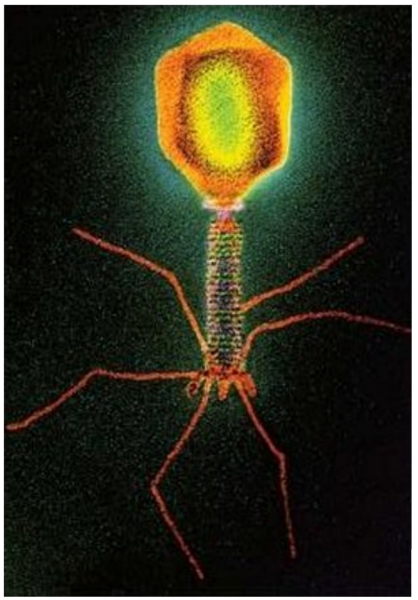
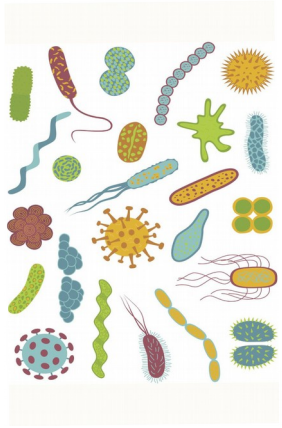
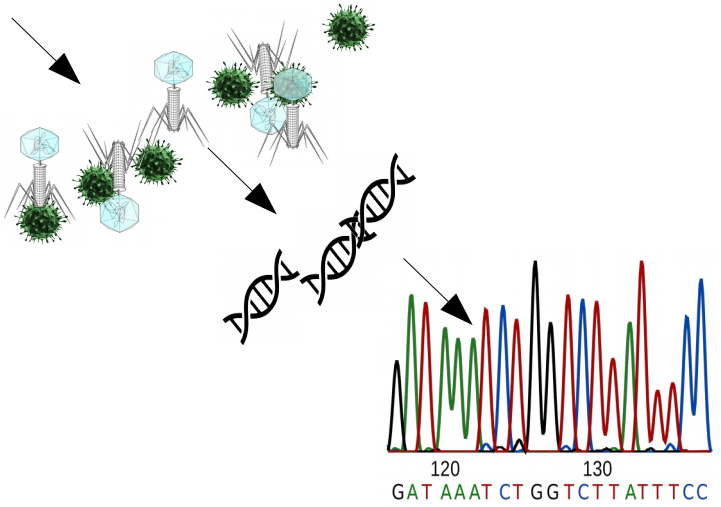
Биоинформатический анализ вирусных последовательностей метагенома

Елизавета Старикова
ФНКЦ ФХМ, лаборатория биоинформатики
estarikova@rcpcm.org



Метагеномика вирусов





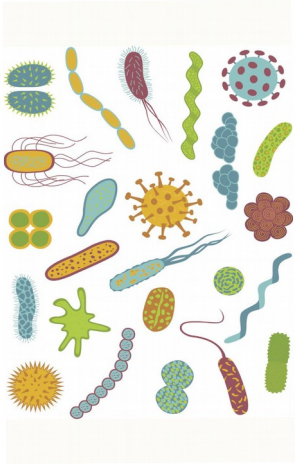
Образец среды

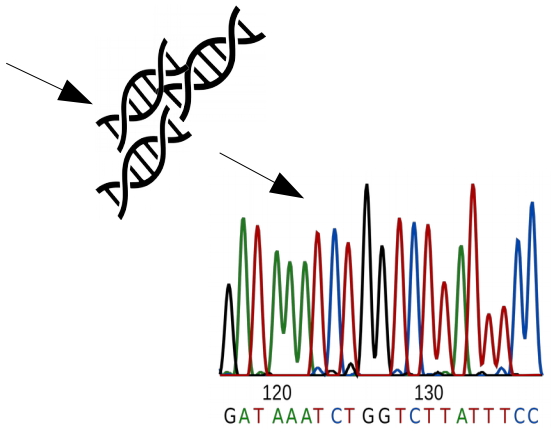
↓

Выделение ВПЧ

↓

Секвенирование





Тотальная ДНК из среды

↓

Секвенирование

↓

Биоинформатическое вылавливание вирусов

Метавиром океана

Метавиром человека: crAssphage

- Некий фаг с неизвестным хозяином (Bacteroides)?
- Распространен в кишечнике человека (до 80% ридов ВПЧ)

