НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ МОЛЕКУЛЯРНОЇ БІОЛОГІЇ ТА ГЕНЕТИКИ

На правах рукопису
УДК 577.112 + 371.24

МАЛЬЧЕНКО СЕРГІЙ ЗАХАРОВИЧ

ДОСЛІДЖЕННЯ ДЕЯКИХ ЗАКОНОМІРНОСТЕЙ В ОРГАНІЗАЦІЇ ВТОРИННОЇ СТРУКТУРИ ГЛОБУЛЯРНИХ БІЛКІВ

Спеціальність 03.00.03 - молекулярна біологія

А в т о р е ф е р а т
дисертації на здобуття наукового ступеню
кандидата біологічних наук

Київ - 1995

Робота виконана у лабораторії біоінженерії Інституту молекулярної біології та генетики НАН України.

Науковий керівник - кандидат біологічних наук Н.О. Чащін

Науковий консультант - Кандидат фізико-математичних наук В.І.Данілов

Офіційні опоненти - доктор біологічних наук Ю.Л. Радавський

кандидат біологічних наук О.І. Корнелюк

Провідна організація - Інститут біохімії ім. О.В.Палладіна НАН України, м.Київ

Захист відбудеться 27.02 1995 р. о 10 год. на засіданні спеціалізованої ради Д.016.11.01 Інституту молекулярної біології та генетики НАН України за адресою: 252143, Київ-143, вул. Заболотного, 150.

З дисертацією можна ознайомитися в бібліотеці Інституту молекулярної біології та генетики НАН України за адресою: 252143, Київ-143, вул. Заболотного, 150.

Автореферат розіслано "24" 01 1995 р.

Вчений секретар
спеціалізованої ради
кандидат біологічних наук Л.Л. Лукаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність проблеми. Вивчення просторової структури білків відіграє важливу роль в рішенні фундаментальних та прикладних питань молекулярної біології, біохімії та біофізики. Однією з найважливіших проблем в цій області є проблема передбачення структури білків, тобто встановлення відповідності між амінокислотною послідовністю білкових ланцюгів та їх унікальним просторовим укладанням.

На сьогодні відсутні точні теоретичні схеми та практичні розробки для прямого передбачення третинной структури білків за амінокислотною послідовністю. В зв'язку з цим здійснюються спроби рішення цієї проблеми через передбачення вторинної структури білків. Проте за останні 10 років точність методів передбачення вторинної структури зросла всього лише на 10% і складає в середньому 65% (табл. I). Тому на сьогоднішній час продовжується пошук та розробка нових підходів і алгоритмів для вирішення цієї проблеми, зокрема із застосуванням експертних систем.

Передбачення вторинної структури білків за їх первинною послідовністю має самостійне значення та широку область практичного застосування при визначенні функціонально важливих ділянок білків з нерозшифрованою просторовою структурою, при встановленні спорідненості первинних та вторинних структур, для визначення належності передбачного білка до конкретної функціональної або структурної родини. Рішення цієї проблеми відкриває нові можливості для сучасної біотехнології та білкової інженерії при використанні білків з заданою просторовою структурою та визначеними якостями.

Мета та завдання дослідження. Метою даної роботи є вивчення та виявлення закономірностей, властивих вторинній структурі глобулярних білків, на основі існуючих баз даних просторових структур, а також розробка пакетів прикладних програм для передбачення вторинної структури білків.

Для досягнення поставленої мети необхідно було розв'язати наступні завдання:

1) провести порівняльний аналіз існуючих методів передбачення вторинної структури білків та спробувати визначити причини недостатньо високої завбачної сили даних методів;

2) удосконалити та адаптувати метод, відомий в області експертних систем як GUHA метод (generation of unary hypoteses automatically - автоматичне утворення унарних гіпотез), для передбачення вторинної структури глобулярних білків, та програмно реалізувати цей метод для IBM PC сумісних комп'ютерів;

3) проаналізувати гомо- та гетерогенні короткі повторюючиєся послідовності амінокислот на предмет виявлення потенційних информаційних залишків;

4) на конкретному прикладі бактеріальных 5-метілцитозин метілтрансфераз (ЕС 2.1.1.73) продемонструвати пристосованність модифікованого GUHA методу та ряду розроблених оригінальних програм для дослідження функціональних сайтів білків та їх співвідношення з просторовою структурою білка.

Наукова новизна та практична цінність роботи. В процесі проведених досліджень отримано наступні результати:

- проаналізовано причини недостатньої завбачної сили існуючих методів передбачення вторинної структури білків, та запропоновано можливі шляхи подальшого прогресу в збільшенні

точності передбачення вторинної структури;

- на основі аналізу схожості первинних послідовностей білків вперше визначено точні межі співпадання вторинної структури при визначеному відсотку гомології первинних послідовностей як для гомологічних, так і для негомологічних білків;

- показано, що існують групи амінокислот, які мають тенденцію до утворення визначеної вторинної структури. При вирівнюванні гомологічних послідовностей та визначенні їх структури дані амінокислоти виграють критичну роль, демонструючи здатність практично при однаковому контексті руйнувати короткі регулярні структури;

- на основі модифікованого GUHA методу створена програма розрахунку вторинної структури глобулярних білків для IBM PC сумісних компьютерів із середньою точністю передбачення 68%;

- використовуючи розроблений нами модифікований GUHA метод, метод пошуку зразків та ряд оригінальних програм, в сайт-впізнаючому домені метілтрансфераз вперше виявлено п'ять мотивів у первинній послідовності цих ферментів та показано їх співпадіння з функціональними мотивами на вторинній структурі, а також показано наявність в сайт-впізнаючому домені метілтрансфераз в-а-в структурного домена, який присутній на багатьох ДНК-зв'язуючих білках.

