Piotr Rotkiewicz

MODELOWANIE MOLEKULARNE BIAŁEK W OPARCIU O PODOBIEŃSTWA SEKWENCYJNE I ANALOGIE STRUKTURALNE

Praca doktorska wykonana
w latach 1998-2002
w Pracowni Teorii Biopolimerów
Wydziału Chemii
Uniwersytetu Warszawskiego

Promotor pracy:
prof. dr hab. Andrzej Koliński

Warszawa 2002

Panu profesorowi
Andrzejowi Kolińskiemu
– najlepszemu promotorowi –
składam serdeczne podziękowania
za inspirujący temat badań,
pomoc i opiekę
podczas pracy nad doktoratem.

Dziękuję profesorowi
Jeffreyowi Skolnickowi
za wiele cennych dyskusji
podczas odbywania staży naukowych
w Danforth Plant Science Center
w St.Louis.

Kolegom
z Pracowni Teorii Biopolimerów
bardzo dziękuję za miłą atmosferę
podczas realizowania pracy doktorskiej.

Pracę dedykuję
Rodzicom.

Spis treści

1 Wstęp

Problem przewidywania struktury przestrzennej białek na podstawie sekwencji aminokwasów jest jednym z najciekawszych zagadnień i największych wyzwań współczesnej biochemii. Jest to problem badawczy bardzo interesujący zarówno ze względów poznawczych, jak i praktycznych. Znajomość trójwymiarowej struktury białek pozwala na zbadanie i wykorzystanie ich biologicznych funkcji i ma podstawowe znaczenie dla medycyny i biotechnologii. Poznanie struktury i funkcji wszystkich białek pochodzących z całego genomu pozwala na lepsze zrozumienie procesów przebiegających w żywych komórkach i może pomóc w odkryciu nowych szlaków metabolicznych. Dzięki temu możliwe jest projektowanie nowych leków. Znajomość funkcji odpowiednich białek roślinnych umożliwia ich modyfikację w celu zwiększenia wartości odżywczych roślin i uzyskania ich odporności na choroby. Badania takie są domeną genomiki strukturalnej, proteomiki i bioinformatyki.

Strukturę przestrzenną białek można badać przy pomocy metod doświadczalnych i metod teoretycznych. Dostępne obecnie techniki eksperymentalne (przede wszystkim dyfraktometria rentgenowska i spektroskopia jądrowego rezonansu magnetycznego) pozwalają na zbadanie struktur stosunkowo niewielkiej części nowo poznawanych białek, głównie ze względu na swoją czasochłonność i wysoki koszt badań [1]. Dlatego bardzo duże znaczenie mają teoretyczne metody przewidywania struktury trzeciorzędowej. Znaczna część tych metod opiera się na wykorzystaniu podobieństwa nowo poznawanych sekwencji do sekwencji białek o wcześniej poznanej strukturze. Do zbudowania modelu konieczna jest wówczas znajomość struktury bardzo podobnego (spokrewnionego ewolucyjnie) białka–wzorca. Takie podobne struktury (białka homologiczne) znane są tylko dla około 40% nowo sekwencjonowanych białek. Wiadomo jednak, że w znanych strukturach białek mogą występować krótsze fragmenty o wysokim stopniu podobieństwa sekwencji, albo fragmenty podobne strukturalnie, ale o niewykrywalnym (przy pomocy standardowych technik) podobieństwie sekwencji. Rozszerzenie zakresu wykorzystania informacji ewolucyjnych w różnych aspektach modelowania białek jest tematem tej pracy.

1.1 Cele pracy

Pierwszym celem pracy było wykorzystanie podobieństw sekwencyjnych do wyprowadzania potencjałów statystycznych bliskiego i dalekiego zasięgu, które zostały następnie użyte w metodzie *ab initio* przewidywania struktury białek z wykorzystaniem uproszczonego modelu białka i dynamiki Monte Carlo.

Drugim celem pracy, ściśle związanym z pierwszym, było zaproponowanie i zbadanie kilku zastosowań wykorzystania informacji ewolucyjnych w modelowaniu białek: odbudowywania brakujących fragmentów struktur białek, poprawiania modeli białek zbudowanych przy pomocy innych metod modelowania, modelowania struktur białek na podstawie niewielkiej liczby więzów.

W ramach pracy stworzono kilka pomocniczych metod i programów służących do analizy i wizualizacji modeli białek oraz do przygotowywania bazy danych struktur białek.

Integralną częścią pracy są publikacje stanowiące opis i ilustrację zastosowań przedstawionych metod (załączone w rozdziale 8). Rysunki struktur białek zamieszczone w tej pracy wykonano używając programu Biodesigner [2]. Pracę przygotowano i złożono do druku wykorzystując pakiet LaTeX.[1]

[1] Archiwum programu TeX: `http://www.ctan.org`

2 Część literaturowa

2.1 Wprowadzenie

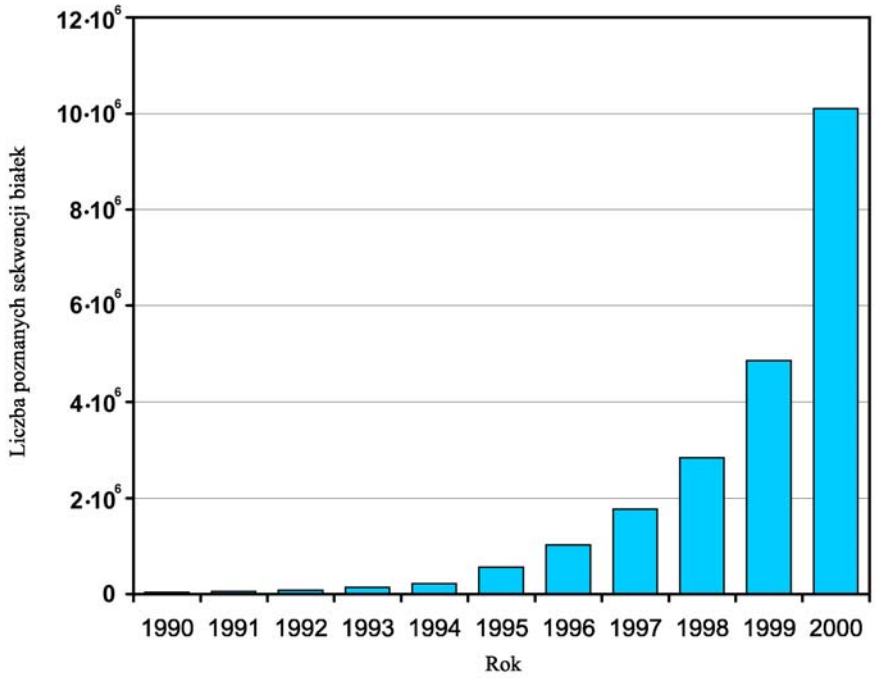
W ostatnich latach obserwujemy lawinowy wzrost liczby znanych sekwencji białek [3]. Zaawansowane projekty sekwencjonowania genomów różnych organizmów, również genomu ludzkiego [4], dostarczają nowych sekwencji białek w bardzo szybkim tempie. Liczba znanych sekwencji białek wzrosła wykładniczo w ciągu ostatnich kilku lat (rysunek 1) i obecnie wynosi ponad dziesięć milionów. Natomiast liczba znanych struktur białek sięga kilkunastu tysięcy (rysunek 2). Znajomość struktury białka jest kluczowa dla poznania i zrozumienia jego funkcji, dlatego niezwykle ważne w biochemii i biologii molekularnej są metody określania struktury białek.

2.2 Eksperymentalne metody określania struktury przestrzennej białek

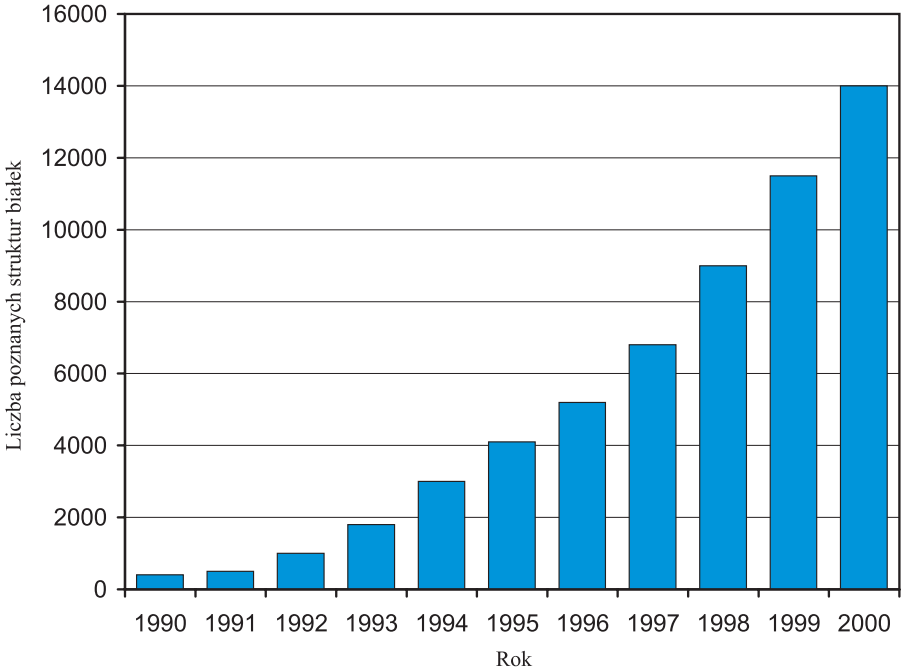
Białka należą do cząsteczek chemicznych o bardzo skomplikowanej strukturze przestrzennej [5, 6]. Jej zbadanie wymaga zastosowania zaawansowanych technik eksperymentalnych. Zwykle stosuje się do tego celu jedną z dwóch metod: dyfraktometrię rentgenowską lub spektroskopię jądrowego rezonansu magnetycznego.

Metodą wykorzystywaną najczęściej jest dyfraktometria rentgenowska (przy jej pomocy rozwiązano około 80% znanych struktur białek [7]). Struktury białek o najwyższej jakości otrzymuje się wykorzystując źródła promieniowania X o bardzo dużej intensywności (synchrotrony). Rozdzielczość eksperymentalna takich struktur (średni błąd kwadratowy oznaczenia położeń wszystkich atomów) wynosi około 1.5 Å.

Spektroskopia NMR umożliwia badanie struktur białek o wielkości do 300 aminokwasów, rozpuszczalnych w wodzie [5, 8]. Zwykle wymaga znakowania białek izotopami węgla ^{13}C i azotu ^{15}N, a w przypadku większych białek również deuteru ^{2}H. W wyniku pomiaru NMR otrzymuje się zbiór więzów (odległości i kątów) pomiędzy określonymi atomami cząsteczki białka, które umożliwiają odtworzenie struktury przestrzennej.



Rysunek 1: Wzrost liczby sekwencji białek poznanych w ciągu ostatnich dziesięciu lat. Źródło: *GenBank*, NCBI, `http://www.ncbi.nlm.nih.gov`



Rysunek 2: Wzrost liczby struktur białek poznanych w ciągu ostatnich dziesięciu lat. Źródło: PDB, `http://www.rcsb.org`

Kolejną metodą eksperymentalną używaną do określania struktury przestrzennej białek jest mikroskopia elektronowa. Jest ona stosowana przede wszystkim do określania struktury białek trudnych do wykrystalizowania, na przykład białek membranowych [6] lub białek otoczek wirusów [9]. Niestety, struktury otrzymane przy pomocy mikroskopii elektronowej są znacznie mniej dokładne, niż struktury otrzymane poprzednio opisanymi sposobami. Inne metody eksperymentalne to m.in. spektroskopia fluorescencyjna i metody biochemiczne, na przykład metoda mutacji cystein, która dostarcza informacji o położeniu mostków dwusiarczkowych [10].

Podczas prób zbadania przestrzennej struktury białek metodami eksperymentalnymi napotyka się wiele problemów. Konieczne jest otrzymanie czystego białka w ilości pozwalającej na przeprowadzenie eksperymentu. Zwykle dokonuje się w tym celu ekspresji genu kodującego białko, a następnie namnaża się kultury bakterii aż do momentu otrzymania białka w odpowiedniej ilości. Już ten początkowy etap jest bardzo trudny, na przykład wymaga bardzo starannego dobrania warunków hodowli bakterii. Kolejnym etapem jest oczyszczenie i zatężenie roztworu białka do stopnia pozwalającego na wykonanie eksperymentu NMR lub – w przypadku dyfraktometrii rentgenowskiej – na otrzymanie kryształu. Krystalizacja białka nastręcza wiele trudności (wymagane jest otrzymanie stosunkowo dużego monokryształu o wysokiej jakości). Jeżeli uda się otrzymać kryształ lub odpowiednio stężony roztwór białka i pomyślnie przeprowadzić eksperyment, kolejne problemy stwarza konieczność analizy zebranych danych w celu zbudowania modelu trójwymiarowej struktury białka. Nierzadko dane eksperymentalne są na tyle zniekształcone lub niekompletne, że zbudowanie modelu białka na ich podstawie jest zadaniem bardzo trudnym.

W ostatnich latach intensywnie rozwija się gałąź nauki zwana genomiką strukturalną [11, 12]. Przedmiotem jej badań jest rozwiązywanie struktur białek metodami eksperymentalnymi na skalę całych genomów. Jednak dysproporcja pomiędzy liczbą znanych sekwencji białek i liczbą znanych struktur wciąż się powiększa. Wobec tego rośnie zapotrzebowanie na wykorzystanie teoretycznych metod przewidywania struktury białek [13].

2.3 Teoretyczne metody przewidywania struktury przestrzennej białek

Metody teoretyczne przewidywania struktury białek wygodnie jest podzielić na trzy grupy: metody wykorzystujące podobieństwo sekwencji, metody wykorzystujące kompatybilność sekwencji i struktury (metoda przewlekania sekwencji, *threading*) i metody bezpośredniego przewidywania struktury na podstawie sekwencji (metody klasy *ab initio*).

2.4 Metody wykorzystujące podobieństwo sekwencji białek

Około 30% wszystkich poznawanych sekwencji białek wykazuje wysoki stopień podobieństwa do sekwencji białek o już znanej strukturze i funkcji (przez wysoki stopień podobieństwa rozumiemy ponad 30% stopień identyczności ich sekwencji) [11, 12, 14]. Gdy sekwencje dwóch białek są do siebie podobne w tak wysokim stopniu, można założyć, że są one ze sobą spokrewnione, a ich struktury również są do siebie podobne [15].[2] Wynika to z faktu, że struktury białek są podczas ewolucji zachowywane lepiej, niż sekwencje. Fakt ten umożliwia wykorzystanie białka o znanej strukturze jako wzorca (*template structure*) służącego do zbudowania modelu białka o strukturze nieznanej (*target structure*). Podejście to określa się mianem modelowania homologicznego (*homology modeling*) lub modelowania porównawczego (*comparative modeling*). Pojęcie homologii oznacza, że dwa różne białka są ze sobą spokrewnione w sensie ewolucyjnym, to znaczy, że pochodzą od wspólnego przodka [6, 17, 18].

Liczba typów struktur (*folds*), jakie mogą przyjmować białka, jest ograniczona [19]. Według różnych oszacowań wynosi ona od 1000 do 2600 [20, 21]. Przypuszcza się, że w ciągu najbliższych dziesięciu lat większość rodzin strukturalnych białek zostanie poznana, co spowoduje dodatkowy wzrost znaczenia metody modelowania homologicznego [22].

[2] Jest interesujące, że ta własność dotyczy przede wszystkim białek występujących w żywych organizmach. Wyprodukowano syntetyczne białka o stopniu identyczności sekwencji równym 60%, a w istotny sposób różniące się strukturą [16].

Większość współcześnie używanych metod modelowania homologicznego struktur białek składa się z następujących etapów:

1. Identyfikacja homologicznego białka–wzorca. Wykorzystuje się w tym celu metody sekwencyjne lub metodę przewlekania sekwencji (opisaną w rozdziale 2.5).

2. Zbudowanie dopasowania sekwencji modelowanego białka do struktury wzorca. Zwykle używa się do tego celu algorytmu programowania dynamicznego (opisanego w rozdziale 2.4.2).

3. Zbudowanie modelu białka (rozdział 2.4.7).

4. Sprawdzenie poprawności zbudowanego modelu – na przykład przez obliczenie energii modelu za pomocą metod mechaniki molekularnej.

2.4.1 Porównywanie sekwencji dwóch białek

Aby porównać sekwencje dwóch białek, konieczne jest wprowadzenie miary ich podobieństwa. Najprostszą miarą podobieństwa jest stopień identyczności obu sekwencji – stosunek liczby identycznych aminokwasów na odpowiadających sobie pozycjach do długości sekwencji:

```
M T Y K L I V G A G C P T I A A K
- + + - - + + + + + + + - - - - +
C T Y L V I V G A G C P A V G R K
```

W powyższym przykładzie stopień identyczności wynosi $10/17 = 58.8\%$ (10 identycznych aminokwasów w sekwencjach o długości 17 aminokwasów).

Zazwyczaj porównywane sekwencje białek różnią się długością. W takim przypadku konieczne jest wprowadzenie przerw (*gaps*) w jednej lub w obu porównywanych sekwencjach. Na przykład w celu porównania poniższych sekwencji:

```
      G G G V V V V V
G G G A A A A V V V V V
```

należy do pierwszej sekwencji wprowadzić przerwę o długości czterech aminokwasów:

```
G G G - - - - V V V V V
+ + + - - - - + + + + +
G G G A A A A V V V V V
```

Stopień identyczności wynosi w tym przypadku 100% (zwykle porównuje się liczbę identycznych aminokwasów z długością krótszej sekwencji). Taki układ sekwencji białek z jednoznacznie przyporządkowanymi pozycjami aminokwasów nosi nazwę dopasowania sekwencji (*sequence alignment*).[3]

Wprowadzenie przerwy do jednej z sekwencji odpowiada obecności motywu strukturalnego nie występującego w jednym z porównywanych białek. Może to być pętla, która pojawiła się w przebiegu zmian ewolucyjnych, albo fragment struktury drugorzędowej, który zmienił długość w wyniku ewolucji. Można wyobrazić sobie następujący, ekstremalny przykład:

```
    A A A A A
A G A G A G A G A
```

Po wprowadzeniu przerw w pierwszej sekwencji otrzymujemy:

```
A - A - A - A - A
+ - + - + - + - +
A G A G A G A G A
```

Mimo stopnia identyczności wynoszącego 100%, trudno w tym przypadku oczekiwać biologicznego sensu znalezionego dopasowania! Dlatego z wprowadzaniem przerw w sekwencjach dopasowywanych białek związany jest parametr zwany kosztem wprowa-

[3] Od pewnego czasu w polskiej literaturze naukowej lansowany jest termin „uliniowienie" jako tłumaczenie angielskiego słowa *alignment*. W tej pracy autor (zgodnie z pracami m.in. Jaroszewskiego [23]) postanowił używać terminu „dopasowanie", który jest bardziej obszerny znaczeniowo (obejmuje również dopasowanie struktur).

dzenia przerwy (*gap penalty*). Zapobiega on tworzeniu nonsensownych (z biologicznego punktu widzenia) dopasowań.

Miara dopasowania s (*alignment score*) dwóch sekwencji jest sumą podobieństw odpowiadających sobie par aminokwasów oraz kosztów wprowadzenia przerw w dopasowaniu:

$$s = \sum_{i=1}^{N} f(S_1(i), S_2(i)) + \sum_{i=1}^{M} (G_O + G_E \cdot L(i)) \tag{1}$$

gdzie N jest długością dopasowania, f jest funkcją oceniającą podobieństwo dwóch aminokwasów, S_1 i S_2 są dopasowywanymi sekwencjami ($(S_1(i)$ jest typem aminokwasu występującym na i-tej pozycji dopasowania), M jest liczbą przerw w dopasowaniu, G_O jest kosztem wprowadzenia przerwy (*gap opening penalty*), G_E jest kosztem wydłużenia przerwy (*gap elongation penalty*), $L(i)$ jest długością i-tej przerwy. W przypadku wykorzystania identyczności aminokwasów jako miary podobieństwa funkcja f ma postać:

$$f(A, B) = \begin{cases} 1 & \text{gdy } A = B, \\ 0 & \text{gdy } A \neq B. \end{cases} \tag{2}$$

W ogólnym przypadku f może mieć postać bardziej złożoną.

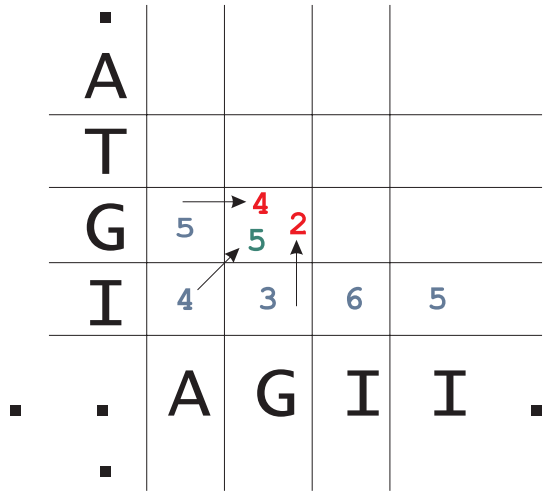
2.4.2 Algorytm programowania dynamicznego

Do obliczania dopasowania sekwencji dwóch białek używa się algorytmu programowania dynamicznego (*dynamic programming*). Gwarantuje on otrzymanie optymalnego dopasowania [24]. Algorytm programowania dynamicznego wymaga zbudowania macierzy o wymiarach $N \times M$, gdzie N, M są długościami sekwencji (każdej komórce macierzy odpowiada para aminokwasów z obu sekwencji). Macierz tę wypełnia się startując z jej lewego dolnego rogu, modyfikując zawartość komórek według prostych reguł:

- ruch wzdłuż przekątnej macierzy – do komórki dodaje się wartość miary podobieństwa danej pary aminokwasów

- ruch w poziomie – do komórki dodaje się wartość kosztu wprowadzenia przerwy w pierwszej sekwencji
- ruch w pionie – do komórki dodaje się wartość kosztu wprowadzenia przerwy w drugiej sekwencji

Wybiera się ten ruch, dla którego wartość komórki osiąga wartość maksymalną (rysunek 3). Po zbudowaniu takiej macierzy, w jej górnym prawym rogu znajduje się wartość optymalnej miary dopasowania. Następnie, analizując ruchy wykonane podczas budowania macierzy (starując z górnego prawego rogu), można zbudować dopasowanie obu sekwencji.



Rysunek 3: Elementarny ruch podczas budowania macierzy w algorytmie programowania dynamicznego. Należy obliczyć wartość środkowej komórki macierzy. Wartości zaznaczone kolorem niebieskim są znane przed wykonaniem ruchu. Ruch w poziomie: dodajemy wartość kosztu wprowadzenia przerwy (równą -1), wynik: $5 - 1 = 4$. Ruch w pionie: podobnie, wynik: $3 - 1 = 2$. Ruch wzdłuż przekątnej: dodajemy wartość funkcji podobieństwa dla pary G-G (równą 1), wynik: $4 + 1 = 5$. Wybieramy najbardziej korzystny ruch, to znaczy wzdłuż przekątnej. Następnie przesuwamy się do kolejnej komórki (w prawo).

Algorytm *dynamic programming* ma złożoność obliczeniową rzędu $O(N^2 log(N))$.[4] Dla niektórych przypadków można zredukować tę złożoność do $O(N^2)$ [26].

Istnieje wiele modyfikacji algorytmu programowania dynamicznego. Najważniejsze z nich sprowadzają się do odmiennego traktowania przerw znajdujących się na początku i na końcu dopasowania. Jeżeli koszty wprowadzenia takich przerw są równe 0, algorytm tworzy lokalne dopasowanie sekwencji (*local sequence alignment*) [27], pozwalające na przykład na dopasowanie pojedynczej domeny białka do sekwencji białka wielodomenowego. Miara lokalnego dopasowania (*Smith-Waterman score*, SWS) jest powszechnie przyjętą w biologii miarą określania podobieństwa dwóch sekwencji. Jeżeli krańcowe przerwy traktujemy w taki sam sposób jak pozostałe, algorytm generuje dopasowanie globalne (*global sequence alignment*) [24]. Stwierdzono, że podczas przeszukiwania bazy danych w celu znalezienia sekwencji białek homologicznych, dopasowanie globalne jest miarą bardziej czułą niż dopasowanie lokalne [28].

