



ЮГОЗАПАДЕН УНИВЕРСИТЕТ
“НЕОФИТ РИЛСКИ” - БЛАГОЕВГРАД
ПРИРОДО-МАТЕМАТИЧЕСКИ ФАКУЛТЕТ
КАТЕДРА “ИНФОРМАТИКА”

Методи Георгиев Трайков

**МАТЕМАТИЧЕСКИ МОДЕЛИ И АЛГОРИТМИ ЗА
ПРЕДСКАЗВАНЕ НА ПРОСТРАНСТВЕНАТА СТРУКТУРА НА
ПРОТЕИНИ**

АВТОРЕФЕРАТ

НА ДИСЕРТАЦИЯ

*за присъждане на образователна и научна степен „доктор” в
област на висше образование: 4. Природни науки, математика и информатика.
Професионално направление 4.6. Информатика и компютърни науки.
Докторска програма “Информатика”*

Научен ръководител: проф. д-р Петър Миланов

Благоевград, 2017г.

Настоящият дисертационен труд съдържа въведение, четири глави, изводи, и цитирана литература. Общият му обем е 160 страници, от които 15 страници използвана литература. Включени са 80 фигури и 17 таблици. Цитираната литература обхваща 186 заглавия.

Дисертационният труд е обсъден на 15.06.2017 г. на заседание на катедрен съвет на катедра „Информатика” към Природо-математически факултет при Югозападен университет „Неофит Рилски” - Благоевград.

Публичната защита на дисертационния труд ще се състои на 14.08.2017 г. от 10:00 часа в зала 412 на Югозападен университет „Неофит Рилски”- Благоевград, пред научно жури в състав:

Научно жури:

академик д. м. н. Петър Кендеров – външен член

проф. д-р Румен Андонов – външен член

доц. д-р Йордан Митев - външен член

проф. д-р Петър Миланов – вътрешен член

проф. д-р Никола Янев – вътрешен член

Материалите по защитата са на разположение на интересувашите се в кабинет № 1426 в Учебен корпус-1 към Югозападен Университет „Неофит Рилски” – Благоевград.

Автор: **Методи Георгиев Трайков**

Тема: **МАТЕМАТИЧЕСКИ МОДЕЛИ И АЛГОРИТМИ ЗА
ПРЕДСКАЗВАНЕ НА ПРОСТРАНСТВЕНАТА
СТРУКТУРА НА ПРОТЕИНИ**

СЪДЪРЖАНИЕ

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД.....	4
1. Актуалност и мотивировка на темата.....	4
2. Цели и задачи на дисертацията.....	5
3. Научна новост.....	6
4. Аprobация.....	6
5. Публикации.....	7
СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД.....	7
ПЪРВА ГЛАВА: ВЪВЕДЕНИЕ В ПРОБЛЕМА ЗА НАГЪВАНЕ НА ПРОТЕИНИ..	7
ВТОРА ГЛАВА: МОДЕЛИ ЗА НАГЪВАНЕ НА ПРОТЕИНИ.....	12
ТРЕТА ГЛАВА: АЛГОРИТМИ ЗА НАГЪВАНЕ НА ПРОТЕИНИ.....	14
ЧЕТВЪРТА ГЛАВА: РЕЗУЛТАТИ И ОБСЪЖДАНИЯ.....	21
ИЗВОДИ.....	40
СПИСЪК С ПУБЛИКАЦИИ ПО ДИСЕРТАЦИЯТА В НАУЧНИ СПИСАНИЯ.....	40
БИБЛИОГРАФИЯ.....	44

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност и мотивировка на темата

Изследванията в областта на биоинформатиката стават все по актуални. Днес компютърните методи за изследване, анализ и предсказване на пространствената структура на протеините са широко използвани в научните изследвания. Чрез тях се постига по-голяма ефективност на финансовите разходи, съкращаване на времето за обработка на тежки оптимизационни задачи и откриване на алтернативи за многобройните тестове с животни.

Математическото и компютърното моделиране е от ключово значение за биоинформатичните изследвания, защото биоинформатиката е прилагането на компютърни технологии за съхранение, организиране и анализиране на тримерната структура на протеините чрез линейното подреждане на техните базови молекули. Чрез компютърното моделиране се постига интегриране на данни под формата на математически обекти и виртуални изображения, което позволява изучаване и моделиране на процесите в биологичните системи, а това допринася за развитието на молекулярната биология. По проблема за предсказване на пространствената структура на протеините в България работят различни групи като: Центъра за съвременни биоинформатични изследвания към ЮЗУ "Неофит Рилски"; Института по математика и информатика – БАН; Центъра по Биоинформатика в Университета в Рен 1, Франция; Център по Биоинформатика в Кьолн, Германия; Институт за геномика и биоинформатика в Грац, Австрия; CLC bio - водещи разработчици на биоинформатичен софтуер; Колеж за наука и технологии във Филаделфия, САЩ и други.

През последните години са публикувани повече от 30 хиляди статии свързани с тази тематика. Съществуват повече от 100 научни списания в тази област. Водещи университети като Харвардски и Оксфордски университет имат специализирани издания по биоинформатика. Актуалността на тези научни изследвания е описана в доклада на Европейския институт за биоинформатични изследвания (Annual Scientific Report 2014, EMBL: <http://www.embl.de>).

Компютърното предсказване на пространствената структура на протеините е от особено практическо значение, тъй като води до създаване на бази от данни, като например Protein Data Bank, които от своя страна водят до компютърно-подпомогне на лекарствения дизайн, може да се използват при откриване на нови лекарства и да се комбинират успешно с експериментални методи, което дава

възможност да се опростяват или да се моделират сложни и скъпи експерименти. По този начин се получава ценна информация за случващото се в организмите под действието на определени вещества.

Изследванията в настоящия дисертационен труд се базират на целочислено оптимизиране и математическо моделиране на НР модела за нагъване на протеини в 2D квадратна и 3D кубична решетка с цел да се предскаже пространствената структура на протеините. Проблемът за нагъване на протеини е НР-пълнен проблем и неговото решение ще окаже огромно влияние в областта на химията и биологията. Разработването на ефективни математически модели и алгоритми за нагъване на протеини е продължителен процес, който включва знания и умения на различни изследователи: математици, химици, биолози, информатици и други. Намирането на компактна третична структура за даден протеин е трудна задача, поради сложността на елементите, от които зависи процеса на нагъване.

Ползата от проведените компютърни симулации е факта, че разработените алгоритми за нагъване на протеини генерират качествени и по-добри нагъвания от описаните в литературата решения за множество от бенчмарк последователности. Тези бенчмарк последователности представляват протеинови последователности, които са широко използвани в литературата за тестване и определяне на качеството на нови алгоритми за нагъване на протеини.

Алгоритмите за нагъване на протеини може да помогнат да се обясни процеса на нагъване на протеини в природата и успешно да се използват за проектиране и анализ на протеини, като по този начин може да се подпомогне проектирането на нови лекарства, като се спестяват много средства, време и най-вече експериментите с животни.

2. Цели и задачи на дисертацията

Цели:

- (1) Да се разработят целочислени модели за проблема за нагъване на протеини в 2D и 3D;
- (2) Да се разработят алгоритми за нагъване на протеини в 2D и 3D НР решетъчни модели и да се реализират компютърна симулация на процеса на нагъване на протеини.

Задачи:

- (1) Да се изгради аксиоматиката на математическия модел за проблема за нагъване на протеини.
- (2) Да се изведат формули за основните ограничения при процеса на нагъване на протеини в двумерното и тримерното пространство.

- (3) Да се изведе целочислена формулировка на проблема за нагъване на протеини.
- (4) Да се разработят алгоритми за нагъване на протеини в 2D и 3D HP решетъчни модели.
- (5) Да се реализират компютърни симулации на процеса на нагъване на протеини с разработените алгоритми и да се верифицират получените резултати.
- (6) Да се направи анализ на разработените алгоритми и да се сравнят получените резултати с известни в литературата алгоритми за нагъване на протеини.
- (7) Да се симулира нагъване на реални протеини с дължина над 200 аминокиселини.

3. Научна новост

Показано е, че представянето на проблема за нагъване на протеин, като оптимизационен проблем от Теория на графите, може да доведе до изграждане на качествени целочислени математически модели, на базата на които да се разработят алгоритми за решаване на проблема, които генерират отлични резултати.

Разработени са математически модели и алгоритми за нагъване на протеини в 2D квадратна и 3D кубична решетка.

С компютърни симулации, базирани на разработените модели и алгоритми успешно са генерирани нагъвания за протеини с различни дължини. Сравненията на тези резултати с резултати, генерирани от известни в литературата алгоритми за решаване на проблема показват, че разработените модели и алгоритми са качествени и надеждни.

Показано е, че резултатите генерирани от предложените в дисертацията математически модели и алгоритми подобряват описаните в литературата резултати получени от алгоритми, като Монте Карло алгоритми и Генетични алгоритми, което от своя страна води до предлагане на нова бенчмарк последователност.

4. Аprobация

Резултатите от дисертационния труд са обсъждани и докладвани на следните научни прояви:

- International conference on Mathematical Methods and Models in Biosciences (Biomath-2016), Blagoevgrad, Bulgaria;
- PhD Student Scientific Session of the FMNS - 2016(15.06.2016г.), Blagoevgrad, Bulgaria;
- 6th International Conference – FMNS 2015, Blagoevgrad, Bulgaria;

- International Conference on BIOMATH 2015, Blagoevgrad, Bulgaria;
- International Congress on Mathematics (MICOM-2015), Athens, Greece;
- Workshop "Mathematical Modelling and Scientific Computing" December 09 - 11, 2015, Velingrad, Bulgaria;
- International Conference "Enhancement of the competitiveness and restructuring of the food subsectors through benchmarking", 2014, Blagoevgrad, Bulgaria;
- ИМИ-БАН, Секция „Изследване на операциите”, 2014, София;

5. Публикации

Получените резултати са публикувани в 2 статии в списания с импакт фактор и 8 в сборници с абстракти от международни конференции. Всички публикации са на английски език.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Настоящият дисертационен труд съдържа въведение, четири глави, изводи и цитирана литература. Общият му обем е 160 страници, от които 15 страници използвана литература. Включени са 80 фигури и 17 таблици. Цитираната литература обхваща 186 заглавия.

ПЪРВА ГЛАВА: ВЪВЕДЕНИЕ В ПРОБЛЕМА ЗА НАГЪВАНЕ НА ПРОТЕИНИ

В първа глава на настоящата дисертация са представени проблемите и обосновката на биоинформатичните изследвания - основните биологични аспекти, приложимостта и ролята на математическото и на компютърното моделиране при решаване на проблема за нагъване на протеини. Формулирани са целите и задачите на дисертацията.

Изследванията в дисертацията се базират на подходи за компютърно и математическо моделиране за търсене на оптимални конформации (максимален брой контакти между хидрофобните аминокиселини) върху двумерни и тримерни решетки при решаване на проблема за нагъване на протеини. Чрез компютърно моделиране на процес на нагъване на протеини се анализират математически модели и алгоритми за решаване на проблема. Протеиновите последователности използвани за изчислителните експерименти в двумерното и тримерното пространство са известни в литературата

бенчмарк последователности, за които са известни най-добрите решения, намерени до момента. Изследваните протеинови последователности са с различна дължина. На Фигура 1.1 е показана оптимална конформация за протеинова последователност с дължина 36 аминокиселини в двумерното пространство. За тази протеинова последователност най-доброто известно и описано в литературата решение в двумерното пространство е 14 контакта, а в тримерното пространство е 18 контакта.



