Биоинформатика: биологические тексты

М.С.Гельфанд

29 октября 2006

Первый фестиваль науки МГУ
Факультет биоинженерии и биоинформатики

Расшифрован геном!

расшифрован геном человека - Поиск в Google - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.google.ru/search?hl=ru&q=%D1%80%D0%B0%D1%81%D1%88%D0%

Go Links

Войти

Google

Веб Картинки Группы Каталог **Дополнительно »**

расшифрован геном человека | Поиск | Расширенный поиск Настройки

Поиск в Интернете | Поиск страниц на русском

Веб Результаты **1** - **10** из примерно **93 000** для **расшифрован геном человека**. (**0,22** секунд)

Расшифрована структура **генома человека**. Статьи. Наука и техника
Расшифрована структура **генома человека**. Татьяна БАТЕНЕВА. С одной стороны, можно считать, что начался отсчет новой эпохи, сулящей победу над болезнями, ...
www.n-t.ru/tp/in/rs.htm - 12k - Сохранено в кэше - Похожие страницы

14.04.2003 **Геном человека расшифровали** до конца / Новости ...
Программа «**Геном человека**» завершена. Сейчас исследователям известен практически весь генетический код **человека**, а дальнейшая его расшифровка, по их словам, ...
83.222.3.160:8101/mednovosti/news/2003/04/14/genome/ - 21k -
Сохранено в кэше - Похожие страницы

Расшифрован геном мыши: отличие от **человека** - не более процента / ROL
Расшифрован геном мыши: отличие от **человека** - не более процента. 5 декабря 2002 г. С точки зрения генетиков, **человек** мало чем отличается от мыши. ...
www.rol.ru/news/med/news/02/12/05_001.htm - 20k -
Сохранено в кэше - Похожие страницы

Журнальный зал | Знамя, 2000 N10 | Лев Киселев / Валерий Сойфер ...
Закончилось начало — **геном человека расшифрован**. Дальше — самое главное, самое интересное, самое трудное — как же все-таки наши **гены** через свои продукты ...
magazines.russ.ru/znamia/2000/10/konfer.html - 97k -
Сохранено в кэше - Похожие страницы

Расшифрован геном человека
Расшифрован геном человека ... После этого возможна публикация окончательной версии **генома человека**. Из заявления Вентера следует, что полные данные ...
www.ma-ma.ru/ru/library/news/28823.php - 26k -
Сохранено в кэше - Похожие страницы

Элементы - новости науки: Полностью **расшифрован геном** собаки
Полностью **расшифрован геном** собаки ... Генетическое сходство с **человеком** (а **геном** собаки ближе к человеческому, чем, например, **геном** мыши) позволит ...
elementy.ru/news/165035 - 29k - Сохранено в кэше - Похожие страницы

Mass Media : **Расшифрован геном** черного тополя
Расшифрован геном черного тополя. 20.09.2006 14:33 | Радио Свобода ... В США объявлен конкурс на доступный способ расшифровки **генома человека** (NEWSru.com) ...
www.rambler.ru/news/science/dolly/8731606.html?t_detail=1 - 41k -
Сохранено в кэше - Похожие страницы

ALTEX Laboratory. Система автоматизации лабораторных исследований

https://www.google.com/accounts/Login?continue=http://www.g | Internet

расшифрован геном - Поиск в Google - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.google.ru/search?hl=ru&q=%D1%80%D0%B0%D1%81%D1%88%D0%B8%D1

Go Links

Войти

Google

Веб Картинки Группы Каталог **Дополнительно »**

расшифрован геном | Поиск | Расширенный поиск Настройки

Поиск в Интернете | Поиск страниц на русском

Веб Результаты **1** - **10** из примерно **228 000** для **расшифрован геном**. (**0,25** секунд)

Элементы - новости науки: Полностью **расшифрован геном** собаки
Консорциум генетиков завершил расшифровку генетического кода собаки — породистого боксера. В процессе работы была также составлена карта однонуклеотидных ...
elementy.ru/news/165035 - 29k - Сохранено в кэше - Похожие страницы

MEMBRANA | Полужизнь: **расшифрован геном** самого загадочного ...
Mimi, или Mimivirus, номинально, самый большой вирус в мире. Но фактически учёные не могут подобрать для него самую подходящую ветку на древе жизни.
www.membrana.ru/articles/global/2004/10/15/214500.html - 54k -
Сохранено в кэше - Похожие страницы

Mass Media : **Расшифрован геном** черного тополя
Расшифрован геном черного тополя. 20.09.2006 14:33 | Радио Свобода. Международный коллектив генетиков, под руководством профессора Джеральда Таскэна (Gerald ...
www.rambler.ru/news/science/dolly/8731606.html - 39k -
Сохранено в кэше - Похожие страницы

"Аккумулятор новостей": самые свежие события и факты – интересно и ...
США: ученые впервые **расшифровали геном** курицы. Ученые впервые опубликовали детальный отчет о расшифровке **генома** курицы. Об этом пишет журнал Nature. ...
news.battery.ru/dossier/?dsrId=346&
from_m=dossier&from_n=0&newsId=23532482 - 43k -
Сохранено в кэше - Похожие страницы

"Аккумулятор новостей": самые свежие события и факты – интересно и ...
Франция: **расшифрован геном** самого загадочного биологического объекта. Mimi, или Mimivirus, номинально, самый большой вирус в мире. ...
news.battery.ru/dossier/?dsrId=346&
from_m=dossier&from_n=0&newsId=20429953 - 47k -
Сохранено в кэше - Похожие страницы
[Дополнительные результаты с news.battery.ru]

Расшифрован геном шимпанзе - Компьюлента
Геном наших ближайших родственников состоит из 2,8 миллиардов пар оснований ДНК. Теперь генетики будут пытаться отыскать различия между геномом шимпанзе и ...
science.compulenta.ru/225318/ - 35k - Сохранено в кэше - Похожие страницы

Lenta.ru: Прогресс: **Расшифрован геном** собаки
В американском институте Брода **расшифровали геном** собаки, сообщает Reuters. Генетики считают, что их открытие может пригодиться при поиске лекарств от ...
lenta.ru/news/2005/12/08/dog/ - 18k - Сохранено в кэше - Похожие страницы

Russian America Online - dating, journal, news, discussions ...

http://64.233.183.104/search?q=cache:YcRTcfUl9MsJ:elementy.ru/news, | Internet

Расшифрован ли геном?

- Перехватить зашифрованное сообщение – еще не значит его понять

Фрагмент генома (0.1% генома *E. coli*)

Геном бактерии: несколько миллионов нуклеотидов

От 600 до 9 тысяч генов (примерно 90% генома кодирует белки)

Фрагмент генома (0.0001% генома человека)

cgtgcacttctgaaggacttcaggtaccggcgtgcccgcgtcctactgtccgcctgctcgcgtcctgggtgccgcctctgagtagggcgggcgaggagg
CAGCCAAGGCGGAGCTGATGGCTGCGCCGAGGGCGGGGCGGGGTGCAGGCTGGAGCCTTCGGGCATGGCGGGCTTTGGGGGGCATTCGCTGGGGGAGGAG
ACCCCGTTTGACCCCTGACCTCCGGGCCCTGCTGACGTCAGGAACTTCTGACCCCCGGGCCCGAGTGACTTATGGGACCCCCAGTCTCTGGGCCCGGTTG
TCTGTTGGGGTCACTGAACCCCGAGCATGCCTGACGTCTGGGACCCCGGGTCCCCGGGCACAACTGACTGCGGTGACCCCAGATACCAGGACCCGGGAGG
CCTCAGAGAACTCTGGAACCCGTTCGCGCGCGTGGCTGGCGGTGGCGCTGGGCGCTGGGGGGGCAGTGCTGTTGTTGTTGTGGGGCGGGGGTCGGGGTCC
TCCGGCCGTCCTCGCCGCCGTCCCTAGCCCGCCGCCCGCTTCTCCCCGGAGTCAGTACAACTTCATCGCAGATGTGGTGGAGAAGACAGCACCTGCCGTG
GTCTATATCGAGATCCTGGACCGgtaatggtgggggtagaccgggaggcactgaagccacaggctggagggcgggcgggtaggaggggtcagagcctcct
cttatctgtgctttccctccatttcagGCACCCTTTCTTGGGCCGCGAGGTCCCTATCTCGAACGGCTCAGGATTCGTGGTGGCTGCCGATGGGCTCATT
GTCACCAACGCCCATGTGGTGGCTGATCGGCGCAGAGTCCGTGTGAGACTGCTAAGCGGCGACACGTATGAGGCCGTGGTCACAGCTGTGGATCCCGTGG
CAGACATCGCAACGCTGAGGATTCAGACTAAGgtgggggctggggtaggccaggtctggttggagctgcttatttgctcgcatcttcagatgacaggtct
cttttacccattctcccttaggagcctctccccacgctgcctctgggacgctcagctgatgtccggcaaggggagtttgttgttgccatgggaagtccct
ttgcactgcagaacacgatcacatccggcattgttagctctgctcagcgtccagccagagacctgggactcccccaaaccaatgtggaatacattcaaac
tgatgcagctattgatgtgcgtcctgataggagagaaatgacaaatgatgggggagggggagaggctgtgtggtacaagcaccaactgatatatggtgg
atgagcctatatagagcttaggctgcaaaaatgtggccacttattcatgggctgagaaagaagagaatttggagaaagtacctacatcctggtatgcccc
cagacttagaatccccagatctctttcatgttttctccttgtcctacagTTTGGAAACTCTGGAGGTCCCCTGGTTAACCTGgtgagtgagacatccttc
cttccaagaatccctgccccaggtcagtgtgggaagggtaggtttcccctaattcaaggatgtttggtcaagtttctgagcagttctttgttggctatct
ctcaatatccaaccagatctccccaacacttgctggtacttttgttcgggtgcccccatcccctactatttgtttaggctagggaactgggggctgtatc
cctgcagGATGGGGAGGTGATTGGAGTGAACACCATGAAGGTCACAGCTGGAATCTCCTTTGCCATCCCTTCTGATCGTCTTCGAGAGTTTCTGCATCGT
GGGGAAAAGAAGAgtgagcctgccttatggggaaacgggttcctttaatgtggtggaaatagggggaagggcattcagtgggacttcctggagggtggtct
actgggagaagagggcagggaaggaaggatgtagctgggtggggctcatttgtccctctgtcacagATTCCTCCTCCGGAATCAGTGGGTCCCAGCGGCG
CTACATTGGGGTGATGATGCTGACCCTGAGTCCCAGgtatgagctttagggacagtgacatgtaatgtgaccagtgtaatcagagggggcacctctatt
gagctttgttctcatttctgtctttatctaagatgaactgtgtcacacttgaaataatcacaagagctgtctcccttcatcatcttgactttcttatccc
actccactttgtacacctgtcaccagattgatttcatcctgttactgctttgatttcaagccttcaatccattaacttggcatttaagggccattttcca
tctgtctgtaaatcaactttctagacctggctgtaataccttcctatatgaatactcagccaacctgatttcctactcccatgtttttatttatgctgt
tcttgttttatatggttaatat gct
attgatttaacactgtttgtca cat
tgtattttgtagtacctagccc ttt
ggctaatagggtgatctgtgta aag
tcatcctgggctcccctgcaca cct
tcctttctctctgtccatttttctctatagGGCTGGTCTGCGGCCTGGTGATGTGATTTTGGCCATTGGGGAGCAGATGGTACAAAATGCTGAAGATGTT
TATGAAGCTGTTCGAACCCAATCCCAGTTGGCAGTGCAGATCCGGCGGGGACGAGAAACACTGACCTTATATGTGACCCCTGAGGTCACAGAATGAATAG
ATCACCAAGAGTATGAGGCTCCTGCTCTGATTTCCTCCTTGCCTTTCTGGCTGAGGTTCTGAGGGCACCGAGACAGAGGGTTAAATGAACCAGTGGGGGC
AGGTCCCTCCAACCACCAGCACTGACTCCTGGGCTCTGAAgAATCACAGAAACACTTTTTATATAAAATAAAATTATACCTAGCaacatattatagtaaa