Апробація роботи. Матеріали дисертації доповідалися на Міжнародних конференціях "Моделирование и компьютерные методы в молекулярной биологии и генетике" (Новосибирск, 1990), "Bioinformatics in the 90's" (Maastricht, The Netherlands, 1991), "FEBS'92" (Dublin, UK, 1992), "Protein engineering and beyond" (Miami, USA, 1993), на семінарі Національного агротехнічного інституту (INRA, Paris, France, 1991).

Особистий внесок автора. В роботі наведено результати досліджень, виконаних безпосередньо автором.

Публікації. За темою дисертації опубліковано 7 робот.

Структура та об'єм дисертації. Дисертація складається з наступних розділів: вступу, трьох глав - "Огляд літератури", "Матеріали та методи", "Результати та обговорення", висновків та списку використаної літератури, який включає 105 бібліографічних посилань. Роботу викладено на 111 сторінках машинопису, вона містить 7 таблиць, 13 малюнків.

ОСНОВНИЙ ЗМІСТ РОБОТИ.

У вступі обгрунтовано актуальність теми дисертації, викладено мету та завдання, відображено наукову новизну та практичну цінність досліджень.

В першій главі "Огляд літератури" розглянуто сучасні уявлення про структурну організацію глобулярних білків, надано опис та аналіз існуючих методів передбачення вторинної структури білків. На основі аналізу літературних даних зроблено висновок, що найбільш точне передбачення вторинної структури білків досягається за допомогою комбінованих методів або експертних систем (табл.1). Однак, навіть настільки складні методи аналізу вторинної структури не підіймаються в точності передбачення вище 70%. Для визначення можливих причин даного процесу проведено порівняльний аналіз методів передбачення. Узагальнюючі дані, отримані різними методами, можна виділити наступні закономірності:

1) Показана статистично значима переважність знаходження амінокислот, дуплетів та ділянок білка в визначенній конформації.

2) З результатів досліджень виходить, що зустрічаються ділянки білка, які несуть визначене інформаційне навантажен-

Таблиця I. Точність передбачення, яку розраховано для трьох конформацій в сучасних методах передбачення вторинної структури білків.

Клас методів	Точність передбачення, %
Статистичні методи	
Чоу-Фасмана	51
GOR (Гарніє та ін.)	55
CORIII (Гібрат та ін.)	63.3
Методи схожості послідовностей	
Нишикава та Ооі	60
Світ	59
Левін та Гарнієр	63
Методи, засновані на фізичних моделях	
Бью та ін.	59
Лім	56
Фількенштейн-Птіцин	63
Холли та Карплюс	63.2
Методи шаблонів	
Тоха та ін.	62
Руман та Водак	64
Комбіновані методи	
Нейронна сіть (Кнеллер та ін.)	64.3
COMBINE (Бью та ін.)	65.5
Соловйов та ін.	63
Нишикава та ін.	67.4
GUHA метод (Братусь та ін.)	68

ня в детермінації вторинної структури.

В останньому розділі глави "Огляд літератури" розглядаються загальні властивості бактеріальних 5-метілцитозин метілтрансфераз та обговорюється сайт-впізнаючий домен метілаз як модельна система для дослідження функціональних сайтів глобулярних білків та їх співвідношення з просторовою структурою.

Глава "Матеріали та методи" складається з двох частин.

2.1. Банки даних та первинні послідовності.

Для вивчення та виявлення закономірностей, властивих вторинній структурі білків, використували інформацію про вторинну структуру 108 білків (20.000 амінокислотних залишків) з рентгеноструктурним дозволом меншим, ніж 0.2 нм. Дані про вторинну структуру білків отримано з Брукхевенського банку даних просторових структур білків.

Для дослідження функціональних сайтів білків та їх співвідношення з просторовою структурою, 26 послідовностей 5-метілцитозин метілтрансфераз отримано з банку даних SWISS-PROT Європейської молекулярно-біологічної лабораторії, версія 18.0.

2.2. Програмне забезпечення.

Аналіз первинних послідовностей та вторинної структури білків проводився з використанням оригінальних авторських програм, а також програм, доступних через базу даних DOS_SOFTWARE (EMBL).

Для вирівнювання первинних послідовностей та визначення гомологічних районів в білках використали програму MULTALIN (Корпет, 1988); в випадку аналізу банка даних вторинних структур для пошуку гомологічних ділянок та інформаційних фрагментів білків - оригінальну програму BankSearch. Оригінальний

метод пошуку образів використовувався для передбачення білкових сайтів. Передбачення вторинної структури глобулярних білків здійснювалося з використанням модифікованого GUHA метода.

