

[КІЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ННЦ «Інститут біології та медицини»

Кафедра біохімії

Зав. кафедри Савчук О.М.

Протокол №_____ засідання кафедри

від “_____” 2018 р.

**ПОШУК ФУНКЦІОНАЛЬНИХ ЛІНІЙНИХ МОТИВІВ ПРОТЕЇНІВ
З ВИКОРИСТАННЯМ СУКУПНОСТІ БЛКОВИХ ВЗАЄМОДІЙ**

Випускна кваліфікаційна робота

студента 2 року магістратури

денної форми навчання

Клещевнікова Віталія Віталійовича

Науковий керівник від кафедри -

доцент кафедри біохімії,

кандидат біологічних наук

Гребінник Дмитро Миколайович

Робота виконана в Європейській молекулярно-біологічній лабораторії – Європейському інституті біоінформатики (EMBL-EBI), Хінкстон, Кембридж, Велика Британія, під керівництвом керівника групи Доктора Євангелії Петсалакі

Оцінка захисту роботи

Київ – 2018 р.

ЗМІСТ

ВСТУП	5
РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ	7
1.1 Характеристика коротких лінійних мотивів	7
1.1.1 Опис модулів білок-білкових взаємодій	7
1.1.2 Опис коротких лінійних мотивів та ділянок молекулярного розпізнавання	8
1.1.3 Класи коротких лінійних мотивів	10
1.2 Експресія білків, еволюція та сплайсинг, що впливають на клітинну функцію шляхом зміни структури мережі взаємодій	13
1.3 Еволюція лінійних мотивів	16
1.4 Проблеми відкриття лінійних мотивів	17
1.5 Обчислювальні методи є необхідними	18
1.6 Відкриття лінійних мотивів людини, що конвергентно еволюціонували у вірусних білках, <i>de novo</i>	19
РОЗДІЛ 2 МАТЕРІАЛИ ТА МЕТОДИ	21
2.1 Робота з базами даних білкової взаємодії	21
2.2 Аналіз розподілу ступенів	23
2.3 Білкові послідовності й передбачення доменів	24
2.3.1 Білкові послідовності	24
2.3.2 Передбачення домену за допомогою InterProScan	24
2.3.3 Видалення повторюваних доменів	25
2.3.4 Поєднання домену людини з даними вірусно-людської взаємодії	25
2.4 Статистичний метод оцінки того, які домени імовірно опосередковують взаємодію	26

2.6 Інструменти та процедура пошуку мотивів	29
2.6.1 Програмне забезпечення для пошуку мотивів	29
2.6.2 Створення наборів даних для пошуку мотивів.....	30
2.6.3 Процедура пошуку мотивів	31
2.7 Порівняльний аналіз екземплярів мотивів до еталонних даних	32
2.7.1 Еталонні дані	32
2.7.2 Процедура порівняльного аналізу.....	33
2.7.3 Приклади відкритих заново та мотивів-кандидатів	34
2.8 Процедура визначення подібності паттерну мотивів	34
2.9 Технічне обладнання	35
2.10 Статистичний аналіз даних	35
РОЗДІЛ 3 РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ	37
3.1 Дослідження мережі взаємодій білків людини між собою та з білками вірусів	37
3.1.1 Дослідження асиметрії вірусно-людської мережі білкових взаємодій	37
3.1.2 Дослідження ефекту упередженості даних на центральність білків людини, що є мішенями вірусів	39
3.3 Дослідження доменів, що імовірно опосередковують взаємодію між білками	42
3.4 Пошук коротких лінійних мотивів	45
3.5 Дослідження ефекту фільтрації за ймовірним доменом розпізнання на чутливість передбачення мотивів.....	51
3.6 Дослідження схожості мотивів знайдених <i>de novo</i> до відомих мотивів	53
3.7 Приклади відкритих заново та мотивів-кандидатів.....	55

3.7.1 Дослідження класів мотивів, що відкриті заново, мотивів-кандидатів та їх імовірних доменів розпізнавання	55
3.7.2 Мотиви PDZ, які відкриті заново	56
3.7.3 Мотиви-кандидати, що зв'язують домен PDZ	59
3.7.4 Мотив-кандидат, що зв'язує домен SH3	62
3.7.5 Мотиви-кандидати, що зв'язують домен WD40	64
3.7.6 Мотиви-кандидати, які розпізнаються доменом, що зв'язує дволанцюгову РНК, та доменом EF-hand	68
3.7.7 Мотив-кандидат, що зв'язує BAG-домен	70
3.8 Майбутні напрямки дослідження	72
3.8.1 Молекулярне стикування мотиву та домену й покращений аналіз мережі людини	72
3.8.2 Інтеграція кількох предикторів	73
3.8.3 Експериментальна перевірка передбачених мотивів	74
ВИСНОВКИ	75
СПИСОК ДЖЕРЕЛ ЛІТЕРАТУРИ	76
ДОДАТКИ	88
Додаток А	88
Додаток Б	89
Додаток В	91
Додаток Г	92
Додаток Д	94

ВСТУП

Лінійні мотиви - це короткі мотиви амінокислотної послідовності, що опосередковують фізичні та селективні білок-білкові взаємодії. Вони, як правило, розташовані в невпорядкованих ділянках білка і, як правило, розпізнаються структурованими глобулярними доменами [1].

Відомо, що взаємодії, опосередковані лінійним мотивом, з'єднують і направляють сигнальні шляхи клітин [2]. Ця функція часто додатково регулюється посттрансляційними модифікаціями й кооперативністю взаємодій [3]. Лінійні мотив-опосередковані взаємодії можуть швидко еволюціонувати і допомагати змінювати сигнальні мережі клітини при видоутворенні, у захворюваннях або взаємодії патогена-хазяїна [4–6].

Ряд лінійних мотивів виявлено з використанням традиційних методів молекулярної біології та гіпотезних досліджень, проте ці методи є трудомісткими, а більшість функціональних лінійних мотивів ще не визначені [1]. Використання обчислювальних інструментів пошуку для ідентифікації лінійних мотивів у гомологічних білках, як правило, передбачує велику кількість нефункціональних мотивів. Деякі підходи показали покращення ефективності виявлення функціональних мотивів: включення даних про взаємодію білків, консервація послідовності у кількох видах та фільтрування для мотивів, розташованих у неструктурзованих ділянках [7, 8]. Методи, такі як фаговий дисплей, були розроблені, щоб допомогти експериментальному виявленню мотивів у масштабах протеому [9]. Однак ми далекі від повної характеризації коду взаємодії доменів-лінійних мотивів і поточні оцінки припускають, що на сьогоднішній день було виявлено лише 1% мотивів, порівняно з очікуваними 15-40% взаємодій [10, 11].

Вірусні білки імітують клітинні лінійні мотиви для взаємодії та модифікації клітинної сигналізації таким чином, що сприяє прогресуванню вірусної інфекції [5]. Ми можемо використовувати цю функціональну залежність для підвищення чутливості обчислювального передбачення

мотивів. Цей аналіз не було зроблено раніше з таким великим набором даних, а також з використанням комбінації вірусно-людських та людських мереж білкових взаємодій. На відміну від інших досліджень інтерактомного масштабу [12], ми використовуємо статистичний метод, щоб оцінити, які домени можуть опосередковувати взаємодію. Обчислювально передбачені пари доменів-лінійних мотивів будуть перевірені за допомогою скріну фагового дисплею в лабораторії, що співпрацює з нашою. Ця робота може сприяти розумінню коду взаємодії доменів-лінійних мотивів та того, як віруси використовують цей механізм.

Метою даного дослідження є використання даних взаємодії вірусних білків з білками хазяїна та доменів, які їх імовірно розпізнають як засобу обмеження простору пошуку для відкриття нових функціональних лінійних мотивів. Відповідно до мети було поставлено задачі:

1. Отримати та обробити дані експериментальної взаємодії з публічних баз даних та вивчити властивості мережі вірусно-людської білкової взаємодії.
2. Використати вірусно-людську мережу, інструменти імовірності пошуку мотивів, щоб відкрити короткі лінійні мотиви *de novo*. Використати послідовності вірусних білків, щоб обмежити простір пошуку.
3. Визначити домени білкової послідовності у всіх вірусних і людських білках. Оцінити домени людини, ймовірно, опосередковують взаємодію з кожним вірусним білком.
4. Оцінити наш метод пошуку мотивів за допомогою еталонного набору даних відомих вірусних мотивів.
5. Реалізувати цей метод пошуку мотиву в статистичній мові програмування R, за допомогою інструментів командного рядка і високопродуктивного обчислювального кластера.

Хочу висловити подяку моєму науковому керівнику доктору Евангелії Петсалакі за допомогу у роботі над проектом та менторство; Європейському Інституту Біоінформатики (EMBL-EBI) за надання обчислювальних ресурсів та фінансування; кафедрі біохімії за дозвіл виконувати роботу в EMBL-EBI.

РОЗДІЛ 1

ОГЛЯД ЛІТЕРАТУРИ

1.1 Характеристика коротких лінійних мотивів

1.1.1 Опис модулів білок-білкових взаємодій

Структура та функції клітин виникають внаслідок взаємодії між молекулами усередині та ззовні клітин [13]. Білки, нуклеїнові кислоти, ліпіди та малі молекули усі можуть утворювати біологічно важливі взаємодії. У нашому дослідженні ми зосереджуємося на взаємодії білків. Ці взаємодії регулюють клітинні процеси та організмові фенотипи від смерті клітини до скорочення м'язів. Зникнення або введення нового білкового контакту може становити молекулярну основу захворювання або еволюційної адаптації [5, 6, 14]. Для створення цих фенотипів білки взаємодіють у специфічних умовах у визначених типах клітин та субклітинних локалізаціях [15]. Таким чином, взаємодії організовують біохімічні та забезпечують структурні функції білка.

Всі аспекти функціонування білків здійснюються модулями, вбудованими в його послідовність. Ці модулі можуть складатися у стабільну тривимірну структуру в нативних умовах (глобулярні домени) або не мати стабільної 3D-структур (невпорядковані ділянки). Глобулярні домени залишають за собою різні функції, що вимагають точного просторового розташування амінокислотних залишків та жорсткої структури: ферментативна, ліганд-зв'язуюча (ДНК, ліпіди, пептиди) або структурні функції. Глобулярні домени являють собою більшість відомих інтерфейсів взаємодії між білками, однак, більшість цих взаємодій є дуже стабільними і не мають динамічних властивостей, необхідних для індукованих і тимчасових взаємодій. Функціональність взаємодій опосередкованих глобулярними доменами доповнюється короткими лінійними мотивами (SLiM або лінійні мотиви), розташованими в гнучких невпорядкованих ділянках [1].

1.1.2 Опис коротких лінійних мотивів та ділянок молекулярного розпізнавання

Короткі лінійні мотиви (SLiMs) - мотиви послідовності 3-15 амінокислотних залишків, що опосередковують фізичну та селективну взаємодію між білками. Лінійна послідовність мотиву, а не його тривимірна структура, вважається важливою для зв'язування. SLiMs, як правило, розташовані в невпорядкованій ділянці білка [16, 17]. Це може бути довга невпорядкована ділянка або коротка петля на поверхні глобулярного домену [18]. Гнучкість цього регіону дозволяє глобулярним доменам взаємодіючого партнера розпізнати SLiM. Тільки 1-5 амінокислотних залишків є необхідними детермінантами специфічності розпізнавання, такими як фосфотирозин у мотиві зв'язування SH2-домену [19]. Амінокислотні залишки в сусідніх позиціях можуть додатково модифікувати специфічність, спорідненість та селективність мотиву. Допустимі послідовності сприяють підвищенню зв'язування, тоді як неприпустимі порушують зв'язування близько до основного сайту. Які амінокислоти є необхідними, допустимими або неприпустимими, в більшості випадків є специфічним до екземпляру домену визнання [19]. Форма та фізико-хімічні властивості зв'язуючої кишени визначають специфічність, спорідненість та селективність домену розпізнавання до послідовності [1].

Родини доменів можуть мати широку специфічність до класу мотивів. Наприклад, домени SH2, SH3 та PDZ зв'язуються відповідно з фосфотирозином, з пролін-багатими або С-кінцевими мотивами. Навпаки, конкретний екземпляр домену в білку може розпізнати більш специфічну послідовність мотивів у обмеженому наборі білків. Наприклад, домен SH2 GRB2 зв'язує фосфорильований мотив pYENV рецепторних тирозинкіназ, що приводить до індуцибельного рекрутування, тоді як SH2 домен Src розпізнає мотив pYEEI у послідовності самого Src, що викликає аутоінгібування кінази [20]. Контекст послідовності навколо фосфорильованих тирозинів визначає,

які білки, що містять SH2 домен, будуть рекрутовані та які сигнальні шляхи будуть активовані. Інші докінг мотиви та їхні домени розпізнавання доповнюють низьку специфічність доменів серин/треонінінкіназ (такі як МАР-кінази) до їх мішеней [21]. Як показано на прикладах, взаємодії опосередковані SLiM можуть бути індуцибельними, тимчасовими і виконувати регулюючі функції. І домени, і мотиви можуть мати інший ступінь специфічності зв'язування щодо своїх партнерів. Кілька доменів можуть розпізнавати одні й ті ж різномірні мотиви, або той самий невибагливий домен може розпізнавати кілька мотивів [1].

Мотив-опосередковані взаємодії є найслабшими з трьох основних типів: взаємодії доменів, взаємодії між доменом та мотивами, та взаємодії між ділянками молекулярного розпізнання (molecular recognition feature або MORF). Ці типи взаємодій відрізняються площею інтерфейсу взаємодії, що, у свою чергу, сприяє спорідненості. Взаємодії між доменами є найсильнішими (пікомолярна спорідненість) і, як правило, беруть участь у формуванні стабільного білкового комплексу. У взаємодіях домен-MORF невпорядкована ділянка одного білка переходить до впорядкованого стану, набуваючи стійку 3-мірну структуру в комплексі. MORFs також називають внутрішньо невпорядкованими доменами. Ці взаємодії мають проміжну міцність (наномолярна спорідненість). Низька спорідненість мотив-опосередкованої взаємодії (низька мікромолярна спорідненість) впливає на динаміку взаємодій, що дає змогу швидко перемикання, або вимагає наявності кількох мотивів для високоавідного біологічно важливого зв'язування. Це, поряд з іншими властивостями SLiM, ідеально підходить для з'єднання білкових комплексів [13, 22], таргетингу білків до певний органел або збирання функціонально різних комплексів навколо інваріантного ядра (протеасома, машинерія ініціації транскрипції). Таким чином, практично всі клітинні процеси залежать від взаємодій, опосередкованих SLiM.

1.1.3 Класи коротких лінійних мотивів

Мотиви можуть бути розділені на 2 загальні групи: мотиви, які опосередковують зв'язування, та мотиви, які є мішеню для посттрансляційної модифікації (PTM). Кожна з цих груп може бути далі поділена. Мотиви, які опосередковують зв'язування SLiM включають ліганд-зв'язуючі, таргетинг-, докінг- та деградаційні мотиви. Мотиви посттрансляційної модифікації підрозділяються на мотиви додавання/видалення фрагментів, або класичні PTM-мотиви, а також мотиви розщеплення [1]. У базі даних ELM, яка збирає екземпляри відомих мотивів з літератури, анотовано 6 типів мотивів [23].

Ліганд-зв'язуючі мотиви. Класичні ліганд-зв'язуючі мотиви є посередниками збірки білкового комплексу - включаючи функціонально різні комплекси навколо одного і того ж інваріантного ядра або скафмолдінг білків, що утворюють один і той же шлях. Наприклад, ядерні рецептори рекрутують набір транскрипційнійних репресорів або активаторів через CoRNR мотив або NR-box мотив, залежно від зв'язування стероїдного гормону - їхнього ліганда [24]. Скафмолд-білки можуть регулювати клітинну сигналізацію кількома способами: від визначення лінійних шляхів завдяки організації кіназ у правильного порядку (наприклад, KSR, який організовує каскад МАР кінази) до інгібування за допомогою титрування скафолду або аллостерічного регулювання [25].

Таргенг-мотиви керують локалізацією білків, направляючи транслокацію білків між субклітинними компартментами за допомогою машинерії транспортування (транспортні мотиви, наприклад сигнал ядерної локалізації, сигнал ядерного експорту), або шляхом утримання білка в правильному компартменті завдяки закріпленню цього білка у комплексі специфічному до компартменту (SxIP-мотив, що розпізнається ЕВН доменом білків, що зв'язують кінці мікротрубочок [26]).

Докінг-мотиви рекрутують модифікуючі ферменти й широко використовуються для підвищення субстратної специфічності ферментів.

Докінг-мотиви часто підводять ферменти, щоб ті модифікували іншу ділянку в субстраті. Докінг-мотиви можуть привести до розпізнавання субстрату у 3 основні способи. Мотиви можуть розпізнаватися сайтом в каталітичному домені, відмінному від каталітичного сайту. Наприклад, докінг-жолоб каталітичного домену MAP кінази розпізнає докінг мотив у MEF2A, MAF2K1 або MKP1 [1]. Крім того, окремий модуль взаємодії, розташований у тому самому білку, може розпізнавати докінг-мотив. Раніше згадані SH2-домени часто виконують цю функцію. Не лише домен SH2 Src-кінази бере участь в аутоінгібуванні, але і в рекрутуванні субстратів, такому як рекрутування FAK1-кінази через мотив pYAEI [27]. Нарешті, домен розпізнання, який зв'язує субстрат, може бути розташований в іншому білку, який утворює комплекс з ферментом. Спочатку цей комплекс має бути зібраний, що може покладатися на взаємодію лінійного мотиву або доменів. Одним із прикладів є CDK (циклінзалежні кінази), які покладаються на домен розпізнання в цикліновому білку для розпізнавання їх мішеней [28].

Окрема група докінг-мотивів, які регулюють стабільність білку, називається деградаційними мотивами або дегронами (DEG в базі даних ELM). Ці мотиви рекрутують убіквітин-лігазу (наприклад, Е3-куліновий комплекс) до своїх субстратів. Прикріплення убіквітину до цих субстратів спрямовує їх на деградацію протеасомою - так звана, убіквітин-протеасомна система. Слід зазначити, що залежно від кількості доданих убіквітинів або структури поліубіквітину, на додаток до деградації, ця мітка може контролювати взаємодію білків та субклітинну локалізацію (поліубіквітин K48 є міткою деградації) [29, 30]. Е3-убіквітин-лігази основними детермінантами специфічності деградації та є найбільш поширеними в геномі людини (> 700 Е3 ферментів, ~ 40 Е2 ферментів, 2 Е1 ферментів) [31, 32, 29, 33].

Мотиви посттрансляційної модифікації. Друга велика група мотивів збігається з сайтом посттрансляційної модифікації (PTM) і опосередковує

розділення субстрату активним сайтом ферменту. Існує 3 основних класи мотивів посттрансляційної модифікації:

1. Мотиви, що розпізнаються ферментом, який каталізує додавання або видалення групи, наприклад фосфату, убіквітину або ліпіду (MOD in ELM).
2. Мотиви, що розпізнаються ферментом розщеплення (CLV).
3. Мотиви, що розпізнаються ферментом, який каталізує цис-транс перетворення пептидного зв'язку пролін. Ці ферменти називаються пептидилпроліл цис-транс ізомерази; найвідомішим прикладом є сімейство білків-циклофілінів та PIN1 [1].

Багато видів PTM типу приєднання-видалення групи були відкриті: фосфорилювання, ацетилювання, метилювання, SUMO-лювання, убіквітінювання, зажорення ліпідів до мембрани і багато інших, менш поширеніх модифікацій [34–39]. Вони широко використовуються в кількох клітинних процесах, але найбільш вивченими є фосфорилюванням-опосередкована сигналізація і епігенетичний контроль експресії генів. Епігенетичний контроль в цьому контексті описує модифікацію невпорядкованих хвостів гістонових білків на різних сайтах, що контролює стан хроматину та транскрипцію [40].

Ці модифікації часто утворюють контекст для ліганд-зв'язуючих мотивів шляхом порушення або створення взаємодії безпосередньо або через кооперативні механізми, що включають структурні зміни, індуковані зміною заряду, кілька мотивів або партнерів взаємодії [1]. Наприклад, домен SH2 GRB2 зв'язується з залишком фосфотирозину тирозинкиназного рецептора після його фосфорилювання [41]. Часто, у нас немає остаточного підтвердження того, який механізм використовується.

Мотиви розщеплення розпізнаються каталітичним доменом протеаз (подібно модифікаційним мотивам) і необоротним чином гідролізуються у ділянці розщеплення (на відміну від сайтів модифікацій). Ці ферменти виконують обмежений протеоліз, порушуючи або іноді уможливлюючи функцію білка. Найбільш відомими мотивами цього класу є ті, що

ропізнаються каспазами - основними регуляторами програмованої загибелі кліти, апоптозу [42], або реакції запалення в міелоїдних клітинах [43]. Апоптоз може бути ініційований імунними клітинами (Т-кіллерами або натуральними кілерами) ззовні клітини або пошкодженням ДНК чи мітохондрій всередині клітини і зазвичай починається з активації регуляторних каспаз. Регуляторні каспази (наприклад, каспаза 8 і 9) активують ефекторні каспази (наприклад, каспазу 3, 6 та 7), розпізнаючи мотив LEHD та розщеплюючи його [44]. Ефекторні каспази розпізнають мотив [DSTE][^{^P}][^{^DEWHFYC}]D[GSAN], CLV_C14_Caspase3-7 в ELM [45], та викликають розщеплення сотен білків, що призводить до апоптозу. Еволюціонувавши цей мотив розпізнавання, білок може стати під контроль шляху апоптозу [1]. Ефекторна каспаза може мати як регуляторний (активацію ферменту розщеплення ДНК) знищуючий (розщеплення цитоскелетних білків) ефект на її субстрати.

Усі ці класи частково перекриваються і не є взаємовиключними, наприклад, ліганд-мотив, що зв'язує, може приєднати білок до комплексу, але також визначати його субклітинну локалізацію (таргетинг). Той же мотив може бути мотивом посттрансляційної модифікації та класичний ліганд-зв'язуючий мотивом (ліганд SH2-домену). Отже, класи мотивів визначаються в контексті взаємодії, а не як властивість послідовності. Така контекстна залежність та функціональне визначення роблять відкриття лінійних мотивів складним [8]. Крім того, як це буде розглянуто в одному з наступних розділів, однакова послідовність амінокислот може бути функціональним мотивом або ні залежно від доступності для зв'язування доменами розпізнавання.

1.2 Експресія білків, еволюція та сплайсинг, що впливають на клітинну функцію шляхом зміни структури мережі взаємодій

Щоб проілюструвати, як еволюція лінійних мотивів може створити нову функцію шляхом зміни структури мережі білкової взаємодії, розглянемо приклад докінг-мотивів. Як описано в попередньому розділі, докінг-мотиви функціонують шляхом розміщення субстрату в безпосередній близькості від

кatalітичного домену, тобто збільшуючи локальну концентрацію субстрату і, таким чином, дозволяючи досягнення специфічності та селективності (ортогональності) сигнальної відповіді (кілька стимулів - одна кіназа - кілька стимул-залежних субстратів) завдяки просторовому розділенню невідповідних ферментів та субстратів. У цьому світлі важливо підкреслити динамічний та залежний характер взаємодій опосередкованих SLiM та кількісний характер реакцій клітинної сигналізації. Незважаючи на те, що цільові субстрати можуть бути фосфорильовані певною мірою, тільки правильні субстрати будуть модифіковані з такою швидкістю, яка достатня для викликання біологічно значущої відповіді (внаслідок просторової близькості) [46]. Модулярність білкової послідовності дозволяє вводити довільні каталітичні домени та докінг-мотиви в послідовність білків, щоб привести іншу каталітичну функцію до відповідного білкового комплексу або клітинної локації, що визначається докінг-мотивом. Додання лінійних мотивів інших класів додає регуляції на іншому рівні (дегрон чи мотив розщеплення).

Той факт, що лінійний-мотив-опосередкована просторова близькість (не просторова структура білка) достатня для багатьох регуляторних взаємодій, збільшує еволюційну пластичність цих взаємодій. Якщо ви можете знайти спосіб розмістити противірусний блок господаря, наприклад, цитидинезаміназу APOBEC3G, в безпосередній близькості від Cullin-E2 убіквітін-лігази (наприклад, шляхом конструювання скаффолд-білка, що містить мотиви для обох), ви можете викрасти власну систему клітини, щоб дозволити вірусну інфекцію [47]. Це є прикладом зміни структури мережі через експресію білка - у цьому випадку вірусного білка, однак той же механізм може контролювати функціональну різноманітність типів клітин за допомогою білків з експресією обмеженою до клітинної лінії.

Іншим процесом, який спирається на просторову близькість, є регуляція генів. Клітини можуть активувати транскрипцію онкогену, об'єднуючи ДНК-зв'язуючий домен, який зв'язується з промоторами цих генів (білок FLI1) до невпорядкованої ділянки, що містить мотиви, які рекрутують машинерію

активації транскрипції (транс-активаційний домен, білки EWSR1) [6]. Такі злиття генів у раку, як правило, порушують взаємодію білків з іншими молекулами: білками, РНК, ДНК. Невпорядковані ділянки білка, що містять лінійні мотиви й сайти посттрансляційної модифікації, можуть бути вибірково виключені в злитому білку, знімаючи регулюючий контроль. Це є прикладом зміни структури мережі шляхом мутагенних подій з наступним фенотипічним відбором (у цьому випадку на здатності безконтрольно проліферувати). Незважаючи на те, що це є приклад еволюції лінійного мотиву за онкологічної хвороби людини, аналогічні процеси можуть діяти, щоб змінити клітинні функції, на більш довгій еволюційній шкалі часу [48].

Наше розуміння ролі лінійних мотивів у функціональних інноваціях покращилося, у результаті аналізу взаємодії білків з різними анатованими функціями та як ці взаємодії еволюціонували уздовж філогенетичного дерева [22]. Ми обговорюємо це в наступному пункті.

Спираючись на дані, Kim та співавт. визначають модулі взаємодії білків, які теоретично відповідають білковим комплексам. Взаємодії, опосередковані SLiM, більш імовірно з'єднують білки між модулями з різною функцією, тоді як взаємодії доменів, більш імовірно з'єднують білки в самих модулях. Взаємодії, опосередковані SLiM- або доменом, були передбачені. Функція модулів визначалася за допомогою анотацій з онтології генів. Kim та співавтори також показали, що складні види вищих тварин набули більше взаємодій, опосередкованих мотивами, ніж взаємодій, опосередкованих доменами [22]. У незалежному дослідженні Hein та співавт. зкомбінували експериментально визначену спорідненість взаємодії з топологією мережі. Вони виявили, що білки всередині модулів пов'язані сильними взаємодіями, проте білкові контакти між модулями були слабкими. Хоча Hein та співавт. не продемонстрували, що ці слабкі взаємодії є опосередкованими SLiM, якщо врахувати те, що ми знаємо про аффінність цих взаємодій, було б справедливим гіпотезувати саме те.

