

Поиск сигналов в последовательности ДНК

План

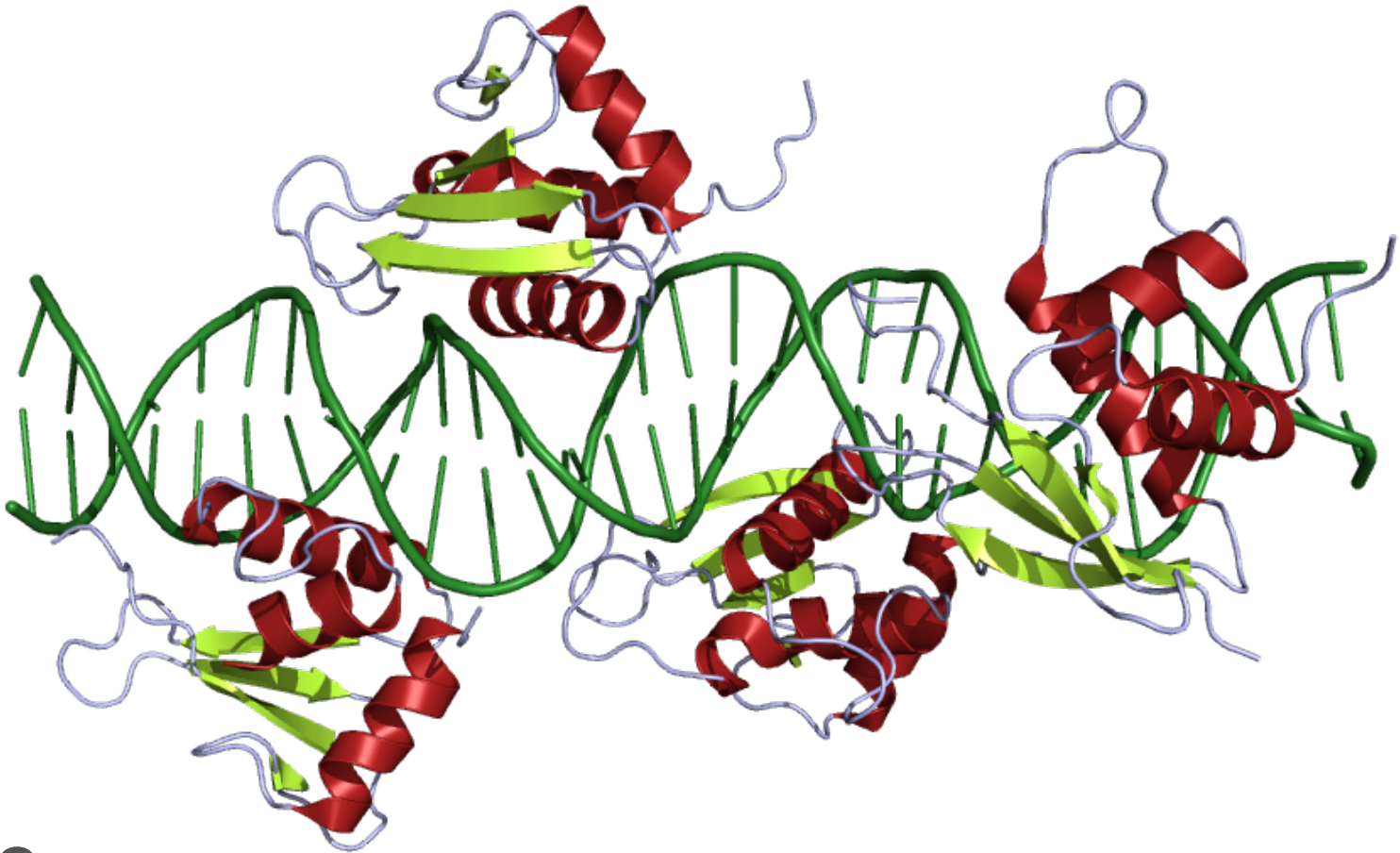
- Что такое сигнал — сайт связывания ТФ, и т. д. Сильные и слабые сайты?
- Умные слова мотив, паттерн, консенсус профиль
- БД мотивов — TransFac, JasPAR, RegTransBase
- Выявление и поиск мотивов (построение мотива по имеющимся данным и поиск нового мотива в последовательности):
 - НММ — требует большой объем данных
 - PFM/PWM — программа MEME