“Earth's virome”

- Из ~3000 метагеномов собрано 700 000 вирусных контигов, хозяева определены лишь для небольшой части

JGI IMG/VR INTEGRATED MICROBIAL GENOMES / VIRUS

Quick Genome Search: Go

My Analysis Carts**: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Genes

Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | My IMG | Data Marts | Help

IMG Viral Content

Viral Datasets	
Isolate Viruses (iVGs)	8382
Metagenomic Viral Contigs (mVCs)	706691
Total Viral Datasets	715073
Viral Clusters	
Viral Clusters	108954
Viral Singletons	260380
With Host	
Isolate Viruses (iVGs)	6912
Metagenomic Viral Contigs (mVCs): spacer hit	20781
Metagenomic Viral Contigs (mVCs): total	34213

Viral/Spacer BLAST

Download VPF Models
Download IMG/VR Database

The **IMG/VR** system (http://nar.oxfordjournals.org/content/early/2016/10/30/nar.gkw1030) serves as a starting point for the sequence analysis of viral fragments derived from metagenomic samples. Virus detection methods and host assignment

MENU nature International journal of science

Access provided by Scientific Research Institute of Physical-Chemical Medicine

Altmetric: 430 More detail »

Article

Uncovering Earth's virome

David Paez-Espino, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova & Nikos C. Kyrpides

Nature **536**, 425–430 (25 August 2016)
doi:10.1038/nature19094
Download Citation

Bacteriophages | Microbial ecology | Phage biology

Received: 23 November 2015
Accepted: 08 July 2016
Published: 17 August 2016

Как найти вирусные последовательности в WGS-метагеноме?

По гомологии:

- *По сходству с известными вирусными геномами*

  Недостатки:

  - текущие базы данных охватывают лишь ничтожную часть вирусного разнообразия

  - огромная вариабельность вирусов

- *По сходству с известными вирусными белками*

  Программы, вычисляющие “профаговые” и “фаговые” области по плотности фаговых белков: PHASTER, Phigaro, VirSorter

По составу:

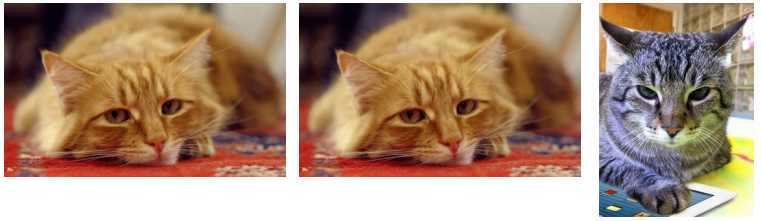
- Анализ кодонового состава, GC-состава и состава k-mer'ов

  Программы: MGTAXA и др.

Скрытые марковские модели (HMM)

