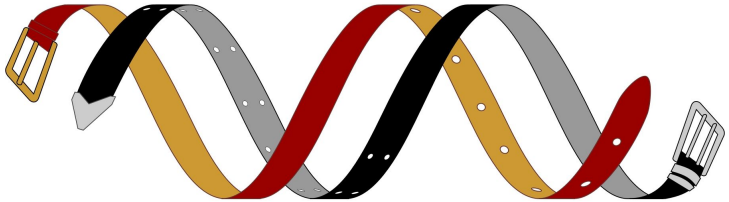
Учебно-Научный Центр



Биоинформатика

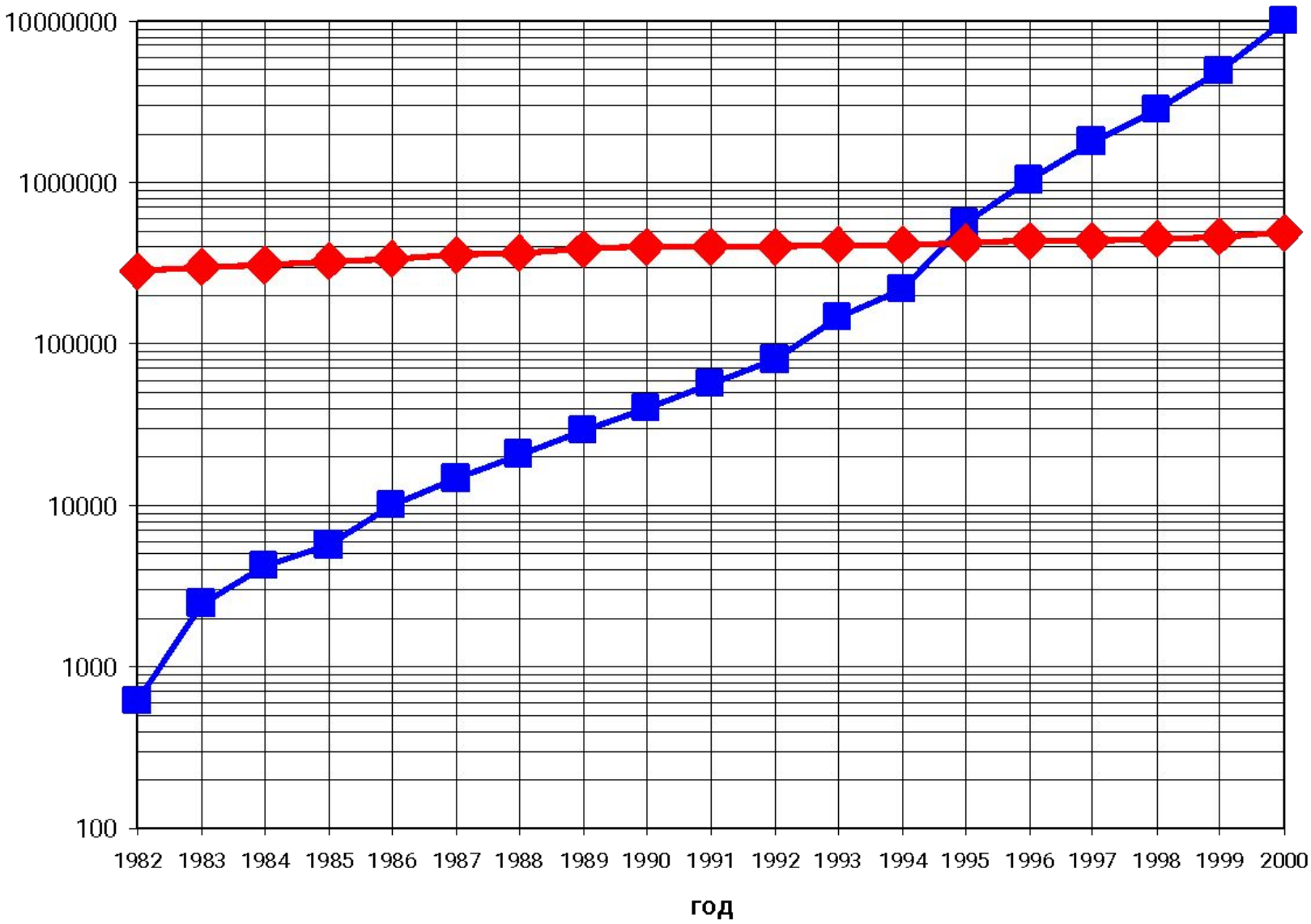
Биоинформатика, или молекулярная биология in silico

М.Гельфанд

Семинар в ИППИ 7 апреля 2006

Пропаганда 1

красный: статьи
синий: последовательности



Анализ индивидуальных генов

- Поиск родственных белков в банках последовательностей – перенос функции от гомологов
- Функциональные сайты (каталитические центры)
- Функциональные участки (трансмембранные сегменты, сигнальные пептиды и т.п.)

- Анализ на уровне индивидуальных генов даёт возможность охарактеризовать 50-75% генов в новом геноме

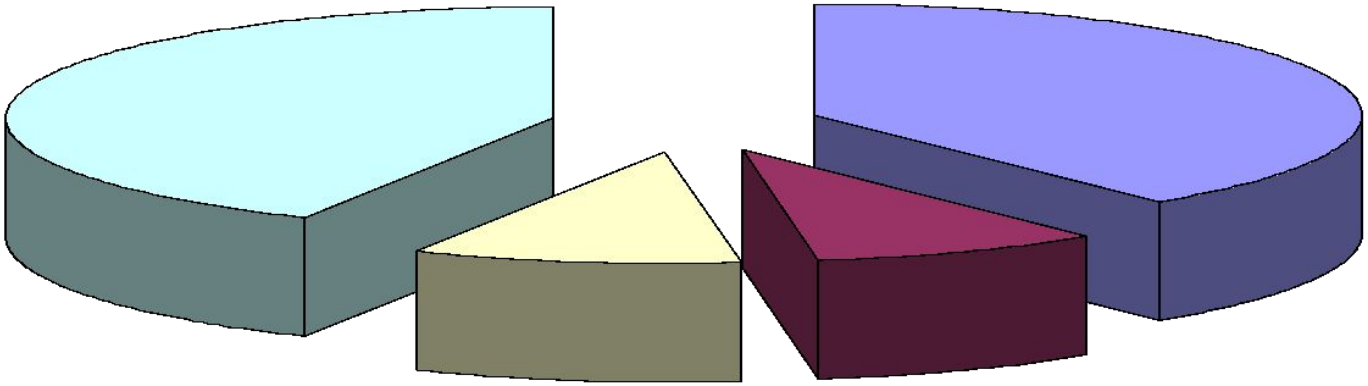
Но:

- ~100 универсально отсутствующих генов (нет ни одного известного гена для известной функции)
- множество функций, для которых неизвестны представители в больших таксонах
- в каждом геноме ~5-10% консервативных генов с неизвестной функцией
- трудно предсказывать специфичность в мультигенных семействах (транспортёры, факторы транскрипции)
- нельзя найти что-то принципиально новое

How much do we know about the *Escherichia coli* proteome?

Characterized experimentally

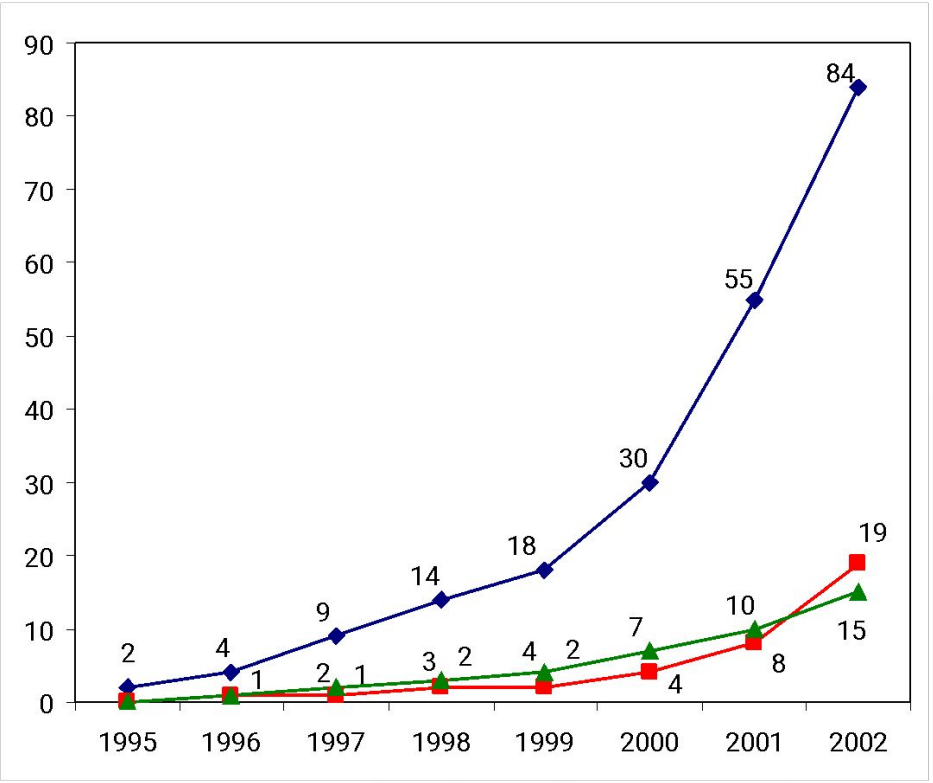
Function inferred by similarity only

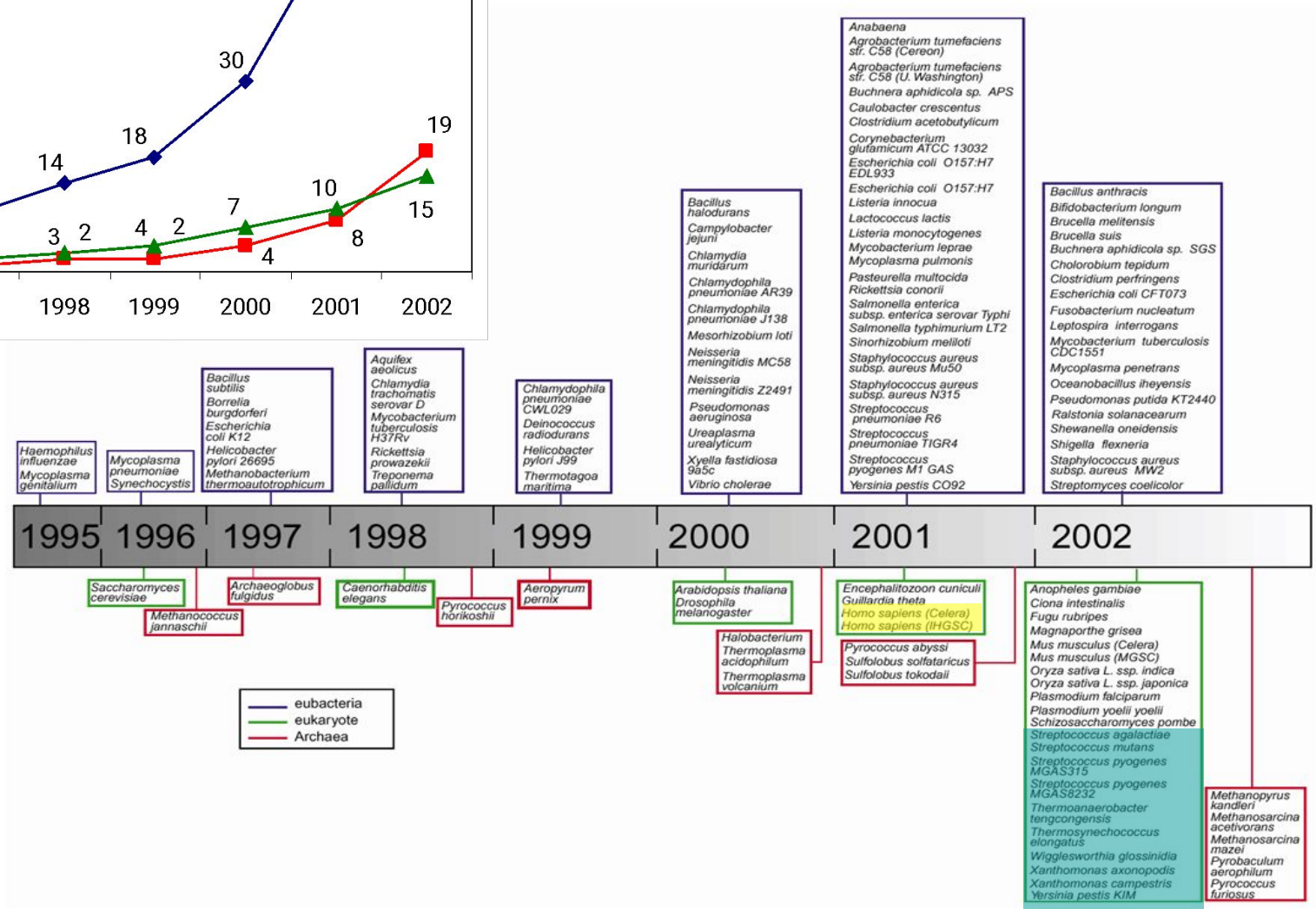


"Hypothetical"

