

Лекция девятая «Контекстный анализ и распознавание сайтов связывания транскрипционных факторов»

Версия для печати

Лектор М.А.Поздняков.

(СЛАЙД 1)

Наша лекция называется Контекстный анализ и распознавание сайтов связывания транскрипционных факторов.

(СЛАЙД 2)

Как известно, все живые организмы в настоящее время делят на прокариот и эукариот.

Прокариоты являются доядерными одноклеточными организмами. Среди них выделяют царство археобактерий и царство эубактерий. Многие археобактерии живут в экстремальных условиях: при высоких концентрациях солей или при высоких температурах, приближающихся к 100 градусам. Анализ нуклеотидных последовательностей показал, что археобактерии отстоят примерно одинаково далеко и от остальных прокариот, и от эукариот. Эукариоты (дословно - настоящие ядерные) имеют ядро. К ним относятся все многоклеточные организмы, включая человека.

(СЛАЙД 3-4)

Основными отличиями эукариот от прокариот является:

1. наличие ядра
2. наличие других компартментов клетки: вакуоли, комплекс Гольджи, лизосомы, митохондрии и др.
3. Наличие окислительного фосфорилирования (дыхательной цепи).
4. Отличия в клеточной границе - (1) наличие клеточной стенки (у растений), (2) отсутствие муреина и других составляющих, свойственных бактериям
5. ДНК линейные (у прокариот - кольцевые)
6. Отсутствие плазмид как средства обмена информацией
7. Диплоидность ДНК
8. Значительно более богатый арсенал репарации ДНК.
9. Прерывистая структура генов.
10. Значительно бо'льшие размеры геномов
11. Намного бо'льшая доля некодирующей ДНК
12. Значительно больше повторяющихся последовательностей: (а) генов - кластеры изофункциональных генов; (б) повторы в некодирующих областях геномов.
13. Значительно бо'льшая упаковка ДНК
14. Есть многоклеточные организмы. Дифференцировка клеток по функциям. Особенности экспрессии эукариотического гена
15. Значительно более сложная регуляция уровня экспрессии генов. И др.

В целом можно заметить, что в отличие от прокариот, эукариоты организованы существенно более сложно. Основными отличиями эукариот от прокариот является наличие ядра, отделяющего цитоплазму клетки от упакованной ДНК. Для большинства видов эукариот характерна диплоидность - каждая молекула ДНК содержится в ядре в двух копиях (клетки прокариот гаплоидны - они имеют молекулы ДНК только в одной копии). ДНК упакована в нуклеопротеиновые структуры хроматина, что позволяет избирательно регулировать функциональную активность генов. Упаковка ДНК также позволяет эукариотам иметь огромный размер геномов по сравнению с прокариотами

Кратко рассмотрим эти пункты.

(СЛАЙД 5)

Прерывистая структура генов.

Схематически строение эукариотического гена приведено на слайде.

Характерная особенность генов эукариот, отличающая их от прокариотических генов, это наличие экзон-интронной или мозаичной структуры, образованной чередованием экзонов (кодирующих, смысловых районов) и интронов (некодирующих районов). За счёт этого гены эукариот могут достигать существенно больших размеров, чем гены прокариот.

(СЛАЙД 6)

Рассмотрим, что включает в себя понятие "функционирование гена".

Экспрессией эукариотического гена называется его проявление - обеспечение какого-то признака у организма. (теория Бидла и Татума)

Основными этапами экспрессии гена является

1. Транскрипция
2. Процессинг РНК
3. Трансляция
4. Посттрансляционный процессинг белка и его транспорт в нужное место клетки

Процессинг РНК:

Сайты, по которым происходит вырезание интронов, называют сайтами сплайсинга.

Сплайсинг одного гена может происходить несколькими способами. Это означает, что в состав зрелых мРНК могут входить разные комбинации экзонов.

Мало генов но много белков:

Блочность организации 5'-регуляторных районов проявляется также и в таком их свойстве, как **наличие для многих генов альтернативных промоторов** (зачастую расположенных на значительном расстоянии один от другого). В зависимости от функционального состояния клетки транскрипция одного и того же генного локуса может осуществляться с различных (альтернативных) промоторов.

Существенно также то, что некоторым промоторам свойственно **наличие множественных стартов трансляции**, находящихся на относительно небольшом расстоянии друг от друга. Считывание с одного гена разных вариантов РНК называется альтернативной транскрипцией. Указанная особенность 5'-регуляторных районов лежит в основе механизма формирования большого разнообразия первичных транскриптов одного и того же генного локуса и, как следствие этого, разнообразия белков, кодируемых одним и тем же генным локусом.

Дополнительное увеличение разнообразия кодируемых белков может достигаться за счет альтернативного сплайсинга, приводящего к объединению в зрелую мРНК различных комбинаций экзонов, соответствующих одному и тому же первичному транскрипту. В настоящее время известно много примеров первичных транскриптов, на основе которых сплайсинг может проходить по десяткам альтернативных путей.