Сайты и белки



Painting published on cover of magazine: Nature Structural and Molecular Biology, September 1994, Volume 1 No. 9.

INTERFERON REGULATORY FACTOR 1



2pi0

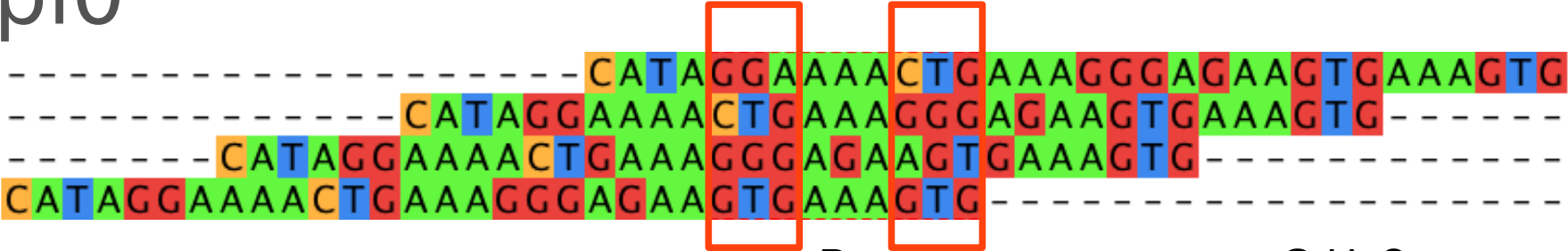
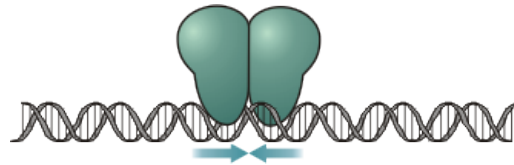


Рисунок предоставлен О.Н. Занегиной

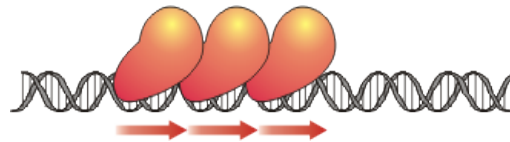
❖ ДНК-связывающие белки и их сигналы

□ Кооперативные однородные

- Палиндромы

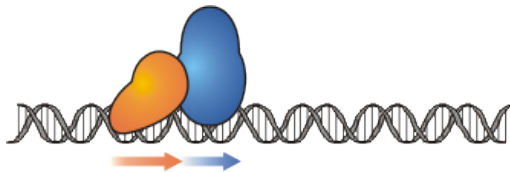


- Прямые повторы



□ Кооперативные неоднородные

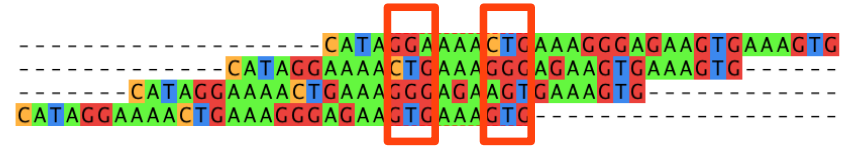
- Кассеты



□ Другие

Некоторые термины

- Мотив - сигнал, который есть у последовательностей, с которыми белок связывается и нет у последовательностей, с которыми белок не связывается.



