

КИЇВСЬКИЙ 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

ЗМІСТ	2
ВСТУП	5
РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ	7
1.1 Короткі лінійні мотиви.....	7
1.1.1 Модулі білок-білкових взаємодій.....	7
1.1.2 Короткі лінійні мотиви та ділянки молекулярного розпізнавання	8
1.1.3 Класи коротких лінійних мотивів	10
1.2 Експресія білків, еволюція та сплайсинг впливають на клітинну функцію шляхом зміни структури мережі взаємодій	14
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 і 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	26

2.4 Статистичний метод оцінки того, які домени ймовірно 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 ймовірно опосередковують взаємодію між білками	41
3.4 Пошук коротких лінійних мотивів	44
3.5 Дослідження ефекту фільтрації за ймовірним доменом розпізнання на чутливість передбачення мотивів	50
3.6 Дослідження схожості мотивів знайдених <i>de novo</i> до відомих мотивів	52

3.7 Приклади відкритих заново та мотивів-кандидатів.....	54
3.7.1 Дослідження класів мотивів, що відкрито заново, мотивів-кандидатів та їх ймовірних доменів розпізнавання	54
3.7.2 Мотивів PDZ, які відкрито заново.....	55
3.7.3 Мотиви-кандидати, що зв'язують домен PDZ	58
3.7.4 Мотив-кандидат, що зв'язує домен SH3	61
3.7.5 Мотиви-кандидати, що зв'язують домен WD40	63
3.7.6 Мотиви-кандидати, що розпізнаються доменом, що зв'язує дволанцюгову РНК, та доменом EF-hand	67
3.7.7 Мотив-кандидат, що зв'язує BAG-домен	69
3.8 Майбутні напрямки дослідження	71
3.8.1 Молекулярне стикування мотиву та домену і покращений аналіз на людській стороні	71
3.8.2 Інтеграція декількох предикторів.....	72
3.8.3 Експериментальна перевірка передбачених мотивів	73
ВИСНОВКИ.....	74
СПИСОК ДЖЕРЕЛ ЛІТЕРАТУРИ.....	76
ДОДАТКИ.....	77
Додаток А	77
Додаток Б	78
Додаток В	79
Додаток Г	80
Додаток Д.....	81
Додаток Е	82
Додаток Ж	84

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 [24926813].

Відомо, що взаємодії, опосередковані лінійним мотивом, з'єднують і направляють сигнальні шляхи клітин [PMC3664230]. Ця функція часто додатково регулюються посттрансляційними модифікаціями і кооперативнотю взаємодій [PMC4666095]. Лінійні мотив-опосередковані взаємодії можуть швидко еволюціонувати і допомагати змінювати сигнальні мережі клітини при видоутворенні, у захворюваннях або взаємодії патогена-хазяїна [27540857, PMC4089993, PMC4654906]

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 [24926813]. Використання обчислювальних інструментів пошуку для ідентифікації лінійних мотивів у гомологічних білках, як правило, передбачує велику кількість нефункціональних мотивів. Декілька підходів показали покращення ефективності виявлення функціональних мотивів: включення даних про взаємодію білків, консервація послідовності у декількох видах та фільтрування для мотивів, розташованих у неструктурзованих ділянках [25555723, PMC4652402]. Методи, такі як фаговий дисплей, були розроблені, щоб допомогти експериментальному виявленню мотивів у масштабах протеому [26297553]. 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 [15943979, 16962311].

Вірусні білки імітують клітинні лінійні мотиви для взаємодії та модифікації клітинної сигналізації таким чином, що сприяє прогресуванню

вірусної інфекції [PMC4089993]. 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 [21879107], ми також 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 за нашою.

Метою даного дослідження є використання даних взаємодії вірусних білків з білками хазяїна та доменів які їх ймовірно розпізнають як засобу обмеження простору пошуку для відкриття нових функціональних лінійних мотивів.

Ця робота може сприяти розумінню коду взаємодії доменів-лінійних мотивів та того, як віруси використовують цей механізм.

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.

РОЗДІЛ 1

ОГЛЯД ЛІТЕРАТУРИ

1.1 Короткі лінійні мотиви

1.1.1 Модулі білок-білкових взаємодій

Структура та функції клітин виникають внаслідок взаємодії між молекулами усередині та ззовні клітин [26496610]. 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 [15993577, 27540857, 24882001]. Для створення цих фенотипів білки взаємодіють у специфічних умовах у визначених типах клітин та субклітинних локалізаціях [28746306]. 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 доповнюються короткими лінійними мотивами (SLiM або лінійні мотиви), розташованими в гнучких невпорядкованих ділянках [24926813].

1.1.2 Короткі лінійні мотиви та ділянки молекулярного розпізнавання

Короткі лінійні мотиви (SLiMs) - мотиви послідовності 3-15 амінокислотних залишків, що опосередковують фізичну та селективну взаємодію між білками. Лінійна послідовність мотива, а не його тривимірна структура, вважається важливою для зв'язування. SLiMs, як правило, розташовані в невпорядкованій ділянці білка [21909575, 17387114]. Це може бути довга невпорядкована ділянка або коротка петля на поверхні глобулярного домену [12384576]. 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 [PMC2984226]. 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 (наприклад, [PMC2984226]). Форма та фізико-хімічні властивості зв'язуючої кишени визначають специфічність, спорідненість та селективність домену розпізнавання до послідовності [24926813].

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 (PMID: 11719057). Контекст послідовності навколо фосфорилюваних 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 MAP-кінази) до їх мішеней [11371562]. 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 [24926813].

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 декількох мотивів для високоавідного біологічно важливого зв'язування. Це, поряд з іншими властивостями SLiM, ідеально підходить для з'єднання білкових комплексів [26496610, PMC4191887], 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 мішенню 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 [24926813]. У базі даних ELM, яка збирає екземпляри відомих мотивів з літератури, анотовано 6 типів мотивів [PMC5753338].

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 транскрипційній репресорів або активаторів через CoRNR мотив або NR-box мотив, залежно від зв'язування стероїдного гормону - їхнього ліганда [15276186]. 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 [21551057].

Таргенг-мотиви керують локалізацією білків, направляючи транслокацію білків між субклітинними компартментами за допомогою конкретних 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 [22885064]).

Докінг-мотиви рекрутують модифікуючі ферментів і широко використовуються для підвищення субстратної специфічності ферментів. Докінг-мотиви часто підводять ферменти для модифікації іншої ділянки в субстраті. Докінг-мотиви можуть привести до розпізнавання субстрату у 3 основних способи. Мотиви можуть розпізнаватися сайтом в каталітичному домені, відмінному від каталітичного сайту. Наприклад, докінг-жолоб каталітичного домену MAP кінази розпізнає докінг мотив у MEF2A, MAP2K1 або MKP1 [24926813]. 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 [11604500]. Нарешті, домен розпізнання, який зв'язує субстрат, може бути розташований в іншому білку, який утворює комплекс з ферментом. Спочатку цей комплекс повинен бути зібраний, що може покладатися на взаємодію лінійного мотива або доменів. Одним із прикладів є CDK (циклінзалежні кінази), які покладаються на домен розпізнання в цикліновому білку для розпізнавання їх мішеней [16707497].

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) [15571809, 20704751]. Е3-убіквітин-лігази основними детермінантами специфічності деградації та є найбільш

поширеними в геномі людини (> 700 E3 ферментів, ~ 40 E2 ферментів, 2 E1 ферментів) [15571809, 25394868, 27015313, 15688063].

Мотиви посттрансляційної модифікації. Друга велика група мотивів збігається з сайтом посттрансляційної модифікації (PTM) і опосередковує розпізнання 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; найвідомішими прикладами є сімейство білків-циклофілінів та PIN1 [24926813].

Багато видів PTM типу приєднання-видалення групи були відкриті: фосфорилювання, ацетилювання, метилювання, SUMO-лювання, убіквітінювання, заякорення ліпідів до мембрани і багато інших, менш поширених модифікацій [17585314, 25053359, 25491103, 28488703, 22781905, 23175280]. Вони широко використовуються в декількох клітинних процесах, але найбільш вивченими є фосфорилюванням-опосередкована сигналізація і епігенетичний контроль експресії генів. Епігенетична контроль в цьому контексті описує модифікацію невпорядкованих хвостів гістонових білків на різних сайтах, що контролює стан хроматину та транскрипцію [PMC3193420].

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, декілька мотивів або партнерів взаємодії [24926813]. Наприклад, домен SH2 GRB2 зв'язується з залишком фосфотирозину тирозинкиназного рецептора після його фосфорилювання [8816475]. Часто, у нас немає достаточного 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 модифікації). Ці ферменти виконують обмежений протеоліз, порушуючи або іноді уможливлюючи функцію білка. Найбільш відомими мотивами цього класу є ті, що розпізнаються каспасами - основними регуляторами програмованої загибелі кліти, апоптозу [23545416], або реакції запалення в мієлойдних клітинах [26121197]. Апоптоз може бути ініційований імунними клітинами (Т-кіллерами або натуральними кілерами) зовні 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 [PMC3721276]. Ефекторні каспази розпізнають мотив [DSTE][[^]P][[^]DEWHFYC]D[GSAN], CLV_C14_Caspase3-7 в ELM [26615199], та викликають розщеплення сотен білків, що призводить до апоптозу. Еволюціонувавши цей мотив розпізнавання, білок може стати під контроль шляху апоптозу [24926813]. Ефекторна каспаза може мати як регуляторний (активацію ферменту розщеплення ДНК) знищуючий (розщеплення цитоскелетних білків) ефект на її субстрати.

Всі 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 [26581338]. Крім того, як це буде розглянуто в більш пізнньому розділі, однакова послідовність амінокислот може бути функціональним мотивом або ні залежно від доступності для зв'язування доменами розпізнавання.

1.2 Експресія білків, еволюція та сплайсинг 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 каталітичного домену, тобто збільшуючи локальну концентрацію субстрату і, таким чином, дозволяючи досягнення специфічності та селективності (ортогональності) сигнальної відповіді (декілька стимулів - одна кіназа - декілька 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) [18339942]. 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 додасть 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 білок 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 [14564014]. 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) [27540857]. Такі злиття генів у раку, як правило, порушує взаємодію білків з іншими молекулами: білками, РНК, ДНК. Невпорядковані ділянки білку, що містять лінійні мотиви і 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 хвороби людини, раку, аналогічні процеси можуть діяти, щоб змінити клітинні функції, на більш довгій еволюційній шкалі часу [18753782].

Наше розуміння ролі лінійних мотивів у функціональних інноваціях покращилося, у результаті аналізу взаємодій білків з різними анотованими функціями та як ці взаємодії розвиваються вздовж філогенетичного дерева [25299147]. Ми обговорюємо це в наступному пункті.

Спираючись на дані, Kim та ін визначають модулі взаємодій білків, які теоретично відповідають білковим комплексам. Взаємодії, опосередковані SLiM, більш ймовірно з'єднують білки між модулями з різною функцією, тоді як взаємодії доменів, більш ймовірно 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, опосередковані

домінами [25299147]. У незалежному дослідженні Hein та ін наклали 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 інші 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 мішенею ферментів [22749400].

1.3 Еволюція лінійних мотивів

Для розглядання еволюційних властивостей лінійних мотивів, краще протиставити їх до глобулярних доменів. Механізми еволюції доменів в значній мірі є загальноприйнятим знанням в той час як повне розуміння еволюції мотивів досі встановлюються. Домени еволюціонують шляхом дуплікації, дивергенції та рекомбінації [21286315]. 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 поділений їх спільним предком [26589632]. 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 декілька або навіть одну подію заміни амінокислот [24773235, 24926813]. Цей контекст забезпечує умови, необхідні для конвергентної еволюції SLiM.

Далі, давайте 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. Поширеною 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) [24882001]. 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 [15585523].

1.4 Проблеми відкриття лінійних мотивів

Незважаючи на те, що очікується, що мотивів багато, їх важко знайти. Низька складність, що підвищує їх функціональність та легкість еволюціонувати, 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 [1373379, 1846030]. Пізніше стало поширеною 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 інші

[26581338]. Та сама амінокислотна послідовність може бути функціональною залежно від структурного контексту. Наприклад, в гідрофобному ядрі глобулярних доменів часто можна знайти послідовність, що відповідає сигналу ядерного експорту, що має 4 гідрофобних залишків. Експериментальний мутагенез такого мотива в ядерному белку викликає його агрегацію, що перешкоджає його експорту з ядра, яке можна помилково вважати свідченням функціонального мотиву [23900254, 26581338]. Цей приклад підкреслює важливість пошуку мотивів в невпорядкованих ділянках.

Як продемонстровано в недавньому дослідженні Hagai та ін., низька складність мотивів є проблемою при передбаченні випадків відомих мотивів в усьому протеомі [24882001]. 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 всі дають додаткові докази для кожного конкретного мотиву [26581338, 25555723]. 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 [26581338, 18460207]. 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 [28805808].

Обчислювальне відкриття мотивів *de novo* спрямоване 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 [26581338, 25555723], 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 іншій стороні [21879107]. 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 [28514442] і розробили кращі інструменти для обчислювальних відкриттів [25792551] та високопродуктивних експериментів *in-vitro* [26297553], функціональна валідація залишається проблемою [26581338].

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 [24882001]. Щікаво, що прокаріотичні віруси часто не використовують мотиви, оскільки бактеріальні білки зазвичай мають менше мотивів [24882001]. Віруси людини не тільки мімікрують клітинний мотив, але й те ж дослідження показало, що ці мотиви, ймовірно, еволюціонували *ex-nihilo*.

У нашому дослідженні ми скористатися цією властивістю вірусних білків. Ми використовуємо експериментальні вірусно-людської дані взаємодії для визначення набору людських або вірусних білків (рисунок 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. [24234451] за допомогою функції `loadIntActFTP`, включеної в пакет `MItools` на мові програмування `R` [<https://github.com/vitkl/MItools>]. Записи в базі даних IntAct були очищені від тегів та текстового опису, щоб полегшити подальший аналіз, використовуючи функцію `cleanMITAB`. Ми використовуємо UniProt записи (`accessions`), щоб називати учасників взаємодії та відфільтрували тільки білок-білкові взаємодії. Спеціальний комп'ютерний код, що враховую 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 (март 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 [26096987].