По последним данным, по крайней мере 30% генов человека характеризуются альтернативным сплайсингом

(СЛАЙД 7)

Значительно бо'льшие размеры геномов

Таблица 1. Размеры геномов некоторых организмов.

Организм	Примерный размер гаплоидного генома
E.coli	$3 \cdot 10^7$
Миксомицеты	$7 \cdot 10^7$
Трипаносома	$8 \cdot 10^7$
Нематода	$8 \cdot 10^7$
Резушник Талля	$7 \cdot 10^7$
Шелкопряд	$5 \cdot 10^8$
Плодовая мушка	$1.65 \cdot 10^8$
Морской еж	$8 \cdot 10^8$
Шпорцевая лягушка	$3 \cdot 10^9$
Протей	$5 \cdot 10^{10}$
Курица	$1.2 \cdot 10^9$
Мышь	$3 \cdot 10^9$
Корова	$3.1 \cdot 10^9$
Человек	$2.9 \cdot 10^9$
Кукуруза	$5 \cdot 10^9$
Лук	$1.5 \cdot 10^{10}$

Из приведенной таблицы видно, что геномы высших эукариот значительно превосходят по размерам геномы прокариот. У высших растений размер геномов может достигать очень больших величин за счет полиплоидии.

Намного бо'льшая доля некодирующей ДНК

Геномы эукариот организованы существенно более сложно, чем геномы прокариот. Главной особенностью геномов эукариот в сравнении с прокариотами является наличие избыточной ДНК.

По результатам экспериментов в рамках геномного проекта по E.Coli геном особей этого вида составляет 4 639 221 нуклеотидное основание, 87.8 % занимают реальные и вероятные белок-кодирующие гены (цистрыны) 0.8 % - гены, кодирующие различные фракции РНК, не связанные с кодированием белков (тРНК, рРНК и др.) 0.7 % - некодирующие повторы. Таким образом, подавляющая доля ДНК E.coli занята генами а межгенные участки составляют относительно малую долю - около 11%. Средний размер межгенных промежутков - около 100 пар оснований. Эти межгенные промежутки часто содержат разнообразные функциональные сайты, то есть выполняют регуляторную функцию.

Иная картина наблюдается у эукариотических организмов. У человека выявлено около 30 000 генов. В то же время размер генома человека составляет $3 \cdot 10^9$ пар оснований. Таким образом, кодирующую часть генома можно оценить как 10-15 % от генома человека. Экспериментальные работы в рамках проекта "Геном человека" показали, что только 1 % генома приходится на кодирующие экзоны, 24 % на некодирующие интроны и 75 % на межгенные промежутки.

Пункт Значительно больше повторяющихся последовательностей

В конце 60-х годов 20-го столетия была открыта фундаментальная особенность геномов эукариот - наличие нуклеотидных последовательностей разной степени повторяемости. Различают следующие фракции последовательностей в геномах эукариот:

1. уникальные последовательности, то есть представленные в одном экземпляре
2. Среднечастотные повторы. Повторяются десятки и сотни раз
3. Высокочастотные повторы, число повторов достигает 10^6 раз.

Уникальные последовательности чаще всего представлены генами. Число генов у эукариот определяют одним из двух способов. Первый способ прямой - прочитывание геномных последовательностей и определение открытых рамок считывания. Этот подход позволяет оценить число генов в рамках геномных проектов, то есть возможно, не все гены. Второй способ - оценочный: размер генома делят на средний размер гена + средний размер межгенного промежутка и получают оценку числа генов данного генома.

Одним из типов среднечастотных повторов являются кластера изофункциональных генов. Изофункциональными называются гены, продукты экспрессии которых характеризуются структурно-функциональным сходством. В качестве примера подобных кластеров можно привести гены рРНК и гистонов.

Эти гены тандемно повторяются в геномах и представлены большим числом идентичных копий.

Гены эукариот отделены друг от друга районами нетранскрибируемой ДНК (межгенными спейсерами).

Основной причиной того, что геномы эукариот имеют низкую кодирующую плотность, является наличие огромной фракции некодирующей ДНК, представленной повторяющимися последовательностями, межгенными спейсерами и интронами. В этом состоит существенное отличие геномов эукариот от прокариот.

В межгенных спейсерах, а также интронах располагаются различные типы повторяющихся последовательностей.

(СЛАЙД 8)

Значительно бо'льшая упаковка ДНК

При наличии у эукариот намного большего по размерам генома (по сравнению с прокариотами) у эукариот возникает проблема компактизации хроматина. Для геномов эукариот характерно наличие сложно организованного хроматина с большим количеством уровней упаковки. Первый (низший) уровень упаковки хроматина соответствует нуклеосома, в составе которой ДНК накручена на октамер, образованный белками гистонами. Одной нуклеосома соответствует примерно 200 пар оснований. Нуклеосома имеет форму цилиндра, вокруг которого ДНК делает два оборота. Участок ДНК между двумя нуклеосомами называется линкером. 150 пар оснований непосредственно намотано на нуклеосому, и еще около 50 - линкер.