- Профиль — способ отображения мотива, основанный на выравнивании



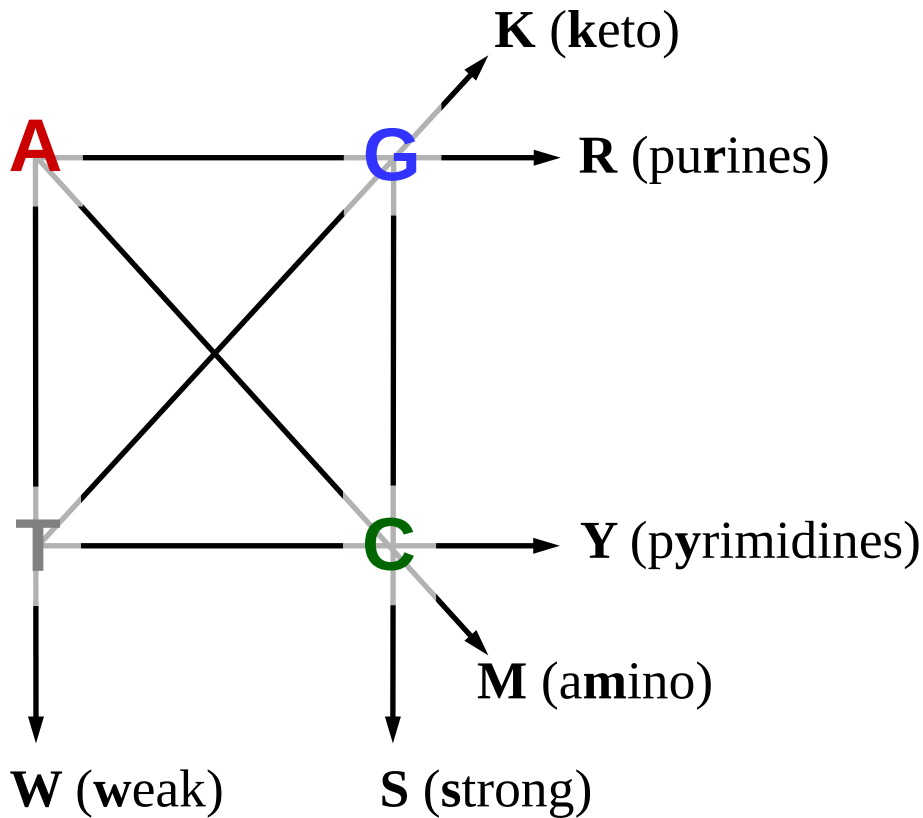
- Паттерн — вид профиля, в котором указываются все основания/аминокислоты, встречающиеся в данной позиции без указания частот



- Консенсус — способ отображения мотива, в котором в каждой позиции указывается самое частое основание/аминокислота



Для справки: Ambiguity codes



C/G/T (“не A”) → **B**

A/G/T (“не C”) → **D**

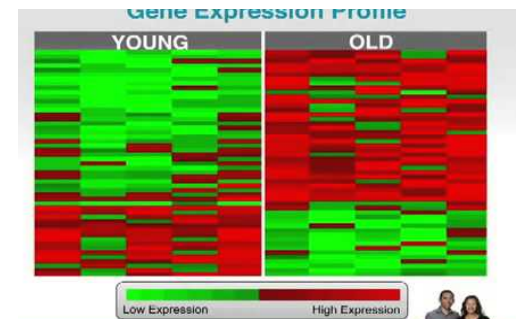
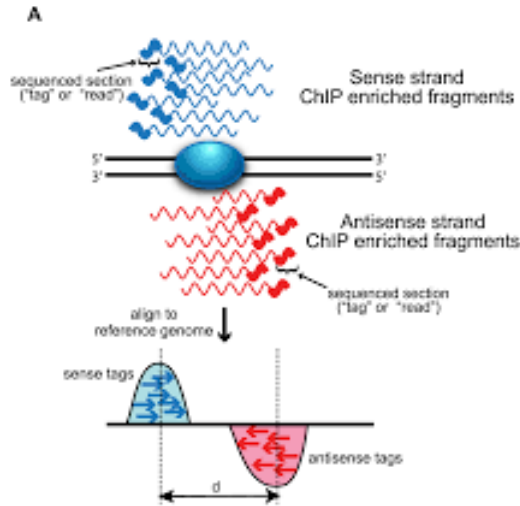
A/C/T (“не G”) → **H**