Глава "Результати та обговорення" складається з п'яти частин.

3.1. Модифікація GUHA метода для передбачення вторинної структури глобулярних білків.

Для передбачення вторинної структури білка використували метод, відомий в області експертних систем як GUHA метод. Раніш було показано, що в традиційному GUHA методі перепускалися важливі знання за рахунок неповного визначення критеріїв перевірки на важливість знання. Вдосконалення GUHA методу містилося в розробці таких критеріїв визначення важливості знання, коли задовольняється основний принцип індукції - висновок усіх суттєвих знань. Суть методу полягає в наступному. Припустимо, досліджувана предметна область подана емпірічними даними у вигляді таблиці і дано два стани предметної області, які описуються за формулами P1 та P2. Кожна формула описує конкретний стан предметної області. Використовуючи таблицю емпірічних даних, підраховується 1) кількість усіх спостережень, які містять два стани; 2) кількість усіх спостережень, які містять один стан; 3) кількість усіх спостережень,які містять інший стан. Після обробки отриманих частот за одним з показників, можна перевірити гіпотезу про взаємозалежність даних станів. Якщо формула P1 описує мету дослідження,а P2 пробігає множину станів предметної області, то можна отримати множину станів, пов'язаних з метою дослідження. Якість гіпотез, що добираються,залежить від критерію,

який застосовується. В даній роботі вибрано точний критерій перевірки на незалежність двох ознак (критерій Фішера). Точне значення критерія Фішера називають критичним рівнем. Критичний рівень відповідає імовірності помилковості даної гіпотези, тому, чим меншим є критичний рівень, тим імовірнішим є взаємозв'язок між станами гіпотези. Цей факт використано при прогнозуванні вторинної структури. Здійснено програмну реалізацію модифікованого GUHA методу для IBM PC сумісних компьютерів та зроблено адаптацію даного методу для передбачення вторинної структури глобулярних білків. Вторинна структура класифікована за трьома конформаціями: а-спіраль (h), в-складка (e) та нерегулярна структура (c). Таким чином, кожному амінокислотному залишку відповідає один з трьох станів вторинної структури (h,e,c). Завдання передбачення вторинної структури білка формулюється наступним чином. Припустимо, задано ділянку білка з m залишків і треба визначити, в яку структуру вбудовується i-й залишок даної ділянки. Для цього з банку даних залучаються все послідовності довжиною m, які містять на i-му місці заданий амінокислотний залишок. Далі в i-е місце кожної послідовності замість залишка заноситься значення вторинної структури цього залишка. Отримані послідовності, записані одна за одною, створюють таблицю емпірічних даних, i-й стовбчик якої складається з значень вторинної структури залишку, який визначається. Інші стовбці складають відповідні залишки (всього - m стовбчиків). Якщо в якості стану мети P1 задати стан а-спіраль, тоді модифікований GUHA метод запропонує всі комбінації контекстних залишків, які пов'язані з вбудовою i-го залишка в а-спіраль. Аналогічно отримуються результати для інших значень вторинної

структури. В підсумку формується множина гіпотез вбудови даного залишка на вторинну структуру. Критерій передбачення такий: значення вторинної структури, що прогнозується, таке, котре дає найменше значення критичного рівня на усіх отриманих гіпотезах. Передбачення для білка довольної довжини здійснюється послідовним прогнозуванням кожного залишка в контексті (m-1) відповідних залишків.

Для практичного запровадження методу необхідно визначити параметри: m-довжину послідовності залишків; i-розположення залишка, який прогнозується. Емпірічним шляхом встановлено наступні оптимальні значення: m = 5 та i = 1. Для визначення точності передбачення тестувались всі 108 білків з бази даних вторинних структур. Прогнозуємий білок вилучається з загального набораy. Оцінка точності передбачення за окремими конформаційними станами розраховувались за формулою (Гарніе, Робсон, 1989): Qs = 1/2[(Ts+ /Ts) + (Ts-+/Ts-)], де s - конформаційний стан (h,e,c); (Ts) - кількість залишків в структурі s за даними рентгеноструктурного (р/с) аналізу; (Ts+) - кількість залишків в структурі s, які співпадають за передбаченням з даними р/с аналіза; (Ts-) - кількість залишків, які не входять до структури s за даним р/с аналізу; (Ts-+) - кількість залишків, що не входять в структуру s та співпадають за передбаченням з даними р/с аналіза. Загальна достовірність оцінювалася за формулою: Q = [(Th+)+(Te+)+(Tc+)]/N, де N - загальна кількість залишків. Середня точність передбачення за використаним в роботі банком даних складала для a-спіралі - 74 %, в-складки - 67 %, нерегулярної структури - 71 % та загальна для трьох конформацій - 68 %.

3.2. Розробка метода пошуку зразків для передбачення біл-

кових сайтів.

