Еще один шаг на пути к пониманию взаимодействий белков в LINE-1

Валентина Чернова, Ольга Хасина, Екатерина Полякова-Георгантас, Дарья Репкина, Ксения Метелева, Марк Масрамон, и Дмитрий Алексеев, Илья Алтухов, Джон ЛаКава

Абстракт

LINE1 (Long Interspersed Nuclear Element-1) - это мобильный элемент в геноме человека который распространяет себя механизмом copy-paste, который известен как ретротранспозиция. Белки ORF-1 и ORF-2, закодированные в последовательности L1, формируют комплекс белков с белками хозяина. Эти комплексы, которые остаются плохо-изученными, участвуют в жизненном цикле L1. Мы изучали белок-белковые взаимодействия с L1 при помощи String DB и BioPlex.

Мы разработали алгоритм, позволяющий определить, какие белки друг с другом взаимодействуют, и где они локализованы относительно центра комплекса. Мы также обнаружили среди этих белков предположительные физические связи, которых на данный момент нет в базах данных.

Введение

Примерно половину генома человека составляют повторяющиеся участки ДНК которые приходящиеся на ретро-элементы. При этом примерно 20% генома занимают специфично L1 участки. Особенно интересно что ретротранспозоны L1 были найдены причастными в развитии нервной системы и L1 экспрессирован в половине всех раков. Примерно в каждом двадцатом младенце есть ново-мобилизованная вставка L1.

Полный ген L1 имеет длину ~6000 нуклеотидов и функционирует как бицисторнная мРНК, а также кодирует два белка ORF1 и ORF2, которые в свою очередь являются связывающим РНК белками с шапероновой и эндонуклеозидной активностями.

Мы использовали недавнюю работу Molloy и соавторов [1], для анализа белковых комплексов L1 которые могут формироваться в живых клетках. Molloy и соавторы определяют набор возможных белковых комплексов основываясь на некотором количестве аффинных протеомных экспериментов.

Материалы и методы

Белковые кластеры LINE-1

Во время работ мы использовали белковые кластеры, предложенный в работе К.Р. Моллой [2]. Суммарно в этих кластерах содержится 31 белок.

I-DIRT эксперименты

24 эксперимента I-DIRT в разных условиях были взяты из неопубликованной информации Дж. ЛаКавы в виде таблицы показателей схожести у 886 белков. Из них только 21 были обнаружены в кластерах из работы К.Моллой.

Базы данных белок-белковых взаимодействий

Для того чтобы определить и описать белок-белковые взаимодействия мы использовали 2 базы данных: String DB [3] и BioPlex V.2.0 [4].В базе данных String мы рассматривали только подтвержденные экспериментально или найденные в других базах данных связи между белками. Мы рассматривали связи со средним уровнем достоверности (0,4).

Визуализация сетей

Сети, полученные из результатов I-DIRT были визуализированы с использованием Cytoscape. Значения корреляции и количество входящих граней были отображены в толщине граней и размером узлов соответственно.

Алгоритм идентификации белок-белковых взаимодействий

Мы использовали RStudio Version 1.0.153 для построения идентификации новых кластеров и разработки алгоритмов для определения взаимодействий, которые были визуализированы в Cytoscape. Во время анализа мы сконцентрировались на 31 белке. Данные из экспериментов Дж.ЛаКавы показывали частоты присутствия белков в разных условиях.

Для определение связи между белками мы создали бинаризованную таблицу, где 1 - белок идентифицирован в эксперименте, а 0 - не идентифицирован. Если один белок присутствовал в любом случае, когда присутствовал другой белок, то мы предполагаем что эти белки связаны. Тот белок который пристутствует только при наличии другого находится дальше от центра комплекса. Также мы создали корреляционную матрицу между белками. Таблица связи между белками и корреляционная матрица были использованы для построения сети взаимодействия (Рис. 3).

Результаты и их обсуждение

Функциональное описание кластеров (Рис. 1)

Кластер 1: Эти белки учатвуют в клеточном росте. Белок ORF2 отсутствует в базе String, в то время как остальные белки взаимодействуют. Исходя из данных в БД BioPlex, нам не удалось найти информации о каких либо взаимодействиях белков ORF2, HSPAA1, и TUBB4B.