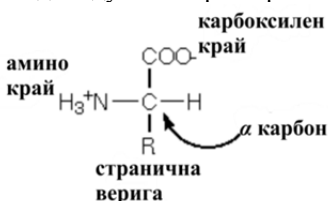
Фигура 1.1: Оптимална конформация за протеинова последователност с дължина 36 аминокиселини в двумерното пространство.

Биологичните аспекти на процеса на нагъване на протеини, както и на самите протеини са разгледани в Раздел 1.1. В Раздел 1.3 е разгледан обобщен модел на проблема за нагъване на протеини, а в Раздел 1.4 са разгледани решетъчни модели, използване за решаване на разглеждания проблем в 2D и 3D. Представени са математически модели за изчисляване на горна граница при решаване на разглеждания проблем (Ahn & Park, 2010). Базирайки се на тази горна граници се показва качеството и надеждността на предлаганите модели и алгоритми. В Разделите 1.3 и 1.4 са въведени необходими понятия и факти. В Раздел 1.5 е направен необходимия литературен обзор.

1.1. Биологични аспекти на изследванията

Протеините представляват линейни полимери, изградени от последователност от аминокиселини (полипептидни остатъци). Всички аминокиселини притежават общи структурни характеристики, включително α -въглерод, който е свързан с аминогрупа, карбоксилна група и странична верига. Основните елементи в структурата на протеините са *аминокиселините*. В човешкото тяло има 20 различни вида аминокиселини. Всяка аминокиселина съдържа аmino група (NH_3) и карбоксилна група (COOH); аминокиселините се различават

само в техните странични вериги (фигура 1.2). Тези странични вериги може да бъдат полярни (хидрофилни), неполярни (хидрофобни) или неутрални. Аминокиселините се свързват с помощта на пептидни връзки. Страничните вериги, които не участват във формирането на пептидните връзки, придават на всяка аминокиселина уникални свойства. Броя на аминокиселините и тяхната последователност, дават на протеина неговите индивидуални характеристики.



Фигура 1.2: Химична формула на една аминокиселина

Структурата на протеините е йерархично организирана на четири нива (фигура 1.3): първична, вторична, третична и четвъртична структура. Всяко ниво е важно и тяхната комбинация води до един всеобхватен поглед над протеините.



Фигура 1.3: Йерархична структура на протеин

Обектите, които представляват интерес на изследване в настоящата дисертация са *бенчмарк протеинови последователности*. Бенчмарк последователност е протеинова последователност, която се използва за тестване и определяна на качеството на описаните в литературата алгоритми за решаване на проблема за нагъване на протеини в 2D и 3D. Поради тази причина използваните в дисертацията бенчмарк протеинови последователности са описани в голямо количество литература и са известни най-добрите получени резултати за тях.

1.2. Обобщен модел

През 1985г. Кен Дил предлага хидрофобен-хидрофилен (НР) модел за решаване на проблема за нагъване на протеини, който след това се описва в огромно количество литература поради своята основна роля в моделите за нагъване на протеини. Ученият класифицира 20-те аминокиселини като хидрофобни (Н) или полярни (Р) със степен на хидрофобност на аминокиселините (Таблица 1.1). Две хидрофобни аминокиселини образуват *контакт*, ако са съседни в нагъването и несъседни в основната последователност. Тъй като целта е формиране на силно компактни хидрофобни ядра, целта на оптимизационната функция е да увеличи броя на контактите между хидрофобните атоми (Н-Н контакти).

Име	Символ	Класификация	Име	Символ	Класификация
Alanine	A	Hydrophobic	Leucine	L	Hydrophobic
Arginine	R	Polar	Lysine	K	Polar
Asparagine	N	Polar	Methionine	M	Hydrophobic
Aspartic Acid	D	Polar	Phenylalanine	F	Hydrophobic
Cysteine	C	Polar	Proline	P	Hydrophobic
Glutamic Acid	E	Polar	Serine	S	Polar
Glutamine	Q	Polar	Threonine	T	Polar
Glycine	G	Polar	Tryptophan	W	Hydrophobic
Histidine	H	Polar	Tyrosine	Y	Polar
Isoleucine	I	Hydrophobic	Valine	V	Hydrophobic

Таблица 1.1: 20-те аминокиселини

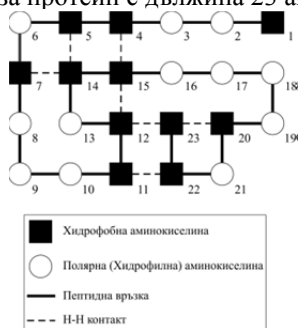
Alberts и неговите сътрудници отбелязват, че хидрофобният ефект сред аминокиселините допринася значителна част от общата енергийна функция, така че това е най-важната сила при определянето на структурата на протеина (Alberts, et al., 1998). Хидрофобността на аминокиселините е мярка от термодинамичното взаимодействие между страничната верига и водата. 20-те типа аминокиселини имат различни характеристики по отношение на размера, заряда и хидрофобността. НР модел опростява проблема за нагъване на протеина, като разглежда само два вида аминокиселини: Н и Р. От друга страна, без НР модела се счита, че 20-те аминокиселини са различни една от друга (Chandru, et al., 2003; Heun, 2003; Lau & Dill, 1989; Lesk, 2001; Thompson, et al., 1994).

Оптималната 3D структурата на протеина като цяло е неговото минимално енергийно състояние. Така че, моделите за нагъване на протеини трябва да представляват енергийното състояние на структурата на протеините. Предложени са няколко енергийни функции (Erman & Dill, 2000; Irback & Sandelin, 1998; Irback, et al., 1997; Miller, et al., 2002). Въпреки това, не е била определена точна

енергийна функция, дори за протеини с по-малко от 50 аминокиселини. Така че, за да се разбере 3D структурите на протеините, често се използват приближени модели. Целта на тези модели е да максимизират броя на контактите между хидрофобните аминокиселини, който не са съседни в линейната последователност (Lau & Dill, 1989), при следните условия:

1. (*Присвояване*) Всяка аминокиселина трябва да заема точно една клетка от решетката.
2. (*Не самопресичащи се*) Две аминокиселини не могат да заемат една и съща клетка в решетката.
3. (*Свързаност*) Всеки две аминокиселини, които са съседни в протеиновата последователност, също трябва да заемат съседни клетки в решетката.

Greenberg и неговите сътрудници, демонстрират, че тези три изисквания дават правилно решение на проблема за нагъване на протеини (Greenberg, et al., 2004). Фигура 1.4 показва изпълнимо решение на проблема за протеин с дължина 23 аминокиселини.



Фигура 1.4: Пример за 2D възможно решение на HP решетъчния модел

За да се отчетат контактите между хидрофобните аминокиселини при моделирането на HP решетъчния модел се използва L_1 разстояние (разстояние по Манхатън) между две точки или аминокиселини, т.е. $d((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$. Също така HP решетъчния модел опростява формата на протеините и дава известно предимство при моделирането.

ВТОРА ГЛАВА: МОДЕЛИ ЗА НАГЪВАНЕ НА ПРОТЕИНИ

В Глава 2 се съдържа описание на методичните подходи, използвани за решаването на поставените задачи. Разгледани са подходи като целочислено моделиране и подравняване на последователности.

2.1. Целочислено моделиране

В този раздел проблема за нагъване на протеини е представен чрез средствата на целочисленото оптимиране. Целочисленото оптимиране (Integer Programming – IP) е успешна техника за решаване на дискретни оптимизационни проблеми, които възникват в различни научни области. При целочисленото оптимиране се използват алгебрични ограничения (линейни уравнения и неравенства с целочислени стойности за променливите), за да се минимизира или максимизира определена целева функция. Тези алгебрични ограничения позволяват разширяване на мощните методики на линейното програмиране, за да се достигне до разглеждане на комбинаторен оптимизационен проблем.

Лау и Дил (Lau & Dill, 1989) показват, че HP решетъчният модел за нагъване на протеин в 2D квадратна или 3D кубична решетка може да бъде описан по следния начин:

Максимизиране:

Броя на Н-Н контактите.

При условия:

1. (*Присвояване*) Всяка аминокиселина трябва да заема точно една клетка от решетката.
2. (*Не самопресичащи се*) Две аминокиселини не могат да заемат една и съща клетка в решетката.
3. (*Свързаност*) Всеки две аминокиселини, които са съседни в протеиновата последователност, също трябва да заемат съседни клетки в решетката.

Нека е даден протеин с аминокиселина последователност S . Аминокиселините са групирани на хидрофилни или полярни ($P \rightarrow 0$) и хидрофобни ($H \rightarrow 1$).

Представения в този раздел математически модел се базира на квадратна решетка с фиксирана големина. Първата стъпка е да се направи групиране на N аминокиселините според тяхната позиция във входната последователност S на четни (*EVEN*) и нечетни (*ODD*). Задачата е да се максимизира броя на контактите между хидрофобните аминокиселини (Н-Н контактите) в 2D квадратен или 3D кубичен решетъчен модел, като се спазват посочените по-горе ограничения.

Размера N на страната на решетките е входен параметър на алгоритмите, например:

- $N = n$;
- $N = 2\sqrt{n}$;
- $N = \frac{n}{2}$.

За по-голяма простота 2D квадратния или 3D кубичния решетъчен модел се преобразува в 1D по следния начин (Yanev, et al., 2011):

- двумерните координати (x, y) се преобразуват като една i координата по следния начин: $i = N*(y-1)+x$;
- тримерните координати (x, y, z) се преобразуват като една i координата по следния начин: $i = N^2*(z-1)+N*(y-1)+x$.

2.2. Подравняване на последователности

В този раздел са разгледани подходи за подравняване на протеинови последователности. Подравняването на последователности има много полезни приложения в изчислителната биология. При подравняване на биологични последователности, като ДНК последователности или протеинови последователности, оценъчната функцията отразява биологичната правдоподобност на различни видове мутации, които се срещат в анализираниите последователностите.

Целта при процеса на подравняване на последователности е множество последователности да бъдат подравнени, така че да се максимизира броя на позициите, където има съвпадение на символи от всяка една от последователностите. За да се получи добро подравняване се вмъкват така наречените *гапове* / *празни пространства* ("-"), при което останалите символи се изместват на по-добри позиции. Оценката на генерираното подравняване се осъществява от оценъчна функция, която дава точки за съвпадение на символи и вмъква наказания за евентуални гапове / празни пространства. Основната цел е да се генерира оптимално подравняване за множество от последователности (Argos, 1987; Dayhoff & Eck, 1968; Dayhoff, et al., 1978; Doolittle, 1979; Feng, et al., 1985; Fitch, 1970; Fitch & Margoliash, 1967; Grantham, 1974; McLachlan, 1971; Miyata, et al., 1979).

Предложени са голям брой стратегии, които да подобрят точността при подравняване на протеинови последователности. Тези стратегии включват множествено подравняване на последователности, подравняване на много последователности чрез итеративни двоични сравнения и подравняване чрез шаблони, профили или модели на аминокиселини (Barton & Sternberg, 1987; Feng & Doolittle, 1987; Feng

& Doolittle, 1990; Fitch, 1970; Fredman, 1984; Jue, et al., 1980; Taylor & Orengo, 1989).