Для передбачення білкових сайтов в основному використовуються три методи - консенсус, метод вагових матриць та пошук патернів (шаблонів). Найбільша точність передбачення сайтів досягається при використанні методу вагових матриць та складає 80-90% істиних та 30-40% помилкових передбачень сайту. Критичним елементом в методі вагових матриць є вибір лімітуючого вагового коефіцієнту, який не є мінімальним для даної множини послідовностей. Це дозволяє значно скоротити кількість помилкових передбачень, але в той же час зменшує кількість істиних передбачень.

Нами розроблено оригінальний метод передбачення білкових сайтів на основі об'єднання методів вагових матриць та пошуку патернів, який названо методом пошуку образів. Важливим елементом даного метода є вбудована експертна система, яка здійснює аналіз результатів, що отримані за допомогою вагової матриці та пошуку патернів. Точність передбачення сайтів при використанні метода пошуку образів складає 95-100% істиних та 0-5% помилкових передбачень сайта, що дозволяє досить точно виділяти сайт, який цікавить дослідника, з будь якої множини послідовностей. Даний метод було використано для точного передбачення лівої та правої межи сайт-впізнаючого домена 5-метілцитозину метилаз.

3.3. Аналіз коротких амінокислотних послідовностей, що повторюються, та пошук інформаційних сайтів білків.

Для пошуку інформаційних сайтів білків нами проведено оці- нку статистичних характеристик білкових послідовностей та аналіз їх стрічання на вторинній структурі на основі бази даних з 108 білків, отриманих з Брукхевенського банку даних.

В даній роботі під гомогенними повторними послідовностями розумі- ються послідовності типу Xn, де X-одна з природних амінокислот; n - ступінь повторності. Під гетерогенними повторними послідовностями розуміється послідовність типу (XY)n, де X,Y -два різних залишки з 20 природних амінокислот, n - ступінь повторності. Гетерогенна повторна послідовність утворюється при об'єднанні трипептидів типу XYX та YXY. Для оцінки інформаційного значення коротких послідовностей, які повторюються, використовувалася така стратегія - послідовність вважалася інформаційно значимою для конформації, якщо всі залишки з цієї послідовності знаходилися в даній конформації. Якщо ж послідовність з таким же амінокислотним складом визначалася і в інтій конформації, тоді досліджувана послідовність вважалася інформаційно незначимою.

У випадку аналіза гомогенних повторюючихся послідовностей, ступінь повторності, виявлений нами у використаній базі даних, складав від 2 до 4, тобто визначені послідовності типу XX, XXX та XXXX (дипептиди, трипептиди та тетрапептиди), відомості про зустрічаємість котрих на вторинній структурі подані в табл.2. Аналіз даних таблиці виявляє ряд цікавих фактів. По-перше, при збільшенні довжини повторних послідовностей виділяються чотири аліфатичних залишка - G, A, V та L, які мають високу здатність до створення нерегулярної конформації, a-спіралі та в-складки, відповідно. По-друге, нами знайдено 13 залишків з повним набором якостей, необхідних для утворення функціональних сайтів білків (гідрофобні та гідрофільні, негативно та позитивно заряджені) та з тенденцією знаходитися переважно в одній з трьох основних

Таблиця 2. Дані про стрічання послідоностей типу X у вторинній структурі білків.

Пептид	Вторинна структура	Кількість в БД	Інформативність
Дипептид			
CC	ee,ce,ec,hh	5 1 1 1	E*
MM	hh,ce,ee,ec	13 2 2 1	H*
HH	cc,hh	4 2	CH
WW	hh,he,cc	5 1 1	HC
FF	hh,ee,hc,ec,cc	15 3 2 2 1	H*
PP	cc,hh,ee	31 1 1	C*
QQ	hh,ee,cc,ce,hc,ch,ec	14 7 3 2 1 1 1	CHE
Трипептид			
GGG	ccc,cce,cee	11 1 1	CE
AAA	hhh,ccc,hhc,eec	14 4 2 2	H*
III	eee,eec	1 1	EC
VVV	eee,cee,eec	6 2 1	EC
SSS	ccc,chh,eee,cee,eec, ecc,cce	9 3 2 1 1 1 1	C*
TTT	ccc,eec,eee,cee	2 2 1 1	CE
LLL	hhh,eee,cee,ecc	9 3 1 1	H*
PPP	ccc	1	C
RRR	ccc,eee	1 1	CE
NNN	ccc	1	C
DDD	ccc	4	C
EEE	hhh	2	H
KKK	hhh,ccc,cch	2 1 1	HC
YYY	ccc,cee,eec	1 1 1	CE
Тетрапептид			
GGGG	cccc	1	C
AAAA	hhhh	1	H
VVVV	ceee	2	E
SSSS	cccc,ceee,eecc	1 1 1	CE
LLLL	ceee	1	E

БД -база даних; H(h), E(e) і C(c) - a-спіраль, в-складка і нерегулярна структура. *Структура має переважне значення, однак в цьому же контексті можуть зустрічатися й інші структури.

конформацій вторинної структури.