Кластер 2: Эти белки ассоциированы со связыванием с мРНК, деградацией мРНК, а также регуляцией трансляции. В БД String нет информации о белке ORF1, в то время как другие белки взаимодействуют - UPF1 с PABPC4 (достоверность 0.4) и UPF1 с MOV10 (достоверность 0.125). В БД BioPlex нет информации о взаимодействиях белков UPF1, PABPC1, PABPC4, и ORF1.

Кластер 3: HSPA8 и HSPA1A белки шапероны, а также участвуют в репрессии транскрипции. MEPCE кэпирующий фермент, который стаблицизирует snRNA, а также отрицательный регулятор связывания с РНК Pol II промотером. В соответствии с БД String, MEPCE не взаимодействует с HSPA8 и HSPA1A, в то время как они взаимодействуют с другом. В BioPlex не информации о каких либо взаимодействиях с белком HSPA1A.

Кластер 4: В отличие от многих других кластеров, у этих белков нет общей функции. Согласно базе данных String, только у NAP1L1 и NAPL4 известны зафиксированные взаимодействия. Белки HAX1 и YME1L1 не включены в БД BioPlex.

Cluster 5: Все белки этого кластера задействованы в связывании белков и ДНК. В базе String DB присутствуют все белки, но взаимосвязи показаны только между двумя парами белков: PURB и PURA, PCNA и PARP1.

Информация представленная выше ясно показывает, что базы данных String и BioPlex являются очень незавершенными и не предоставляют никакой высококачественной информации о тех белковых комплексах которые мы исследуем. Более того, эти базы данных не предоставляют ни направления взаимодействий ни контекста в плане механизмов этих комплексов.

Идентификация белок-белковых взаимодействий и их локализация

Мы разработали алгоритм, который позволяет анализировать связанные наборы данных I-DIRT, которые используются для прогнозирования новых белок-белковых взаимодействий(далее PPI) и, таким образом, более широких комплексов. Это позволяет нам лучше понять не только существование этих взаимодействий, но и их локализацию в комплексе и корреляцию их совместного появления. Эта частота корреляции может быть воспринята в формате гистограммы в приложении 4.

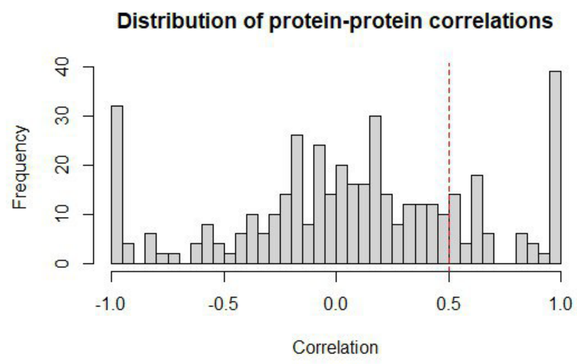

Рисунок 4: Распределение корреляций между белками на основе I-DIRT экспериментов.

Наш алгоритм сгенерировал 21 узел (белок) и 138 направленных взаимодействий на всех уровнях корреляции, из них 44 имеют корреляцию выше 0.5. Анализируя сеть созданную с помощью этого алгоритма в Cytoscape предоставило нам не только приблизительное значение расстояния к центру комплекса но и относительное положение элементов этого комплекса к друг другу.

Это создает некоторый контекст и более полное понимание природы этих комплексов, которые просто не представлены в общепринятых базах данных.