Геном человека: 3 000 000 000 нуклеотидов

Примерно 25 тысяч генов, < 5% генома кодирует белки

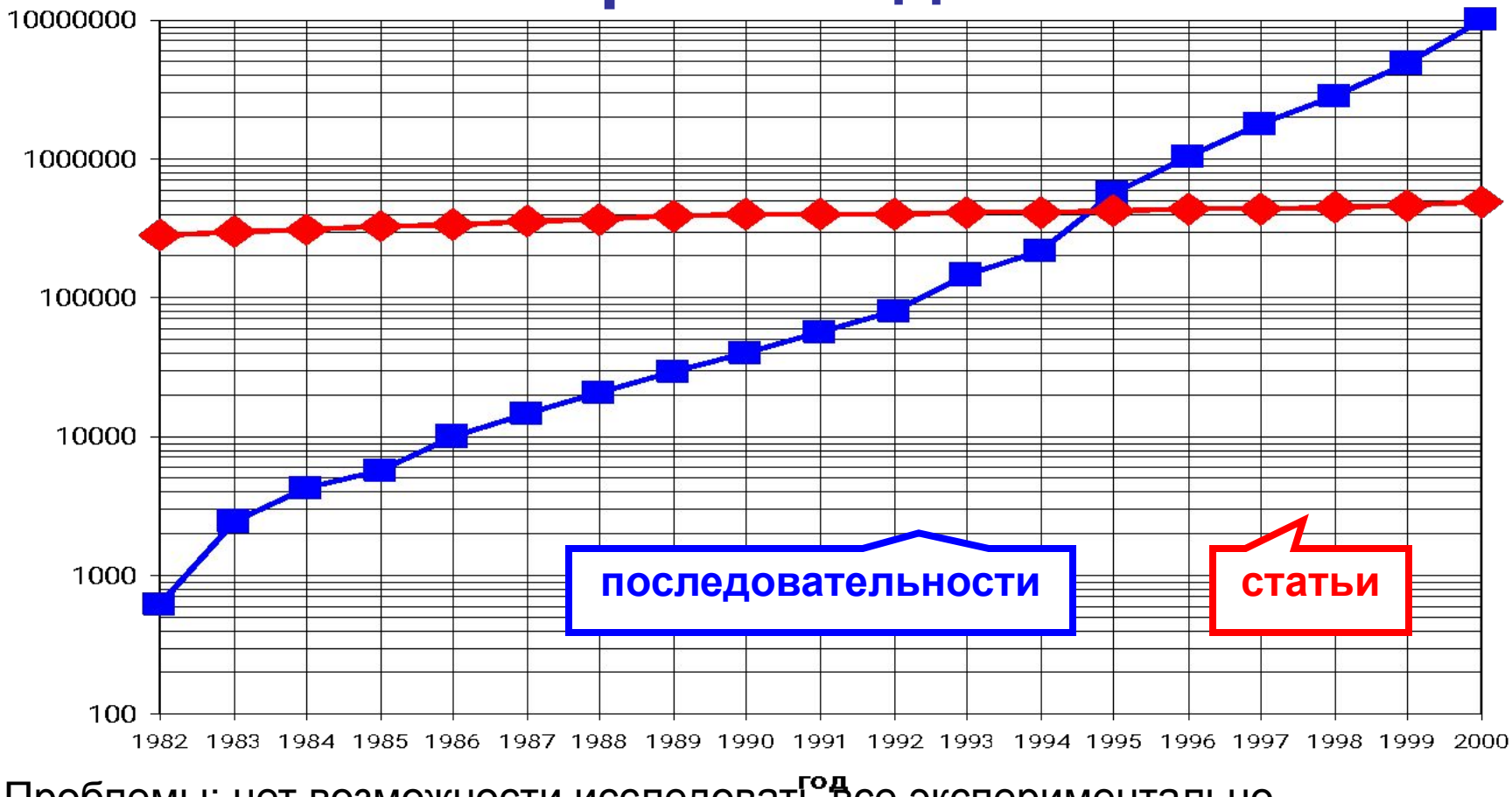
Что же мы хотим понять?

- Где
  - картировать гены в геноме и определить аминокислотные последовательности кодируемых белков
- Что
  - предсказать функции генов (кодируемых белков)
- Когда
  - описать регуляцию генов, зависимость экспрессии от внешних условий и внутреннего состояния клетки
- Где – 2
  - определить локализацию белка в клетке (или вне её)

TTT	F	TCT	S	TAT	Y	TGT	C
TTC	F	TCC	S	TAC	Y	TGC	C
TTA	L	TCA	S	TAA	*stop*	TGA	*stop*
TTG	L	TCG	S	TAG	*stop*	TGG	W
CTT	L	CCT	P	CAT	H	CGT	R
CTC	L	CCC	P	CAC	H	CGC	R
CTA	L	CCA	P	CAA	Q	CGA	R
CTG	L	CCG	P	CAG	Q	CGG	R
ATT	I	ACT	T	AAT	N	AGT	S
ATC	I	ACC	T	AAC	N	AGC	S
ATA	I	ACA	T	AAA	K	AGA	R
ATG	M/ *start*	ACG	T	AAG	K	AGG	R
GTT	V	GCT	A	GAT	D	GGT	G
GTC	V	GCC	A	GAC	D	GGC	G
GTA	V	GCA	A	GAA	E	GGA	G
GTG	V	GCG	A	GAG	E	GGG	G

Таблица генетического кода

Пропаганда



- Проблемы: нет возможности исследовать все экспериментально
- Возможности: можно использовать методы сравнительной геномики
  - > 1000 геномов бактерий (~400 полных)
  - простейшие: малярийный плазмодий, инфузория, лейшмания, …
  - растения: арабидопсис, тополь, рис
  - пара дюжин дрожжей и другие грибы
  - насекомые: дюжина дрозофил, комар, пчела, …
  - позвоночные: человек, шимпанзе, мышь, крыса, собака, курица, 2 рыбы, …