(СЛАЙД 9)

Второй уровень - уровень нуклеопротеиновых нитей. Нуклеосомы располагаются по спирали друг относительно друга. **Третий уровень - петлевой (супербидный).**

Специальные белки "перехватывают" нуклеопротеиновую нить, образуя петли. Четвертый уровень компактизации хроматина - это состояние хромосом при делении клетки. Каждый уровень укорачивает длину хромосомы примерно на порядок. Упаковка ДНК в ядре дала эукариотическим организмам возможность существенного увеличения размера генома. В свою очередь, это обеспечило огромную информационную емкость геномов эукариот и возможность кодирования сложных паттернов регуляции экспрессии генов (Kolchanov et al., 2000). Последнее особенно необходимо для обеспечения ткане-специфичной экспрессии генов в организмах многоклеточных эукариот.

Пункт Есть многоклеточные организмы. Дифференцировка клеток по функциям.

Особенности экспрессии эукариотического гена

Среди эукариот есть большое количество многоклеточных сложноорганизованных организмов, включая высших животных (земноводных, рептилий, птиц, млекопитающих). Эти организмы состоят из сотен тканей, состоящих из дифференцированных клеток, ткани образуют органы, органы образуют системы, которые обеспечивают функционирование сложного организма. В подавляющем большинстве случаев дифференцированные клетки содержат весь генетический материал, который достался организму от родителей. Однако в каждой специализированной клетке в каждый момент времени функционируют далеко не все ее гены.

На уровень активности гена влияют

1. стадии клеточного цикла,
2. функционального состояния клетки,
3. ткани,
4. органа,
5. стадии индивидуального развития,
6. действия внешних индукторов и др.

Это обеспечивается сложными и разнообразными механизмами дифференциальной регуляции экспрессии эукариотических генов. Существует **регуляция уровня экспрессии гена на всех этапах экспрессии гена** (рассмотренных выше). А именно:

(СЛАЙД 10)

Упаковка ДНК	1) Гены в сильноупакованных участках генома не экспрессируются или экспрессируются слабо. 2) Играть роль также химические модификации ДНК (метилирование), 3) вероятность посадки нуклеосомы, 4) химические модификации белков нуклеосомы, др.
транскрипция	Сложный аппарат регуляции транскрипции
трансляция	1) Время жизни РНК 2) Использование кодонов, которым соответствуют часто- или редко-встречающиеся тРНК в цитоплазме клетки, др.

Экспрессия генов эукариот может регулироваться на различных уровнях их организации и функционирования, связанных с особенностями нуклеосомной упаковки хроматина, интенсивностью сплайсинга, полиаденилирования, стабильностью мРНК в цитоплазме. Однако, именно транскрипции принадлежит ключевая роль в регуляции экспрессии генов (СЛАЙД 11)

Транскрипция является первой стадией реализации генетической информации, на которой нуклеотидная последовательность ДНК копируется в виде нуклеотидной последовательности РНК. В основе механизма этого копирования лежит принцип комплементарного взаимодействия нуклеотидных оснований. Транскрипция осуществляется ферментами ДНК-зависимыми РНК-полимеразами, которые синтезируют РНК на матрице ДНК. Синтез РНК начинается в определенных местах ДНК, называемых промоторами, а завершается в терминаторах. Участок ДНК, ограниченный промотором и терминатором, является единицей транскрипции, его иногда называют транскриптом. В пределах одного транскрипта копируется только одна из двух нитей ДНК, которая называется матричной или значащей. Во всех транскриптах, считываемых в одном направлении, значащей является одна нить ДНК; в транскриптах, считываемых в обратном направлении, значащей является другая нить ДНК. Соседние транскрипты могут быть отделены друг от друга нетранскрибируемыми участками ДНК, а могут и перекрываться, в частности, так, что значащими становятся обе цепи ДНК. Разбиение ДНК на множество транскриптов обеспечивает возможность независимого считывания разных генов, их независимого включения и выключения.

(СЛАЙД 12-13)

У прокариот транскрипты чаще называют **оперонами** (что означает единица регуляции). Многие опероны прокариот содержат по несколько генов, обычно функционально связанных, например, оперон, кодирующий ферменты синтеза определенной аминокислоты. Существуют опероны, содержащие гены, не кодирующие белков - продуктом такого гена является РНК. Это рибосомальные РНК, транспортные РНК и др. Также у прокариот есть смешанные опероны.

Строение 5'-регуляторных районов генов прокариот.

