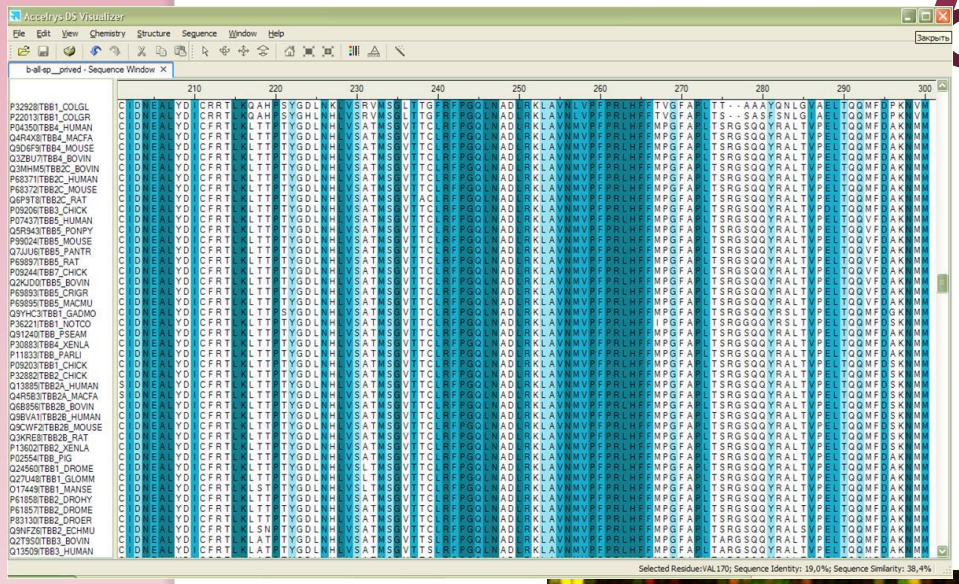
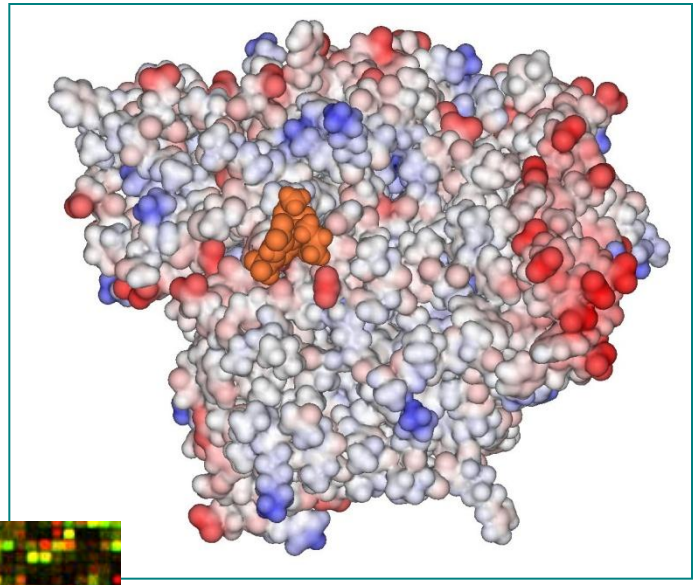
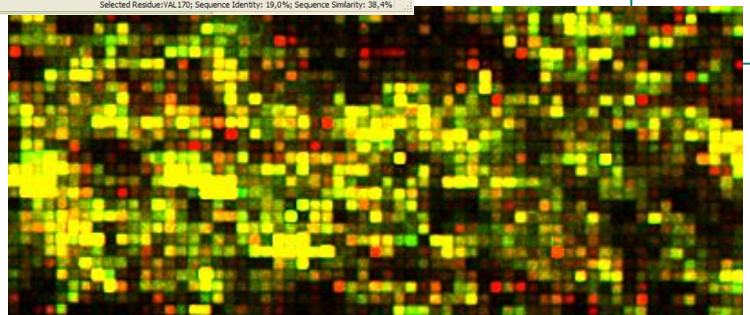
Київський національний університет ім. Тараса Шевченка
Інститут високих технологій

БІОІНФОРМАТИК

к.б.н. Нилорко О.А.







Лавиноподібне накопичення даних молекулярної та структурної біології, яке відбувається протягом останніх 20 років, кардинальним чином змінило характер біологічних (насамперед, молекулярно-біологічних) досліджень, спричинивши розвиток нових комплексних дисциплін, що дістали сукупну назву **«омік»-технологій** (геноміка, транскриптоміка, метаболоміка, протеоміка, феноміка та ін.).

«Омік»-технології дозволяють генерувати та оперувати даними в надзвичайно широкому діапазоні, починаючи з досліджень цілих геномів з послідуючим аналізом експресії генів за допомогою мікроматриць, мас-спектрометрією білків та метаболітів та закінчуючи візуалізацією біологічних процесів та розробкою конкретних заходів по охороні здоров'я

Вузьке місце в біологічних науках зсунулося з отримання первинних результатів до їх зберігання, препроцесінгу, аналізу та інтерпретації.

Поточним викликом є видалення цього вузького місця шляхом комбінації наук про життя з інформаційними технологіями.

накопичення великої кількості біологічних даних стимульовало розвиток особливої наукової дисципліни, що дозволяє інтегрувати і обробляти їх - біоінформатики

множинність визначення біоінформатики

- вся сукупність і методів обчислювальної біології (синоніми – обчислювальна біологія, інформаційна біологія)
- сукупність програм та методів розробки баз даних для зберігання і маніпулювання геномною інформацією
- методи і програми аналізу послідовностей макромолекул

приклад розгорнутого визначення (за Altman, 1998)

- **Біоінформатика досліджує два інформаційних потоки в молекулярній біології:**

1. **передачу інформації на будь-якій стадії центральної догми, включаючи організацію і контроль генів в ДНК-послідовностях, ідентифікацію одиниць транскрипції, передбачення структури білків за їх послідовністю, аналіз молекулярних функцій**
2. **передачу інформації в межах експериментальної процедури, включаючи системи генерації гіпотез,**

власне визначення ;)

- **біоінформатика – спроба інтерпретації біологічних “текстів”, прикладом яких є послідовності макромолекул в живих системах**

- **Біоінформатика – наука про закономірності зберігання, передачі і реалізації інформації на молекулярному, субклітинному та клітинному рівні організації живого**

будь-які визначення біоінформатики як правило охоплюють застосування комп’ютерних наближень на рівні не вище клітинного

основні розділи біоінформатики

області інтересу комп'ютерних фахівців в біології

1. біоінформатика **послідовностей** – класична біоінформатика

2. **структурна** біоінформатика – обчислювальна структурна біологія

3. **комп'ютерна** геноміка

біоінформатика послідовностей – класична біоінформатика

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

Статистика надходжень нуклеотидних послідовностей в GenBank

http://www.ncbi.nlm.nih.gov/genbank/

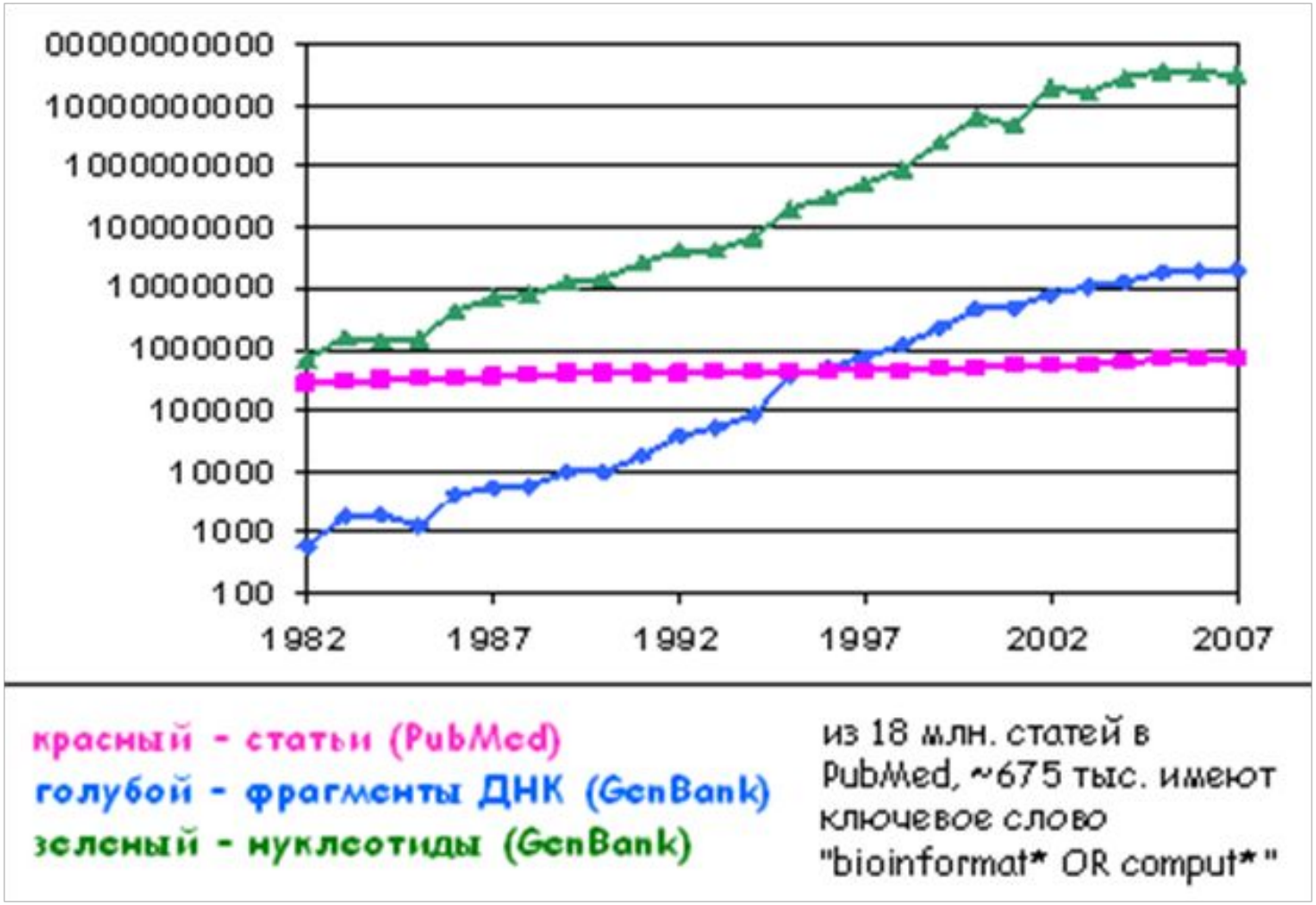
На момент свого заснування в 1982 році містив 606 послідовностей, які складалися з 680 338 літер.