Поиск генов если известен белок: просто

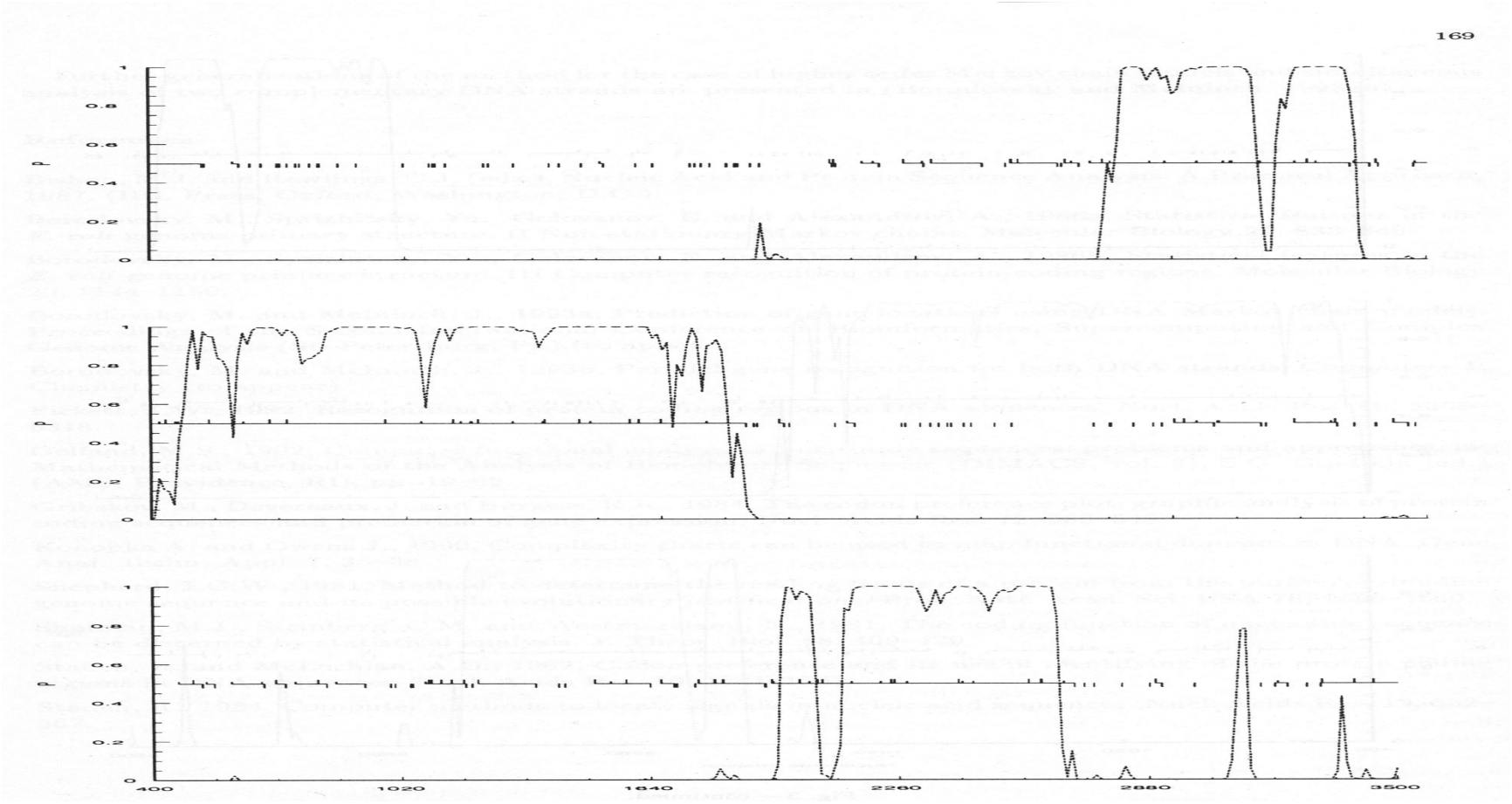
```
aaacgccctggccttctatagccactattatcagccgctgccgttggtactgcgcggata

tggtgcgggcaatgacgttacagctgccggtgtctttgctgatctgctacgtaccctctc

                        M  V  K  V  Y  A  P  A  S  S  A  N  M
atggaagttaggagtctgacatggttaaagtttatgccccggcttccagtgccaatatga

S  V  G  F  D  V  L  G  A  A  V  T  P  V  D  G  A  L  L  G
gcgtcgggtttgatgtgctcggggcggcggtgacacctgttgatggtgcattgctcggag

D  V  V  T  V  E  A  A  E  T  F  S  L  N  N  L  G  R  F  G
atgtagtcacggttgaggcggcagagacattcagtctcaacaacctcggacgctttgggt

A  D  K  L  P  S  E  P  R  E  N  V  Y  Q  C  W  E  R  F  C
ccgataagctgccgtcagagccacgggaaaatgtttatcagtgctgggagcgtttttgcc

Q  E  L  G  K  Q  I  P  V  A  M  T  L  E  K  N  M  P  I  G
aggagcttggcaagcaaattccagtggcgatgactctggaaaagaatatgccgatcggtt

F  V  H  I  C  R  L  D  T  A  G  A  R  V  L  E  N
ttgttcatatttgccggctggatacggcgggcgcacgagtactggaaaactaaatgaaac

tctacaatctgaaagatcacaatgagcaggtcagctttgcgcaagccgtaacccaggggt

tgggcaaaaatcaggggctgtttttcccgcacgacctgccggaattcagcctgactgaaa
```

… или родственный белок: тоже просто

```
aaacgccctggccttctatagccactattatcagccgctgccgttggtactgcgcggata

tggtgcgggcaatgacgttacagctgccggtgtctttgctgatctgctacgtaccctctc

                        M  V  V  V  Y  A  P  A  S  I  G  N  V
atggaagttaggagtctgacatggttaaagtttatgccccggcttccagtgccaatatga

S  V  G  F  D  V  L  G  A  A  V  S  P  I  D  G  S  L  L  G
gcgtcgggtttgatgtgctcggggcggcggtgacacctgttgatggtgcattgctcggag

D  R  V  E  V  A  A  G  E  Q  P  F  T  L  K  C  V  G  D  F
atgtagtcacggttgaggcggcagagaca---ttcagtctcaacaacctcggacgctttg

V  A  K  L  P  V  E  Q  E  E  N  V  Y  H  C  W  L  V  F  A
ccgataagctgccgtcagagccacgggaaaatgtttatcagtgctgggagcgttttgcc

R  E  L  D  K  K  V  P  V  S  M  T  L  E  K  N  M  P  I  G
aggagcttggcaagcaaattccagtggcgatgactctggaaaagaatatgccgatcggtt

F  V  H  V  C  R  L  D  S  T  G  S  K  V  L  E  N
ttgttcatatttgccggctggatacggcgggcgcacgagtactggaaaactaaatgaaac

tctacaatctgaaagatcacaatgagcaggtcagctttgcgcaagccgtaacccaggggt

tgggcaaaaatcaggggctgtttttcccgcacgacctgccggaattcagcctgactgaaa
```

Статистические особенности

- Отличия в частотах олигонуклеотидов в кодирующих и некодирующих областях



GenMark

- Стартовые кодоны

Начала генов *Bacillus subtilis*

dnaN	ACATTATCCGTTAGGAGGATAAAA**ATG**
gyrA	GTGATACTTCAGGGAGGTTTTTTA**ATG**
serS	TCAATAAAAAAAGGAGTGTTTCGC**ATG**
bofA	CAAGCGAAGGAGATGAGAAGATTC**ATG**
csfB	GCTAACTGTACGGAGGTGGAGAAG**ATG**
xpaC	ATAGACACAGGAGTCGATTATCTC**ATG**
metS	ACATTCTGATTAGGAGGTTTCAAG**ATG**
gcaD	AAAAGGGATATTGGAGGCCAATAA**ATG**
spoVC	TATGTGACTAAGGGAGGATTCGCC**ATG**
ftsH	GCTTACTGTGGGAGGAGGTAAGGA**ATG**
pabB	AAAGAAAATAGAGGAATGATACAA**ATG**
rplJ	CAAGAATCTACAGGAGGTGTAACC**ATG**
tufA	AAAGCTCTTAAGGAGGATTTTAGA**ATG**
rpsJ	TGTAGGCGAAAAGGAGGGAAAATA**ATG**
rpoA	CGTTTTGAAGGAGGGTTTTAAGTA**ATG**
rplM	AGATCATTTAGGAGGGGAAATTCA**ATG**

Участок связывания рибосом

dnaN	ACATTATCCGTTAGGAGGATAAAAATG
gyrA	GTGATACTTCAGGGAGGTTTTTTAATG
serS	TCAATAAAAAAAGGAGTGTTTCGCATG
bofA	CAAGCGAAGGAGATGAGAAGATTCATG
csfB	GCTAACTGTACGGAGGTGGAGAAGATG
xpaC	ATAGACACAGGAGTCGATTATCTCATG
metS	ACATTCTGATTAGGAGGTTTCAAGATG
gcaD	AAAAGGGATATTGGAGGCCAATAAATG
spoVC	TATGTGACTAAGGGAGGATTCGCCATG
ftsH	GCTTACTGTGGGAGGAGGTAAGGAATG
pabB	AAAGAAAATAGAGGAATGATACAAATG
rplJ	CAAGAATCTACAGGAGGTGTAACCATG
tufA	AAAGCTCTTAAGGAGGATTTTAGAATG
rpsJ	TGTAGGCGAAAAGGAGGGAAAATAATG
rpoA	CGTTTTGAAGGAGGGTTTTAAGTAATG
rplM	AGATCATTTAGGAGGGGAAATTCAATG

Сравнение генов в родственных геномах

```
Sty  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Stm  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Sen  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Eco  TTGCCCG--TGCCAGACGGCAGATTATCTCCCTGACCTGGTGGTTGCCCAGGAGGAGGGCCGGAAATAGGTTGTATCATT
Kpn  ----CGG--TGGCGCAGTGCCTGATGGG-CCTCGCCCTGGAGGACGGTCTGGCAT---ATCAGCAAGGGGGTGCGTCATG
Ype  TTGTTAGAACAGGGGAAAACGGTAAACAGTGTGGCATTAGATGTCGGTTATAGCT-----CCGCCTCTGCTTTTATCGCC
           *        *                     * *  *  *              * *     *  *   *

Sty  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCTTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Stm  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCCTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Sen  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCCTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Eco  ACGTATCCTTATAC----------CTGAAATCTTCGCAAG--TATGCCTGGCCGCGAGATTATGGC--ACACTTGTCCGG
Kpn  ATTCATCCTTTCGATATCGCGGTGCTGGAACCAGGTGATGAGTATGCCTGGCGGCCAGATTATGGC--ACACTTCCCCAG
Ype  ATGTTTCAGCAAATAT--------CGGGTACCA-CGCCTGAGCGTTTCCGGCGGGGCAATAGTGGCTTATACTAAGCCCC
     *    **                 *    * *            *    *** *    **  ****  * ***   **

Sty  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Stm  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Sen  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Eco  TTAACTCTCGT--CTCATACAG------GTAACACAAAC--GTGAAAATCCTTGTTGATGAAAATATGCCTTATGCCCGC
Kpn  TTAACTCTCGTT-CTCAGACAG------GTACTGAACT---GTGAAAATCCTCGTTGATGAAAATATGCCCTATGCCCGT
Ype  CTGTTTTTCATCTGTATGGCAGTTCGCTGTCGGAGAGTAAAGTGAAAATTCTGGTTGATGAAAATATGCCGTACGCTGAG
      *   * ** *   *    ***      **     *     ******** ** **************** ** **
```

- Гены консервативнее, чем межгенные области (точнее, особенности эволюции другие)