Нарешті, структура мережі взаємодій може бути змінена шляхом сплайсингу мотив-кодуючих послідовностей, щоб змінити локалізацію білка або його здатність бути мішенню ферментів [49].

1.3 Еволюція лінійних мотивів

Для розглядання еволюційних властивостей лінійних мотивів, краще протиставити їх глобулярним доменам. Механізми еволюції доменів значною мірою є загальноприйнятим знанням в той час як повне розуміння еволюції мотивів досі встановлюється. Домени еволюціонують шляхом дуплікації, дивергенції та рекомбінації [50]. Навпаки, SLiMs часто еволюціонують *de novo* або *ex nihilo* в послідовностях як не-гомологічних, так і гомологічних білків. Не-гомологічні білки можуть здобути той самий мотив. Гомологічний білок може еволюціонувати нові класи мотивів, не поділені їх спільним предком [4]. Гомологічні білки можуть втратити мотив, який вони поділяли, а замість цього здобути той самий мотив у послідовності того ж невпорядкованого регіону. Це явище називається оборотом мотивів. Щоб краще зрозуміти ці явища, необхідно враховувати контекст у послідовності і структурі, в якому мотиви еволюціонують.

Послідовність невпорядкованих ділянок не обмежується структурним контекстом, що дозволяє швидку еволюцію послідовності. Заміни амінокислот не мають руйнівного впливу на структуру білків, якщо вони відбуваються у невпорядкованій ділянці, що призводить до зниження вартості кількох послідовних замін. Невпорядковані ділянки забезпечують контекст в якому короткий лінійний мотив може еволюціонувати в кілька або навіть одну подію заміни амінокислот [51, 1]. Цей контекст забезпечує умови, необхідні для конвергентної еволюції SLiM.

Далі розглянемо, як селективні сили діють на мотиви, що еволюціонують *ex-nihilo*. Якщо новий мотив ніколи не розпізнається в потрібному контексті та не дає еволюційної переваги або невигоди, цей мотив буде втрачено через той самий процес випадкової мутації. Позитивний вибір дозволить зберегти

мотив, який надає корисні методи регуляції. Модель, за якою 2 білки можуть незалежно набути певну послідовність, називається конвергентною. Поширилою стратегією для еволюції короткого лінійного мотиву є конвергенція (на відміну від доменної структури та доменної архітектури білків). Про це свідчать гомологічні вірусні білки, які поділяють еволюційне походження, але втратили успадковані та здобули нові лінійні мотиви (докладніше розглянуто у наступному розділі) [5]. Крім того, не тільки мотиви можуть конвергентно еволюціонувати в не-гомологічних білках, але комбінації мотивів також можуть еволюціонувати таким чином, припускаючи, що (на відміну від архітектури домену) функціональна необхідність є більш важливою, ніж еволюційне походження [52].

1.4 Проблеми відкриття лінійних мотивів

Незважаючи на те, що за нашими очікуваннями мотивів багато, їх важко знайти. Низька складність, що підвищує їх функціональність та легкість в еволюціонуванні, призводить до проблем при виявленні цих мотивів. Перші SLiM були виявлені за допомогою ретельно розроблених експериментальних досліджень злиття генів (*gene fusion*), наприклад сигнал затримання у ER або циклін-дегрон [53, 54]. Пізніше стало поширилою практикою шукати нові екземпляри відомих мотивів. Можна сканувати весь протеом чи білки, що представляють інтерес, для передбачення мотивів для подальшого експериментального дослідження. Проте існує чимало небезпек використання цього дуже спрощеного підходу, які нещодавно розглянули Gibson та співавт. [8]. Та сама амінокислотна послідовність може бути функціональною залежно від структурного контексту. Наприклад, в гідрофобному ядрі глобуллярних доменів часто можна знайти послідовність, що відповідає сигналу ядерного експорту, що має 4 гідрофобних залишків. Експериментальний мутагенез такого мотиву в ядерному білку викликає його агрегацію, що перешкоджає його експорту з ядра, яке можна помилково вважати свідченням

функціонального мотиву [8, 55]. Цей приклад підкреслює важливість пошуку мотивів в невпорядкованих ділянках.

Як продемонстровано в недавньому дослідженні Nagai та співавт., низька складність мотивів є проблемою при передбаченні випадків відомих мотивів в усьому протеомі [5]. Мотиви низької складності можна знайти за аналогічними частотами як в істинних, так і в рандомізованих вірусних послідовностях, що призводить до великої кількості помилково-позитивних мотивів. Наприклад, лише 1-6% відомих мотивів, що мають відповідну послідовність у білках 2 видів вірусів та 2 вірусних родин, зустрічаються у менш ніж 0,1% випадкових послідовностей.

1.5 Обчислювальні методи є необхідними

Інтеграція багатьох джерел даних, обчислювальних методів є важливою для пошуку нових екземплярів відомих мотивів та виявлення нових мотивів. Різні способи обмеження простору пошуку дозволяють вирішити проблеми, висвітлені в попередньому розділі. Визнання структурного контексту, консервації залишків, відомі взаємодії білків - усі дають додаткові докази для кожного конкретного мотиву [7, 8]. Мотиви, що мають суперечливі докази такі, як розташування в глобулярному домені або відсутність консервації залишків навіть серед споріднених видів, не повинні піддаватися подальшому експериментальному дослідженню. Останній аспект важко вирішити правильно через те, що мотиви можуть змінювати позицію у білку, через низьку якість даних більшості послідовностей білків та те, що програми вирівнювання не дуже добре вирівнюють неструктуровані послідовності [56, 8]. Ще однією проблемою для обчислювального відкриття SLiM є регіони низької складності: довгі ділянки з однієї амінокислоти. Однак, маскування цих ділянок може вводити хибно негативні спрацювання, оскільки ці регіони часто служать субстратом для еволюції мотивів [57].

Обчислювальне відкриття мотивів *de novo* спрямоване на пошук раніше невідомих мотивів, а також на вирішення проблеми низької складності

мотивів. Визначення пошукового простору - є дуже важливим. Набори білків, які, як вважається, містять мотиви, можуть бути отримані з даних наборів гомологічних послідовностей або взаємодій білків [7, 8], обидва можуть бути досить шумними. Білки взаємодіють з багатьма іншими білками через різні ділянки у їх послідовності. Отже, рідко всі відомі інтерактори містять один і той же мотив. Крім того, існує проблема справжніх, але off-target мотивів. Використовуючи будь-яку мережу білкових взаємодій, ми можемо виявити справжні мотиви використовуючи неправильні взаємодії без домену розпізнання на іншому боці [12]. Нарешті, можуть існувати нефункціональні мотиви, які можуть бути обчислювально відкритими і розпізнаватися правильними доменами *in vitro*, проте білки ніколи не взаємодіють *in-vivo*. Щоб вирішити цю проблему, дослідник може подивитися, чи білки коли-небудь експресуються у тому ж типі клітин або типі клітин інтересу перед подальшим експериментальним дослідженням.

Незважаючи на те, що наукова спільнота продукувала більш якісні геноми, дані про взаємодію з білками [58] і розробили кращі інструменти для обчислювальних відкриттів [59] та високопродуктивних експериментів *in vitro* [9], функціональна валідація залишається проблемою [8].

1.6 Відкриття лінійних мотивів людини, що конвергентно еволюціонували у вірусних білках, *de novo*

Еукаріотичні віруси спираються на конвергентну еволюцію мотивів для взаємодії та викрадення клітинних функцій. Це продемонстровано в численних цільових дослідженнях (розглянутих у розділі 3.7) та недавньому систематичному обчислювальному передбаченні мотивів у всіх вірусних білках [5]. Цікаво, що прокаріотичні віруси часто не використовують мотиви, оскільки бактеріальні білки зазвичай мають менше мотивів [5]. Віруси людини не тільки мімікрують клітинний мотив, але й те ж дослідження показало, що ці мотиви, ймовірно, еволюціонували *ex-nihilo*.

У нашому дослідженні ми скористалися цією властивістю вірусних білків. Ми використовуємо експериментальні вірусно-людські дані взаємодії для визначення набору людських або вірусних білків (рис 3.4.1 В або С), які можуть містити мотиви інтересу. Потім ми обмежили пошуковий простір мотивів, шукаючи лише мотиви, що конвергентно еволюціонували у вірусних білках. Ми дотримуємося найкращої практики маскування невпорядкованих ділянок; однак, ми також оцінюємо домени, які можуть посприяти взаємодії для підвищення чутливості та інтерпретабельності передбачених мотивів.

РОЗДІЛ 2

МАТЕРІАЛИ ТА МЕТОДИ

2.1 Робота з базами даних білкової взаємодії

Дані про білкову та білкову взаємодію (PPI) завантажено з бази даних IntAct випуску від 13 листопада 2017 р. [60] за допомогою функції `loadIntActFTP`, включеної в пакет `MItools` мовою програмування R. Записи в базі даних IntAct були очищені від тегів та текстового опису, щоб полегшити подальший аналіз, використовуючи функцію `cleanMITAB`. Ми використали UniProt записи (accessions), щоб називати учасників взаємодії та відфільтрували тільки білок-білкові взаємодії. Спеціальний комп’ютерний код, що враховує топологію таксономічного дерева, було створено та використано для визначення того, які взаємодії в базі даних є людина-людина (таксономія ID 9606, функція `FullInteractome`), вірусно-вірусна (таксономія ID 10239) і яка взаємодія між людиною та всіма вірусними таксонами (таксономія ID 10239, `interSpeciesInteractome` функція). Дані таксономії були завантажені за допомогою Uniprot REST API (березень 2018, функція `loadTaxIDAllLower`). Ми зберігаємо ізоформи та пост-трансляційно оброблені ланцюги, а не обираємо канонічну послідовність за умовчанням. Це може бути особливо важливо для деяких вірусів, чиї білки транслюються як єдиний поліпептидний ланцюг, але потім розщеплюються на функціональні білки [61].

На додаток до бази даних IntAct ми використовували дані проекту BioPlex ([58] та неопубліковані дані), що включають близько 7500 експериментів з аффінної-очистки-масс-спектрометрії (AP-MS) для виявлення більш ніж 70000 взаємодій. Дані завантажено з веб-сайту BioPlex від 1 грудня 2017 р. (BioPlex 2.3) за допомогою функції `loadBioplex`. Ми використали маппінг ідентифікатора генів Entrez до UniProt accession, що надається BioPlex. Цей маппінг включає маппінг один ген до багатьох білків і багато генів для одного білка. У результаті, мережа взаємодії має певні взаємодії, які насправді не

перевірені. BioPlex може мати більш високий показник off-target мотивів (обговорюється в розділі 1.5).

Ми використали кілька підмножин даних у базі даних IntAct: 2 великомасштабні дослідження та 2 способи виявлення взаємодії (табл. 2.1).

Великі дослідження. Дані двох великомасштабних досліджень були відібрані за допомогою функції subsetMITABbyPMID: набір даних групи Mann [13] та набір даних групи Vidal [62].

Дослідження групи Mann створило 1330 стабільних клітинних ліній HeLa, які експресують 1155 різних білків-наживок (bait), які будуть використовуватися для AP-MS. Дослідження групи Vidal проводили з використанням двогібридного методу дріжджів [62].

Метод виявлення взаємодій: на основі методу виявлення взаємодії були створені дві підмножини бази даних IntAct: двогібридний та афінне очищення-мас-спектрометрія [13, 63]. Ми визначаємо двогібридний метод використовуючи онтологію PSI-MI: метод виявлення "аналіз комплементації транскрипції" (MI:0018) - усі методи, що належать до цього типу (які є в дитячих термінах в онтології). Ми ідентифікуємо метод AP-MS з використанням двох термінів онтології PSI-MI: метод виявлення "технологія афінної хроматографії" (MI:0004) та методика ідентифікації учасника "часткова ідентифікація білкової послідовності" (MI:0433). Використання термінів онтології для пошуку взаємодії дозволяє вказати лише один термін, а не перелік кожного окремого методу виявлення. Щоб визначити, які методи були включені, можна переглянути службу пошуку онтологій (<https://www.ebi.ac.uk/ols/ontologies/mi>). Функція subsetMITABbyMethod використовує онтологію для визначення всіх дочірніх термінів категорій, описаних вище, для фільтрування даних взаємодії.

Рандомізована мережа. Щоб отримати контрольний набір даних для пошуку мотивів, ми створили мережу, яка є ідентичною в кількості граней і ступеня для кожного взаємодіючого білка, але містить випадкові взаємодії. Для цього ми пермутували (переставили) взаємодії на другій позиції

(IDs_interactor_B). Ми рандомізували BioPlex та вірусно-людську мережу, які використовувались в пошуку мотивів.

Таблиця 2.1

Набори даних про взаємодію білків

Набір даних	Кількість білків	Кількість унікальних взаємодій
Вірусно-людська мережа	882 viral / 4544 human	14484
Мережа людини: всі дані IntAct	19573	156732
Мережа людини: BioPlex	12070	73665
Мережа людини: дані групи Mann	4952	15601
Мережа людини: дані групи Vidal	8638	44747
Мережа людини: двогібридні дані	13552	69298
Мережа людини: дані афінне очищення-мас-спектрометрія	11707	59153

2.2 Аналіз розподілу ступенів

Ми проаналізували розподіл ступеня людських та вірусних білків в мережі взаємодій білків людини-віруса, людини-людини та віруса-віруса. Ступінь - це кількість взаємодіючих партнерів білка. Ми завантажили набори даних взаємодій, як описано в розділі 2.1, використовуючи функцію loadHumanViralPPI. Ми підрахували кількість взаємодіючих партнерів вірусних білків, людських білків або вірусних білків у кожній мережі або її підгрупі за методом або дослідженням (функція humanViralDegree). Ми обчислили медіану кожного розподілу та візуалізували кожний розподіл, використовуючи розподіл щільності (N 3.3.1 та Додаток А). Крім того, ми

розрахували кількість взаємодій та кількість білків у кожній мережі. Результати цього аналізу обговорюються в розділі 3.1.

2.3 Білкові послідовності й передбачення доменів

2.3.1 Білкові послідовності

Ми використали Uniprot REST API (інтерфейс прикладного програмування) та Uniprot FTP [64] для завантаження послідовностей у форматі FASTA (20 жовтня 2017 р.). Ми використали функцію downloadFastaMixed (пакет MItools) для завантаження послідовностей всіх білків, включаючи ізоформи білків та послідовно оброблені ланцюжки. Ця функція завантажує всі канонічні та ізоформні послідовності в SwissProt за допомогою UniProt FTP. Потім вона завантажує послідовності non SwissProt один за одним, використовуючи UniProt REST API (функція downloadFasta). Нарешті, друга функція завантажує позицію пост-трансляційно процесованого регіону в послідовності білків та виділяє послідовність білка у використовуючи положення цього регіону (функція downloadFastaPostproc).

2.3.2 Передбачення домену за допомогою InterProScan

Для всіх людських та вірусних білків, що мають дані про взаємодію, доступні домени були ідентифіковані / передбачені за допомогою програмного забезпечення InterProScan та сигнатур послідовності InterPro. InterPro - це мета-база даних, яка збирає сигнатури послідовності доменів, родин доменів, сайтів та повторів [65]. Ми запускаємо InterProScan версії 5.25-64.0 в автономному режимі на кластері обчислень LSF x86_64-pc-linux-gnu (64-роздрядний, 8-ядерний, 16 Гб оперативної пам'яті), що працює під операційною системою Red Hat Enterprise Linux Server 7.3 (Maipo). Ми використовували наступні версії всіх баз даних: CDD-3.16 [66], Coils-2.2.1, Gene3D-4.1.0 [67], Hmap-201701.18 [68], MobiDBLite-1.0, Pfam-31.0 [69], PIRSF-3.02 [70], PRINTS-42.0 [71], ProDom-2006.1 [72], ProSitePatterns-20.132 та

ProSiteProfiles-20.132 [73], SFLD-2 [74], SMART-7.1 [75], SUPERFAMILY-1,75 [76], TIGRFAM-15,0 [77]. Код командного рядка, що було використано для запуску InterProScan, наданий у додатку Д.

Вивідна інформація була збережена у стандартному форматі GFF3 для зберігання діапазонів послідовності.

2.3.3 Видалення повторюваних доменів

Більшість баз даних InterPro, що входять до складу InterPro, містять сигнатури багатьох типів (домени, родини, сайти та повторення), тому потрібно було фільтрувати тільки сигнатури домену. Ми завантажили анотації типу сигнатур дляожної сигнатури InterPro з InterPro FTP за допомогою функції `getInterProEntryTypes`. Ми проаналізували вихідний файл InterPro за допомогою функції `readInterProGFF3`, об'єднали ці файли (функція `addInterProEntryTypes`) та вибрали вибрані домени (функція `SubsetByInterProEntryType`).

Деякі бази даних члени InterPro можуть містити сигнатури, що описують по суті один і той же домен послідовності. Ми спиралися на роботу InterPro з інтеграції сигнатур з різних баз даних, щоб видалити цю надмірність. Якщо дві бази даних члени InterPro забезпечують сигнатуру, яка відповідає одному і тому ж домену, InterPro вказує на це, надаючи єдиний ідентифікатор InterPro (наприклад, IPR002048). Ми використали цей метод та функцію `collapseByInterProID` для збереження лише одного регіону домену на один білок та ідентифікатор InterPro. Зауважте, цей метод зберігає всі домени, які належать до однієї родини. Наприклад, загальний протеїн кіназний домен та тирозин або серін-треонін-домени протеїн кіназні домени.

Далі ми перетворили діапазони доменів білкових послідовностей та їх анотації (об'єкт класу `Granges` у R) на таблиці домен-білкових пар. Ми називаємо це мережею домен-білок.

2.3.4 Постання домену людини з даними вірусно-людської взаємодії

На наступній стадії обробки даних ми об'єднали мережу домен-білок та мережу взаємодії вірусних та людських білків. Після цього ми розрахували кілька описових статистичних даних. Вони включали скільки білків містить кожен домен (фонове число); фонову частоту домену; скільки людських білків є мішенями кожного вірусного білка; скільки людських взаємодій вірусного протеїну містять кожен домен (також називають доменним числом); яке збагачення домену серед інтеракторів вірусного білка. Вивчено залежності між цими мірами (не показано, але включене до файлу .Rmd). Ці міри використовуються для оцінки того, які домени, ймовірно, опосередковують взаємодію.

2.4 Статистичний метод оцінки того, які домени імовірно опосередковують взаємодію

Edwards та співавт. продемонстрували, що справжні мотиви часто передбачаються неправильними даними про взаємодію, "off-target" мотиви [12]. Експериментально визначені інтерактори білка A дають мотив, проте цей мотив не опосередковує взаємодію білку A з білками, що містять мотив. Замість цього білок B розпізнає цей мотив у підмножині взаємодій білка A. Теоретично, оцінка того, які домени, ймовірно, опосередковують взаємодію, повинна поліпшити передбачення "off-target" мотивів. Ми можемо оцінювати мотиви шляхом вивчення їхніх доменів-кандидатів розпізнання (розділ 3.7). Крім того, ми можемо відфільтровувати набори даних для пошуку мотивів, для яких надійне передбачення домену розпізнавання (рис. 2.4).

Ми розробили метод ідентифікації доменів, збагачених серед білків людини що є мішенню одного вірусного білка. Збагачені домени можуть служити в якості проксі для доменів що опосередковують взаємодію. Ці домени можуть розпізнавати SLiM у вірусних білках, зв'язувати вірусні білки через їх взаємодію між доменами або показати збагачення з функціональних причин.

Замість того, щоб обчислювати ймовірність виявлення певного домену N разів серед інтеракторів вірусного білка, з огляду на його фонове число, ми вирахували ймовірність будь-якого домену бути присутнім N разів серед інтеракторів цього вірусного білка (рис 3.3, функція permutationPval) . Ми обчислили це шляхом рандомізації людських мішеней вірусних білків, зберігаючи ступінь та загальну кількість взаємодій незміненими; як якщо б вірусні білки вибирали людські білки незалежно від їхнього доменного складу. Для кожного вірусного білку ми обчислили, скільки разів ми бачили кожен домен. Тоді ми розраховували, як часто частота пермутованих доменів більша або дорівнює кількості спостережених доменів. Це надає емпіричну величину p-value - ймовірність того, що кількість доменів буде також висока чи вище за нульову гіпотезу. Фон за пермутацій розраховується для кожного вірусного білка, щоб пояснити різну кількість взаємодій цих білків.

Прямо не включаючи фонову частоту домену в обчислення ми підвищуюмо надійність збагачення домену порівняно з гіпергеометричним тестом. Edwards та співавт. [12] обговорили проблеми використання гіпергеометричного тесту для пошуку збагачених мотивів: відсутність композиційної рівномірності протеома, різниця в довжині білка. У нашому випадку рідкісні домени повинні обов'язково збагатитись у будь-якому наборі білків людини-вірусних мішеней навіть у кількості 1 або 2 через незначну кількість білкових взаємодій що білки, як правило, мають.

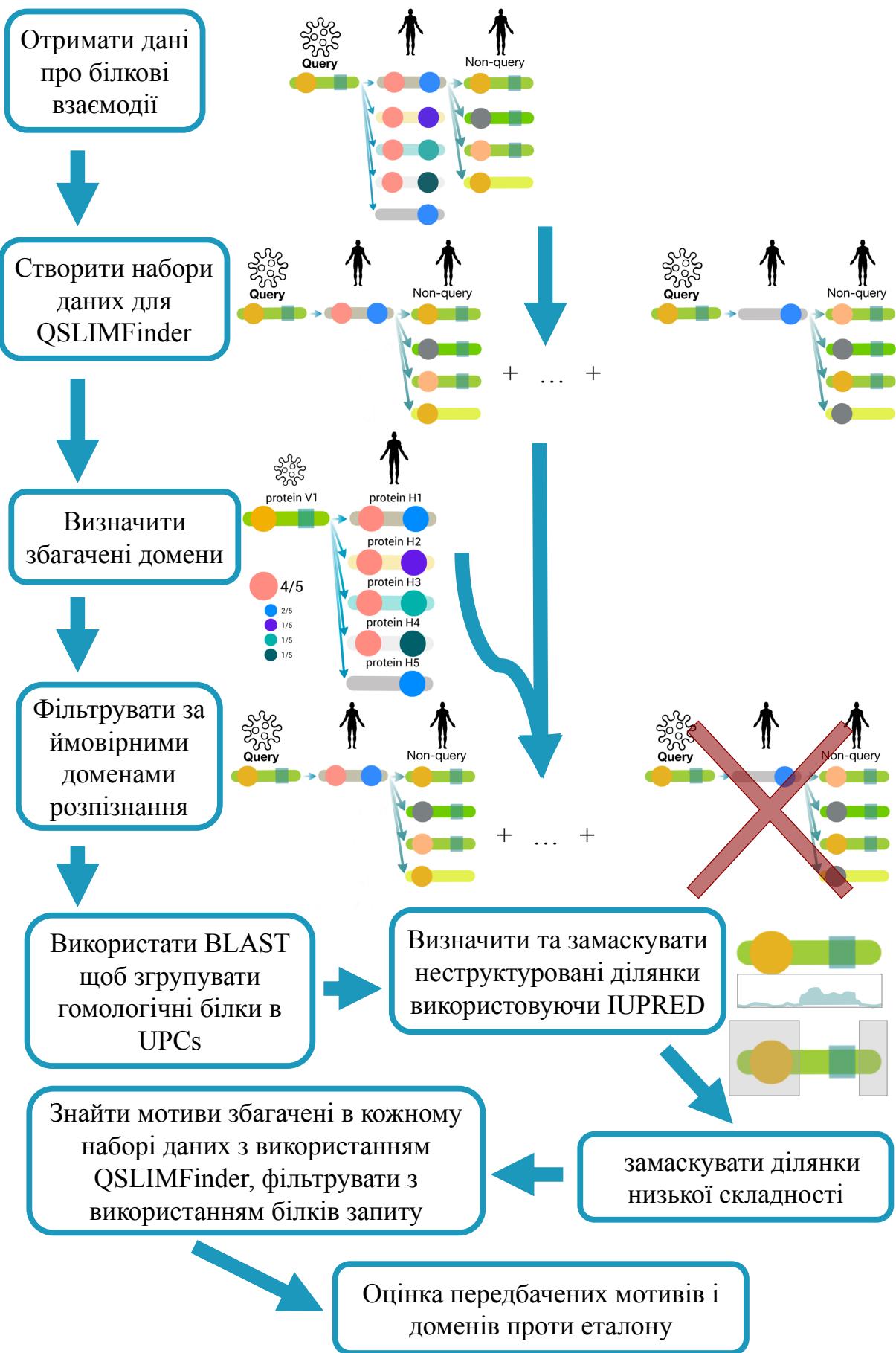


Рис 2.4. Схема процедури пошуку мотивів

2.6 Інструменти та процедура пошуку мотивів

QSLiMFinder [59] - це інструмент командного рядка, який потребує послідовності білків як текстовий файл FASTA, назву білка запиту як текстового файлу, інші параметри, не специфічні для кожного набору даних. У нашому аналізі кожен набір даних визначається комбінацією білка людини - мішені вірусу (називається seed) та вірусним білком, який використовується як запит (query, рис 2.4). Ми використовували послідовності вірусних білків (рис. 3.4.1 В) або людських та вірусних білків (рис. 3.4.1 С), які взаємодіють з цим єдиним білком людини - мішенню віруса. Якщо цей людський білок має більше однієї вірусної взаємодії, кожна з цих взаємодій використовується як запит. Інструмент QSLiMFinder виключає послідовність запитів із набору послідовностей, що використовуються для розрахунку статистики збагачення мотивів. Мотиви представлені як регулярні вирази (regular expressions, regex). Розраховується ймовірність спостереження за рядом випадків сумісних регулярних виразів у заданому наборі білкових послідовностей. Послідовність білків запитів використовується для фільтрації сукупності оцінюваних регулярних виразів. Це покращує чутливість, зменшуючи кількість протестованих гіпотез.

Таблиця 4.1 у додатку Г підсумовує комбінації наборів даних та інших параметрів, які ми оцінили за допомогою тестування проти еталонного набору даних.