Через 10 років кількість послідовностей збільшилася до 78 608 (101 008 486 літер),

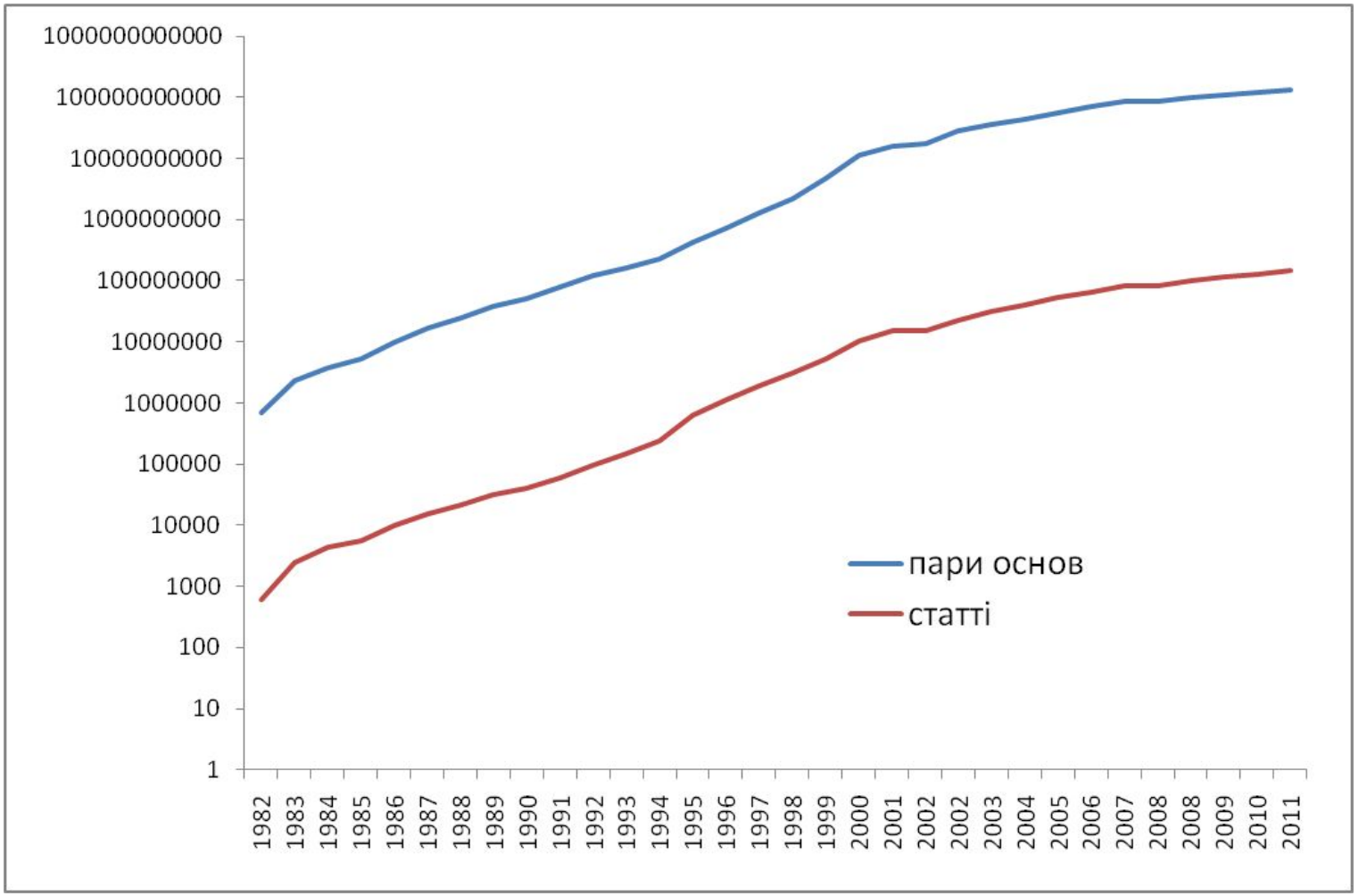
Через 20 років – до 22 318 883 (28 507 990 166 літер).

На кінець 2011 GenBank містив 135 117 731 375 літер в 129 902 276 послідовностях при загальному розмірі файлів 468 Гб.

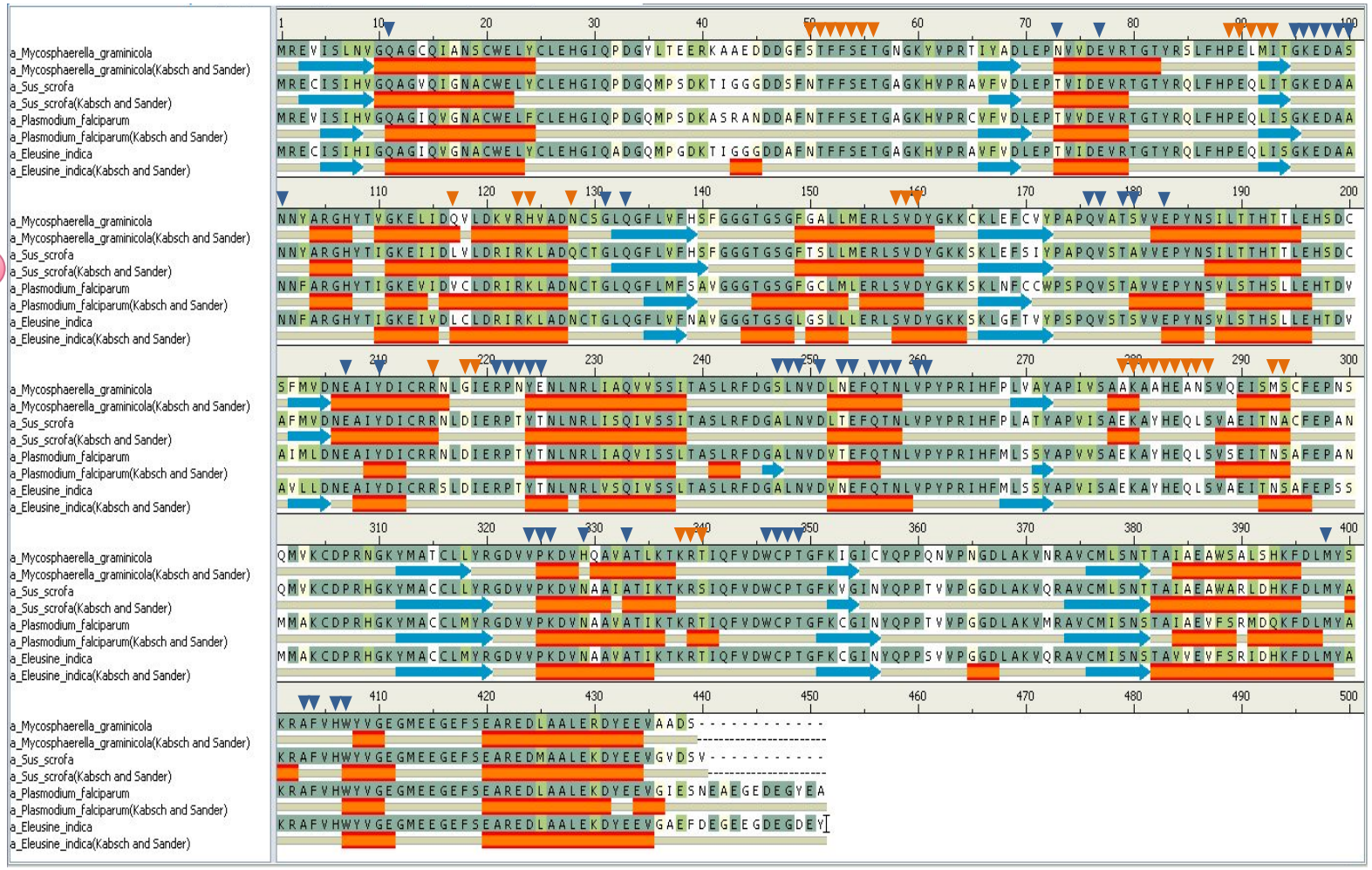
Статистика надходжень нуклеотидних послідовностей в GenBank (2007)



Статистика надходжень нуклеотидних послідовностей в GenBank (2012)



детальна статистика доступна за адресою
ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt



Дані щодо послідовностей – розвиток алгоритмів для парного та множинного вирівнювання послідовностей, визначення та дослідження мотивів, використання імовірнісних моделей для пошук генів, вирівнювання послідовностей,

