

Биоинформатика

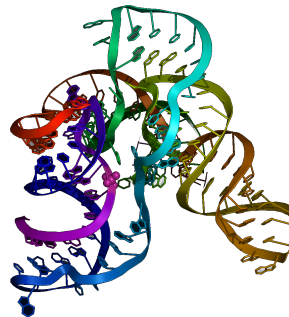
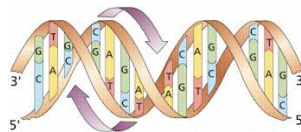
Андрей Александрович Миронов

Факультет биоинженерии и биоинформатики

РНК - одноцепочечная полинуклеотидная цепь

ДНК vs РНК

- в РНК используется рибоза вместо дезокси-рибозы
- в РНК обычно U вместо T
- ДНК – обычно двунитевая полинуклеотидная цепь
- РНК – обычно однонитевая полинуклеотидная цепь
- нуклеотиды в РНК образуют комплементарные пары.
- РНК может образовывать сложную структуру



Информационные функции РНК

Информационные функции

- перенос информации от ДНК к белку
- т-РНК — соответствие между кодоном и аминокислотой
- геномы вирусов
 - двунитевая РНК: ротавирусы, вирусы без симптомов
 - РНК+ полиомиелит, менго, риновирусы (насморк), ящур
 - РНК- грипп,

Каталитические функции РНК

Каталитические функции

- Синтез белка — катализируется РНК
- Процессинг РНК
 - сплайсинг — катализируется РНК
 - авто-сплайсинг (самосплайсирующиеся интроны)
 - Рибонуклеаза Р — фермент процессинга тРНК
 - Другие рибозимы
- Гипотеза РНК-мира

Направляющие функции РНК

Направляющие функции

- редактирование РНК
- модификация рРНК, тРНК
- микро-нуклеус → макро-нуклеус
- внутренняя инициация трансляции — IRES
- репарация ДНК

Регуляторные функции РНК

Регуляторные функции

- терминация транскрипции прокариот
- аттенюаторы
- секвесторы трансляции
- рибопереключатели
- Т-боксы
- микро РНК
- структуры, регулирующие сплайсинг
- PIWI - РНК
- компенсация дозы гена X-хромосомы
- контроль репликации
- ...

Прочие функции РНК

Прочие функции

- теломеразная РНК
- тмРНК
- затравки для репликации

Неизвестные функции РНК

Неизвестные функции

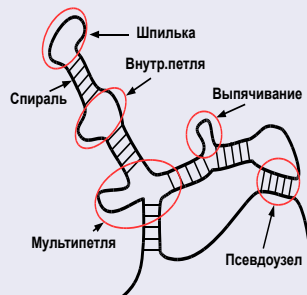
- длинные некодирующие РНК
- очень длинные некодирующие РНК
- короткие некодирующие РНК
- энхансерные и промоторные РНК

Вторичная структура РНК

Вторичная структура - совокупность спаренных оснований

Элементы вторичной структуры

- **спираль** = посл. спаренных оснований
- **шпилька** = петля, 1 спираль
- **внутр.петля** = 2 спирали
- **мультипетля** >2 спиралей
- **выпячивание** = 2 спирали, по одной нити пары подряд



Экспериментальные данные

Эксперименты

- специфическая модификация спаренных или неспаренных оснований
- обратная транскрипция
 - на модифицированных основаниях срывается — смотрим где чаще срывы
 - на модифицированных основаниях делает ошибки — смотрим частоту ошибок
- массовое секвенирование и анализ результатов → вероятность спаривания

Золотой стандарт

Золотой стандарт

- Экспериментально определенные структуры
- Консервативные структуры (например, тРНК)

Алгоритм Нуссинофф

Оптимальная структура на сегменте $[i, j]$

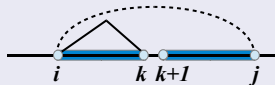
i не спарено



i спарено с j



i спарено с $i < k < j$



рекурсия

$$S(i, j) = \max \begin{cases} S(i, j - 1) \\ S(i + 1, j - 1) + 1 \\ S(i + 1, k) + 1 + S(k + 1, j) \end{cases}$$