2.4.3 Macierze podstawień

Stopień identyczności jest najprostszą miarą podobieństwa aminokwasów. Dużo bardziej efektywna miara wykorzystuje chemiczne i fizyczne podobieństwo aminokwasów. Na przykład leucyna i izoleucyna są do siebie podobne – mają niepolarny charakter i zbliżoną wielkość. Również kwas glutaminowy i kwas asparaginowy mają zbliżony charakter. Natomiast fenyloalanina i arginina nie są do siebie podobne. Stopień podobieństwa par aminokwasów może być zapisany w postaci macierzy o rozmiarach 20×20 zwanej macierzą podstawień (*substitution matrix* lub *mutation matrix*) – rysunek 4.

Funkcja podobieństwa f służąca do zbudowania dopasowania ma w tym przypadku postać:

$$f(A, B) = M[A, B] \tag{3}$$

gdzie M jest macierzą podstawień, A i B są rodzajami aminokwasów.

[4] Czasowa złożoność obliczeniowa, $f(N)$, jest liczbą podstawowych operacji wykonywanych przez program w jednostce czasu. N jest parametrem charakteryzującym rozmiar problemu, na przykład długością sekwencji białka. Mówimy, że algorytm jest rzędu $O(g(N))$, jeżeli zaczynając od pewnego N_0 mamy $f(N) < cg(N)$ dla pewnej stałej c [25].

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1

Rysunek 4: Macierz podstawień BLOSUM80 [29].
Wartości dodatnie odpowiadają parom aminokwasów o podobnych własnościach.

Większość macierzy podstawień została zbudowana na podstawie analizy dopasowań sekwencji białek spokrewnionych ewolucyjnie. Można założyć, że aminokwasy ulegające mutacjom w podobnych sekwencjach nie powodują zmiany struktury przestrzennej białka – a więc są podobne do siebie w sensie zachowywania struktury. Macierze zbudowane w ten sposób to m.in. macierze PAM [30] i Gonnet [31]. Inną macierzą jest macierz BLOSUM [29] zbudowana w oparciu o ciągłe (nie zawierające przerw) fragmenty dopasowań. Tworzy się także macierze podstawień wykorzystując bezpośrednio podobieństwo strukturalne białek lub optymalizując macierz tak, aby otrzymać najlepsze rezultaty w metodzie, dla której została przeznaczona. Używa się do tego celu m.in. algorytmów genetycznych [32]. [5]

2.4.4 Inne metody porównywania sekwencji białek

Algorytm programowania dynamicznego pozwala na obliczenie optymalnej wartości miary dopasowania (zgodnie z założonymi wartościami kosztów wprowadzenia przerw w dopa-

[5] Algorytm genetyczny jest heurystyczną metodą optymalizacji globalnej, stosowaną szczególnie do rozwiązywania problemów kombinatorycznych. „Populacja" możliwych rozwiązań problemu jest modyfikowana przy pomocy operacji genetycznych: mutacji i krzyżowania, a rozwiązania lepiej spełniające wstępne założenia mają wyższe prawdopodobieństwo „przeżycia" [33].

sowaniu). Ma on jednak stosunkowo dużą złożoność obliczeniową i dlatego nie nadaje się do przeszukiwania baz danych zawierających miliony sekwencji białek. W tym celu używa się innych, szybszych algorytmów – najbardziej rozpowszechnione metody to BLAST [34, 35] i FASTA [36, 37].

Metoda BLAST (*Basic Local Alignment Search Tool*) opiera się na spostrzeżeniu, że dwie sekwencje podobne do siebie zawierają zwykle trójki aminowkwasów, które są identyczne lub prawie identyczne. Przyjęcie takiego założenia i odpowiednie zaprojektowanie algorytmu pozwala na przyspieszenie przeszukiwania bazy danych o kilka rzędów wielkości w stosunku do algorytmu *dynamic programming*. Wynikiem zastosowania metody BLAST jest dopasowanie lokalne obejmujące otoczenie znalezionego podciągu trzech aminokwasów.

Również metoda FASTA jest oparta na poszukiwaniu krótkich (dwuaminokwasowych) podciągów w porównywanych sekwencjach. Znajdowanych jest kilka takich lokalnie podobnych fragmentów i następnie budowane jest dopasowanie globalne. Metoda FASTA jest wolniejsza i mniej czuła niż algorytm BLAST.

2.4.5 Dopasowanie wielu sekwencji

Wykorzystanie informacji ewolucyjnej zawartej w znanych sekwencjach białek wymaga równoczesnego porównania wielu sekwencji [38]. Dzięki temu możliwe jest wskazanie aminokwasów, które nie uległy mutacjom podczas ewolucji. W wyniku porównania wielu sekwencji uzyskuje się dopasowanie wielu sekwencji (*multiple sequence alignment*). Korzyści wynikające z zastosowania dopasowania wielu sekwencji ilustruje poniższy (ekstremalny) przykład. Trudno jest dopasować następujące sekwencje:

```
S E S E S E S E
A V A V A V A V
```

Po uwzględnieniu trzeciej sekwencji i zbudowaniu dopasowania:

```
S E S E S E S E
+ - + - + - + -
S V S V S V S V
- + - + - + - +
A V A V A V A V
```

otrzymuje się 50% zgodność dopasowania.

Algorytm programowania dynamicznego nie nadaje się bezpośrednio do porównywania większej liczby sekwencji, gdyż jego złożoność obliczeniowa rośnie wykładniczo wraz z liczbą porównywanych sekwencji. Dostępne rozwiązania polegają na znalezieniu optymalnych dopasowań par sekwencji, a następnie na łączeniu tych par (algorytm typu *branch–and–bound*) [39]. W ten sposób działa najpopularniejszy program budujący dopasowania wielu sekwencji – CLUSTAL W [40]. Inne znane metody [41] wykorzystują m. in. ukryte modele Markowa (*HMM, Hidden Markov Models*) [42] i algorytmy genetyczne.

2.4.6 Wykorzystanie informacji ewolucyjnej

W procesie ewolucji żywych organizmów struktury białek zachowywane są w stopniu znacznie większym, niż ich sekwencje. Wynika to z mechanizmu działania doboru naturalnego. Niekorzystne mutacje, w wyniku których zmianie ulega struktura białka, powodują na ogół śmierć organizmu (a więc takie zmiany nie są przekazywane następnym pokoleniom) [6]. Natomiast mutacje nie wpływające niekorzystnie na strukturę białka (a więc również nie modyfikujące w znaczący sposób jego biologicznej funkcji) mogą zostać bezpiecznie przekazane. Można stąd wyciągnąć wniosek, że odcinki sekwencji nie ulegające zmianie w procesie ewolucji odpowiadają istotnym funkcjonalnie fragmentom struktury lub fragmentom determinującym jej stabilność. Najczęściej zachowywane są aminokwasy należące do hydrofobowego rdzenia białka.

Wykorzystanie informacji ewolucyjnej pozwala na zwiększenie czułości metod sekwencyjnych. W praktyce tworzy się dopasowanie wielu sekwencji w oparciu o sekwencyjną bazę danych, a następnie buduje profil sekwencyjny wykorzystywany do kolejnych przeszukiwań bazy danych [43]. Profil sekwencyjny ma postać macierzy o wymiarach

$20 \times N$, gdzie N jest długością sekwencji. Każda kolumna takiej macierzy zawiera częstości występowania różnych aminokwasów na danej pozycji (rysunek 5). W opisany sposób działa najpopularniejsza obecnie metoda sekwencyjna – PSI–BLAST (*Position Specific Iterated BLAST*) [35].

```
V F A N K N A N G L D F L        V F A N K N A N G L D F L
V F A N K N A N G D A F L   GLY 0 0 0 0 0 0 0 0 7 1 0 0 0
V Y A N K D A N G D A F L   ALA 0 0 8 0 0 0 2 0 0 2 0 0 0
V F A N K D A N G D N F L   SER 0 0 1 0 2 0 0 0 0 0 0 0 0
V F A N S D A N G A S F L   CYS 0 0 0 0 0 0 0 0 0 0 0 0 0
V F A N S A A N G D S F L   VAL 9 0 0 0 0 0 0 0 0 2 0 1 0
V Y S N Y E T S G V D I L   THR 0 0 0 0 0 3 4 0 0 0 0 0 0
V Y S N Y E T T G V D I L   ILE 0 0 0 0 0 0 0 0 0 0 0 4 0
V - - - - K H N E V D I L   PRO 0 0 0 0 0 0 0 0 0 0 0 0 0
V - - - - S H N E V D I L   MET 0 0 0 0 0 0 0 2 0 0 0 0 0
V - - - - K H N E V D I L   ASP 0 2 0 4 0 1 0 0 0 1 3 0 0
V - - - - S H N E V D I L   ASN 0 0 0 4 0 0 0 4 0 0 0 0 0
V E A D Y T T I G G L V L   LEU 0 0 0 0 0 0 0 0 0 0 3 0 9
V E A D Y T S H G A L V L   LYS 0 0 0 0 2 0 0 0 0 0 0 0 0
V E A D Y T T M G A L V L   GLU 0 2 0 0 0 0 0 0 1 0 0 0 0
V E A D S T T M G A L V L   GLN 0 0 0 0 0 0 0 0 0 0 0 0 0
- D A D S T T M G A L I L   ARG 0 0 0 0 0 0 0 0 0 0 0 0 0
- D A D Y T T M G A L I L   HIS 0 0 0 0 0 0 1 0 0 0 0 0 0
- D A D Y T T M G G L I L   PHE 0 3 0 0 0 0 0 0 0 0 0 2 0
- D A D Y T T M G G L I L   TYR 0 1 0 0 4 0 0 0 0 0 0 0 0
- D A D Y T T M G G L I L   TRP 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Rysunek 5: Przykład dopasowania wielu sekwencji (fragment sekwencji mioglobiny 1mba) i odpowiadający mu profil sekwencyjny (częstości występowania aminokwasów na danej pozycji przeskalowano do wartości z przedziału $\langle 0, 9 \rangle$).

Dodatkowe zwiększenie czułości można uzyskać budując profil sekwencyjny zarówno dla sekwencji badanego białka, jak też dla sekwencji umieszczonych w bazie danych. Tego typu podejście jest wykorzystywane przez kilka dostępnych metod sekwencyjnych uznawanych obecnie za najbardziej wiarygodne.[6] Należą do nich metody BASIC i PDB–BLAST.[7]

[6]Porównanie wielu metod przewidywania struktury białek znajduje się na stronie WWW `http://www.bioinfo.pl`

[7]Strona WWW programu PDB–BLAST: `http://bioinformatics.ljcrf.edu/pdb_blast/`

2.4.7 Budowanie modelu białka

Poszczególne programy wykorzystują różne algorytmy do budowania modeli białek. Na przykład program MODELLER odczytuje ze struktur homologicznych więzy (odległości i kąty pomiędzy atomami) i na ich podstawie tworzy rozmyty opis położenia wszystkich atomów w białku (zwany funkcją gęstości prawdopodobieństwa, *probability density function*) [44]. Następnie, zgodnie z tym opisem, generowany jest model cząsteczki białka. Takie podejście pozwala na wygodne łączenie danych pochodzących z kilku struktur białek homologicznych oraz na stosowanie innych więzów (na przykład pochodzących z metod eksperymentalnych). Inne programy korzystają z metody *distance geometry* w celu konwersji więzów (odległości pomiędzy atomami) do postaci współrzędnych kartezjańskich [45]. Po zbudowaniu modelu białka jest on optymalizowany przy pomocy mechaniki molekularnej i pełnoatomowego pola siłowego. Program MODELLER używa w tym celu pola siłowego CHARMM [46].

Inne programy (na przykład SWISS–MODEL) bezpośrednio „kopiują" fragmenty struktury białka–wzorca dopasowane do sekwencji modelowanego białka. Następnie dokonują „mutacji" odpowiadających sobie aminokwasów, to znaczy zamieniają grupy boczne aminokwasów z białka–wzorca na grupy boczne aminokwasów białka modelowanego. Fragmenty struktury odpowiadające częściom dopasowania zawierającym przerwy (są to zwykle pętle łańcucha białkowego) są uzupełniane przy pomocy pasujących fragmentów z bazy danych.

Dopasowania tworzone przez metody sekwencyjne są zazwyczaj dobrej jakości (pod warunkiem odnalezienia wzorca charakteryzującego się wysokim, ponad 30% podobieństwem sekwencyjnym) i pozwalają na zbudowanie modelu białka o rozdzielczości porównywalnej z rozdzielczością struktury eksperymentalnej (praca [47], załącznik 4). Standardowe programy służące do modelowania homologicznego (na przykład MODELLER [44], COMPOSER [48], SWISS–MODEL [49]) odczytują współrzędne atomów cząsteczki białka–wzorca i na podstawie dopasowania tworzą model nowego białka. Dlatego fragmenty źle dopasowane (na przykład pętle) są zwykle błędnie odbudowywane przez te programy. Podobnie, niedokładne dopasowanie (w przypadku niskiego podobieństwa se-

kwencji wzorca i sekwencji modelowanego białka) bywa przyczyną generowania modeli struktur białek o niskiej jakości. Program MODELLER wykazuje silną tendencję do zapętlania dłuższych nieustrukturyzowanych fragmentów łańcucha białkowego.

2.5 Metody przewlekania sekwencji (*threading*)

Metody wykorzystujące podobieństwo sekwencji pomijają całkowicie informacje strukturalne pochodzące z białka–wzorca. Dlatego stworzono grupę metod zwanych metodami przewlekania sekwencji (*threading*) [50]. Metody przewlekania generują dopasowanie (*alignment*) sekwencji modelowanego białka do struktury białka–wzorca. Miara dopasowania nie jest w tym przypadku prostą sumą podobieństw aminokwasów, ale wykorzystuje funkcję oceniającą kompatybilność sekwencji i struktury:

$$f(A, G) = V(A, G) \tag{4}$$

gdzie $V(A, G)$ jest wartością funkcji oceniającej zgodność sekwencji i struktury, A jest typem aminokwasu sekwencji białka modelowanego, G jest opisem fragmentu struktury białka–wzorca.

Przykładowo, jeżeli funkcja oceniająca zgodność typu aminokwasu i rodzaju struktury drugorzędowej ma następującą postać (wykorzystuje fakt, że alanina często występuje we fragmentach posiadających strukturę helikalną):

$$V(A, G) = \begin{cases} 1 & \text{gdy } A = \text{ALA i } G = \text{struktura helikalna (H)}\ , \\ 0 & \text{w przeciwnym wypadku.} \end{cases} \tag{5}$$

to wartość miary następującego dopasowania:

```
Sekwencja:  A A A A A A A A V V V
Struktura:  E E C C H H H H H H H
Zgodność :  0 0 0 0 1 1 1 1 0 0 0
```

jest równa 4.

Jako funkcję oceniającą kompatybilność sekwencji i struktury używa się często potencjałów statystycznych [51, 52].

Nazwa „metoda przewlekania sekwencji" pochodzi stąd, że *de facto* sekwencja białka o nieznanej strukturze jest „przewlekana" przez każde białko strukturalnej bazy danych. Dla każdej pary sekwencja–struktura budowane jest dopasowanie i obliczana miara dopasowania. Zwykle wykorzystuje się do tego algorytm programowania dynamicznego. Lista miar dopasowania uporządkowana malejąco opisuje zgodność sekwencji białka modelowanego (*target*) i kolejnych struktur–wzorców *templates* z bazy danych. Wartość miary dopasowania mówi o tym, czy znalezione dopasowanie jest wiarygodne i czy można użyć go do zbudowania modelu białka. Wiarygodność miary dopasowania ocenia się wykorzystując parametry rozkładu miar dopasowań (co opisano w kolejnym rozdziale).

Większość współcześnie wykorzystywanych metod przewlekania sekwencji to metody hybrydowe, łączące podejście czysto sekwencyjne z metodami opartymi na potencjałach statystycznych [23]. Jak wykazały testy przeprowadzone podczas konkursu CASP4 [53], jedne z najlepiej działających metod przewlekania sekwencji to 3D-PSSM [54], GenTHREADER [55] i FFAS (*Fold and Function Assignment System*) [56, 57].

Metoda przewlekania sekwencji posiada poważne ograniczenie. Algorytm programowania dynamicznego używany do tworzenia dopasowania jest algorytmem działającym lokalnie. Dlatego nie jest możliwe bezpośrednie wykorzystanie potencjałów dwu– i wielociałowych [58]. Stosowany sposób ominięcia tego ograniczenia polega na obliczeniu oddziaływania każdego aminokwasu z „uśrednioną" strukturą (założenie, że otoczenie aminokwasu nie ulega zmianie niezależnie od stworzonego dopasowania, tak zwane przybliżenie *frozen approximation*) [59, 60]. Inne rozwiązanie tego problemu zostało zaproponowane w programie PROSPECTOR, w którym proces budowania dopasowania sekwencji i struktury powtarza się wielokrotnie, co prowadzi do samouzgodnienia dopasowania [28]. Do tworzenia dopasowania próbuje się również używać innych algorytmów (na przykład metodę Monte Carlo [61]).

2.5.1 Ocena wiarygodności miary dopasowania sekwencji

Bezwzględna wartość miary dopasowania dwóch sekwencji (lub sekwencji do struktury) nie wystarcza do oceny wiarygodności dopasowania. Wartość miary dopasowania zależy m.in. od składu aminokwasowego porównywanych sekwencji, użytej funkcji podobieństwa aminokwasów, kosztów wprowadzenia i wydłużenia przerwy, długości porównywanych sekwencji. Dlatego konieczne jest wprowadzenie sposobu oceny wiarygodności miary dopasowania sekwencji.

Do oceny wiarygodności miary dopasowania sekwencji stosuje się często ocenę standardową (*z-score*):

$$z = \frac{\bar{x} - x}{\sigma(x)} \tag{6}$$

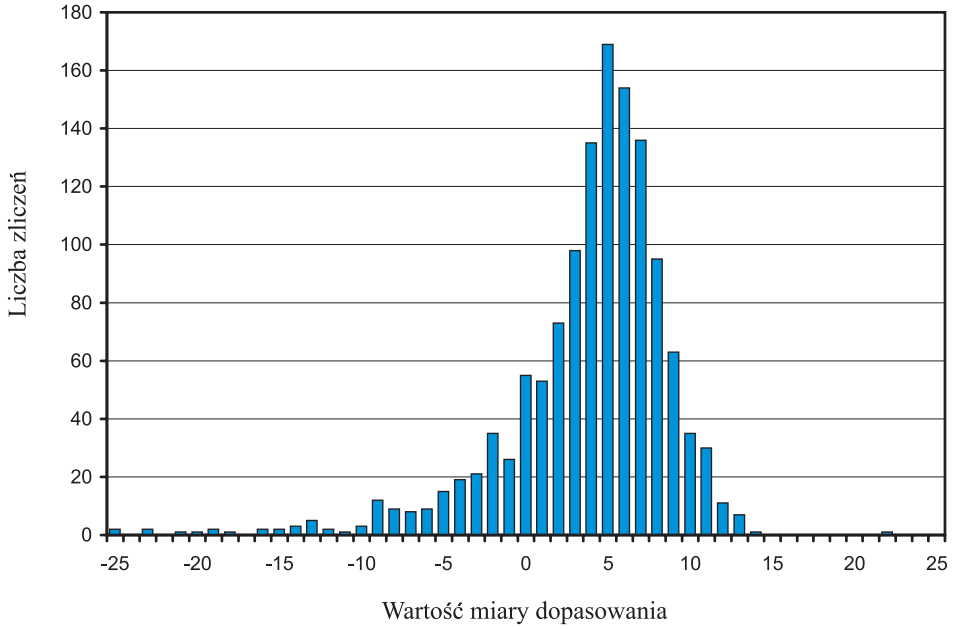
gdzie x jest miarą dopasowania, $\bar{x}$ jest średnią miarą dopasowania, $\sigma(x)$ jest odchyleniem standardowym wszystkich zebranych wyników.

Zakładając określony próg oceny standardowej, możemy ocenić wiarygodność miary dopasowania. Wybierając ocenę standardową zakłada się, że rozkład wartości miar dopasowania jest rozkładem normalnym. W rzeczywistości histogram wartości miar dopasowania daje się dobrze opisać przy pomocy rozkładu wartości maksymalnej, co przedstawiono na rysunkach 6 i 7 (niekiedy wykorzystuje się również rozkład Poissona) [62]. Dzięki temu możliwe jest obliczenie prawdopodobieństwa utworzenia danego dopasowania w sposób przypadkowy (im mniejsze prawdopodobieństwo zbudowania dopasowania w sposób przypadkowy, tym większa jest jego wiarygodność) [63]:

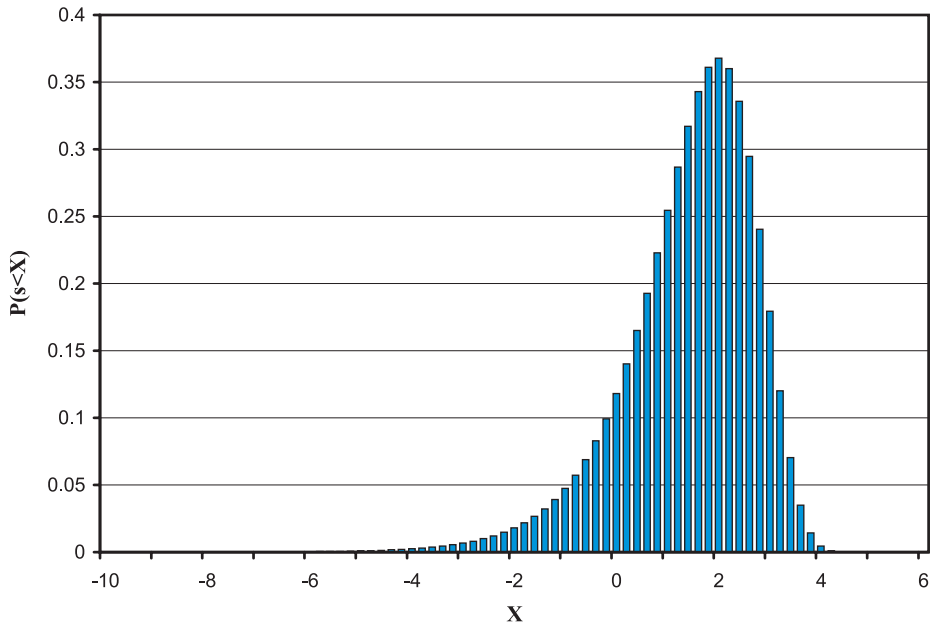
$$\mathrm{P}(s < X) = e^{-\gamma N M \xi^X} \tag{7}$$

gdzie $\mathrm{P}(s < X)$ jest prawdopodobieństwem przypadkowego utworzenia dopasowania o wartości miary dopasowania mniejszej niż X zgodnie z rozkładem wartości maksymalnej; N i M są długościami dopasowywanych sekwencji; γ i ξ są parametrami rozkładu.

Parametry γ i ξ można obliczyć na podstawie rzeczywistego rozkładu miar dopasowania [63, 64].



Rysunek 6: Przykładowy histogram wartości miar dopasowania – wyników metody przewlekania sekwencji.



Rysunek 7: Przykładowy histogram wartości obliczonych zgodnie z rozkładem wartości maksymalnej.

2.6 Metody bezpośredniego przewidywania struktury białek na podstawie sekwencji

Przedstawione dotychczas metody modelowania struktur białek wykorzystują znajomość struktur białek homologicznych. Metody sekwencyjne i metoda przewlekania sekwencji są w stanie znaleźć białka–wzorce dla około 40–50% poznawanych sekwencji. Jednak dla pozostałej części nowo sekwencjonowanych białek nie jest możliwe znalezienie odpowiednich homologów w strukturalnych bazach danych [65, 66]. Zawodzą obie grupy metod [11]. Próbę rozwiązania tego problemu stanowią metody bezpośredniego przewidywania struktury białek na podstawie sekwencji aminokwasów.

Większość metod bezpośrednich (metod klasy *ab initio*) korzysta z hipotezy termodynamicznej Anfinsena [67]. Mówi ona, że białka w stanie natywnym znajdują się w minimum energii swobodnej. Metody bezpośrednie korzystają zatem z funkcji potencjału tak skonstruowanej, aby przyjmował minimum dla konformacji białka odpowiadającej strukturze natywnej. Potencjały stosowane w metodach przewidywania struktury białek można podzielić na dwie grupy. Pierwszą grupę stanowią potencjały wyprowadzane na podstawie obliczeń kwantowochemicznych lub analizy danych spektroskopowych [68]. Tego typu potencjały są stosowane w obliczeniach mechaniki i dynamiki molekularnej [46, 69]. Drugą grupę potencjałów stanowią potencjały statystyczne, wyprowadzane w oparciu o analizę baz danych zawierających znane struktury białek (*knowledge–based potentials*) [70].

Znalezienie konformacji natywnej białka, nawet przy założeniu, że potencjał jest dobrze skonstruowany i rzeczywiście osiąga minimum dla struktury natywnej, jest nadal zadaniem bardzo trudnym – ze względu na ogromną dostępną przestrzeń konformacyjną [71]. Metoda dynamiki molekularnej jest niepraktyczna, ponieważ przy jej pomocy można symulować procesy trwające najwyżej dziesiątki nanosekund, w wyjątkowych wypadkach do 1 mikrosekundy [72]. Proces zwijania białek (*protein folding*) trwa w rzeczywistości od kilku milisekud do kilku minut, a więc o kilka rzędów wielkości dłużej. Ponadto, przeszkodę stanowi ogromny rozmiar symulowanego układu (kilkanaście lub kilkadziesiąt tysięcy atomów wchodzących w skład białka i otaczających go cząste-

czek wody). Dotychczas przeprowadzono nieliczne pomyślne próby symulacji procesu zwijania się niewielkich peptydów i białek [73].

Do przeszukiwania przestrzeni konformacyjnej używa się różnych metod obliczeniowych, m.in. metodę Monte Carlo (opisaną szerzej w rozdziale 2.6.1), algorytmy genetyczne [74], budowanie modeli z niewielkich (kilkuaminokwasowych) fragmentów struktur pochodzących ze struktur natywnych [75].

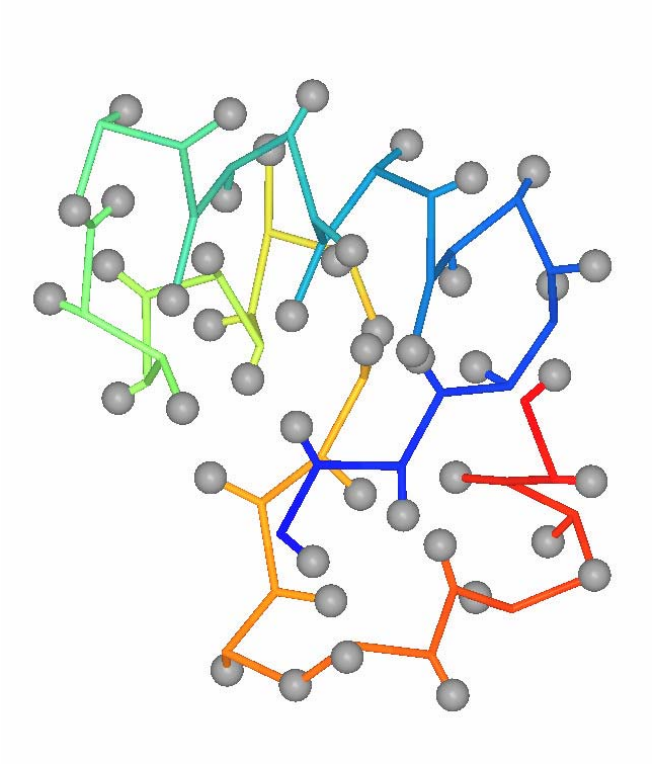
Bardzo ważnym aspektem metod przewidywania struktury białek jest sposób reprezentacji łańcucha białkowego [76]. Grupy boczne aminokwasów (determinujące strukturę białka) reprezentowane w postaci punktowych centrów oddziaływań pozwalają dramatycznie zmniejszyć konieczną do przeszukania przestrzeń konformacyjną. Dalsze ograniczenie przestrzeni konformacyjnej można uzyskać przez jej dyskretyzację, na przykład dzięki użyciu siatki [77].

2.6.1 Metoda dynamiki siatkowej Monte Carlo

Od wielu lat w grupie prof. Andrzeja Kolińskiego rozwijana jest metoda przewidywania struktury białek z wykorzystaniem uproszczonego modelu siatkowego i metody Monte Carlo [77, 78, 79, 80, 81]. Metoda ta jest z powodzeniem wykorzystywana do przewidywania trzeciorzędowej struktury białek bezpośrednio na podstawie sekwencji i została pomyślnie zweryfikowana w konkursach CASP3 [82] i CASP4 [83, 84]. W tej pracy wykorzystano jeden z wariantów tej metody, nazwany SICHO (SIde CHain Only). Przeznaczeniem metody jest przewidywanie struktury jednodomenowych białek globularnych o rozmiarach do 250 aminokwasów.

Poniżej przedstawiono najważniejsze założenia modelu SICHO. Dokładny opis modelu zawarty jest w pracach [80, 81, 85] (załączniki 1, 2 i 5).

Uproszczona reprezentacja siatkowa łańcucha białkowego Każdy aminokwas traktowany jest jak punkt umieszczony w środku masy grupy bocznej (powiększonej o atom Cα). Glicyna reprezentowana jest przez punkt odpowiadający położeniu węgla α (rysunek 8).



Rysunek 8: Uproszczony model łańcucha polipeptydowego (białko krambina, PDB 1crn). Szare kulki oznaczają pozycje środków mas grup bocznych aminokwasów. Kolorowy łańcuch łączy kolejne atomy Cα.

Punkty reprezentujące środki mas grup bocznych są umieszczone na siatce sześciennej o krawędziach długości 1.45 Å (jest to rozmiar jednostki siatkowej). Dopuszczalna odległość pomiędzy grupami bocznymi waha się od 4.3 do 7.9 Å i jest definiowana przez 646 możliwych wektorów łączących dwie kolejne grupy boczne. Ciąg tych wektorów reprezentuje łańcuch białkowy. Wyłączonej objętości grupy bocznej odpowiada 16 najbliższych węzłów siatki otaczających środek masy grupy bocznej. Pozycje atomów Cα są w przybliżony sposób odbudowywane w trakcie symulacji na podstawie dokładnie znanych pozycji grup bocznych (praca [81], załącznik 2, strona 295, rysunki 1–3).

Algorytm Monte Carlo Podstawowy algorytm Monte Carlo polega na wielokrotnym wykonywaniu następujących kroków:

- modyfikacja konformacji białka
- obliczenie energii zmodyfikowanego układu
- obliczenie prawdopodobieństwa zmiany konformacji zgodnie z kryterium Metropolisa; jeżeli konformacja zostaje odrzucona, następuje powrót układu do poprzedniego stanu

Według kryterium Metropolisa [86] prawdopodobieństwo przejścia ze stanu o energii E_1 do stanu o energii E_2 wynosi:

$$P(A,B) = \begin{cases} 1 & \text{gdy } E_2 <= E_1, \\ e^{\frac{E_1 - E_2}{k_B T}} & \text{gdy } E_2 > E_1. \end{cases} \tag{8}$$

gdzie k_B jest stałą Boltzmanna, T jest temperaturą symulacji (K).

Często wykorzystuje się schemat symulowanego schładzania (*simmulated annealing*), polegający na stopniowym zmniejszaniu temperatury podczas symulacji. Dzięki temu próbkowanie przestrzeni konformacyjnej jest coraz dokładniejsze w miarę zbliżania się do konformacji natywnej. Istnieje wiele udoskonaleń metody Monte Carlo, między innymi metoda ESMC (*entropy sampling Monte Carlo*) i metoda wymiany replik, REM (*replica exchange method*) [87].

Ruchy łańcucha W programie dynamiki siatkowej Monte Carlo zaimplementowane zostały następujące rodzaje ruchów łańcucha (N oznacza liczbę aminokwasów):

- ruch pojedynczej grupy bocznej: N wywołań podczas elementarnego kroku symulacji
- ruch dwóch grup bocznych: $N - 1$ wywołań
- ruch końców łańcucha: 2 wywołania

- ruch większych fragmentów łańcucha: jedno wywołanie podczas elementarnego kroku symulacji

Pole siłowe Pole siłowe wykorzystuje potencjały statystyczne otrzymane w wyniku analizy bazy danych struktur białek. W prezentowanym modelu użyto następujących potencjałów (szczegółowo opisanych w pracy [88], załącznik 5, na stronie 595):

1. Potencjał poprawiający lokalną geometrię łańcucha. Potencjał ten modyfikuje rozkład odległości w swobodnym łańcuchu (zbliżony do rozkładu Gaussa) tak, że przypomina on rozkład odległości w rzeczywistym łańcuchu białkowym. W zdecydowany sposób poprawia to jakość lokalnej geometrii łańcucha i ogranicza przestrzeń konformacyjną, zwiększając efektywność symulacji [89].

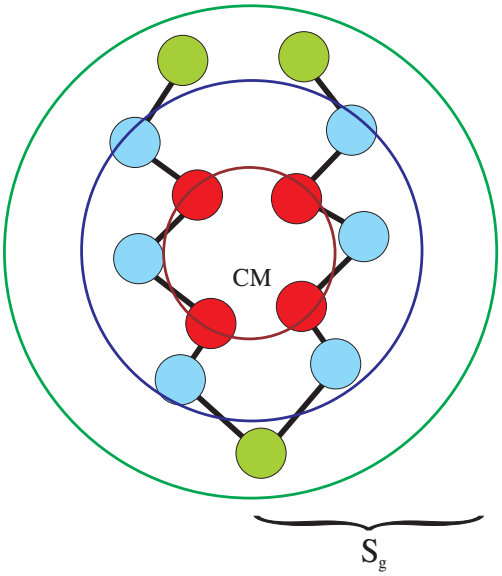
2. Potencjał hydrofobowy (typu *burial*), opisujący „preferencje" aminokwasów do przebywania w kontakcie ze środowiskiem zewnętrznym (definiujący *implicite* obecność rozpuszczalnika w układzie).

3. Potencjał centrosymetryczny, poprawiający strukturę białek globularnych wskutek wymuszania odpowiedniego rozkładu gęstości reszt aminokwasowych wewnątrz globuli białkowej (rysunek 9). Dla innych typów białek może on zostać odpowiednio zmodyfikowany (na przykład dla białek fibrylarnych może wymuszać symetrię cylindryczną).

4. Potencjał wiązań wodorowych, umożliwiający tworzenie się stabilnej struktury drugorzędowej i naddrugorzędowej.

5. Potencjał wielociałowy (typu profilu kontaktowego), opisujący lokalne otoczenie aminokwasów wewnątrz cząsteczki białka (zależny od liczby kontaktów z aminokwasami różnych typów i od orientacji kontaktów, co wyjaśniono bliżej w rozdziale 3.3).

6. Potencjał bliskiego zasięgu zależny od sekwencji, który opisuje „preferencje" aminokwasów do tworzenia określonej struktury drugorzędowej łańcucha.



Rysunek 9: Schemat rozkładu gęstości reszt aminokwasowych wewnątrz cząsteczki białka. *CM* oznacza środek masy cząsteczki białka, S_g – promień bezwładności (który dla białek globularnych może być w bardzo dokładny sposób przybliżony równaniem $S_g = 2.2N^{0.38}$, gdzie N jest liczbą aminokwasów).

7. Potencjał dalekiego zasięgu (kontaktowy) zależny od sekwencji, opisujący częstość występowania różnych aminokwasów blisko siebie w przestrzeni.

Bardzo ważnym aspektem dotyczącym pola siłowego jest ustalenie odpowiednich udziałów (wag) poszczególnych potencjałów w całkowitej energii układu (praca [88], załącznik 5, strona 598).

2.6.2 Zastosowanie modeli białek średniej rozdzielczości

Proces zwijania białek symulowany metodą Monte Carlo z wykorzystaniem uproszczonego modelu siatkowego pozwala niekiedy otrzymać struktury białek, które różnią się o 2 – 3 Å (RMSD) od struktury natywnej, co stanowi jakość zbliżoną do wyników metod eksperymentalnych [81] (załącznik 2). Często jednak metoda ta (podobnie jak inne metody klasy *ab initio* przewidywania struktury białek) tworzy modele o średniej rozdzielczości – odległe od struktury natywnej o 4 – 7 Å. Modele białek o takiej rozdzielczości trudno jest bezpośrednio zastosować do szczegółowej analizy funkcji białka (na przykład oddziaływania liganda z miejscem aktywnym receptora). Istnieją jednak

zastosowania, dla których modele białek o średniej rozdzielczości są wystarczające.

Przewidywanie funkcji białek Ocena biologicznej funkcji białka jest niekiedy możliwa już po przewidzeniu klasy strukturalnej (*fold*), do której należy białko, a więc na podstawie struktury o rozdzielczości 6 – 8 Å. W ostatnich latach powstała automatyczna metoda wykorzystująca „rozmyty" opis miejsc aktywnych w białkach do wyznaczania funkcji – *Fuzzy Functional Forms*, FFF [90, 91]. Zaletą tej metody jest możliwość wykorzystania modeli białek o średniej rozdzielczości (4 – 6 Å) do przewidywania funkcji. Metoda oparta jest na uproszczonym opisie geometrii aminokwasów tworzących miejsce aktywne w białku. Takie definicje umieszczone w bazie danych są używane do analizy modeli białek. Zgodność geometrii miejsca aktywnego i sekwencji budujących je aminokwasów pozwala określić funkcję białka. Metoda FFF została z powodzeniem zastosowana do analizy funkcji białek z rodziny oksydoreduktaz dwusiarczkowych (w genomie *E. coli*) oraz rybonukleaz-T1 [92].

Dokowanie ligandów do modeli białek o średniej rozdzielczości Techniki obliczeniowe pozwalające na dopasowanie (dokowanie) modeli niewielkich cząsteczek chemicznych (ligandów) do struktur białek (receptorów) są powszechnie wykorzystywane w praktyce i pozwalają na badanie mechanizmu procesu wiązania liganda przez receptor [93]. Możliwa jest również ocena ewentualnych efektów biochemicznych modyfikacji cząsteczki liganda bądź receptora. Standardowe metody dokowania (na przykład AutoDock [94], DOCK [95], FlexiDock [96]) wymagają użycia modeli receptorów o rozdzielczości porównywalnej z rozdzielczością struktur eksperymentalnych. W przeciwnym razie metody te nie są w stanie dostarczyć wiarygodnych wyników.

Ostatnio powstała metoda przewidywania konformacji układów ligand – receptor wykorzystująca uproszczone modele zarówno cząsteczki liganda, jak też białka receptorowego [97]. Struktura cząsteczki białka i cząsteczki liganda jest dopasowana do siatki sześciennej. Energia oddziaływania liganda i receptora jest obliczana przy użyciu prostych potencjałów statystycznych. Przeszukiwana jest cała możliwa przestrzeń konformacyjna (dzięki dyskretyzacji przy pomocy siatki odbywa się to stosunkowo szybko).

Metoda jest efektywna: pozwala trafnie ocenić położenie miejsca aktywnego w 2/3 modeli białek o rozdzielczości od 4 do 6 Å. Dla 20% modeli metoda poprawnie określa orientację liganda w miejscu aktywnym receptora.

Przewidywanie czwartorzędowej struktury białek Poznanie mechanizmu oddziaływań pomiędzy białkami jest kluczowe dla zrozumienia procesów zachodzących w komórce. Dlatego przewidywanie czwartorzędowej struktury białek jest bardzo ważnym i interesującym zastosowaniem modeli białek niskiej rozdzielczości. Powstały metody oparte na dynamice Monte Carlo uproszczonych modeli białek pozwalające z dużą dokładnością przewidywać strukturę kompleksów białkowych i aglomeratów większej liczby białek. Metody tego typu są szybkie i wykorzystują struktury białek o rozdzielczości do 7 Å [98, 99]. Stworzono również podobne metody oparte o algorytmy genetyczne [100]. Spektakularne wyniki (porównywalne z rezultatami metod eksperymentalnych) osiągnięto podczas symulacji *ab initio* struktur kompleksów krótkich białek helikalnych (struktur typu zamka leucynowego, *leucine zipper*) [101].

Poprawianie jakości przewidzianej struktury cząsteczki białka Struktury białek średniej i niskiej rozdzielczości mogą stanowić informację wejściową dla metod poprawiających jakość przewidzianej struktury białek. Metody wykorzystujące klasyczne pola siłowe są z powodzeniem stosowane do poprawiania modeli białek średniej rozdzielczości. To podejście jest używane w programach budujących modele białek w oparciu o podobieństwa sekwencyjne. Program MODELLER wykorzystuje w tym celu pole siłowe CHARMM [46] i metodę dynamiki molekularnej.

Dodatkowe informacje można uzyskać analizując całą trajektorię symulacji dynamiki procesu zwijania białek. Często zdarza się, że kolejne struktury w trajektorii „oscylują" wokół struktury natywnej białka. Można pogrupować podobne do siebie struktury i obliczyć strukturę średnią. Okazuje się, że często jest ona bardziej zbliżona do struktury natywnej białka, niż którakolwiek ze struktur zawartych w trajektorii [85]. Do takich obliczeń używa się metodę budowania klastrów [102] lub metodę *distance geometry*, opisaną bliżej w rozdziale 4.6 i w pracy [85] (załącznik 7).

3 Wykorzystanie podobieństw sekwencyjnych do wyprowadzania potencjałów statystycznych

Potencjały wyprowadzane w oparciu o analizę znanych struktur białek można podzielić na dwie grupy. Do pierwszej z nich należą potencjały wykorzystujące rozkłady parametrów geometrycznych białek (na przykład odległości pomiędzy atomami lub kontaktów grup bocznych aminokwasów) [103, 104, 105, 106]. Rozkłady te są konwertowane do postaci potencjału dzięki wykorzystaniu prawa Boltzmanna lub w oparciu o teorię Bayesa [107]. Drugą grupę potencjałów stanowią potencjały konstruowane w taki sposób, aby osiągały minimum dla struktury natywnej białka, a konformacje odbiegające od struktury natywnej (*decoys*) miały wyższe energie [108]. W tej pracy wykorzystano pierwszy sposób obliczania potencjałów.

Zgodnie z prawem Boltzmanna, różnica energii dwóch stanów o obsadzeniach N_1 i N_2, ΔE, jest równa:

$$\Delta E = -k_B T \ln \left(\frac{N_1}{N_2} \right) \tag{9}$$

gdzie k_B jest stałą Boltzmanna, T jest temperaturą (K).

Liczba obsadzeń N_1 odpowiada częstości występowania określonej konformacji w bazie danych. N_2 odpowiada spodziewanej częstości występowania określonej konformacji i jest określane mianem stanu odniesienia. Zastosowanie prawa Boltzmanna do obliczenia wartości potencjału statystycznego wymaga spełnienia kilku założeń [70]. Baza danych powinna być reprezentatywna dla wszystkich znanych struktur białek. Struktury znajdujące się w bazie danych powinny odpowiadać strukturom równowagowym (białkom w stanie natywnym). Rozkład częstości występowania określonych konformacji w bazie danych powinien być zgodny z rozkładem Boltzmanna.

3.1 Przygotowanie bazy danych

Potencjały statystyczne wyprowadzane są w oparciu o strukturalną bazę danych białek. Trudno jest wykorzystać do tego celu bezpośrednio bazę PDB, ponieważ jest ona redun-

dantna – występują w niej struktury bardzo podobne (na przykład homologiczne białka pochodzące z różnych organizmów). Ponadto, około 30% białek zawartych w bazie PDB posiada błędy (przerwy w łańcuchu polipeptydowym, brakujące aminokwasy, nieścisłości w numeracji atomów). Dlatego do wyprowadzania potencjałów konieczne jest zastosowanie reprezentatywnego podzbioru bazy PDB, pozbawionego struktur zawierających błedy. Do wyprowadzania potencjałów w tej pracy używano dwóch reprezentatywnych podzbiorów bazy PDB: bazy PDBSELECT98, zawierającej 1025 struktur białek [109], oraz bazy przygotowanej przy pomocy programu PDBREF, zawierającej 2860 struktur (opis programu PDBREF znajduje się w rozdziale 6.2.2).

3.2 Potencjał bliskiego zasięgu

Poszczególne aminokwasy występują z różną częstością we fragmentach łańcucha polipeptydowego o różnej strukturze drugorzędowej. Na przykład w helikalnych fragmentach łańcuchów białkowych często występuje alanina. Walina „preferuje" fragmenty łańcucha posiadające strukturę β.

Strukturę drugorzędową białek często reprezentuje się przy pomocy trójliterowego kodu: litera H odpowiada strukturze helikalnej (*helical*) α, litera E strukturze typu β (*extended*), litera C (lub znak '-') – pozostałym typom struktury drugorzędowej (*coil*). Reprezentacja łańcucha białkowego w metodzie Monte Carlo wykorzystuje bardziej dokładny sposób opisu lokalnej geometrii. Wykorzystuje się do tego celu odległości pomiędzy środkami mas grup bocznych aminokwasów. Zapis $r_{i,i+x}(A,B)$ odpowiada odległości pomiędzy środkiem masy i-tego i $i+x$-tego aminokwasu (i jest pozycją aminokwasu w sekwencji białka). A i B odpowiadają typom i-tego i $i+x$-tego aminokwasu. Na przykład $r_{i,i+1}(\mathrm{GLY},\mathrm{ALA})$ oznacza odległość pomiędzy sąsiadującymi ze sobą glicyną i alaniną. Dokładna reprezentacja lokalnej geometrii wymaga jednoczesnego użycia kilku odległości: $r_{i,i+1}$, $r_{i,i+2}$, $r_{i,i+3}$ i $r_{i,i+4}$. W przypadku odległości pomiędzy i-tym a i+3-cim aminokwasem, wartość odległości dodatkowo pomnożono przez wartość odpowiadającą chiralności łańcucha białkowego, wynoszącą 1 dla łańcucha prawoskrętnego i -1 dla łańcucha lewoskrętnego (parametr oznaczono symbolem $r^*_{i,i+3}$). Zapobiega to tworzeniu

konformacji nie występujących w rzeczywistych strukturach białek (na przykład lewoskrętnych helis). Przykładowe rozkłady parametru $r^*_{i,i+3}$ przedstawiono na rysunkach 10 i 11.

Stanem odniesienia dla potencjału bliskiego zasięgu są rozkłady odległości uśrednione po wszystkich aminokwasach. Rozkłady te przedstawiono na w pracy [80] (załącznik 1) na stronie 119, rysunek 5. Rozkłady odległości zdyskretyzowano, dzieląc je na kilka (od 3 do 14) przedziałów. Na przykład odległość $r_{i,i+1}$ podzielono na trzy przedziały: $r < 5$ Å, $r \in \langle 5\text{ Å}, 6.5\text{ Å}\rangle$, $r > 6.5$ Å. Potencjał bliskiego zasięgu w każdym z przedziałów obliczono w następujący sposób:

$$V_{i,i+x,m}(A,B) = -k_B T \ln\left(\frac{r_{i,i+x,m}(A,B)}{\langle r_{i,i+x,m}\rangle}\right) \tag{10}$$

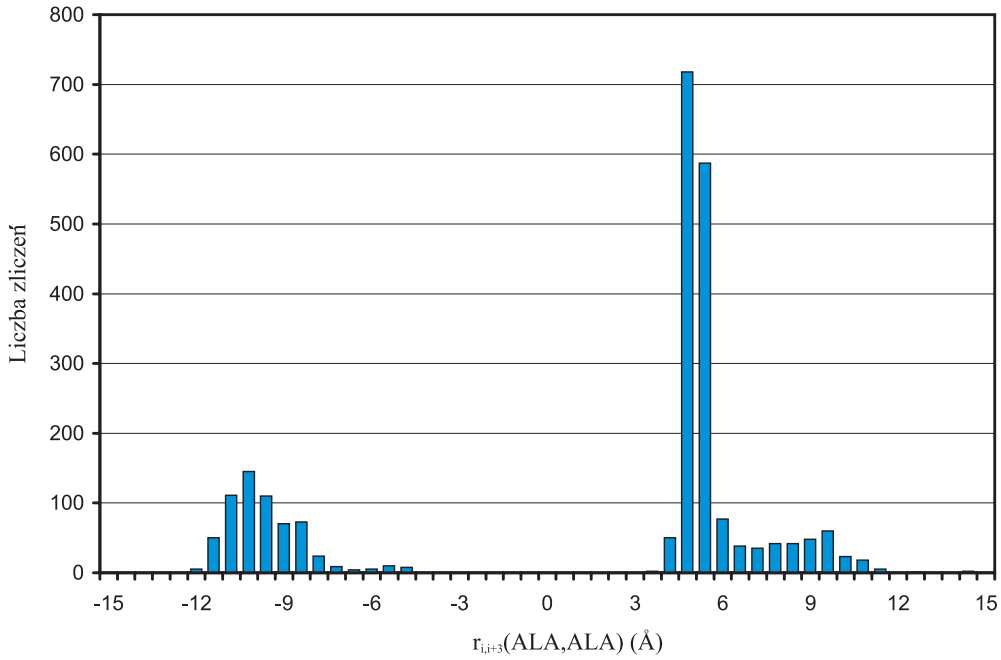
gdzie $r_{i,i+x,m}(A,B)$ jest zaobserwowaną częstością występowania odległości $r_{i,i+x}$ w przedziale m, pomiędzy aminokwasami typów A i B; $\langle r_{i,i+x,m}\rangle$ jest częstością występowania odległości $r_{i,i+x}$ w przedziale m, bez względu na typ aminokwasu.

Poniżej przedstawiono wartości potencjału $r^*_{i,i+3}$ obliczonego dla par alanina-alanina i walina-walina. Potencjał jest podzielony na 14 przedziałów, obejmując zakres od -14 do 14 Å, co 2 Å.

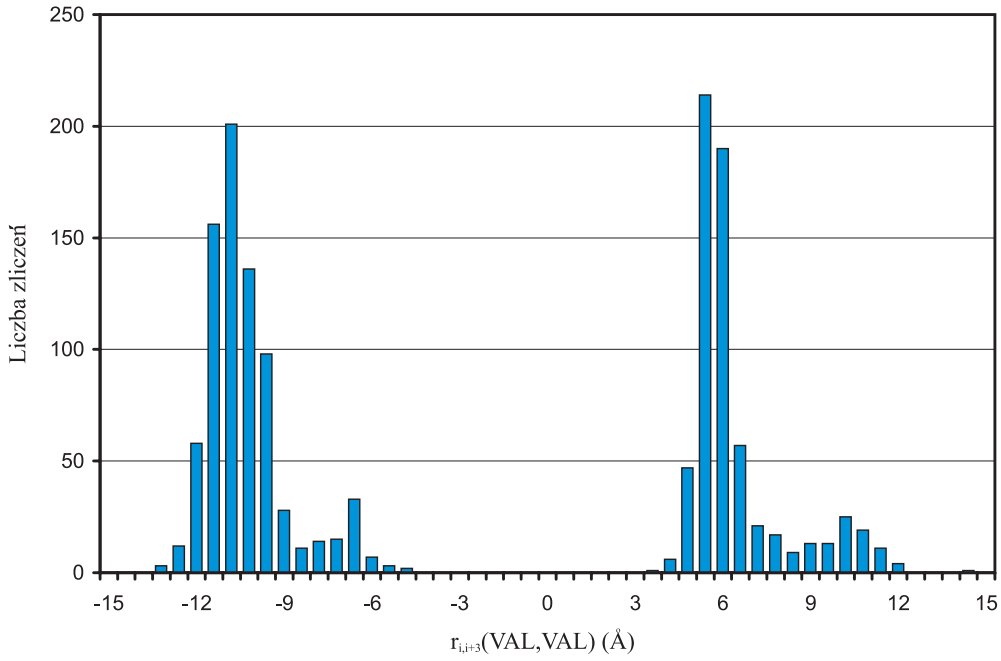
```
ALA-ALA
2.00 -0.73 -1.28 0.51 1.44 2.00 2.00 2.00 2.00 -1.17 -0.54 -0.68 0.54 2.00