На додаток до бази даних IntAct ми використовували дані проєкту BioPlex [28514442 та неопубліковані дані], що включає в себе 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. В 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 декілька 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 [26496610] та набір даних групи Vidal [25416956, невказано1304].

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 двохгібридного методу дріжджів [25416956].

Метод виявлення взаємодій: на основі методу виявлення взаємодії були створені дві підмножини бази даних IntAct: двогібридний та афінне очищення-мас-спектрометрія [26681426, 26496610]. 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 участника "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. Я завантажив набори даних взаємодій, як описано в розділі 2.1, використовуючи функцію loadHumanViralPPI. Я підрахував кількість взаємодіючих партнерів вірусних білків, людських білків або вірусних білків у кожній мережі або її підгрупі за методом або дослідженням (функція humanViralDegree). Я обчислив медіану кожного розподілу та візуалізував кожний розподіл, використовуючи розподіл

щільності (Малюнок 3.3.1 та Додатковий малюнок 1). Крім того, я розрахував кількість взаємодій та кількість білків у кожній мережі. Результати цього аналізу обговорюються в розділі 3.1.

2.3 Білкові послідовності і передбачення доменів

2.3.1 Білкові послідовності

Я використав Uniprot REST API (інтерфейс прикладного програмування) та Uniprot FTP [PMC5210571] для завантаження послідовностей у форматі FASTA (20 жовтня 2017 р.). Я використав 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 він 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 білку 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 [PMC5210578]. Ми запускаємо InterProScan версії 5.25-64.0 в автономному режимі на кластері обчислень LSF x86_64-pc-linux-gnu (64-розрядний, 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 [27899674], Coils-2.2.1, Gene3D-4.1.0 [PMC5210570], Hamap-201701.18 [25348399], MobiDBLite-1.0, Pfam-31.0

[26673716], PIRSF -3.02 [19455212], PRINTS-42.0 [https://doi.org/10.1002/047001153X.g306301.pub2], ProDom-2006.1 [12230033], ProSitePatterns-20.132 та ProSiteProfiles-20.132 [23161676], SFLD-2 [24271399].], SMART-7.1 [29040681], SUPERFAMILY-1,75 [11697912], TIGRFAM-15,0 [12520025]. Код командної строки, що було використано для запуску InterProScan, наданий у додатку Д.

Вивідна інформація була збережена в 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. Я завантажив анотації типу сигнатур для кожної сигнатури InterPro з InterPro FTP за допомогою функції `getInterProEntryTypes`. Я проаналізував вихідний файл InterPro за допомогою функції `readInterProGFF3`, об'єднав ці файли (функція `addInterProEntryTypes`) та вибрали 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 це, надаючи єдиний ідентифікатор InterPro (наприклад, IPR002048). Ми використали цей метод та функцію `collapseByInterProID` для збереження лише одного регіону домену на один блок 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'ект класу Granges у R) до таблиці домен-білкових пар. Ми називаємо це мережею домен-білок.

2.3.4 Посуднання 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 містять кожен домен (фонове число); фонову частоту домену; скільки людських білків є мішенями кожного вірусного білку; скільки людських взаємодій вірусного протеїну містять кожен домен (також називають доменним числом); яке зображення домену серед інтеракторів вірусного білку. 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 ймовірно опосередковують взаємодію

Edwards та ін продемонстрували, що справжні мотиви часто передбачаються неправильними даними про взаємодію, "off-target" мотиви [21879107]. 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, я вирахував ймовірність будь-якого домену бути присутнім N раз серед інтеракторів цього вірусного білка (рис 3.3, функція `permutationPval`) . Я обчислив це шляхом рандомізації людських мішень вірусних білків, зберігаючи ступінь та загальну кількість взаємодій незміненими; як якщо б вірусні білки вибирали людські білки незалежно від їх доменного складу. Для кожного вірусного білку я обчислив, скільки разів я бачив кожен домен. Тоді я розраховував, 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 нульової гіпотези. Фон за пермутації розраховується для кожного вірусного білка, щоб пояснити різну кількість взаємодій цих білків.

Прямо не включаючи фонову частоту домену в обчислення ми підвищуємо надійність збагачення домену порівняно з гіпергеометричним тестом. Edwards та ін. [21879107] 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 мішень навіть у кількості 1 або 2 через незначну кількість білкових взаємодій що білки, як правило, мають.

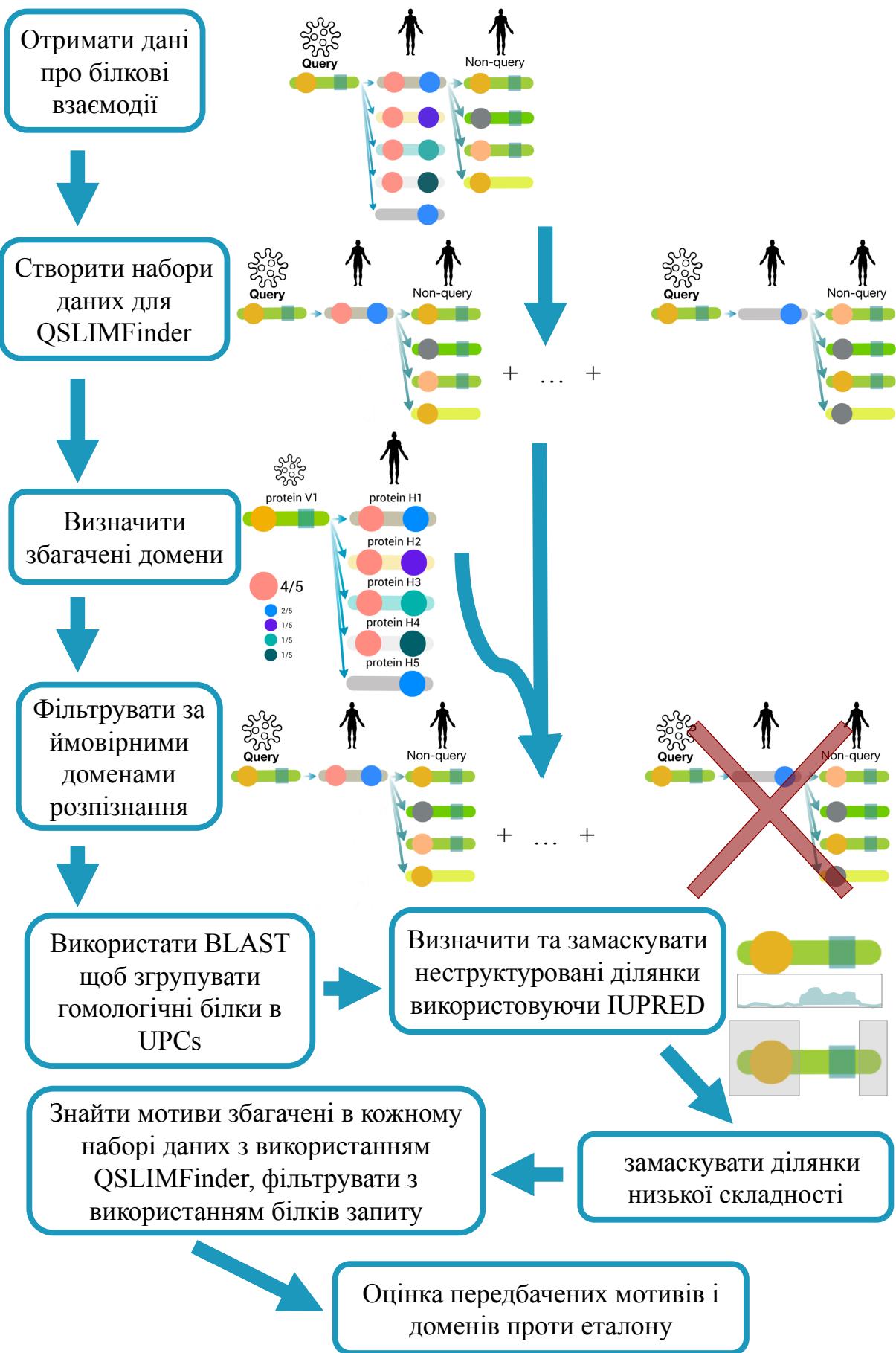


Рис 2.4. Схема процедури пошуку мотивів

2.6 Інструменти та процедура пошуку мотивів

QSLiMFinder [25792551] - це інструмент командного рядка, який потребує: послідовності білків як текстовий файл FASTA, назву білку запиту як текстового файлу, інші параметри, не специфічні для кожного набору даних. У нашому аналізі кожен набір даних визначається комбінацією білка людини - мішені віруса (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 Е 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 [17912346, 25555723]. Гомологічні послідовності, швидше за все, містять ті ж самі паттерни амінокислотної послідовності, і тому можуть штучно підсилювати підтримку для кожного мотиву. QSLIMFinder групував гомологічні послідовності з використанням NCBI BLAST 2.6.0 [20003500] 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. [15769473]. Я зкомпілював 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 мотива 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 білку 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. Цей 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 по послідовності білків (розділ 2.3.1).

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

спісок наборів даних для QSLIMFinder був відфільтрований, щоб включити ті, де query білок 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 забережені.

2.6.3 Процедура пошуку мотивів

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

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). Більш довгі мотиви можна виявити як набір з декількох коротших мотивів. Я обмежив 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 [17912346]. Коли ми додали ці мотиви, ми виявили більше вірних 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 схожі 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 аnotовані базі даних Eukaryotic Linear Motif (ELM) станом на листопад 2017 року [26615199]. 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 С).

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 відкликанням. 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, я завантажив дані набору збагачених доменів і набори даних мотивів, підготовлені для QSLIMFinder. Я необов'язково фільтрував обидва набори даних по ймовірності домену (розділ результатів 3.5-3.7). Далі я виділив *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. Я не об'єднував два мотиви, якщо тип мотива був іншим. Наступним кроком я розрахував 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.

Я використовував p-value для кожного унікального мотиву, як предиктор двоїчного результату: відповідність відому му істинному мотиву проти

помилкового позитивного або нового мотиву-кандидата. Кілька прогнозованих мотивів можуть відповідати одному відомому мотиву, наприклад, 3 варіанти мотива PDZ на рис 3.7.2.

Я проаналізував продуктивність на різних порогів, використовуючи пакет ROCR R та функцію mBenchmarkMotifsROC для організації моого аналізу. Я досліджував точність, відкликання, істинно позитивну швидкість, хибну позитивну швидкість при кількох порогах. Я використав цей аналіз, щоб вибрати три значення порогу p-value: м'який порог за 0,3; оптимальний порог при мінімальному р-значенні, коли точність більша, ніж відкликання (змінюється в різних наборів даних); і суворий порог, 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). Я оцінив, чи домени, які ймовірно опосередковують взаємодії для кожного вірусного білка, що містить мотив, є відомими SLIM-зв'язуючими доменами. Для того, щоб візуалізувати результати за допомогою Cytoscape, я перетворив результати порівняльного аналізу на направлену мережу: людський білок -> область визнання -> мотив -> вірусний білок. Для масштабування розміру вузла я використовував p-value мотиву та домену. Cytoscape-файл, що містить мережі, показані в розділі 3.7: https://github.com/vitkl/viral_project/blob/master/results/thesis%20example%20plots.cys.

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 [18375965]. Складність мотивів можна описати, використовуючи інформаційний вміст (IC). IC описує, наскільки зменшення невизначеності забезпечується мотивом. Я запускаю Comparimotif як інструмент командного рядка, включений в SlimSuite (обговорюється в розділі 2.6.1) з налаштуваннями за замовчуванням. Результат з цієї програми завантажено в Cytoscape. Я використовував евристичну оцінку 1,162 (кількість співпадаючих позицій х Нормалізований IC) для фільтрування мережі подібності мотивів [https://github.com/vitkl/viral_project/blob/master/qslimfinder.Full_IntAct3.FALSE/result/comparimotif.compare.cys].

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 так методи описані у відповідних розділах чи оригінальних публікаціях. Де була потреба, власні методи були

запрограмовані і були включені до пакету R під назвою MItools [<https://github.com/vitkl/MItools>]. 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 пакеті R під назвою viral_project [https://github.com/vitkl/viral_project]. 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 системому рівні, ми розглянули тенденції щодо кількості взаємодій 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).

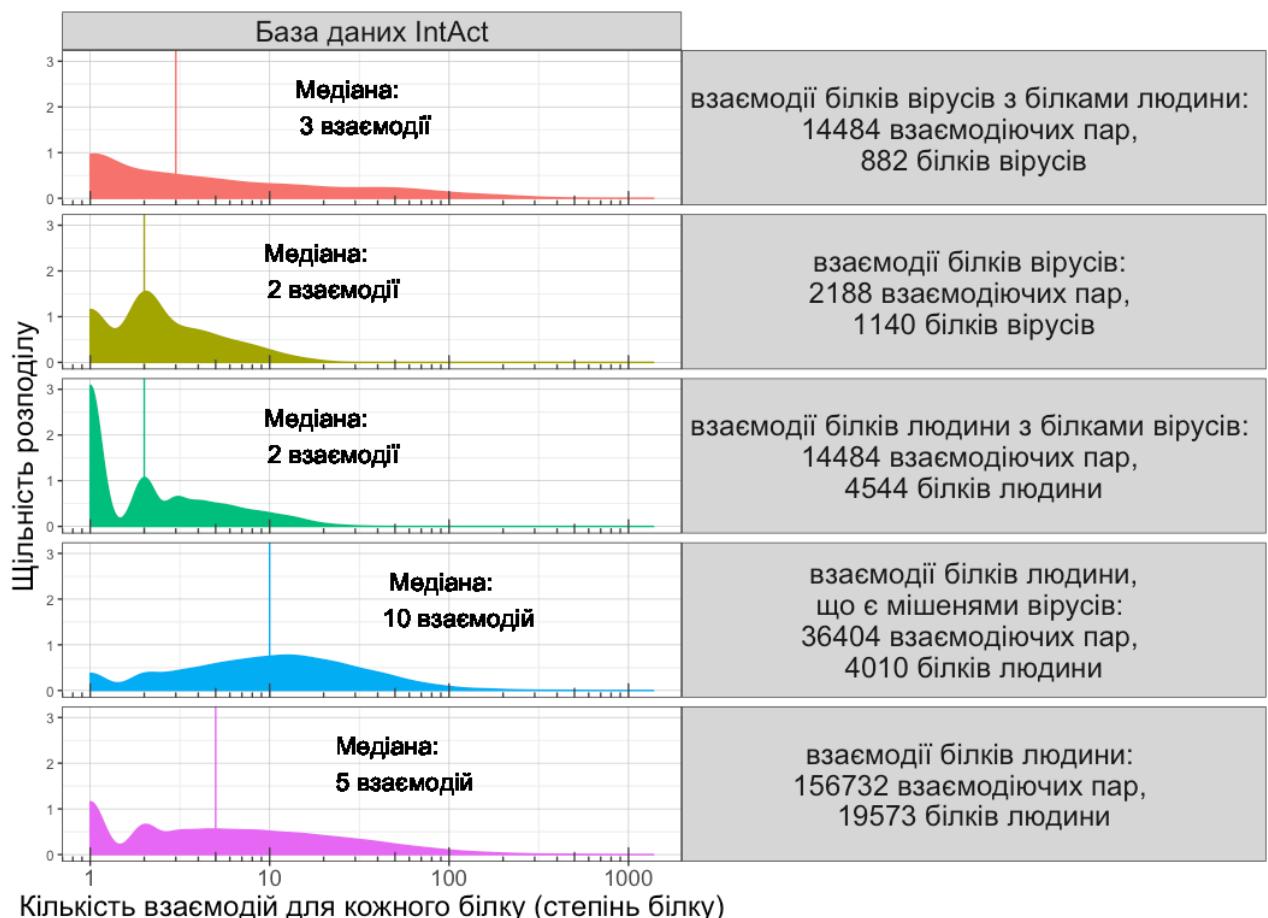


Рис 3.1.1. Графік, що показує щільність розподілу числа взаємодій кожного людського чи вірусного протеїну в кожній мережі, яку ми використовували для нашого аналізу. Для кожного білку ось X показує

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 вірусно-людській мережі для вірусних та людських білків відповідно.

