + 3$  (odpowiadających za  $\alpha$ -helisy) oraz połowę dalszych kontaktów - kontakty typu bb w  $\beta$  kartkach mają bimodalny rozkład odległości, jak to ilustruje Rys. 2.7.

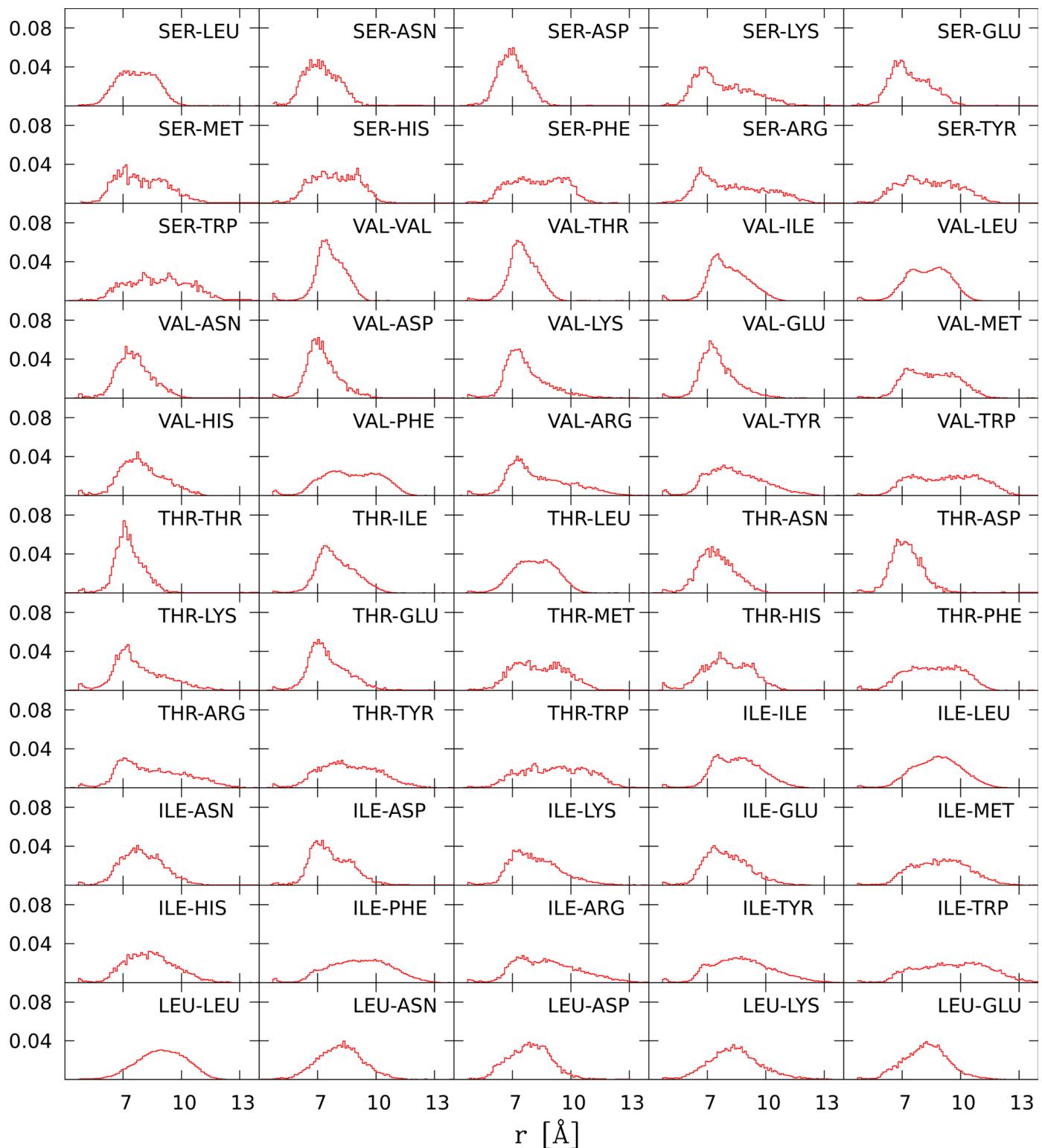
Rysunki 2.8, 2.9 i 2.10 zawierają rozkłady odległości  $C_\alpha - C_\alpha$  dla kontaktów ss, z uwzględnieniem tożsamości aminokwasów biorących udział w kontakcie. Rozkłady te dotyczą tylko sytuacji, gdzie kryteria kierunkowe są spełnione. Szerokość słupka histogramu to 0.1 Å, wysokości słupków są unormowane aby sumowały się do 1.

Odległości podane w tabeli 2.3 odpowiadają użytym minimom potencjału  $r_{min}$ . Są one w zakresie od 6.42 Å (Ala-Ala) do 10.85 Å (Trp-Trp). Ze względu na duże wartości  $r_{min}$  dla par Trp-Trp, pary tych aminokwasów nie mogą tworzyć kontaktów  $i, i + 3$ . Warto zauważyć, że w bazie UniProt najdłuższy fragment polyW złożony wyłącznie z tryptofanów ma długość 6 (patrz tabela 3.1), natomiast w bazie PDB najdłuższy taki fragment ma długość 4 (dla struktury 3N85) i odpowiada skrętowi.

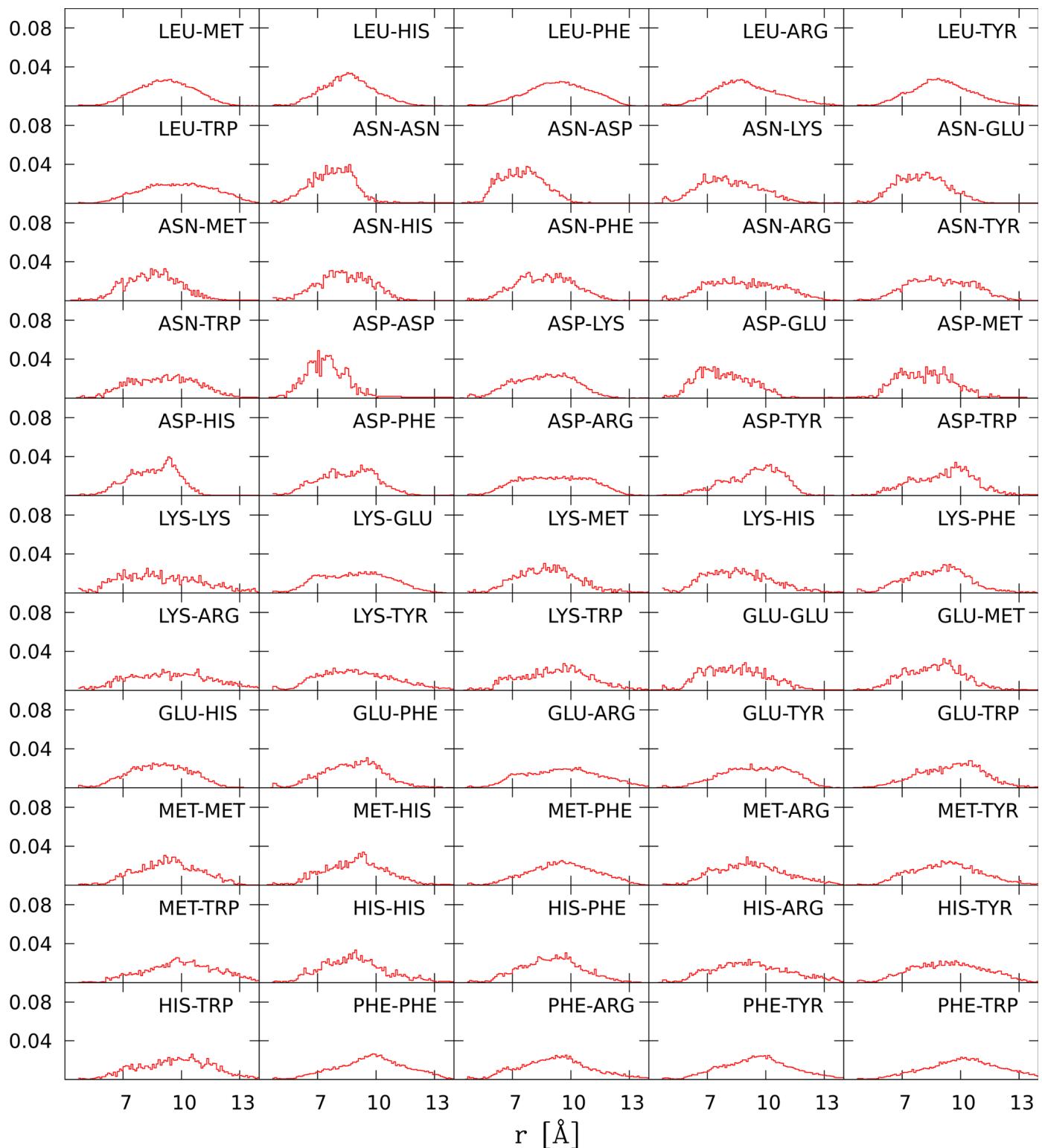
Dla kontaktów bb i bs minima  $r_{min}$  wynoszą odpowiednio 5.0 i 6.8 Å, bez uwzględnienia tożsamości aminokwasów. Uwzględnione zostały za to kontakty  $i, i + 3$  (minima dla kontaktów bb zostały ustalone wyłącznie na podstawie takich kontaktów, minima dla kontaktów bs nie uwzględniają ich w ogóle), ponieważ średnie wartości nie odpowiadają dobrze ani geometrii kontaktów  $i, i + 3$ , ani tych dalszych.



Rysunek 2.8: Rozkłady odległości  $C_{\alpha} - C_{\alpha}$  w kontaktach ss dla danych par aminokwasów, część 1. Oparte na rys. S1 z artykułu [I].



Rysunek 2.9: Rozkłady odległości C<sub>α</sub> – C<sub>α</sub> w kontaktach ss dla danych par aminokwasów, część 2. Oparte na rys. S2 z artykułu [I].



Rysunek 2.10: Rozkłady odległości C<sub>α</sub> – C<sub>α</sub> w kontaktach ss dla danych par aminokwasów, część 2. Oparte na rys. S3 z artykułu [I].

	Gln	Cys	Ala	Ser	Val	Thr	Ile	Leu	Asn	Asp	Lys	Glu	Met	His	Phe	Arg	Tyr	Trp
<b>Gln</b>	8.63																	
<b>Cys</b>	7.72	7.56																
<b>Ala</b>	7.39	6.97	6.42															
<b>Ser</b>	7.64	6.97	6.53	6.65														
<b>Val</b>	7.81	7.56	7.06	7.17	7.65													
<b>Thr</b>	7.77	7.40	6.94	6.97	7.54	7.30												
<b>Ile</b>	8.24	7.95	7.45	7.52	8.06	7.93	8.53											
<b>Leu</b>	8.44	8.07	7.65	7.68	8.29	8.12	8.77	8.93										
<b>Asn</b>	8.19	7.49	7.02	7.18	7.54	7.46	7.96	8.14	7.74									
<b>Asp</b>	8.15	7.18	6.73	6.99	7.22	7.19	7.65	7.86	7.50									
<b>Lys</b>	8.69	7.83	7.26	7.73	7.69	7.79	8.16	8.39	8.11	8.59								
<b>Glu</b>	8.41	7.45	7.04	7.41	7.50	7.51	7.97	8.20	8.00		8.90							
<b>Met</b>	8.84	8.29	7.91	7.94	8.48	8.33	8.95	9.14	8.49	8.15	8.80	8.61	9.29					
<b>His</b>	8.64	8.17	7.50	7.88	7.92	7.98	8.37	8.57	8.36	8.50	8.58	8.84	8.93	8.83				
<b>Phe</b>	8.95	8.50	8.17	8.24	8.69	8.58	9.11	9.34	8.65	8.51	8.79	8.75	9.55	8.98	9.73			
<b>Arg</b>	9.26	8.24	7.99	8.27	8.31	8.50	8.76	8.98	8.87	9.12		9.52	9.27	9.23	9.26			
<b>Tyr</b>	9.27	8.26	8.02	8.36	8.39	8.58	8.78	9.02	8.96	9.35	9.04	9.48	9.28	9.38	9.56	9.51	9.34	
<b>Trp</b>	9.58	8.95	8.65	8.75	9.22	9.14	9.57	9.79	9.11	9.10	9.21	9.48	10.02	9.66	10.17	9.82	10.08	10.85

Tabela 2.3: Średnie odległości (w Å) dla kontaktów ss z bazy CATH. Podkreślenia dotyczą aminokwasów naładowanych różnoimiennie.

Tak jak w podobnych modelach [108, 145], kontakty typu bb są opisywane w sposób specjalny: ich studnia potencjału ma głębokość  $2\epsilon$  (każdy inny kontakt ma energię  $-\epsilon$ ) i nie mogą tworzyć się między parami  $i, i+4$ . To podejście uwzględnia “przenumerowanie” [138] wiązań wodorowych tworzonych przez łańcuch główny w gruboziarnistej reprezentacji  $\alpha$ -helisy: kontakty  $i, i+4$  rodzaju bb, które zostały zidentyfikowane przez kryterium OV (dla białek z bazy CATH), mają odległości  $C_\alpha - C_\alpha$  bliskie 6 Å, i ponad 98% z nich jest także kontaktem rodzaju bs (patrz podpunkt 2.6.3). Dlatego wystarczy podwoić amplitudę kontaktów bb  $i, i+3$ . Pozwala to też zachować balans między liczebnością kontaktów danego rodzaju (typu ss jest 2 razy więcej niż bb w białkach uporządkowanych [146]).

Równania ruchu dla pseudoatomów zmieniają efektywne położenia atomów  $C_\alpha$ . Podczas ruchu dany kontakt może zniknąć albo się pojawić. Kiedy kontakt jest quasi-adiabatycznie wyłączany po przekroczeniu odległości  $f\sigma_{i,j}$  (gdzie  $f = 1.5$ ),  $\sigma_{i,j}$  zależy od rodzaju kontaktu, ponieważ jest powiązane z odległością  $r_{min}$ . Jeśli para aminokwasów jest połączona kontaktem jednego rodzaju, nie może być między nimi jednocześnie kontaktu innego rodzaju. Nowy rodzaj kontaktu może być utworzony dopiero kiedy poprzedni potencjał zostanie quasi-adiabatycznie wygaszony do zera (tak że nigdy nie działają naraz potencjały o dwóch różnych  $r_{min}$ ). W praktyce takie zmiany zdarzają się rzadko.

Wszystkie pseudoatomy oddziałują także potencjałem czysto odpychającym, aby łańcuch nigdy nie przechodził sam przez siebie. Potencjał ten jest zawsze włączony i odpowiada potencjałowi LJ z głębokością  $\epsilon$  i  $r_{min} = 5$  Å, jednak potencjał ten działa tylko dla  $r < r_{min}$  i jest przesunięty tak, aby był ciągły dla  $r = r_{min}$ . Dla białek uporządkowanych [142] przyjętą wartością takiego odpychającego potencjału są 4 Å, jednak dużo mniej restykcyjny potencjał sztywności łańcucha dla białek nieuporządkowanych wymaga wzmacnienia sztywności poprzez zwiększenie efektywnej objętości aminokwasów. Można ją także powiązać z otoczką hydratacyjną, która często otacza białka nieuporządkowane [3].

### 2.5.3 Kryteria odległościowe

Aby utworzyć kontakt, musi być spełnione parę kryteriów. Pierwsze z nich dotyczy odległości  $r_{on}$ , dla której potencjał LJ między pseudoatomami jest włączany. Sprawdzane były różne warianty  $r_{on}$ , jednak najsukuteczniejsza okazała się wersja, gdzie zawsze  $r_{on} = r_{min}$ . Ma ona dodatkową zaletę: dla  $r = r_{on}$  siła byłaby ciągła, nawet gdyby nie było quasi-adiabatycznego włączania kontaktów. Pozostałe dwa kryteria zostaną przedstawione w podpunktach 2.5.4 i 2.5.5.

Już utworzony kontakt jest wyłączany jeśli odległość  $C_\alpha - C_\alpha$  przekracza  $f \sigma_{i,j}$ , gdzie  $f=1.5$ . Niedawne testy wykazały, że wybór  $f=1.3$  może zwiększyć zgodność z symulacjami  $\alpha$ -synukleiny prowadzonymi na superkomputerze ANTON [147]. Zmniejszenie  $f$  może lepiej odzwierciedlić szybkie zrywanie i odtwarzanie kontaktów w hydrofilowych aminokwasach takich jak glutamina, odzwierciedlając efekt cząsteczek wody, które współzawodniczą z innymi aminokwasami w tworzeniu z nimi wiązań wodorowych. Jednak w pozostałej części rozprawy użyte będzie  $f = 1.5$ . Prowadzi to do problemów przy symulacji wielu łańcuchów, omówionych w podpunkcie 5.2.1. Można ich uniknąć używając  $f = 1.3$ . Tego rodzaju komplikacje były motywacją do opracowania modelu opartego na niewłaściwych kątach dihedralnych, który jest przedstawiony w podrozdziale 2.6.

### 2.5.4 Kryteria kierunkowe

Kontakt bb może powstać jeśli atom azotu z łańcucha głównego  $i$ -tego aminokwasu stworzy wiązanie wodorowe z atomem tlenu łańcucha głównego  $j$ -tego aminokwasu (lub odwrotnie). Takie wiązanie może być utworzone tylko jeśli atomy są ku sobie skierowane. Aby odtworzyć to wymaganie używając wyłącznie położen atomów  $C_\alpha$ , użyte zostały (znane już wcześniej [146, 148, 149]) wektory pomocnicze  $\mathbf{h}_i$ , które są zaczepione tam gdzie atom  $C_\alpha$  aminokwasu  $i$ . Wektory te są prostopadłe do płaszczyzny wyznaczonej przez aminokwasy  $i - 1$ ,  $i$  oraz  $i + 1$ , tzn. są równoległe do wektora  $\mathbf{v}_i \times \mathbf{v}_{i+1}$ , gdzie  $\mathbf{v}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$  oraz  $\mathbf{r}_{i,j} = \mathbf{r}_j - \mathbf{r}_i$ . Przykłady tych wektorów są podane na dolnym panelu Rys. 2.11. W przypadku wiązania wodorowego wektory te powinny być prawie równoległe (lub antyrównoległe w przypadku antyrównoległych  $\beta$ -kartek). To kryterium kierunkowe oddaje trzy warunki, które są konieczne do utworzenia kontaktu bb [149]:

- $|\cos(\mathbf{h}_i, \mathbf{r}_{i,j})| > 0.92$  (kąt graniczny  $23^\circ$ )
- $|\cos(\mathbf{h}_j, \mathbf{r}_{i,j})| > 0.92$
- $|\cos(\mathbf{h}_i, \mathbf{h}_j)| > 0.75$  (kąt graniczny  $41^\circ$ )

Wartości kątów granicznych są dobrane na podstawie statystycznych rozkładów tych kątów [149].

Kryteria kierunkowe związane z kontaktami ss mogą być zdefiniowane poprzez wprowadzenie wektora normalnego [145]:

$$\mathbf{n}_i = \frac{\mathbf{r}_{i-1} + \mathbf{r}_{i+1} - 2\mathbf{r}_i}{|\mathbf{r}_{i-1} + \mathbf{r}_{i+1} - 2\mathbf{r}_i|} . \quad (2.5)$$

Wektor przeciwny do niego ( $-\mathbf{n}$ ) w przybliżeniu wskazuje kierunek od atomu węgla  $C_\alpha$  do  $C_\beta$  (jak na górnym panelu Rys. 2.11).

Aby utworzył się kontakt ss, łańcuchy boczne muszą być skierowane ku sobie. Aby tak było, muszą być spełnione dwa warunki [145]:

- $\cos(\mathbf{n}_i, \mathbf{r}_{i,j}) < 0.5$
- $\cos(\mathbf{n}_j, \mathbf{r}_{j,i}) < 0.5$  (kąt graniczny  $60^\circ$ ).

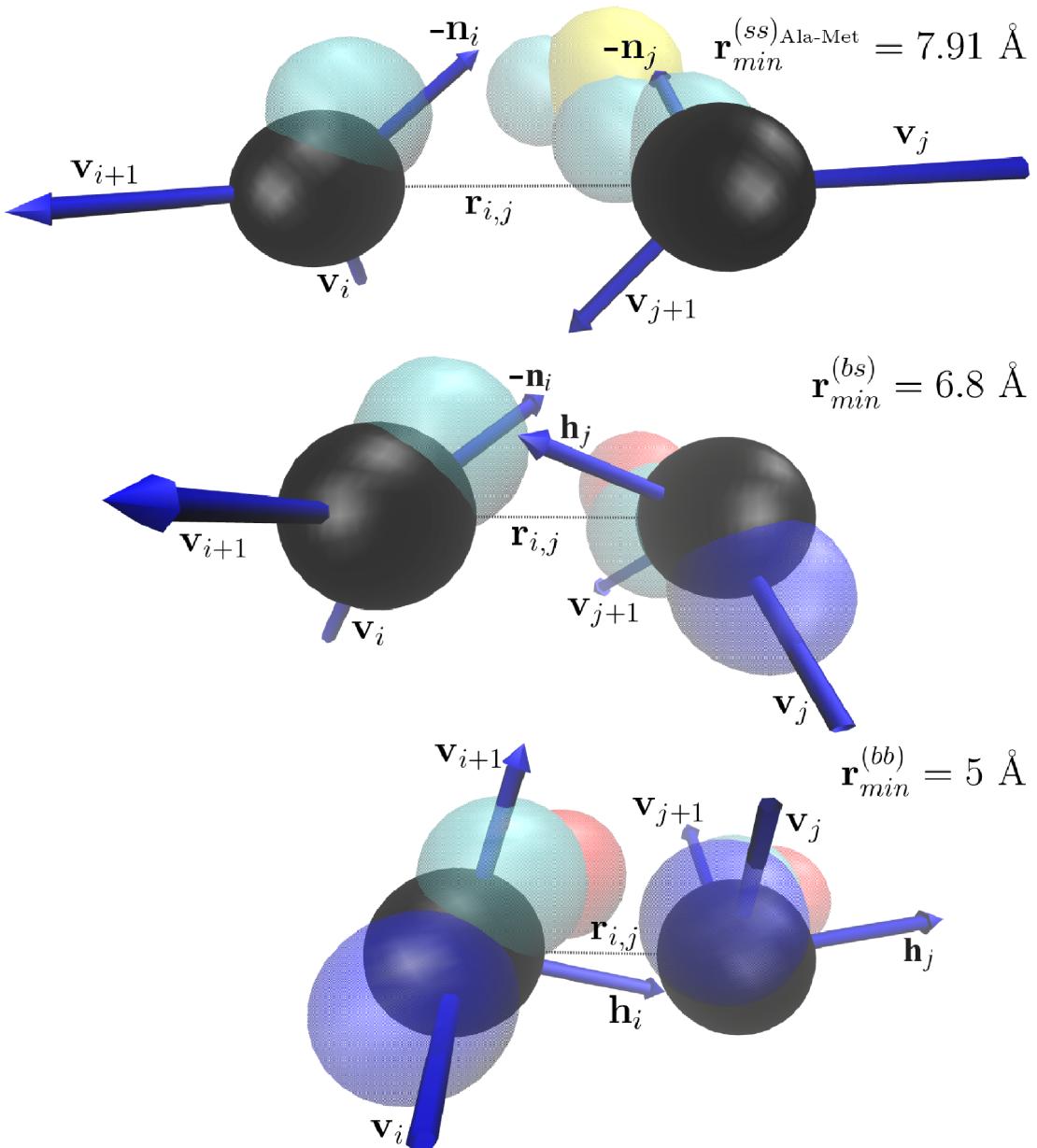
Warunki te są jakościowo różne od tych dla kontaktów bb, ponieważ natura kontaktów ss jest dużo bardziej różnorodna (a więc odpowiednie kryteria są mniej restrykcyjne), w przeciwieństwie do ściśle określonego wzorca wiązań wodorowych łańcucha głównego [149].

Powyzsze kryteria kierunkowe dla kontaktów ss są spełnione dla 97% kontaktów wyznaczonych z bazy CATH na podstawie kryterium przekrywania. Zastosowanie w tych kryteriach dokładniejszego (ale bardziej skomplikowanego [150, 151]) wyrażenia na położenie atomu  $C_\beta$  powoduje, że kryteria spełnia tylko 82% tych kontaktów, dlatego użyte zostało prostsze wyrażenie (na  $-\mathbf{n}$ ).

Podobne rozważania prowadzą do następujących dwóch warunków na utworzenie kontaktu bs (łańcuch boczny  $i$ -tego aminokwasu powinien być w kontakcie z łańcuchem głównym  $j$ -tego aminokwasu, patrz środkowy panel Rys. 2.11):

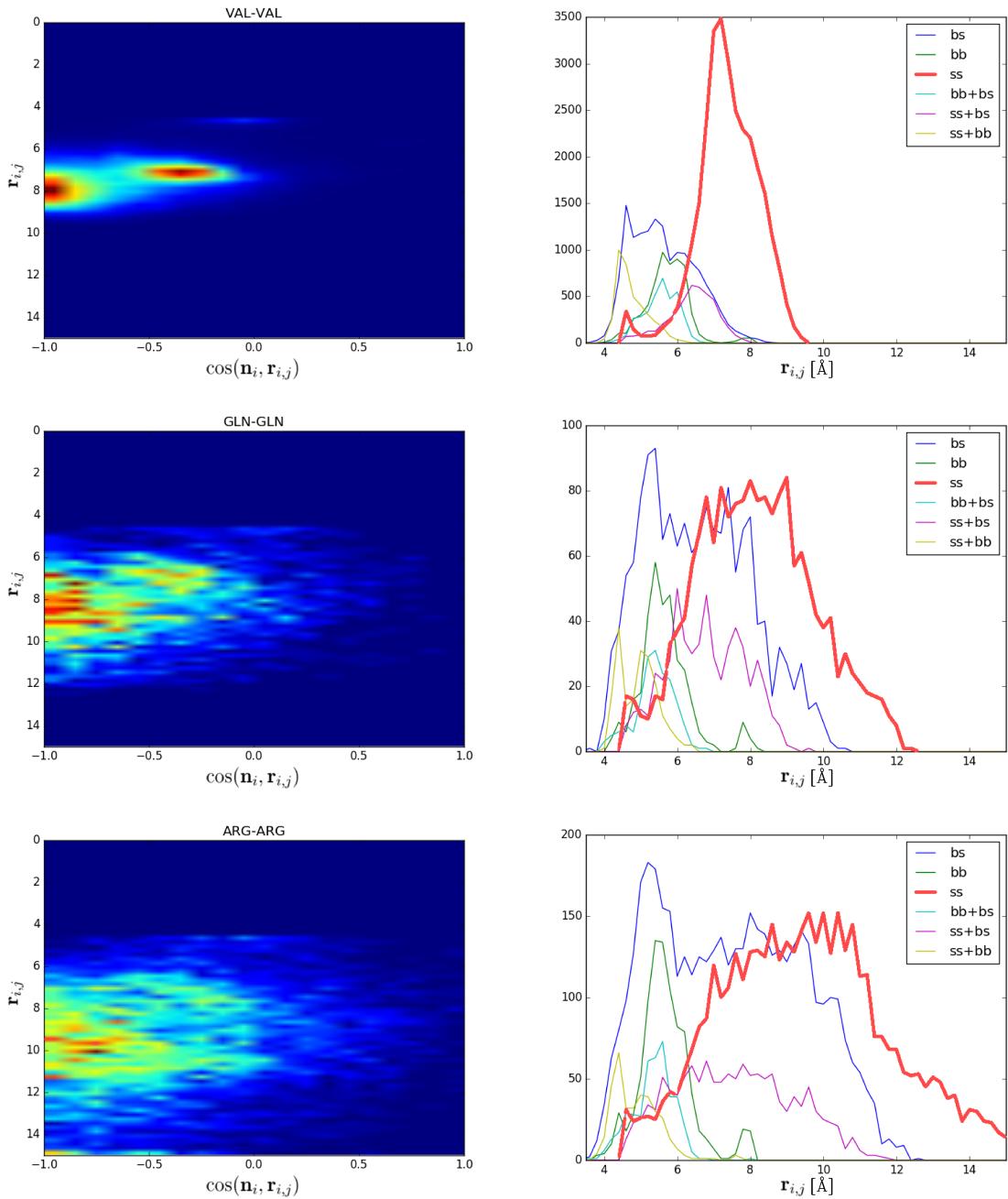
- $\cos(\mathbf{n}_i, \mathbf{r}_{i,j}) < 0.5$
- $|\cos(\mathbf{h}_j, \mathbf{r}_{j,i})| > 0.92$

Jak pokazuje Rys. 2.6, większość kontaktów znalezionych tylko na podstawie kryterium przekrywania dla białek uporządkowanych spełnia kryteria kierunkowe podane wyżej. Wyjątki dotyczą głównie kontaktów bs, których natura może nie być do końca uchwycona przez zmieszanie kryteriów kierunkowych łańcucha głównego i bocznego. Co ciekawe, liczba kontaktów między łańcuchem głównym i hydrofobowymi łańcuchami bocznymi jest porównywalna do tej z polarnymi. Dlatego trudno mówić o tym, co powoduje w tym przypadku przyciąganie, więc  $r_{min}$  dla kontaktów bs należy traktować z większą rezerwą niż dla pozostałych przypadków. Kiedy kryteria kierunkowe zostaną zastosowane do kontaktów bs, średnia odległość w rozkładzie zmniejsza się z 6.8 Å do 5.4 Å.



Rysunek 2.11: Przykłady kontaktów między alaniną (po lewej) a metioniną (po prawej). Odległości między atomami  $C_\alpha$  (czarne sfery) wynoszą (idąc od górnego panelu) 7.91, 6.03 and 5.4 Å. Panele przedstawiają (idąc od góry) kontakty typu ss, bs oraz bb. Dwa górne panele pochodzą z pełnoatomowej symulacji wykonanej w NAMD, dolny panel to fragment  $\alpha$ -helisy ze struktury o kodzie PDB 14GS. Wszystkie ciężkie atomy (poza czarnymi  $C_\alpha$ ) są pokolorowane wg schematu CPK. Nie pokazano atomów wodoru. Atomy łańcuchów bocznych są pokazane jako półprzezroczyste sfery. Strzałki oznaczają wektory  $\mathbf{h}$  (dla oddziaływań łańcucha głównego) bądź  $\mathbf{n}$  (dla oddziaływań łańcuchów bocznych). Wektor  $r_{i,j}$  łączący atomy  $C_\alpha$  jest zaznaczony kropkowaną linią. Po prawej zapisane są graniczne wartości, poniżej których dany kontakt jest włączany. Wektory  $\mathbf{v}_i$  są zdefiniowane jako  $\mathbf{r}_i - \mathbf{r}_{i-1}$ . Oparte na rys. 2 z artykułu [I].

Dozwolone wartości  $\cos(\mathbf{n}_i, \mathbf{r}_{i,j})$  wydają się być niezależne od odległości  $r_{i,j}$ . Przykłady dwuwymiarowych rozkładów (gdzie na jednej osi jest  $\cos(\mathbf{n}_i, \mathbf{r}_{i,j})$ , a na drugiej  $r_{i,j}$ ) są pokazane na Rys. 2.12. Widać na nim, że korelacja między odlegością  $C_\alpha$ - $C_\alpha$  a cosinusem jest bardzo niewielka. Jest tak dla przypadku łańcuchów bocznych krótkich (Val-Val), średnich (Gln-Gln) i długich (Arg-Arg).



Rysunek 2.12: Lewe panele przedstawiają dwuwymiarowe rozkłady, gdzie pokazana jest liczba kontaktów o danej odległości  $C_\alpha$ - $C_\alpha$  oraz danym  $\cos(\mathbf{n}_i, \mathbf{r}_{i,j})$  (niebieski oznacza mniej kontaktów, czerwony – więcej) dla kontaktów ss. Prawe panele to zwykłe rozkłady odległości  $C_\alpha$ - $C_\alpha$ . Oparte na rys. S4 z artykułu [I].

Warto zauważyć, że po utworzeniu kontaktu żaden potencjał nie utrzymuje kryteriów kierunkowych, ponieważ potencjał LJ jest sferycznie symetryczny i zależy tylko od odległości między pseudoatomami. Kryteria kierunkowe są sprawdzane tylko w momencie tworzenia kontaktu. Inna sytuacja zachodzi w omawianym później drugim modelu, w którym hamiltonian nie zależy od czasu.

### 2.5.5 Kryteria związane z liczbą koordynacyjną

Aminokwasy zostały podzielone na 6 klas:

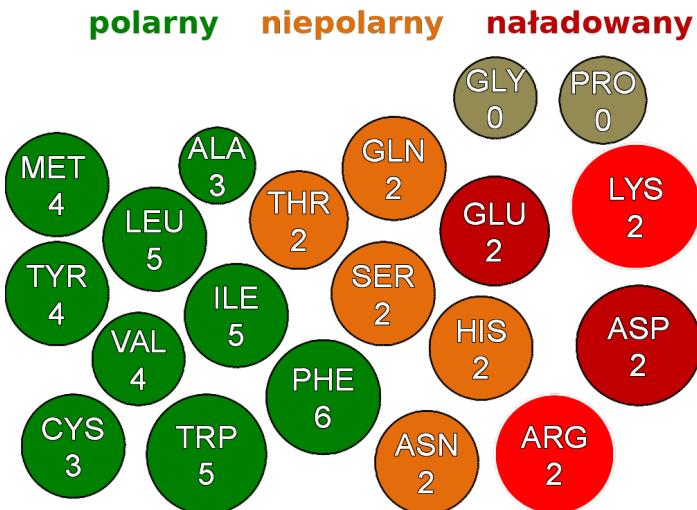
1. Gly
2. Pro
3. hydrofobowe: Ala, Cys, Val, Ile, Leu, Met, Phe, Tyr, Trp
4. polarne: Gln, Ser, Thr, Asn, His
5. ujemnie naładowane: Asp, Glu
6. dodatnio naładowane: Arg, Lys

Podział ten jest zastosowany w tabeli 2.4. Ostatnie dwie klasy (naładowane) liczą się jako polarne. Podział na aminokwasy hydrofobowe i polarne odpowiada (z wyjątkiem alaniny) podziałowi opartemu na wartościach własnych macierzy Miyazawa i Jernigana [152]. Kontakty typu ss mogą być utworzone między dwoma hydrofobowymi aminokwasami [145], ale także między dwoma polarnymi (co jest kluczowe dla modelowania polyQ) oraz między polarnym i hydrofobowym. Różnoimiennie naładowane aminokwasy mogą utworzyć kontakt ss (rozumiany jako mostek solny), w przeciwieństwie do jednoimiennie naładowanych. Głębokość potencjału LJ jest taka sama dla każdego rodzaju kontaktu ss. Jest to duża różnica w stosunku do zwykłych modeli typu "HP" [145], gdzie kontakty polarnych aminokwasów są słabsze lub w ogóle się nie tworzą. Dlatego, mimo że opisywany model nie może uchwycić detali wszystkich możliwych oddziaływań (np. między elektronami  $\pi$  pierścieni aromatycznych), są one uwzględnione w sposób statystyczny.

Poza omówionymi później mostkami dwusiarczkowymi [153, 154] cysteiny mogą tworzyć między sobą także zwyczajne kontakty typu ss.

Każdy aminokwas może utworzyć najwyżej  $z_s$  kontaktów. Limit  $z_s$  zależy od tego jaki to aminokwas i wynosi  $z_s = n_b + \min(s, n_H + n_P)$ , gdzie  $n_b$  to dozwolona liczba kontaktów dla łańcucha głównego,  $s$  dla łańcucha bocznego, a  $n_H$  i  $n_P$  to maksymalne liczby kontaktów jakie łańcuch boczny może utworzyć odpowiednio z hydrofobowymi i polarnymi łańcuchami bocznymi innych aminokwasów. Wartości tych

parametrów są przedstawione w tabeli 2.4 i schematycznie na rysunku 2.13. Kontakt  $b$  zmniejsza limit dla łańcucha głównego jednego aminokwasu i limit dla łańcucha bocznego drugiego aminokwasu. Nie zmniejsza jednak limitu oddziaływań polarnych  $n_P$  ani hydrofobowych  $n_H$ , dlatego wpływa tylko na limit  $s$  danego łańcucha bocznego. Traktowanie łańcucha głównego jako tworzącego kontakty polarne zmieniłoby tylko proporcje między typami kontaktów, ale nie wpłynęłoby znacząco na wyniki.



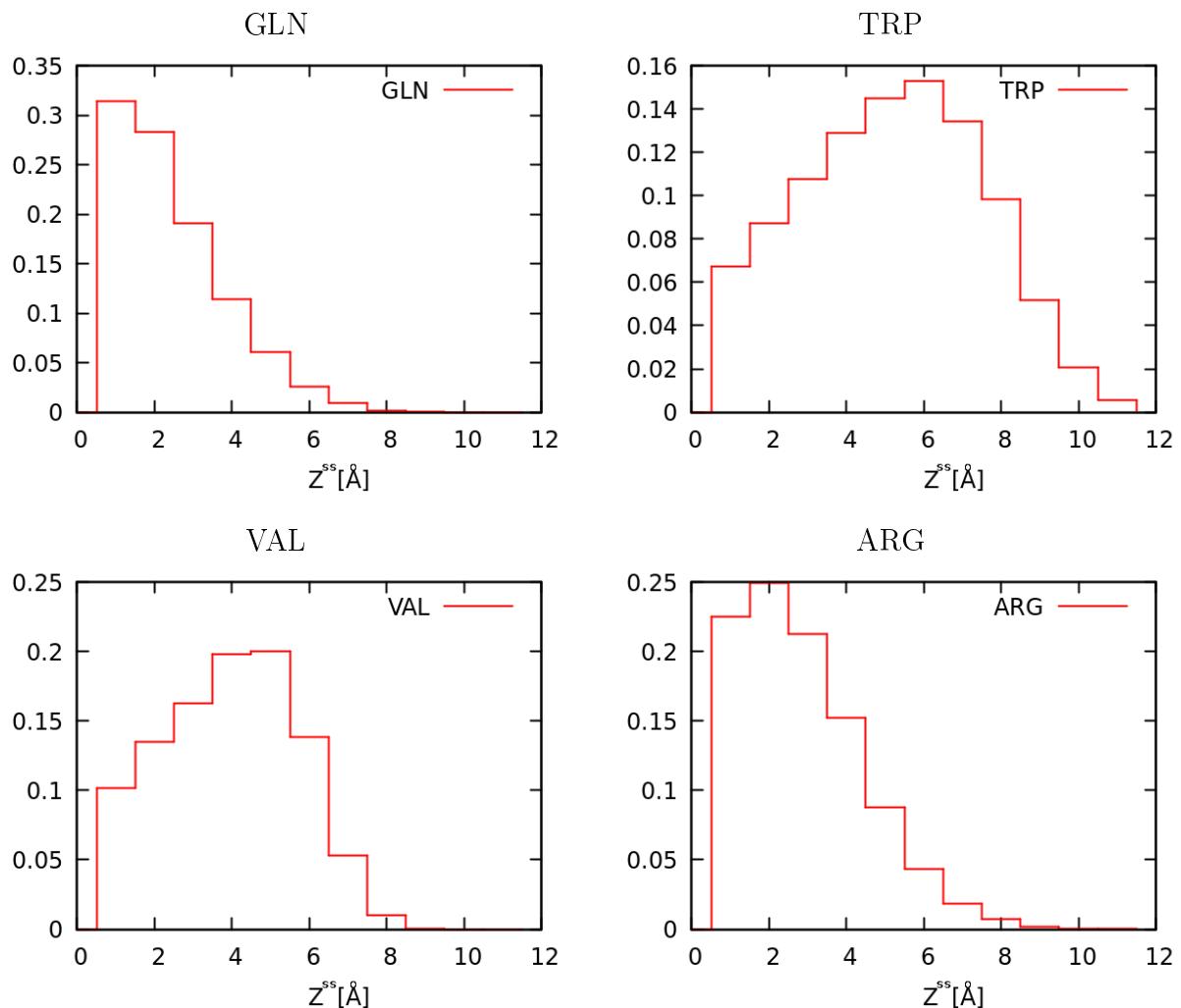
Rysunek 2.13: Schematyczne przedstawienie rodzajów aminokwasów. Liczby oznaczają limit kontaktów łańcucha bocznego  $s$  dla danego aminokwasu.

Limit dla łańcucha bocznego  $n_b$  jest zawsze równy 2 (z wyjątkiem proliny, dla której  $n_b = 1$ ), co odpowiada dwóm możliwym wiązaniom wodorowym (jedno dla atomu tlenu, drugie dla azotu).

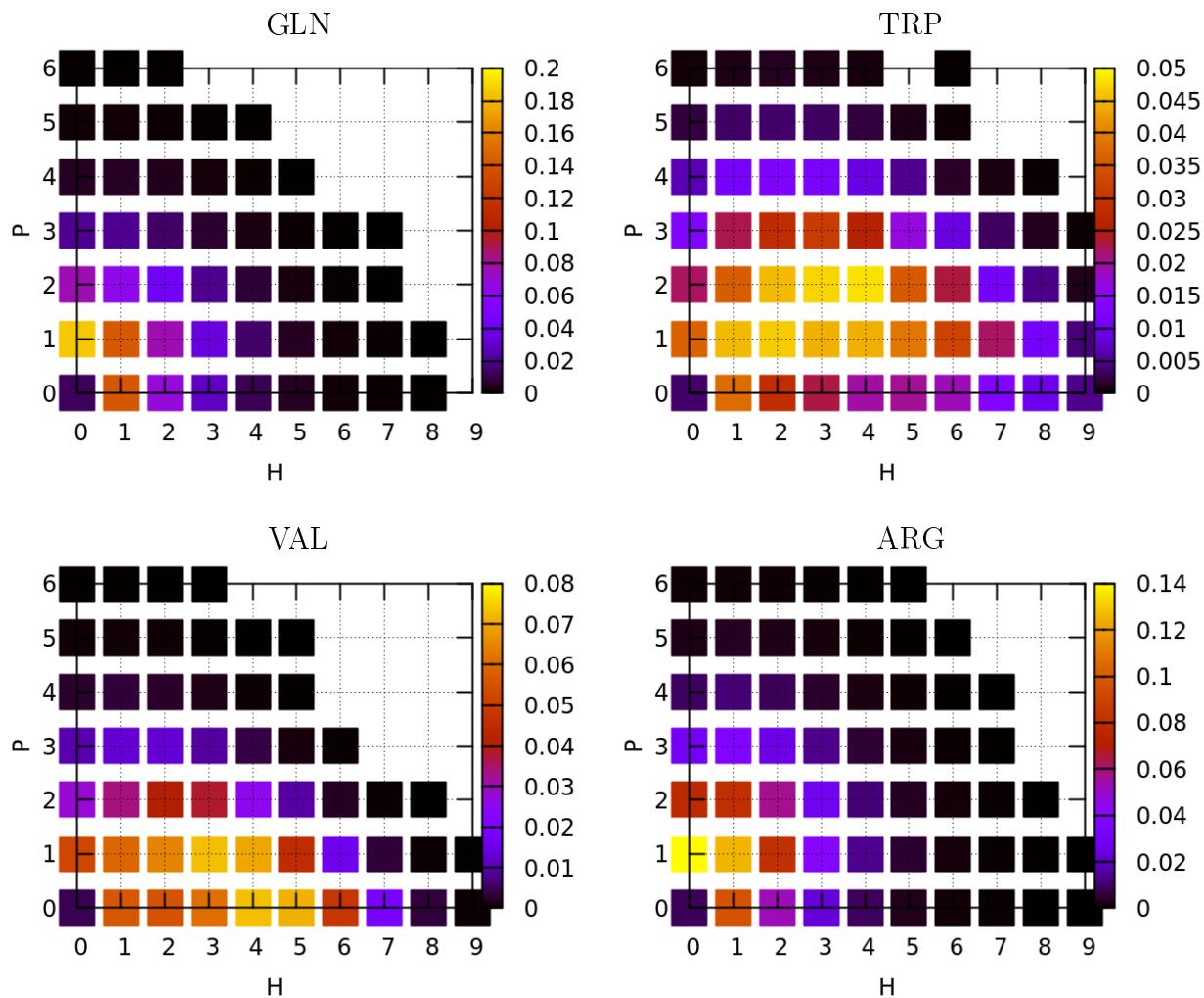
Liczby  $n_H$ ,  $n_P$  oraz  $s$  zależą od aminokwasu i zostały ustalone na podstawie bazy kontaktów przez obliczenie dla każdego aminokwasu liczby kontaktów ss ( $Z^{ss}$ ), czyli liczby koordynacyjnej jego łańcucha bocznego.  $s$  jest powiązane z  $Z_{max}^{ss}$  - wartościami  $Z^{ss}$  odpowiadającymi maksimum rozkładu. Rozkłady  $Z^{ss}$  dla 4 wybranych aminokwasów są przedstawione na Rys. 2.14.

Dla aminokwasów hydrofobowych rozkłady są szerokie i  $s = Z_{max}^{ss}$ . Dla polarnych, rozkłady są wąskie i  $s = Z_{max}^{ss} + 1$ . Zwiększenie o 1 jest korektą związaną z tym, że w białkach uporządkowanych aminokwasy polarne znajdują się zwykle blisko powierzchni białka i część ich możliwych kontaktów jest „zajęta” przez cząsteczki wody, stąd niedoszacowanie  $Z_{max}^{ss}$ . Z tego samego powodu nie wybieramy  $s$  dla aminokwasów polarnych jako możliwej do utworzenia liczby wiązań wodorowych. Wartości  $n_H$ ,  $n_P$  zostały ustalone podobną metodą, przez obliczenie dwuwymiarowych rozkładów, gdzie na jednej osi jest liczba kontaktów hydrofobowych, a na drugiej hydrofilowych (z maksimami  $H_{max}$ ,  $P_{max}$ ).

Rozkłady te są przedstawione na Rys. 2.15. Wtedy,  $n_H$  odpowiada maksimum  $H_{max}$  takiego rozkładu (najczęstsza liczba hydrofobowych kontaktów ss dla danego aminokwasu), natomiast  $n_P$  odpowiada  $P_{max} + 1$ . Wartości  $s$ ,  $n_H$ ,  $n_P$  znajdują się w tabeli 2.4.



Rysunek 2.14: Rozkłady liczby kontaktów ss ( $Z^{ss}$ ) dla glutaminy, tryptofanu, waliny i arginin, unormowane tak, aby suma słupków była równa 1.



Rysunek 2.15: Dwuwymiarowe rozkłady liczby hydrofobowych (H) i polarnych (P) kontaktów ss dla glutaminy, tryptofanu, waliny i argininy, unormowane tak, aby suma kwadratów była równa 1.

nazwa	Gly	Pro	Gln	Cys	Ala	Ser	Val	Thr	Ile	Leu
klasa	-	-	P	P	H	P	H	P	H	H
s	0	0	2	3	3	2	4	2	5	5
$n_H$	0	0	0	2	1	0	4	0	4	4
$n_P$	0	0	2	2	1	2	1	2	2	2
nazwa	Asn	Asp	Lys	Glu	Met	His	Phe	Arg	Tyr	Trp
klasa	P	P-	P+	P-	H	P	H	P+	H	H
s	2	2	2	2	4	2	6	2	4	5
$n_H$	0	0	0	0	1	0	4	0	2	4
$n_P$	2	2	2	2	1	2	2	2	2	3

Tabela 2.4: Limity kontaktów jakie mogą tworzyć aminokwasy. Klasa może być żadna (Gly i Pro), hydrofobowa (H) lub polarna (P). Końcówka + lub – mówi o tym czy aminokwas jest naładowany. Liczba  $s$  to maksymalna liczba kontaktów jakie może utworzyć łańcuch boczny. Maksymalna liczba kontaktów z hydrofobowymi łańcuchami bocznymi to  $n_H$ , a z polarnymi  $n_P$ . Histydyna jest traktowana jako nienalałdowana.

## 2.5.6 Elektrostatyka

Oddziaływanie elektrostatyczne mogą mieć znaczący wpływ na dynamikę białek nieuporządkowanych, zwłaszcza jeśli mają duży średni ładunek przypadający na jeden aminokwas [155]. Nie jest to przypadek poliglutaminy ani glutenu, jednak ponieważ model ma działać dla jak największej liczby białek nieuporządkowanych, uwzględnia on (oprócz kontaktów bb, bs i ss) oddziaływanie elektrostatyczne. Naładowane aminokwasy oddziałują ze sobą zmodyfikowanym potencjałem Debye'a-Hueckel'a (DH) [156]:

$$V_{D-H} = \frac{e^2 \exp(-r/\lambda)}{4\pi\epsilon\epsilon_0 r} \quad (2.6)$$

gdzie długość ekranowania  $\lambda = 10 \text{ \AA}$ . Tak jak w modelu grupy Tozzini i in. [112], względna przenikalność elektryczna  $\epsilon$  zależy od odległości między aminokwasami:  $\epsilon = 4 \text{ \AA}^{-1} r$  (w rzeczywistości  $\epsilon$  powinna zmieniać się wraz z odległością sigmoidalnie, tak aby we wnętrzu białka wynosić około 4, a dla dużych odległości dążyć do 80;;  $\epsilon_0$  to przenikalność elektryczna próżni). Przyjęcie takiego wzoru na  $\epsilon$  prowadzi do następującego potencjału oddziaływań elektrostatycznych:

$$V_{el}(r) = \frac{85 \exp(-r/\lambda) \epsilon \text{ \AA}^2}{r^2} \quad (2.7)$$

Oddziaływanie te są długozasięgowe i nie wliczają się do limitu  $s$  ani  $n_P$ . Kiedy dwa przeciwnie naładowane aminokwasy utworzą kontakt ss, jest on opisywany zwykłym potencjałem LJ z minimum  $r_{min}$  pochodzącym z tabeli 2.3, analogicznie do tego jak mostki solne są opisane w modelach gruboziarnistych opartych na strukturze natywnej [53]. Kiedy kontakt ss reprezentujący mostek solny jest quasi-adiabatyczniełączony, pozostałe oddziaływanie elektrostatyczne tworzone przez dwa aminokwasy wchodzące w skład mostka są w tym samym czasie quasi-adiabatycznie wyłączane, ponieważ ładunek w mostku solnym jest ekranowany i nie powinien być „widoczny” dla pozostałych aminokwasów. Gdyby nie uwzględniać tego ekranowania i pozostawić oddziaływanie elektrostatyczne włączone, nie wpłynęłoby to znacząco na wyniki.

Jednoimiennie naładowane aminokwasy nie mogą tworzyć kontaktów ss, ale mogą tworzyć ze sobą kontakty typu bb lub bs. Ładunek jest zwykle zlokalizowany na końcu długiego łańcucha bocznego, w związku z tym oddziaływanie z łańcuchem głównym nie powinno być zaburzone.

Możliwe jest także całkowite zrezygnowanie z opisywania oddziaływań różnoimiennie naładowanych aminokwasów przez potencjał DH i zamiast tego traktować je jako zwykłe kontakty typu ss (jeśli spełnione są warunki na ich utworzenie). Taka zmiana polepsza zgodność z wynikami eksperymentalnymi dla białek uporządkowanych i jest szerzej opisana w podrozdziale 2.6 dotyczącym modelu z hamiltonianem niezależnym od czasu. Nie ma jednak dużego wpływu na wyniki dla poliglutaminy i glutenu, więc większość wyników została uzyskana przy użyciu modelu gdzie potencjał DH opisuje zarówno przyciągające jak i odpychające oddziaływanie elektrostatyczne.

Pierwszy i ostatni aminokwas w łańcuchu nie mogą tworzyć kontaktów bb, bs ani ss, których kryteria geometryczne wymagają położen sąsiednich pseudoatomów w łańcuchu. W podobnym modelu używającym metody Monte Carlo [149] takie kontakty mogą być tworzone, ale ich amplituda jest zmniejszona. W modelach pełnoatomowych końce łańcucha są naładowane, ale w modelach gruboziarnistych jest to często zaniedbywane [53]. Ten model dopuszcza obie możliwości **[VI]**, jednak w tej rozprawie zastosowano wersję gdzie końce nie są naładowane.

### 2.5.7 Mostki dwusiarczkowe

Aby modelować mostki dwusiarczkowe można używać różnych potencjałów. Jednym z nich jest potencjał harmoniczny, który może być włączany i wyłączany, co odpowiada tworzeniu i zrywaniu mostka [157]. Ta metoda jest zaimplementowana w modelu **[VI]**, ale okazało się, że równie dobrze można użyć potencjału LJ z głębokością  $4\epsilon$  i minimum  $r_{min}^{SS} = 5.9 \text{ \AA}$ . Mgr Mariusz Raczkowski sprawdził, że taki potencjał poprawnie odtwarza zwijanie i termiczną denaturację krambiny i ubikwityny.

Niezależnie od tego jaki potencjał je opisuje, mostki dwusiarczkowe są quasi-adiabatycznie włączane i wyłączane tak jak pozostałe kontakty typu ss (z tą samą skalą czasową  $10\tau$  i tymi samymi kryteriami). Prawdziwe skale czasowe tworzenia i zrywania mostków dwusiarczkowych są dużo dłuższe [158]. Jedynym dodatkowym kryterium jest ograniczenie zapewniające, że jedna cysteina może tworzyć tylko jeden mostek. W związku z tym dwie cysteiny tworzące mostek nie będą już mogły utworzyć mostka z innymi cysteinami (dopóki mostek między nimi nie zostanie zerwany).

### 2.5.8 Zgodność z doświadczeniem i symulacjami pełnoatomowymi

Pomimo wielu przybliżeń zastosowanych w modelu, jest on zgodny z wieloma poprzednimi wynikami symulacji i doświadczeń dotyczących średnich wielkości opisujących geometrię łańcucha.

Głównym celem przedstawionych poniżej symulacji jest wykazanie ich zgodności z doświadczeniem. Używamy głównie trzech parametrów opisujących geometrię łańcucha:

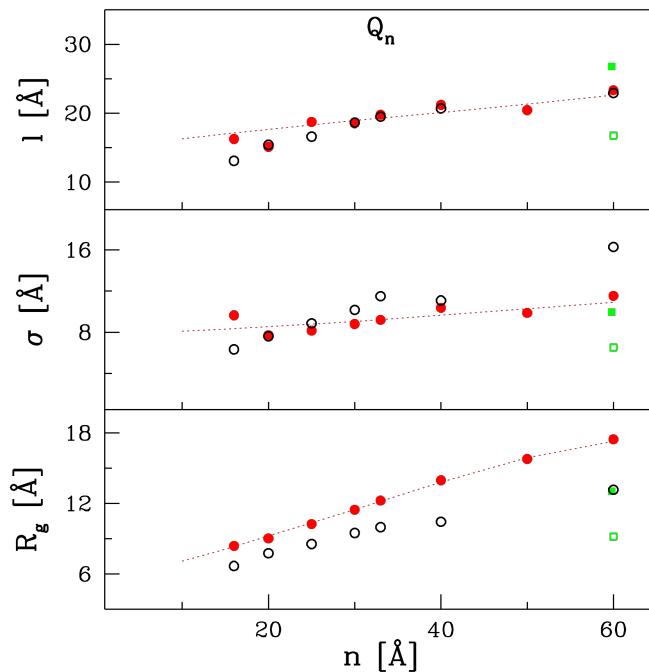
- średni promień bezwładności  $R_g = \sqrt{\langle r_g^2 \rangle}$ , gdzie  $r_g$  to promień chwilowy, liczony dla jednej konformacji
- średnia odległość między końcami łańcucha  $l = \langle d_{ee} \rangle$ , gdzie  $d_{ee}$  to odległość chwilowa
- dyspersja  $d_{ee}$ , czyli  $\sigma = \sqrt{l^2 - \langle d_{ee} \rangle^2}$

Gruboziarnista natura modelu nie pozwala na przewidywania dotyczące parametrów pełnoatomowych takich jak wykres Ramachandrana czy sprzężenia NMR, w związku z tym nie możemy porównywać ich z doświadczeniem.

W przypadku porównań do symulacji pełnoatomowych też używamy tylko średnich parametrów geometrycznych bądź ich rozkładów. Z wyników doświadczalnych wybraliśmy te, które używając do badania białek nieuporządkowanych metod NMR, FRET i SAXS przyjmują jak najmniej założeń teoretycznych (aby porównanie z doświadczeniem było jak najbardziej bezpośrednie). Porównania pokazują rysunki 2.16, 2.19 oraz 2.20. Pierwsze dwa dotyczą porównań z symulacjami pełnoatomowymi, ostatnie - z wynikami doświadczalnymi.

Średnie parametry geometryczne zostały obliczone na podstawie 100 niezależnych symulacji, każda trwająca 100 000  $\tau$  (z czego pierwsze 50 000  $\tau$  nie było uwzględniane, aby dać układowi czas na dojście do równowagi). Konformacje były zapisywane co 100  $\tau$ . Ze względu na brak struktury natywnej, początkową konformację w symulacji był łańcuch utworzony przez samoomijające błędzenie przypadkowe (ang. *self-avoiding random walk*), z długością kroku 3.8 Å.

#### 2.5.8.1 Poliglutamina i poliwalina



Rysunek 2.16: Wyniki otrzymane przy użyciu modelu gruboziarnistego dla poliglutaminy  $Q_n$  (pełne czerwone kółka) oraz poliwaliny  $V_n$  (pełne zielone kwadraty) w funkcji liczby aminokwasów  $n$ . Górnego panelu pokazuje średnią odległość między końcami  $l$ , środkowy jej dyspersję  $\sigma$ , a dolny średni promień bezwładności  $R_g$ . Puste kółka odpowiadają wynikom symulacji pełnoatomowych [34] dla  $Q_n$ . Pusty zielony kwadrat (tylko dla  $n=60$ ) odpowiada wynikom symulacji pełnoatomowych [20] dla  $V_{60}$ . Proste pokazują tylko trend. Rozmiary symboli są rzędu błędu średniej. Oparte na rys. 4 z artykułu [I].

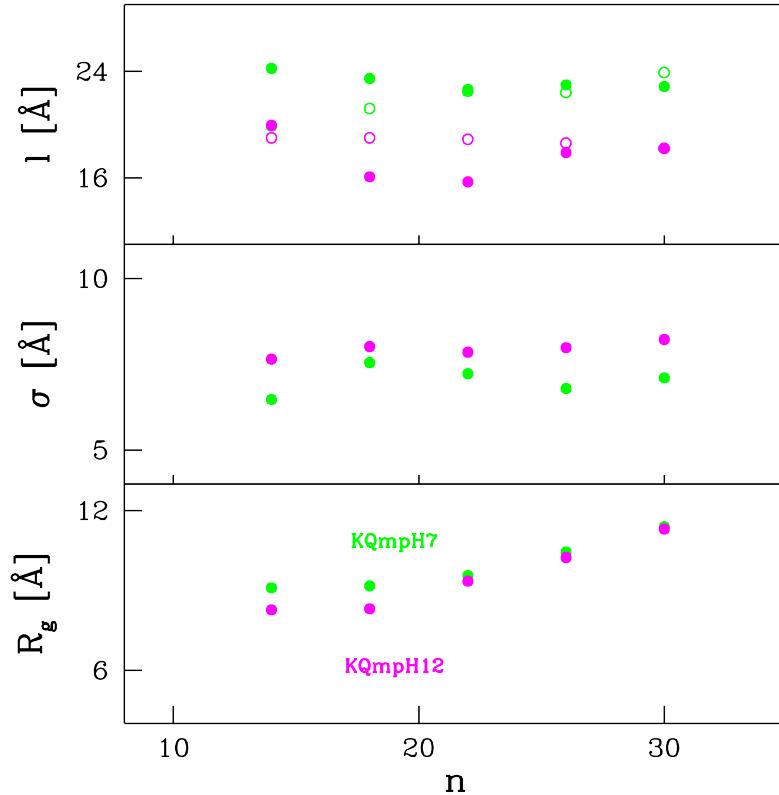
Rys. 2.16 pokazuje porównanie z wynikami dla  $Q_n$ ,  $V_n$ , gdzie  $n \geq 60$ . Grupa Cossio i in. [20] wygenerowała 30 063 statystycznie niezależnych konformacji dla  $V_{60}$  używając meta-dynamiki obejmującej wymianę replik i metodę ukrytego rozpuszczalnika (był on reprezentowany przez metodę GBSA (ang. *Generalized Born Surface Area*). Tylko niewielki ułamek tych konformacji daje się zidentyfikować jako konformacje obecne w bazie CATH. Białka uporządkowane wykorzystują zatem tylko niewielką część możliwych konformacji. Do stworzenia wykresu 2.16 wykorzystaliśmy 7 077 reprezentatywnych konformacji  $V_{60}$  udostępnionych przez autorów. Model gruboziarnisty daje parametry nawet dwa razy większe dla  $n=60$ , ale dla mniejszych  $n$  różnice powinny być mniejsze.

Warto zauważyć, że długie fragmenty  $V_n$  nie istnieją w naturze (patrz tabela 3.1), podczas gdy długie fragmenty  $Q_n$  są dobrze znane (choćby w chorobie Huntingtona). Gómez-Sicilia [34] zastosował podejście Cossio i in. [20] dla  $Q_n$ , gdzie  $n$  jest w zakresie od 16 do 80. Udostępnione zostało 308, 491, 330, 422, 479, 322, 269, 246, i 108 konformacji, odpowiednio dla  $n=16, 20, 25, 30, 33, 38, 40, 60$ , i 80. Ze względu na małą statystykę nie rozważamy  $n=80$ .

Rys. 2.16 pokazuje, że parametry geometryczne uzyskane przez Gómez-Sicilia i in. [34] zgadzają się z modelem gruboziarnistym, zwłaszcza w przypadku  $l$ . Zgodność ta wystarczy do modelowania glutenu (zestawu białek bogatych w glutaminę [159]). Zgodność z danymi pełnoatomowymi można zwiększyć, zwiększając wartość limitu kontaktów  $s$  z 2 na 3. Zwiększało to także liczbę zawężonych konformacji. Zachowaliśmy jednak  $s = 2$ , aby nie tworzyć wyjątków w metodzie budowania modelu. Analogicznie można by zwiększyć zgodność zwiększając limit dla kontaktów łańcucha głównego,  $n_b$ , z 2 do 3, jednak taka zmiana nie byłaby fizycznie uzasadniona.

Jednym z osiągnięć Gómez-Sicilia i in. [34] było odkrycie, że duża część (9.3%) statystycznie niezależnych konformacji  $Q_{60}$  zawiera węzeł, a długość węzła jest większa lub równa 36 aminokwasom. Takie zawężone konformacje mogą zatykać proteasom i prowadzić do toksyczności [79]. W doświadczeniu [160] toksyczność pojawia się powyżej długości 35 aminokwasów. W modelu gruboziarnistym także pojawiają się zawężone konformacje dla  $Q_{60}$  (węzły są jednak płytkie), natomiast brak ich dla  $V_{60}$ . W modelach pełnoatomowych  $V_{60}$  tworzyło węzły około 3 razy rzadziej niż  $Q_{60}$ , więc także pod względem tworzenia węzłów wyniki są zgodne. Przykład zawężonej konformacji pokazuje prawy panel Rys. 2.18. Najwięcej węzłów tworzyły łańcuchy polyW.

Doświadczalne wyniki FRET dla poliglutaminy o 8, 12, 16, 20 oraz 24 aminokwasach zostały uzyskane przez Walters'a i Murphy'ego [141] dla 2 wartości pH: 7 i 12. Badane przez nich łańcuchy były



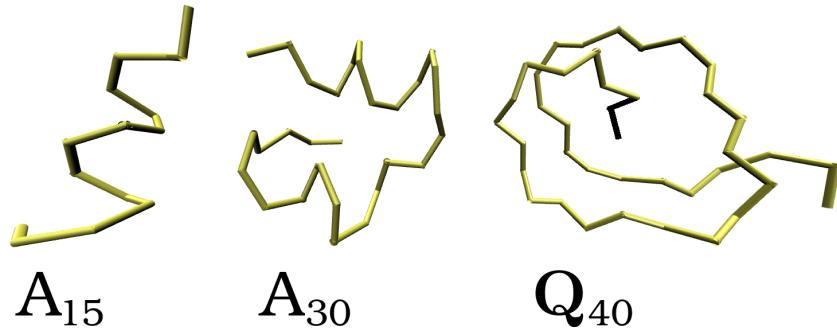
Rysunek 2.17: Wyniki otrzymane modelem grubobiarnistym dla  $Q_m$  zakończonych tryptofanem i lizynami (pełne zielone kółka) bądź asparaginą (pełne purpurowe kółka) w funkcji liczby aminokwasów  $n$ . Górnego panelu pokazuje  $l$ , środkowy  $\sigma$ , dolny  $R_g$ . Puste kółka oznaczają wyniki eksperymentalne [141]. Oparte na rys. S10 z artykułu [I].

jednak zakończone sekwencją KKW od N-końca oraz sekwencją AKK od C-końca, tak że liczba aminokwasów w łańcuchu  $n = m+6$ . Dlatego wyniki te nie są porównywane do symulacji  $Q_n$ , lecz do symulacji zawierających dodatkowe aminokwasy (przez co uwzględnione zostają oddziaływanie elektrostatyczne lizyn). Rys. 2.17 i 2.20 pokazują porównanie wyników modelu do wyników doświadczenia dla pH=7. Oznaczenie układu to KQmpH7, dane zostały pokolorowane na zielono. Jest duża zgodność wyników dla  $l$ , ale duża różnica dla  $\sigma$ , której źródła nie można zweryfikować, ponieważ nie wiadomo jak liczona była niepewność w metodzie doświadczalnej [141]. Wydaje się jednak, że  $\sigma$  z doświadczenia jest zbyt mała, zwłaszcza dla  $n=30$ , gdzie fluktuacje końców powinny być duże. W takim razie  $\sigma$  w doświadczeniu oznacza prawdopodobnie błąd średniej, a nie dyspersję. Wyniki różnią się jednak znacznie od tych dla "czystego"  $Q_n$ , co można przypisać znaczeniu naładowanych lizyn. Jeśli przyjmiemy, że zmiana pH z 7 do 12 zmienia głównie stan protonacyjny lizyny (zgodnie z Enciso i in. [161]), powinno to być w przybliżeniu równoważne zamianie lizyny na neutralną asparaginę. Rys. 2.17 pokazuje, że symulacje gdzie zamiast lizyny pojawia się asparagina są zgodne z danymi dla pH=12.