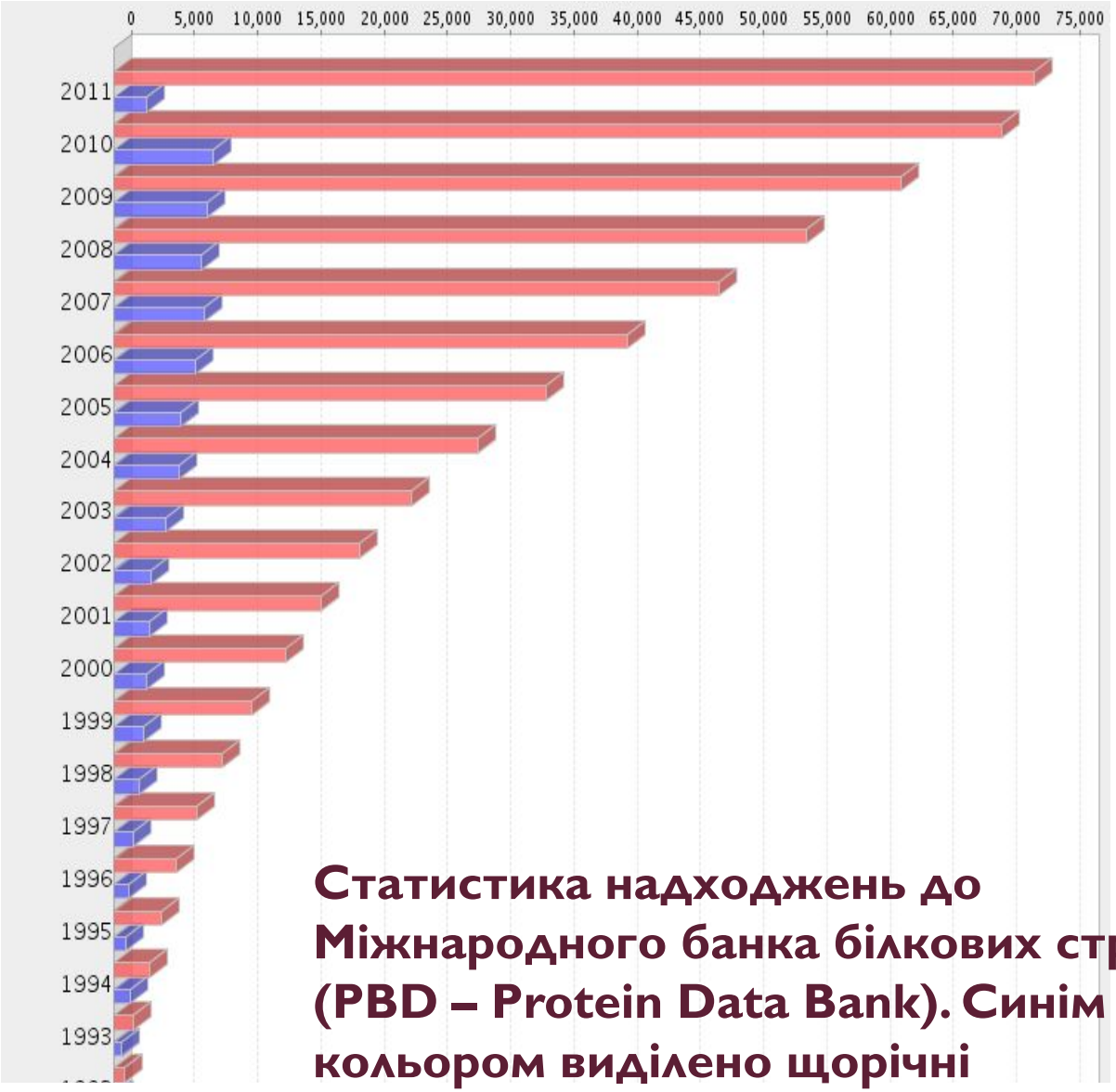
точки застосування класичної біоінформатики

- Вирівнювання й визначення подібності двох послідовностей
- Побудова множинних вирівнювань
- Розпізнавання генів
- Передбачення сайтів зв'язування регуляторних білків
- Передбачення вторинної структури РНК
- Молекулярна філогенія

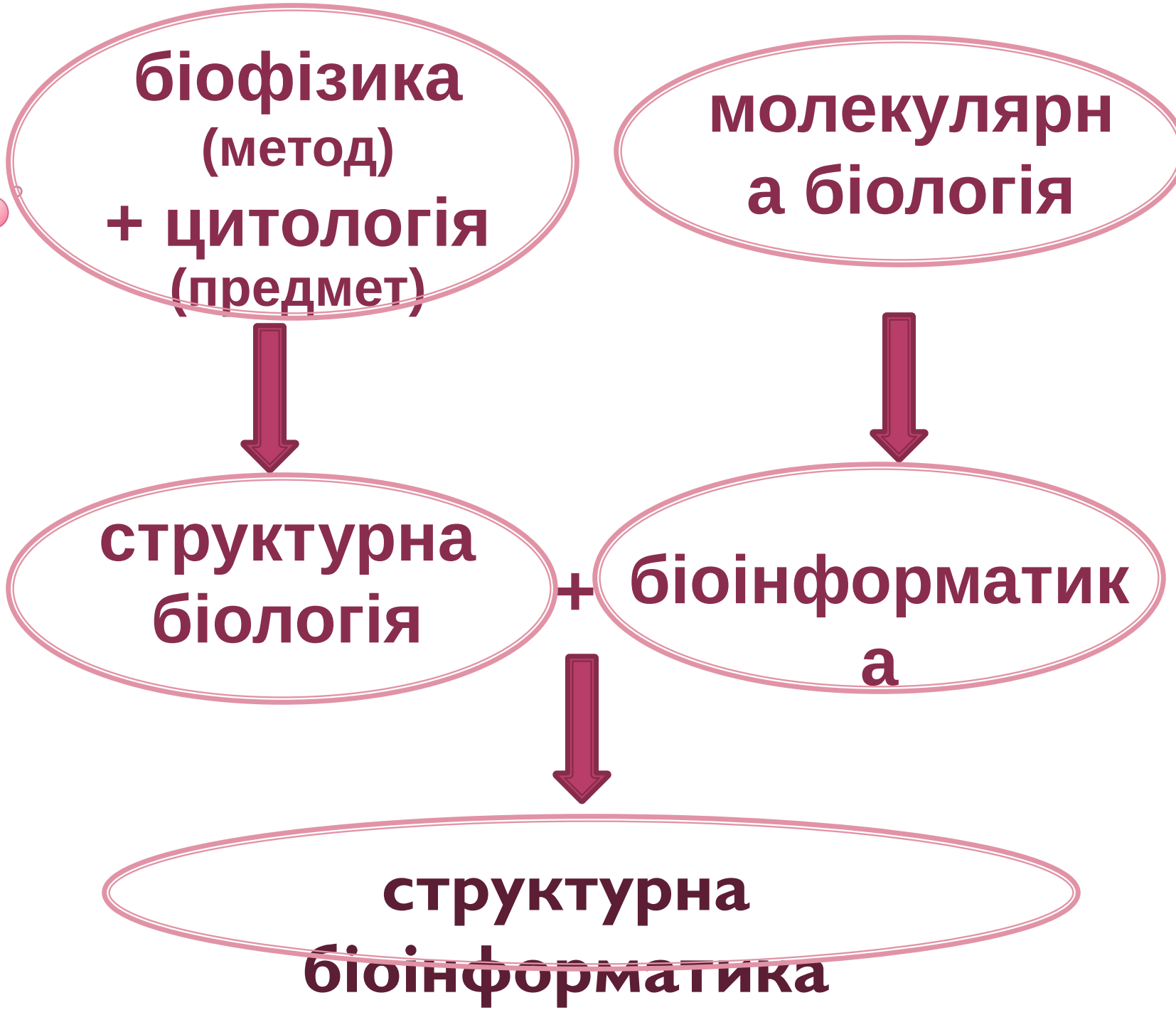
структурна біоінформатика – обчислювальна структурна біологія

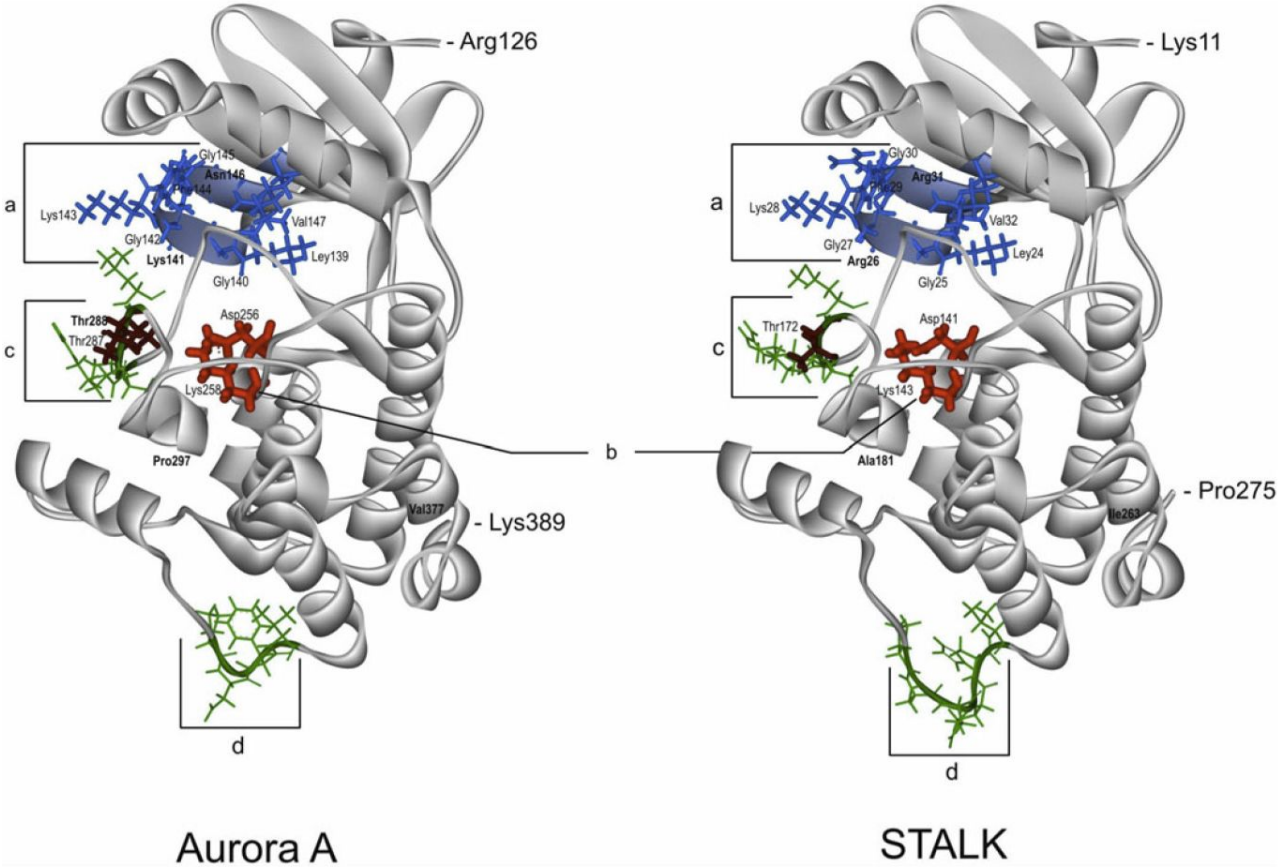
структурна біоінформатика

- з точки зору біоінформатики – підрозділ біоінформатики, що фокусується на представленні, зберіганні, запиті, аналізі та відтворенні структурної інформації в атомному та субклітинному просторовому масштабі
- з точки зору структурної біології – обчислювальний апарат, що застосовується для визначення, представлення, зберігання, запиту, аналізу та відтворення просторової структури макромолекул та субклітинних утворень



Статистика надходжень до Міжнародного банка білкових структур (PBD – Protein Data Bank). Синім кольором виділено щорічні надходження, червоним – загальна кількість статей в банку


біофізика
(метод)
+ цитологія
(предмет)
молекулярна біологія
структурна біологія
+
біоінформатика
структурна біоінформатика



Структурні дані – розвиток обчислювальної геометрії , комп'ютерної графіки, алгоритмів для аналізу кристалографічних даних та даних ЯМР і наступної розробки правдоподібних моделей макромолекул.