У прокариот процесс транскрипции осуществляется с помощью "голоэнзима" - полного фермента, РНК-полимеразы, состоящего из собственно РНК-полимеразы и присоединяющегося к ней сигма-фактора. РНК-полимераза осуществляет основную реакцию полимеризации рибонуклеотидов (копирование РНК с одной нити на другую) сигма-фактор нужен для опознавания промотора. Промоторы имеют определенные последовательности нуклеотидов, "узнаваемые" РНК-полимеразой. Обнаружили следующие постоянные особенности в последовательностях промоторов прокариот:

1. наличие стартовой точки транскрипции Более 90% случаев - ген начинается с пурина.
2. особая последовательность нуклеотидов, начиная с позиции -10. 6 оснований, обнаруживают почти во всех промоторах. (Домен Прибнова).
3. особая последовательность нуклеотидов в районе -35. Тоже 6 оснований.
4. Расстояние, разделяющее эти консервативные последовательности, составляет 16 и 18 пар оснований в 90% случаев. В виде исключений могут быть 15, 19, 20. Эти особенности хорошо согласуются с размерами молекулы РНК-полимеразы.

Вначале РНК-полимераза контактирует с районом -55 +20 пар оснований.

Иницирующий комплекс содержит РНК-полимеразу и сигма-фактор, который слабо связывается с участком промотора в положении -35, контролируя посадку РНК-полимеразы именно на промотор. Затем РНК-полимераза связывается с доменом Прибнова - районом вблизи положения -10. Одновременно с этим расплетается почти два витка двойной спирали ДНК (длиной в 17 пар оснований). Присутствующие в домене Прибнова многочисленные пары АТ существенно облегчают такое расплетание. Затем сигма-фактор диссоциирует, транскрипционный комплекс превращается в комплекс элонгации.

(СЛАЙД 14)

ТРАНСКРИПЦИЯ

Цикл транскрипции можно разделить на четыре основные стадии, каждая из которых состоит из более мелких этапов:

1. связывание РНК-полимеразы с ДНК
2. инициация цепи РНК
3. рост (элонгация) цепи РНК
4. терминация цепи РНК

Этот цикл у прокариот удается полностью осуществить в простой бесклеточной системе, состоящей из ДНК-матрицы и очищенной РНК-полимеразы, без каких-либо дополнительных факторов. Это, конечно, не значит, что прокариотическая РНК-полимераза является единственным белком, участвующим в транскрипции. В ней могут участвовать и разные регуляторные белки. Однако их роль вспомогательная: они мешают или помогают РНК-полимеразе на разных стадиях цикла транскрипции, которые она осуществляет и без них.

У эукариот транскрипция изучена намного меньше, чем у прокариот. Это связано, в том числе и с тем, что очищенные РНК-полимеразы эукариот не способны осуществить полный цикл транскрипции. Для этого всегда нужны дополнительные белковые факторы, лишь часть из которых получена в очищенном виде. Пока что эукариотический цикл транскрипции удается осуществить лишь в частично очищенных клеточных экстрактах, содержащих неохарактеризованные компоненты. В основных чертах транскрипция про- и эукариот сходны.

Цикл транскрипции начинается с присоединения РНК-полимеразы к промотору - строго определенному участку ДНК, с которого начинается синтез РНК. Механизм поиска РНК-полимеразой промоторов изучен недостаточно. Предполагается, что молекулы РНК-полимеразы присоединяются к случайным местам двойной спирали ДНК, некоторое время перемещаются по ней, отсоединяются и снова присоединяются к ДНК до тех пор, пока не окажутся на промоторах.

Оказавшись на промоторе, РНК-полимераза образует с ним так называемый закрытый промоторный комплекс, в котором ДНК сохраняет двуспиральную структуру. В закрытом комплексе РНК-полимераза еще не способна к синтезу РНК. Этот комплекс нестабилен и легко диссоциирует при повышении температуры.

Закрытый комплекс обратимо превращается в открытый комплекс, в котором РНК-полимераза расплетает примерно один виток двойной спирали ДНК в районе **стартовой точки** - нуклеотида, с которого начинается комплементарное копирование матрицы. В открытом комплексе связь РНК-полимеразы с промотором становится значительно более прочной, чем в закрытом.

Инициация - следующая стадия транскрипции, требует наличия субстратов РНК-полимеразы - нуклеозидтрифосфатов, и заключается в синтезе первых нескольких звеньев цепи РНК. Первый нуклеотид входит в состав цепи, сохраняя свою трифосфатную группу, а последующие присоединяются к 3'-ОН группе предыдущего с освобождением пироглютата. На стадии инициации РНК-продукт связан с матрицей непрочной и с высокой вероятностью может освобождаться из комплекса. В этом случае РНК-полимераза не покидая промотора снова иницирует РНК. Такой синтез коротких олигонуклеотидов называют **абортивной инициацией** в противоположность **продуктивной** - то есть завершающейся образованием полноценного РНК-продукта.