ТРЕТА ГЛАВА: АЛГОРИТМИ ЗА НАГЪВАНЕ НА ПРОТЕИНИ

В Глава 3 се разгледани някои от известните в литературата алгоритми за предсказване на третичната структура на протеините, като алгоритмите на Харт и Истраил, алгоритъма на Newaman и евристични алгоритми (Монте Карло алгоритми, Генетични алгоритми и други). Резултатите от тези алгоритми се сравняват с резултатите получени от предложените в дисертацията евристични алгоритми за нагъване на протеини, като по този начин се подчертава качеството на разработените алгоритми. Също така в тази глава накратко са описани техниките, използвани в различните типове алгоритми за нагъване на протеини, а именно метод на Разклоненията и границите, Динамично програмиране и само-избягващи се пътища.

3.1. Приблизени алгоритми

В този раздел от дисертацията са разгледани някои приближени алгоритми за НР модела за нагъване на протеини. Резултатите от тези алгоритми се отнасят за решетъчни модели, за линейно-верижни и странично-верижни НР модели, както и за обобщени извън-решетъчни и допирателно-сферични НР модели.

Когато е доказано, че дадена оптимизационна задача е изчислително неразрешима, т.е. НР пълна, следващата стъпка, е да се помисли за съществуването на алгоритми, които могат да генерират решения, които са близки до оптималното решение. Тези алгоритми се наричат *Приблизени алгоритми*. За количественото изразяване на така наречените „*решения, близки до оптималните*” се използва следната терминология. Един α -приближен алгоритъм за даден проблем е полиномиален алгоритъм, чийто изход е в съотношение най-много α пъти от оптималното решение. За проблем за минимизация, изходът е най-много α пъти стойността на оптималното решение, а за проблем за максимизация, изхода е най-малко $1/\alpha$ пъти стойността на оптималното решение. Стойността α се нарича *приближено съотношение* или *приближена гаранция* за проблема.

3.1.1. Алгоритъм на Харт и Истраил

Нека е даден проблем за максимизиране на целева функция C , която максимизира броя на контактите в дадено нагъване.

Първата стъпка при проектирането на приближен алгоритъм за този максимизиращ проблем е да се намери горна граница на

оптималната стойност на целевата функция, която да може да бъде изчислена в полиномиално време.

Основната идея на първия в литературата строг приближен алгоритъм с доказуема приближена гаранция за проблема за нагъване на протеин в HP модела за 2D квадратна и 3D кубична решетки, е установена от Харт и Истраил (Hart & Istrail, 1995). На входа на този алгоритъм се задава протеинова последователност S с хидрофобни (H) и хидрофилни (P) аминокиселини. За дадена хидрофобна аминокиселина се казва, че е *нечетна* ако тя се намира на нечетна позиция във входната последователност S и *четна*, ако тя се намира на четна позиция. Нека $\mathcal{O}(S)$ да показва броя на нечетните хидрофобни аминокиселини в S , а $\mathcal{E}(S)$ да показва броя на четните хидрофобни аминокиселини в S . Алгоритма на Харт и Истраил за нагъване на протеини в 2D или 3D решетъчен HP модел използва следната оценка за оптималния брой контакти: за всяко нагъване $F(S)$ за последователността S имаме:

$$\#contacts \left(F(S) \leq C_{2D}(S)(C_{3D}(S)) \right)$$

където

$$C_{2D}(S) = 2 \min\{\mathcal{O}(S), \mathcal{E}(S)\} + k \text{ или } C_{3D}(S) = 4(\min\{\mathcal{O}(S), \mathcal{E}(S)\}) + k \text{ в 3D случая}$$

$$k = \begin{cases} 0, & \text{ако няма H аминокиселини на първа и последна позиция} \\ 1, & \text{ако има H аминокиселини на първа или последна позиция} \\ 2, & \text{ако има H аминокиселини на първа и последна позиция} \end{cases}$$

Горната граница $C_{2D}(S)$ ($C_{3D}(S)$) е в пряка зависимост от свойствата на основната квадратна (кубична) решетка, по-специално H-H контакт в нагъването може да се образува само ако аминокиселините са на различни позиции в последователността.

Нюман е дал пример за последователност, за която $C_{2D}(S)$ е в рамките на множител 2 за оптималния брой на контактите (Newman, 2002). Следователно всеки приближен алгоритъм (който по дефиниция работи за всички протеинови последователности), за който може да се гарантира, че генерираното от него решение е близко до оптималната граница и се базира на неравенството

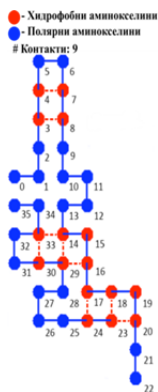
$$\#contacts(F(S) \leq C_{2D}(S))$$

не може да постигне приближено съотношение строго по-голяма от $\frac{1}{2}$.

3.1.2. Алгоритъм на Нюман

Нюман разработва подобрен приближен алгоритъм с приближено съотношение $\frac{1}{3}$ за 2D квадратна решетка (Newman, 2002). Този алгоритъм използва допълнителни извивки в последователността,

които подобряват броя на контактите. Фигура 3.1 показва нагъване получено чрез алгоритъма на Нюман за протеин с дължина 36 аминокиселини.



Фигура 3.1: Нагъване получено от алгоритъма на Нюман

3.2. Евристични алгоритми

В този раздел от дисертацията са разгледани някои от известните в литературата евристични методи за 2D или 3D HP модела за нагъване на протеини. Разработените в дисертацията алгоритми за проблема за нагъване на протеини в 2D и 3D са евристични алгоритми. Това е причината да се разгледат известните в литературата евристични алгоритми (Монте Карло методи, Генетични алгоритми, алгоритъма на Мравките, метода на Симулирано закаляване), като целта е да се направи сравнение между разработените в дисертацията алгоритми и известни в литературата алгоритми за нагъване на протеини.

3.2.1 Монте Карло методи

Монте Карло (Monte Carlo - MC) методите са много известни методи използвани при изчисляване на сложни системи. Решетъчно-базираните Монте Карло симулации за проблема за нагъване на протеини са едни от най-успешните методи за нагъване на протеини. Тези методи използват опростено описание на последователността от аминокиселини. Unger и Moult опростяват допълнително тези методи, като използват двумерен модел и представят аминокиселините като единични атоми (Unger & Moult, 1992).

3.2.2. Генетични алгоритми

Методите с генетични алгоритми се базират на оптимизационни процедури подобни на естествената генетична еволюция, при които популацията постепенно се подобрява чрез селекция. Приложението

на генетичните алгоритми към проблема за нагъване на протеини може да се разгледа като разширение на Монте Карло (Monte Carlo – MC) методи, които включват обмен на информация между множество от паралелни симулации. Генетичните алгоритми поддържат популация от еволюиращи конформации. По време на процеса на нагъване, който е еквивалентен на процеса на мутация, всяка конформация се променя независимо за определено период от време чрез Монте Карло процедурата на Metropolis (Metropolis, et al., 1953).

3.2.3. Алгоритъм на мравките

Оптимизационния алгоритъм на мравките (Ant Colony Optimization – ACO), е еволюционна техника, вдъхновена от поведението на мравките при търсене на храна. Първите, които използват алгоритъма на мравките за проблема за нагъване на протеини са Shmygelska и Hoos (Shmygelska & Hoos, 2005). Техния алгоритъм се основава на три фази: изграждане, локално търсене и феромонно актуализиране. В първата фаза, алгоритъма на мравките конструира нагъване в решетка, започвайки от произволно място. Следващата стъпка е използване на алчна процедура за локално търсене, която е базирана на метод за широко-обхватна мутация, създаден от Shmygelska и Hoos (Shmygelska & Hoos, 2005). След това, алгоритъма актуализира феромонната матрица, използвайки два основни механизма: еднородно съотношение на преобразуване и укрепване на локалните нагъвания. Shmygelska и Hoos прилагат алгоритъма на мравките за няколко бенчмарк последователност в 2D и 3D HP модели.

3.2.4. Симулирано закаляване

В този раздел са разгледани алгоритъм за нагъване на протеини, базиран на метода Симулирано закаляване (Aarts & Lenstra, 2003; Michalewicz & Fogel, 2004). На фигура 3.2 е показана обобщена блок-схема на начина на работа на този тип евристични алгоритми за намиране на оптимална третична структура на протеините.



Фигура 3.2: Метод на Симулирано закаляване

Множеството $N(x)$ съдържа съседни решения за решението x , а $f(x)$ показва стойността на решението x . Съседните решения за дадено решение представляват структури, които може да се генерират от един ход на локално търсене.

3.3. Метод на разклоненията и границите

В този раздел е направено описание на метода на разклоненията и границите. Метода на Разклоненията и Границите е проектиран за решаване на дискретни и комбинаторни оптимизационни проблеми. По същество този метод се явява вариация на техника за пълно изчерпване, изключваща подмножества от допустими решения, които не съдържат оптималното решение за дадения проблем. Разглеждания проблем се представя чрез дървовидна структура (Сигал & Иванова, 2002).

Метода на Разклоненията и Границите, се състои от систематично изброяване на кандидат решения чрез обхождане на пространството за търсене: множеството от кандидат решенията се разглежда като дърво, в което всеки клон представлява едно кандидат решение, а пълното множество от кандидат решения представлява корена на дървото. Метода изследва отделните клонове на това дърво, които представляват подмножества на множеството от решения. Избора на дадено кандидат решение става чрез проверка на очакваните горни и долни граници за оптималното решение. Едно кандидат решение се изхвърля, ако не е по-добро от най-доброто намерено до момента решение.

Този метод зависи от ефективната оценка на долната и горната граница на разклоненията в пространството за търсене и подходи за

търсене чрез пълно изчерпване, тъй като размерът на разклоненията клони към нула.

Метода на Разклоненията и Границите работи съгласно два принципа:

- Рекурсивно разделяне на пространството за търсене на по-малки пространства, след това минимизиране на функцията $f(x)$ за тези по-малки пространства. Разделянето се нарича *разклоняване*;
- Разклоняването само по себе си би означава тестване и изброяване чрез метода на грубата сила на кандидат решенията. За да се подобри ефективността на търсенето чрез груба сила, метода на Разклоненията и Границите запазва пътя до границите на минимума или максимума, които се опитва да намери и използва тези граници, за да "реже" пространството за търсене, премахвайки кандидат решения, за които може да докаже, че не съдържат оптималното решение на проблема.

Превръщането на тези принципи в конкретен алгоритъм за конкретен оптимизационен проблем изисква някаква структура на данните, която да представлява множеството от кандидат решения.

3.4. Динамично програмиране

Динамичното програмиране (Dynamic Programing – DP) е рекурсивна техника, подобна на метода Разделяй и Владей, която решава даден проблем чрез разделяне на проблема на подпроблеми и комбиниране на решенията за всеки един подпроблем (Cormen, et al., 2001). Разликата между Динамичното програмиране и метода Разделяй и Владей, е в това че динамичното програмиране решава всеки подпроблем само веднъж и запазва полученото решение в таблица, като по този начин избягва повтарящите се изчисления всеки път, когато възниква подпроблем.

Динамичното програмиране може да се обобщи, както следва:

1. Създаване на рекурсивни свойства, базирани на "принципа на оптималността", които описват оптималното решение на даден проблем от гледна точка на оптималните решения на подпроблемите за този проблем.
2. Решава проблемите от долу на горе т.е., първо изчислява оптималните решения за по-малките по размер проблеми и след това използва тези решения, за да изчисли решението на по-големия проблем.

Динамичното програмиране води началото си с Белман (1957) и за първи път е използвано при подравняване на биологични

последователности от Needleman и Wunsch (Needleman & Wunsch, 1970).