Молекулярна графіка – одне з перших

структурна біоінформатика

Більш глибоке розуміння, як біологічна функція обумовлена просторовою структурою.

Чи можна передбачити просторову структуру, базуючись виключно на інформації про послідовність?

задачі структурної біоінформатики

- класифікація білків за особливостями просторової структури, аналіз та/або передбачення активних сайтів
- оцінка якості тривимірних структур;
- дослідження кореляції різних типів структурної інформації, зберігання структур в базах даних, інтеграція структурних даних з даними інших джерел
- створення інфраструктури для побудови структурних моделей з окремих компонентів (передбачення структури модульних білків, реконструкція різних ділянок білка за різними матрицями)
- дизайн білків з новими функціональними властивостями та розуміння принципів їх згортки (фолдінгу)
- принципи дизайну біологічно-активних нових сполук на основі структурних особливостей їх мішеней
- розробка нових моделей відтворення поведінки макромолекул для поглибленого розуміння їх функцій

точки застосування структурної біоінформатики

- вибір білків-мішеней
- трекінг умов кристалізації
- аналіз кристалографічних даних
- аналіз даних ЯМР
- анотування і оцінка тривимірних структур
- зберігання структур в базах даних
- дослідження кореляції різних типів структурної інформації
- візуалізація даних
- класифікація білкових структур

труднощі структурно-біоінформатичних обчислень

- структурні дані є нелінійними, взаємодії між атомами також нелінійні – необхідність використання складних алгоритмів
- структурний простір, в якому ведуться обчислення, є мінливим
- фундаментальний зв’язок між молекулярною структурою та фізикою – спроби спростити модель приводять до ускладнень в розумінні процесів взаємодії

труднощі структурно-біоінформатичних обчислень

- візуалізація даних – одночасно превага і недолік: вона спрямована на людину і неефективно розуміється комп'ютером
- структурні дані гнучкі, динамічні і містять достатньо велику кількіть шуму
- просторова структура консервативніша за послідовність – проблема переходу від однієї структури до іншої
- недостатня кількість інформації щодо мембранних та фібрілярних білків
- нестаток інформації щодо асоціації білкових доменів

комп’ютерна геноміка

•**Обчислювальна геноміка** фокусується (як цілком зрозуміло з назви) на розмітці та порівняльному аналізі організації геномів різного походження, а також досліджує взаємодію різних компонентів геному.

•Згідно даних Національного центру біотехнологічної інформації США (NCBI), на сьогодні виконується секвенування (визначення послідовності) 4742 геномів бактерій, 91 геному архей та 1215 – по еукаріотах, крім того, повністю завершено 104 геномні проекти по археях, 1439 – по бактеріях та 39 – по еукаріотах.

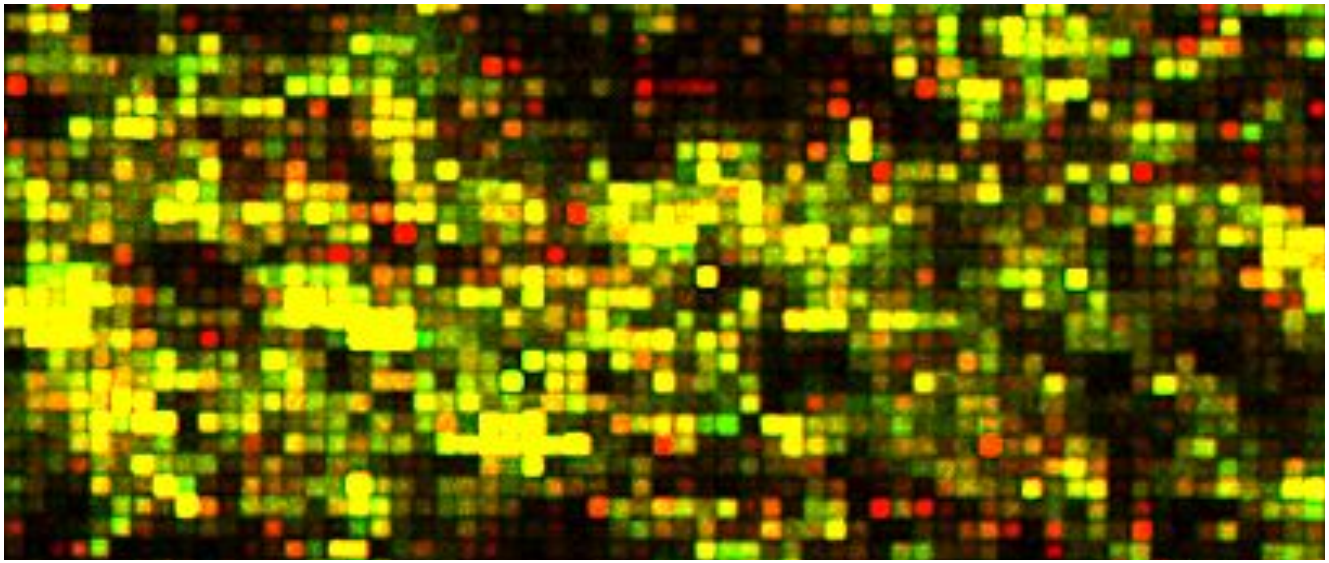
точки застосування комп’ютерної геноміки

- Передбачення генів у послідовностях. При цьому в деяких випадках вдається навіть знайти помилки в послідовності.
- Попередня анотація по подібності й іншим особливостям білкових послідовностей.
- Порівняльний аналіз геномів.
- Дослідження регуляції роботи генів.
- Пошук пропущених генів.
- Метаболічна реконструкція

•Задача метапболічної реконструкції є спільною як для обчислювальної геноміки, так і для **системної біології** – науки, що досліджує шляхи та мережі взаємодії між різними компонентами біологічних систем на різних рівнях їх організації. Реконструкція повної мережі метаболічних шляхів клітини в ряді також розглядається як предмет окремої дисципліни – метаболоміки

структурна біоінформатика – основа структурної геноміки

Структурна геноміка – високопропускне визначення просторової структури макромолекул (в першу чергу, білків!) в масштабі цілого геному.



Нова область інтересу – аналіз даних експресії. Необхідність обробки вельми зашумлених даних спричинила розвиток відповідних алгоритмів статистичного аналізу та машинного навчання, зокрема в методах угруповань та класифікаційних техніках.


GenBank Overview - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/Genbank/
noc molecular
Файл Правка Вид Избранное Сервис Справка
Избранное
GenBank Overview
Gmail - Входящие (3...
Страница Безопасность Сервис
NCBI
GenBank Overview
PubMed Entrez BLAST OMIM Books Taxonomy Structure
Search Entrez for Go
NCBI Home
NCBI Site Map
Submit to GenBank
Submit an update
Search GenBank
GenBank and RefSeq: a comparison
BLAST
What is GenBank?
GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan;36(Database issue):D25-30). There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.
The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.
An example of a GenBank record may be viewed for a Saccharomyces cerevisiae gene.
In The News: 2009 H1N1 Flu Virus (Swine Flu)
The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the NCBI Flu Resource
NLM/NCBI 2009 H1N1 Flu Resources:
Newest 2009 H1N1 influenza A sequences
Citations recently added to PubMed
MedlinePlus (consumer health information)
Enviro-Health Links
FLU.GOV
Know what to do about the flu.
VISIT FLU.GOV
SHARE THIS WIDGET
Submissions to GenBank
Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper. There are several options for submitting data to GenBank:
BankIt, a WWW-based submission tool for convenient and quick submission of sequence data
Sequin, NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is available by
Интернет
135%


UniProt - Windows Internet Explorer
http://www.uniprot.org/
noc molecular
Файл Правка Вид Избранное Сервис Справка
Избранное
UniProt
Gmail - Входящие (3...
Страница Безопасность Сервис
UniProt
Downloads · Contact · Documentation/Help
Search
Blast
Align
Retrieve
ID Mapping
Search in
Protein Knowledgebase (UniProtKB)
Query
Search
Clear
Fields »
WELCOME
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
What we provide
UniProtKB
Protein knowledgebase, consists of two sections:
Swiss-Prot, which is manually annotated and reviewed.
TrEMBL, which is automatically annotated and is not reviewed.
Includes Complete Proteome Sets.
UniRef
Sequence clusters, used to speed up similarity searches.
UniParc
Sequence archive, used to keep track of sequences and their identifiers.
Supporting data
Literature citations, taxonomy, keywords and more.
UniProt
NEWS
UniProt release 15.15 – Mar 2, 2010
Bacillus subtilis, a Gram-positive model bacterium fully annotated in UniProtKB/Swiss-Prot · Cross-references to EuPathDB, ProtClustDB and SUPFAM · Change to cross-references to HOVERGEN
› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives
SITE TOUR
Learn how to make best use of the tools and data on this site.
PROTEIN SPOTLIGHT
love at first smell
March 2010
The making of life is demanding. Take any form from fungus to bacteria, and plants to humans the creation of progeny does not just happen. It takes a lot of molecular dialogue to divide E…
© 2002–2010 UniProt Consortium | License & Disclaimer | Contact
EMBL-EBI
PIR
SIB
Интернет
135%

RCSB Protein Data Bank - Windows Internet Explorer

http://www.rcsb.org/pdb/home/home.do

noc molecular

Файл Правка Вид Избранное Сервис Справка

Избранное | RCSB Protein Dat... | Gmail - Входящие (3...