Когда РНК-продукт достигает критической длины (3-9 оснований в зависимости от промотора), abortивная инициация полностью прекращается, транскрибируемый комплекс стабилизируется и уже не распадается до тех пор, пока синтез молекулы РНК не будет

(СЛАЙД 15)

Элонгация транскрипции начинается с того момента, когда расплетено примерно 18 пар оснований. При этом примерно 12 оснований ДНК образуют комплементарные пары с растущей цепью РНК. По мере продвижения РНК-полимеразы по матрице впереди РНК-полимеразы происходит расплетание, а позади - восстановление двойной спирали ДНК. При восстановлении пары оснований ДНК происходит освобождение от водородных связей остатка очередного РНК.

В процессе элонгации РНК-полимераза движется по ДНК с непостоянной скоростью. Завершается синтез РНК в строго определенных участках ДНК - **терминаторах**, где происходит отделение от ДНК и готовой РНК, и РНК-полимеразы, которая может вступить в следующий цикл транскрипции.

В состав первичного транскрипта - пре-мРНК входят как экзоны, так и интроны (некодирующие районы). В процессе сплайсинга интроны вырезаются из пре-мРНК. Оставшиеся же части - экзоны объединяются в зрелую матричную РНК (мРНК), которая может транслироваться в белок.

Район, узнаваемый РНК-полимеразой как финиш транскрипции называется районом терминации транскрипции. В 3' нетранслируемом районе гена - 3'UTR (рис. 1-2) может находиться также контекстный сигнал полиаденилирования. Этот сигнал состоит из гексамера ААТААА (иногда встречается АТТААА)

(СЛАЙД 16)

Транскрипция у эукариот

У эукариот в составе транскриптона, как правило, входит один ген.

Происходит сложнее, чем у прокариот, в связи со следующими причинами:

1. Транскрипцию осуществляет 3 разные РНК-полимеразы.
2. РНК-полимераза эукариот не может самостоятельно инициировать транскрипцию. Для этого нужно большое число белков, называемых общими (базальными) факторами транскрипции. Эти белки должны объединиться в комплекс, который запустит транскрипцию. Формирование такого комплекса - многоступенчатый процесс, от скорости его образования зависит скорость инициации транскрипции и, следовательно, уровень экспрессии эукариотического гена. Многие регуляторные белки (транскрипционные факторы) действуют, влияя главным образом на процесс сборки транскрипционного комплекса.
3. Многие регуляторные белки у эукариот могут влиять на скорость транскрипции, даже если эти белки связываются с участками ДНК, расположенными за тысячи пар оснований от промотора.

(СЛАЙД 17)

Транскрипция у эукариот, в отличие от прокариот, обеспечивается тремя различными ДНК-зависимыми РНК-полимеразами:

1. РНК-полимераза I транскрибирует гены рибосомных РНК (5,8S РНК, 18S РНК и 28S рРНК).
2. РНК-полимераза II обеспечивает транскрипцию генов, кодирующих белки, а также малых ядерных РНК (за исключением гена U6).
3. РНК-полимераза III транскрибирует гены тРНК, 5S рРНК, 7SL РНК и U6 РНК. Особые РНК-полимеразы обеспечивают транскрипцию клеточных органелл эукариот - хлоропластов и митохондрий. РНК-полимераза хлоропласт сходна с РНК-полимеразой бактерий. РНК-полимеразы митохондрий очень простые, по простоте сходны с РНК-полимеразами некоторых вирусов бактерий. Ген, кодирующий митохондриальную РНК-полимеразу, располагается в ядре.

РНК-полимеразы архебактерий похожи на РНК-полимеразы эукариот.

(СЛАЙД 18)

Базальные факторы транскрипции необходимы для инициации транскрипции всех эукариотических генов. Они связываются с РНК-полимеразой II, формируя комплекс, окружающий стартовую точку. Они также определяют сайт инициации. Базальные транскрипционные факторы вместе с РНК-полимеразой составляют базальный транскрипционный комплекс.

Базальные факторы транскрипции в настоящее время очищены и выделены. Это белки TFIIA, TFIIB, TFIID, TFIIЕ, TFIIF, TFIIN. Также выделяют базальный транскрипционный фактор TFIIJ. "II" - потому что РНК-полимераза II. Вместе с РНК-полимеразой II эти белки могут иницировать транскрипцию во внеклеточных системах *in vitro*.

Формирование белкового комплекса на промоторной последовательности начинается с того, что фактор транскрипции TFIID связывается с ТАТА-боксом. Затем к этому комплексу присоединяется фактор TFIIB, который связывается с РНК-полимеразой и TFIIF - образуется так называемый минимальный комплекс инициации транскрипции.

Для начала транскрипции РНК-полимераза должна освободиться от комплекса факторов транскрипции. Для этого к транскрипционному комплексу присоединяются транскрипционные факторы TFIIN и TFIIЕ. Фактор TFIIN фосфорилирует РНК-полимеразу, после чего она изменяет конформацию, освобождается от всех факторов транскрипции и начинает транскрипцию.