### 2.5.8.2 Polialanina i poliprolina

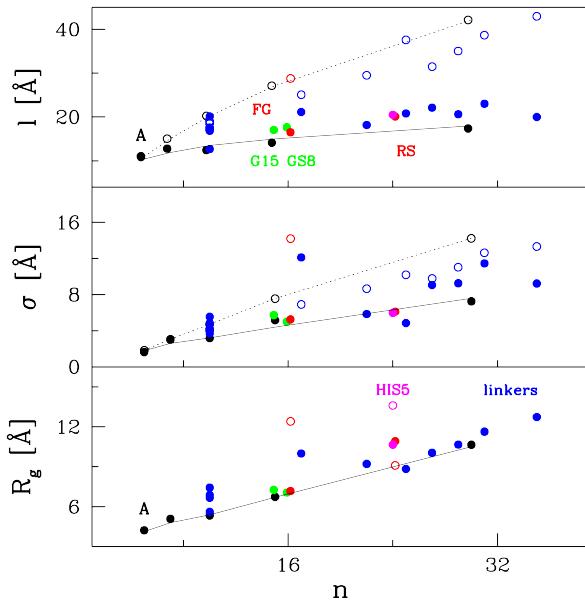
Kolejnym interesującym układem jest polyA. Symulacje pełnoatomowe, używające modelu TIP3P dla wody [162] i pola siłowego CHARMM36 [97] pozwoliły wyznaczyć  $l$  i  $\sigma$  dla wielu wartości  $n$ . Porównanie z modelem gruboziarnistym pokazuje Rys. 2.19. Dla  $n \leq 15$  zgodność jest bardzo dobra, ponieważ w obu przypadkach konformacje są helikalne (co jest omówione także w [20]), jak pokazuje lewy panel Rys. 2.18. Dla większych wartości  $n$  dane gruboziarniste zaczynają różnić się od pełnoatomowych. Jednak w symulacjach pełnoatomowych (dla których konformacją początkową była  $\alpha$ -helisa) całkowity czas symulacji to 100 ns, podczas gdy w symulacji gruboziarnistej jest on 500 razy większy. W tym drugim przypadku łańcuch może odwiedzić o wiele więcej punktów w przestrzeni fazowej i przyjąć konformacje takie jak na środkowym panelu Rys. 2.18. Takie sklepione konformacje pozwalają na mniejszą odległość między końcami łańcucha niż w przypadku helisy. Zostały one zaobserwowane w symulacjach  $A_n$  dla  $n > 40$ , używających metody ukrytego rozpuszczalnika, a zatem obejmujących dłuższe skale czasowe [163].

Graniczna wartość  $n$  powyżej której pojawiają się sklepione konformacje może zależeć od wielu czynników, takich jak sztywność łańcucha – w modelu gruboziarnistym jest ona dosyć mała, więc graniczne  $n$  może być mniejsze. Niezależnie od dokładnej wartości granicznego  $n$ , konformacja  $\alpha$ -helisy nie może być jedynym preferowanym stanem dla odpowiednio długich białek (o ile nie jest stabilizowana dodatkowymi oddziaływaniami) ze względu na dużą giętkość i mały wkład do entropii [164].

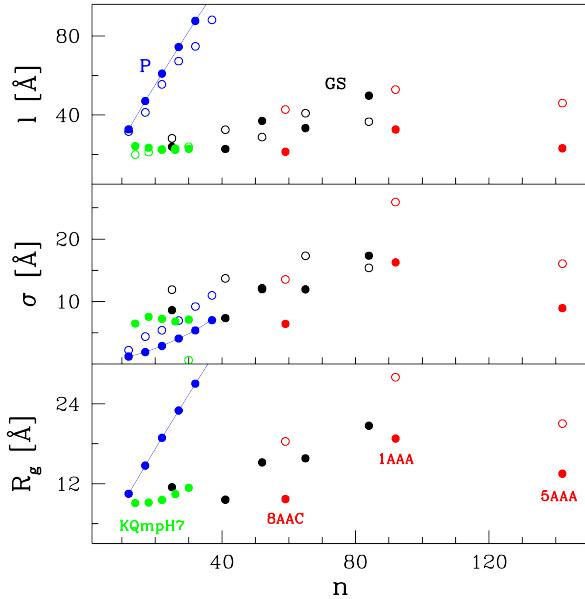


Rysunek 2.18: Przykłady konformacji trzech homopeptydów: polialanin  $A_{15}$  (po lewej) i  $A_{30}$  (pośrodku) oraz poliglutaminy  $Q_{40}$  (po prawej). Ostatnia konformacja jest zawęziona: N-koniec  $Q_{40}$  (pokazany na czarno) przechodzi przez pętle do środka, tworząc płytke węzły. Modyfikacja rys. 7 z artykułu [I].

Wyniki dla polyP pokazuje Rys. 2.20. Zgodność z danymi uzyskanymi metodą FRET [14] dla  $l$  oraz  $\sigma$  jest bardzo dobra. Wszystkie parametry geometryczne szybko rosną wraz z  $n$ , co odzwierciedla fakt że dla polyP kształt łańcucha wynika głównie z jego dużej sztywności, tak więc oddziaływanie nielokalne takie jak kontakty mają mniejsze znaczenie.



Rysunek 2.19: Panele jak na Rys. 2.16, tyle że dla innych układów. Pełne kółka oznaczają wyniki modelu gruboziarnistego, puste kółka wyniki symulacji pełnoatomowych. Czarne kółka dotyczą polyA, niebieskie wybranych łączników [97]. Wybrane zostały łączniki z  $n = 10$  oraz te z  $n > 15$ . Czerwone kółka odpowiadają układom oznaczonym jako FG oraz RS. W przypadku FG, dane dostępne były tylko dla  $R_g$ . Zielone kółka to układy G15 oraz GS8. Fioletowe punkty odnoszą się do His5. Oparte na rys. 5 z artykułu [I].



Rysunek 2.20: Panele jak na Rys. 2.16 i 2.19, ale porównanie dotyczy wyników eksperymentalnych (puste kółka). Niebieskie kółka to polyP, czarne to układ oznaczony jako GS, czerwone odpowiadają białkom z bazy PEDb [19]. Zielone punkty dotyczą układu KQmpH7. Oparte na rys. 6 z artykułu [I].

### 2.5.8.3 Inne peptydy

Białka bogate w glicynę [15, 21] często posiadają fragmenty zawierające także serynę (takie jak GGS czy GGGGS, razem nazywane “GS”). Ze względu na wysoką zawartość glicyny białka te powinny być bardzo giętkie. Faktycznie  $\sigma$  na Rys. 2.19 jest duże. Dane te dobrze zgadzają się z modelem gruboziarnistym. Jednak odległość  $l$  z artykułów [15, 21] nie była obliczona bezpośrednio na podstawie wyników FRET: posłużyła jako wiąz do odtworzenia rozkładów odległości przy użyciu prostych modeli.

Symulacje pełnoatomowe nieuporządkowanych łączników (o sekwencjach podanych w tabeli 1 w artykule [97]) obejmowały także atomy rozpuszczalnika. Niestety mieliśmy dostęp tylko do danych dotyczących  $l$ , a nie  $R_g$ . W modelu gruboziarnistym odwzorowane zostały wszystkie łączniki, ale dla zwiększenia czytelności na Rys. 2.19 pokazane zostały tylko łączniki z  $n=10$  oraz  $n \geq 17$  (niebieskie kółka). Dane mają duży rozrzut, ale wyniki są zadowalająco zgodne.

Nawet bardzo popularne pełnoatomowe pola siłowe mogą nie nadawać się do symulacji białek nieuporządkowanych, ponieważ trzeba w takich symulacjach uwzględnić zwiększoną rolę oddziaływań z cząsteczkami wody w porównaniu z białkami ustrukturyzowanymi [165, 166, 167]. Zaproponowane niedawno pole siłowe CHARMM36m [167] powinno prowadzić do lepszych wyników zarówno dla białek nieuporządkowanych jak i tych ustrukturyzowanych. Pole to zostało wykorzystane do symulacji kilku białek. Jednymi z nich były nukleoporyna FG i peptyd RS, oba nieuporządkowane. Rys. 2.19 pokazuje, że model gruboziarnisty daje wyniki porównywalne z tymi uzyskanymi przez nowe pole siłowe (dla peptydu RS dostępne były  $l$  i  $\sigma$ , a dla FG-nukleoporyny tylko  $R_g$ ).

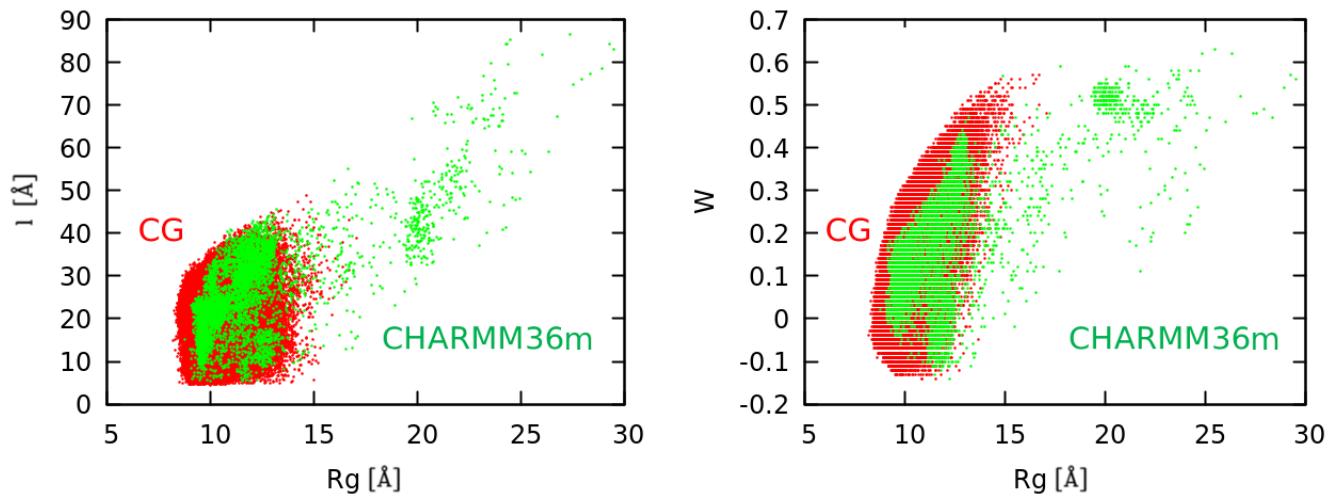
Zgodność jest gorsza dla histatyny-5 (His5), użytej do parametryzacji pól Amber i Gromos [22]). Wynika to prawdopodobnie z tego, że w modelu gruboziarnistym histydyna jest traktowana jak aminokwas nienaładowany. Dodanie cząstkowego (lub losowo przydzielonego części histydyn) ładunku powinno poprawić zgodność (odpychanie elektrostatyczne zwiększyłoby  $R_g$ ).

Baza danych PEDb (ang. *Protein Ensemble Databank*) [19] zawiera wiele danych strukturalnych uzyskanych dzięki symulacjom białek nieuporządkowanych z wiązami doświadczalnymi (dane eksperymentalne pochodzą głównie z metod NMR i SAXS). Białka te mają  $n$  znaczaco większe od rozważanych dotychczas. Rys. 2.20 pokazuje, że choć dane dla  $R_g$  nie zgadzają się z modelem gruboziarnistym, rząd wielkości jest ten sam. Jednak dane z PEDB zawierają nie tylko wyniki eksperymentów, ale także założenia modeli użytych do wygenerowania struktur, a zatem brak zgodności w tym przypadku nie jest tak niekorzystny dla modelu gruboziarnistego jak byłoby w pozostałych przypadkach.

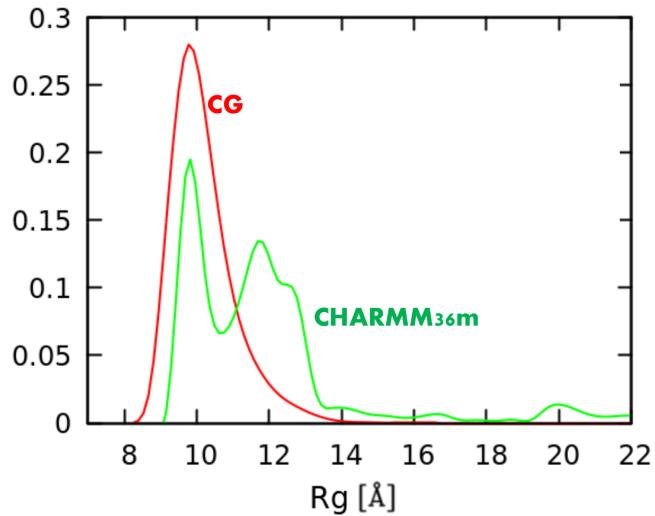
#### 2.5.8.4 Białko A $\beta$ 42

W przypadku symulacji pełnoatomowych można porównywać nie tylko wartości średnie, ale także to jak wygląda rozkład danej wielkości. Dzięki uprzejmości prof. D Thompsona i dr S. Bhattacharya z Uniwersytetu Limerick miałem dostęp do trwających ponad 100 ns symulacji pełnoatomowych białka A $\beta$ 42, kluczowego w rozwoju choroby Alzheimera [168]. Rys. 2.21 przedstawiają wykresy, na których zaznaczone jest  $R_g$  oraz  $l$  dla każdej konformacji bądź  $R_g$  i parametr asferyczności  $w$ . Parametr  $w$  jest równy 0 dla idealnej sfery, a 1 dla idealnego pręta [169]. Jest zdefiniowany jako  $w = \frac{\Delta R}{\bar{R}}$ , gdzie  $\bar{R} = \frac{1}{2}(R_1 + R_3)$ , zaś  $\Delta R = R_2 - \bar{R}$ ; wartości własne tensora bezwładności  $R_1$ ,  $R_2$  i  $R_3$  są ponumerowane od najmniejszej do największej.

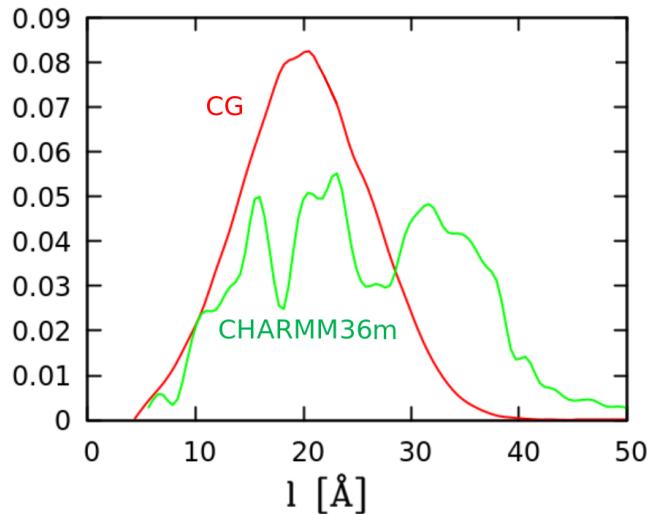
Widać, że na obu panelach Rys. 2.21 chmury odpowiadające największej liczbie punktów się pokrywają, jednak pełnoatomowe pole siłowe produkuje także bardziej rozwinięte konformacje. Trudno oszacować na ile są one liczne (punkty mogą nachodzić na siebie), w związku z tym wykreślone zostały także rozkłady tych parametrów ( $R_g$  na Rys. 2.22,  $l$  na Rys. 2.23). Poza głównym maksimum (które pokrywa się z wynikami modelu gruboziarnistego) na rozkładach widoczne jest także drugie maksimum, które reprezentuje konformacje rozwinięte i w dużym stopniu otoczone rozpuszczalnikiem. W modelu z ukrytym rozpuszczalnikiem odtworzenie takich konformacji byłoby bardzo trudne, ponieważ niewysyczane kontakty oznaczają wyższą energię (w modelu pełnoatomowym jest to zrównoważone korzystnym oddziaływaniem z wodą).



Rysunek 2.21: Porównanie symulacji białka A $\beta$ 42 modelem gruboziarnistym (czerwony) z wynikami uzyskanymi polem siłowym CHARMM36m (zielony). Każda kropka oznacza jedną konformację.

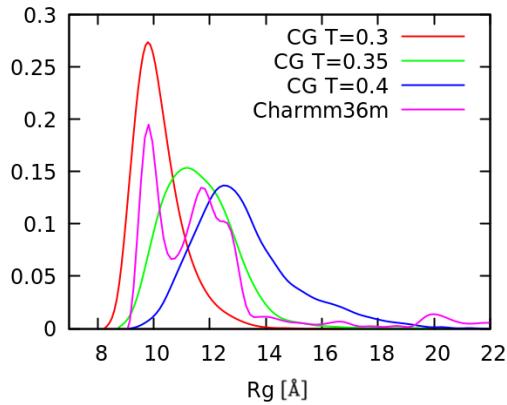


Rysunek 2.22: Rozkład wartości  $R_g$  dla białka  $\text{A}\beta_{42}$  symulowanego modelem gruboziarnistym (czerwony) oraz polem siłowym CHARMM36m (zielony).

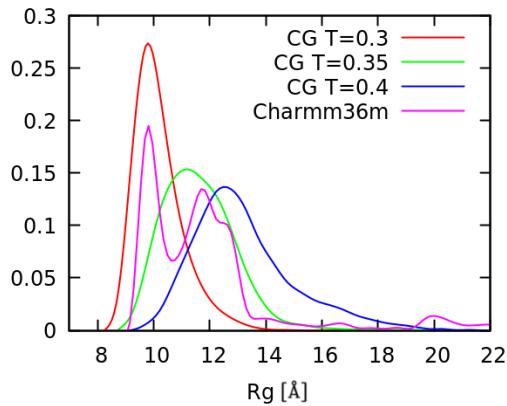


Rysunek 2.23: Rozkład wartości  $l$  dla białka  $\text{A}\beta_{42}$  symulowanego modelem gruboziarnistym (czerwony) oraz polem siłowym CHARMM36m (zielony).

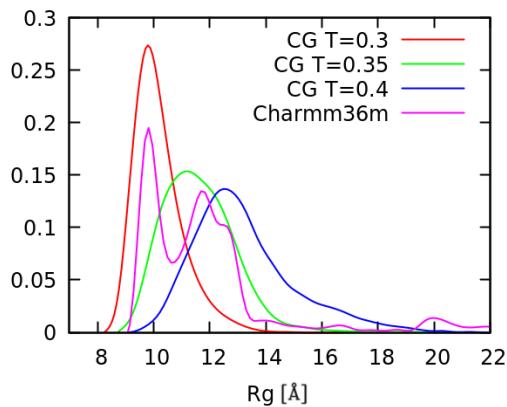
Interesujące jest także to, jak rozkłady zmieniają się wraz ze zmianą temperatury symulacji. Rys. 2.24, 2.25 i 2.26 przedstawiają porównanie rozkładów odpowiednio  $R_g$ ,  $l$  i  $w$  dla temperatur  $0.3 \epsilon/k_B$  (czerwony),  $0.35 \epsilon/k_B$  (zielony) i  $0.4 \epsilon/k_B$  (niebieski). Rozkłady  $R_g$  dla  $T = 0.3 \epsilon/k_B$  odpowiadają głównemu maksimum rozkładu pełnoatomowego, dla  $T = 0.35 \epsilon/k_B$  maksimum w modelu gruboziarnistym przesuwa się do drugiego maksimum rozkładu pełnoatomowego, a dla  $T = 0.4 \epsilon/k_B$  jest już na prawo od drugiego maksimum. Oznacza to, że nie ma temperatury dla której rozkład byłby, tak jak dla symulacji pełnoatomowych, bimodalny. Warto jednak zauważyć, że zgodność rozkładów dla  $w$  oraz  $l$  jest bardzo dobra.



Rysunek 2.24: Rozkład wartości  $R_g$  dla białka  $\text{A}\beta 42$  symulowanego polem siłowym CHARMM36m (fioletowy) oraz modelem gruboziarnistym w temperaturach  $0.3 \epsilon/k_B$  (czerwony),  $0.35 \epsilon/k_B$  (zielony) i  $0.4 \epsilon/k_B$  (niebieski).



Rysunek 2.25: Rozkład wartości  $l$  dla białka  $\text{A}\beta 42$  symulowanego polem siłowym CHARMM36m (fioletowy) oraz modelem gruboziarnistym w temperaturach  $0.3 \epsilon/k_B$  (czerwony),  $0.35 \epsilon/k_B$  (zielony) i  $0.4 \epsilon/k_B$  (niebieski).



Rysunek 2.26: Rozkład wartości  $w$  dla białka  $\text{A}\beta 42$  symulowanego polem siłowym CHARMM36m (fioletowy) oraz modelem gruboziarnistym w temperaturach  $0.3 \epsilon/k_B$  (czerwony),  $0.35 \epsilon/k_B$  (zielony) i  $0.4 \epsilon/k_B$  (niebieski).

## 2.6 Model z Hamiltonianem niezależnym od czasu

Kluczową ideą prezentowanego w tej pracy podejścia jest rozróżnienie kiedy oddziaływała łańcuch boczny, a kiedy główny, tylko na podstawie położen atomów  $C_\alpha$ . Określenie rodzaju kontaktu między 2 aminokwasami w modelu quasi-adiabatycznym wymagało znajomości położenia 6 aminokwasów, ale kiedy kontakt był już włączony, siły działały tylko między parą aminokwasów w kontakcie (i były określone zwykłym sferycznie symetrycznym potencjałem LJ). Co prawda chwilowa mapa kontaktów powinna utrzymywać łańcuch w geometrii zbliżonej do takiej, jaką mogłoby przyjąć prawdziwe białko (tak dzieje się w modelach opartych na strukturze natywnej), jednak w ten sposób dynamika układu musi zależeć od jego historii (bo to ona określa chwilową mapę kontaktów).

Z tego powodu quasi-adiabatyczne przełączanie kontaktów nie spełnia postulatu równowagi szczególowej i może prowadzić do nierównowagowych stanów stacjonarnych, które były badane np. w kontekście błon biologicznych [170, 171, 172]. Odpowiednio dobrany czas przełączania powinien sprawić, że problem ten nie będzie istotny w skalach czasowych istotnych dla symulowanych procesów (patrz Rys. 2.1). Jednak całkowicie uniknąć tych problemów można tylko w modelu, który w sposób ciągły utrzymuje prawidłową geometrię łańcucha.

Aby cały czas kontrolować, czy kryteria kierunkowe są spełnione, można albo wrócić do próbkowania Monte Carlo (dla którego kryteria te były pierwotnie wypracowane [145, 149]), albo wprowadzić do hamiltonianu człon wielociałowy. Człony takie są kluczowe w odtwarzaniu więzów geometrycznych, które nie wynikają już z kształtu aminokwasów, ponieważ ich detale strukturalne zostały utracone w reprezentacji gruboziarnistej [173, 174]). Jest to szczególnie ważne, gdy aminokwas jest reprezentowany tylko przez jedną kulę. Przykładem takiego wielociałowego członu jest wzór opisujący oddziaływanie dipol-dipol, który może być użyty do modelowania wiązań wodorowych tworzonych przez łańcuch główny [175].

W tym podrozdziale prezentowany jest nowy, oparty na doświadczeniu i przeznaczony do dynamiki molekularnej model, w którym oddziaływanie krótkozasięgowe są reprezentowane przez niezależne od czasu człony 4-ciałowe. Punktewyjście jest 4-ciałowy potencjał dwuścienny, używany do zapewnienia odpowiedniej sztywności łańcucha białkowego<sup>1</sup> [176]. Ogranicza on kąt, jaki może powstać między dwiema płaszczyznami, każda wyznaczona przez trzy aminokwasy.

Potencjał ten może być użyty w “niewłaściwej” (ang. *improper*) postaci, aby opisywać sztywność łańcucha bocznego w modelu z 2 kulkami na aminokwas (bądź w każdym innym przypadku, gdy jedna kula tworzy więcej niż 2 wiązania). Gdy dwa aminokwasy A i B oddziałują ze sobą, obraz ten

---

<sup>1</sup> W pierwotnej postaci dla białek odtwarza on detale wiązania peptydowego, i jest potrzebny nawet w pełnoatomowych polach siłowych aby lepiej oddać częściowo podwójną naturę tego wiązania.

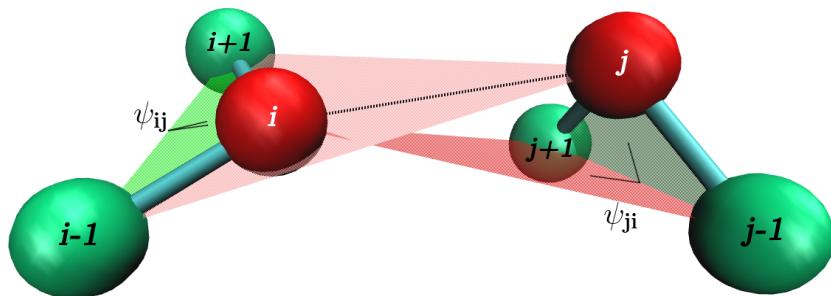
może być jeszcze bardziej uproszczony poprzez usunięcie kulki reprezentującej łańcuch boczny aminokwasu A i wstawienie na jej miejsce kulki reprezentującej aminokwas B. I odwrotnie: łańcuch boczny aminokwasu B zostaje zastąpiony kulką reprezentującą aminokwas A. W ten sposób “niewłaściwy” potencjał dwuścienny może być użyty w modelu z jedną kulką na aminokwas. Płaszczyzny utworzone w takim przypadku przez odpowiednie trójki aminokwasów są pokazane na Rys. 2.27. Dzięki tej procedurze każde oddziaływanie zachodzi między konkretną parą aminokwasów (mimo użycia członów 4-ciałowych) i jest opisywane przez dwa “pseudo-niewłaściwe” kąty dwuścienne, po jednym dla każdego z aminokwasów. Kąty te będą oznaczone skrótem PID (ang. *Pseudo-Improper-Dihedral*).

To, jakie kąty PID są najczęściej obecne w białkach, zostało ustalone na podstawie wspomnianej w podpunkcie 2.5.2 bazy struktur PDB. Okazało się, że rozkład kąta PID jest całkiem inny w zależności od tego, czy przyporządkowany mu aminokwas oddziałuje z drugim przy pomocy łańcucha głównego, czy bocznego. Dzięki temu potencjał PID może być użyty do odróżnienia tych dwóch przypadków.

Użycie nowego 4-ciałowego potencjału zamiast potencjału LJ przełączanego quasi-adiabatycznie wymagało reparametryzacji wszystkich aspektów modelu. Użyty w tym celu zestaw danych eksperymentalnych jest inny niż ten opisany w podpunkcie 2.5.8. Nowy zestaw danych składa się z 23 promieni bezwładności ( $R_g$ ) wyznaczonych eksperymentalnie dla 23 białek nieuporządkowanych. Model quasi-adiabatyczny także został sprawdzony dla tych danych.

Aby sprawdzić, który wariant modelu po reparametryzacji najlepiej zgadza się z eksperymentem, wykorzystany został współczynnik Pearsona, który pozwala określić rozbieżność między wynikami symulacji i eksperymentu. Najlepsze wartości tego współczynnika zostały osiągnięte przez model PID.

Innym sposobem porównania modeli jest sprawdzenie rozkładów energii, które dla modelu quasi-adiabatycznego wykazują odchylenia od rozkładu Boltzmanna. Nie są one obecne w modelu PID. Nowy model wymaga jednak znaczowo większych mocy obliczeniowych.



Rysunek 2.27: Idea kątów PID. Potencjał oddziaływania między aminokwasami  $i$  oraz  $j$  zależy od kąta  $\Psi_{ij}$  (określonego przez pseudoatomy  $i - 1, i, i + 1$  oraz  $j$ ), kąta  $\Psi_{ji}$  (określonego przez pseudoatomy  $j - 1, j, j + 1$  oraz  $i$ ) oraz od odległości  $r$  między pseudoatomami  $i$  oraz  $j$ . Oparte na rys. 1 z artykułu [II].

## 2.6.1 Dane z bazy PDB

Prawie wszystkie białka składają się z tych samych 20 rodzajów aminokwasów, więc natura oddziaływania między parą aminokwasów powinna być podobna w przypadku białek uporządkowanych i nieuporządkowanych, nawet jeśli częstość występowania tych oddziaływań się różni (np. wskutek tego, że białka nieuporządkowane zawierają mniej białek hydrofobowych).

Dlatego aby określić jakie kąty PID najczęściej pojawiają się przy oddziaływaniu par aminokwasów wykorzystany został zestaw 21 090 uporządkowanych białek z bazy CATH [143] (o podobieństwie sekwencji nieprzekraczającym 40%: cath-dataset-nonredundant-S40.pdb).

Wartości kątów PID zostały obliczone dla każdej pary aminokwasów w kontakcie, gdzie kontakt jest zdefiniowany na podstawie kryterium przekrywania ciężkich atomów (patrz podpunkt 2.5.2).

Ciężkie atomy mogą należeć do łańcucha bocznego albo do głównego. Jeśli ciężki atom z łańcucha głównego jednego z aminokwasów przekrywa się z ciężkim atomem łańcucha bocznego drugiego aminokwasu, jest to kontakt typu bs. Analogicznie przekrywanie łańcuchów bocznych skutkuje kontaktem typu ss, a kiedy przekrywają się łańcuchy główne, powstaje kontakt bb. Dla danej pary aminokwasów może przekrywać się więcej niż jedna para atomów, wtedy kontakt jest więcej niż jednego typu naraz.

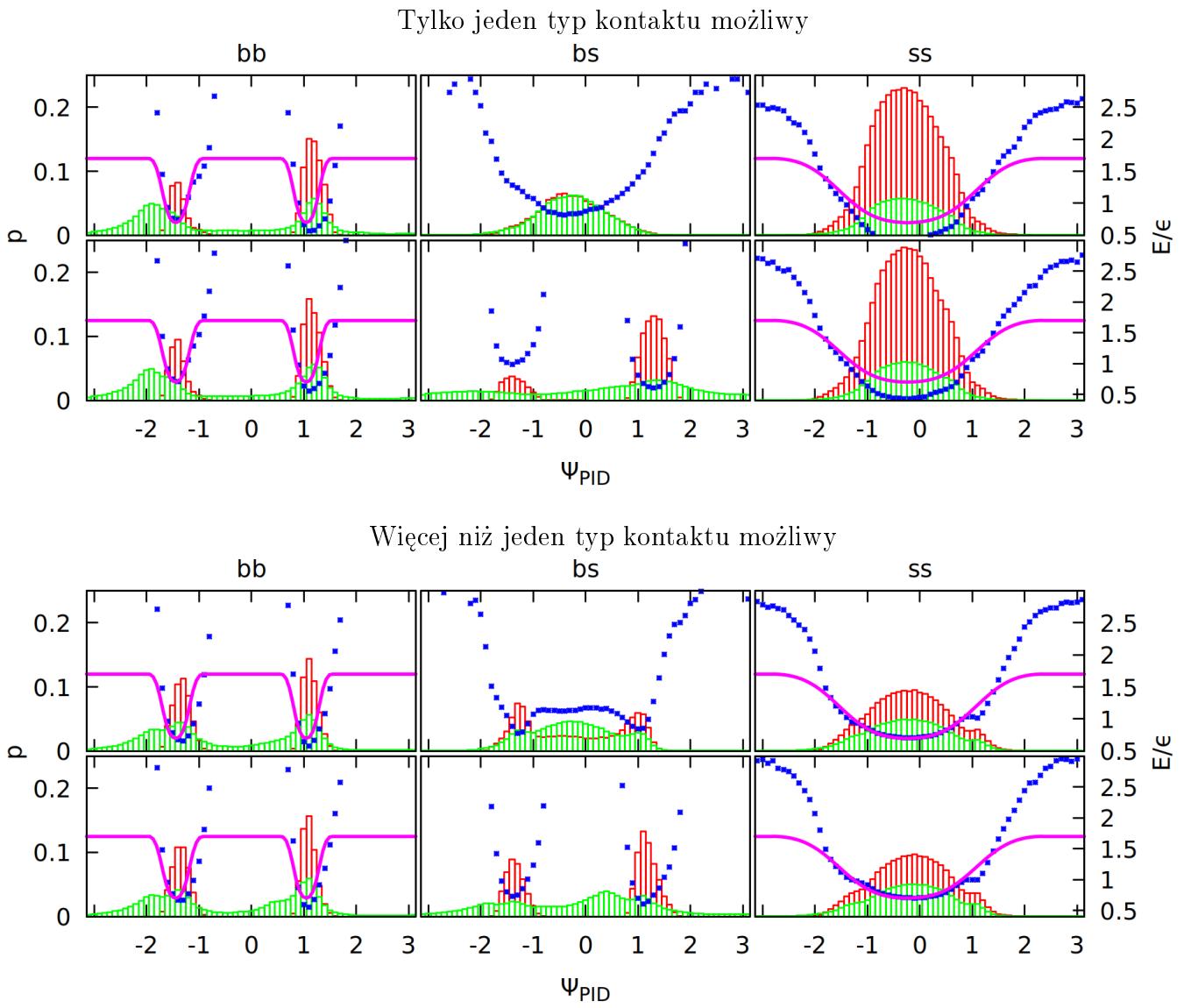
Aby przyporządkować kąty PID i odległości charakterystyczne dla jednego typu kontaktu, użyte zostały rozkłady odpowiadające tylko tym kontaktom, w których przekrywanie atomów prowadzi do kontaktu tylko jednego typu (np. tylko ss). Tabela 2.5 pokazuje ile kontaktów każdego typu zostało wykrytych w użytym zestawie białek.

Typ kontaktu	bb	bs	ss
Liczba par aminokwasów z kontaktem <b>tylko</b> tego typu	624699	953782	1870746
Liczba par aminokwasów z kontaktem tego typu	3742271	3872292	4297919

Tabela 2.5: Liczba kontaktów na podstawie kryterium przekrywania dla wszystkich białek z użytego zestawu (w sumie 7974804).

Rys. 2.28 pokazuje rozkłady kątów PID dla przypadków ss, bs i bb. Zielone histogramy pokazują rozkład kontaktów określonych na podstawie kryterium przekrywania, czerwone histogramy uwzględniają tylko te z kontaktów, które spełniają kryteria kierunkowe określone w podpunkcie 2.5.4.

Różnice między rozkładami kontaktów tylko jednego typu (góra Rys.2.28) a rozkładami uwzględniającymi także kontakty, które są więcej niż jednego typu naraz (dół Rys. 2.28) są niewielkie i dotyczą głównie przypadku bs. Potencjał statystyczny uzyskany na podstawie tych rozkładów (czerwone histogramy) dzięki inwersji boltzmannowskiej  $V_B = -k_B T \ln(p(\Psi_{PID}))$  został dopasowany do funkcji analitycznej opisanej w podpunkcie 2.6.5.

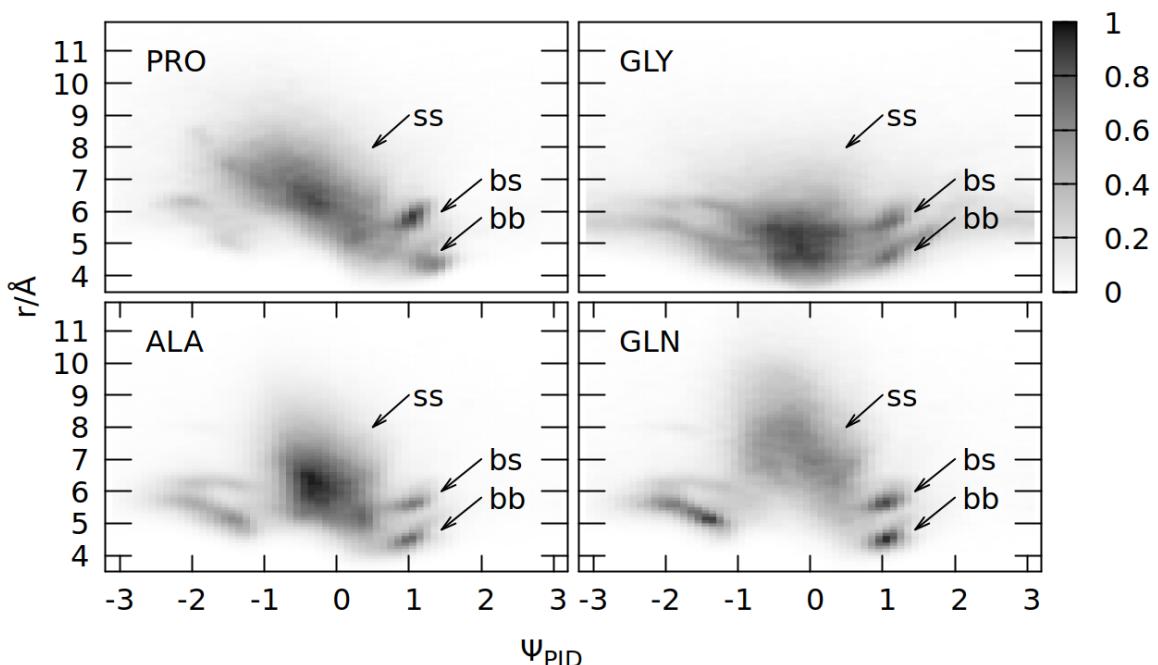


Rysunek 2.28: Rozkłady kątów PID z kontaktów **tylko** jednego typu (górny obrazek) bądź więcej niż jednego typu (dolny obrazek). Kontakt  $i, i+3$  oraz  $i, i+4$  nie są uwzględnione. Każdemu kontaktowi odpowiadają dwa kąty PID. Rozkłady pierwszego ( $\Psi_{PID}^{ij}$ ) są na górnym panelach, a drugiego ( $\Psi_{PID}^{ji}$ ) na dolnych. Czerwone histogramy uwzględniają tylko kontakty spełniające kryteria kierunkowe, zielone uwzględniają wszystkie. Funkcja analityczna (purpurowa linia) została dopasowana do potencjału wynikającego z odwrócenia boltzmannowskiego (niebieskie kropki, jednostka energii  $\epsilon \approx 1.5$  kcal/mol). Oparte na rys. 2 oraz S2 z artykułu [II].

Kontakty tworzone przez łańcuch boczny i łańcuch główny są powiązane z innymi wartościami kątów PID. Dwa minima potencjału dla przypadku bb odpowiadają równoległym bądź antyrównoległym  $\beta$ -kartkom albo prawo- bądź lewo-skrętnym  $\alpha$ -helisom (dla kontaktów  $i, i+3$  obecne jest tylko jedno minimum, co odpowiada przewadze prawoskrętności, patrz podpunkt 2.6.2).

Rys. 2.29 pokazuje 4 przypadki rozkładów dwuwymiarowych, gdzie kąt PID  $\Psi_{PID}$  jest na jednej osi, a odległość  $C_\alpha$ - $C_\alpha$ , czyli  $r$ , jest na drugiej osi. Różne łańcuchy boczne odpowiadają innym rozkładom  $r$  (jak widać np. dla GLN i ALA), ale rozkłady kątów PID są podobne (z wyjątkiem PRO i GLY). Kontakty łańcucha bocznego dają szerokie rozkłady dla  $\Psi_{PID} \approx 0$  rad i  $r > 6 \text{ \AA}$ , podczas gdy kontakty łańcucha głównego odpowiadają mniejszym, lepiej określonym wartościom  $r < 6 \text{ \AA}$  i  $\Psi_{PID} \approx \pm 1$  rad.

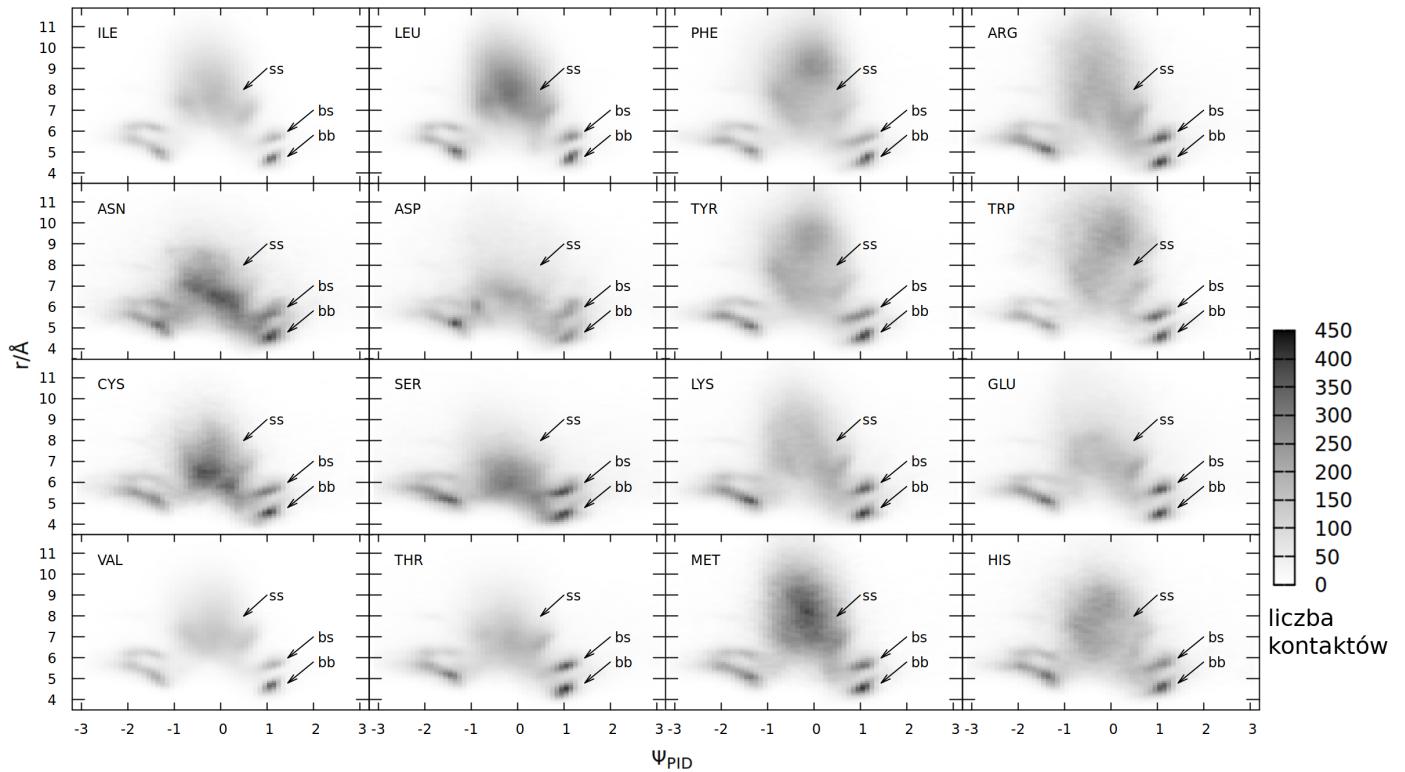
Dwuwymiarowe rozkłady dla pozostałych 16 przypadków pokazuje Rys. 2.30. Charakterystyczne odległości dla kontaktów ss ( $r_{min}^{ss}$ , podane w tabeli 2.3) zostały wyznaczone na podstawie jednowymiarowych rozkładów, pokazanych na Rys. 2.8, 2.9, 2.10.



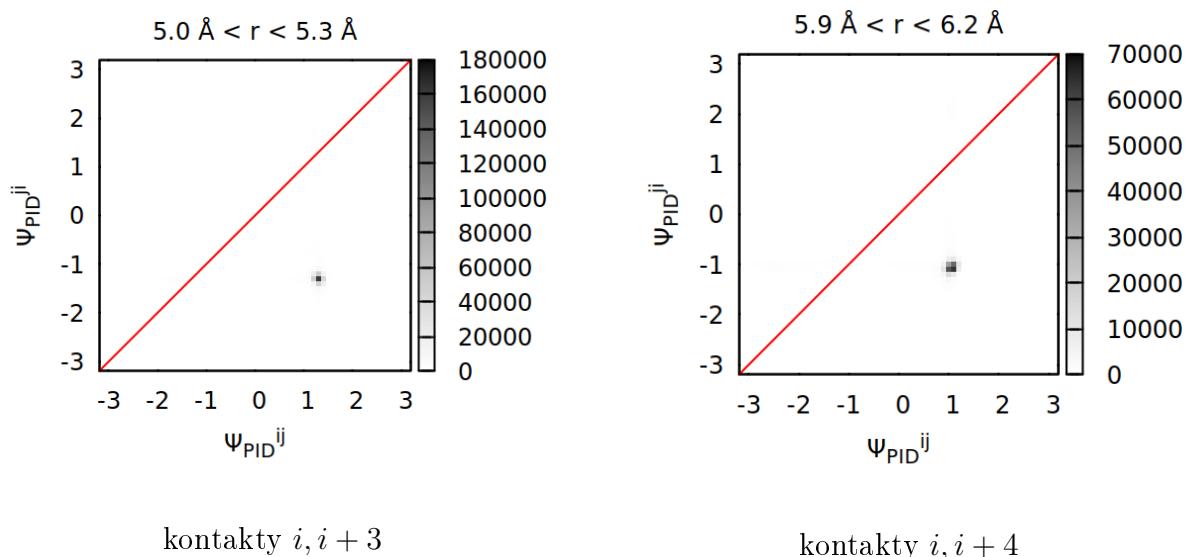
Rysunek 2.29: Dwuwymiarowe rozkłady z kątem PID ( $\Psi_{PID}$ , w radianach) na osi rzędnych i odlegością  $C_\alpha$ - $C_\alpha$  ( $r$ ) na osi odciętych. Kontakty  $i, i+3$  oraz  $i, i+4$  nie są uwzględnione. Rozkłady dotyczą każdego kontaktu z zestawu białek, gdzie choć jeden aminokwas w parze był danego rodzaju (PRO, GLY, ALA, GLN). Kąt PID ( $\Psi_{PID}$ ) dla każdego kontaktu jest tym związanym z danym aminokwasem (jeśli kontakt jest między takimi samymi aminokwasami, odpowiadają mu w rozkładzie dwa różne kąty PID). Oparte na rys. 3 z artykułu [II].

## 2.6.2 Prawoskrętność

Dla kontaktów  $i, i+3$  oraz  $i, i+4$  rozkłady kątów PID dla pierwszego i drugiego aminokwasu (licząc od N-końca) różnią się. Wynika to z prawoskrętności większości  $\alpha$ -helis. Różnicę pokazuje Rys. 2.31. Rozkłady dla kontaktów  $i, i+5$  i dalszych są w większości symetryczne ze względu na zamianę kątów, z wyjątkiem przypadku bs (patrz podpunkt 2.6.3).



Rysunek 2.30: Dwuwymiarowe rozkłady z kątem PID ( $\Psi_{PID}$ ) na osi rzędnych i odległością  $C_\alpha$ - $C_\alpha$  ( $r$ ) na osi odciętych. Kontakty  $i, i + 3$  oraz  $i, i + 4$  nie są uwzględnione. Rozkłady jak na Rys. 2.29, tyle że dla pozostałych 16 aminokwasów. Oparte na rys. S4 z artykułu [II].

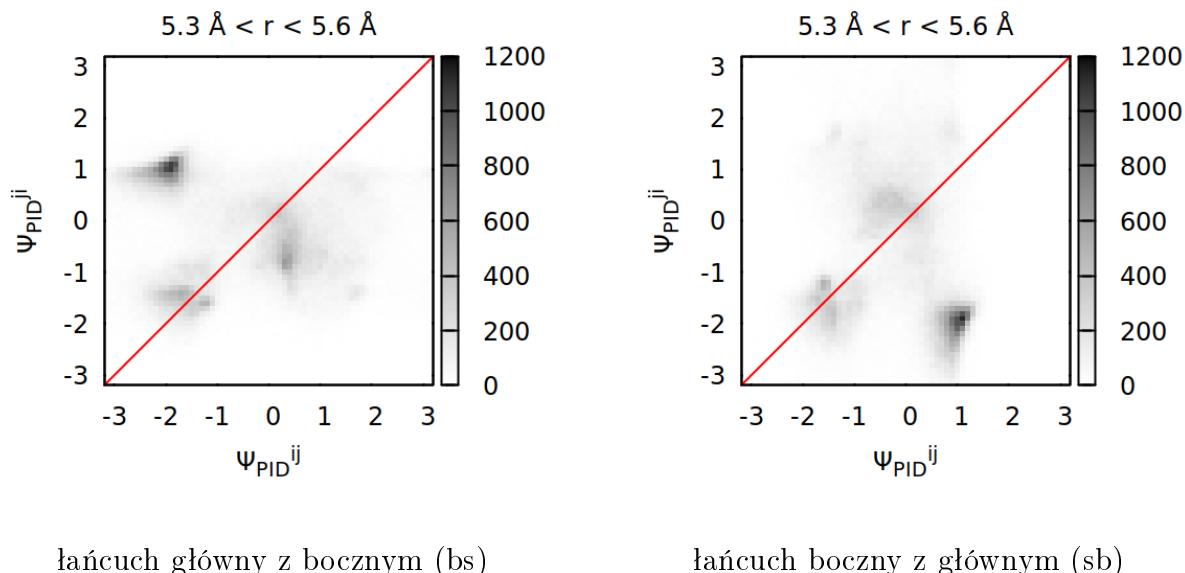


Rysunek 2.31: Dwuwymiarowe rozkłady dla kontaktów bb, gdzie pierwszy kąt PID w parze jest na osi odciętych, a drugi (licząc od N-końca) na osi rzędnych. Asymetryczność zdaje się wynikać z prawoskrętności  $\alpha$ -helix. Rozkłady dotyczą odległości  $r$  z zakresu o szerokości 0.3 Å aby uniknąć szumu pochodzącego od dalszych odległości. Dla wybranych odległości rozkłady są bardzo wąskie (jednemu słupkowi histogramu odpowiada zakres kątów 0.1 rad x 0.1 rad). Oparte na rys. S4 z artykułu [II].

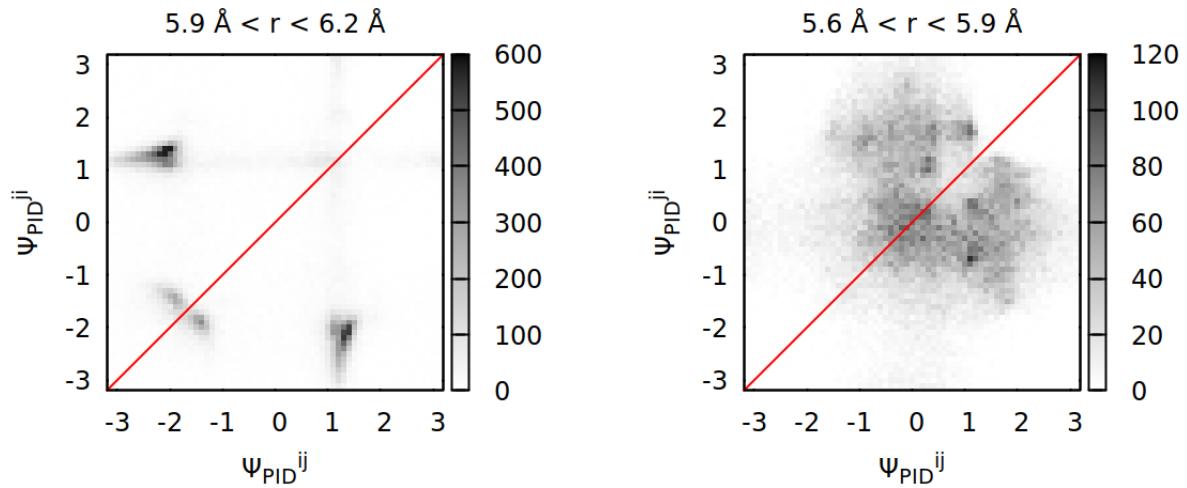
### 2.6.3 Kontakty typu bs

Kontakty bs mają bardzo szerokie rozkłady kątów PID i odległości C<sub>α</sub>-C<sub>α</sub>. Dlatego warto odróżnić przypadki, w których pierwszy aminokwas pochodzi z łańcucha bocznego, a drugi z bocznego (bs) oraz przypadek przeciwny (sb). Największa różnica (dla rozkładów gdzie kontakt może być więcej niż jednego typu naraz) zachodzi dla odległości  $5.3 \text{ \AA} < r < 5.6 \text{ \AA}$ , a maksimum rozkładu zdaje się zależeć od rodzaju kontaktu (Rys. 2.32). Jednak po obejrzeniu przykładowych struktur okazało się, że maksimum to odpowiada kontaktom bb, a łańcuch boczny jest tylko zawadą steryczną skutkującą tym, że tylko jedno maksimum jest możliwe. Potwierdza to rozkład kontaktów **tylko jednego typu**: rozkłady wyłącznie typu bb zawierają oba maksima (lewa część Rys. 2.33). Większość kontaktów typu bs jest także typu bb (patrz tabela 2.5), dlatego właśnie te mieszane kontakty zostały zaznaczone strzałką na Rys. 2.29 i 2.30.

Rozkład zawierający tylko kontakty bs (oraz sb) też jest bardzo szeroki (prawa część Rys. 2.33), będąc sumą dwóch możliwych przypadków:  $\Psi_{PID}^{ij} \approx 0 \text{ rad}$  i  $\Psi_{PID}^{ji} \approx 1.5 \text{ rad}$  bądź odwrotnie:  $\Psi_{PID}^{ji} \approx 0 \text{ rad}$  i  $\Psi_{PID}^{ij} \approx 1 \text{ rad}$ . Jest to zgodne z wynikami z Rys. 2.28, gdzie jeden aminokwas jest donorem łańcucha bocznego, a drugi głównego. Zakres odległości dla takich rozkładów to  $5.5 \text{ \AA} < r < 7 \text{ \AA}$  (dla większych odległości rozkład staje się jeszcze szerszy, patrz Rys. 2.34).



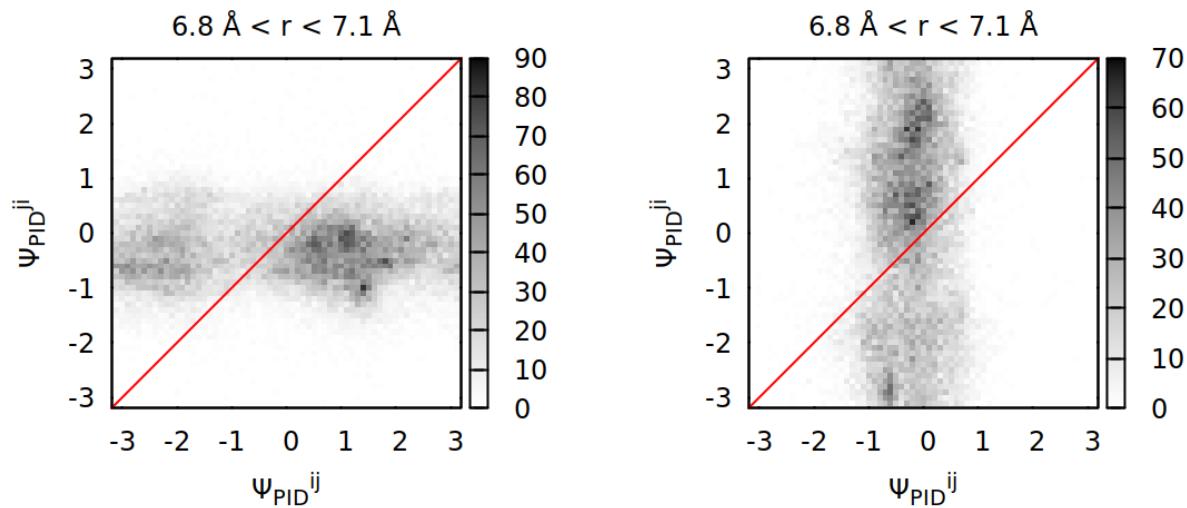
Rysunek 2.32: Dwuwymiarowe rozkłady dla kontaktów bb, gdzie pierwszy kąt PID w parze jest na osi odciętych, a drugi (licząc od N-końca) na osi rzędnych. Kontakty mogą być więcej niż jednego typu (w tym bb lub ss). Zakres odległości  $r$  jest na górze każdego wykresu. Oparte na rys. S5 z artykułu [II].



tylko łańcuch główny (bb)

tylko łańcuch boczny z głównym (bs i sb)

Rysunek 2.33: Dwuwymiarowe rozkłady kontaktów, gdzie pierwszy kat PID w parze jest na osi odciętych, a drugi (licząc od N-końca) na osi rzędnych. Kontakt mogą być tylko jednego typu (nie licząc rozróżnienia bs i sb). Zakres odległości  $r$  jest na górze każdego wykresu. Oparte na rys. S6 z artykułu [II].



tylko łańcuch główny z bocznym (bs)

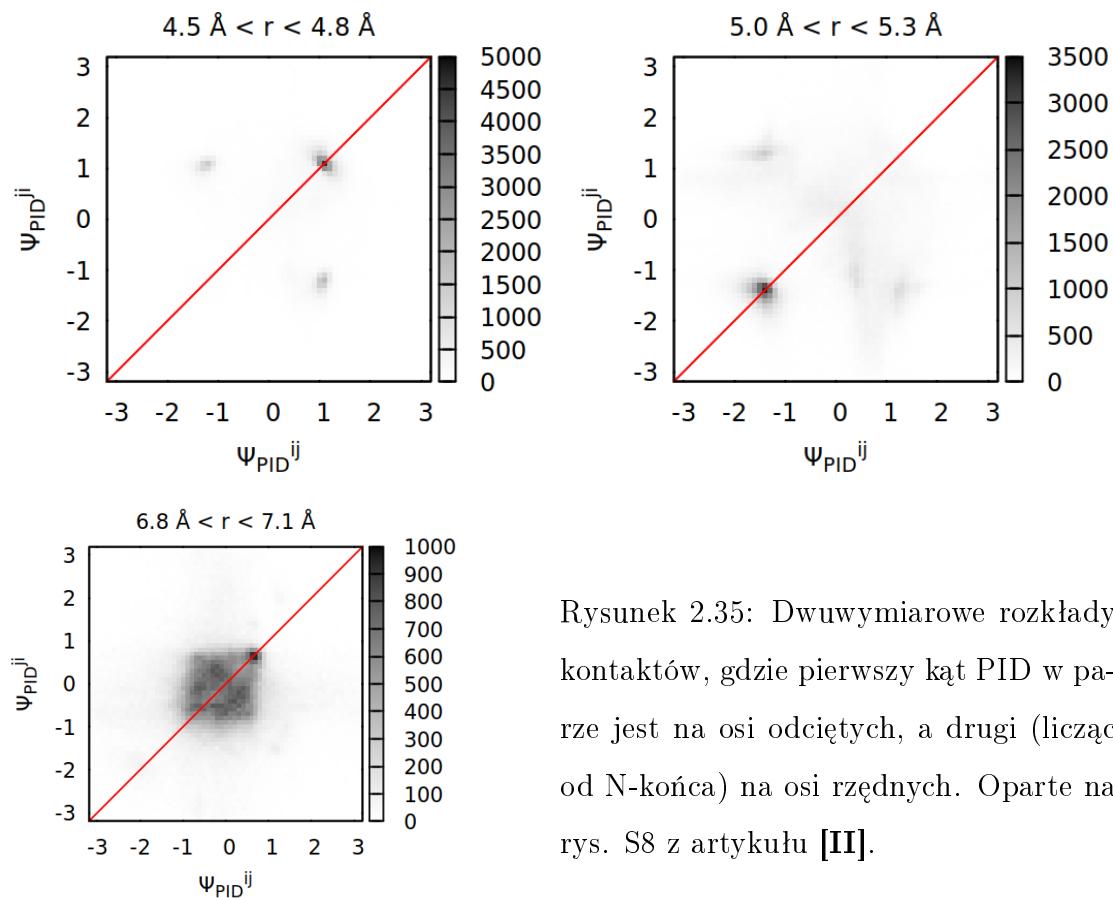
tylko łańcuch boczny z głównym (sb)

Rysunek 2.34: Dwuwymiarowe rozkłady kontaktów, gdzie pierwszy kat PID w parze jest na osi odciętych, a drugi (licząc od N-końca) na osi rzędnych. Kontakt mogą być tylko jednego typu (bs po lewej, sb po prawej). Zakres odległości  $r$  jest na górze każdego wykresu. Bezwzględna liczba kontaktów jest dużo mniejsza niż na pozostałych rysunkach. Oparte na rys. S7 z artykułu [II].

## 2.6.4 Dane bez podziału na kontakty typu bb, bs i ss

Znalezienie charakterystycznych odległości i kątów PID dla kontaktów bs jest bardzo skomplikowane, dlatego warto sprawdzić na ile są one istotne. Rys. 2.35 pokazuje rozkłady kontaktów bez podziału na typ. Okazuje się, że maksima odpowiadające kontaktom bb oraz ss są na tych wykresach dobrze rozdzielone nawet bez użycia kryteriów kątowych ani informacji o tym, które ciężkie atomy się przekrywają. Kontakty typu bb mają dwa maksima odpowiadające  $\psi_0^{bb+} = 1.05$  rad oraz  $\psi_0^{bb-} = -1.44$  rad, podczas gdy kontaktom ss odpowiada jedno szersze maksimum  $\psi_0^{ss} = -0.23$  rad. Są to te same maksima co na Rys. 2.28.

Warto zauważyc, że kąt  $\psi_{PID} \approx +1$  rad jest częstszy dla mniejszych odległości C<sub>α</sub>-C<sub>α</sub> niż kąt  $\psi_{PID} \approx -1$  rad, co ma odzwierciedlenie w potencjale PID ( $r_{min}^{bb+} = r_{min}^{bb-} - 0.6 \text{ \AA}$ ).



Rysunek 2.35: Dwuwymiarowe rozkłady kontaktów, gdzie pierwszy kąt PID w parze jest na osi odciętych, a drugi (licząc od N-końca) na osi rzędnych. Oparte na rys. S8 z artykułu [II].

## 2.6.5 Implementacja potencjału PID

Potencjał oddziaływanego między parą aminokwasów w modelu PID powinien zależeć od odległości między nimi i od dwóch kątów PID, które tworzą. Najprostszym sposobem wprowadzenia takiej zależności jest iloczyn trzech czynników:  $V(\psi_A, \psi_B, r) = \lambda_A(\psi_A)\lambda_B(\psi_B)\phi(r)$ , gdzie  $\psi_A$  jest pierwszym kątem PID w parze,  $\psi_B$  jest drugim kątem, a  $r$  odpowiada odległości C <sub>$\alpha$</sub> -C <sub>$\alpha$</sub> . Jako pierwsze przybliżenie dla  $\lambda$  została użyta funkcja cosinus, a dla  $\phi$  potencjał LJ:  $\phi(r) = \epsilon^{LJ} \left[ \left( \frac{r_{min}}{r} \right)^{12} - 2 \left( \frac{r_{min}}{r} \right)^6 \right]$ , gdzie  $r_{min}$  to minimum, a  $\epsilon^{LJ}$  to głębokość potencjału (dyskutowana w podpunkcie 2.6.6). Z powodu szerokości rozkładu bs (zielone histogramy na Rys. 2.28) model PID opisuje tylko kontakty typu bb oraz ss (patrz podpunkty 2.6.3 i 2.6.4). Jest to duża różnica względem modelu quasi-adiabatycznego, gdzie kontakty bs były uwzględnione [I]. Dla kontaktów bb oraz ss rozkłady mają wyraźne maksima, do których można dopasować potencjał statystyczny (oddzielnie dla kątów i odległości). Każdy rozkład ma inne maksimum i szerokość, dlatego szczegółowa postać funkcji  $\lambda$  to:

$$\lambda(\psi) = \begin{cases} 0.5 \cdot \cos[\alpha(\psi - \psi_0)] + 0.5 & \text{gdy } -\pi < \alpha(\psi - \psi_0) < \pi \\ 0 & \text{w przeciwnym przypadku} \end{cases} \quad (2.8)$$

Każdemu kontaktowi ss odpowiada  $r_{min}^{ss}$  z tabeli 2.3. Ponieważ dla  $r < r_{min}$  potencjał  $\phi(r)$  staje się silnie odpychający, parametry  $\alpha$  dla kontaktów ss oraz bb powinny być dobrane tak, aby  $\lambda_{ss} = 0$  gdy  $\lambda_{bb} \neq 0$  i odwrotnie.