Дополнительная информация о пятом кластере: не все белки из кластера 5 присутствуют в таблице для алгоритма. По полученным результатам ясно, что белок PARP1 находится в центре, а белки LARP7 и PURA находятся на периферии. Мы обнаружили две отсутствующие в базах данных связи с высоким уровнем корреляции, LARP7 и PARP1 (0,95), PURA и PARP1(0,8)

Описание всей сети взаимодействий

Предполагаемую относительную локализацию белков в комплексе, определенная по количеству входящих в них граней можно увидеть в таблице 1. Тот факт, что HAX1- самый центральный белок, согласно этой модели, довольно неожиданно, так как ожидаемо было бы, чтобы это был белок ORF1. Несмотря на это, так как был мечен ORF2, вполне вероятно, что ORF1 мог оторваться в раствор быстрее, чем HAX1. Тенденция группирования по кластерам очень интересна во множестве случаев, и может показать, что локализация в комплексе влияет на поведение при эксперименте вместе с функциональными отношениями

Существуют несколько высоко коррелирующих связей между разными кластерами, смоделированные сетью, кроме связей с HIST1H2BO. В них включены HAX1, находящийся ближе к центру, чем ORF1 с корреляцией 0,87, DDX6 ближе к центру, чем HSPA1A с корреляцией 0,82, MOV10 дальше от центра, чем LARP7 с корреляцией 0,99.

С корреляцией 0,65, ниже установленного порога, белки из кластера 1 очень активно соединяются с PABPC4.

Заключение и перспективы

В свете перехода от подхода “garbage in-garbage out” к большим данным, появляется необходимость в разработки новых алгоритмов анализа данных. Разработанный нами алгоритм является одним из премером, кроме того он применять не только по отношению к комплексам LINE-1, но к другим белковы взаимодействиям.

Предсказания и модели предложенные нами имеют значительный вклад, однако требуют подтверждения с помощью дополнительных экспериментов. Для валидации алгоритма требуется больше экспериментальных с уже известными и еще не изученными белковыми комплексами.

Дальнейшие шаги в изучении белок-белковых взаимодействие внесут значительный вклад в биомедицине, например в разработке новых биомаркеров заболеваний и потенциальных лекарственных мишеней. Интерактомика находится в начале своего развития, и мы верим что создание алгоритмов внесет значительный вклад.

Number of incoming edges	1	2	3	4	5	6	7	8	9	10
Proteins	PCNA			PURA	HSPA 1A	HSPA 8	LARP 7	DDX6	IPO7	HAX1
				PARP 1	HSPA 90AA 1	MOV1 0	TUBB 4B	UPF1	L1RE 1	
					HIST1 H2BO	ORF2		PABP C1		
						HSPA 90AB 1		PABP C4		
								TUBB		
								RPS2 7A		

Таблица1: По столбцам - количество связанных белков с данным белком, по строкам - белки из кластера LINE-1.

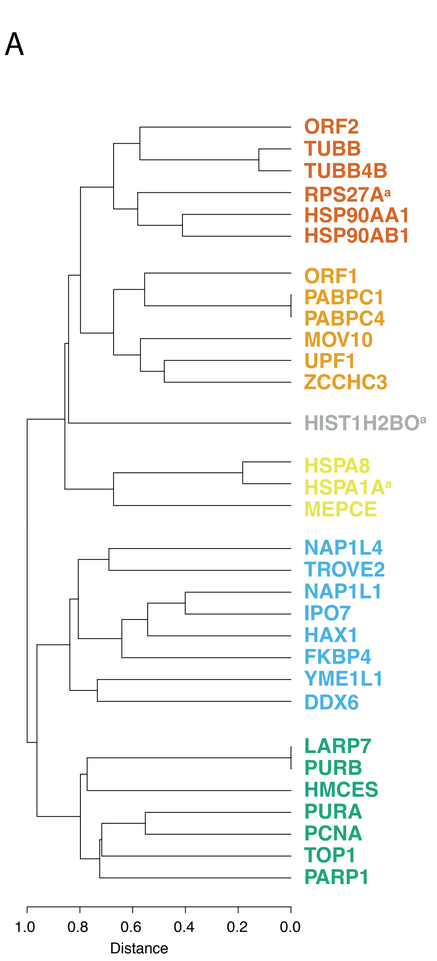

Рисунок 1: Ось X - различные эксперименты, ось У - значение аффинности для белков.