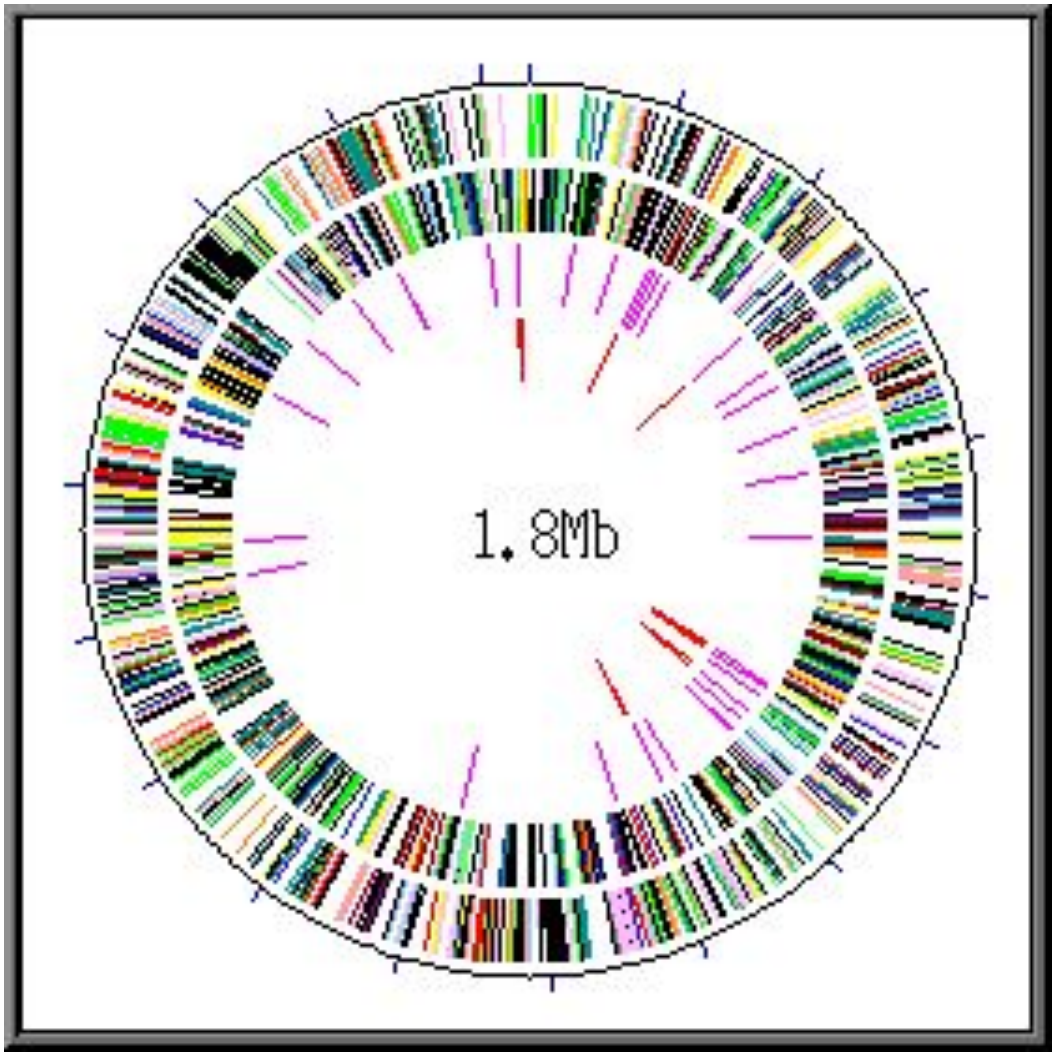
"Conserved hypothetical"