Мораль

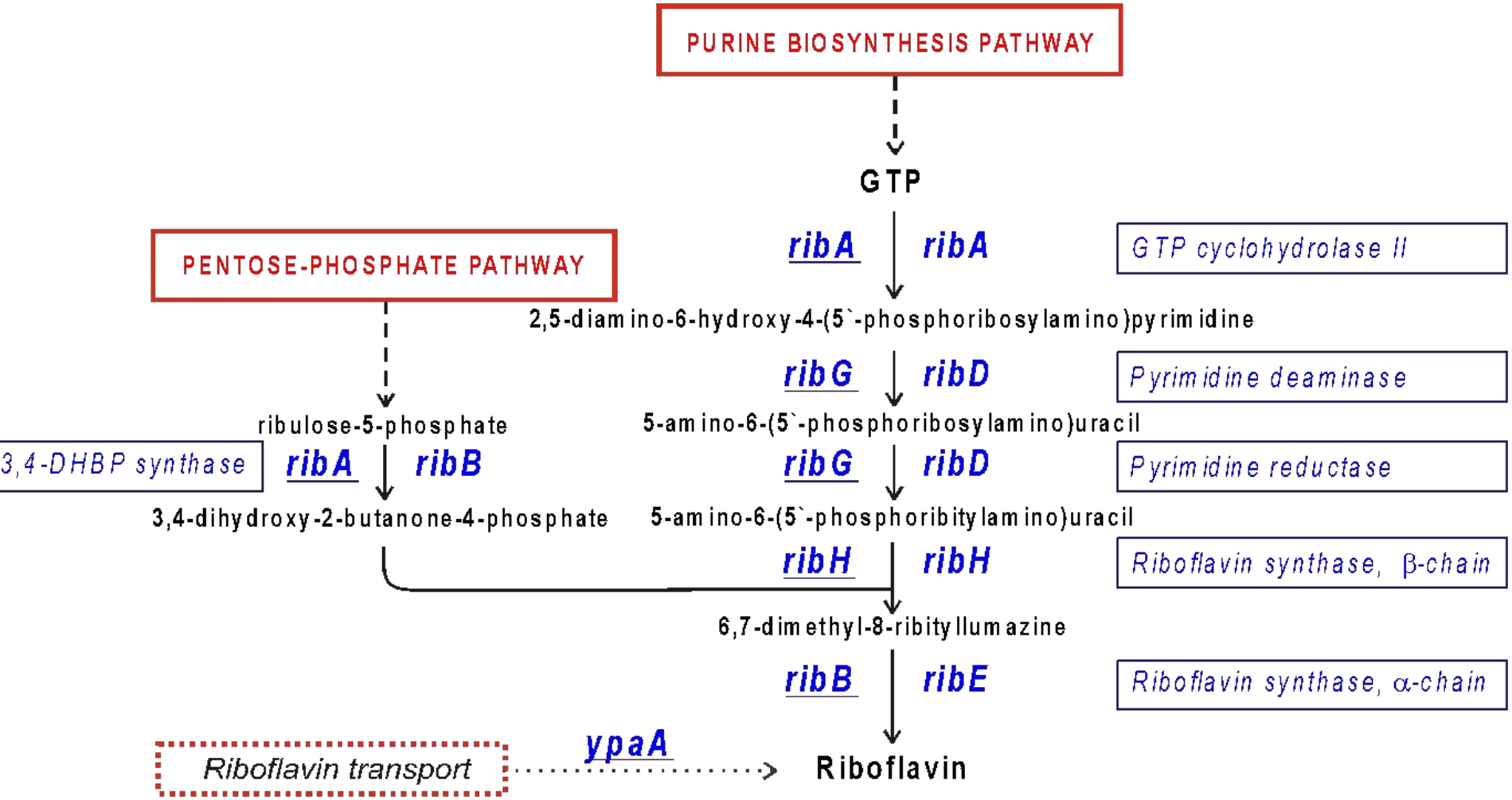
- ***Комплексный подход***: использование многих разнородных соображений, каждое из которых по отдельности – слабое

- ***Сравнительный подход***: одновременный анализ множества геномов (находящихся на различных эволюционных расстояниях друг от друга)

Как предсказывать функции

- Белки, похожие по последовательности, имеют сходные функции
- Если нет родственных белков с известной функцией, то:
  - предсказав структурные особенности, можно определить функциональный класс
  - изучение геномного контекста позволяет отнести белок к функциональной подсистеме

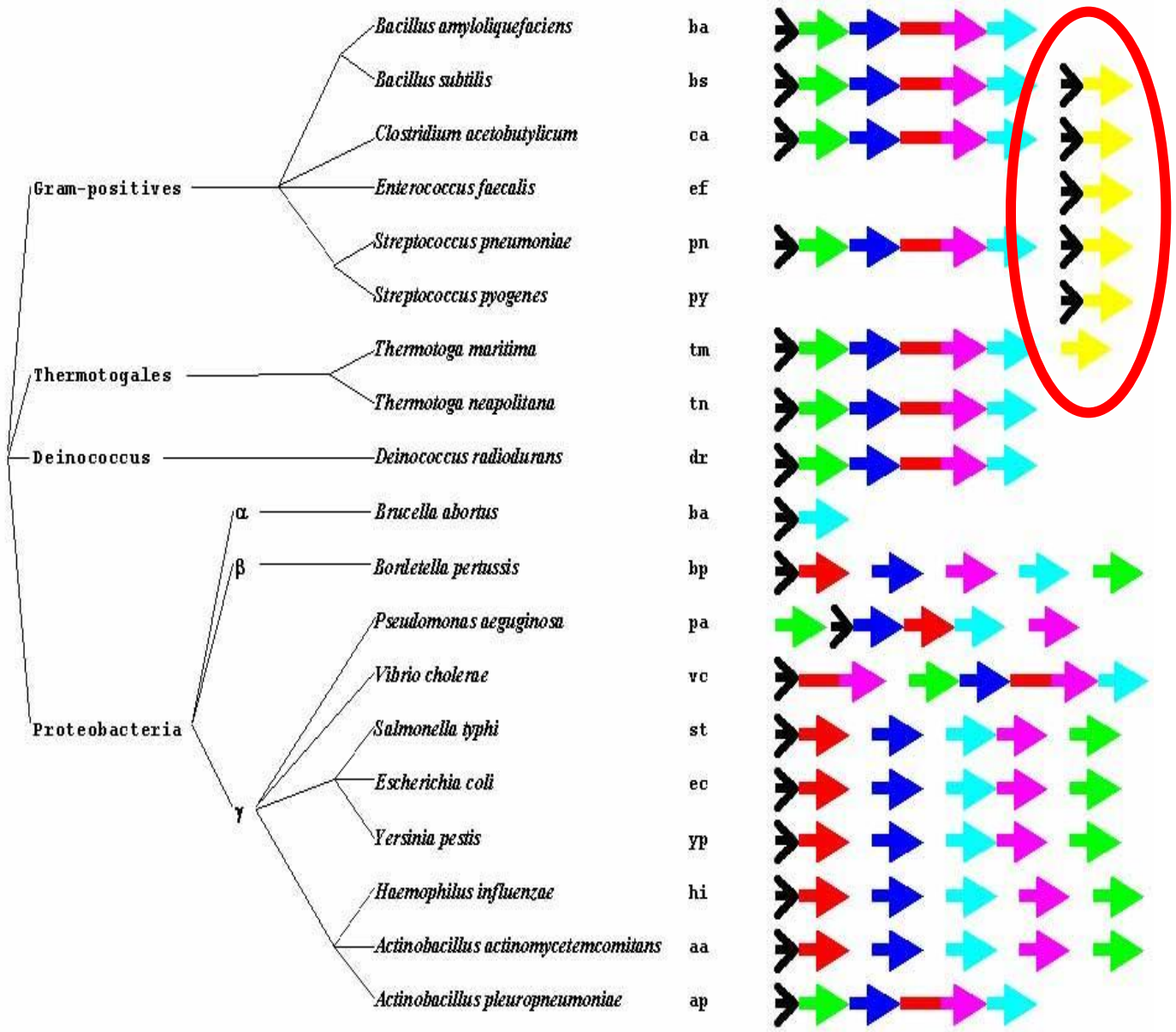
Метаболический путь биосинтеза рибофлавина (витамина B_2)