A/C/G (“не T”) → **V**

A/C/G/T → **N** (nucleotide)

Источник: РГМ

Исходные данные для поиска мотива (примеры)



- Данные экспериментов ChIP-Seq

- Upstream области коэкспрессирующихся генов

CCTACGCAAACGTTTTCTTTTT
GTCTCGCAAACGTTTGCTTTCC
CACACGCAAACGTTTTCGTTTA
TCCACGCAAACGGTTTCGTCAG
GCCACGCAACCGTTTTCTTGC
GATACGCAAACGTGTGCGTCTG
CCGACGCAATCGGTTACSTTGA
GTTGCGCAAACGTTTTCTGTTAC

Методы обнаружения МОТИВОВ

Поиск точных последовательностей

Ищем подпоследовательности:

- которые часто встречаются в наборе последовательностей, связывающихся с белком, и
- не встречаются в контрольном наборе

Недостатки — ищет более-менее точное совпадение, в то время как большинство сайтов связывания ТФ устроены более сложно

Поиск мотива с использованием позиционно-весовой матрицы

Информационное содержание и энтропия Шеннона

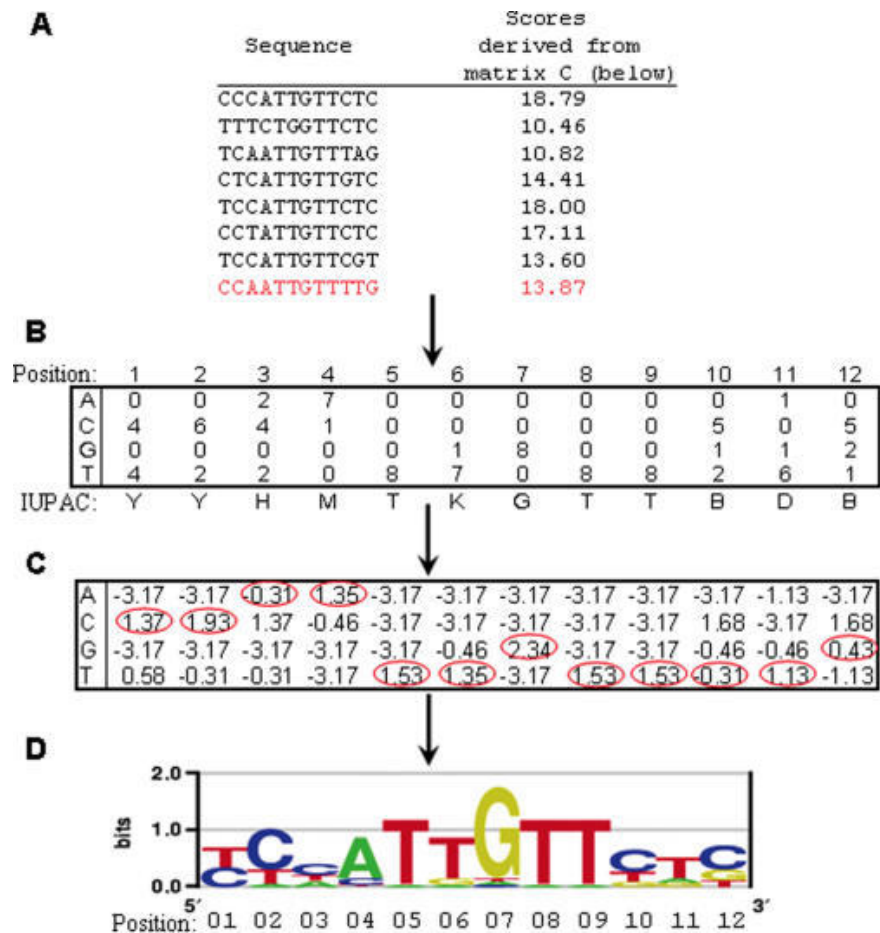
$$H = - \sum_{i=1}^N p_i \cdot \log_2 (p_i)$$