VAL-VAL
1.28 -1.45 -1.13 0.22 1.40 2.00 2.00 2.00 2.00 -0.39 -0.65  0.11 0.01 2.00
```

Dla pary alanina–alanina wartość potencjału osiąga minimum dla konformacji prawoskrętnej helisy ($r^*_{i,i+3} \in (4\text{Å}, 6\text{Å}\rangle$), a dla pary walina–walina – dla konformacji lewostrętnej beta–wstęgi ($r^*_{i,i+3} \in (-12\text{Å}, -10\text{Å}\rangle$), co jest zgodne z rozkładem statystycznym parametru $r^*_{i,i+3}$ (rysunki 10 i 11).



Rysunek 10: Rozkład wartości parametru $r^*_{i,i+3}$ par alanina–alanina. Bezwzględna wartość parametru odpowiada odległości pomiędzy środkami mas grup bocznych aminokwasów. Znak parametru mówi o chiralności łańcucha (znak ujemny: łańcuch lewoskrętny, znak dodatni: prawoskrętny).



Rysunek 11: Rozkład wartości parametru $r^*_{i,i+3}$ par walina–walina.

3.3 Potencjał dalekiego zasięgu (kontaktowy)

Potencjał dwuciałowy dalekiego zasięgu (kontaktowy) opisuje „preferencje" poszczególnych aminokwasów do przebywania blisko innych aminokwasów (do kontaktowania się). Sformułowanie „dalekiego zasięgu" oznacza aminokwasy oddalone od siebie w sekwencji, ale leżące blisko siebie w przestrzeni. Bliskość aminokwasów w przestrzeni wynika z oddziaływań pomiędzy ich grupami bocznymi. Istotną rolę odgrywają tutaj oddziaływania elektrostatyczne i hydrofobowe [110]. Potencjał użyty w tej pracy ma charakter prostokątnej „studni" (przyjmuje pewną określoną wartość dla pary kontaktujących się aminokwasów i wartość 0 dla aminokwasów nie kontaktujących się ze sobą).

Podczas wyprowadzania potencjału kontaktowego przyjęto następującą definicję kontaktu: dwa aminokwasy kontaktują się ze sobą, jeżeli odległość pomiędzy ich dwoma dowolnymi atomami (z wyłączeniem wodorów) jest mniejsza, niż 4.5 Å. Zbiór wszystkich kontaktów w cząsteczce białka nosi nazwę mapy kontaktów [111]. W modelu SICHO definicja kontaktu jest inna, ze względu na uproszczony sposób reprezentacji łańcucha polipeptydowego: dwa aminokwasy kontaktują się ze sobą, jeżeli odległość pomiędzy środkami mas ich grup bocznych jest mniejsza niż pewien próg, wynoszący około 6.5 Å i zależny od typu aminokwasów, obliczony w taki sposób, aby uzyskać jak najlepszą zgodność z „pełnoatomową" definicją kontaktu.

W obliczeniach pominięto najbliższych sąsiadów (odległych w sekwencji o ± 1 aminokwas), którzy zawsze kontaktują się ze sobą (ponieważ odległość pomiędzy dwoma kolejnymi atomami Cα wynosi średnio 3.8 Å).

Potencjał dalekiego zasięgu zastosowany w tej pracy obliczono w następujący sposób:

1. Dla każdego typu aminokwasu znaleziona została całkowita liczba oddziaływań (kontaktów) S_A w bazie danych:

$$S_A = \sum_{i=1}^{D} \sum_{j=1}^{N_i} c_A(i,j) \tag{11}$$

gdzie D jest liczbą struktur białek w bazie danych, N_i jest długością sekwencji i-tego białka, $c_A(i,j)$ jest liczbą kontaktów j-tego aminokwasu w białku i ze

wszystkimi aminokwasami typu A w tym białku.

2. Dla każdego typu aminokwasu znaleziona została średnia liczba kontaktów q_A :

$$q_A = \frac{S_A}{\sum_{i=1}^{D} n_A(i)} \tag{12}$$

gdzie S_A jest liczbą oddziaływań aminokwasu typu A, D jest liczbą struktur białek w bazie danych, $n_A(i)$ jest liczbą aminokwasów typu A w białku i.

3. Dla każdego typu aminokwasu policzony został ułamek liczby kontaktów $X_A(i)$, niezależnie w każdym z białek w bazie danych:

$$X_A(i) = \frac{q_A \cdot n_A(i)}{\sum_{B=1}^{T} q_B \cdot n_B(i)} \tag{13}$$

gdzie $X_A(i)$ jest ułamkiem liczby kontaktów aminokwasu typu A w białku i, q_A jest średnią liczbą kontaktów aminokwasu typu A, $n_A(i)$ jest liczbą aminokwasów typu A w białku i, T jest liczbą typów aminokwasów (równą w tym przypadku 20).

4. Spodziewana liczba oddziaływań $c_{AB}^{expected}(i)$ pomiędzy aminokwasami dwóch typów wyrażona została w następujący sposób (jest to tak zwane przybliżenie *quasi–*chemiczne, [103, 112]):

$$c_{AB}^{expected}(i) = X_A(i) \cdot X_B(j) \cdot C(i) \tag{14}$$

gdzie $c_{AB}(i)$ jest spodziewaną liczbą oddziaływań pomiędzy aminokwasami typów A i B w białku i, $X_A(i)$ jest ułamkiem liczby kontaktów aminokwasu typu A w białku i, $C(i)$ jest całkowitą liczbą kontaktów pomiędzy wszystkimi aminokwasami w białku i.

5. Stan odniesienia dla potencjału dalekiego zasięgu stanowi spodziewana sumaryczna liczba oddziaływań $C_{AB}^{expected}$ we wszystkich białkach w bazie danych:

$$C_{AB}^{expected} = \sum_{i=1}^{D} c_{AB}^{expected}(i) \tag{15}$$

gdzie D jest liczbą struktur białek w bazie danych, $c_{AB}^{expected}(i)$ jest spodziewaną liczbą oddziaływań pomiędzy aminokwasami typów A i B w białku i.

6. Obliczono rzeczywistą liczbę oddziaływań pomiędzy aminokwasami typów A i B:

$$C_{AB} = \sum_{i=1}^{D} c_{AB}(i) \tag{16}$$

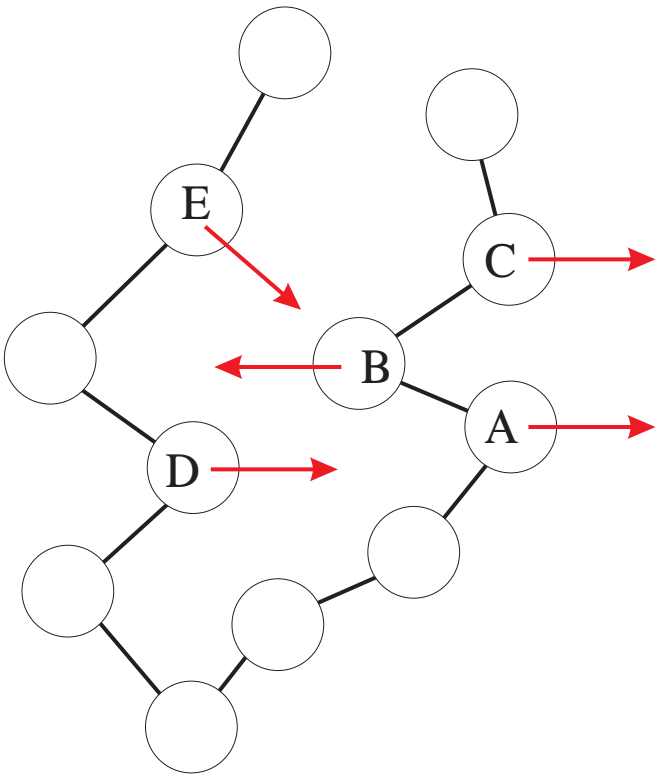
gdzie D jest liczbą struktur białek w bazie danych, $c_{AB}(i)$ jest liczbą oddziaływań pomiędzy aminokwasami typów A i B w białku i.

7. Potencjał dalekiego zasięgu został obliczony zgodnie ze wzorem Boltzmanna (9):

$$V_{AB}^{long} = -k_B T \ln\left(\frac{C_{AB}}{C_{AB}^{expected}}\right) \tag{17}$$

gdzie V_{AB}^{long} jest wartością potencjału dla pary aminokwasów typów A i B, C_{AB} jest rzeczywistą sumaryczną liczbą oddziaływań we wszystkich białkach w bazie danych, $C_{AB}^{expected}$ jest spodziewaną sumaryczną liczbą oddziaływań we wszystkich białkach w bazie danych.

Potencjał dalekiego zasięgu wykorzystywany w tej pracy został dodatkowo usprawniony dzięki uwzględnieniu różnych orientacji geometrycznych kontaktujących się aminokwasów (rysunek 12) [88]. Wszystkie kontakty pomiędzy aminokwasami podzielono na trzy klasy – w zależności od kąta pomiędzy wektorami łączącymi węgiel Cα i środek masy grupy bocznej kontaktujących się aminokwasów. Dla każdej klasy strukturalnej niezależnie obliczono potencjał dalekiego zasięgu. Obliczony potencjał przedstawiono na rysunku 13. Zauważyć można cechy potencjału odpowiadające charakterystycznym własnościom poszczególnych aminokwasów. Na przykład para lizyna i kwas asparaginowy „preferują" orientację równoległą (występują często na powierzchni białka), a para cysteina–cysteina – antyrównoległą.



Rysunek 12: Schemat możliwych orientacji kontaktujących się aminokwasów w łańcuchu białkowym. Strzałki oznaczają kierunki wektorów łączących atomy Cα ze środkami mas grup bocznych. Aminokwasy A i C – orientacja równoległa (kąt pomiędzy wektorami $\in (0, 60^{\mathrm{o}}\rangle$), B i E – orientacja pośrednia (kąt $\in (60, 120^{\mathrm{o}}\rangle$), B i D – orientacja antyrównoległa (kąt $\in (120, 180^{\mathrm{o}}\rangle$).

PAR	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
GLY	0.1	0.3	0.0	-0.2	0.1	0.0	0.1	0.4	0.2	0.2	0.1	0.2	0.4	0.4	0.2	0.3	0.3	0.1	0.0	0.1
ALA	0.3	-0.1	-0.2	-0.5	-0.4	-0.1	-0.4	0.4	-0.3	0.1	0.1	-0.4	0.4	0.4	0.2	0.3	0.0	-0.2	-0.1	0.2
SER	0.0	-0.2	-0.5	-0.5	-0.2	-0.5	0.0	0.3	0.0	-0.5	-0.4	0.0	-0.1	-0.2	-0.2	-0.8	-0.9	-0.8	-0.8	-0.8
CYS	-0.2	-0.5	-0.5	-1.6	-0.9	-0.4	-0.8	-0.1	-0.7	0.2	-0.1	-0.9	0.5	0.6	0.0	-0.6	-1.3	-1.5	-1.1	-1.3
VAL	0.1	-0.4	-0.2	-0.9	-1.0	-0.3	-0.9	0.2	-0.8	0.4	0.1	-1.0	0.1	0.2	0.0	-0.6	-0.9	-1.4	-1.1	-1.1
THR	0.0	-0.1	-0.5	-0.4	-0.3	-0.5	-0.3	0.2	-0.2	-0.4	-0.4	-0.3	-0.2	-0.4	-0.4	-0.9	-0.9	-0.8	-0.8	-0.8
ILE	0.1	-0.4	0.0	-0.8	-0.9	-0.3	-1.0	0.3	-0.8	0.4	0.2	-1.0	0.2	0.2	0.0	-0.6	-0.7	-1.4	-1.2	-1.2
PRO	0.4	0.4	0.3	-0.1	0.2	0.2	0.3	0.8	0.2	0.5	0.3	0.4	0.6	0.4	0.3	-0.5	-0.6	-0.6	-0.8	-0.9
MET	0.2	-0.3	0.0	-0.7	-0.8	-0.2	-0.8	0.2	-1.0	0.4	0.2	-0.9	0.3	0.3	0.0	-0.6	-0.9	-1.4	-1.1	-1.1
ASP	0.2	0.1	-0.5	0.2	0.4	-0.4	0.4	0.5	0.4	-0.2	-0.6	0.5	-0.8	0.1	-0.3	-1.2	-1.0	-0.3	-0.7	-0.5
ASN	0.1	0.1	-0.4	-0.1	0.1	-0.4	0.2	0.3	0.2	-0.6	-0.6	0.2	-0.2	-0.2	-0.3	-0.8	-0.8	-0.6	-0.8	-0.8
LEU	0.2	-0.4	0.0	-0.9	-1.0	-0.3	-1.0	0.4	-0.9	0.5	0.2	-1.2	0.3	0.2	-0.2	-0.6	-0.8	-1.4	-1.1	-1.1
LYS	0.4	0.4	-0.1	0.5	0.1	-0.2	0.2	0.6	0.3	-0.8	-0.2	0.3	0.3	-1.0	-0.3	-0.4	-0.6	-0.2	-0.8	-0.5
GLU	0.4	0.4	-0.2	0.6	0.2	-0.4	0.2	0.4	0.3	0.1	-0.2	0.2	-1.0	0.2	-0.2	-1.4	-1.0	-0.4	-0.7	-0.6
GLN	0.2	0.2	-0.2	0.0	0.0	-0.4	0.0	0.3	0.0	-0.3	-0.3	-0.2	-0.3	-0.2	-0.3	-1.0	-0.9	-0.6	-0.8	-0.9
ARG	0.3	0.3	-0.8	-0.6	-0.6	-0.9	-0.6	-0.5	-0.6	-1.2	-0.8	-0.6	-0.4	-1.4	-1.0	-0.8	-0.9	-0.6	-1.0	-0.7
HIS	0.3	0.0	-0.9	-1.3	-0.9	-0.9	-0.7	-0.6	-0.9	-1.0	-0.8	-0.8	-0.6	-1.0	-0.9	-0.9	-1.5	-0.9	-1.2	-1.1
PHE	0.1	-0.2	-0.8	-1.5	-1.4	-0.8	-1.4	-0.6	-1.4	-0.3	-0.6	-1.4	-0.2	-0.4	-0.6	-0.6	-0.9	-1.5	-1.3	-1.4
TYR	0.0	-0.1	-0.8	-1.1	-1.1	-0.8	-1.2	-0.8	-1.1	-0.7	-0.8	-1.1	-0.8	-0.7	-0.8	-1.0	-1.2	-1.3	-1.1	-1.3
TRP	0.1	0.2	-0.8	-1.3	-1.1	-0.8	-1.2	-0.9	-1.1	-0.5	-0.8	-1.1	-0.5	-0.6	-0.9	-0.7	-1.1	-1.4	-1.3	-1.6

MID	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
GLY	0.1	0.4	0.1	-0.3	0.3	0.2	0.3	0.5	0.2	0.5	0.2	0.3	0.7	0.9	0.5	0.4	0.4	0.2	0.2	0.2
ALA	0.4	0.1	0.3	-0.2	-0.2	0.2	-0.1	0.5	0.0	0.6	0.4	-0.2	0.8	0.8	0.4	0.5	0.5	0.0	0.2	0.3
SER	0.1	0.3	0.1	-0.1	0.4	0.2	0.4	0.5	0.4	0.2	0.3	0.4	0.7	0.5	0.3	-0.4	-0.6	-0.6	-0.6	-0.7
CYS	-0.3	-0.2	-0.1	-1.7	-0.3	0.0	-0.2	0.1	-0.3	0.3	0.3	-0.4	0.8	0.8	0.3	-0.6	-1.1	-1.2	-1.0	-1.1
VAL	0.3	-0.2	0.4	-0.3	-0.2	0.3	-0.2	0.3	-0.2	0.8	0.7	-0.4	1.0	0.9	0.6	-0.3	-0.4	-1.0	-0.7	-0.9
THR	0.2	0.2	0.2	0.0	0.3	0.3	0.4	0.5	0.3	0.3	0.4	0.4	0.8	0.5	0.4	-0.3	-0.5	-0.6	-0.5	-0.6
ILE	0.3	-0.1	0.4	-0.2	-0.2	0.4	-0.3	0.4	-0.2	0.9	0.7	-0.3	1.0	0.8	0.7	-0.2	-0.5	-1.1	-0.8	-1.0
PRO	0.5	0.5	0.5	0.1	0.3	0.5	0.4	0.5	0.3	0.8	0.6	0.4	1.1	0.7	0.5	-0.4	-0.5	-0.6	-0.8	-0.8
MET	0.2	0.0	0.4	-0.3	-0.2	0.3	-0.2	0.3	-0.5	0.9	0.6	-0.3	1.0	0.9	0.5	-0.4	-0.6	-1.2	-1.0	-1.0
ASP	0.5	0.6	0.2	0.3	0.8	0.3	0.9	0.8	0.9	0.6	0.2	0.8	0.1	1.1	0.6	-0.8	-0.6	-0.2	-0.5	-0.5
ASN	0.2	0.4	0.3	0.3	0.7	0.4	0.7	0.6	0.6	0.2	0.1	0.6	0.6	0.6	0.3	-0.3	-0.5	-0.4	-0.6	-0.5
LEU	0.3	-0.2	0.4	-0.4	-0.4	0.4	-0.3	0.4	-0.3	0.8	0.6	-0.5	0.8	0.8	0.4	-0.4	-0.5	-1.2	-0.9	-1.0
LYS	0.7	0.8	0.7	0.8	1.0	0.8	1.0	1.1	1.0	0.1	0.6	0.8	1.6	0.2	0.8	0.2	0.0	-0.1	-0.3	-0.2
GLU	0.9	0.8	0.5	0.8	0.9	0.5	0.8	0.7	0.9	1.1	0.6	0.8	0.2	1.2	0.7	-0.7	-0.4	-0.1	-0.4	-0.3
GLN	0.5	0.4	0.3	0.3	0.6	0.4	0.7	0.5	0.5	0.6	0.3	0.4	0.8	0.7	0.6	-0.4	-0.2	-0.3	-0.4	-0.5
ARG	0.4	0.5	-0.4	-0.6	-0.3	-0.3	-0.2	-0.4	-0.4	-0.8	-0.3	-0.4	0.2	-0.7	-0.4	-0.2	-0.4	-0.4	-0.5	-0.6
HIS	0.4	0.5	-0.6	-1.1	-0.4	-0.5	-0.5	-0.5	-0.6	-0.6	-0.5	-0.5	0.0	-0.4	-0.2	-0.4	-1.0	-0.6	-0.7	-0.8
PHE	0.2	0.0	-0.6	-1.2	-1.0	-0.6	-1.1	-0.6	-1.2	-0.2	-0.4	-1.2	-0.1	-0.1	-0.3	-0.4	-0.6	-1.4	-1.0	-1.2
TYR	0.2	0.2	-0.6	-1.0	-0.7	-0.5	-0.8	-0.8	-1.0	-0.5	-0.6	-0.9	-0.3	-0.4	-0.4	-0.5	-0.7	-1.0	-0.8	-1.0
TRP	0.2	0.3	-0.7	-1.1	-0.9	-0.6	-1.0	-0.8	-1.0	-0.5	-0.5	-1.0	-0.2	-0.3	-0.5	-0.6	-0.8	-1.2	-1.0	-1.1

ANT	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
GLY	0.1	0.1	0.1	-0.4	0.1	0.2	0.2	0.4	0.2	0.5	0.3	0.3	1.0	1.0	0.6	0.7	0.5	0.3	0.4	0.3
ALA	0.1	-0.1	0.3	-0.3	-0.5	0.2	-0.4	0.4	-0.2	1.0	0.7	-0.4	1.0	1.0	0.6	0.6	0.4	-0.3	0.0	0.1
SER	0.1	0.3	0.3	-0.1	0.3	0.5	0.4	0.5	0.6	0.7	0.5	0.3	1.3	0.9	0.7	-0.2	-0.6	-0.7	-0.6	-0.5
CYS	-0.4	-0.3	-0.1	-2.3	-0.3	-0.1	-0.2	-0.1	-0.4	0.6	0.2	-0.4	0.4	0.7	0.1	-0.7	-1.2	-1.4	-1.2	-1.2
VAL	0.1	-0.5	0.3	-0.3	-0.6	0.3	-0.5	0.3	-0.2	1.0	0.7	-0.4	0.7	0.8	0.4	-0.3	-0.7	-1.3	-1.0	-1.1
THR	0.2	0.2	0.5	-0.1	0.3	0.6	0.3	0.6	0.5	0.9	0.5	0.3	1.2	1.0	0.6	-0.1	-0.6	-0.8	-0.6	-0.8
ILE	0.2	-0.4	0.4	-0.2	-0.5	0.3	-0.5	0.4	-0.3	0.9	0.7	-0.4	0.7	0.7	0.3	-0.4	-0.8	-1.4	-1.1	-1.2
PRO	0.4	0.4	0.5	-0.1	0.3	0.6	0.4	0.4	0.3	1.0	0.6	0.4	1.3	1.0	0.6	-0.1	-0.6	-0.7	-0.8	-0.8
MET	0.2	-0.2	0.6	-0.4	-0.2	0.5	-0.3	0.3	-0.7	1.1	0.5	-0.3	0.7	1.1	0.3	-0.3	-0.7	-1.4	-0.9	-1.1
ASP	0.5	1.0	0.7	0.6	1.0	0.9	0.9	1.0	1.1	1.8	0.8	1.0	0.9	2.1	1.2	-0.2	-0.3	-0.1	-0.2	-0.2
ASN	0.3	0.7	0.5	0.2	0.7	0.5	0.7	0.6	0.5	0.8	0.6	0.8	1.4	1.2	0.8	0.1	-0.2	-0.4	-0.3	-0.4
LEU	0.3	-0.4	0.3	-0.4	-0.4	0.3	-0.4	0.4	-0.3	1.0	0.8	-0.4	0.6	0.7	0.3	-0.4	-0.6	-1.4	-1.1	-1.1
LYS	1.0	1.0	1.3	0.4	0.7	1.2	0.7	1.3	0.7	0.9	1.4	0.6	2.5	1.4	1.5	1.0	0.3	-0.2	0.0	0.0
GLU	1.0	1.0	0.9	0.7	0.8	1.0	0.7	1.0	1.1	2.1	1.2	0.7	1.4	2.3	1.6	-0.1	0.0	-0.1	-0.1	-0.2
GLN	0.6	0.6	0.7	0.1	0.4	0.6	0.3	0.6	0.3	1.2	0.8	0.3	1.5	1.6	0.8	0.2	-0.1	-0.4	-0.3	-0.4
ARG	0.7	0.6	-0.2	-0.7	-0.3	-0.1	-0.4	-0.1	-0.3	-0.2	0.1	-0.4	1.0	-0.1	0.2	0.4	0.0	-0.5	-0.3	-0.5
HIS	0.5	0.4	-0.6	-1.2	-0.7	-0.6	-0.8	-0.6	-0.7	-0.3	-0.2	-0.6	0.3	0.0	-0.1	0.0	-0.9	-0.8	-0.6	-0.9
PHE	0.3	-0.3	-0.7	-1.4	-1.3	-0.8	-1.4	-0.7	-1.4	-0.1	-0.4	-1.4	-0.2	-0.1	-0.4	-0.5	-0.8	-1.6	-1.2	-1.2
TYR	0.4	0.0	-0.6	-1.2	-1.0	-0.6	-1.1	-0.8	-0.9	-0.2	-0.3	-1.1	0.0	-0.1	-0.3	-0.3	-0.6	-1.2	-0.9	-1.1
TRP	0.3	0.1	-0.5	-1.2	-1.1	-0.8	-1.2	-0.8	-1.1	-0.2	-0.4	-1.1	0.0	-0.2	-0.4	-0.5	-0.9	-1.2	-1.1	-1.4

Rysunek 13: Potencjał dalekiego zasięgu zależny od wzajemnej orientacji kontaktujących się aminokwasów. Od góry: potencjał odpowiadający orientacji równoległej, pośredniej i antyrównoległej.

3.4 Wprowadzanie informacji ewolucyjnych do potencjałów statystycznych

Przedstawione dotychczas potencjały są zależne od typów par aminokwasów. Wyższą specyficzność potencjałów statystycznych można uzyskać wprowadzając do nich informacje ewolucyjne, dzięki uwzględnieniu lokalnego podobieństwa sekwencyjnego do białek o znanej strukturze, a następnie budując potencjał oddzielnie dla każdej sekwencji modelowanych białek. Poniżej opisano sposób wprowadzania informacji ewolucyjnych do potencjałów statystycznych bliskiego i dalekiego zasięgu. Usprawnione w ten sposób potencjały zostały wykorzystane w metodzie SICHO oraz w metodzie przewlekania sekwencji PROSPECTOR [28].

Sekwencja modelowanego białka jest porównywana z nieredundantną bazą danych sekwencji białek (zawierającą około 500 tysięcy sekwencji) przy pomocy programu PSI–BLAST. W wyniku otrzymuje się dopasowanie wielu sekwencji, które konwertowane jest do postaci profilu sekwencyjnego (opisanego w rozdziale 2.4.6):

$$P_A(i) = \frac{n_A(i)}{M}, i = 1, 2, ...N \tag{18}$$

gdzie $P_A(i)$ jest wartością profilu sekwencyjnego dla aminokwasu typu A na i-tej pozycji sekwencji, $n_A(i)$ jest liczbą wystąpień aminokwasu typu A w wejściowym dopasowaniu wielu sekwencji, M jest liczbą dopasowanych sekwencji, N jest długością sekwencji modelowanego białka.

Następnie reprezentatywana baza struktur białek jest przeszukiwana w celu odnalezienia fragmentów struktur o podobnej sekwencji. Odbywa się to przy użyciu odcinka sekwencji o długości kilkunastu aminokwasów („okna" sekwencji). Dla kolejnych pozycji i sekwencji modelowanego białka i pozycji j sekwencji białka o znanej strukturze, oblicza się wartość lokalnego dopasowania $s(i, j)$:

$$s(i,j) = \frac{\sum_{k=-W}^{W} \sum_{l=1}^{20} P_{A_l}(i+k) \cdot f(A_l, D(j+k))}{2W+1}, i = 1, 2, ...N, j = 1, 2, ...M \tag{19}$$

gdzie $2W + 1$ jest rozmiarem okna, P jest profilem sekwencyjnym, A_l jest l-tym

typem aminokwasu (A_1 odpowiada glicynie, A_2 alaninie, itd.), f jest funkcją podobieństwa pary aminokwasów, D jest sekwencją białka o znanej strukturze, $D(j+k)$ jest typem aminokwasu występującym na $j+k$-tej pozycji sekwencji D, N jest długością sekwencji modelowanego białka, M jest długością sekwencji D. Dla wartości $i+k$, $j+k$ wykraczających odpowiednio poza przedziały $\langle 1, N\rangle$ i $\langle 1, M\rangle$, profil sekwencyjny P i funkcja podobieństwa aminokwasów f przyjmują wartość 0.

Jako funkcję podobieństwa typów aminokwasów zastosowano macierz mutacji BLOSUM80 [29]. Rozmiar okna sekwencji wynosił 15 aminokwasów. Potencjały statystyczne zmodyfikowano w następujący sposób:

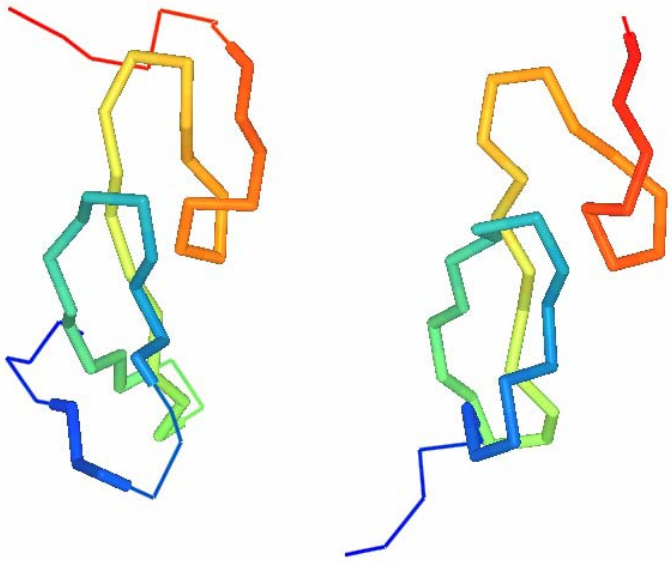
Potencjał bliskiego zasięgu Wartość potencjału obliczono niezależnie dla każdej pozycji i sekwencji modelowanego białka. Dla każdej pozycji i znaleziono 200 najlepiej dopasowanych fragmentów struktur, o najwyższych wartościach miar dopasowania $s(i,j)$. Miarę lokalnego dopasowania $s(i,j)$ zastosowano jako wagę podczas obliczania częstości występowania odległości r w przedziale m. Następnie potencjał bliskiego zasięgu obliczono według równania 10.

Potencjał dalekiego zasięgu Wartość potencjału obliczono niezależnie dla każdej pary aminokwasów i_1, i_2 sekwencji modelowanego białka. Dla każdej pary i_1, i_2 znaleziono wartości miar dopasowania $s(i_1, j_1)$, $s(i_2, j_2)$. Miary dopasowania zastosowano jako wagi podczas obliczania liczby kontaktów c. Następnie potencjał bliskiego zasięgu obliczono zgodnie z równaniami 11 – 17.

3.5 Ocena specyficzności potencjałów statystycznych

Poprawnie skonstruowana funkcja potencjału powinna osiągać minimum dla struktury natywnej białka, z którego pochodzi sekwencja testowa, a wartości zbliżone – dla struktur białek homologicznych. Do testowania potencjałów statystycznych wykorzystano uproszczoną metodę przewlekania sekwencji, bez wprowadzania przerw w dopasowaniu (*gapless threading*). Dla pary sekwencja–struktura zawierających odpowiednio N i M aminokwasów, możliwe jest zbudowanie $|N-M+1|$ takich dopasowań. Procedurę powta-

rza się dla wszystkich białek w strukturalnej bazie danych. Parametry wyprowadzania potencjałów (rozmiar okna sekwencji, rodzaj funkcji oceniającej podobieństwo par aminokwasów) dobrano w taki sposób, aby zmaksymalizować wartość oceny standardowej (*z-score*) miary dopasowania do struktury natywnej. Właściwym testem potencjałów wzbogaconych o informacje ewolucyjne było ich wykorzystanie podczas symulacji procesu zwijania białek metodą SICHO.