3.5. Само-избягваща се разходка

В този раздел на дисертацията са описани методи за подравняване на нагъвания на протеини чрез само-избягващи се разходки (Self-Avoiding Walks – SAW) в 2D и 3D пространство. Тези методи се фокусират върху правилното дефиниране на контакти между двойки от аминокиселини, а множеството от всички тези контакти се нарича карта на контакти. Един от измерителите за сходство, който се използва при подравняване на нагъвания е Припокриваща Карта на Контактите (Contact-Map Overlap - CMO).

3.5.1. 1 SAW = 2 Стека + 1 Опашка

Една картата на контакти за само-избягваща се разходка може да се разложи на два "стека" и една "опашка". Оказва се, че стековете (наподобяващи α -спирали) съответстват на вериги от аминокиселини в частично подредени структури и представляват обобщения на вторичните структури, разглеждани като графи. От двоична гледна точка, опашките (наподобяващи β -листа) съответстват на анти-вериги в частично подредени структури.

3.5.2. Контактни дървета на Schmitt-Waterman

Работата на Waterman, осигурява ясно формулирани аналитични формули за преброяване на общия брой на протеинови структури с определена дължина. Тези аналитични формули биха могли да осигурят по-голяма яснота при изчисляване на строги приближения за функцията на разпределение за дадено нагъване на протеин в NP модели, чрез споменатото по-горе разлагане (фигурите по-долу) (Schmitt & Waterman, 1994; Waterman, 1978).



Фигура 3.3: Стек



Фигура 3.4: контактно дърво
на Schmitt-Waterman



Фигура 3.5: Опашка

3.5.3. Подравняване на нагъвания чрез Припокриваща Карта на Контакти

Задачата за подравняване на протеинови последователности е важен проблем, който намира приложения при класифицирането на известни нагъвания, с помощта на които се предсказват нови

нагъвания и се измерва качеството на алгоритмите за предсказване на нагъвания (Eidhammer, et al., 2000; Holm, et al., 2008; Murzina, et al., 1995; Orengo, et al., 1997). С цел да се измери качеството на алгоритмите за нагъване на протеини, е необходимо да се определи измерител за сходство на нагъването. Проучени са няколко такива измерители, включително средно квадратично отклонение (RMSD), сходство с карта на разстоянието и припокриваща карта на контактите (Backofen, et al., 2000; Fitch, 1970).

Проблема за припокриващата карта на контактите представлява проблем за намиране на карта, която прилага подредбата на аминокиселините в един протеин към аминокиселините на втори протеин, с цел да се увеличи общия брой на контактите във втория протеин.

През последното десетилетие този измерител (който е NP-пълн (Goldman, et al., 1999)), е получил огромно внимание при количественото определяне на приликата между тримерните количествени структури. Въпреки това, известните в литературата алгоритми показват скромно представяне и не са приложими при мащабни сравнения. Най-добрите получени резултати в тази област са на Yanev et al., който предлагат нов целочислен модел и точен Branch-and-Bound алгоритъм с граници, получени чрез Лагранжова релаксация. Ефективността на този подход е демонстрирана над известно в литературата бенчмарк множество (множество на Сколник, 40 домейна), за което алгоритъма на Yanev et al. значително надминава най-добрите съществуващи точни алгоритми. Също така, с този алгоритъм са решени за първи път много трудни СМО примери (Argos, 1987; Backofen & Will, 2001; Caprara & Lancia, 2002; Caprara, et al., 2004; Marinari & Parisi, 1992; Yanev, et al., 2008).

ЧЕТВЪРТА ГЛАВА: РЕЗУЛТАТИ И ОБСЪЖДАНЯ

В Глава 4 са представени и обсъдени получените резултати от разработените математически модели и алгоритми за решаване на проблем за нагъване на протеини в 2D квадратна и 3D кубична решетка.

Реализирани са компютърни симулация на процеса на нагъване на протеини с различна дължина. Протеините използвани в симулациите са известни в литературата бенчмарк последователности (таблица 4.1). Символите H_i , P_i и $(\dots)_i$ в таблица 4.1 показват i повторения на съответния символ или подпоследователност.

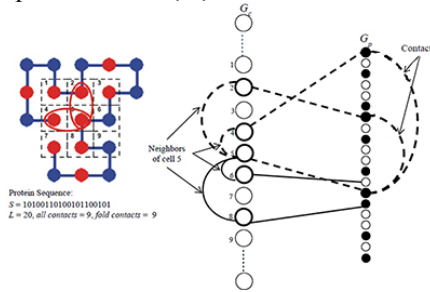
Нека $G = G_S \cup G_c$ да бъде пълен двуделен граф с множество от върхове $V = V_S \cup V_c$. Сдвояването (в този случай едно-към-едно съпоставяне на V_S към V_c) $M = \{e_1, e_2, \dots, e_n\}$ с $|M| = n$ е **допустимо** ако покритите върхове в V_c дефинират само-избягващ се път (фигура 4.1). Дефинира се функцията $z(e_i, e_j) = z_{i,k,j,l} = 1$, ако $(i, j) \in E_S$, $(k, l) \in E_c$ и $z_{i,k,j,l} = 0$ в противен случай. Накрая, се дефинира

$$v(M) = \sum_{e_i, e_j \in M} z(e_i, e_j) \quad (4.1)$$

Нека $\tilde{M}$ да бъде множеството от допустими съвпадения. Тогава проблема за намиране на оптимално нагъване над избрана решетка е:

Проблем за нагъване в граф (Folding problem on graph – FPG):

За $M \in \tilde{M}$ да се намери $v = \max v(M)$.



Фигура 4.1: Едно-към-едно изчертаване на V_S към V_c

Преобразуването на НР проблема за нагъване на протеини към оптимизационен проблем от графи позволява да се изградят различни целочислени модели. Повечето от тях включват двоични променливи, да кажем $x_{i,k}$, за моделиране на възможните съвпадения от горните 0-1 решения до прости линейни ограничения:

(Присвояване)

$$\sum_{i=1}^{N^2} x_{i,k} = 1 \quad k = 1, \dots, n. \quad (4.2)$$

(Не самопресичащи се)

$$\sum_{k=1}^n x_{i,k} \leq 1 \quad i = 1, \dots, N^2. \quad (4.3)$$

(Свързаност)

$$x_{i,k} \leq \sum_{j \in G(i)} x_{j,k+1} \quad i = 1, \dots, N^2; k = 1, \dots, n-1. \quad (4.4)$$

където $G(i)$ е множеството от съседите на i -та клетка.

Целевата функция може да се изрази по следния начин: чрез линеаризация на $z_{i,k,j,l} = x_{i,k}x_{j,l}$ и/или чрез разделяне на сумата по z на под суми по различен начин.

4.1.1. 2D квадратна решетка

На базата на първият модел е разработен ефикасен евристичен алгоритъм за нагъване на протеини в хидрофобен-полярен (Hydrophobic-Polar – HP) модел за 2D квадратна решетка. Математическият модел разполага всяка аминокиселина в точка от 2D квадратна решетка (Dill, 1985; Dill, et al., 1995; Traykov, et al., 2016).

4.1.1.1. Математически модел

Математическият модел за проблема за нагъване на протеини в 2D е базиран на квадратна решетка с фиксиран размер N на страната:

- $N = n$;
- $N = 2\sqrt{n}$;
- $N = \frac{n}{2}$.

Подобни решетки лесно може да се представят като шахматна дъска, където съседите за всяка бяла (черна) клетка са четирите черни (бели) обграждащи го клетки. Клетките по границите на квадратната решетка са от степен по-малка от 4.

Математическият модел изглежда по следния начин:

$$\max \sum_{e_i, e_j \in M} z(e_i, e_j)$$

при условия:

$$\begin{aligned} \sum_{i=1}^{N^2} x_{i,k} &= 1, \quad k = 1, \dots, n \\ \sum_{k=1}^n x_{i,k} &\leq 1, \quad i = 1, \dots, N^2 \\ x_{ik} &\leq \sum_{j \in G(i)} x_{j,k+1}, \quad i = 1, \dots, N^2; k = 1, \dots, n-1 \\ x_{ik} &\in \{0,1\}, \quad i = 1, \dots, N^2; k = 1, \dots, n \end{aligned}$$

В този модел само-избягващ се път с дължина n се представя, като последователност от ходове $(a_2, a_3, \dots, a_{n-1})$, $a_i \in \{s, r, l\}$, където s -направо, r -надясно и l -наляво (в зависимост от предходния ход), при условие, че първите два елемента от последователността са фиксирани в две съседни клетки в центъра на решетката.

4.1.1.2. Алгоритъм за решаване на проблема в 2D

Нека $i = \{2, 3, \dots, n\} = \cup^k S_i S_i \cap S_{i+1} = \emptyset$ и S_i^* - подниз с дължина = $len(S_i)$ над $\{s, r, l\}$ азбука. За дадена квадратна решетка $SL = \{-n \leq i, j \leq n\}$ с елемент $a[0]$ поставен на позиция $SL[0, 0]$ и елемент $a[1]$ на $SL[0, 1]$, дадена последователност е възможна ако нейното поставяне в SL е само-избягващ се път. Нека $T_{i-1} = \cup^{i-1} S_i^*$ и $S_{i,j} = T_{i-1} \oplus st[j]$ ($\oplus$ за свързване), където $st[j]$ е низ с дължина j . Накрая, ако $j = len(S_i)$ функцията $cont(S_{i,j})$ показва броя на контактите в съответния само-избягващ се път (Traykov, et al., 2016).

Алгоритъм GREEDYLIKE:

```

1: for i in {1, 2, ..., k} do
2:   rec := 0
3:   PATHFINDER(i, 0)
4: end for
def PATHFINDER(i, m)  /Генерира всички поднизове с определена дължина.
5: if m > len(S_i) then
6:   if cont(S_{i,m}) > rec then          /* cont() брой контактите за
7:     rec := cont(S_{i,m})              получената конформация. */
8:     Запазва S_{i,m}
9:   end if
10:  return
11: end if
12: for a in {s, r, l} do
13:   st[m] := a
14:   if check(S_{i,m}) then              /* check() проверява дали получената
15:     PATHFINDER(i, m+1)               конформация е само-избягващ се път. */
16:   end if
17: end for

```

4.1.2. 3D кубична решетка

В тази част от дисертацията са представени нова целочислена формулировка на проблема за нагъване на протеини, точен алгоритъм и два евристични алгоритъма за решаване на проблема в 3D кубична решетка с размер на страната N .

4.1.2.1. Математически модел

Преобразуването на HP проблема за нагъване на протеини към оптимизационен проблем от Теория на графите позволява да се изградят различни целочислени модели. Една подходяща платформа за изграждане на целочислени модели е така наречения подравнен (Alignment) граф $G = (V_c \otimes V_s, E)$ с множество от ребра $E = \{i, k, j, l\}$, където $(i, j) \in E_c$, $(k, l) \in E_s$ (Yanev, et al., 2017).