где H — энтропия, p_i —
вероятность i -го события

Информационное содержание (I) —
мера уменьшения неопределенности
после получения некоторого
сообщения

$$I = H_{\text{before}} - H_{\text{after}}$$

Поиск мотива с использованием позиционно-весовой матрицы



Вес ($I(b_j)$) основания b в данной позиции j
 $I(b_j) = f(b_j) \log f(b_j) - p(b) \log p(b)$,
 где $f(b_j)$ — частота основания b в позиции j выравнивания, $p(b)$ — фоновая частота основания b

Вес позиции — сумма по столбцу,
 вес мотива — сумма весов позиций

Поиск мотивов программой MEME

Дано: длина искомого сигнала;
набор последовательностей, в которых ищется сигнал

1. Последовательно берем фрагмент заданной длины в каждой последовательности, ищем похожие фрагменты в других последовательностях, строим выравнивание. Берем базовые частоты букв из дополнения.
2. Для каждого выравнивания получаем PWM с максимальным весом, используя алгоритм EM (Expectation maximization)
3. Выбираем заданное число PWM с лучшим весом

Алгоритм EM (Expectation maximization)

- Строим PWM по выравниванию:
- По очереди удаляем фрагмент из выравнивания, и заменяем его на лучший по PWM фрагмент в соответствующей последовательности
- Находим максимальный вес, записываем PWM с максимальным весом

Ограничения MEME

1. Предположение о независимости позиций выравнивания
2. Находит мотив без гэпов
3. Последовательности должны быть как можно короче и содержать минимум шума
4. После 40 последовательностей, включение дополнительных последовательностей не улучшает работу алгоритма

Набор программ для работы с МОТИВАМИ

Introduction - MEME Suite - Google Chrome

Бх Мi Se Se Pc A: A: со A: 40 jo Ge As Dε Inl Ev Ar Pr Inl Ml FlI m FlI M M St lir Pε A M Pε Bi Bi H(Pl (A x Anna

meme-suite.org

Сервисы Яндекс.Словари Расписание рейс National Center for Biotechnology Information BBC - Homepage home Official REBASE Home Import to Mendelius Другие закладки

The MEME Suite

Motif-based sequence analysis tools

```
graph TD; A[Your DNA, RNA or protein sequences] --> B[Motif Discovery: MEME, DREME, MEME-ChIP, GLAM2]; B --> C[Discovered motifs (de novo)]; D[Motif databases] --> E[Motif Enrichment: CentriMo, AME, SpaMo, GOMo]; F[Your DNA, RNA or protein motifs] --> E; G[GO databases] --> E; E --> H[Annotated motifs: GO function, GO compartment, GO process]; C --> I[Sequence databases]; H --> I; I --> J[Motif Scanning: FIMO, MAST, MCAST, GLAM2SCAN]; J --> K[Annotated sequences]; L[Motif databases] --> M[Motif Comparison: Tomtom]; K --> M; M --> N[Aligned motifs];
```

Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.