В випадку аналіза гетерогенних повторних послідовностей нами виявлені наступні інформаційні характеристики залишків (табл.3): 1) залишки G, A та V є найсильнішими (за тенденцією до утворення визначеної конформації) в своїх структурних групах триплетів; 2) виявлено інформаційно значущі залишки для а-спіралі - A,E та F; в-складки - V; нерегулярної структури - G, D, P, K та N. Ці дані повністю співпадають з аналогичною інформацією, отриманою для гомогенних послідовностей, які повторюються. Таким чином, можна зробити висновок про те, що існує ряд залишків, (табл. 2, 3), які несуть більше інформаційне навантаження в детермінації вторинної структури як для гомо- так і для гетерогенних повторних послідовностей залишків. Такі амінокислоти було нами названо іформаційно значимими.

Отримані дані про іформаційно значимі амінокислоти можна використовувати при аналізуванні гомології послідовностей білків та співпадіння їх вторинних структур, наприклад, для вивчення впливу даних амінокислот на утворення або руйнування коротких елементів вторинної структури білків. Для цього необхідно було провести аналіз гомології послідовностей білків та співпадіння їх вторинних структур.

3.4. Аналіз гомології послідовностей білків та співпадіння їх вторинних структур.

На сьогодішній день сильне заперечення проти методів передбачення гомології послідовностей міститься в роботі (Кабаш, Сандер, 1983), в якій показано, що ідентичні пентапептиди можуть знаходитися одночасно в різних конформаціях вторинної структури з імовірністю, яка перевищує випадкове спі-

Таблиця 3. Дані про стрічання трипептидов типу XYX і YXY, які утворюють регулярну послідовність типу (XY)

Структура (трипептид)	Кількість в БД	1	2	Інформативність
1	2	3	4	5
а-спіраль				
AMA ACA AEA AWA AYA	2 1 7 2 4	16	5	+
LAL LQL	4 1	5	2	
ICI IDI IPI	1 1 1	3	3	
VMV VDV	1 1	2	2	
TMT	2	2	1	
MVM MSM MQM MRM	1 2 1 1	5	4	?
CNC	1	1	1	
DRD	1	1	1	
ELE EQE EHE EYE	3 3 1 1	8	4	+
QLQ QEQ QRQ QFQ	2 1 1 1	5	4	?
KQK KYK	1 1	2	1	
RTR RKR RHR RNR	1 2 1 2	6	4	?
HAH HTH	1 1	2	2	
WGW WRW	1 1	2	2	
FAF FLF FMF FKF	3 1 1 2	7	4	+
YDY	4	4	1	
PFP	1	1	1	
NFN	1	1	1	
в-складка				
LHL	1	1	1	
IAI IMI IWI	1 1 1 1	4	3	?
VIV VQV VYV	3 2 1	6	3	+
TDT TFT	1 1	2	2	
MTM MKM	1 1	2	2	
CWC	1	1	1	
DHD	1	1	1	
KIK KWK KPK	2 1 1	4	3	?
RGR RSR	1 1	2	2	
HVH HFH	1 1	2	2	

Продовження табл.3

1	2	3	4	5
FRF FHF FYF	1 1 1	3	3	?
YTY YCY YFY	1 1 1	3	3	?
NYN	1	1	1	
Нерегулярна структура				
GLG GIG GTG GMG GQG GKG	5 5 4 2 3 5	37	10	+
GRG GWG GFG GPG	5 2 3 3			
LCL	1	1	1	
SES SQS SRS SWS SPS	2 3 1 1 2	8	5	?
TET TWT TPT	1 1 4	6	3	
MGM MCM	1 1	2	2	
CYC	1	1	1	
DGD DSD DTD DCD DED	9 1 2 1 2	25	10	+
DQD DWD DYD DPD DND	3 1 1 3 2			
ESE ENE	3 1	4	2	
QSQ QPQ	1 2	3	2	
KSK KHK KFK KNK	3 3 4 1	11	4	?
RIR RER RPR	1 1 1	3	3	
HGH HEH HKH	1 1 1	3	3	
WPW	1	1	1	
FDF FPF	4 1	5	2	
YVY YSY YEY	2 2 1	5	3	
PGP PAP PLP PIP PSP PTP	1 1 8 2 2 2	25	13	+
PMP PDP PQP PKP PHP PWP	1 2 2 2 1 1			
PYP	1			
NGN NSN NCN NDN NEN NQN	1 3 1 1 3 1	11	7	+
NKN	1			

В рядках 1 1 2 вказано загальну кількість трипептидів даного типу в БД та кількість типів трипептидов в БД для даної амінокислоти, відповідно. В ряку "Інформативність" знаком "+" позначено яскраво виражену тенденцию до утворення визначеної структури; знаком "?"- можливий вплив залишків цього типу на детермінацію вторинної структури. Для оцінки інформативности крім загальної кількості трипептидів в БД важливе значення має кількість типів трипептидів.