Предположим, что изначальные состояния кота равновероятны

Состояния кота как марковская модель

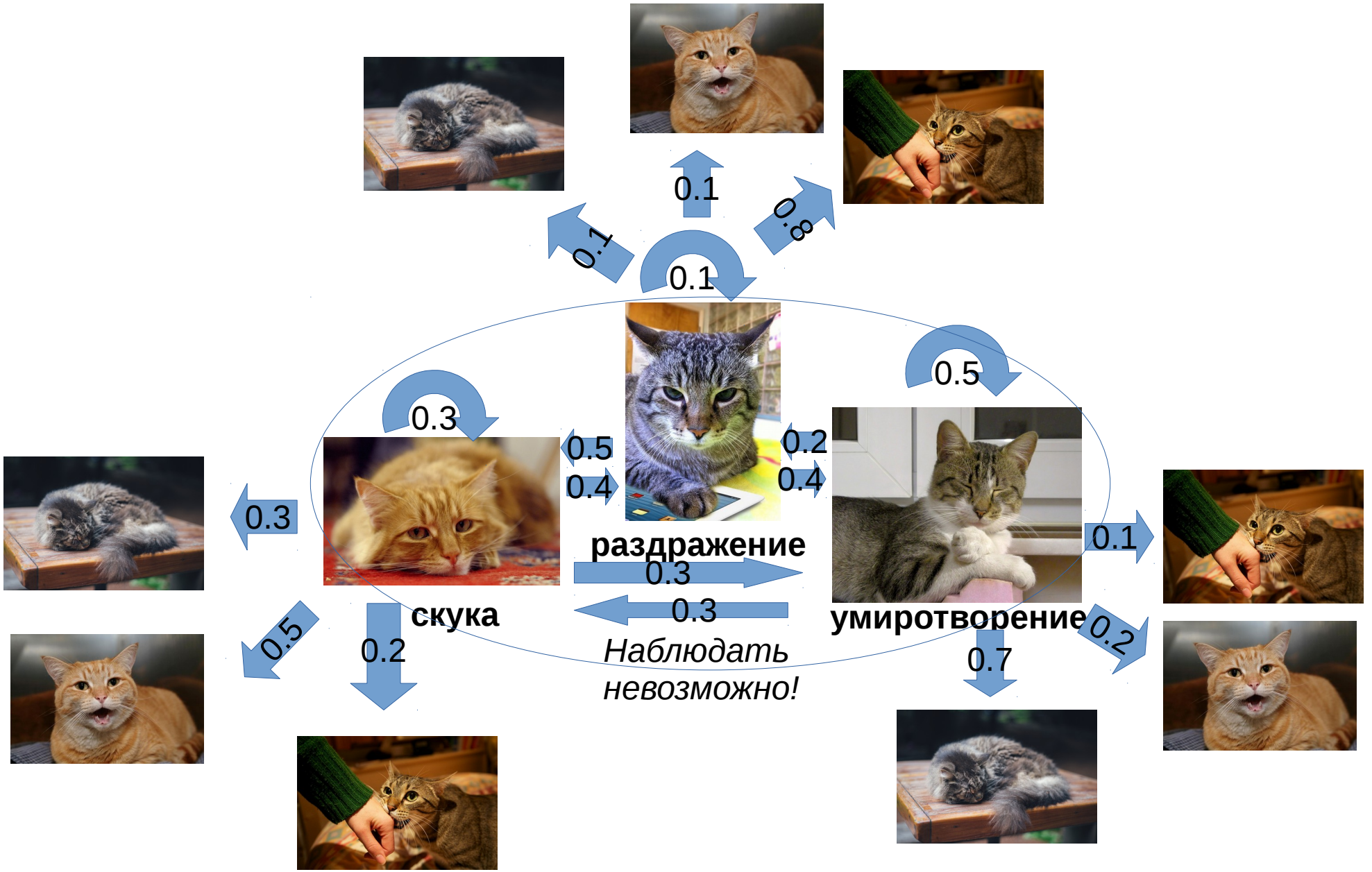
Вероятность, что  = 0.333*0.3*0.4 = <u>0.04</u>

Скрытые марковские модели (HMM)

Эти состояния скрыты!

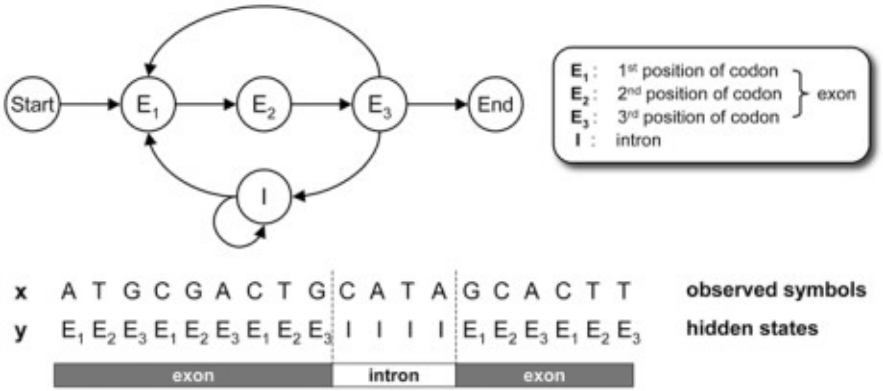
Наблюдаемые состояния:

Скрытые марковские модели (HMM)



Применение HMM в биоинформатике

- Предсказание кодирующих последовательностей



- Предсказании функции белка

- Предсказание вторичной структуры белка

и всякое другое...

Применение HMM в биоинформатике

- Соответствует ли новая последовательность имеющемуся профилю?