- Motif Discovery**
 - MEME
 - DREME
 - MEME-ChIP
 - GLAM2
- Motif Enrichment**
- Motif Scanning**
- Motif Comparison**
 - Tomtom
- Manual**
 - OVERVIEW
 - Motif Discovery**
 - MEME
 - DREME
 - MEME-ChIP
 - GLAM2
 - Motif Enrichment**
 - CentriMo
 - AME
 - SpaMo
 - GOMo
 - Motif Scanning**
 - FIMO
 - MAST
 - MCAST
 - GLAM2Scan
 - Motif Comparison**

MEME Multiple Em for Motif Elicitation	CentriMo Local Motif Enrichment Analysis	FIMO Find Individual Motif Occurrences
DREME Discriminative Regular Expression Motif Elicitation	AME Analysis of Motif Enrichment	MAST Motif Alignment & Search Tool
MEME-ChIP Motif Analysis of Large Nucleotide Datasets	SpaMo Spaced Motif Analysis Tool	MCAST Motif Cluster Alignment and Search Tool
GLAM2 Gapped Local Alignment of Motifs	GOMo Gene Ontology for Motifs	GLAM2Scan Scanning with Gapped Motifs
Tomtom Motif Comparison Tool	GT-Scan Identifying Unique Genomic Targets	

PMC1524905....png (Advances in P....pdf (Advances in P....pdf Ошибка: Не удалось ска chipseq_loos.pdf Показать все

From: **MEME Suite: tools for motif discovery and searching**

Nucleic Acids Res. 2009;37(suppl_2):W202-W208. doi:10.1093/nar/gkp335

TOMTOM OUTPUT	
Query File: ./output/tomtom_52109112_85850420/query.meme	
Target File: /www/tomcat/meme/db/motif_databases/macisaac_yeast.v1.meme	
Distance Measure: pearson	
All Motif Matches with <i>q</i> value at most: 0.5	
Target Motif: RGT1 Target Description: Query Motif: query Query Description: <i>p</i> -value: 0.00035 <i>q</i> -value: 0.087 Overlap: 6 Query Offset: 4 Orientation: + Figures: [EPS] [PNG]	
Total matches under <i>q</i> -value threshold: 0	
Command line:	<pre>/www/tomcat/meme/bin/tomtom -oc ./output/tomtom_52109112_85850420 -query ./output/tomtom_52109112_85850420/query.meme -target /www/tomcat/meme/db/motif_databases/macisaac_yeast.v1.meme -target-url-type macisaac -dist pearson</pre>

Figure Legend:

The figure shows the Tomtom output from searching a single DNA motif against a collection of yeast transcription factor binding site motifs identified via ChIP-seq (9). Tomtom shows that the query motif closely resembles the binding motif for transcription factor RGT1.

•

From: **MEME Suite: tools for motif discovery and searching**

Nucleic Acids Res. 2009;37(suppl_2):W202-W208. doi:10.1093/nar/gkp335

Sequence Analysis with fimo

Pattern Name	Sequence Name	Start	Stop	Score	p-value	q-value	Matched Sequence
1	NP_418484.4lyjcB	281	298	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418485.1lyjcC	149	132	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418031.1lyiaJ	175	158	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418032.1lyiaK	26	43	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418535.1lproP	37	54	19.7078	4.3e-08	0.0173	ATGTGTGAAGTTGATCAC
1	NP_414666.1lgcd	126	143	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC
1	NP_414667.4lhpt	80	63	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC

Figure Legend:

Полезные ссылки

- Анализ данных ChiPSeq
https://books.google.ru/books?hl=ru&lr=&id=YC2K_v1mficC&oi=fnd&pg=PA135&dq=makeev+vsevolod&ots=uSo84sL8A6&sig=xOTJH2RcPWhsjL7cBELQqkCkyfl&redir_esc=y#v=onepage&q=makeev%20vsevolod&f=true
- Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research. 2006;34(Web Server issue):W369-W373. doi:10.1093/nar/gkl198.
- Tran NTL, Huang C-H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biology Direct. 2014;9:4. doi:10.1186/1745-6150-9-4.
- Kulakovskiy IV, Makeev VJ. DNA sequence motif: a jack of all trades for ChIP-Seq data. Adv Protein Chem Struct Biol. 2013;91:135-71.

Сравнение двух PWM

- Как в MEME
- Vorontsov et co-authors
<https://almob.biomedcentral.com/articles/10.1186/1748-7188-8-23>