Як вірусно-людська мережа, так і людська мережа є найбільшими серед тих, що були аби-коли 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 недавня розробка спеціалізованого інструменту Palopoli та співавт. [25792551] мотивували наш підхід до пошуку мотивів у людській мережі, використовуючи послідовності кожного вірусного білка як фільтр, а не шукати 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 взаємодії протеїнів. Додаток 2 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 хаби людської мережі (білки з багатьма взаємодіями) [PMC3593624, 25417202]. Проте останнім часом з'явилися декілька 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 [26300911]. Ця упередженість означає, що краще вивчені білки мають більше взаємодій, що перешкоджає корисності ступеня білку як міру функціональної важливості білка. Для вирішення цієї проблеми та поліпшення покриття мереж білкової взаємодії проводиться кілька систематичних досліджень взаємодій між ~ 17000 білками людини [26496610, 28514442, 25416956]. 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, мішені вірусів, як правило, мають набагато більше взаємодіючих партнерів (медіана 10 у порівнянні зі середнім значенням 5 для білків, не мішеней вірусів). Це може бути пояснено функціональною різницею та поведінкою вірусів, атакувати білки-хаби, 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 частину вищий ступінь. Інша можливість полягає в тому, що білки-мішенні вірусів, більш вивчені в цілому.

На малюнку 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 упередженості 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. Ця тенденція вірусних білків взаємодіяти з найбільш вивченими людськими білками може бути особливо сильною, оскільки багато з цих найбільш вивчених білків (у тому числі Р53

[218111]) були вперше виявлені через їх вірусні взаємодії (основний метод відкриття людських білків перед тим як першим геном людини був секвенований).

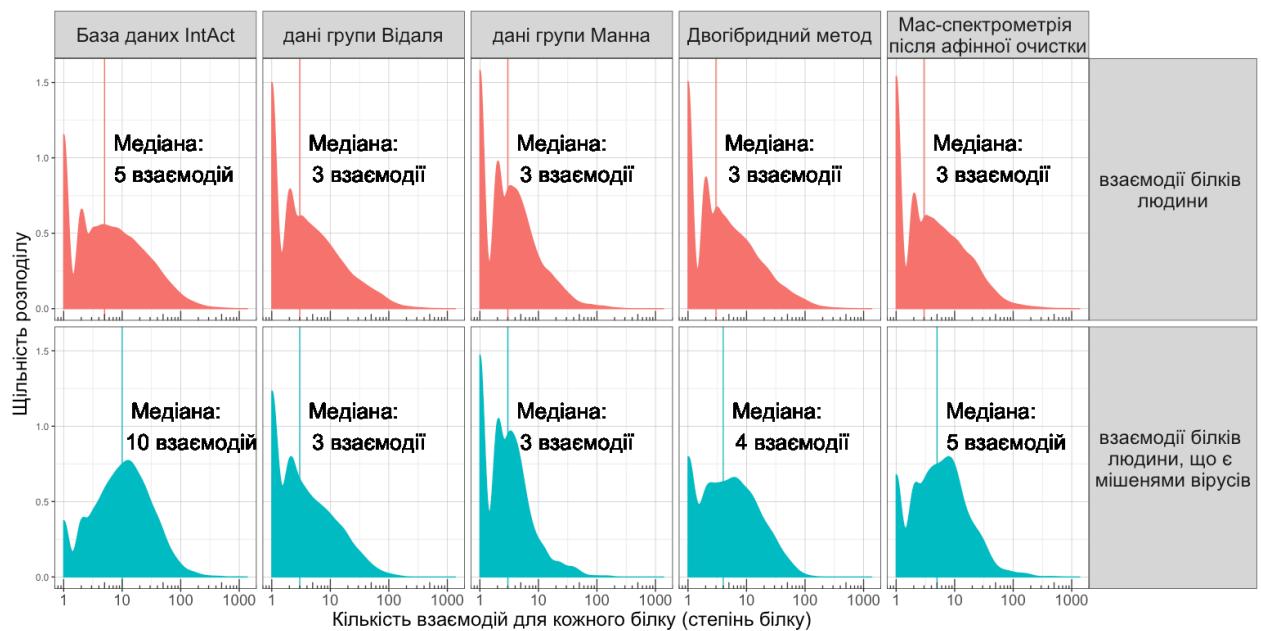


Рис 3.1.2. Графік, що показує щільність розподілу кількості взаємодій для всіх білків людини або білків людини-мішеней вірусів, для всіх даних, або обраних методів виявлення взаємодій протеїнів та широкомасштабних неупереджених досліджень. Логарифмічно трансформована ось X показує кількість білків, що взаємодіють з кожним білком (степінь білку), вісь Y показує щільність розподілу.

3.3 Дослідження доменів, що ймовірно 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, зображені у білках людини, які взаємодіють з кожним вірусним білком.

Хоча, загальноприйнято передбачати взаємодій між доменами використовуючи дані про взаємодію з білками [28514442], 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 [25555723, 25207816].

Я пропоную, 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-value $<2.2\text{e-}16$). Відомі домени що розпізнають

SLIM мають переважно низькі p-value (рис. 3.3 В). Це дозволяє нам використовувати збагачення домену як проксі для визначення домену, який, ймовірно, опосередковує взаємодію (включаючи мотив-опосередковану взаємодію).

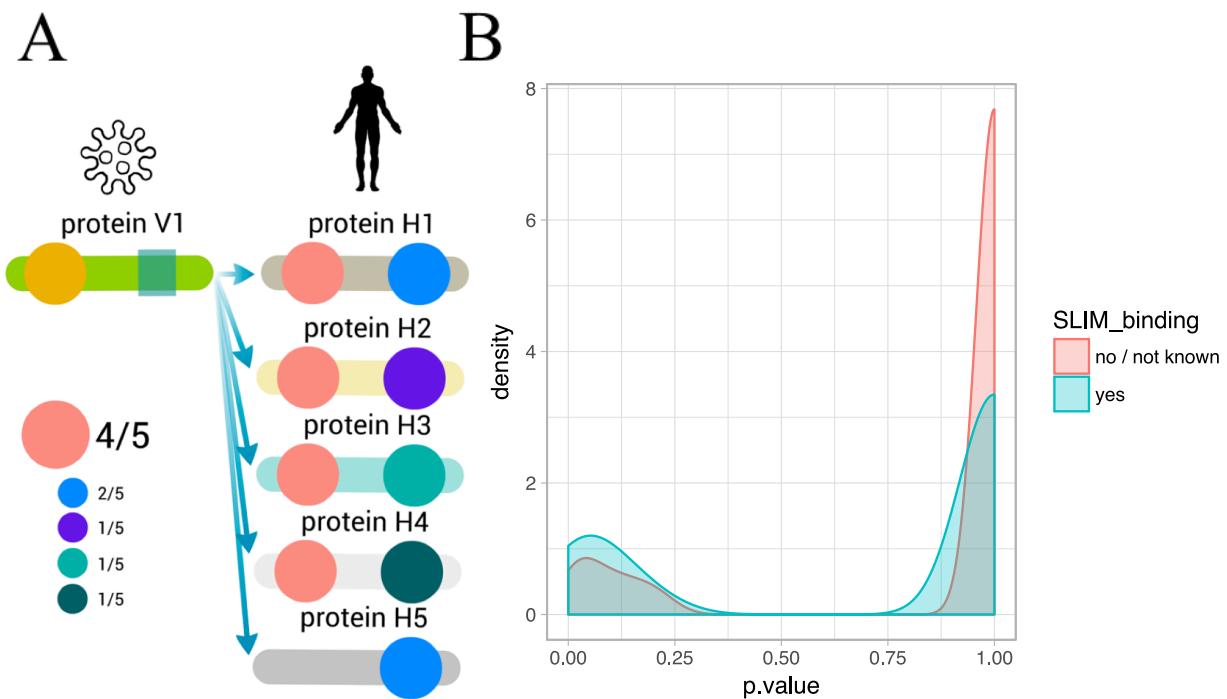


Рис 3.3. А. Схема, що ілюструє, як визначено збагачені домени. Ми знаходимо домени в людських білках, що є мішенями кожного вірусного білку. Ми підраховуємо, скільки разів спостерігається кожен домен. Далі ми використовуємо підхід на основі перестановок для обчислення емпіричного значення p-value для будь-якого домену, що з'являється багато разів у білках, обраних як мішень білком V1. В. Щільність розподілу емпіричних значень p-value для відомих доменів, що зв'язують SLIM або для всіх інших доменів. Ось X показує значення p-value; Ось Y показує щільність розподілу. SLIM-зв'язуючі домени, як правило, мають більше низьких значень p-value, ніж всі інші домени.

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 ймовірно 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 доменів 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 Edwards і ін (QSLIMFinder) [PMC4495300]. 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 мотива 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 мотива.

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 [24234451].

Ми також використовували дані про взаємодію білків з великого неупередженого скріну, зробленого групою Vidal [25416956] (додаток 3). 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 [21879107, 20055997]. 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.

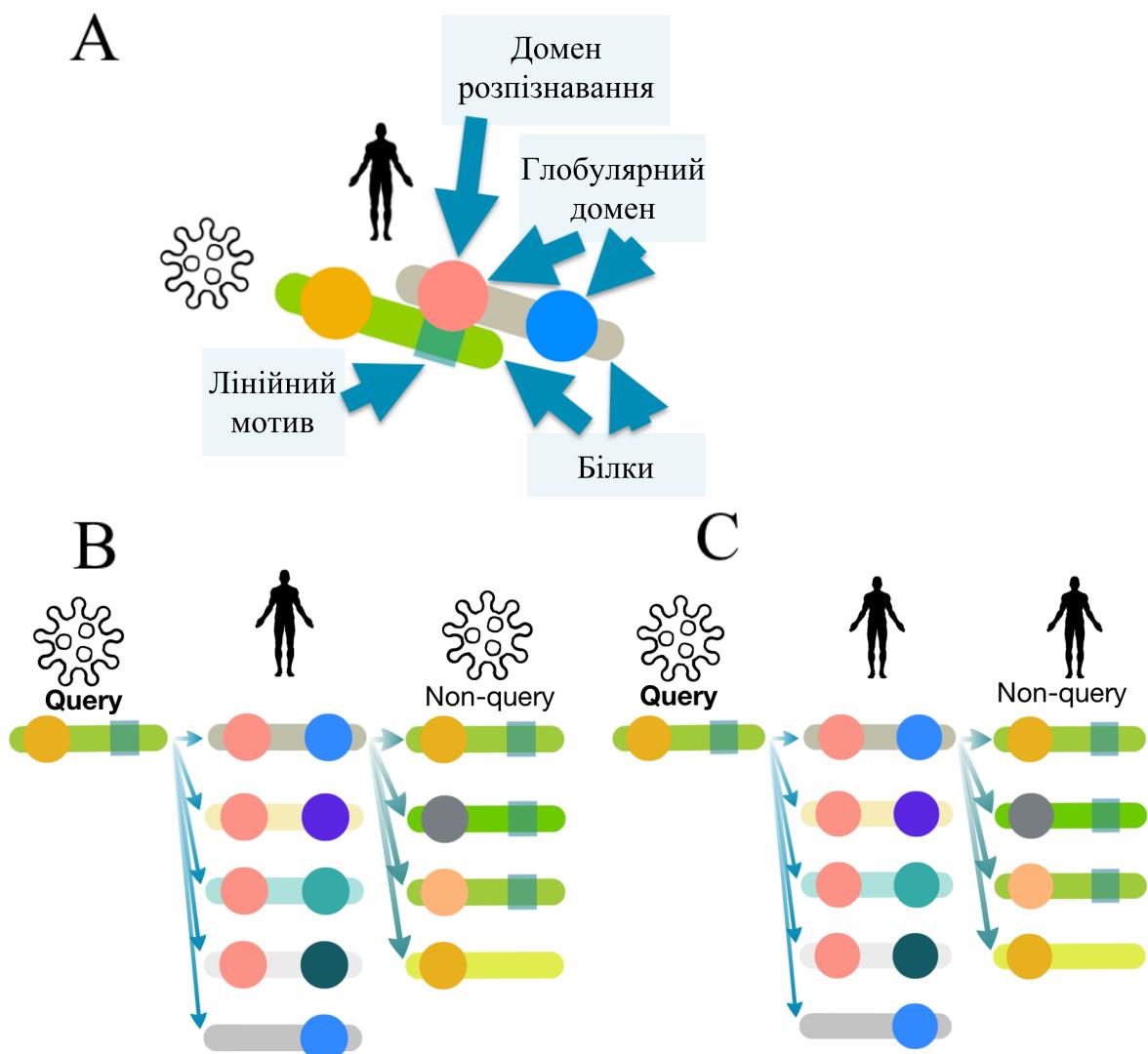


Рис 3.4.1. Схема, яка показує, як будуються набори даних для пошуку мотивів. Білки не запиту (non-query) використовувались для пошуку мотивів, які повинні бути присутніми в білку запиту (query). Кожен набір даних складається з усіх взаємодіючих партнерів одного білка людини та одного білка запиту (query). 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.