Rysunek 14: Przykład analogii strukturalnej. Białka 1egf (czynnik wzrostu, z lewej) i 1apo (czynnik koagulacji krwi), o długościach sekwencji odpowiednio 42 i 52 aminokwasy, nałożono przy pomocy programu SAL (przedstawionego w rozdziale 6.2.3). Grubą linią zaznaczono fragmenty o podobnej strukturze, RMSD Cα pomiędzy nimi wynosi 2.3 Å, identyczność sekwencji jest równa 16%.

Dzięki zastosowaniu potencjałów statystycznych można wykryć podobieństwa strukturalne w białkach niespokrewnionych ze sobą. Mówimy w takim przypadku o analogii strukturalnej (rysunek 14).

4 Modelowanie struktur białek w oparciu o podobieństwa sekwencyjne i analogie strukturalne

4.1 Modelowanie *ab initio* struktur białek z wykorzystaniem niewielkiej liczby więzów

Metody eksperymentalne, a zwłaszcza metody NMR, często nie są w stanie dostarczyć informacji wystarczających do zbudowania modelu białka. Klasyczne programy wykorzystywane w badaniach białek metodą NMR wymagają kilkunastu więzów (bliskiego i dalekiego zasięgu) przypadających na jeden aminokwas do zbudowania poprawnego modelu białka [8]. Nie zawsze udaje się eksperymentalnie uzyskać wystarczającą liczbę więzów. Stąd potrzeba stworzenia metody korzystającej z mniejszej ich liczby. Zastosowania takiej metody nie są ograniczone wyłącznie do wspomagania metod eksperymentalnych. Więzy mogą być odczytane ze struktur białek homologicznych. Wiarygodność takich więzów można poprawić wykorzystując kilka struktur homologicznych i tworząc odpowiedni konsensus. Modelowanie struktur białek w oparciu o niewielką liczbę więzów przedstawiono w pracy [80] (załącznik 1, strona 125).

W pracy wykorzystano wcześniejszą wersję modelu SICHO (wykorzystującą nieco inny rodzaj siatki). Potencjał statystyczny bliskiego zasięgu będący częścią tego modelu wyprowadzono w oparciu o podobieństwa sekwencyjne. Metodę przetestowano na kilku białkach należących do różnych klas strukturalnych, o rozmiarach od 56 do 146 aminokwasów. Wykorzystano więzy kontaktowe pomiędzy środkami mas grup bocznych aminokwasów. Wyniki przeprowadzonego testu wskazują, że metoda jest w stanie zbudować modele białek zbliżone jakością do struktur eksperymentalnych (RMSD Cα $\in \langle 2.6\text{Å}, 4.6\text{Å}\rangle$). Wymaga to użycia jednego więzu przypadającego na 5–7 aminokwasów. Białka, w których przeważa struktura drugorzędowa typu β, wymagają użycia większej liczby więzów, ze względu na większą dostępną przestrzeń konformacyjną. Liczba więzów niezbędnych do zbudowania poprawnego modelu białka jest zatem znacząco mniejsza, niż w przypadku standardowo wykorzystywanych programów do obliczeń NMR.

Zaprezentowana metoda została z powodzeniem wykorzystana do budowania modeli białek w oparciu o rzeczywiste więzy NMR [113].

4.2 Poprawianie modeli zbudowanych przy pomocy metody przewlekania

Istotnym problemem napotykanym podczas prób praktycznego zastosowania wyników metody przewlekania sekwencji jest niezadowalająca jakość tworzonych przez nią dopasowań, a w konsekwencji budowanych modeli białek. W dopasowaniach sekwencyjno–strukturalnych można napotkać następujące typy błędów:

- Przesunięcia (*misalignments*), polegające na niewłaściwym dopasowaniu fragmentu sekwencji do fragmentu struktury. Często spotykane są niewielkie przesunięcia (na przykład o dwa aminowkwasy w beta–wstędze), albo błędne dopasowania całych motywów strukturalnych, niszczące poprawną strukturę budowanego modelu białka.

- Fragmenty niedopasowane (*unaligned regions*), wynikające zwykle z różnic strukturalnych pomiędzy strukturą modelowanego białka a strukturą wzorca.

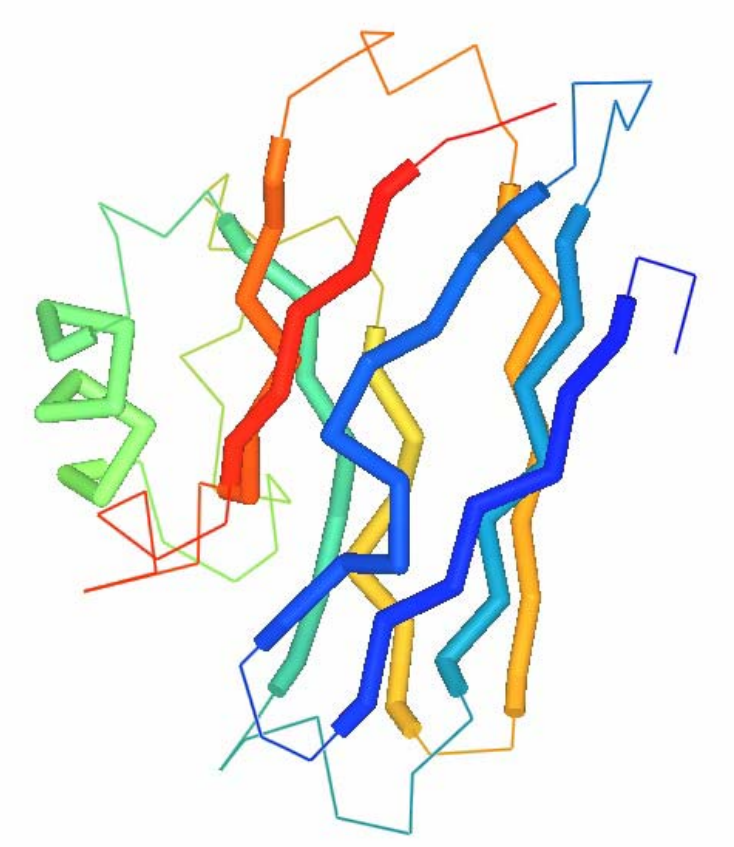
- Przerwy występujące we fragmentach dopasowanych do ciągłych elementów struktury drugorzędowej.

- Niekiedy struktura białka–wzorca i struktura białka modelowanego mogą znacznie się różnić, pomimo stosunkowo dużego podobieństwa sekwencyjnego. Do zbudowania poprawnego modelu może być wówczas wymagana na przykład zmiana kolejności motywów strukturalnych (co zwykle stanowi poważną barierę dla standardowych programów stosowanych w modelowaniu porównawczym).

W pracy zaproponowano dwie strategie poprawiania modeli tworzonych przy pomocy metody przewlekania sekwencji: metodę heurystyczną opartą na analizie i zmianie dopasowań, oraz metodę wykorzystującą model SICHO. Pierwsze podejście można zastosować do poprawiania dopasowań jeszcze przed zbudowaniem modelu białka. Metodę drugą można wykorzystać podczas budowania modelu lub w celu poprawienia struktury już zbudowanego modelu białka.

4.2.1 Poprawianie dopasowań sekwencyjno-strukturalnych przy pomocy metody heurystycznej z wykorzystaniem potencjałów statystycznych

Potencjały statystyczne (potencjał bliskiego zasięgu i potencjał kontaktowy) mogą być wykorzystane do oceny fragmentów dopasowań. Projektując opisaną metodę wykorzystano obserwację, że często najlepiej zachowywanym podczas ewolucji fragmentem struktury białka globularnego jest jego hydrofobowy rdzeń. Rdzeń białka zdefiniowano jako ciągłe elementy struktury drugorzędowej, zawarte w sferze o promieniu równym promieniowi bezwładności struktury białka (przykład przedstawiono na rysunku 15). Jest to dość dobre przybliżenie dla jednodomenowych białek globularnych.



Rysunek 15: Przykład automatycznego odszukiwania rdzenia hydrofobowego w strukturze białka (azuryna, kod PDB 2aza). Pogrubioną linią zaznaczono fragmenty sklasyfikowane jako rdzeń hydrofobowy.

Algorytm poprawiania dopasowania zaprojektowano w taki sposób, aby wyeliminować możliwie wiele błędów występujących w dopasowaniu. Algorytm składa się z dwóch etapów:

- przesuwanie ciągłych fragmentów sekwencji dopasowanej do struktury w taki sposób, aby wyeliminować przerwy w obrębie rdzenia hydrofobowego białka–wzorca
- odrzucanie fragmentów ocenionych przy pomocy potencjałów statystycznych jako źle dopasowane

W wyniku uzyskuje się dopasowanie krótsze, ale o większej zgodności sekwencji modelowanego białka ze strukturą białka–wzorca.

Przeprowadzono test metody dla 31 par sekwencji i struktur białek homologicznych. Początkowe dopasowania zostały wygenerowane przy pomocy metody przewlekania sekwencji [88]. Wyniki zamieszczono w tablicy 1. Średnie odchylenie RMSD Cα (pomiędzy strukturą natywną modelowanego białka i strukturą białka–wzorca, zmierzone w obrębie dopasowania) wyniosło przed zastosowaniem metody 7.29 Å, średnia długość dopasowania: 120. Po poprawieniu dopasowań przy pomocy opisanej metody, średni RMSD spadł do 4.39 Å, a średnia długość dopasowania spadła do 68. Można zadać pytanie, czy dokładne dopasowanie obejmujące tylko niewielki fragment struktury jest lepsze, niż dopasowanie obejmujące większość struktury, ale niedokładne. Zwykle nie jest to prawdą w przypadku programów wykorzystywanych do modelowania porównawczego (MODELLER, COMPOSER), które nie są w stanie zbudować poprawnego modelu, gdy dopasowanie nie obejmuje większości struktury. Natomiast informacja, które części dopasowania są poprawne, jest bardzo cenna dla programów *ab initio*, ponieważ umożliwia odczytanie dokładnych więzów ze struktury białka–wzorca. Warto zauważyć, że poprawne dopasowania nie są w istotny sposób zmieniane przez opisaną metodę. Innym ważnym zastosowaniem opisanej metody może być detekcja rdzenia hydrofobowego w sekwencjach białek o nieznanej strukturze. Zaletą metody jest szybkość jej działania (dopasowania są poprawiane praktycznie w sposób natychmiastowy).

Tablica 1: Rezultaty poprawiania dopasowań sekwencyjno–strukturalnych przy pomocy metody heurystycznej wykorzystującej potencjały statystyczne

Para białek homologicznych	Przed poprawianiem RMSD Cα	Dlugość dopasowania	Po poprawianiu RMSD Cα	Długość dopasowania
1aaj_ 1paz_	6.74	87	4.60	52
1aba_ 1ego_	6.53	80	5.20	42
1bbhA 2ccyA	2.74	123	2.04	55
1bbt1 2plv1	12.55	175	3.57	58
1c2rA 1ycc_	4.35	99	1.53	41
1cauB 1cauA	5.18	163	3.31	63
1cewI 1molA	4.85	76	4.05	47
1chrA 2mnr_	3.50	344	2.66	319
1dxtB 1hbg_	2.74	136	2.26	127
1fxiA 1ubq_	10.94	59	7.50	32
1gp1A 2trxA	11.48	101	8.54	57
1hip_ 2hipA	3.55	68	2.07	52
1hom_ 1lfb_	5.56	55	1.40	44
1hrhA 1rnh_	7.15	108	4.22	53
1isuA 2hipA	6.06	59	2.94	22
1mup_ 1rbp_	5.56	147	2.11	87
1onc_ 7rsa_	3.81	102	3.94	89
1pfc_ 3hlaB	3.84	91	3.31	59
1ten_ 3hhrB	5.60	84	3.85	30
2azaA 1paz_	7.60	81	5.59	53
2hpdA 2cpp_	6.44	392	3.77	217
2mtaC 1ycc_	14.35	96	11.55	45
2pia_ 1fnr_	15.72	255	6.66	98
2pna_ 1shaA	10.69	52	9.10	42
2sarA 9rnt_	6.36	88	4.37	38
2sas_ 2scpA	6.45	160	5.09	88
3cd4_ 2rhe_	7.06	95	2.83	56
3chy_ 4fxn_	6.07	111	3.89	40
3hlaB 2rhe_	10.30	84	4.34	33
5fd1_ 2fxb_	10.95	59	6.42	42
8i1b_ 4fgf_	11.31	108	3.39	69

4.2.2 Poprawianie modeli białek przy pomocy metody *ab initio*

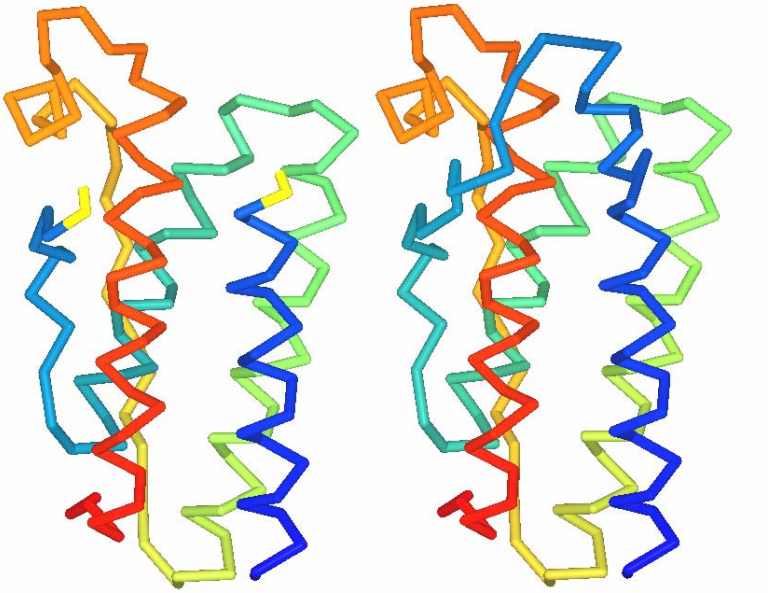
Modele struktur białek stworzone przy pomocy metody przewlekania sekwencji mogą być wykorzystane jako struktury startowe dla symulacji metodą SICHO. Informacje odczytane ze struktury białka–wzorca mogą służyć jako dodatkowe więzy. Metoda SICHO jest w stanie zmodyfikować pewne fragmenty łańcucha białkowego, jednocześnie zachowując geometrię fragmentów o poprawnej strukturze, w konsekwencji poprawiając jakość modelu. Projekt ten opisano szczegółowo w pracy [88] (załącznik 5).

Wykorzystano potencjały bliskiego i dalekiego zasięgu wzbogacone o informacje ewolucyjne, wyprowadzone w opisany wcześniej sposób. Metodę przetestowano na dziewięciu przykładowych parach sekwencja–struktura, wykorzystując dopasowania wygenerowane przez metodę przewlekania sekwencji. W sześciu przypadkach osiągnięto istotną poprawę jakości modeli (odchylenie RMSD Cα od struktury natywnej spadło o 2–3 Å). Uzyskano znaczącą poprawę jakości struktury. Pewne problemy stwarzała automatyczna ocena jakości poprawionych modeli. Kryterium energetyczne nie sprawdza się najlepiej, ponieważ zarówno początkowe modele, jak również poprawione struktury są często stosunkowo odległe od struktury natywnej białka. Zauważono jednak, że w przypadku modeli wysokiej jakości (o niskim odchyleniu od struktury natywnej), poprawne fragmenty dopasowań są stosunkowo mało mobilne podczas obliczeń dynamiki Monte Carlo. Fakt ten wykorzystano do oceny wyników modelowania (praca [88], załącznik 5, strona 608, rysunki 11 i 12). Metoda sprawdza się bardzo dobrze podczas modelowania porównawczego w oparciu o odległe ewolucyjnie białka–wzorce (*distant homology modeling*), a otrzymane wyniki są jakościowo lepsze, niż wyniki uzyskane przy pomocy standardowych narzędzi (MODELLER).

Opisaną metodę zastosowano w praktyce do poprawienia modelu domeny katalitycznej endonukleazy z rodziny GIY–YIG [114].

4.3 Rekonstrukcja brakujących fragmentów białek

W łańcuchach polipeptydowych około 20% białek zawartych w bazie PDB występują przerwy (rysunek 16). Przerwy te odpowiadają fragmentom łańcucha, których struktura nie mogła być otrzymana przy pomocy metod eksperymentalnych. Wynika to bądź z niedoskonałości metody eksperymentalnej (na przykład niewystarczającej jakości dyfraktogramu rentgenowskiego), bądź z mobilności danego fragmentu łańcucha. Często zdarza się, że brakujący fragment ma istotne znaczenie biologiczne, na przykład uczestniczy w procesie wiązania liganda lub w oddziaływaniach z innymi białkami. Dlatego pożądana jest próba znalezienia struktury takich brakujących fragmentów. Projekt opisano w pracy [115] (załącznik 3).



Rysunek 16: Niekompletna struktura białka z bazy PDB (1ax8), brakuje 13 aminokwasów – kolorem żółtym zaznaczono początek i koniec przerwy. Z prawej strony przedstawiono zrekonstruowaną strukturę – wynik zastosowania opisanej metody.

Problem odbudowania brakującego fragmentu struktury białka można przedstawić jako specjalny przypadek budowania modelu białka w oparciu o homologię. Niekompletna struktura jest odpowiednikiem białka–wzorca. W prezentowanej pracy użyto programu opartego na modelu SICHO, wykorzystującego schemat Monte Carlo wy-

miany replik [87]. Strukturę startową zbudowano przy pomocy programu MODELLER. Dzięki temu możliwe było porównanie wyników prezentowanej metody z wynikami zastosowania standardowego narzędzia modelowania homologicznego.

Tablica 2: Wyniki testu metody uzupełniania przerw w strukturach białek.

Nazwa białka	Długość białka	Długość przerwy	RMSD Cα całej struktury	RMSD Cα brakującego fragmentu	RMSD Cα modelu zbudowanego przy pomocy MODELLERa
1ad2_	224	19	2.57	3.84	4.51
1ag4_	103	15	2.19	3.46	2.60
1ahk_	129	16	2.30	3.57	3.20
1ail_	70	18	3.25	4.19	4.10
1bfg_	126	16	2.49	3.18	3.02
1bovA	69	25	4.74	4.35	5.88
1cne_	260	14	2.25	2.47	4.17
1ctf_	68	11	2.04	1.06	2.52
1cyo_	88	19	4.36	4.61	6.73
1fdr_	245	20	2.24	4.64	5.21
1fts_	295	25	2.98	4.30	5.65
1fts_	295	22	1.84	2.20	3.91
1gifA	115	19	5.35	7.78	5.41
1hfh_	120	27	3.92	5.03	5.61
1ife_	91	18	2.33	2.09	3.53
1jer_	110	17	3.80	4.71	6.44
1latA	71	20	5.47	4.77	6.86
1np4_	184	22	4.44	8.23	7.21
1plc_	99	10	1.92	1.84	4.02
1sro_	76	15	3.57	4.29	5.72
1ubq_	76	15	3.32	2.77	5.69
1vhh_	157	18	2.72	3.71	6.11
2azaA	129	33	4.01	5.74	8.20
3cd4_	178	17	2.61	2.51	3.88

Wyniki testu zaprezentowanej metody przedstawiono w tablicy 2. Średnie odchylenie atomów $C\alpha$ zbudowanych modeli od struktury natywnej wyniosło 3.2 Å. Program MODELLER zbudował modele o średnim RMSD Cα równym 5.0 Å (co jest wynikiem gorszym jakościowo).

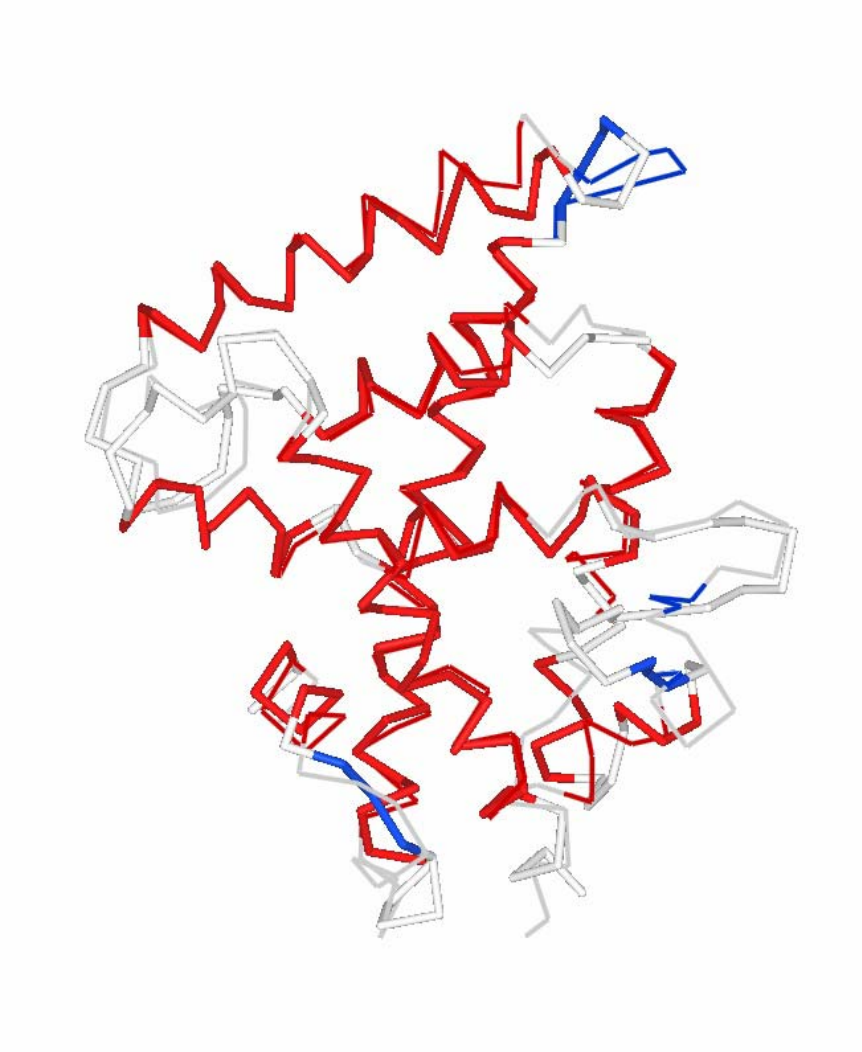
4.3.1 Przykład praktycznego zastosowania metody: budowanie modelu receptora witaminy D

W trakcie pracy nad doktoratem zrealizowano projekt modelowania struktury domeny wiążącej ligand receptora witaminy D. Projekt opisano szczegółowo w pracy [47] (załącznik 4). Wykorzystano standardowe narzędzia modelowania molekularnego (PSI–BLAST [35], MODELLER [44]). Był to projekt bardzo interesujący, gdyż wkrótce po jego zakończeniu opublikowana została rzeczywista (krystalograficzna) struktura receptora, co umożliwiło bezpośrednią konfrontację wyników modelowania teoretycznego z wynikami eksperymentalnymi [116].

Receptor witaminy D należy do rodziny receptorów jądrowych (podobnie jak receptor estrogenu, progesteronu, kwasu retinowego). Domena wiążąca ligand receptora witaminy D jest białkiem o długości 308 aminokwasów, o strukturze drugorzędowej prawie wyłącznie helikalnej.

Program PSI–BLAST zastosowany do przeszukania bazy PDB odnalazł pięć sekwencji białek homologicznych o znanych strukturach, o stopniu identyczności od 20 do 35%. Do zbudowania wstępnego modelu przy użyciu MODELLERA wykorzystano dopasowania sekewencji wyprodukowane przez program PSI–BLAST. Sekwencja receptora witaminy D posiada jednak stosunkowo długi, 40–aminokwasowy fragment nie posiadający odpowiednika w białkach homologicznych. Przesłanki wynikające z badań doświadczalnych wskazują, że może on odgrywać istotną rolę w procesie wiązania się liganda z receptorem. Do odbudowania tego brakującego fragmentu zastosowano program do symulacji uproszczonych modeli białek metodą Monte Carlo. Jakość zbudowanego modelu poprawiono wykorzystując program Sybyl i pole siłowe Tripos [96]. W dalszej kolejności zbudowano kompletny model kompleksu ligand – receptor wykorzystując program FlexiDock z pakietu Sybyl.

Średnie odchylenie pozycji atomów Cα (RMSD) w porównaniu ze strukturą natywną wyniosło 2.4 Å (rysunek 17). Udało się poprawnie przewidzieć 4 z 6 biologicznie istotnych kontaktów pomiędzy ligandem a aminokwasami tworzącymi miejsce aktywne receptora. Dobra jakość zbudowanego modelu i wysoka zgodność z danymi biologicz-



Rysunek 17: Przewidziana struktura domeny wiążącej ligand receptora witaminy D. Cienką linią zaznaczono strukturę rzeczywistą otrzymaną metodami krystalograficznymi [116]. RMSD Cα = 2.4 Å.

nymi potwierdziła przydatność zastosowanej metodologii. Obecnie praca ta jest kontunuowana w celu wykorzystania zbudowanego modelu receptora do oceny aktywności syntetycznych analogów witaminy D [117].

4.4 Odbudowanie pełnoatomowego modelu cząsteczki białka

Aminokwasy w modelu SICHO reprezentowane są w postaci centrów oddziaływań odpowiadających środkom mas grup bocznych. W wielu zastosowaniach niezbędna jest znajomość położeń wszystkich atomów cząsteczki białka.

PULCHRA[8] (PowerfUL CHain Restoration Algorithm) to program pozwalający na skonstruowanie pełnoatomowego modelu cząsteczki białka na podstawie modelu uproszczonego. Informację wejściową stanowią pozycje środków mas grup bocznych aminokwasów. Program odbudowuje współrzędne wszystkich atomów cząsteczki białka.

Pierwszym krokiem metody jest odbudowanie położeń atomów Cα. Przyjęto, że atom Cα leży w płaszczyźnie tworzonej przez trzy kolejne środki mas grup bocznych, w pewnej odległości równowagowej od centralnej grupy bocznej. Odległość ta jest zależna od typu aminokwasu. Następnie pozycje węgli Cα są optymalizowane przy pomocy algorytmu największego spadku gradientu i pola siłowego wykorzystującego proste potencjały harmoniczne. Pole to ma następującą postać:

$$V = \sum_{i=1}^{N-1} k_{C\alpha}(r_{C\alpha}(i) - r^0_{C\alpha})^2 + \sum_{i=1}^{N} k_{SC}(r_{SC}(i) - r^0_{SC})^2 \tag{20}$$

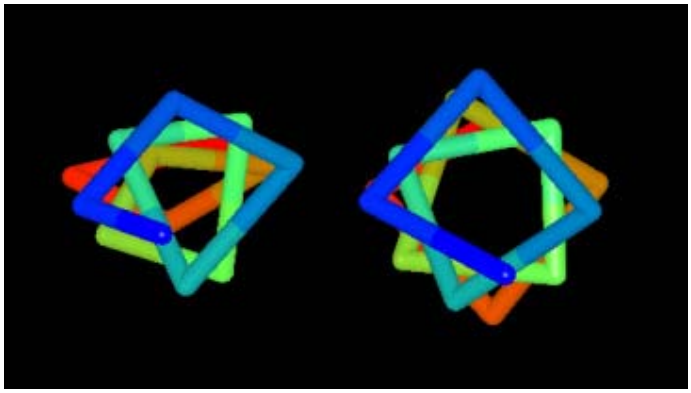