An aligned sequence family or “region of interest”

```
... ACA --- ATG ...
... TCA ACT ATC ...
... ACA C-- AGC ...
... AGA --- ATC ...
... ACC G-- ATC ...
```

A “classic” **profile** summarizing the sequence family

	1	2	3	4	5	6	7	8	9
A:	0.8	.	0.8	0.2	.	.	1.0	.	.
C:	.	0.8	0.2	0.2	0.2	.	.	.	0.8
G:	.	0.2	.	0.2	.	.	.	0.2	0.2
T:	0.2	.	.	.	.	0.2	.	0.8	.
-:	.	.	.	0.4	0.8	0.8	.	.	.
	A	C	A	-	-	-	A	T	C

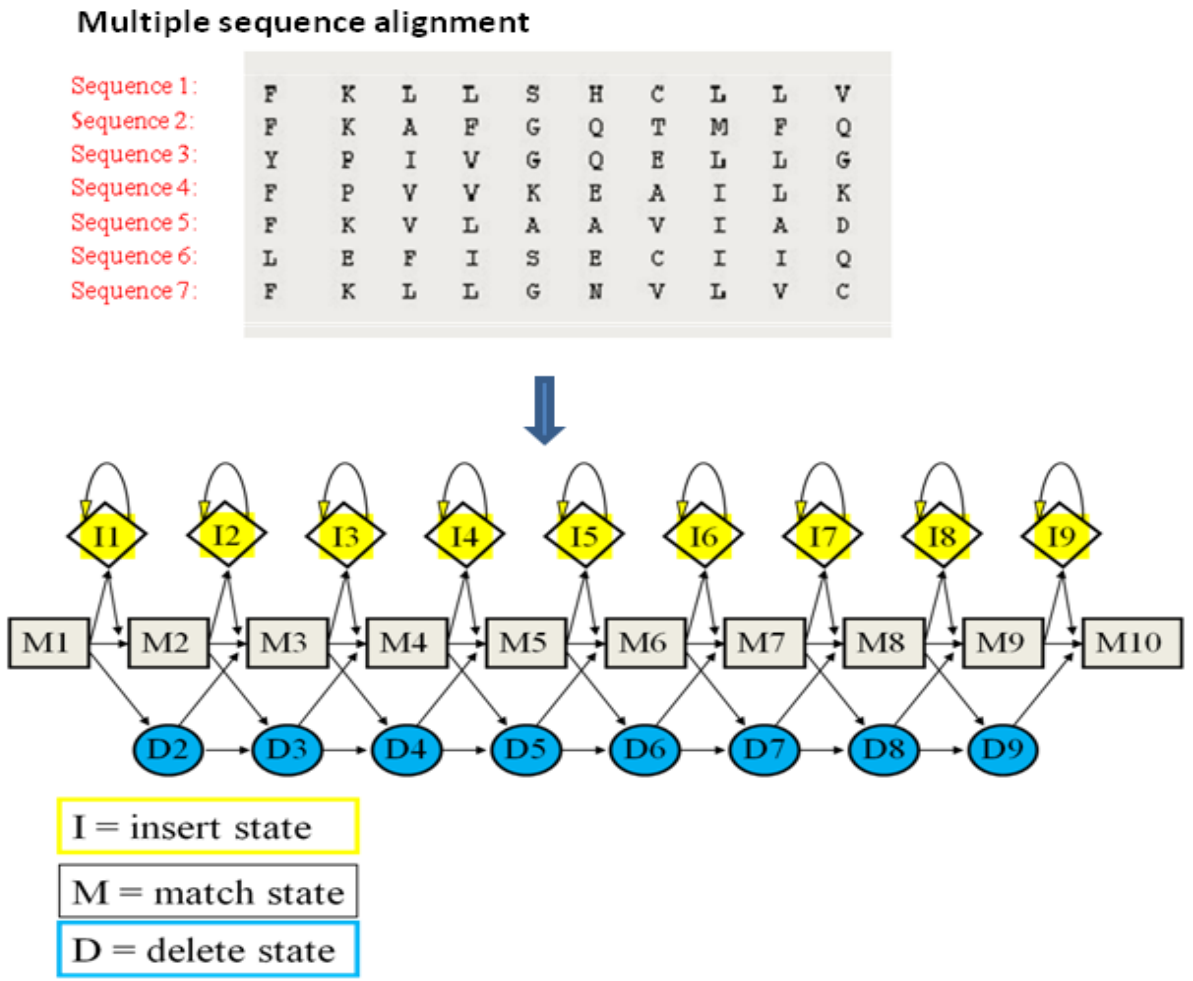
ACAATC is the consensus sequence

Новая последовательность:

ACAACTAGG

Можно вычислить по профилю, с какой вероятностью она относится к данной группе.