Оптимизация энергии

Энергия вторичной структуры

- Энергия водородных связей — связи между основаниями на разных цепях
- Энергия стэкинга — взаимодействие оснований в стопке
- Энергия петель — энтропийный вклад

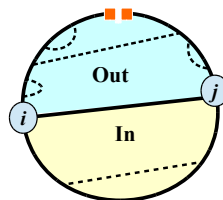
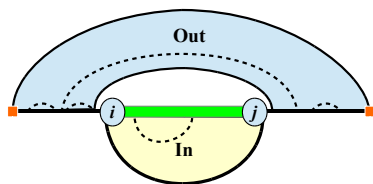
Задачи и алгоритмы

- минимизация полной энергии вторичной структуры
 $\Delta G \rightarrow \min$
- вычисление вероятностей спаривания
$$Z = \sum_{structures} \exp\left(-\frac{\Delta G}{kT}\right); p(ij) = \frac{Z(ij)}{Z}$$
- поиск субоптимальных структур
- учет экспериментальных данных

Статистическая сумма

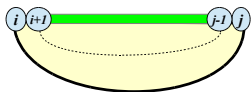
Условная стат.сумма: сумма по всем структурам, где $i \diamond j$:

$$Z(i, j) = \sum_{struct|i \diamond j} \exp\left(-\frac{\Delta G}{kT}\right) = \sum_{struct|i \diamond j} \exp\left(-\frac{(\Delta G_{in} + \Delta G_{out})}{kT}\right)$$

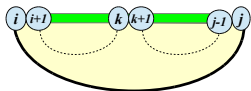


$$Z(i, j) = \sum_{In\ struct} \exp\left(-\frac{\Delta G_{In}}{kT}\right) \cdot \sum_{Out\ struct} \exp\left(-\frac{\Delta G_{Out}}{kT}\right)$$

Статистическая сумма



$$In(i + 1, j - 1) \cdot \exp\left(-\frac{\Delta G(i, j)}{kT}\right)$$



$$(In(i + 1, k) + In(k + 1, j - 1)) \cdot \exp\left(-\frac{\Delta G(i, j)}{kT}\right)$$

Внутренняя стат. сумма

$$In(i, j) = In(i + 1, j - 1) \cdot \exp\left(-\frac{\Delta G(i, j)}{kT}\right) + \sum_k (In(i + 1, k) + In(k + 1, j - 1)) \cdot \exp\left(-\frac{\Delta G(i, j)}{kT}\right)$$

Статистическая сумма

Вероятность спаривания:

$$P(i \diamond j) = \frac{In(i, j) \cdot Out(i, j)}{In(0, L)}$$

Сравнительный анализ

Сравнительный анализ

Если есть много последовательностей, для которых мы ожидаем схожую структуру, то надо искать общую структуру для них

- Строим выравнивание, потом ищем общую структуру.
- Ищем общую структуру вместе с построением выравнивания.

Сравнительный анализ

Предсказания

- транспортные РНК (Robert W. Holley, Hans Zachau)
- рибосомные РНК
- рибопереключатели
- рибозимы

Рибопереключатели

Исходные данные

- Есть перед одним из генов биосинтеза флавинов область ок. 150 нукл.:
 - некоторые мутации снимают регуляцию
 - некоторые мутации не меняют регуляцию
- Есть похожие области перед генами синтеза флавинов почти во всех геномах