Консервативная последовательность перед генами рибофлавинового пути

```
                1           2          2'    3     Add.      3'      Variable        4          4'    5          5'            1'
                ==========> ==>      <==   ==>   ->-<-   <==     ->     <-       ==>        <==   ==>      <==      <==========
BS   TTGTATCTTCGGGG-CAGGGTGGAAATCCCGACCGGCGGT  21 AGCCCGTGAC--   8   4   8 -----TGGATTCAGTTTAA-GCTGAAGCCGACAGTGAA-AGTCTGGAT-GGGAGAAGGATGAT
BQ   AGCATCCTTCGGGG-TCGGGTGAAATTCCCAACCGGCGGT  19 AGTCCGTGAC--   8   5   8 -----TGGATCTAGTGAAACTCTAGGGCCGACAGT-AT-AGTCTGGAT-GGGAGAAGGATATG
BE   TGCATCCTTCGGGG-CAGGGTGAAATTCCCGACCGGCGGT  20 AGCCCGCGA---   3   4   3 -----AGGATCCGGTGCGATTCCGGAGCCGACAGT-AT-AGTCTGGAT-GGGAGAAGGATGCC
HD   TTTATCCTTCGGGG-CTGGGTGGAAATCCCGACCGGCGGT  19 AGTCCGTGAC--  10   4  10 -----TGGACCTGGTGAAAATCCGGGACCGACAGTGAA-AGTCTGGAT-GGGAGAAGGAAACG
Bam  TGTATCCTTCGGGG-CTGGGTGAAATCCCGACCGGCGGT  23 AGCCCGTGAC--   8   4   8 -----TGGATTCAGTGAAAAGCTGAAGCCGACAGTGAA-AGTCTGGAT-GGGAGAAGGATGAG
CA   GATGTTCTTCAGGG-ATGGGTGAAATTCCCAATCGGCGGT   2 AGCCCGCAA---   3   4   3 ------AGATCCGGTTAAACTCCGGGGCCGACAGTTAA-AGTCTGGAT-GAAAGAAGAAATAG
DF   CTTAATCTTCGGGG-TAGGGTGAAATTCCCAATCGGCGGT   2 AGCCCGCG----   7   6   7 --------ATTTGGTTAAATTCCAAAGCCGACAGT-AA-AGTCTGGAT-GGAAGAAGATATTT
SA   TAATTCTTTCGGGG-CAGGGTGAAATTCCCAACCGGCAGT   6 AGCCTGCGAC--  11   3  11 -----CTGATCTAGTGAGATTCTAGAGCCGACAGTTAA-AGTCTGGAT-GGGAGAAGAATGT
LLX  ATAAATCTTCAGGG-CAGGGTGTAATTCCCTACCGGCGGT   2 AGCCCGCGA---   4   4   4 -----ATGATTCGGTGAAACTCCGAGGCCGACAGT-AT-AGTCTGGAT-GAAAGAAGATAATA
PN   AACTATCTTCAGGG-CAGGGTGAAATTCCCTACCGGTGGT   2 AGCCCACGA---   3   4   3 -----ATGATTTGGTGAAATTCCAAAGCCGACAGT-AT-AGTCTGGAT-GAAAGAAGATAAAA
TM   AAACGCTCTCGGGG-CAGGGTGGAATTCCCGACCGGCGGT   3 AGCCCGCGAG--   5   4   5 -----TTGACCCGGTGGAATTCCGGGGCCGACGGTGAA-AGTCCGGAT-GGGAGAGAGCGTGA
DR   GACCTCTTTCGGGG-CGGGGCGAAATTCCCCACCGGCGGT  15 AGCCCGCGAA--   8  12   9 -----CCGATGCCGCGCGCAACTCGGCAGCCGACGGTCAC-AGTCCGGAC-GAAAGAAGGAGGAG
TQ   CACCTCCTTCGGGG-CGGGGTGGAAGTCCCCACCGGCGGT   3 AGCCCGCGAA--   5   4   5 -----CCGACCCGGTGGAATTCCGGGGCCGACGGTGAA-AGTCCGGAT-GGGAGAAGGAGGGC
AO   AATAATCTTCAGGG-CAGGGTGAAATTCCCGATCGGCGGT   2 AGTCCGCGA---   7   7   7 -----AGGAACCGGTGAGATTCCGGTACCGACAGT-AT-AGTCTGGAT-GGAAGAAGATGAAA
DU   TTTAATCTTCAGGG-CAGGGTGAAATTCCCGATCGGTGGT   2 AGTCCGCGA---  13   4  12 -----AGGAACTAGTGAAATTCTAGTACCGACAGT-AT-AGTCTGGAT-GGAAGAAGAGCAGA
CAU  GAAGACCTTCGGGG-CAAGGTGAAATTCCTGATCGGCGGT  20 AGCCCGCGA---   3   4   3 -----AGGACCCGGTGTGATTCCGGGGCCGACGGT-AT-AGTCCGGAT-GGGAGAAGGTCGGC
EN   TAAAGTCTTCAGGG-CAGGGTGAAATTCCCGACCGGTGGT   2 AGTCCACG----   5   4   5 -------GATTTGGTGAAATTCCAAAACCGACAGT-AG-AGTCTGGAT-GGGAGAAGAATTAG
TFU  ACGCGTGCTCCGGG-GTCGGTGAAAGTCCGAACCGGCGGT   3 AGTCCGCGAC--   8   5   8 -----TGGAACCGGTGAAACTCCGGTACCGACGGTGAA-AGTCCGGAT-GGGAGGTAGTACGTG
SX   -AGCGCACTCCGGG-GTCGGTGAAAGTCCGAACCGGCGGT   3 AGTCCGCGAC--   8   5   8 -----TTGACCAGGTGAAATTCCTGGACCGACGGTTAA-AGTCCGGAT-GGGAGGCAGTGCGCG
BU   GTGCGTCTTCAGGG-CGGGGTGAAATTCCCCACCGGCGGT  30 AGCCCGCGAGCG     137   GTCAGCAGATCTGGTGAGAAGCCAGAGCCGACGGTTAG-AGTCCGGAT-GGAAGAAGATGTGC
BPS  GTGCGTCTTCAGGG-CGGGGCGAAATTCCCCACCGGCGGT  21 AGCCCGCGAGCG   8   4   8 GTCAGCAGATCTGGTCCGATGCCAGAGCCGACGGTCAT-AGTCCGGAT-GAAAGAAGATGTGC
REU  TTACGTCTTCAGGG-CGGGGTGCAATTCCCCACCGGCGGT  31 AGCCCGCGAGCG   7   5   7 GTCAGCAGATCTGGTGAGAGGCCAGGGCCGACGGTTAA-AGTCCGGAT-GAAAGAAGATGGGC
RSO  GTACGTCTTCAGGG-CGGGGTGGAATTCCCCACCGGCGGT  21 AGCCCGCGAGCG  11   3  11 GTCAGCAGATCCGGTGAGATGCCGGGGCCGACGGTCAG-AGTCCGGAT-GGAAGAAGATGTGC
EC   GCTTATTCTCAGGG-CGGGGCGAAATTCCCCACCGGCGGT  17 AGCCCGCGAGCG   8   4   8 GACAGCAGATCCGGTGTAATTCCGGGGCCGACGGTTAG-AGTCCGGAT-GGGAGAGAGTAACG
TY   GCTTATTCTCAGGG-CGGGGCGAAATTCCCCACCGGCGGT  67 AGCCCGCGAGCG   8   3   8 GTCAGCAGATCCGGTGTAATTCCGGGGCCGACGGTTAA-AGTCCGGAT-GGGAGAGGGTAACG
KP   GCTTATTCTCAGGG-CGGGGCGAAATTCCCCACCGGCGGT  20 AGCCCGCGAGCG   8   4   8 GTCAGCAGATCCGGTGTAATTCCGGGGCCGACGGTTAA-AGTCCGGAT-GGGAGAGAGTAACG
HI   TCGCATTCTCAGGG-CAGGGTGAAATTCCCTACCGGTGGT   2 AGCCCACGAGCG  26   9  30 GTCAGCAGATTTGGTGAAATTCCAAAGCCGACAGT-AA-AGTCTGGAT-GAAAGAGAATAAAA
VK   GCGCATTCTCAGGG-CAGGGTGAAATTCCCTACCGGTGGT  14 AGCCCACGAGCG  11   9  11 GTCAGCAGATTTGGTGAGAATCCAAAGCCGACAGT-AT-AGTCTGGAT-GAAAGAGAATAAGC
VC   CAATATTCTCAGGG-CGGGGCGAAATTCCCCACCGGTGGT  13 AGCCCACGAGCG   5   4   5 GTCAGCAGATCTGGTGAGAAGCCAGGGCCGACGGTTAC-AGTCCGGAT-GAGAGAGAATGACA
YP   GCTTATTCTCAGGG-CGGGGTGAAAGTCCCCACCGGCGGT  40 AGCCCGCGAGCG  16   6  16 GTCAGCAGACCCGGTGTAATTCCGGGGCCGACGGTTAT-AGTCCGGAT-GGGAGAGAGTAACG
AB   GCGCATTCTCAGGG-CAGGGTGAAAGTCCCTACCGGTGGT  25 AGCCCACGAGCG  16   4  27 GTCAGCAGATTTGGTGCGAATCCAAAGCCGACAGTGAC-AGTCTGGAT-GAAAGAGAATAAAA
BP   GTACGTCTTCAGGG-CGGGGTGCAATTCCCCACCGGCGGT  18 AGCCCGCGAGCG  10   4  10 GTCAGCAGACCTGGTGAGATGCCAGGGCCGACGGTCAT-AGTCCGGAT-GAGAGAAGATGTGC
AC   ACATCGCTTCAGGG-CGGGGCGTAATTCCCCACCGGCGGT  16 AGCCCGCGAGCA  10   3  11 ---CGCAGATCTGGTGTAAATCCAGAGCCGACGGT-AT-AGTCCGGAT-GAAAGAAGACGACG
Spu  AACAATTCTCAGGG-CGGGGTGAAACTCCCCACCGGCGGT  34 AGCCCGCGAGCG   6   6   6 GTCAGCAGATCTGGTG 52 TCCAGAGCCGACGGT 31 AGTCCGGAT-GGAAGAGAATGTAA
PP   GTCGGTCTTCAGGG-CGGGGTGTAAGTCCCCACCGGCGGT  13 AGCCCGCGAGCG   7   3   7 GTCAGCAGATCTGGTGCAACTCCAGAGCCGACGGTCAT-AGTCCGGAT-GAAAGAAGGCGTCA
AU   GGTTGTTCTCAGGG-CGGGGTGCAATTCCCCACCGGCGGT  17 AGCCCGCGAGCG   7   9   7 GTCAGCAGATCCGGTGAGAGGCCGGAGCCGACGGT-AT-AGTCCGGAT-GGAAGAGGACAAGG
PU   AAACGTTCTCAGGG-CGGGGTGCAATTCCCCACCGGCGGT  19 AGCCCGCGAGCG  19   4  18 GTCAGCAGACCCGGTGTGATTCCGGGGCCGACGGTCAC-AGTCCGGATGAAGAGAGAACGGGA
PY   TAACGTTCTCAGGG-CGGGGTGCAACTCCCCACCGGCGGT  19 AGCCCGCGAGCG  15   4  16 GTCAGCAGACCCGGTGTGATTCCGGGGCCGACGGTCAT-AGTCCGGATGAAGAGAGAGCGGGA
PA   TAACGTTCTCAGGG-CGGGGTGAAAGTCCCCACCGGCGGT  19 AGCCCGCGAGCG  14   4  13 GTCAGCAGACCCGGTGCGATTCCGGGGCCGACGGTCAT-AGTCCGGATAAAGAGAGAACGGGA
MLO  TAAAGTTCTCAGGG-CGGGGTGAAAGTCCCCACCGGCGGT  16 AGCCCGCGAGCG   8   5   8 GTCAGCAGATCCGGTGTGATTCCGGAGCCGACGGTTAG-AGTCCGGAT-GAAAGAGGACGAAA
SM   AAGCGTTCTCAGGG-CGGGGTGAAATTCCCCACCGGCGGT  34 AGCCCGCGAGCG   8   3   8 GTCAGCAGATCCGGTCGAATTCCGGAGCCGACGGTTAT-AGTCCGGAT-GGAAGAGAGCAAGC
BME  GCTTGTTCTCGGGG-CGGGGTGAAACTCCCCACCGGCGGT  17 AGCCCGCGAGCG  10  15  10 GTCAGCAGATCCGGTGAGATGCCGGAGCCGACGGTTAA-AGTCCGGAT-GGAAGAGAGCGAAT
BS   ATCAATCTTCGGGG-CAGGGTGAAATTCCCTACCGGCGGT  18 AGCCCGCGA---   5   4   5 -----AGGATTCGGTGAGATTCCGGAGCCGACAGT-AC-AGTCTGGAT-GGGAGAAGATGGAG
BQ   GTCTATCTTCGGGG-CAGGGTGAAAATCCCGACCGGCGGT  27 AGCCCGCGA---   3   5   3 -----AGGATTTGGTGTGATTCCAAAGCCGACAGT-AT-AGTCTGGAT-GGGAGAAGATGGAG
BE   ATTCATCTTCGGGG-CAGGGTGAAATTCCCGACCGGCGGT  20 AGCCCGCGA---   3   4   3 -----AGGATCCGGTGCGAGTCCGGGAGCCGACAGT-AT-AGTCTGGAT-GGGAGAAGATGAAG
CA   AATGATCTTCAGGG-CAGGGTGAAATTCCCTACCGGCGGT   2 AGCCCGCGAG--   3   4   3 ----TATGATCCGGTTTGATTCCGGAGCCGACAGT-AA-AGTCTGGAT-GAAAGAAGATATAT
DF   GAAGATCTTCGGGG-CAGGGTGAAATTCCCTACCGGCGGT   2 AGCCCGCG----   6   4   6 -------GATTTGGTGAGATTCCAAAGCCGACAGT-AA-AGTCTGGAT-GAGAGAAGATATTT
EF   GTTCGTCTTCAGGGGCAGGGTGTAATTCCCGACCGGTGGT   3 AGTCCACGAC--   5   3   5 ----ATTGAATTGGTGTAATTCCAATACCGACAGT-AT-AGTCTGGAT--AAAGAAGATAGGG
LLX  AAATATCTTCAGGG-CACCGTGTAATTCGGGACCGGCGGT  21 ACTCCGCGAT--   4   4   4 -----TTGAAGCAGTGAGAATCTGCTAGCGACCGACAGT-AA-AGTCTGGAT-GGAAGAAGATGAAC
LO   GTTCATCTTCGGGG-CAGGGTGCAATTCCCGACCGGTGGT   3 AGTCCACGAT--   3  10   3 ----TTGACTCTGGTGTAATTCCAGGACCGACAGT-AT-AGTCTGGAT-GGGAGAAGATGTTG
PN   AAGAGTCTTCAGGG-CAGGGTGAAATTCCCGACCGGCGGT 125 AGTCCGTG----   3   4   3 -------GATGTGGTGAGATTCCACAACCGACAGT-AT-AGTCTGGAT-GGGAGAAGACGAAA
ST   AAGTGTCTTCAGGG-CAGGGTGTGATTCCCGACCGGCGGT  14 AGTCCGCG----   3   4   3 -------GATGTGGTGTAACTCCACAACCGACAGT-AT-AGTCTGGAT-GAGAGAAGACCGGG
MN   AAGTGTCTTCAGGG-CAGGGTGAGATTCCCGACCGGCGGT 104 AGTCCGCG----   3   4   3 -------GATGTGGTGAAATTCCACAACCGACAGT-AA-AGTCTGGAT-GGGAGAAGACTGAG
SA   ATTCATCTTCGGGG-TCGGGTGTAATTCCCAACCGGCAGT   6 AGCCTGCGAC--  11   3  11 -----CTGATCTAGTGAGATTCTAGAGCCGACAGT-AT-AGTCTGGAT-GGGAGAAGATGGAG
AMI  TCACAGTTTCAGGG-CGGGGTGCAATTCCCCACTGGCGGT  14 AGCCCGCGC---   5   5   5 ------TGATCTGGTGCAAATCCAGAGCCAACGGT-AT-AGTCCGGAT-GGAAGAAACGGAGC
DHA  ACGAACCTTCGAGG-TAGGGTGAAATTCCCGACCGGCGGT  20 AGCCCGCAAC--  11   4  11 --CGACTGACTTGGTGAGACTCCAAGGCCGACGGT-AT-AGTCCGGAT-GGGAGAAGGTACAA
EN   AATAATCTTCGGGG-CAGGGTGAAATTCCCGACCGGTGGT   2 AGTCCACG----   4   6   4 -------GATTTGGTGAAATTCCAAAACCGACAGT-AG-AGTCTGGAT-GAGAGAAGAAAAGA
GLU  ---TGTTCTCAGGG-CGGGGCGAAATTCCCCACCGGCGGT  28 AGCCCGCGAGCG  10   4  10 GTCAGCAGATCCGGTTAAATTCCGGAGCCGACGGTCAT-AGTCCGGAT-GCAAGAGAACC---
```