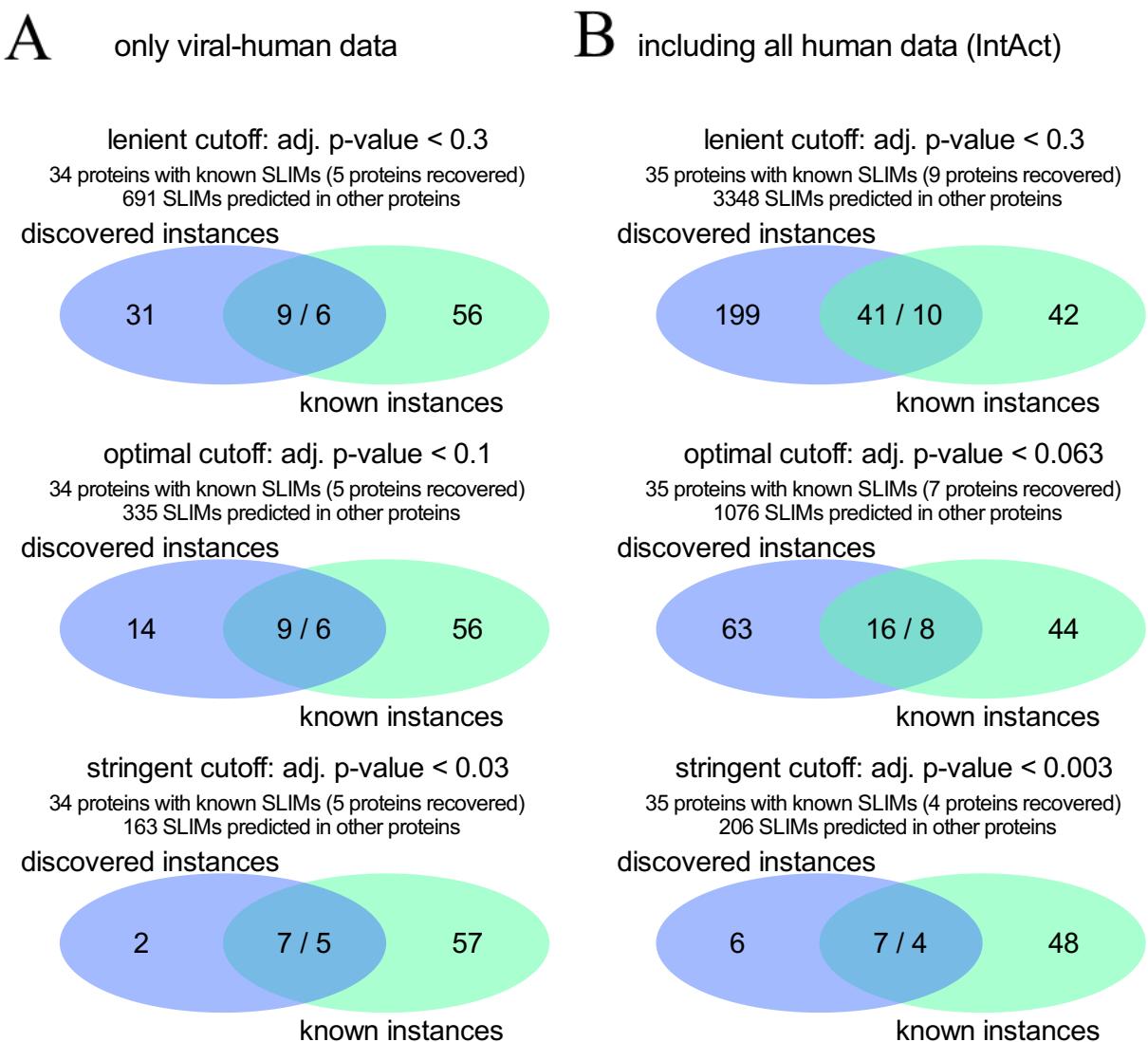


Рис 3.4.2. Діаграми Вена, що показують кількість виявлених мотивів-кандидатів та відомі мотиви ми відкрили заново. Наведено дві стратегії побудови наборів даних (A та B) та 3 пороги значень p-value. Блакитне коло показує кількість екземплярів мотивів, передбачених, але невідомих. Зелене коло показує 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 мотива N-глікозилування (MOD_N-GLC_1) [26615199]. Вірусні білки містять не більше 3 відомих мотивів різних класів (ранній блок Е1А людського аденовірусу С, P03255) [26615199]. 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 ретинобластому білок (LIG_Rb_LxCxE_1), у білку E7 людського віrusу папіломи та раннього E1A-протеїну людського аденоvіrusу С [26615199]. Використовуючи лише вірусний набір даних, ми пропустили відомий мотив - сигналу ядерної локалізації в основному білковому полімерази 2 білка віrusу грипу А та фрагмент відомого мотива, що зв'язує домен PDZ, в протеїні Е6 людського папіломавіrusу [26615199]. 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. [25416956, та неопубліковані дані], 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 3). 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 [28514442], яка потенційно зберігає взаємодію з опосередкованим мотивом краще, ніж двогібридні скріни Vidal. У розділах 3.6 та 3.7 я зосереджуся 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.

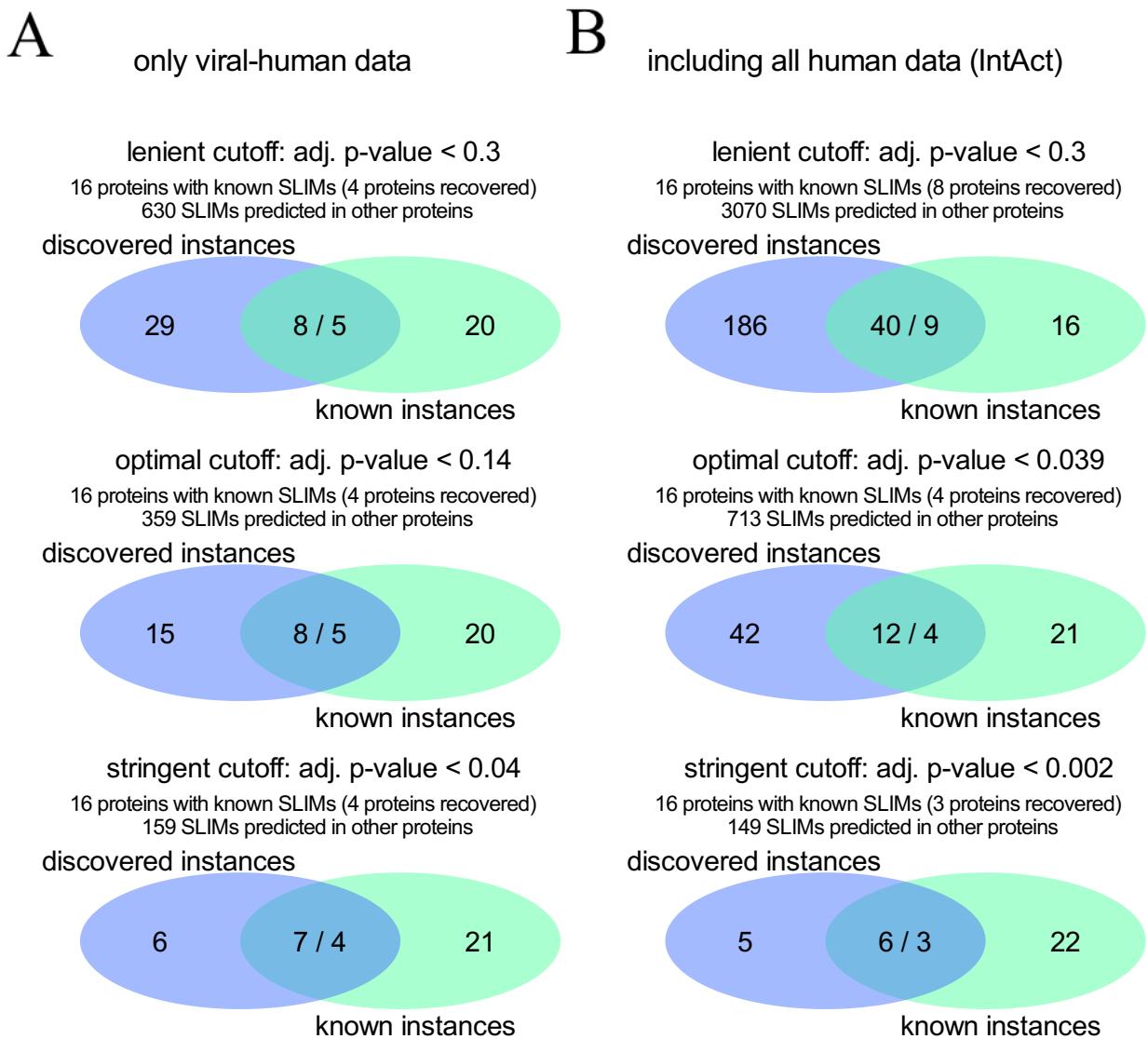


Рис 3.5. Діаграми Вена, що показують кількість знайдених мотивів-кандидатів та відкритих заново відомих мотивів при фільтруванні по домену. Наведено дві стратегії побудови наборів даних (A та B) та пороги 3 значень p-value. Синє коло показує кількість екземплярів мотивів, передбачених, але невідомих. 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). 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.

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 скupчення 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.

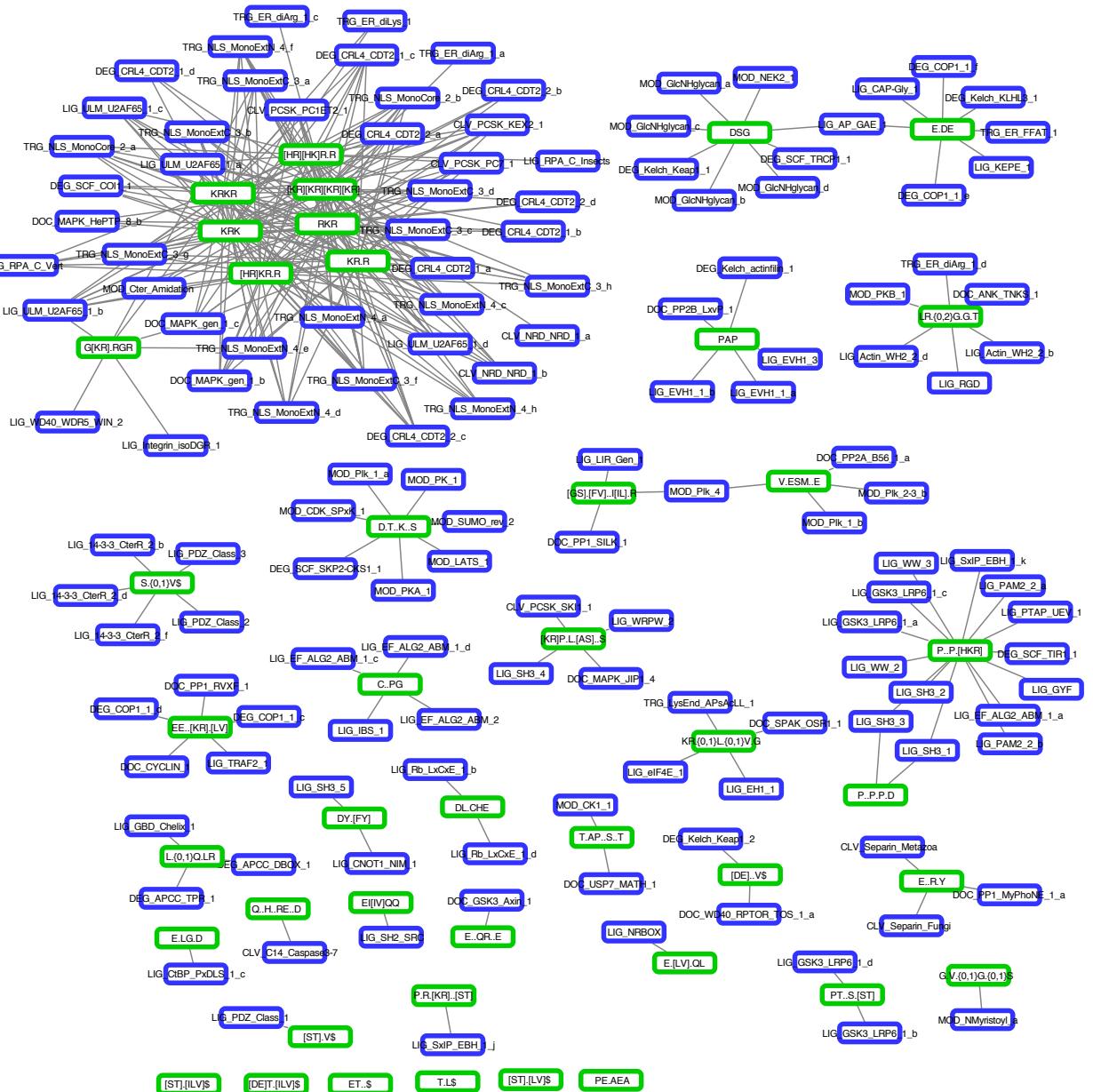


Рис 3.6.1. Схема, що показує подібність виявлених коротких лінійних мотивів до відомих мотивів. Зелені вузли - це послідовності виявлених мотивів, сині вузли - це відомі мотиви в базі даних ELM, ребра показують наявність подібності між мотивами вище порога 1,162 (Score, Comparomtif3).