Ако дъгите в E се нарекат z дъги, тогава проблема се свежда до намиране на път в G , който да активира максимален брой z дъги. Чрез

разделяне на сумата (4.1), би могло да се получат различни целочислени модели като следващия: нека без ограничения в общността да се приеме, че множеството $EVEN$ е с по-малък брой елементи и те са присвоени към множеството от върхове V_b . Нека $y_{i,j} \in \{0,1\}$, $i \in V_b, j \in V_w$, да съответства на сумата от z дъги между i -ти и j -ти ред (фигура 4.2). Тогава проблема е да се максимизира:

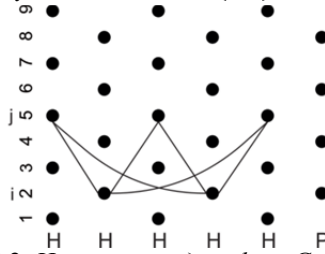
$$\sum y_{i,j}, \quad (4.1')$$

При допълнителни (към (4.2),(4.3),(4.4)) условия, определящи кога y_{ij} да бъде равно на 1:

$$v_i = \sum_{k \in EVEN} x_{i,k} \quad i \in V_b, \quad (4.5)$$

$$v_j = \sum_{k \in ODD} x_{j,k} \quad j \in V_w, \quad (4.6)$$

$$v_j \geq y_{i,j} \leq v_i. \quad (4.7)$$



Фигура 4.2: Частичен подграф на G

Частичния подграф на G , показан на фигура 4.2, има за цел да разясни връзката между $y_{i,j}$ и z дъгите (контакти), т.е. $y_{i,j}$ се използва за отчитане на сумата за тези дъги и може да приема стойност 0 или 1 (Yanev, et al., 2017).

Завършеният модел изглежда по следния начин:

$$\max \sum y_{i,j}$$

при условия:

$$\sum_{i=1}^{N^3} x_{i,k} = 1, \quad k = 1, \dots, n$$

$$\sum_{k=1}^n x_{i,k} \leq 1, \quad i = 1, \dots, N^3$$

$$x_{i,k} \leq \sum_{j \in G(i)} x_{j,k+1}, \quad i = 1, \dots, N^3; k = 1, \dots, n-1$$

$$\begin{aligned}
v_i &= \sum_{k \in \text{EVEN}} x_{i,k}, \quad i \in V_b \\
v_j &= \sum_{k \in \text{ODD}} x_{j,k}, \quad j \in V_w \\
v_j &\geq y_{i,j} \leq v_i \\
x_{i,k} &\in \{0,1\}, \quad i = 1, \dots, N^3; k = 1, \dots, n \\
y_{i,j} &\in \{0,1\}, \quad i \in V_b, j \in V_w.
\end{aligned}$$

4.1.2.2. Алгоритми за решаване на проблема в 3D

Общия вход е кубична подрешетка $d_1 \geq x \geq 0, d_2 \geq y \geq 0, d_3 \geq z \geq 0$ свързана с дължината n на входната НР последователност S . За евристичните алгоритми **HE-1** и **HE-2** последователността S се разделя на сегменти с предварително определена дължина, т.е. $S = S_1 \cup S_2 \cup \dots \cup S_k, S_i \cap S_{i+1} = \emptyset$ (Yanev, et al., 2017).

Точен алгоритъм (Exact algorithm – EA):

- 1: **If** $\text{length}(S) < t$ **then**
- 2: Стъпка 1: Изгражда се $.lp$, който съдържа модела (4.1')-(4.7).
- 3: Стъпка 2: Модела се решава чрез MIP софтуер, като CPLEX.
- 4: **End if**

Параметърът t зависи от размера на модела (броя на двоичните променливи е $mn/4$) и параметрите на компютъра.

Евристичния алгоритъм **HE-1** по-долу се базира на концепцията за разделяна на последователност S на сегменти и разделяне на куба на равнини. Евристичния алгоритъм **HE-2** се базира на концепцията за разделяна на последователността S на сегменти.

Евристичен алгоритъм 1 (Heuristic algorithm-1 – HE-1)

- 1: **For** $i = 1, 2, \dots, k$ **do**:
- 2: **If** $i = 1$ **then**
- 3: Създава се и се решава модела за вход S_i и $z = i$ (2D подрешетка)
- 4: **Else**:
- 5: Нека $v_j^* = 1$ за $j \in J$ в оптималното решение за S_{i-1} .
- 6: Стъпка 1: Създава се модела за вход S_i и $z = i$ (2D подрешетка). Добавя $\sum v_j$ към целевата функция (1). За да се запази свързаността се добавя $x_{s,1} = 1$ към множеството от ограничения, където s е номера на реда (в правления граф или координати) за последния символ от S_{i-1} в оптималното решение.
- 7: Стъпка 2: Решава се модела.
- 8: **End if**
- 9: **End for**

Евристичен алгоритъм 2 (Heuristic algorithm-2 – HE-2):

```
1: For  $i = 1, 2, \dots, k$  do:
2:   If  $i = 1$  then
3:     Създава се и се решава модела за вход  $S_i$ .
4:   Else:
5:     Стъпка 1: Създава се модела за входа  $T_i = S_1 + S_2 + \dots + S_i$ . За всяко
        $x_{i,k} = 1$  в оптималното решение за входа  $T_{i-1}$  се добавя (към
       множеството от ограничения за входа  $T_i$ )  $x_{i,k} = 1$ .
6:     Стъпка 2: Решава се модела.
7:   End if
8: End for
```

4.1.3. Дискусия

В този раздел се дискутират разработените модели и алгоритми за решаване на проблема в двумерното и тримерното пространство. Основната концепция на предлаганите евристични алгоритми за нагъване на протеини в 2D и 3D в HP модела, а именно разделянето на входната последователност на сегменти с предварително определена дължина, позволява да се достигне до решение за протеинова последователност с всякаква дължина.

Евристичният алгоритъм GREEDYLIKE (описан в раздел 4.1.1) за нагъване на протеини в 2D и евристичният алгоритъм HE-1 (описан в раздел 4.1.2) за нагъване на протеини в 3D се изпълняват в квадратна решетка (равнина в 2D), която може да бъде представена като квадратна дъска, всяка клетка от която е последователно номерирана.

4.2. Изчислителни експерименти

В тази част от дисертацията са представени резултатите генерирани от точния алгоритъм и евристичните алгоритми за двумерното и тримерното пространство описани в раздел 4.1, направен е анализ и сравнение на получените резултати с резултати, генерирани от известни в литературата алгоритми за нагъване на протеини в 2D и 3D. Изчислителните експерименти са направени с посочените в таблица 4.1 протеинови последователности.

4.2.1. Тестови примери 1: $n = 20, 24, 25, 36$

За първите изчислителни експерименти се използват къси протеинови последователности с дължини 20, 24, 25, 36 аминокиселини.

4.2.1.1. Оценка на изпълнението на алгоритмите

Първите изчислителни експерименти са проведени с евристичния алгоритъм за нагъване на протеини в 2D квадратна решетка (алгоритъма GREEDYLIKE). Изчислителните експерименти показват, че предложеният алгоритъм е коректен, тъй като получените решение

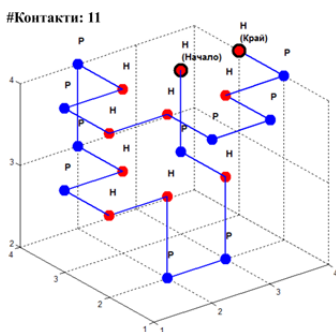
(оптималните конформации) съдържат всички аминокиселини на тестовите протеинови последователности, аминокиселините са разположени правилно в 2D квадратната решетка и тяхната първоначална последователност е запазена.

Времето за изпълнение на алгоритъма и получените контакти за всяка от разглежданите НР последователности е обобщено в таблица 4.2.

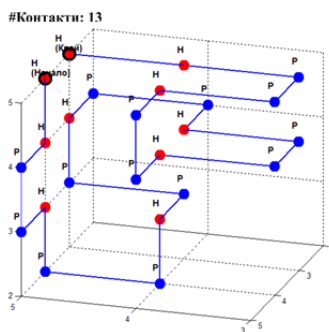
Дължина	НР Последователност	Контакти	Размер на сегментите	Време (сек.)
20	НННННННННННННННННННН	9	5	0.48
24	НННННННННННННННННННННН	9	10	11
25	РРНННННННННННННННННННННН	8	4	42
36	РРНННННННННННННННННННННН НННННННННННННН	13	10	51

Таблица 4.2: Изчислителни резултати за последователности с дължина 20, 24, 25 и 36 аминокиселини

Следващите фигури показват оптималните конформации получени при решаване на задачата в 3D за разглежданите последователности, чрез алгоритмите описани в раздел 4.1.2

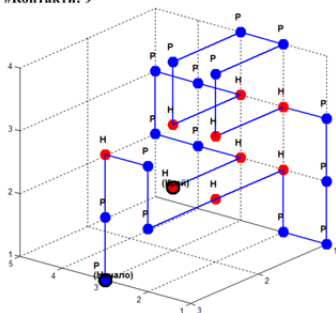


Фигура 4.2: Нагъване на последователност с дължина 20



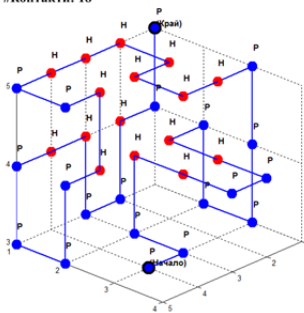
Фигура 4.3: Нагъване на последователност с дължина 24

#Контакти: 9



Фигура 4.4: Нагъване на последователност с дължина 25

#Контакти: 18



Фигура 4.5: Нагъване на последователност с дължина 36

Следващата таблица (Таблица 4.3) обобщава получените резултати от проведените числови експерименти за протеинови последователности с дължина 20, 24, 25 и 36 аминокиселини в 2D квадратна и 3D кубична решетка в HP модела.

Дължина	Контакти	
	2D квадратна решетка	3D кубична решетка
20	9	11
24	9	13
25	8	9
36	13	18

Таблица 4.3: Получени резултати в 2D и 3D за протеини с дължина 20, 24, 25 и 36

4.2.1.2. Сравнение с получени резултати от известни в литературата алгоритми

В този раздел от дисертацията се прави сравнение на получени резултати от известни в литература алгоритми за нагъване на протеини с резултатите получени от предложените алгоритми за нагъване на протеини в 2D квадратна и 3D кубична решетка в HP модела.

В Таблица 4.4 се прави сравнение между резултати генерирани от алгоритъма GREEDYLIKE (колоната **GREEDYLIKE** (Traykov, et al., 2016)), две реализации на Монте Карло алгоритъм (колоните **MC1** (Thachuk, et al., 2007) и **MC2** (Chen & Huang, 2005)), Генетичен алгоритъм (Chen & Huang, 2005; Toma & Toma, 1996)(колоната **GA**) и Mixed Search алгоритъм (Chen & Huang, 2005) (колоната **MS**). Колоната **BKS** показва най-добрите известни до

момента решения в двумерното пространство за протеинови последователности с дължина 20, 24, 25 и 36.

Дължина	Контакти					
	BKS	GREEDYLIKE	MC1	MC2	GA	MS
20	9	9	9	9	9	9
24	9	9	9	9	9	9
25	8	8	8	7	8	8
36	14	13	14	12	14	14

Таблица 4.4: Сравнение на получените резултати в 2D квадратна решетка за HP последователности с дължини 20, 24 и 25

В тримерния случай резултатите получени, чрез описаните в Раздел 4.1.2 алгоритми се сравняват с известни в литературата алгоритми като Оптимизационен алгоритъм на мравките, Генетичен алгоритъм и Еволюционен алгоритъм с връщане назад (Таблица 4.5).