Пропаганда – 2
Полные геномы

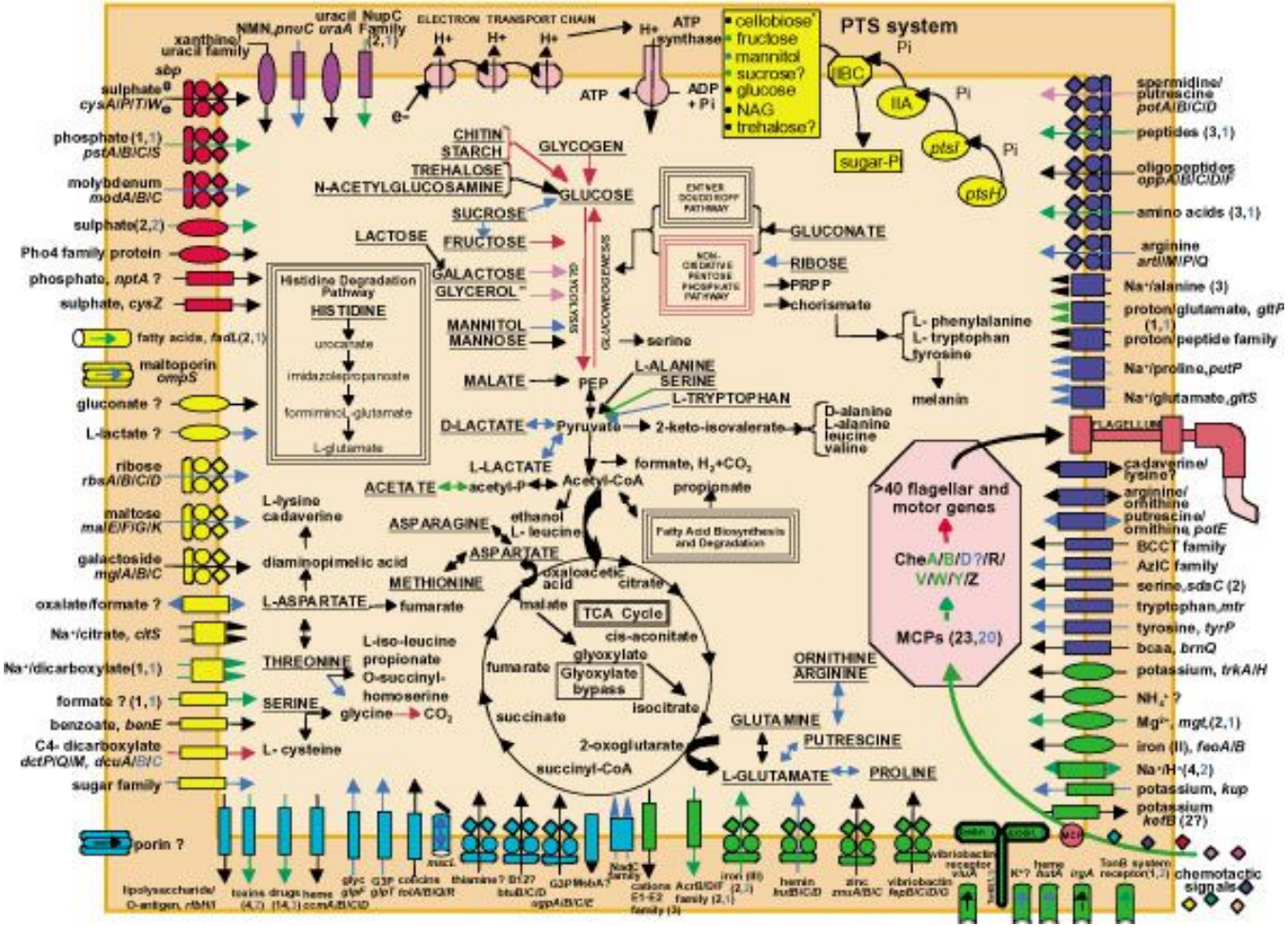




Haemophilus influenzae, 1995



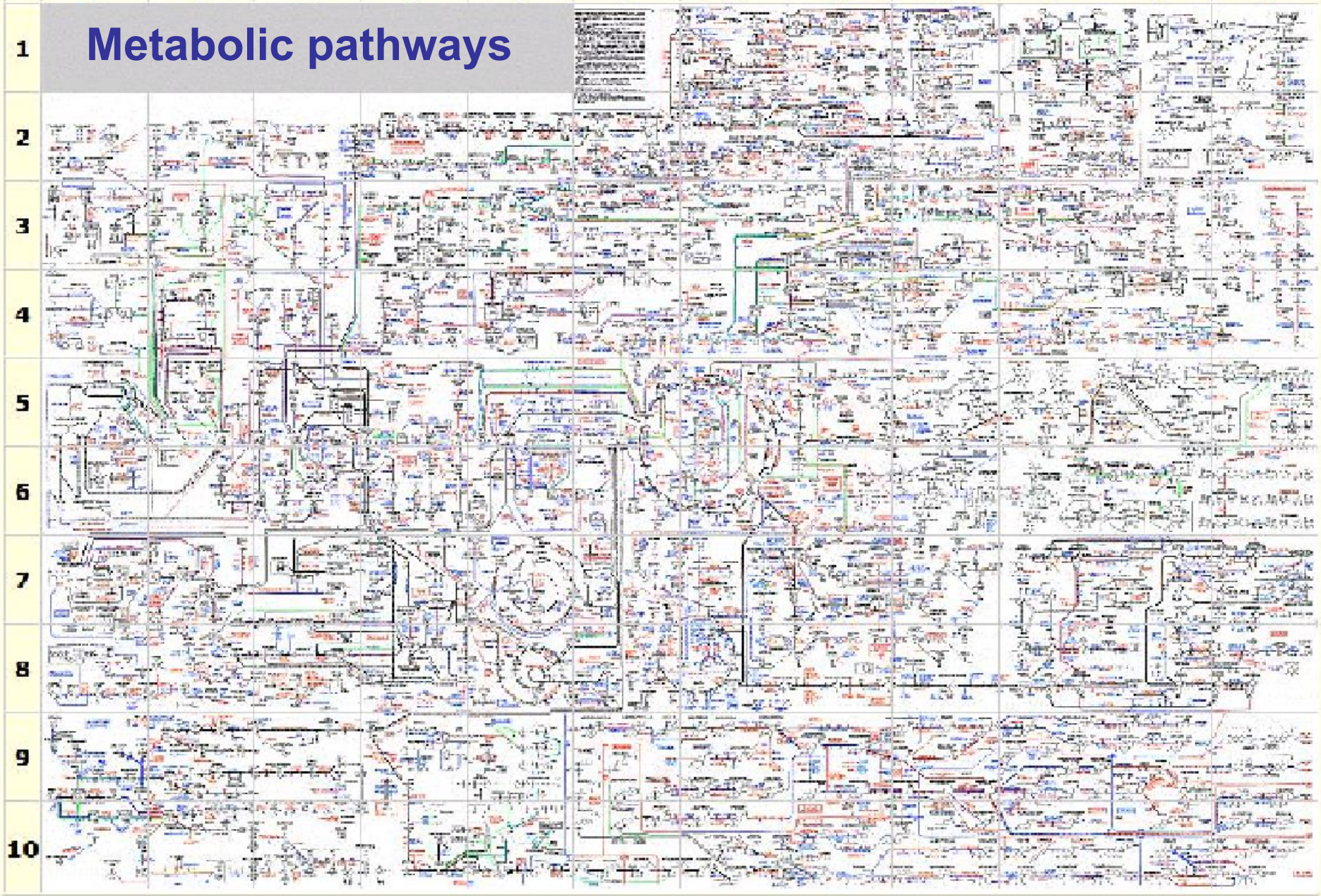
Vibrio cholerae, 2000



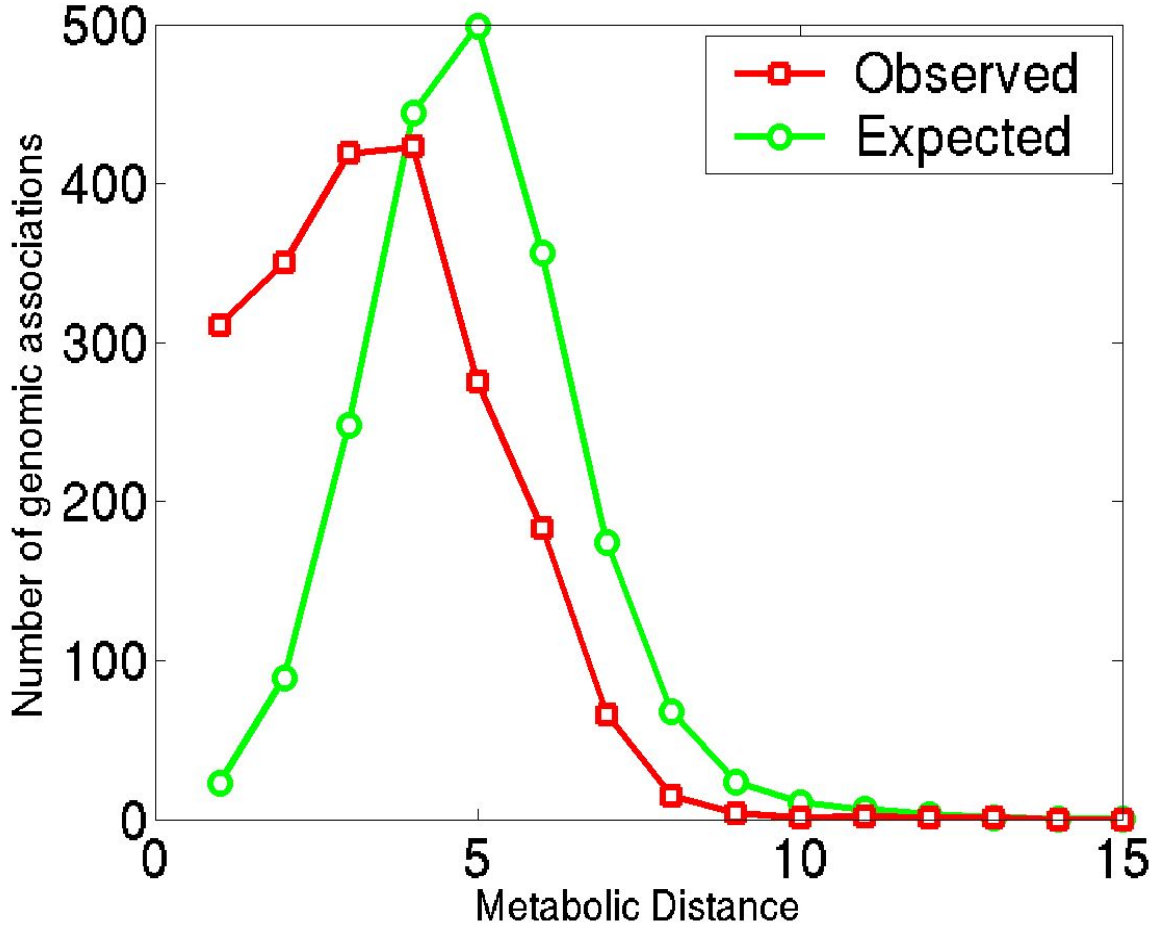
Сравнительно-геномные подходы

- Positional clustering
- Phylogenetic profiling
- Gene fusions

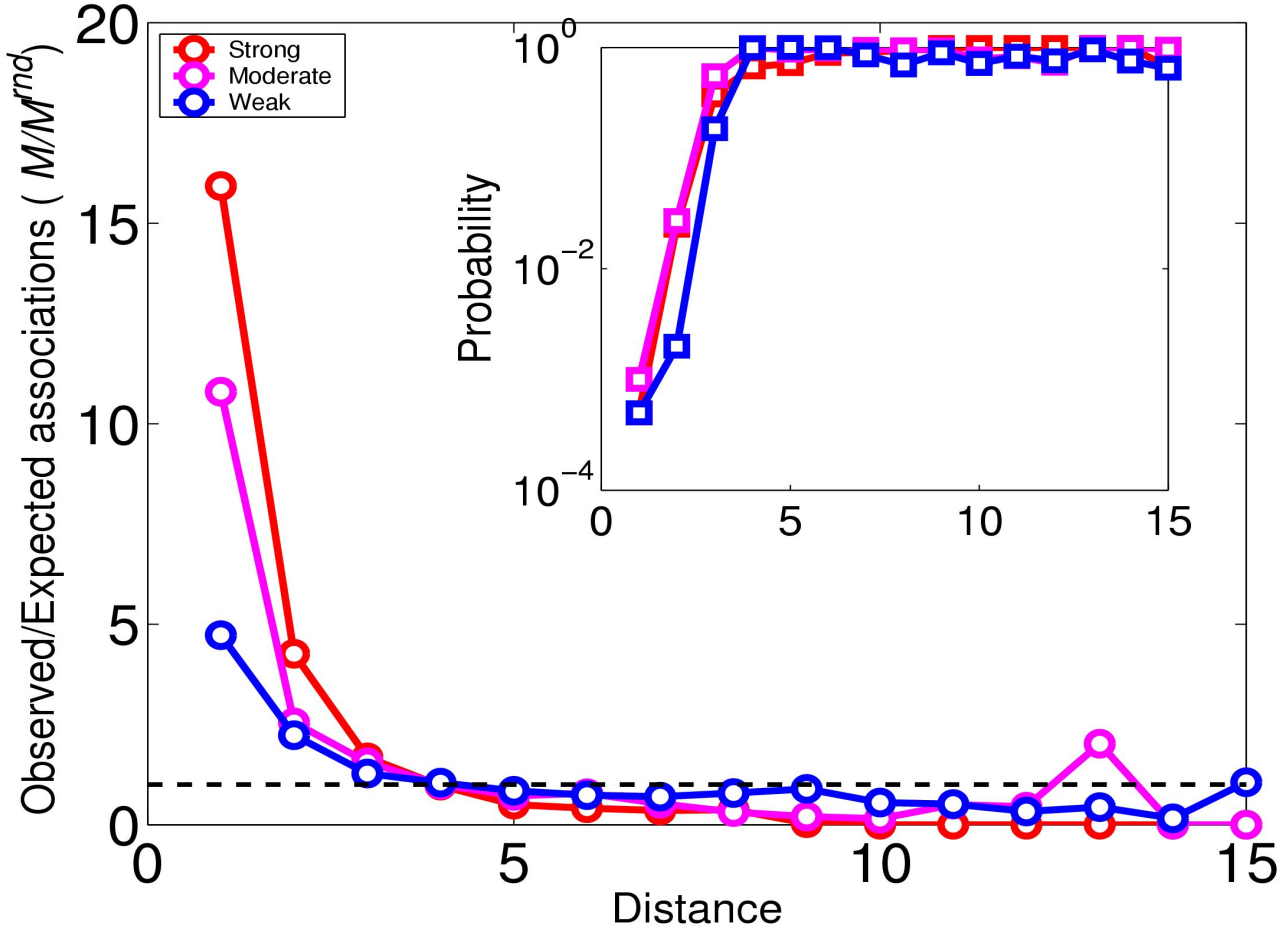
Metabolic pathways



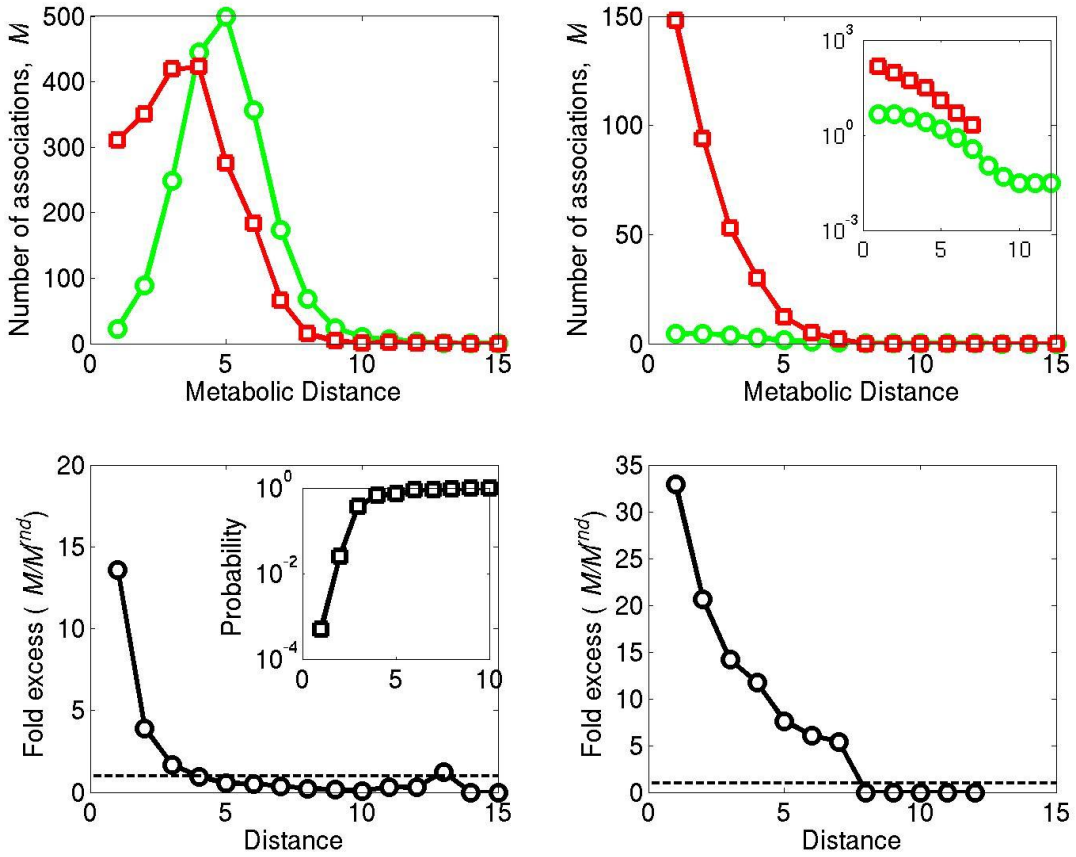
Functionally dependent genes tend to cluster on chromosomes in many different organisms



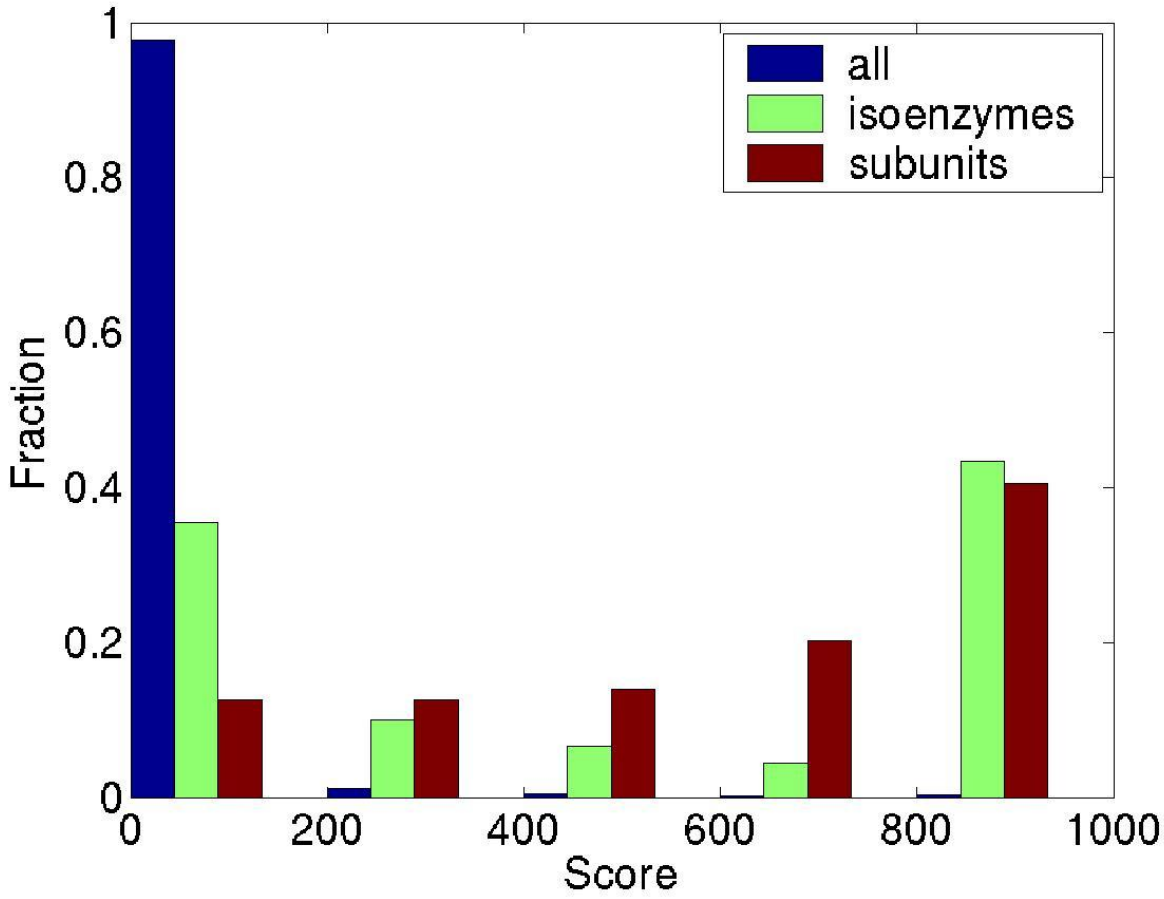
More genomes (stronger links) => highly significant clustering