Профильные скрытые марковские модели (HMM) для предсказания функции белков



Пример пептида:

FKVVSACILV

Решаем, принадлежит ли он к данному семейству на основе вероятностей последовательности аминокислот, прописанных в модели

Prokaryotic Virus Orthologous Groups (pVOG)

NEW **New features added to the pVOGs Database:** **Viral Quotient measurements** JS **Searchable javascript tables** (it may take a moment to load when a large content is displayed)

Welcome to the pVOGs Database

"*Viruses are the most abundant biological entities on earth and encompass a vast amount of genetic diversity. The recent rapid increase in the number of sequenced viral genomes has created unprecedent structure and evolution of the virosphere.*" Kristensen DM et al., 2013.

This website provides access to the most recent database **9,518** orthologous groups shared among nearly **3,000** thousand complete genomes of viruses that infect bacteria and archaea (**P**rokaryotic **V**irus **O**r keeps growing and may have a wide range of applications in different areas of life sciences including taxonomy, evolutionary genetics, genomic epidemiology, metagenomics, systems biology, ecology, among othe

This webpage is a resource portal which provides access to the updated set of pVOGs[1], previously known as POGs (Phage Orthologous Groups):

You will find information regarding all prokaryotic virus genomes used for this update in the following **Genome Table**. General description of all pVOGs obtained for this update will be find at the **VOG Table**. Access identified pVOG may be found following the links present in both tables. The Viral Quotient (VQ) measurements have been added to the individual VOG tables, such as **VOG0008** showing a VQ of 1 (a gene only **Downloads** page provides VOG alignments, HMM profiles and files of all VOG tables in tabular delimited format for downloading. The history of updates and new releases can be followed at **News & Updates** pa in this website is also provided.

[1] Would you like to know more about how pVOGs were obtained? Read about them here Grazziotin et al., 2017.

[2] Read more about Viral Quotient measurements: methodology was previously described by Kristensen et al., 2013.

Importance and Applications of pVOGs

✔ To understand the history of viral protein families

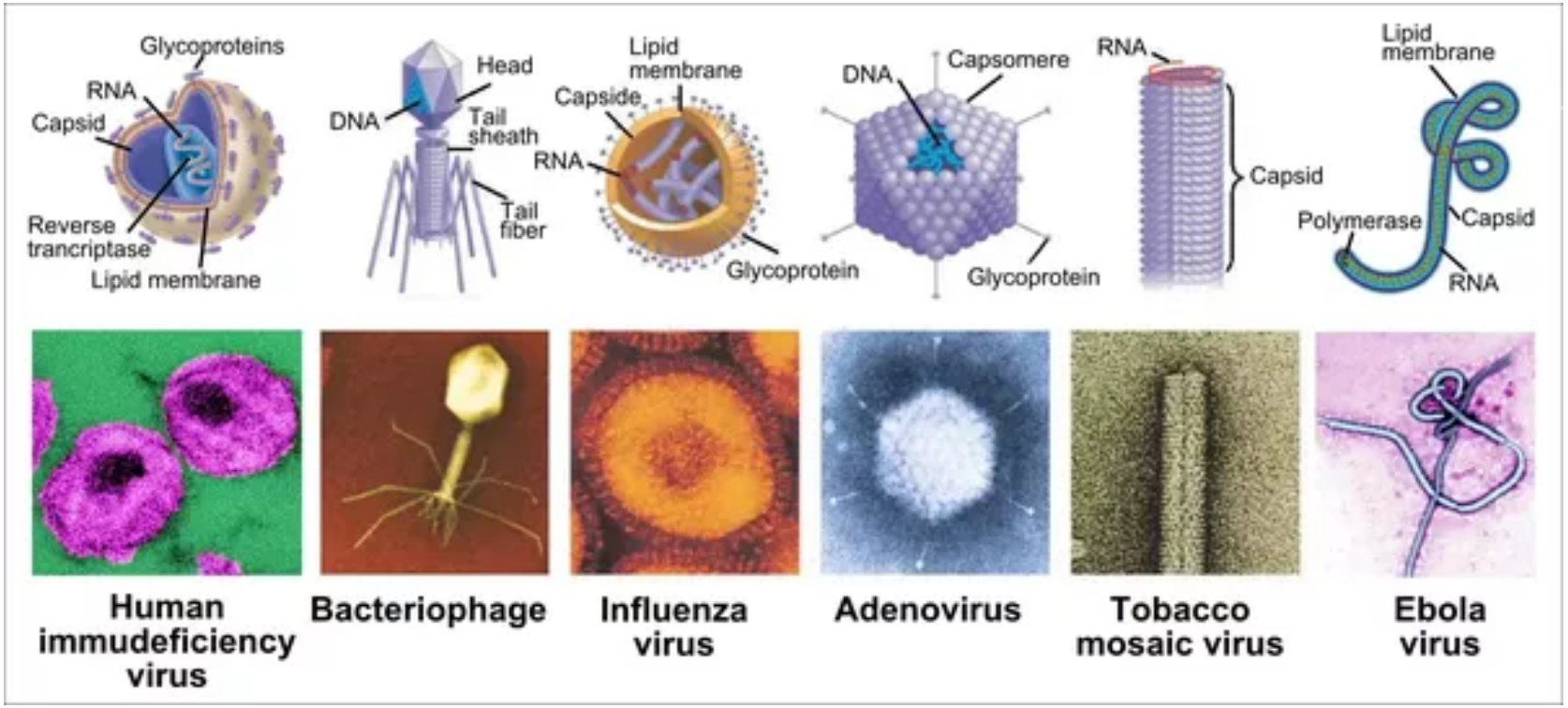
✔ To aid in evolutionary classification of known phages