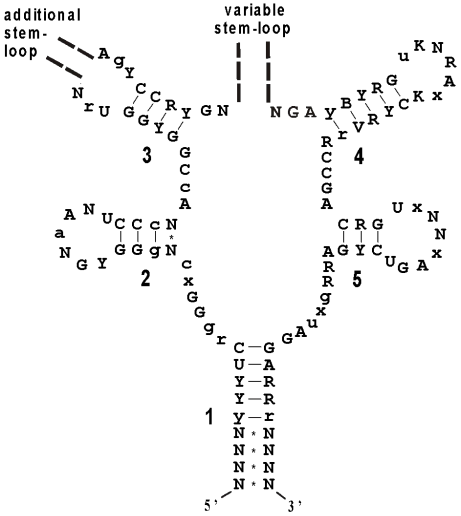
… и еще перед одним геном (*upaA*)



цветные стрелки – гены пути

желтые стрелки – *upaA*, ген с неизвестной функцией

черные стрелки – регуляторный элемент



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* spp.:
  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, Lee et al., 2001; прямой эксперимент, Winkler et al., 2002).

Молекулярная эволюция

Symp Soc Exp Biol. 1958;12:138-63.

On protein synthesis.

CRICK FH.

J Theor Biol. 1965 Mar;8(2):357-66.

Molecules as documents of evolutionary history.

Zuckerkandl E, Pauling L.

После расхождения видов изменения в генах накапливаются почти случайно
=> гены в близких видах более похожи, чем в далеких
=> степень различия генов можно использовать для датировки расхождений

Молекулярная палеонтология: как видели динозавры

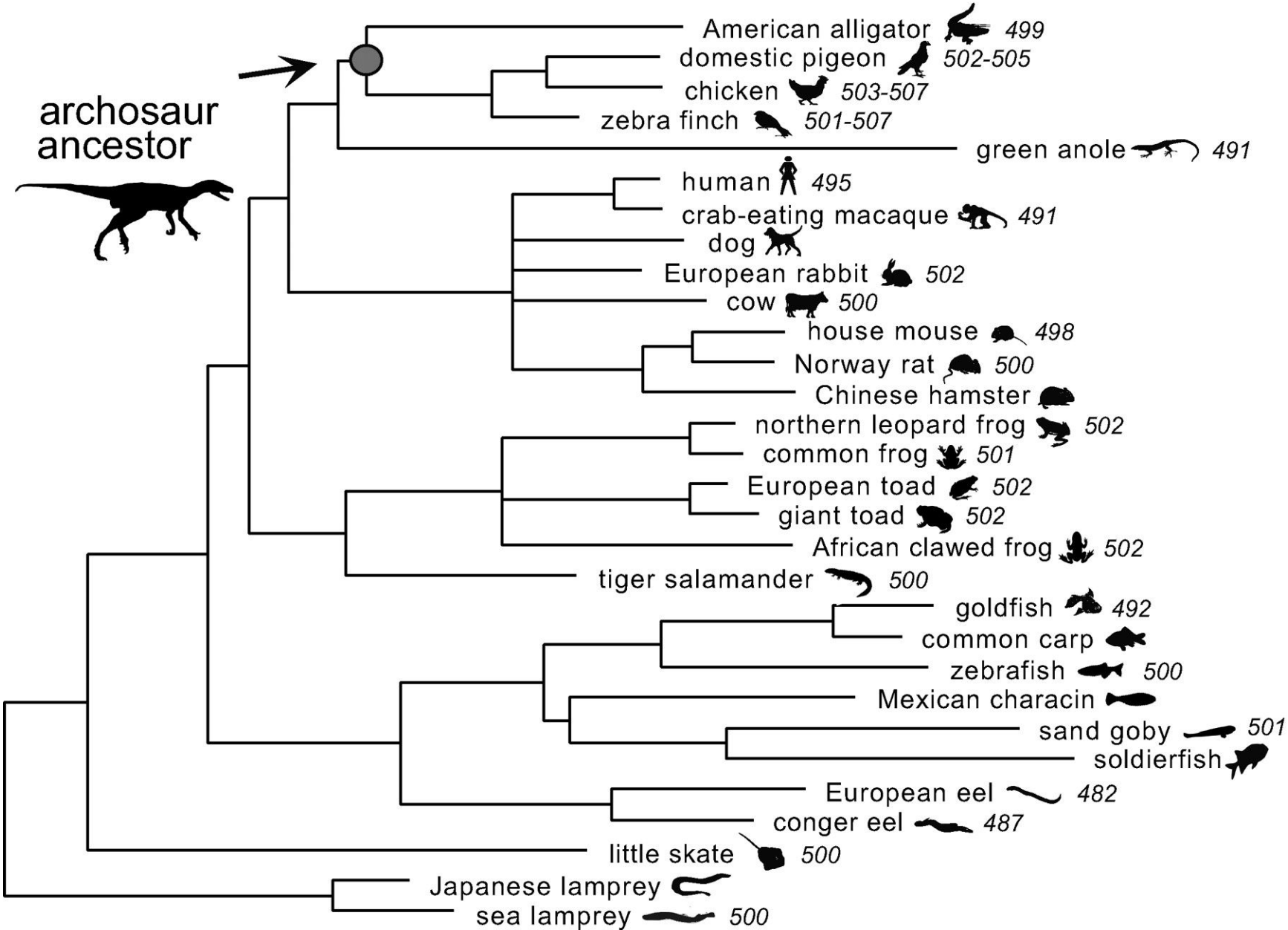
Recreating a Functional Ancestral Archosaur Visual Pigment

Belinda S. W. Chang, Karolina Jönsson,* Manija A. Kazmi,* Michael J. Donoghue,† and Thomas P. Sakmar**