Ponieważ rozkład kąta PID dla kontaktów bb ma dwa maksima ( $\psi_0^{bb+}$  oraz  $\psi_0^{bb-}$ ), potencjał bb ma dwa człony, odpowiadające każdemu z nich. Dlatego:

$$\lambda_{bb}(\psi) = \begin{cases} 0.5 \cdot \cos[\alpha^{bb+}(\psi - \psi_0^{bb+})] + 0.5 & \text{gdy } -\pi < \alpha^{bb+}(\psi - \psi_0^{bb+}) < \pi \\ 0.5 \cdot \cos[\alpha^{bb-}(\psi - \psi_0^{bb-})] + 0.5 & \text{gdy } -\pi < \alpha^{bb-}(\psi - \psi_0^{bb-}) < \pi \\ 0 & \text{w pozostałych przypadkach} \end{cases} \quad (2.9)$$

Odpychająca część potencjału LJ powinna zawsze być włączona dla małych odległości, aby uniknąć samoprzecinania się łańcucha (efekt wyłączonej objętości). Dlatego użyta postać potencjału bb jest jeszcze bardziej skomplikowana:

$$V^{bb}(\psi_A, \psi_B, r) = \begin{cases} \lambda^{bb}(\psi_A)\lambda^{bb}(\psi_B)\phi^{bb}(r) & \text{gdy } r > r_{min}^{bb} \\ \phi^{bb}(r) + (1 - \lambda^{bb}(\psi_A)\lambda^{bb}(\psi_B))\epsilon^{bb} & \text{w przeciwnym przypadku} \end{cases} \quad (2.10)$$

Taki wzór zapewnia nieprzecinanie się łańcucha, ponieważ dla  $r < r_{min}^{bb}$  potencjał  $\phi^{bb}$  nie jest już mnożony przez czynnik  $\lambda^{bb}$ . Dla kontaktów ss,  $V^{ss}(\psi_A, \psi_B, r) = \lambda^{ss}(\psi_A)\lambda^{ss}(\psi_B)\phi^{ss}(r)$ , zaś  $r_{min}^{ss}$  zależy od tego, które aminokwasy są w kontakcie (dalej szczegółowo  $V^{ss}$  są omówione w podpunkcie 2.6.6). Całkowity potencjał PID to:  $V = V^{ss} + V^{bb}$ .

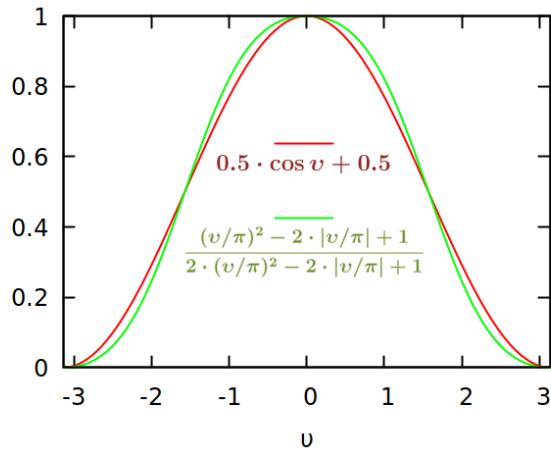
Przedstawione Rys. 2.28 dopasowanie funkcji do potencjału statystycznego opartego na rozkładach kontaktów tylko jednego typu (bb, bs albo ss), które spełniają kryteria kierunkowe (czerwone histogramy), dało następujące parametry:  $\alpha^{bb+} = 6.4$ ,  $\alpha^{bb-} = 6.0$ ,  $\alpha^{ss} = 1.2$ ,  $\psi_0^{ss} = -0.23$  rad,  $\psi_0^{bb+} = 1.05$  rad,  $\psi_0^{bb-} = -1.44$  rad. Dopasowanie do rozkładów, w których kontakty nie musiały spełniać kryteriów kierunkowych (zielone histogramy) skutkowało modelami, które gorzej zgadzały się z eksperymentem (patrz paragraf 2.6.8.3).

Wartości  $r_{min}^{ss}$  są podane w tabeli 2.3. W podpunkcie 2.6.8 model PID został oznaczony literą P, a model quasi-adiabatyczny literą A.

### 2.6.5.1 Algebraiczne przybliżenie funkcji cosinus

Aby przyspieszyć obliczenia, funkcja cosinus została zastąpiona jej algebraicznym przybliżeniem, składającym się z dwóch połączonych ze sobą algebraicznych funkcji sigmoidalnych [177] (tak że jedna jest odbiciem drugiej):

$$0.5 \cdot \cos v + 0.5 \approx \frac{(v/\pi)^2 - 2 \cdot |v/\pi| + 1}{2 \cdot (v/\pi)^2 - 2 \cdot |v/\pi| + 1} \quad (2.11)$$



Rysunek 2.36: Porównanie funkcji cosinus z jej algebraicznym przybliżeniem. Oparte na rys. S1 z artykułu [II].

### 2.6.6 Forma potencjału Lenard-Jonesa

Jednostka energii  $\epsilon = 110 \text{ pN}\cdot\text{\AA} \approx 1.5 \text{ kcal/mol}$  została przejęta z modelu quasi-adiabatycznego [I], który z kolei przejął ją z modeli dla białek ustrukturyzowanych [53]. Chociaż  $\epsilon$  odpowiada energii kontaktu w tych modelach, nie musi być to odpowiednia wielkość dla potencjału PID. W przypadku kontaktów bb  $\epsilon^{bb}$  jest równe  $\epsilon$ , ponieważ wielkość ta odpowiada w przybliżeniu jednemu wiązaniu wodorowemu w łańcuchu głównym [178], jednak zróżnicowana natura kontaktów ss może wymagać reparametryzacji. Sprawdzone zostały wartości między 0 i 1 dla potencjału ze stałym  $\epsilon^{ss}$  (ten przypadek jest oznaczony

jako ME). Wypróbowane były też dwa warianty, w których  $\epsilon^{ss}$  zależy od aminokwasu: klasyczna macierz Miyazawa-Jernigana na podstawie statystyk PDB z roku 1996 [71] (oznaczona jako MJ) oraz macierz MDCG oparta na simulacjach pełnoatomowych [179] (oznaczona MD). Każda z tych macierzy także może być przeskalowana przez czynnik między 0 a 1 (oznaczony indeksem dolnym). We wszystkich trzech przypadkach (ME, MJ oraz MD)  $\epsilon^{ss} = 0$  dla proliny i glicyny.

Rozkłady odległości dla niektórych par aminokwasów są bardzo szerokie (Rys. 2.8, 2.9, 2.10). Wynika to z giętkości długich łańcuchów bocznych. Giętkość ta jest tracona w modelach z jedną kulką na aminokwas. Aby zrekompensować tę stratę, można wyobrazić sobie formę potencjału LJ zmodyfikowaną na potrzeby kontaktów ss:

$$\phi^{ss}(r) = \begin{cases} \epsilon^{ss} \left[ \left( \frac{r_{min}^{ss}}{r} \right)^{12} - 2 \left( \frac{r_{min}^{ss}}{r} \right)^6 \right] & \text{gdy } r > r_{min}^{ss} \\ -\epsilon^{ss} & \text{gdy } r_{min}^{ss} > r > r_{min}^{bb} \\ \epsilon^{ss} \left[ \left( \frac{r_{min}^{bb}}{r} \right)^{12} - 2 \left( \frac{r_{min}^{bb}}{r} \right)^6 \right] & \text{gdy } r_{min}^{bb} > r \end{cases} \quad (2.12)$$

Sprawdzona została zarówno ta zmodyfikowana forma (oznaczona literą F) oraz zwykła postać potencjału LJ (oznaczona literą L).

### 2.6.7 Oddziaływanie elektrostatyczne

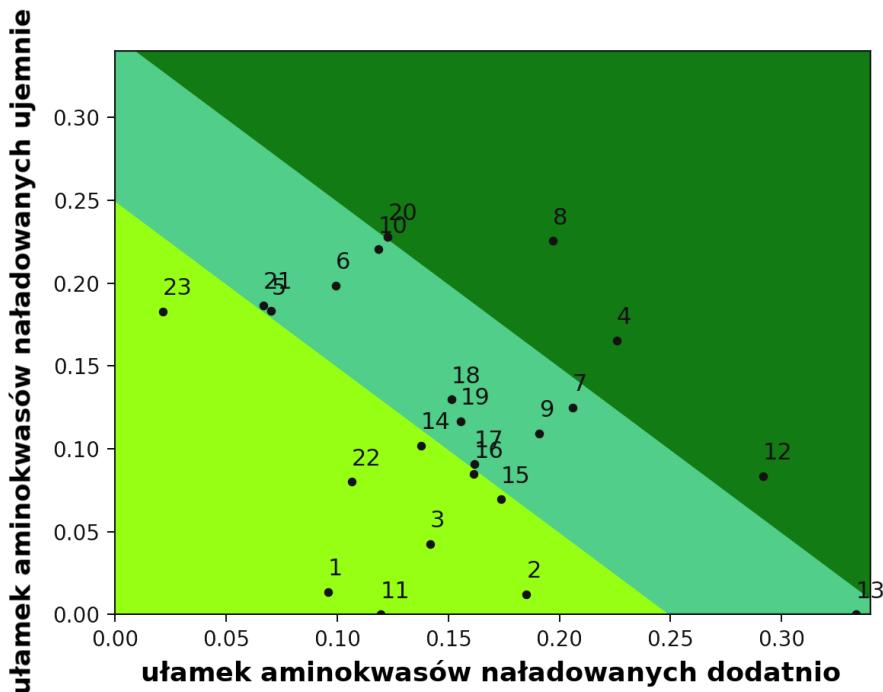
Model quasi-adiabatyczny wykorzystuje ekranowany potencjał Debye'a-Hueckle'a ze względną przenikalnością elektryczną zależną od odległości  $r$  między pseudoatomami:  $\varepsilon = 4 \text{ \AA}/r$ , zgodnie z podejściem Tozzini i in. [112] (dlatego ta wersja oddziaływań elektrostatycznych jest oznaczona literą T). To podejście zostało wypracowane dla białek uporządkowanych, gdzie przenikalność wewnętrz hydroforego rdzenia jest znaczco niższa niż w wodzie. Białka nieuporządkowane nie posiadają takiego rdzenia i są bardziej wystawione na oddziaływanie z rozpuszczalnikiem, dlatego wypróbowany został prostszy wariant ze stałą przenikalnością  $\varepsilon = 80$  (oznaczony literą C). W obu wariantach oddziaływanie elektrostatyczne opisuje potencjał:  $V_{D-H}(r) = \frac{q_1 q_2 \exp(-r/s)}{4\pi\varepsilon\varepsilon_0 r}$  gdzie  $s$  to długość ekranowania zależna od siły jonowej. Histydyna nadal ma przypisany ładunek 0.

### 2.6.8 Porównanie wariantów modelu z doświadczeniem

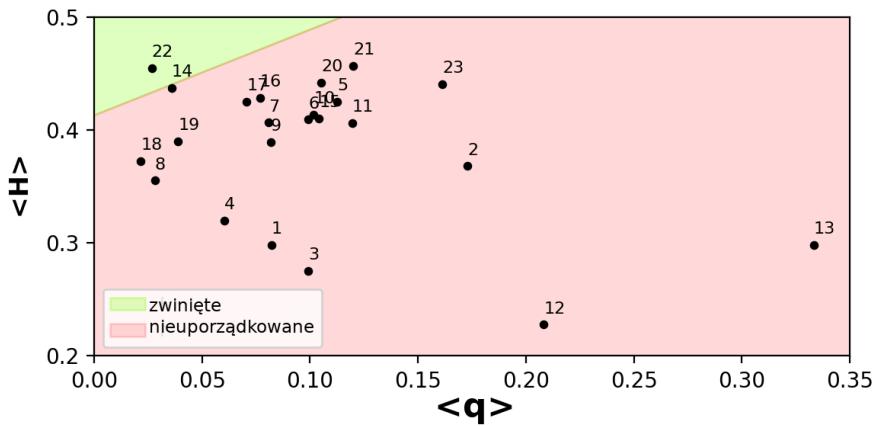
Ponad 200 wariantów modelu zostało przetestowanych na zestawie 23 białek nieuporządkowanych, których promień bezwładności został wyznaczony w doświadczeniach niskokątowego rozpraszania promieni X (SAXS, ang *Small Angle X-ray Scattering*) [16, 19, 23, 69, 165, 180, 181, 182, 183, 184, 185].

#### 2.6.8.1 Białka użyte do parametryzacji

Białka użyte do parametryzacji różnią się liczbą aminokwasów i zajmują różne obszary na diagramie Das-Pappu (Rys. 2.37) i na wykresie Uverskiego (Rys. 2.38) [186]. Obszary diagramu Das-Pappu dotyczące białek z wysoką zawartością aminokwasów naładowanych nie są pokazane (większość białek mieści się w obszarach zawartych na Rys. 2.37 [187]).



Rysunek 2.37: Diagram Das-Pappu dla 23 białek nieuporządkowanych użytych do parametryzacji, oznaczonych wg numeracji z tabeli 2.6. Modyfikacja rys. S9 z artykułu [II].



Rysunek 2.38: Wykres Uverskiego dla 23 białek nieuporządkowanych użytych do parametryzacji, oznaczonych wg numeracji z tabeli 2.6. Oś rzędnych używa skali hydropatyczności wg Kyte, Doolittle i in. [188], przeskalowanej do zakresu od 0 do 1. Modyfikacja rys. S10 z artykułu [II].

Użyta długość ekranowania (patrz podpunkt 2.6.7) może znaczco wpływać na wyniki, dlatego w symulacji każdego z białek została użyta długość  $s$  odpowiadająca sile jonowej z doświadczenia, w którym został wyznaczony promień bezwładności danego białka. Stała długość  $s = 10 \text{ \AA}$  skutkowała mniejszą zgodnością z doświadczeniem: najlepszym modelem okazał się wtedy idealny łańcuch Gaussowski z parametrem  $b = 6.7 \text{ \AA}$  (dane niepokazane, definicja parametru  $b$  jest podana pod koniec paragrafu 2.6.8.2).

nr	id	$n$	$R_g/\text{\AA}$	$s/\text{\AA}$	źródło
1	IB5	73	$27.9 \pm 1.0$	13.6	[23]
2	Ash1	81	$28.4 \pm 3.4$	7.9	[23]
3	II1ng	141	$41.1 \pm 1.0$	13.6	[23]
4	RNaseE	248	$52.6 \pm 0.3$	7.85	[23]
5	ACTR	71	$25.1 \pm 1.3$	6.8	[189]
6	NHE1	131	$36.3 \pm 1.8$	6.8	[189]
7	sNase	136	$21.2 \pm 1.0$	23.3	[189]
8	5AAA	142	$22.15 \pm 0.87$	4.3	[19]
9	6AAA	110	$28.1 \pm 0.1$	7.7	[19]
10	8AAC	59	$14.6 \pm 0.5$	13.6	[19]
11	9AAA	92	$29.9 \pm 0.3$	7.55	[19]
12	his5	24	$13.6 \pm 0.2$	7.85	[16]
13	RS	24	$12.62 \pm 0.07$	6.7	[190]
14	tauK10	168	$40.0 \pm 1.0$	7.4	[180, 185]
15	tauK17	144	$36.0 \pm 2.0$	7.4	[180, 185]
16	tauK18	130	$38.0 \pm 3.0$	7.4	[180, 185]
17	tauK19	99	$35.0 \pm 1.0$	7.4	[180, 185]
18	tauK25	185	$41.0 \pm 2.0$	7.4	[180, 185]
19	tauK44	283	$52.0 \pm 2.0$	7.4	[180, 185]
20	RNF4	57	$25.8 \pm 3.9$	8.5	[181, 185]
21	NRG1	75	$26.8 \pm 1.1$	7.4	[182, 185]
22	PIR	75	$26.5 \pm 0.5$	7.8	[183, 185]
23	p53	93	$28.7 \pm 0.3$	6.6	[184, 185]

Tabela 2.6: Właściwości 23 białek nieuporządkowanych użytych do parametryzacji: numer porządkowy, krótkie oznaczenie (id), liczba aminokwasów  $n$ , promień bezwładności  $R_g$ , długość ekranowania  $s$  oraz źródło z którego pochodzą te wartości.

IB5	SARSPPGKPQGPPQQEGNKPQGPPPGKPQGPPPAGGNPQQPQAPPAGKP QGPPPPPQGGRPPRPAQGQQPPQ
Ash1	GASASSSPSPSTPTKSGKMRSSSPVRPKAYTPSPRSPNYHRFALDSPPQS PRRSSNSSITKKGSRRSSGSSPTRHTRVCV
II1ng	ISGKPVGRRPQGGNQPQRPPPPGKPQGPPPQGGWQSQGPPPPGKPEGR PPQGRNQSQGPPPQHGKPERPPPQGSQGTTPPPGKPERPPPQGGNQSHR PPPPPGKPERPPPQGGNQSRGPPPQHGKPEGPPPQEGNKS
RNaseE	ERQQDRRKPRQNRRDRNERRDTRSERTEGSDNREENRNRRQAQQQTAE TRESRQQAEKARIADTADEQQAPRTERSRRNDDKRQAQQEAALKNEEQ SVQETEQEERVRPVQPRRKQRQLNKVRYEQSVAEEAVVAPVVEETVAAE PIVQEAPAPRTTELVKVPLPVVAQTAPEQQEENNADNRDNGGMRRSRRSP RHLRVSGQRRLRYRDCRYPIQSPMPLTVACASPELASGVWIRVPIVR
ACTR	GTQNRPLLRSNLDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEI DRALGIPELVNQGQALEPKQD
NHE1	MVPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVN EELKGKVGLSLRDPAKVAEEDEDGGIMMRSKETSSPGTDDVFTPAPSD SPSSQRIQRCLSDPGPHPEPGEPEFFFPGQ
sNase	ATSTKKLHKEPATLIKAIDGDTVKLMLYKGQPMTRFLLVDTPETKHPKG VEKYGPEASAFKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKVN EALVRQGLAKVAYVYKPNNTHEQHLRKSEAQAKKEK
5AAA	MDYKDDDDKNRALSPMVSEFETIEQENSYNEWLRAKVATSLADPRPAIPH DEVERRMAERFAKMRKERSKQMDYKDDDDKNRALSPMVSEFETIEQENSY NEWLRAKVATSLADPRPAIPHDEVERRMAERFAKMRKERSKQ
6AAA	VRTKADSVPGTYRKVVAARAPRKVLGSSTSATNSTVSSRKAENKYAGGN PVCVRPTPKWQKGIGEFRLSPKDSEKENQIPEEAGSSGLGAKRKACPL QPDHTNDEKE
8AAC	MEAIAKHDFSATADDELSFRKTQILKILNMEDDSNWTRELDGKEGLIPS NYIEMKNHD
9AAA	GSMTPSTPPRSRGTRYLAQPSGNTSSALMQGQKTPQKPSQNLVPVTPST TKSFKNAPLAPPNSNMGMTSPFNGLTSPQRSPFPKSSVKRT
his5	DShAKRHHGYKRKFHEKHHSHRGY
RS	GAMGPSYGRSRSRSRSRSRSRSRS
tauK10	QTAPVPMPLKNVSKIGSTENLKHQPGGGKVQIVYKPVDSLKVTSKCGS LGNIHHKPGGGQVEVKSEKLDKDRVQSKIGSLDNITHVPGGNKKIETH KLTFRENAAKTDHGAIEIVYKSPVVGDTSPRHLNSVSSTGSIDMVDSPQ LATLADEVASASLAKQGL
tauK17	SSPGSPGTPGSRSRTPSLPTPPTRPKVAVVRTPPKSPSSAKSRLQTAP VPMPDLKNVSKIGSTENLKHQPGGGKVQIVYKPVDSLKVTSKCGSLGNI HHKPGGGQVEVKSEKLDKDRVQSKIGSLDNITHVPGGNKKIE
tauK18	MQTAPVPMPLKNVSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCG SKDNIKHVPGGGSVQIVYKPVDSLKVTSKCGSLGNIHHKPGGGQVEVKSE KLDFKDRVQSKIGSLDNITHVPGGNKKIE
tauK19	MQTAPVPMPLKNVSKIGSTENLKHQPGGGKVQIVYKPVDSLKVTSKCG SLGNIHHKPGGGQVEVKSEKLDKDRVQSKIGSLDNITHVPGGNKKIE
tauK25	MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGI GDTPSLEDEAAGHVTQARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAA PPGQKGQANATRIPAKTPPAPKTPSSGEPPKGDRSGYSSPGSPGTPGS RSRTPSLPTPPTRPKVAVVRTPPKSPSSAKSRL
tauK44	MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGI GDTPSLEDEAAGHVTQARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAA PPGQKGQANATRIPAKTPPAPKTPSSGEPPKGDRSGYSSPGSPGTPGS RSRTPSLPTPPTRPKVAVVRTPPKSPSSAKSRL KIGSTENLKHQPGGGKVQIVYKPVDSLKVTSKCGSLGNIHHKPGGGQVE KSEKLDKDRVQSKIGSLDNITHVPGGNKKIE
RNF4	GHMGSWEAEPIELVETAGDEIVDLTCSLEPVVVDLTHNDSVVIVDERRR PRRNARR
NRG1	MEIYSPDMSEVAEAERSSSPSTQLSADPSLDGLPAAEDMPEPQTEDGRTPG LVGLAVPCCACLEAERLRGCLNSEK
PIR	QSVSPMRSVSENLSVAMDFSGQKTRVIDNPTEALSVAVEGLAWRKKGCL RLGNHGSPTAPSQSSAVNMALHRSQ
p53	MEEPQSDPSVEPPLSQETFSIDLWKLPENNVLSPLPSQAMDDMLSPDDI EQWFTEDEPGPDEAPRMPEAAPPVAPAPAAPTAAAPAPAPSPL

Tabela 2.7: Sekwencje 23 białek nieuporządkowanych użytych do parametryzacji.

### 2.6.8.2 Modele użyte do porównania

Każde z 23 białek było symulowane przez ponad 200 wariantów modelu gruboziarnistego: z potencjałem PID (litera P) albo z potencjałem quasi-adiabatycznym (litera A), z możliwością tworzenia przyciągających kontaktów  $i, i+4$  (indeks górnny  $+$ ) bądź bez tej możliwości (indeks  $-$ ), ze standardową (litera L) bądź spłaszczoną (litera F) formą potencjału LJ, z różnymi głębokościami tego potencjału, określonymi jedną z 3 macierzy: ME (stałe), MJ (Miyazawa-Jernigan) oraz MD (symulacje pełnoatomowe). Każda z tych macierzy była przeskalowana przez czynnik w indeksie dolnym. Przenikalność dielektryczna mogła być stała (litera C) bądź zależna od odległości (litera T).

Oznaczenie każdego wariantu składa się zatem z 4 symboli (pełna legenda podana jest w tabeli 2.8), np.  $A^+ L ME_1 T$  oznacza model z potencjałem quasi-adiabatycznym, możliwością tworzenia przyciągających kontaktów  $i, i+4$ , zwykłym potencjałem LJ ze stałym  $\epsilon^{LJ} = \epsilon$  oraz z przenikalnością elektryczną wg Tozzini i in. [112] (a zatem  $A^+ L ME_1 T$  oznacza model opisany w podrozdziale 2.5).

$P^- F MD_{0.1} C$  oznacza model z potencjałem PID, brakiem możliwości tworzenia kontaktów  $i, i+4$ , spłaszczonym potencjałem LJ z głębokością określoną przez macierz MDCG przeskalowaną przez czynnik 0.1 oraz stałą przenikalnością elektryczną równą 80.

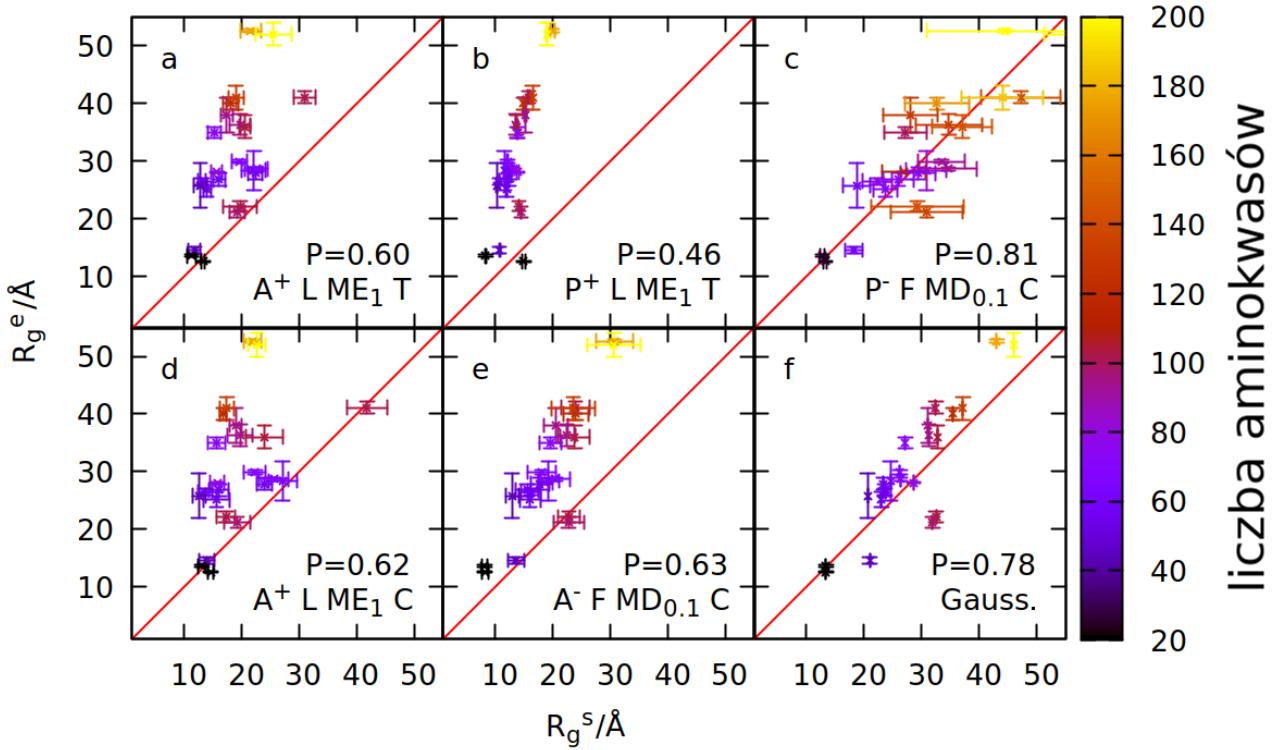
Dla każdego modelu obliczony został jego współczynnik Pearsona [53], zdefiniowany jako:

$$P = 1 - \sqrt{\frac{1}{N} \sum_{p=1}^N \left( \frac{R_g^{exp} - R_g^{sim}}{R_g^{exp}} \right)_p^2} \quad (2.13)$$

gdzie  $R_g^{exp}$  to promień bezwładności z eksperymentu,  $R_g^{sim}$  to promień z symulacji, a suma dotyczy każdego z  $N = 23$  białek. Tabela 2.9 zawiera pełną listę modeli z ich współczynnikami Pearsona (a także współczynnikami  $\chi^2$ ).

Rys. 2.39 pokazuje tylko 5 wybranych wariantów, w tym oryginalny model quasi-adiabatyczny (panel a). Zastosowanie potencjału PID zamiast potencjału quasi-adiabatycznego bez zmiany innych parametrów modelu skutkowało pogorszeniem zgodności z doświadczeniem (panel b), dlatego 5 opisanych powyżej parametrów modelu zostało zreparametryzowanych. Najlepszy z modeli po reparametryzacji (panel c) zgadza się z doświadczeniem dużo lepiej niż wyjściowy model quasi-adiabatyczny (panel a).

Zwiększenie zgodności mogło być skutkiem samej zmiany parametrów, a nie wprowadzenia nowego potencjału, dlatego te same parametry zostały zastosowane do modelu quasi-adiabatycznego (paneły d i e). Poprawiły one jednak jego zgodność jedynie nieznacznie. Dlatego potencjał PID faktycznie okazał się lepszy od quasi-adiabatycznego przełączania kontaktów. Panel f pokazuje wynik dla modelu łańcucha Gaussowskiego [191], który dany jest wzorem  $R_g^2 = \frac{1}{6}nb^2$ , gdzie  $n$  to liczba aminokwasów, a  $b$  to efektywna długość Kuhna, jedyny parametr tego modelu. Wartość wyznaczona przez dopasowanie metodą najmniejszych kwadratów do zestawu 23 białek nieuporządkowanych wynosi  $b = 6.7 \text{ \AA}$ .



Rysunek 2.39: Porównanie promienia bezwładności wyznaczonego w symulacji i w eksperymencie dla 23 białek nieuporządkowanych. Każdy panel dotyczy jednego z 6 modeli. Oznaczenia modeli określa tabela 2.8. Oznaczone są także współczynniki Pearsona. Liczba aminokwasów w białku została oznaczona skalą koloru. Modyfikacja rys. 4 z artykułu [II].

Dla każdego białka i każdego wariantu modelu wykonano 20 niezależnych symulacji. Dla modelu quasi-adiabatycznego (litera A) całkowity czas symulacji wynosił 150 000  $\tau$ , z czego połowę stanowił czas dochodzenia do równowagi ( $R_g$  było uśredniane na podstawie drugiej połowy symulacji). Dla modelu PID (litery P oraz W) obydwa czasy były 10 razy mniejsze (zostały wyznaczone w oparciu o czas, po jakim wartość  $R_g$  przestaje zależeć od warunków początkowych). Ta różnica może wskazywać na to, że rzeczywista skala czasowa modelu PID jest dłuższa, dlatego efektywność modelu PID na Rys. 2.42 może być niedoszacowana.

### 2.6.8.3 Szczegółowy ranking modeli

Tabela 2.9 zawiera wszystkie modele posortowane wg ich współczynnika Pearsona. Wyznaczony został także współczynnik  $\chi^2 = \frac{1}{N} \sum_{p=1}^N \left( \frac{(R_g^{exp} - R_g^{sim})^2}{\sigma_{sim}^2 + \sigma_{exp}^2} \right)_p$ , gdzie  $\sigma_{sim}$  to błąd średniej z wyników symulacji określony wg metody próbkowania przy użyciu podzbiorów (ang. *jackknife resampling*), natomiast  $\sigma_{exp}$  to niepewność eksperymentu wg tabeli 2.6. Niepewności są dużo mniejsze od różnicy  $R_g^{exp} - R_g^{sim}$ , dlatego wartości  $\chi^2$  są bardzo duże (te większe od 1000 zostały oznaczone jako “>1000”).

Dokładna liczba sprawdzonych wariantów modelu to 246. Wszystkie symbole użyte w ich oznaczeniach definiuje tabela 2.8. Niektóre z nich nie zostały zdefiniowane w paragrafie 2.6.8.2:

wariant oznaczony literą W oznacza model PID oparty na potencjale statystycznym wg wszystkich kontaktów (zielone histogramy na Rys. 2.28), a nie tylko tych spełniających kryteria kątowe.

Ponieważ białka nieuporządkowane nie mają hydrofobowego rdzenia i są w dużej mierze otoczone wodą, warto było rozważyć potencjał LJ z górką, która miała odtwarzać barierę potencjału wynikającą z otoczki hydratacyjnej. Potencjał ten oznaczony jest indeksem dolnym  $B$ .

W aminokwasach naładowanych ładunek znajduje się zwykle blisko końca łańcucha bocznego, dlatego sprawdzony został także wariant (oznaczony literą D), gdzie potencjał PID dotyczy także oddziaływań elektrostatycznych.

Najlepsze dopasowanie modelu łańcucha Gaussowskiego do danych eksperymentalnych odpowiada długości Kuhna  $b = 6.7 \text{ \AA}$ , jednak inne wartości  $b$  też zostały sprawdzone.

P	potencjał PID (dopasowanie do czerwonych histogramów z górnej części Rys. 2.28)
A	potencjał quasi-adiabatyczny
W	potencjał PID (dopasowanie do zielonych histogramów z górnej części Rys. 2.28)
indeks +	możliwość tworzenia przyciągających kontaktów $i, i + 4$
indeks -	brak możliwości tworzenia przyciągających kontaktów $i, i + 4$
L	zwykły potencjał Lenard-Jones'a
F	potencjał Lenard-Jones'a z płaskim regionem między $r_{min}^{bb}$ a $r_{min}^{ss}$ dla kontaktów ss
$L_B$	potencjał określony równaniem $\phi(r) = \epsilon^{LJ} \left[ \left( \frac{r_{min}}{r} \right)^{12} - \frac{9}{2^{1/6}} \left( \frac{r_{min}}{r} \right)^7 + \frac{13}{2} \left( \frac{r_{min}}{r} \right)^6 \right]$
$F_B$	Tak jak $L_B$ , ale z płaskim regionem między $r_{min}^{bb}$ a $r_{min}^{ss}$ dla kontaktów ss
ME	głębokość potencjału ss jest taka sama dla każdej pary aminokwasów (domyślnie $1 \text{ \epsilon}$ )
MJ	głębokość potencjału ss określa macierz Miyazawa-Jernigana [71]
MD	głębokość potencjału ss określa macierz MDCG [179]
indeks $_{0-1}$	czynnik skalujący macierz oddziaływania (od $_0$ do $_1$ )
C	elektrostatyka wg potencjału Debye'a-Hueckel'a z przenikalnością $\epsilon = 80$
T	elektrostatyka wg potencjału Debye'a-Hueckel'a z przenikalnością $\epsilon = 4 \text{ \AA}/r$ [173]
R	elektrostatyka jak w C, ale tylko dla aminokwasów jednoimiennie naładowanych (aminokwasy naładowane przeciwnie oddziałują tak jakby nie były naładowane)
D	potencjał elektrostatyczny także zależy od kątów PID: $\lambda_A(\psi_A)\lambda_B(\psi_B)V_{D-H}(r)$ , gdzie $V_{D-H}(r) = \frac{\pm e^2 \exp(-r/s)}{4\pi\epsilon\epsilon_0 r}$ (z przenikalnością $\epsilon = 80$ )
GC	łańcuch Gaussowski z długością Kuhna $b$

Tabela 2.8: Wytlumaczenie oznaczeń wszystkich wariantów modelu.

P <sup>-</sup> F MD <sub>0.1</sub> C	0.814	13.3	P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.726	40.1	P <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> C	0.682	720.9
P <sup>-</sup> F MD <sub>0.4</sub> C	0.813	14.0	A <sup>-</sup> L MD <sub>0.1</sub> C	0.725	99.8	W <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> C	0.681	138.5
P <sup>+</sup> F MD <sub>0.1</sub> C	0.812	13.7	A <sup>-</sup> L MD <sub>0.4</sub> C	0.725	99.5	W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.679	156.6
P <sup>-</sup> F ME <sub>0.0</sub> C	0.812	14.4	A <sup>-</sup> L <sub>B</sub> MD <sub>0.4</sub> C	0.725	99.5	P <sup>+</sup> F <sub>B</sub> MD <sub>1</sub> T	0.679	>1000
P <sup>+</sup> F MD <sub>0.2</sub> C	0.811	14.3	P <sup>+</sup> F MD <sub>0.1</sub> R	0.723	21.2	W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.677	147.3
P <sup>-</sup> F ME <sub>0.0</sub> T	0.809	11.6	W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.723	130.0	P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.677	32.9
P <sup>-</sup> F MD <sub>0.2</sub> C	0.809	13.9	A <sup>+</sup> L MD <sub>0.1</sub> C	0.721	137.0	P <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> C	0.675	32.2
P <sup>-</sup> F MD <sub>0.3</sub> R	0.808	29.6	A <sup>+</sup> L MD <sub>0.4</sub> C	0.721	136.7	W <sup>+</sup> F <sub>B</sub> MD <sub>0.3</sub> R	0.674	173.5
P <sup>+</sup> F MD <sub>0.4</sub> C	0.806	16.7	A <sup>+</sup> L <sub>B</sub> MD <sub>0.4</sub> C	0.721	136.7	A <sup>+</sup> F MD <sub>0.2</sub> R	0.673	197.3
P <sup>+</sup> F MD <sub>0.4</sub> D	0.803	13.6	A <sup>+</sup> L ME <sub>0.5</sub> R	0.721	80.7	A <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> R	0.673	197.3
P <sup>-</sup> F MD <sub>0.4</sub> D	0.802	10.7	P <sup>+</sup> F MD <sub>0.3</sub> C	0.720	113.1	A <sup>+</sup> F MD <sub>0.1</sub> R	0.673	197.5
P <sup>+</sup> F MD <sub>0.1</sub> D	0.799	13.0	P <sup>+</sup> F <sub>B</sub> MD <sub>0.3</sub> C	0.720	27.0	A <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.673	197.5
P <sup>+</sup> F MD <sub>0.2</sub> D	0.790	13.1	W <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> R	0.718	136.4	A <sup>+</sup> F MD <sub>0.4</sub> R	0.673	202.1
GC, $b = 6.7 \text{ \AA}$	0.788	89.1	P <sup>-</sup> F MD <sub>0.5</sub> R	0.718	386.9	A <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.673	202.1
P <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.788	17.9	P <sup>-</sup> F MD <sub>0.1</sub> R	0.718	386.9	P <sup>+</sup> F <sub>B</sub> MD <sub>1</sub> C	0.673	>1000
P <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.788	17.9	P <sup>-</sup> F <sub>B</sub> MJ <sub>0.1</sub> R	0.716	24.2	W <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> D	0.673	125.2
P <sup>-</sup> F <sub>B</sub> MJ <sub>0.1</sub> C	0.786	17.3	P <sup>+</sup> F <sub>B</sub> MD <sub>1</sub> R	0.713	525.2	P <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> D	0.672	29.4
P <sup>-</sup> F ME <sub>0.0</sub> D	0.786	13.6	P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.713	29.0	P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.672	27.7
P <sup>-</sup> F MD <sub>0.1</sub> D	0.786	13.4	P <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.712	24.5	W <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> D	0.670	146.0
P <sup>-</sup> F <sub>B</sub> ME <sub>0.1</sub> C	0.785	61.6	P <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.712	24.5	W <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.669	123.0
P <sup>-</sup> F MD <sub>0.2</sub> D	0.783	12.5	P <sup>-</sup> F MD <sub>0.1</sub> R	0.707	25.2	P <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.669	34.0
P <sup>+</sup> F MD <sub>0.3</sub> R	0.783	43.1	P <sup>-</sup> F MJ <sub>0.1</sub> R	0.707	333.2	W <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> D	0.668	127.7
A <sup>-</sup> L MD <sub>0.4</sub> R	0.765	62.8	P <sup>-</sup> F ME <sub>0.0</sub> R	0.704	27.6	P <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.666	34.4
A <sup>-</sup> L MD <sub>0.1</sub> R	0.764	62.5	P <sup>-</sup> F MD <sub>0.2</sub> R	0.702	25.9	P <sup>-</sup> F ME <sub>0.1</sub> R	0.666	677.3
A <sup>+</sup> L MD <sub>0.4</sub> R	0.752	94.2	GC, $b = 5.2 \text{ \AA}$	0.700	513.4	P <sup>+</sup> F <sub>B</sub> MD <sub>0.3</sub> R	0.665	35.1
A <sup>+</sup> L <sub>B</sub> MD <sub>0.4</sub> R	0.752	95.0	W <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.698	144.3	P <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> D	0.663	29.8
A <sup>+</sup> L MD <sub>0.1</sub> R	0.751	95.3	P <sup>+</sup> F MD <sub>0.5</sub> R	0.698	211.5	P <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.663	33.8
P <sup>-</sup> F <sub>B</sub> ME <sub>0.1</sub> R	0.744	19.8	P <sup>+</sup> F MD <sub>0.1</sub> R	0.698	211.5	P <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.659	35.9
W <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.744	97.0	W <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.698	118.2	P <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> C	0.658	35.1
P <sup>-</sup> F MD <sub>0.3</sub> C	0.738	214.6	P <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.692	34.1	P <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> R	0.656	36.9
P <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> C	0.733	24.1	P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.692	34.1	P <sup>-</sup> F MD <sub>0.5</sub> C	0.655	629.9
W <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> R	0.730	101.2	W <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> R	0.688	127.8	P <sup>-</sup> F MD <sub>0.1</sub> C	0.655	629.9
W <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.730	79.6	P <sup>+</sup> F <sub>B</sub> MD <sub>1</sub> D	0.688	689.2	P <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> D	0.653	34.1
P <sup>-</sup> F MD <sub>0.4</sub> R	0.729	21.2	P <sup>+</sup> F MJ <sub>0.1</sub> R	0.687	254.1	P <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> D	0.651	35.8
P <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.728	483.6	W <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.686	99.0	A <sup>-</sup> F MD <sub>0.1</sub> R	0.650	167.4
P <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> R	0.728	483.6	W <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> D	0.686	148.9	A <sup>-</sup> F MD <sub>0.4</sub> R	0.649	167.3
P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> T	0.727	29.9	A <sup>+</sup> L ME <sub>0.5</sub> C	0.685	100.2	A <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.649	167.3
P <sup>+</sup> F MD <sub>0.4</sub> R	0.727	22.4	A <sup>+</sup> L <sub>B</sub> MD <sub>0.2</sub> C	0.685	100.2	A <sup>+</sup> F ME <sub>0.5</sub> R	0.647	158.5
P <sup>+</sup> F MD <sub>0.2</sub> R	0.727	22.3	A <sup>+</sup> L <sub>B</sub> MD <sub>0.1</sub> C	0.685	100.2	P <sup>-</sup> F MJ <sub>0.1</sub> C	0.643	527.3
W <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.726	127.2	W <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> C	0.682	156.9	A <sup>+</sup> F MD <sub>0.2</sub> C	0.643	247.4
P <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.726	40.1	P <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.682	720.9	A <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> C	0.643	247.4

Tabela 2.9: Współczynniki Pearsona oraz  $\chi^2$  dla wszystkich wariantów modelu, część 1.

W <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.643	179.7	P <sup>-</sup> F <sub>B</sub> MJ <sub>0.3</sub> C	0.558	>1000	W <sup>+</sup> F MD <sub>0.3</sub> C	0.458	712.7
A <sup>+</sup> F MD <sub>0.1</sub> C	0.642	247.5	P <sup>-</sup> F MD <sub>0.5</sub> R	0.534	>1000	W <sup>-</sup> F MD <sub>0.3</sub> C	0.455	743.2
A <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.642	247.5	P <sup>-</sup> F MD <sub>0.3</sub> R	0.534	>1000	W <sup>+</sup> F MD <sub>0.5</sub> R	0.450	901.0
P <sup>+</sup> F MD <sub>0.1</sub> D	0.642	603.0	W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.531	933.3	W <sup>+</sup> F MD <sub>0.1</sub> R	0.450	901.0
A <sup>+</sup> F MD <sub>0.4</sub> C	0.642	252.2	GC, $b = 3.8 \text{ \AA}$	0.531	>1000	P <sup>-</sup> F MJ <sub>0.3</sub> R	0.450	>1000
A <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.642	252.2	P <sup>-</sup> F <sub>B</sub> ME <sub>0.3</sub> R	0.527	>1000	W <sup>-</sup> F MD <sub>0.5</sub> R	0.448	638.4
P <sup>+</sup> F MD <sub>0.1</sub> T	0.641	682.1	P <sup>-</sup> F MD <sub>0.5</sub> C	0.512	>1000	W <sup>+</sup> F MD <sub>0.1</sub> D	0.444	961.5
P <sup>+</sup> F MD <sub>0.5</sub> C	0.634	439.5	P <sup>-</sup> F MD <sub>0.3</sub> C	0.512	>1000	P <sup>-</sup> F MJ <sub>0.3</sub> C	0.444	>1000
P <sup>+</sup> F MD <sub>0.1</sub> C	0.634	439.5	P <sup>+</sup> F MD <sub>1</sub> R	0.511	>1000	W <sup>+</sup> F MD <sub>0.5</sub> C	0.443	970.3
A <sup>-</sup> F MD <sub>0.4</sub> C	0.634	211.5	W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> D	0.510	>1000	W <sup>+</sup> F MD <sub>0.1</sub> C	0.443	970.3
A <sup>-</sup> F MD <sub>0.1</sub> C	0.633	211.9	W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.504	>1000	P <sup>+</sup> F MJ <sub>0.3</sub> R	0.440	>1000
A <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> C	0.633	211.5	W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> T	0.500	>1000	W <sup>+</sup> F MD <sub>0.1</sub> T	0.439	873.8
P <sup>+</sup> F ME <sub>0.1</sub> R	0.632	639.0	P <sup>-</sup> F <sub>B</sub> MJ <sub>0.5</sub> R	0.499	>1000	W <sup>-</sup> F MD <sub>0.5</sub> C	0.438	830.8
W <sup>-</sup> F <sub>B</sub> MD <sub>0.3</sub> C	0.631	176.3	P <sup>-</sup> F <sub>B</sub> ME <sub>0.3</sub> C	0.498	>1000	P <sup>+</sup> F MJ <sub>0.3</sub> D	0.435	>1000
P <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.631	42.8	P <sup>+</sup> F MD <sub>1</sub> D	0.497	>1000	P <sup>+</sup> F MJ <sub>0.3</sub> C	0.434	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> R	0.630	230.0	P <sup>+</sup> F MD <sub>1</sub> C	0.492	>1000	P <sup>+</sup> F MJ <sub>0.3</sub> T	0.431	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.630	230.0	P <sup>+</sup> F MD <sub>0.2</sub> C	0.492	>1000	P <sup>-</sup> F <sub>B</sub> ME <sub>0.5</sub> R	0.427	>1000
P <sup>+</sup> F MJ <sub>0.1</sub> D	0.629	333.9	W <sup>+</sup> F MD <sub>0.1</sub> R	0.488	599.1	P <sup>-</sup> F ME <sub>0.3</sub> R	0.421	>1000
P <sup>+</sup> F <sub>B</sub> MD <sub>0.2</sub> R	0.628	41.8	P <sup>+</sup> F MD <sub>1</sub> T	0.486	>1000	P <sup>-</sup> F <sub>B</sub> ME <sub>0.5</sub> C	0.420	>1000
A <sup>+</sup> F ME <sub>1</sub> T	0.625	282.3	P <sup>+</sup> L ME <sub>1</sub> C	0.485	>1000	P <sup>-</sup> F ME <sub>0.3</sub> C	0.416	>1000
P <sup>+</sup> F MJ <sub>0.1</sub> T	0.623	407.8	W <sup>+</sup> F MD <sub>0.2</sub> R	0.485	548.7	W <sup>+</sup> F MD <sub>0.5</sub> R	0.414	>1000
P <sup>+</sup> F MJ <sub>0.1</sub> C	0.621	512.9	W <sup>-</sup> F MD <sub>0.1</sub> R	0.485	482.5	P <sup>-</sup> F MJ <sub>0.5</sub> R	0.413	>1000
P <sup>+</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.620	43.1	W <sup>-</sup> F MD <sub>0.4</sub> R	0.482	432.4	P <sup>+</sup> F ME <sub>0.3</sub> R	0.412	>1000
P <sup>-</sup> F <sub>B</sub> MD <sub>0.2</sub> R	0.619	43.2	W <sup>+</sup> F MD <sub>0.4</sub> R	0.481	531.5	P <sup>+</sup> F ME <sub>0.3</sub> D	0.412	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.3</sub> C	0.618	220.8	W <sup>-</sup> F MD <sub>0.2</sub> R	0.481	606.3	W <sup>+</sup> F MD <sub>0.5</sub> C	0.412	>1000
P <sup>-</sup> F <sub>B</sub> MD <sub>0.1</sub> R	0.618	45.7	P <sup>-</sup> F <sub>B</sub> MJ <sub>0.5</sub> C	0.480	>1000	P <sup>-</sup> F MJ <sub>0.5</sub> C	0.410	>1000
A <sup>+</sup> L ME <sub>1</sub> C	0.617	299.9	W <sup>+</sup> F MD <sub>0.2</sub> C	0.477	550.4	W <sup>+</sup> F MD <sub>0.5</sub> D	0.409	>1000
A <sup>+</sup> F ME <sub>0.5</sub> C	0.612	183.8	W <sup>+</sup> F MD <sub>0.1</sub> C	0.475	740.2	P <sup>+</sup> F ME <sub>0.3</sub> C	0.408	>1000
P <sup>-</sup> F <sub>B</sub> MD <sub>0.4</sub> R	0.610	44.7	W <sup>+</sup> F MD <sub>0.4</sub> C	0.474	678.3	W <sup>+</sup> F MD <sub>0.5</sub> T	0.408	>1000
P <sup>-</sup> F ME <sub>0.1</sub> C	0.602	>1000	W <sup>+</sup> F MD <sub>0.1</sub> D	0.472	602.1	P <sup>+</sup> F MJ <sub>0.5</sub> R	0.406	>1000
A <sup>+</sup> L ME <sub>1</sub> T	0.601	334.4	W <sup>+</sup> F MD <sub>0.4</sub> D	0.472	536.2	P <sup>+</sup> F ME <sub>0.3</sub> T	0.406	>1000
P <sup>+</sup> F ME <sub>0.1</sub> D	0.596	923.2	W <sup>-</sup> F MD <sub>0.2</sub> C	0.471	532.9	P <sup>+</sup> F MJ <sub>0.5</sub> D	0.403	>1000
W <sup>-</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.595	286.2	W <sup>-</sup> F MD <sub>0.1</sub> C	0.471	513.4	P <sup>+</sup> F MJ <sub>0.5</sub> C	0.402	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> C	0.586	399.6	W <sup>+</sup> F MD <sub>0.2</sub> D	0.471	726.9	P <sup>+</sup> F MJ <sub>0.5</sub> T	0.401	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> C	0.586	399.6	W <sup>-</sup> F MD <sub>0.4</sub> C	0.470	576.1	P <sup>-</sup> F ME <sub>0.5</sub> R	0.389	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> D	0.584	466.4	W <sup>-</sup> F MD <sub>0.4</sub> D	0.468	606.8	P <sup>-</sup> F ME <sub>0.5</sub> C	0.386	>1000
P <sup>+</sup> F ME <sub>0.1</sub> C	0.582	>1000	W <sup>-</sup> F MD <sub>0.2</sub> D	0.468	597.7	P <sup>+</sup> F ME <sub>0.5</sub> D	0.384	>1000
P <sup>-</sup> F <sub>B</sub> MJ <sub>0.3</sub> R	0.581	>1000	W <sup>+</sup> F MD <sub>0.3</sub> R	0.465	770.1	P <sup>+</sup> F ME <sub>0.5</sub> R	0.384	>1000
W <sup>+</sup> F <sub>B</sub> MD <sub>0.1</sub> T	0.579	380.2	W <sup>-</sup> F MD <sub>0.1</sub> D	0.465	536.0	GC, $b = 2.7 \text{ \AA}$	0.382	>1000
P <sup>+</sup> F ME <sub>0.1</sub> T	0.577	>1000	W <sup>-</sup> F MD <sub>0.3</sub> R	0.464	546.1	P <sup>+</sup> F ME <sub>0.5</sub> C	0.381	>1000
GC, $b = 4.1 \text{ \AA}$	0.570	>1000	P <sup>+</sup> L ME <sub>1</sub> T	0.459	>1000	P <sup>+</sup> F ME <sub>0.5</sub> T	0.381	>1000

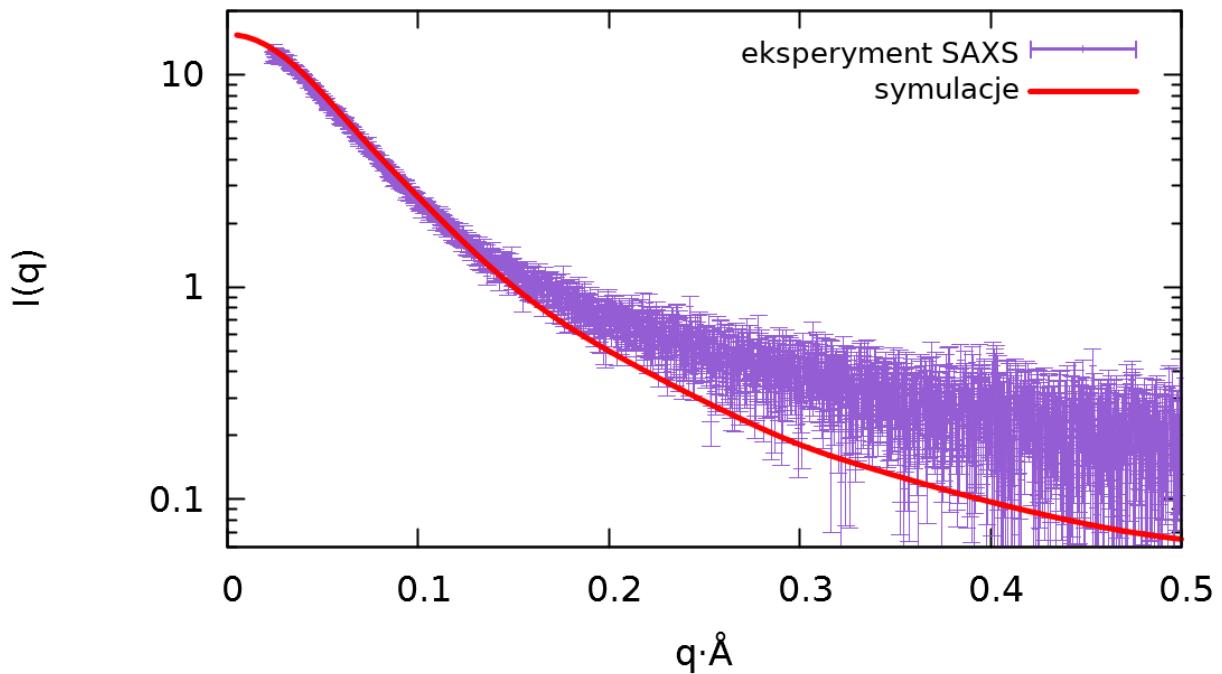
Tabela 2.9: Współczynniki Pearsona oraz  $\chi^2$  dla wszystkich wariantów modelu, część 2.

#### 2.6.8.4 Bezpośrednie porównanie z wynikami SAXS

Wszystkie wartości promieni bezwładności  $R_g$  dla zestawu 23 białek nieuporządkowanych zostały wyznaczone na podstawie eksperymentów SAXS. Bezpośrednim wynikiem takiego eksperymentu jest krzywa rozpraszenia, zawierająca informacje nie tylko o promieniu bezwładności, ale także o średnim kształcie danego białka. Odtworzenie takiej krzywej na podstawie zestawu struktur z symulacji jest możliwe [18], ale porównanie jej z doświadczeniem wymaga dopasowania pewnych parametrów i znajomości krzywej otrzymanej w doświadczeniu. Nie było to możliwe dla wielu z 23 wybranych białek.

Promień bezwładności  $R_g$  jest natomiast jedną liczbą, którą bardzo łatwo porównać. Z tych powodów model był reparametryzowany tylko na podstawie wartości  $R_g$ . Jednak porównanie z bezpośrednimi wynikami stanowi dodatkową weryfikację modelu.

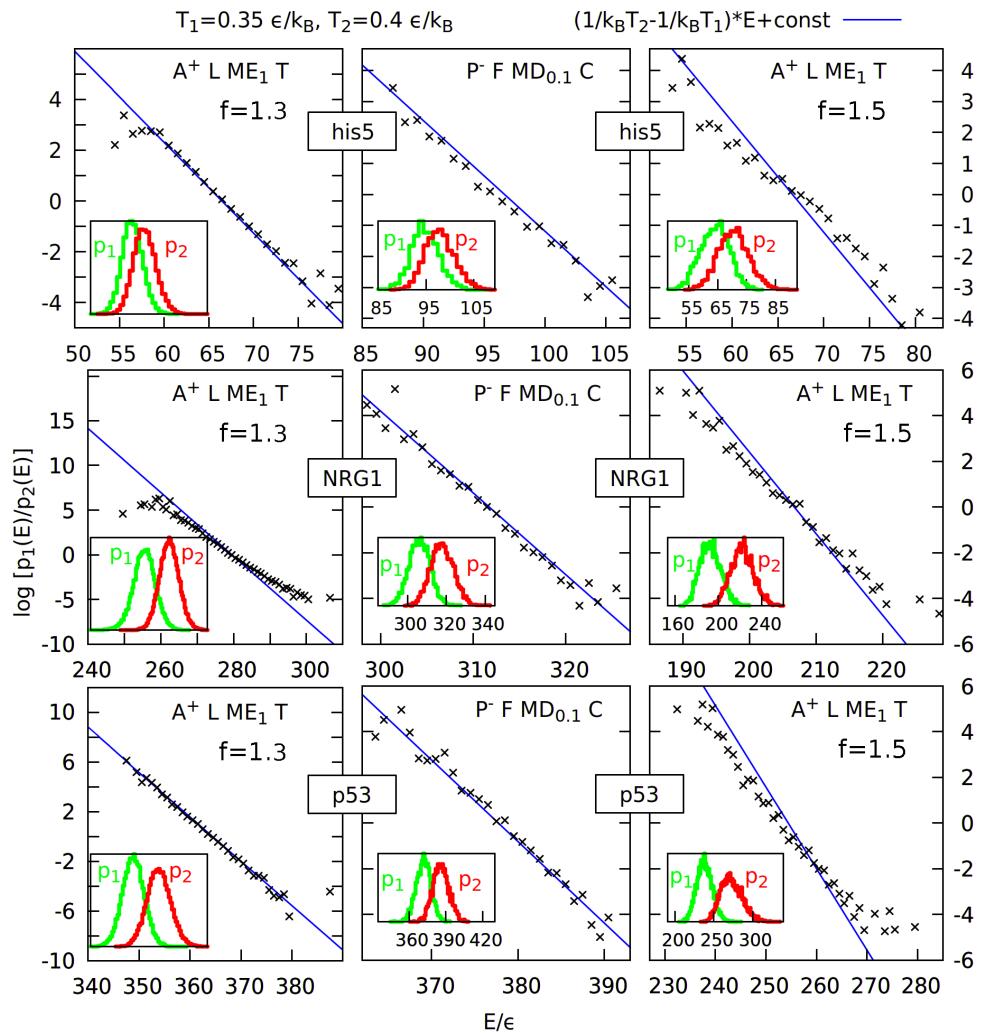
Rys. 2.40 jest przykładem takiego porównania między krzywą rozpraszenia SAXS dla białka 6AAA (czerwony) i wynikiem najlepszego modelu z tabeli 2.9 (fioletowy). Najlepsza zgodność występuje dla zakresu najwyższych kątów rozpraszenia, z wektorem rozproszenia  $q < 0.15 \text{ \AA}^{-1}$ . Zakres ten zawiera większość informacji o kształcie i rozmiarze białka.



Rysunek 2.40: Intensywność rozpraszenia SAXS  $I$  jako funkcja wektora rozpraszenia  $q$ , mierzona eksperymentalnie dla białka 6AAA [17] (fioletowy), porównana do wyników symulacji (czerwony). Symulacja używała wariantu P<sup>-</sup> F MD<sub>0.1</sub> C modelu gruboziarnistego. Krzywa rozpraszenia została obliczona dla struktur zapisywanych co 5000  $\tau$ , przy użyciu algorytmu który powstał wraz z metodą EROS [18]. Zostały użyte domyślne parametry otoczki hydratacyjnej dla powierzchni białka. Intensywność (w jednostkach umownych) została przeskalowana tak, aby zgadzać się z doświadczeniem. Modyfikacja rys. S11 z artykułu [II].

### 2.6.8.5 Test histogramów

Różne warianty modelu można porównać także testem histogramów [192], który sprawdza czy wyniki symulacji są niesprzeczne z rozkładem Boltzmanna: stan o energii  $E$  powinien pojawiać się z prawdopodobieństwem  $p(E) = \Omega(E) \exp(-E/k_B T)/Q$ , gdzie  $\Omega(E)$  to gęstość stanów, a  $Q$  jest czynnikiem normalizacyjnym. Jeśli przeprowadzi się symulacje używając dwóch różnych temperatur  $T_1$  i  $T_2$ , można obliczyć następującą wielkość:  $\log(p_1(E)/p_2(E)) = \log(Q_1/Q_2) + E \cdot (1/k_B T_2 - 1/k_B T_1)$ . Wielkość ta powinna liniowo zależeć od energii, a współczynnik kierunkowy prostej powinien wynosić  $(1/k_B T_2 - 1/k_B T_1)$ . Wyznaczono tę zależność (traktując wyraz wolny  $\log(Q_1/Q_2)$  jako parametr dopasowania) dla modelu quasi-adiabatycznego ( $A^+ L ME_1 T$ ) i dla najlepszego wariantu modelu PID ( $P^- F MD_{0.1} C$ ), dla białek his5 (24 aminokwasy), NRG1 (75 aminokwasów) oraz p53 (93 aminokwasy), patrz Rys. 2.41. Punkty dla modelu PID leżą bliżej prostej  $E(1/k_B T_2 - 1/k_B T_1) + \text{const}$  niż dla modelu z  $f = 1.5$ , ale dalej niż dla modelu z  $f = 1.3$  (czynnik  $f$  jest omawiany w podpunkcie 2.5.3).



Rysunek 2.41: Test histogramów [192] dla najlepszego wariantu modelu PID ( $P^- F MD_{0.1} C$ , pośrodku) oraz dla modelu quasi-adiabatycznego ( $A^+ L ME_1 T$ ) z czynnikiem  $f = 1.3$  (po lewej) oraz  $f = 1.5$  (po prawej), na podstawie symulacji białek his5, NRG1 oraz p53, dla temperatur  $0.35$  oraz  $0.4 \epsilon/k_B$ . Małe panele pokazują histogramy energii (szerokość słupka to  $1 \epsilon$ ), użyte do wyznaczenia wielkości  $\log(p_1(E)/p_2(E))$  pokazanej na dużych panelach w funkcji energii. Modyfikacja rys. 5 z artykułu [II].

## 2.7 Podsumowanie

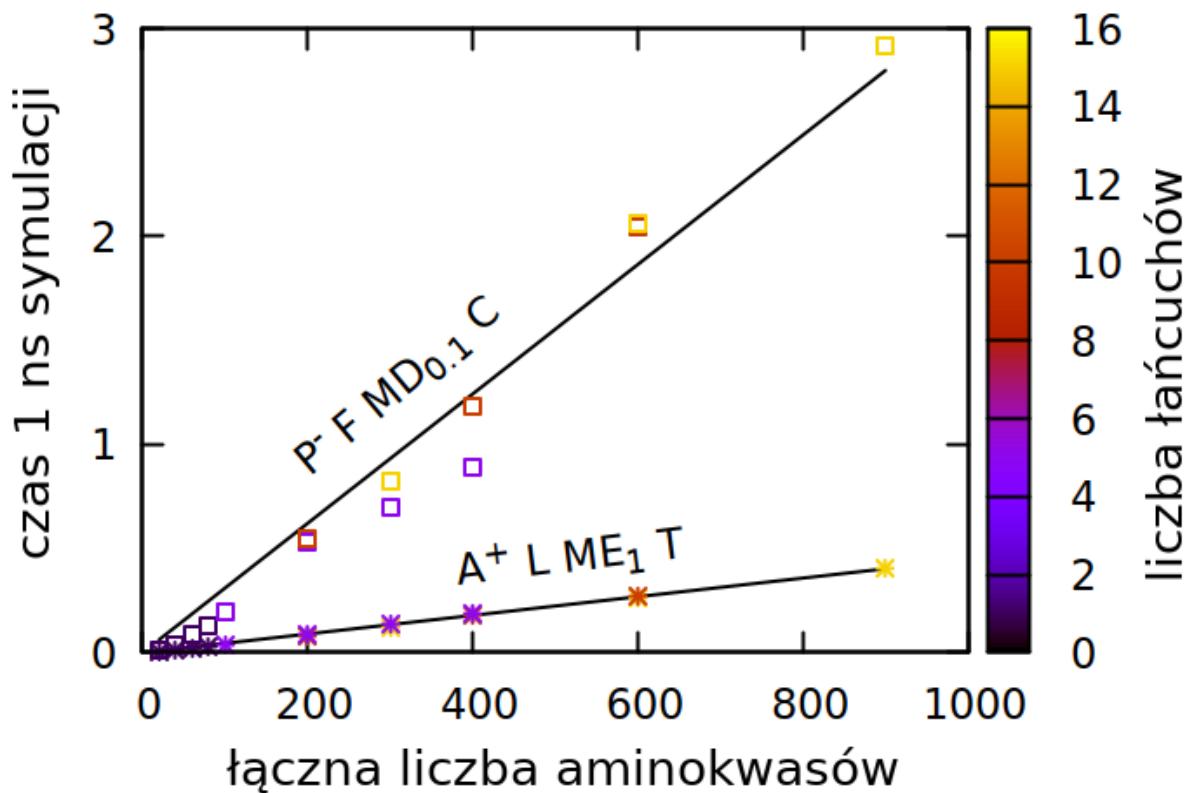
Przedstawiony został model oparty na dynamice molekularnej pseudoatomów, których położenia odpowiadają atomom  $C_\alpha$  białek nieuporządkowanych. Każdy wariant tego modelu umożliwia dostęp do dużo dłuższych skali czasowych niż symulacje pełnoatomowe. Mimo użycia tylko jednego pseudoatomu na aminokwas, nawet pierwotny wariant ( $A^+ L ME_1 T$ ) jest zgodny z wieloma danymi pochodzącyymi z eksperymentu i z symulacji pełnoatomowych. Potwierdza to słuszność założenia, że można odtworzyć właściwości geometryczne łańcucha, korzystając jedynie z położen atomów  $C_\alpha$ , a następnie wykorzystać je do dynamiki molekularnej.

W przypadku modelu quasi-adiabatycznego na podstawie tych położen tworzona jest dynamiczna mapa kontaktów, które mogą znikać i pojawiać się ponownie. Kontakty mogą wynikać z oddziaływań łańcucha bocznego z bocznym, bocznego z głównym bądź głównego z głównym, a oprócz nich uwzględnione są także oddziaływanie elektrostatyczne [112]. Model ten zgadza się nie tylko ze średnimi wielkościami dotyczącymi całego łańcucha, ale także z ich rozkładami otrzymanymi na podstawie symulacji pełnoatomowych. Nie jest to zatem tylko przypadkowa zbieżność, lecz faktyczne odwzorowanie kształtu i dynamiki białka. Zaletą modelu quasi-adiabatycznego jest także duża szybkość, ponieważ dynamika obejmuje tylko oddziaływanie dwuciałowe (efekty wielociałowe pojawiają się tylko przy aktualizowaniu chwilowej mapy kontaktów).