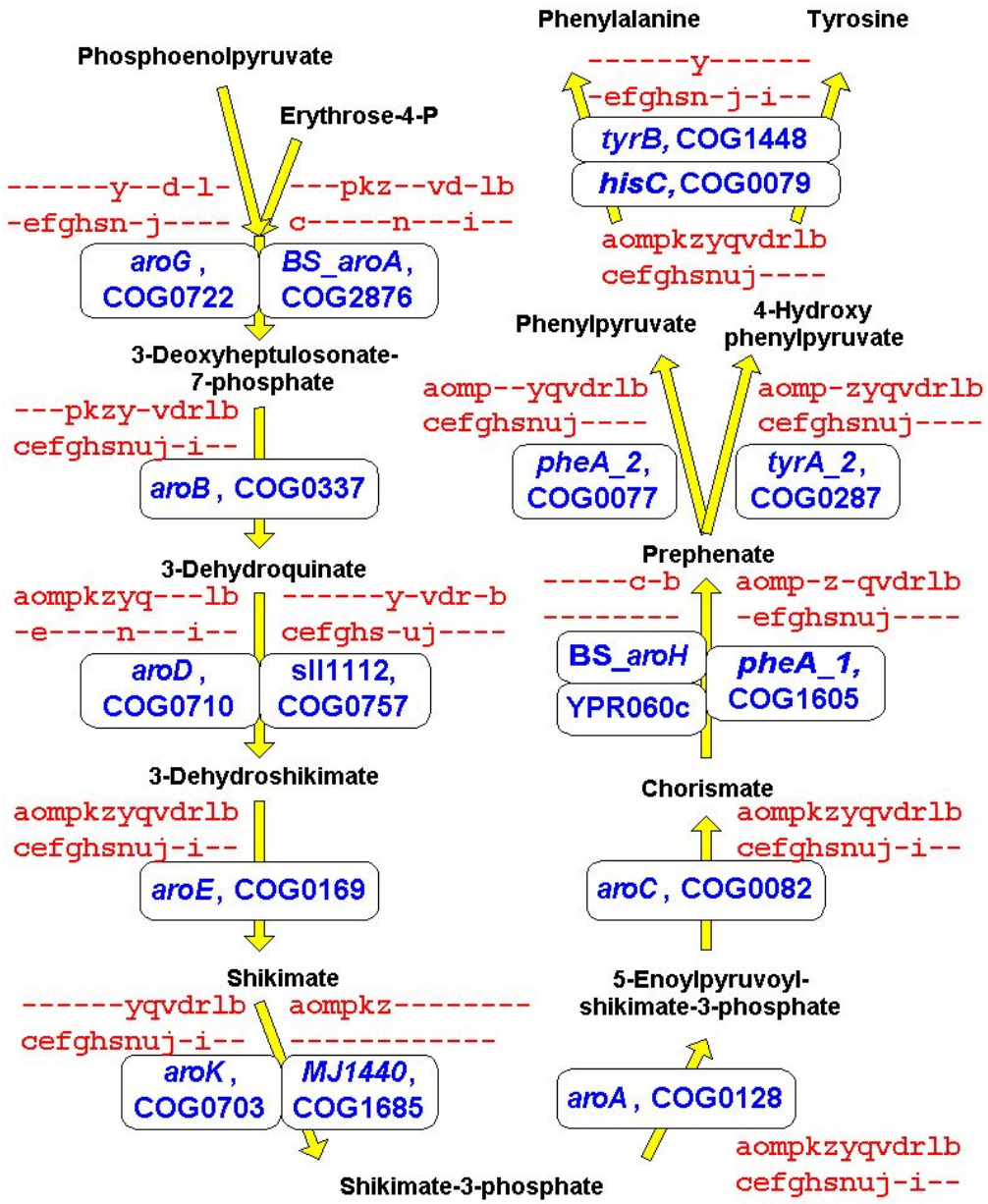
… особенно в линейных путях (справа)



Распределение уровней связи (бимодальное для изоферментов, монотонное для субъединиц)



Phyletic profiles in the Phe/Tyr pathway



Arithmetics of phyletic patterns

3-dehydroquinate dehydratase (EC 4.2.1.10):

Class I (AroD) COG0710 `aompkzyq---lb-e----n---i--`

Class II (AroQ) COG0757 `------y-vdr-bcefghs-uj----`

+

Two forms combined `aompkzyqvdrlbcefghsnuj-i--`

Shikimate dehydrogenase (EC 1.1.1.25):

AroE COG0169 `aompkzyqvdrlbcefghsnuj-i--`

Shikimate kinase (EC 2.7.1.71):

Typical (AroK) COG0703 `------yqvdrlbcefghsnuj-i--`

Archaeal-type COG1685 `aompkz-------------------`

+

Two forms combined `aompkzyqvdrlbcefghsnuj-i--`

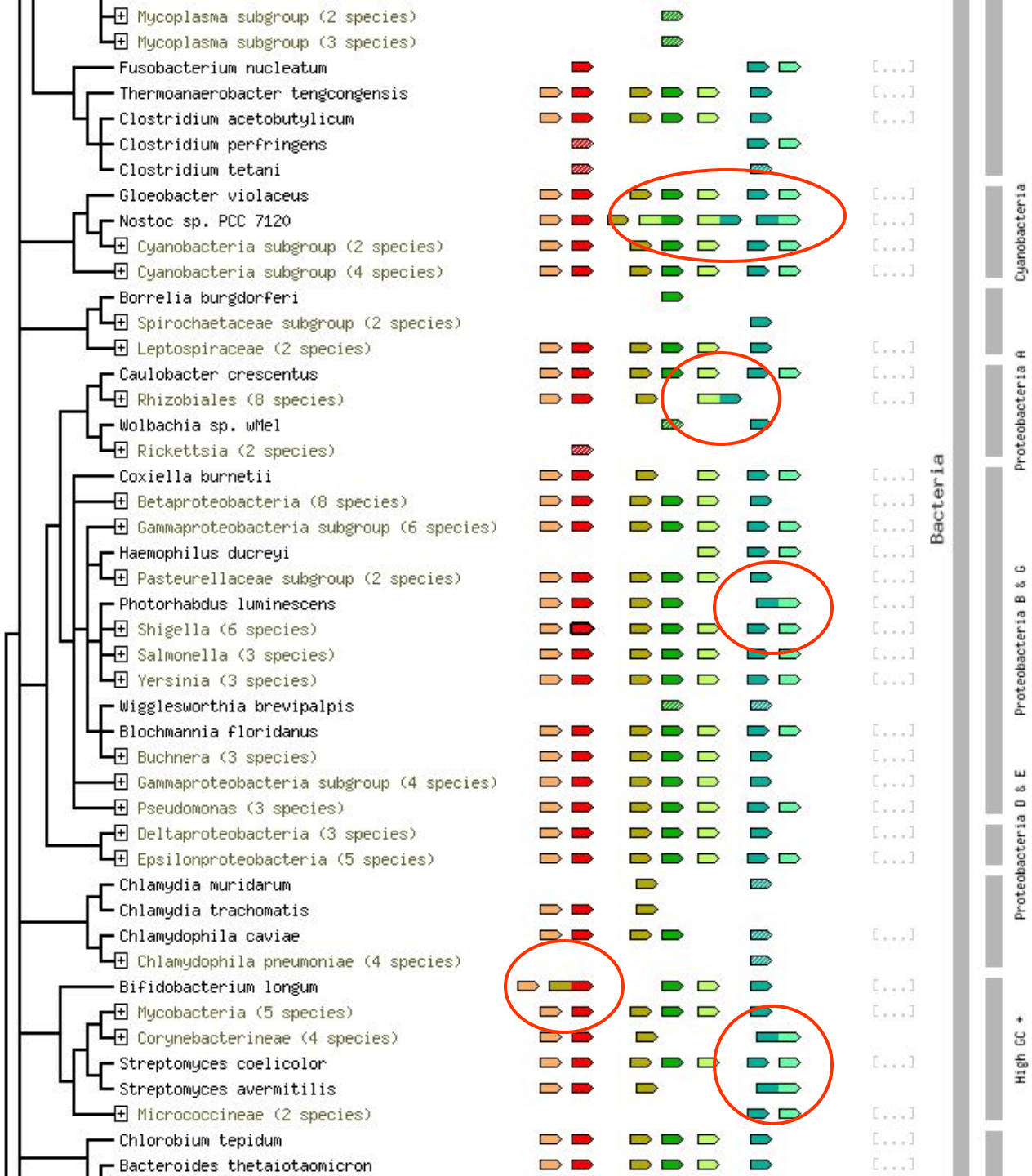
5-enolpyruvylshikimate 3-phosphate synthase (EC 2.5.1.19)