✔ To be applied in the reconstruction of ancestral phage genomes

✔ To help annotate orthologs from poorly characterized genomes

✔ To identify virus-specific genes that could be used as diagnostic markers of prokaryotic viruses presence in a given environment

Вирус не может обойтись без капсида



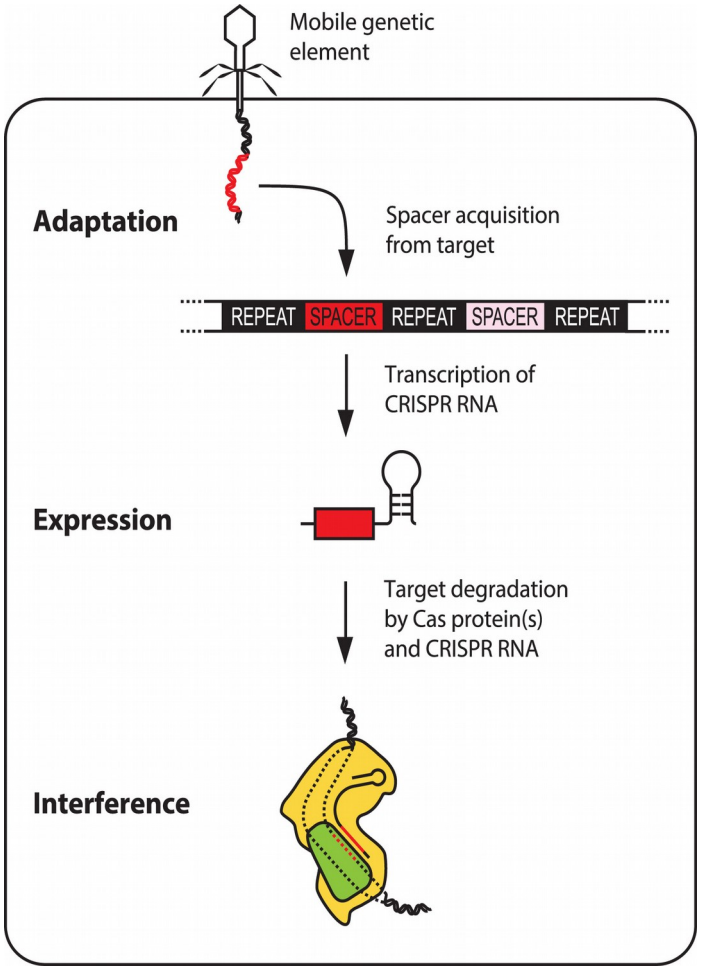
На практике будем использовать HMM-профили капсидных белков