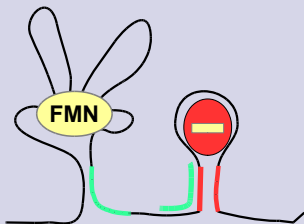
Рибоперекляатели

	1	2	2'	3	Add.	3'	Variable	4	4'	5	5'	1'	
	====>>>>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	<====>>>	
BS	TTGATCTCTGAGGGG	-CA	TGGAAAT	GAGCGG	CGT	21	AGC	CGAC	--	8	4	8	----TGGATTCAATTAA--CTGAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BQ	AGCATCTCTGAGGGG	-TC	TGAAATG	TAACCG	CGT	19	AGC	CGAC	--	8	5	8	----TGGATTCAATTAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BT	TTGATCTCTGAGGGG	-CA	TGGAAAT	GAGCGG	CGT	20	AGC	CGAC	--	8	4	8	----TGGATTCAATTAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BU	TTGATCTCTGAGGGG	-CA	TGGAAAT	GAGCGG	CGT	19	AGC	CGAC	--	10	4	10	----TGGATTCAATTAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CA	TGATCTCTGAGGGG	-CT	TGAAAT	GAGCGG	CGT	23	AGC	CGAC	--	8	4	8	----TGGATTCAATTAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CAN	GATCTCTGAGGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	3	4	3	----AGATCGGTTAAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CF	CTTACTCTGAGGGG	-TA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	3	4	3	----AGATCGGTTAAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CG	TTTACTCTGAGGGG	-CA	TGAAAT	TACCGG	CGT	7	AGC	CGAC	--	11	5	11	----CTGTCTGTAGATTTAGACCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
LLX	ATGAACTCTGAGGG	-A	TGTAAT	TACCGG	CGT	2	AGC	CGAA	--	4	4	4	----ATGATTCGTTGAAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
FN	AACATCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	3	4	3	----ATGTTTGTGAAATTCAAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
FW	AAAGCGCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	3	AGC	CGAA	--	5	4	5	----TTTCCCGCTGGGTAATTCGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BK	GCCTCTCTGAGGG	-B	TGAAAT	TACCGG	CGT	15	AGC	CGAA	--	8	12	9	----CCGTCGCTGCAGACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
TO	CACTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	3	AGC	CGAA	--	5	4	5	----CCGACCTCGGTAATTCGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
AO	AATAATCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	7	7	7	----AGGAACTGTGAGTTTGGTACCGGCAATGAA-AGTTGGATGGGAGAGATGAT
FU	TGTAATCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	13	4	12	----AGGACTGTGAAATTCAGTACCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CGI	GAAACTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	20	AGC	CGAA	--	3	4	3	----AGGAACTGTGAGTTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
DN	TAAACTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAA	--	5	4	5	----GATTGTGAAATTTAAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
TFU	ACCGCTCTGAGGG	-CT	TGAAAT	TACCGG	CGT	3	AGC	CGAC	--	8	5	8	----TGGAACTGTGAAACTTGGCCCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
SB	-AAGCGCACTCTGAGGG	-CTG	TGAAAT	TACCGG	CGT	3	AGC	CGAC	--	8	5	8	----TTGACCGATTGAAATTCGACCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BD	TTGGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	30	AGC	CGAGCG		137			GTGAGCACTCTGTGAGAACTTGAAGCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
BH	TEGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	21	AGC	CGAGCG		8	4	8	GTGAGCACTCTGTGAGAACTTGAAGCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
BEU	TTAGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	31	AGC	CGAGCG		7	5	7	GTGAGCACTCTGTGAGAACTTGAAGCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
RBC	TEAGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	21	AGC	CGAGCG		11	3	11	GTGAGCACTCTGTGAGAACTTGAAGCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
ESG	GCCTATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	17	AGC	CGAGCG		8	4	8	GACGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAGTAGTAG
FX	GCCTATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	67	AGC	CGAGCG		3	8		GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAGTAGTAG
BF	CTTACTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	20	AGC	CGAGCG		8	4	8	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAGTAGTAG
HL	TCCACTCTCTGAGGG	-A	TGAAAT	TACCGG	CGT	2	AGC	CGAGCG		26	9	30	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
VK	GCCTATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	14	AGC	CGAGCG		11	9	11	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
VC	CAATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	13	AGC	CGAGCG		5	4	5	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAGTAGTAG
YF	CCCTATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	40	AGC	CGAGCG		16	14		GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAGTAGTAG
AB	GCCTATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	25	AGC	CGAGCG		16	4	27	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
BIP	TEAGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	18	AGC	CGAGCG		10	4	10	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
AC	ACATCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	16	AGC	CGAGCG		10	3	11	----CCGACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
Spu	ACAAATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	34	AGC	CGAGCG		6	6		GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
FP	CTCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	13	AGC	CGAGCG		7	3		GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
AU	GCCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	17	AGC	CGAGCG		7	9	7	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
FU	AAAGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	19	AGC	CGAGCG		19	4	18	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
MY	TAAATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	19	AGC	CGAGCG		15	4	16	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
HLA	TTACTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	19	AGC	CGAGCG		14	4	13	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
ME	TAAATCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	16	AGC	CGAGCG		8	5	8	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
SME	AAAGCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	34	AGC	CGAGCG		8	3	8	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
BN	GCCTCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	17	AGC	CGAGCG		10	15	10	GTGAGCACTCTGTGATTTGGCCCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT
BS	AATCACTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	18	AGC	CGAGCG		5	4	5	----AGTTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BQ	TTGATCTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	27	AGC	CGAGCG		3	3		----AGTTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
BE	ATTCACTCTCTGAGGG	-A	TGAAAT	TACCGG	CGT	20	AGC	CGAGCG		3	4	3	----AGGATCGGTTGGGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
CF	AATCACTCTCTGAGGG	-A	TGAAAT	TACCGG	CGT	2	AGC	CGAGCG		3	4	3	----TATGATCGTTTGTGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
DA	GAATCTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAGCG		6	4	6	----CTTGTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
DF	CTTACTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	14	AGC	CGAGCG		5	3		----ATGATTTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
LXO	AAATATCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	21	AGC	CGAGCG		4	4	4	----TTGATGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
LL	GTTACTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	3	AGC	CGAGCG		3	10	3	----TTCAGCTCTGTGATTTGACGACCGGCAATGAA-AGTTGGATGGGAGAGATGAT
FN	AGAGCTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	125	AGC	GTG		--	3	4	----GATGTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
ST	AAATCTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	14	AGC	CGAGCG		3	4	3	----CTTGTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
HN	AAATCTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	10	AGC	CGAGCG		3	4	3	----ATGATTTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
SA	ATTCACTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	6	AGC	CGAGCG		11	3	11	----CTGTCTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
AM	TCCAGTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	14	AGC	CGCC		--	5	5	----TGATCTCTGTGAACTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
DLA	ACCAATCTCTGAGGG	-TA	TGAAAT	TACCGG	CGT	20	AGC	CGAAC		--	11	11	----CCGACTCTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
FN	AATCACTCTCTGAGGG	-CA	TGAAAT	TACCGG	CGT	2	AGC	CGAGCG		4	4	4	----CTTGTGTGAGATTTGAAGCGCGGCAATGAA-AGTTGGATGGGAGAGATGAT
GBU	TTTTCTCTCTGAGGG	-C	TGAAAT	TACCGG	CGT	28	AGC	CGAGCG		10	4	10	GTGAGCACTCTGTGATTTAACTTGAAGCGCGGCAATGAGTAGCGGATGGATGGAAAGATGAT

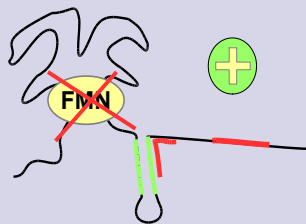
Рибопереключатели

Механизм работы

Есть флаavin → запрет



Нет флавина → разрешение



Экспериментальные подтверждения

- разрушение структуры разрушает регуляцию
- рентгеноструктурный анализ

Рибопереключатели

Рибопереключатели

- Cobalamin
- FMN – flavin
- Glutamine
- Glycine
- Lysine
- PreQ1 – pre-queuosine1
- Purine riboswitches
- SAH – S-adenosyl homocysteine
- SAM – S-adenosyl methionine
- TPP – thiamin pyrophosphate
- ...

рибопереключатель Lis