AroA COG0128 `aompkzyqvdrlbcefghsnuj-i--`

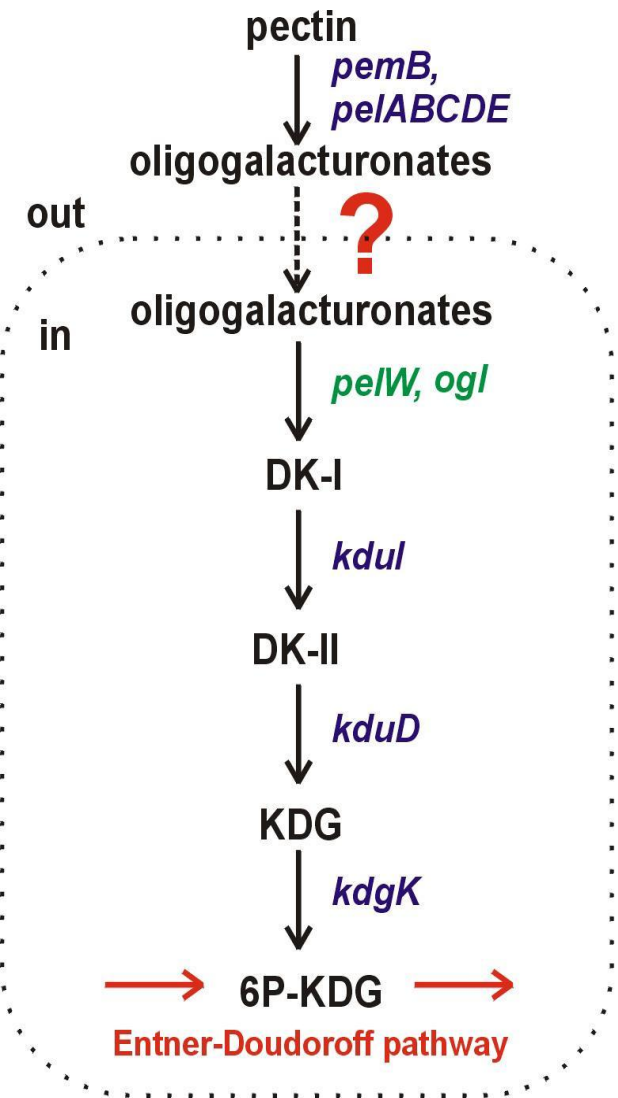
Chorismate synthase (EC 2.5.1.19) AroC

COG0082 `aompkzyqvdrlbcefghsnuj-i--`

STRING: *trpB* – fusions

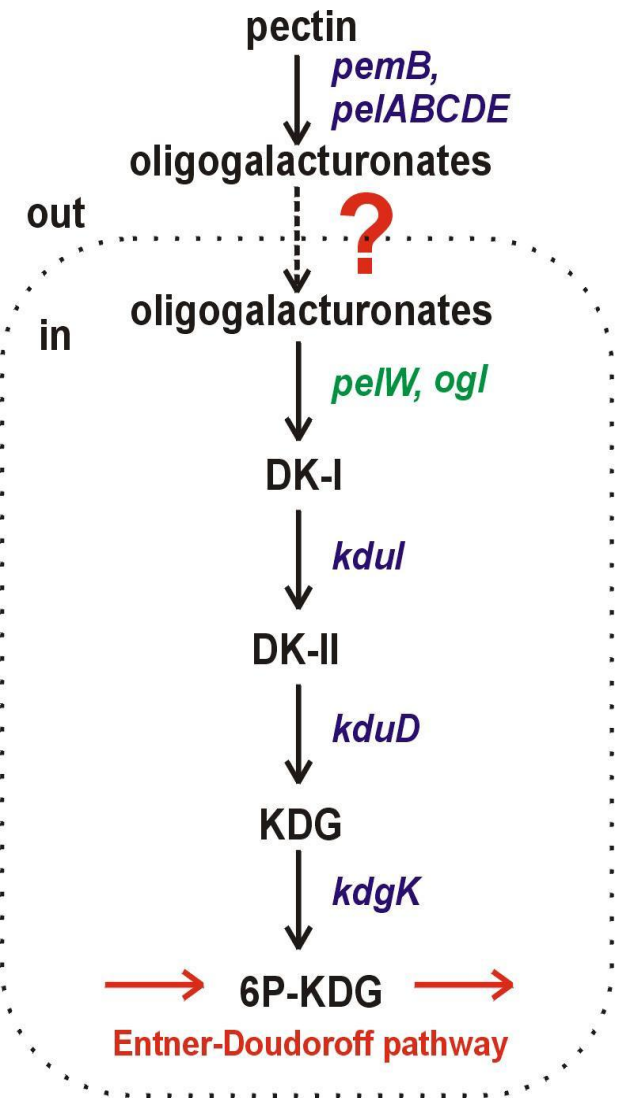


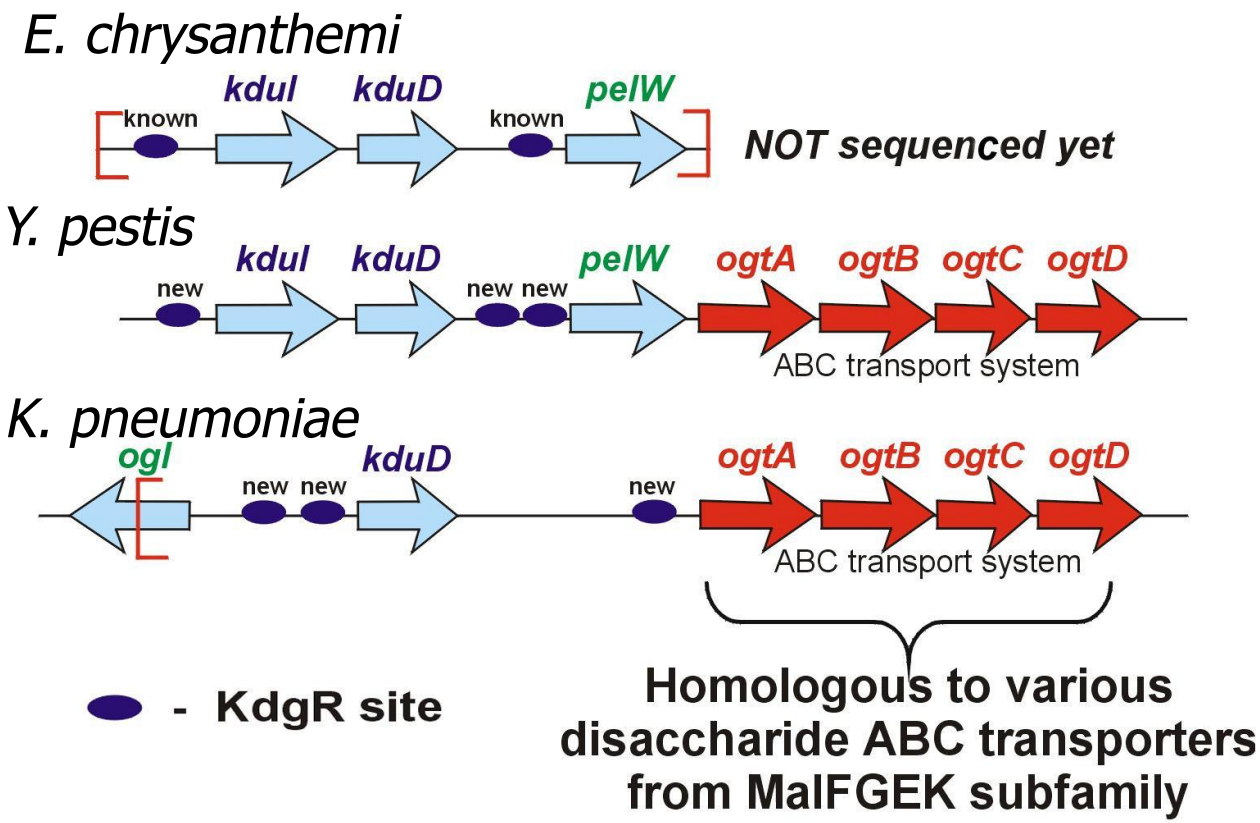
Утилизация пектина



E. chrysanthemi

… и транспорт олигогалактуронатов





YpaA: транспортёр рибофлавина

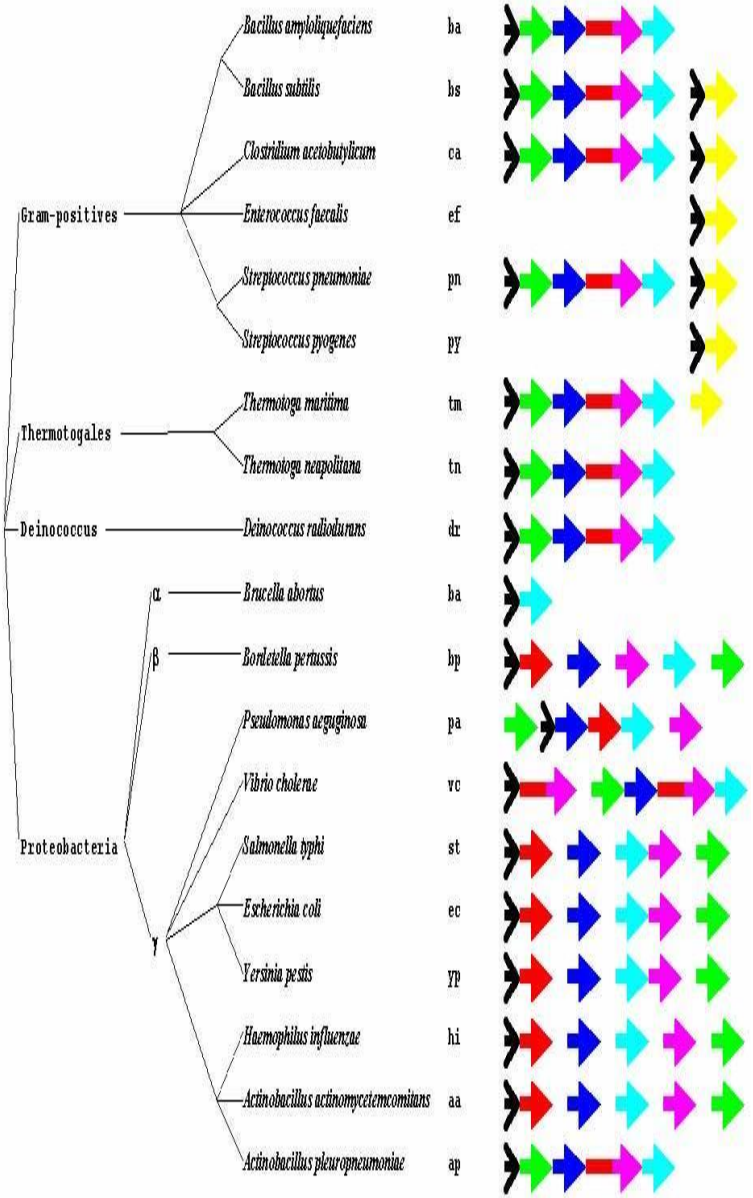
- 5 предсказанных TM-сегментов => потенциальный транспортёр
- регуляторный RFN-элемент => ко-регуляция с генами метаболизма рибофлавина => транспорт рибофлавина или предшественника
- *S. pyogenes*, *E. faecalis*, *Listeria*: есть *ypaA*, нет генов биосинтеза рибофлавина => транспорт рибофлавина

Предсказание:

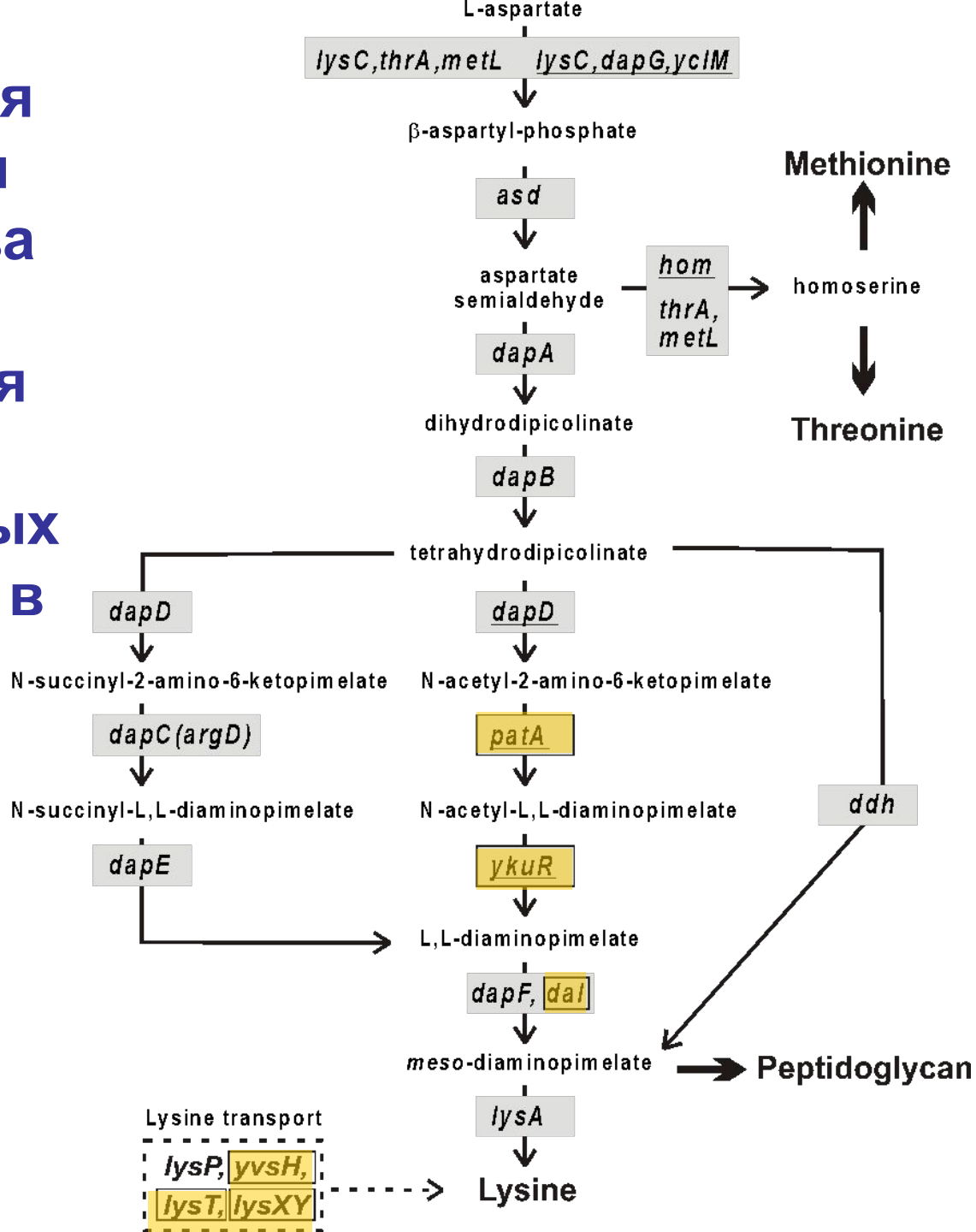
YpaA – рибофлавиновый транспортёр (Gelfand et al., 1999)

Проверка:

- YpaA переносит рибофлавин (генетический анализ, Кренева и др., 2000)
- *ypaA* регулируется рибофлавином (анализ экспрессии на микрочипах



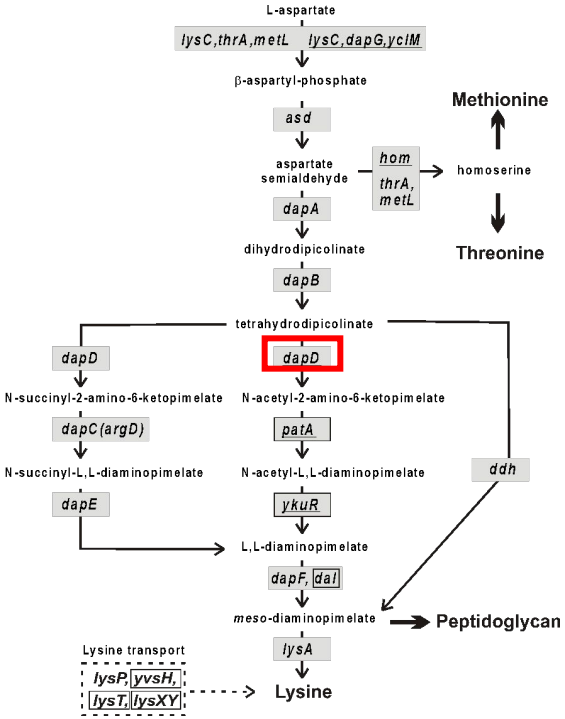
Метаболическая реконструкция пути биосинтеза лизина: Идентификация пути ацетилированных интермедиатов в *B. subtilis* и родственных бактериях