2.6.1 Програмне забезпечення для пошуку мотивів

Ми використовували інструмент командного рядка QSLIMFinder, який є частиною версії SLIMSuite випущеної групою Edwards 2016-09-12 [7, 78]. Гомологічні послідовності, швидше за все, містять ті ж самі паттерни амінокислотної послідовності, і тому можуть штучно підсилювати підтримку для кожного мотиву. QSLIMFinder групували гомологічні послідовності з використанням NCBI BLAST 2.6.0 [79] для отримання неспоріднених білкових

кластерів (UPC або UP). Крім того, короткі лінійні мотиви, як правило, розташовані в неупорядкованих областях, тому невпорядковані ділянки були масковані, використовуючи програмне забезпечення передбачення невпорядкованих ділянок білка IUPRED, отримане 4 вересня 2017 р. [80]. Ми скомпілювали це програмне забезпечення з джерела на кластері обчислень LSF x86_64-pc-linux-gnu під управлінням Red Hat Enterprise Linux Server 7.3 (Maipo).

2.6.2 Створення наборів даних для пошуку мотивів

Як описано в розділі 2.6, на рисунках 2.4 та 3.4.1, набори даних для пошуку мотиву визначаються білками seed та query. Ми повинні були створити два файли для кожного набору даних для QSLIMFinder: файл FASTA, що містить послідовності білків, які взаємодіють із seed, і текстовий файл, що містить ідентифікатор білка query. Щоб створити ці файли та створити BASH команди, які запускають QSLIMFinder з цими файлами та додатковими параметрами, ми інтегрували дані про білкову взаємодію, дані послідовності білків та дані домену. Ця послідовність дій реалізована як функція PPInetwork2SLIMFinder.

Ми використовували як seed всі білки людини-мішені вірусних білків або білки, що, принаймні, мають один домен передбачений необхідним для взаємодії з принаймні одним вірусним білком при значенні порогового p-value 0,5. Цей поріг був обраний емпірично на основі порівняльного аналізу з еталоном. Вибір більш жорстких порогових значень не покращив від cliкання так само, як і видалення білків без ймовірних доменів, що опосередковують взаємодію (розділ 3.5). З цих seed білків ми вибирали ті, що мали дані про послідовності білків (розділ 2.3.1).

Список фільтрованих seed білків потім використовувався для створення наборів даних для QSLIMFinder, як показано на рисунку 2.4 (функція listInteractionSubsetFASTA). Коли seed білок людини мав більше однієї вірусного взаємодії, ми додали інші вірусні білки до non-query набору. Далі

цей список наборів даних для QSLIMFinder був відфільтрований, щоб включити ті, де query блок має збагачений домен за заданого порогове значення та мінімальну кількість послідовностей у кожному наборі даних (1 вірусний query і 2 вірусних або 2 людських non-query). Нарешті, файли, що містять послідовності та імена білків запитів, були створені та шляхи до цих файлів збережені.

2.6.3 Процедура пошуку мотивів

Для створення команд BASH, які запускають QSLIMFinder використовуючи кожний набір даних, ми поєднали список директорії файлів із директоріями до програмного забезпечення QSLIMFinder та інших параметрів (функція mQSLIMFinderCommand). Приклад команди надано у Додатку Д.

Параметри, які ми використовували, будуть розглянуті в цьому абзаці (всі інші використано за замовчуванням). Використовувалося маскування невпорядкованих ділянок (dismask = T), за умовчанням - 0.2 порог іупред значень. Маскування консервацією не використовувалося, оскільки пошук мотивів здійснювався за допомогою не-гомологічних вірусних білків. Всі мотиви нижче порогу ймовірності QSLIMFinder Sig 0.3 були збережені (probcut = 0.3). Ми використовували довжину мотиву за замовчуванням (кількість визначених позицій, slimlen = 5) та кількість послідовних невизначених позицій (minwild = 0 maxwild = 2). Довгі мотиви можна виявити як набір з кількох коротших мотивів. Ми обмежили кількість послідовностей в одному наборі даних до 800, що пояснюється обмеженнями часу роботи (maxseq = 800).

Ми протестували варіант обмеження результату до клауду з 1+ фіксованим мотивом (cloudfix = T), а ні (cloudfix = F). Мотив клауд - це групи мотивів, які перекриваються в 2 визначених позиціях. Деякі клауди включають лише один мотив і той є двозначний, і Edwards рекомендує видалити ці мотиви [78]. Коли ми додали ці мотиви, ми виявили більше правильних мотивів на

більш м'яких порогах. З іншого боку, цей підхід додало більше помилково-позитивних/нових мотивів кандидатів, що робить метрики точності та відкликання дуже схожими для обох варіантів.

2.7 Порівняльний аналіз екземплярів мотивів до еталонних даних

2.7.1 Еталонні дані

Щоб оцінити, чи зможемо ми передбачити короткі лінійні мотиви, ми перевірили, наскільки добре ми передбачаємо набір відомих лінійних мотивів у вірусних білках. Ми зібрали всі лінійні мотиви в вірусних білках, які були анатовані базі даних Eukaryotic Linear Motif (ELM) станом на листопад 2017 року [45]. Цей набір даних містить регулярні вирази, які визначають мотиви та екземпляри 243 мотивів у 143 вірусних білках. З них ми обрали лінійні мотиви у вірусних білках, які, як відомо, взаємодіють з людськими білками. Ми включили ліганд-зв'язуючі, пост-трансляційно модифікаційні та докінг мотиви, але виключили дегрони, мотиви розщеплення та таргетингу. Ці типи мотивів, як правило, є більш загальними та присутні у багатьох білках. Наприклад, мотиви таргетингу можуть бути присутніми у вірусних і людських білках через їх спільну локалізацію, але не тому, що вони опосередковують взаємодію з білком інтересу, що робить їх легкими для виявлення, але не є актуальними для нашого дослідження.

Остаточний еталонний набір даних для порівняння містить 51 вірусний білок. Для кожного набору даних для пошуку мотивів, набір даних для порівняльного аналізу ще більше скорочується, щоб включати лише ті білки, в яких ми шукали мотиви. Найбільший набір для порівняння, який ми використовували, містить 52 мотиви з 35 вірусних білків. Даний набір даних побудований з використанням взаємодій між вірусними та білками людини, а також між білками людини-мішенями вірусу та іх партнерами в мережі людини (рис 3.4.1 C).

Для тестування наших передбачень доменів, які, ймовірно, опосередковують вірусно-людську взаємодію, ми використовували список відомих мотив-зв'язуючих доменів, котрі анотовані в базі даних ELM. Основною функцією цих доменів є опосередкування взаємодій. Ми сподіваємось, що правильна процедура передбачення доменів, які можуть опосередковувати взаємодії, повинна передбачати SLIM-зв'язуючі домени як ті, що опосередковують взаємодію частіше, ніж інші домени. 118 з цих доменів присутні в 1016 людських білках-мішенях 597 вірусних білків.

2.7.2 Процедура порівняльного аналізу

Метою порівняльного аналізу було визначити, які параметри пошуку мотивів найкраще працюють, і вибрати порогову позицію з прийнятною точністю та відкліканням. Для цього ми виявили, які екземпляри мотивів виявлені за низкого порогу QSLIMFinder Sig в 0,3 збігаються з відомими прикладами з бази даних ELM. Виявлені унікальні мотиви (по позиції регіону в білку) повинні відповідати принаймні 2 позиціям амінокислот відомих мотивів. Це можна було б ще покращити, оцінюючи визначені позиції в регулярних виразах.

По-перше, ми завантажили дані набору збагачених доменів і набори даних мотивів, підготовлені для QSLIMFinder. Ми необов'язково фільтрували обидва набори даних за ймовірністю домену (розділ результатів 3.5-3.7). Далі ми виділили *de novo* відкриті мотиви, які були виявлені з використанням відфільтрованих наборів даних QSLIMFinder та екземпляри ELM, які могли бути виявлені за допомогою цих наборів даних. Екземпляри ELM були відфільтровані для певних типів мотивів. Ми не об'єднували два мотиви, якщо тип мотиву був іншим. Наступним кроком ми розрахували спільний предиктор, який включає значення p-values домену та мотиву. Цей предиктор не покращив відкриття відомих мотивів (результати не показані), що пропонує більш складний підхід до їх інтеграції (обговорюється в розділі 3.8.2). Ми використовували p-value для мотиву у всіх аналізах.

Ми використовували p-value для кожного унікального мотиву, як предиктор двоїчного результату: відповідність відомому істинному мотиву проти помилкового позитивного або нового мотиву-кандидата. Кілька прогнозованих мотивів можуть відповідати одному відомому мотиву, наприклад, 3 варіанти мотиву PDZ на рис 3.7.2.

Ми проаналізували продуктивність на різних порогів, використовуючи пакет ROCR R та функцію mBenchmarkMotifsROC для організації моого аналізу. Ми досліджували точність, відкликання, істинно позитивну швидкість, хибну позитивну швидкість при кількох порогах. Ми використали цей аналіз, щоб вибрати три значення порогу p-value: м'який поріг за 0,3; оптимальний поріг при мінімальному р-значенні, коли точність більша, ніж відкликання (змінюється в різних наборів даних); і суворий поріг, коли точність більше 0,5 (змінюється в різних наборів даних).

2.7.3 Приклади відкритих заново та мотивів-кандидатів

Ми вирішили вивчити відкриті заново та мотиви-кандидати, передбачені при суворому порозі, використовуючи комбінацію мереж взаємодії білків вірусів-людини та людини (IntAct) та фільтрування за доменом (рис 3.4.1 С та 2.4). Ми оцінили, чи домени, які імовірно опосередковують взаємодії для кожного вірусного білка, що містить мотив, є відомими SLIM-зв'язуючими доменами. Для того, щоб візуалізувати результати за допомогою Cytoscape, ми перетворили результати порівняльного аналізу на спрямовану мережу: людський білок -> область визнання -> мотив -> вірусний білок. Для масштабування розміру вузла ми використовували p-value мотиву та домену.

2.8 Процедура визначення подібності паттерну мотивів

Для порівняння подібності паттерну мотивів для всіх мотивів, виявлених за жорстким порогом з використанням набору даних IntAct (qslimfinder.Full_IntAct3.FALSE), ми порівняли регулярний вираз, що

визначає відкриті мотиви до всіх відомих мотивів у базі даних ELM. Ми використовували програмне забезпечення Comparimotif V3.13.0 для виконання всіх попарних порівнянь та зберегли подібність мотивів [81]. Складність мотивів можна описати, використовуючи інформаційний вміст (IC). IC описує, наскільки зменшення невизначеності забезпечується мотивом. Ми запускали Comparimotif як інструмент командного рядка, включений в SlimSuite (обговорюється в розділі 2.6.1) з налаштуваннями за замовчуванням. Результати з цієї програми завантажено в Cytoscape. Ми використовували евристичну оцінку 1,162 (кількість співпадаючих позицій х Нормалізований IC) для фільтрування мережі подібності мотивів.

2.9 Технічне обладнання

Для аналізу був використаний обчислювальний кластер LSF x86_64-pc-linux-gnu, операційна система Red Hat Enterprise Linux Server 7.3 (Maipo) конфігурації є різною для кожного завдання та зазначена у відповідних розділах. Альтернативно був використаний комп’ютер даної конфігурації: процесор - 2.9 GHz Intel Core I5; Пам’ять - 16 GB 1867 MHz DDR3; операційна система: MAC OS Sierra V10.12.

2.10 Статистичний аналіз даних

Статистичний аналіз та обробка даних виконувалися за допомогою мови статистичного програмування R. Для передбачення доменів у послідовностях білків, передбачення доменів, що опосередковують взаємодії, передбачення мотивів, оцінки схожості мотивів та порівняльного аналізу використовувалися складні статистичні моделі та методи, описані у відповідних розділах чи оригінальних публікаціях. Де була потреба, власні методи були запрограмовані і включені до пакету R під назвою MItools [82]. Цей пакет є публічно доступним. Всі етапи аналізу були виконані та описані з використанням відтворюваних документів R Markdown, де код аналізу

доповнюється текстовим описом (*.Rmd). Будь-які рисунки чи інші результати, отримані за допомогою коду аналізу, а також подробиці про середовище R, були включені у вихідні документи (*.html), коли аналіз був виконаний. Ці аналітичні документи, деякі вхідні дані, вихідні дані та результати проекту були організовані в проекті языку програмування R [83]. Наступні файли охоплюють аналіз, описаний у попередніх розділах:

- interactions_and_sequences.Rmd: секції 2.1, 2.3.1 and 2.3.2
- remove_redundant_domains.Rmd: секція 2.3.3
- map_domains_to_human_viral_network.Rmd: секція 2.3.4
- what_we_find_VS_ELM_count_justFisher.Rmd: секція 2.4
- Motif_search_strategies_IntAct_Vidal_viral2.Rmd: секція 2.6
- compr_benchmarking_strateg_IntAct_Vidal.Rmd: секція 2.7, 2.8
- compr_benchmarking_strateg_cloudfixF_IntAct_BioPlex.Rmd: секція 2.7
- compr_benchmarking_venn.Rmd: секції 2.7.2, 2.7.3
- Degree_distribution_in_the_network.Rmd: секція 2.2

РОЗДІЛ 3

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

3.1 Дослідження мережі взаємодій білків людини між собою та з білками вірусів

3.1.1 Дослідження асиметрії вірусно-людської мережі білкових взаємодій

Щоб краще зрозуміти взаємодію вірусних та людських білків на системному рівні, ми розглянули тенденції щодо кількості взаємодій, що ці білків утворюють. Крім того, Рис 1 підсумовує дані про білкову взаємодію, які ми використовували в нашому дослідженні (кількість білків, кількість взаємодій та їх розподіл між білками).

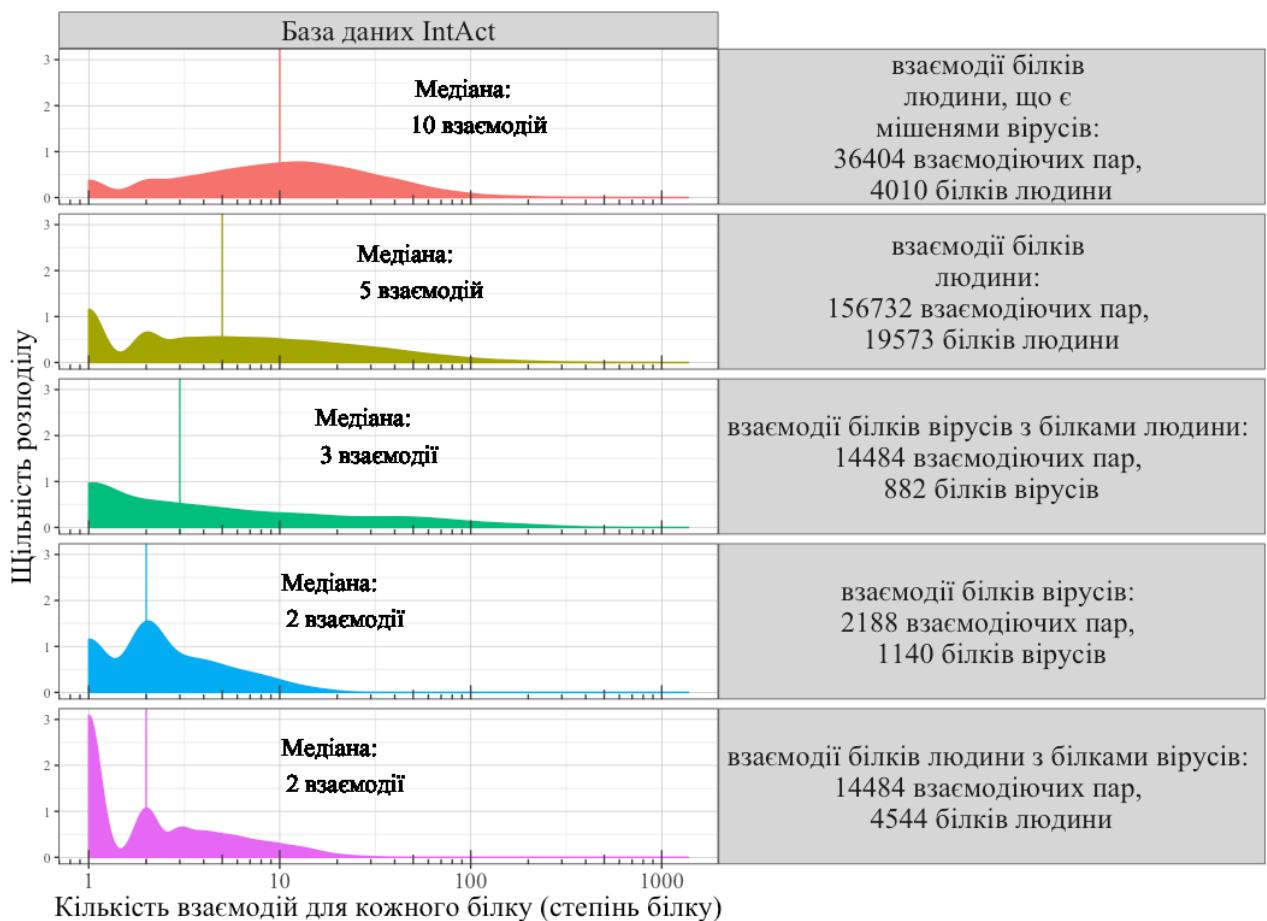


Рис 3.1.1. Графік, що показує щільність розподілу числа взаємодій кожного людського чи вірусного протеїну в кожній мережі, яку ми використовували для нашого аналізу. Для кожного білка вісь X показує

кількість білків що з ним взаємодіють, вісь Y показує щільність розподілу. Різні мережі та різні білки (вірусні чи людські) показані в рядках. Верхній рядок показує, що 19399 білків людини утворюють мережу з 155702 взаємодіями з 5 взаємодіями на білок в середньому (медіана). Рядки 2 і 3 показують розподіли кількості взаємодій у вірусно-людській мережі для вірусних та людських білків відповідно.

Як вірусно-людська мережа, так і людська мережа є найбільшими серед тих, що були абиcoli використані для виявлення мотивів на сьогоднішній день. Ми не використовували жодних підмножин вірусно-людських даних, щоб зберегти якомога більше взаємодій, хоча і за певного зниження якості.

Ми спостерігаємо чотири основні тенденції:

- Як людські, так і вірусні білки в вірусно-людській мережі мають у середньому меншу кількість анатованих взаємодій, ніж людські білки, у мережі людини. Людсько-вірусні взаємодії менш вивчені. Менше систематичних досліджень проводилося для виявлення всіх взаємодій вірусних білків.
- Вірусні білки взаємодіють з більшою кількістю білків людини, ніж білки людини взаємодіють з вірусними білками. Це буде обговорено пізніше.
- Білки людини, мішені віруса, мають в середньому в 2 рази більше взаємодій. Це буде обговорено пізніше.
- Людсько-вірусні дані дуже неповні: половина людських білків мають 2 або менше взаємодій з вірусними білками, половина вірусних білків мають 3 або менше взаємодій з людськими білками. Це дозволяє припустити, що багато з цих вірусно-людських взаємодій не можуть забезпечити достатньо інформації для виявлення мотивів. Це та нещодавня розробка спеціалізованого інструменту Palopoli та співавт. [59] мотивували наш підхід до пошуку мотивів у людській мережі, використовуючи послідовності кожного вірусного білка як фільтр, а не шукаючи мотиви тільки у вірусних білках.

Давайте обговоримо другу тенденцію більш докладно. Ми виявили, що вірусні білки, як правило, взаємодіють з багатьма білками людини, тоді як людські білки взаємодіють лише з кількома вірусними білками (рис 3.1.1, рядки 2 і 3). Це може відображати біологічну потребу вірусів перешкоджати роботі кількох клітинних процесів. Крім того, ця різниця може відображати технічний аспект вивчення взаємодій між вірусами: більшість вірусних білків можливо були використані як приманка (*bait*), оскільки для виявлення цих взаємодій необхідна вірусна інфекція або екзогенна експресія вірусних білків. Крім того, ми бачимо загальну тенденцію до того, що вірусні білки мають менше взаємодій ніж білки людини у середньому, що може відображати той самий упереджений вплив: на сьогоднішній день було проведено набагато менш високопродуктивних досліджень білків людини та вірусів. У 36 дослідженнях людей було використано понад 50 приманок, тоді як лише у 5 вірусних. Це означає, що основна частина даних походить з невеликих цільових експериментів, а не скрінів взаємодій протеїнів. Додаток А показує, що при дослідженні взаємодій між вірусними та людськими білками з використанням двогібридного методу виявлення взаємодій (*two-hybrid*) вірусні білки мають значно більше ідентифікованих взаємодій, ніж людські білки, і надалі підтримують гіпотезу про те, що менша кількість вірусних взаємодій білків людини може бути частково пояснена технічними причинами.

3.1.2 Дослідження ефекту упередженості даних на центральність білків людини, що є мішенями вірусів

Віруси обирають як мішені людські білки, що є центральними, лише у даних упереджених присутністю добре вивчених білків. Література часто говорить, що вірусні білки цілять центральні білки людської мережі (білки з багатьма взаємодіями) [84, 85]. Проте останнім часом з'явилися кілька досліджень, які свідчать про те, що достатньо визнана асоціація між

релевантністю для захворювань та великою кількістю взаємодій у мережі білкові взаємодії можуть бути завищенні, якщо дослідники враховують дослідницьку упередженість даних [86]. Ця упередженість означає, що краще вивчені білки мають більше взаємодій, що перешкоджає корисності ступеня білка як міри функціональної важливості білка. Для вирішення цієї проблеми та поліпшення покриття мереж білкової взаємодії проводиться кілька систематичних досліджень взаємодій між ~ 17000 білками людини [13, 58, 62]. Ми можемо використовувати ступінь білку в кожному з цих досліджень як кращу міру справжньої функціональної важливості білків.

У попередньому розділі ми побачили, що людські білки, мішенні вірусів, як правило, мають набагато більше взаємодіючих партнерів (медіана 10 у порівнянні зі середнім значенням 5 для білків, не мішених вірусів). Це може бути пояснено функціональною різницею та поведінкою вірусів, обираючих як мішень центральні білки, або технічним та дослідницьким упередженням. Методи очищення афінністю, як правило, витягають та вимірюють білкові комплекси, а не прямі та бінарні взаємодії, що призводить до того, що білки мають більше взаємодій при вимірюванні за допомогою цих методів. Багато взаємодій між вірусами та людьми в нашому наборі даних походить з цього типу досліджень, що може пояснити частково вищий ступінь. Інша можливість полягає в тому, що білки-мішенні вірусів, більш вивчені в цілому.

На рисунку 3.1.2, ми бачимо, що в той час як метод аффінного очищення з подальшою мас-спектрометрією (AP-MS) був використаний для виявлення більшої кількості взаємодій білків-мішенні вірусів ніж дигибридний метод, метод виявлення взаємодій не повністю пояснює більш високі кількості взаємодій. Дослідницька упередженість пояснює це: як дані Mann і співавт., так і Vidal і співавт., мають ідентичні медіани та ідентичну форму розподілу для білків-мішенні вірусів і всіх людських білків.

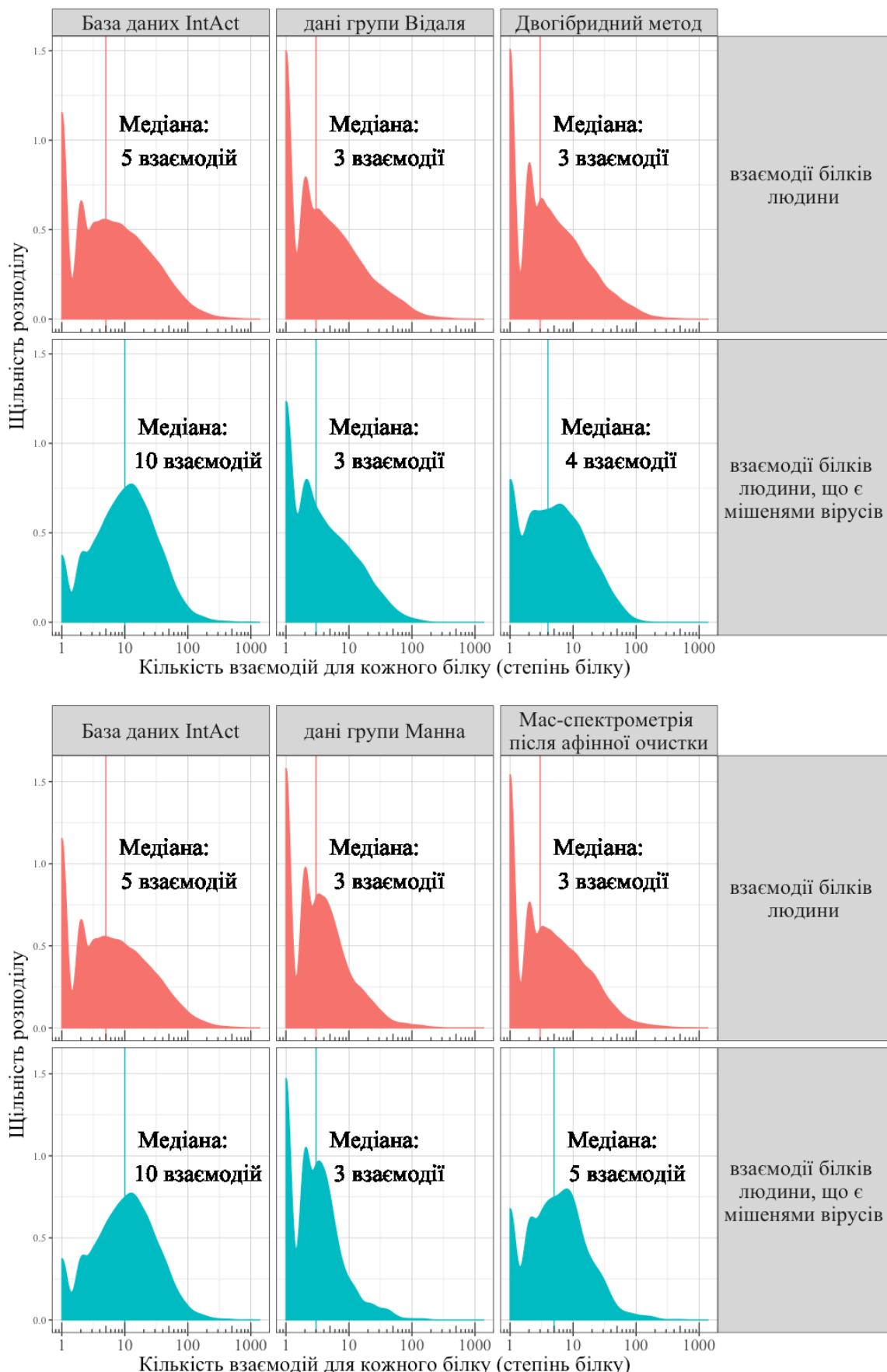


Рис 3.1.2. Графік, що показує щільність розподілу кількості взаємодій для всіх білків людини або білків людини-мішеней вірусів, для всіх даних, або

обраних методів виявлення взаємодії протеїнів та широкомасштабних неупереджених досліджень. Логарифмічно трансформована вісь X показує кількість білків, що взаємодіють з кожним білком (ступінь білка), вісь Y показує щільність розподілу.