випадіння. Для пояснення результатів автори запропонували гіпотезу високої конформаційної адаптивності коротких пептидів. В межах цієї гіпотези висунуто три припущення: 1) існують ідентичні пентапептиди, конформація яких повністю визначається навколишнім амінокислотним контекстом; 2) існують ідентичні пентапептиди, конформація яких не залежить від контекста; 3) ідентичність пентапептидів не відображує схожість вторинних структур або еволюційну спорідненність білків, які порівнюються. Для підтвердження своїх висновків автори наводять 25 прикладів ідентичних пентапептидів, 13 з яких підтверджують перше припущення, а 12 - друге. Використовуючи розроблену нами програму BankSearch, проведено пошук та аналіз ідентичних пентапептидів в базі даних з кількістю амінокислот, в два рази перевищуючою таку в роботі Кабаша та Сандера. Нами знайдено 143 ідентичних пентапептида з негомологічних білків, конформація яких з однаковою імовірністю була схожою та відмінною. Також виділено 162 пентапептида в гомологічних білках, конформація яких повністю ідентична. Дослідження цих прикладів демонструє, що , як в випадку неспівпадіння конформацій, так і в випадку їх співпадіння, фактором, який визначає конформацію, є наявність в найближчому оточенні (я5 залишків) пентапептидів інформаційно значимих амінокислот. Ми припускаємо, що "конкурентна боротьба" між інформаційно значимими амінокислотами, які детермінують визначену конформацію, в основному визначає локальную вторинну структуру ділянки білка.

Отримані дані свідчать про те, що локальна вторинна структура (ділянки білка довжиною до 15 залишків) визначається в основному наявністю інформаційно значимих аміно-

кислот. Одним з критичних елементів в методах передбачення вторинної структури на основі гомології послідовностей є визначення точних меж відповідності гомології послідовностей та схожості їх вторинних структур.Тому наступним кроком було вивчення впливу інформаційно значимих амінокислот на конформацію ділянок білка довжиною більшою, ніж 15 залишків.

Всі методи передбачення вторинної структури на основі гомології послідовностей засновані на простому принципі: короткі пептиди з високим ступенем гомології повинні мати схожу вторинну структуру. Однак, до цього часу було зроблено тільки одну спробу (Стернберг, Ислам, 1990) визначити, наскільки короткими повинні бути пептиди та яким високим повинен бут и ступінь гомології. Узагальнюючи отримані дані, ми прийшли до висновку: для фрагментів білків довжиною від 20 до 262 залишків ступінь гомології 25% є недостатньою для того, щоб вони мали ідентичну конформацію. Нами проведено аналіз (з обліком впливу інформаційно значимих амінокислот на утворення вторинної структури) ділянок білків довжиною від 15 до 200 залишків зі ступенем гомології >25% на предмет визначення точних меж (у відсотках) відповідності гомології послідовностей та схожості їх вторинних структур. На основі отриманих даних визначені наступні межі: для фрагментів, ступінь гомології яких не перевищує 30% - вторинна структура не співпадає; з гомологією від 30 до 47% - вторинна структура частково співпадає, з гомологією вище 47% - практично повне співпадіння вторинної структури . Також було визначено білки, які нами названо "самогомологічними" білками. Вони складаються з кількох високогомологічних блоків, що мають ідентичну вторинну структуру (мал.1A,B,D). При об'єднан-

A. Лектин (аглютінін)

```
RCGEQGSNMECPNNLCCSQYGYCGMGGDYCGKGCQDGACWTSK 1-43
*** *     **** ****** ** * :*** *** * *
RCGSQAGGATCPNNHCCSQYGHCGFGAEYCGAGCQGGPCRADI 44-86
:**** **  **** ****:* ** * *:** *** * *  *
KCGSQSGGKLCPNNLCCSQWGSCGLGSEFCGGGCQSGACSTDK 87-129
 **  : **:;* ** ***  **** *  :** ***** *
PCGGDAGGRVCTNNYCCSAGGSCGIGPGYCGAGCQSGGCDG   130-170
  hhhh          EE     EE  HHHH
```

B. Кальцій-зв'язуючий білок

```
TEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQD 1-47
      * :*** :*****:* *:  **  **  **:  *: *:::
DTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDE 74-119
 hhhHHHHHHHHHHH       EE HHHHHHHHHH       HHHHHHHHHH

MINEVDADGNGTIDFPEFLTMMARKMK 48-73
** * : **:* ::: **: **
MIREANIDGDGQVNYEEFVQMMTA  120-143
    EEHHHHHHHHhhhhh
```

C. Имуноглобулін FAB (людська мієлома)

```
DYWGQGTPVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVK   114-156
**:    *****  *    *     *     * *   *  :*
DYFP..QPVTVSWNSGALTSGVHTFP.AVLQSSGLYSLSSVVT   157-196
   eeeEEee       eeEEE   ee       EEEEEEEEe
```

D. Гама-кристалін

```
                    ч
GKITFYEDRGFQGHCYECSSDCPNLQPYFSRCN..SIRVDSGCWM 1-45
     **   * *   * : *** **  *   :  *: *  * *
FRMRIYERDDFRGQMSEITDDCPSLQDRFHLSEVHSLNVLEGSWV 88-133
 EEEEEee   eeeEEEE      hhhhh      EEEEEE EEE

LYERPNYQGHQYFLRRGDYPDYQQWMGFNDSIRSCRLIPQHTGT  46-87
*** * * *:** ** *:*  * :*    *  : *   : :
LYEMPSYRGRQYLLRPGEYRRYLDWGAMNAKVGSLRRVMDFY    89-176
EEEE  EEEEEEE  EEE  HHHH           EEEEe
```