Podczas parametryzacji 246 wariantów modelu, niektóre warianty korzystały z potencjału quasi-adiabatycznego. Miały one mniejszy współczynnik Pearsona niż najlepsze warianty modelu PID, należy jednak pamiętać że wszystkie parametryzowane warianty korzystały z czynnika  $f = 1.5$  (patrz podpunkt 2.5.3). Zmiana  $f$  na 1.3 może znacznie poprawić efektywność tego modelu (patrz paragraf 2.6.8.5 i podpunkt 4.2.1), niestety czynnik ten został wzięty pod uwagę już po wykonaniu wszystkich symulacji, wymagających milionów godzin obliczeń. Kolejna reparametryzacja uwzględniająca różne  $f$  może przynieść interesujące rezultaty.

Być może po kolejnej reparametryzacji lepiej modelowane będą białka częściowo nieuporządkowane. Pewna ich kategoria staje się ustrukturyzowana po związaniu z substratem [193]. Omawiany tu model nie jest na tyle dokładny, aby przewidzieć strukturę przyjętą przez białko w takiej sytuacji (nie potrafi np. przewidzieć struktury białka PUMA po związaniu z kompleksem Mcl-1 [194]). Ani potencjał PID, ani potencjał quasi-adiabatyczny nie nadają się też do zwijania białek uporządkowanych, a także do utrzymywania ich w konformacji natywnej (patrz podrozdział 3.5). Dla białek, które w całości lub w części są ustrukturyzowane, można jednak ustalić mapę kontaktów opartą na strukturze natywnej i prowadzić dzięki niej efektywne symulacje zwijania **[VI]**.

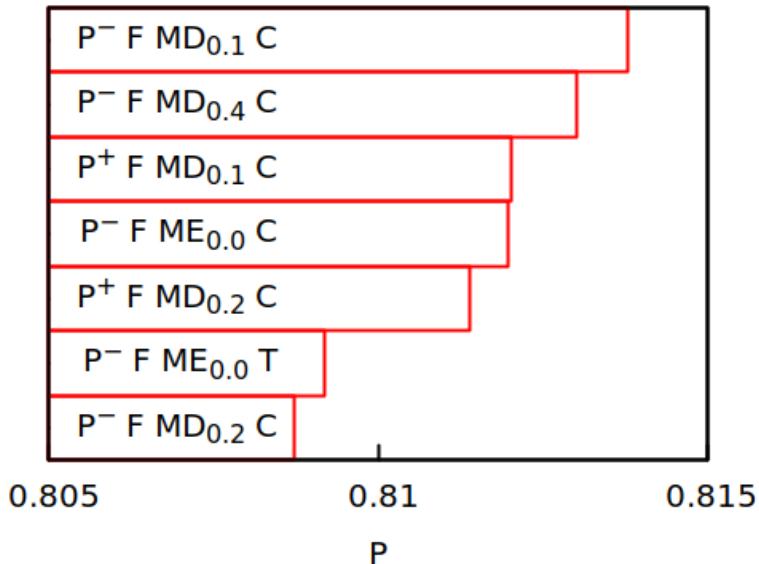
Mimo że szybkość symulacji skala się prawie liniowo wraz z rozmiarem układu, model PID jest co najmniej 5 razy wolniejszy (Rys. Fig. 2.42), przy założeniu że we wszystkich modelach jednostka  $\tau$  odpowiada 1 nanosekundzie. Dużo mniejszy czas potrzebny na dojście do stanu równowagowego w modelu PID (patrz koniec paragrafu 2.6.8.2) wskazuje na konieczność powtórzenia dla tego modelu procedury użytej do wyznaczenia skali czasowej symulacji [132]. Model PID trudniej także wykorzystać do symulacji zrównoległych na wielu rdzeniach (ze względu na człyne wielociałowe).



Rysunek 2.42: Rzeczywisty czas jednej nanosekundy symulacji na 1 rdzeniu dla modelu PID ( $P^- F MD_{0.1} C$ , kwadraty) i quasi-adiabatycznego ( $A^+ L ME_1 T$ , gwiazdki), w funkcji liczby symulowanych aminokwasów. Liczba łańcuchów w symulacji jest zaznaczona kolorem. Gęstość układu dla symulacji z wieloma łańcuchami wynosiła  $1 \text{ aminokwas/nm}^3$  (aa/nm<sup>3</sup>). Dopasowane linie mają współczynniki odpowiednio  $0.003 \text{ s/aminokwas}$  i  $0.0005 \text{ s/aminokwas}$ , z błędem dopasowania rzędu  $0.0001 \text{ s/aminokwas}$ . Modyfikacja rys. 6 z artykułu [II].

W najlepszych 7 wariantach modelu z tabeli 2.9 macierz oddziaływań jest przeskalowana przez czynnik mniejszy od  $1/2$  (Rys. 2.43). Warto zauważyć że dwa spośród siedmiu najlepszych wariantów mnożą macierz przez 0, co oznacza, że aminokwasy nie oddziałują ze sobą, z wyjątkiem sztywności łańcucha, oddziaływań elektrostatycznych i odpychania na małych odległościach (zapewniającego nieprzecinanie się łańcucha). Bardzo prosty model łańcucha Gaussowskiego także sprawdza się całkiem dobrze (panel f na Rys. 2.39).

Jest to dobrym potwierdzeniem tego, że woda faktycznie jest dobrym rozpuszczalnikiem dla białek nieuporządkowanych, a oddziaływanie aminokwasów odległych od siebie w sekwencji wpływają na wymiary łańcucha tylko w niewielkim stopniu [1]. Fakt ten został niedawno użyty w hierarchicznym podejściu do symulowania białek nieuporządkowanych, gdzie tylko niewielkie fragmenty łańcucha są szczegółowo modelowane, a cały łańcuch jest konstruowany na podstawie wyników modelowania tych fragmentów [197].



Rysunek 2.43: Najlepsze 7 wariantów modelu PID (wg ich współczynników Pearsona). Oparte na rys. 7 z artykułu [II].

Ze względu na większą szybkość działania (i wątpliwości co do skali czasowej modelu PID), w pozostałych rozdziałach pracy to model quasi-adiabatyczny ( $A^+ L ME_1 T$ ) będzie używany do badania dynamiki konformacyjnej białek nieuporządkowanych i ich agregacji (z wyjątkiem podrozdziału 3.5). Model quasi-adiabatyczny może być łatwo połączony z modelem dla białek ustrukturyzowanych, ponieważ oba opierają się na mapie kontaktów. Takie połączenie może być użyte choćby do badania giętkich łączników w białkach wielodomenowych [97, 198]. Każda z domen mogłaby być utrzymywana przez mapę kontaktów [53], a każdy z łączników byłby opisany przez potencjał quasi-adiabatyczny<sup>2</sup>. Takie łączone podejście zostało zastosowane w rozdziale 5 dotyczącym symulacji białek z ziaren pszenicy, kukurydzy i ryżu.

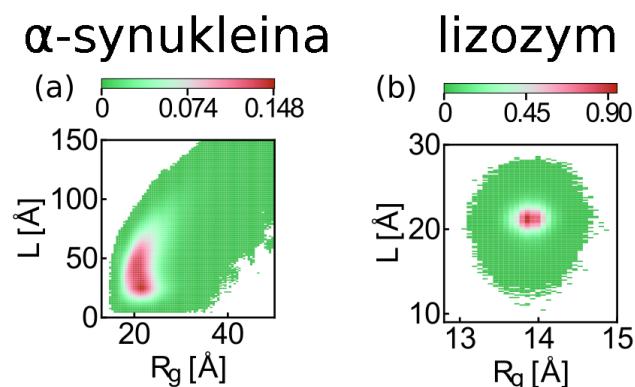
<sup>2</sup>bądź potencjał PID

# Rozdział 3

## Wyniki symulacji pojedynczych łańcuchów

### 3.1 Wprowadzenie

Białka nieuporządkowane odwiedzają dużo większy obszar przestrzeni konformacji niż uporządkowane [III]. Jednym ze sposobów przedstawiania tej przestrzeni jest podawanie dwuwymiarowych rozkładów promienia bezwładności  $R_g$  oraz odległości między końcami  $l$  (oznaczanej także jako  $L$ ). Takie rozkłady wykonane na podstawie symulacji modelem gruboziarnistym dla nieuporządkowanej  $\alpha$ -synukleiny (presentowanym tu modelem dla białek nieuporządkowanych) oraz uporządkowanego lizozymu (modelem na podstawie struktury natywnej [53]) zostały przedstawione na Rys. 3.1. Widać na nim, że dla białka uporządkowanego rozkład jest skoncentrowany wokół jednego punktu (struktury natywnej), a dla  $\alpha$ -synukleiny odwiedzane są różne konformacje. Warto zwrócić uwagę, że skala wykresów jest całkiem inna (mimo że oba białka mają około 140 aminokwasów). Rozkłady  $R_g, l$  oraz prawdopodobieństwa przejścia między stanami w zdyskretyzowanym rozkładzie  $R_g, l$  wyglądają podobnie dla dużego zakresu skal czasowych symulacji (o ile symulacja jest prowadzona w równowadze) [III]. Oznacza to, że równowagowe konformacje charakteryzują się podobnym rozkładem  $R_g, l$  niezależnie od czasu życia.



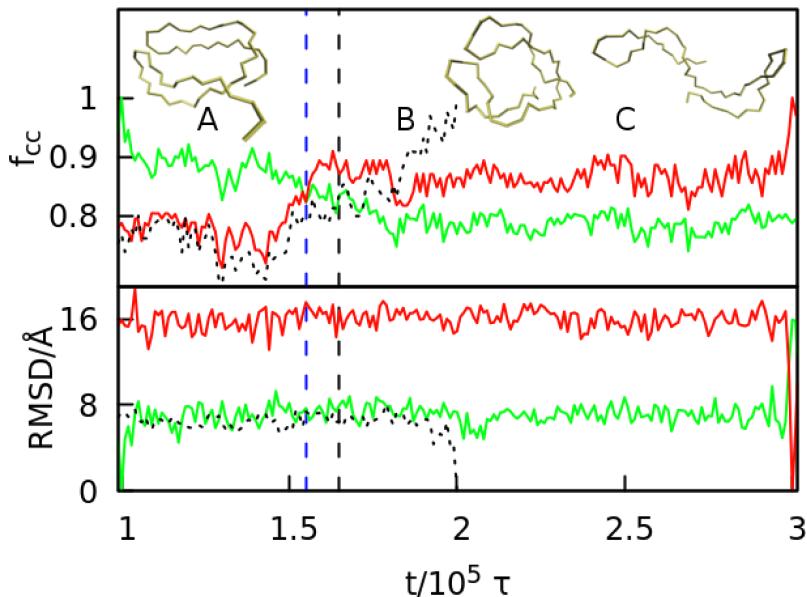
Rysunek 3.1: Dwuwymiarowy rozkład  $R_g$  oraz  $l = L$  dla  $\alpha$ -synukleiny oraz lizozymu. Zmodyfikowany fragment rys. 6 z artykułu [III].

### 3.2 Badanie dynamiki konformacyjnej homopeptydów

Ponieważ białka nieuporządkowane zmieniają dynamicznie konformacje z jednej na drugą, interesujące jest pytanie jak długo pozostają w jednej konformacji. Jedną z metod odpowiedzi jest próbkowanie

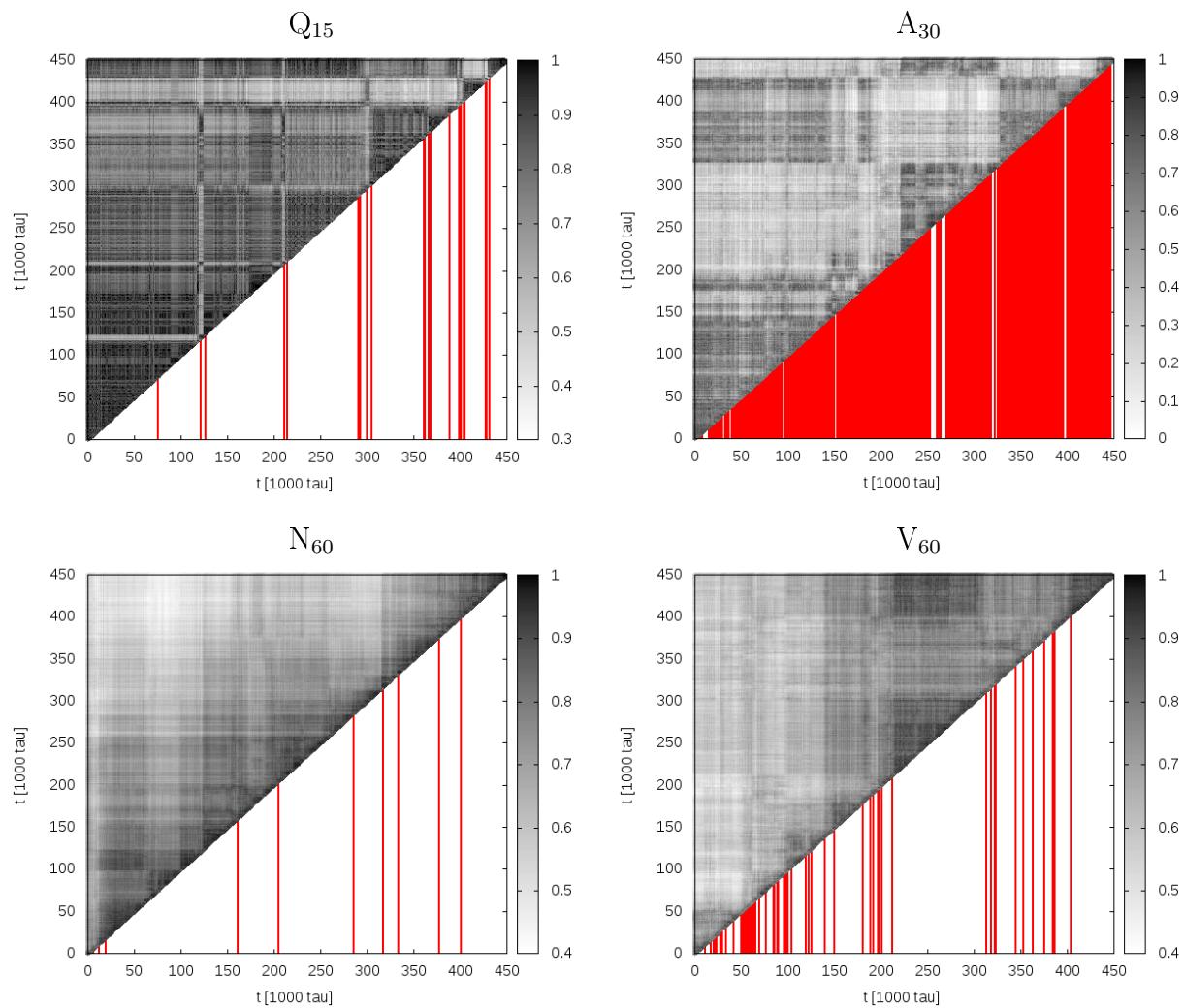
układu co  $1 \tau$  i wyznaczanie czasu, aż RMSD (pierwiastek ze średniego kwadratowego odchylenia położenia atomów  $C_\alpha$ , ang. *root mean square deviation*) liczone względem zadanej konformacji przekroczy  $5 \text{ \AA}$ . Dla  $Q_{60}$  rozkład tych czasów rozciąga się do  $65 \tau$ , ma średnią  $24 \tau$  i odchylenie standardowe  $10 \tau$ . Tak zdefiniowany typowy czas życia konformacji w symulacji pełnoatomowej [34] jest rzędu  $20 \text{ ns}$ , a więc porównywalny z symulacjami gruboziarnistymi.

Innym sposobem na scharakteryzowanie przejść konformacyjnych jest użycie ułamka  $f_{cc}$  kontaktów, które są wspólne z daną konformacją odniesienia. RMSD zależy od kształtu całego białka, zaś  $f_{cc}$  od oddziaływań zachodzących w białku. Na przykład zgięcie łańcucha na pół powoduje dużą zmianę RMSD, ale dopóki obie połówki nie utworzą między sobą kontaktów,  $f_{cc}$  wiele się nie zmieni, co będzie świadczyć o zachowaniu dotychczasowej mapy kontaktów. Zależność obu wielkości (RMSD i  $f_{cc}$ ) od czasu dla symulacji  $Q_{60}$  pokazuje Rys. 3.2. Rozważane są 3 konformacje odniesienia: pierwsza (nazwana A i oznaczona zielonymi liniami) pochodzi z chwili  $t_A = 10^5 \tau$ , druga (nazwana B i oznaczona czarnymi liniami) z chwili  $t_B = 2 \cdot 10^5 \tau$ , a trzecia (nazwana C i oznaczona niebieskimi liniami) z chwili  $t_C = 3 \cdot 10^5 \tau$ . Czerwone pionowe linie oznaczają czas, w którym, wg ułamka  $f_{cc}$ , konformacja zaczyna bardziej przypominać konformację C niż A. Dzieje się to dla  $t_{A/C} = 1.53 \cdot 10^5 \tau$ , co oznacza, że zmiany w mapie kontaktów zachodzą w skali mikrosekund. Podobne przejście do konformacji podobnej do B (oznaczone czarną pionową linią), dzieje się w późniejszej chwili  $t_{A/B} = 1.67 \cdot 10^5 \tau$ . Natomiast RMSD przekracza próg  $5 \text{ \AA}$  dużo szybciej - w ciągu pierwszych  $0.05 \cdot 10^5 \tau$ .



Rysunek 3.2: Zależność od czasu ułamka  $f_{cc}$  (góra) i RMSD (dół) dla pojedynczej symulacji  $Q_{60}$ . Wielkości są obliczane co  $1000 \tau$  w odniesieniu do trzech wybranych konformacji uzyskanych dla  $t_A = 10^5 \tau$  (A),  $t_B = 2 \cdot 10^5 \tau$  (B, przerywana linia) i  $t_C = 3 \cdot 10^5 \tau$  (C). Struktury odniesienia są przedstawione na górze. Czerwona pionowa linia oznacza przejście (dla  $f_{cc}$ ) między A i C. Czarna pionowa linia: między A i B. Modyfikacja rys. 3 z artykułu [I].

Rys. 3.3 pokazuje macierze  $f_{cc}$  dla 4 różnych homopolimerów o różnej długości i składzie. Ułamek wspólnych kontaktów ( $f_{cc}$ ) między dwiema konformacjami jest liczony jako liczba wspólnych kontaktów podzielona przez całkowitą liczbę kontaktów w tej konformacji, która ma ich więcej. Każdy punkt na macierzy  $f_{cc}$  oznacza  $f_{cc}$  między dwiema konformacjami (z chwil oznaczonych na osiach wykresu), zaznaczone w skali szarości. Macierz jest symetryczna, więc pokazano tylko fragment nad diagonalą. Duże zmiany w  $f_{cc}$  odpowiadają przejściom konformacyjnym.

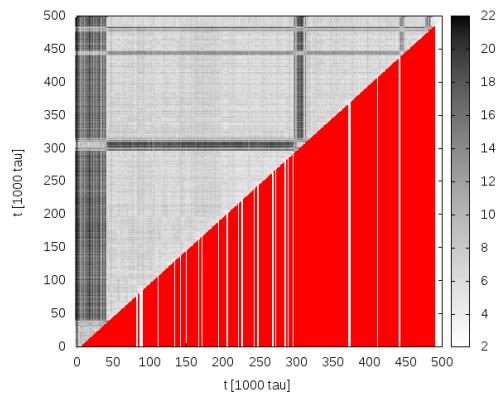


Rysunek 3.3: Macierze  $f_{cc}$  (ułamka wspólnych kontaktów dla konformacji z różnych czasów, zapisywanych co  $1000 \tau$ ) dla różnych homopeptydów. Oparte na rys. S9 z artykułu [I].

Aby opisać te przejścia ilościowo, został wymyślony autorski algorytm klastrowania konformacji: dwie konformacje są w jednym klastrze jeśli  $f_{cc}$  pomiędzy nimi jest większe od pewnego progu (50 %) i sąsiadują one ze sobą w czasie (algorytm działa poprzez łączenie sąsiednich klastrów, a  $f_{cc}$  między klastrami jest liczone jako średnie  $f_{cc}$  między strukturami z obu klastrów). W przeciwnieństwie do znanych algorytmów klastrowania, takich jak k-means, które nie są w żaden sposób sprzężone z czasem symulacji, ten algorytm potrafi nie tylko podzielić symulację na reprezentatywne konformacje, ale także powiedzieć ile każda z nich trwa (i nie zależy to od wyboru reprezentanta).

Czerwone linie na Rys. 3.3 oznaczają granice między klastrami. Próg wynosi 0.5, chociaż powinien zależeć od układu (np.  $A_{30}$  bardzo często zmienia konformacje, a  $N_{60}$  pozostaje przez większość czasu w jednym kształcie). Dla niektórych układów duże zmiany konformacyjne zachodzą w skali setek mikrosekund. PolyA wciąż tworzy helisę, rozwija ją i następnie do niej wraca, w przeciwnieństwie do symulacji pełnoatomowych, w których helisa wydaje się być trwała. Takie przejścia mogą wynikać z przybliżonej natury modelu gruboziarnistego, ale mogą także ukazywać prawdziwe zachowanie polyA w dużych skalach czasowych, niedostępnych dla symulacji pełnoatomowych.

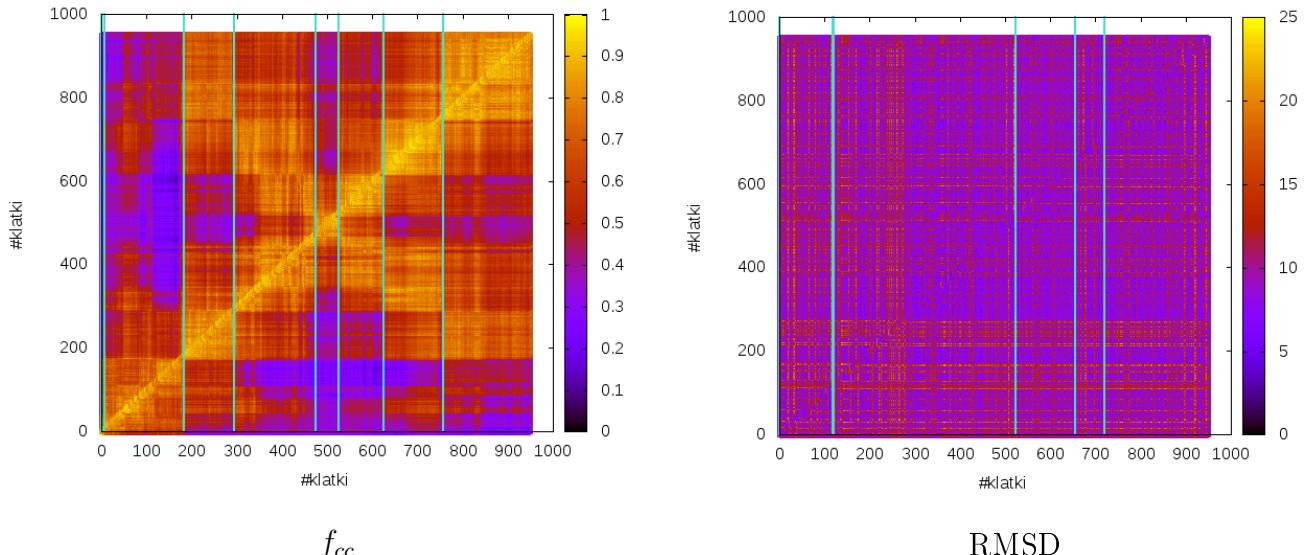
Opisanej wyżej metody klastrowania można użyć dla każdego deskryptora, który mówi na ile dwie konformacje są ilościowo podobne. W przypadku  $f_{cc}$  podobieństwo musi być powyżej 0.5, natomiast dla RMSD można zastosować próg 5 Å. Ponieważ progi zależą od układu, dużo wygodniej jest zmodyfikować algorytm tak, aby łączył sąsiednie klatki w klastry, aż zostanie osiągnięta zadana liczba klastrów. W ten sposób nie trzeba szukać progu dla każdego układu oddzielnie. Nieprzemyślane zastosowanie progu 5 Å do tak dużego układu jak  $Q_{60}$  skutkuje podziałem na zbyt wiele klastrów, jak na Rys. 3.4.



Rysunek 3.4: Macierz RMSD (dla konformacji zapisywanych co  $1000 \tau$ ) dla  $Q_{60}$ . Zlewające się czerwone linie są objaśnione w tekście.

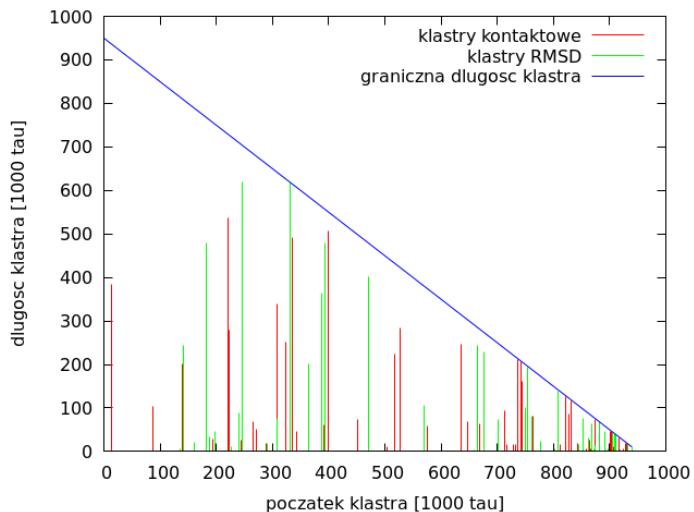
Rys. 3.5 pokazuje efekt działania algorytmu zadaną liczbą klastrów (10) dla tej samej symulacji  $Q_{60}$  (klatki były zapisywane co  $1000 \tau$ ), jednak w jednym przypadku klastrowana jest macierz  $f_{cc}$ , a w drugim RMSD. Widać, że  $f_{cc}$  tworzy dużo wyraźniej podzieloną macierz, a przejścia są lepiej widoczne i dobrze zaznaczone. Tymczasem algorytm klastrowania dla RMSD pokazuje przejścia w całkiem innych miejscach, co potwierdza, że te dwa deskryptory są różnymi, komplementarnymi metodami opisu dynamiki białka.

Interesujące jest zbadanie czasów trwania konformacji wyznaczonych bez zadanego z góry progu RMSD lub  $f_{cc}$ , lecz jedynie zadaną z góry liczbą klastrów (10). Wtedy można porównywać ze sobą długości czasu trwania konformacji wyznaczonych przy pomocy różnych deskryptatorów (w przeciwnym



Rysunek 3.5: Porównanie macierzy  $f_{cc}$  (po lewej) i macierzy RMSD (po prawej) dla tej samej symulacji  $Q_{60}$ . Obie macierze zostały podzielone na 10 klastrów (rozdzielonych przez zielone pionowe linie).

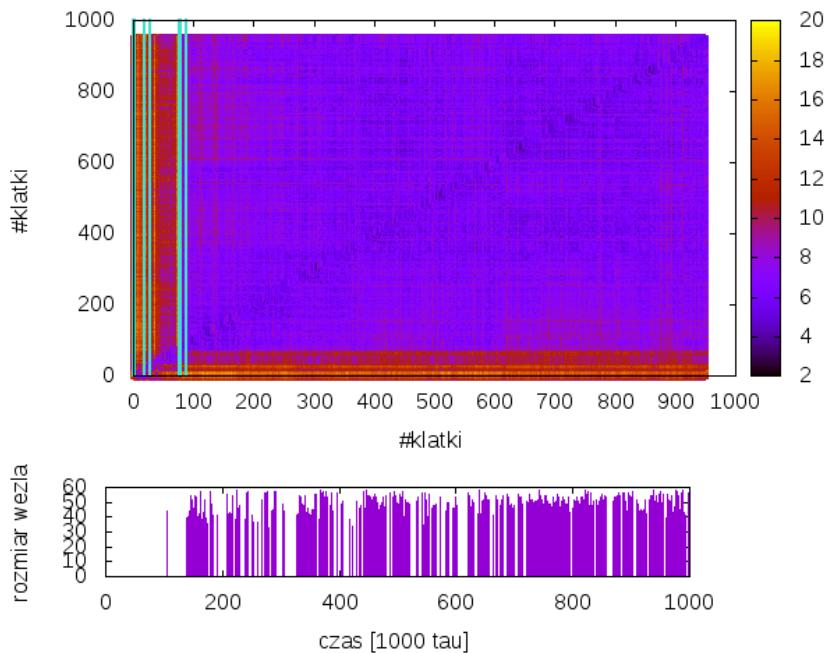
razie czas trwania zależałby od wybranego progu). Oczywiście im większa liczba klastrów na jaką ma być podzielona symulacja, tym krótszy czas trwania pojedynczego klastra. Jednak liczba klastrów jest wielkością niezależną od deskryptora. Rys. 3.6 pokazuje czasy trwania klastrów z 10 niezależnych symulacji  $Q_{60}$ , każda trwająca  $1\ 000\ 000 \tau$ . Pierwsze  $50\ 000 \tau$  jest przeznaczone na dojście układu do stanu równowagi i nie jest brane pod uwagę przy klastrowaniu, w związku z tym początek klastra  $0 \tau$  oznacza czas  $50\ 000 \tau$  od początku symulacji. Ponieważ koniec symulacji oznacza automatycznie koniec klastra, nie znamy tak naprawdę czasów trwania klastrów dotykających zielonej linii, dlatego nie należy ich brać pod uwagę. Widać, że niektóre klastry trwają nawet  $600\ 000 \tau$  (czas bardzo trudny do osiągnięcia dla symulacji pełnoatomowych). Statystyka jest za mała aby stworzyć rozkład czasów trwania klastra, jednak Rys. 3.6 sugeruje, że czasy trwania uzyskane wg RMSD są porównywalne do uzyskanych wg  $f_{cc}$ .



Rysunek 3.6: Czasy trwania klastrów z 10 niezależnych symulacji  $Q_{60}$ , każda trwająca  $1\ 000\ 000 \tau$ . Każda symulacja została podzielona na 10 klastrów.

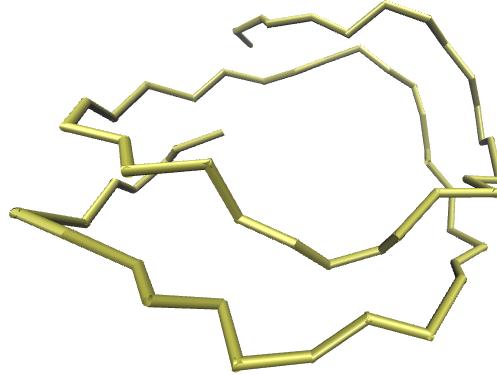
### 3.3 Tworzenie węzłów

Początek i koniec węzła były zawsze wyznaczane przy pomocy algorytmu KMT [199]. Węzły tworzyły się na tyle rzadko (mniej niż 2 % klatek), że trudno pokazać dla nich wyniki ilościowe. Jednak połączenie badania dynamiki tworzenia się węzła z badaniem macierzy RMSD daje poszlakę, że węzły mogą zwiększać czas trwania konformacji. Rys. 3.7 pokazuje macierz RMSD dla  $Q_{60}$  oraz wykres rozmiaru węzła w czasie. Wszystkie klatki, w których jest węzeł (z wyjątkiem jednej) należą do jednego klastra (macierz RMSD została podzielona na 10 klastrów). Klaster ten trwa dużo dłużej od pozostałych. Ponieważ jest to tylko jedna symulacja, nie można na jej podstawie wysnuwać daleko idących wniosków (zwłaszcza że węzeł, przedstawiony na Rys. 3.8, jest płytka i czasem znika). Bardziej systematyczna analiza zawęzłonych konformacji (przedstawiona w publikacji [VII] dla przypadku  $\alpha$ -synukleiny) potwierdza jednak, że węzeł znacznie ogranicza liczbę stanów konformacyjnych dostępnych dla białka.

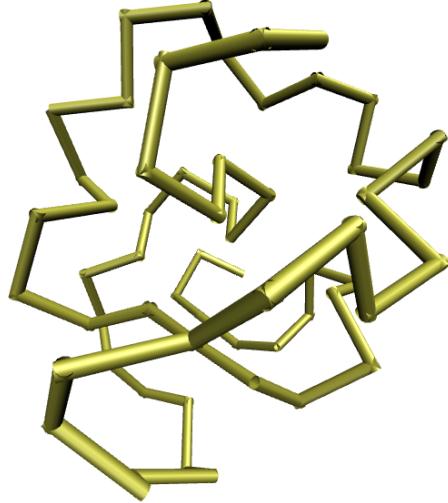


Rysunek 3.7: Macierz RMSD dla  $Q_{60}$  podzielona na 10 klastrów (góra) oraz wykres rozmiaru węzła w czasie (dół): każdy słupek oznacza rozmiar węzła dla jednej klatki. Jeśli słupka nie ma, w danej klatce nie było węzła. Pierwsze 50 000  $\tau$  symulacji to dochodzenie do równowagi, dlatego macierz RMSD kończy się dla 950 000  $\tau$  (wykres rozmiaru węzła uwzględnia także dojście do równowagi, zatem kończy się w 1 000 000  $\tau$ ).

Zwiększenie maksymalnej liczby kontaktów, jakie może utworzyć łańcuch boczny glutaminy z 2 na 3 (patrz poprzedni rozdział) powoduje, że węzłów tworzy się więcej (ale wciąż mniej niż 9 % klatek), w dodatku są one głębsze. Przykład głębokiego węzła przedstawia Rys. 3.9.



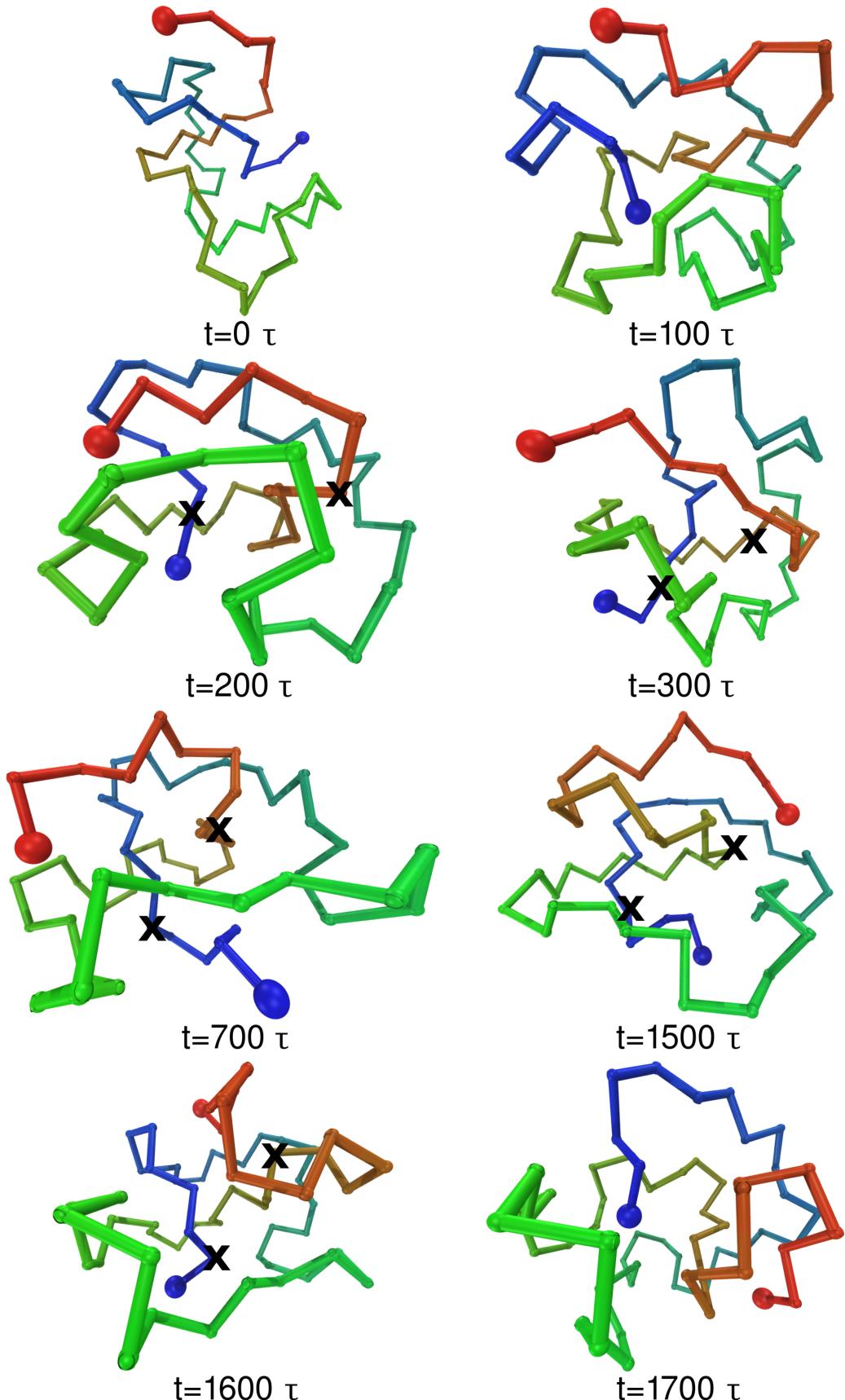
Rysunek 3.8: Płytki węzeł dla  $Q_{60}$ , którego ewolucję w czasie przedstawia Rys. 3.7.



Rysunek 3.9: Przykład głębokiego węzła dla  $Q_{60}$ , uzyskany podczas symulacji gdzie maksymalna liczba kontaktów łańcucha bocznego została zwiększena do  $s = 3$ .

Interesujące jest zbadanie, jak taki głęboki węzeł się tworzy i znika. Przedstawia to Rys. 3.10. Widać na nim, że biało się zawęża, gdy niebieski N-koniec przechodzi przez zieloną pętlę, a znika, gdy boki pętli zamkają się pod nim. Jest to zatem węzeł podobny do tych płytowych, pokazanych na Rys. 2.18 i 3.8, różniący się od nich rozmiarem pętli i tym, jak daleko koniec łańcucha przez nią przechodzi.

Wszystkie przedstawione w tej rozprawie węzły reprezentują najprostszy rodzaj węzła,  $3_1$ , którego rzut na płaszczyznę daje krzywą przecinającą się w 3 miejscach. W białach (także tych uporządkowanych) występują bardziej skomplikowane topologie węzłów, szerzej dyskutowane w [VII].

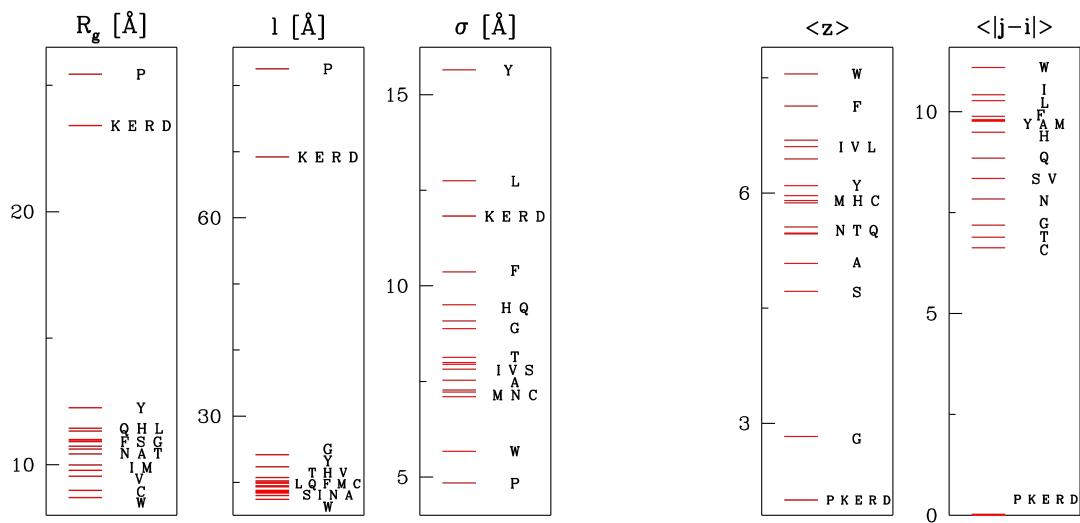


Rysunek 3.10: Samorzutne tworzenie (pierwsze 4 panele) i rozplątywanie (ostatnie 4 panele) głębokiego węzła w  $Q_{60}$  (symulacja dla  $s = 3$ ). Łańcuch pokolorowano tak, aby N-koniec był niebieski, a C-koniec czerwony. Końce węzła (jeśli istnieje) zostały zaznaczone czarnym X. Końce łańcucha oznaczono kulką. Oparte o rys. 2 z rozdziału w książce [VII].

### 3.4 Wyniki dla pojedynczych homopeptydów

Model gruboziarnisty został wykorzystany do przewidzenia właściwości geometrycznych wszystkich 20 homopeptydów (dla liczby aminokwasów  $n = 30$ ). Duże różnice w wynikach potwierdzają, że każdy aminokwas jest modelowany inaczej, a oddziaływanie między aminokwasami grają dużą rolę. Dla każdego homopeptydu wykonano 100 symulacji trwających 200 000  $\tau$ .

Podstawowe parametry geometryczne pokazują pierwsze 3 panele Rys. 3.11, a parametry związane z kontaktami pozostałe 2 panele. Te ostatnie to  $\langle z \rangle$ , czyli średnia liczba koordynacyjna; oraz  $\langle |j - i| \rangle$ , czyli średnia liczba aminokwasów dzieląca dwa aminokwasy ( $i$  oraz  $j$ ) będące w kontakcie ze sobą. Liczba koordynacyjna uwzględnia wszystkie kontakty oraz wiązania peptydowe wiążące ze sobą aminokwasy w łańcuchu. Jeśli podzielimy  $\langle |j - i| \rangle$  przez  $n$ , otrzymamy parametr znany jako porządek kontaktów (ang. *contact order*).



Rysunek 3.11: Promień bezwładności  $R_g$ , odległość między końcami  $l$ , jej dyspersja  $\sigma$ , średnia liczba koordynacyjna  $\langle z \rangle$  i średnia odległość aminokwasów z jednego kontaktu w sekwencji  $\langle |j - i| \rangle$ , dla homopeptydów składających się z 30 aminokwasów. Względny błąd średniej nie przekracza 1, 3, 2, 0.5 i 3.3% odpowiednio dla  $R_g$ ,  $l$ ,  $\sigma$ ,  $\langle z \rangle$  i dla  $\langle |j - i| \rangle$ . Aminokwasy są podpisane wg konwencji jednoliterowej. Gdyby histydynę traktować jako naładowaną, jej parametry byłyby takie same jak dla K,E,R oraz D. Modyfikacje rys. 8 i 9 z artykułu [I].

Wszystkie homopeptydy wydają się być nieuporządkowane, ponieważ  $\sigma$  jest zawsze znacząca i przekracza 5 Å. Zgodnie z [1] i [155], istnieje znacząca różnica między łańcuchami naładowanymi (polyK, polyE, polyR, i polyD) a dużo bardziej sklebioną resztą. Naładowane homopeptydy mają największe  $R_g$  oraz  $l$ , ponieważ nie tworzą żadnych kontaktów ss. PolyP jest następne, ponieważ także

nie może tworzyć kontaktów ss (ale aminokwasy polyP nie odpychają się już dalekozasięgowo). W związku z dużą sztywnością łańcucha polyP jest najszytniejszym homopeptydem i ma najmniejsze  $\sigma$ . Jednak to polyW wydaje się być najbardziej ekstremalnym przypadkiem: ma najmniejsze wartości  $R_g$  i  $l$  oraz największe wartości  $$  i  $|j - i|$ . Te cechy wynikają z dużych  $r_{min}$  oraz  $s$ , a przez to z dominacji kontaktów ss (każdy tryptofan może stworzyć 4 takie). W symulowanych konformacjach polyW prawie nie ma kontaktów bb, jest za to najwięcej węzłów: występują w 16% wszystkich symulacji i trwają przynajmniej przez 5% całkowitego czasu symulacji. Inne hydrofobowe homopeptydy są zwykle bardziej zwinięte i mniej sztywne niż te polarne, z wyjątkiem polyG i polyA, które mają bardzo małe łańcuchy boczne.

Inną metodą porównania homopeptydów jest badanie tego, jak skaluje się promień bezwładności dla wielu  $n$ : rozważane były homopeptydy między 10 a 100 aminokwasów. Dla  $n \neq 30$  każda ze 100 symulacji trwała  $50\,000 \tau$ .  $R_g$  dobrze skaluje się zgodnie z prawem potęgowym:  $R_g \sim n^\nu$

Wartości wykładnika  $\nu$  są podane w tabeli 2.4. Dla dużych hydrofobowych aminokwasów wykładnik  $\nu$  jest mniejszy (co odpowiada reżimowi słabego roztwarzalnika [164]) niż dla małych i polarnych aminokwasów (dla których woda jest być dobrym roztwarzalnikiem). Nasz model wydaje się więc dobrze oddawać kluczowe różnice między aminokwasami. Trzeba jednak pamiętać, że dopasowanie dotyczyło tylko zakresu  $n < 100$ . Dla większych  $n$  wykładnik  $\nu$  może dążyć do wartości  $\nu_G = 1/2$ , charakterystycznej dla łańcucha Gaussowskiego (ang. *Gaussian chain*). Wg szacunkowych obliczeń Flory'ego [164] wykładnik  $\nu$  dla polyA powinien być bliski  $1/2$  dla  $n > 200$ , natomiast dla polyP może znaczco różnić się od  $1/2$  nawet dla  $n$  rzędu 10 000.

Podczas poszukiwania informacji na temat homopolimerów, w bazie UniProt [200] znaleziono naj-dłuższe fragmenty złożone wyłącznie z jednego aminokwasu. Liczba aminokwasów w takim fragmencie,  $n_{max}$ , zmienia się od 6 (dla Ile, Tyr, Trp) do 79 (dla Gln).  $n_{max}$  także jest podane w tabeli 2.4. Mimo że dla niektórych aminokwasów  $n_{max}$  jest małe, można teoretycznie rozważać dłuższe łańcuchy (choćby aby wyznaczyć wspomniane prawa skalowania). Przewidywania modelu gruboziarnistego można by potwierdzić, syntetyzując sztucznie homopeptydy o długościach przekraczających naturalne  $n_{max}$ .

Tabela 3.1 pokazuje m.in. znane maksymalne długości fragmentów homopeptydowych z bazy UniProt [200]. Najdłuższe z nich odpowiadają glutaminie (79), serynie (58) i asparaginie (46). Dla alaniny ta długość to 24, ale dla niewiele większej waliny już tylko 9. Tym niemniej warto badać dłuższe homopeptydy (jak pokazują choćby wyniki dla  $V_{60}$  [20]).

Większa sztywność i mniejsza skłonność do zwijania się może tłumaczyć czemu polarne homopeptydy znalezione w bazie UniProt są dłuższe niż hydrofobowe (ponownie z wyjątkiem polyG i polyA, które mogą występować jako giętkie łączniki.)

nazwa	Gly	Pro	Gln	Cys	Ala	Ser	Val	Thr	Ile	Leu
$n_{max}$	23	27	79	11	24	58	9	24	6	13
ID <sub>UniProt</sub>	P10275	Q9NP73	Q156A1	Q03751	Q8R089	O15417	P32867	Q869S5	Q95US5	Q80YA8
$\nu$	0.56	0.93	0.61	0.48	0.57	0.54	0.46	0.56	0.51	0.49
nazwa	Asn	Asp	Lys	Glu	Met	His	Phe	Arg	Tyr	Trp
$n_{max}$	46	45	11	33	7	23	9	14	6	6
ID <sub>UniProt</sub>	Q54XG7	Q08438	Q8CI03	Q6PCN3	Q01668	Q6ZQ93	Q3S2U2	P38835	Q9C5D3	P86690
$\nu$	0.54	0.81	0.81	0.81	0.60	0.62	0.52	0.81	0.51	0.44

Tabela 3.1:  $n_{max}$  oznacza liczbę aminokwasów w najdłuższym fragmencie homopeptydowym znalezionym w recenzowanych wpisach bazy Uniprot (baza tych wpisów oznaczona jest jako SProt), z najwyższym poziomem pewności PE=1 (patrz [200]). ID<sub>UniProt</sub> to kod białka zawierającego najdłuższy fragment w bazie UniProt. Wykładnik  $\nu$  dotyczy prawa potęgowego określającego zależność  $R_g(n)$ .

### 3.5 Próba zastosowania modeli dla białek uporządkowanych

Podczas sprawdzania wariantów modelu PID oraz modelu quasi-adiabatycznego niektóre z nich zostały wykorzystane do symulacji białek uporządkowanych. Startowa konformacja pochodziła ze struktury natywnej lub błędzenia przypadkowego, a RMSD względem struktury natywnej posłużyło jako kryterium oceny, który model lepiej sprawdza się dla białek uporządkowanych. Okazało się, że w tej kategorii najlepiej radzą sobie całkiem inne warianty niż te na górze tabeli 2.9 - tabela 3.2 pokazuje, że najmniejsze RMSD jest osiągane dla macierzy oddziaływania pomnożonych przez czynnik 1, podczas gdy mnożnik ten jest mniejszy niż 0.5 w najlepszych wariantach dla białek nieuporządkowanych. Potwierdza to, że oddziaływanie między aminokwasami są w takich białkach dużo słabsze.

1L2Y			1ERY			1UBQ		
Model id	F <sub>RMSD</sub>	N <sub>RMSD</sub>	Model id	F <sub>RMSD</sub>	N <sub>RMSD</sub>	Model id	F <sub>RMSD</sub>	N <sub>RMSD</sub>
P <sup>+</sup> F <sub>B</sub> ME <sub>1</sub> T	6.0 Å	6.0 Å	W <sup>-</sup> F <sub>B</sub> MJ <sub>1</sub> T	6.0 Å	7.5 Å	P <sup>+</sup> F MJ <sub>1</sub> T	11.1 Å	7.0 Å
P <sup>-</sup> F <sub>B</sub> ME <sub>1</sub> T	6.0 Å	6.0 Å	P <sup>-</sup> F <sub>B</sub> MJ <sub>1</sub> T	6.1 Å	7.6 Å	P <sup>+</sup> F <sub>B</sub> ME <sub>1</sub> T	11.3 Å	7.7 Å
P <sup>-</sup> F ME <sub>1</sub> T	6.1 Å	6.1 Å	P <sup>-</sup> F <sub>B</sub> ME <sub>1</sub> T	6.3 Å	7.0 Å	P <sup>-</sup> F MJ <sub>1</sub> T	11.4 Å	7.3 Å
P <sup>-</sup> F MD <sub>0.1</sub> C	6.2 Å	6.1 Å	W <sup>+</sup> F <sub>B</sub> ME <sub>1</sub> T	6.3 Å	7.0 Å	W <sup>-</sup> F ME <sub>1</sub> T	11.4 Å	7.3 Å
P <sup>-</sup> F MD <sub>0.4</sub> C	6.2 Å	6.1 Å	P <sup>+</sup> F <sub>B</sub> ME <sub>1</sub> T	6.6 Å	7.0 Å	W <sup>+</sup> F <sub>B</sub> MJ <sub>0.5</sub> T	11.4 Å	8.3 Å
P <sup>-</sup> F <sub>B</sub> MJ <sub>1</sub> T	6.2 Å	6.2 Å	W <sup>+</sup> L ME <sub>1</sub> T	6.7 Å	7.0 Å	P <sup>-</sup> F <sub>B</sub> ME <sub>1</sub> T	11.5 Å	8.1 Å
P <sup>-</sup> F MJ <sub>1</sub> T	6.3 Å	6.3 Å	W <sup>-</sup> F ME <sub>1</sub> T	6.7 Å	7.3 Å	P <sup>+</sup> F MD <sub>1</sub> T	12.2 Å	9.3 Å
P <sup>+</sup> F <sub>B</sub> MJ <sub>1</sub> T	6.4 Å	6.4 Å	W <sup>-</sup> F MJ <sub>1</sub> T	6.8 Å	7.3 Å	W <sup>-</sup> L ME <sub>1</sub> T	12.6 Å	-
P <sup>+</sup> F MJ <sub>1</sub> T	6.4 Å	6.5 Å	P <sup>+</sup> L ME <sub>1</sub> T	6.8 Å	7.5 Å	P <sup>-</sup> F MD <sub>0.1</sub> C	12.7 Å	7.8 Å
W <sup>-</sup> L <sub>B</sub> MD <sub>1</sub> T	8.6 Å	8.8 Å	W <sup>+</sup> F <sub>B</sub> MD <sub>0.5</sub> T	7.4 Å	8.0 Å	P <sup>-</sup> L ME <sub>1</sub> T	12.8 Å	-

Tabela 3.2: RMSD dla 3 białek ustrukturyzowanych (1L2Y, 1ERY i 1UBQ) dla symulacji modelem gruboziarnistym rozpoczynających się z konformacji błędzenia przypadkowego (symulacje zwijania, F<sub>RMSD</sub>) i ze struktury natywnej (N<sub>RMSD</sub>).

Symulacje były prowadzone dla 3 białek (1L2Y: 20 aminokwasów, 1ERY: 39 aminokwasów, 1UBQ: 76 aminokwasów) i trwały 150 000  $\tau$ . W przypadkach symulacji startujących z konformacji błędzenia przypadkowego  $R_{\text{MSD}}$  jest na tyle duże, że trudno mówić o zwijaniu. Dlatego przedstawiany tu model gruboziarnisty (niezależnie od wariantu) przeznaczony jest tylko dla białek nieuporządkowanych. Można go jednak łatwo połączyć z modelem opartym na strukturze natywnej, aby symulować białka częściowo uporządkowane i częściowo nieuporządkowane [VI]. Postępowanie to jest opisane w rozdziale dotyczącym symulacji glutenu.

## 3.6 Podsumowanie

Model gruboziarnisty został wykorzystany do badania przejść konformacyjnych w homopeptydach, w tym przejść związań z tworzeniem węzłów. Pokazał on, że przejścia te mogą zachodzić w bardzo długich skalach czasowych, co potwierdza zasadność jego stosowania: bardziej dokładne modele mają duże trudności z osiągnięciem takich skal czasowych. Jednak w przypadku pojedynczych łańcuchów białkowych te trudności mogą być przezwyciężone przy pomocy zaawansowanych metod próbkowania, takich jak wymiana replik, w związku z tym główną zaletą tego modelu jest możliwość wyznaczenia fizycznego czasu trwania danej konformacji (pamiętając że zależność  $1 \tau = 1 \text{ ns}$  jest jedynie przybliżona).

Model został wykorzystany do symulacji 20 homopeptydów o różnych długościach. Wyniki wydają się być zgodne z dotychczasowym stanem wiedzy, ale zawierają także przewidywania, które mogą być potwierdzone doświadczalnie (np. wykładniki prawa potęgowego dla  $R_g$ ). Warto zauważyć, że w modelach używanych do interpretacji eksperymentów FRET białka nieuporządkowane są często traktowane jako idealny łańcuch Gaussowski, który jest dobrym przybliżeniem tylko dla białek o bardzo wielu aminokwasach ( $n > 100$ ). Wykładniki dla homopolimerów różnią się od  $1/2$ , ale zawierają się w przewidzianym przez Flory'ego [164] zakresie między  $1/3$  a  $3/5$  (z wyjątkiem homopolimerów naładowanych i polyPro, w których elektrostatyczne odpychanie lub sztywność prowadzą do dużo bardziej rozprostowanych łańcuchów). Mimo że symulowane były tylko homopolimery, inne białka powinny mieć wykładniki w podobnym zakresie [201].

Niestety tak prosty model gruboziarnisty nie może odwzorować właściwości białek uporządkowanych, ale to ograniczenie można obejść łącząc go z modelem opartym na strukturze natywnej [VI].

# Rozdział 4

## Symulacje wielu łańcuchów poliglutaminy i polialaniny

### 4.1 Wprowadzenie

Ten rozdział przedstawia wyniki symulacji wielu identycznych łańcuchów poliglutaminy  $Q_{20}$ ,  $Q_{40}$ , i  $Q_{60}$  oraz polialaniny  $A_{20}$ , gdzie wskaźnik dolny oznacza długość jednego łańcucha  $N$ . Białka poliglutaminy są znane ze skłonności do agregacji [75], a choroby neurozwyrodnieniowe związane z fragmentami  $Q_N$  pojawiają się, gdy  $N \gtrsim 40$  [202], dlatego wybrane zostały długości poniżej, powyżej i w pobliżu tego progu. Polialanina jest trzecim najczęściej występującym w białkach homopeptydowym fragmentem po poliglutaminie i poliasparaginie<sup>1</sup>. Sztucznie zsyntetyzowane homopeptydy  $A_{19}$  także potrafią agregować tworząc fibryle [203]. Z tych powodów jako czwarty układ zostały wybrane właśnie łańcuchy  $A_{20}$ .

Agregacja białek może prowadzić do separacji faz. Ogólnie przyjęty model separacji dwóch faz ciekłych w cieczy van der Waalsa zakłada istnienie obszaru współlistnienia, ograniczonego skierowaną w dół asymetryczną parabolą na wykresie gęstości ( $\rho$ ) i temperatury ( $T$ ), zaś wierzchołek paraboli oznacza temperaturę krytyczną ( $T_c$ ), powyżej której nie można odróżnić faz [204].

Sytuacja fizyczna dla kropel złożonych z białek jest dużo bardziej skomplikowana [205]: zamiast cząsteczek jednoatomowych agregują polimery, a agregacja zachodzi w środowisku wodnym [206]. Diagram fazowy może zależeć od sekwencji białka, a w szczególności od rozkładu aminokwasów naładowanych [207, 208]. Konstrukcja diagramów fazowych białek nieuporządkowanych mogłaby być tematem oddzielnej rozprawy, dlatego na potrzeby tej pracy symulowana była jedynie agregacja nienaładowanych homopeptydów  $A_N$  i  $Q_N$ .

Ważne kroki w badaniu separacji faz w białkach przy pomocy symulacji podjęli Dignon i in. [69], też używając prostego modelu gruboziarnistego z jedną kulką na aminokwas. W modelu tym oddziaływanie między aminokwasami opisywane są potencjałem LJ<sup>2</sup> oraz potencjałem Debye'a-Hückela.

<sup>1</sup> Jednoimiennie naładowane aminokwasy poliasparaginy się odpychają, więc ten homopeptyd nie jest tak interesującym obiektem badań jak polialanina i poliglutamina.

<sup>2</sup> W wariancie z amplitudą zależną od hydrofobowości aminokwasów [70] bądź od macierzy Miyazawa-Jernigana [49].

Dla białek FUS (ang. *Fused in Sarcoma*) model ten przewiduje diagram podobny do cieczy van der Waalsa [209]. Białka te były symulowane w pudełku z periodycznymi warunkami brzegowymi, o kształcie graniastosłupa wydłużonego w jednym kierunku. Profil gęstości wzduż tego kierunku był sigmoidalny, a gęstości w dolnej i górnej części sigmoidy posłużyły do wyznaczenia krzywej współlistnienia (symulacje były powtarzane dla różnych temperatur).

Oddziaływanie między aminokwasami w przedstawionym tu modelu są bardziej skomplikowane niż w modelu [69] i obejmują np. podział na kontakty tworzone przez łańcuchy boczne i łańcuch główny (choć w obu modelach jeden pseudoatom reprezentuje jeden aminokwas). W związku z tym w tej rozprawie na wykresie gęstości i temperatury badane są inne wielkości, oparte o liczbę kontaktów: łańcuchy połączone między sobą kontaktami tworzą agregaty (stacjonarne rozkłady tych agregatów opisane są w podpunkcie 4.3.1). Użyte zostało jednak pudełko symulacyjne w kształcie sześciianu (z periodycznymi warunkami brzegowymi), więc powstające agregaty szybko “parowały”, wymieniając się łańcuchami z resztą układu. Kinetyka tego procesu została opisana na podstawie analizy czasu trwania kontaktu między łańcuchami (co przedstawia podpunkt 4.3.3).

Dla niskich temperatur wymiana łańcuchów zachodzi wolniej, a agregaty przyjmują kształt podobny do włókien amyloidowych (których istnienie jest potwierdzone doświadczalnie zarówno dla poliglutaminy [84] jak i polialininy [203]). Dynamika takich agregatów odpowiada fazie “szkła amyloidowego”, które zostało ilościowo opisane w paragrafie 4.3.1.1. Faza ta nie występuje w modelu Dignona i in. [69] - analiza kontaktów między łańcuchami nie musi prowadzić do takiego samego wykresu fazowego jak przeprowadzona w [69] analiza gęstości. Dlatego przedstawione tu wyniki są komplementarne do tych z [69, 210] i prezentują inne podejście do badania agregacji białek.

To, jak dla różnych gęstości i temperatur zmieniają się właściwości pojedynczych łańcuchów i ich par przedstawia podpunkt 4.3.2.

## 4.2 Protokół symulacji

Symulacje prowadzone są modelem quasi-adiabatycznym (opisanym w podrozdziale 2.5) w zespole NVT, z periodycznymi warunkami brzegowymi. Dla każdej opisywanej tu kombinacji gęstości  $\rho$  i temperatury  $T$  przeprowadzone zostało parę niezależnych symulacji (co najmniej dwie), z których każda trwała  $1\ 500\ 000 \tau$ . Tylko ostatnia  $\frac{1}{3}$  czasu symulacji została wykorzystana do zbierania wyników (pierwsze  $1\ 000\ 000 \tau$  było przeznaczone na dojście układu do stanu równowagi).

Liczba łańcuchów jest oznaczona jako  $n_m$ , a liczba aminokwasów w jednym łańcuchu jako  $N$ . Większość wyników uzyskano dla układów składających się z 1800 aminokwasów, podzielonych następująco:  $n_m = 90$  i  $N = 20$ ,  $n_m = 45$  i  $N = 40$  bądź  $n_m = 30$  i  $N = 60$ .

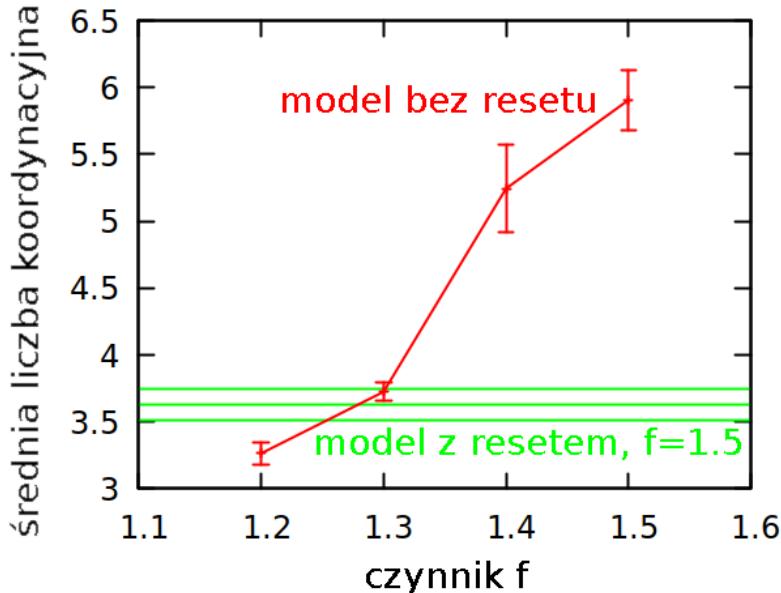
Symulacje kontrolne dla 2400 aminokwasów potwierdziły, że wyniki nie zmieniają się znacząco w zależności od rozmiaru układu (mimo że agregaty dla większego układu nadal “parowały”, a mniej niż sto łańcuchów to układ daleki od granicy termodynamicznej).

Na początku symulacji łańcuchy są ułożone losowo w sześciennym pudełku do symulacji, a ich konformacje wyznacza samoomijające błędzenie przypadkowe. Rozmiar pudełka zależy od przyjętej gęstości  $\rho$  (liczonej w aminokwasach na  $\text{nm}^3$ , w skrócie  $\text{aa}/\text{nm}^3$ ), ponieważ liczba aminokwasów jest stała. Dla  $\rho > 1.2 \text{ aa}/\text{nm}^3$  początkowy rozmiar pudełka wynosi tyle ile wynika z gęstości  $1.2 \text{ aa}/\text{nm}^3$ , a następnie (po upływie  $100\ 000 \tau$  na dojście pojedynczych łańcuchów do stanu równowagi) ściany pudełka zbliżają się do siebie aż do osiągnięcia zadanej gęstości z szybkością  $0.001 \text{ nm}/\tau$ . Jest to konieczne, ponieważ samoomijające błędzenie przypadkowe skutkuje łańcuchami o dużo większych rozmiarach (mierzonych np. promieniem bezwładności) niż rozmiary dla białek w równowadze.

#### 4.2.1 Problemy z symulacją wielu łańcuchów

W użytym modelu quasi-adiabatycznym każdy aminokwas może stworzyć jedynie ograniczoną liczbę kontaktów (patrz podpunkt 2.5.5). Na początku symulacji każdy aminokwas będzie oddziaływał z aminokwasami z tego samego łańcucha z dużo większym prawdopodobieństwem w stosunku do aminokwasów z innych łańcuchów. Może to doprowadzić do wysycenia liczby koordynacyjnej tego aminokwasu przez kontakty utworzone w ramach jednego łańcucha, zanim spotka on inne łańcuchy. łańcuchy nie będą wtedy tworzyć między sobą kontaktów. Aby temu zaradzić, w symulacjach wielu łańcuchów wprowadzone zostało resetowanie listy kontaktów co  $5000 \tau$ : podczas resetu kryteria kątowe i odległościowe są na nowo stosowane dla wszystkich aminokwasów, tak aby umożliwić utworzenie kontaktów między łańcuchami. Konieczność wprowadzenia takiego resetu jest oczywistym ograniczeniem modelu. Dlatego rozwinięty został model z hamiltonianem niezależnym od czasu (opisany w podrozdziale 2.6). Okazał się on zbyt wolny na potrzeby symulacji tak wielu łańcuchów jak opisano tutaj (zaś wyniki dla małych układów sugerują że łańcuchy mogą się w nim zbyt mocno szczepiać).