3.7 Приклади відкритих заново та мотивів-кандидатів

3.7.1 Дослідження класів мотивів, що відкрито заново, мотивів-кандидатів та їх ймовірних доменів розпізнавання

Для вивчення хітів, отриманих за допомогою комбінації людсько-вірусної та повної мережі людини (IntAct), я вибрал 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 еволюціонував у вірусному білку, він прогнозує мотиви в протеїнах людини, які зв'язують один і теж ж домен розпізнавання. Я розглянув 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 мотивами, 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 декількох 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 декількох 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 Мотивів PDZ, які відкрито 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 мотива білку 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 мотива 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 [PMC4970744]. HPV 16 і HPV 18 E6 націлені на найбільшу кількість білків людини. Деякі людські білки, такі як DLG1, звязуються всіма типами білків HPV E6. Мішені білків E6, у тому числі DLG1, PTPN3 та SCRIB, зазвичай убіквітинуються і деградуються протеасомою [PMC2748042, 10523825, PMC1865939]. E6 білки виконують це, діючи як білки-скафолди, щоб залучають E3 убіквітин-лігазу E6-AP (UBE3A) через мотив LXXLL, розташований в лігазі [17023019, PMC1865939]. 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 [15371341]. Інші дослідження також показують, що E6 має більш складні зв'язки з його мішенями: білок SCRIB позитивно регулює транскрипцію та швидкість трансляції білка E6 [29074188]. 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.

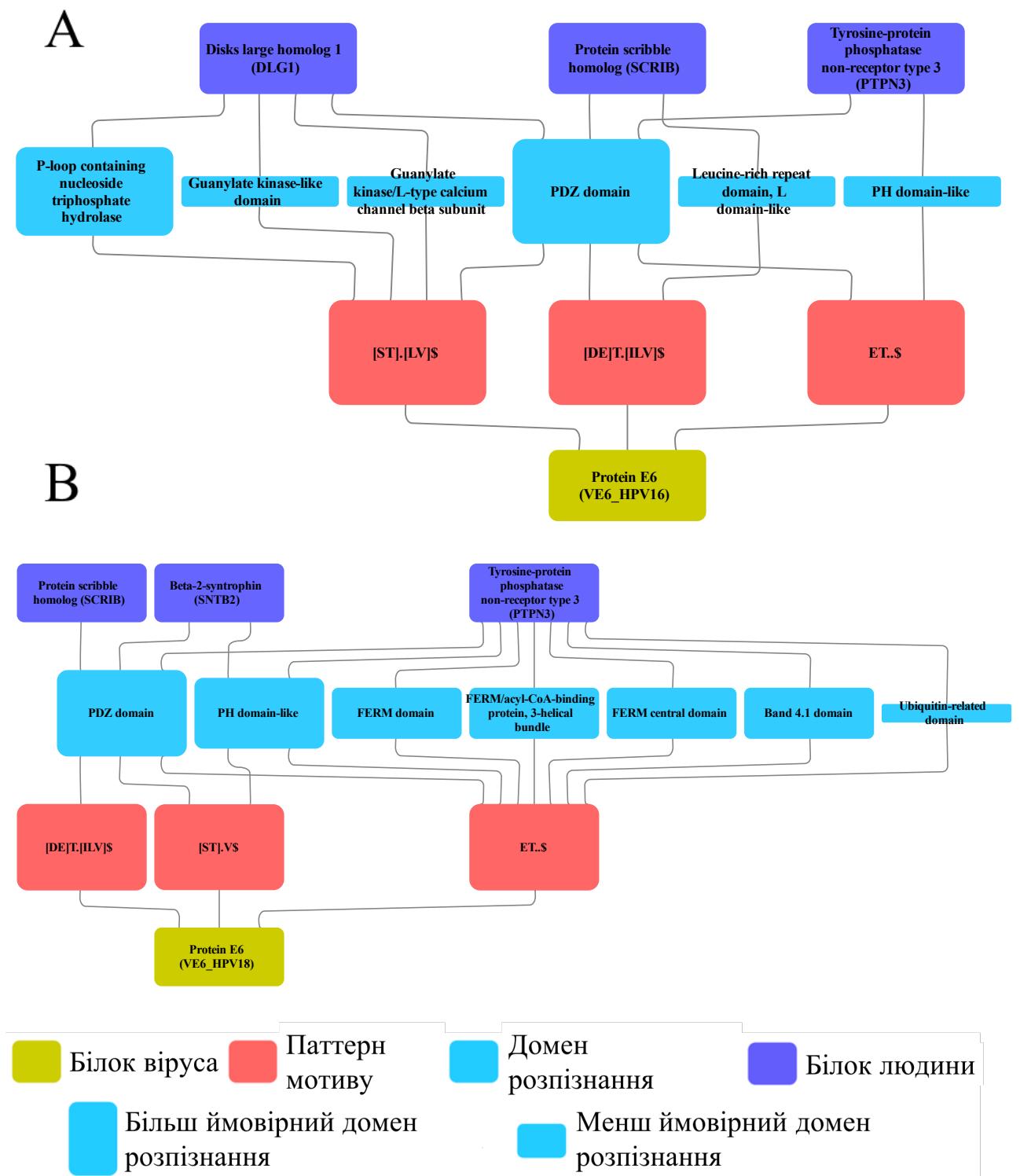


Рис 3.7.2. Схема мережі, що показує відомі мотиви PDZ в білку E6 людського папіломавірусу 16 та 18. Ми показуємо варіанти мотива та ті домени в білках людини, які можуть бути відповідальними за зв'язування

протеїну E6. Три варіанти цього мотива 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 E6 і є доменом, що опосередковує взаємодію з цим відомим мотивом. А. Мережа цього мотива в людському папіломавірусі 16. Б. Мережа цього мотива 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 E6, 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 E6 зв'язує меншу кількість білків, ніж HPV-16 або HPV-18, про які йшлося раніше [PMC4970744]. Згідно з їх результатами, мотив HPV-70 E6 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.) [PMC4970744]. Це може вказувати, що взаємодія білка E8 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 мотива PDZ в білку HPV-70 E6, але воно вказує на

потенційну невідповідність в даних про білкову взаємодію, які використовуються для нашого дослідження.

З точки зору функції, ERBIN слугить як адаптерний протеїн, який зв'язує нефосфорильований receptor ERBB2, тим самим стабілізуючи цей стан [16203728]. Він є важливим для локалізації ERBB2 на базолатеральну сторону епітеліальних клітин [10878805]. З огляду на те, що HPV також зв'язує й інші білки, пов'язані з апікально-базальної полярністю клітини, такими як DLG1 та SCRIB, що було обговорено раніше, ERBIN може представляти реальну мішень.

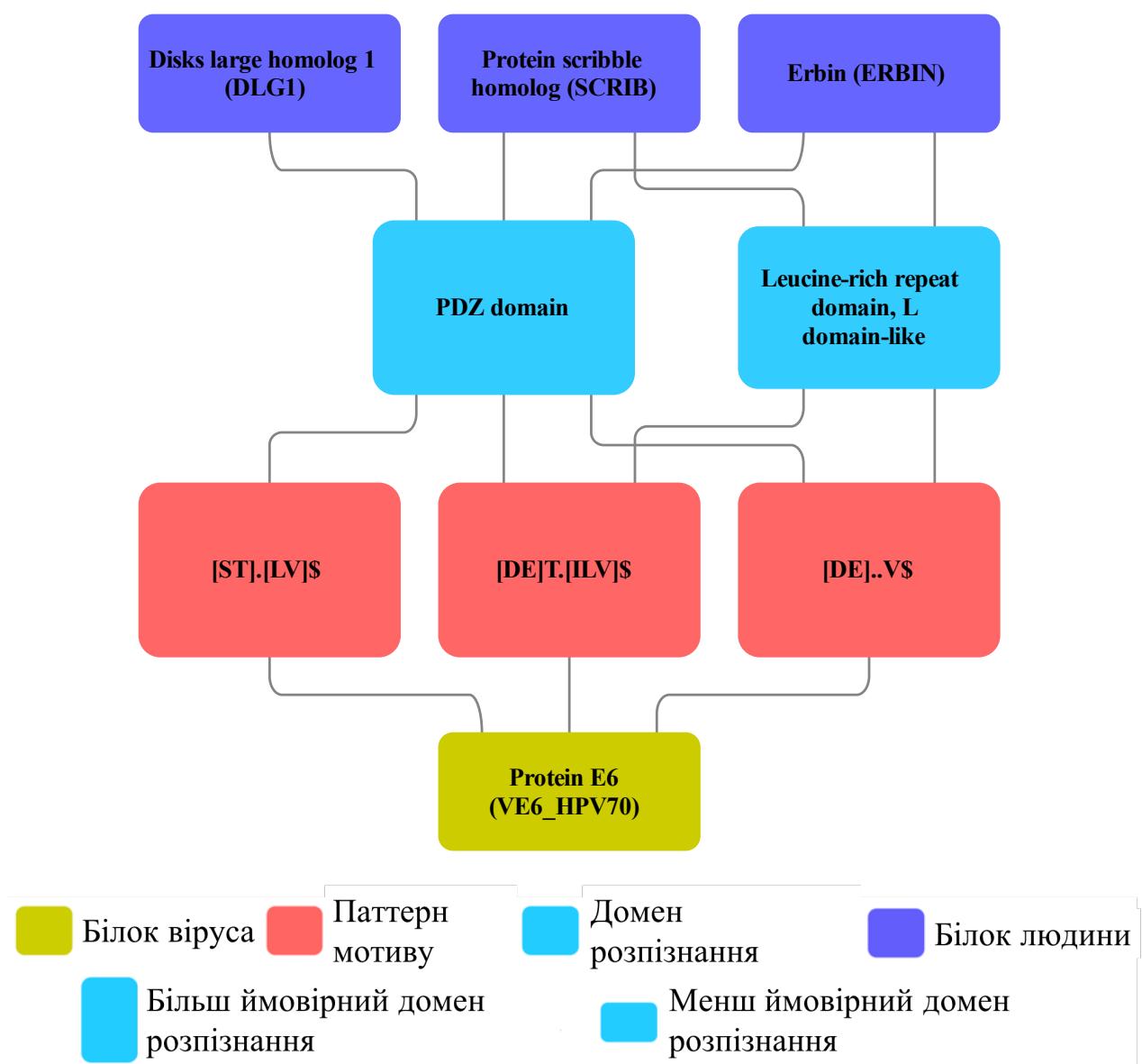


Рис 3.7.3.1. Схема мережі, що показує мотиви-кандидати PDZ в білках E6 людського папіломавірусу 70 були підтвердженні в попередньому дослідженні,

але не анотовані в ELM. Ми показуємо варіанти мотивів і ті домени в білках людини, які можуть бути відповідальними за зв'язування протеїну Е6. Три варіанти цього мотива були передбачені у білку Е6 і 74 білках людини, які взаємодіють з білками-мішенями вірусів DLG1, SCRIB та ERBIN. домен PDZ є найбільш збагаченим серед мішень білка Е6.

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 [21849460]. Паттерн мотиву, що ми ідентифікуємо, загально схожий на паттерн високо патогенного пташиного (RS.V), але не людського (ES.V) вірусів грипу А [21247458]. Вірус H5N1 викрадає проапоптичну функцію SCRIB, використовуючи домен PDZ мотив, для зміни його субклітинної локалізації [PMC2953166]. 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.

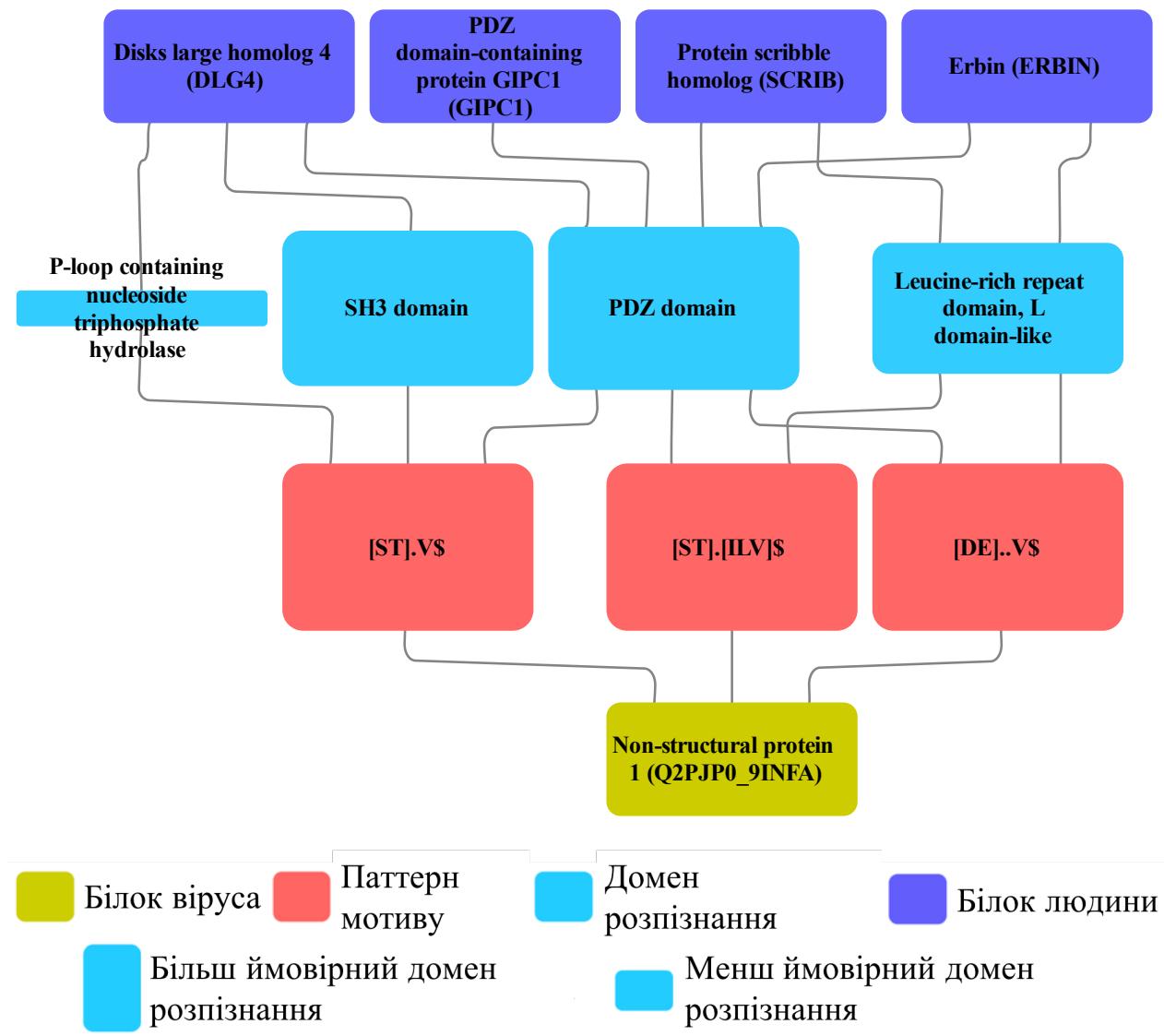


Рис 3.7.3.2. Схема мережі, що показує мотиви-кандидати PDZ в неструктурному білку 1 (NS1) вірусу H5N1 грипу були підтвердженні в попередньому дослідженні, але не були анотовані в ELM. Ми показуємо варіанти мотивів і ті домени в білках людини, які можуть бути відповідальними за зв'язування NS1. Три варіанти цього мотива 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 *de novo* виявили екземпляр домен SH3-зв'язуючого мотиву 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 P..P, нагадує канонічний ліганд домену SH3 [7953536]. 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 P..P.[HKR] дозволяє Nef зв'язати домен SH3 з сімейства SRC-кіназ для їх активації та сприяння вірусній патогенності [PMC398106, 16849330].