E. Цитохром С3

```
      eee       eee          hhhh  ee  ee
APKAPADGLKMDKTKQPVVFNHSTHKAVKCGDCHHPVNGKENYQKCATAGCHDNMDKK   1-59
                     *   *  **    * :  *    **:   :
     DKSAKGYYHAMHDKGTKFKSCVGCHLETAGADAAKKKELTGCKGSKCHS  60-108
```

F. Феродоксин

```
   ee          u
  AYVINDSCIACGACKPECPVNIIQGSI   1-27
    :  **** ** *   ***   :
  YAIDADSCIDCGSCASVCPVGAPNPED  28-54
          hhhhh      ee
```

Мал 1. Порівняння послідовностей, які мають різну ступінь гомології, та їх вторинних структур: A, B, D - самогомологічніе білки з дво- і чотирисиметричною структурою гомологічних блоків; C- гомологічний блок; E, F - дуплікація функціональных сайтів. Символами E і H позначено співпідіння а-складки та в-спіралі в усіх вирівнених послідовностях в однакових позиціях; e і h - структури присутякі присутні тільки в одному з вирівнених білків. На E та F элементи вторинних структур (e,h) позначено над відновідними їм послідовностями; символами "*" і ":" позначено відповідно ідентичні та функціонально схожі амінокислоти; вертикальною стрілкою - можливий вплив інформаційно значимих амінокислот на утворення і руйнування локальної вторинної структури в вирівнених послідоностях. Ступень гомології послідовностей дорівнює - для A - 60-70%; B - 63%; C - 37%; D - 47%; E - 22%; F - 44%. Цифри в кінці послідовностей, які порівнюються, відповідають позиції фрагмента в білку.

ні таких блоків утворюється симетрична структура білка (як в первинній послідовності, так і у вторинній структурі),що може мати біологічне значення для збільшення спорідненності при зв'язуванні з субстратом, або константи зв'язування з субстратом за рахунок збільшення кількості активних сайтів.

3.5. Аналіз сайт-впізнаючих доменів 5-метілцитозин метілі трансфераз з використанням розробленого пакета програм.

Практичне запровадження модифікованого GUHA методу та аналізу інформаційних сайтів білків було здійснено на прикладі аналізу сайт-впізнаючих доменів метілтрансфераз. Показано їх співпадіння з функціональними мотивами у вторинній структурі цих ферментів, виявлене на основі сумісного передбачення вторинної структури доменів.

Для аналізу варіабельних доменів було отримано 26 послідовностей 5-метілцитозин метілаз з банка даних SWISS-PROT. Визначення варіабельних доменів проведено за допомогою розробленого нами метода пошуку образів шляхом пошуку VIII-го та IX-го консервативних блоків, після чого амінокислотна послідовність між ними приймалася за послідовність варіабельного домена. За допомогою цього же методу визначені п'ять консервативних блоків (субдоменів) усередині варіабельного домена (таблиця 4), вторинна структура яких визначена за допомогою модифікованого GUHA методу та використання стратегії інформаційних сайтів білків.

Всі п'ять субдоменів демонструють консервативність на рівні вторинної структури білка і асоційовані, в основному, з в-структурою. Однак, перший субдомен має наступну передбачну вторинну структуру:

EEEEEECCCCHHHHHHHHHHHHHCCCCCCEEEEEEEE, де H - а-спіраль; Е-в -складка; С - нерегулярна структура. Таку стуктуру знай-

Таблиця 4. П'ять консервативних субдоменів, знайдених у варіабельному домені 5-метілцитозин метілтрансфераз. В квадратных дужках позначено залишки, які стрічаються в даній позиції субдомена. Символ X - залишок, цифри в дужках після символу X означають ступінь повторності або довжину інтервала між сусідніми залишками.

Субдомен	Послідовність
1	[VIL]X(3)[VL]T[PG]X(4)K[RH]X(0,1)Q[NE]X(3)F X(0,1)K[DE][DN]GX(1)PX(2,3)[VI][DN]X(6)V
2	[RK]X(3)[WY][DEN]XT[VIL]X[STA]SX(5)[LI]H
3	[VIL]LEX(2)VX(2)KY[YI]X(0,1)LX(2)[DEQ]X(1,2) [DENQ]XL
4	QQXLX(8,9)EY[STA]X(2)[DE][QE]X[LI]X(2,3)[LI] [QE]X[PG]EX(3,4)[QE]X(2)P[VI]
5	GX[VI][DN]X(2)GX(6)VYX(3)G[VL][SA]PT[LI]T[TS] X(1,2)GXG

дено на багатьох ДНК-зв'язуючих білках та означено як в-а-в структурний домен або складка Россмана (Россман та ін., 1974). Таким чином, вперше в сайт-впізнаючому домені 5-метілцитозин метілаз знайдено мотив, відповідний в-а-в структурному домену ДНК-зв'язуючих білків.