Страница Безопасность Сервис

MyPDB Login

A MEMBER OF THE PDB

RCSB PDB PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures

As of **Tuesday Mar 23, 2010 at 5 PM PDT** there are 64229 Structures | **PDB Statistics**

HELP | PRINT | PDB ID or Text | Search | Advanced Search

Home Hide
- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- New Website Features

Deposition Hide
- All Deposit Services
- Electron Microscopy
- X-ray | NMR
- Validation Server
- BioSync Beamline
- Related Tools

Search Hide
- Advanced Search
- Latest Release
- Latest Publications
- Sequence Search
- Chemical Components
- Unreleased Entries
- Browse Database
- Histograms

Tools Hide
- File Downloads
- FTP Services
- File Formats
- Services: RESTful | SOAP
- Widgets
- Compare Structures

Education Hide
- Understanding PDB Data
- Molecule of the Month
- Educational Resources

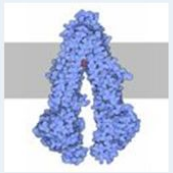
A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the **wwPDB**, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Hide Welcome Message

Featured Molecules (Previous Features: MOM | PSI) Hide

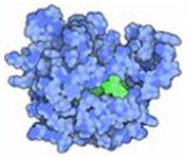


Molecule of the Month:

P-glycoprotein

Our environment is filled with toxic substances that attack our molecular machinery. Our cells protect themselves from these dangers in many ways. In some cases, they use enzymes to convert them into harmless compounds. In other cases, they sequester them safely out of the way. For others, cells build specialized pumps that find toxins and eject them outside, for safe disposal.

Full Article...



Protein Structure Initiative Featured Molecule:

Phytochrome

In two new NMR structures of an unusually small phytochrome, researchers have revealed how plants see light and shade. The structures reveal for the first time the complex motion of the chromophore after absorbing red light.

Full Article...

New user? Try the browser **compatibility check** and information on **Getting Started** or the **narrated tutorial**

Customize This Page

New Features Hide

Bidirectional Sorting

Read more about the releases:

Website Release Archive:

News Hide

Weekly | Quarterly | Yearly

Statement on Retraction of PDB Entries

2010-03-23

West Windsor Plainsboro High School North Wins New Jersey Science Olympiad Protein Modeling State Finals

Read More

San Diego Science Festival Expo Day: March 27

Read More

Previous Weekly News:

- Exhibiting at NSTA
- Improved Ligand Searching
- New Website Features
- Online Narrated Tutorial Demonstrates How to Use the RCSB PDB

Ошибка на странице. | Интернет | 135%

Основний спосіб визначити схожість двох послідовностей - вирівняти їх

```
>EC_Tr  :  MQNRLTI KDI ARLSGVGKSTVSRVLNNEYR
>EC_Fr  :  MKLDEI ARLAGVSRTTASYVI NGKAKQYR
```

- При аналізі первинних структур процедура вирівнювання виявляє сходство між послідовностями (**sequence similarity**), яке може свідчити про гомологію **(homology)**, тобто еволюційну спорідненість макромолекул.

Геп – пропуск в послідовності

```
>EC_Tr  :  MQNRLTIKDIARLSGVGKSTVSRVLNNE---YR
>EC_Fr  :  ----MKLDEIARLAGVSRTTASYVINGKAKQYR
```

Гомологичные последовательности – последовательности, имеющие общее происхождение (общего предка).

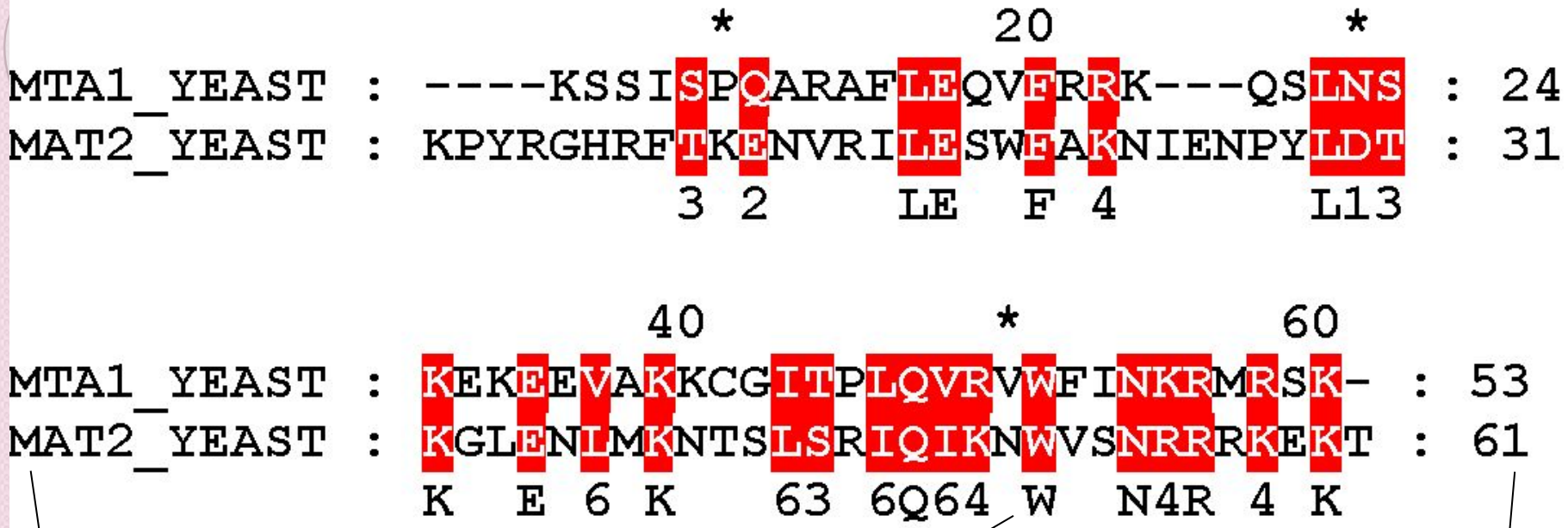
Признаки гомологичности белков

сходная 3D-структура

в той или иной степени похожая аминокислотная последовательность

- **разные другие соображения…**

Что изображено?



Номер столбца выравнивания

Название последовательности

Консервативный остаток

Функционально консервативная позиция

Номер последнего в строке остатка ИЗ ЭТОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

```
      домовой
скупидом
    водомерка ?
```

Гэп – пропуск в последовательности

```
лесовоз      ?--лесо---воз
ледоход       лед---оход---
```

Попарное выравнивание:

```
                                    *                   20
XYLR_ECOLI  :  GYPSLQYFYSVFKKAYDTTPKEYR  :  24
XYLR_HAEIN  :  GYPSIQYFYSVFKKEFEMTPKEFR  :  24
```

Множественное выравнивание:

```
                                *                   20
APPY_ECOLI  :  GYNSTSYFICAFKDYYGVTPSHYF  :  24
CELD_ECOLI  :  GYSSPSLFIKTFKKLTSFTPKSYR  :  24
CFAD_ECOLI  :  GISSASYFIRVFNKHYGVTPKQFF  :  24
ENVY_ECOLI  :  GYSSTSYFISVFKAFYGLTPLNYL  :  24
FAPR_ECOLI  :  GYTSVSYFIKTFKEYYGVTPKKFE  :  24
MELR_ECOLI  :  GFRSSSRFYSTFGKYVGMSPQQYR  :  24
RHAS_ECOLI  :  GFSDSNHFSTLFRREFNWSPRDIR  :  24
ROB_ECOLI   :  RFDSQQTFTRAFKKQFAQTPALYR  :  24
TETD_ECOLI  :  QFDSQQSFTRRFKYIFKVTPSYYR  :  24
XYLR_ECOLI  :  GYPSLQYFYSVFKKAYDTTPKEYR  :  24
XYLR_HAEIN  :  GYPSIQYFYSVFKKEFEMTPKEFR  :  24
               g  s   F   Fk     tP
```

Паттерн – регулярное выражение UNIX'а:

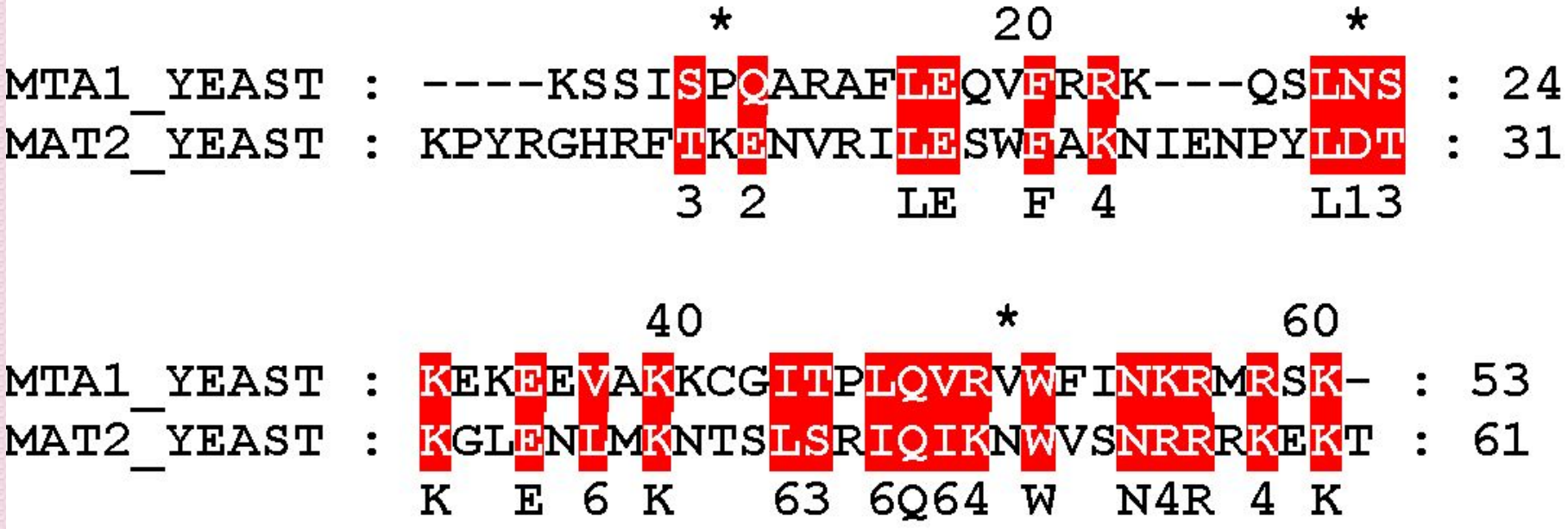
[AC]-x-V-x(4)-{ED}

Ala или Cys- x-Val- x- x- x - x- (любой, но не Glu и не Asp)

F	K	L	L	S	H	C	L	L	V
F	K	A	F	G	Q	T	M	F	Q
Y	P	I	V	G	Q	E	L	L	G
F	P	V	V	K					
F	K	V	L	A					
L	E	F	I	S					
F	K	L	L	G					

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Правильно ли выровнены последовательности?



В чем *биологический* смысл выравнивания?