gdzie N jest liczbą aminokwasów w łańcuchu białkowym, $k_{C\alpha}$ i k_{SC} są czynnikami skalującymi, $r_{C\alpha}$ jest odległością pomiędzy dwoma kolejnymi atomami Cα, $r^0_{C\alpha}$ jest odległością równowagową (wynoszącą 3.8 Å), r_{SC} jest odległością pomiędzy atomem Cα a środkiem masy grupy bocznej, r^0_{SC} jest odległością równowagową zależną od rodzaju aminokwasu i lokalnej konformacji łańcucha (od kątów płaskich pomiędzy wektorami łączącymi kolejne atomy Cα).

Kolejnym etapem metody jest odbudowanie pozycji atomów tworzących wiązania peptydowe. Położenia płytek peptydowych ustalane są na podstawie wzajemnej orientacji trzech kolejnych wektorów łączących węgle α łańcucha białkowego.

Następnym krokiem algorytmu jest odbudowanie pozycji atomów łańcuchów bocznych aminokwasów. W programie użyta została biblioteka konformacji rotamerów, za-

[8] łac. *pulchra* – piękna, z założenia ma być to metoda służąca do otrzymywania „pięknych", kompletnych struktur białek.

wiarająca od jednej (dla małych aminokwasów) do 20 (dla argininy) dopuszczalnych konformacji grupy bocznej. Ponieważ położenia środka masy grupy bocznej i węgla α są z góry ustalone, jedynym dopuszczalnym stopniem swobody jest obrót wokół wektora łaczącego atom Cα i środek masy grupy bocznej.



Rysunek 18: Poprawa jakości lokalnej geometrii łańcucha polipeptydowego w wyniku zastosowania metody PULCHRA. Z lewej strony – fragment łańcucha białkowego (pozycje atomów Cα) bezpośrednio po rekonstrukcji z pozycji środków mas grup bocznych, z prawej strony – wynik zastosowania opisanej metody.

Po odbudowaniu atomów grup bocznych następuje etap usuwania konfliktów sterycznych. Tworzona jest lista grup bocznych będących w konflikcie sterycznym z innymi atomami białka (gdy odległość między dwoma niezwiązanymi ze sobą atomami jest mniejsza niż 1.5 Å). Następnie z tej listy losowana jest grupa boczna, która obracana jest o pewien niewielki kąt w taki sposób, by wyeliminować konflikt steryczny. Proces ten powtarzany jest w sposób iteracyjny, aż do chwili usunięcia wszystkich konfliktów.

Średnia dokładność odbudowania położeń wszystkich atomów po konwersji z modelu uproszczonego (opartego na środkach mas grup bocznych) wynosi 1.2 Å. Poprawie ulega jakość struktury drugorzędowej (rysunek 18). Jak wykazały przeprowadzone testy, sposób reprezentacji zredukowanego modelu białka przy pomocy środków mas grup bocznych pozwala na bardziej wierne odtworzenie pełnoatomowej struktury, niż w przypadku modelu opartego na atomach Cα. Przedstawioną metodę rozbudowano i zintegrowano z metodą mechaniki molekularnej opartą na polu siłowym CHARMM [46]. Zostało to dokładnie opisane w pracy [118] (załącznik 6).

4.5 Szybka metoda przeszukiwania bazy struktur białek

Najpopularniejszą obecnie metodą przeszukiwania baz sekwencji białek jest metoda BLAST i jej odmiany (PSI–BLAST, PHI–BLAST). Niekwestionowaną zaletą tej metody jest szybkość działania, przekraczająca o kilka rzędów wielkości klasyczne metody oparte na algorytmie programowania dynamicznego. Odbywa się to jednak kosztem zmniejszenia czułości. Ponadto, w odróżnieniu od metody przewlekania sekwencji, programy BLAST pomijają całkowicie informację pochodzącą z dostępnych struktur białek. Dlatego bardzo interesująca wydaje się możliwość uzupełnienia metody typu BLAST o informacje strukturalne.

Istota algorytmu typu BLAST opiera się na spostrzeżeniu, że dwie podobne do siebie sekwencje białek zawierają często krótkie (trójaminokwasowe) fragmenty, które są prawie identyczne w obu sekwencjach. Jeżeli zawęzi się obszar poszukiwań do obszaru zawierającego wyłącznie takie bardzo podobne do siebie fragmenty, można odrzucić zdecydowaną część przeszukiwanej bazy sekwencji. Kolejny krok algorytmu polega na rozszerzeniu znalezionego krótkiego podciągu w celu oceny wiarygodności domniemanego obszaru podobieństwa dwóch sekwencji. Buduje się w tym celu lokalne dopasowanie startując od środkowego aminokwasu trójaminokwasowego podciągu i używając algorytmu programowania dynamicznego. Do oceny miary tego dopasowania używa się obliczonych wcześniej parametrów rozkładu wartości maksymalnej.

Algorytm BLAST porównuje jednocześnie trójki aminokwasów. Rozszerzenie algorytmu o informacje strukturalne wymagało więc stworzenia zestawu nowych potencjałów statystycznych opartych na trójaminokwasowych fragmentach struktur. Wykorzystano dwa rodzaje potencjałów: potencjał bliskiego zasięgu typu $r_{i,i+2}(A, B, C)$ oraz potencjał hydrofobowy $B_n(A, B, C)$. Potencjał bliskiego zasięgu wyprowadzono zgodnie z opisem w rozdziale 3.2, jedyna różnica polegała na „wymuszeniu" podobieństwa trzech, a nie dwóch aminokwasów. Potencjał hydrofobowy (*burial*) opisuje „preferencje" aminokwasów do kontaktu ze środowiskiem zewnętrznym. Aminokwasy podzielono na trzy klasy, w zależności od liczby kontaktów z innymi aminokwasami (im większa liczba kontaktów z innymi aminokwasami, tym mniejszy kontakt z rozpuszczalnikiem). Potencjał

obliczono w następujący sposób:

$$B_n(A,B,C) = -k_B T \ln\left(\frac{b_n(A,B,C)}{b(A,B,C)}\right) \tag{21}$$

gdzie $b_n(A,B,C)$ jest częstością, z jaką trójka aminokwasów A,B,C występuje w klasie n ($n = 1,2,3$), $b(A,B,C)$ jest częstością występowania trójki aminokwasów A,B,C w bazie danych.

Wiarygodność dopasowania wygenerowanego przy pomocy opisanego algorytmu oceniano na podstawie wartości prawdopodobieństwa znalezienia dopasowania w sposób przypadkowy, zgodnie z opisem przedstawionym w rozdziale 2.5.1. Założono, że dopasowania, dla których obliczone zgodnie z rozkładem wartości maksymalnej prawdopodobieństwo jest mniejsze niż 0.001, są wiarygodne.

Do przetestowania metody wykorzystano zbiór Fischera, zawierający 68 par białek homologicznych (zbiór Fischera przedstawiony jest w pracy [85], załącznik 7, na stronie 139). Metoda wykorzystująca wyłącznie informacje sekwencyjne znalazła 34 białka homologiczne, metoda rozszerzona o potencjały statystyczne – 52 białka homologiczne.

Zaprezentowaną metodę wykorzystano również do wyszukania białek homologicznych w kilkunastu genomach różnych organizmów. Wyniki przedstawiono w tablicy 3. Metoda wzbogacona o potencjały statystyczne jest w stanie zidentyfikować średnio o 7% więcej białek homologicznych, niż metoda czysto sekwencyjna. Wyniki nie są tak spektakularne, jak w przypadku zbioru testowego Fischera. Wynika to stąd, że dla każdego z białek w bazie Fischera istnieje odpowiednia struktura białka homologicznego. Nie jest to prawdą w przypadku rzeczywistych genomów (szczególnie widoczne jest to w przypadku archebakterii, których białka są stosunkowo słabo poznane).

Pod koniec 2001 roku pojawiło się kilka prac poświęconych wykorzystaniu metod przewlekania sekwencji do przeszukiwania całych genomów [119, 120]. Zaprezentowana metoda pozwala osiągnąć porównywalne wyniki. Jej zaletą jest szybkość działania oraz wykorzystanie sprawdzonego i wiarygodnego algorytmu przeszukiwania bazy danych.

Tablica 3: Porównanie wyników szybkiego przeszukiwania bazy danych z wykorzystaniem metody czysto sekwencyjnej oraz metody uzupełnionej o potencjały statystyczne.

Nazwa organizmu	Liczba białek w genomie	Liczba białek znalezionych przez metodę sekwencyjną	Liczba białek znalezionych przez metodę wzbogaconą o informacje strukturalne
Archaea			
Aeropyrum pernix	2694	231	386
Archaeoglobus fulgidus	2409	335	447
Methanococcus jannaschii	1773	338	566
Pyrococcus abyssii	1765	273	318
Procaryota			
Aquifex aeolicus	1522	325	425
Bacillus halodurans	4066	853	1121
Bacillus subtilis	4367	817	1326
Chlamydia muridarum	953	288	421
Deinococcus radiodurans	3116	604	938
Escherichia coli	4290	914	1475
Haemophilus influenzae	1707	432	597
Helicobacter pylori	1575	259	560
Mycoplasma genitalium	479	151	273
Rickettsia prowazekii	834	298	393
Synechocystis sp.	3169	657	847
Ureaplasma urealyticum	611	193	314
Vibrio cholerae	3837	799	1030
Xylella fastidiosa	2766	417	628
Eucaryota			
Arabidopsis thaliana (chromosom 2)	4038	687	906
Caenorhabditis elegans	21965	3252	4050
Oryza sativa (chromosom 10)	268	37	91
Plasmodium falciparum (chromosom 2)	210	38	95
Saccharomyces cerevisiae	6203	1047	1532

4.6 Analiza trajektorii dynamiki Monte Carlo

Trajektoria dynamiki Monte Carlo zawiera struktury mniej lub bardziej odległe od struktury natywnej, a ich odróżnienie na podstawie wartości energii jest trudne. Można pogrupować struktury podobne do siebie i obliczyć strukturę średnią. Okazuje się, że często jest ona bardziej zbliżona do struktury natywnej białka, niż którakolwiek ze struktur zawartych w trajektorii. W pracy zaproponowano sposób generowania modeli białek na podstawie trajektorii dynamiki Monte Carlo przy pomocy metody *distance geometry* (wykorzystując pakiet TINKER [121]). Opis zamieszczono w pracy [85] (załącznik 7).

Metoda *distance geometry* umożliwia odtworzenie współrzędnych kartezjańskich atomów na podstawie zbioru więzów – odległości pomiędzy parami atomów. Metoda ta jest wykorzystywana m.in. w modelowaniu homologicznym [45], do budowania modeli białek na podstawie więzów NMR [122], do konstruowania modeli białek na podstawie krótkich fragmentów białek homologicznych [123]. Algorytm składa się z trzech etapów:

- Sprawdzenie, czy wszystkie więzy spełniają regułę trójkąta (to znaczy, czy dla dowolnych trzech atomów o współrzędnych A, B i C spełniona jest nierówność $\overline{AB} + \overline{BC} \geq \overline{AC}$). Jeżeli nierówność nie jest spełniona, więzy są odpowiednio modyfikowane (jest to proces zwany wygładzaniem więzów, *triangle smoothing*).
- Odbudowanie współrzędnych kartezjańskich (*coordinate embedding*).
- Poprawianie jakości zbudowanego modelu (minimalizując odchylenie końcowego modelu od początkowych więzów, na przykład przy pomocy metody największego spadku gradientu).

Do odbudowania struktury wykorzystano uśrednione więzy (odległości pomiędzy parami węgli α) odczytane z kolejnych struktur trajektorii. Więzy pochodzące ze struktur o niższej energii miały większy udział w sumarycznych więzach. Ponadto więzy zmodyfikowano tak, aby lokalna geometria była zgodna z przewidzianą strukturą drugorzędową. Metoda *distance geometry* jest w stanie wygenerować model białka o odchyleniu RMSD Cα od struktury natywnej średnio o 0.3 Å niższym, niż odchylenie najlepszego modelu w trajektorii. W nielicznych przypadkach poprawa jakości struktury wynosi 1 – 2 Å.

5 Podsumowanie i wnioski końcowe

Tematem pracy doktorskiej było wykorzystanie informacji ewolucyjnych (w postaci podobieństw sekwencyjnych i analogii strukturalnych) w różnych aspektach modelowania struktury przestrzennej białek. W pracy zaproponowano nowe techniki rozszerzające zakres metod modelowania homologicznego.

Opisano sposób wykorzystania podobieństw sekwencyjnych do poprawy specyficzności potencjałów statystycznych. Potencjały wyprowadzono w taki sposób, aby umożliwić ich ścisłą integrację z metodą przewidywania struktury białek SICHO. W ten sposób uzupełniono metodę bezpośredniego przewidywania struktury białek o informacje ewolucyjne. Potencjały te zostały również wykorzystane w metodzie przewlekania sekwencji PROSPECTOR.

Sprawdzono kilka praktycznych zastosowań informacji ewolucyjnych w modelowaniu białek. Potencjały statystyczne zostały użyte do poprawiania dopasowań sekwencyjno – strukturalnych i modeli białek zbudowanych przy pomocy metody przewlekania sekwencji. Przedstawiono sposób rekonstrukcji brakujących fragmentów białek i zastosowano go do stworzenia modelu receptora witaminy D. Przetestowano metodę budowania modeli białek w oparciu o niewielką liczbę więzów. Zaproponowano i zbadano nową metodę przewlekania sekwencji opartą na szybkim algorytmie heurystycznym typu BLAST i wypróbowano ją na kilkunastu genomach różnych organizmów. Przedstawiono również sposób rekonstrukcji pełnoatomowej struktury białka na podstawie uproszczonego modelu opartego na środkach mas grup bocznych aminokwasów. Ponadto stworzono kilka użytecznych programów komputerowych przeznaczonych do analizy i wizualizacji struktur białek. Programy zaprezentowano w dodatku 6.2.

Wyniki przedstawione w pracy wskazują, że wykorzystanie informacji ewolucyjnych w modelowaniu struktur białek stwarza szerokie perspektywy rozwoju. Interesujący wydaje się pomysł bezpośredniego wykorzystania homologii sekwencyjnej podczas symulacji procesu zwijania białek. Fragmenty łańucha polipeptydowego byłyby zastępowane odpowiednimi fragmentami białek homologicznych. Powodowałoby to regularyzację lokalnej geometrii łańcucha, poprawianie kontaktów pomiędzy grupami bocznymi (w przy-

padku zastępowania całych motywów strukturalnych) oraz bardziej efektywne przeszukiwanie przestrzeni konformacyjnej. Informacje ewolucyjne mogą być użyte do poprawy specyficzności wszystkich potencjalów wykorzystywanych w modelu SICHO. Szczególnie interesująca jest możliwość wyprowadzenia potencjału hydrofobowego w oparciu o podobieństwa sekwencyjne. Metoda przewlekania sekwencji (rozwinięcie algorytmu typu BLAST) może być udoskonalona dzięki wykorzystaniu innym potencjałów statystycznych oraz – analogicznie do programu PSI–BLAST – wprowadzeniu iterowanych profili sekwencyjno–strukturalnych.

Przedstawione i rozwinięte w pracy metody modelowania struktury białek w oparciu o podobieństwa sekwencyjne i analogie strukturalne zastosowano z powodzeniem w praktyce. Stanowi to inspirację do dalszych badań.

6 Dodatki

6.1 Wyjaśnienie skrótów i oznaczeń używanych w pracy

BLAST – Basic Local Alignments Search Tool, jedna z najpopularniejszych i najszybszych metod przeszukiwania sekwencyjnych baz danych. [9]

CAFASP – Critical Assessment of Fully Automated Structure Prediction, konkurs automatycznych metod przewidywania struktury białek, jest organizowany równolegle z konkurencją CASP.[10]

CASP – Critical Assessment of Protein Structure Prediction methods, organizowany co dwa lata (od 1994 roku) konkurs metod predykcji struktury białek. Zestaw kilkudziesięciu sekwencji białek o nieznanej strukturze jest publikowany w Internecie, następnie grupy teoretyczne próbują przewidzieć ich struktury. W międzyczasie struktury są również rozwiązywane metodami eksperymentalnymi, po czym następuje porównanie wyników przewidywania struktury z wynikami eksperymentalnymi. W ostatniej edycji konkursu (CASP4, 2000) wzięło udział ponad 150 grup badawczych zajmujących się metodami przewidywania struktur białek [124].[11]

distance geometry – metoda obliczeniowa służąca do obliczania współrzędnych kartezjańskich zbioru punktów na podstawie zbioru odległości pomiędzy nimi; w pracy, wobec braku polskiego odpowiednika, autor postanowił konsekwentnie używać nazwy angielskiej.

FASTA – szybka metoda przeszukiwania baz danych sekwencji białek. [12]

FFF – Fuzzy Functional Forms, sposób opisu miejsca aktywnego białka umożliwiający przewidywanie funkcji biologicznej struktur białek niskiej rozdzielczości [90, 92].

[9] Strona WWW programu BLAST: `http:/www.ncbi.nlm.nih.gov/BLAST/`
[10] Strona WWW konkursu CAFASP: `http://www.cs.bgu.ac.il/~ dfischer/ CAFASP2/`
[11] Strona WWW konkursu CASP: `http://predictioncenter.llnl.gov`
[12] Strona WWW programu FASTA: `http://alpha10.bioch.virginia.edu/ fasta/`

Opiera się na pomiarze odległości pomiędzy trzema arbitralnie wybranymi aminokwasami kluczowymi dla funkcji białka.

PDB – Protein Data Bank, baza struktur białek, największa na świecie baza danych gromadząca informacje o poznanych strukturach białek (i innych biopolimerów), zarządzana obecnie przez San Diego Supercomputing Center [7, 125].[13] Obecnie (pod koniec 2001 roku) baza PDB zawiera struktury ponad 15000 białek. Również: format pliku, który wykorzystywany jest do przechowywania informacji w tej bazie. Pliki PDB identyfikowane są za pomocą czteroznakowego kodu: pierwszy znak określa wersję struktury białka, trzy kolejne litery jednoznacznie identyfikują biąlko (na przykład zapis 3mba oznacza trzecią wersję białka mioglobiny). Do oznaczeń kodów PDB w tej pracy dodano piąty znak, odpowiadający symbolowi łańcucha (w przypadku białek złożonych z jednego łańcucha jest zastępowany znakiem '_').

PDBREF – PDB REFerence set, program służący do tworzenia reprezenatywnego podzbioru bazy PDB (opisany w rozdziale 6.2.2).

RMSD – Root Mean Square Deviation, miara odchylenia położeń atomów pomiędzy dwiema porównywanymi cząsteczkami (zwykle RMSD Cα, to znaczy tylko pomiędzy atomami Cα).

SAL – Structural Alignment, program służący do porównywania ze sobą struktur białek różniących się długością sekwencji (opisany w rozdziale 6.2.3).

SICHO – SIde CHain Only, przybliżony sposób reprezentacji łańcucha białkowego. Aminokwasy reprezentowane są w postaci centrów oddziaływań odpowiadających środkom mas grup bocznych [89]. Rownież: nazwa metody bezpośredniego przewidywania struktury białek na podstawie sekwencji, wykorzystująca siatkową dynamikę Monte Carlo uproszczonego modelu białka.

z-score – ocena standardowa (opisana bliżej w rozdziale 2.5.1).

[13] Strona WWW bazy PDB: `http://www.rcsb.org`

6.2 Opis innych programów stworzonych w trakcie przygotowywania pracy

Podczas realizowania pracy doktorskiej powstało kilka użytecznych programów komputerowych. Wszystkie opisane poniżej programy dostępne są za pośrednictwem stron WWW Pracowni Teorii Biopolimerów Wydziału Chemii UW. [14]