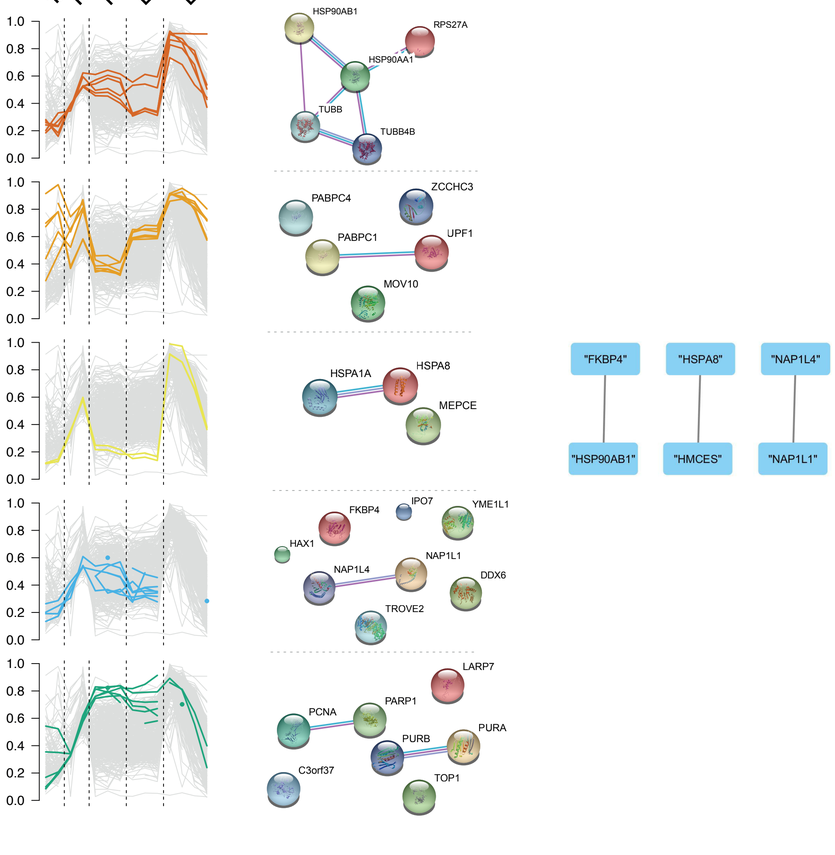

Рисунок 2: Взаимодействия между белками согласно String и BioPlex

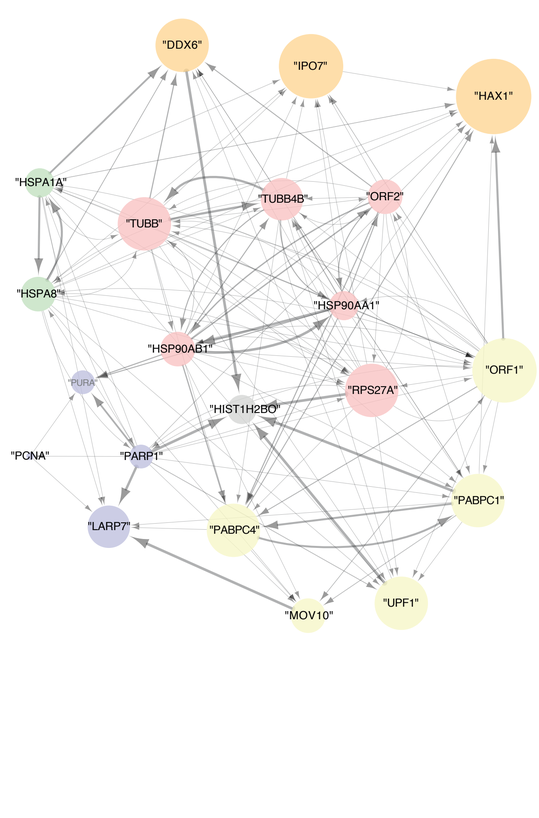

Рисунок 3: Сеть взаимодействий построенная на основе разработанного алгоритма - для каждого кластера в отдельности и вместе. Тольщина связи указывает на корреляционный коэффеициент. Размер узла отображает количество связанных белков с данным белком (входящие ребра).