- *Буквы в одной колонке определяют сопоставление аминокислотных остатков двух белков*
- Сопоставленные остатки, по идее, должны иметь что-то общее в молекулах белка; что???

Предложение: биологический смысл имеет сопоставление одинаковых или функционально сходных остатков белка.
Эти остатки играют сходную роль.
Сопоставление непохожих остатков не имеет смысла.

Какое выравнивание “правильнее”?

```
                          *          20              *
MTA1_YEAST  :  ----KSSISPQARAFLEQVFRRK---QSLNS  : 24
MAT2_YEAST  :  KPYRGHRFTKENVRILESWFAKNIENPYLDT  : 31
                        3 2     LE   F 4       L13

                       40             *          60
MTA1_YEAST  :  KEKEEVAKKCGITPLQVRVWFINKRMRSK-  : 53
MAT2_YEAST  :  KGLENLMKNTSLSRIQIKNWVSNRRRKEKT  : 61
               K   E 6 K    63  6Q64 W  N4R 4 K
```

12 консервативных остатков

```
                     *         20         *          4
MTA1_YEAST  :  K----SSISPQA-R------A------F-----LEQVFR  : 17
MAT2_YEAST  :  KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLENLMK  : 39
               K       3 2  R      A      5     LE 6 4

               0          *         60         *
MTA1_YEAST  :  RKQSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK-  : 53
MAT2_YEAST  :  NT-SL-SR------------IQIKNWVSNRRRKEKT  : 61
                  SL S4             6Q64 W  N4R 4 K
```

13 “консервативных” остатков

Чтобы понять смысл выравнивания, вернемся к тому, что такое последовательность аминокислотных остатков и что такое белок

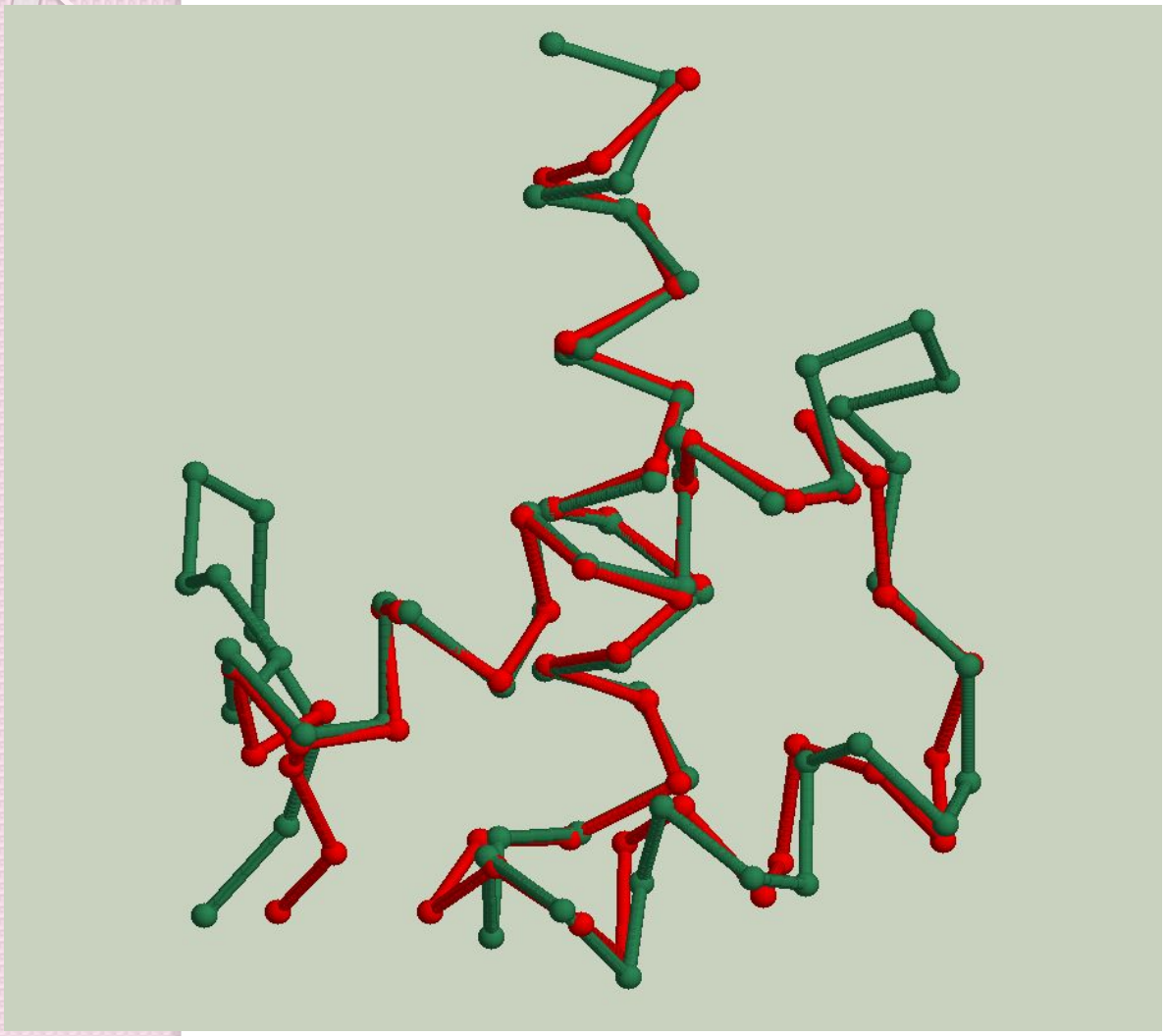
(i)Последовательность – удобный способ закодировать структурную (химическую) формулу молекулы белка (до посттрансляционных модификаций)

(ii) Белок – это большая молекула, сохраняющая в живой клетке постоянную пространственную структуру, т.е.- взаимное расположение ковалентно связанных атомов (конформацию)

(**iii) Последовательность однозначно определяет** в какую пространственную **структуру** свернется белок в клетке

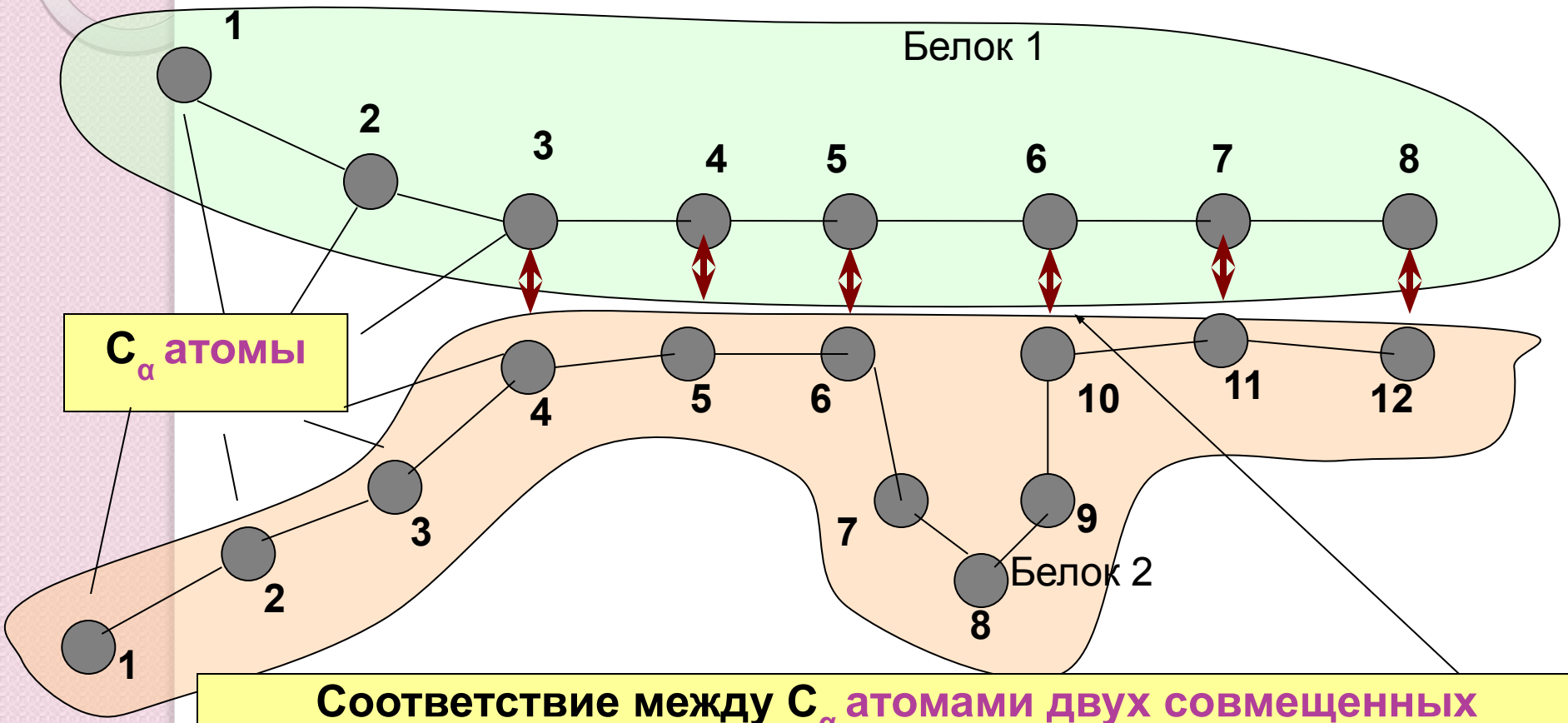
(iv) Функция белка в клетке **проявляется** только при сохранении **уникальной пространственной структуры**