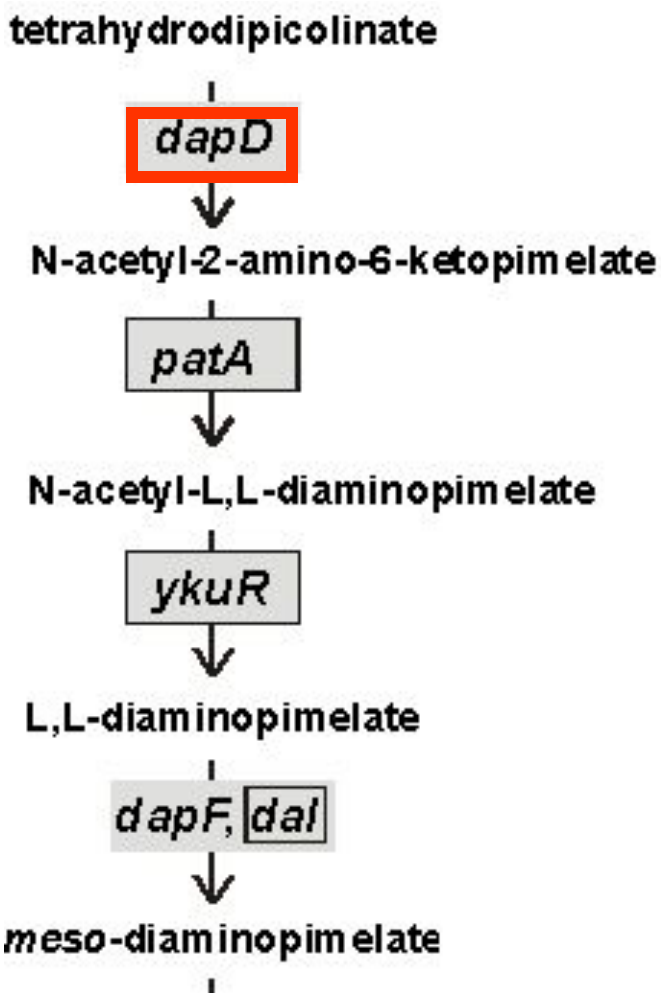


Идентификация пути ацетилированных интермедиатов - 0

dapD (*yquQ*):

- ортолог известного гена *E. coli*

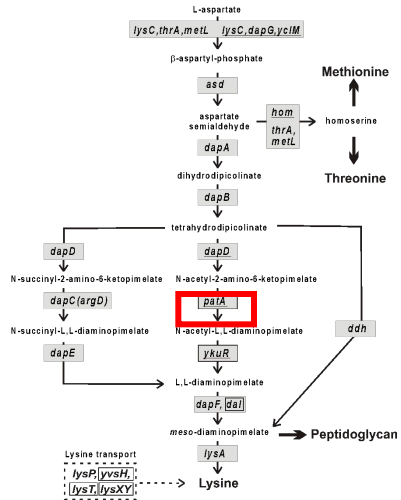


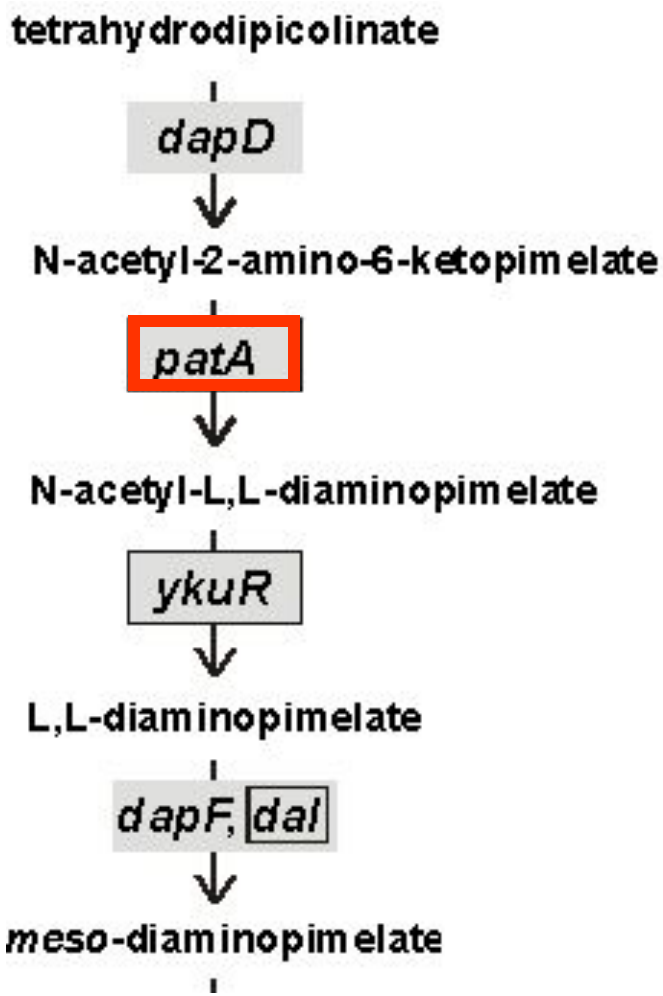


Идентификация пути ацетилированных интермедиатов - 1

patA:

- пиридоксаль-фосфат-зависимая аминотрансфераза (по гомологии)
- ко-локализуется и ко-регулируется с генами биосинтеза лизина во многих грам-положительных бактериях

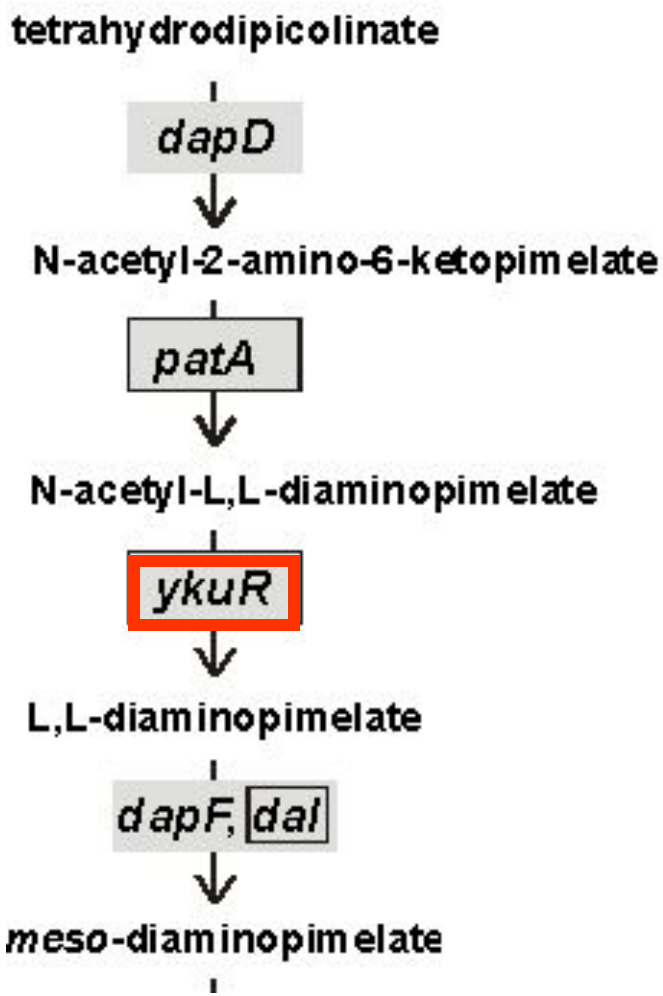




Идентификация пути ацетилированных интермедиатов - 2

ykuR:

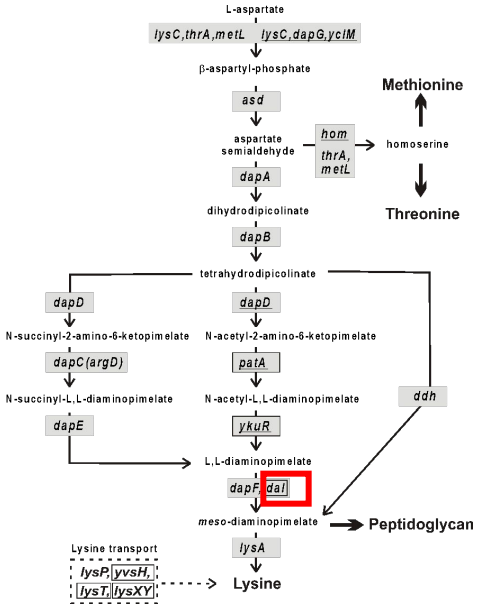
- N-ацил-L-аминокислота амидогидролаза (по гомологии)
- ко-локализуется и ко-регулируется с геном биосинтеза лизина *dapD* во многих грам-положительных бактериях
- в некоторых случаях принадлежит к большому лизиновому оперону, регулируемому LYS-элементом

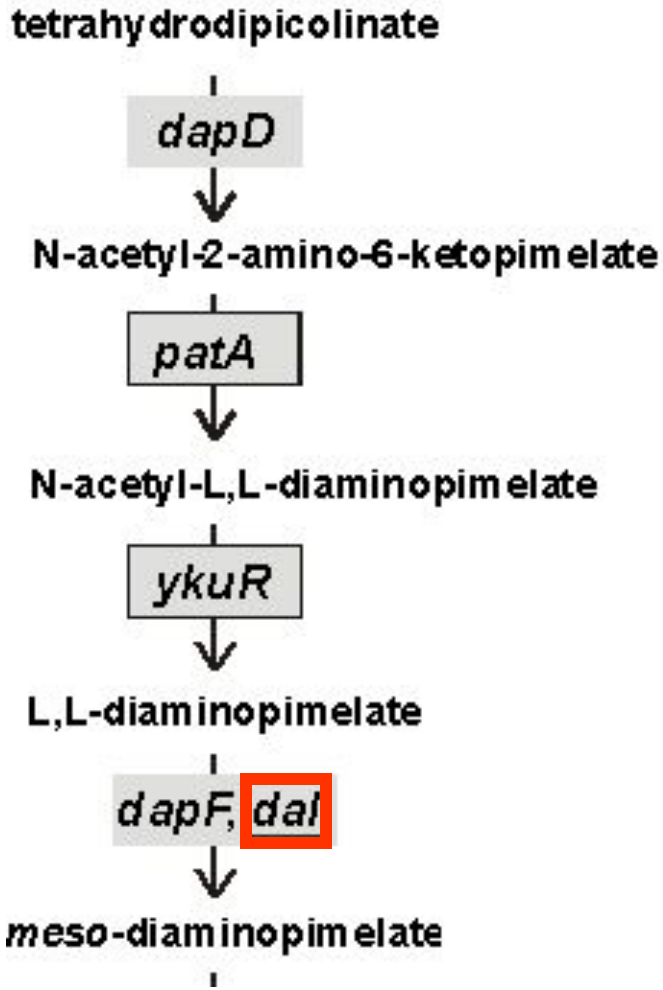


Идентификация пути ацетилированных интермедиатов - 3

dapX:

- *dapF* отсутствует у некоторых бактерий (*Staphylococcus aureus, Oenococcus oeni, Leuconostoc mesenteroides*)
- во **всех** этих геномах есть *dapX*, гомологичный аланиновой рацемазе и другим эпимеразам
- в *S. aureus dapX* принадлежит к большому лизиновому оперону
- в *O. oeni* оперон *dapX-asd* регулируется LYS-элементом





Сравнительная геномика систем утилизации цинка

Две роли цинка в бактериях:

- Структурная в ДНК-полимеразах, праймазах, рибосомных белках
- Каталитическая в протеазах и других белках