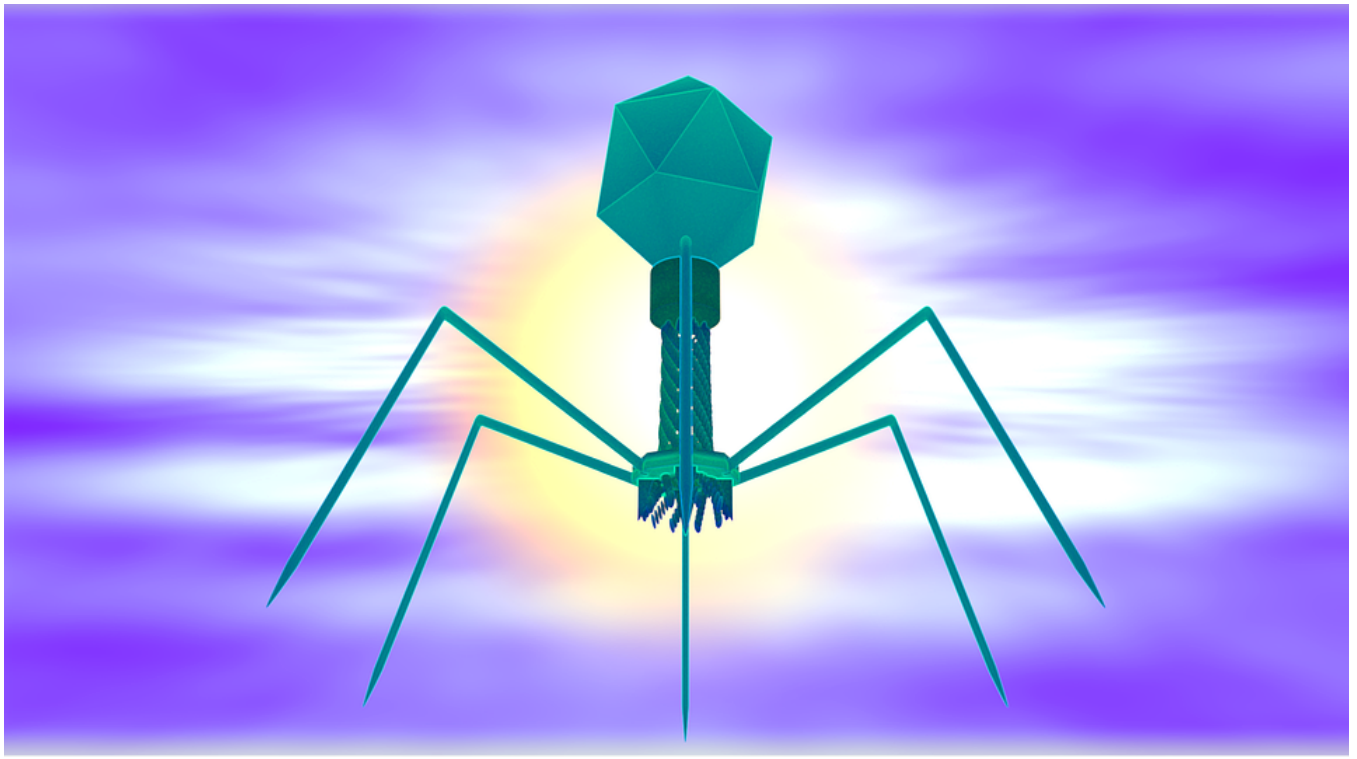
Как предсказать хозяина?

- BLAST (если повезет)
- Корреляции представленности (если есть много метагеномов)
- Если есть сходство с известным вирусом, можно спроецировать информацию о его хозяине
- Соответствие последовательности вируса CRISPR-спейсерам бактерии-хозяина

Как предсказать хозяина? CRISPR

- Спейсеры длиной 24-50 п.н. в геноме хозяина – фрагменты фага
- Есть базы спейсеров известных бактерий
- Сравнение неизвестного фага с базой таких спейсеров дает вероятность найти хозяина





Спасибо за внимание!

Обсудить метагеномику, бактериофагов и вирусные белки можно по почте:

estarikova@rcpcm.org