ВИСНОВКИ

1. Удосконалено та адаптовано метод, відомий в області експертних систем як GUHA метод. На основі цього методу створено програму розрахунку вторинної структури білків для IBM PC сумісних компьютерів. Середня точність передбачення модифікованого GUHA метода для трьох конформацій вторинної структури склала 68%, що є однією з самих високих в сучасних методах передбачення вторинной структури білків.

2. Показано, що існують групи "інформаційно значимих" амінокислот, які мають тенденцію до утворення визначеної вторинної структури - A, E, та F - а-спіралі; V - в-складки; G, D, K, P та N - нерегулярної конформації. Вони мають повний набор фізикохімічних якостей (гідрофільність, гідрофобність, позитивний та негативний заряд), необхідних для створення функціональних сайтів.

3. Встановлено, що послідовності довжиною від 20 залишків та вище зі ступенем гомології до 30%, як правило, не мають схожої вторинної структури, зі ступенем гомології від 30 до 47% - частково гомологічні, вище 47% - мають практично повністю співпадаючу вторинну структуру.

4. При вирівнюванні гомологічних послідовностей та визначенні їх структури важливу роль відіграють "інформаційно значимі" амінокислоти, здатні при практично однаковому кон-

тексті руйнувати короткі регулярні структури.

5. На прикладі бактеріальних 5-метілцитозин метілтрансфераз продемонстровано пристосованність модифікованого GUHA -методу та ряду оригінальних програм для дослідження функціональних сайтів глобулярних білків та їх співвідношення з просторовою структурою білка.

6. В сайт-впізнаючому домені 5-метілцитозин метілтрансфераз вперше виявлено п'ять функціональних мотивів, які корелюють з консервативними елементами просторової структури цих ферментів.

7. Показано присутність в сайт-впізнаючому домені метілтрансфераз в-а-в структурного домена, знайденого у багатьох ДНК-зв'язуючих білків.

Основний зміст дисертації опубліковано в роботах:

1. Мальченко С.З.,Чащин Н.А. Предсказание вторичной структуры белков // Биополимеры и клетка. -1992. -Т.8. -N 4. -С. 21-30.

2. Мальченко С.З., Чащин Н.А. О некоторых закономерностях в организации пространственной структуры белков // Биополимеры и клетка. - 1993. -Т.9. -N 1. -С.3-9.

3. Братусь А.В., Мальченко С.З., Чащин Н.А. Предсказание вторичной структуры белков модифицированным GUHA-методом // Биополимеры и клетка. -1993. -Т.9. -N 5. -С.61-66.

4. Чащин Н.А., Братусь А.В., Мальченко С.З. Использование GUHA- метода для предсказания вторичной структуры белков // Междунар. конф. "Моделирование и компьютерные методы в молекулярной биологии и генетике". Тез.докл. - Новосибирск, 1990. -С.129.

5. Maltchenko S.Z., Bratus A.V., Chashchin N.A. A new method for the prediction of secondary protein structures with using of se- lfstuding expert system // Internat. conf. "Bioinforma- tics in the 90's". Abstr.papers. - Maastricht. - 1991. - p.13.

6. Maltchenko S.Z. Conservative subdomains in target regontion domain domain from m5C methylases // Internat. conf. "FEBS'92". Abstr.papers. - Dublin. - 1992. -p.104.

7. Maltchenko S.Z. The target recognition domains from methyltra- nsferases of II class have a common subdomains // Proc.1993 Miami Bio/Technology Winter Symposium. Miami Short Reports. - Miami. -1993. -V.3. - p.46.

Аннотация

Мальченко С.З. Исследование некоторых закономерностей в организации вторичной структуры глобулярных белков.

Диссертация на соискание степени кандидата биологических наук по специальности 03.00.03 - молекулярная биология, Институт молекулярной биологии и генетики НАН Украины, Киев, 1995.

Защищается семь научных работ, которые содержат результаты изучения и выявления закономерностей, присущих вторичной структуре глобулярных белков, из имеющихся на сегодняшний день баз данных пространственных структур. На основе полученных данных разработан пакет прикладных программ для предсказания вторичной структуры глобулярных белков.

Ключові слова: вторинна структура білків, гомологія послідовностей, 5-метилцитозин метилтрансферази, функціональні мотиви білків, експертна система.

Maltchenko S.Z. The studying of some regulatories in the secondary structure organization of globular proteins.

Dissertation is defended as a candidat work for the biological sciences, spetiality of 03.00.03 - molecular biology, Institute of Molecular Biology and Genetics NAS of Ukraine, Kiev, 1995.

The seven scientific articles that contain a result of studying and revealing of regulatories peculiar to a secondary structure of globular proteins based on a currently available protein spatial structure databases are defended in the thesis presented. The software package has been worked up for a globular protein secondary structure prediction based on a data obtained.

Підписано до друку 13.01.95р формат 60x84/16
Папір друк. Умов. друк. л. 1,0. Тираж 100 примірник. Заказ №83
Надруковано ЦУОП ДНПП "Плодвинконсерв" м. Київ, Саксаганського,1