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

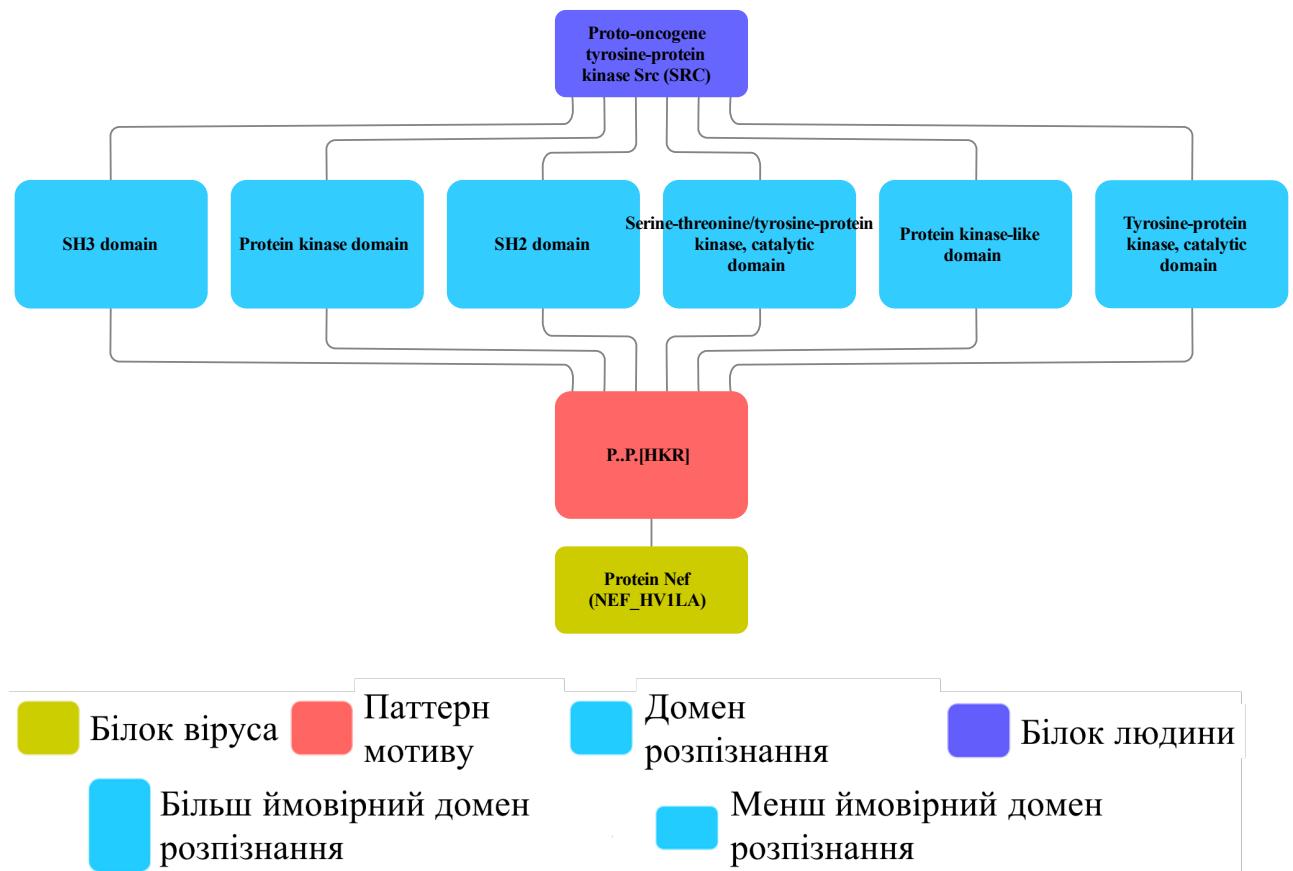


Рис 3.7.4. Схема мережі, що показує мотив-кандидат SH3 в білку Nef вірусу Імунодефіциту людини типу 1 був підтверджений у попередньому дослідженні, але не був анотований в ELM. Ми показуємо ті домени в білках людини, які можуть бути відповідальними за зв'язування Nef. Один із варіантів

цього мотива 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 мітить білки для протеасомної деградації [10648623, 10066435]. SCF (комплекс FBXW11 або BTRC) є частиною сигнальних шляхів, включаючи шлях Wnt-beta-catenin і NF-kappaB, де він мітить або бета-катенін (fosфорилюваний за допомогою GSK3beta), або IкappaB для деградації. У свою чергу, це пригнічує (бета-катенін) або активує (NF-kappaB) транскрипційний фактор в кінці шляху [10321728, 10437795].

Vpu має відомий екземпляр мотива DSG..S, який дозволяє ВІЛ викрасти SCF убіквітин лігазу для деградації білків хозяїна, таких як противірусний фактор тетерін/BST-2 і CD4 [PMC2729927, 9660940]. 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 [9660940]. NSP1 ротавірусу А також має відомий мотив DSG..S [28851847]. Цей білок використовує убіквітин лігази хозяїна для деградації ключових факторів, що активують вироблення інтерферону, таких як IRF3, IRF5 або IRF7 [27009959, 17251580]. Інтерферон

звичайно виробляється у відповідь на вірусну інфекцію та допомагає обмежити інфекцію до сусідніх клітин [24751921]. 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) [PMC353757]. Регулятор P53 убіквітин лігаза MDM2 містить відомий мотив DSG і сам деградується β -TRCP1/2, що обговорювалося раніше [PMC3494375]. 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 [PMC3017642].

Щоб краще зрозуміти структурний аспект цієї взаємодії, я виконав докінг трьох пептидів з β -TRCP1 / FBXW11, використовуючи PepSite 2 (PDB 1P22, ланцюг А). Цей аналіз показує, що короткий мотив DSG, який ми передбачаємо, 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.

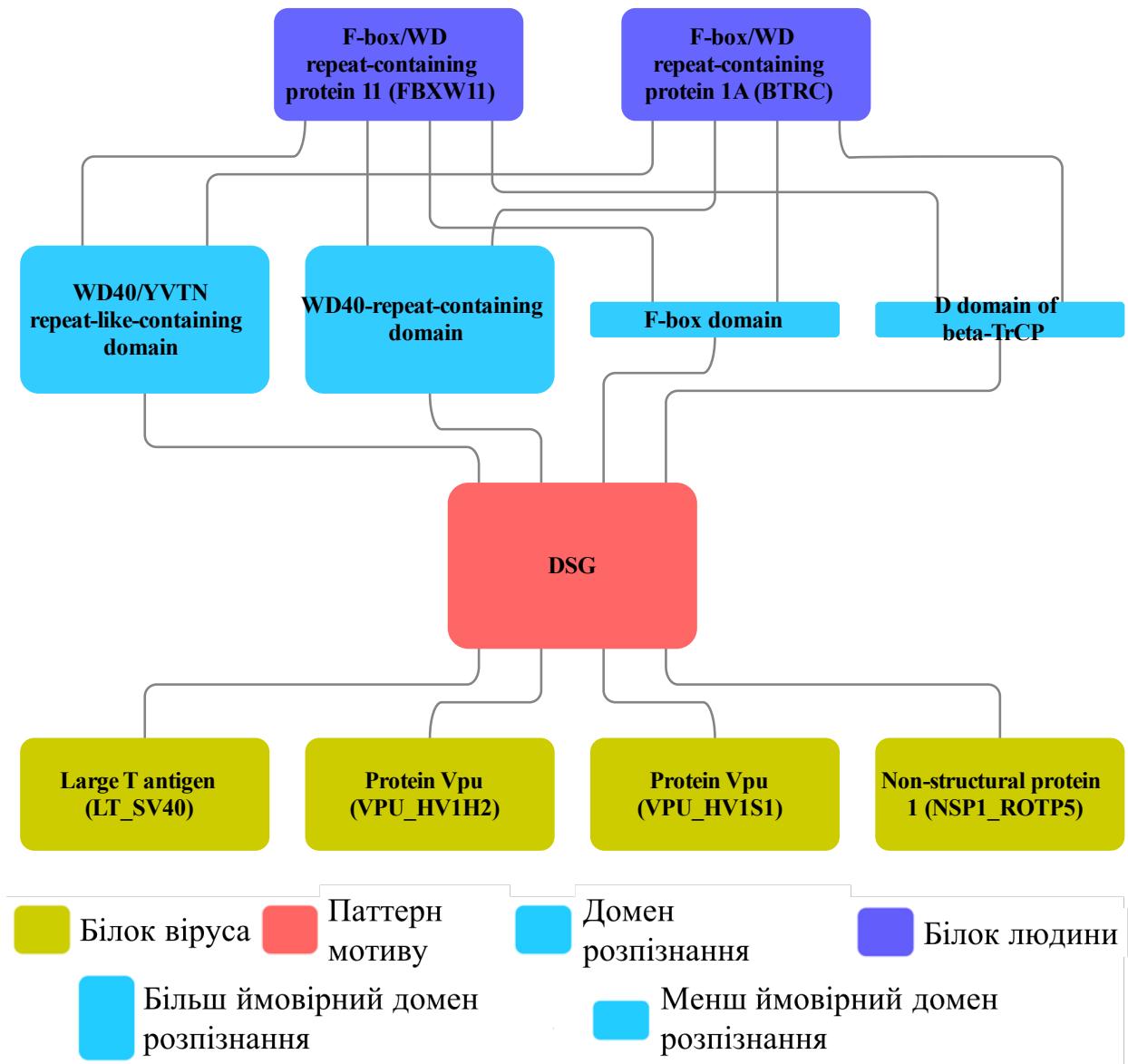


Рис 3.7.5.1. Схема мережі, що показує мотиви-кандидати DSG. Ці мотиви були передбачені в 4 вірусних білках. Всі вони обрали як мішень 2 субодиниці розпізнавання субстрату у комплексі SCF Е3 убіквітин лігази людини, FBXW11 та BTRC. 3 екземпляри мотиву були підтвердженні в попередньому дослідженні, але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

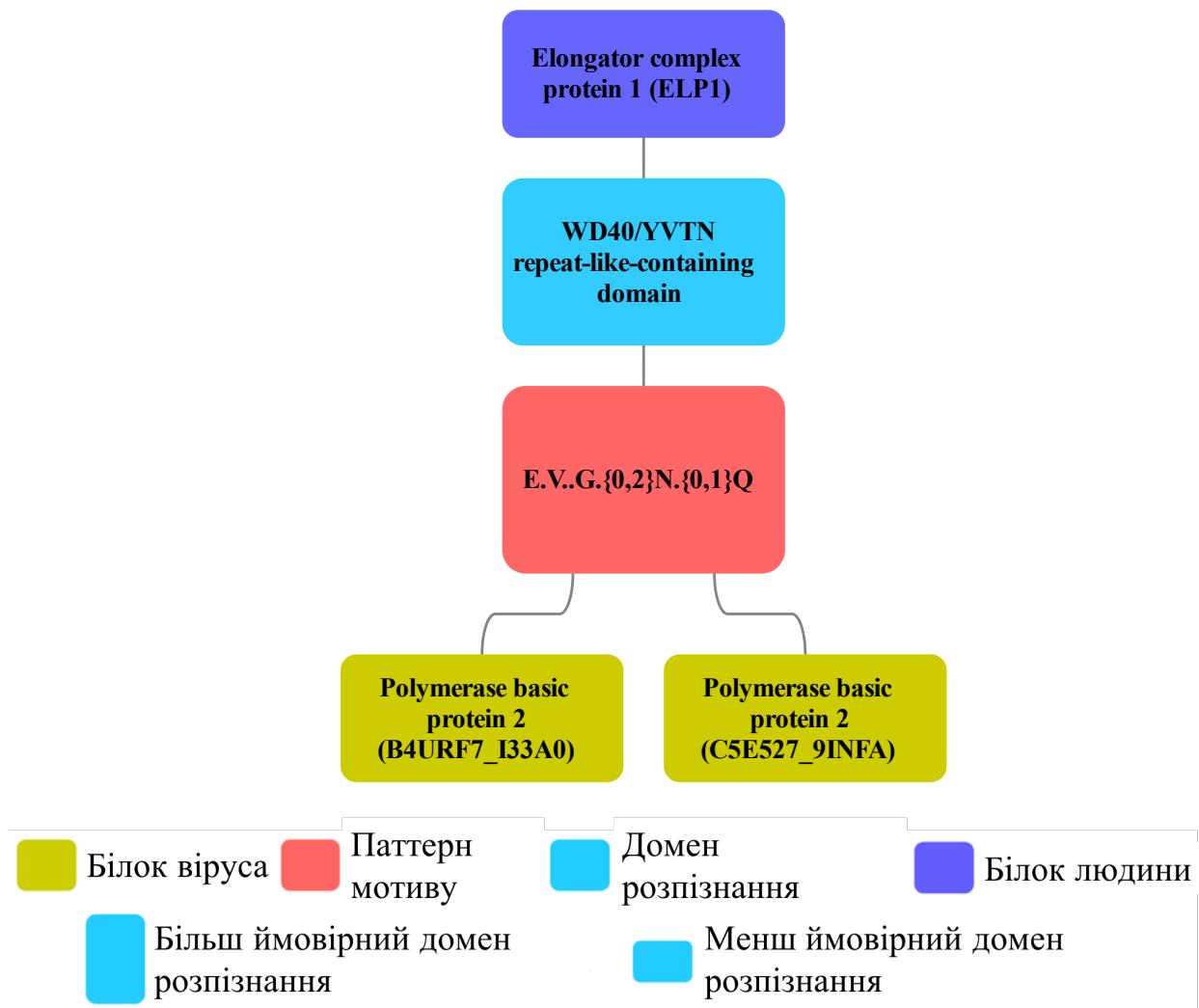


Рис 3.7.5.2. Схема мережі, що показує мотив-кандидат E.V..G.{0,2}N.{0,1}Q передбачений в полімеразному лужному білку 2 у 2-х штамів грипу А. Ми передбачаємо, що цей мотив розпізнається доменом WD40 в людському протеїні ELP1.