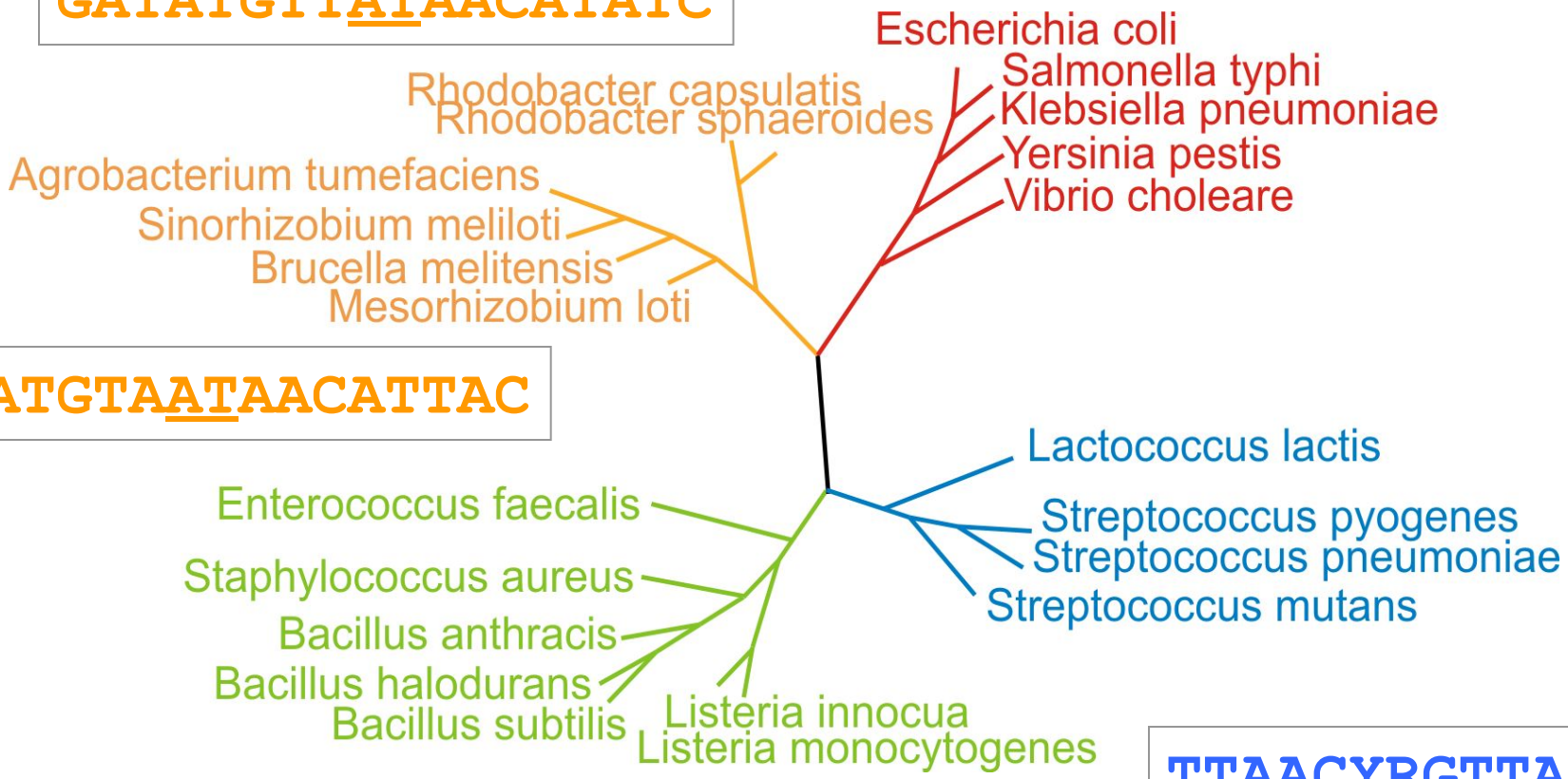
Регуляторы и сигналы

nZUR-α

nZUR-γ

GAAATGTTATANTATAACATTTC

GATATGTTATAACATATC



GTAATGTAATAACATTAC

TTAACYRGTTAA

pZUR

TAAATCGTAATNATTACGATTTA

AdcR

Цинк и паралоги белков рибосом

	L36	L33	L31	S14
nZUR				
E. coli, S.typhi	–	–	– +	–
K. pneumoniae	–	–	– –	–
Y. pestis,V. cholerae	– ×	–	– +	–
pZUR				
B subtilis	–	– + –	– +	– +
S. aureus	–	– – –	–	– +
Listeria spp.	–	– –	–	– +
AdcR				
E. faecalis	–	– × – –	–	– + –
S. pne., S. mutans	–	– – –	–	–

(в скобках – мотив «цинковая лента»)

		L36	L33	L31	S14
nZUR	*E. coli, S.typhi*	(–)	–	(–) +	–
	K. pneumoniae	(–)	–	(–) –	–
	Y. pestis,V. cholerae	(–) ×	–	(–) +	–
pZUR	*B subtilis*	(–)	(–) + –	(–) +	(–) +
	S. aureus	(–)	(–) – –	–	(–) +
	Listeria spp.	(–)	(–) –	–	(–) +
AdcR	*E. faecalis*	(–)	(–) × – –	–	(–) + –
	S. pne., S. mutans	(–)	(–) – –	–	(–)

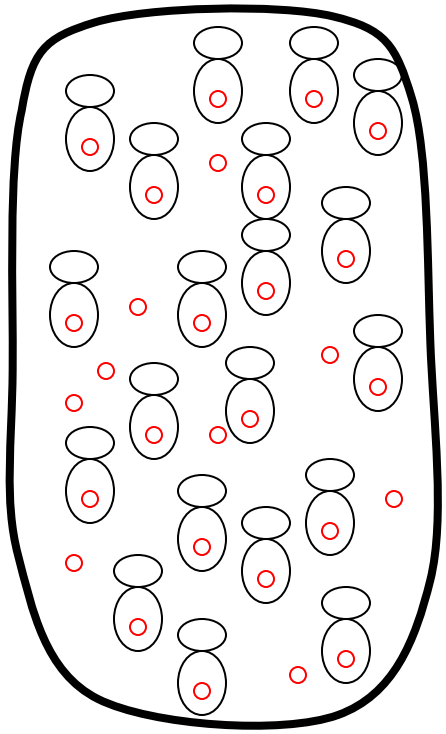
Сводка наблюдений:

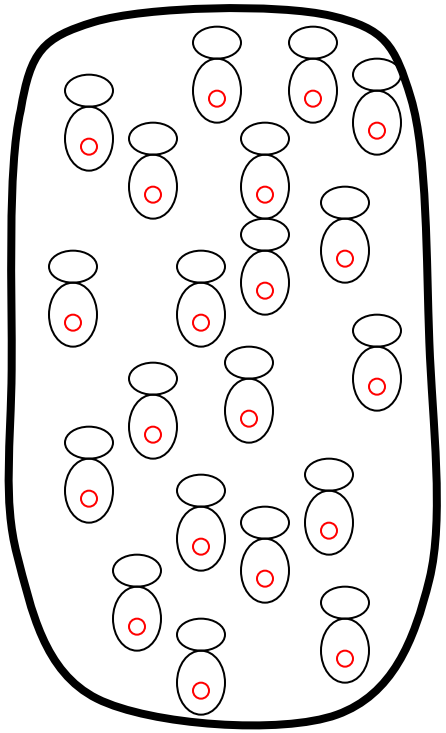
- Makarova-Ponomarev-Koonin, 2001:
  - L36, L33, L31, S14 – это единственные рибосомные белки, дуплицированные более, чем в одном геноме
  - L36, L33, L31, S14 – четыре из семи рибосомных белков, содержащих мотив цинковой ленты (четыре цистеина)
  - Из двух (или более) копий L36, L33, L31, S1, обычно одна содержит мотив цинковой ленты, а другая – нет
- **Среди генов, кодирующих паралоги рибосомных белков, как правило одни регулируется цинковым репрессором, а соответствующий белок никогда не имеет мотива цинковой ленты**

Плохой сценарий

недостаточно цинка: весь цинк потреблен рибосомами, ферменты голодают

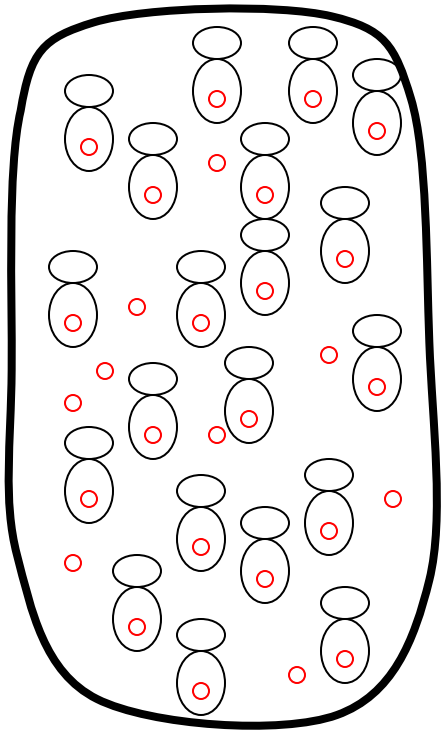
достаточно цинка



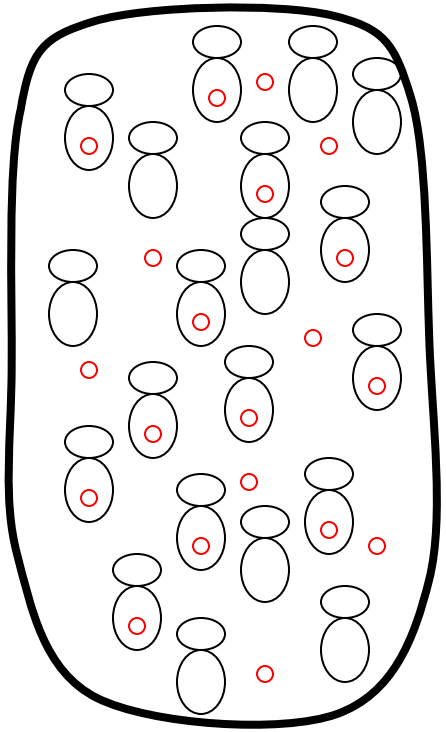


Хороший сценарий

достаточно цинка

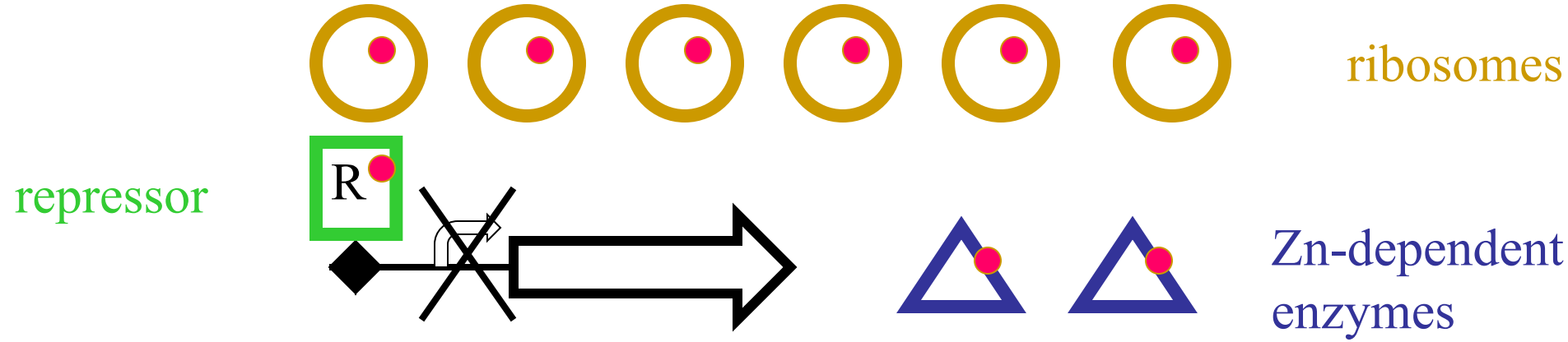


недостаточно цинка: часть рибосом включает белки, не содержащие цинка – остается для ферментов

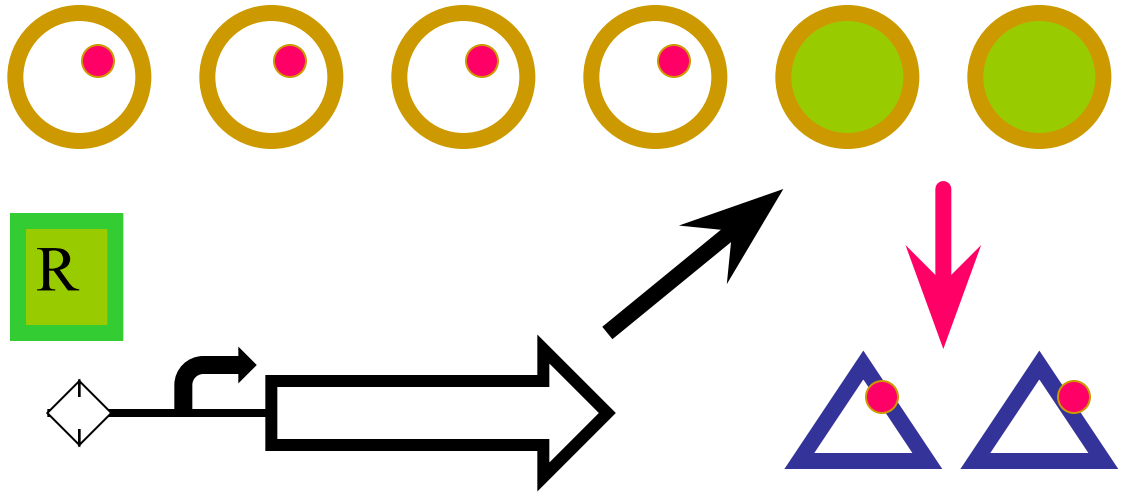


Регуляторный механизм

Sufficient Zn



Zn starvation



Предсказание …

(Proc Natl Acad Sci U S A. 2003 Aug 19;100(17):9912-7.)

Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins

Ekaterina M. Panina*†, Andrey A. Mironov*‡, and Mikhail S. Gelfand*‡§

… и подтверждения

Zinc is a key factor in controlling alternation of two types of L31 protein in the *Bacillus subtilis* ribosome

Hideaki Nanamiya,[1†] Genki Akanuma,[1†]
Yousuke Natori,[1] Rikinori Murayama,[1] Saori Kosono,[2]
Toshiaki Kudo,[2] Kazuo Kobayashi,[3]
Naotake Ogasawara,[3] Seung-Moon Park,[4] Kozo Ochi[4]
and Fujio Kawamura[1*]

(Mol Microbiol.
2004 Apr;52(1):273-83.)