Дължина	Контакти						
	BKS	GA	Backtracking-EA	ACO-metaheuristic	EA	HE-1	HE-2
20	11	11	11	10	11	11	11
24	13	13	13	8	13	13	13
25	9	9	9	6	9	9	9
36	18	18	18	10	18	18	18

Таблица 4.5: Сравнение на получените резултати в 3D кубична решетка за HP последователности с дължина 20, 24 и 25

Колоната **BKS** (Best Known Solution) в таблица 4.5 показва най-добрите известни до момента решения за всяка от тестовите HP последователности в тримерното пространство. Колоната **GA** показва резултати генерирани от Генетичен алгоритъм (Lin & Su, 2011), колоната **Backtracking-EA** показва резултати получени от Еволюционен алгоритъм с връщане назад (Cotta, 2003), а колоната **ACO-Metaheuristic** показва резултати получени чрез Оптимизационен алгоритъм на мравките (Thilagavathi & Amudha, 2015; Shmygelska & Hoos, 2005). Резултатите получени от предлаганите алгоритми се съдържат в колони **EA** (Точен алгоритъм), **HE-1**(Евристичен алгоритъм-1) и **HE-2**(Евристичен алгоритъм-2). Тук отново се вижда, че предлаганите алгоритми (**EA**, **HE-1** и **HE-2**) генерират най-добрите известни до момента решения, както останалите алгоритми.

4.2.1.3. Дискусия

В този раздел се дискутират получените резултати за къси протеинови последователности в двумерното и тримерното пространство.

Изчислителната експериментите показват, че евристичния алгоритъм GREEDYLIKE и предлагания целочислен модел са

ефикасни при решаване на задачата за къси протеинови последователности. Когато входната последователност е голяма, обикновено се получава голямо пространство за търсене, което води до генериране на много решения, които не могат да бъдат минимални енергийни конфигурации.

Предложения целочислен модел за нагъване на протеини в 3D решетъчен HP модел, се използва успешно в точен алгоритъм и е ефективен компонент за два евристични алгоритма. Изчислителни експерименти проведени с тези алгоритми показват, че без допълнителни подобрения, точният алгоритъм и двете евристики са в състояние да намерят (чрез използване на CPLEX (www.ibm.com/software/products/en/ibmilogcpleoptistud)) оптималния брой контакти за къси протеинови последователности.

4.2.2. Тестови примери 2: $n = 46, 48, 50, 60, 64$

За вторите изчислителни експерименти с HP последователности с дължини 45, 48, 50, 60 и 64 аминокиселини. Целта е да се тестват моделите и алгоритмите със средни по дължина последователности.

4.2.2.1. Оценка на изпълнението на алгоритмите

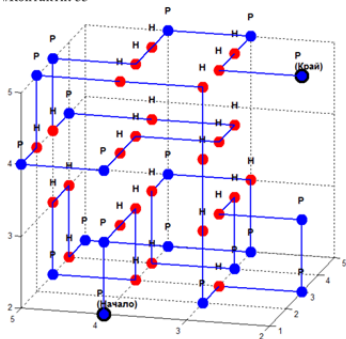
Първите изчислителни експерименти са проведени с евристичния алгоритъм GREEDYLIKE за нагъване на протеини в 2D квадратна решетка. В следващата таблица се обобщава времето за изпълнение на алгоритъма, както и броя на получените контакти за всяка една от тестовите последователности.

Дължина	HP Последователност	Контакти	Размер на сегментите	Време (сек.)
46	RRNNRRNNRRRRNNRRNNRRNNRRN NNNRNRRNNNNNNRRNNRRNNRR	21	10	26
48	RRNRNRRNNRRRRRRNNNNNNNN NNRRRRRRNNRRNNRRNNRRNNNN	23	12	143
50	NNRRNRNRRNNNNRRRRRRRRRRN RRNRNRRRRRRNNRRNNNNRRNRNRR NN	20	7	43
60	P(PH ₃) ₂ H ₅ P ₃ H ₁₀ PHR ₃ H ₁₂ P ₄ H ₆ PH ₂ PHR	35	7	4
64	H ₁₂ PHRPHR ₂ H ₂ P ₂ H ₂ P ₂ HR ₂ H ₂ P ₂ H ₂ P ₂ HR 2H ₂ P ₂ H ₂ P ₂ HRPH ₁₂	35	7	4

Таблица 4.6: Изчислителни резултати за последователности с дължина 48, 50, 60 и 64

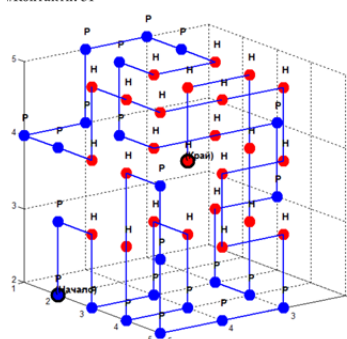
Следващите фигури показват получените нагъвания за тези тестови примери в 3D кубична решетка, генерирани чрез разработените модели и алгоритмите за нагъване на протеини описани в раздел 4.1.2.

#Контакти: 33



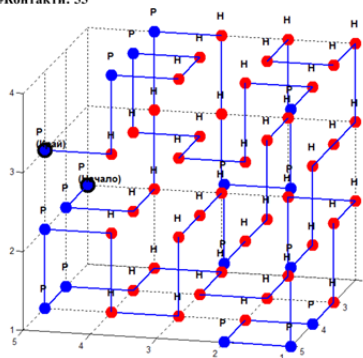
Фигура 4.6: Нагъване на протеин с дължина 46

#Контакти: 31



Фигура 4.7: Нагъване на протеин с дължина 48

#Контакти: 55



Фигура 4.8: Нагъване на протеин с дължина 60

Следващата таблица (Таблица 4.7) обобщава получените резултати от изпълнението на предлаганите алгоритми за нагъване на протеини в 3D кубична решетка в HP модела над последователности с дължина 46, 48, 50 и 60 аминокиселини.

ACO-Metaheuristic) (Shmygelska & Hoos, 2005; Thilagavathi & Amudha, 2015), Генетичен алгоритъм (колоната **GA**) (Lin & Su, 2011) и Еволюционен алгоритъм с връщане назад (колоната **Backtracking-EA**) (Cotta, 2003), както вече бе споменато в Раздел 4.2.1 (таблица 4.9).

Дължина	Контакти						
	BKS	GA	Backtracking-EA	ACO-metaheuristic	EA	HE-1	HE-2
46	32	–	–	21	33	31	31
48	29	25	25	–	31	31	28
50	26	23	23	–	28	28	28
60	49	37	39	–	55	55	52

Таблица 4.9: Сравнение на получените резултати в 3D кубична решетка за последователности с дължина 46, 48, 50 и 60

Таблица 4.9 ясно показва, че генерираните резултати от описаните алгоритми в Раздел 4.1.2 значително превъзхождат резултатите получени от останалите алгоритми. Нещо повече, за последователности с дължини 50 и 60 аминокиселини Точния алгоритъм, Евристичен алгоритъм-1 и Евристичен алгоритъм-2 генерират решения, които подобряват известните до момента решения. Също така Точния алгоритъм и Евристичен алгоритъм-1 подобряват известното до момента решение за последователност с дължина 48 аминокиселини.

4.2.2.3. Дискусия

В този раздел се дискутират получените резултати за протеинови последователности с дължини 46, 48, 50 и 60 аминокиселини в двумерното и тримерното пространство.

Изчислителните експерименти в 3D случай показват, че предлагания математически модел и базираните на него алгоритми за нагъване на протеини, генерират нагъвания, в които броя на контактите надвишава, както решенията генерирани от споменатите алгоритми за нагъване на протеини в 3D кубична решетка, така и най-добрите известни до момента решения, които са описани в литература. Това показва, че концепцията за разделя на протеиновата последователност на сегменти в комбинация с представянето и решаването на проблема за нагъване на протеини като проблем от теория на графите води до значително подобряване на получените нагъвания за къси и средни по дължина протеини.

В двумерния случай почти всички алгоритми (включително алгоритъма GREEDYLIKE) генерират решения (нагъвания) близки до най-добрите известни решения (таблица 4.8). Изключение прави само първата реализация на Монте Карло алгоритъма (колоната **MC1**),

който генерира най-добрите известни до момента решения за тези последователности.

4.2.3. Тестови примери 3: $n = 102, 123, 136$

В този раздел от дисертацията са описани изчислителни експерименти с НР последователности с дължини 102, 123 и 136 аминокиселини. Целта е да се тества качеството и изпълнението на предлаганите модели и алгоритми при работа с дълги протеинови последователности. По този начин ще се покаже до каква степен може да се разчита на предлаганите алгоритми или с други думи може ли предлаганите алгоритми да се приложат над реални протеини с голяма дължина.

4.2.3.1. Оценка на изпълнението на алгоритмите

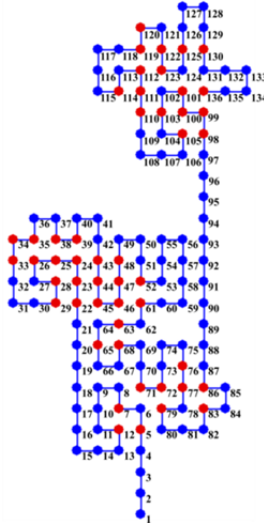
Първите тестове за тези последователности са проведени с евристичния алгоритъм GREEDYLIKE за нагъване на протеини в 2D квадратна решетка, тъй като този алгоритъм се изпълнява в двумерното пространство и вероятността да се стигне до коректно решение в разумно време е по-голяма в сравнение с алгоритмите за нагъване на протеини в 3D кубична решетка. Таблица 4.10 показва броя на получените контакти и времето за изпълнение на алгоритъма GREEDYLIKE за всяка една от тестовите последователности.

Дължина	НР Последователност	Контакти	Размер на сегментите	Време (сек.)
102	PH ₂ P ₅ H ₂ P ₂ H ₂ PHR ₂ HR ₇ HP ₃ H ₂ PH ₂ P ₆ HP ₂ HP HR ₂ HR ₅ H ₃ P ₄ H ₂ PH ₂ P ₅ H ₂ P ₄ H ₄ PHR ₈ H ₅ P ₂ HR ₂	28	7	7
123	P ₂ H ₃ PHR ₄ HP ₅ H ₂ P ₄ H ₂ P ₂ H ₂ P ₄ HR ₄ HR ₂ HR ₂ H ₂ P ₃ H ₂ PHR ₃ P ₄ H ₃ P ₆ H ₂ P ₂ HR ₂ PHR ₂ HR ₇ HR ₂ H ₃ P ₄ HR ₃ H ₅ P ₄ H ₂ PHR ₂ PHR ₂ PHR ₂	35	10	100
136	HR ₅ HR ₄ HR ₂ PH ₂ P ₄ HR ₃ P ₄ PHR ₂ HR ₄ P ₁₁ HR H ₂ HR ₃ PHR ₂ P ₃ H ₂ P ₂ HR ₂ PHR ₂ PHR ₈ HR ₃ H ₆ P ₃ H ₂ P ₂ H ₃ P ₃ H ₂ PH ₅ P ₉ HR ₄ PHR ₄	37	10	68

Таблица 4.10: Изчислителни резултати за последователности с дължина 102, 123 и 136

Фигурата по-долу показва получения резултат за последователност с дължина 136 аминокиселини след успешното приключване на алгоритъма GREEDYLIKE.

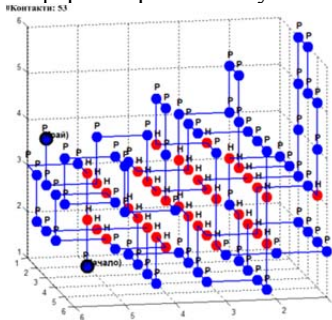
#Контакти: 37



Фигура 4.9: Нагъване на протеин с дължина 136 аминокиселини в 2D квадратна решетка

За тестване на способностите на алгоритмите за нагъване на протеини в 3D се използва само последователност с дължина 102 аминокиселини. Причината за това е големия размер на тестовите примери разглеждани в тази част от дисертацията. Тези опити ясно показват сложността на проблема и трудностите, който трябва да се преодолеят при решаване на проблема за нагъване на протеини.