3.7.6 Мотиви-кандидати, що розпізнаються доменом, що зв'язує дволанцюгову 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 [8525619], а також інгібують TRIM25-опосередковане убіквітинування, що є частиною антивірусної відповіді [23209422]. Ми припускаємо, що цей мотив імітує РНК, яку цей домен людини розпізнає.

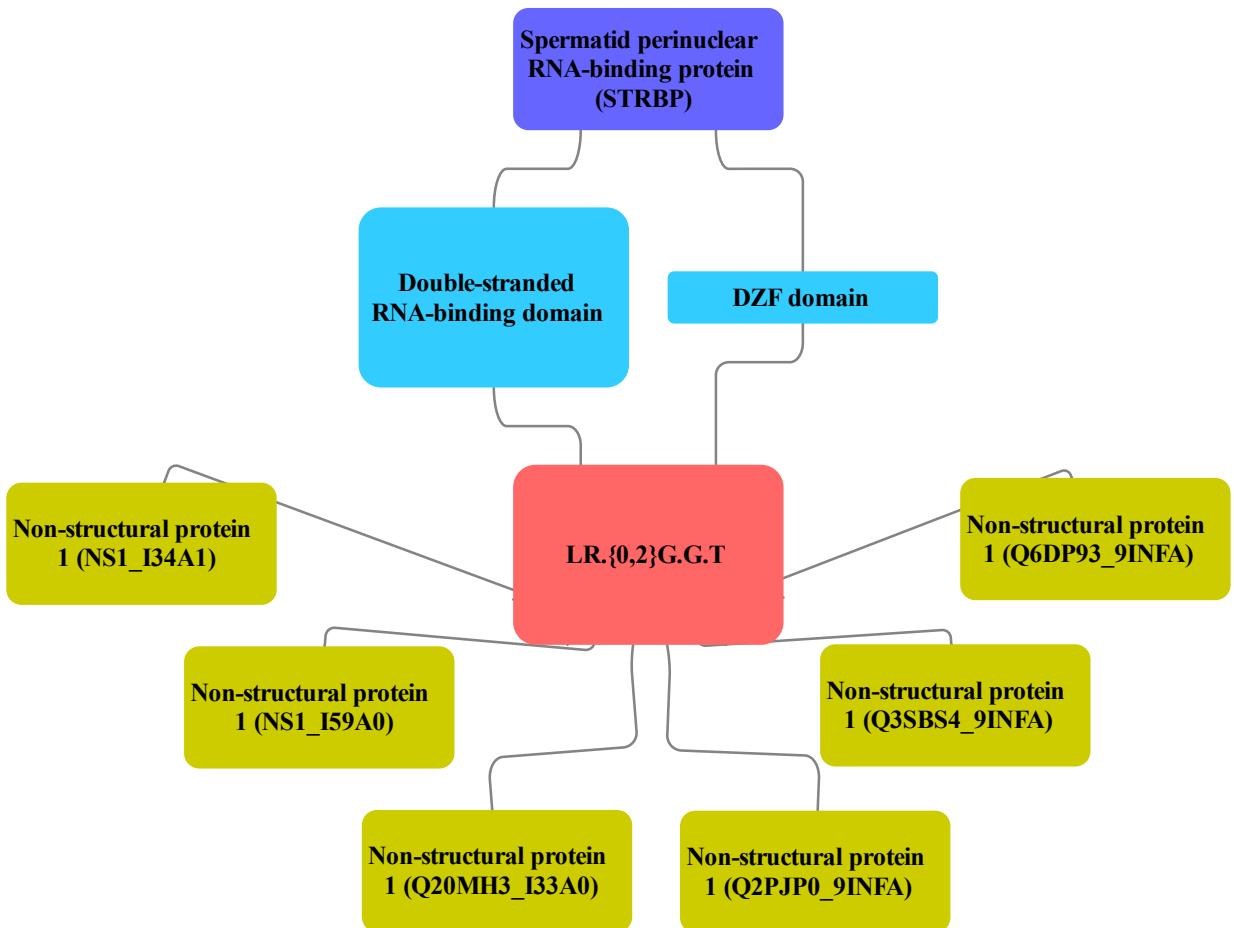




Рис 3.7.6.1. Схема мережі, що показує мотив-кандидат $LR.\{0,2\}G.G.T$ передбачений в 6 неструктурних білках різних штамів грипу А, що включають як пташину, так і людську лінію. Ми передбачаємо, що цей мотив розпізається дволанцюговим РНК-зв'язуючим доменом людського білка STRBP.

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 [27138253] і відправляють їх на плазматичну мемрану / позаклітинний простір. Цей білок, як відомо, не взаємодіє з вірусами, окрім недавніх високопродуктивних робіт, що проаналізували інтерактоми декількох штамів віrusу грипу А [28169297]. 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 віrusу грипу та є важливим компонентом механізму транскрипції віrusу [23600869]. Білки РНК-полімерази (включаючи PB1) залишаються пов'язаними з вірусною РНК та упаковуються у вірусні частинки. 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 PB1-зв'язуючих білків (рис. 13). Дослідження 3 білків людини, в яких передбачено цей мотив, може прояснити, наскільки цей мотив може бути справжнім хітом.

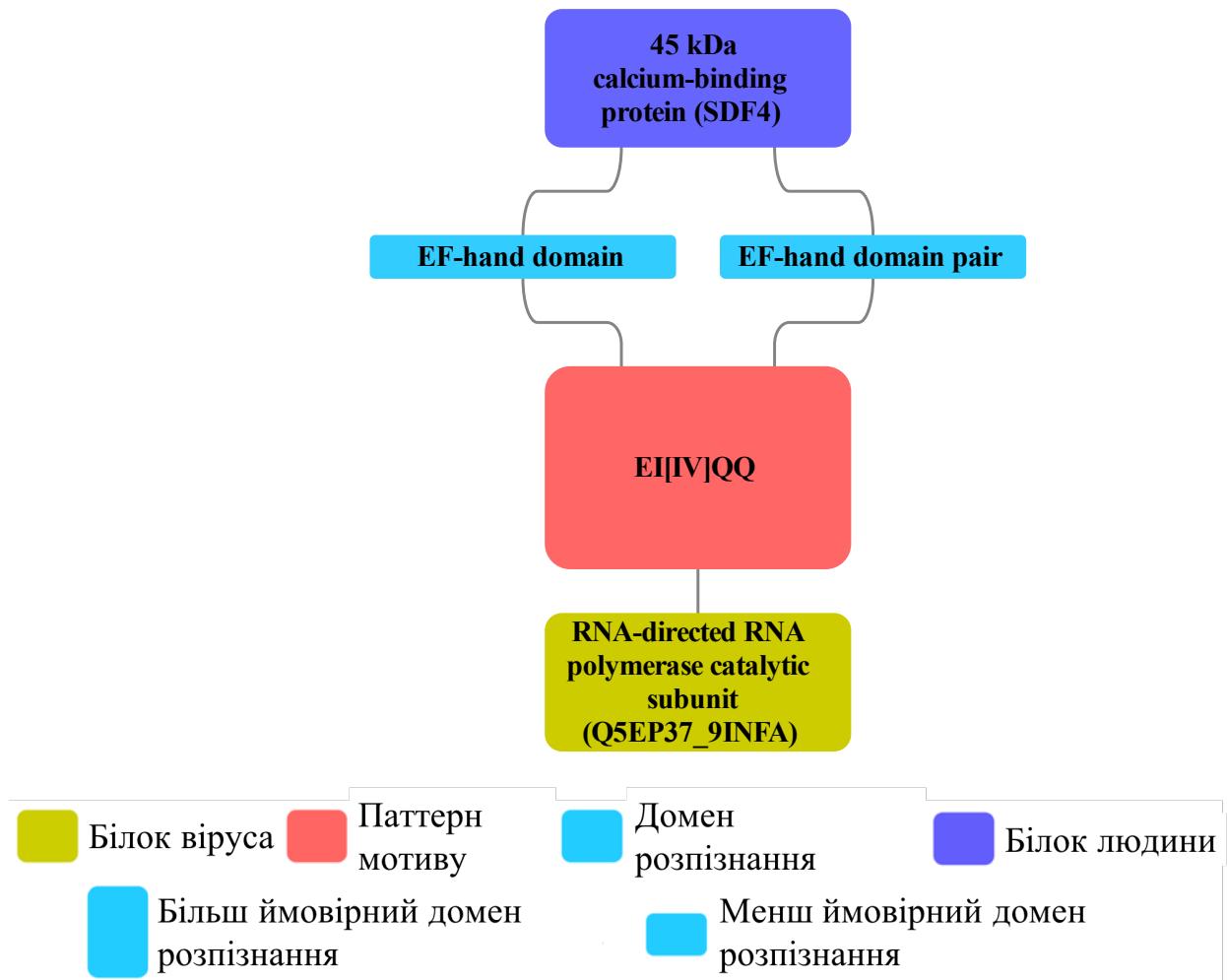


Рис 3.7.6.2. Схема мережі, що показує мотив-кандидат EI[IV]QQ розташований в РНК-залежній РНК -полімеразі грипу А і потенційно розпізнаного доменом EF-hand білка людини SDF4.

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 низкого 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 [29462212, 16177824]. BAG2 і BAG3 є ко-шапероновими білками HSP70 та

HSC70 і працюють як фактор обміну нуклеотидів [9873016]. Таким чином, EBNA5 може посилювати діяльність ко-шаперонів HSP70 та HSC70 або впливати на проліферацію клітин або апоптоз через функції BAG2 або BAG3. Дослідуючи 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 [22600738] не свідчить про наявність сильного сайту зв'язування у цьому домені.

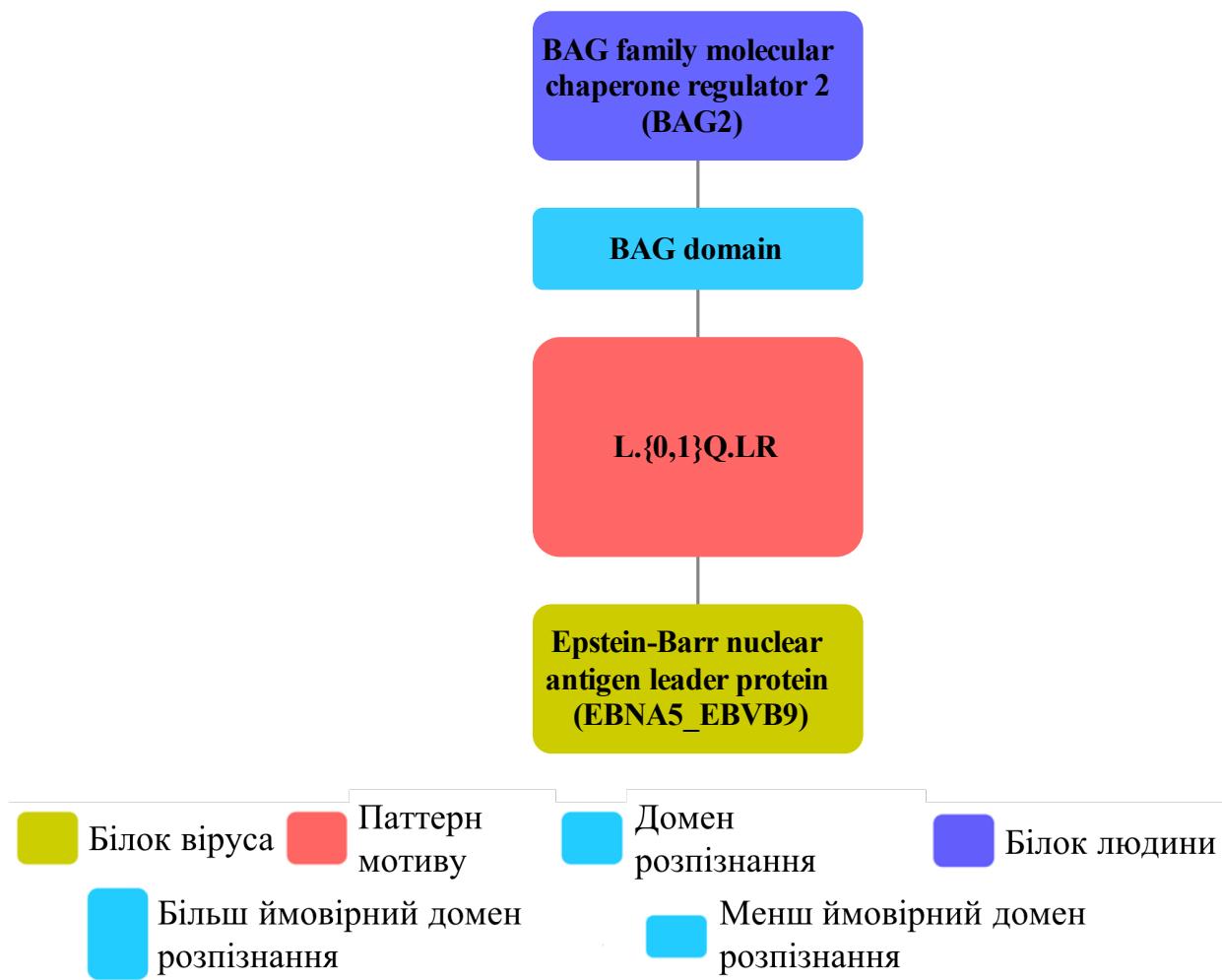


Рис 3.7.7. Схема мережі, що показує мотив-кандидат L.{0,1}Q.LR в білку EBNA5 віrusу Епштейна-Барр потенційно розпізнається доменом BAG у ко-шапероновому білку людини BAG2.