Итак, сборка инициаторного транскрипционного комплекса (транскриптосомы) начинается со связывания с промотором фактора TFIID, затем последовательно присоединяются TFIIB, РНК-полимераза II и TFIIF генов. В состав белкового фактора TFIID входят TBP - TATA-box Binding Protein и 8 ассоциированных факторов ТАФ (TAF - Transcription Associated Factors). В дальнейшем к комплексу присоединяются факторы TFIIЕ, TFIIN, а также коактиваторы (TFIIA и белки TAF). В состав инициаторного комплекса входит около 30 полипептидов

Комплекс РНК-полимеразы и базальных транскрипционных факторов может функционировать с низкой скоростью. Необходимый уровень активности требует участия специфических факторов транскрипции. Это регуляторные белки, которые контролируют процесс образования базального транскрипционного комплекса.

Специфические транскрипционные факторы обладают двумя важными свойствами:

1. опознают специфические последовательности, расположенные в промоторах, энхансерах и других регуляторных элементах гена
2. связываются с другими компонентами транскрипционного комплекса после своего присоединения к ДНК. Поэтому у специфических транскрипционных факторов есть специальные домены, отвечающие каждой из этих функций.

(СЛАЙД 19)

Рассмотрим теперь более подробно строение 5'-регуляторных районов генов эукариот. Каждый 5'-регуляторный район эукариот содержит сайт инициации транскрипции - участок, начиная с которого по матрице ДНК начинается синтез РНК. Кроме того, 5'-регуляторный район содержит большое количество сайтов связывания транскрипционных факторов и других типов регуляторных элементов. Каждый эукариотический ген, кодирующий белок, имеет 5'-регуляторный район, необходимый для формирования мультимерного транскрипционного комплекса. Обязательной компонентой этого комплекса является РНК-полимераза II, обеспечивающая ДНК-зависимый синтез РНК.

Строгий консенсус ТАТА-бокса связывающего белка (ТВР) имеет вид ТАТААААА, тогда как его расширенный аналог в вырожденном 15-буквенном коде имеет вид STWTAWADRSSSSSS

Анализ нуклеотидных последовательностей, обладающих повышенным сродством к ТВР, выявил еще более длинный расширенный консенсус, имевший вид NNNCNGSSSSCCCTTTWWWWAAAGSSSSSSSCNNG.

Заметно, что (А+Т)-богатое "ядро" консенсуса (подчеркнуто внизу) окружено G+C богатыми флангами. Можно предположить, что наблюдаемые в вырожденном коде закономерности важны для правильного взаимодействия ДНК ТАТА-бокса с белком ТВР.

Фактором транскрипции называется любой белок, который необходим для инициации транскрипции, но не является собственно РНК-полимеразой. Регуляторные районы выявляют экспериментально с помощью мутаций, которые изменяют скорость транскрипции, а также другими методами. Регуляторные последовательности ДНК необходимы как для инициации транскрипции, так и для регулирования ее скорости.

Промоторный район, таким образом, состоит из участка, на котором образуется комплекс из РНК-полимеразы II и базальных факторов транскрипции, а также многочисленных регуляторных элементов последовательности, с которыми связываются регуляторные белки, называемые специфическими факторами транскрипции. Регуляторные элементы гена могут быть расположены как выше, так и ниже точки начала транскрипции.

Коровый промотор содержит старт транскрипции и охватывает область от -60 до +40 п.о. по отношению к нему.

Коровый промотор содержит в своем составе ряд коротких функционально значимых мотивов размером до 5-25 п.о. Среди них наиболее полно изучены ТАТА-бокс, называемый также блоком Хогнесса, инициатор (Inr-элемент), СААТ-бокс и GC-бокс.

СААТ-бокс имеет консенсусную последовательность GGCCAATCT. GC-бокс имеет консенсусную последовательность GGCCGG. Некоторые промоторы имеют более одного элемента СААТ и GC.

Промоторы могут содержать различные комбинации регуляторных элементов, но ни один элемент не встречается во всех промоторах.

ТАТА-бокс представляет собой А/Т-богатую последовательность, находящуюся на расстоянии 25-35 п.о. выше старта транскрипции. Канонический консенсус ТАТА-бокса имеет последовательность ТАТААААА. Обычно ТАТА-бокс с обеих сторон окружён G/C-богатыми районами. По наличию или отсутствию ТАТА-бокса все промоторы делятся на две группы: ТАТА-содержащие (ТАТА+) и ТАТА-несодержащие (ТАТА-). ТАТА-бокс не является обязательным элементом промоторов эукариот. Для человека, например, около 70% промоторов генов обладают ТАТА-боксом

Inr-элемент непосредственно содержит старт транскрипции (+1 нуклеотид транскрипта). По сравнению с ТАТА-боксом он имеет менее выраженный консенсус. Для человека он имеет вид YYA₊₁NWYY. В работе проведена классификация Inr-элементов на группы по сходству последовательностей. Показано, что в отсутствии ТАТА-бокса, именно Inr-элемент имеет ключевое значение для правильного позиционирования РНК-полимеразы.