Пространственное совмещение полипептидных цепей белков mta1_yeast и mat2_yeast



На плоской картинке видно плохо ☹

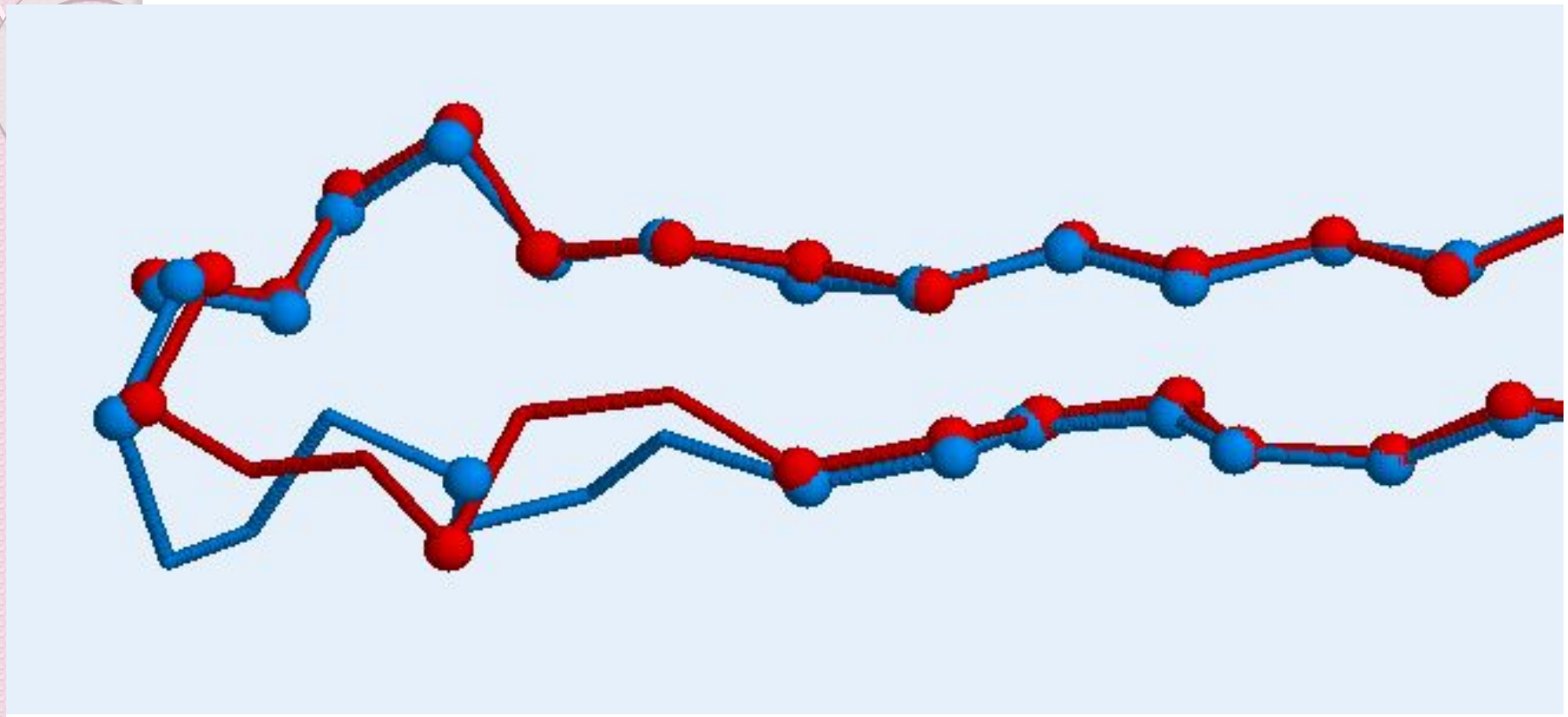
Схематическое изображение совмещенных структур



Соответствие между C_α атомами двух совмещенных структур,
основанное на близости в пространстве

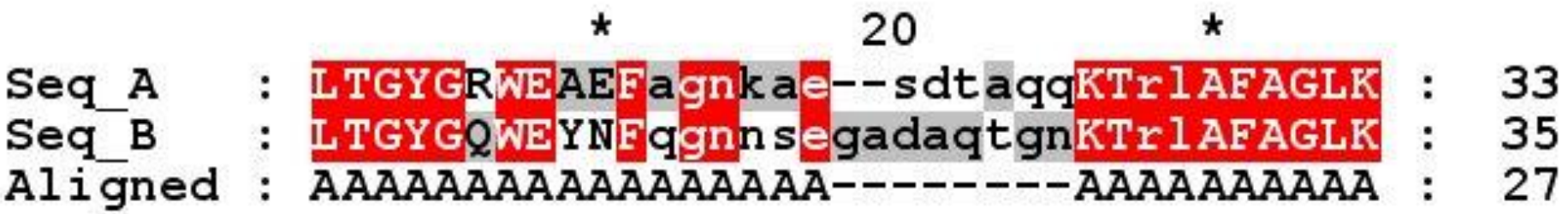
Другой способ отобразить совмещение полипептидных цепей называется структурным выравниванием последовательностей

Совмещение структур и выравнивание последовательностей



```
                        *          20          *
Seq_A    :  LTGYGRWEAEFagnkae--sdtaqqKTrlAFAGLK  :  33
Seq_B    :  LTGYGQWEYNFqgnnsegadaqtgnKTrlAFAGLK  :  35
Aligned  :  AAAAAAAAAAAAAAAAA--------AAAAAAAAAA  :  27
```

Еще раз: разметка по совмещенным структурам

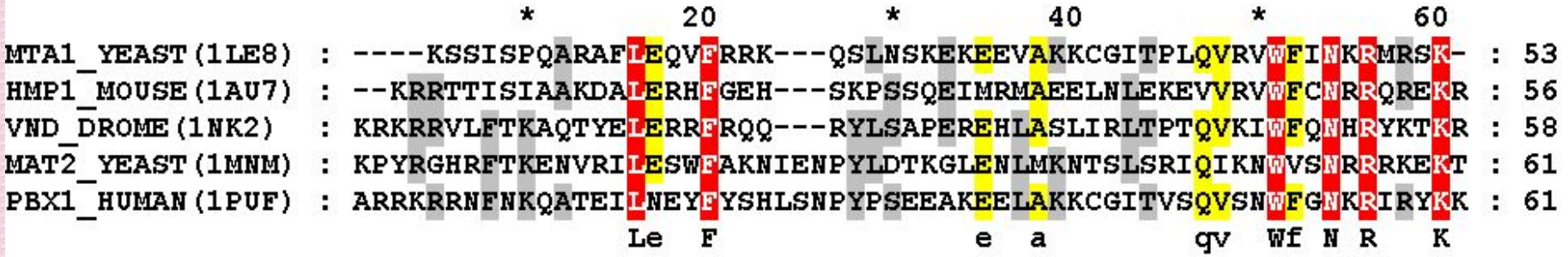


Биологически обоснованное выравнивание гомеодоменов

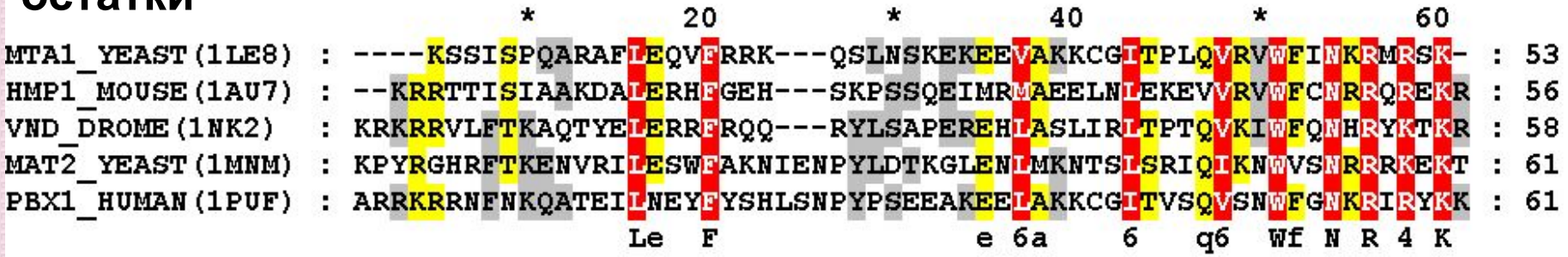
```
                                  *          20          *
MTA1_YEAST(1LE8:A)  :  ----KSSISPQARAFLEQVFRRK---QSLNS  :  24
MAT2_YEAST(1MNM:C)  :  KPYRGHRFTKENVRILESWFAKNIENPYLDT  :  31
Aligned             :  -------AAAAAAAAAAAAAAAA---AAAAA  :  21

                              40          *          60
MTA1_YEAST(1LE8:A)  :  KEKEEVAKKCGITPLQVRVWFINKRMRSK-  :  53
MAT2_YEAST(1MNM:C)  :  KGLENLMKNTSLSRIQIKNWVSNRRRKEKT  :  61
Aligned             :  AAAAAAAAAAAAAAAAAAAAAAAAAAAAA-  :  50
```