Ця тенденція вірусних білків взаємодіяти з найбільш вивченими людськими білками може бути особливо сильною, оскільки багато з цих найбільш вивчених білків (у тому числі P53 [87]) були вперше виявлені через їх вірусні взаємодії (основний метод відкриття людських білків перед тим як першим геном людини був секвенований).

3.3 Дослідження доменів, що імовірно опосередковують взаємодію між білками

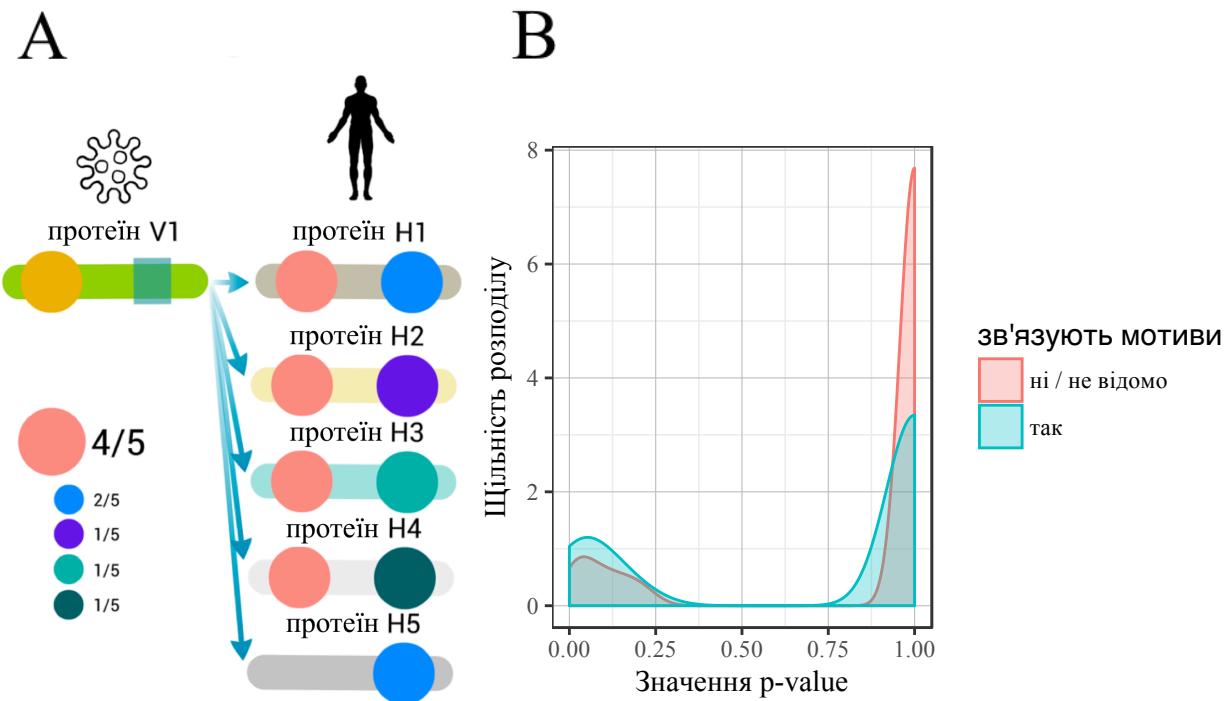
Знаючи, що вірусний білок обирає як мішень білки людини, що містять певний домен, можна фільтрувати дані, щоб збільшити співвідношення сигнал-шум у пошуку мотивів. Щоб оцінити це, ми знайшли домени, збагачені у білках людини, які взаємодіють з кожним вірусним білком.

Хоча, загальноприйнято передбачати взаємодій між доменами використовуючи дані про взаємодію з білками [58], наскільки мені відомо, ніяких спроб прогнозувати взаємодії домену з білком не було опубліковано. Гіпергеометричний розподіл зазвичай використовується для розрахунку ймовірності перекриття в елементах між категоріями. У цьому випадку перекриваються взаємодіючі партнери вірусного протеїну та білки з певним доменом за нульовою гіпотезою. Однак цей тест дає p-value для збагачення конкретного домену, але не для збагачення будь-якого домену, оскільки цей метод залежить від порівняння частоти білків з певним доменом в наборі тих, які взаємодіють з певним вірусним білком до фонової частоти цього домену. Показано, що підходи, які залежать від фонової частоти, наприклад в повному протеомі, зазвичай погані при оцінці фонового розподілу для виявлення

збагачених мотивів через неоднорідний склад послідовностей білків протеома [7, 88].

Ми вважаємо, що гіпергеометричний розподіл також непридатним для прогнозуванні доменів взаємодії. Ми бачили, що низька кількість взаємодій може штучно піднімати частоту домену в наборі (не показана). Коли білок взаємодіє з лише трьома іншими білками, мінімальна частота будь-якого домену становитиме 0,33, що призведе до того, що навіть найрозповсюдженіші домени будуть збагаченими в цьому наборі. Наприклад, нуклеозидтрифосфатна гідролаза що містить Р-петлю (P-loop containing nucleoside triphosphate hydrolase), є найбільш поширеним доменом у фоновому наборі білків-мішеней вірусів, але його частота становить лише 0,01; що означає, що домен буде в 5 разів збагачений, навіть якщо він присутній лише у 1 з 20 білків. Ці проблеми роблять гіпергеометричний розподіл непридатним для ідентифікації збагачених областей.

Для боротьби з цими проблемами ми розробили процедуру на основі перестановок для розрахунку вірогідності бачити будь-який домен N числа разів серед білків взаємодіючих з вірусним білком (рис. 3.3 А). На відміну від тесту Фішера, наша процедура виділяє домени, збагачені відомими доменами що розпізнають SLIM (однобічний тест Колмогорова-Смірнова з двома зразками, $D^- = 0.13548$, $p\text{-value} < 2.2e-16$). Відомі домени що розпізнають SLIM мають переважно низькі $p\text{-value}$ (рис. 3.3 В). Це дозволяє нам використовувати збагачення домену як проксі для визначення домену, який, ймовірно, опосередковує взаємодію (включаючи мотив-опосередковану взаємодію).



Ми бачимо 2 основних обмеження цього підходу:

1. Вірусні білки можуть обирати як мішень функціонально пов'язані білки людини. Ці білки можуть мати спільну доменну архітектуру, тому ми, можливо, не зможемо розрізнати, який з доменів більш імовірно опосередковує взаємодію. Одним із прикладів є можливий домен SH3 - зв'язуючий білок, який зв'язує 4 кінази з ідентичною доменною архітектурою (розглянута пізніше). Інший приклад - це домен нуклеозидтрифосфатної

гідролази, що містить Р-петлю. Якщо він збагачений, він відображає перевагу вірусу зв'язувати білок з GTP-ase активністю; однак інший набір доменів може бути відповідальним за зв'язування.

2. Деякі людсько-вірусні взаємодії опосередковуються взаємодіями між доменами. Багато збагачених доменів буде опосередковувати зв'язування, але не допоможе відкрити мотиви.

Ми вибрали значення p-value 0.5, щоб виключити всі домени, які, ймовірно, не забезпечують взаємодії з вірусними білками. Під час побудови набору даних для пошуку мотивів ми використовували всі інші пари доменнів та білків: ми шукаємо лише мотив у вірусному білку, який має ймовірний домен розпізнавання в людському білку. Після фільтрування ми залишили 5379 взаємодій між 396 вірусними білками та 754 збагаченими доменами людини.

3.4 Пошук коротких лінійних мотивів

Ми визначили (SLIMs), що конвергентно еволюціонували в вірусних білках з використанням імовірнісного методу, розробленого Edwardsi співавт. (QSLIMFinder) [59]. Ці мотиви є збагаченими у білках, які взаємодіють з білками-мішенями вірусів. Для цього аналізу ми припустили, що кожен вірусний білок має мотив, що розпізнається глобулярним доменом в білку людини. Один і той же домен може розпізнавати екземпляри цього мотиву в білках людини (рис 3.4.1 C) або в інших вірусних білках (рис 3.4.1 B). Метою обох підходів є відкриття послідовності цього мотиву.

Замість того, щоб покладатися на значення p-value, скориговане за частотою помилкового відкриття (FDR), яке було надано програмою QSLIMFinder як показник частоти помилкового відкриття, ми оцінили ефективність нашого підходу, порівнявши передбачені мотиви з відомими мотивами на трьох різних порогах статистичної значимості (рис. 3.4.2). Відомі мотиви були взяті з бази даних ELM, як описано в розділі 3.2. Ми оцінюємо

ефективність передбачення мотивів у вірусних білках-запитах, проте ми також передбачаємо мотиви в людській мережі.

Ми можемо відкрити заново відомі мотиви, використовуючи обидві стратегії, показані на рисунку 3.4.1 В та С. Рис 3.4.2 наводить розбивку кількості виявлених мотивів-кандидатів та відомих мотивів, які ми відкрили заново. Хоча ми застосували ті самі критерії до встановленого порогу, підхід, що використовує лише вірусні дані, вимагає більш низького скорегованого значення p-value для відкриття заново тієї ж частки відомих мотивів з тією самою частотою помилки, ніж підхід, що включає всі дані людини з бази даних IntAct [60].

Ми також використовували дані про взаємодію білків з великого неупередженого скріну, зробленого групою Vidal [62] (Додаток Б). Ці дані працюють гірше, ніж всі дані з бази даних IntAct. Це узгоджується з попереднім дослідженням Edwards та співавт, які показали, що їх метод пошуку мотивів (SLIMFinder) чутливіший до відсутності сигналу, ніж до наявності шуму [89, 12]. Це означає, що важливіше зберігати якомога більше білків з мотивом, навіть за рахунок додавання більшої кількості білків, що не мають мотиву. Краще мати 10/100 ніж 3/6 білків, що містять мотив.

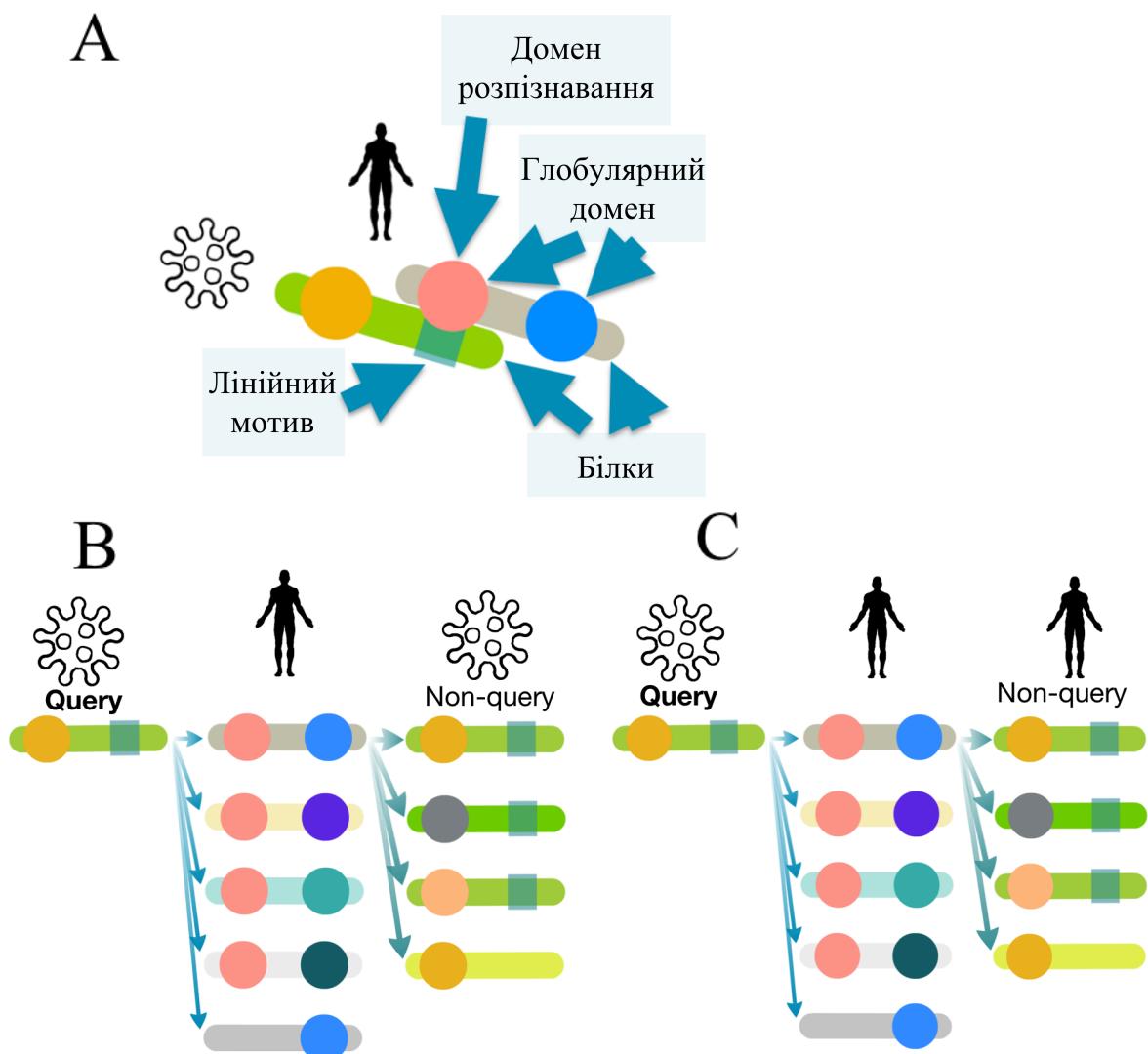


Рис 3.4.1. Схема, яка показує, як будуються набори даних для пошуку мотивів. Білки не запиту (non-query) використовувались для пошуку мотивів, які повинні бути присутніми в білку запиту (query). Кожен набір даних складається з усіх взаємодіючих партнерів одного білка людини та одного білка запиту (query). А. Легенда. В. Набори даних можуть бути побудовані з використанням білків людини, які взаємодіють з декількома вірусними білками. С. Білок-запит може бути вірусним білком, що мімікрує мотив, присутній у білках-не-запиту людини. Додавання цих білків-не запиту може забезпечити більшу потужність і інтерпретативність по відношенню до сухо вірусного набору даних.

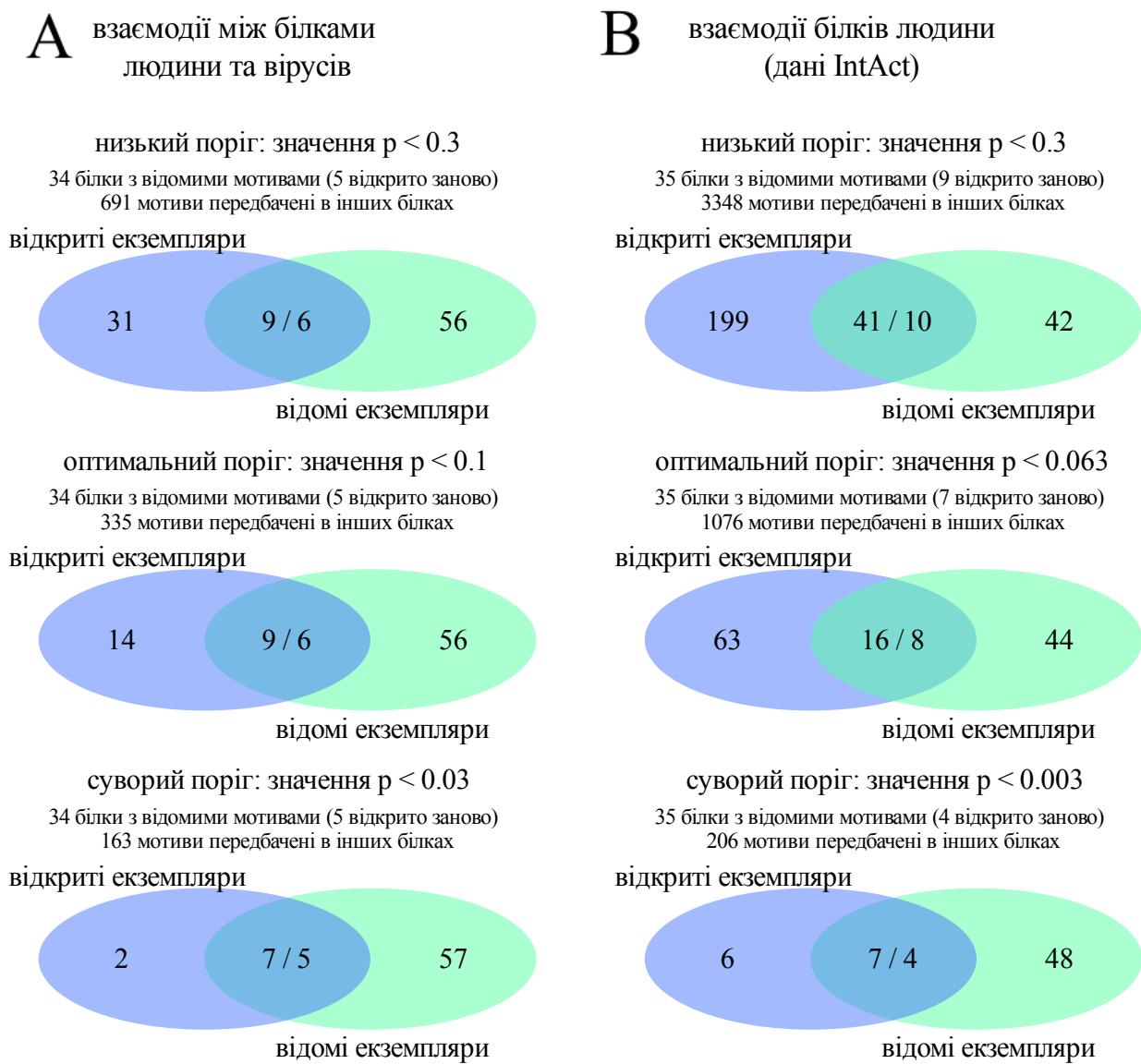


Рис 3.4.2. Діаграми Вена, що показують кількість виявлених мотивів-кандидатів та відомі мотиви, які відкрили заново. Наведено дві стратегії побудови наборів даних (A та B) та 3 пороги значень p -value. Блакитне коло показує кількість екземплярів мотивів, передбачених, але невідомих. Зелене коло показує кількість екземплярів мотивів, відомих, але не відкритих заново. Накладання показує кількість виявлених екземплярів, які відповідають відомим екземплярам (передбачені / відомі). Ці цифри відрізняються, оскільки кілька схожих передбачених екземплярів мотивів можуть співпадати з одним відомим мотивом у тому ж місці в послідовності білка (наприклад, див. розділ 3.7.2). Корректоване значення p -value - це QSLIMFinder Sig, який є вірогідністю спостереження N числа мотивів у випадковій послідовності,

скоригованої для числа тестів всіх можливих мотивів. Низький поріг відображає ймовірність такого високого помилкового відкриття, як ми максимально готові допустити. При оптимальному порозі, точність (precision), частка відкритих випадків, які відповідають відомим, приблизно дорівнює відкликанню (recall), частка відомих випадків, які ми відкрили заново. За жорсткого порогу, точність становить 0,5 або вище, ми виявляємо в середньому або новий мотив-кандидат, або один помилковий мотив, для кожного відомого мотиву у кожному вірусному білку. Для кожного набору даних і порогу ми показуємо, скільки мотивів було виявлено у білках, які не містять відомих мотивів.

Використовуючи підхід, який включав всі дані людини, ми відкрили заново 1/5 відомих мотивів, присутніх у білках-запитах, які ми могли б знайти. Однак ми також виявили безліч мотивів, які не збігаються з відомим мотивом: у середньому 5 нових мотивів-кандидатів або 5 помилкових мотивів на кожен відомий мотив кожного білка. Це число нових мотивів на білок є малоймовірним: вірусні білки містять у відомих випадках найбільш як 4 мотиви. Наприклад, геном поліпротеїну вірусу гепатиту С (P27958) має 4 випадки мотиву N-глікозилування (MOD_N-GLC_1) [45]. Вірусні білки містять не більше 3 відомих мотивів різних класів (ранній білок E1A людського аденовіrusу С, P03255) [45]. З цієї причини ми розглядали ще два суворіших порогових значення. Ми могли б обміняти меншу силу, щоб виявити справжні мотиви, на вищу точність. За оптимального порогу ми пропустили ще 2 відомих мотивів, але зменшили кількість потенційних помилкових мотивів; проте ми як і раніше прогнозували так багато як 4 нових мотивів-кандидатів або 4 помилкові мотиви на кожен з відомих. Нарешті, ми обрали жорсткий поріг, за яким ми відновили лише відомі 4 випадки вірусних білків, але мали нижчий потенційний нову / хибно-позитивну частоту. Ми розглянемо ці мотиви докладно в розділах 3.7.1 та 3.7.2.

Щоб проілюструвати, як порівняльний аналіз використовуючий відомі випадки, є корисним для вибору порогу, давайте розглянемо значення p-value, кориговані FDR, за найбільш суворого порогу. Для підходу, який використовує тільки віруси, суворим порогом є значення QSLIMFinder Sig $p < 0,03$, що відповідає $< 0,3$ після коригування FDR. Значення Sig p-value скориговане FDR, для підходу, який включає дані людини, також перевищує традиційне порогове значення $p < 0,05$ (0,078). Це свідчить про те, що статистична модель на основі FDR може не відображати FDR на реальних даних. Крім того, різні набори даних про взаємодію з білками повертають істинні мотиви з різними значеннями p-value, але все одно на вершині списку.

При суворому порозі обидва підходи виявляють перекриті, але не ідентичні набори мотивів. З поєднаних 10 мотивів ми відкрили заново 7 мотивів із використанням вірусного набору даних (відповідають 5 відомим) та 7 мотивів, додавши інформацію про взаємодію з людьми (відповідають 4 відомим). Використовуючи набір даних, що включав в себе мережу взаємодії білків людини, ми пропустили відомий мотив, що зв'язує ретинобластома білок (LIG_Rb_LxCxE_1), у білку E7 людського віrusу папіломи та раннього E1A-протеїну людського аденоvіrusу C [45]. Використовуючи лише вірусний набір даних, ми пропустили відомий мотив - сигналу ядерної локалізації в основному білковому полімерази 2 білка віrusу грипу А та фрагмент відомого мотиву, що зв'язує домен PDZ, в протеїні E6 людського папіломавіrusу [45]. Це говорить про те, що, хоча в деяких випадках мережа білків людини забезпечувала сигнал, в інших випадках вона додавала шум.

Нарешті, ми порівняли людську мережу, отриману одним неупередженим дослідженням Vidal та співавт. [62], до повного набору даних IntAct. За рівної суворості порогових значень ми можемо виявити трохи меншу кількість мотивів, а однакові порог точності потребують меншого значення p-value (додаток Б). Це дозволяє припустити, що дані Vidal та співавт можуть бути збіднінimi на SLIM-опосередковані взаємодії у порівнянні з усіма даними білкових взаємодій людини. Двогібридні скріни дріжджів виймають два білки

з клітинного контексту, який можуть знадобитися для зв'язування мотивів (наприклад, фосфорилювання). Крім того, група Vidal продемонструвала, що їх метод визначає взаємодії, які в середньому сильніші (більша аффінність зв'язування), ніж ті, що можуть бути визначеними іншими методами, такими як мас-спектрометрія афінної очистки білків [неопубліковані дані].

Наступним кроком ми поєднуємо мотиви, знайдені за допомогою тільки вірусних даних, або всіх людських даних, і використовуємо менш шумна мережа взаємодій білків людини, таку як BioPlex [58], яка потенційно зберігає взаємодію з опосередкованим мотивом краще, ніж двогібридні скріни Vidal. У розділах 3.6 та 3.7 ми зосереджуємося на результатах, отриманих з використанням повної мережі IntAct, розглянутої в цьому розділі.

Як показані в цьому розділі, навіть при суворому порозі ми можемо відкрити заново відомі мотиви і передбачити 206 екземплярів мотивів у вірусних білках, що не містять відомих мотивів. Далі ми хотіли б дізнатись, чи фільтрування даних взаємодій за наявністю ймовірного домену, що опосередковує взаємодію, може покращити нашу здатність викрити мотиви.

3.5 Дослідження ефекту фільтрації за ймовірним доменом розпізнання на чутливість передбачення мотивів

Багато відомих вірусних мотивів не мають достатньої підтримки в даних взаємодій білків для того, щоб бути відкрити заново. Ці мотиви не є відкритими заново навіть при низькому порозі (рис. 3.4.2). Для багатьох інших мотивів у даних взаємодії недостатньо інформації, щоб вказати на ймовірний розпізніваючий домен. Серед тих, у кого достатньо - ми можемо виявити більшу частку відомих мотивів (рис. 3.5). Ми відновили більшу частку відомих мотивів порівняно з підходом без фільтрування для доменів. Це показує, що підхід працює, і навіть за суворим порогом ми можемо відкрити заново справжні мотиви. У наступних розділах 3.6 та 3.7 ми обговоримо відкриті заново мотиви та ряд мотивів кандидатів, які ми виявили, використовуючи

дані про взаємодію білків людини. По-перше, ми будемо розглядати схожість мотивів, потім детально розглядати конкретні випадки.