С Inr-элементом способны специфично взаимодействовать различные транскрипционные факторы. В группе ТАТА-несодержащих промоторов *Drosophila melanogaster* вблизи позиции +30 найден DPE-элемент, являющийся функциональным аналогом ТАТА-бокса

(СЛАЙД 20-22)

В целом, для 5'-регуляторных районов генов характерно очень сложное строение и наличие большого количества регуляторных элементов. Обязательным элементом, абсолютно необходимым для инициации транскрипции, является коровый (базальный) промотор. Под коровым промотором понимают минимальную последовательность ДНК, необходимую для правильной инициации транскрипции гена *in vitro*. Коровый промотор - ключевой элемент 5'-регуляторного района гена, обеспечивающий сборку базального транскрипционного комплекса и инициацию транскрипции на базальном уровне.

Важную роль в транскрипционном контроле генов эукариот играют удалённые регуляторные элементы (энхансеры, сайленсеры), располагающиеся на расстоянии в тысячи п.о. от старта транскрипции. По этой причине все регуляторные элементы разделяют на проксимальные (располагающиеся непосредственно вблизи старта транскрипции) и дистальные (удалённые). В частности, базальный промотор относится к группе проксимальных регуляторных элементов.

Регуляторные районы генов эукариот имеют блочно-иерархическую организацию. Экспрессия гена может контролироваться следующими регуляторными районами гена: коровым промотором, энхансерами (усилителями транскрипции), или сайленсерами (подавляющими транскрипцию районами), которые могут быть расположены за многие тысячи п.о. от старта транскрипции. Каждый из перечисленных выше регуляторных районов содержит в своем составе сайты связывания определенных транскрипционных факторов (ССТФ). Два соседних ССТФ могут представлять композиционный элемент (рис. 1-4). В этом случае их совместное действие согласовано, то есть его эффект значительно отличается от действия каждого ССТФ в отдельности. Встречаемость и расположение ССТФ в 5'-регуляторных районах генов часто отражает ткане- или стадие-специфичные особенности регуляции их экспрессии. Разнообразие строения промоторных районов генов создаёт наибольшие трудности при создании программ распознавания промоторов.

Блочно-иерархическая организация, характерная для регуляторных районов генов эукариот (Dyana, 1989; Arnone and Davidson, 1997), обеспечивает возможность гибкой регуляции транскрипции. Стадии клеточного цикла, функционального состояния клетки, ткани, органа, стадии индивидуального развития, действия внешних индукторов определяют присутствие в ядре клетки некоторого набора транскрипционных факторов, которые взаимодействуют с регуляторными элементами конкретного гена. В результате этого формируется уникальный транскрипционный комплекс, обеспечивающий необходимый уровень транскрипции гена в конкретной клеточной ситуации.

Существенная особенность 5'-регуляторных районов генов эукариот - их большая длина, иногда достигающая десятков тысяч п.о., что на порядки больше максимального размера регуляторных районов прокариот, который, например, для *E.coli* имеет длину не более 450 п.о.

(СЛАЙД 23-24)

Потенциально на основе одного и того же набора регуляторных элементов (блоков), входящих в состав определенного регуляторного района возможно формирование огромного разнообразия транскрипционных комплексов и, как следствие этого, реализация соответствующего разнообразия паттернов экспрессии.

(СЛАЙД 25)

Лавинообразный рост объема информации по нуклеотидным последовательностям в большой степени превышает возможности экспериментальной функциональной аннотации секвенированных последовательностей. В то же время, функциональная аннотация совершенно необходима для понимания механизмов функционирования генома, что, в свою очередь, имеет неоспоримую ценность как для фундаментальной и прикладной науки, так и в клинической практике, профилактике и лечении заболеваний человека, в том числе наследственных заболеваний. В связи с этим, представляет большую актуальность разработка компьютерных методов функциональной аннотации новых генов и целых геномов, основывающаяся на основе последовательностей ДНК и информации, накапливаемой в базах данных по регуляции экспрессии генов эукариот. Большие объемы как секвенированных последовательностей, так и информации по регуляции экспрессии эукариотических генов, а также сложность и большая биологическая специфика этой информации накладывают особые требования на разрабатываемые методы компьютерного анализа регуляторных районов, контролирующих экспрессию генов эукариот, и предсказания этих районов вдоль сверхдлинных нуклеотидных последовательностей. Эти требования включают в себя адекватное отражение биологической специфики и сложного строения данных по регуляции экспрессии генов, способность работать с гетерогенными, разнокачественными и содержащимися в различных базах данными, и могут быть выполнены только при привлечении современных компьютерных технологий и методов анализа данных.