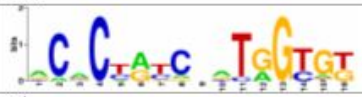
Journal of Bacteriology, April 2006, p. 2715-2720, Vol. 188, No. 7
0021-9193/06/$08.00+0 doi:10.1128/JB.188.7.2715-2720.2006
Copyright © 2006, American Society for Microbiology. All Rights Reserved.

Liberation of Zinc-Containing L31 (RpmE) from Ribosomes by Its Paralogous Gene Product, YtiA, in *Bacillus subtilis*

Genki Akanuma, Hideaki Nanamiya, Yousuke Natori, Naofumi Nomura, and Fujio Kawamura*

Регуляторная система «с нуля под ключ»

- Консервативный сигнал перед генами рибонуклеотид-редуктаз
- Потенциальный регулятор (через филогенетический паттерн + домены)

1	2	3	4	5
α-proteobacteria		±	none in *Wolbacia*, *Rickettsia* spp.	
β-proteobacteria		+	all	
γ-proteobacteria		±	none in *Buchnera*, *Wigglesworthia*	*topA* in *Pseudomonas* spp.; *dnaA* in *Shewanella* spp.
δ-proteobacteria		±	none in *Desulfovibrio* spp.	*dnaA* in *Myxococcus xanthus*, *Desulfotalea psychrophila*; *parA* in *Desulfuromonas* spp.
ε-proteobacteria	n/a	–	none	
Cyanobacteria		±	none in *Nostoc* sp.	
Bacteroidetes/Chlorobi	n/a	–	none	
Bacillus/Clostridium		±	none in *Leuconostoc mesenteroides*, *Oenococcus oeni*	*dgk-pnuC* in lactobacilli; *nucA* in *Lactococcus lactis*; *yvdD-yvdC* in *Bacillus* spp.; *ligA* in *Clostridium acetobutylicum*
Mycoplasmatales	n/a	–	none	
Actinobacteria		+	all	
Spirochaetes	n/a	±	only in *Treponema denticola*	
Thermus/Deinococcus		+	all	
Thermotogales		+	all	
Chlamydiales		+	all	
Other	n/a	±	in *Pirellula* sp., *Chloroflexus aurantiacus*, *Fusobacterium nucleaticum*, none in *Aquifex aeolicus*	

Другие члены регулона

- Реутилизация дезоксирибонуклеотидов

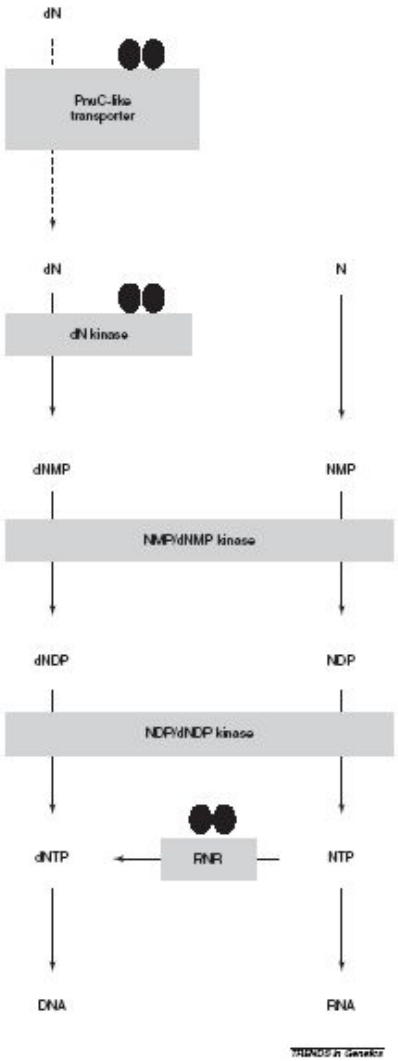
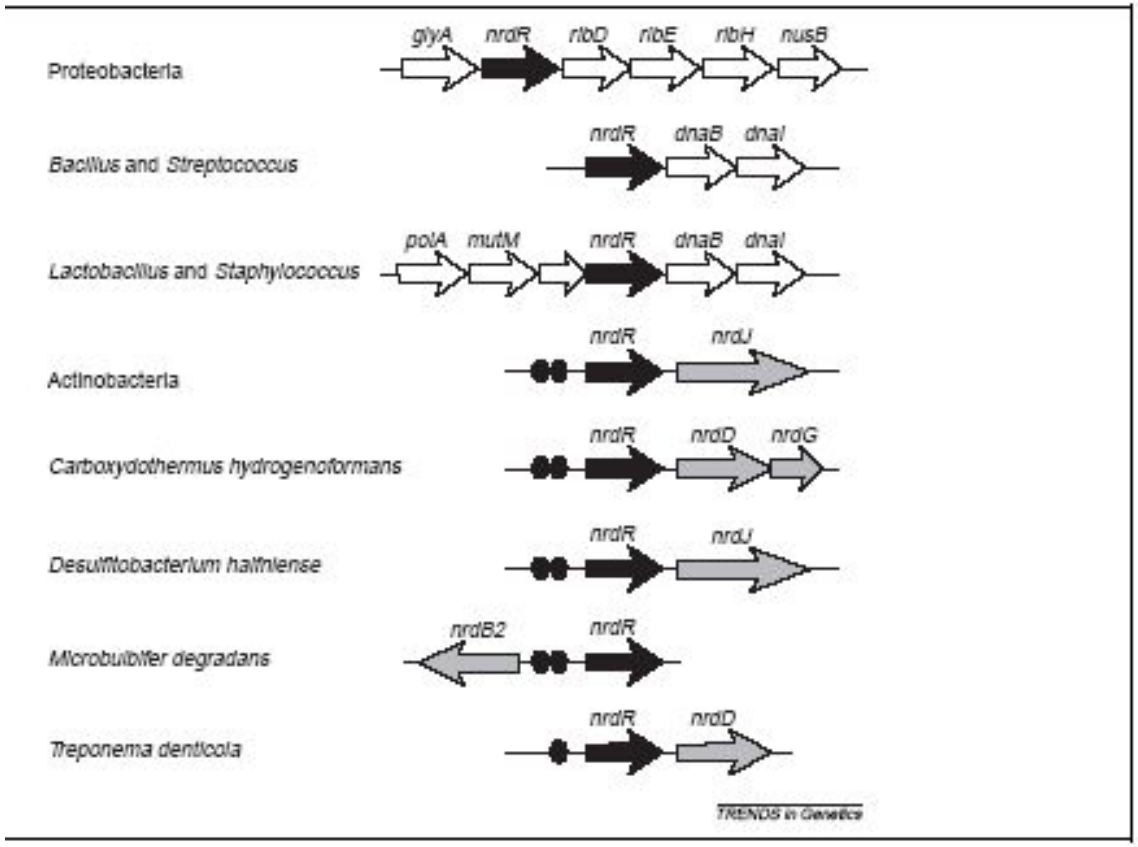


Figure 84. Two alternative pathways of the deoxyribonucleotides synthesis which are predicted to be regulated by NrdR in Gram-positive lactobacilli. dN/N (deoxy)nucleoside, dNMP/NMP (deoxy)nucleoside monophosphate, dNDP/NDP (deoxy)nucleoside diphosphate, dNTP/NTP (deoxy)nucleoside triphosphate. Dots represent candidate NrdR-boxes.

- Репликация (ДНК-лигазы, топоизомеразы, ДНК-полимеразы



tion of the nrdR-containing loci in some bacterial genomes. Genes encoding the predicted ribonucleotide reductase regulator NrdR and the ripponents are shown in black and grey, respectively. The black circles indicate the predicted NrdR-sites. The direction of transcription is indicated

Как регулируется: репрессия в результате кооперативного связывания

A.

```
EC   TGCTTTTTACTTTGAGCTACATCAAAAAAAGCTCAAACATCCTTGATGCAAAGCACTATATATAGACTTTAAAATGCGTCCCAACCCAATATGTTGTATT
ST   TGATTTTTACCTTGTTCTACATCAATAAAATTGCAAACATCCTTGATGCAAATCACTACATATAGACTTTAAAATGCACGCCGACCCAATATGTTGTATT
KP   ACCTTTTTACCTTGTTCTGGGTCAATAAAATCGCAAACATCTTTGATGCAAATCACTACATATAGAACTTAAAATGCGCCTCGGCCCAACATATTGTATT
        ******* ***  **   **** ****   ******** ********** ***** *******  ********    *  ***** ** *******

EC   AATCGACTATAATTGCTACTACAGCTCCCCACG--AAAAAGGTGCGGCGTTGTGGATAAGC-GGATGGCGATTGCGGA-AAGCACCGGAAAACGAAACGA
ST   AATTGACTACAATTGCTACAACACCTGTTCACT--CGACACAAGGTGAATTGTGGATAACCTGGGTCAGGATTGCGGG-AAGTCATTGGAAAAGAGATGA
KP   AATCGTCTATTAT-GTCACCATATCTTGTCGATGTCTGGCGGTGATGAGATGTGGATAAAACGGGCCGGATCCGAAGGTAAACAGCACGAGCCGTAGCGT
     *** * ***  ** *  ** * * **   *             *  *   *********   **         *  *  **        *   *    *

EC   AAAAACCGGAAAACGCCTTTCCCAATTTCTGTGGATAACCTGTTCTTAAAAATATGGAGCGATCATGACACCGCATGTGATGAAACGAGA
ST   ATAAACCTGTTA-TGGCTTCCCCGGCCTCTGTGGATAACCTGTTCTTACAAATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
KP   GCAGCGCCTTCG-GGATAACCTCCGCCTCTGTGGATAACCTGTTCT---ATATATGGAGTGATCATGACACCGCATGTGATGAAACGTGA
       *   *       *     * *    *******************   * ******** ***************************** **
```

B.

```
YP   AACAGGGAATAACCC-TAACGCC--AATTTCCTTGTTCTAGGTCAACAATATTGGCTATCAGTTGACTGTCACTCATCCAGATACCCATATATAGTGTCT
YE   AACAGGGAATAAACC-TAAAGCT--GATTTCCTTGTTCTAGGTCAATTAT-----------GTTGACTGTCACTTCTGCCATTACCCATATATAGTGTCT
Eca  AAGTTCGATTTATCTACTAGGGAGGAATTTCTTTGCTCTACATCAATTTTGCAGCGATAAAAGTGCAAACACCCCTACGCAATTTCAATATATAGTGCCT
Ech  AAGACTGATTTCTCTACGATGCCGGAATTTCCTTGAGCCAGGTCAATTCTAACGCAATAAAACCGGGTCCCCCTCCAGGCGAATTCAATATATAGTGTCT
     **    ** *   *    * *     ***** ***  * *  ****   *             *       *           * ********** **

YP   ATAATATTTTAAGCATCTATATGTAGTAGTTATCCACAAAAGCATCCACATCCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAA-----
YE   ATAGTAATTACGATACCTATATGTAGTAGTTATCCACAAAACCATCCACA-CCCCCTCGCAGCCCTGATGTGCTGCGGTTTGC-CTTGTGGATAAGATGG
Eca  ATCCTGTAAACATTACCTACATATAGTGTTTATCCACAAAGTCATCCACA-GCCCTCTGTAACCCTTGCCAGTTACGGTTCTCGCCTGTGGATAAC----
Ech  ATCCTTTTACAAGAACCTACATATTGTGTTTATCCACAAGAACATCCACA-GCC-TCCGCACCCGTTGTCAGCCGCGGCTTCCGTCTGTGGATAAC----
     **  *         * *** ** * **  **********   ********  **    * * ** *     *   *** *  *   *********

YP   ------------CCCTATGCGGCGGTATACAGGAGTGACATTGTGAAAACAGTAGTGATTAAACGGGACGGCTGCCAGGT
YE   TTTTGGGGCTAATCCTACGCGGCAGGATACAGGAGCGACATTGTGAAAACAGTAGTGATTAAACGGGACGGTTGTCAGGT
Eca  ------------CTTTCCAG----------AGGAAGAA-AACGTGAAACCAGTAGTGATTAAACGGGACGGTTGCCAGGT
Ech  ---------ATCAACAAAGGAAGAACACCGAGGAACAAC---ATGAAACCAGTAGTGATTAAACGGGACGGATGTCAGGT
                                 ****   *     ***** ********************** ** *****
```

Что осталось за кадром

- Эукариоты
- Структуры
- Молекулярная эволюция
  - Гены
  - Геномы
  - Метаболические и регуляторные системы
- Другие виды данных и что с ними делать
  - Экспрессия
  - Белок-ДНКовые взаимодействия
  - Белок-белковые взаимодействия
  - Структура хроматина (метилирование, гистоны и их модификации и т.д.)
- «Системная биология»