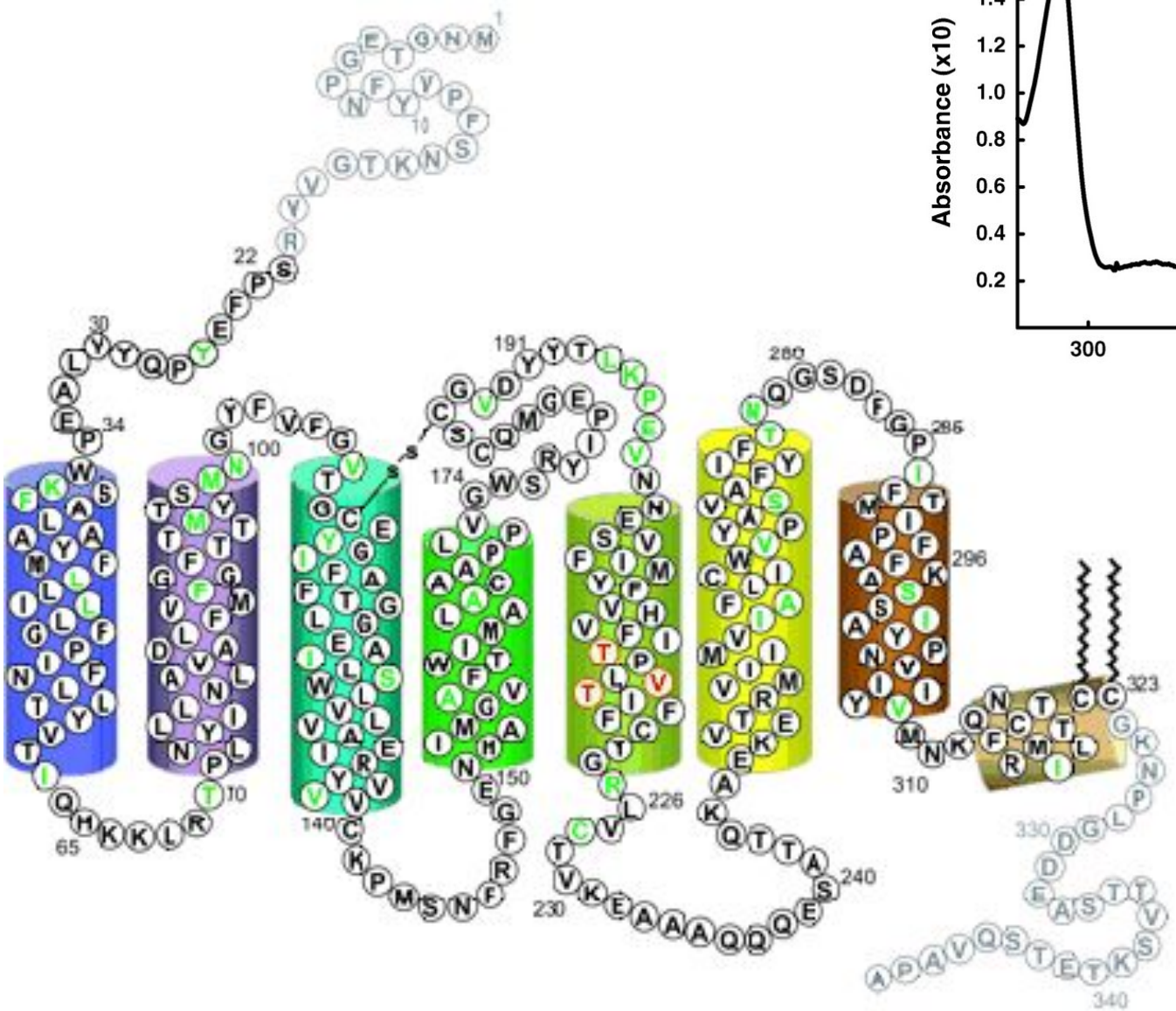
*Laboratory of Molecular Biology & Biochemistry, The Rockefeller University, New York; and †Department of Ecology & Evolutionary Biology, Yale University

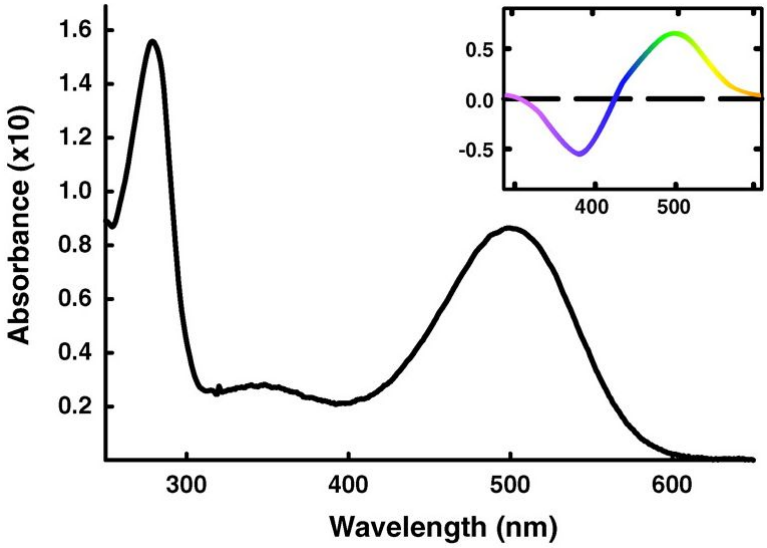
Mol. Biol. Evol. 19(9):1483–1489. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Таксономическое дерево



Реконструированный белок и его свойства



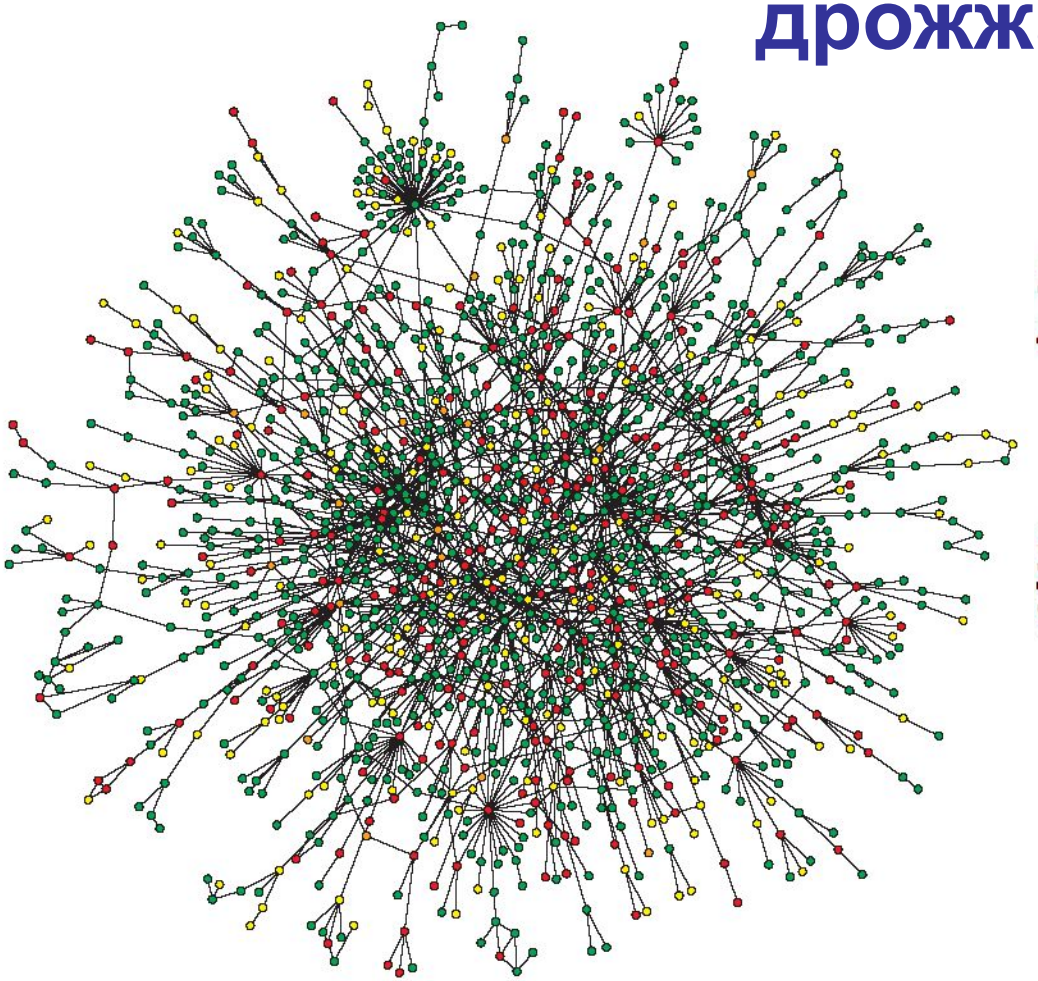


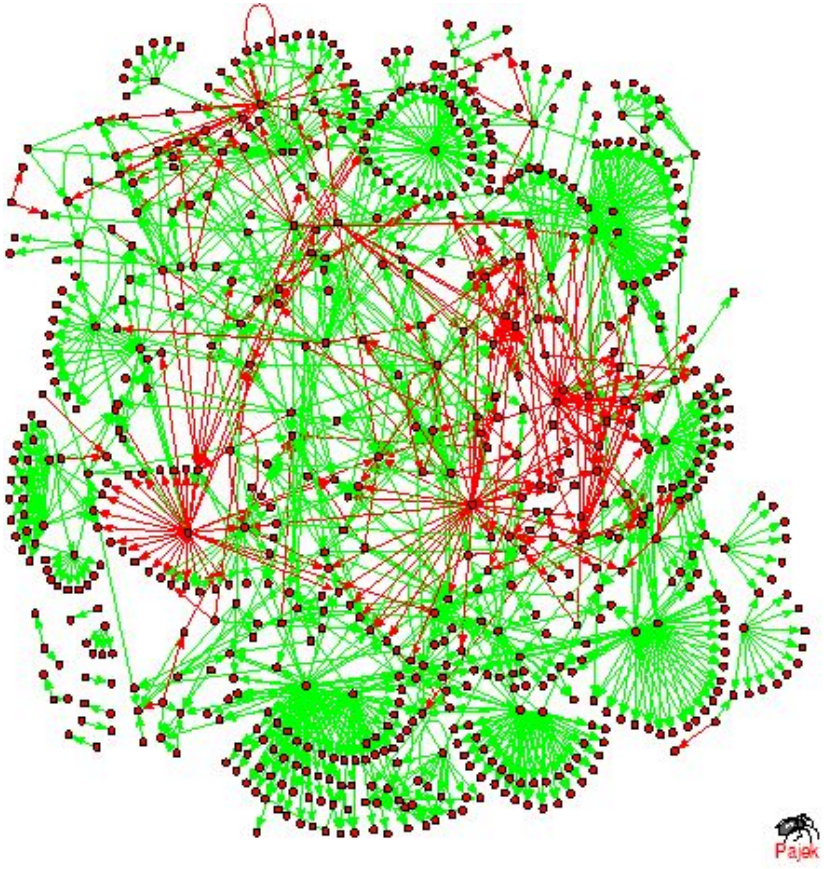
Не только тексты

Можно использовать данные, которые порождаются другими типами массовых экспериментов