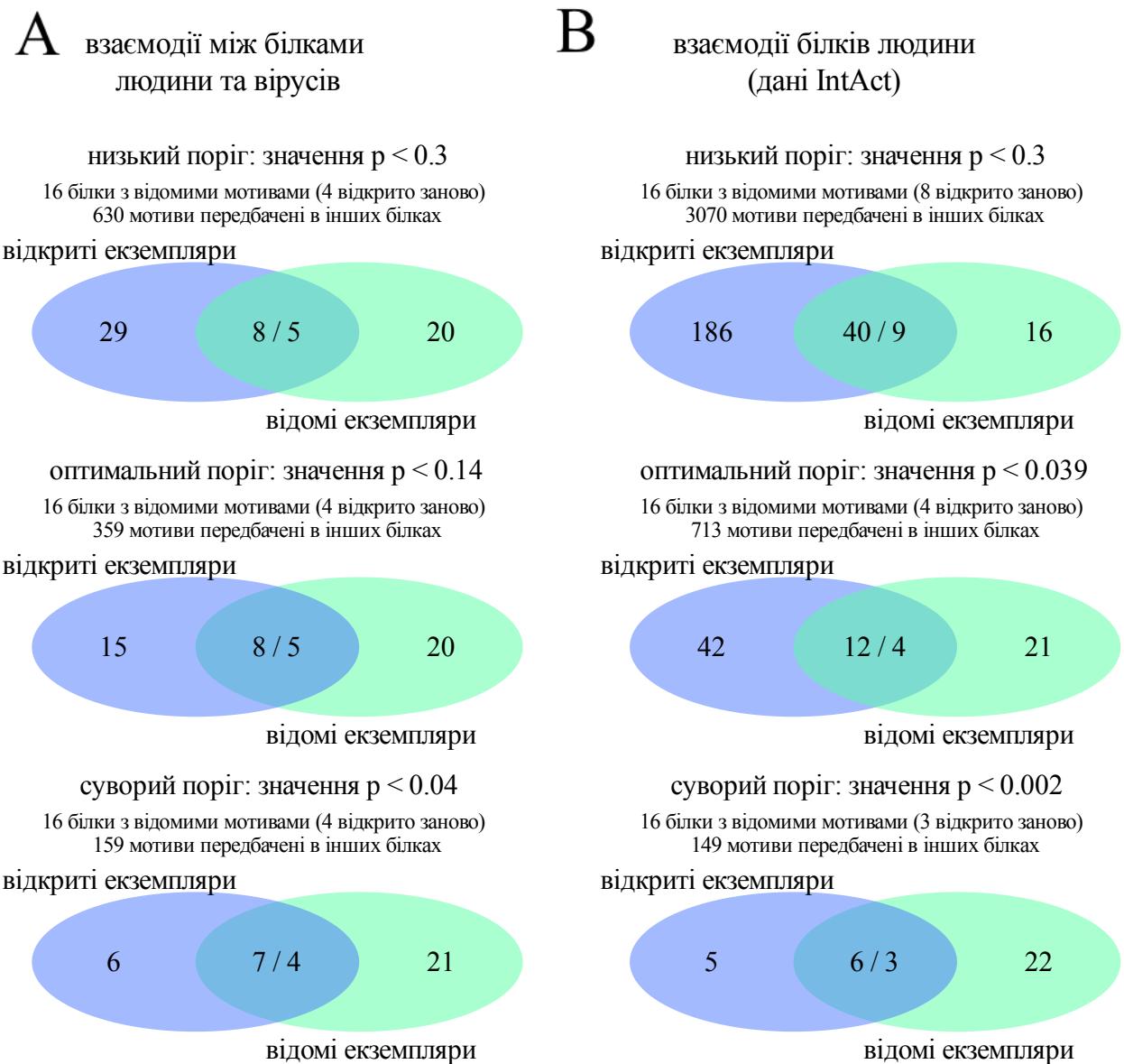


Рис 3.5. Діаграми Вена, що показують кількість знайдених мотивів-кандидатів та відкритих заново відомих мотивів при фільтруванні по домену. Наведено дві стратегії побудови наборів даних (A та B) та пороги 3 значень p -value. Синє коло показує кількість екземплярів мотивів, передбачених, але невідомих. Зелене коло показує кількість екземплярів мотивів, відомих, але не відкритих заново. Накладання показує кількість виявлених екземплярів, які відповідають відомим (передбачених / відомих). Ці цифри відрізняються, оскільки кілька схожих передбачених екземплярів мотивів можуть співпадати з одним відомим мотивом у тому ж місці в послідовності білка (наприклад,

див. розділ 3.7.2). Корректоване значення p-value - це QSLIMFinder Sig, який є вірогідністю спостереження N числа мотивів у випадковій послідовності, скоригованої для числа тестів всіх можливих мотивів. Низький поріг відображає ймовірність такого високого помилкового відкриття, як ми максимально готові допустити. При оптимальному порозі, точність (precision), частка відкритих випадків, які відповідають відомим, приблизно дорівнює відкликанню (recall), частка відомих випадків, які ми відкрили заново. За жорсткого порогу, точність становить 0,5 або вище, ми виявляємо в середньому або новий мотив-кандидат, або один помилковий мотив, для кожного відомого мотиву у кожному вірусному білку. Для кожного набору даних і порогу ми показуємо, скільки мотивів було виявлено у білках, які не містять відомих мотивів.

3.6 Дослідження схожості мотивів знайдених *de novo* до відомих мотивів

Щоб з'ясувати, які короткі лінійні мотиви-кандидати ми виявили, ми дослідили, які відомі мотиви є схожими до виявлених вибраних за суворим порогом. Всі ці мотиви нагадують якийсь відомий мотив у базі даних ELM (відповідні послідовності зі значенням інформаційного змісту 0.5). Ми підбиваємо підсумки результатів подібності, які були фільтровані вище балів 1.162, щоб уникнути надмірної кількості відповідностей на рисунку 3.6.1. Ми бачимо чітке скупчення сигналів ядерної локалізації (таргетингу) мотивів, що мають KR-паттерн. Ми також бачимо, що мотиви, багаті на пролін (P..P.[HKR] і P..P.P.D), визнані як ліганди домену SH3.

Порівняння подібності мотивів є дуже схильним до над-передбачення і в той же час не може спіймати мотиви низького інформаційного змісту (дуже мало визначених позицій). Наприклад, класичний С-термінальний мотив для доменного зв'язування PDZ не був визначений. Для визначення, чи співпадають з інші збіжності паттернів мотивів кандидатів з правильним класом мотивів, потрібне подальше дослідження. Більш надійний спосіб

ідентифікації класів мотивів полягає у поєднанні подібності паттерну мотивів та чи мають ці мотиви правильний відповідний домен розпізнавання.

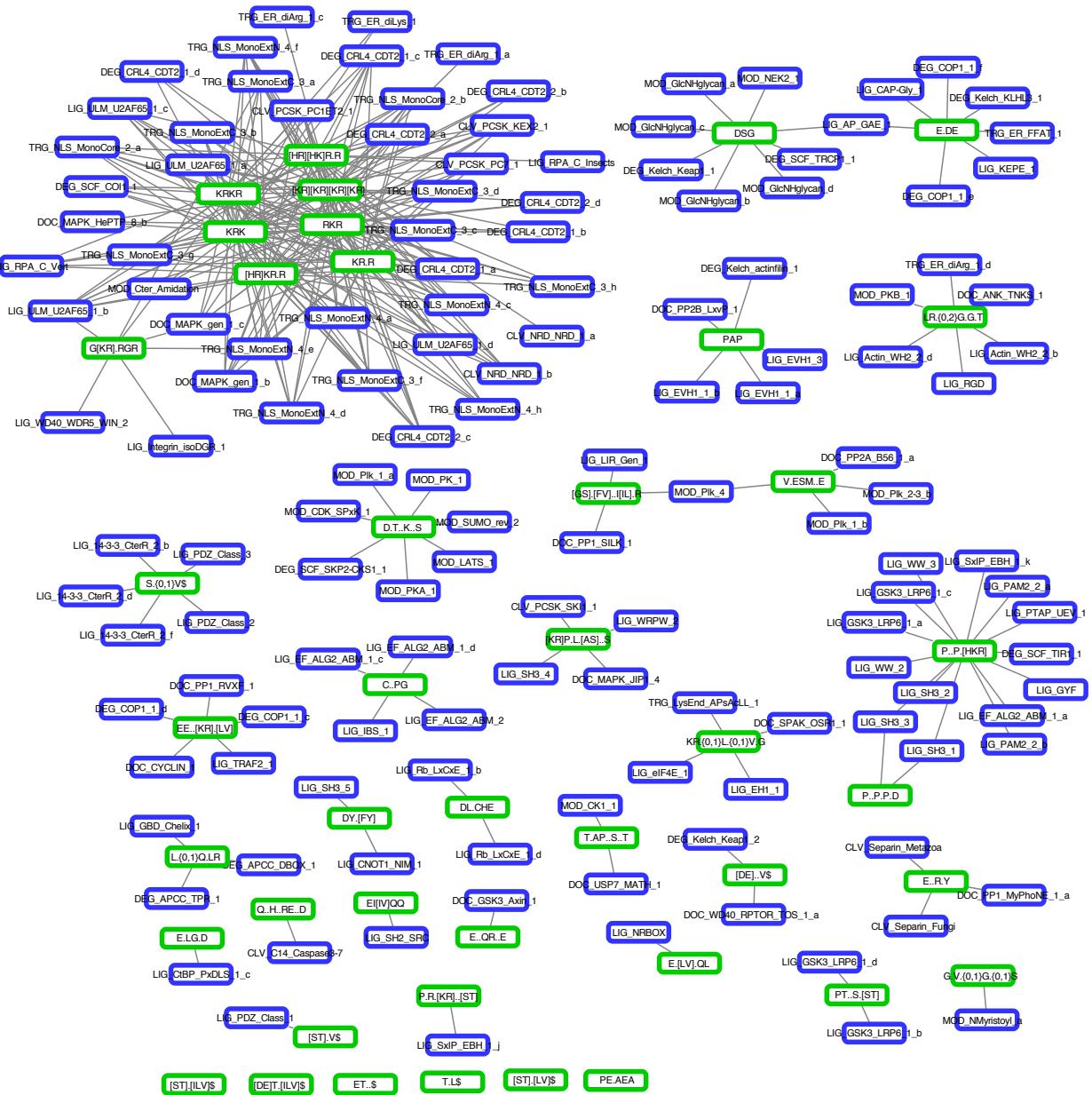


Рис 3.6.1. Схема, що показує подібність виявлених коротких лінійних мотивів до відомих мотивів. Зелені вузли - це послідовності виявлених мотивів, сині вузли - це відомі мотиви в базі даних ELM, ребра показують наявність подібності між мотивами вище порога 1,162 (Score, Comparomtf3).

3.7 Приклади відкритих заново та мотивів-кандидатів

3.7.1 Дослідження класів мотивів, що відкриті заново, мотивів-кандидатів та їх імовірних доменів розпізнавання

Для вивчення хітів, отриманих за допомогою комбінації людсько-вірусної та повної мережі людини (IntAct), ми вибрали мотиви-кандидати під найсуворішим порогом (точність $> 0,5$ або 1 мотив-кандидат на кожен відомий мотив). Цей набір даних є легшим для інтерпретації, оскільки на додаток до мотиву, який конвергентно еволюціонували у вірусному білку, він прогнозує мотиви в протеїнах людини, які зв'язують один і теж ж домен розпізнавання. Ми розглянули передбачені мотиви та їх найбільш ймовірні домени розпізнавання.

Заохочувально, найпоширеніша група мотивів кандидатів не була таргетинг-мотивами, але класичні ліганд-зв'язуючими мотивами ([ST].[LV]\$) С-кінцевими ми, що розпізнаються доменом PDZ. 26 варіантів цих мотивів ([ST].[LV]\$, [DE]T.[ILV]\$, ET..\$, [ST].V\$, [DE]..V\$, T.L\$, [ST].[ILV]\$) були передбачені на С-кінці 16 вірусних білків, які зв'язуються з 7 білками, що містять домен PDZ. 2 екземпляри цих мотивів вже були відомі і будуть розглянуті в наступному розділі. Для всіх цих випадків, крім 3 мотивів в 2х білках), домен PDZ був правильно ідентифікований як найбільш вірогідний або один з найбільш вірогідних доменів, що опосередковують взаємодію.

Як і очікувалось, однією з найпоширеніших груп мотивів кандидатів було 12 сигналів ядерної локалізації (таргетинг, TRG, багатий на KR амінокислоти) мотиви, присутні в 11 вірусних білках, які зв'язуються з 4 білками людини (апарату ядерного імпорту), кожен з яких містить Armadillo-подібний домен - правильно визначений процедурою збагачення домену. Багато інших вірусних білків, які використовуються в нашому дослідженні, локалізуються в ядрі, але не були виявлені взаємодіючими з апаратом ядерного імпорту, що свідчить про те, що зафіксовані мотиви можуть бути посередниками для більш стабільної взаємодії. Альтернативне пояснення полягає в тому, що вірусні

протеїни, що містять мотиви, є присутніми у достатній кількості для зв'язування апарату ядерного імпорту для виявлення цих взаємодій, гіпотеза підтверджується тим, що 5 з цих мотивів знаходяться в капсидних білках. Доменне збагачення також підхопило Armadillo-подібний домен, як найбільш імовірний для кількох мотивів, які не нагадують сигнал ядерної локалізації (E..QR..E, DT.K..S, PT..S.[ST] , V.ESM..E). Ми виявили два з цих мотивів, що взаємодіють з білками людини, які не належать до апарату ядерного імпорту (не-АТФазна регуляторна субодиниця 1 26S протеасоми - E..QR..E мотив; Е3 убіквітин-лігаза HUWE1 - PT..S.[ST] мотив). Подальше вивчення цих мотивів-кандидатів необхідне для того, щоб визначити, чи Armadillo-подібний домен цих білків може звязувати неканонічні ліганди.

Передбачено екземпляри кількох інших класів ліганд-зв'язуючих мотивів: 4 WD40 мотиви-кандидати, 1 SH3мотив, 1 EF-hand мотив, 1 РН-домен-подібний мотив. Деякі з цих мотивів докладніше розглянуті в наступних розділах. Деякі з цих мотивів-кандидатів мають домени, які не є відомими доменами, що зв'язують SLiM, але позначені як найімовірніші, включаючи 1 цикліновоподібний домен, 1 кератинову головку типу 2, 2 Gro-EL-подібні та 7-ти ВАG-домен. Проте й мотив, й домени можуть бути виявлені помилково, тому подальше дослідження є необхідним перед експериментальною перевіркою.

Далі ми будемо обговорювати, як індивідуальні мотиви підтримуються нашим аналізом та незалежною літературою.

3.7.2 Мотиви PDZ, які відкрили заново

При строгому порозі достовірності (точності $> 0,5$) ми відкрили заново 2 відомі мотиви, що зв'язують домен PDZ, білка Е6 вірусу папіломи людини (ВПЛ) типу 16 і 18 (рис 3.7.2 А і В, відповідно).

У цьому прикладі той самий білок у двох пов'язаних віrusах обрав як мішень перекриваючий набір білків людини. Білок-гомолог скрібл (SCRIB) і

тиrozин-протеїн фосфатазний receptor типу 3 (PTPN3) є мішенями обох вірусів.

Анотація цих мотивів у базі даних ELM базується на структурних доказах: взаємодія мотиву E6 з доменом PDZом людського білка MAGI1. Хоча ми шукали мотиви в білках, які зв'язують MAGI1 як до, так і після фільтрації за доменом, ми не могли відкрити заново їх навіть за низького порогу. Тим не менш цей екземпляр ELM мотиву зв'язуючого домен PDZ є відкритими заново з використанням 3 інших білків, що мають домен PDZ. Давайте розглянемо, як HPV може порушити їх.

Білок E6 HPV зв'язує цілий ряд білків, що містять домен PDZ. PDZ-зв'язуючий С-кінцевий регіон змінюється в різних типах цих вірусів та диктує переважне звязування [90]. HPV 16 і HPV 18 E6 націлені на найбільшу кількість білків людини. Деякі людські білки, такі як DLG1, зв'язуються всіма типами білків HPV E6. Мішені білків E6, у тому числі DLG1, PTPN3 та SCRIB, зазвичай убіквітинуються і деградуються протеасомою [91–93]. E6 білки виконують це, діючи як білки-скафолди, щоб залучають E3 убіквітин-лігазу E6-AP (UBE3A) через мотив LXXLL, розташований в лігазі [94, 92]. Основна функція білка E6, що має значення для хвороби, полягає в активації теломерази, щоб імморталізувати інфіковані клітини. Ця функція також спирається на роль білка E6 у приєданні убіквітину, щоб знищити репресор транскрипції теломерази NFX1-91 [95]. Інші дослідження також показують, що E6 має більш складні зв'язки з його мішенями: білок SCRIB позитивно регулює транскрипцію та швидкість трансляції білка E6 [96]. На цьому етапі також не цілком зрозуміло, наскільки звязування білків, що беруть участь у встановленні апікально-базальної клітинної полярності та внутрішньоклітинних контактів, є корисним для цього вірусу. У будь-якому випадку, взаємодія HPV E6 з PDZ-білками людини є вирішальними для інфікування та пухлинного генезу.

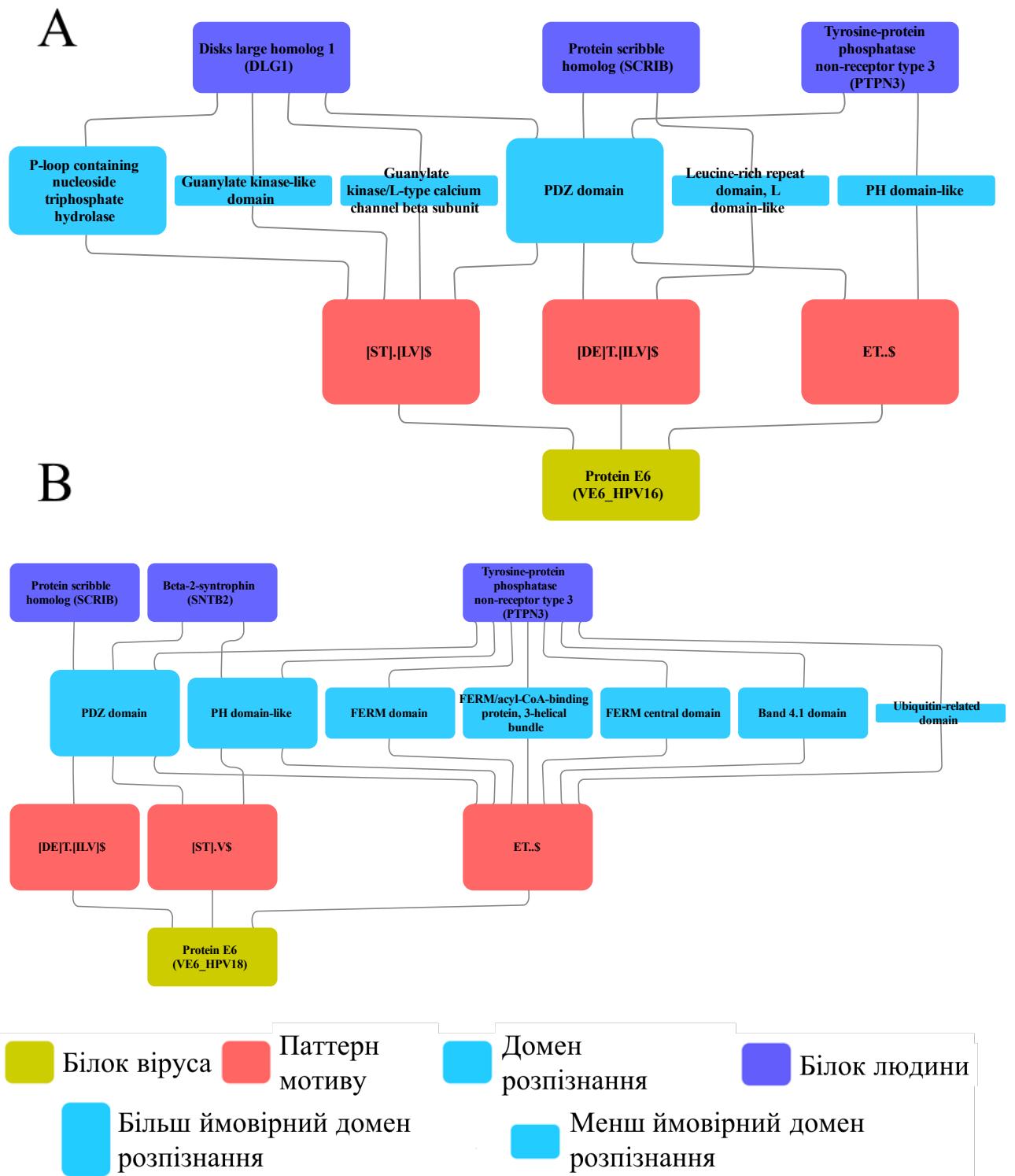


Рис 3.7.2. Схема мережі, що показує відомі мотиви PDZ в білку E6 людського папіломавірусу 16 та 18. Ми показуємо варіанти мотиву та ті домени в білках людини, які можуть бути відповідальними за зв'язування протеїну E6. Три варіанти цього мотиву були передбачені білком E6 та

білками людини, які взаємодіють з білками-мішенями вірусних білків, які називаються DLG1, SNTB2, SCRIB та PTPN3. домен PDZ є найбільш збагаченим серед мішеней білка Е6 і є доменом, що опосередковує взаємодію з цим відомим мотивом. А. Мережа цього мотиву в людському папіломавірусі 16. Б. Мережа цього мотиву в людському папіломавірусі 18.

3.7.3 Мотиви-кандидати, що зв'язують домен PDZ

Як зазначено в розділі 3.7.1, мотиви домену PDZ є найбільш поширеними в нашому наборі знайдених мотивів. Ми відкрили заново 2 екземпляри аnotatedовані в ELM і 14 інших екземплярів. Тут ми розглянемо мотиви-кандидати, які мають найбільшу підтримку.

Перший мотив, як і всі відомі мотиви, міститься у білку Е6, але в іншому типі HPV: HPV-70 (рис 3.7.3.1). Хоча й не аnotatedований в ELM, цей мотив також відомий. Відповідно до дослідження Thomas та інших С-кінцевий пептид HPV-70 Е6 зв'язує меншу кількість білків, ніж HPV-16 або HPV-18, про які йшлося раніше [90]. Згідно з їх результатами, мотив HPV-70 Е6 дійсно зв'язується з DLG1, однак він не зв'язує SCRIB, і жоден із досліджуваних мотивів не зв'язує ERBIN (рис. 3.7.3.1.) [90]. Це може вказувати, що взаємодія білка Е8 HPV-70 з SCRIB є непрямим чи помилковим. Ми все ще можемо ідентифікувати мотив домену PDZ у цьому білку, використовуючи взаємодії SCRIB, оскільки SCRIB дійсно зв'язує мотив домену PDZ; однак, область PDZ SCRIB може бути достатньо селективною, щоб уникнути зв'язування HPV-70. Відсутність пептидної взаємодії з ERBIN неможлива, оскільки жоден з пептидів, протестованих у тому дослідженні, не зв'язує ERBIN, що може свідчити про те, що ERBIN не експресується в клітинній лінії (HaCat), де дослідники виконували пептидний pull-down. В цілому, це дослідження, проведене Томасом та ін, служить підтвердженням цього екземпляру мотиву PDZ в білку HPV-70 Е6, але воно вказує на потенційну невідповідність в даних про білкову взаємодію, які використовуються для нашого дослідження.

З точки зору функції, ERBIN служить як адаптерний протеїн, який зв'язує нефосфорильований receptor ERBB2, тим самим стабілізуючи цей стан [97]. Він є важливим для локалізації ERBB2 на базолатеральну сторону епітеліальних клітин [98]. З огляду на те, що HPV також зв'язує й інші білки, пов'язані з апікально-базальною полярністю клітини, такими як DLG1 та SCRIB, що було обговорено раніше, ERBIN може представляти реальну мішень.

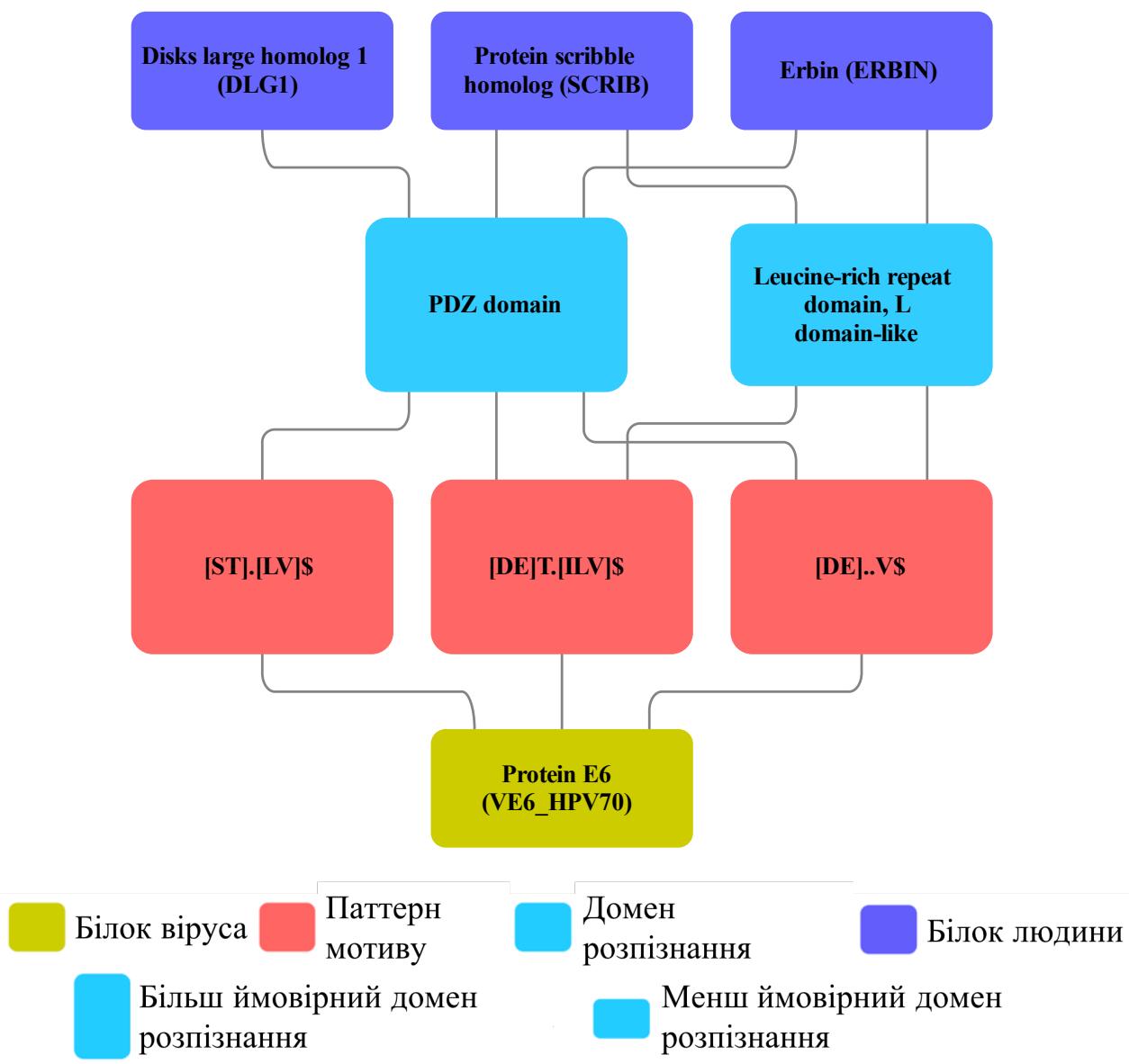


Рис 3.7.3.1. Схема мережі, що показує мотиви-кандидати PDZ в білках E6 людського папіломавірусу 70 були підтвердженні в попередньому дослідженні, але не анотовані в ELM. Ми показуємо варіанти мотивів і ті домени в білках людини, які можуть бути відповідальними за зв'язування протеїну E6. Три

варіанти цього мотиву були передбачені у білку E6 і 74 білках людини, які взаємодіють з білками-мішенями вірусів DLG1, SCRIB та ERBIN. домен PDZ є найбільш збагаченим серед мішеней білка E6.