Следващите фигури показват полученото решения в 3D кубична решетка за последователност с дължина 102 аминокиселини в аналитична и графична форма. Броя на получените контакти е 53.



Фигура 4.10: Графично решение за протеин с дължина 102 аминокиселини в 3D кубична решетка

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	H	N	P	P	P	P	H	N	P	P	H	N	P	P	H	N	P	P	H	N	P	P	P	P	P	H	N	P	P
z	z	y	x	y	z	y	x	x	z	y	z	y	x	y	x	z	x	y	x	z	x	z	x	y	z	x	y	z	z

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
H	N	P	H	N	P	P	P	P	P	H	N	P	P	H	N	P	H	N	P	P	P	P	P	P	H	N	P	P	P
z	x	y	z	z	y	x	x	y	y	x	z	y	z	y	z	x	y	x	z	z	y	z	x	z	z	x	z	y	z

61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
P	H	N	P	H	N	P	P	P	P	P	H	N	P	P	H	N	P	H	N	P	P	P	P	H	N	P	P	P	P
x	y	z	y	z	y	x	z	y	x	y	y	z	z	z	y	x	y	z	y	x	z	x	z	x	y	z	z	y	z

91	92	93	94	95	96	97	98	99	100	101	102
P	H	N	P	H	N	P	H	N	P	P	P
z	z	x	x	y	x	y	z	x	z	x	z

Фигура 4.11: Аналитично решение за протеин с дължина 102 аминокиселини в 3D кубична решетка

4.2.3.3. Дискусия

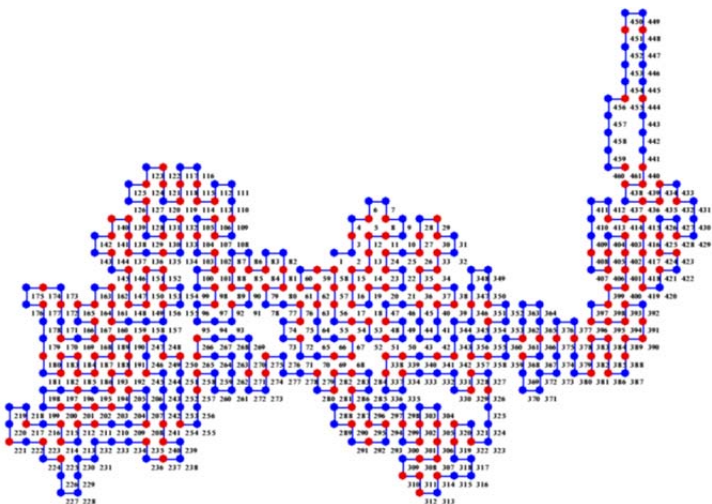
В този раздел се дискутират получените резултати за протеинови последователности с дължини 102, 123, и 136 аминокиселини в 2D квадратна и 3D кубична решетка.

Тестовите примери, които се разглеждат в този раздел не са бенчмарк последователности, т.е. в литературата няма обосновани резултати за тези последователности получени от други алгоритми за нагъване на протеини в двумерния и тримерния случай. Вземайки в предвид и факта, че моделите и алгоритми за нагъване на протеини в 3D кубична решетка в HP модела, който са описание в раздел 4.1.2 подобряват най-добрите решения за последователности с дължина 46, 48, 50 и 60, то полученото решение за последователност с дължина 102 представлява най-доброто известно до момента решение и тази последователност може да се категоризира, като нова бенчмарк последователност в 3D.

4.2.4. Тестови примери 4: Реални протеини от PDB ($n = 461, 464$)

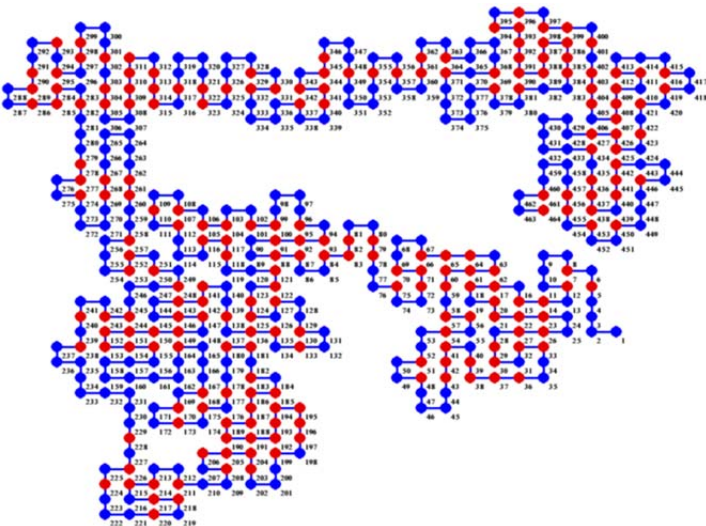
В тази част от дисертацията са описани изчислителни експерименти с реални протеини от Protein Data Bank (PDB, <http://www.rcsb.org/pdb>). Дължината на тестовите протеини е 461 и 464 аминокиселини. Поради голямата дължина на тези протеини изчислителните експерименти са проведени само в 2D квадратна решетка.

Първият експеримент е проведен с протеин с дължина 461. В Protein Data Bank този протеин е под ID 4EJ4. Той представлява делта опиоиден рецептор. Следващата фигура (Фигура 4.12) показва получения резултат след изпълнението на алгоритъма GREEDYLIKE в 2D квадратна решетка. Боря на получените Н-Н контакти е 172.



Фигура 4.12: Нагъване на протеин с PDB ID 4EJ4 в 2D квадратна решетка

Следващия експеримент е проведен с протеин с дължина 464 аминокиселини. В Protein Data Bank този протеин е под ID 4DKL. Той представлява мю опиоиден рецептор. Фигура 4.13 показва получения резултат след изпълнението на алгоритъма GREEDYLIKE. Боря на получените контакти е 166.



Фигура 4.13: Нагъване на протеин с PDB ID 4DKL

Евристичния алгоритъм GREEDYLIKE, който е описан в раздел 4.1.1 генерира решения за изследваните протеини за 2525 секунди за протеин с дължина 461 и за 2179 секунди за протеин с дължина 464.

ИЗВОДИ

1. Проблема за нагъване на протеини е представен като оптимизационен проблем от Теория на графите;
2. Предложен е целочислен математически модел за решаване на проблема в 2D и 3D;
3. Разработени са алгоритми за решаване на проблема базирайки се предложения целочислен математически модел, а именно: Евристичен алгоритъм за решаване на проблема в 2D, точен алгоритъм за решаване на проблема в 3D и два евристични алгоритъма за решаване на проблема в 3D;
4. Базирайки се на предложените алгоритми са разработени програми за намиране на оптимални конформации в 2D квадратна и 3D кубична решетка;
5. Разработена е софтуер за визуализация на оптимални конформации в 3D кубична решетка;
6. В резултат от разработения математически модел и алгоритми за решаване на проблема в 3D са подобрени описани в литература резултата за някои известни бенчмарк последователности;
7. Предложена е нова бенчмарк последователност, а именно протеин с дължина 102 аминокиселини.
8. Изчислителните експерименти показват, че идеята за разделяне на проблема на подпроблеми работи добре за дълги протеинови последователности;

СПИСЪК С ПУБЛИКАЦИИ ПО ДИСЕРТАЦИЯТА В НАУЧНИ СПИСАНИЯ

1. Yanev, N., Traykov, M., Milanov, P., Yurukov, B., Protein folding prediction in a cubic lattice in hydrophobic-polar model, *Journal of Computational Biology*, 2017, 24(5), 412-421, (Impact factor: **1.537**).
2. Traykov, M., Angelov, S., Yanev, N., A new heuristic algorithm for protein folding in the HP model. *Journal of computational biology*, 2016, 23(8), 662-668, (Impact factor: **1.537**).

СПИСЪК С ПУБЛИКАЦИИ ПО ДИСЕРТАЦИЯТА В СБОРНИЦИ ОТ НАУЧНИ КОНФЕРЕНЦИИ

1. Traykov, M., Yanev, N. and Milanov, P. Protein folding prediction in 3D cubic lattice in HP model. Proceedings of the Biomath Communications – BIOMATH2016, 2016, Blaqoevgrad, BULGARIA, 3(1), p. 67.
2. Yanev, N., Milanov, P., Yurukov, B. Trenchev, I., Traykov, M. Heuristic algorithm for HP folding in HP model. Proceedings of the MASSEE International Congress on Mathematics - MICOM-2015, 2015, Athens, GREECE, p. 31.
3. Todorin, I., Trenchev, I., Mavrevski, R., Stoilov, A., Traykov, M., Yurukov, B. Computer graphic's application in bioinformatics. Proceedings of the MASSEE International Congress on Mathematics - MICOM-2015, 2015, Athens, GREECE, p. 35.
4. Todorin, I., Trenchev, I., Stoilov, A., Mavrevski, R., Traykov, M., A model for HP folding prediction using increasing constrain for spreading in the process of making conformations. Proceedings of the Biomath Communications – BIOMATH2015, 2015, Blaqoevgrad, BULGARIA, 2(1), p. 88.
5. Todorin, I., Traykov, M., Mavrevski, R., Stoilov, A., Trenchev, I. 3D visualization of the biological structures. Proceedings of the Biomath Communications – BIOMATH2015, 2015, Blaqoevgrad, BULGARIA, 2(1), p. 89
6. M. Traykov, R. Mavrevski, A. Stoilov. Pricing of goods and services through partial differential equations. Proceedings of the Scientific Conference "Enhancement of the competitiveness and restructuring of the food subsectors through benchmarking", 15 May 2014, Blaqoevgrad, BULGARIA, 1, pp. 74-83.

УЧАСТИЕ В НАУЧНО-ИЗСЛЕДОВАТЕЛСКИ ПРОЕКТИ

1. *"Bioinformatics research: protein folding, docking and prediction of biological activity"*, NSF I02/16, 12.12.14, фонд "Научни изследвания", ръководител доц. д-р Иван Тренчев.
2. *"Компютъризирана система за получаване, изследване и анализ на електрокардиографски сигнали (ЕКГ)"*, Договор № RP-A19/16, Финансиран от: ЮЗУ „Неофит Рилски“, 2016
3. *"Математически модели и статистически методи за изследване на предприемачеството и конкурентоспособността на бизнеса в*

Югозападна България", ЮЗУ "Неофит Рилски" по Наредба 9, ръководител доц. д-р Иван Тренчев.

УЧАСТИЕ В СПЕЦИАЛИЗАЦИИ

1. Едномесечна специализация по DAAD, 15.10-15.11.2014г. в Института по математика и информатика, БАН в София, при проф. д-р Никола Янев.
2. Участие в DAAD Intensive course - New Aspects of the Time Frequency Analysis involving Frechet Frames, 28.09.16-03.10.16, Нови Сад, Сърбия.
3. Двумесечна специализация по Еразъм+, 01.03-31.04.2017г. в Университета в Рен 1, Рен, Франция, при проф. д-р Румен Андонов (IRISA).
4. Участие в ежеседмични семинари в Центъра за съвременни биоинформатични изследвания, Благоевград, България.
5. Участие в организационния комитет на Международна конференция „БИОМАТ 2015”, 14-19.06.2015г. УЦ „Бачиново”, Благоевград, България.
6. Участие в организационния комитет на Международна конференция „БИОМАТ 2016”, 19-24.06.2016г. УЦ „Бачиново”, Благоевград, България.