Jedną z metod pozwalających na uniknięciu okresowego resetu kontaktów jest zmiana czynnika  $f$ , który określa odległość powyżej której kontakt jest quasi-adiabatycznie wyłączany. Wynosi ona  $2^{-1/6}fr_{min}$ , gdzie  $r_{min}$  to minimum potencjału LJ dla danego kontaktu. Domyślana wartość  $f$  to 1.5 (dla której potencjał LJ osiąga tylko ok. 30% swojej maksymalnej amplitudy), jednak wartość 1.3 pozwala na odtworzenie właściwości modelu z resetem przy użyciu modelu bez resetu (jak pokazuje Rys. 4.1 dla średniej liczby koordynacyjnej). Zmniejszoną wartość  $f$  można uzasadnić mniejszą trwałością kontaktów w bezpośrednim sąsiedztwie wody, która często otacza białka nieuporządkowane, w szczególności tak hydrofilowe jak poliglutamina. Dla  $f = 1.3$  potencjał LJ osiąga już ponad 60% maksymalnej amplitudy, co może utrudniać quasi-adiabatyczne przełączanie.



Rysunek 4.1: Zależność od  $f$  średniej liczby koordynacyjnej dla symulacji 8 łańcuchów  $Q_{20}$  trwającej 250 000  $\tau$ .

#### 4.2.2 Algorytm grupowania łańcuchów w agregaty

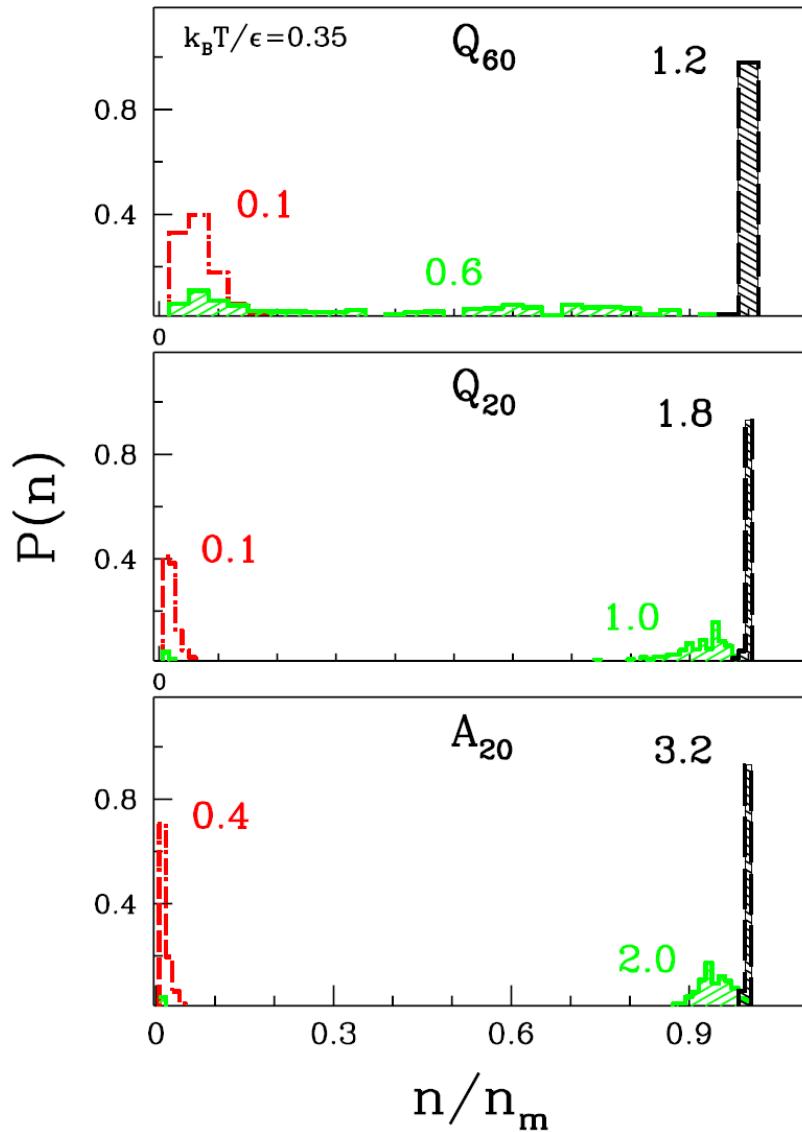
Aby opisać agregację białek pomocne jest pojęcie agregatu. Dwa łańcuchy należą do tego samego agregatu jeśli posiadają choć jedną wspólną parę aminokwasów, które są w odległości mniejszej niż  $r_{min}^{ss}$ . Odległość ta odpowiada minimum potencjału LJ dla kontaktów typu ss (czyli 8.6 Å dla dwóch glutamin oraz 6.4 Å dla dwóch alanin). A zatem łańcuch należy do danego agregatu jeśli posiada choć jeden aminokwas odległy od niego o mniej niż  $r_{min}^{ss}$ .

W trakcie symulacji monitorowane jest prawdopodobieństwo tego, że losowo wybrany łańcuch będzie należał do agregatu o wielkości  $n$  (tzn. zawierającego  $n$  łańcuchów). Prawdopodobieństwo to jest oznaczone jako  $P(n)$  i jest wyznaczone na podstawie wszystkich klatek z ostatnich 500 000  $\tau$  symulacji (w tym czasie rozkład wielkości agregatów jest już stacjonarny). Warto pamiętać że  $P(n)$  nie jest tym samym co rozkład wielkości agregatów  $p(n)$ . Np. jeśli istnieją 3 agregaty o wielkości 2 oraz jeden agregat o wielkości 6 (czyli łącznie jest 12 łańcuchów), to  $p(2) = 3/4$ , natomiast  $P(2) = 3 \cdot 2/12 = 1/2$ .

### 4.3 Wyniki

#### 4.3.1 Agregacja dla różnych temperatur i gęstości

Przykłady rozkładów  $P(n)$  są pokazane na Rys. 4.2. Dla małych  $\rho$ , rozkłady mają maksimum w pobliżu  $n=1$ , co odpowiada fazie o niskiej gęstości. Mimo że łańcuchy poruszają się w wodzie (która w symulacji reprezentowana jest tylko przez szum termiczny i tłumienie), faza ta będzie oznaczona jako G (gazowa) ze względu na zaniedbywalną rolę oddziaływań między łańcuchami.



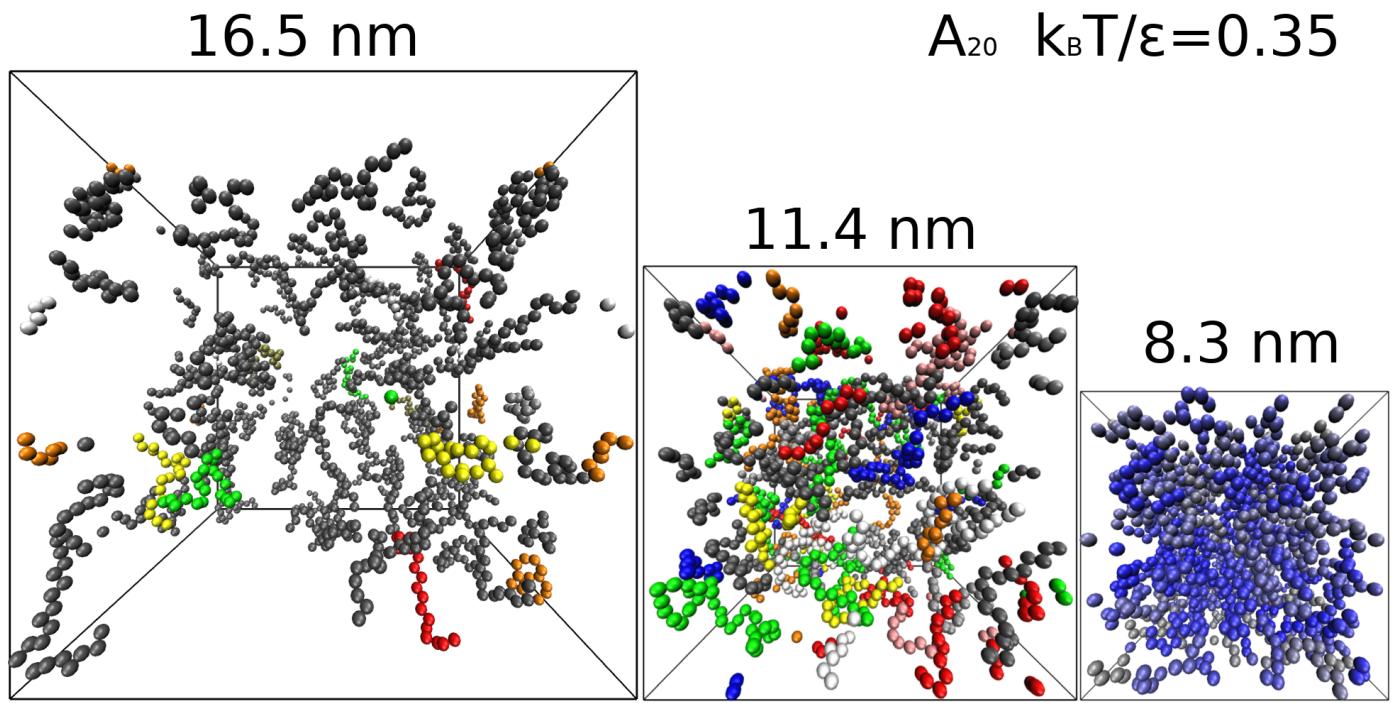
Rysunek 4.2: Uśrednione po czasie stacjonarne rozkłady  $P(n)$  dla podanych układów i gęstości, otrzymane dla  $k_B T = 0.35 \epsilon$ . Liczby przy słupkach oznaczają gęstość (w aminokwasach/nm<sup>3</sup>). Czerwony odpowiada fazie G, czarny fazie C, a zielony sytuacji pośredniej. Modyfikacja rys. 1 z artykułu [IV].

Dla dostatecznie dużych  $\rho$  praktycznie wszystkie łańcuchy tworzą jeden agregat. Jest to faza C (“ciecz”). Dla pośrednich wartości  $\rho$  pojawiają się agregaty wszystkich możliwych rozmiarów. Granice między fazami zależą od układu. Aby je określić, przyjęte zostały następujące kryteria:

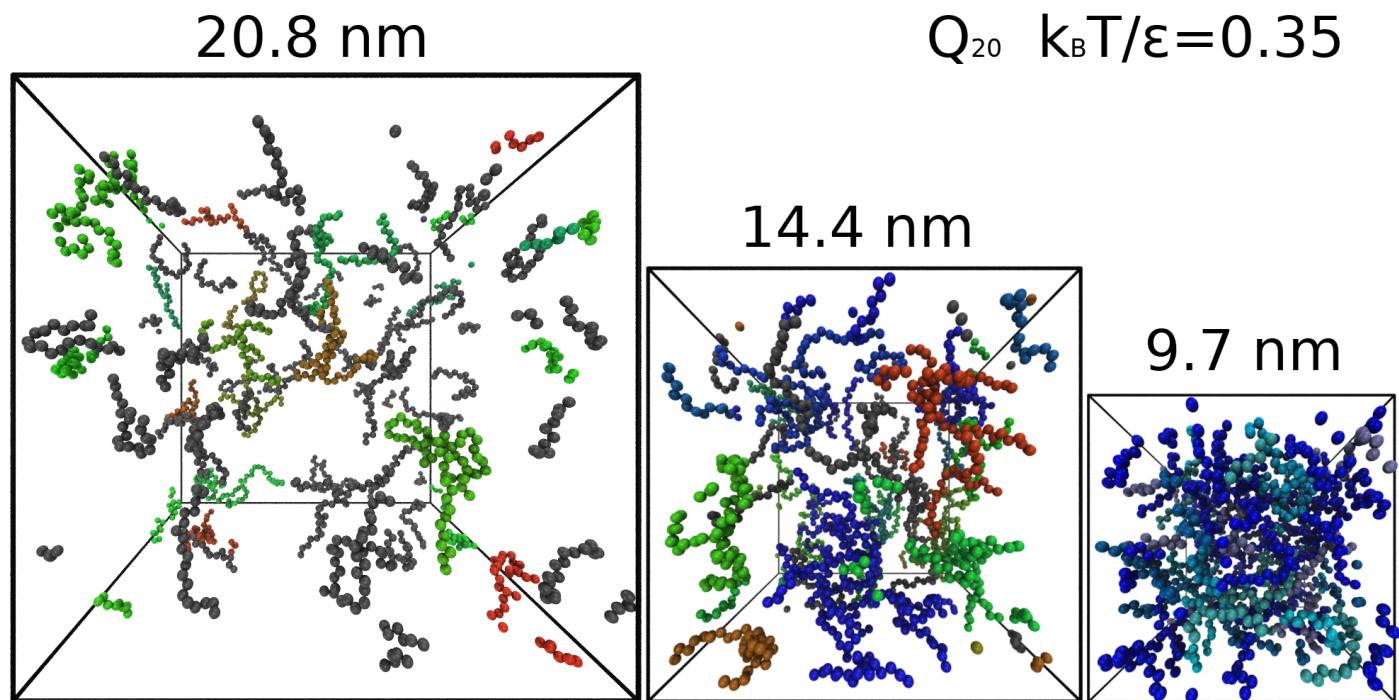
1. Faza G odpowiada  $P(1) \geq 0.5$
2. Faza C odpowiada  $P(n_{max}) \geq 0.95$
3. Granica B przebiega dla  $P(n_{max}) = 0.5$

gdzie  $n_{max}$  oznacza największy zaobserwowany rozmiar agregatu dla danego układu (w fazie C jest zwykle równy liczbie łańcuchów  $n_m$  lub trochę mniejszy, ze względu na wspomniane parowanie łańcuchów). Cały obszar pomiędzy fazami G i C odpowiada sytujom pośrednim. Jednak łańcuchy częściej grupują się w duże agregaty dla  $P(n_{max}) \geq 0.5$ , dlatego została dodatkowo zaznaczona granica B.

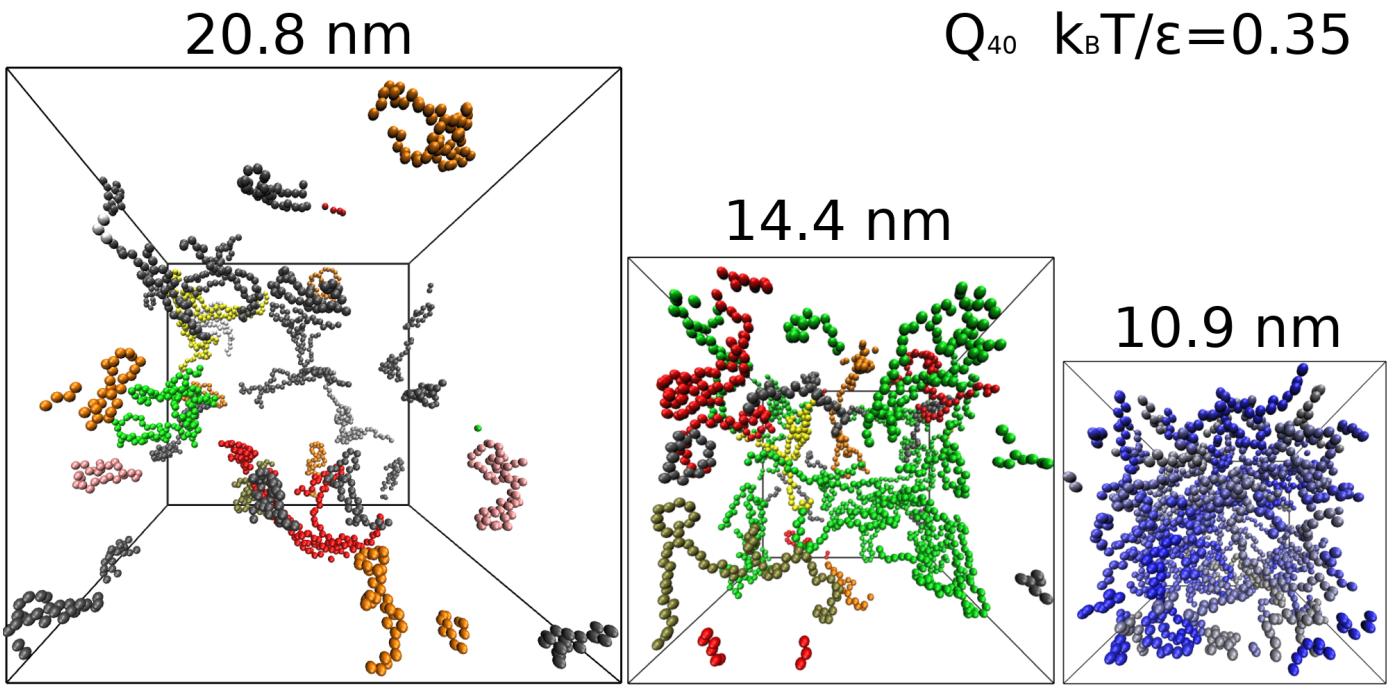
Rys. 4.3-4.6 ilustrują jak symulowane układy wyglądają w każdej z tych 3 faz (G, C oraz pośrednia).



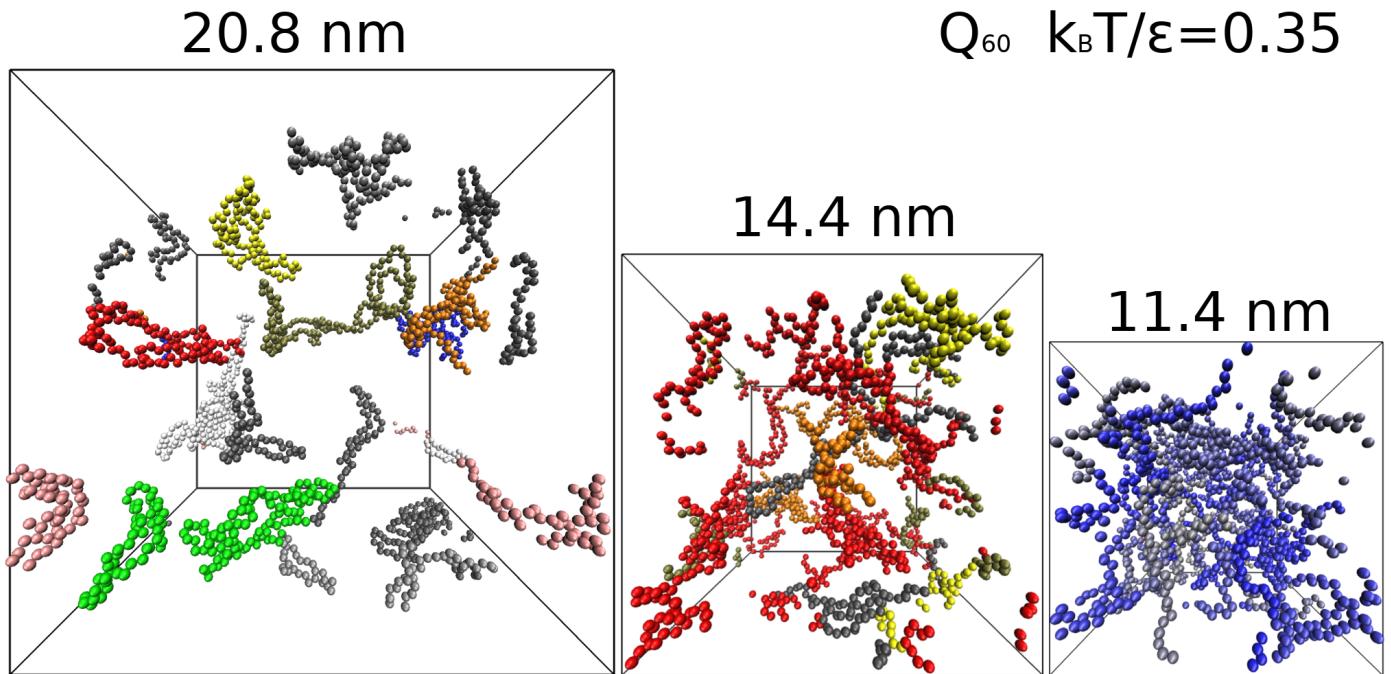
Rysunek 4.3: 90 łańcuchów  $A_{20}$  dla gęstości  $0.4 \text{ aa}/\text{nm}^3$  (faza G),  $1.2 \text{ aa}/\text{nm}^3$  (faza pośrednia) oraz  $3.2 \text{ aa}/\text{nm}^3$  (faza C). Podpisy oznaczają długość boku pudełka oraz temperaturę. Oparte na rys. S1 z artykułu [IV].



Rysunek 4.4: 90 łańcuchów  $Q_{20}$  dla gęstości  $0.2 \text{ aa}/\text{nm}^3$  (faza G),  $0.6 \text{ aa}/\text{nm}^3$  (faza pośrednia) oraz  $2.0 \text{ aa}/\text{nm}^3$  (faza C). Pojedyncze łańcuchy są pokolorowane na szaro, kolory oznaczają agregaty co najmniej 2 łańcuchów. Prawy panel zawiera tylko jeden agregat, pokolorowany odcieniami niebieskiego. Podpisy oznaczają długość boku pudełka oraz temperaturę. Oparte na rys. 2 z artykułu [IV].



Rysunek 4.5: 45 łańcuchów  $Q_{40}$  dla gęstości  $0.2 \text{ aa/nm}^3$  (faza G),  $0.6 \text{ aa/nm}^3$  (faza pośrednia) oraz  $1.4 \text{ aa/nm}^3$  (faza C). Podpisy oznaczają długość boku pudełka oraz temperaturę. Oparte na rys. S2 z artykułu [IV].



Rysunek 4.6: 30 łańcuchów  $Q_{60}$  dla gęstości  $0.2 \text{ aa/nm}^3$  (faza G),  $0.6 \text{ aa/nm}^3$  (faza pośrednia) oraz  $1.2 \text{ aa/nm}^3$  (faza C). Podpisy oznaczają długość boku pudełka oraz temperaturę. Oparte na rys. S3 z artykułu [IV].

#### 4.3.1.1 Anizotropowa agregacja i tworzenie złogów amyloidowych

Dla niskich temperatur agregaty zmieniają się dużo wolniej i po pewnym czasie zaczynają przypominać włókna amyloidowe (patrz Rys. 4.7). Dzieje się tak niezależnie od układu (choć temperatura przy jakiej agregaty amyloidowe się pojawiają wydaje się rosnąć wraz z liczbą aminokwasów w łańcuchu, patrz Rys. 4.11). Niezależnie od gęstości istnieje zwykle więcej niż jeden taki agregat - dla dużych gęstości "włókna" o różnych orientacjach znajdują się na siebie (formalnie tworząc jeden agregat), a czas potrzebny na ich reorientację jest bardzo długi - dlatego fazę tę można nazwać szkłem amyloidowym (i oznaczyć literą A).

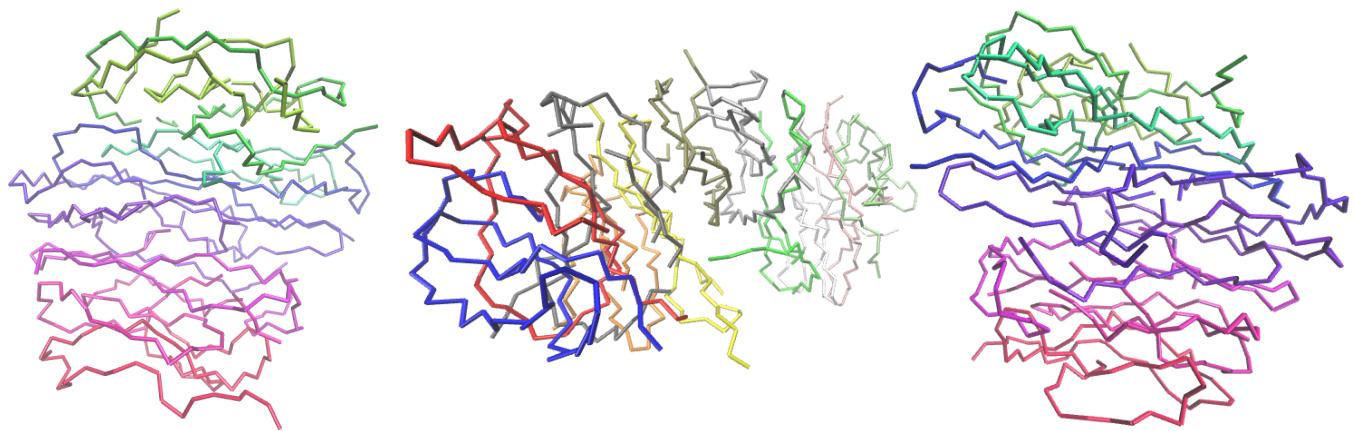
Łańcuchy główne w amyloidach układają się równolegle bądź antyrównolegle do siebie (patrz Rys. 4.8). Układ staje się zatem lokalnie anizotropowy (ponieważ zwykle istnieje więcej niż jeden agregat, uporządkowanie nie jest globalne). Pojedyncze łańcuchy w szkle amyloidowym przypominają  $\beta$ -nici lub (dla Q<sub>60</sub>)  $\beta$ -spinki. Prawdziwe agregaty polyQ na poziomie molekularnym mogą tak wyglądać [211].

Aby scharakteryzować tę lokalnie anizotropową fazę wprowadzona została wielkość  $C = |\cos \alpha|$ , gdzie kąt  $\alpha$  jest zdefiniowany tak jak dopełnienie kąta płaskiego, tyle że nie między dwoma sąsiednimi fragmentami jednego łańcucha, a między fragmentami z różnych łańcuchów (patrz panel e na Rys. 4.8). Wielkość  $C$  została obliczona tylko dla tych par łańcuchów, które mają między sobą przynajmniej jeden kontakt ("odległościowy", tzn. taki jak zdefiniowano w podpunkcie 4.2.2).

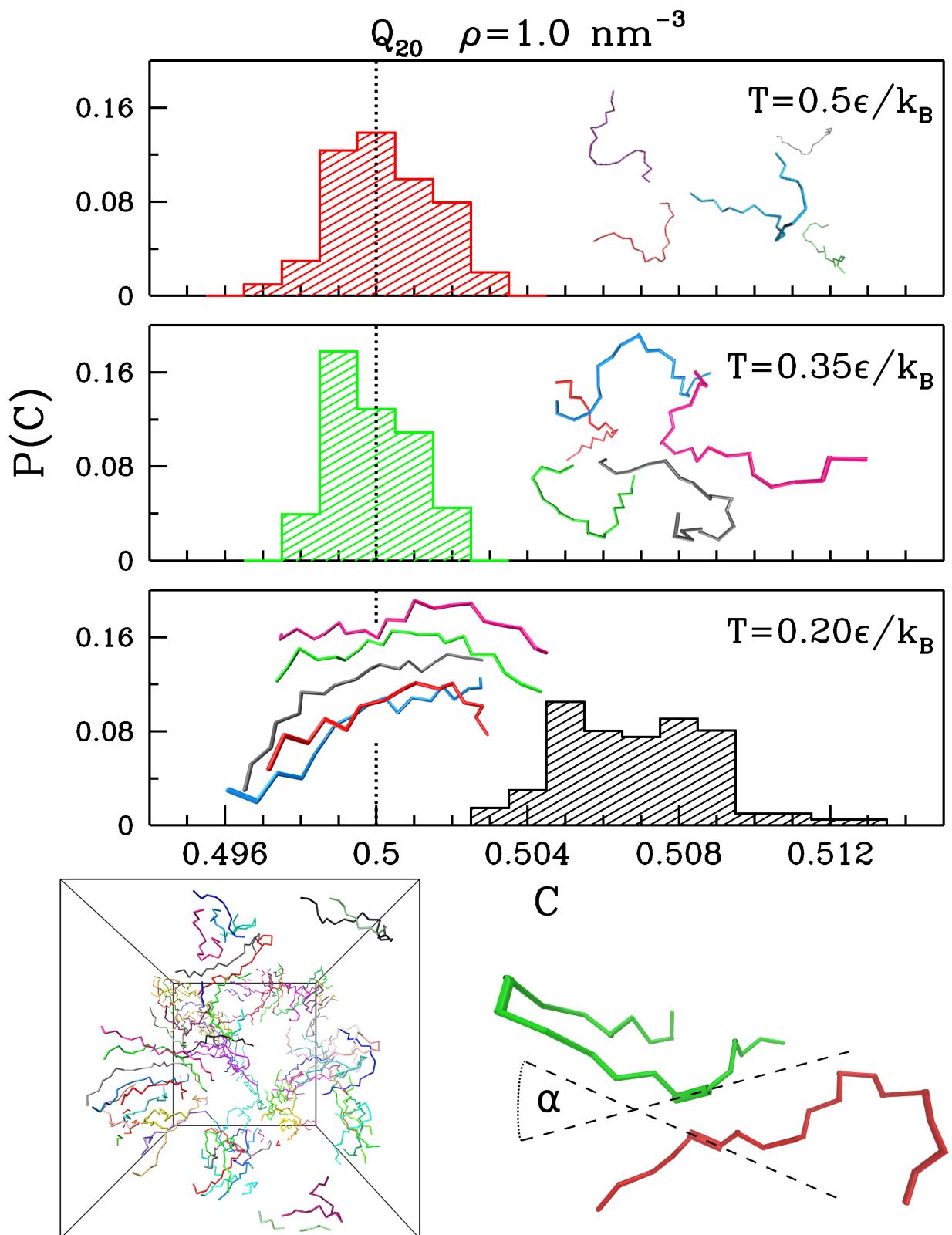
Dla dostatecznie wysokiej temperatury rozkład  $C$  jest w przybliżeniu symetryczny, z maksimum w  $\langle C \rangle = \frac{1}{2}$  (co odpowiada średniej wielkości  $|\cos \alpha|$  dla losowego<sup>3</sup> kąta  $\alpha$ ). Dla niższych temperatur rozkład zaczyna przesuwać się na prawo, co oznacza anizotropię (patrz panele a,b,c na Rys. 4.8).

---

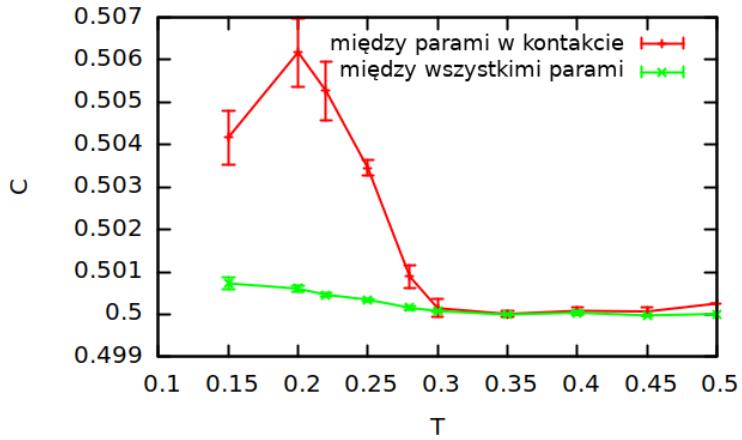
<sup>3</sup>"Losowy" kąt oznacza przypadkowe wzajemne ułożenie dwóch fragmentów, a zatem prawdopodobieństwo wybranego kąta  $\alpha$  nie jest określone rozkładem jednostajnym. Wykazanie, że faktycznie  $\langle C \rangle = \frac{1}{2}$  dla tego przypadku wymagało użycia programu Mathematica (za co dziękuję doktorantowi Nguyen Minh) i rozwiązania następującej całki:

$$\int_0^{2\pi} \int_0^{2\pi} \int_{-1}^1 \int_{-1}^1 \left| \sqrt{(1-a^2)(1-b^2)} \cos(x-y) + ab \right| \frac{da db dx dy}{16\pi^2} = \frac{1}{2}$$


Rysunek 4.7: Przykłady 3 agregatów przypominających włókna amyloidowe, uzyskane dla temperatury  $0.2 \epsilon/k_B$  i gęstości  $0.2 \text{ aa}/\text{nm}^3$ . Dwa agregaty po prawej pochodzą z tej samej klatki symulacji.



Rysunek 4.8: Panele a, b oraz c pokazują rozkłady wielkości  $C$  w układzie  $Q_{20}$  dla podanych temperatur i gęstości. Są ilustrowane konformacjami 5 bliskich sobie łańcuchów pochodzących z odpowiednich symulacji.  $C$  jest zdefiniowane przy pomocy kąta  $\alpha$  między fragmentami dwóch łańcuchów (co pokazuje panel e). Panel d pokazuje cały symulowany układ dla  $k_BT/\epsilon=0.2$ . Bok pudełka ma długość 12.2 nm. Modyfikacja rys. 4 z artykułu [IV].



Rysunek 4.9: Zależność parametru  $C$  uśrednionego po wszystkich możliwych parach fragmentów łańcuchów (zielony) i tylko między parami fragmentów łańcuchów w kontakcie (czerwony). Dane dla symulacji Q<sub>20</sub> o gęstości 1 aa/nm<sup>3</sup>.

Średnia wartość  $C$  zmienia się dużo mniej, jeśli zamiast tylko tych par łańcuchów które są ze sobą w kontakcie<sup>4</sup>, zostaną uwzględnione wszystkie możliwe pary łańcuchów (patrz Rys. 4.9). Potwierdza to, że anizotropia fazy A nie ma charakteru globalnego (inaczej wszystkie łańcuchy byłyby w pewnym stopniu równoległe do siebie), a jedynie lokalny (tylko łańcuchy w kontakcie ze sobą<sup>4</sup> są równoległe). To, że w fazie A współistnieją różnie zorientowane agregaty ilustruje panel d na Rys. 4.8.

Rys. 4.9 pokazuje zależność  $\langle C \rangle(T)$ . Ma ona kształt sigmoidalny, i można wyróżnić na niej dwie charakterystyczne temperatury: najniższą temperaturę dla której  $\langle C \rangle$  zaczyna różnić się od 1/2 ( $k_B T / \epsilon = 0.3$ ) oraz temperaturę odpowiadającą punktowi przegięcia sigmoidy ( $k_B T / \epsilon = 0.25$ ). Prowadzi to do 2 różnych kryteriów granicy fazy A: kryterium 1/2 i kryterium przegięcia. Można spodziewać się, że kształt sigmoidy wynika ze skończonych rozmiarów układu (tak jak dla modelu Isinga [212]): wtedy punkt przegięcia sigmoidy odpowiada temperaturze przejścia fazowego w granicy termodynamicznej.

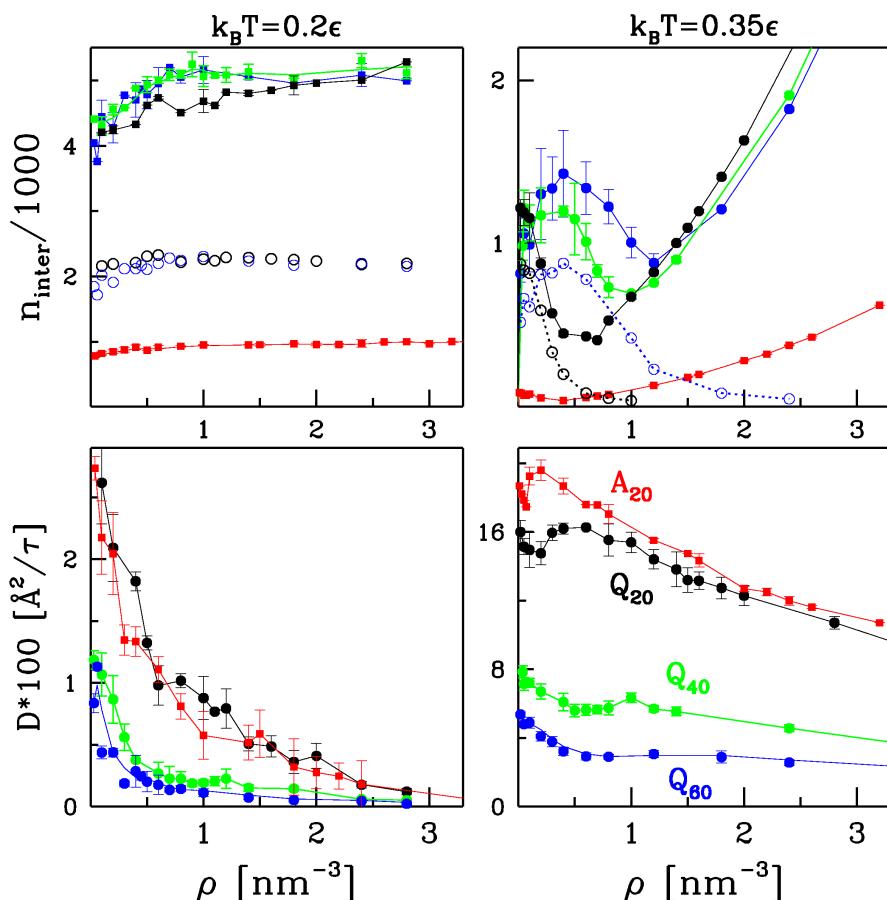
Faza A może być nazywana “szkłem spinowym” z dwóch powodów: globalna amorficzność agregatów, które lokalnie są izotropowe, oraz znaczne zmniejszenie współczynnika dyfuzji  $D$  (liczonego dla środków masy łańcuchów) w porównaniu do faz G oraz C. Na Rys. 4.10 współczynnik  $D$  spada monotonicznie w funkcji  $\rho$  dla każdej izotermy  $k_B T / \epsilon = 0.2$ , natomiast dla  $k_B T / \epsilon = 0.35$  spadek jest znacznie mniej stromy.

---

<sup>4</sup>Tzn. choć jedna para ich aminokwasów jest w odległości  $r < r_{min}^{ss}$  od siebie.

Dla  $T = 0.35 \epsilon/k_B$  i małych gęstości różnica między kontaktami “odległościowymi” i “dynamicznymi” na Rys. 4.10 jest niewielka, zaś dla dużych gęstości część aminokwasów zawsze będzie tworzyć kontakty “odległościowe” ze względu na duże zatłoczenie, które z kolei utrudnia spełnienie kryteriów kątowych (poprzez deformację łańcuchów). Może to tłumaczyć czemu liczba kontaktów najpierw rośnie wraz z  $\rho$ , potem spada (zbyt duże  $\rho$  utrudnia tworzenie kontaktów), a w przypadku kontaktów “odległościowych” ponownie rośnie. Ta niemonotoniczność ma pewne odzwierciedlenie we współczynnikach dyfuzji (więcej kontaktów między łańcuchami spowalnia dyfuzję).

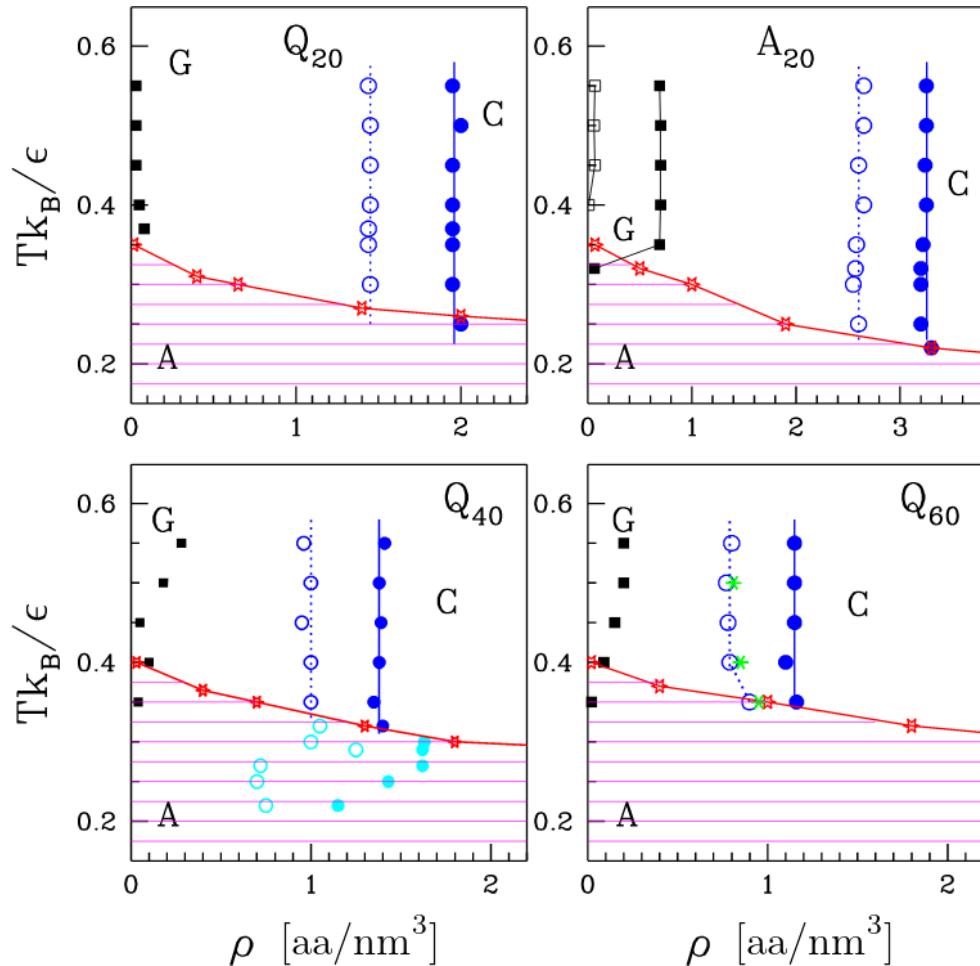
Dla  $T = 0.2 \epsilon/k_B$  każdy z badanych układów znajduje się w fazie A (niezależnie od tego czy zastosujemy kryterium przegięcia czy 1/2). Współczynnik dyfuzji  $D$  jest w tym przypadku o rząd wielkości mniejszy niż dla temperatury pokojowej. Gęstość nie jest na tyle duża aby układ zachowywał się jak ciało stałe, choć dla dużych gęstości  $D$  spada już prawie do zera.



Rysunek 4.10: Dolne panele pokazują współczynnik dyfuzji  $D$  w funkcji gęstości  $\rho$  dla układów A<sub>20</sub> (czerwony), Q<sub>20</sub> (czarny), Q<sub>40</sub> (zielony) i Q<sub>60</sub> (niebieski). Lewe panele odpowiadają  $T = 0.2 \epsilon/k_B$ , prawe  $T = 0.35 \epsilon/k_B$ . Górnne panele pokazują odpowiednie liczby kontaktów między łańcuchami. Ciągłe krzywe i pełne kółka odpowiadają kontaktom “odległościowym”, a przerywane krzywe i puste kółka (tylko dla Q<sub>20</sub> i Q<sub>60</sub>) “dynamicznym” (tzn. spełniającym kryteria kątowe, odległościowe i koordynacyjne). Modyfikacja rys. S7 z artykułu [IV].

#### 4.3.1.2 Diagramy fazowe

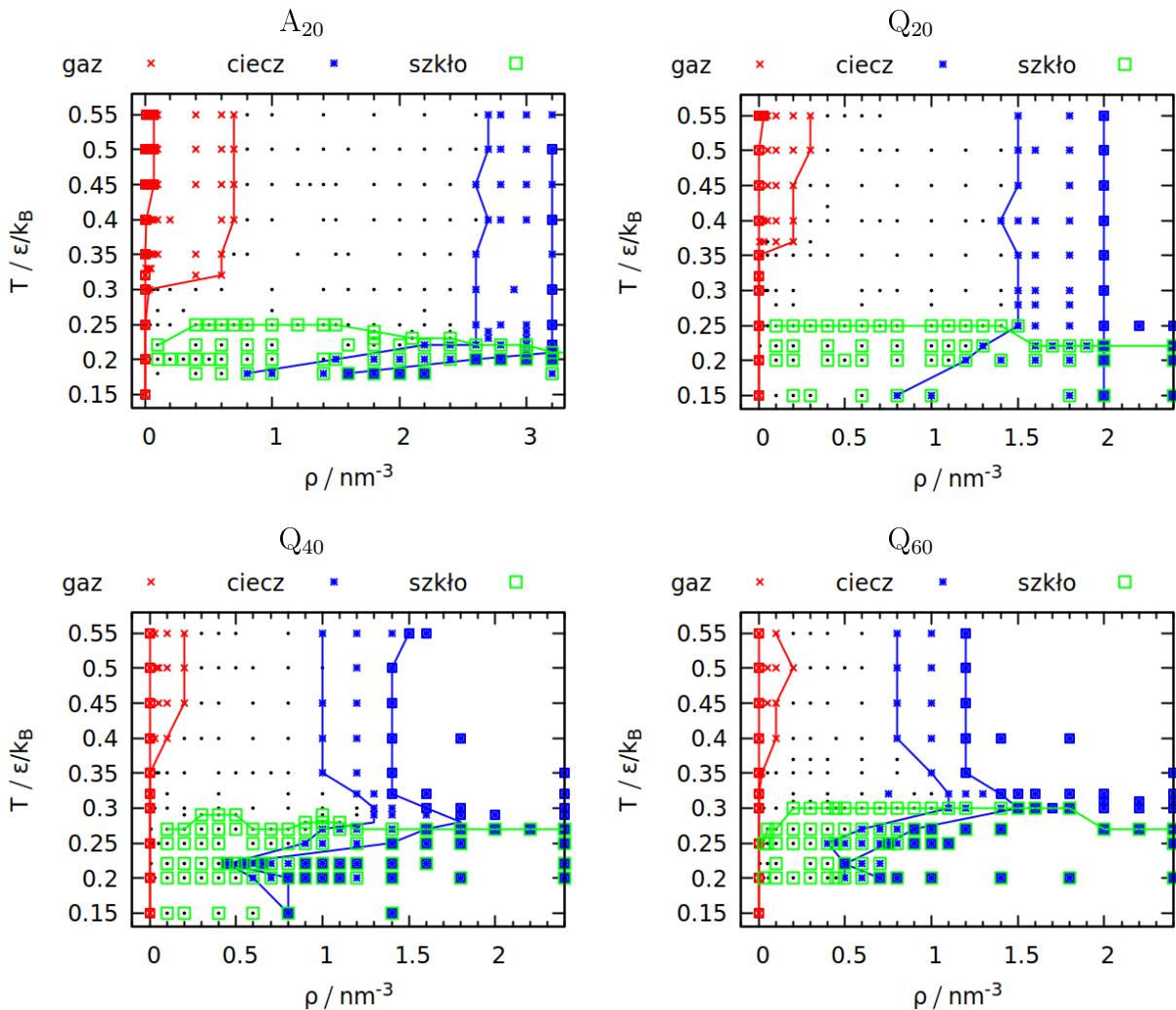
Najbardziej ogólny diagram fazowy pokazuje Rys. 4.11. Rozkład faz jest podobny dla każdego z 4 układów, ale granice między fazami przebiegają dla różnych gęstości: im większe  $N$ , tym mniejszy próg  $\rho$  powyżej którego pojawia się faza C (osiągnięcie fazy C dla  $A_{20}$  wymaga 3 razy większej gęstości niż dla  $Q_{60}$ ). Model quasi-adiabatyczny został sparametryzowany dla temperatury pokojowej i jest oparty na bazie kontaktów uzyskanych ze struktur, które w większości też odpowiadają temperaturze pokojowej. Symulacje powyżej temperatury  $0.6 \epsilon/k_B$  mogą być więc niemiarodajne, dlatego diagramy fazowe pokazują jedynie mniejsze temperatury.



Rysunek 4.11: Diagramy fazowe dla 4 badanych układów. Czarne kwadraty, pełne niebieskie kółka oraz czerwone gwiazdki określają odpowiednio granice faz G, C, oraz A (wg kryterium 1/2). Puste niebieskie kółka określają granicę B. Granice B oraz C są trudne do wyznaczenia dla niskich temperatur, dlatego nie są dla nich pokazane (z wyjątkiem morskich kółek dla  $Q_{40}$ ). Zielone gwiazdki oznaczają granicę B dla układu 2400 aminokwasów (zamiast 1800). Modyfikacja rys. 3 z artykułu [IV].

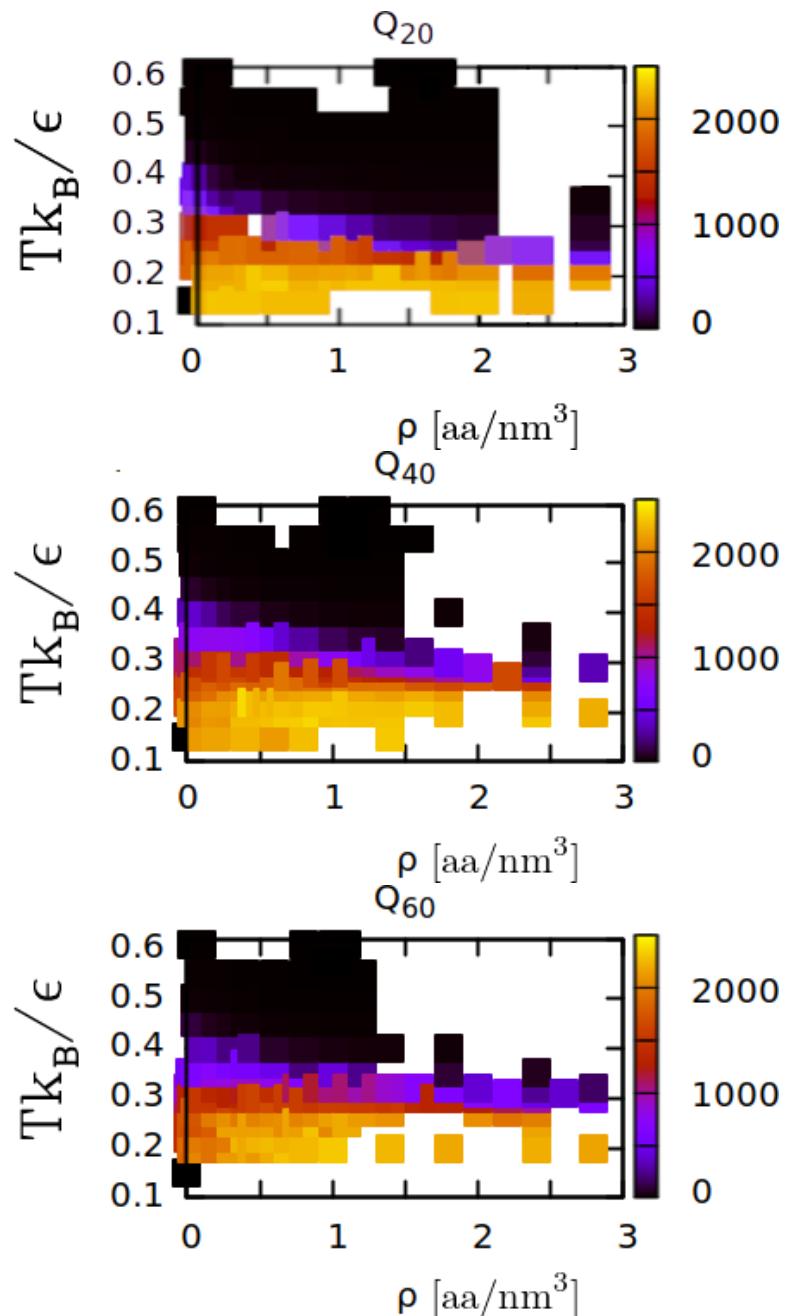
Rys. 4.11 nie pokazuje granic faz G i C poniżej granicy fazy A (wyznaczonej przez kryterium 1/2). Nie zawiera zatem pełnej informacji o diagramie fazowym. Dlatego zamieszczony został tu także Rys. 4.12, w którym każdy punkt odpowiada jednej kombinacji  $\rho, T$ . Dla każdej kombinacji przeprowadzone zostało parę symulacji, a na ich podstawie obliczone zostały uśrednione wielkości  $P(1)$ ,  $P(n_{max})$  oraz  $\langle C \rangle$ . Można dzięki nim dokładnie wykreślić granice faz, tym razem używając dla fazy A uproszczonego kryterium przegięcia sigmoidy: dla wszystkich układów jej górną granicą waha się między 0.504 a 0.508 (dolina to 0.5), dlatego kryterium przegięcia w przybliżeniu odpowiada  $\langle C \rangle \geq 0.502$ .

Gęstość zero odpowiada symulacji pojedynczego łańcucha (granica nieskończonego rozcieńczenia). Na Rys. 4.12 graniczna temperatura, przy której pojawia się faza A rośnie wraz z gęstością w obszarze  $\rho < 0.5 \text{ aa/nm}^3$ . Ten (bardzo niewielki) wzrost nie jest widoczny na Rys. 4.11, jednak może on sugerować, że tylko w tym zakresie gęstości agregacja jest limitowana dyfuzyjnie. Innym argumentem za tą tezą jest wzrost liczby kontaktów "dynamicznych" tylko dla  $\rho < 0.5$  na Rys. 4.10. Dla większych gęstości liczba ta jest stała ( $T = 0.2 \epsilon/k_B$ ) lub maleje ( $T = 0.35 \epsilon/k_B$ ).



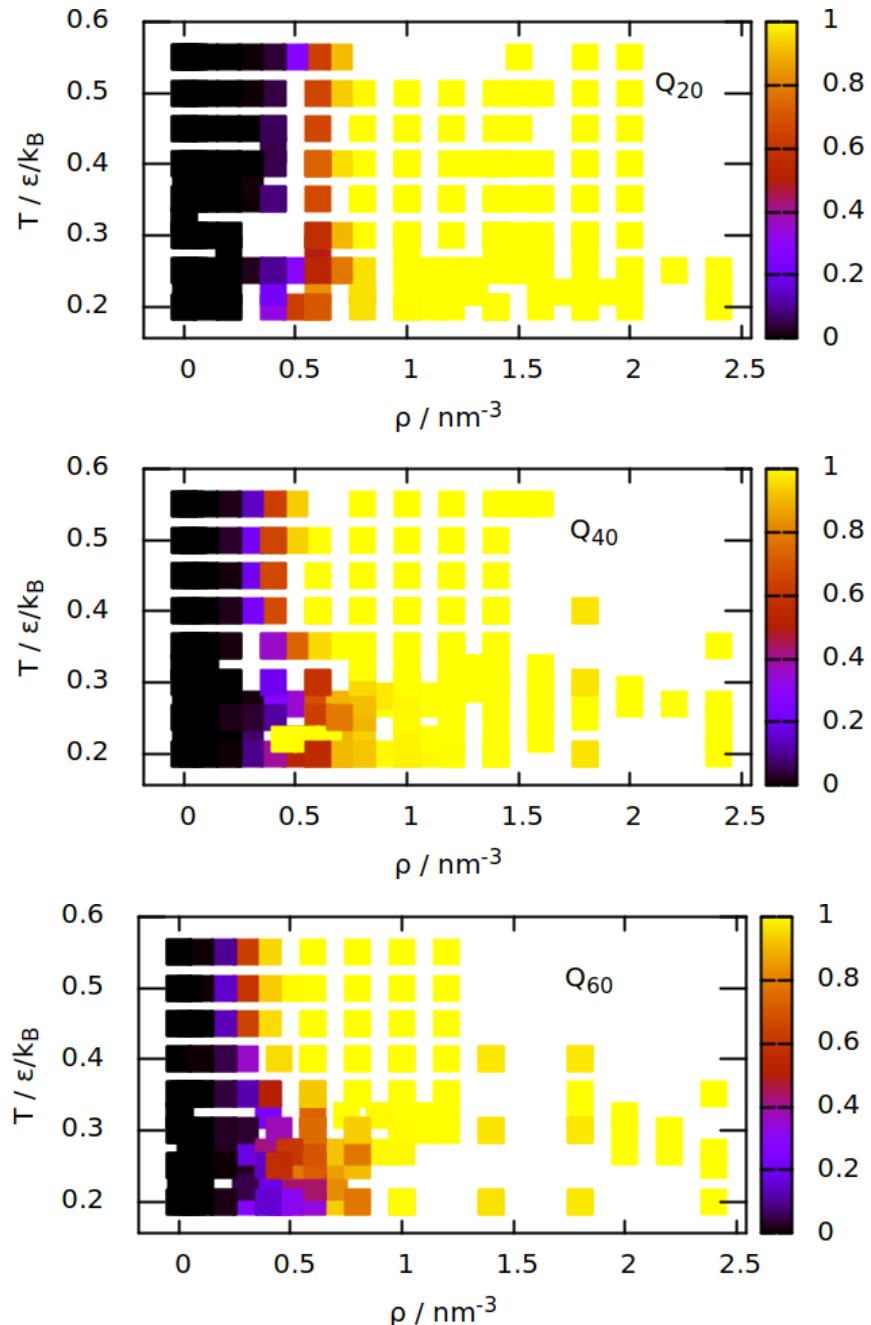
Rysunek 4.12: Diagramy fazowe pokazujące wszystkie symulowane układy i kombinacje  $\rho, T$  (poza tymi wychodzącymi poza obszar wykresu). Każdy punkt odpowiada jednej kombinacji. Czerwone krzyżyki odpowiadają  $\langle P(1) \rangle \geq 0.5$ , czerwone kwadraty  $\langle P(1) \rangle \geq 0.95$ , niebieskie gwiazdki  $\langle P(n_{max}) \rangle \geq 0.5$ , niebieskie kwadraty  $\langle P(n_{max}) \rangle \geq 0.95$ , a zielone kwadraty  $\langle C \rangle \geq 0.502$ .

Inną miarą agregacji łańcuchów jest liczba kontaktów “dynamicznych” między łańcuchami. Tylko te kontakty są opisywane potencjałem LJ z minimum o głębokości  $\epsilon$ , a więc tylko one odpowiadają za faktyczne przyciąganie się łańcuchów. Rys. 4.10 sugeruje, że liczba ta maleje dla dużych gęstości. Rys. 4.13 pokazuje, że wielkość ta zależy przede wszystkim od temperatury, a obszar dla którego kontaktów “dynamicznych” jest najwięcej odpowiada fazie A. Dlatego tylko dla fazy A można mówić o trwałych agregatach, w których białka utrzymywane są razem przez oddziaływanie przyciągające. W fazie C efektywne przyciąganie może wynikać ze splatania się łańcuchów (patrz paragraf 4.3.2.2).



Rysunek 4.13: Diagramy dla  $Q_{20}$ ,  $Q_{40}$  i  $Q_{60}$ , pokazujące w funkcji  $\rho, T$  liczbę kontaktów “dynamicznych” między łańcuchami. Każdej kombinacji  $\rho, T$  odpowiada prostokąt, którego kolor oznacza liczbę kontaktów. Prostokąty znajdują na siebie.

Jeszcze inną metodą wyznaczenia diagramu fazowego jest badanie prawdopodobieństwa perkolacji  $P_p$ , tzn. istnienia agregatu, który przechodzi przez dwie naprzeciwległe ściany pudełka do symulacji. To prawdopodobieństwo w funkcji  $\rho, T$  przedstawia Rys. 4.14. Jest ono bliskie 1 nawet po lewej stronie granicy B (gdzie  $\langle P(n_{max}) \rangle < 0.5$ ). Warto zauważyć, że  $P_p$  zależy od  $T$  niemonotonicznie, podobnie jak granice B i C na Rys. 4.12. Może to wynikać z tego, że dla dużych temperatur łańcuchy mają większe rozmiary (patrz paragraf 4.3.2.1), dla  $k_B T / \epsilon \approx 0.3$  są zwinięte i skupione w wielu małych agregatach (co przesuwa granice na prawo), a dla mniejszych temperatur agregują w duże agregaty amyloidowe (co ponownie przesuwa granicę na lewo).



Rysunek 4.14: Diagramy dla  $Q_{20}$ ,  $Q_{40}$  i  $Q_{60}$ , pokazujące prawdopodobieństwo perkolacji  $P_p$  w funkcji  $\rho, T$ . Każdej kombinacji  $\rho, T$  odpowiada prostokąt, którego kolor oznacza wartość  $P_p$ . Prostokąty nachodzą na siebie.

## 4.3.2 Właściwości pojedynczych łańcuchów i ich par

### 4.3.2.1 Pojedyncze łańcuchy

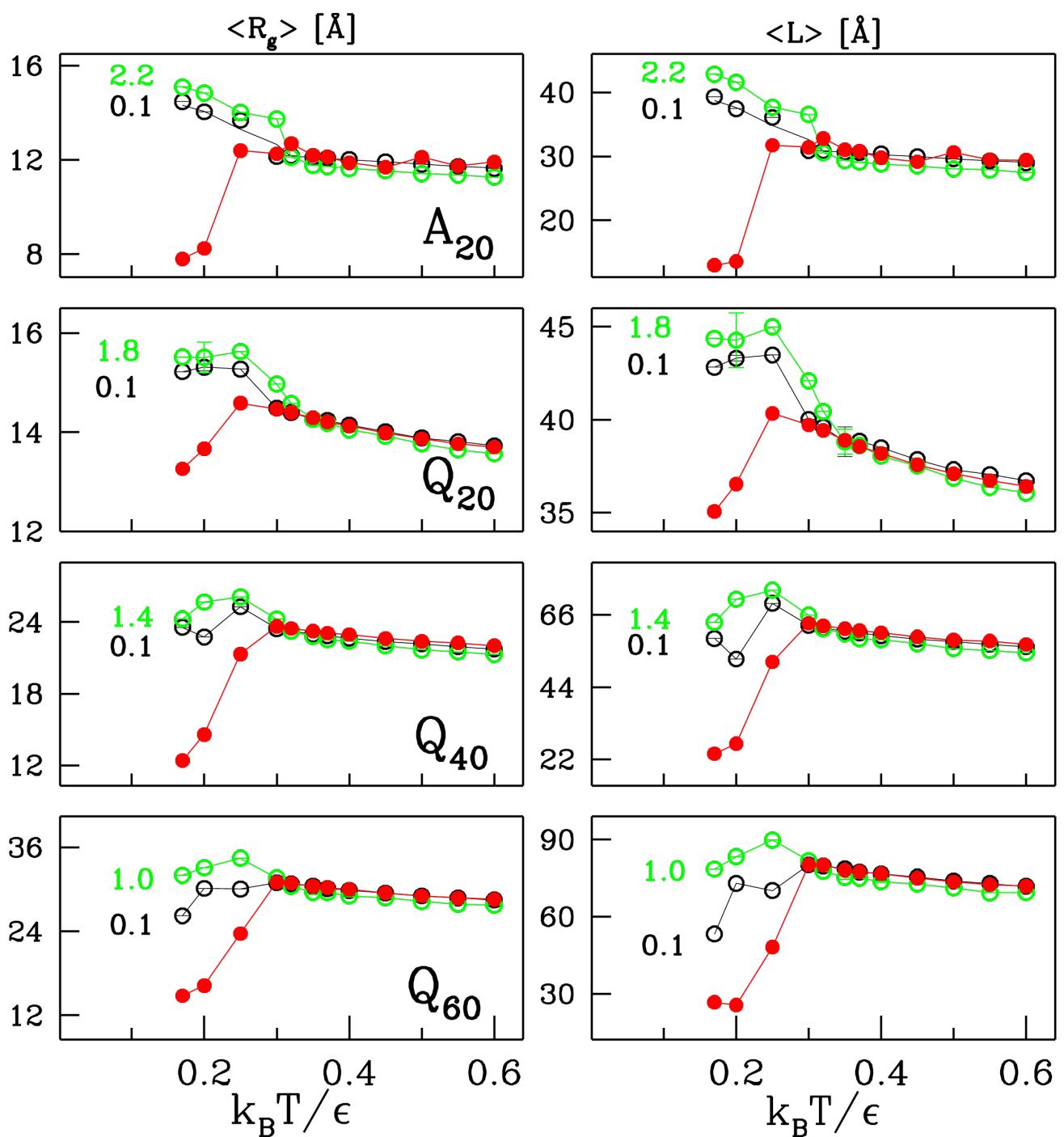
Kształt symulowanych białek zmienia się w zależności od warunków, jak pokazano na Rys. 4.15 dla średnich wartości promienia bezwładności  $R_g$  i odległości między końcami  $L$  w funkcji temperatury.

Czerwone krzywe na Rys. 4.15 odpowiadają simulacjom pojedynczych łańcuchów i wykazują dużą zależność od  $T$  (zwłaszcza w pobliżu temperatury pokojowej). Czarne krzywe odpowiadają fazie G, zielone fazie pośredniej między granicą B a granicą C (wielkości są uśrednione po wszystkich łańcuchach). Zależność od  $T$  jest dużo słabsza, a różnice między krzywymi wskazują na duży wpływ obecności innych białek na kształt badanych homopeptydów.