Другий мотив-кандидат, що зв'язує домен PDZ, який ми передбачаємо розташований в неструктурних білків вірусу грипу А H5N1 (рис 3.7.3.2). Цей мотив виявлений з використанням наборів даних з 4 білків людини: SCRIB і ERBIN, DLG4 і GIPC1. Хоча він не анотований в ELM, цей мотив також відомий. Крім того, було продемонстровано, що цей мотив дозволяє H5N1 порушувати шільні контакти через його взаємодію з SCRIB і DLG1 [99]. Паттерн мотиву, що ми ідентифікуємо, загально схожий на паттерн високо патогенного пташиного (RS.V), але не людського (ES.V) вірусів грипу А [100]. Вірус H5N1 викрадає проапоптичну функцію SCRIB, використовуючи домен PDZ мотив, для зміни його субклітинної локалізації [101]. Це запобігає апоптотичній смерті інфікованих клітин. Взаємодія NS1 з ERBIN, DLG4 та GIPC1 не є детально описаною на сьогоднішній день.

Підбиваючи підсумки, в попередніх дослідженнях було описано два мотиви PDZ, які не були анотовані в ELM, які мали найбільшу підтримку в нашому аналізі. Це служить підтвердженням того, що наша процедура для відкриття мотивів *de novo*, працює. Далі, давайте розглянемо ряд менш поширених мотивів.

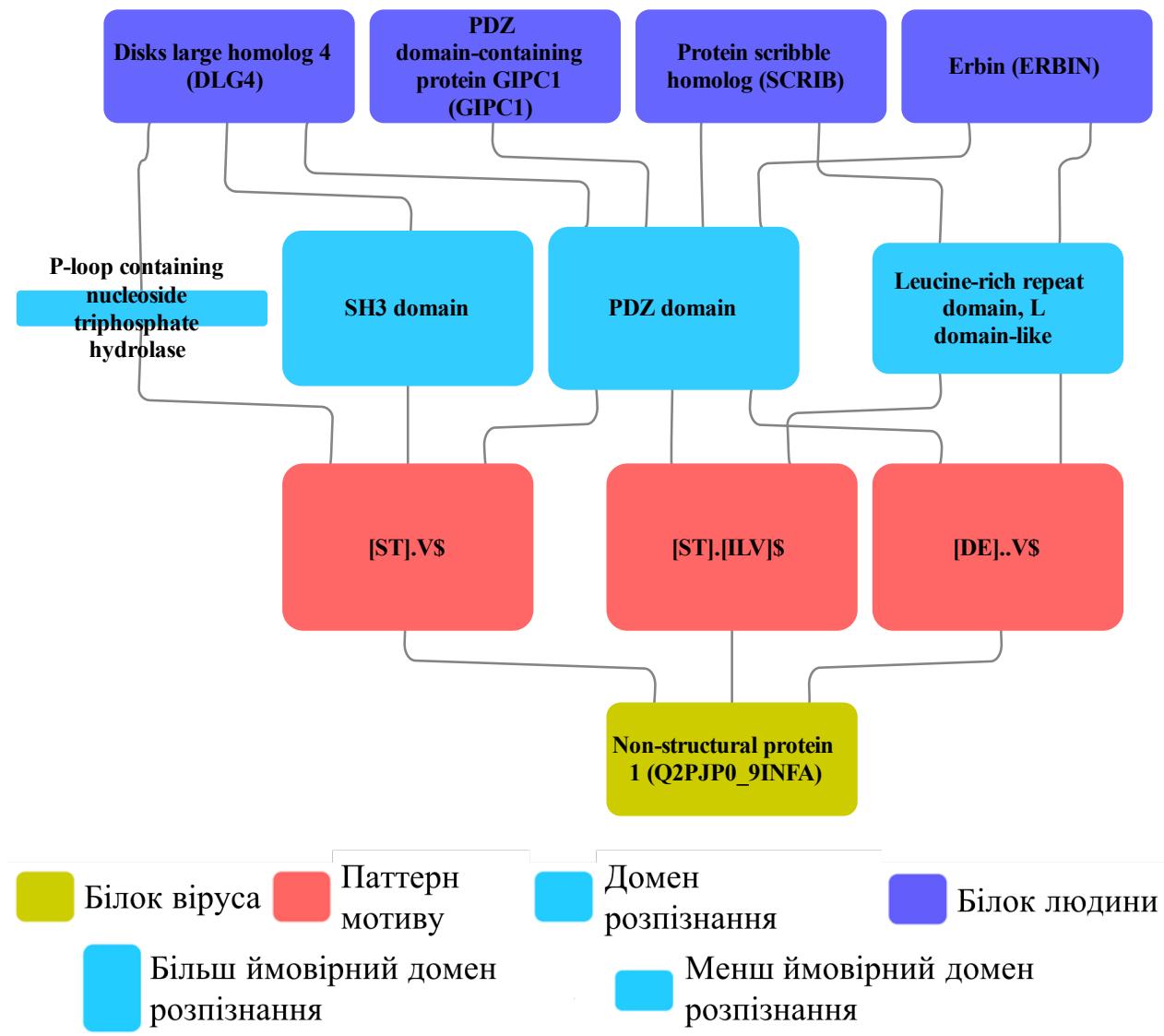


Рис 3.7.3.2. Схема мережі, що показує мотиви-кандидати PDZ в неструктурному білку 1 (NS1) вірусу H5N1 грипу були підтвердженні в попередньому дослідженні, але не були анотовані в ELM. Ми показуємо варіанти мотивів і ті домени в білках людини, які можуть бути відповідальними за зв'язування NS1. Три варіанти цього мотиву були передбачені у білку NS1 і у 86 білками людини, які взаємодіють з білками DLG4, GIPC1, SCRIB та ERBIN, що є мішенями вірусів. домен PDZ є найбільш збагаченим серед мішеней білка Е6. Високе збагачення домену SH3 і домену лейцін-багатого повтору може відображати функціональну перевагу NS1.

3.7.4 Мотив-кандидат, що зв'язує домен SH3

Ми виявили екземпляр домен SH3-зв'язуючого мотиву *de novo* в білку Nef віруса імунодефіциту людини типу 1 (рис. 3.7.4). Хоча ми не змогли ідентифікувати єдиного домену взаємодії, ми бачили, що послідовність, що містить Р..Р, нагадує канонічний ліганд домену SH3 [102]. Відомо, що Nef взаємодіє лише з 5 білками людини, з яких 4 поділяють доменну архітектуру SRC-кінази. Цей мотив дійсно є ще одним відомим прикладом, не зазначеним в базі даних ELM. Як підтверджено дослідженнями мутагенезу, мотив Р..Р.[HKR] дозволяє Nef зв'язати домен SH3 з сімейства SRC-кіназ для їх активації та сприяння вірусній патогенності [103, 104].

Тому мотив, що зв'язує домен SH3, в білку Nef є ще одним підтвердженим мотивом, який не був включений в наші навчальні дані.

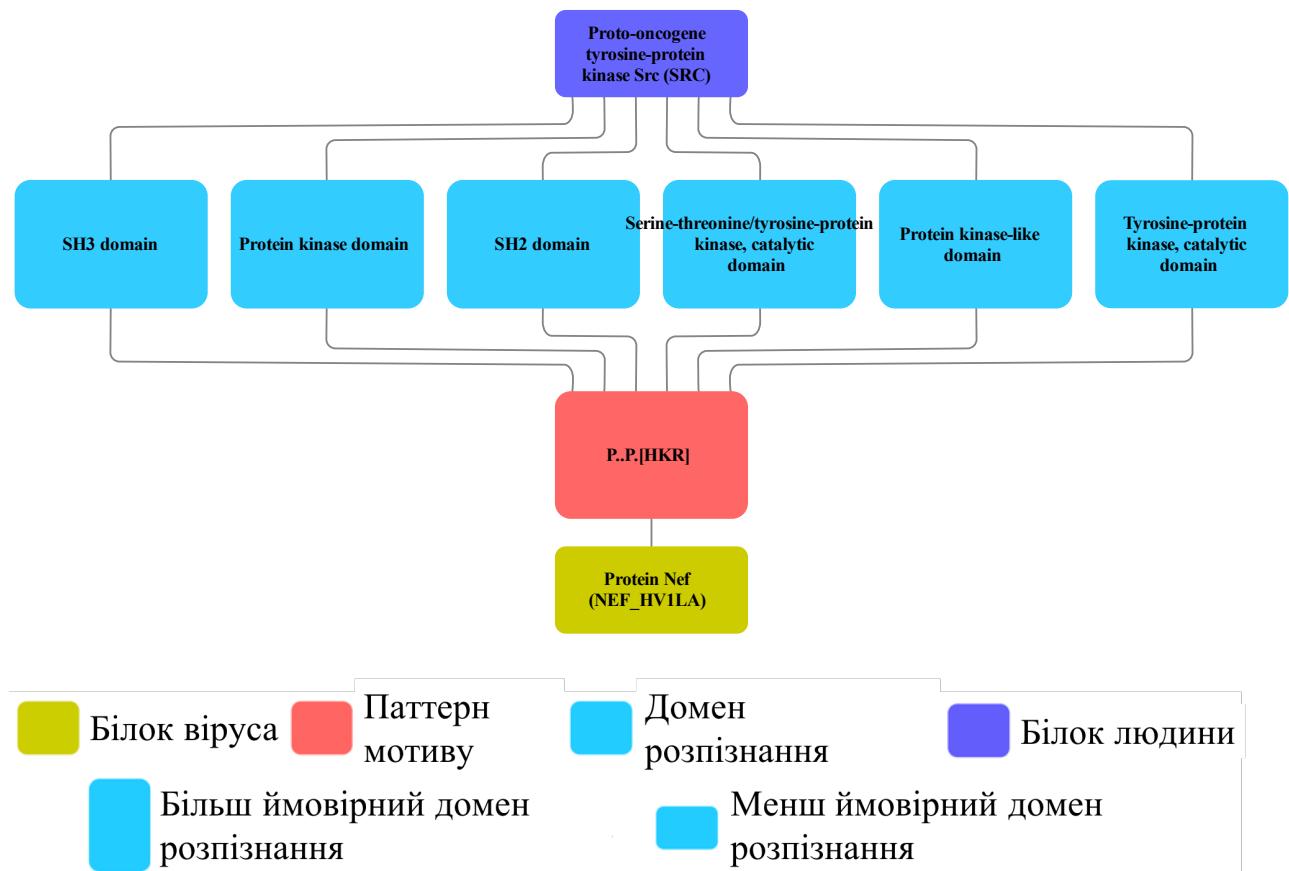


Рис 3.7.4. Схема мережі, що показує мотив-кандидат SH3 в білку Nef вірусу Імунодефіциту людини типу 1 був підтверджений у попередньому дослідженні, але не був анотований в ELM. Ми показуємо ті домени в білках людини, які можуть бути відповідальними за зв'язування Nef. Один із варіантів

цього мотиву був передбачений у білку Nef та у 93 білках людини, які взаємодіють з SRC. Усі домени, крім каталітичного домену тирозин-протеїнкінази, є однаково збагаченими, що може відображати функціональні переваги Nef або упередження в доступних даних, оскільки відомо, що Nef зв'язує лише 5 білків людини.

3.7.5 Мотиви-кандидати, що зв'язують домен WD40

Ми передбачили DSG мотив в якості мотиву, що зв'язує WD40 (рис 3.7.5.1), розташованого в чотирьох вірусних білків трьох вірусних видів: Vpu білок вірусу імунодефіциту людини (VPU_HV1H2 і VPU_HV1S1), Великий Т-антиген вірусу SV40 (LT_SV40) і неструктурний білок Rotavirus A (NSP1_ROT5). Цей мотив розпізнається двома білками з F-box / WD-повтором: FBXW11, також відомий як β-TRCP1, і BTRC, також відомий як β-TRCP2. Обидва вони служать субодиницею розпізнавання субстрату комплексу E3 убіквітин-білок-лігази SCF (білок SKP1-CUL1-F-box). Цей комплекс убіквітинує і містить білки для протеасомної деградації [105, 106]. SCF (комплекс FBXW11 або BTRC) є частиною сигнальних шляхів, включаючи шлях Wnt-beta-catenin і NF-kappaB, де він містить або бета-катенін (fosфорилюваний за допомогою GSK3beta), або IкаппаB для деградації. У свою чергу, це пригнічує (бета-катенін) або активує (NF-kappaB) транскрипційний фактор в кінці шляху [107, 108].

Vpu має відомий екземпляр мотиву DSG..S, який дозволяє ВІЛ викрасти SCF убіквітин лігазу для деградації білків хозяїна, таких як противірусний фактор тетерін/BST-2 і CD4 [109, 110]. Цей мотив (коли фосфорилюється по обох серинах) зазвичай розпізнається FBXW11 та BTRC, що направляє білок, що містить мотив, на деградацію. Очевидно, Vpu знайшов спосіб уникнути самодеградації [110]. NSP1 ротавірусу А також має відомий мотив DSG..S [111]. Цей білок використовує убіквітин лігази хозяїна для деградації ключових факторів, що активують вироблення інтерферону, таких як IRF3, IRF5 або IRF7 [112, 113]. Інтерферон звичайно виробляється у відповідь на

вірусну інфекцію та допомагає обмежити інфекцію до сусідніх клітин [114]. Ці випадки служать підтвердженням нашої процедури відкриття мотивів: справжній мотив, не представлений в базі даних ELM - даних, які ми використовували для вибору оптимальних параметрів і порога..

Великий Т антиген (TAg) вірусу SV40, не має відомого мотиву DSG. Цей протеїн взаємодіє з супресором пухлин та детектором пошкодження ДНК P53 (так було відкрито P53) [115]. Регулятор P53 убіквітин лігаза MDM2 містить відомий мотив DSG і сам деградується β -TRCP1/2, що обговорювалося раніше [116]. Незважаючи на те, що TAg вірусу SV40 не має перевіреного мотиву DSG..N (зверніть увагу на заміщення останнього серину на аспарагін), його гомолог TAg білок вірусу JC містить мотив DSG..S [117].

Щоб краще зрозуміти структурний аспект цієї взаємодії, ми виконали докінг трьох пептидів з β -TRCP1 / FBXW11, використовуючи PepSite 2 (PDB 1P22, ланцюг А). Цей аналіз показує, що короткий мотив DSG, який ми передбачуємо, може мати сайт для зв'язування в FBXW11, однак, не з дуже високою статистичною значимістю. Крім того, передбачається, що мотив DSG зв'язує F-box, а не WD-40 домен, який ми прогнозуємо за допомогою процедури збагачення доменів. Дивно, що докінг повної послідовності відомого мотиву з Vpu (DSGNES) або мотиву TAg з SV40 (DSGHET) також не має сайту зв'язування передбаченого PepSite 2 з високою значимістю.

Обидва DSG мотиви мають дуже сильну підтримку 24/36 (FBXW11) або 29/56 (BTRC) білків з мотивом серед негомологічних білків (UPC, див. секцію 2.4), які взаємодіють з FBXW11 або BTRC.

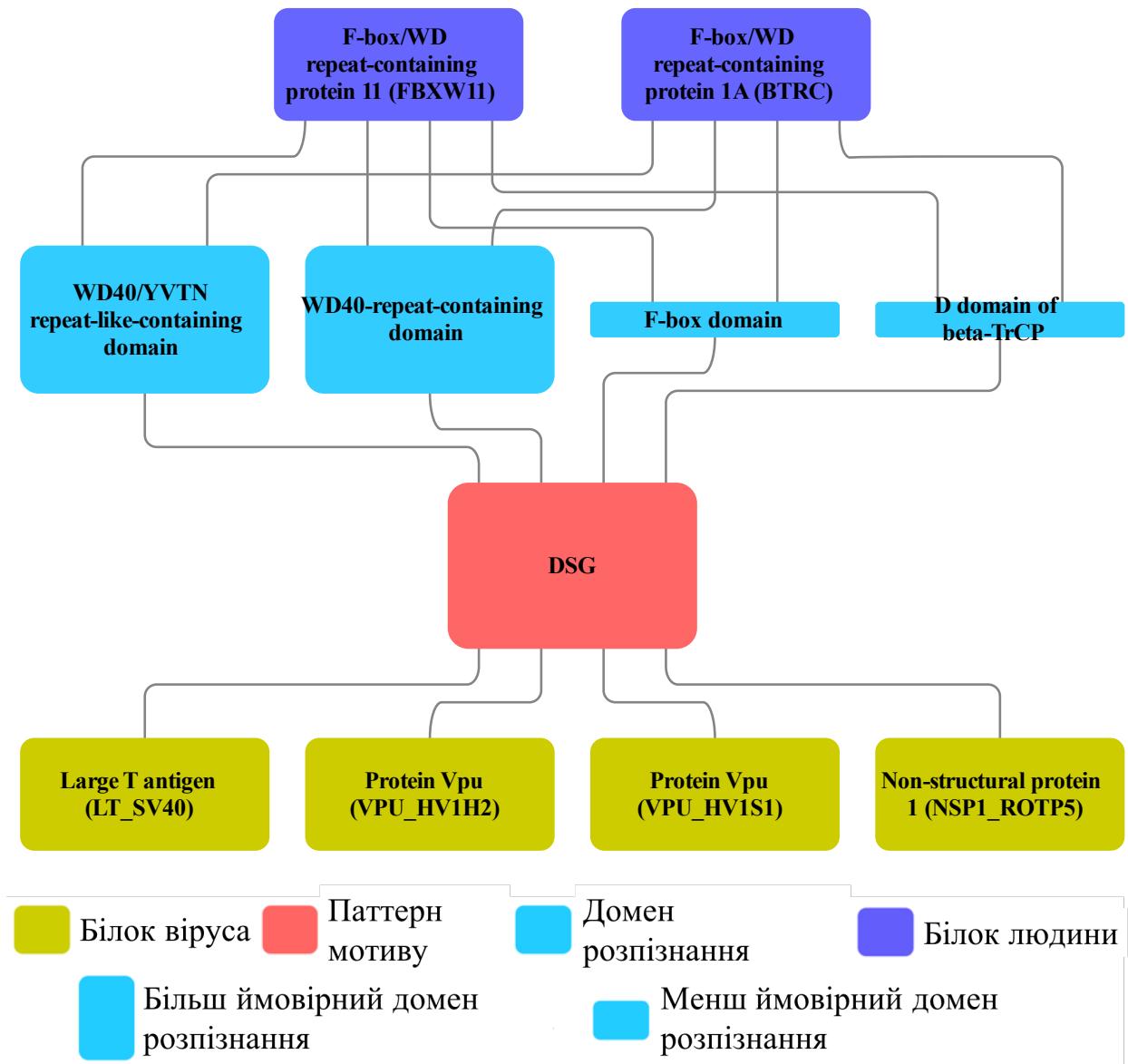


Рис 3.7.5.1. Схема мережі, що показує мотиви-кандидати DSG. Ці мотиви були передбачені в 4 вірусних білках. Всі вони обрали як мішень 2 субодиниці розпізнавання субстрату у комплексі SCF Е3 убіквітин лігази людини, FBXW11 та BTRC. 3 екземпляри мотиву були підтвердженні в попередньому дослідженні, але не були анотовані в ELM. Виняток становить LT у SV40. Домен WD40 найбільш збагачений серед мішеней вірусних білоків VPU_HV1H2 та LT_SV40.

Другий мотив-кандидат, що зв'язує WD40 (рис 3.7.5.2), був передбачений в полімеразному лужному білку 2 (PB2) РНК-полімерази у 2 штамів вірусу грипу А (білок B4URF7 у штамі A/WS/1933 H1N1, білок C5E527 в A/New

York/1682/2009 H1N1). Ми також передбачаємо цей мотив у 4 людських білках, які всі зв'язують білок 1 елонгаторного комплексу людини (ELP1). ELP1 бере участь у елонгації транскрипції РНК-полімеразою 2 у складі комплексу, який відіграє роль в ремоделюваннях хроматину та ацетилює гістон H3 [22854966]. WD40 передбачено як найбільш імовірний домен, що підтримується 9/210 білками або 5/140 білками, що містять цей домен (для кожного штаму віруса відповідно). З огляду на РНК-полімеразну функцію PB2, ми можемо припустити, що він також викрадає фактори елонгації хазяїна використовуючи мотив E.V..G.{0,2}N.{0,1}Q для полегшення цього процесу.



Рис 3.7.5.2. Схема мережі, що показує мотив-кандидат E.V..G.{0,2}N.{0,1}Q передбачений в полімеразному лужному білку 2 у 2-х штамів грипу А. Ми передбачаємо, що цей мотив розпізнається доменом WD40 в людському протеїні ELP1.

3.7.6 Мотиви-кандидати, які розпізнаються доменом, що зв'язує дволанцюкову РНК, та доменом EF-hand

Ми передбачаємо мотив LR. $\{0,2\}$ G.G.T, який може бути розпізнаний дволанцюжковим РНК-зв'язуючим доменом у Q96SI9 - людському сперматидному перинуклеарному білку, який розпізнає вірусну РНК. Ми прогнозуємо цей мотив у 6 неструктурних вірусних білках з 4 штамів грипу А та 4 білках людини (рис 3.7.6.1). Ці вірусні білки беруть участь у блокуванні трансляції мРНК хазяїна [118], а також інгібують TRIM25-опосередковане убіквітинування, що є частиною антивірусної відповіді [119]. Ми припускаємо, що цей мотив імітує РНК, яку цей домен людини розпізнає.

Мотив, який зв'язується з доменом EF-hand, показаний на рисунку 3.7.6.2. Cab45 є EF-hand доменним і Ca⁽²⁺⁾ зв'язувальним білком, необхідним для сортування секреторних білків у мережі trans-Golgi. Олігомери Cab45 зв'язують секреторні та плазматичні мембрани білки [120] і відправляють їх на плазматичну мембрану / позаклітинний простір. Цей білок, як відомо, не взаємодіє з вірусами, окрім недавніх високопродуктивних робіт, що проаналізували інтерактоми кількох штамів вірусу грипу А [121]. Cab45 є мішеню 12 різних вірусних білків з 6 вірусних таксонів, хоча, ці взаємодії були профілізовані методами очищення афінності, які вимірюють як безпосередні, так і непрямі взаємодії. Вірусний білок (PB1, Q5EP37), який зв'язує Cab45 і містить мотив EI[IV]QQ, є однією з РНК-залежних РНК-полімераз вірусу грипу та є важливим компонентом механізму транскрипції вірусу [122]. Білки РНК-полімерази (включаючи PB1) залишаються пов'язаними з вірусною РНК та упаковуються у вірусні частинки.

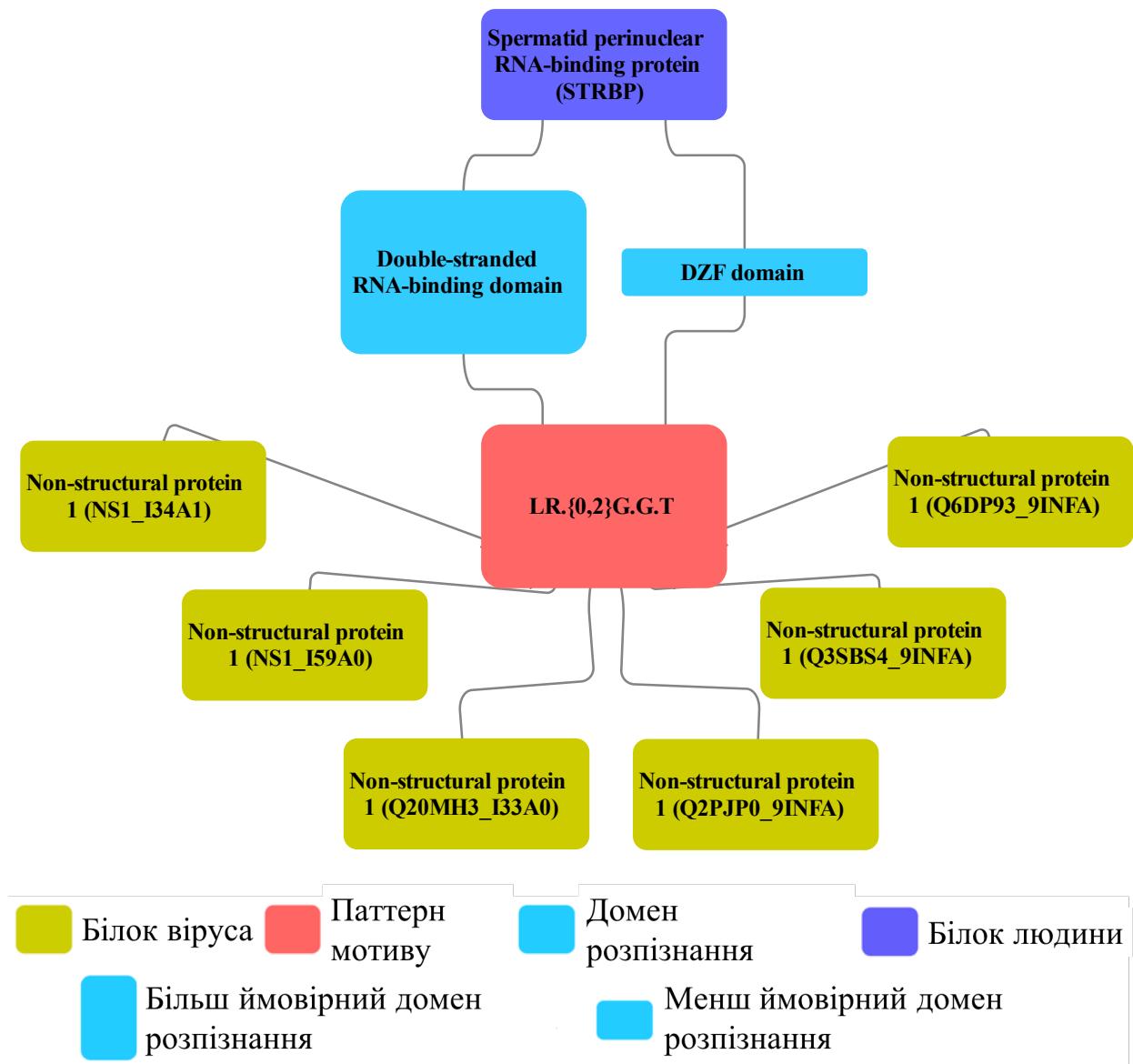


Рис 3.7.6.1. Схема мережі, що показує мотив-кандидат $LR.\{0,2\}G.G.T$ передбачений в 6 неструктурних білках різних штамів грипу А, що включають як пташину, так і людську лінію. Ми передбачаємо, що цей мотив розпізається дволанцюговим РНК-зв'язуючим доменом людського білка STRBP.