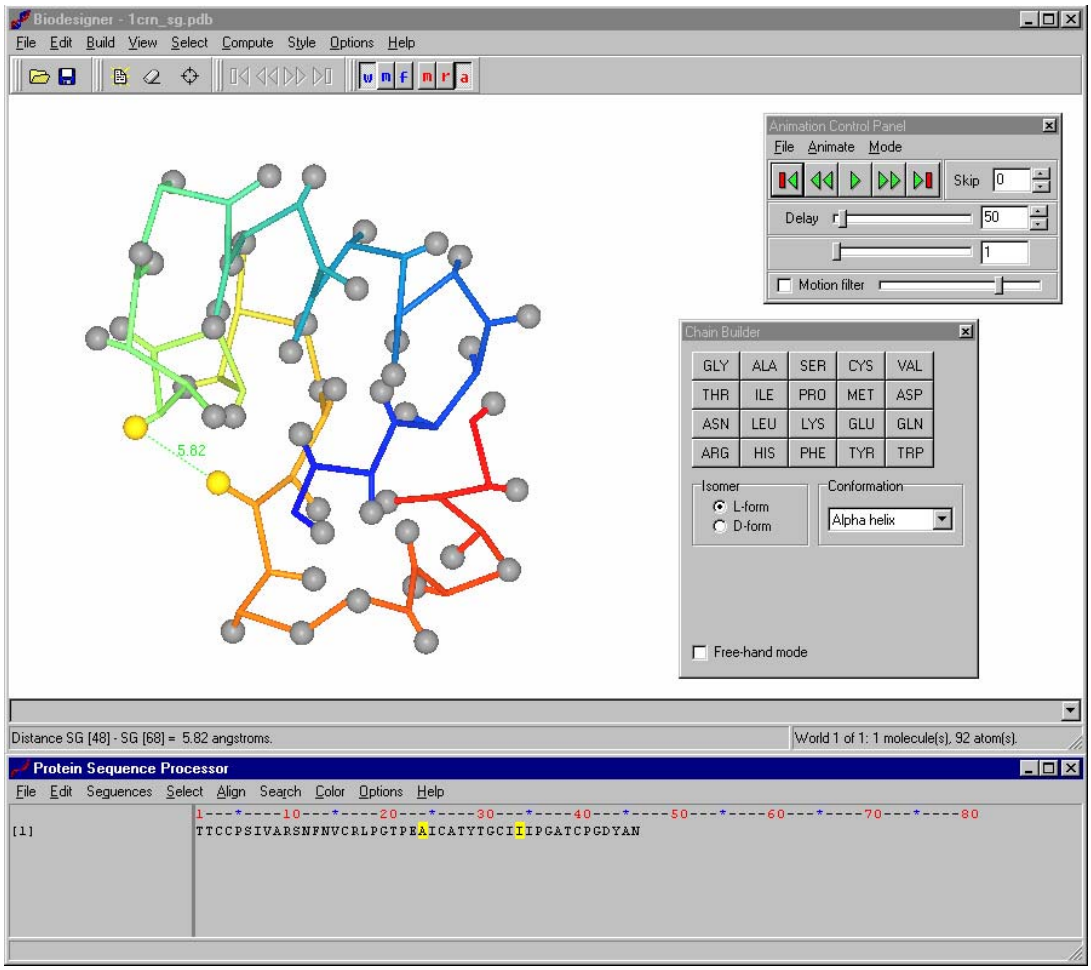
6.2.1 Biodesigner – wizualizacja i modelowanie struktury białek

Biodesigner jest programem przeznaczonym do modelowania i wizualizacji cząsteczek związków chemicznych, w szczególności cząsteczek białek. Podczas projektowania programu położono nacisk na integrację wbudowanych metod analizy i edycji sekwencji białkowych z narzędziami służącymi do wizualizacji struktur. Dzięki temu w poglądowy sposób może zostać zaprezentowany związek pomiędzy strukturą a sekwencją białka (rysunek 19).

Biodesigner automatycznie rozpoznaje wiele popularnych formatów plików (PDB, MONSSTER, Alchemy, Hyperchem, Insight, Sybyl). Umożliwia to łatwe przenoszenie plików pomiędzy różnymi środowiskami obliczeniowymi. Program pozwala zapisywać pliki we własnym formacie BIO oraz w formacie PDB.

Program posiada szerokie możliwości w zakresie wizualizacji cząsteczek chemicznych (rysunek 20). Dostępne są popularne style wyświetlania, takie jak model drutowy, model typu „kulki i pręciki", model czaszowy, model wstęgowy i inne. Ponadto uproszczone modele białek i kwasów nukleinowych mogą być wyświetlane przy pomocy stylu typu „drabinka". Poszczególne style wyświetlania mogą być w łatwy sposób łączone ze sobą. Kolory rysunków odpowiadają różnym własnościom atomów i aminokwasów i mogą być dowolnie modyfikowane. Stworzone rysunki mogą zostać zapisane w formacie bitmapowym (do celów publikacji w Internecie lub interaktywnych prezentacji) lub w formacie wektorowym typu PostScript (do druku). Ponadto możliwe jest tworzenie wysokiej klasy rysunków fotorealistycznych przy pomocy dodatkowych programów wykorzystujących technikę typu *ray-tracing*.

[14] Pracownia Teorii Biopolimerów: `http://biocomp.chem.uw.edu.pl`

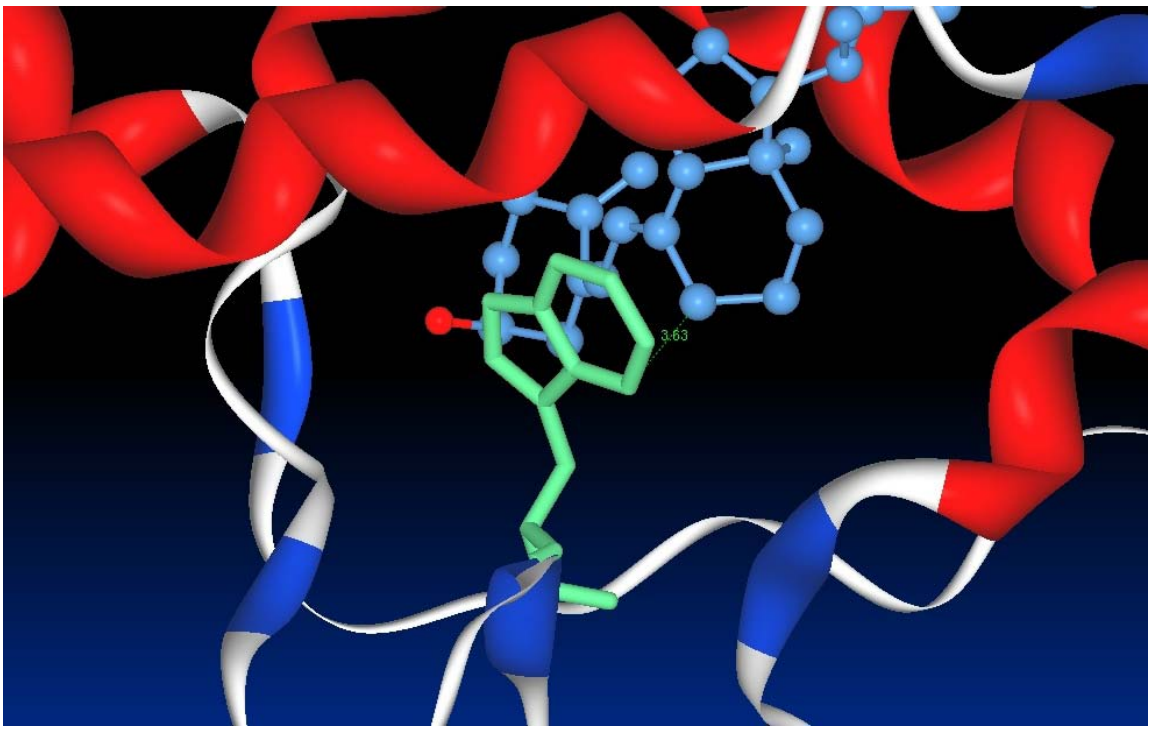


Rysunek 19: Typowy wygląd ekranu podczas pracy z programem Biodesigner

Biodesigner pozwala również na wyświetlanie i analizowanie trajektorii pochodzących na przykład z symulacji dynamiki Monte Carlo. Ruchy o wysokiej częstotliwości mogą zostać odfiltrowane, w ten sposób możliwe jest pokazanie ruchów całych motywów strukturalnych. Animacje mogą zostać następnie zapisane w postaci filmu (pliku graficznego typu AVI).

Program pozwala na wygodne manipulowanie zaznaczonymi fragmentami łańcucha polipeptydowego (na przykład wybranymi elementami struktury drugorzędowej, pojedynczymi rotamerami lub pojedynczymi atomami).

Program wyposażony jest w edytor sekwencji białek. Edytor sprzężony jest z modułem wyświetlającym struktury w taki sposób, że każda zmiana struktury odzwiercie-



Rysunek 20: Przykład wizualizacji cząsteczki białka przy pomocy programu Biodesigner (kompleks witaminy D z jej receptorem, kod PDB 1db1). Łańcuch białka przedstawiono przy pomocy modelu wstęgowego i pokolorowano według typu struktury drugorzędowej. Cząsteczkę liganda przedstawiono przy pomocy modelu „kulki i patyczki" i pokolorowano według typu atomów. Kolorem zielonym zaznaczono grupę boczną tryptofanu ważną dla wiązania liganda i wyświetlono najmniejszą odległość pomiędzy tryptofanem a ligandem.

dlana jest w oknie z wyświetlonymi sekwencjami i *vice versa*. Edytor sekwencji pozwala na modyfikowanie sekwencji i dopasowań, tworzenie dopasowań dwóch i większej liczby sekwencji, translację kodu DNA do sekwencji białkowych, kolorowanie wyświetlonych sekwencji. Moduł rozpoznaje wiele formatów plików sekwencyjnych (m.in. BLAST, CLUSTAL-W, FASTA, PIR, TXT). Dodatkowo program umożliwia przeszukiwanie sekwencyjnych baz danych w poszukiwaniu białek homologicznych przy pomocy programu PSI–BLAST. Po znalezieniu struktury białka homologicznego i stworzeniu dopasowania, program pozwala na zbudowanie modelu białka. Zbudowany model cząsteczki białka może być poprawiony przy pomocy algorytmu typu PULCHRA. Biodesigner pozwala również na ocenę jakości zbudowanego modelu białka.

Program Biodesigner został napisany w języku C++ (ponad 15000 linii kodu). Naj-

nowsza wersja programu (przeznaczona dla systemu Windows) jest dostępna publicznie za pośrednictwem Internetu.[15] Powstała również uproszczona wersja programu (**TraX**) przeznaczona dla systemu Linux. Do chwili obecnej (koniec roku 2001) program Biodesigner został pobrany ze strony WWW ponad 5000 razy.

6.2.2 PDBREF – tworzenie reprezentatywnej bazy danych struktur białek

Baza PDB zawiera obecnie (pod koniec roku 2001) ponad 16000 struktur białek. Jednak informacje zawarte w bazie PDB są często redundantne. Dostępne są struktury tych samych białek otrzymane różnymi metodami, jak również białka pochodzące z odmiennych orgranizmów, ale nie różniące się strukturą w istotny sposób. Liczne struktury w bazie PDB są niekompletne (brakuje niektórych atomów lub całych fragmentów łańcuchów). Dlatego tworzone są bazy danych będące reprezentatywnym podzbiorem bazy PDB, na przykład baza PDBSELECT [126, 109] lub HOMSTRAD [127]. Bazy tego typu są jednak zwykle nieaktualne, z uwagi na fakt, że oryginalna baza PDB jest uaktualniana bardzo często (co tydzień). Ponadto kryteria tworzenia reprezentatywnej bazy struktur białek są arbitralnie przyjęte przez jej autorów i niekiedy nie są optymalne w konkretnych zastosowaniach. Dlatego podczas przygotowywania pracy stworzone zostało narzędzie służące do generowania reprezentatywnego podzbioru bazy PDB: program PDBRFEF.

Program PDBREF (PDB REFerence set) wykorzystuje następujące kryteria tworzenia reprezentatywnego zbioru struktur białek:

- stopień identyczności sekwencji, powyżej którego sekwencje dwóch białek są traktowane jak identyczne, zwykle 75%,
- stosunek długości sekwencji, poniżej którego sekwencje dwóch białek traktowane są jak różne (niezależnie od ich podobieństwa sekwencyjnego), zwykle 50%,
- jakość struktury pozwalająca na zaliczenie białka do reprezentatywnej bazy danych (rodzaj metody eksperymentalnej, przy pomocy której otrzymywano strukturę, oraz czynnik residualny R)[16]

[15] Strona WWW programu Biodesigner: `http://www.pirx.com/biodesigner/`

[16] Czynnik residualny – wartość opisująca jakość struktury krystalograficznej, wyrażająca odchylenie

Budowanie reprezentatywnej bazy danych (pierwsze uruchomienie programu) jest procesem stosunkowo długotrwałym (trwa kilkanaście godzin), kolejne uruchomienia w celu uaktualnienia bazy są wykonywane bardzo szybko. Dodatkowo program PDBREF dzieli pliki z bazy PDB na niezależne łańcuchy białkowe. Posiada też możliwość dzielenia białek wielodomenowych na pojedyncze domeny. Reprezentatywny podzbiór bazy PDB (edycja z 14 stycznia 2001) wygenerowany przy pomocy programu PDBREF zawiera 5041 łańcuchów białkowych (dla stopnia identyczności wynoszącego 75%), lub 2860 łańcuchów białkowych (dla stopnia identyczności wynoszącego 35%). Pierwsza podzbiór może być używany w metodzie przewlekania, gdzie nawet niewielkie różnice pomiędzy strukturami z bazy danych mogą mieć znaczenie. Drugi podzbiór był używany do wyprowadzania potencjałów homologicznych ze względu na wymaganą niewielką redundancję zbioru struktur.

6.2.3 SAL – porównywanie struktur białek

Powszechnie używaną miarą podobieństwa dwóch łańcuchów białkowych jest średnie odchylenie kwadratowe pozycji atomów Cα (RMSD) po uprzednim optymalnym nałożeniu obu struktur [128]:

$$RMSD = \frac{\sqrt{\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{y}_i)^2}}{N} \tag{22}$$

gdzie N jest liczbą porównywanych atomów w każdej z obu cząsteczek, $\mathbf{x}_i$ są współrzędnymi i-tego atomu pierwszej cząsteczki, $\mathbf{y}_i$ są współrzędnymi i-tego atomu drugiej cząsteczki.

RMSD jest miarą pozwalającą określić podobieństwo dwóch łańcuchów białkowych o tej samej długości. Natomiast określanie stopnia podobieństwa dwóch białek różniących się długością łańcucha jest nietrywialne. Obliczenie RMSD takich białek wymaga wcześniejszego utworzenia ich dopasowania w taki sposób, aby liczby porównywanych atomów były identyczne.

przewidzianej struktury od struktury rzeczywistej, waha się od 0.0 (idealne dopasowanie) do ok. 0.6 (struktura losowa). Wartość 0.2 i niższa odpowiada strukturze krystalograficznej dobrej jakości.

Istnieje wiele metod porównywania struktur białek. Do najbardziej znanych należą:

- metoda CE [129] badająca podobieństwo odległości i kątów w dwóch strukturach
- metoda DALI [130, 131] oparta na analizie podobieństwa map kontaktów dwóch białek
- metoda SSAP [132] wykorzystuje algorytm podwójnego programowania dynamicznego
- metoda VAST [133] (*Vector Alignment Search Tool*) odszukująca niewielkie fragmenty obu struktur o wysokim stopniu podobieństwa

Metody te dostępne są publicznie w postaci serwisów WWW. W trakcie realizowania pracy doktorskiej stworzony został progam SAL (Structural ALignment) umożliwiający ocenę podobieństwa strukturalnego dwóch łańcuchów białkowych. W programie wykorzystano zmodyfikowany algorytm iterowanego programowania dynamicznego [134, 135] polegający na wielokrotnym powtarzaniu następujących kroków:

- tworzona jest dwuwymiarowa macierz odległości pomiędzy wszystkimi parami atomów węgli Cα obu cząsteczek białek
- przy pomocy algorytmu programowania dynamicznego (opisanego w rozdziale 2.4.2) zastosowanego do zbudowanej w poprzednim kroku macierzy tworzone jest dopasowanie obu struktur (*structural alignment*)
- struktury są następnie obracane w taki sposób, aby optymalnie nałożyć na siebie dopasowane w porzednim kroku fragmenty

Kroki te powtarza się aż do momentu zbiegnięcia algorytmu (gdy nowo wygenerowana orientacja obu łańcuchów nie różni się od poprzednio wygenerowanej) lub do przekroczenia założonego z góry limitu liczby kroków algorytmu. Modyfikacje w stosunku do oryginalnie opisanej metody polegały na starannym dobraniu początkowej orientacji

obu łańcuchów (próbie wstępnej optymalizacji) oraz na zmiennych parametrach kar za wprowadzenie przerw w dopasowaniu.

Algorytm zastosowany w programie SAL wykorzystuje podejście heurystyczne, nie dające gwarancji znalezienia optymalnego dopasowania. Dopasowanie zależy od założonych arbitralnie parametrów kar za wprowadzenie przerw. Dlatego bardzo ważna jest statystyczna ocena podobieństwa dwóch struktur. Opracowując program SAL zbadano parametry rozkładu wartości RMSD podczas porównywania struktur białek o różnych długościach. Dzięki temu program podaje też wartość oceny standardowej (*z-score*). Program SAL był używany m.in. do oceny jakości modeli białek w konkurencji CAFASP2.

Bibliografia

[1] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, pp. 595–602, 1996.

[2] P. Rotkiewicz, "Biodesigner - visual homology modeling program," *J. Mol. Graphics*, p. wysłano do redakcji, 2001.

[3] S. Rastan and L. Beeley, "Functional genomics: Going forwards from databases," *Curr. Opin. Genet. Devel.*, vol. 7, pp. 777–783, 1997.

[4] C. Dennis, R. Gallagher, and P. Campbell, "Everyone's genome," *Nature*, vol. 409, p. 813, 2001.

[5] C. Branden and J. Tooze, *Introduction to Protein Structure*. Garland Publishing, New York, 1991.

[6] T. E. Creighton, *Proteins. Structures and Molecular Properties*. W.H.Freeman and Co, New York, 1993.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalow, and P. E. Bourne, "The Protein Data Bank," *Nuc. Acids Res.*, vol. 28, pp. 235–242, 2000.

[8] G. Wider, "Structure Determination of Biological Macromolecules in Solution Using NMR Spectroscopy," *BioTechniques*, vol. 29, pp. 1278–1294, 2000.

[9] J. T. Finch and A. Klug, "The structure of viruses of the papilloma polyoma type 3. Structure of rabbit papilloma virus, with an appendix on the topography of contrast in negative staining for electron microsocopy," *J. Mol. Biol.*, vol. 1, pp. 1–13, 1965.

[10] R. Kreisberg, V. Buchner, and D. Arad, "Paired natural cysteine mutation mapping: Aid to constraining models of protein tertiary structure," *Prot. Sci.*, vol. 4, pp. 2405–2410, 1995.

[11] A. Fiser, R. Sanchez, F. Melo, and A. Sali, "Comparative protein structure modeling," `http://guitar.rockefeller.edu/ andras/`, 2000.

[12] A. Sali, "100,000 protein structures for the biologist," *Nature Struct. Biol.*, vol. 5, pp. 1029–1032, 1998.

[13] J. Skolnick, A. Kolinski, P. Rotkiewicz, and B. Ilkowski, "Prediction of protein structure and function on a genomic scale," *Abstr. Pap. Am. Chem. Soc.*, vol. 221, pp. 58–COMP Part 1, 2001.

[14] Z. Li and H. Scheraga, "Monte Carlo Minimization Approach to the Multiple Minima Problem in Protein Folding," *Proc. Natl. Acad. Sci. USA*, vol. 84, pp. 6611–6615, 1987.

[15] R. Rodriguez and G. Vriend, "Professional gambling (sometimes incorrectly called homology modeling)," `http://www.sander.embl-heidelberg.de/future/articles/text/gambling.html`, 1997.

[16] R. F. Service, "Amino acid alchemy trasmutes sheets to coils," *Science*, vol. 277, pp. 179–180, 1997.

[17] R. Lewin, "When does homology mean something else?," *Science*, vol. 237, pp. 1570–1573, 1987.

[18] P. Rotkiewicz, "Komputerowe modelowanie białek z wykorzystaniem homologii sekwencyjnej," Master's thesis, Uniwersytet Warszawski, Czerwiec 1998. Praca magisterska.

[19] Z. T. Zhang, "Relations of the numbers of protein sequences, families, and folds," *Prot. Eng.*, vol. 10, pp. 757–761, 1997.

[20] C. Chotia, "One thousand families for the molecular biologist," *Nature*, vol. 360, pp. 543–544, 1992.

[21] J. Liu and B. Rost, "Comparing function and structure between entire proteoms," *Prot. Sci.*, vol. 10, pp. 1970–1979, 2001.

[22] R. Sanchez and A. Sali, "Advances in comparative protein–structure modeling," *Curr. Opin. Biotech.*, vol. 6, pp. 437–451, 1997.

[23] L. Jaroszewski, *Hybrydowe metody przewidywania struktury białek globularnych.* PhD thesis, Uniwersytet Warszawski, 1998. Praca doktorska.

[24] S. B. Needleman and C. D. Wunsh, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, 1970.

[25] A. Aho, J. Hopcroft, and J. Ullman, *Projektowanie i analiza algorytmów komputerowych.* PWN, 1985.

[26] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences," *J. Mol. Biol.*, vol. 162, pp. 705–708, 1982.

[27] T. F. Smith and M. S. Waterman, "Identification of Common Molecular Subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.

[28] J. Skolnick and D. Kihara, "Defrosting the frozen approximation: PROSPECTOR: a new approach to threading," *Proteins*, p. w druku, 2001.

[29] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices form protein blocks," *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 10915–10919, 1992.

[30] M. O. Dayhoff, R. M. Schwarz, and B. C. Orcutt, "A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results," in *Atlas of Protein Sequence and Structure*, pp. 353–358, National Biomedical Research Foundation, Washington D.C., 1979.

[31] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database," *Science*, vol. 56, pp. 1443–1445, 1992.

[32] C. Notredame and D. G. Higgins, "SAGA – sequence alignment by genetic algorithm," *Nucl. Acid. Res.*, vol. 24, pp. 1515–1524, 1996.

[33] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Pub. Co., 1989.

[34] S. F. Altschul, W. Gish, W. Miller, and E. W. Myers, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[35] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucl. Acid. Res.*, vol. 25, pp. 3389–3402, 1997.

[36] D. Lipman and W. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, pp. 1435–1441, 1985.

[37] D. Lipman and W. Pearson, "Improved tools for biological sequence comparison," *Proc. Natl. Acad. Sci. USA*, vol. 85, pp. 2444–2448, 1988.

[38] J. L. Thorne, H. Kishino, and J. Felsenstein, "An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences," *J. Mol. Evol.*, vol. 33, pp. 114–124, 1991.

[39] W. R. Taylor, "Multiple sequence alignment by a pairwise algorithm," *Comput. Appl. Biosci.*, vol. 3, pp. 81–87, 1987.

[40] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.

[41] M. A. McClure, T. K. Vasi, and W. M. Fitch, "Comparative Analysis of Multiple Protein-Sequence Alignment Methods," *Mol. Biol. Evol.*, vol. 11, no. 4, pp. 571–592, 1994.

[42] P. Baldi, A. Chauvin, T. Hunkapiller, and H. A. McClure, "Hidden Markov Models of biological primary sequence information," *Proc. Natl. Acad. Sci. USA*, vol. 91, pp. 1059–1063, 1994.

[43] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Protein Analysis: Detection of Distantly Related Proteins," *Proc. Natl. Acad. Sci. USA*, vol. 84, pp. 4355–4358, 1987.

[44] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *J. Mol. Biol.*, vol. 234, pp. 779–815, 1993.

[45] A. Aszodi and W. R. Tylor, "Homology modeling by distance geometry," *Fold. Des.*, vol. 1, pp. 325–334, 1996.

[46] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy minimalization and dynamics calculations," *J. Comp. Chem.*, vol. 4, pp. 187–217, 1983.

[47] P. Rotkiewicz, W. Sicińska, A. Kolinski, and H. F. D. Luca, "Model of Three-dimensional Structure of Vitamin D Receptor and its Binding Mechanism with 1α-dihydroxyvitamin D_3," *Proteins*, vol. 44, pp. 188–199, 2001.

[48] B. N. Srinivasan and T. L. Blundell, "An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure," *Prot. Eng.*, vol. 6, pp. 501–512, 1993.