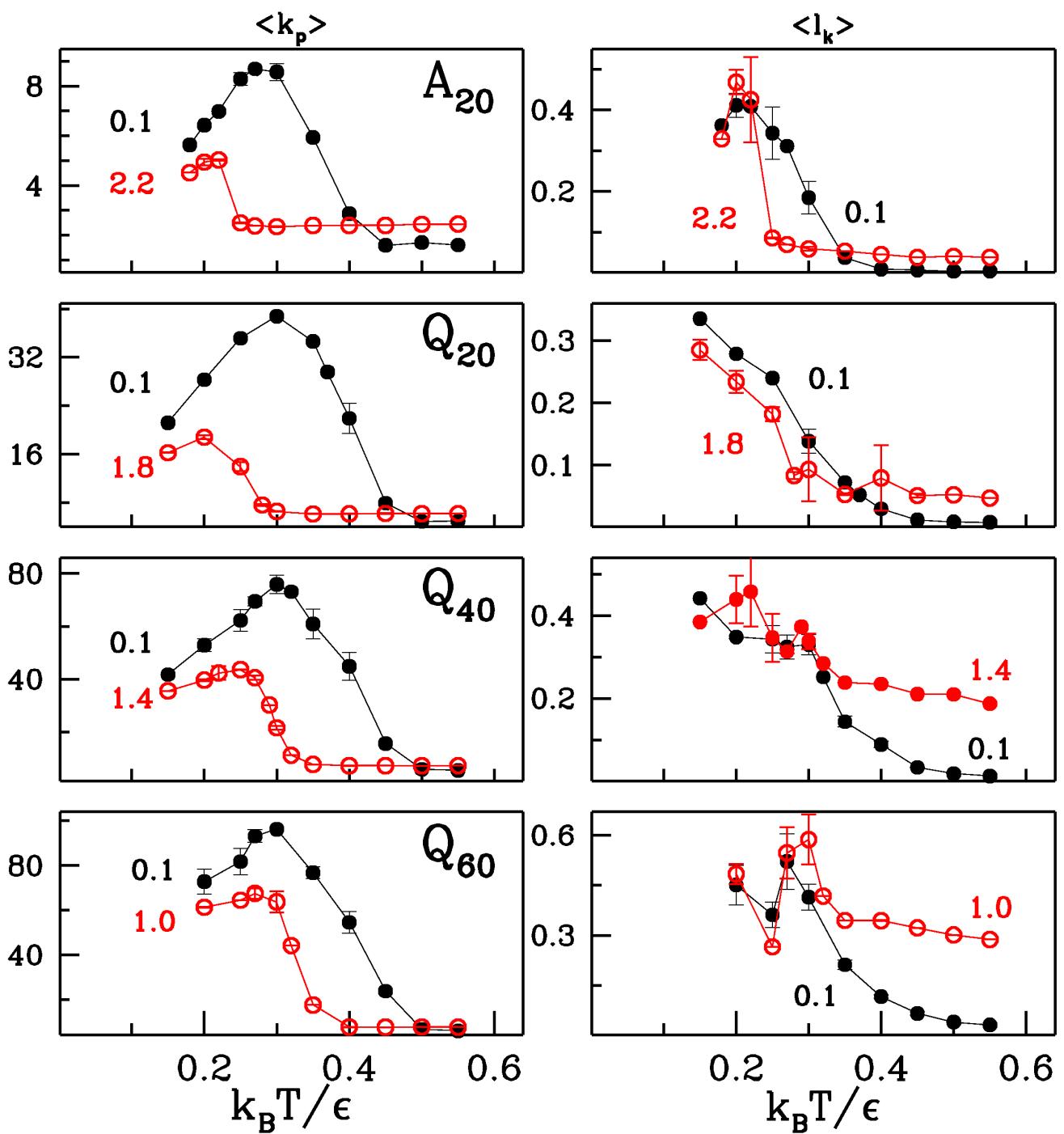
### 4.3.2.2 Pary

Rys. 4.16 opisuje właściwości charakteryzujące pary łańcuchów które są ze sobą w kontakcie (“odległościowym”).  $\langle k_p \rangle$  to liczba takich kontaktów uśredniona po wszystkich parach we wszystkich klatkach każdej symulacji. Maksimum  $\langle k_p \rangle$  przypada dla  $k_B T / \epsilon$  pomiędzy 0.35 a 0.3 (w zależności od  $n_m$ ), spadek dla wyższych temperatur jest spowodowany zanikiem fazy A i zrywaniem kontaktów “dynamicznych” (patrz Rys. 4.13). Co ciekawe, dla niskich temperatur  $\langle k_p \rangle$  jest niższe niż na samej granicy fazy A. Możliwe że stykanie się łańcuchów w uporządkowanych, “płaskich” agregatach utrzymywanych głównie przez kontakty typu bb skutkuje mniejszą liczbą kontaktów niż w agregatach amorficznych. Inną możliwością jest zachodzenie “zimnej denaturacji” dla tak niskich temperatur.

$\langle l_k \rangle$  oznacza średnią liczbę splątań między parą łańcuchów [117], obliczoną algorytmem [117]. Splątanie zachodzi, gdy łamana łącząca pierwszy i ostatni aminokwas w łańcuchu nie może być skrócona do jednego odcinka bez przecinania innej takiej łamanej [118, 119]. Po zminimalizowaniu długości wszystkich łamanych (z warunkiem nieprzecinania), każde skrzyżowanie łamanych oznacza jedno splątanie [120]. Liczba splątań spada wraz ze wzrostem  $T$ : spadek jest monotoniczny dla Q<sub>20</sub> i Q<sub>40</sub>, natomiast dla A<sub>20</sub> i Q<sub>60</sub> występuje maksimum  $\langle l_k \rangle$  na granicy fazy A (dla niskich temperatur agregaty zmieniają się z amorficznych w uporządkowane i splątań jest mniej). Dla  $N > 20$  i  $\rho > 0.5$  aa/nm<sup>3</sup> liczba splątań nie spada do zera dla dużych temperatur. Splątania te mogą wtedy powodować efektywne przyciąganie między łańcuchami, mimo braku kontaktów “dynamicznych”.



Rysunek 4.15: Właściwości pojedynczych łańcuchów dla 4 symulowanych układów w funkcji temperatury dla zaznaczonych gęstości (w jednostkach aa/nm<sup>3</sup>). Średni promień bezwładności  $R_g$  jest po lewej, średnia odległość między końcami łańcucha  $L$  po prawej. Oparte na rys. S8 z artykułu [IV].



Rysunek 4.16: Właściwości par łańcuchów dla 4 symulowanych układów w funkcji temperatury dla zaznaczonych gęstości (w jednostkach  $\text{aa}/\text{nm}^3$ ). Po lewej średnia liczba kontaktów “odległościowych” między parą łańcuchów w kontakcie  $k_p$ , po prawej średnia liczba splatań na parę łańcuchów  $l_k$ . Oparte na rys. S9 z artykułu [IV].

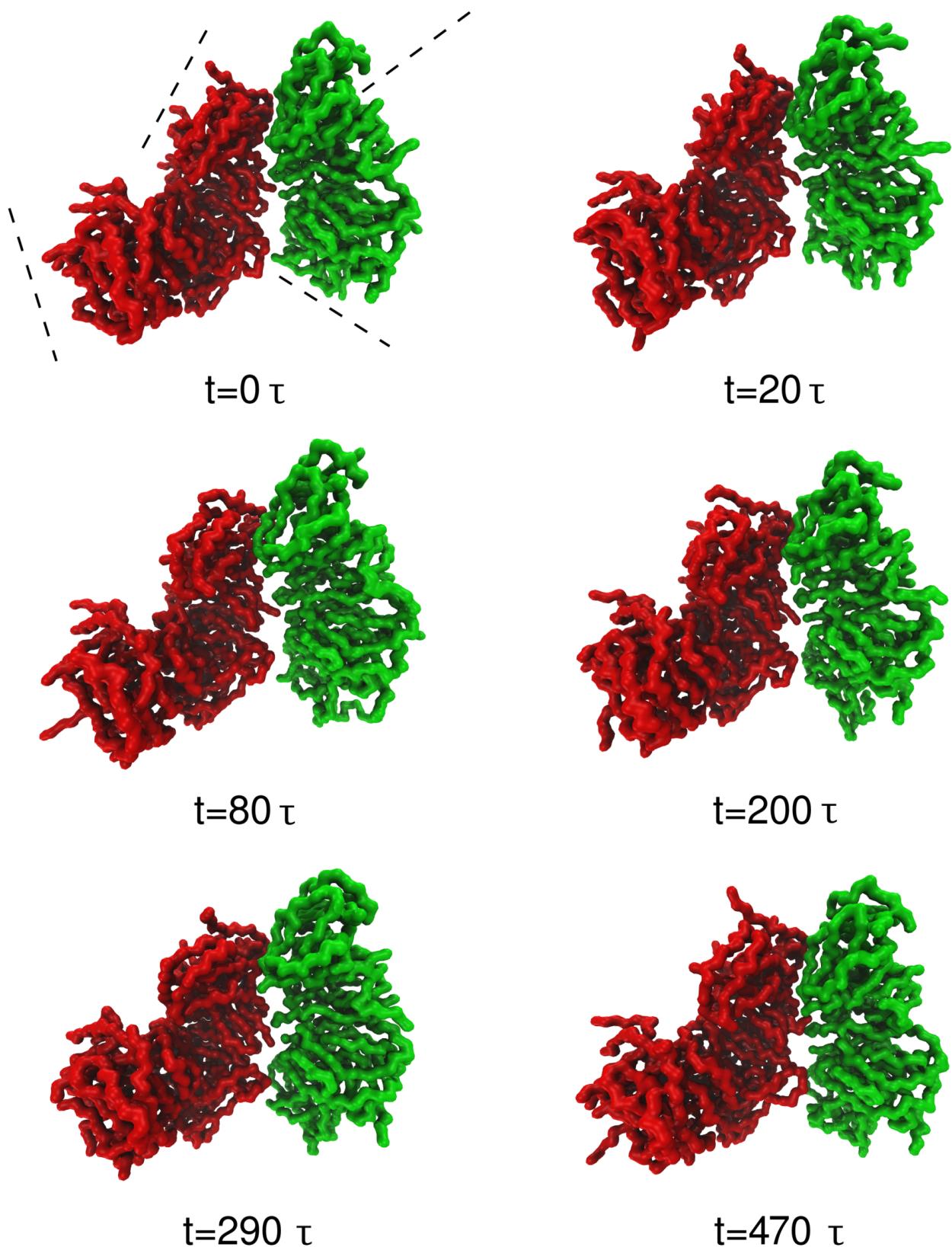
### 4.3.3 Dynamika łączenia się i rozpadu klastrów

Mały rozmiar symulowanych układów uniemożliwia badanie dynamiki dugożyjących kropel. Można jednak badać procesy łączenia się i rozpadania dużych agregatów w obszarze na lewo od granicy C. Rys. 4.17 i 4.18 pokazują przykłady takich procesów dla Q<sub>60</sub>. Rys. 4.17 ilustruje łączenie się dwóch agregatów w fazie A. Natomiast Rys. 4.18 pokazuje rozpad agregatu dla temperatury pokojowej. Oba rysunki zostały otrzymane dzięki nierównowagowym symulacjom 80 łańcuchów Q<sub>60</sub>, początkowo zgrupowanych w dwa agregaty po 40 łańcuchów, każdy pochodzący z symulacji równowagowej w pudełku o gęstości  $\rho_{int} = 1.1 \text{ aa/nm}^3$  (która można traktować jako gęstość jednego agregatu). Pudełko w którym zostały umieszczone oba agregaty miało efektywną gęstość  $\rho_{eff} = 0.3 \text{ aa/nm}^3$ . Dzięki takim warunkom początkowym można badać dynamikę większych klastrów niż podczas symulacji równowagowych.

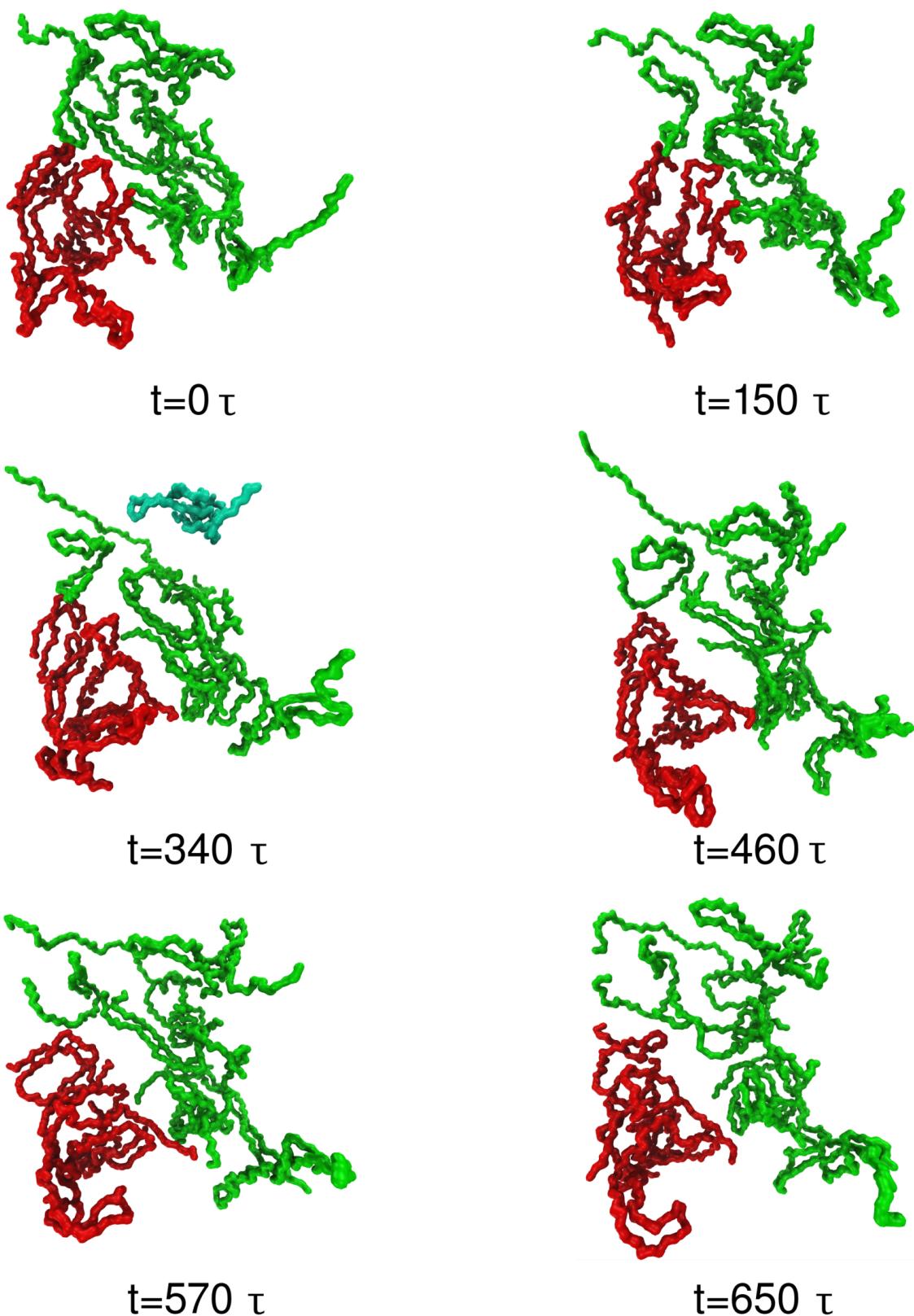
Proces łączenia na Rys. 4.17 zaczyna się w ciągu pierwszych  $20\tau$  od początkowego stanu w którym klastry były rozdzielone (brak kontaktów “odległościowych” ani “dynamicznych”). Połączenie nastąpiło w ciągu  $470\tau$ , a więc bardzo szybko w porównaniu do skali czasowej całej symulacji. “Orientacja” obu agregatów (zaznaczona na Rys. 4.17 przerywanymi liniami) nie zmieniła się mimo ich połączenia.

Proces rozpadu zachodzi w ciągu  $650\tau$  dla przypadku z Rys. 4.18. W trakcie rozpadu niewielkie agregaty oddzielają się i ponownie przyłączają do dwóch dużych agregatów. Pokazuje to, jak dynamiczne są procesy “parowania” i “skraplania” łańcuchów w symulacji. Te same procesy w dłuższych skalach czasowych są pokazane na Rys. 4.19 i Rys. 4.20.

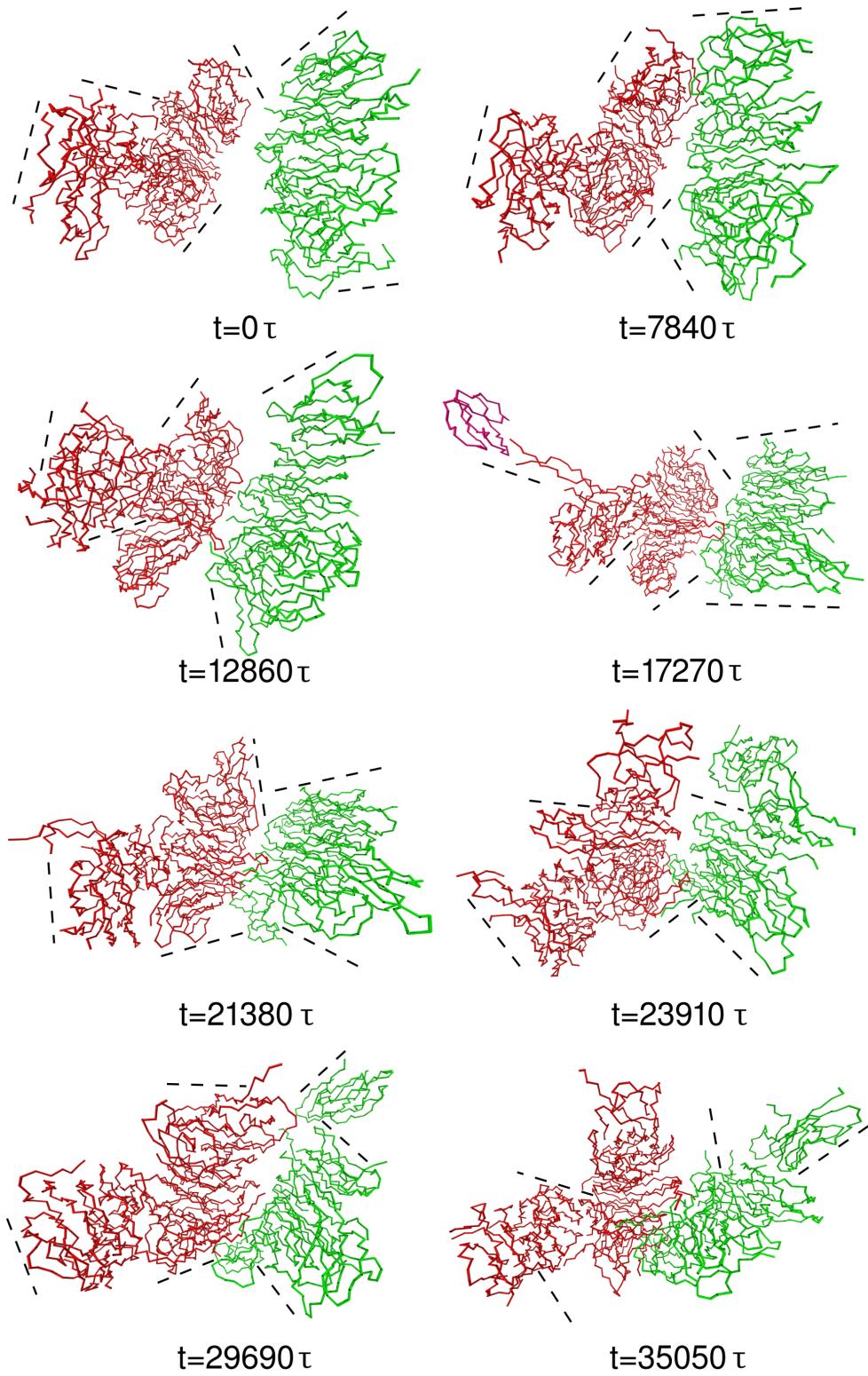
Warto zauważyć że przedstawione tu procesy dotyczą na tyle niewielkiej liczby łańcuchów, że nie można traktować ich jako hydrodynamicznego łączenia i dzielenia się kropel, a tempo wymiany łańcuchów jest bardzo duże. Dla cieczy van der Waalsa właściwości hydrodynamiczne łączenia i dzielenia się kropel można zaobserwować już dla układów o rozmiarach rzędu 10 nm [213] (około 400 cząsteczek). W przypadku białek podobne rozmiary obejmują mniej niż 100 łańcuchów, a zatem granica powyżej której pojawiają się efekty hydrodynamiczne wydaje się zależeć bardziej od liczby swobodnych cząsteczek niż od bezwzględnych rozmiarów układu. Najmniejsze krople złożone z białek dla których badano efekty hydrodynamiczne miały średnice rzędu mikrometra [66], a więc około 100 razy większe niż agregaty rozpatrywane tutaj. Dignon i in. [69] sprytnie obeszli ten problem, rozszerzając pudełko do symulacji tylko w jednym kierunku, co pozwoliło na badanie większych skal odległości przy stosunkowo niewielkim zwiększeniu liczby białek w układzie.



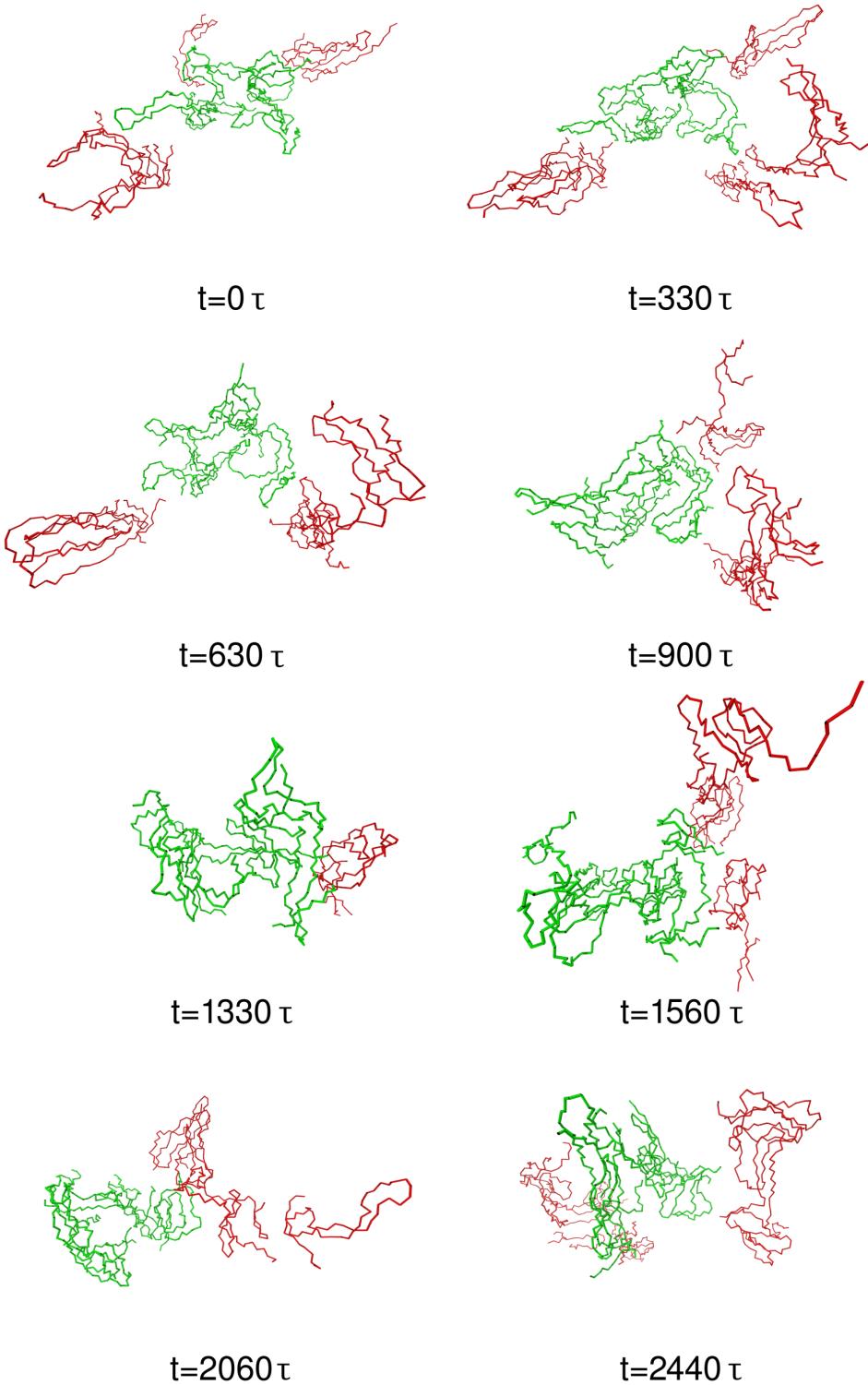
Rysunek 4.17: Proces łączenia się 2 agregatów Q<sub>60</sub> w fazie A. Czerwony i zielony agregat składają się odpowiednio z 18 oraz 12 łańcuchów. Przerywane linie na lewym górnym panelu pokazują lokalną orientację łańcuchów. Modyfikacja rys. 6 z artykułu [IV].



Rysunek 4.18: Proces rozpadu 2 agregatów  $Q_{60}$  w fazie pośredniej. Czerwony agregat składa się z 7 łańcuchów  $Q_{60}$ . Liczba łańcuchów w zielonym agregacie zmienia się między 9 a 15 w trakcie procesu, ponieważ zachodzą w nim inne procesy rozpadania i łączenia (jeden z łańcuchów, który odłączył się tylko na jednym panelu, jest pokazany w kolorze morskim). Modyfikacja rys. 7 z artykułu [IV].



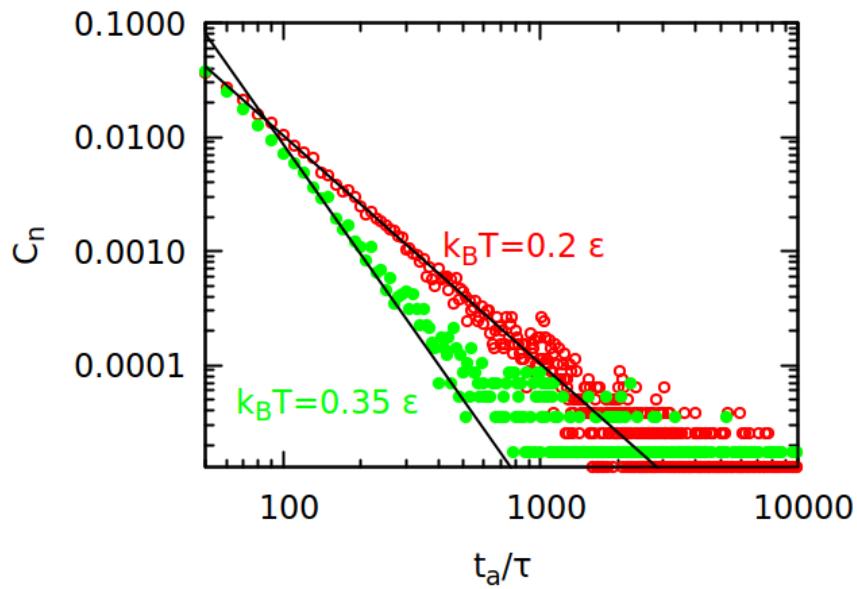
Rysunek 4.19: Ewolucja w czasie dwóch agregatów  $Q_{60}$  w ciągu  $35\ 050 \tau$ . Temperatura  $T = 0.2 \epsilon/k_B$ . Na pierwszym panelu zielony i czerwony agregat nie są ze sobą połączone. Na ostatnim panelu są już jednym agregatem (jednak na wszystkich panelach są pokolorowane wg podziału z 1. panelu). Pierwsze połączenie między klastrami, które zaszło pomiędzy pierwszym i drugim panelem, zostało pokazane na Rys. 4.17. Na czwartym panelu (po prawej), niewielki agregat (fioletowy) oddzielił się na krótko od czerwonego, aby ponownie stać się jego częścią na piątym panelu. Modyfikacja rys. S10 z artykułu [IV].



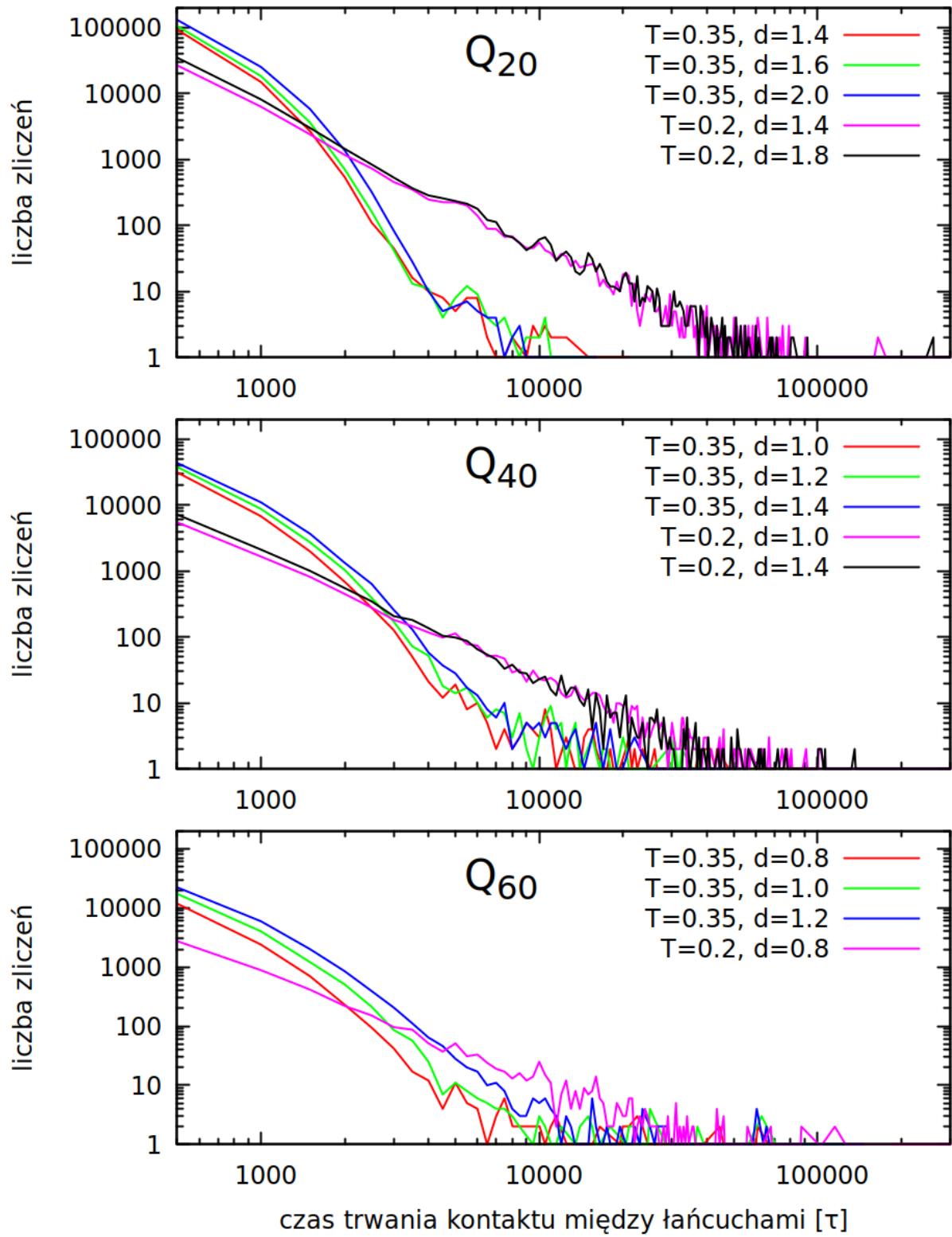
Rysunek 4.20: Ewolucja w czasie dwóch agregatów  $Q_{60}$  w ciągu  $2\ 440\ \tau$ . Temperatura  $T = 0.35\ \epsilon/k_B$ . Rdzeń zielonego klastra tworzy 9 łańcuchów, które zawsze są razem, podczas gdy inne łańcuchy (czerwone) oddzielają się i powracają do agregatu w serii procesów fuzji i rozpadów. Modyfikacja rys. S11 z artykułu [IV].

#### 4.3.3.1 Prawa potęgowe w dynamice agregatów

Podziałom łączenia i podziału dużych agregatów towarzyszy dołączanie i odłączanie się od nich mniejszych agregatów i pojedynczych łańcuchów (patrz Rys. 4.20). Sugeruje to pewne samopodobieństwo tych procesów (duże agregaty dzielą się na mniejsze, a te na jeszcze mniejsze), któremu okazuje się odpowiadać prawo potęgowe. Dotyczy ono czasów życia kontaktów między łańcuchami. Nie są to jednak pojedyncze kontakty między aminokwasami. Kontakt między łańcuchami trwa tak długo, jak długo przynajmniej jedna para ich aminokwasów jest ze sobą w kontakcie (“odległościowym”). Rys. 4.21 pokazuje histogramy czasu życia tak zdefiniowanych kontaktów w układzie 80 łańcuchów Q<sub>60</sub> dla  $T = 0.2$  oraz dla  $0.35 \epsilon/k_B$ . Wykładnik prawa potęgowego (odpowiednio 2.0 i 3.2 dla powyższych temperatur) zależy od temperatury, natomiast prawie nie zależy od gęstości (patrz Rys. 4.22 i 4.23). Rys. 4.21 został uzyskany dla symulacji nierównowagowych (opisanych na początku podpunktu 4.3.3), natomiast Rys. 4.22 dla symulacji równowagowych. Prawo potęgowe działa jedynie w ograniczonym zakresie czasów życia. Szum dla największych czasów wynika ze skończonych rozmiarów układu, natomiast odchylenia dla najmniejszych czasów są skutkiem niedoskonałego próbkowania układu (mapy kontaktów “odległościowych” są wyznaczane co  $10 \tau$  dla Rys. 4.21 i co  $500 \tau$  dla Rys. 4.22).



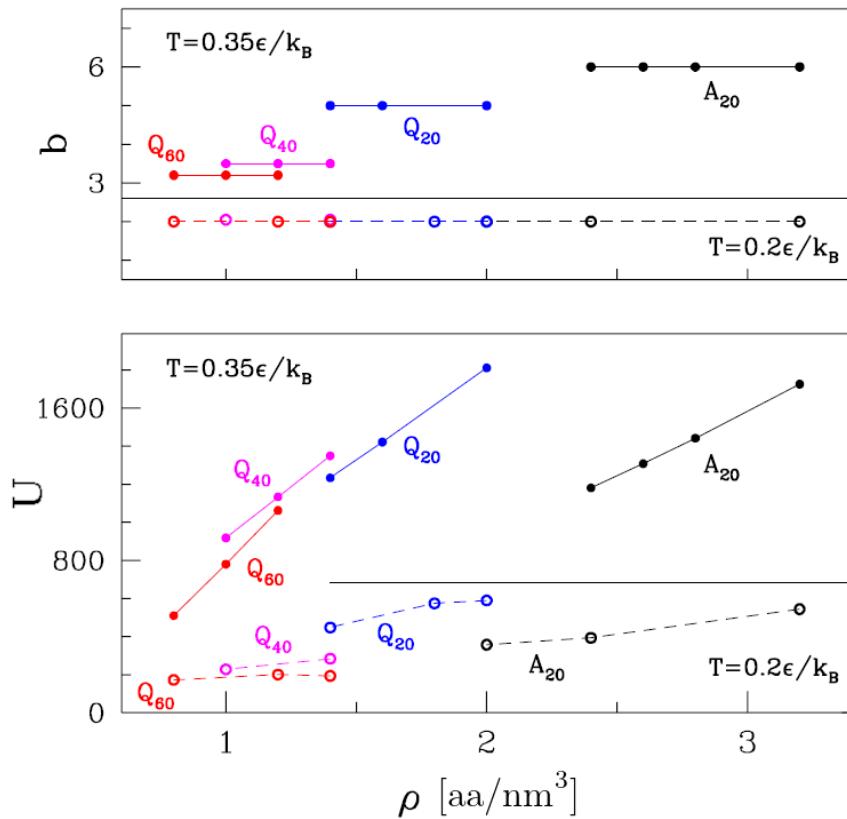
Rysunek 4.21: Histogramy czasu  $t_a$  trwania kontaktu między łańcuchami z szerokością słupka  $10 \tau$  dla temperatur  $k_B T = 0.2 \epsilon$  (puste kółka) oraz  $k_B T = 0.35 \epsilon$  (pełne kółka). Dwa łańcuchy są w kontakcie jeśli między ich aminokwasami jest choć jeden kontakt “odległościowy”. Histogramy są znormalizowane przez całkowitą liczbę zliczeń (78867 dla  $k_B T = 0.2 \epsilon$  oraz 57268 dla  $k_B T = 0.35 \epsilon$ ). Rozkłady pochodzą z ostatnich  $90\ 000 \tau$  nierównowagowej symulacji 80 łańcuchów Q<sub>60</sub>. Wykładniki dopasowanych funkcji potęgowych to -2 dla  $k_B T = 0.2 \epsilon$  oraz -3.2 dla  $k_B T = 0.35 \epsilon$ . Oparte na rys. S12 z artykułu [IV].



Rysunek 4.22: Histogramy czasu  $t_a$  trwania kontaktu między łańcuchami z szerokością słupka 500  $\tau$  dla podanych temperatur i gęstości (odpowiednio w jednostkach  $\epsilon/k_B$  i aa/nm<sup>3</sup>). Rozkłady pochodzą z ostatnich 500 000  $\tau$  równowagowych symulacji 30 łańcuchów Q<sub>60</sub>, 45 łańcuchów Q<sub>40</sub> lub 90 łańcuchów Q<sub>20</sub>.

Łączna liczba zliczeń na histogramach takich jak na Rys. 4.22 zawiera informację o tym, jak często pojawiał się kontakt między łańcuchami (czas symulacji dla wszystkich układów na Rys. 4.22 jest taki sam). Układy różniły się jednak liczbą łańcuchów, dlatego liczba zliczeń musi być znormalizowana, aby można było porównywać ją między układami. Liczba zliczeń podzielona przez liczbę łańcuchów jest oznaczona jako  $U$  i pokazana w funkcji gęstości na Rys. 4.23. Im większa gęstość, tym większe  $U$ , ponieważ szanse na zderzenie między łańcuchami są większe (z tych samych powodów  $U$  jest większe dla  $k_B T/\epsilon=0.35$  niż dla 0.2). To, że  $U$  jest największe dla krótkich łańcuchów sugeruje, że liczba zliczeń zależy od liczby łańcuchów w potędze większej niż jeden. W końcu dla  $n_m$  łańcuchów możliwe do utworzenia jest  $n_m(n_m - 1)/2$  kontaktów między nimi. Zbadanie zależności liczby zliczeń od  $n_m$  wymagałoby symulacji większej liczby układów o różnych  $n_m$ .

Rys. 4.23 przedstawia także wykładnik  $b$  dla temperatury pokojowej i dla temperatury w której występuje wyłącznie faza A. Wykładnik nie zależy od gęstości, i dla  $k_B T/\epsilon=0.2$  wynosi 2 niezależnie od układu, co prawdopodobnie wynika z właściwości fazy A (dalsza dyskusja w [VIII]). Mały wykładnik oznacza, że duża część histogramu liczby zliczeń przypada na długie czasy życia. Tłumaczy to dlaczego  $b$  dla  $k_B T/\epsilon=0.35$  jest mniejsze dla długich łańcuchów: ich dynamika jest wolniejsza niż krótkich, zatem ich kontakty mają dłuższe czasy życia.



Rysunek 4.23: Znormalizowana liczba zliczeń  $U$  oraz wykładnik  $b$  w funkcji gęstości dla temperatur  $k_B T/\epsilon=0.2$  oraz 0.35 (rozdzielone linią ciągłą). Wykres opiera się na ostatnich 500 000  $\tau$  równowagowych symulacji 30 łańcuchów  $Q_{60}$ , 45 łańcuchów  $Q_{40}$  lub 90 łańcuchów  $Q_{20}$  i A<sub>20</sub>. Modyfikacja rys. 5 z rozdziału w książce [VIII].

## 4.4 Podsumowanie

Wyznaczone zostały diagramy fazowe dla 4 wybranych układów złożonych z homopeptydów o różnej długości. Diagramy okazały się być jakościowo podobne do siebie.

Peskett i in. [84] zidentyfikowali doświadczalnie dwa rodzaje kropel złożonych z łańcuchów Q<sub>25</sub>: “ciemne” i “jasne” (te drugie charakteryzujące się dużo mniejszą ruchliwością łańcuchów). Faza “jasna” może odpowiadać opisanej tu fazie “szkła amyloidowego” (A), która została jakościowo opisana i ilościowo zanalizowana (w szczególności odpowiadał jej dużo mniejszy współczynnik dyfuzji  $D$ , tak samo jak dla fazy “jasnej” w [84]). Faza A nie mogłaby powstać w modelu Dignona i in. [69], ponieważ kontakty między aminokwasami są w nim wyłącznie izotropowe (a użyty model quasi-adiabatyczny korzysta z anizotropowych kryteriów kierunkowych).

Inna topologia wykresów fazowych świadczy o tym, że badane wielkości (prawdopodobieństwo przynależności do klastra o danej wielkości  $P(n)$ , średni moduł cosinusa kąta między fragmentami łańcuchów w kontakcie  $\langle C \rangle$ , prawdopodobieństwo perkolacji  $P_p$  i liczba kontaktów  $n_{inter}$ ) opisują agregację białek w innym ujęciu niż model separacji dwóch ciekłych faz. Podejście to można zastosować do układów zbyt małych, aby zachodziły w nich efekty hydrodynamiczne takie jak pojawienie się kropel o wyraźnie zaznaczonej powierzchni (w przeprowadzonych simulacjach agregaty wymieniały się łańcuchami na tyle szybko i były na tyle małe, że nie było możliwe wyznaczenie powierzchni rozdzielającej fazę gęstą i rzadką)<sup>5</sup>.

Obecność innych białek w zatłoczonym środowisku okazała się znaczco wpływać na kształt łańcuchów (opisany promieniem bezwładności  $R_g$  i odlegością między końcami  $L$ ), a łańcuchy w agregatach okazały się być splecione ze sobą.

Badania doktoranta Pedro Carvalho ze Środowiskowego Laboratorium Fizyki Biologicznej wykazały, że skonstruowanie takich samych diagramów fazowych (także opartych o kontakty między łańcuchami i kryteria zdefiniowane w podpunkcie 4.3.1) dla cieczy van der Waalsa daje wykresy o bardzo podobnej topologii do tych przedstawionych w tym rozdziale. W szczególności granice faz C i G nie schodzą się w jednym punkcie nawet dla bardzo wysokich temperatur. Potwierdza to, że podejście zastosowane tutaj różni się od podejścia Dignona i in. [69] i jest nowym sposobem opisu agregacji białek nieuporządkowanych.

---

<sup>5</sup>Do wyznaczenia powierzchni agregatów został wykorzystany program Spaceball [115], jednak powierzchnia ta zmieniała się na tyle szybko, że dane były zbyt zaszumione aby prowadzić do interesujących wyników.

# Rozdział 5

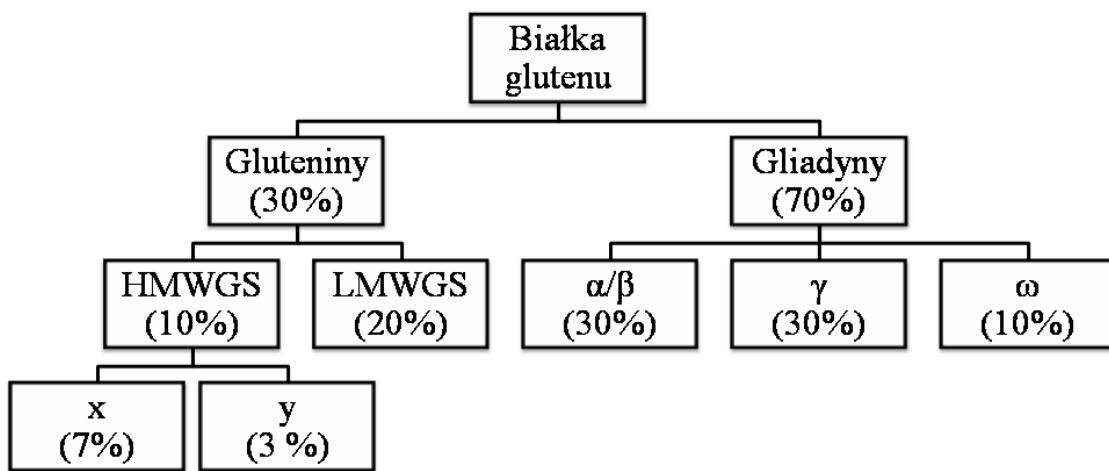
## Symulacje białek glutenu, kukurydzy i ryżu

### 5.1 Wprowadzenie

Gluten składa się z nierozpuszczalnych w wodzie białek pochodzących z ziaren pszenicy (*Triticum spp.*). Można otrzymać go poprzez wymycie wodą z mąki pszennej pozostałych składników (głównie skrobi i rozpuszczalnych białek globularnych) [137]. Jeśli wstawimy ciasto z wody i mąki pod strumień bieżącej wody, ponad 75% masy tego co zostanie w sitku stanowią właśnie białka [214]. Czasami za gluten uznaje się wszystkie składniki mąki, które nie rozpuszczają się w wodzie, jednak w tej rozprawie za definicję glutenu stanowią wyłącznie nierozpuszczalne białka ziaren pszenicy.

Podczas wyrabiania ciasta białka glutenu stanowią mniej niż 20% jego masy, lecz to właśnie niezwykłe właściwości lepkospłejyste tych białek są kluczowe dla jakości powstałego pieczywa [90, 215]. Białka glutenu zwykle mają sekwencje o niskiej złożoności, zawierające wiele powtarzających się fragmentów bogatych w glutaminę i prolinę [137]. Po łacinie *gluten* oznacza klej, a nazwa *glutamina* wynika właśnie z wysokiej zawartości tego aminokwasu w glutenie (ponad 30% [137]). Glutamina jest aminokwasem polarnym, ale bogate w nią białka faktycznie mogą tworzyć nierozpuszczalne agregaty [84, 85, 86] - skojarzenie z klejem jest zatem jak najbardziej uzasadnione. Agregacja poliglutaminy została omówiona w poprzednim rozdziale. W tym rozdziale symulacje zostaną wykorzystane do próby odpowiedzi na pytanie, co czyni gluten zarazem tak kleistym i elastycznym.

Białka glutenu są nieuporządkowane [214, 216, 217], i dzielą się na 2 frakcje: gliadyny (około 70% całkowitej masy glutenu) i gluteniny (pozostałe 30%) [137]. Gliadyny są rozpuszczalne w 50% roztworze alkoholu, w przeciwieństwie do glutenin, które mogą tworzyć kompleksy o masie rzędu megadaltonów, łącząc się ze sobą kowalencyjnie przy pomocy mostków dwusiarczkowych [137]. Gluteniny dzielą się dalej na te o niskiej i wysokiej masie cząsteczkowej (LMWGS i HMWGS ang. *low/high molecular weight glutenins*). Pojedynczy łańcuch HMWGS może zawierać nawet ponad 800 aminokwasów (patrz tabela 5.1). Gluteniny mogą też tworzyć więcej międzyłańcuchowych mostków dwusiarczkowych, dlatego są uważane za kluczowe dla elastyczności glutenu (a przez to całego ciasta) [94].



Rysunek 5.1: Skład glutenu wg [137]. Udział wyrażono jako procent masowy.

Gliadyny są krótsze, tworzą mniej mostków i uważa się, że odpowiadają za lepkość glutenu [90]. Szczegółowy podział gliadyn i glutenin przedstawia Rys. 5.1.

Istniejące teorie starają się wyjaśnić elastyczność glutenu trzema głównymi czynnikami (nie wykluczają się one wzajemnie, lecz każdy z nich składa się na elastyczność glutenu):

1. tworzenie mostków dwusiarczkowych i wiązań wodorowych: łańcuchy boczne glutenin mogą tworzyć między sobą wiązania wodorowe (patrz Rys. 1.1), a mostki mogą utrzymywać białka glutenu razem, tworząc coś w rodzaju żelu polimerowego [94];
2. model “pętli i ciągów” [214, 218] zakłada istnienie w glutenie “pętli” między sąsiednimi łańcuchami: są to puste przestrzenie, wypełnione przez rozpuszczalnik (np. wodę). Podczas rozciągania takie wnęki znikają, a znajdujące się bliżej siebie łańcuchy łączą się wiązaniem wodorowym, tworząc “ciągi”, które są równoległe do kierunku rozciągania i utrudniają dalszą deformację;
3. ponieważ łańcuchy glutenin zawierają setki aminokwasów, mogą tworzyć sieć splatań, która może zwiększać odporność na rozciąganie niezależnie od natury chemicznej tych łańcuchów [219].

Podczas doktoratu wykonane zostały symulacje samych gliadyn, samych glutenin oraz glutenu (czyli gliadyn i glutenin w odpowiednich proporcjach). W ten sposób można ocenić właściwości każdej frakcji oddziennie i oszacować ich wkład do elastyczności glutenu. Białka glutenu pełnią w ziarnie funkcję materiału zapasowego: są magazynem energii i składników budulcowych (m.in. azotu, stąd duża zawartość bogatej w azot glutaminy). Białka o podobnej funkcji występują w ziarnach kukurydzy i ryżu, jednak w doświadczeniu nie wykazują tak niezwykłych lepkosprężystych właściwości [220]. Te białka także zostały zasymulowane - stanowią one “próbkę kontrolną”, która ma pokazać, że model gruboziarnisty potrafi odtworzyć wyjątkowość glutenu.

Dokładny skład symulowanych białek podany jest w podrozdziale 5.2. Podczas symulacji są one umieszczone w pudełku, którego periodyczne deformacje mają imitować “ugniatanie” ciasta. Szczegóły omawia podrozdział 5.3. Wyniki (podrozdział 5.4) dotyczą tego, jak zmienia się elastyczność układu dzięki poddaniu go takim deformacjom, co dzieje się z nim na poziomie molekularnym, a wreszcie czy da się na podstawie takich oscylacji wyznaczyć moduł ścinania glutenu. Rozważana jest także zasadność obliczania krzywych rozpraszania SAXS dla symulacji glutenu.

Ze względu na duże rozmiary kompleksów tworzonych przez białka glutenu, a także długość pojedynczych białek (patrz tabela 5.1) użyta została szybsza, quasi-adiabatyczna wersja modelu.

## 5.2 Badane układy

Skład każdego z badanych układów: gliadyn, glutenin, glutenu oraz białek ziaren ryżu i kukurydzy został wyznaczony w oparciu o literaturę [91, 94, 137, 215, 221, 222, 223, 224, 225, 226]. Z kolei sekwencje każdego z białek pochodzą z bazy UniProt [227]. Sekwencje podane w tabeli 5.1 różnią się od tych z UniProt, ponieważ nie zostały uwzględnione sekwencje sygnałowe, które w dojrzałym białku są odcięte (a białka w mące są już jak najbardziej dojrzałe).

Proporcje gliadyn i glutenin w glutenie (podane za [91] i [137]) zależą od odmiany pszenicy [215]. Skład glutenu z tabeli 5.1 odpowiada odmianie z wysoką zawartością glutenin, ponieważ właśnie te białka są kluczowe dla elastyczności glutenu [94].

Białka glutenu są tylko częściowo nieuporządkowane, ponieważ zawierają fragmenty o wysokiej złożoności, które (w przeciwnieństwie do całego białka) mogą posiadać nawet strukturę trzeciorzędową (patrz podpunkt 5.2.1). Liczba aminokwasów w pojedynczym łańcuchu jest zatem sumą długości fragmentów nieuporządkowanych i uporządkowanych (sumowanie w tabeli 5.1 jest od N-końca do C-końca). To, które fragmenty są uporządkowane także zostało ustalone dzięki bazie UniProt.

Białka z ziaren kukurydzy (*Zea mays*) [221, 222, 223] oraz ryżu (*Oryza sativa*) [224, 225, 226] wydawały się nie mieć tak wyraźnego podziału na fragmenty uporządkowane i nieuporządkowane (choć może wynikać to z niewiedzy autora), zatem są symulowane jako całkowicie nieuporządkowane.

We wszystkich przypadkach starano się wybrać białko o sekwencji jak najbardziej reprezentatywnej dla całej grupy. Z tego powodu została odrzucona np. gliadyna  $\gamma$  o kodzie UniProt P04729, która znacznie różni się sekwencją od pozostałych gliadyn  $\gamma$ .

## Gluten

Rodzaj białka (gluten)	Kod Uniprot	Liczba łańcuchów	Liczba aminokwasów $N$	Suma aminokwasów $\Sigma N$
LMWGS, typ 1D1	P10386	3	<b>114+170</b>	852
HMWGS, typ DX5	P10388	1	<b>89+696+42</b>	827
HMWGS, typ DY10	P10387	1	<b>119+466+42</b>	627
Gliadyna $\alpha/\beta$ , typ MM1	P18573	3	<b>130+157</b>	861
Gliadyna $\gamma$ , typ B	P06659	3	<b>117+155</b>	843
Gliadyna $\omega$ , typ 1,2	Q9FUW7	1	261	261
Suma		12		4271

## Gluteniny

Rodzaj białka (gluteniny)	Kod Uniprot	Liczba łańcuchów	Liczba aminokwasów $N$	Suma aminokwasów $\Sigma N$
LMWGS, typ 1D1	P10386	6	<b>114+170</b>	1704
HMWGS, typ DX5	P10388	2	<b>89+696+42</b>	1654
HMWGS, typ DY10	P10387	1	<b>119+466+42</b>	627
Sum		9		3985

## Gliadyny

Rodzaj białka (gluten)	Kod Uniprot	Liczba łańcuchów	Liczba aminokwasów $N$	Suma aminokwasów $\Sigma N$
Gliadyna $\alpha/\beta$ , typ MM1	P18573	7	<b>130+157</b>	2009
Gliadyna $\gamma$ , typ B	P06659	6	<b>117+155</b>	1632
Gliadyna $\omega$ , typ 5	Q402I5	1	420	420
Gliadyna $\omega$ , typ 1,2	Q9FUW7	1	261	261
Suma		15		4322

Tabela 5.1: Skład glutenu, glutenin i gliadyn użyty podczas symulacji. Suma aminokwasów ( $\Sigma N$ ) oznacza liczbę aminokwasów w pojedynczym łańcuchu ( $N$ ) pomnożoną przez liczbę łańcuchów. Jeśli występują ustrukturyzowane domeny, liczba aminokwasów w pojedynczym łańcuchu jest przedstawiona jako suma fragmentów nieuporządkowanych i uporządkowanych (te drugie pogrubione).

## Kukurydza

Rodzaj białka (kukurydza)	Kod Uniprot	Liczba łańcuchów	Liczba aminokwasów $N$	Suma aminokwasów $\Sigma N$
$\alpha$ -zeina, typ 4	O48966	3	245	735
$\alpha$ -zeina, typ 16	P04700	4	242	968
$\alpha$ -zeina, typ 19C2	P06677	3	219	657
$\gamma$ -zeina	P08031	2	164	328
Glutelina-2	P04706	4	204	816
Suma		16		3504

## Ryż

Rodzaj białka (ryż)	Kod Uniprot	Liczba łańcuchów	Liczba aminokwasów $N$	Suma aminokwasów $\Sigma N$
Glutelina, typ-A 1	P07728 chain 1	2	282	564
Glutelina, typ-A 1	P07728 chain 2	2	193	386
Glutelina, typ-A 3	Q09151 chain 1	1	281	281
Glutelina, typ-A 3	Q09151 chain 2	1	191	191
Glutelina, typ-B 1	P14323 chain 1	2	278	556
Glutelina, typ-B 1	P14323 chain 2	2	197	394
Glutelina, typ-D 1	Q6K508 chain 1	1	258	258
Glutelina, typ-D 1	Q6K508 chain 2	1	199	199
Prolamina 14E	Q0DJ45	1	131	131
prolamina C	P17048	1	137	137
10 kDa prolamina	Q0DN94	1	110	110
Prolamina 14P	Q42465	1	132	132
19 kDa globulina	P29835	1	164	164
Suma		17		3503

Tabela 5.2: Skład białek ryżu i kukurydzy użytych podczas symulacji. Suma aminokwasów ( $\Sigma N$ ) oznacza liczbę aminokwasów w pojedynczym łańcuchu ( $N$ ) pomnożoną przez liczbę łańcuchów.

### 5.2.1 Fragmenty uporządkowane

Dla wielu białek glutenu początek i koniec ich sekwencji ma dużo większą złożoność niż reszta. Np. domena A, znajdująca się na N-końcu gluteniny HMW, wykazuje wysoki stopień homologii do roślinnego inhibitora enzymów trawiennych [228]. Podobne położenie par cystein w sekwencji sugeruje, że tak jak w tych inhibitorach tworzą się stałe mostki w ramach jednego łańcucha. Te i podobne więzy strukturalne zostały uwzględnione poprzez wygenerowanie struktur o rozdzielcości pełnoatomowej dla uporządkowanych regionów, a następnie wykorzystanie ich do utworzenia map kontaktów, tak jak w modelu Go [229]. Potencjał sztywności łańcucha jest dla ustrukturyzowanych fragmentów taki sam jak dla nieustrukturyzowanych (tzn. taki jak opisany w podrozdziale 2.3).

Fragmenty ustrukturyzowane zostały wybrane na podstawie literatury [228] i podziału sekwencji na domeny wg bazy UniProt [227]. Rezultat podziału przedstawia tabela 5.1.

Struktury pełnoatomowe dla ustrukturyzowanych fragmentów zostały wygenerowane metodą modelowania homologicznego przy pomocy serwera ITASSER [230, 231]. Struktury te zostały wykorzystane do znalezienia mapy kontaktów, opartej na kryterium przekrywania ciężkich atomów (szczegółły konstrukcji mapy kontaktów są podane w [229] oraz skrótnie w podpunkcie 2.5.2).

Dzięki temu, że do symulacji glutenu został użyty model quasi-adiabatyczny, można bez trudu połączyć zadaną z góry mapę kontaktów dla fragmentów ustrukturyzowanych z chwilową mapą kontaktów obliczaną w programie. Kontakty przewidziane przez serwer ITASSER są “zawsze włączone”, tzn. odpowiednie pary aminokwasów zawsze się przyciągają. Dlatego można kontakty te określić jako “statyczne”, w przeciwieństwie do “dynamicznych” kontaktów dla fragmentów nieuporządkowanych. Kontakty “statyczne” są opisane potencjałem LJ z minimum odpowiadającym odległości między atomami  $C_\alpha$  w pełnoatomowej strukturze wygenerowanej serwerem ITASSER. Aminokwasy tworzące “statyczne” kontakty mogą tworzyć także kontakty “dynamiczne” z innymi (o ile nie należą do tego samego fragmentu ustrukturyzowanego). Istnienie kontaktu “statycznego” wlicza się do limitu  $s$  danego aminokwasu.

## 5.3 Protokół symulacji

Pudełko, w którym symulowany jest gluten posiada (przez cały czas symulacji) periodyczne warunki brzegowe dla ścian prostopadłych do osi  $X$  oraz  $Y$ , natomiast ściany prostopadłe do osi  $Z$  nie pozwalają na przechodzenie przez nie aminokwasów (i są przyciągające bądź odpychające).

Początkowe wszystkie łańcuchy są losowo<sup>1</sup> rozmieszczone w sześciennym pudełku<sup>2</sup> o gęstości 0.1 aminokwasu/nm<sup>3</sup> (aa/nm<sup>3</sup>), a ich kształt jest określony przez samoomijające się błądzenie przypadkowe<sup>1</sup>.

Losowe rozmieszczenie białek jest umotywowane tym, że w ziarnach białka glutenu też są od siebie oddzielone [232]. Ich agregacja zachodzi dopiero podczas mieszania mąki w środowisku wodnym. Jest to odzwierciedlone w protokole symulacji: najpierw układ ma 125 000 ns (przyjmujemy  $1 \tau \approx 1 \text{ ns}$ ) na dojście do stanu równowagi (w ramach pojedynczych białek) w małej gęstości (ponad 30 razy mniejszej niż gęstość glutenu). Gęste układy polimerów mają bardzo długie czasy relaksacji [233], dlatego ważne jest, aby białka glutenu miały możliwość “zwinięcia się” (zwłaszcza jeśli mają fragmenty uporządkowane) przed agregacją. Oczywiście losowo dyfundujące łańcuchy mogą przypadkowo zagregować przed przyjęciem właściwej konformacji, dlatego wszystkie podawane tu wyniki są średnią z co najmniej trzech symulacji. Takie przypadki są jednak mało prawdopodobne, patrz panel 1 na Rys. 5.2.

Ściany prostopadłe do osi  $Z$  są na początku symulacji wyłącznie odpychające. Odległość między tymi dwiema ścianami (a zarazem długość boku pudełka) jest oznaczona jako  $s$ . Odpychanie aminokwasów przez ściany zapewnia ucięty potencjał LJ z głębokością  $4\epsilon$  i minimum 0.5 nm (dla  $r > 0.5 \text{ nm}$  potencjał wynosi zero, a minimum jest podniesione tak, aby również odpowiadało energii zero).

Po 125  $\mu\text{s}$  przeznaczonych na dojście do równowagi, ściany pudełka zaczynają zbliżać się do siebie z prędkością 2 mm/s, aż zostanie osiągnięta gęstość  $\rho_0$  (w większości symulacji odpowiadająca gęstości glutenu [234]  $\rho_0 \approx 3.5 \text{ aa/nm}^3$ , patrz podpunkt 5.3.2). Następnie układ ma kolejne 150  $\mu\text{s}$  na dojście do równowagi (co przedstawia panel 2 na Rys. 5.2). Na tym etapie długość boku pudełka ma wartość  $s = s_0$ , określoną przez gęstość  $\rho_0$  i liczbę aminokwasów w pudełku.

Kolejnym etapem po oczekaniu 150  $\mu\text{s}$  jest włączenie przyciągania dla ścian prostopadłych do osi  $Z$ . Kiedy odległość między ścianą a dowolnym aminokwasem stanie się mniejsza niż 0.5 nm, przyciągająca część potencjału LJ o głębokości  $4\epsilon$  jest quasi-adiabatycznie włączana (tak jak dla mostków dwusiarczkowych, patrz podpunkt 2.5.7). Potencjał LJ jest określony dla odległości  $r_{i,w}$  między aminokwasem  $i$  oraz punktem  $w$  na ścianie. Współrzędna  $Z$  punktu  $w$  odpowiada położeniu ściany, natomiast współrzędne  $X$  oraz  $Y$  są takie same jak współrzędne aminokwasu  $i$ , kiedy po raz pierwszy przekroczył próg  $r < 0.5 \text{ nm}$ .

Kiedy aminokwas  $i$  oddali się od punktu  $w$  dalej niż na 2 nm, jest uznawany za odłączonego od ściany i kontakt między aminokwasem  $i$  i punktem  $w$  na ścianie jest quasi-adiabatycznie wyłączany. Aminokwas może potem przyłączyć się do ściany w innym miejscu  $w'$ .

---

<sup>1</sup>Nie jest to całkowicie losowe rozmieszczenie ani losowe błądzenie, ponieważ aminokwasy muszą być w odległości  $r > 4 \text{ \AA}$  od siebie, a kąt płaski między każdą trójkątową aminokwasem w błądzeniu przypadkowym musi być większy od  $60^\circ$ .

<sup>2</sup>Jeśli początkowa konfiguracja białka przechodzi przez ścianę, pudełko jest zwiększone we wszystkich kierunkach, więc rzeczywista gęstość początkowa może być mniejsza.

Odpychająca część potencjału LJ jest zawsze włączona dla  $r < 0.5$  nm, aby żaden aminokwas nie przeszedł przez ścianę prostopadłą do osi  $Z$ . Siła działająca na ściany jest sumą sił z jaką działają na nie wszystkie aminokwasy. Jest ona uśredniana co 100 ns, aby zmniejszyć znaczenie szumu termicznego (aminokwasy odbijają się od ściany). Powody, dla których została wybrana właśnie ta metoda realizacji przyciągania przez ściany, są przedstawione w podpunkcie 5.3.1.

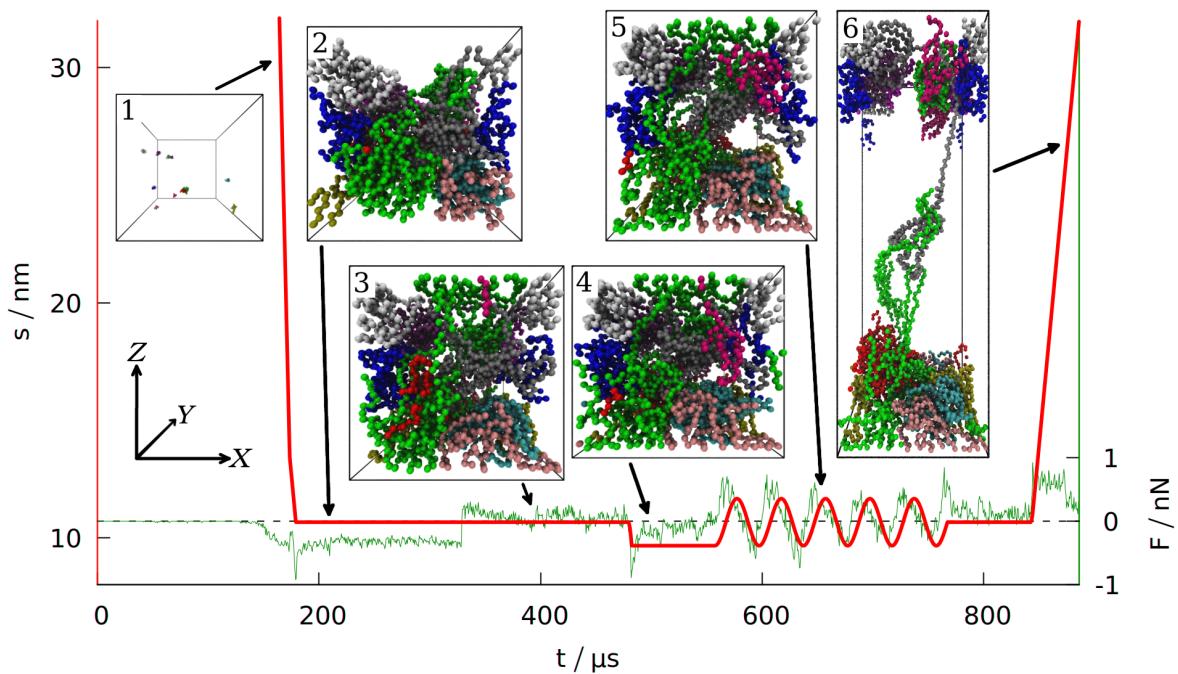
Kiedy przyciąganie ścian zostanie włączone, układ ma następuje  $150 \mu\text{s}$  na dojście do równowagi (co odpowiada panelowi 3 na Rys. 5.2). To, co stanie się po tej (trzeciej już) przerwie zależy od tego, którego rodzaju simulacja jest wykonywana: rodzi to odkształcenie normalne, ścinające oraz symulacja równowagowa. Dla odkształcenia normalnego obie ściany zbliżają się do siebie wzdłuż osi  $Z$  aż osiągną odległość  $s = s_0 - A$ , gdzie  $A = 1$  nm jest amplitudą oscylacji. Ściany prostopadłe do osi  $X$  i  $Y$  pozostają w odległości  $s_0$ , zatem pudełko do symulacji przestaje być sześcianem.<sup>3</sup> Po osiągnięciu  $s = s_0 - A$  (co pokazuje panel 4 na Rys. 5.2), odległość między ścianami zmienia się periodycznie:  $s(t) = s_0 - A \cos(\omega t)$ . Podczas eksperymentów  $\omega \sim 1 \text{ Hz}$  [90], czego nie da się osiągnąć w symulacji. Domyślny okres oscylacji to  $40 \mu\text{s}$ , co odpowiada  $f \approx 25 \text{ kHz}$ . Największe  $s$  osiągnięte podczas oscylacji pokazuje panel 5 na Rys. 5.2.