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 і 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 декількох постхокових прикладах (DSG, BAG-доменний мотив) [22600738]. 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 [21238461], 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 ймовірно 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 [PMC4652402], однак, вірусні білки еволюціонують швидко, тому фільтр консервації може видалити справжні мотиви [PMC4089993]. Тим не менш, ми можемо використовувати консерваційний фільтр для білків людини: мотив повинен бути присутнім у людини та декількох 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 декількох 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 знайти. 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 мотива 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* [28002650]. Наш обчислювальний прогноз підкреслює екземпляри доменів в білках людини, які варто перевірити проти невпорядкованих ділянок вірусних білків.

ВИСНОВКИ

1. Я отримав та обробив дані експериментальної взаємодії з публічних баз даних та літератури. Я вивчив властивості мережі вірусно-людської взаємодії. Білки людини - мішені вірусів ввиджуються як центральні, але цей ефект може бути результатом упередженості дослідження в сукупному наборі даних білкових взаємодій.

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

3. Ми визначили домени білкової послідовності у всіх вірусних і 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 мотива 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. Я розробив цю процедуру пошуку мотивів на статистичній мові програмування R, використовуючи інструменти командного рядка та обчислювальний кластер. Ця процедура може бути використана групою (і науковою спільнотою) для передбачення мотивів, у той час, як створюються нові дані про взаємодію білків.

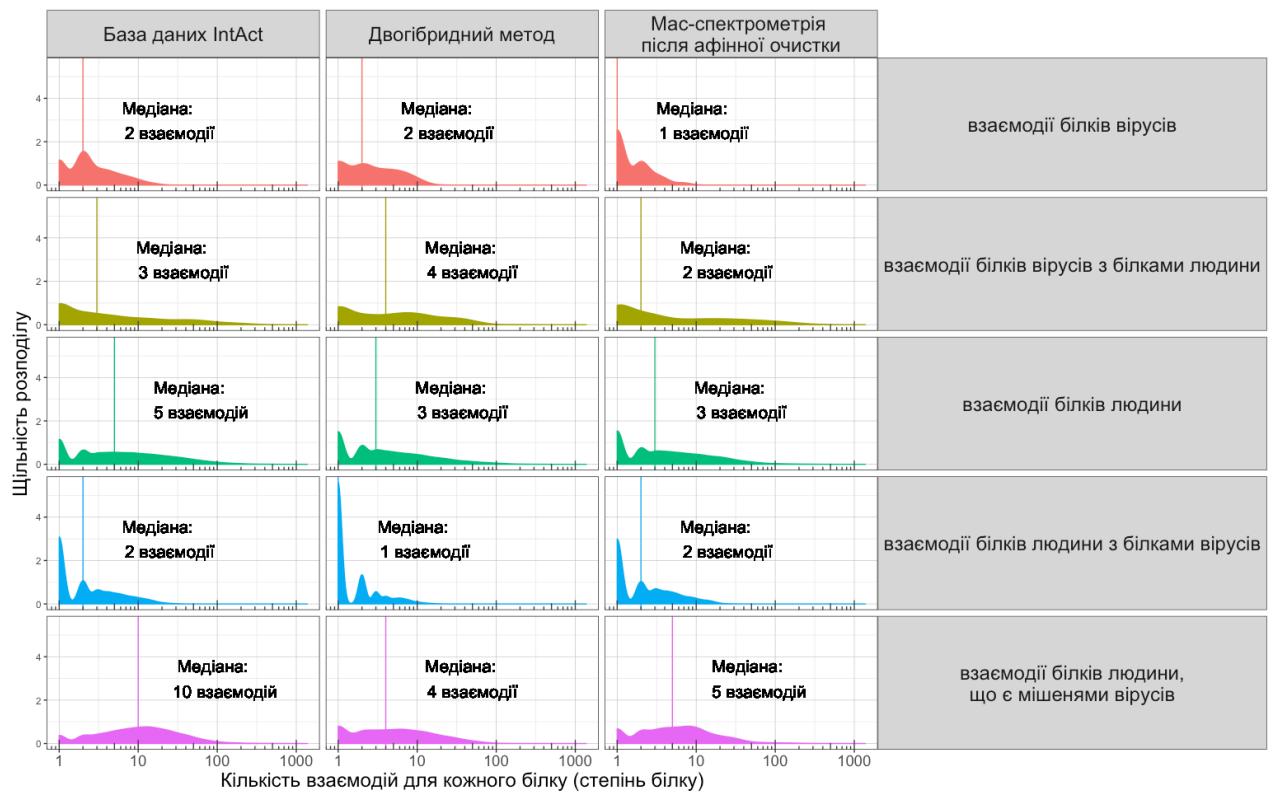
6. Ця робота сприяє нашему розумінню коду взаємодії лінійного мотива - домену розпізнавання та спрямовує вибір доменів людини-вірусних мішеней

для подальших експериментальних досліджень невпорядкованого вірусного протеома.

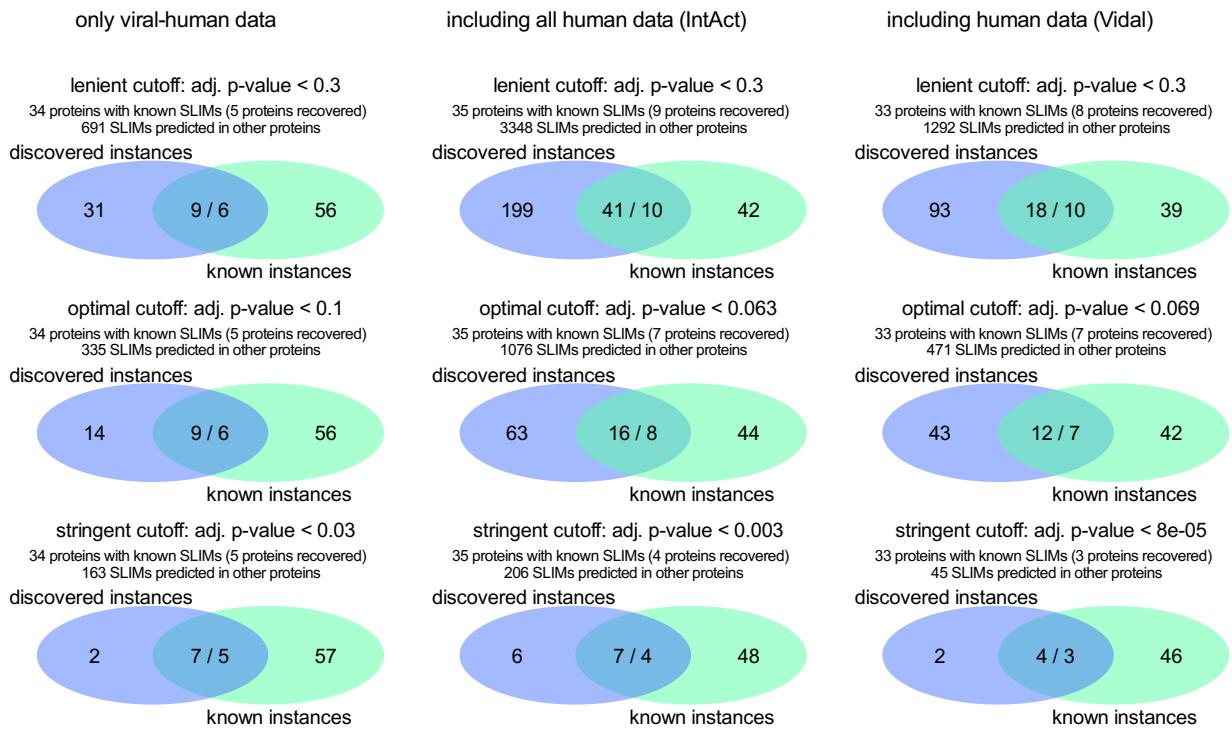
СПИСОК ДЖЕРЕЛ ЛІТЕРАТУРИ

ДОДАТКИ

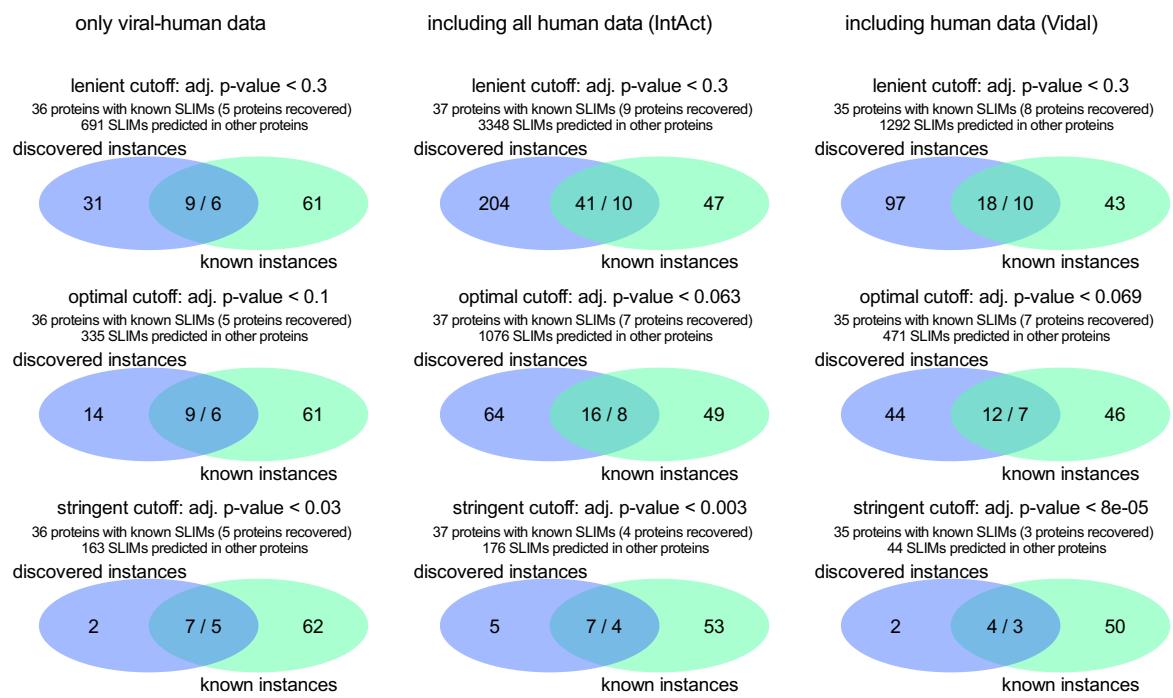
Додаток А



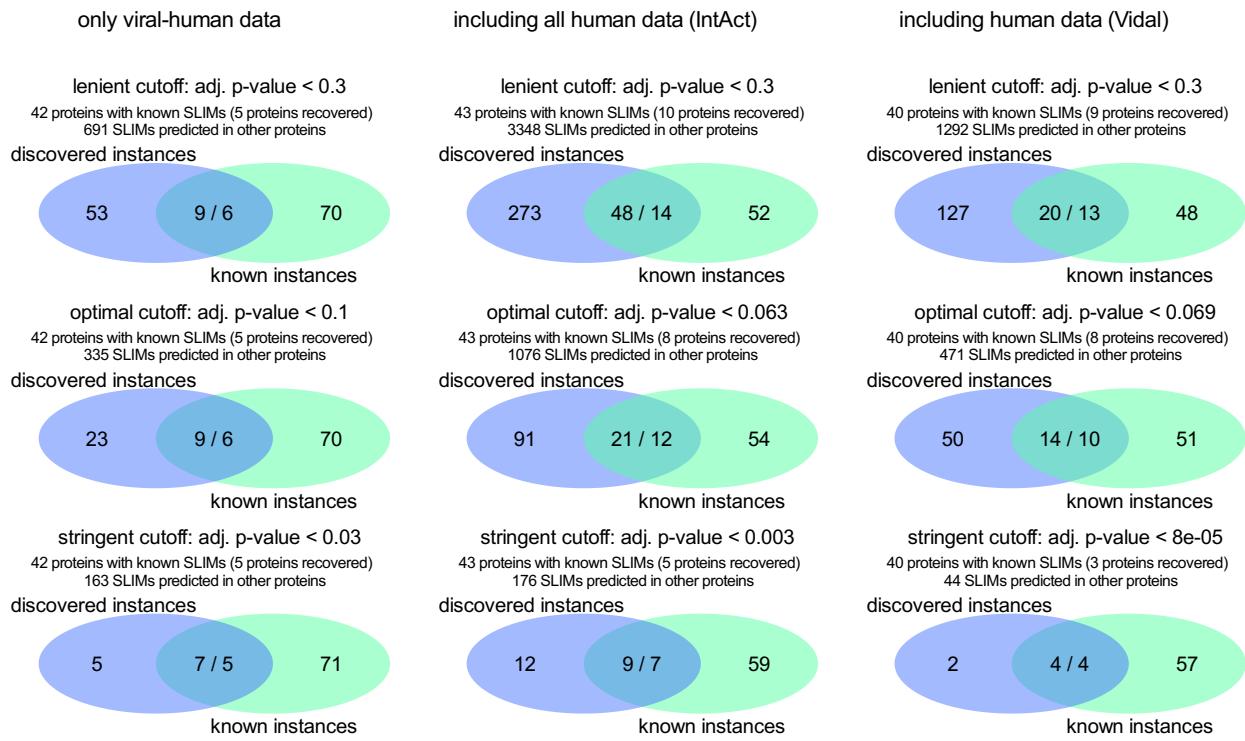
ДІАГРАМА ВЕНА ЩО ПОКАЗУЄ



ДІАГРАМА ВЕНА, ЩО



ДІАГРАМА ВЕНА



ІНСТРУКЦІЯ ДЛЯ ЗАПУСКУ ПРОГРАМИ ДЛЯ ПЕРЕДБАЧЕННЯ ДОМЕНІВ

Наступний код командної строки BASH був використаний для запуску програми InterProScan:

```
/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
```

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

Наступний код командної строки BASH був використаний для запуску програми QSLIMFinder:

```
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
```