[49] M. C. Peitsch, "PROMOD and SWISS-MODEL – Internet-based tools for automated comparative protein modeling," *Biochem. Soc. Trans.*, vol. 24, pp. 274–279, 1996.

[50] D. T. Jones and J. M. Thornton, "Protein Fold Recognition," *Journal of Computer-Aided Molecular Design*, vol. 7, pp. 439–456, 1993.

[51] J. U. Bowie, R. Luthy, and D. Eisenberg, "A Method to Identify Protein Sequences that Fold into a Known Three-dimensional Structure," *Science*, vol. 253, pp. 164–170, 1991.

[52] J. U. Bowie and D. Eisenberg, "Inverted protein structure prediction," *Curr. Op. Struct. Biol.*, vol. 3, pp. 437–444, 1993.

[53] M. J. Sippl, P. Lackner, F. S. Domingues, A. Prlic, R. Malik, A. Andreeva, and M. Wiederstein, "Assessment of the CASP4 Fold Recognition Category," *Proteins*, p. w druku, 2002.

[54] L. A. Kelley, R. M. MacCallum, and M. J. E. Sternberg, "Enhanced Genome Annotation using Structural Profiles in the Program 3D–PSSM," *J. Mol. Biol.*, vol. 299, pp. 501–522, 2000.

[55] D. T. Jones, "GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences," *J. Mol. Biol.*, vol. 287, pp. 797–815, 1999.

[56] L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik, "Fold and Function Assignment System," *Prot. Sci.*, vol. 9, pp. 232–241, 2000.

[57] L. Rychlewski and A. Godzik, "FFAS," `http://bioinformatics.burnham-inst.org/FFAS`, 2001.

[58] R. H. Lathrop, "The Protein Threading Problem With Sequence Amino Acid Interaction Preferences is NP-Complete," *Protein Eng.*, vol. 7, no. 9, pp. 1059–1068, 1994.

[59] A. Godzik and J. Skolnick, "Topology Fingerprint Approach to the Inverse Protein Folding Problem," *J. Mol. Biol.*, vol. 226, pp. 000–000, 1992.

[60] B. Zhang, L. Jaroszewski, L. Rychlewski, and A. Godzik, "Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies," *Fold. Des.*, vol. 2, pp. 307–317, 1997.

[61] S. H. Bryant and S. F. Altschul, "Statistics of Sequence-structure Threading," *Curr. Op. Struct. Biol.*, vol. 5, pp. 236–244, 1995.

[62] M. S. Waterman and M. Vingron, "Sequence Comparizon Significance and Poisson Approximation," *Stat. Sci.*, vol. 9, pp. 367–381, 1994.

[63] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc. Natl. Acad. Sci. USA*, vol. 87, pp. 2264–2268, 1990.

[64] S. Karlin and S. F. Altschul, "Applications and statistics for multiple high-scoring segments in molecular sequences," *Proc. Natl. Acad. Sci. USA*, vol. 90, pp. 5873–5877, 1993.

[65] L. Rychlewski, B. Zhang, and A. Godzik, "Fold and function predictions for M*ycoplasma genitalium* proteins," *Fold. Des.*, vol. 3, pp. 229–238, 1998.

[66] R. Sanchez and A. Sali, "Large scale protein structure modeling of the S*accharomyces cerevisiae* genome," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 13597–13602, 1998.

[67] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–238, 1973.

[68] M. Levitt, M. Hirshberg, R. Sharon, and V. Dagget, "Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution," *Comp. Phys. Comm.*, vol. 91, pp. 215–231, 1995.

[69] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, 1995.

[70] E. S. Huang, R. Samudrala, and B. H. Park, "Scoring functions for ab initio folding," in *Predicting Protein Structure: Methods and Protocols*, Humana Press, 2000.

[71] M. Kardar, "Which Came First, Protein Sequence or Structure?," *Science*, vol. 273, p. 610, 1996.

[72] Y. Duan and P. A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science*, vol. 282, pp. 740–744, 1998.

[73] Y. Duan and P. A. Kollman, "Computational protein folding: From lattice to all-atom," *IBM Systems Journal*, vol. 40, pp. 297–309, 2001.

[74] R. Unger and J. Moult, "Genetic Algorithms for Protein Folding Simulations," *J. Mol. Biol.*, vol. 231, pp. 75–81, 1993.

[75] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions," *J. Mol. Biol.*, vol. 268, pp. 209–225, 1997.

[76] A. Kolinski and J. Skolnick, *Lattice models of protein folding, dynamics and thermodynamics*. R. G. Landes, 1996.

[77] A. Kolinski and J. Skolnick, "Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme," *Proteins*, vol. 18, pp. 338–352, 1994.

[78] A. Kolinski and J. Skolnick, "Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin," *Proteins*, vol. 18, pp. 353–366, 1994.

[79] J. Skolnick, A. Kolinski, and A. Ortiz, "MONSSTER: A method for folding globular proteins with a small number of distance restraints," *J. Mol. Biol.*, vol. 265, pp. 217–241, 1997.

[80] A. Kolinski, P. Rotkiewicz, and J. Skolnick, "Application of high coordination lattice model in protein structure prediction," in *Monte Carlo Approach to Biopolymers and Protein Folding* (P. Grassberger, G. T. Barkema, and W. Nadler, eds.), pp. 100–130, World Scientific, Singapore/London, 1998.

[81] A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, "Protein Folding: Flexible Lattice Models," *Progress in Theoretical Physics Suppl.*, vol. 138, pp. 292–300, 1999.

[82] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, "*Ab initio* folding of proteins using restraints derived from evolutionary information," *Proteins Suppl. (CASP3 Proceedings)*, vol. 3, pp. 177–185, 1999.

[83] A. M. Lesk, L. L. Conte, and T. J. P. Hubbard, "Assessment of Novel Fold Targets in CASP4: Predictions of Three-dimensional Structures, Secondary Structures, and Interresidue Contacts," *Proteins*, p. w druku, 2002.

[84] J. Skolnick, A. Kolinski, D. Kihara, M. R. Betancourt, P. Rotkiewicz, and M. Boniecki, "Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement," *Proteins*, p. w druku, 2002.

[85] A. Kolinski, M. R. Betancourt, D. Kihara, P. Rotkiewicz, and J. Skolnick, "Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement," *Proteins*, vol. 44, pp. 133–149, 2001.

[86] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1953.

[87] D. Gront, A. Kolinski, and J. Skolnick, "Comparison of three Monte Carlo search strategies for a proteinlike homopolymer model: folding thermodynamics and identification of low energy structures," *J. Chem. Phys.*, vol. 113, pp. 5065–5071, 2000.

[88] A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, "A Method for the Improvement of Threading-Based Protein Models," *Proteins*, vol. 37, pp. 592–610, 1999.

[89] A. Kolinski, L. Jaroszewski, P. Rotkiewicz, and J. Skolnick, "An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass," *J. Phys. Chem.*, vol. 102, pp. 4628–4637, 1998.

[90] J. S. Fetrow, A. Godzik, and J. Skolnick, "Functional analysis of the E*scherichia coli* genome using the sequence-structure-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxireductase activity," *J. Mol. Biol.*, vol. 282, pp. 703–711, 1998.

[91] J. Skolnick, J. S. Fetrow, and A. Kolinski, "Structural genomics and its importance to gene function analysis," *Nature Biotechnology*, vol. 18, pp. 283–287, 2000.

[92] J. S. Fetrow and J. Skolnick, "Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1-ribonucleases," *J. Mol. Biol.*, vol. 281, pp. 949–968, 1998.

[93] D. M. Lorber and B. M. Shoichet, "Flexible ligand docking using conformational ensembles," *Prot. Sci.*, vol. 7, pp. 938–950, 1998.

[94] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function," *J. Comp. Chem.*, vol. 19, pp. 1639–1662, 1998.

[95] C. M. Oshiro and D. I. Kuntz, "Flexible ligand docking using a genetic alorithm," *J. Comput.-Aided Mol. Design*, vol. 9, pp. 113–130, 1995.

[96] "SYBYL Version 6.5," *Tripos Inc.*, 2000.

[97] M. Wojciechowski and J. Skolnick, "Docking of small ligands to low-resolution and theoretically predicted receptor structures," *J. Comp. Chem.*, vol. 23, pp. 189–197, 2002.

[98] I. A. Vakser, "Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex," *Proteins Suppl.*, vol. 1, pp. 226–230, 1997.

[99] I. A. Vakser, O. G. Matar, and C. F. Lam, "A systematic study of low-resolution recognition in protein-protein complexes," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 8477–8482, 1999.

[100] E. J. Gardiner, P. Willett, and P. J. Artymiuk, "Protein Docking Using a Genetic Algorithm," *Proteins*, vol. 43, pp. 44–56, 2001.

[101] M. Vieth, A. Kolinski, C. L. Brooks, and J. Skolnick, "Prediction of the Folding Pathways and Structure of the GCN4 Leucine Zipper," *J. Mol. Biol.*, vol. 237, pp. 361–367, 1994.

[102] M. R. Betancourt and J. Skolnick, "Finding the needle in a haystack: Educing native folds from ambiguous ab initio protein structure prediction," *J. Comput. Chem.*, vol. 22, pp. 339–353, 2000.

[103] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation," *Macromolecules*, vol. 18, pp. 534Ű–552, 1985.

[104] M. J. Sippl, "Knowledge-based Potentials for Proteins," *Curr. Opin. Stuct. Biol.*, vol. 5, pp. 229–235, 1995.

[105] B. A. Reva, A. V. Finkelstein, M. F. Sanner, and A. J. Olson, "Accurate Mean-Force Pairwise-Residue Potentials for Discrimination of Protein Folds," *Pacific Symposium on Biocomputing*, vol. 2, pp. 373–384, 1997.

[106] L. A. Mirny and E. I. Shaknovich, "How to Derive a Protein Folding Potential? A New Approach to an Old Problem," *J. Mol. Biol.*, vol. 264, pp. 1164–1179, 1996.

[107] R. I. Dima, J. R. Banavar, and A. Maritan, "Scoring functions in protein folding and design," *Protein Sci.*, vol. 9, pp. 812–819, 2000.

[108] V. N. Maiorov and G. M. Crippen, "Contact potential that recognizes the correct folding of globular proteins," *J. Mol. Biol.*, vol. 227, pp. 876–888, 1992.

[109] U. Hobohm and C. Sander, "Enlarged representative of protein structures," *Protein Sci.*, vol. 3, pp. 522–524, 1994.

[110] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, "Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?," *Protein Sci.*, vol. 6, pp. 676–688, 1997.

[111] A. Kolinski, A. Godzik, and J. Skolnick, "Contact map," in *Encyclopedia of Computational Biology*, pp. 567–571, T. Creighton Ed. John Wiley & Sons, 1999.

[112] J. Skolnick, , A. Kolinski, and A. Ortiz, "Derivation of Protein-Specific Pair Potentials Based on Weak Sequence Fragment Similarity," *Proteins*, vol. 38, pp. 3–16, 2000.

[113] A. Kolinski and J. Skolnick, "Assembly of protein structure from sparse experimental data: An efficient Monte Carlo model," *Proteins*, vol. 32, pp. 475–489, 1998.

[114] J. Bujnicki, P. Rotkiewicz, A. Kolinski, and L. Rychlewski, "Three-dimensional fold prediction and *ab initio* modeling of the I-*tev*I homing endonuclease catalytic domain, a GIY-YIG superfamily member," *Protein Eng.*, p. w druku, 2000.

[115] A. Kolinski, P. Rotkiewicz, and J. Skolnick, "Structure of Proteins: New Approach to Molecular Modeling," *Polish J. Chem.*, vol. 75, pp. 587–599, 2001.

[116] N. Rochel, J. M. Wurtz, A. Mitschler, B. Klaholz, and D. Moras, "The structure of vitamin D receptor," *Molecular Cell*, vol. 5, pp. 173–179, 2000.

[117] R. R. Sicinski, A. Kolinski, P. Rotkiewicz, W. Sicinska, J. M. Prahl, C. M. Smith, and H. F. D. Luca, "2-Ethyl and 2-Ethylidene Analogs of 1α,25-Dihydroxy-19-norvitamin D3: Synthesis, Conformational Analysis, Biological Activities, and Docking to the Modeled rVDR Ligand Binding Domain," *J. Med. Chem.*, p. wysłano do redakcji, 2002.

[118] M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, and C. L. Brooks, "Accurate Reconstruction of All-Atom Protein Representations from Side Chain Based Low Resolution Models," *Proteins*, vol. 41, pp. 86–97, 2000.

[119] S. Dietmann and L. Holm, "Identification of homology in protein structure classification," *Nature Struct. Biol.*, vol. 8, pp. 953–957, 2001.

[120] L. Salwiński and D. Eisenberg, "Motif-based fold assignment," *Prot. Sci.*, vol. 10, pp. 2460–2469, 2001.

[121] R. V. Pappu, R. K. Harn, and J. W. Ponder, "Analysis and Application of Potential Enegry Smoothing for Global Optimization," *J. Phys. Chem. B*, vol. 102, pp. 9725–9742, 1998.

[122] W. Braun, "Distance geometry and related methods for protein structure determination from NMR data," *Q. Rev. Biophys.*, vol. 19, pp. 115–157, 1987.

[123] E. S. Huang, R. Samudrala, and J. W. Ponder, "*Ab initio* fold prediction of small helical proteins using distance geometry and knowledge–based scoring functions," *J. Mol. Biol.*, vol. 290, pp. 267–281, 1999.

[124] E. E. Lattman, "CASP4," *Proteins*, vol. 44, p. 399, 2001.

[125] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures," *J. Mol. Biol.*, vol. 112, pp. 535–542, 1977.

[126] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, "Selection of representative protein data sets," *Protein Sci.*, vol. 1, pp. 409–417, 1992.

[127] K. Mizuguchi, T. L. Blundell, and J. P. Overington, "HOMSTRAD: a database of protein structure alignments for homologous families," *Prot. Sci.*, vol. 7, pp. 2469–2471, 1998.

[128] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-I 9, no. 5, pp. 698–700, 1987.

[129] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Eng.*, vol. 11, pp. 739–747, 1998.

[130] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *J. Mol. Biol.*, vol. 233, pp. 123–138, 1993.

[131] L. Holm and C. Sander, "DALI/FSSP classification of 3D protein folds," *Nucl. Acid Res.*, vol. 25, pp. 231–234, 1995.

[132] C. A. Orengo and W. R. Taylor, "SSAP: Sequential Structure Alignment Program for Protein Structure Comparison," *Methods in Enzymology*, vol. 266, pp. 617–635, 1996.

[133] J.-F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," *Curr. Op. Struct. Biol.*, vol. 6, pp. 377–385, 1996.

[134] M. Gerstein and M. Levitt, "Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures," *Proc. of ISMB-96*, vol. 1, pp. 59–67, 1996.

[135] M. Gerstein and M. Levitt, "Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins," *Prot. Sci.*, vol. 7, pp. 445–456, 1998.

8 Spis publikacji należących do pracy

1. Andrzej Kolinski, Piotr Rotkiewicz, Jeffrey Skolnick, *Application of a high coordination lattice model in protein structure prediction*, w *Proceedings of Workshop on Monte Carlo Approach to Biopolymers and Protein Folding*, P. Grassberger, G.T. Barkema and W.Nadler, Eds., World Scientific Publishing Co., Singapore, 100–130 (1998)

   Udział autora polegał na opracowaniu potencjałów statystycznych bliskiego zasięgu, przeprowadzeniu części symulacji modelowania białek z wykorzystaniem niewielkiej liczby więzów, oraz na analizie wyników symulacji.

2. Andrzej Kolinski, Piotr Rotkiewicz, Bartosz Ilkowski, Jeffrey Skolnick, *Protein Folding: Flexible Lattice Models*, Progress in Theoretical Physics Suppl., **138**, 292–300 (1998)

   Autor zbudował modele kilku białek w oparciu o struktury białek homologicznych i dopasowania pochodzące z metody przewlekania sekwencji, wykorzystując standardowy protokół modelowania homologicznego (program MODELLER) i metodę SICHO. Przeprowadzono analizę i porównanie wyników otrzymanych obiema metodami.

3. Andrzej Kolinski, Piotr Rotkiewicz, Jeffrey Skolnick, *Structure of Proteins: New Approach to Molecular Modeling*, Polish J. Chem., **75**, 587–599 (2001)

   Autor przeprowadził test metody rekonstrukcji brakujących fragmentów białek (symulacje Monte Carlo i analizę wyników), a następnie zastosował metodę dla kilku rzeczywistych przypadków niekompletnych białek z bazy PDB.

4. Piotr Rotkiewicz, Wanda Sicińska, Andrzej Kolinski, Hector F. De Luca, *Model of Three-dimensional Structure of Vitamin D Receptor and its Binding Mechanism with 1α-dihydroxyvitamin D_3*, Proteins, **44**, 188–199 (2001)

   Autor zbudował model domeny wiążącej ligand receptora witaminy D (wykorzystując programy BLAST, MODELLER i metodę SICHO) i przeprowadził oblicze-

nia konformacji kompleksu ligand–receptor używając programu dokującego FlexiDock.

5. Andrzej Kolinski, Piotr Rotkiewicz, Bartosz Ilkowski, Jeffrey Skolnick, *A Method for the Improvement of Threading-Based Protein Models*, Proteins, **37**, 595–610 (1999)

   Udział autora polegał na opracowaniu programu tłumaczącego wyniki metody przewlekania sekwencji (dopasowania sekwencji do struktur białek homologicznych) do postaci więzów i potencjałów statystycznych używanych w programie dynamiki siatkowej białek metodą Monte Carlo. Autor przeprowadził również analizę wyników symulacji.

6. Michael Feig, Piotr Rotkiewicz, Andrzej Kolinski, Jeffrey Skolnick, Charles L. Brooks III, *Accurate Reconstruction of All-Atom Protein Representations from Side Chain Based Low Resolution Models*, Proteins, **41**, 86–97 (2000)

   Autor zaproponował szybką metodę poprawiania pozycji atomów Cα, odbudowywania pozycji płytek peptydowych i położeń grup bocznych na podstawie modelu uproszczonego.

7. Andrzej Kolinski, Marcos R. Betancourt, Daisuke Kihara, Piotr Rotkiewicz, Jeffrey Skolnick, *Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement*, Proteins, **44**, 133–139 (2001)

   Autor napisał część kodu programu służącego do przewlekania sekwencji, zaimplementował sposób poprawiania jakości struktury białek z wykorzystaniem metody *distance geometry*, opracował potencjały bliskiego i dalekiego zasięgu używane w metodzie przewlekania sekwencji i w metodzie SICHO, przeprowadził część symulacji przewidywania struktury białek, wykonał modelowanie porównawcze przy pomocy programu MODELLER.