Таким образом, Компьютерный анализ геномных последовательностей приобретает в последнее время всё большее значение. Он может не только объяснять особенности структурно-функциональной организации известных районов геномов, но также дает возможность предсказания особенностей структурно-функциональной организации вновь секвенированной геномной ДНК.

(СЛАЙД 26-29)

Одной из важных проблем современной информационной биологии является построение компьютерных методов распознавания функциональных ДНК-сайтов.

Методы распознавания функциональных сайтов Основаны на предположении, что существуют какие-то особенности последовательности (сигналы), общие для всех сайтов данного типа, которые распознаются соответствующими ДНК- или РНК- связывающими белками или рибонуклеопroteинами. Основными такими сигналами принято считать предпочтения определенных нуклеотидных оснований в определенных позициях сайтов; распределения непозиционированных олигонуклеотидов, элементы вторичной структуры РНК, параметры двухцепочечной ДНК, симметричные последовательности и др.

Свойства ДНК

ДНК - гетерогенная полимерная молекула, обеспечивающая хранение и передачу генетической информации в ряду поколений.

Наиболее распространена *in vivo* В-форма ДНК, представляющая собой правозакрученную двойную спираль, состоящую из двух цепей, положение которых относительно друг друга фиксируется за счет формирования водородных связей между парами комплиментарных оснований. Одному шагу спирали соответствует одна комплиментарная пара оснований. Для В-формы ДНК на один виток спирали приходится в среднем 10.5 п.о.; поперечный размер спирали составляет 19 Å.

(СЛАЙД 30-33)

Установлено, что локальные конформационные особенности двойной спирали ДНК, а также определяющие их параметры зависят от нуклеотидного контекста ДНК. Это относится и к физико-химическим свойствам ДНК, например, температуре плавления.

Для всех приведенных параметров характерна выраженная зависимость их величины от динуклеотидного контекста. Например, максимальное значение параметра Twist - 39.5° характерно для динуклеотида ТА, а минимальное - 29.3° характерно для динуклеотида АТ. При этом различие между минимальной и максимальной величинами составляет 10.2° . Столь же выраженные зависимости от локального нуклеотидного контекста свойственны и для других рассмотренных в качестве примеров конформационных и физико-химических свойств.

(СЛАЙД 34)

В Институте цитологии и генетики СО РАН создана уникальная, не имеющая аналогов в мировой науке [Интернет-доступная база данных TRRD по транскрипционным регуляторным районам генов эукариот](http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/) (<http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/>). База TRRD (Transcription Regulatory Regions Database) предназначена для накопления экспериментальных данных по регуляции транскрипции генов эукариот и включает информацию о регуляторных районах, составляющих разные уровни сложной иерархической системы регуляции транскрипции - сайтах связывания транскрипционных факторов, композиционных элементах, промоторах, энхансерах, сайленсерах, протяженных районах регуляции транскрипции.

(СЛАЙД 35)

В нашей лаборатории разработана программа TRRD-Pars, реализованная в среде Visual C++. Она осуществляет динамическую верификацию информации о нуклеотидных последовательностях и позициях сайтов связывания транскрипционных факторов из TRRDSITES используя последовательности, из EMBL/GenBank, ссылки на которые записаны в TRRD. Используя информацию, содержащуюся в TRRD, EMBL, GenBank программа создает выборки последовательностей сайтов связывания различных транскрипционных факторов.

(СЛАЙД 36)

Поиск сайтов по консенсусу

Одним из простейших методов распознавания сайтов является поиск сайтов по консенсусу. Чтобы построить консенсус, рассматривают набор выровненных последовательностей. В каждой позиции находят основание, наиболее часто встречающееся по сравнению с ожидаемой частотой. Для этого используют разные статистические критерии или информационное содержание. Рассмотрение может проходить в 4-х или 15-буквенном коде. Консенсусы используют, когда число известных сайтов мало. Этот метод дает очень грубую функциональную разметку последовательности.

Консенсус является распространенным способом описания функциональных сайтов, включая сайты связывания транскрипционных факторов. При составлении консенсуса рассматривается выборка выровненных сайтов. Каждой позиции сайта приписывается наиболее консервативный нуклеотид (если он имеется в этой позиции). В том случае, когда для данной позиции консервативным является определенное сочетание нуклеотидов, для описания консенсуса используется 15-ти буквенный код IUPAC-IUB (Bishop, 1994). Например, выборке цис-элементов NF-IL6:

```
TGTAGTAAG
TTATGGAAT
TTCCGCAAT
TGGTGAAAT
TTACGCAAG
TGTAGTAAG
TTCGGAAAT
TTGGGCAAG
TGATGGAAG
```

соответствует консенсус

TKNNGNAAK

основным недостатком консенсуса является грубость описания им реальных сайтов, что приводит, как это будет показано ниже (см. раздел 2.3), к большим ошибкам при распознавании сайтов.