ЧЛЕНСТВО В НАУЧНИ ОРГАНИЗАЦИИ:

Членство в Съюза на математиците в България (СМБ)

БЛАГОДАРНОСТИ

Настоящият дисертационен труд е разработен с помощта на специалисти от Центъра за съвременни биоинформатични изследвания и Центъра за функционални изследвания в спорта и кинезитерапията, към ЮЗУ "Н. Рилски" в Благоевград. Тези центрове са създадени и финансирани по проекти DVU 01/197/2008г. от НФ "Научни изследвания" и DOO2-135/2009г. (COST ACTION 0801), ICoSCIS, YMC(H)A - проекти по ПТС „България-Гърция 2007-2013”.

Изказвам своята благодарност на научния ми ръководител проф. д-р Петър Миланов за помощта и подкрепата при разработването и оформянето на настоящия дисертационен труд, както и за съветите, конструктивната критика и търпението, без които този труд ни би бил възможен.

Изказвам своята благодарност на проф. д-р Никола Янев от Института по математика и информатика към БАН, за оказаното съдействие и помощ при извършване на експерименталните изследвания и съвместната ни работа при изготвянето на публикации.

БИБЛИОГРАФИЯ

1. Aarts, E. & Lenstra, K., 2003. *Local Search in Combinatorial Optimization*. н.м.:Princeton University Press.
2. Ahn, N. & Park, S., 2010. Finding an upper bound for the number of contacts in hydrophobic-hydrophilic protein structure prediction model. *Journal of Computational Biology*, Том 17, pp. 647-656.
3. Alberts, В. и др., 1998. *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. н.м.:Garland Publishing Inc..
4. Argos, P., 1987. A sensitive procedure to compare amino acid sequences. *J. Mol. Biol.*, Том 193, pp. 385-396.
5. Backofen, R. & Will, S., 2001. *Optimally compact finite sphere packings hydrophobic cores in the FCC*. Berlin, н.а.
6. Backofen, R., Will, S. & Clote, P., 2000. *Algorithmic approach to quantifying the hydrophobic force contribution in protein folding*. Honolulu, н.а., pp. 93-106.
7. Barton, G. & Sternberg, M., 1987. A strategy for the rapid multiple alignment of proteins. *J. Mol. Biol.*, 198(1), pp. 329-337.
8. Caprara, A. и др., 2004. 1001 optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap. *Journal of Computational Biology*, Том 11, pp. 27-52.
9. Caprara, A. & Lancia, G., 2002. *Structural alignment of large-size proteins via lagrangian relaxation*. н.м., н.а., pp. 100-108.
10. Chandru, V., Dattasharma, A. & Kumar, V., 2003. The algorithmics of folding proteins on lattices. *Discrete Applied Mathematics*, Том 127, pp. 145-161.
11. Chen, M. & Huang, W., 2005. A branch and bound algorithm for the protein folding problem in the HP Lattice Model. *Genomics Proteomics Bioinformatics*, 3(1), pp. 225-230.
12. Corman, H., Leiserson, E., Rivest, L. & Stein, C., 2001. *Introduction to Algorithms*. Cambridge(Massachusetts): The MIT Press.
13. Cotta, C., 2003. Protein structure prediction using evolutionary algorithms hybridized with backtracking. *Artificial Neural Nets Problem Solving Methods*, Том 2687, pp. 321-328.
14. Dayhoff, M. & Eck, R., 1968. *A model forevolution change in proteins..* Washington, DC, National Biomedical Research Foundation, pp. 33-41.
15. Dayhoff, M., Schwartz, R. & Orcutt, B., 1978. *A model for evolutionary change*. Washington, DC, National Biomedical Research Foundation, pp. 345-358.
16. Dill, K., 1985. Theory for the folding and stability of globular proteins. *Biochemistry*, Том 24, pp. 1501-1509.

17. Dill, K. и др., 1995. Principles of protein folding - a perspective from simple exact models. *Protein Sci*, Том 4, pp. 561-602.
18. Doolittle, R., 1979. Protein evolution. От: *The Proteins*. New York: Academic Press, pp. 1-118.
19. Eidhammer, I., Jonassen, I. & Taylor, W., 2000. Structure comparison and structure patterns. *Journal of Computational Biology*, Том 7, pp. 685-716.
20. Erman, B. & Dill, K., 2000. Gaussian model of protein folding. *Journal of Chemical Physics*, Том 112, pp. 1050-1056.
21. Feng, D. & Doolittle, R., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, Том 25, pp. 351-360.
22. Feng, D. & Doolittle, R., 1990. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.*, Том 183, pp. 375-387.
23. Feng, D., Johnson, M. & Doolittle, R., 1985. Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, Том 21, pp. 112-125.
24. Fitch, W., 1966b. An improved method of testing for evolutionary homology. *J. Mol. Biol.*, Том 16, pp. 9-16.
25. Fitch, W., 1970. Further improvements in the method of testing for evolutionary homology among proteins. *J. Mol. Biol.*, Том 49, pp. 1-14.
26. Fitch, W. & Margoliash, E., 1967. Construction of phylogenetic trees. *Science*, Том 15, pp. 299-304.
27. Fredman, M., 1984. Computing evolutionary similarity measures with length independent gap penalties. *Bull. Math. Biol.*, Том 46, pp. 553-566.
28. Goldman, D., Istrail, S. & Papadimitriou, C., 1999. *Algorithmic aspects of protein structure similarity*. Н.М., IEEE Computer Society Press.
29. Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. *Science*, Том 185, pp. 862-864.
30. Greenberg, H., Hart, W. & Lancia, G., 2004. Opportunities for combinatorial optimization in computational biology. *INFORMS Journal on Computing*, Том 16, pp. 211-231.
31. Hart, W. & Istrail, S., 1995. *Fast protein folding in the Hydrophobic-Hydrophilic model within three-eighths of optimal (extended abstract)*. Н.М., Н.А., pp. 157-168.

32. Heun, V., 2003. Approximate protein folding in the HP side chain model on extended cubic lattices. *Discrete Applied Mathematics*, Том 127, pp. 163-177.
33. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A., 2008. Searching protein structure databases with DaliLite v.3. Том 24, p. Bioinformatics.
34. Holm, L. & Sander, C., 1995. *3-D lookup:fast protein structure database seqches at 90 reliability*. Cambrige, UK., н.а., pp. 179-187.
35. Irback, A., Peterson, C., Potthast, F. & Sommelius, O., 1997. Local interactions and protein folding: A three-dimensional off-lattice approach. *Journal of Chemical Physics*, Том 107, pp. 273-282.
36. Irback, A. & Sandelin, E., 1998. Local interactions and protein folding: A model study on the square and triangular lattices. *Journal of Chemical Physics*, Том 108, pp. 2245-2250.
37. Jue, R., Woodbury, N. & Doolittle, R., 1980. Sequence homologies among E. coli ribosomal proteins: evidence for evolutionary related groupings and internal duplications. *J. Mol. Evol.*, Том 15, pp. 129-148.
38. Lau, K. & Dill, K., 1989. A lattice statistical mechanics model of the conformation and sequence spaces of proteins. *Macromolecules*, Том 22, pp. 3986-3997.
39. Lesk, A., 2001. *Introduction to Protein Architecture*. Oxford: Oxford.
40. Lin, C. & Su, S., 2011. Protein 3D HP Model Folding Simulation Using a Hybrid of Genetic Algorithm and Particle Swarm Optimization. *Int. J. Fuzzy Syst.*, Том 13, pp. 140-147.
41. Marinari, E. & Parisi, G., 1992. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, Том 19, pp. 451-458.
42. McLachlan, A., 1971. Tests for comparing related amino acid sequences: cytochrome c and cytochrome. *Mol. Biol.*, Том 61, pp. 409-424.
43. Metropolis, N. и др., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, Том 21, pp. 1087-1092.
44. Michalewicz, Z. & Fogel, B., 2004. *How to Solve It: Modern Heuristics*. Berlin: Springer.
45. Miller, J., Zeng, C., Wingreen, N. & Tang, C., 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. *PROTEINS: Structures, Function, and Genetics*, Том 47, pp. 506-512.
46. Miyata, T., Miyazawa, S. & Yasunaga, T., 1979. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, Том 12, pp. 219-236.

47. Murzina, A., Brenner, S., Hubbard, T. & Chothia, C., 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, Том 247, pp. 536-540.
48. Needleman, S. & Wunsch, C., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, Том 48, pp. 442-453.
49. Newman, A., 2002. *A new algorithm for protein folding in the HP model*. н.м., SODA, pp. 876-884.
50. Orengo, C. и др., 1997. CATH: A hierarchic classification of protein domain structures. *Structure*, pp. 1093-110.
51. Schmitt, W. & Waterman, M., 1994. Linear trees and rna secondary structure. *Discrete Applied Mathematics*, Том 51, pp. 317-323.
52. Shmygelska, A. & Hoos, H., 2005. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics*, Том 6, pp. 30-52.
53. Taylor, W. & Orengo, C., 1989. Protein structure alignment. *J. Mol. Biol.*, Том 208, pp. 1-22.
54. Thachuk, C., Shmygelska, A. & Hoos, H., 2007. A replica exchange Monte Carlo algorithm for protein folding in the HP model. *BMC Bioinformatics*, Том 8, pp. 342-362.
55. Thilagavathi, N. & Amudha, T., 2015. Aco-metaheuristic for 3D-hp protein folding optimization. *ARPJ Journal of Engineering and Applied Sciences*, Том 10, pp. 4948-4953.
56. Thompson, J., Higgins, D., Gibson, T. & Clustal, W., 1994. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position - specific gap penalties and weight matrix choice. *Nucleic Acids Research*, Том 22, pp. 4673-4680.
57. Toma, L. & Toma, S., 1996. Contact interactions method: A new algorithm for protein folding simulations. *Protein Sci.*, Том 5, pp. 147-153.
58. Traykov, M., Angelov, S. & Yanev, N., 2016. A new heuristic algorithm for protein folding in the hp model. *J. Comp. Biol.*, Том 23, pp. 662-668.
59. Unger, R. & Moult, J., 1992. Genetic Algorithms for Protein Folding Simulations. *J. Mol. Biol.*, Том 231, pp. 75-81.
60. Waterman, M., 1978. Combinatorics of rna hairpins and cloverleaves. *Studies in Appl. Math.*, Том 60, pp. 91-96.
61. Yanev, N., Andonov, R., Veber, P. & Balev, S., 2008. Lagrangian Approaches for a class of Matching Problems in Computational Biology. *Comput. Math. Appl.*, Том 55, pp. 1054-1067.

- 62. Yanev, N., Milanov, P. & Mirchev, I., 2011. Integer programming approaches to HP folding. *Serdica J. Computing*, Том 5, pp. 359-366.
- 63. Yanev, N., Traykov, M., Milanov, P. & Yurukov, B., 2017. Protein folding prediction in a cubic lattice in hydrophobic-polar model. *J. Comp. Biol.*, Том 24, pp. 412-421.
- 64. Сигал, И. & Иванова, А., 2002. *Введение в прикладное дискретное программирование: модели и вычислительные алгоритмы*. Москва: ФИЗМАТЛИТ.

ДЕКЛАРАЦИЯ ЗА ОРГИНАЛНОСТ

Декларирам, че настоящата дисертация съдържа оригинални резултати, получени при проведени от мен научни изследвания (с подкрепата и съдействието на научните ми ръководители). Резултатите, които са получени, описани и/или публикувани са описани в раздел „Библиография”.

Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Подпис:.....
/Методи Трайков/