- Уровень экспрессии:
  - Концентрации мРНК
  - Концентрации белков
  - Время жизни мРНК и белков
- Взаимодействия:
  - Белок-ДНКовые
  - Белок-белковые
- Структура генома
  - Метилирование ДНК
  - Положение и модификация нуклеосом
- Функционально-генетические
  - Летальность мутаций
  - Фенотип
  - Синтетические летали

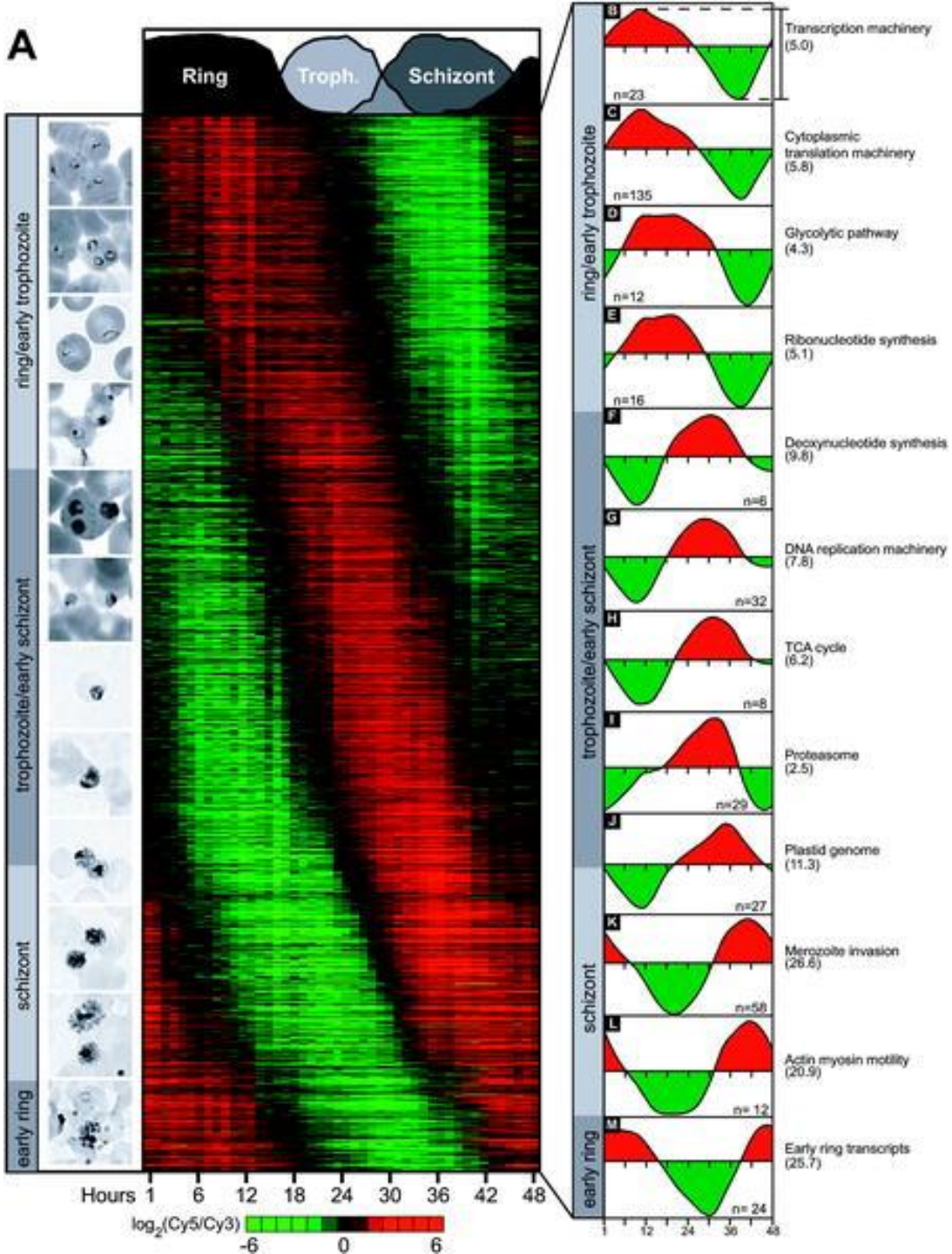
Графы белок-белковых (структурных, сигнальных и др.) и белок-ДНКовых (регуляторных) взаимодействий в дрожжах





Экспрессия (уровень работы) генов

Цикл развития малярийного плазмодия



The Transcriptome
of the Intraerythrocytic Developmental Cycle
of *Plasmodium falciparum*

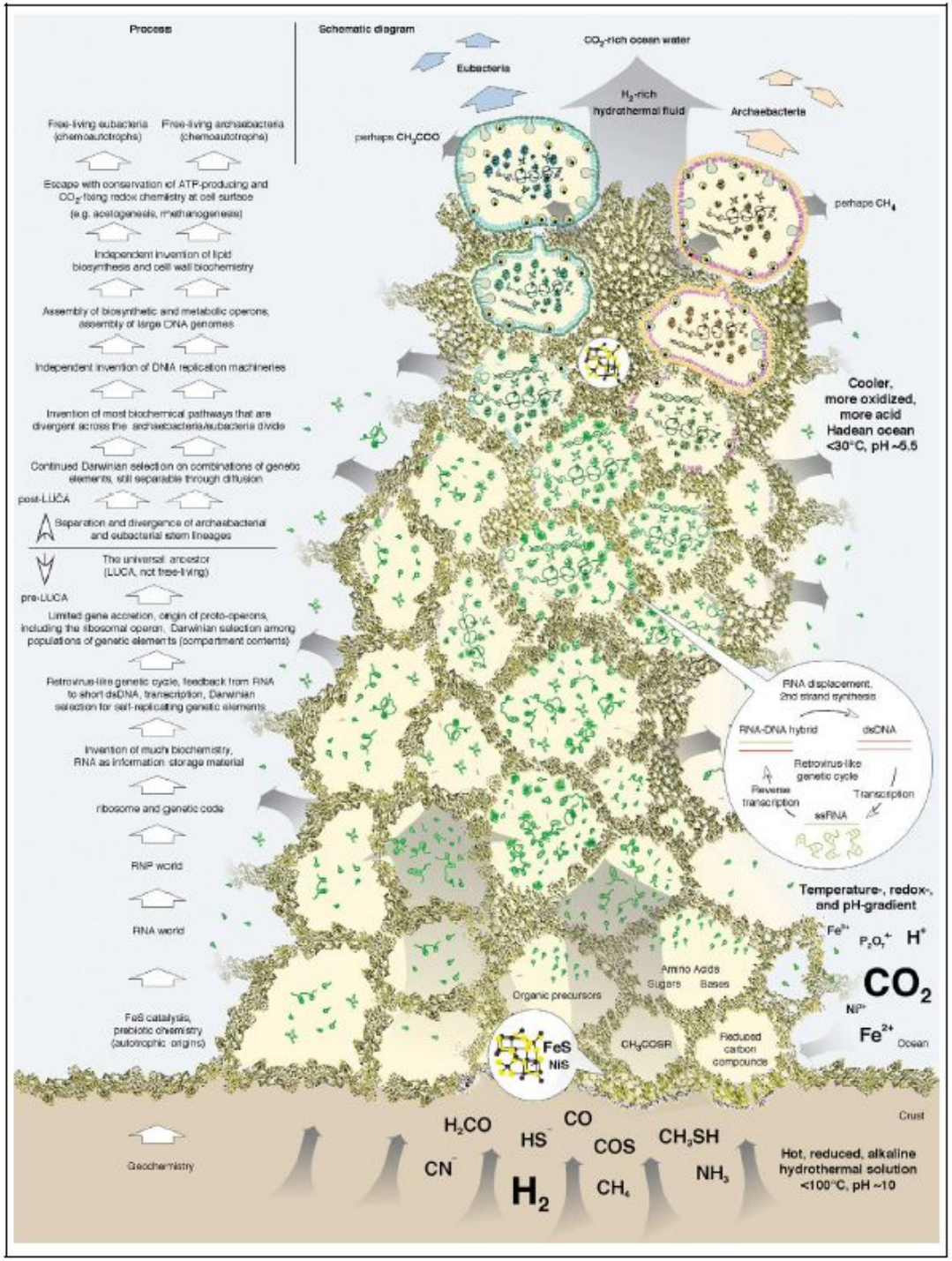
Zbynek Bozdech[1], Manuel Llinás[1], Brian Lee Pulliam[1], Edith D. Wong[1], Jingchun Zhu[2], Joseph L. DeRisi[1]

Биоинформатика

- Биоинформатика – это биология *in silico*
- Кто этим занимается:
  - биологи, математики, физики, химики
  - … биоинформатики
- Хорошая биоинформатика связана с биологией
  - обработка результатов
  - проверка предсказаний
  - в хороших современных проектах эти шаги повторяются несколько раз, начиная с этапа планирования

Происхождение жизни

On the origin of genomes and cells within inorganic compartments

Eugene V. Koonin[a], and William Martin[b]

Trends in Genetics
Volume 21, Issue 12, December 2005, Pages 647-654