Dla symulacji odkształceń ścinających ściany także przesuwają się o taką samą amplitudę  $A$ , jednak oscylacje zachodzą w kierunku  $X$  i są opisywane zależnością  $s'(t) = -A \cos(\omega t)$  (gdzie  $s' = 0$  odpowiada ścianom prostopadłym do osi  $Z$  ustawnionym jedna nad drugą). Odkształcenia ścinające są przedstawione na Rys. 5.3.

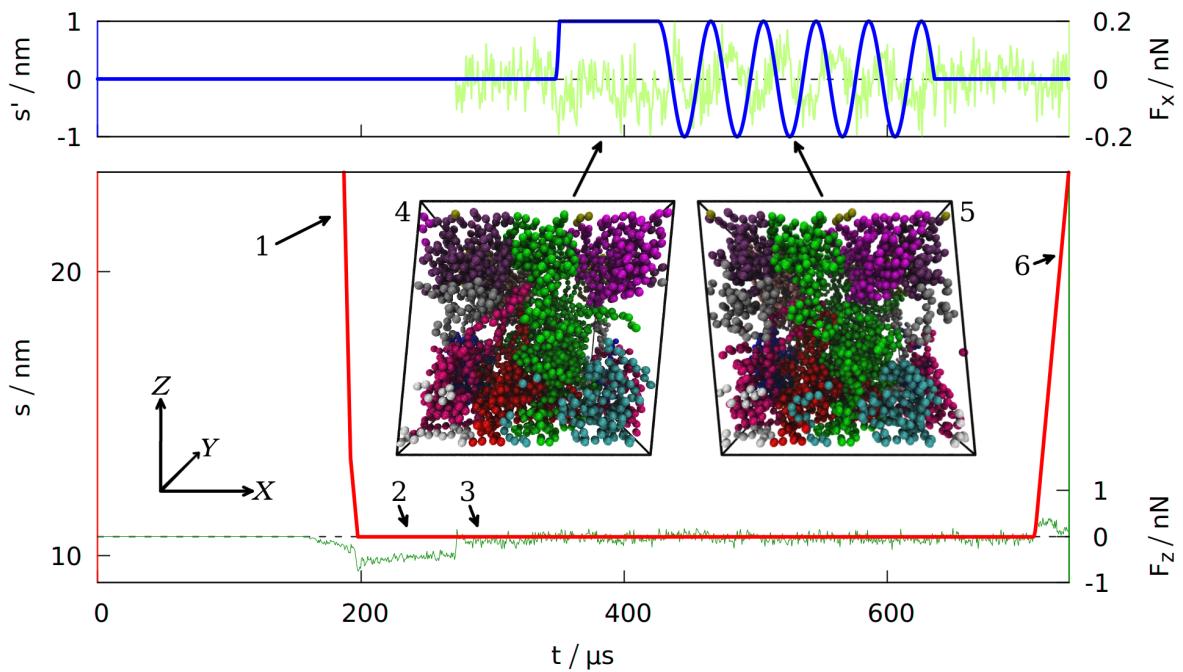
Trzeci, równowagowy rodzaj symulacji zakłada dodatkowe  $100 \mu\text{s}$  czasu na dochodzenie do równowagi zamiast periodycznych deformacji. W pierwszych dwóch rodzajach symulacji takich odkształcających (normalnie bądź ścinająco) oscylacji jest pięć. Po 5 oscylacjach (lub po oczekaniu  $100 \mu\text{s}$ ) zakończonych powrotem do  $s = s_0$  następuje kolejne  $75 \mu\text{s}$  czekania na dojście do stanu równowagi. Po nich następuje ostatni już etap symulacji: ściany przyciągające oddalają się od siebie wzdłuż osi  $Z$  aż do osiągnięcia odległości  $s = 2s_0$ . Umożliwia to obliczenie całkowitej pracy potrzebnej do takiego rozciągnięcia, a także największej siły potrzebnej do rozerwania układu (co odpowiada eksperymentowi rozciągania próbki wzdłuż jednej osi [238]). Panel 6 na Rys. 5.2 pokazuje takie rozciąganie. Szybkość rozciągania to  $0.5 \text{ mm/s}$ , co odpowiada typowym szybkościom rozciągania mikroskopem sił atomowych [199, 229]. Największa szybkość ścian osiągana podczas oscylacji ( $v_{max} = A\omega$ ) jest jeszcze mniejsza.

---

<sup>3</sup>Została też wypróbowana wersja programu, w której ściany (reprezentujące periodyczne warunki brzegowe) wzdłuż  $X$  i  $Y$  także się przesuwają, aby zachować stałą objętość pudełka, jednak ta wersja nie wpływa znaczco na wyniki. Współczynnik Poissona dla chleba wynosi ok. 0.1 [235], dla większości białek roślinnych ok. 0.25 [236], a dla ciasta przyjęta została wartość 0.46 [237], zatem nie ma jednoznacznego eksperymentalnego wskazania za lub przeciw utrzymywaniu stałej objętości.



Rysunek 5.2: Zależność od czasu odległości  $s$  między dwiema ścianami prostopadłymi do osi  $Z$  (czerwony) i łącznej siły, z jaką aminokwasy oddziałują na te 2 ściany (uśrednionej co 100 ns, zielony). Dane dotyczą symulacji glutenu (skład wg tabeli 5.1), o gęstości  $3.5 \text{ aa/nm}^3$ , z periodycznym odkształcaniem normalnym o okresie  $40 \mu\text{s}$ . 6 paneli pokazuje stan symulacji podczas jej kolejnych etapów (każda kulka reprezentuje 1 aminokwas, kolory rozróżniają łańcuchy). Ten sam układ współrzędnych (na górze po lewej) jest użyty dla wszystkich 6 paneli. Modyfikacja rys. 1 z artykułu [V].



Rysunek 5.3: Górnny panel pokazuje zależność od czasu wychylenia ścian  $s'$  z położenia równowagi w kierunku  $X$  (niebieski) oraz składowej  $X$  siły działającej na ściany (zielonkawy). Dolny panel i reszta oznaczeń (w tym numeracja etapów symulacji) jak na Rys. 5.2, tyle że z periodycznym odkształcaniem ścinającym (pokazanym na panelach 4 i 5) zamiast normalnego. Modyfikacja rys. S1 z artykułu [V].

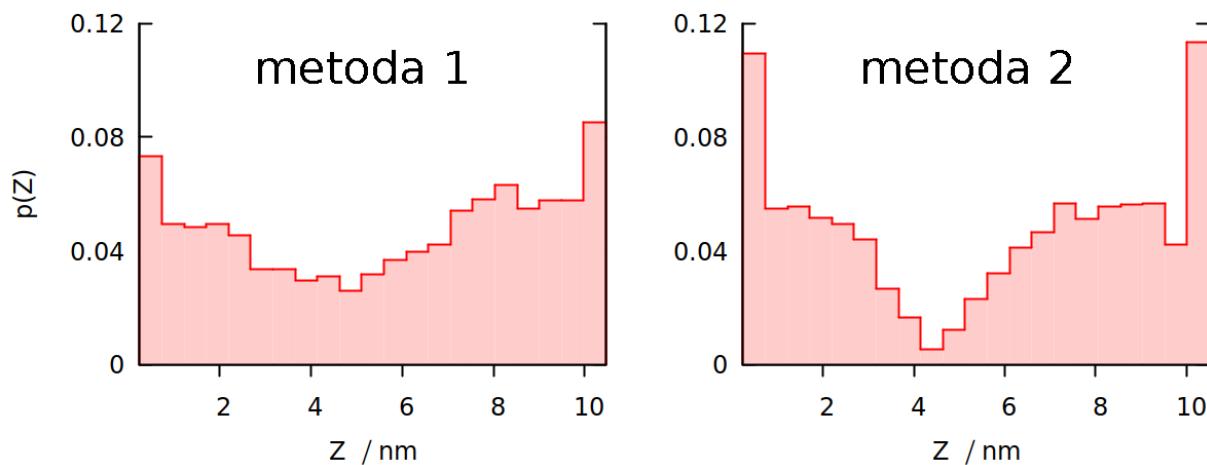
### 5.3.1 Wybór oddziaływania ze ścianami

Sposób oddziaływania aminokwasów ze ścianą jest kluczowy dla wyznaczenia siły, z jaką układ odpowiada na deformację. Rozważone zostały dwie metody:

1. metoda opisana na początku podrozdziału 5.2;
2. ściany utworzone z pseudoatomów ułożonych jak w dwóch warstwach sieci kubicznej (przekrój wzdłuż kierunku krystalograficznego [111]) o stałej sieci  $3 \text{ \AA}$ .

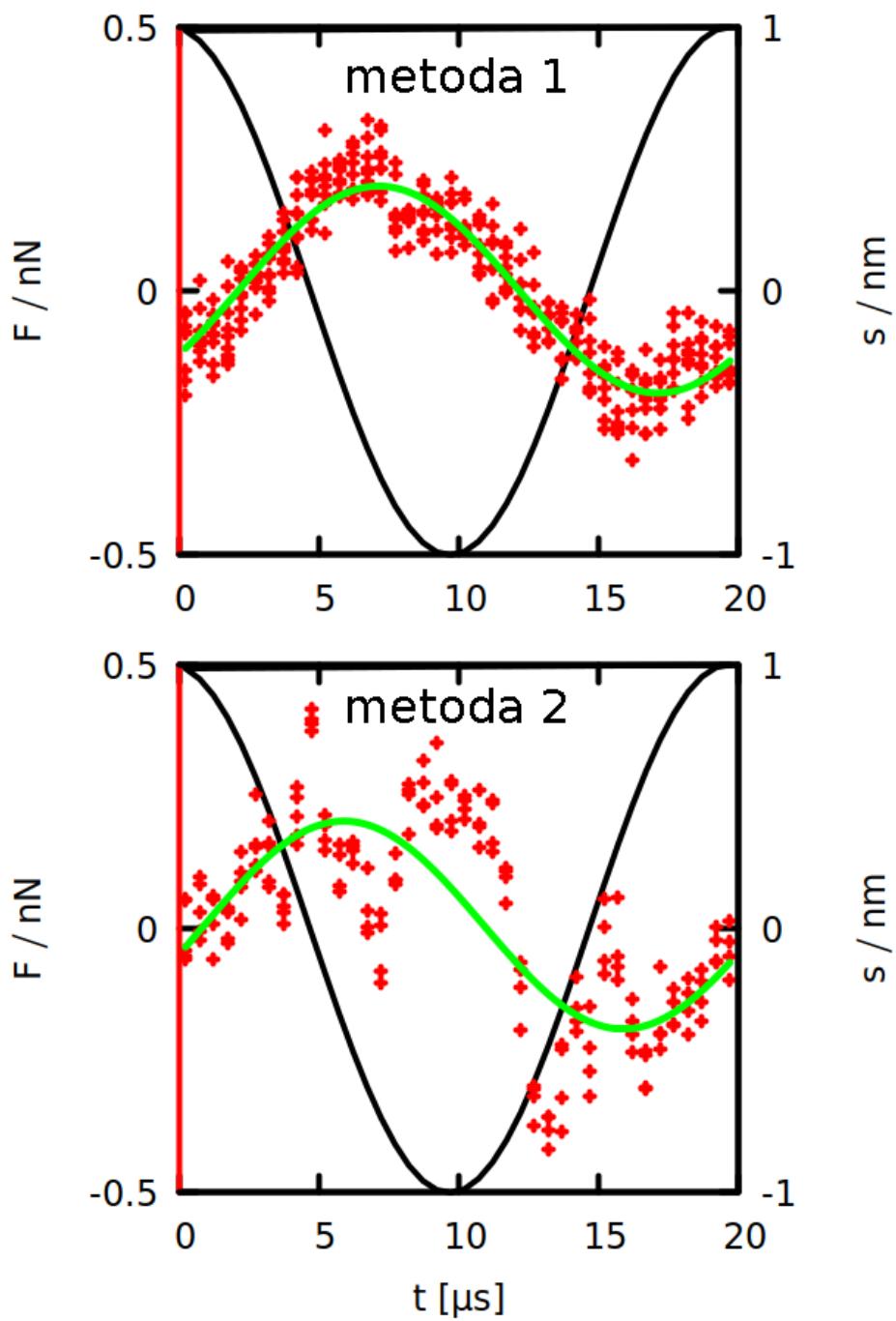
W obydwa metodach aminokwasy oddziałują ze ścianą potencjałem LJ (przyczepianie ich do ściany wiązaniem harmonicznym prowadziło do niestabilności numerycznych i uniemożliwiało odczepienie aminokwasu od ściany). Metoda 2 wymaga znacznie więcej mocy obliczeniowych, aby symulować wszystkie pseudoatome ściany. Mimo że są one nieruchome (lub poruszają się tak jak ściana), przy obliczaniu sił trzeba rozważyć każdą parę bliskich sobie aminokwasów i pseudoatomów ściany. Przyciąganie jest przy tym silniejsze: w metodzie 1 aminokwas jest zawsze przyciągany przez jedno centrum interakcji  $w$ , podczas gdy w metodzie 1 jest przyciągany przez wszystkie pobliskie pseudoatome ściany.

Profile gęstości wzdłuż osi  $Z$  (Rys. 5.4) pokazują, że dla metody 2 aminokwasy zbierają się blisko ścian, czyniąc rozkład gęstości niejednorodnym, co może wpływać na wyniki. Ten efekt nie jest obserwowany dla metody 1.



Rysunek 5.4: Profile gęstości wzdłuż osi  $Z$  (liczba aminokwasów została zsumowana w kierunkach  $X$  i  $Y$ ). Profile pochodzą z symulacji glutenu (o danej gęstości  $\rho$ ), z etapu po periodycznej deformacji pudełka z częstością  $20 \mu\text{s}$ , ale przed rozciąganiem. Modyfikacja rys. S9 z artykułu [V].

Siła, z jaką aminokwasy oddziałują na ściany podczas oscylacji dużo bardziej przypomina sinusoidę dla metody 1 (patrz Rys. 5.5), ponieważ aminokwasy mogą się “prześlizgiwać” po pseudoatomach z metody 2. Z podanych wyżej powodów w symulacjach została użyta metoda 1.

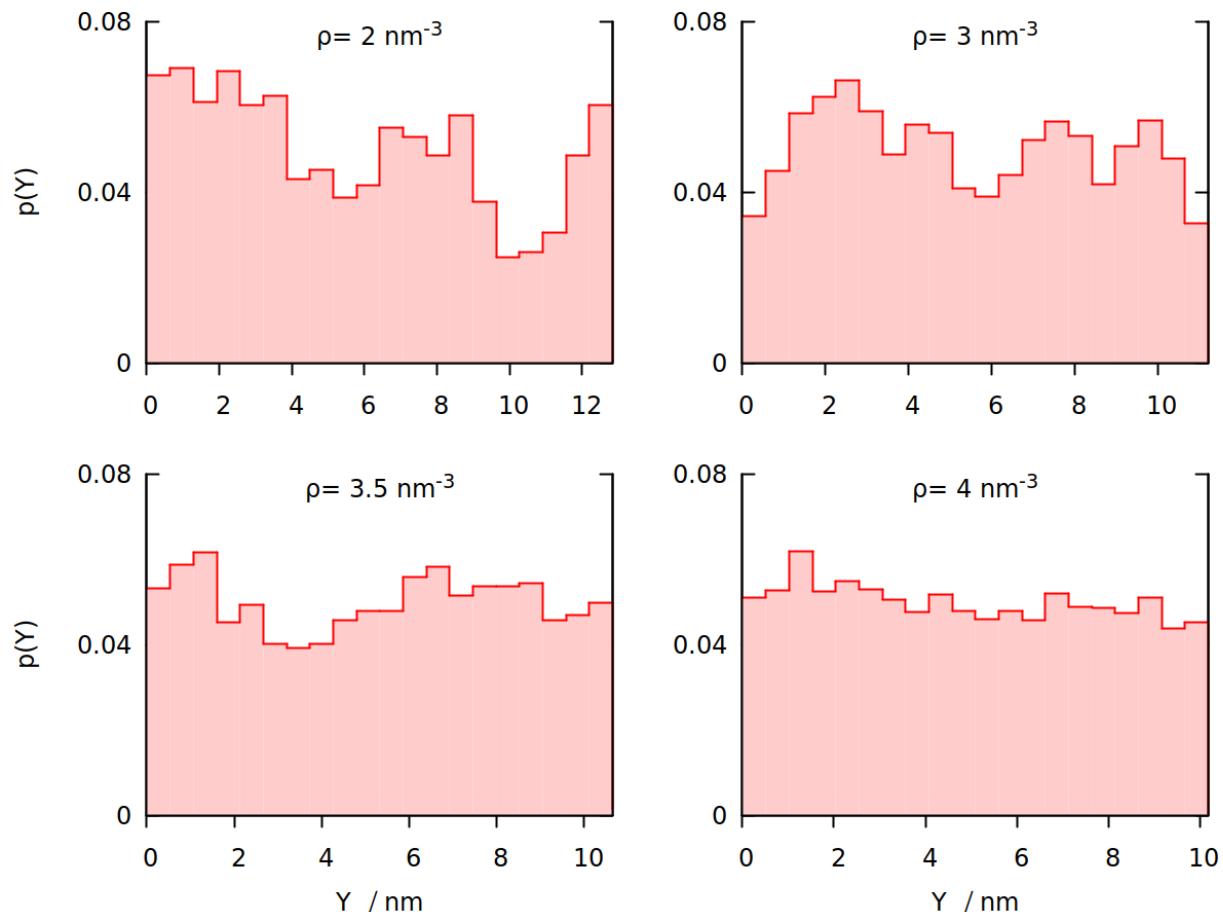


Rysunek 5.5: Siła, z jaką aminokwasy oddziałują na ściany podczas symulacji odkształceń ścinających glutenu z okresem  $20 \mu\text{s}$  dla metody 1 (górnny panel) oraz metody 2 (dolny panel). Każdy czerwony punkt to jedna wartość siły uśredniona co 100 ns. Kolejne oscylacje są nałożone na siebie, tak że czas jest podany z dokładnością do wielokrotności okresu oscylacji. Zielona linia pokazuje dopasowanie użyte do wyznaczenia modułu ścinania. Modyfikacja rys. S10 z artykułu [V].

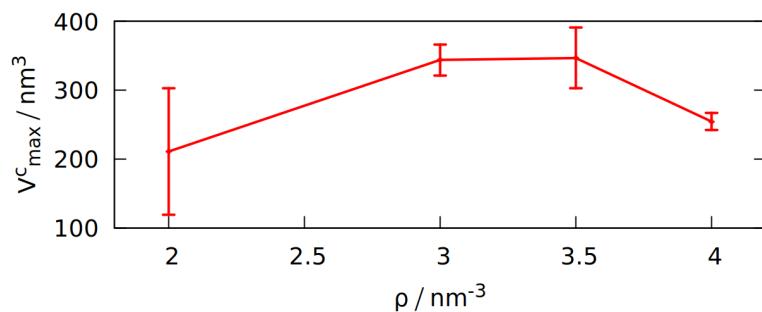
### 5.3.2 Wybór gęstości

Prawdziwe próbki glutenu zwykle zawierają znaczne ilości wody, dlatego w modelu z ukrytym rozpuszczalnikiem efektywna gęstość glutenu może być większa niż rzeczywista. Jednakże w modelu z jedną kulką na aminokwas łańcuchy boczne nie są reprezentowane, dlatego wyłączona objętość może być mniejsza: aminokwasy zajmują mniej miejsca niż w rzeczywistości. Dlatego zasadność wyboru gęstości  $\rho_0 \approx 3.5 \text{ aa/nm}^3$  musi być potwierdzona w inny sposób. Jeden z metod jest porównanie profili gęstości dla różnych  $\rho$ . Rzeczywista gęstość glutenu  $\rho_0$  okazuje się być bliska wartości progowej, powyżej której profile gęstości stają się jednorodne (patrz Rys. 5.6).

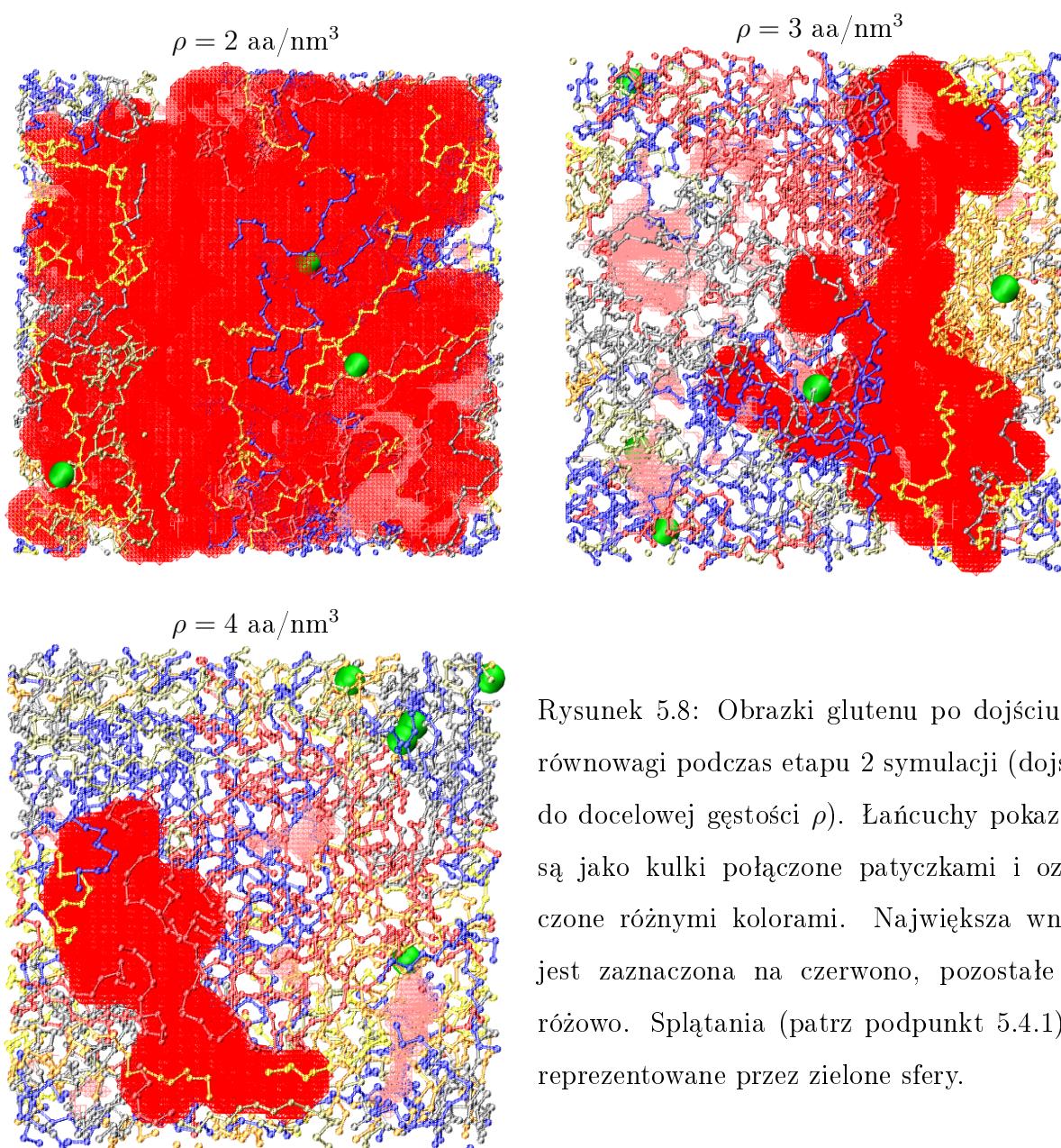
Badane były gęstości pomiędzy  $2 \text{ aa/nm}^3$  a  $4 \text{ aa/nm}^3$ . Im większa gęstość, tym mniej pustej przestrzeni, a więc objętość największej wnęki w układzie powinna się zmniejszać. Rys. 5.7 pokazuje jednak, że dla małych gęstości jest odwrotnie: gęstość  $2 \text{ nm}^{-3}$  jest na tyle niska, że algorytm Spaceball użyty do obliczania rozmiaru wnęek [115] nie traktuje pustej przestrzeni jako wnęki (pusta przestrzeń przechodzi przez cały układ, patrz Rys. 5.3.2). Objętość  $V_{\max}^c$  zaczyna się zmniejszać dopiero dla gęstości  $\rho > \rho = 3.5 \text{ nm}^{-3}$ , dlatego  $\rho_0 = 3.5 \text{ nm}^{-3}$  jest właściwą gęstością, przy której układ białek glutenu można uważać za jednorodny, jest w nim jednak dość wolnego miejsca aby tworzyły się wnęki.



Rysunek 5.6: Profile gęstości wzdłuż osi  $Y$  (liczba aminokwasów została zsumowana w kierunkach  $X$  i  $Z$ ). Profile pochodzą z symulacji glutenu (o danej gęstości  $\rho$ ), z etapu po periodycznej deformacji pudełka z częstotliwością  $20 \mu\text{s}$ , ale przed rozciąganiem. Oparte na rys. S7 z artykułu [V].



Rysunek 5.7: Zależność od gęstości maksymalnej objętości wnęki  $V_c^{\max}$ , policzona dla etapu po periodycznej deformacji pudełka z częstotliwością  $20 \mu\text{s}$ , ale przed rozciąganiem. Oparte na rys. S8 z artykułu [V].



Rysunek 5.8: Obrazki glutenu po dojściu do równowagi podczas etapu 2 symulacji (dojście do docelowej gęstości  $\rho$ ). Łańcuchy pokazane są jako kulki połączone patyczkami i oznaczone różnymi kolorami. Największa wnęka jest zaznaczona na czerwono, pozostałe na różowo. Splątania (patrz podpunkt 5.4.1) są reprezentowane przez zielone sfery.

## 5.4 Wyniki

### 5.4.1 Właściwości białek glutenu, kukurydzy i ryżu

Tabela 5.3 podsumowuje podstawowe właściwości każdego z omawianych układów. Duża liczba cystein na łańcuch powinna zwiększaćczęstość występowania międzyłańcuchowych mostków dwusiarczkowych; z kolei dłuższe łańcuchy powinny ułatwiać tworzenie się splątań. Wreszcie “lepkość” układu jest odzwierciedlona w liczbie koordynacyjnej  $z$ , która jest zdefiniowana jako średnia liczba kontaktów (zarówno “statycznych” jak i “dynamicznych”, patrz podpunkt 5.2.1) utworzonych przez każdy z aminokwasów.

układ	$\langle n_{\text{Cys}} \rangle$	$\langle N \rangle$	$\Sigma N$	$z$
kukurydza	6.1	221	3504	$5.5 \pm 0.1$
gliadyny	6	290	4322	$6.88 \pm 0.03$
gluteniny	7.2	445	3985	$6.55 \pm 0.08$
gluten	6.5	384	4271	$6.65 \pm 0.07$
ryż	3.6	208	3503	$5.62 \pm 0.09$

Tabela 5.3: Średnia liczba cystein na łańcuchu  $\langle n_{\text{Cys}} \rangle$ , średnia liczba aminokwasów w łańcuchu  $\langle N \rangle$ , całkowita liczba aminokwasów  $\Sigma N$  oraz średnia liczba koordynacyjna  $z$  dla wszystkich rozpatrywanych układów. Tylko  $z$  została wyznaczona poprzez symulacje równowagowe (bez oscylacji), pozostałe właściwości wynikają z samej sekwencji białek.

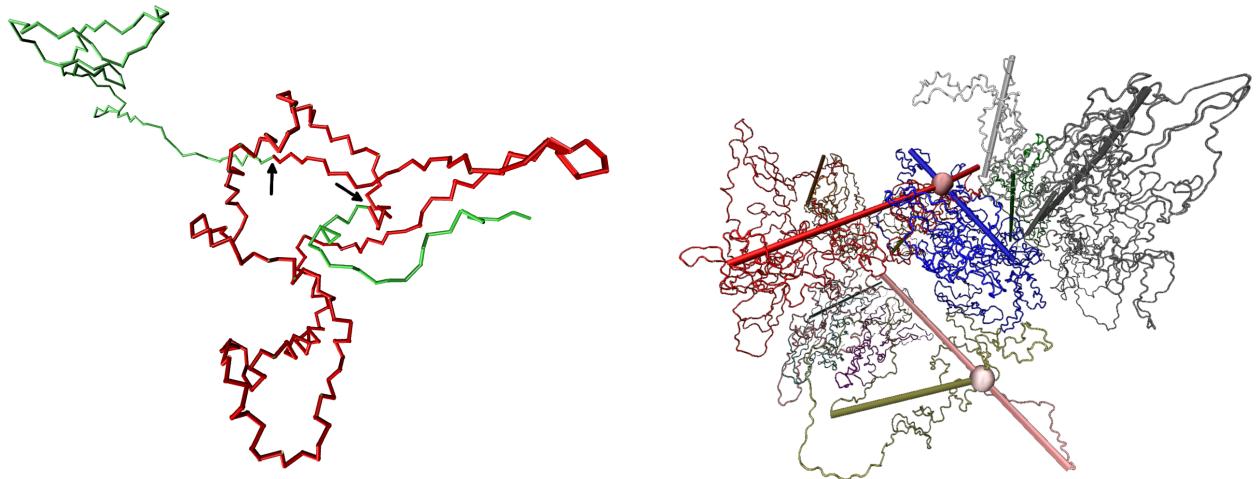
Przeprowadzone zostały 3 rodzaje symulacji:

1. równowagowe, bez żadnych oscylacji;
2. z periodycznym odkształceniem normalnym (w kierunku  $Z$ );
3. z periodycznym odkształceniem ścinającym (w kierunku  $X$ );

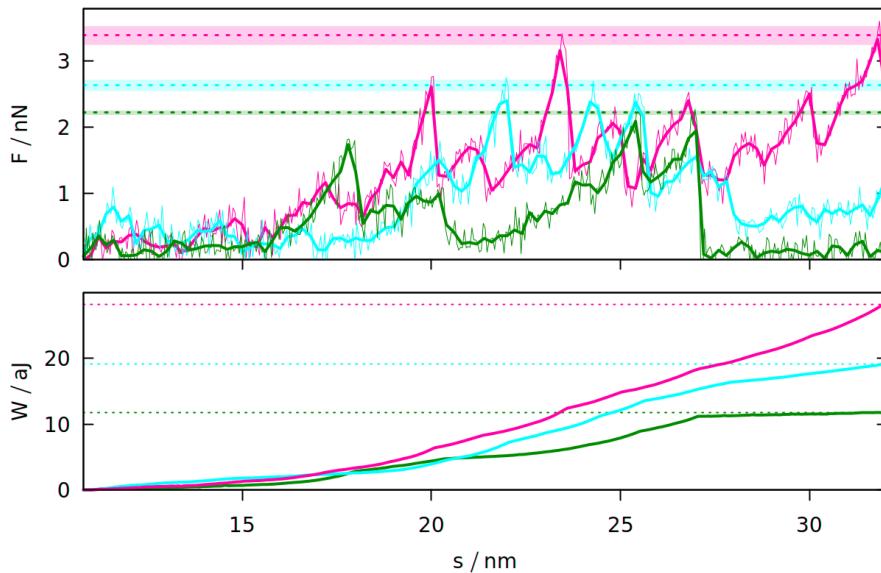
Wszystkie właściwości obliczone na podstawie symulacji (takie jak  $z$ ) są obliczane jako średnia liczona dla etapu symulacji po oscylacjach (a w przypadku braku oscylacji<sup>4</sup> po upływie 100 000  $\tau$ ). Liczba koordynacyjna  $z$  wydaje się nie zależeć od rodzaju symulacji, w przeciwnieństwie do wartości 5 innych właściwości. Są nimi:

<sup>4</sup>Odczekanie dodatkowego czasu ma duże znaczenie przy porównaniu z wynikami symulacji z oscylacjami: wiadomo wtedy, że faktycznie różnica dotyczy oscylacji, a nie samego czasu symulacji: nawet setki tysięcy  $\tau$  to mniej niż najdłuższe czasy relaksacji w tak gęstym układzie [233]. Z powodu ograniczeń technicznych (jedna symulacja trwa ponad miesiąc) trzeba się pogodzić z tym, że symulacje nie są całkowicie równowagowe. Czasy dochodzenia do równowagi (rzędu 100 000  $\tau$ ) zostały dobrane tak, aby średnie wartości obliczanych wartości (takich jak  $z$ ) przestały zmieniać się w czasie.

- liczba splatań  $l_k$ , wyznaczona algorytmem Z1 [117] (opisana w paragrafie 4.3.2.2). Przykłady splatań w glutenie pokazują prawy panel Rys. 5.9 oraz panel 6 na Rys. 5.2, gdzie koniec szarego łańcucha (LMWGS) przechodzi przez pętlę utworzoną przez zielony łańcuch (HMWGS).  $l_k$  jest uśrednione po wszystkich strukturach zapisanych po oscylacjach, a przed końcowym rozciąganiem.
- maksymalna siła podczas końcowego rozciągania układu  $F_{\max}$  (któremu odpowiada panel 6 na Rys. 5.2). Zielona krzywa na Rys. 5.2 pokazuje jak siła najpierw rośnie, a potem spada po rozerwaniu układu przy rozciąganiu. Wysokość maksimum siły zależy w pewnym stopniu od losowego ułożenia łańcuchów. Procedurę wyznaczania średniego  $F_{\max}$  wyjaśnia rysunek 5.10.
- praca mechaniczna  $W_{\max}$  wykonana podczas końcowego rozciągania układu. W tym ostatnim etapie symulacji ściany przyciągające rozchodzą się wzduż osi  $Z$  ze stałą prędkością, rozciągając białka w 2 przeciwnych kierunkach. Siła z jaką białka są ciągnięte może być całkowana po odległości przebytej przez ściany, dając pracę potrzebną do rozciągnięcia układu.
- liczba kontaktów między łańcuchami  $n_{\text{inter}}$ , tak jak  $l_k$  uśredniona po wszystkich strukturach zapisanych po oscylacjach, a przed końcowym rozciąganiem.
- RMSF (średnia kwadratowa fluktuacji położenia, ang. *root mean square fluctuation*), liczona względem średniego położenia aminokwasu, uśredniona po wszystkich aminokwasach i po wszystkich strukturach zapisanych po oscylacjach, a przed końcowym rozciąganiem.



Rysunek 5.9: Po lewej przykład zawężonej gluteniny HMW (dla symulacji glutenu o gęstości  $\rho = 4 \text{ aa/nm}^3$ ). Węzeł zaznaczono na czerwono, a jego końce pokazują czarne strzałki. Modyfikacja rys. S5 z artykułu [V]. Po prawej przedstawienie splatanych łańcuchów glutenu: walce łączą pierwszy i ostatni aminokwas w splatonym łańcuchu, a duża kula oznacza splatanie dwóch łańcuchów.

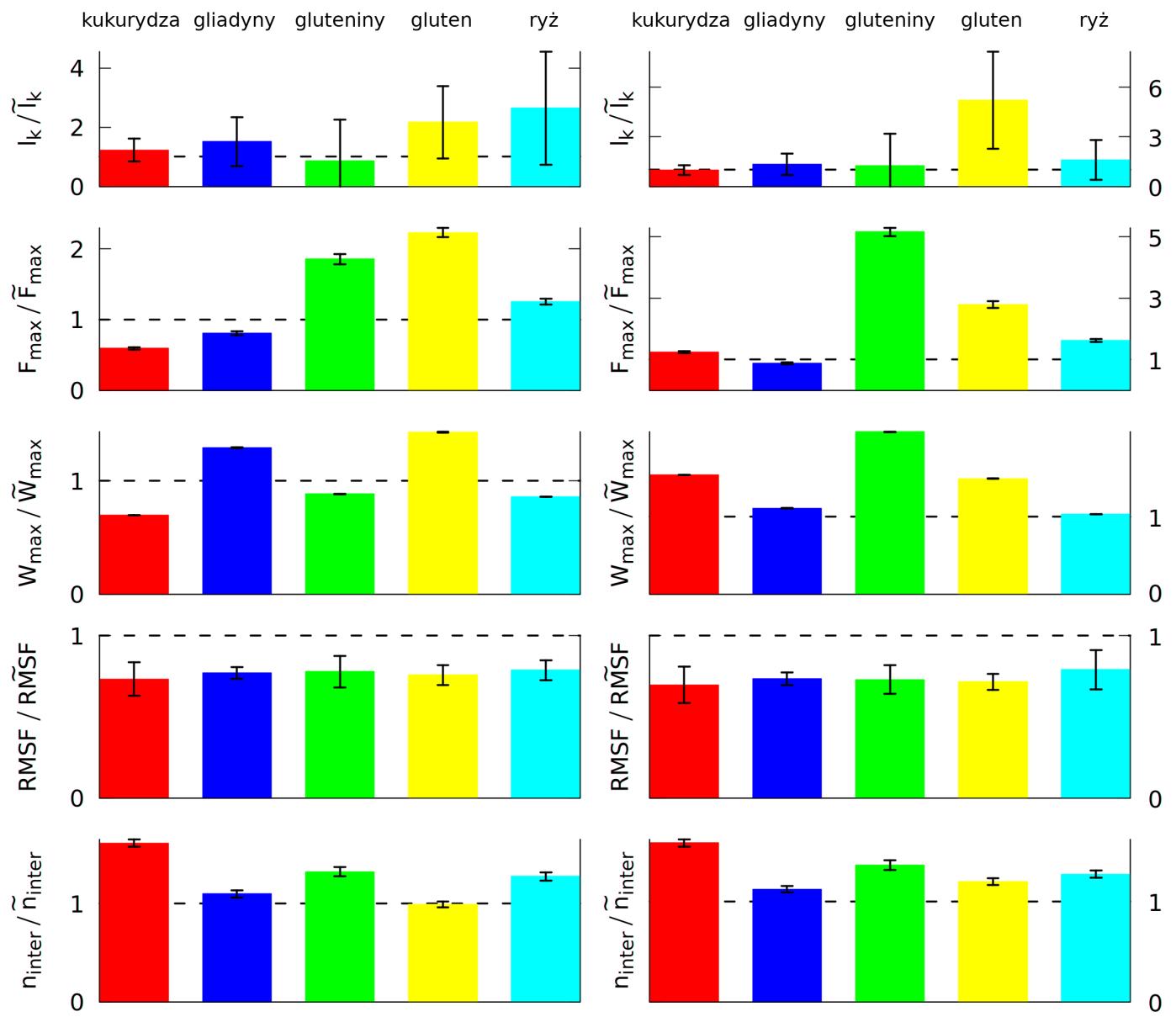


Rysunek 5.10: Siła wywierana na gluten o gęstości  $3.5 \text{ aa/nm}^3$  (górnny panel) podczas ostatniego etapu symulacji oraz wykonana przez nią praca mechaniczna (dolny panel) w funkcji odległości  $s$  między ścianami. Pokazane są 3 niezależne symulacje, każda innym kolorem. Przerywane poziome linie odpowiadają  $F_{max}$  oraz  $W_{max}$  dla danej symulacji. Każde  $F_{max}$  zostało obliczone jako średnia po 5 punktach z największą siłą na wykresie  $F(s)$  dla każdej z 3 symulacji. Paski wokół przerywanych linii reprezentują odchylenie standardowe średniej.  $F_{max}$  pokazane na Rys. 5.11 jest średnią ważoną z tych 3 wartości, gdzie wagi są odwrotnością wariancji średnich. Okres oscylacji to  $40 \mu\text{s}$ . Dla  $F$ , cienkie linie pokazują surowe dane z symulacji, a grube krzywe są wygładzone. Modyfikacja rys. 3 z artykułu [V].

Każdy układ był symulowany w pudełku o trochę innym rozmiarze (aby utrzymać tę samą gęstość dla układów o różnej liczbie aminokwasów), a właściwości układu zależą od losowych warunków początkowych, dlatego porównywanie wielkości absolutnych jest obarczone pewnym błędem. Siłę i pracę można podzielić przez powierzchnię ściany pudełka, a liczbę splatań i kontaktów przez liczbę aminokwasów, zaś wszystkie wielkości uśrednić po dużej liczbie niezależnych symulacji - pozwoliłoby to na porównanie absolutnych wielkości. Istnieje jednak prostszy sposób: można porównać wyniki symulacji rodzaju 2. lub 3. (z oscylacjami) z wynikami symulacji rodzaju 1. (bez oscylacji, oznaczone tyldą  $\sim$ ). Rys. 5.11 pokazuje stosunek 5 wymienionych wyżej wielkości liczonych w symulacjach z periodycznymi odkształceniami (ścinającymi bądź normalnymi) i bez nich. Warto podkreślić, że nie jest to porównanie “przed i po”, lecz “zamiast”: układ po oscylacjach mógłby się zmienić nie z powodu oscylacji, lecz samego upływu czasu. Symulacje rodzaju 1. używały tego samego ziarna dla generatora liczb losowych co symulacje rodzajów 2. i 3. dlatego wpływ warunków początkowych także jest zaniedbywalny, a Rys. 5.11 faktycznie pokazuje jak wyniki zmieniają się dzięki oscylacjom.

odkształcenia normalne

odkształcenia ścinające



Rysunek 5.11: Stosunek 5 różnych wielkości obliczonych po 5 oscylacjach (w liczniku) oraz tych samych wielkości obliczonych bez oscylacji (w mianowniku, oznaczone przez  $\tilde{\cdot}$ ). Lewy i prawy panel odpowiadają periodycznym odkształceniom normalnym i ścinającym. Porównywane wielkości to średnia liczba splatań  $l_k$ , maksymalna siła  $F_{max}$  i praca  $W_{max}$  uzyskane podczas końcowego rozciągania, liczba kontaktów między łańcuchami  $n_{inter}$  oraz RMSF uśrednione po aminokwasach. Stosunek 1 oznaczono linią przerywaną. Każdy kolor odpowiada innemu układowi (podpis na górze) wg tabeli 5.1. Gęstość to  $3.5 \text{ aa}/\text{nm}^3$ . Połączenie zmodyfikowanych rys. 3 i S3 z artykułu [V].

Wyniki pokazane na Rys. 5.11 wykorzystują wartości uśrednione po wielu symulacjach (co najmniej 3), lecz błędy średniej są dla liczby splatań dość wysokie. Tym niemniej dla glutenu (a także dla ryżu w przypadku odkształceń normalnych) liczba splatań wyraźnie rośnie pod wpływem oscylacji. Samo-splatań (węzły) nie zostały uwzględnione, jednak także są obecne w glutenie: niektóre długie białka mogą tworzyć bardzo głębokie węzły, takie jak ten pokazany na lewym panelu Rys. 5.9. Mogą one utrzymywać się przez ponad  $100 \mu\text{s}$ , jednak żaden węzeł nie trwał dłużej niż  $130 \mu\text{s}$ .

Różnice między symulowanymi układami są najbardziej wyraźne w przypadku  $F_{\max}$ , które dla glutenin rośnie aż pięciokrotnie na Rys. 5.11 po wprowadzeniu oscylacji, natomiast dla gliadin wzrost jest bardzo niewielki. Wzrost dla glutenu jest znaczco większy niż dla białek kukurydzy i ryżu, a dla odkształceń normalnych przewyższa nawet (choć niewiele) gluteniny.

Maksymalna<sup>5</sup> praca  $W_{\max}$  potrzebna do rozciągnięcia układu także rośnie po oscylacjach. Ponownie wzrost jest największy dla glutenu i jego frakcji, choć dla glutenin widoczny jest nawet niewielki spadek w przypadku odkształceń normalnych. Wynika to z tego, że wysoka siła często odpowiada rozerwaniu układu, po którym siła może spaść blisko zera, a więc całkowita praca będzie niewielka (jak dla zielonej krzywej na Rys. 5.10). Kiedy rozerwanie nie jest kompletne, praca rośnie także po osiągnięciu  $F_{\max}$  (jak na morskiej krzywej z Rys. 5.10). Tym niemniej zwykle praca  $W_{\max}$  jest dodatnio skorelowana z  $F_{\max}$ . Warto zauważyć, że praca potrzebna do rozciągnięcia glutenin zmniejszyła się po deformacjach normalnych, które mogły rozplatać gluteniny tak, by łatwiej się rozdzielały podczas końcowego rozciągania. Oscylacje ścinające skutkują także większym wzrostem  $F_{\max}$ .

Wysokie  $F_{\max}$  może wynikać z trudnego do zerwania splatania lub z wielu kontaktów między łańcuchami, które trzeba rozerwać podczas rozciągania (patrz podpunkt 5.4.2). Dlatego warto zbadać liczbę kontaktów między łańcuchami,  $n_{\text{inter}}$ . Periodyczne deformacje najbardziej zwiększały  $n_{\text{inter}}$  dla białek kukurydzy, która jednak sumarycznie tworzą mniej kontaktów (jak pokazuje ich średnia liczba koordynacyjna  $z$  podana w tabeli 5.3).

RMSF jest jedyną wielkością, która wyraźnie zmniejsza się dla symulacji po periodycznych deformacjach. Kiedy wzrośnie liczba splatań i kontaktów między łańcuchami, białka są ściślej ze sobą powiązane, a przez to trudniej im się swobodnie poruszać, co powoduje zmniejszenie RMSF.

Silniej powiązane białka wykazują większą odporność na rozciąganie, ale rozerwanie układu może wtedy zajść dla mniejszej odległości między ścianami  $s$ . Średnia odległość  $s_F$  dla której występuje  $F_{\max}$  jest wyznaczona ze zbyt małą dokładnością (duży błąd średniej) aby umieszczać je na Rys. 5.11. Z tych samych powodów nie została pokazana średnia liczba mostków dwusiarczkowych.

<sup>5</sup>Praca jest nazywana maksymalną, mimo że gdyby prowadzić rozciąganie dalej, praca dalej by rosła. Z powodu fluktuacji termicznych siła może stać się ujemna (aminokwasy ciągną zamiast stawiać opór) - wtedy praca przestałaby być monotoniczna. W praktyce zdarza się to tylko dla ścian symulowanych metodą 2 z podpunktu 5.3.1, a więc nie ma znaczenia dla przedstawianych tu wyników. Wpływ ciśnienia atmosferycznego także jest zaniedbywalny, ponieważ odpowiada ono sile 0.01 nN działającej na ściany.

#### 5.4.2 Elastyczność glutenu na poziomie molekularnym

Zgodnie z obrazem “pętli i ciągów”, odpowiedź glutenu na deformację powinna być dwuetapowa: podczas pierwszego etapu białka rozciągają się, a puste przestrzenie między nimi znikają, co wymaga mniejszej siły niż rozplątanie i rozdzielenie białek, które zachodzi w drugim etapie [218, 239]. Przewidywania te zostały sprawdzone w symulacji poprzez śledzenie zmian 6 różnych wielkości podczas końcowego rozciągania układu (patrz Rys. 5.12). Ilustracje na tym rysunku pokazują formowanie się “ciągów”, a przesuwanie się względem siebie czerwonego i zielonego łańcucha powoduje, że siła potrzebna do dalszego rozciągania układu nie spada do zera wskutek konieczności zrywania kolejnych kontaktów międzyłańcuchowych. Ich liczba maleje na początku rozciągania, aby potem znów wzrosnąć podczas formowania się “ciągów”. łańcuchy stają się bardziej rozciągnięte, co potwierdza parametr asferyczności  $w$  [169] (uśredniony po wszystkich łańcuchach).<sup>6</sup>

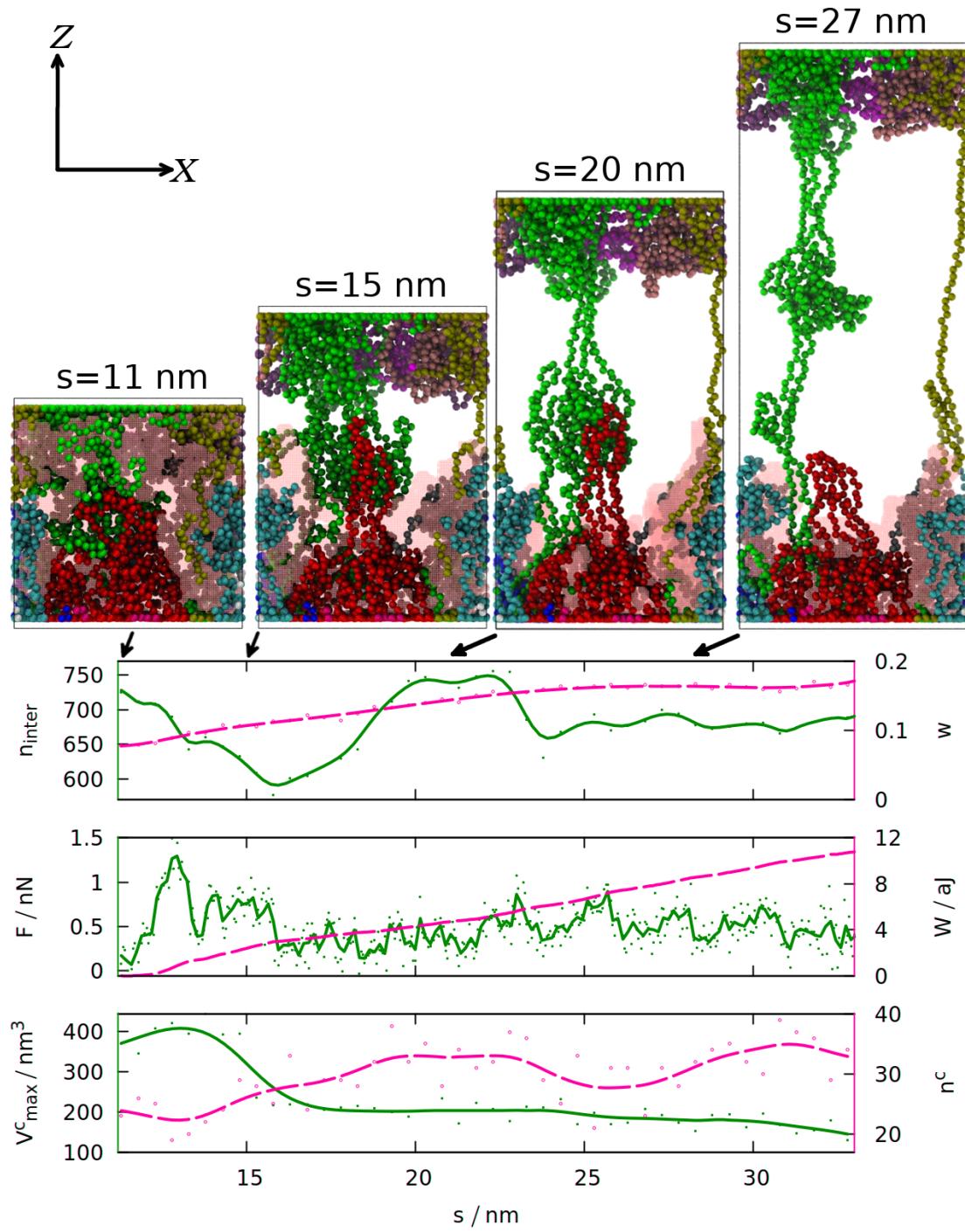
Utworzenie “ciągu” (dwóch równoległych fragmentów łańcuchów, połączonych wiązaniami wodorowymi) nie odpowiada maksimum siły, ponieważ  $F_{\max}$  występuje na początku rozciągania i jest związane z zerwaniem dużej liczby kontaktów między łańcuchami naraz. Tym niemniej siła nie spada do zera, a praca wykonana podczas rozciągania układu wciąż rośnie, ponieważ łańcuchy z przeciwnych końców pudełka nadal są połączone, co prowadzi do dalszego oporu przeciw rozciąganiu.

Aby określić liczbę i rozmiar wnęk, jakie tworzą się między aminokwasami, użyty został algorytm Spaceball [115, 116], który używa próbnika o zadanym promieniu, aby znaleźć trójwymiarowy kontur układu. Objętość wnęk jest obliczana poprzez wypełnienie pudełka siatką takich próbników i obracanie siatki względem układu. Aminokwasów jest dużo, dlatego oczka siatki odpowiadają stałej sieci  $1 \text{ \AA}$ , a rozmiar próbnika to  $3.8 \text{ \AA}$  (co odpowiada odległości między dwoma sąsiednimi atomami  $C_\alpha$ ). Siatka jest obracana tylko raz. Największa wnęka (różowy) obejmuje całą dolną połowę pudełka. Kiedy białka w pobliżu dolnej ściany rozciągają się coraz bardziej, rozmiar wnęki maleje i wydziela się z niej wiele mniejszych wnęk, co odzwierciedla wzrost  $n_c$ .

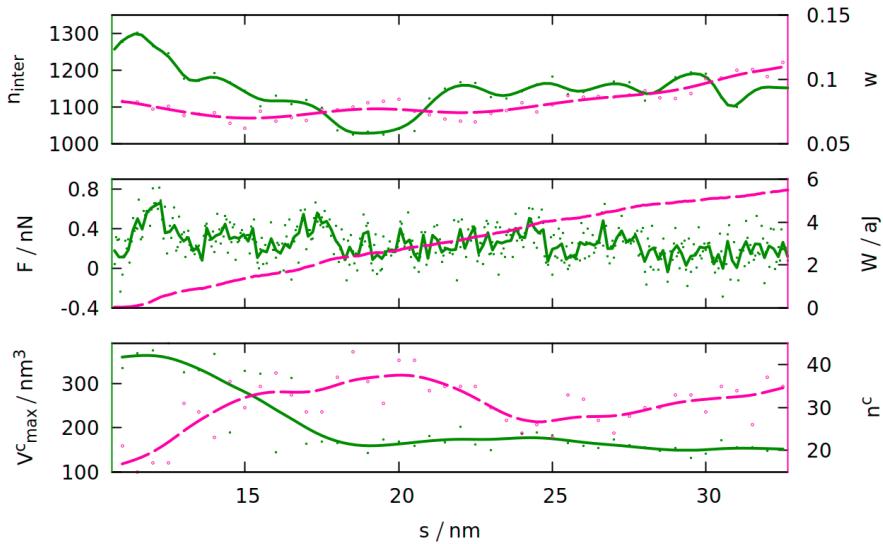
Rys. 5.12 pokazuje przebieg tylko jednej symulacji. Tą samą symulację dla innych losowych warunków początkowych (inne ziarno generatora liczb losowych) pokazuje Rys. 5.13.

Średnie z wielu symulacji także zdają się potwierdzać obraz “ciągów i pętli”: w większości przypadków objętość największej wnęki zmniejsza się wskutek rozciągania (patrz Rys. 5.14). łańcuchy stają się bardziej wydłużone nie tylko podczas końcowego rozciągania, ale także podczas oscylacji (patrz Rys. 5.15): najwyraźniej rozciągnięte białka nie wracają do poprzedniego kształtu, lecz pozostają wydłużone w kierunku deformacji (jest to swego rodzaju efekt pamięci: historia układu wpływa na jego stan).

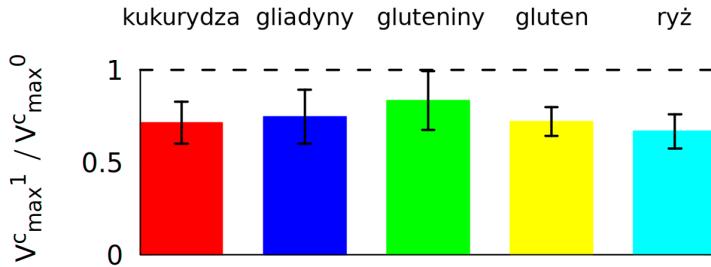
<sup>6</sup>Definicja  $w$  jest podana w paragrafie 2.5.8.4.



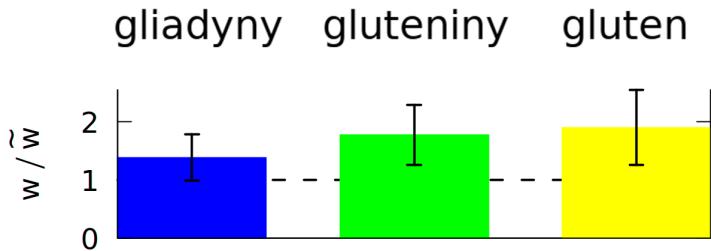
Rysunek 5.12: Zależność od czasu 6 wielkości obliczanych podczas ostatniego etapu symulacji glutenu (skład wg tabeli 5.1). Ciągłe zielone (lub purpurowe) linie odpowiadają wielkościom po lewej (lub prawej) stronie osi rzędnych. Linie są wygładzone, pełne (lub puste) kropki pokazują surowe dane z symulacji. Analizowane wielkości to: liczba kontaktów między łańcuchami  $n_{\text{inter}}$  i średni parametr asferyczności  $w$ , siła  $F$  i praca  $W$ , maksymalna objętość wnęki  $V_c^{\text{max}}$  i liczba wnęk  $n_c$ . 4 panele podobne do tych z Rys. 5.2 (tyle że w rzucie ortograficznym) odpowiadają odległościom  $s$  w podpisach, zaznaczonym takŜe strzałkami. Ten sam układ współrzędnych (na górze po lewej) jest użyty dla wszystkich ilustracji. Największa wnęka jest zaznaczona różową mgiełką. Modyfikacja rys. 4 z artykułu [V].



Rysunek 5.13: Oznaczenia jak na Rys. 5.12, tyle że dla symulacji z innym ziarnem generatora liczb losowych. Modyfikacja rys. S6 z artykułu [V].



Rysunek 5.14: Stosunek maksymalnej objętości wnęki na początku i końcu rozciągania podczas ostatniego etapu symulacji. Wielkość  $V_{\max}^0$  to średnia po pierwszej połowie procesu rozciągania, natomiast  $V_{\max}^1$  to średnia z drugiej połowy. Stosunek 1 oznaczono linią przerywaną. Każdy kolor odpowiada innemu układowi (podpis na górze) wg tabeli 5.1. Gęstość to  $3.5 \text{ aa/nm}^3$ . Modyfikacja rys. S4 z artykułu [V].



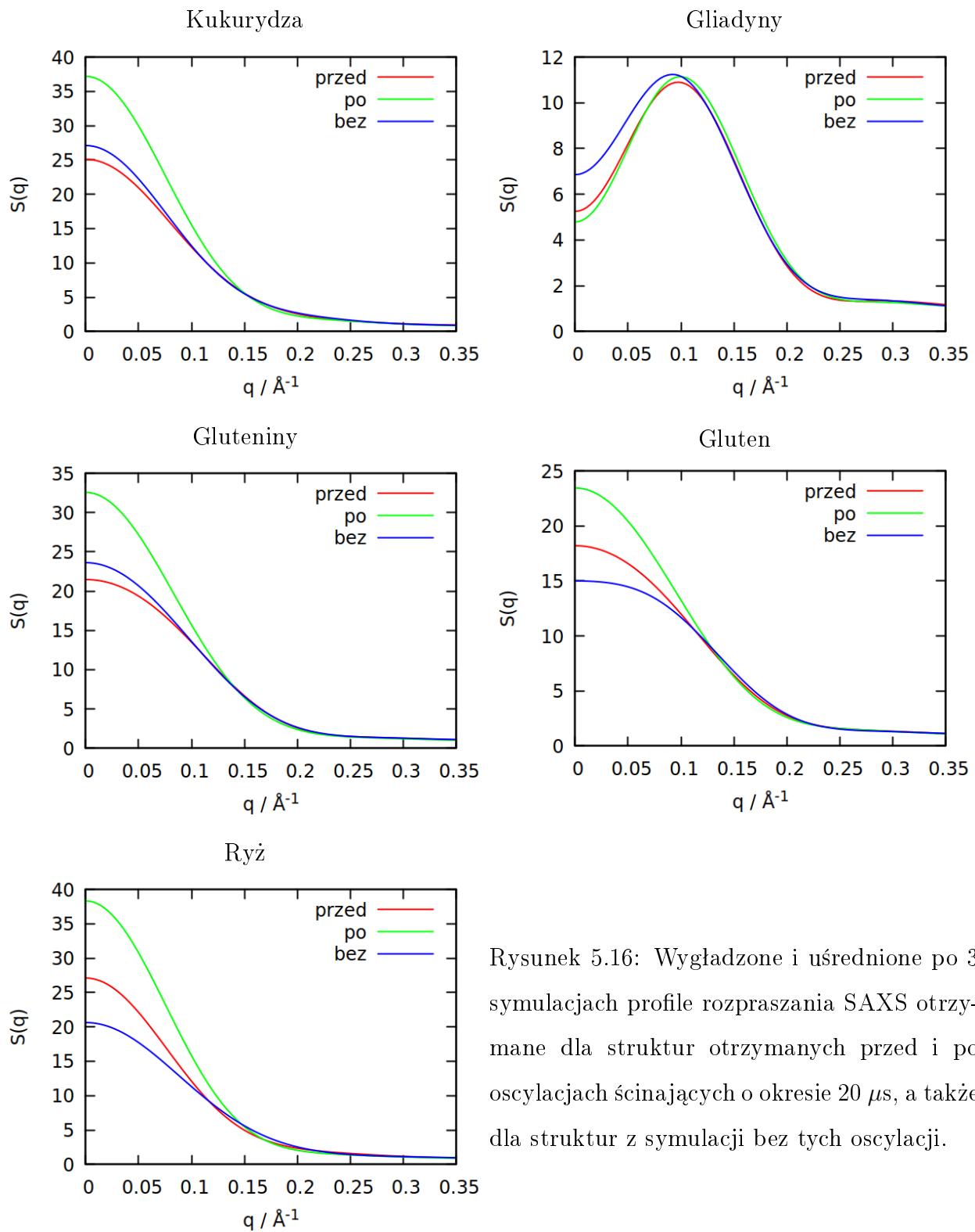
Rysunek 5.15: Stosunek średniego parametru asferyczności  $w$  [169] obliczonego po 5 oscylacjach ścinających (w liczniku) oraz bez żadnych oscylacji (w mianowniku, oznaczony przez  $\tilde{\cdot}$ ). Okres oscylacji to  $70 \mu\text{s}$  (krótsze okresy nie miały znaczącego wpływu na  $w$ ). Stosunek 1 oznaczono linią przerywaną. Każdy kolor odpowiada innemu układowi (podpis na górze) wg tabeli 5.1. Gęstość to  $3.5 \text{ aa/nm}^3$ . Modyfikacja rys. S2 z artykułu [V].

### 5.4.3 Próba odtworzenia krzywych SAXS

W paragrafie 2.6.8.4 użyta została metoda odtworzenia profilu rozpraszania SAXS ze struktur uzyskanych w symulacji komputerowej. Ponieważ profil SAXS jest transformatą Fouriera dwupunktowej funkcji korelacji  $g(r)$  [240], która określa prawdopodobieństwo, że w odległości  $r$  od jednego atomu znajdzie się inny atom, krzywą SAXS można teoretycznie odtworzyć dla każdego układu, dla którego można policzyć  $g(r)$ . W szczególności krzywe takie zostały odtworzone dla roztworów białek nieuporządkowanych o dużym stężeniu [241]. Dla oszacowania profili SAXS dla glutenu użyta została uproszczona wersja programu Fast-SAXS [242], w której przyjęto, że każdy aminokwas to jeden pseudoatom<sup>7</sup>, którego czynnik rozpraszania jest stały i równy 1. Ponieważ w symulacji pudełko do symulacji ma skończone rozmiary, profil rozpraszania będzie pokazywał niefizyczne efekty brzegowe, wynikające z tego, że wiele aminokwasów jest przyczepionych do ścian prostopadłych do osi  $Z$ , a w pozostałych kierunkach warunki brzegowe są periodyczne. W związku z tym przed wykonaniem transformaty Fouriera część krzywej  $g(r)$  dla  $r > R$  (gdzie  $R$  to połowa długości boku pudełka) została przemnożona przez czynnik  $\exp(-2(1 - r/R)^2)$ , co powoduje, że dla  $r > R$  krzywa  $g(r)$  gaussowsko zanika do zera. Podejście to nie bierze pod uwagę fizycznego sensu eksperymentu SAXS, w którym promienie X są rozpraszane w próbce o makroskopowych rozmiarach. Wytlumione  $g(r)$  odpowiada zatem roztworowi, w którym znajdowałoby się wiele identycznych, rozdzielonych od siebie agregatów glutenu, zamiast jednego makroskopowego agregatu (nie są uwzględnione duże odległości  $r$ ). W symulacji połowa boku pudełka  $R \approx 5$  nm, czyli warunek  $r < R$  odpowiada wektorowi rozpraszania  $q > 2\pi/R \approx 0.13 \text{ \AA}^{-1}$  (i ściśle rzecz biorąc tylko ten zakres  $q$  ma znaczenie fizyczne). Co więcej, anizotropowość warunków brzegowych i ściany przyciągającej mogą powodować kolejne zniekształcenia  $g(r)$  (patrz Rys. 5.4).