Ми можемо припустити, що Cab45 служить для полегшення цього процесу. Незважаючи на те, що це можливо, довіра до цього мотиву зменшується, оскільки на відміну від мотиву DSG, описаного раніше, цей мотив-кандидат, що зв'язує EF-hand, був виявлений лише в 4 з 33 білкових послідовностей (партнери Cab45), і релевантність домену підтримується тільки 2 з 45 РВ1-зв'язуючих білків (рис. 13). Дослідження 3 білків людини, в

яких передбачено цей мотив, може прояснити, наскільки цей мотив може бути справжнім хітом.

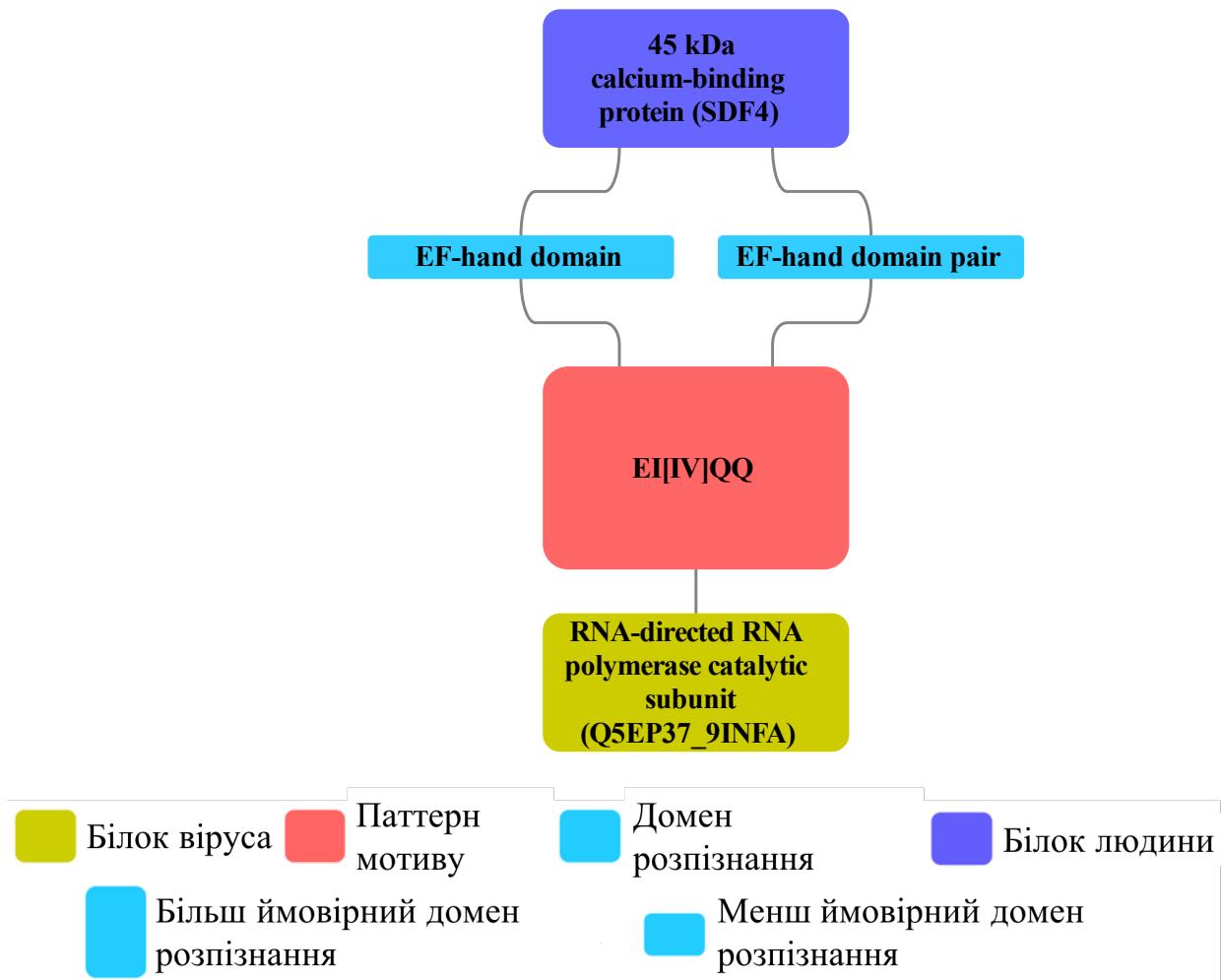


Рис 3.7.6.2. Схема мережі, що показує мотив-кандидат EI[IV]QQ розташований в РНК-залежній РНК -полімеразі грипу А і потенційно розпізнаного доменом EF-hand білка людини SDF4.

3.7.7 Мотив-кандидат, що зв'язує BAG-домен

Ми знайшли мотив-кандидат $(L.\{0,1\}Q.LR)$, який потенційно розпізнається доменом BAG у семи повтореннях у Епштейн-Барр ядерному білку антиген-лідеру (рис. 3.7.7). Цей мотив міститься в 13 інших білках, які зв'язуються з ко-шапероном людини BAG2, і також є передбаченим як посередник взаємодії зі спорідненим білком BAG3, але за низького порогу значущості. Епштейн-Барр ядерний білок антиген-лідер 5 (EBNA5), є одним з

перших білків, виявлених під час інфікування EBV, і є необхідним для трансформації В-клітин, діючи як транскрипційний ко-активуючий агент [123, 124]. BAG2 і BAG3 є ко-шапероновими білками HSP70 та HSC70 і працюють як фактор обміну нуклеотидів [125]. Таким чином, EBNA5 може посилювати діяльність ко-шаперонів HSP70 та HSC70 або впливати на проліферацію клітин або апоптоз через функції BAG2 або BAG3. Досліджуючи ще 13 білків, в яких цей мотив був передбачений, може надати більше доказів, чи BAG-домен дійсно може розпізнати мотив L.{0,1}Q.LR. Проте передбачення зв'язування вірусного пептиду (LGQLLR) з PDB структурою BAG3 миші (1uk5) або домену людського BAG1 (3fzf) з використанням PepSite 2 [126] не свідчить про наявність сильного сайту зв'язування у цьому домені.

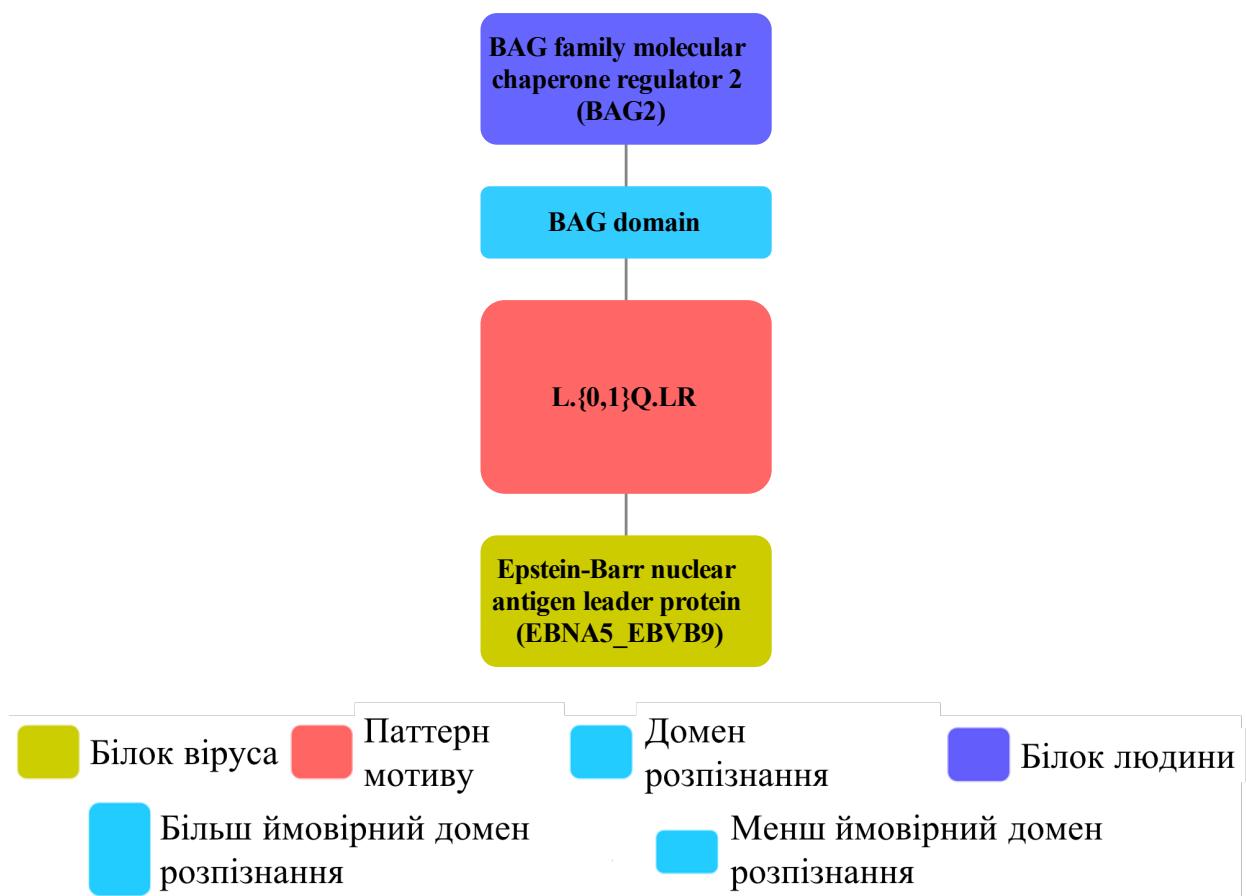


Рис 3.7.7. Схема мережі, що показує мотив-кандидат L.{0,1}Q.LR в білку EBNA5 віrusу Епштейна-Барр потенційно розпізнається доменом BAG у ко-шапероновому білку людини BAG2.

3.8 Майбутні напрямки дослідження

Ми показали, що ми можемо використовувати дані про взаємодію білків та властивість вірусних білків конвергентно еволюціонувати мотиви хазяїна, щоб відкрити заново 3 відомі мотиви з бази даних ELM та 6 інших мотивів, підтверджених попередніми дослідженнями. Також ми знаходимо приклади невідомих мотивів кандидатів та передбачаємо можливі домени розпізнавання. У деяких випадках відкритий мотив нагадує той, який, як відомо, пов'язує найбільш імовірний домен (домен PDZ, домен SH3, мотив DSG), але в багатьох випадках це не так. Більше роботи можна зробити для підвищення точності передбачення домену, а також точності та чутливості передбачення мотивів. У цьому розділі ми обговоримо підходи, які ми можемо використати.

3.8.1 Молекулярне стикування мотиву та домену й покращений аналіз мережі людини

Щоб покращити передбачення як мотиву, так і домену, ми можемо використовувати молекулярне стикування мотиву до домену, використовуючи PepSite2, як це було зроблено на кількох постхокових прикладах (DSG, BAG-доменний мотив) [126]. Це може дозволити приоритетизацію мотивів, які мають гарний структурний зв'язок з доменом, але також забезпечити незалежний спосіб оцінювання найбільш імовірного домену. Двома основними обмеженнями цього підходу є наявність доменних структур і низька чутливість методу. Наприклад, домен PDZ MAGI-1 був ко-кристалізованим з білком E6 HPV-16 [127], однак, PepSite2 не передбачав сильного сайта зв'язування PDZ-мотиву на поверхні цього домену (<http://pepsite2.russelllab.org/match?molvis=jsmol&pdb=2KPL&chain=A&ligand=RRETQL>). Обчислювальна швидкість не є обмеженням, PepSite2 достатньо швидкий для того, щоб дозволити докінг в масштабі інтеракту (принаймні,

не повільніше, ніж QSLIMFinder), щоб оцінити якомога більше пар-мотивів доменів.

Ми можемо провести більш детальний аналіз мережі взаємодій білків людини, щоб поліпшити передбачення домену взаємодії, яке в даний час здійснюється виключно на основі вірусно-людської мережі. По суті, ми передбачаємо домени тільки для вірусних білків, однак ми також ідентифікуємо мотиви в людських білках. Ми можемо вдосконалити прогнозування домену, розглядаючи як вірусні, так і людські білки, які поділяють один і той же мотив. Якщо 4 з 5 білків з мотивом мають один і той же домен, як збагачений - цей домен більше імовірно опосередковує взаємодію. Якщо всі 5 не співпадають, це може означати, що сам мотив не є функціональним.

У нашому аналізі ми не використовували консервацію послідовностей білків для обмеження областей білків, де ми шукаємо мотиви. Ця консервація, як правило, рекомендується [8], однак, вірусні білки еволюціонують швидко, тому фільтр консервації може видалити справжні мотиви [5]. Тим не менш, ми можемо використовувати консерваційний фільтр для білків людини: мотив повинен бути присутнім у людини та кількох інших тварин з добре анатованими геномами. Можливою проблемою може бути те, що використання мотиву вірусом може привести до відбору на мотивів людини, що може збільшити еволюційну швидкість зміни цих мотивів, що робить фільтр консервації неефективним. Однак це ніколи не було продемонстровано.

Ці 3 пропозиції можуть покращити окремі етапи нашої процедури для відкриття мотивів. Далі ми зможемо інтегрувати предикторів/передбачення з кожного кроку більш розумним чином.

3.8.2 Інтеграція кількох предикторів

Ми можемо застосувати підхід машинного навчання до інтеграції ймовірності мотиву, домену та їх взаємодії, передбаченого PepSite2. Кожна з цих ймовірностей надає корисну інформацію про мотив, який ми хочемо

знати. Поєднуючи це, ми можемо покращити як чутливість, так і специфічність прогнозування мотивів.

Найпростіший підхід до інтеграції полягає в припущені незалежності нашого передбачення та у множенні значень p-values, що надаються кожним з методів. Лінійні моделі забезпечують аналогічне рішення: зважена сума значень p-value. Обидві моделі мають недолік, коли сильний сигнал мотиву знижується слабким домена або слабким передбаченням від PepSite2. Оскільки ці слабкі передбачення можуть бути обумовлені відсутністю даних, аніж справжньої біологією, ми можемо обмежити нашу здатність відкрити мотиви. Для боротьби з цим ми можемо використовувати метод на основі дерева рішень, такий як random forest (випадковий ліс) або BART, який є стійким до відсутніх значень. Ці методи можуть вивчити сильний сигнал з одного джерела та об'єднати 3 слабких сигнали.

3.8.3 Експериментальна перевірка передбачених мотивів

Кінцевим кроком є експериментальна перевірка взаємодій нових мотивів-доменів. Як було обговорено в огляді літератури, різні класи мотивів вимагатимуть різних експериментів для функціональної перевірки. Однак, по-перше, нам потрібно перевірити фізичну взаємодію. Ми плануємо використовувати фагові дисплеї для кількох доменів проти невпорядкованого вірусного протеому для визначення специфічності зв'язування. У цьому дослідженні пептиди, які взаємодіють з доменом, ідентифікуються за допомогою NGS-секвенування фагонових геномів, що забезпечує високопродуктивну ідентифікацію взаємодій доменів-лінійних мотивів. Основним обмеженням є те, що кожний домен повинен синтезуватися *in vitro* [128]. Наш обчислювальний прогноз підкреслює екземпляри доменів в білках людини, які варто перевірити проти невпорядкованих ділянок вірусних білків.

ВИСНОВКИ

1. Ми отримали та обробили дані експериментальної взаємодії з публічних баз даних. Ми вивчили властивості мережі вірусно-людської взаємодії. Білки людини - мішені вірусів виступають як центральні, але цей ефект може бути результатом дослідницької упередженості в сукупному наборі даних білкових взаємодій.

2. За допомогою вірусно-людської мережі, ймовірнісних інструментів пошуку мотиву і послідовності вірусних білків, щоб обмежити область пошуку, ми можемо відкрити заново відомі приклади коротких лінійних мотивів у вірусних білках і передбачати нові мотиви-кандидати.

3. Ми визначили домени у всіх людських білках. Ми оцінили, які домени людини, ймовірно, опосередковують взаємодію з кожним вірусним білком. Ці домени є збагаченими на відомі домени розпізнавання мотивів. Фільтрація можливих доменів покращує відкликання. Інтеграція передбачення домену та мотиву підвищує інтерпретацію результатів.

4. За жорсткого порогу з точністю 50% ми можемо відкрити заново *de novo* 3 відомі екземпляри мотивів з нашого навчального набору, відкривши 6 екземпляри відомих мотивів, які не були в нашему тренувальному наборі. Ми передбачаємо 43 екземпляри нових мотивів кандидатів. Ці мотиви та їхні ймовірні домени розпізнавання будуть експериментально перевірятися за допомогою фагового дисплею.

5. Ми розробили цю процедуру пошуку мотивів статистичною мовою програмування R, використовуючи інструменти командного рядка та обчислювальний кластер. Ця процедура може бути використана групою для передбачення мотивів, коли будуть отримані нові дані про взаємодію білків.

6. Ця робота сприяє нашему розумінню коду взаємодії лінійного мотиву - домену розпізнавання та спрямовує вибір доменів людини-вірусних мішень для подальших експериментальних досліджень невпорядкованого вірусного протеому.

СПИСОК ДЖЕРЕЛ ЛІТЕРАТУРИ

1. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation / Van Roey K. [et al.]. // Chem. Rev. – 2014. – Vol. 114, № 13 – P. 6733–6778.
2. Design principles of regulatory networks: searching for the molecular algorithms of the cell / Lim W.A. [et al.]. // Mol. Cell – 2013. – Vol. 49, № 2 – P. 202–212.
3. Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation / Van Roey K. [et al.]. // Cell Commun. Signal – 2015. – Vol. 13 – P. 45.
4. Short linear motifs - ex nihilo evolution of protein regulation / Davey N.E. [et al.]. // Cell Commun. Signal – 2015. – Vol. 13 – P. 43.
5. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions / Hagai T. [et al.]. // Cell Rep – 2014. – Vol. 7, № 5 – P. 1729–1739.
6. Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer / Latysheva N.S. [et al.]. // Mol. Cell – 2016. – Vol. 63, № 4 – P. 579–592.
7. Computational prediction of short linear motifs from protein sequences / Edwards R.J. [et al.]. // Methods Mol. Biol. – 2015. – Vol. 1268 – P. 89–141.
8. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad / Gibson T.J. [et al.]. // Cell Commun. Signal – 2015. – Vol. 13 – P. 42.
9. High-throughput methods for identification of protein-protein interactions involving short linear motifs / Blikstad C. [et al.]. // Cell Commun. Signal – 2015. – Vol. 13 – P. 38.
10. Peptides mediating interaction networks: new leads at last / Neduvia V. [et al.]. // Curr. Opin. Biotechnol. – 2006. – Vol. 17, № 5 – P. 465–471.

11. Linear motifs: evolutionary interaction switches / Neduvva V. [et al.]. // FEBS Lett. – 2005. – Vol. 579, № 15 – P. 3342–3345.
12. Interactome-wide prediction of short, disordered protein interaction motifs in humans / Edwards R.J. [et al.]. // Mol Biosyst – 2012. – Vol. 8, № 1 – P. 282–295.
13. A human interactome in three quantitative dimensions organized by stoichiometries and abundances / Hein M.Y. [et al.]. // Cell – 2015. – Vol. 163, № 3 – P. 712–723.
14. Protein-protein interactions in human disease / Ryan D.P. [et al.]. // Curr. Opin. Struct. Biol. – 2005. – Vol. 15, № 4 – P. 441–446.
15. In vivo FRET-FLIM reveals cell-type-specific protein interactions in Arabidopsis roots / Long Y. [et al.]. // Nature – 2017. – Vol. 548, № 7665 – P. 97–102.
16. Attributes of short linear motifs / Davey N.E. [et al.]. // Mol Biosyst – 2012. – Vol. 8, № 1 – P. 268–281.
17. Local structural disorder imparts plasticity on linear motifs / Fuxreiter M. [et al.]. // Bioinformatics – 2007. – Vol. 23, № 8 – P. 950–956.
18. Structure of a regulatory complex involving the Abl SH3 domain, the Crk SH2 domain, and a Crk-derived phosphopeptide / Donaldson L.W. [et al.]. // Proc. Natl. Acad. Sci. U.S.A. – 2002. – Vol. 99, № 22 – P. 14053–14058.
19. SH2 domains recognize contextual peptide sequence information to determine selectivity / Liu B.A. [et al.]. // Mol. Cell Proteomics – 2010. – Vol. 9, № 11 – P. 2391–2404.
20. SH2 domains, interaction modules and cellular wiring / Pawson T. [et al.]. // Trends Cell Biol. – 2001. – Vol. 11, № 12 – P. 504–511.
21. Docking sites on substrate proteins direct extracellular signal-regulated kinase to phosphorylate specific residues / Fantz D.A. [et al.]. // J. Biol. Chem. – 2001. – Vol. 276, № 29 – P. 27256–27265.

22. Linear motif-mediated interactions have contributed to the evolution of modularity in complex protein interaction networks / Kim I. [et al.]. // PLoS Comput. Biol. – 2014. – Vol. 10 – № 10 – P. e1003881.
23. The eukaryotic linear motif resource - 2018 update / Gouw M. [et al.]. // Nucleic Acids Res. – 2018. – Vol. 46, № D1 – P. D428–D434.
24. Mechanism of the nuclear receptor molecular switch / Nagy L. [et al.]. // Trends Biochem. Sci. – 2004. – Vol. 29, № 6 – P. 317–324.
25. Scaffold proteins: hubs for controlling the flow of cellular information / Good M.C. [et al.]. // Science – 2011. – Vol. 332, № 6030 – P. 680–686.
26. A Proteome-wide screen for mammalian SxIP motif-containing microtubule plus-end tracking proteins / Jiang K. [et al.]. // Curr. Biol. – 2012. – Vol. 22, № 19 – P. 1800–1807.
27. Mechanisms of CAS substrate domain tyrosine phosphorylation by FAK and Src / Ruest P.J. [et al.]. // Mol. Cell. Biol. – 2001. – Vol. 21, № 22 – P. 7641–7652.
28. The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition / Cheng K.-Y. [et al.]. // J. Biol. Chem. – 2006. – Vol. 281, № 32 – P. 23167–23179.
29. Ubiquitin: structures, functions, mechanisms / Pickart C.M. [et al.]. // Biochim. Biophys. Acta – 2004. – Vol. 1695, № 1–3 – P. 55–72.
30. Mechanisms of mono- and poly-ubiquitination: Ubiquitination specificity depends on compatibility between the E2 catalytic core and amino acid residues proximal to the lysine / Sadowski M. [et al.]. // Cell Div – 2010. – Vol. 5 – P. 19.
31. Types of Ubiquitin Ligases / Morreale F.E. [et al.]. // Cell – 2016. – Vol. 165, № 1 – P. 248-248.e1.
32. Function and regulation of cullin-RING ubiquitin ligases / Petroski M.D. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2005. – Vol. 6, № 1 – P. 9–20.
33. SCF ubiquitin ligase-targeted therapies / Skaar J.R. [et al.]. // Nat Rev Drug Discov – 2014. – Vol. 13, № 12 – P. 889–903.

34. Non-histone protein methylation as a regulator of cellular signalling and function / Biggar K.K. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2015. – Vol. 16, № 1 – P. 5–17.
35. The growing landscape of lysine acetylation links metabolism and cell signalling / Choudhary C. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2014. – Vol. 15, № 8 – P. 536–550.
36. Function and regulation of SUMO proteases / Hickey C.M. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2012. – Vol. 13, № 12 – P. 755–766.
37. H₂S signalling through protein sulfhydration and beyond / Paul B.D. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2012. – Vol. 13, № 8 – P. 499–507.
38. Mechanisms of specificity in protein phosphorylation / Ubersax J.A. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2007. – Vol. 8, № 7 – P. 530–541.
39. Protein O-GlcNAcylation: emerging mechanisms and functions / Yang X. [et al.]. // Nat. Rev. Mol. Cell Biol. – 2017. – Vol. 18, № 7 – P. 452–465.
40. Regulation of chromatin by histone modifications / Bannister A.J. [et al.]. // Cell Res. – 2011. – Vol. 21, № 3 – P. 381–395.
41. Evidence for in vivo phosphorylation of the Grb2 SH2-domain binding site on focal adhesion kinase by Src-family protein-tyrosine kinases / Schlaepfer D.D. [et al.]. // Mol. Cell. Biol. – 1996. – Vol. 16, № 10 – P. 5623–5633.
42. Caspase functions in cell death and disease / McIlwain D.R. [et al.]. // Cold Spring Harb Perspect Biol – 2013. – Vol. 5, № 4 – P. a008656.
43. Inflammasomes: mechanism of action, role in disease, and therapeutics / Guo H. [et al.]. // Nat. Med. – 2015. – Vol. 21, № 7 – P. 677–687.
44. Caspase substrates and inhibitors / Poreba M. [et al.]. // Cold Spring Harb Perspect Biol – 2013. – Vol. 5, № 8 – P. a008680.
45. ELM 2016--data update and new functionality of the eukaryotic linear motif resource / Dinkel H. [et al.]. // Nucleic Acids Res. – 2016. – Vol. 44, № D1 – P. D294–300.

46. Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics / Bashor C.J. [et al.]. // Science – 2008. – Vol. 319, № 5869 – P. 1539–1543.
47. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex / Yu X. [et al.]. // Science – 2003. – Vol. 302, № 5647 – P. 1056–1060.
48. Evolution of protein-protein interaction network / Makino T. [et al.]. // Genome Dyn – 2007. – Vol. 3 – P. 13–29.
49. Tissue-specific splicing of disordered segments that embed binding motifs rewrites protein interaction networks / Buljan M. [et al.]. // Mol. Cell – 2012. – Vol. 46, № 6 – P. 871–883.
50. The nature of protein domain evolution: shaping the interaction network / Bagowski C.P. [et al.]. // Curr. Genomics – 2010. – Vol. 11, № 5 – P. 368–376.
51. Classification of intrinsically disordered regions and proteins / van der Lee R. [et al.]. // Chem. Rev. – 2014. – Vol. 114, № 13 – P. 6589–6631.
52. Convergent evolution of domain architectures (is rare) / Gough J. [et al.]. // Bioinformatics – 2005. – Vol. 21, № 8 – P. 1464–1471.
53. Cyclin is degraded by the ubiquitin pathway / Glotzer M. [et al.]. // Nature – 1991. – Vol. 349, № 6305 – P. 132–138.
54. Analysis of the BiP gene and identification of an ER retention signal in *Schizosaccharomyces pombe* / Pidoux A.L. [et al.]. // EMBO J. – 1992. – Vol. 11, № 4 – P. 1583–1591.
55. The transience of transient overexpression / Gibson T.J. [et al.]. // Nat. Methods – 2013. – Vol. 10, № 8 – P. 715–721.
56. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences / Chica C. [et al.]. // BMC Bioinformatics – 2008. – Vol. 9 – P. 229.

57. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins / Chavali S. [et al.]. // Nat. Struct. Mol. Biol. – 2017. – Vol. 24, № 9 – P. 765–777.
58. Architecture of the human interactome defines protein communities and disease networks / Huttlin E.L. [et al.]. // Nature – 2017. – Vol. 545, № 7655 – P. 505–509.
59. QSLiMFinder: improved short linear motif prediction using specific query protein data / Palopoli N. [et al.]. // Bioinformatics – 2015. – Vol. 31, № 14 – P. 2284–2293.
60. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases / Orchard S. [et al.]. // Nucleic Acids Res. – 2014. – Vol. 42, № Database issue – P. D358-363.
61. Unstructural biology of the Dengue virus proteins / Meng F. [et al.]. // FEBS J. – 2015. – Vol. 282, № 17 – P. 3368–3394.
62. A proteome-scale map of the human interactome network / Rolland T. [et al.]. // Cell – 2014. – Vol. 159, № 5 – P. 1212–1226.
63. Fundamentals of protein interaction network mapping / Snider J. [et al.]. // Mol. Syst. Biol. – 2015. – Vol. 11, № 12 – P. 848.
64. UniProt: the universal protein knowledgebase / The UniProt Consortium [et al.]. // Nucleic Acids Res. – 2017. – Vol. 45, № D1 – P. D158–D169.
65. InterPro in 2017-beyond protein family and domain annotations / Finn R.D. [et al.]. // Nucleic Acids Res. – 2017. – Vol. 45, № D1 – P. D190–D199.
66. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures / Marchler-Bauer A. [et al.]. // Nucleic Acids Res. – 2017. – Vol. 45, № D1 – P. D200–D203.
67. CATH: an expanded resource to predict protein function through structure and sequence / Dawson N.L. [et al.]. // Nucleic Acids Res. – 2017. – Vol. 45, № D1 – P. D289–D295.

68. HAMAP in 2015: updates to the protein family classification and annotation system / Pedruzzi I. [et al.]. // Nucleic Acids Res. – 2015. – Vol. 43, № Database issue – P. D1064-1070.
69. The Pfam protein families database: towards a more sustainable future / Finn R.D. [et al.]. // Nucleic Acids Res. – 2016. – Vol. 44, № D1 – P. D279-285.
70. PIRSF family classification system for protein functional and evolutionary analysis / Nikolskaya A.N. [et al.]. // Evol. Bioinform. Online – 2007. – Vol. 2 – P. 197–209.
71. The PRINTS protein fingerprint database: functional and evolutionary applications / Attwood T.K. [et al.]. // Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics / ed. Jorde L.B. et al. – 2006.
72. ProDom: automated clustering of homologous domains / Servant F. [et al.]. // Brief. Bioinformatics – 2002. – Vol. 3, № 3 – P. 246–251.
73. New and continuing developments at PROSITE / Sigrist C.J.A. [et al.]. // Nucleic Acids Res. – 2013. – Vol. 41, № Database issue – P. D344-347.
74. The Structure-Function Linkage Database / Akiva E. [et al.]. // Nucleic Acids Res. – 2014. – Vol. 42, № Database issue – P. D521-530.
75. 20 years of the SMART protein domain annotation resource / Letunic I. [et al.]. // Nucleic Acids Res. – 2018. – Vol. 46, № D1 – P. D493–D496.
76. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure / Gough J. [et al.]. // J. Mol. Biol. – 2001. – Vol. 313, № 4 – P. 903–919.
77. The TIGRFAMs database of protein families / Haft D.H. [et al.]. // Nucleic Acids Res. – 2003. – Vol. 31, № 1 – P. 371–373.
78. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins / Edwards R.J. [et al.]. // PLoS ONE – 2007. – Vol. 2, № 10 – P. e967.
79. BLAST+: architecture and applications / Camacho C. [et al.]. // BMC Bioinformatics – 2009. – Vol. 10 – P. 421.

80. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins / Dosztányi Z. [et al.]. // J. Mol. Biol. – 2005. – Vol. 347, № 4 – P. 827–839.
81. CompaMotif: quick and easy comparisons of sequence motifs / Edwards R.J. [et al.]. // Bioinformatics – 2008. – Vol. 24, № 10 – P. 1307–1309.
82. R package MItools / Kleshchevnikov V. [et al.]. // – 2018.
83. R project viral_project / Kleshchevnikov V. [et al.]. // – 2018.
84. Molecular principles of human virus protein-protein interactions / Halehalli R.R. [et al.]. // Bioinformatics – 2015. – Vol. 31, № 7 – P. 1025–1033.
85. Data-warehousing of protein-protein interactions indicates that pathogens preferentially target hub and bottleneck proteins / Schleker S. [et al.]. // Front Microbiol – 2013. – Vol. 4 – P. 51.
86. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types / Schaefer M.H. [et al.]. // Front Genet – 2015. – Vol. 6 – P. 260.
87. T antigen is bound to a host protein in SV40-transformed cells / Lane D.P. [et al.]. // Nature – 1979. – Vol. 278, № 5701 – P. 261–263.
88. Fast and accurate discovery of degenerate linear motifs in protein sequences / Kelil A. [et al.]. // PLoS ONE – 2014. – Vol. 9, № 9 – P. e106081.
89. Estimation and efficient computation of the true probability of recurrence of short linear protein sequence motifs in unrelated proteins / Davey N.E. [et al.]. // BMC Bioinformatics – 2010. – Vol. 11 – P. 14.
90. Analysis of Multiple HPV E6 PDZ Interactions Defines Type-Specific PDZ Fingerprints That Predict Oncogenic Potential / Thomas M. [et al.]. // PLoS Pathog. – 2016. – Vol. 12, № 8 – P. e1005766.
91. Oncogenic human papillomavirus E6 proteins target the discs large tumour suppressor for proteasome-mediated degradation / Gardiol D. [et al.]. // Oncogene – 1999. – Vol. 18, № 40 – P. 5487–5496.

92. Degradation of tyrosine phosphatase PTPN3 (PTPH1) by association with oncogenic human papillomavirus E6 proteins / Jing M. [et al.]. // J. Virol. – 2007. – Vol. 81, № 5 – P. 2231–2239.
93. The human papillomavirus (HPV) E6* proteins from high-risk, mucosal HPVs can direct degradation of cellular proteins in the absence of full-length E6 protein / Pim D. [et al.]. // J. Virol. – 2009. – Vol. 83, № 19 – P. 9863–9874.
94. Association of E6AP (UBE3A) with human papillomavirus type 11 E6 protein / Brimer N. [et al.]. // Virology – 2007. – Vol. 358, № 2 – P. 303–310.
95. Identification of a novel telomerase repressor that interacts with the human papillomavirus type-16 E6/E6-AP complex / Gewin L. [et al.]. // Genes Dev. – 2004. – Vol. 18, № 18 – P. 2269–2282.
96. The high-risk HPV E6 target scribble (hScrib) is required for HPV E6 expression in cervical tumour-derived cell lines / Kranjec C. [et al.]. // Papillomavirus Res – 2016. – Vol. 2 – P. 70–77.
97. A role for Erbin in the regulation of Nod2-dependent NF- κ B signaling / McDonald C. [et al.]. // J. Biol. Chem. – 2005. – Vol. 280, № 48 – P. 40301–40309.
98. ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor / Borg J.P. [et al.]. // Nat. Cell Biol. – 2000. – Vol. 2, № 7 – P. 407–414.
99. The avian influenza virus NS1 ESEV PDZ binding motif associates with Dlg1 and Scribble to disrupt cellular tight junctions / Golebiewski L. [et al.]. // J. Virol. – 2011. – Vol. 85, № 20 – P. 10639–10648.
100. Analysis of the PDZ binding specificities of Influenza A virus NS1 proteins / Thomas M. [et al.]. // Virol. J. – 2011. – Vol. 8 – P. 25.
101. The ESEV PDZ-binding motif of the avian influenza A virus NS1 protein protects infected cells from apoptosis by directly targeting Scribble / Liu H. [et al.]. // J. Virol. – 2010. – Vol. 84, № 21 – P. 11164–11174.

102. SH3 domains. Molecular “Velcro” / Morton C.J. [et al.]. // *Curr. Biol.* – 1994. – Vol. 4, № 7 – P. 615–617.

103. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4 / Saksela K. [et al.]. // *EMBO J.* – 1995. – Vol. 14, № 3 – P. 484–491.

104. HIV-1 Nef selectively activates Src family kinases Hck, Lyn, and c-Src through direct SH3 domain interaction / Trible R.P. [et al.]. // *J. Biol. Chem.* – 2006. – Vol. 281, № 37 – P. 27029–27038.

105. IkappaBalphа ubiquitination is catalyzed by an SCF-like complex containing Skp1, cullin-1, and two F-box/WD40-repeat proteins, betaTrCP1 and betaTrCP2 / Suzuki H. [et al.]. // *Biochem. Biophys. Res. Commun.* – 1999. – Vol. 256, № 1 – P. 127–132.

106. The SCF(HOS/beta-TRCP)-ROC1 E3 ubiquitin ligase utilizes two distinct domains within CUL1 for substrate targeting and ubiquitin ligation / Wu K. [et al.]. // *Mol. Cell. Biol.* – 2000. – Vol. 20, № 4 – P. 1382–1393.

107. HOS, a human homolog of Slimb, forms an SCF complex with Skp1 and Cullin1 and targets the phosphorylation-dependent degradation of IkappaB and beta-catenin / Fuchs S.Y. [et al.]. // *Oncogene* – 1999. – Vol. 18, № 12 – P. 2039–2046.

108. A complex containing betaTrCP recruits Cdc34 to catalyse ubiquitination of IkappaBalphа / Vuillard L. [et al.]. // *FEBS Lett.* – 1999. – Vol. 455, № 3 – P. 311–314.

109. HIV-1 Vpu neutralizes the antiviral factor Tetherin/BST-2 by binding it and directing its beta-TrCP2-dependent degradation / Mangeat B. [et al.]. // *PLoS Pathog.* – 2009. – Vol. 5, № 9 – P. e1000574.

110. A novel human WD protein, h-beta TrCp, that interacts with HIV-1 Vpu connects CD4 to the ER degradation pathway through an F-box motif / Margottin F. [et al.]. // *Mol. Cell* – 1998. – Vol. 1, № 4 – P. 565–574.

111. Rotavirus NSP1 Requires Casein Kinase II-Mediated Phosphorylation for Hijacking of Cullin-RING Ligases / Davis K.A. [et al.]. // MBio – 2017. – Vol. 8, № 4.
112. The Rotavirus Interferon Antagonist NSP1: Many Targets, Many Questions / Arnold M.M. [et al.]. // J. Virol. – 2016. – Vol. 90, № 11 – P. 5212–5215.
113. Zinc-binding domain of rotavirus NSP1 is required for proteasome-dependent degradation of IRF3 and autoregulatory NSP1 stability / Graff J.W. [et al.]. // J. Gen. Virol. – 2007. – Vol. 88, № Pt 2 – P. 613–620.
114. Interferon-λ in the context of viral infections: production, response and therapeutic implications / Hermant P. [et al.]. // J Innate Immun – 2014. – Vol. 6, № 5 – P. 563–574.
115. Cul7/p185/p193 binding to simian virus 40 large T antigen has a role in cellular transformation / Ali S.H. [et al.]. // J. Virol. – 2004. – Vol. 78, № 6 – P. 2749–2757.
116. The Many Faces of MDM2 Binding Partners / Riley M.F. [et al.]. // Genes Cancer – 2012. – Vol. 3, № 3–4 – P. 226–239.
117. Interaction and co-localization of JC virus large T antigen and the F-box protein β-transducin-repeat containing protein / Reviriego-Mendoza M.M. [et al.]. // Virology – 2011. – Vol. 410, № 1 – P. 119–128.
118. Binding of the influenza virus NS1 protein to double-stranded RNA inhibits the activation of the protein kinase that phosphorylates the eIF-2 translation initiation factor / Lu Y. [et al.]. // Virology – 1995. – Vol. 214, № 1 – P. 222–228.
119. Species-specific inhibition of RIG-I ubiquitination and IFN induction by the influenza A virus NS1 protein / Rajsbaum R. [et al.]. // PLoS Pathog. – 2012. – Vol. 8, № 11 – P. e1003059.
120. Secretory cargo sorting by Ca²⁺-dependent Cab45 oligomerization at the trans-Golgi network / Crevenna A.H. [et al.]. // J. Cell Biol. – 2016. – Vol. 213, № 3 – P. 305–314.

121. Comparative influenza protein interactomes identify the role of plakophilin 2 in virus restriction / Wang L. [et al.]. // *Nat Commun* – 2017. – Vol. 8 – P. 13876.
122. The RNA polymerase of influenza a virus: mechanisms of viral transcription and replication / Fodor E. [et al.]. // *Acta Virol.* – 2013. – Vol. 57, № 2 – P. 113–122.
123. Mediation of Epstein-Barr virus EBNA-LP transcriptional coactivation by Sp100 / Ling P.D. [et al.]. // *EMBO J.* – 2005. – Vol. 24, № 20 – P. 3565–3575.
124. Epstein-Barr virus nuclear antigen EBNA-LP is essential for transforming naïve B cells, and facilitates recruitment of transcription factors to the viral genome / Szymula A. [et al.]. // *PLoS Pathog.* – 2018. – Vol. 14, № 2 – P. e1006890.
125. An evolutionarily conserved family of Hsp70/Hsc70 molecular chaperone regulators / Takayama S. [et al.]. // *J. Biol. Chem.* – 1999. – Vol. 274, № 2 – P. 781–786.
126. PepSite: prediction of peptide-binding sites from protein surfaces / Trabuco L.G. [et al.]. // *Nucleic Acids Res.* – 2012. – Vol. 40, № Web Server issue – P. W423-427.
127. The structural and dynamic response of MAGI-1 PDZ1 with noncanonical domain boundaries to the binding of human papillomavirus E6 / Charbonnier S. [et al.]. // *J. Mol. Biol.* – 2011. – Vol. 406, № 5 – P. 745–763.
128. Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome / Davey N.E. [et al.]. // *FEBS J.* – 2017. – Vol. 284, № 3 – P. 485–498.

ДОДАТКИ

Додаток А

ГРАФІК, ЩО ПОКАЗУЄ ЩІЛЬНІСТЬ РОЗПОДІЛУ ЧИСЛА ВЗАЄМОДІЙ ЛЮДСЬКИХ ЧИ ВІРУСНИХ ПРОТЕЇНІВ

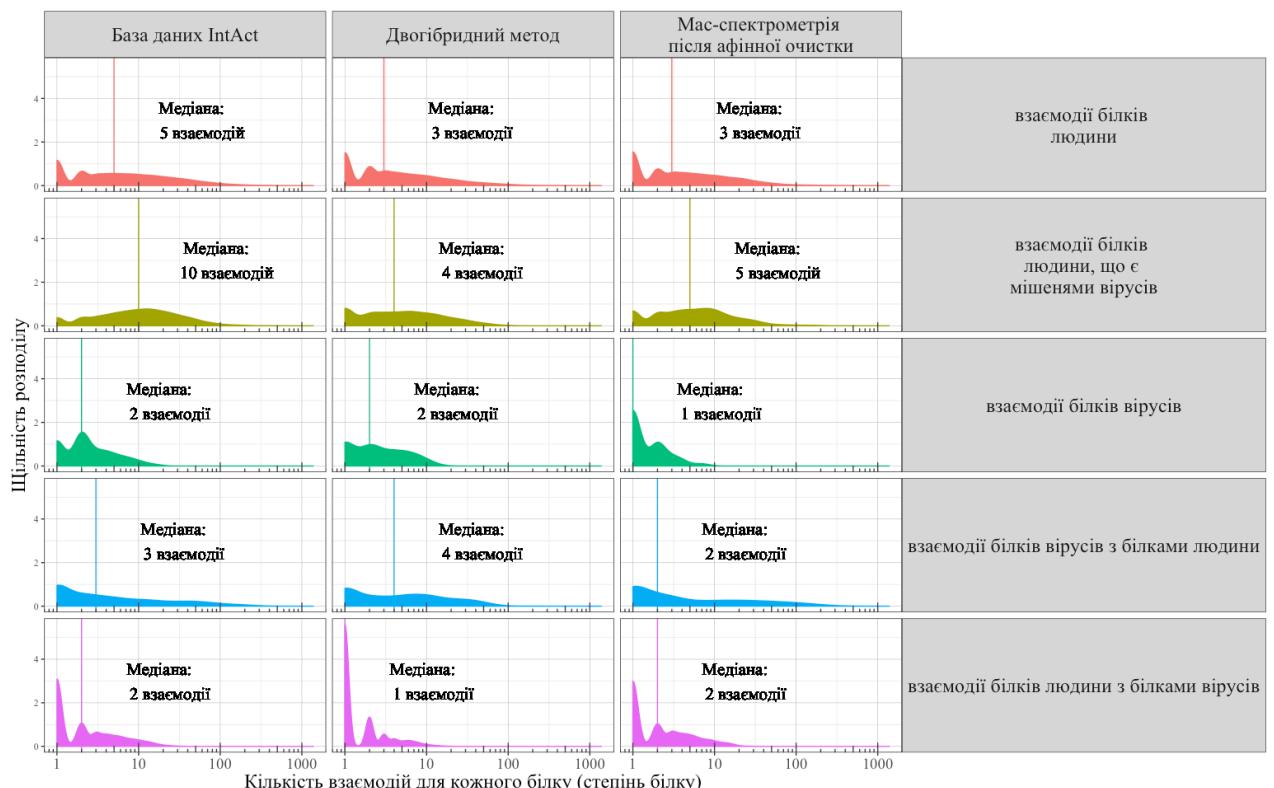


Рис 4.1. Графік, що показує щільність розподілу числа взаємодій кожного людського чи вірусного протеїну в кожній мережі, яку ми використовували для нашого аналізу. Для кожного білка вісь X показує кількість білків що з ним взаємодіють, вісь Y показує щільність розподілу. Різні мережі та різні білки (вірусні чи людські) показані в рядках. Різні методи визначення білкових взаємодій, або всі наявні дані у базі даних IntAct, зазначені у стовбцях.

ДІАГРАМА ВЕНА ЩО ПОКАЗУЄ

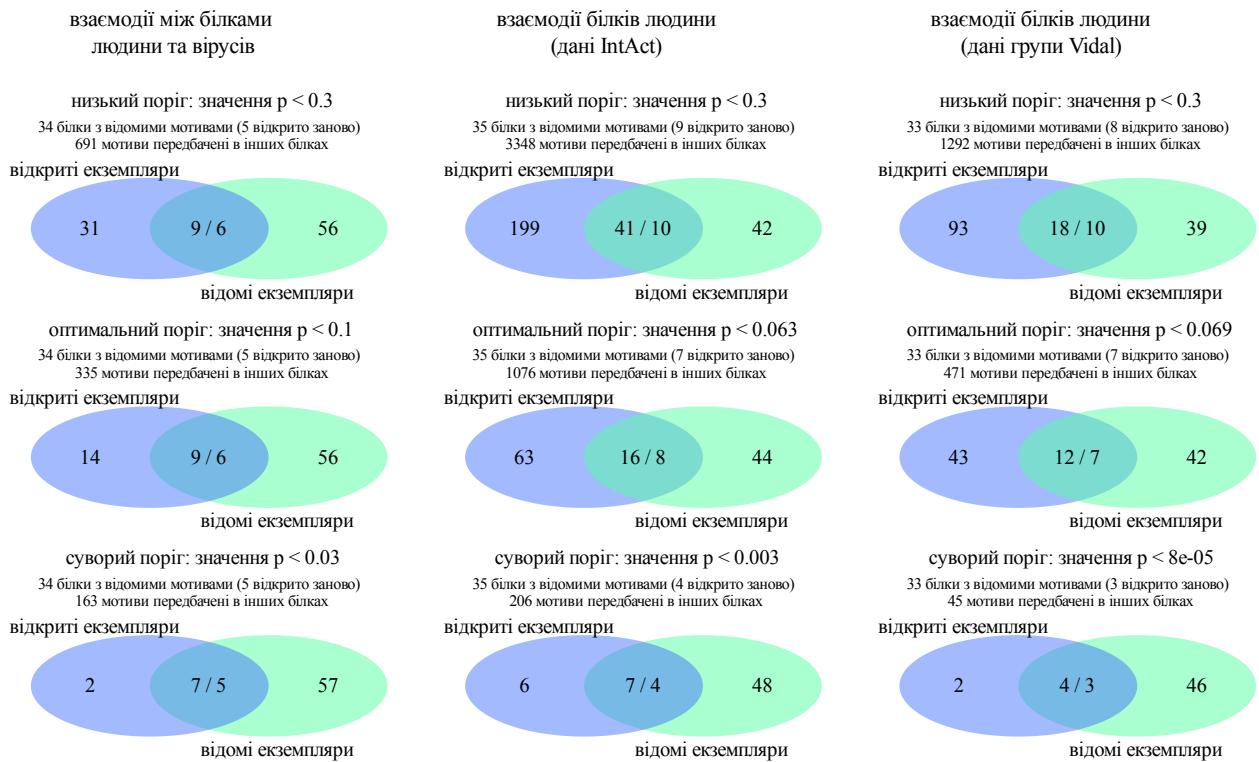


Рис 3.4.2. Діаграми Вена, що показують кількість виявлених мотивів-кандидатів та відомі мотиви ми відкрили заново. Наведено три стратегії побудови наборів даних (стовбці) та 3 пороги значень p -value (рядки). Блакитне коло показує кількість екземплярів мотивів, передбачених, але невідомих. Зелене коло показує кількість екземплярів мотивів, відомих, але не відкритих заново. Накладання показує кількість виявлених екземплярів, які відповідають відомим екземплярам (передбачені / відомі). Ці цифри відрізняються, оскільки кілька схожих передбачених екземплярів мотивів можуть співпадати з одним відомим мотивом у тому ж місці в послідовності білка (наприклад, див. розділ 3.7.2). Коректоване значення p -value - це QSLIMFinder Sig, який є вірогідністю спостереження N числа мотивів у випадковій послідовності, скоригованої для числа тестів всіх можливих мотивів. Низький поріг відображає ймовірність такого високого помилкового відкриття, як ми максимально готові допустити. При оптимальному порозі,

точність (precision), частка відкритих випадків, які відповідають відомим, приблизно дорівнює відкликанню (recall), частка відомих випадків, які ми відкрили заново. За жорсткого порогу, точність становить 0,5 або вище, ми виявляємо в середньому або новий мотив-кандидат, або один помилковий мотив, для кожного відомого мотиву у кожному вірусному білку. Для кожного набору даних і порогу ми показуємо, скільки мотивів було виявлено у білках, які не містять відомих мотивів.

Додаток В**ІНСТРУКЦІЯ ДЛЯ ЗАПУСКУ ПРОГРАМИ ДЛЯ ПЕРЕДБАЧЕННЯ
ДОМЕНІВ**

```
/path_to/interproscan-5.25-64.0/interproscan.sh -i ./data_files/all_human_viral_proteins.fasta -f gff3 -iprlookup -goterms -b ./processed_data_files/all_human_viral_protein_domains102017
```

Рис 4.1. Код командної строки BASH був використаний для запуску програми InterProScan

ПІДСУМОК НАБОРІВ ДАНИХ ДЛЯ ПОШУКУ МОТИВІВ

Таблиця 4.1

Набори даних для QSLIMFinder, які були протестовані

ID набору даних	Query, запит мережа	Головна мережа	cloudfix	Оцінка продуктивності
qslimfinder. Full_IntAct3 cloudfixF.FALSE	Вірусно-людська мережа	Всі дані IntAct	FALSE	2
qslimfinder. BioPlex3 cloudfixF.FALSE	Вірусно-людська мережа	BioPlex	FALSE	2
qslimfinder. all_viral_interaction3 cloudfixF.FALSE	Вірусно-людська мережа	Вірусно-людська мережа	FALSE	1
qslimfinder. randomised_BioPlex3 cloudfixF.FALSE	Вірусно-людська мережа	Рандомізованій Bioplex	FALSE	4
qslimfinder.randomise d _all_viral_interaction3 cloudfixF.FALSE	Рандомізована вірусно-людська мережа	Рандомізована вірусно-людська мережа	FALSE	4
qslimfinder. Full_IntAct3.FALSE	Вірусно-людська мережа	Всі дані IntAct	TRUE	2

qslimfinder. Vidal3.FALSE	Вірусно- людська мережа	Дані групи Vidal	TRUE	3
qslimfinder. all_viral_interaction3. FALSE	Вірусно- людська мережа	Вірусно- людська мережа	TRUE	1

ІНСТРУКЦІЯ ДЛЯ ЗАПУСКУ ПРОГРАМИ ДЛЯ ПЕРЕДБАЧЕННЯ МОТИВІВ

```
bsub -n 1 -q research-rh7 -M 100 -R \"rusage[mem=100]\" python path_to.slimsuite/tools/qslimfinder.py blast+path=path_to.ncbi_blast_2.6.0/bin/ iupath=path_to/iupred/iupred dismask=T consmask=F cloudfix=F probcut=0.3 minwild=0 maxwild=2 slimlen=5 alphahelix=F maxseq=800 savespace=0 iuchdir=T extras=2 resdir=path_to/output/interactors_of.A0FGR8.P0DOE9./ resfile=path_to/output/interactors_of.A0FGR8.P0DOE9./main_result seqin=path_to/input/fasta/interactors_of.A0FGR8.P0DOE9.fas query=path_to/input/query/interactors_of.A0FGR8.P0DOE9.fas
```

Рис 4.2. Код командної строки BASH був використаний для запуску програми QSLIMFinder