Mając te ograniczenia na uwadze warto zauważać, że uzyskane z symulacji profile rozpraszania z Rys. 5.4.3 zgadzają się jakościowo z eksperymentem [91], w którym tylko dla gliadyn profil rozpraszania był niemonotoniczny i posiadał maksimum (białka kukurydzy i ryżu nie były w nim rozpatrywane). Niestety krzywa rozpraszania dla gliadyn na Rys. 5.4.3 posiada maksimum w obszarze  $q < 0.13 \text{ \AA}^{-1}$ , co oznacza, że ta zbieżność wyników z doświadczeniem może nie wynikać z poprawności metody tylko być skutkiem efektów brzegowych. Wątpliwe jest także zwiększenie się intensywności rozpraszania w granicy  $q \rightarrow 0$  po oscylacjach. Intensywność dla  $q = 0$  jest proporcjonalna do ściśliwości izotermicznej, a zatem po oscylacjach powinna maleć a nie rosnąć (układ po obróbce mechanicznej trudniej rozciągnąć). Prawdopodobnie wzrost intensywności dla małych  $q$  wynika z tego, że po oscylacjach więcej aminokwasów jest przyczepionych do ściany, w wyniku czego mogą znaleźć się one bliżej siebie.

<sup>7</sup>A funkcja  $g(r)$  to histogram wszystkich odległości między aminokwasami.



Rysunek 5.16: Wygładzone i uśrednione po 3 symulacjach profile rozpraszania SAXS otrzymane dla struktur otrzymanych przed i po oscylacjach ścinających o okresie  $20 \mu\text{s}$ , a także dla struktur z symulacji bez tych oscylacji.

Obliczenia krzywych rozproszeniowych zostały wykonane przy użyciu programu napisanego na podstawie Fast-SAXS przez dr hab. Bartosza Różyckiego.

#### 5.4.4 Moduł ścinania glutenu

Jedną z niewielu wielkości charakteryzujących reologię glutenu niezależnie od kształtu i wielkości próbki jest dynamiczny moduł ścinania  $G^* = G' + iG''$ , mierzony dla odkształcenia ścinającego zmieniającego się sinusoidalnie w czasie (dla odkształcenia normalnego odpowiada mu dynamiczny moduł Younga  $E^*$ ) [238]. Część rzeczywista ( $G'$ ) odpowiada za naprężenia zgodne w fazie, a część urojona ( $G''$ ) za odpowiedź przesuniętą w fazie (różnica faz  $\delta$  jest określona jako  $\tan \delta = \frac{G''}{G'}$ ). Pomiar  $G^*$  zwykle nie niszczy próbki, ponieważ względna amplituda odkształcenia jest mała (zwykle do kilku procent [90]). Całkowita siła  $F_X(t)$  oddziaływanego na ściany w kierunku  $X$  (patrz Rys. 5.3), podzielona przez powierzchnię ściany  $S$ , definiuje naprężenie ścinające  $\phi(t) = F_X(t)/S$  (przyjęty symbol naprężenia ścinającego  $\tau$  oznacza jednostkę czasu, więc został użyty symbol  $\phi$ ). Zakłada się, że odkształcenie postaci  $\gamma = \gamma_0 \cos(\omega t)$  wywoła naprężenie podobnej postaci, przesunięte w fazie o  $\delta$ :  $\phi(t) = \phi_0 \cos(\omega t + \delta)$ . Dynamiczny moduł ścinania  $G^* = G' + iG''$  jest wtedy zdefiniowany jako [243]:

$$G' = \frac{\phi_0}{\gamma_0} \cos \delta \quad G'' = \frac{\phi_0}{\gamma_0} \sin \delta \quad (5.1)$$

gdzie  $\phi_0$  oraz  $\gamma_0$  to odpowiednio amplitudy naprężenia i odkształcenia. Można wyznaczyć  $G^*$  poprzez zastosowanie odkształcenia  $s'(t) = A \cos(\omega t)$  i dopasowanie do siły działającej na ściany funkcji  $F(t) = F_0 \cos(\omega t + \delta)$ , gdzie  $\delta$  to różnica faz,  $\gamma_0 = A/s_0$ ,  $\phi_0 = F_0/S$ . Jest to bardzo prosta metoda wyznaczania  $G^*$ , istnieją dużo bardziej zaawansowane sposoby [244].

Tabela 5.4 i Rys. 5.17 pokazują porównanie symulacji z doświadczeniem. Tabela 5.4 potwierdza, że gliadyny są odpowiedzialne za lepkość glutenu (duże  $\delta$ ), a gluteniny za elastyczność (małe  $\delta$ ). Pomimo dużej rozbieżności danych z symulacji i doświadczenia, rozróżnienie między gliadynami i gluteninami wciąż jest widoczne w symulacji, dla której zastosowano najdłuższy użyty okres oscylacji (70  $\mu\text{s}$ ):  $\tan \delta_{\text{sim}}$  jest największy dla gliadyn i najmniejszy dla glutenin.

System	$G'$ [kPa]	$G''$ [kPa]	$\tan \delta$	$G'_{\text{sim}}$ [MPa]	$G''_{\text{sim}}$ [MPa]	$\tan \delta_{\text{sim}}$
Gluten	2.5	1.3	0.5	$6.0 \pm 1.6$	$6.1 \pm 1.7$	$1.0 \pm 0.1$
Gliadyny	0.5	0.6	1.2	$5.6 \pm 1.7$	$8.0 \pm 2.3$	$1.4 \pm 0.1$
Gluteniny	40	9	0.2	$8.3 \pm 2.2$	$7.9 \pm 2.1$	$0.95 \pm 0.1$

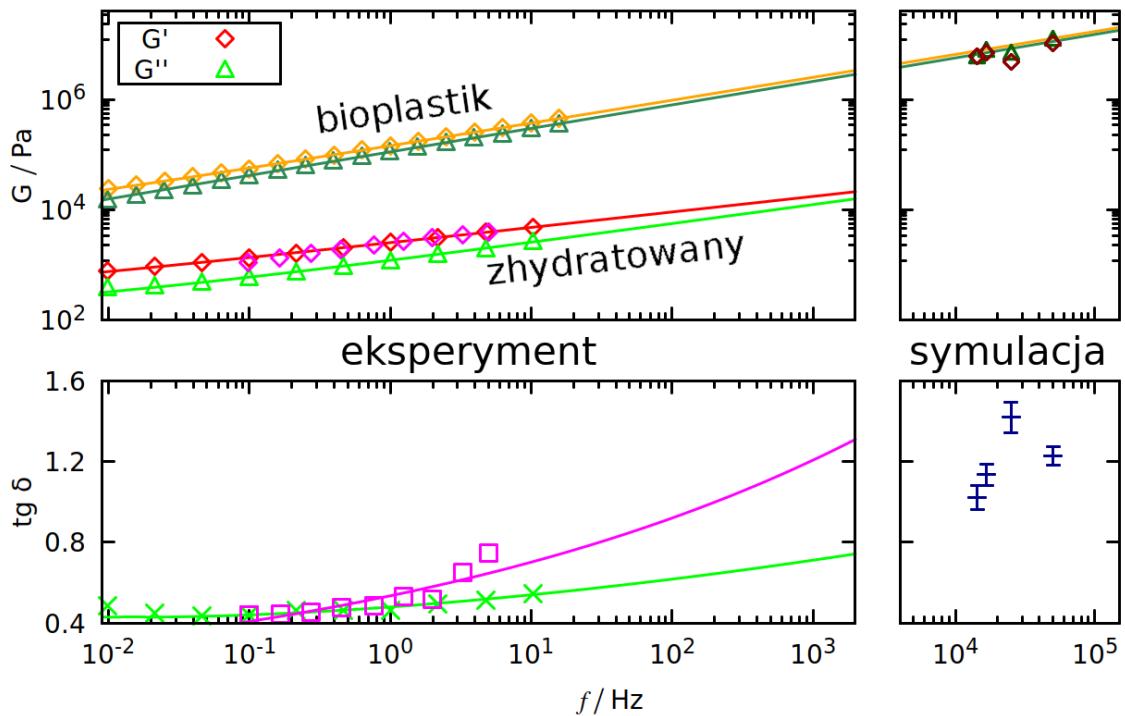
Tabela 5.4: Dynamiczny moduł ścinania wysuszonego glutenu i jego frakcji, wyznaczony w eksperymencie z częstotliwością oscylacji  $f = 1 \text{ Hz}$  [90] oraz w symulacji (oznaczony indeksem  $\text{sim}$ ) z częstotliwością  $f \approx 14 \text{ kHz}$  (odpowiadającej okresowi 70  $\mu\text{s}$ ).

Użyte częstotliwości są 4 razy większe niż w eksperymencie, jednak można dokonać ekstrapolacji na podstawie tego jak doświadczalny moduł  $G^*$  zależy od częstotliwości. Rys. 5.17 pokazuje takie zależności dla 3 niezależnych eksperymentów [246, 247, 248]. Jego dolny panel pokazuje, że  $\tan \delta$

powinno rosnąć wraz z częstotliwością. Uzyskany w symulacji tan  $\delta_{\text{sim}}$  znajduje się w zakresie między ekstrapolacją przewidzianą przez [246] (zielona krzywa) oraz przez [247] (purpurowa krzywa). Duże różnice w ekstrapolacjach doświadczalnych wynikają z różnic w składzie glutenu, które znacznie wpływają na wyniki (purpurowe punkty na Rys. 5.17 pochodzą z zestawu danych Aubaine z artykułu [247]).

Rosnąca  $\delta$  oznacza, że  $G'$  i  $G''$  zbiegają do tej samej wartości, jak pokazuje górnego panelu Rys. 5.17. Zarówno  $G'$  jak i  $G''$  także rosną wraz z częstotliwością, jednak ekstrapolacja tego wzrostu nie odpowiada danym z symulacji dla danych dotyczących zhydratowanego glutenu [246, 247]. Być może dla dużych częstotliwości zależność przestaje być potęgowa. Za różnicę może też odpowiadać bardzo mały rozmiar symulowanego układu i ograniczenia zastosowanego modelu. Metoda ukrytego rozpuszczalnika nie może uwzględnić różnic w poziomie hydratacji glutenu, które mogą znacznie wpływać na wyniki (nawet "wysuszony" gluten jest wciąż zhydratowany) [246].

Z drugiej strony wyniki symulacji są w dobrej zgodności z doświadczeniami wykonanymi na bioplastikach opartych na glutenie: białka glutenu zostały w nich zmieszane z glicerolem i stopione w temperaturze ponad 380 K [248]. Jednak nawet dla bioplastków wytrzymałość na rozciąganie (czyli naprężenie potrzebne do rozerwania próbki) jest mniejsza niż 1 MPa [248], podczas gdy  $F_{\text{max}}$  podzielone przez powierzchnię ściany (około 100 nm<sup>2</sup>) może być ponad 10 razy większe (patrz Rys. 5.10).



Rysunek 5.17: Dynamiczny moduł ścinania (górnego panelu) oraz tan  $\delta$  (dolnego panelu) w funkcji częstotliwości oscylacji  $f$ . Wyniki z symulacji są po prawej, z eksperymentu po lewej (purpurowe punkty pochodzą z [247], czerwone i zielone z [246], pomarańczowe i ciemnozielone z [248])). Dopasowane krzywe mają postać  $a \cdot f^b + c$ , gdzie  $a$ ,  $b$  oraz  $c$  to współczynniki dopasowania. Modyfikacja rys. 5 z artykułu [V].

## 5.5 Podsumowanie

Model gruboziarnisty z ukrytym rozpuszczalnikiem pozwolił na zrozumienie elastyczności glutenu na poziomie mikroskopowym. Ugniatanie i mieszanie próbek glutenu zwiększa ich wytrzymałość na rozciąganie [245]. W symulacjach te czynności były reprezentowane przez periodyczne deformacje. W wyniku tych deformacji gluten i gluteniny stawały się znacznie odporniejsze na rozciąganie, co potwierdza wzrost  $F_{\max}$  oraz  $W_{\max}$ .

Model był także w stanie poprawnie odróżnić gluten od białek innych roślin, a także uchwycić różnicę między wytrzymałyymi gluteninami i lepkimi gliadynami. Różnice te wydają się pochodzić z połączonego efektu splatań i kontaktów między łańcuchami (które w przypadku glutenu odpowiadają głównie wiązaniom wodorowym). Pod wpływem obróbki mechanicznej białka kukurydzy tworzyły więcej kontaktów między łańcuchami niż białka ryżu (co tłumaczy czemu to mąka kukurydziana jest bardziej przydatna przy tworzeniu ciasta bez glutenu [249]). Białka kukurydzy były jednak za krótkie, aby poprzez splatania zwiększyć swoją odporność na rozciąganie.

Rola splatań i wiązań wodorowych między łańcuchami jest dobrze znana w literaturze dotyczącej elastyczności glutenu [94, 219], jednak przedstawione w tej rozprawie symulacje po raz pierwszy pokazują tę rolę na poziomie pojedynczych białek. Symulacje te stanowią dodatkowe potwierdzenie obrazu "pętli i ciągów" [218]. Podczas rozciągania zmniejszeniu objętości największej wnęki towarzyszył wzrost liczby kontaktów między łańcuchami.

Z powodu małego rozmiaru układu tylko kilka mostków dwusiarczkowych między różnymi łańcuchami było naraz obecnych podczas symulacji (najwięcej dla białek kukurydzy) i nie wydawały się one znacznie wpływać na wyniki. Możliwe, że mostki stają się ważniejsze w próbkach o większych rozmiarach, gdzie kowalencyjne wiązania między łańcuchami wpływają na takie właściwości jak rozpuszczalność [137]. Dla około 20 białek glutenu splatania i oddziaływanie niekowalencyjne okazały się wystarczająco dobrze tłumaczyć elastyczność układu.

Periodyczne deformacje zostały wykorzystane do obliczenia dynamicznego modułu ścinania  $G^*$ . Jego zależność od częstości pokazana na Rys. 5.17 zgadza się z wynikami doświadczalnymi dla bioplastiku z glutenu [248] lepiej niż z wynikami dla zwykłego, niemodyfikowanego glutenu [246, 247]. W modelu z ukrytym rozpuszczalnikiem glicerol może w teorii zastąpić wodę jako rozpuszczalnik (model był jednak parametryzowany z myślą o wodzie jako rozpuszczalniku **[I]**). Doświadczalny zakres częstości jest o parę rzędów mniejszy niż ten z symulacji. Aby dojść do dłuższych skal czasowych potrzeby jest jeszcze bardziej uproszczony opis. Innym powodem otrzymania w symulacji

tak dużego  $G^*$  może być sama gruboziarnistość modelu: efektywna gęstość może być inna (patrz podpunkt 5.3.2), co może prowadzić do różnic w elastyczności. Podobną rozbieżność zaobserwowano w przypadku elastyczności wirusów [250].

Pomimo decydującej roli glutenu w określaniu jakości ciasta, moduł ścinania glutenu  $G^*$  nie jest z tą jakością silnie skorelowany [90]. Podczas pieczenia bąble gazu wewnętrz ciasta rozciągają je daleko poza limit odkształceń liniowych. Dlatego pomiary wykonane podczas rozciągania glutenu do ponad 200% początkowego rozmiaru dają bardziej wartościowe informacje [90]. Ostatni etap przedstawionych tu symulacji odpowiadał właśnie takim pomiarom. Jednak wielkości takie jak  $W_{\max}$  czy  $F_{\max}$  zależą od geometrii rozciągania i wielkości próbki, dlatego trudno je bezpośrednio porównać z eksperymentem.

Inną hipotezą, która mogła być sprawdzona dzięki symulacjom, było tworzenie przez białka glutenu struktury  $\beta$ -spirali składającej się z sąsiednich  $\beta$ -skrętów tworzonych przez powtarzające się fragmenty takie jak QPGQ [93] (są one szczególnie częste w gluteninach [214]). Pomiary skaningowym mikroskopem tunelowym wykazały istnienie odpowiadającego  $\beta$ -spirali rowka o okresowości 1.5 nm w niektórych gluteninach [214]. Proces przygotowania próbek w tych doświadczeniach uniemożliwiał sprawdzenie, czy taka spirala pojawia się także w mieszaninie wielu białek glutenu. Modele pojedynczych białek glutenu składające się z następujących po sobie  $\beta$  [251] i  $\gamma$  [92] skrętów były opisywane w literaturze, jednak podobne struktury nie zostały zaobserwowane w opisywanych tu symulacjach (choć  $\beta$ -skręty były obecne).

Nie została zbadana zależność właściwości glutenu od temperatury, ponieważ dane z eksperymentów kalorymetrycznych są niejednoznaczne [216]. Temperatura przejścia szklistego dla niezhydratowanych białek glutenu (ze zredukowanymi mostkami dwusiarczkowymi) to około 400 K. Wraz ze zwiększeniem stopnia hydratacji temperatura przejścia może spaść poniżej 300K [252]. Ze względu na tak silną zależność od stopnia hydratacji trudno byłoby porównać tę temperaturę z eksperymentem.

Nie było też dyskutowane oddziaływanie białek glutenu z białkami układu odpornościowego człowieka, chociaż jest to bardzo interesujący temat badań [253, 254]. Takie oddziaływanie także mogą być symulowane przedstawionym tu modelem. Można także obliczyć, które białka glutenu są lepiej dostępne dla rozpuszczalnika, a przez to dla enzymów trawiennych w żołądku, co może rzucić nowe światło na problem nietolerancji glutenu [255]. Inną modyfikacją modelu jest uwzględnienie “utleniających” i “redukujących” warunków tworzenia mostków dwusiarczkowych (poprzez zmianę potencjału je opisującego) i sprawdzenie jak zmiana tych warunków wpływa na właściwości glutenu. Np. gluten ugniatany w atmosferze beztlenowej ma niższą jakość [232].

# Bibliografia

- [1] V. N. Uversky, Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta*, 2013, **1834**, 932-951.
- [2] A. L. Fink, Natively unfolded proteins. *Curr. Opin. Struct. Biol.* 2005, **15**, 35-41.
- [3] V. N. Uversky, A. K. Dunker, Understanding proteins non-folding. *Biochem. Biophys. Acta*, 2010, **1804**, 1231-1264.
- [4] V. N. Uversky, Natively unfolded proteins: a point where biology waits for physics. *Prot. Sci.* 2002, **11**, 739-756.
- [5] H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 2005, **6**, 197-208.
- [6] A. K. Dunker, I. Silman, V. N. Uversky, J. L. Sussman, Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 2008, **18**, 756-764.
- [7] V. N. Uversky, A. K. Dunker, Understanding protein non-folding. *Biochem. Biophys. Acta*, 2010, **1804**, 1231-1264.
- [8] A. C. M. Ferreon, C. R. Moran, Y. Gambin, A. A. Deniz, Single-molecule fluorescence studies of intrinsically disordered proteins. *Methods Enzymol.*, 2010, **472**, 179-204.
- [9] M. M. Babu, R. van der Lee, N. S. de Groot, J. Gsponer, Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* 2011, **21**, 432-440.
- [10] P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 2015, **6**, 18-29.
- [11] P. E. Wright, H. J. Dyson, Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999, **293**, 321-331.
- [12] V. N. Uversky, Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) is the Limit. *Int. J. Mol. Sci.* 2019, **20**, 4186.
- [13] A. M. Monzon, M. Necci, F. Quaglia, I. Walsh, G. Zanotti, D. Piovesan, S. C. E. Tosatto, Experimentally Determined Long Intrinsically Disordered Protein Regions Are Now Abundant in the Protein Data Bank. *Int. J. Mol. Sci.* 2020, **21**(12), 4496.

- [14] B. Schuler, E. A. Lipman, P. J. Steinbach, M. Kumke, W. A. Eaton, Polyproline and the “spectroscopic ruler” revisited with single-molecule fluorescence. *Proc. Natl. Acad. Sci. (USA)*, 2005, **102**(8), 2754–2759.
- [15] S. Sanyal, D. F. Coker, D. MacKernan, How flexible is a protein: simple estimates using FRET microscopy. *Mol. BioSyst.* 2016, **12**(4), 2988–2991.
- [16] C. Cragnell, D. Durand, B. Cabane, M. Skepo, Coarse-grained modeling of the intrinsically disordered protein histatin 5 in solution: Monte carlo simulations in combination with SAXS. *Proteins*, 2016, **84**, 777-791.
- [17] A. De Biasio, A. Ibáñez de Opakua, T. N. Cordeiro, F. J. Blanco i in. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophys. J.* 2014, **106**(4), 865–874. Dane SAXS odzyskane z: <https://web.archive.org/web/20160911130751/http://pedb.vib.be/accession.php?ID=PED6AAA> (12.05.2020).
- [18] B. Różycki, Y. C. Kim, G. Hummer, SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure*, 2011, **19**(1), 109-116.
- [19] M. Varadi, S. Kosol, P. Lebrun, A. K. Dunker, D. I. Svergun, V. N. Uversky, M. Vendruscolo, D. Wishart, P. E. Wright, P. Tompa i in. PE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucl. Acids Res.* 2014, **42**(D1), D326-D335.
- [20] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, A. Laio, Exploring the universe of protein structures beyond the Protein Data Bank. *PLoS Comp. Biol.* 2010, **6**, e1000957.
- [21] A. Möglich, K. Joder, T. Kieffhaber, End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci. (USA)*, 2006, **103**(33), 12394–12399.
- [22] J. Henriques, C. Cragnell, M. Skepo, Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theor. Comp.* 2015, **11**(7), 3420–3431.
- [23] C. Cragnell, E. Rieloff, M. Skepö, Utilizing coarse-grained modelling and monte carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. *J. Mol. Biol.* 2018, **430**(16), 2478-2492.
- [24] J. A. McCammon, B. R. Gelin, M. Karplus, Dynamics of folded proteins. *Nature*, 1977, **267**, 585.

- [25] M. Levitt, A. Warshel, Computer simulations of protein folding. *Nature*, 1975, **253**, 694-698.
- [26] J. A. McCammon, S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press, Cambridge, 1987.
- [27] J. Gao, K. Kuczera, B. Tidor, M. Karplus, Hidden thermodynamics of mutant proteins: a molecular dynamics analysis. *Science*, 1989, **244**, 1069-1072.
- [28] T. Schlick, *Molecular Modeling and Simulations: An interdisciplinary guide*, 2nd Edition; Springer, New York, 2010.
- [29] Y. Duan, P. A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 1998, **282**, 740-744.
- [30] A. Sethi, J. Tian, D. M. Vu, S. Gnanakaran, Identification of minimally interacting modules in an intrinsically disordered protein. *Biophys. J.* 2012, **103**, 748-757.
- [31] A. Vitalis, X. Wang, R. V. Pappu, Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories. *Biophys. J.* 2007, **93**, 1923-1937.
- [32] L. Esposito, A. Paladino, C. Pedone, L. Vitagliano, Insights into structure, stability, and toxicity of monomeric and aggregated polyglutamine models from molecular dynamics simulations. *Biophys. J.* 2008, **94**, 4031-40.
- [33] H. Ogawa, M. Nakano, H. Watanabe, E. B. Starikov, S. M. Rothstein, S. Tanaka, Molecular dynamics simulation study on the structural stabilities of polyglutamine peptides. *Comp. Biol. Chem.* 2008, **32**, 102-110.
- [34] Á. Gómez-Sicilia, M. Sikora, M. Cieplak, M. Carrión-Vázquez, An exploration of the universe of polyglutamine structures *PLoS Comp. Biol.* 2015, **11**, e1004541.
- [35] C. Czaplewski, S. Kalinowski, A. Liwo, H. A. Scheraga, Application of Multiplexed Replica Exchange Molecular Dynamics to the UNRES Force Field: Tests with  $\alpha$  and  $\alpha+\beta$  Proteins. *J. Chem. Theory Comput.* 2009 **5**(3), 627-640.
- [36] L. Duan, T. Zhu, C. Ji, Q. Zhang, J. Z. H. Zhang, Direct folding simulation of helical proteins using an effective polarizable bond force field. *Phys. Chem. Chem. Phys.* 2017, **19**, 15273-15284.
- [37] S. Rauscher, R. Pomès, Molecular simulations of protein disorder. *Biochem. Cell Biol.* 2010, **88**(2), 269-290.

- [38] T. Terakawa, S. Takada, Multiscale ensemble modeling of intrinsically disordered proteins: p53 n-terminal domain. *Biophys. J.* 2011, **101**(6), 1450-1458.
- [39] X. Liu, J. Chen, Hyres: a coarse-grained model for multi-scale enhanced sampling of disordered protein conformations. *Phys. Chem. Chem. Phys.* 2017, **19**(48), 1463-9076.
- [40] Y. Wang, G. A. Voth, Molecular dynamics simulations of polyglutamine aggregation using solvent-free multiscale coarse-grained models. *J. Phys. Chem. B*, 2010, **114**(26), 8735-8743.
- [41] A. Dawid, D. Gront, A. Koliński, SURPASS Low-Resolution Coarse-Grained Protein Modeling, *J. Chem. Theory Comput.* 2017, **13**(11), 5766-5779.
- [42] C. Czaplewski, A. Karczynska, A. K. Sieradzan, A. Liwo, UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucl. Acids Res.* 2018, **46**(W1), W304-W309.
- [43] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 2007, **111**, 7812-7824.
- [44] M. Baaden, S. J. Marrink, Coarse-grain modelling of protein-protein interactions. *Curr. Op. Struct. Biol.* 2013, **23**, 878-886.
- [45] T. Bereau, M. Deserno, Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* 2009, **130**, 235106.
- [46] A. B. Poma, M. Cieplak, P. E. Theodorakis, Combining the MARTINI and Structure-Based Coarse-Grained Approaches for the Molecular Dynamics Studies of Conformational Transitions in Proteins. *J. Chem. Theory Comput.* 2017, **13**(3), 1366-1374.
- [47] H. Wu, P. G. Wolynes, G. A. Papoian, Awsem-idp: A coarse-grained force field for intrinsically disordered proteins. *J. Phys. Chem. B*, 2018, **122**(49), 11115-11125.
- [48] J. Gu, F. Bai, H. Li, X. Wang, A generic force field for protein coarse-grained molecular dynamics simulations. *Int. J. Mol. Sci.* 2012, **13**, 14451-14469.
- [49] Y. C. Kim, G. Hummer, Coarse-grained models for simulation of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* 2008, **375**, 1416-1433.
- [50] C. Clementi, H. Nymeyer, J. N. Onuchic, Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation of small globular proteins. *J. Mol. Biol.* 2000, **298**, 937-953.

- [51] T. X. Hoang, M. Cieplak, Molecular dynamics of folding of secondary structures in Go-like models of proteins. *J. Chem. Phys.* 2000, **112**, 6851-6862.
- [52] J. Karanicolas, C. L. Brooks III, The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 2002, **11**(10), 2351-2361.
- [53] J. I. Sułkowska, M. Cieplak, Selection of optimal variants of Go-like models of proteins through studies of stretching. *Biophys. J.* 2008, **95**, 3174-3191.
- [54] M. Cieplak, M. O. Robbins, Nanoindentation of 35 virus capsids in a molecular model: Relating mechanical properties to structure. *PLOS ONE*, 2013, **8**, e63640.
- [55] G. Polles, G. Indelicato, R. Potestio, P. Cermelli, R. Twarock, C. Micheletti, Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition. *PLOS Comp. Biol.* 2013, **9**, e1003331.
- [56] K. Wołek, M. Cieplak, Self-assembly of model proteins into virus capsids. *J. Phys. Condens. Matter*, 2017, **47**, 474003.
- [57] M. Sikora, J. I. Sułkowska, M. Cieplak, Mechanical strength of 17 134 model proteins and cysteine spliknots. *PLoS Comp. Biol.* 2009, **5**, e1000547.
- [58] A. Valbuena, J. Oroz, R. Hervas, A. M. Vera, D. Rodriguez, M. Menendez, J. I. Sułkowska, M. Cieplak, M. Carrion-Vazquez, On the remarkable mechanostability of scaffoldins and the mechanical clamp motif. *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 13791-13796.
- [59] Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science*, 2017, **357**, eaaf4382.
- [60] A. E. Posey, A. S. Holehouse, R. V. Pappu, Phase separation of intrinsically disordered proteins. *Methods Enzymol.* 2018, **611**, 1-30.
- [61] B. Monterroso, S. Zorrill, M. Sobrinos-Sanguino, C. D. Keating, German Rivas, Microenvironments created by liquid-liquid phase transition control the dynamic distribution of bacterial FtsZ protein. *Sci. Reports*, 2016, **6**, 35140.
- [62] J. Berry, S. C. Weber, N. Vaidya, M. Haataja, C. P. Brangwynne, RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. USA*, 2015, **112**, E5237-E5245.

- [63] F.-M. Boisvert, S. van Koningsbruggen, J. Navascues, A. I. Lamons, The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.* 2007, **8**, 574-585.
- [64] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege i in. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 2009, **324**, 1729-1732.
- [65] C P. Brangwynne, T. J. Mitchison, A. A. Hyman, Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes. *Proc. Natl. Acad. Sci. USA*, 2011, **108**, 4334-4339.
- [66] C. M. Caragine, S. C. Haley, A. Zidovska, Surface fluctuation and coalescence of nucleolar droplets in the human cell nucleus. *Phys. Rev. Lett.* 2018, **121**, 148101.
- [67] R. Kurita, H. Tanaka, On the abundance and general nature of the liquid-liquid phase transition in molecular systems. *J. Phys.: Condens. Matter*, 2005, **17**, L293-L-302.
- [68] M. W. C. Dharmawardana, D. D. Klug, R. C. Remsing, Liquid-Liquid Phase Transitions in Silicon. *Phys. Rev. Lett.* 2020, **125**, 075702.
- [69] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* 2018, **14**, e1005941.
- [70] H. S. Ashbaugh, H. W. Hatch, Natively unfolded protein stability as a coil-to-globule transition in charge/hydrophobicity space. *J. Am. Chem. Soc.* 2008, **130**, 9536-9542.
- [71] S. Miyazawa, R. L. Jernigan, Residue-residue potentials with a favourable contact pair term and an unfavourable high packing density term, for simulation and threading. *J. Mol. Biol.* 1996, **256**, 623-644.
- [72] M. DiFiglia, E. Sapp, K. O. Chase, S. W. Davies, G. P. Bates, J. P. Vonsattel, N. Aronin, Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, 1997, **277**, 1990-1993.
- [73] J. K. Cooper, G. Schilling, M. F. Peters, V. L. Dawson, T. M. Dawson, C. A. Rose i in. Truncated N-terminal fragments of huntingtin with expanded glutamine repeats form nuclear and cytoplasmic aggregates in cell culture. *Hum. Mol. Genet.* 1998, **7**, 783-790.
- [74] M. Y. Tang, C. J. Proctor, J. Woulfe, D. A. Gray, Experimental and computational analysis of polyglutamine-mediated cytotoxicity. *PLoS Comp. Biol.* 2010, **6**, e1000944.
- [75] D. Bak, M. Milewski, Choroby związane z agregacją białek. *Postępy Biochemii*, 2005, **51**, 297-307.

- [76] H. Zhang, S. Elbaum-Garfinkle, E. M. Langdon, N. Taylor, P. Occhipinti, A. Bridges, C. P. Branwynne, A. S. Gladfelter, RNA controls polyQ protein phase transitions. *Mol. Cell*, 2015, **60**, 220-230.
- [77] G. Schlissel, M. K. Krzyzanowski, F. Caudron, Y. Barral, J. Rine, Aggregation of the Whi3 protein, not loss of heterochromatin, causes sterility in old yeast cells. *Science*, 2017, **355**, 1184-1187.
- [78] M. E. MacDonald, J. F. Gusella, Huntington's disease: translating a CAG repeat into a pathogenic mechanism. *Curr. Opin. Neurobiol.* 1996, **6**, 638-643.
- [79] M. Wojciechowski, Á. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, Unfolding knots by proteasome-like systems: simulations of the behavior of folded and neurotoxic proteins. *Mol. BioSyst.* 2016, **12**, 2700-2712.
- [80] A. San Martin, P. Rodriguez-Aliaga, J. A. Molina, A. Martin, C. Bustamante, M. Baez, Knots can impair protein degradation by ATP-dependent proteases. *Proc. Natl. Acad. Sci. USA*, 2017, **114**, 9864-9869.
- [81] N. F. Bence, R. M. Sampat, R. R. Kopito, Impairment of the ubiquitin-proteasome system by protein aggregation. *Science*, 2001, **292**, 1552-1555.
- [82] C. L. Wellington, L. M. Ellerby, A. S. Hackam, R. L. Margoils, M. A. Trifiro, C. A. Ross, D. W. Nicholson, D. E. Bredesen, M. R. Hayden i in. Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J. Biol. Chem.* 1998, **273**, 9158-9167.
- [83] I. Sanchez, C.-J. Xu, P. Juo, A. Kakizaka, J. Blenis, J. Yuan, Caspase-8 is required for cell death induced by expanded polyglutamine repeats. *Neuron*, 1999, **22**, 623-33.
- [84] T. R. Peskett, F. Rau, J. O'Driscoll, R. Patani, A. R. Lowe, H. R. Saibil, A liquid to solid phase transition underlying pathological huntingtin exon1 aggregation. *Mol. Cell*, 2018, **70**, 588-601.
- [85] S. Elbaum-Garfinkle, Matter over mind: Liquid phase separation and neurodegeneration. *JBC Reviews*, 2019, **294**, 7160-7168.
- [86] N. Scarafone, C. Pain, A. Fratamico, G. Gaspard, N. Yilmaz, P. Filee, M. Galleni, A. Matagne, M. Dumoulin, Amyloid-like fibril formation by polyQ proteins: a critical balance between the polyQ length and the constraints imposed by host protein. *PLoS ONE*, 2012, **7**, e31253.

- [87] J. Wen, D. R. Scoles, J. C. Facelli, Molecular dynamics analysis of the aggregation propensity of polyglutamine segments. *PLoS ONE*, 2017, **12**, e0178333.
- [88] A. J. Marchut, C. K. Hall, Effects of chain length on the aggregation of model polyglutamine peptides: molecular dynamics simulations. *Proteins*, 2007, **66**, 96-109.
- [89] K. M. Ruff, S. J. Khan, R. V. Pappu, A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. *Biophys. J.* 2014, **107**, 1226-1235.
- [90] R. Kieffer. The role of gluten elasticity in the baking quality of wheat. w D. Lafiandra, S. Masci, R. D'Ovidio, ed. *The gluten proteins*. Royal Society of Chemistry, Cambridge, 2004.
- [91] F. Rasheed, W. Newson, T. Plivelic, R. Kuktaite, M. Hedenqvist, M. Gällstedt, E. Johansson. Structural architecture and solubility of native and modified gliadin and glutenin proteins, non-crystalline molecular and atomic organization. *Roy. Soc. Chem. Adv.* 2014, **4**(4), 2051–2060.
- [92] D. D. Kasarda. Contrasting molecular models for a hmw-gs. w *Proceedings of the International Meeting on Wheat Kernel Proteins Molecular and Functional Aspects*, Witerbo, 1994. Universita degli Studi della Tuscia.
- [93] N. Matsushima, C. E. Creutz, R. H. Kretsinger. Polyproline,  $\beta$ -turn helices. novel secondary structures proposed for the tandem repeats within rhodopsin, synaptophysin, synexin, gliadin, rna polymerase ii, hordein, and gluten. *Proteins*, 1990, **7**(2), 125–155.
- [94] P. Shewry, Y. Popineau, D. Lafiandra, P. S. Belton. Wheat glutenin subunits and dough elasticity, findings of the eurowheat project. *Trends Food Sci. Technol.* 2000, **11**(12), 433–441.
- [95] E. A. Bayer, J. P. Belaich, Y. Shoham ancd R. Lamed, *Annu. Rev. Microbiol.* 2004, **58**, 521-554.
- [96] B. Rózycki, M. Cieplak M. Czjzek, Large conformational fluctuations of the multi-domain Xylanase Z of *Clostridium thermocellum*. *J. Struct. Biol.* 2015, **191**, 68-75.
- [97] B. Rózycki, P.-A. Cazade, S. O'Mahony, D. Thompson, M. Cieplak, The length but not the sequence of peptide linker modules exerts the primary influence on the conformations of protein domains in cellulosome multi-enzyme complexes. *Phys. Chem. Chem. Phys.* 2017, **19**, 21414-21425.
- [98] D. Souza. *Science: A closer look at gluten*. <https://www.youtube.com/watch?v=zDEcvSc2UKA>, 2013.
- [99] Y. C. Kim, G. Hummer, Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* 2008, **375**, 1416-1433.

- [100] T. Frembgen-Kesner, C. T. Andrews, S. Li, N. A. Ngo, S. A. Shubert, A. Jain, O. J. Olayiwola, M. R. Weishaar, A. H. Elcock, Parametrization of Backbone Flexibility in a Coarse-Grained Force Field for Proteins (COFFDROP) Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of All Possible Two-Residue Peptides. *J. Chem. Theor. Comp.* 2015, **11**(5), 2341–54.
- [101] M. Cheon, I. Chang, C. K. Hall, Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins: Str. Func. Bioinf.* 2010, **78**(14), 2950–2960.
- [102] V. A. Wagoner, M. Cheon, I. Chang, C. K. Hall, Computer simulation study of amyloid fibril formation by palindromic sequences in prion peptides. *Proteins: Str. Func. Bioinf.* 2011, **79**, 2132-2145.
- [103] M. Bixon, S. Lifson, Potential functions and conformations in cycloalkanes. *Tetrahedron*, 1967, **23**, 769-784.
- [104] S. Lifson, A. Warshel, Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *J. Chem. Phys.* 1968, **49**, 5116-5129.
- [105] M. Levitt, S. Lifson, Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* 1969, **46**, 269-279.
- [106] D. De Sancho, R. B. Best, Modulation Of An IDP Binding Mechanism and Rates By Helix Propensity And Non-Native Interactions: Association Of Hif1 $\alpha$  With CBP. *Mol. BioSyst.* 2012, **8**(1), 256-267.
- [107] D. Ganguly, W. Zhang, J. Chen, Electrostatically Accelerated Encounter and Folding For Facile Recognition Of Intrinsically Disordered Proteins. *PLoS Comp. Biol.* 2013, **9**(11), e1003363.
- [108] M. Enciso, A. Rey, Improvement of Structure-Based Potentials for Protein Folding by Native and Nonnative Hydrogen Bonds. *Biophys. J.* 2011, **101**(6). 1474–82.
- [109] M. Levitt, A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 1976, **104**, 59-107.
- [110] A. Kolinski, ed. *Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies*, Springer, New York, 2010.
- [111] A. Liwo, ed. *Computational methods to study the structure and dynamics of biomolecules and biomolecular processes - from bioinformatics to molecular quantum mechanics*, Springer, Heidelberg, 2014.

- [112] V. Tozzini, J. Trylska, C. Chang, J. A. McCammon, Flap Opening Dynamics in HIV-1 Protease Explored with a Coarse-Grained Model. *J. Struct. Biol.* 2007, **157**(3), 606–15.
- [113] G. B. Brandani, S. Takada, Chromatin remodelers couple inchworm motion with twist-defect formation to slide nucleosomal DNA. *PLOS Comp. Biol.* 2018, **14**(11): e1006512.
- [114] S. Das, A. N. Amin, Y.-H. Lin, H. S. Chan, Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Phys. Chem. Chem. Phys.* 2018, **20**, 28558-28574.
- [115] M. Chwastyk, M. Jaskólski, M. Cieplak. The volume of cavities in proteins and virus capsids. *Proteins*, 2016.
- [116] M. Chwastyk, M. Jaskólski, M. Cieplak. Structure-based analysis of thermodynamic and mechanical properties of cavity-containing proteins - case study of plant pathogenesis-related proteins of class 10. *FEBS J.* 2014, **281**(1), 416–429.
- [117] M. Kröger, Shortest multiple disconnected path for the analysis of entanglements in two- and three-dimensional polymeric systems. *Comput. Phys. Commun.* 2005, **168**, 209-232.
- [118] S. Shanbhag, M. Kröger, Primitive Path Networks Generated by Annealing and Geometrical Methods: Insights into Differences. *Macromolecules*, 2007, **40**(8), 2897-2903.
- [119] N. C. Karayiannis, M. Kröger, Combined Molecular Algorithms for the Generation, Equilibration and Topological Analysis of Entangled Polymers: Methodology and Performance. *Int. J. Mol. Sci.* 2009, **10**, 5054-5089.
- [120] R. S. Hoy, K. Foteinopoulou, M. Kröger, Topological analysis of polymeric melts: Chain-length effects and fast-converging estimators for entanglement length. *Phys. Rev. E*, 2009, **80**, 031803.
- [121] Y. Ueda, H. Taketomi, N. Go, Studies on protein folding, unfolding and fluctuations by computer simulations. *Biopolymers*, 1978, **17**, 1531-1548.
- [122] I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan, J. R. Banavar, Lattice model for rapidly folding protein-like heteropolymers. *Proc. Natl. Acad. Sci. USA* 1995, **92**, 9206-9209.
- [123] N. Koga, S. Takada, Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* 2001, **313**, 171-180.
- [124] J. D. Bryngelson, P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* 1987, **84**, 7524-7528.

- [125] D. Baker, A surprising simplicity to protein folding. *Nature*, 2000, **405**, 39-42.
- [126] N. Go, Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 1983, **12**, 183-210.
- [127] H. Abe, N. Go, Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers*, 1981, **20**, 1013-1031.
- [128] A. Sali, E. Shakhnovich, M. Karplus, How does a protein fold. *Nature*, 1994, **369**, 248-251.
- [129] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* 1999, **290**, 253-266.
- [130] G. Settanni, T. X. Hoang, C. Micheletti, A. Maritan, Folding pathways of prion and doppel. *Biophys. J.* 2002, **83**, 3533-3541.
- [131] K. Wołek, Á. Gómez-Sicilia, M. Cieplak, Determination of contact maps in proteins: a combination of structural and chemical approaches. *J. Chem. Phys.* 2015, **143**, 243105.
- [132] P. Szymczak, M. Cieplak, Stretching of proteins in a uniform flow. *J. Chem. Phys.* 2006, **125**(16), 164903.
- [133] T. Veitshans, D. Klimov, D. Thirumalai, Protein folding kinetics: time scales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding Des.* 1997, **2**, 1-22.
- [134] A. B. Poma, M. Chwastyk, M. Cieplak, Polysaccharide-protein complexes in a coarse-grained model. *J. Phys. Chem. B.* 2015, **119**, 12028-12041.
- [135] M. P. Allen D. J. Tildesley, *Computer simulation of liquids*, Oxford University Press, New York, 1987.
- [136] A. Ghavani, E. van der Giessen, P. R. Onck, Coarse-grained potentials for local interactions in unfolded proteins. *J. Chem. Theor. Comp.* 2013, **9**, 432-440.
- [137] H. Wieser. Chemistry of gluten proteins. *Food Microbiol.* 2007, **24**(2), 115–119.
- [138] A. Kolinski, Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol.* 2004, **51**(2), 349-71.
- [139] A. B. Poma, M. Chwastyk, M. Cieplak, Polysaccharide-protein complexes in a coarse-grained model. *J. Phys. Chem. B.* 2015, **119**, 12028-12041.

- [140] K. Wołek, M. Cieplak, Criteria for folding in structure-based models of proteins. *J. Chem. Phys.* 2016, **144**, 185102.
- [141] R. H. Walters, R. M. Murphy, Examining Polyglutamine Peptide Length: A Connection between Collapsed Conformations and Increased Aggregation. *J. Mol. Biol.* 2009, **393**(4), 978–992.
- [142] J. I. Sułkowska, M. Cieplak, Mechanical stretching of proteins – a theoretical survey of the Protein Data Bank. *J. Phys.: Condens. Matter*, 2007, **19** 283201.
- [143] N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, I. Sillitoe, CATH: an expanded resource to predict protein function through structure and sequence. *Nucl. Acids Res.* 2016, **45**, D289-D295.
- [144] G. Settanni, T. X. Hoang, C. Micheletti, A. Maritan, Folding pathways of prion and doppel. *Biophys. J.* 2002, **83**, 3533-3541.
- [145] N. B. Hung, D.-M. Le, T. X. Hoang, Sequence dependent aggregation of peptides and fibril formation. *J. Chem. Phys.* 2017, **147**, 105102.
- [146] N.-V. Buchete, J. E. Straub, D. Thirumalai, On the development of coarse-grained protein models: Importance of relative side-chain orientations, backbone interactions. *Coarse-Graining of Condensed Phase, Biomolecular System*, ed. G. A. Voth, CRC Press, Boca Raton, 2009. rozdz. 10, 141-156.
- [147] P. Robustelli, S. Piana, D. E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA*, 2018, **115**(21), E4758–E4766.
- [148] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, A. Maritan, Geometry and symmetry prescript the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. USA*, 2004, (USA) **101**, 7960-7964.
- [149] M. Enciso, A. Rey, A refined hydrogen bond potential for flexible protein models. *J. Chem. Phys.* 2010, **132**, 235102.
- [150] D. G. Covell, R. L. Jernigan, Conformation of folded proteins in restricted spaces. *Biochem.* 1990, **29**, 3287-94.
- [151] C. Micheletti, F. Seno, J. R. Banavar, A. Maritan, Learning effective amino acid interactions through iterative stochastic techniques. *Proteins Struct. Funct. Genet.* 2001, **42**, 422-31.
- [152] M. Cieplak, N.S. Holter, A. Maritan, J. R. Banavar, Amino acid classes and the protein folding problem. *J. Chem. Phys.* 2001, **114**, 1420.

- [153] A. Korkut, W. A. Hendrickson, A Force Field For Virtual Atom Molecular Mechanics Of Proteins. *Proc. Natl. Acad. Sci. (USA)*, 2009, **106**(37), 15667-15672.
- [154] M. Qin, W. Wang, D. Thirumalai. Protein Folding Guides Disulfide Bond Formation. *Proc. Natl. Acad. Sci. (USA)*, 2015, **112**(36), 11241-11246.
- [155] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, R. V. Pappu, Net charge per residue modulates conformational ensembles of intrinsically disordered proteins, *Proc. Natl. Acad. Sci. (USA)*, 2010, **107**, 8183-8188.
- [156] P. Debye, E. Hueckel, Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen, *Phys. Zeitschrift*, 1923, **24**, 185-206.
- [157] M. Qin, W. Wang, D. Thirumalai. Protein folding guides disulfide bond formation. *Proc. Natl. Acad. Sci. USA*, 2015, **112**(36), 11241–11246.
- [158] K. Tilley, R. Benjamin, K. Bagorogoza, B. Okot-Kotber, O. Prakash, H. Kwen. Tyrosine cross-links, Molecular basis of gluten structure and function. *J. Agric. Food. Chem.* 2001, **49**(5), 2627–2632.
- [159] H. Wieser, Chemistry of gluten proteins. *Food Microbiol.* 2007, **24**, 115-119.
- [160] J. Petruska, M. J. Hartenstine, M. F. Goodman, Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J. Biol. Chem.* 1998, **273**, 5204-5210.
- [161] M. Enciso, C. Schütte, L. Delle Site. pH-Dependent Coarse-Grained Model of Peptides. *Soft Matter*, 2013, **9**(26), 6118-6127.
- [162] W. L. Jorgenson, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 1983, **79**, 926-935.
- [163] P. Palencar T. Bleha, Molecular dynamics simulations of the folding of poly(alanine) peptides. *J. Mol. Model.* 2011, **17**, 2367-23745.
- [164] P. J. Flory, Spatial Configuration Of Macromolecular Chains. *Brit. Polym. J.* 1976, **8**(1), 1-10.
- [165] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, H. Grubmueller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theor. Comp.* 2015, **11**, 5513-5524.

- [166] W. Wang, W. Ye, C. Jiang, R. Luo, H.-F. Chen, New force field on modeling intrinsically disordered proteins. *Chem. Biol. Drug Des.* 2014, **84**, 253-269.
- [167] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmueller, A. D. MacKerell Jr, CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 2017, **14**, 71-73.
- [168] S. Bhattacharya, L. Xu, D. Thompson, Long-range Regulation of Partially Folded Amyloidogenic Peptides. *Sci. Rep.* 2020, **10**, 7597.
- [169] A. Starzyk, M. Cieplak. Denaturation of proteins near polar surfaces. *J. Chem. Phys.* 2011, **135**, 235103.
- [170] B. Różycki, T. Weikl, R. Lipowsky, Adhesion of Membranes with Active Stickers. *Phys. Rev. Lett.* 2006, **96**(4), 048101.
- [171] B. Różycki, T. Weikl, R. Lipowsky, Adhesion of membranes via switchable molecules. *Phys. Rev. E*, 2006, **73**(6), 061908.
- [172] B. Różycki, T. Weikl, R. Lipowsky, Stochastic resonance for adhesion of membranes with active stickers. *Eur. Phys. J. E*, 2007, **22**, 97–106.
- [173] V. Tozzini, Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophys.* 2010, **43**(3), 333-371.
- [174] A. Liwo, C. Czaplewski, J. Pillardy, H. A. Scheraga, Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* 2001, **115**(5), 2323-2347.
- [175] D. Alemani, F. Collu, M. Cascella, M. Dal Peraro, A nonradial coarse-grained potential for proteins produces naturally stable secondary structure elements. *J. Chem. Theory Comput.* 2010, **6**, 315-324.
- [176] W. C. Swope, D. M. Ferguson, Alternative expressions for energies and forces due to angle bending and torsional energy. *J. Comp. Chem.* 1992, **13**(5), 585-594.
- [177] <https://blob.pureandapplied.com.au/sigmoid-a-post-about-an-algebraic-function-im-having-too-much-fun/> (dostęp 15.01.2020).
- [178] S.-Y. Sheu, D.-Y. Yang, H. L. Selzle, E. W. Schlag, Energetics of hydrogen bonds in peptides. *Proc. Natl. Acad. Sci. USA*, 2003, **100**(22), 12683-12687.

- [179] M. R. Betancourt, S. J. Omovie, Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.* 2009, **130**(19), 195103.
- [180] E. Mylonas, A. Hascher, P. Bernado, M. Blackledge, E. Mandelkow, D. I. Svergun, Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochem.* 2008, **47**(39), 10345–10353.
- [181] C. C. Kung, M. T. Naik, S. Wang, H. Shih, C. Chang, L. Lin, C. Chen, C. Ma, C. Chang, T. Huang, Structural analysis of poly-sumo chain recognition by the rnf4-sims domain. *Biochem. J.* 2014, **462**, 53-65.
- [182] M. Chukhlieb, A. Raasakka, S. Ruskamo, P. Kursula, The N-terminal cytoplasmic domain of neuregulin 1 type III is intrinsically disordered. *Amino Acids*, 2015, **47**(8), 1567-1577.
- [183] K. Moncoq, I. Broutin, V. Larue, D. Perdereau, K. Cailliau, E. Browaeys-Poly, A.-F. Burnol, A. Ducruix, The pir domain of grb14 is an intrinsically unstructured protein: implication in insulin signaling. *FEBS Lett.* 2003, **554**(3), 240-246.
- [184] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, A. R. Fersht, Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 240-246.
- [185] T. N. Cordeiro, F. Herranz-Trillo, A. N. Urbanek, A. N. Esta na, J. Cortés, N. Sibille, P. N. Bernadó, Structural Characterization of Highly Flexible Proteins by Small-Angle Scattering. *Adv. Exper. Med. Biol.* 2017, **1009**, 107-129.
- [186] A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. G. Richardson, R. V. Pappu, CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* 2017, **112**(1), 16-21.
- [187] R. D. Requião, L. Fernandes, H. de Souza, S. Rossetto, T. Domitrovic, F. L. Palhano, Protein charge distribution in proteomes and its impact on translation. *PLOS comp. biol.* 2017, **13**(5), e1005549.
- [188] J. Kyte, R. Doolittle, A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157**, 105-132.
- [189] G. Dignon, W. Zheng, Y. Kim, R. Best, J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Comp. Biol.* 2018, **14**(1), p.e1005941. [Tabela S2]

- [190] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. De Groot, H. Grubmüller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513-5524.
- [191] I. Teraoka, Chapter 1. *Models of Polymer Chains*; John Wiley & Sons, 2002; 1-67.
- [192] W. Janke, *Monte Carlo Methods in Classical Statistical Physics*; Springer Berlin Heidelberg, 2008.
- [193] L. X. Peterson, A. Roy, C. Christoffer, G. Terashi, D. Kihara, Modeling disordered protein interactions from biophysical principles. *PLoS Comp. Biol.* 2017, **13**, e1005485.
- [194] C. L. Day, C. Smits, F. C. Fan, E. F. Lee, W. D. Fairlie, M. G. Hinds, Structure of the BH3 Domains from the p53-Inducible BH3-Only Proteins Noxa and Puma in Complex with Mcl-1, *J. Mol. Biol.* 2008, **380**(5), 958-971.
- [195] M. Chwastyk, A. Poma Bernaola, M. Cieplak, Statistical radii associated with amino acids to determine the contact map: Fixing the structure of a type, cohesin domain in the *Clostridium thermocellum* cellulosome. *Phys. Biol.* 2015, **12**, 046002.
- [196] P. R. Shewry, N. G. Halford, P. S. Belton, A. S. Tatham, The structure and properties of gluten: an elastic protein from wheat grain. *Phil. Trans. Roy. Soc. B*, 2002, **357**, 133-142.
- [197] L. M. Pietrek, L. S. Stelzl, G. Hummer, Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations. *J. Chem. Theory Comput.* 2020, **16**(1), 725-737.
- [198] B. Różycki, M. Cieplak, Stiffness of the C-terminal disordered linker affects the geometry of the active site in endoglucanase Cel8A. *Mol. BioSyst.* 2016, **12**, 3589-3599.
- [199] Sikora, J. I. Sułkowska, M. Cieplak, Mechanical strength of 17 134 model proteins and cysteine slipknots, *PLoS Comput. Biol.* 2009, **5**, e1000547.
- [200] The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucl. Acids Res.* 2017, **45**, D158-D169.
- [201] L. Hong, J. Lei, Scaling Law for the Radius of Gyration of Proteins and Its Dependence on Hydrophobicity. *J. Polym. Sci. Pol. Phys.* 2009, **47**, 207 - 214.
- [202] R. Wetzel, Physical Chemistry of Polyglutamine: Intriguing Tales of a Monotonous Sequence, *J. Mol. Biol.* 2012, **421**(4–5), 466-490.

- [203] J. P. Bernacki, R. M. Murphy, Length-dependent aggregation of uninterrupted polyalanine peptides. *Biochemistry*, 2011, **50**(43), 9200-9211.
- [204] J. P. Hensen, I. R. McDonald, Theory of simple liquids, Academic Press, New York (1973).
- [205] H.-X. Zhou, V. Nguemaha, K. Mazarakos, S. Qin. Why do disordered and structured proteins behave differently in phase separation. *Trends Biochem. Sci.* 2018, **43**, 499-516.
- [206] J. McCarty, K. T. Delaney, A. P. O. Danielsen, G. H. Fredrickson, J.-E. Shea, Complete phase diagram for liquid-liquid phase separation of intrinsically disordered proteins. *Phys. Chem. Lett.* 2019, **10**, 1644-1652.
- [207] Y.-H. Lin, J. P. Brady, J. D. Forman-Kay, H. S. Chan, Charge pattern matching as a ‘fuzzy’ mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New J. Phys.* 2017, **19**, 115003.
- [208] S. Boyko, X. Qi, T.-H. Chen, K. Surewicz, W. Surewicz, Liquid-liquid phase separation of tau protein: The crucial role of electrostatic interactions. *J. Biol. Chem.* 2019, **294**, 11054-11059.
- [209] Y. Zhou, S. Liu, G. Liu, A. Oztuerk, G. G. Hicks, ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulations. *PLoS Genet.* 2013, **9**, E1003895.
- [210] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, J. Mittal, Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. USA*, 2018, **115**, 9929-9934.
- [211] D. Sharma, L. M. Shinchuk, H. Inouye, R. Wetzel, D. A. Kirschner, Polyglutamine homopolymers having 8-45 residues form slablike  $\beta$ -crystallite assemblies. *Proteins*, 2005, **61**, 398-411.
- [212] E. Ibarra-García-Padilla, C. G. Malanche-Flores, F. J. Poveda-Cuevas, The hobbyhorse of magnetic systems: the Ising model. *Eur. J. Phys.* 2016, **37**, 065103.
- [213] J. Koplik, J. R. Banavar, Continuum deductions from molecular hydrodynamics. *Annu. Rev. Fluid Mech.* 1995, **27**, 257-292.
- [214] P. Shewry, N. Halford, P. S. Belton, A. Tatham. The structure and properties of gluten, an elastic protein from wheat grain. *Phil. Trans Roy. Soc. B, Biol. Sci.* 2002, **357**(1418), 133–142.
- [215] F. Anjum, M. Khan, A. Din, M. Saeed, I. Pasha, M. Arshad. Wheat gluten, High molecular weight glutenin subunits, structure, genetics, and relation to dough elasticity. *J. Food Sci.* 2007, **72**(3), 56–63.

- [216] J. Ahmed, H. Ramaswamy, V. Raghavan. Dynamic viscoelastic, calorimetric and dielectric characteristics of wheat protein isolates. *J. Cereal Sci.* 2008, **47**(3), 417–428.
- [217] E. Blanch, D. Kasarda, L. Hecht, K. Nielsen, L. Barron. New insight into the solution structures of wheat gluten proteins from Raman optical activity. *Biochemistry*, 2003, **42**(19), 5665–5673.
- [218] P. S. Belton. The molecular basis of dough rheology. *Bread making, Improving quality*, 2003, **13**, 273–287.
- [219] H. Singh, F. MacRitchie. Application of polymer science to properties of gluten. *J. Cereal Sci.* 2001, **33**(3), 231–243.
- [220] P. R. Shewry, N. G. Halford. Cereal seed storage proteins, structures, properties and role in grain utilization. *J. Exp. Bot.* 2002, **53**(370), 947–958.
- [221] Y. Wu, J. Messing. Proteome balancing of the maize seed for higher nutritional value. *Front. Plant Sci.* 2014, **5**.
- [222] Y. Wu, W. Wang, J. Messing. Balancing of sulfur storage in maize seed. *BMC Plant Biol.* 2012, **12**(1), 77.
- [223] C. Y. Tsai. Genetics of storage protein in maize. w Jules Janick, editor, *Plant Breeding Reviews, Volume 1*, pp. 103–138. Springer US, Boston, MA, 1983.
- [224] P. Chen, Z. Shen, L. Ming, Y. Li, W. Dan, G. Lou, B. Peng, B. Wu, Y. Li, and D. et al. Zhao. Genetic basis of variation in rice seed storage protein (albumin, globulin, prolamin, and glutelin) content revealed by genome-wide association analysis. *Front. Plant Sci.* 2018, **9**, 612.
- [225] C. Jiang, Z. Cheng, C. Zhang, T. Yu, Q. Zhong, J. Q. Shen, X. Huang. Proteomic analysis of seed storage proteins in wild rice species of the oryza genus. *Proteome Sci.* 2014, **12**(1).
- [226] D. G. Muench, T. W. Okita. The storage proteins of rice and oat. w B. A. Larkins, I. K. Vasil, ed. *Cellular and Molecular Biology of Plant Seed Development*, pp. 289–330. Springer Netherlands, Dordrecht, 1997.
- [227] The UniProt Consortium. Uniprot, a worldwide hub of protein knowledge. *Nucl. Acids Res.* 2018, **47**(D1), D506–D515.
- [228] R. Cazalis. Homology modeling and molecular dynamics simulations of the n-terminal domain of wheat high molecular weight glutenin subunit 10. *Protein Sci.* 2003, **12**(1), 34–43.

- [229] J. I. Sułkowska, M. Cieplak. Selection of optimal variants of go-like models of proteins through studies of stretching. *Biophys. J.* 2008, **95**, 317491.
- [230] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang. The i-tasser suite, Protein structure and function prediction. *Nat. Methods*, 2015, **12**, 7–8.
- [231] J. Yang, Y. Zhang. I-tasser server, new development for protein structure and function predictions. *Nucl. Acids Res.* 2015, **43**, W174–W181.
- [232] H. Belitz, R. Kieffer, W. Seilmeier, H. Wie. Structure and function of gluten proteins. *Cereal Chem.* 1986, **63**, 336–341.
- [233] K. Kremer, G. S. Grest. Dynamics of entangled linear polymer melts, A molecular-dynamics simulation. *J. Chem. Phys.* 1990, **92**(8), 5057.
- [234] Engineering Toolbox. Densities of some common materials. [http://www.engineeringtoolbox.com/density-materials-d\\_1652.html](http://www.engineeringtoolbox.com/density-materials-d_1652.html).
- [235] B. Lagrain, E. Wilderjans, C. Glorieux i in. Importance of Gluten and Starch for Structural and Textural Properties of Crumb from Fresh and Stored Bread. *Food Biophys.* 2012, **7**, 173–181.
- [236] J. Horabik, J. Wiącek, P. Parafiniuk, M. Stasiak, M. Bańda, R. Kobyłka, M. Molenda, Discrete Element Method Modelling of the Diametral Compression of Starch Agglomerates. *Materials* 2020, **13**, 932.
- [237] M. N. Charalambides, L. Wanigasooriya, G. J. Williams i in. Biaxial deformation of dough using the bubble inflation technique. I. Experimental. *Rheol Acta* 2002, **41**, 532–540.
- [238] D. N. Abang Zaidel, N. L. Chin, Y. A. Yusof. A review on rheological properties and measurements of dough and gluten. *J. Appl. Sci.* 2010, **10**, 2478–2490.
- [239] N. Wellner, E. Mills, G. Brownsey, R. Wilson, N. Brown, J. Freeman, N. Halford, P. Shewry, P. S. Belton. Changes in protein secondary structure during gluten deformation studied by dynamic Fourier transform infrared spectroscopy. *Biomacromolecules*, 2005, **6**(1), 255–261.
- [240] J. Kofinger, G. Hummer, Atomic-resolution structural information from scattering experiments on macromolecules in solution, *Phys. Rev. E*, 2013, **87**, 052712.
- [241] E. Fagerberg, S. Lenton, M. Skepo, Evaluating Models of Varying Complexity of Crowded Intrinsically Disordered Protein Solutions Against SAXS, *J. Chem. Theory Comput.* 2019, **15**(12), 6968–6983.

- [242] S. Yang, S. Park, L. Makowski, B. Roux, A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes, *Biophys. J.* 2009, **96**, 4449–4463.
- [243] J. J. Aklonis, W. J. MacKnight. *Introduction to Polymer Viscoelasticity* Hoboken. Wiley-Interscience, NJ, 2005.
- [244] G. S. Grest, M. Putz, R. Everaers, K. Kremer. Stress-strain relation of entangled polymer networks. *J. Non-Cryst. Solids*, 2000, **274**(1-3), 139–146.
- [245] B. Dobraszczyk, M. Morgenstern. Rheology and the breadmaking process. *J. Cereal Sci.* 2003, **38**(3), 229–245.
- [246] C. Létang, M. Piau, C. Verdier. Characterization of wheat flour–water doughs. part i, Rheometry and microstructure. *J. Food Eng.* 1999, **41**, 121–132.
- [247] B. S. Khatkar. Dynamic rheological properties and bread-making qualities of wheat gluten, effects of urea and dithiothreitol. *J. Sci. Food Agric.* 2005, **85**(2), 337–341.
- [248] Y. Song, Q. Zheng, Q. Zhang. Rheological and mechanical properties of bioplastics based on gluten- and glutenin-rich fractions. *J. Cereal Sci.* 2009, **50**(3), 376–380. Special Section, Enzymes in Grain Processing.
- [249] H. Sánchez, C. Osella, M. A. Torre. Optimization of gluten-free bread prepared from cornstarch, rice flour, and cassava starch. *J. Food Sci.* 2006, **67**, 416–419.
- [250] M. Cieplak, M. O. Robbins. Nanoindentation of virus capsids in a molecular model. *J. Chem. Phys.* 2010, **132**(1), 015101.
- [251] O. Parchment, P. Shewry, A. Tatham, D. Osguthorpe. Molecular modeling of unusual spiral structure in elastomeric wheat seed protein. *Cereal Chem.* 2001, **78**(6), 658–662.
- [252] T. Noel, R. Parker, S. Ring, A. Tatham. The glass-transition behaviour of wheat gluten proteins. *Int. J. Biol. Macromol.* 1995, **17**(2), 81–85.
- [253] P. Koehler, H. Wieser, K. Konitzer. *Celiac disease and gluten*. Elsevier Science, Burlington, 2014.
- [254] M. Dziuba, J. Dziuba, A. Iwaniak. Bioinformatics-aided characteristics of the structural motifs of selected potentially celiac-toxic proteins of cereals and leguminous plants. *Polish J. Food Nutr. Sci.* 2007, **57**(4), 405–414.
- [255] S. McAdam. Getting to grips with gluten. *Gut*, 2000, **47**(6), 743–745.