Совмещение 5-и гомеодоменов

Множественное выравнивание гомеодоменов

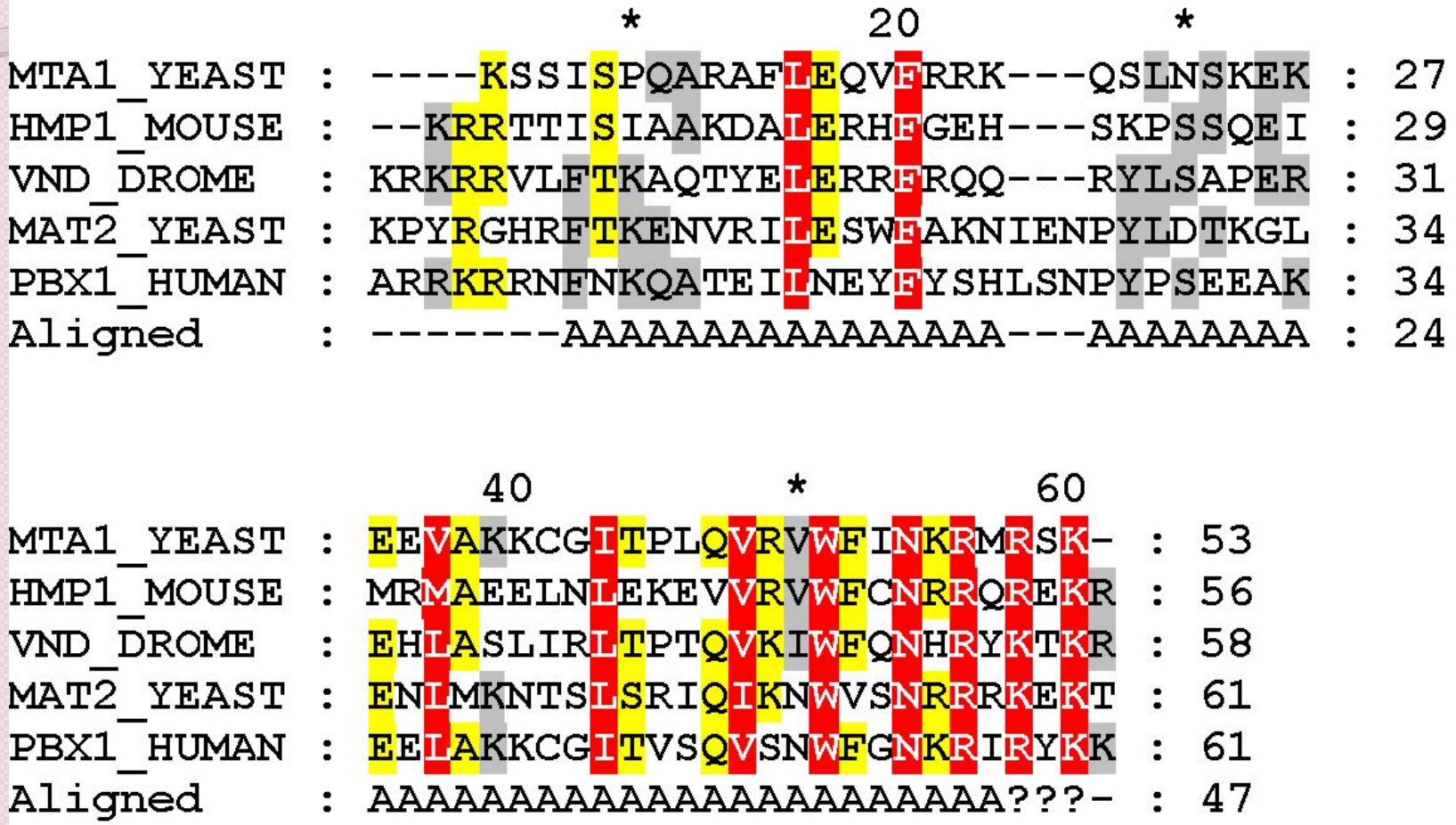


Красным выделены консервативные (одинаковые у всех) остатки;
желтым – на 80% консервативные (одинаковые почти у всех) остатки

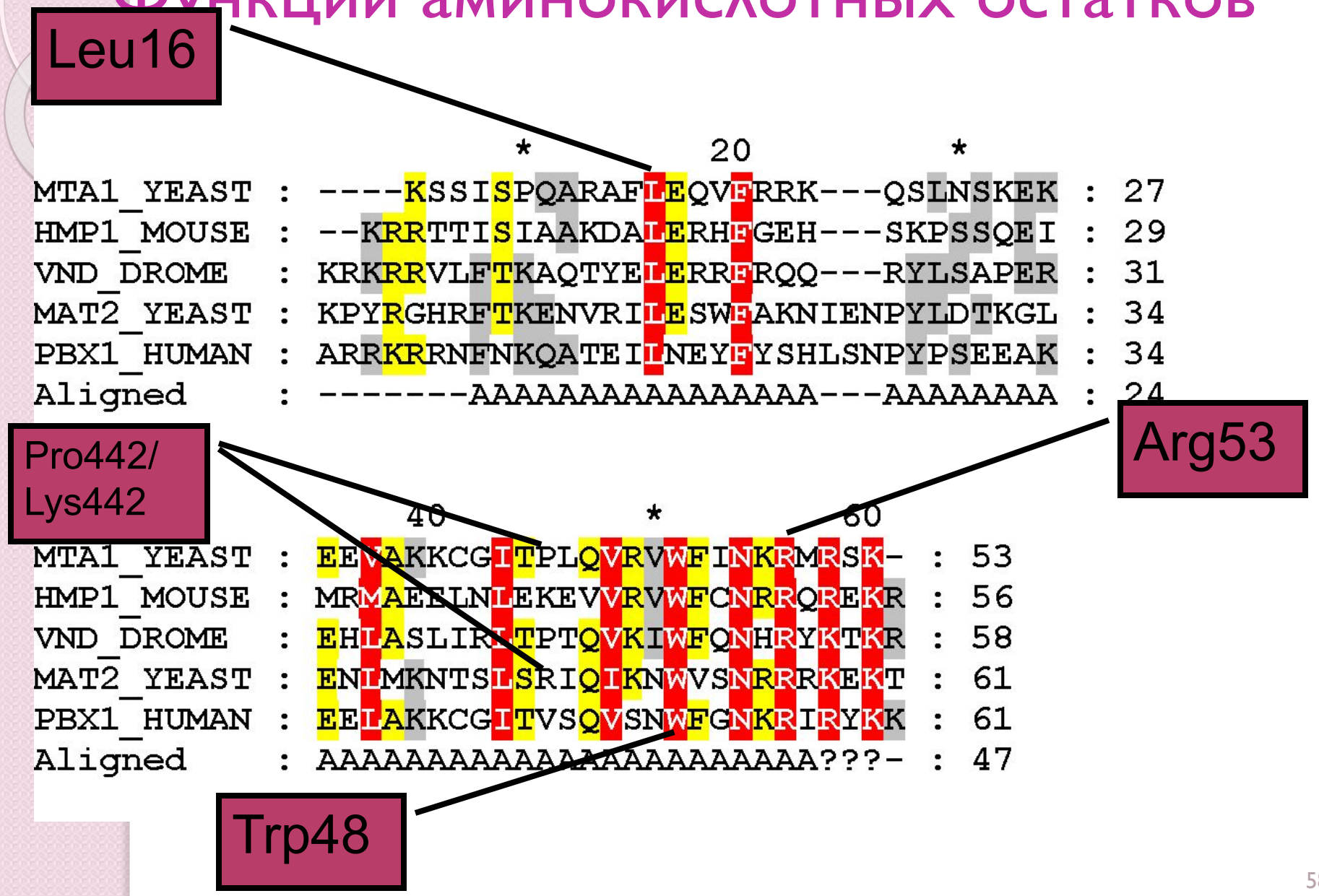


Красным выделены консервативные и функционально консервативные остатки

Размеченное множественное выравнивание



Функции аминокислотных остатков



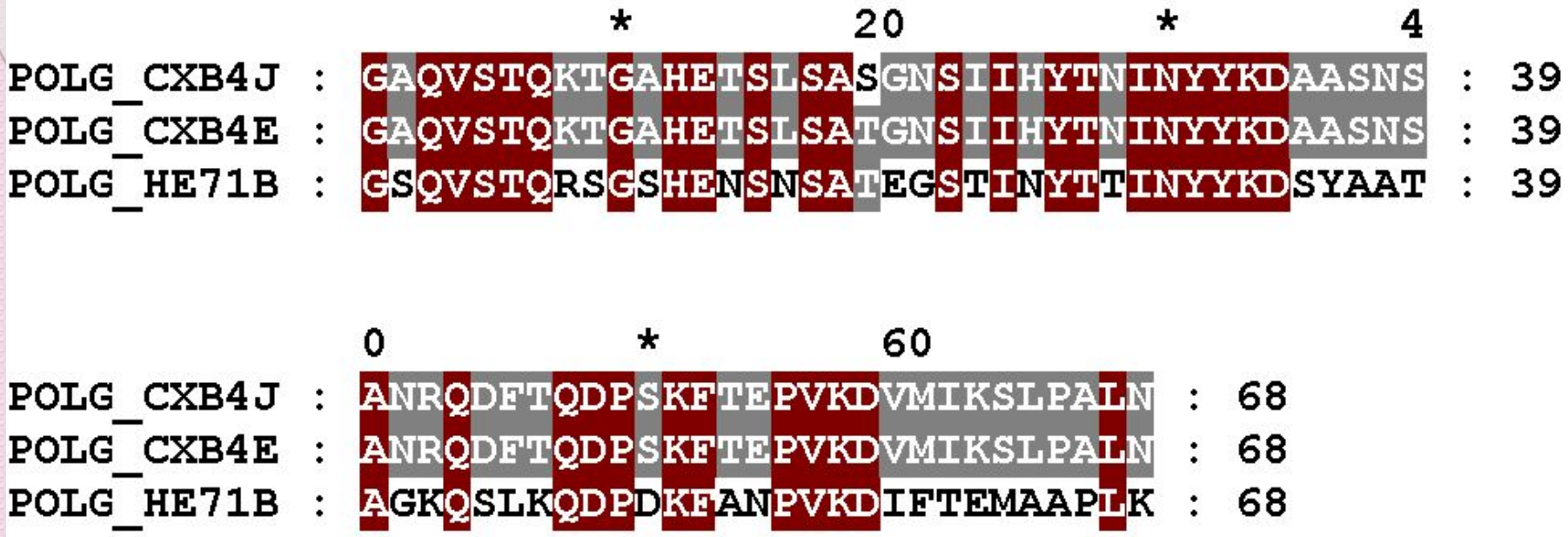
В “правильном” выравнивании много консервативных аминокислотных остатков и функционально консервативных позиций

Выравнивание и эволюция

```
                          *         20          *          4
POLG_CXB4J  :  GAQVSTQKTGAHETSLSASGNSIIHYTNINYYKDAASNS  :  39
POLG_CXB4E  :  GAQVSTQKTGAHETSLSATGNSIIHYTNINYYKDAASNS  :  39

               0          *         60
POLG_CXB4J  :  ANRQDFTQDPSKFTEPVKDVMIKSLPALN  :  68
POLG_CXB4E  :  ANRQDFTQDPSKFTEPVKDVMIKSLPALN  :  68
```

Последовательности белка оболочки из двух штаммов вируса Коксаки

..



Последовательности белка оболочки из двух штаммов вируса Коксаки и энтеровируса человека

Аминокислотные остатки в одной колонке биологически обоснованного выравнивания, как правило, "произошли" из одного и того же остатка - их общего предка

ПРОБЛЕМА: как построить “правильное” выравнивание последовательностей белков если структуры белков неизвестны?

На сегодня известны:

- более 10 млн(!!!) последовательностей белков (включая фрагменты и трансляты)
- пространственные структуры около 65 тыс. белков

Алгоритмические решения проблемы воплощены в программах

Программы выравнивания последовательностей тестируются путем сравнения с биологически обоснованными – построенными по совмещению структур – выравниваниями

Существуют базы данных структурных выравниваний последовательностей (BAliBAse и др.)

Предположим, известны структуры родственных белков и, значит, *биологически обоснованное выравнивание последовательностей*

- При > 60% совпадающих букв любая современная программа даст *(почти)* правильный результат
- При < 20% совпадающих букв (такие примеры существуют) ни одна программа не даст правильного выравнивания
- Между 20% и 60% , обычно, результат программы частично правилен

(*) Справедливы ли положения с предыдущего слайда для выравнивания

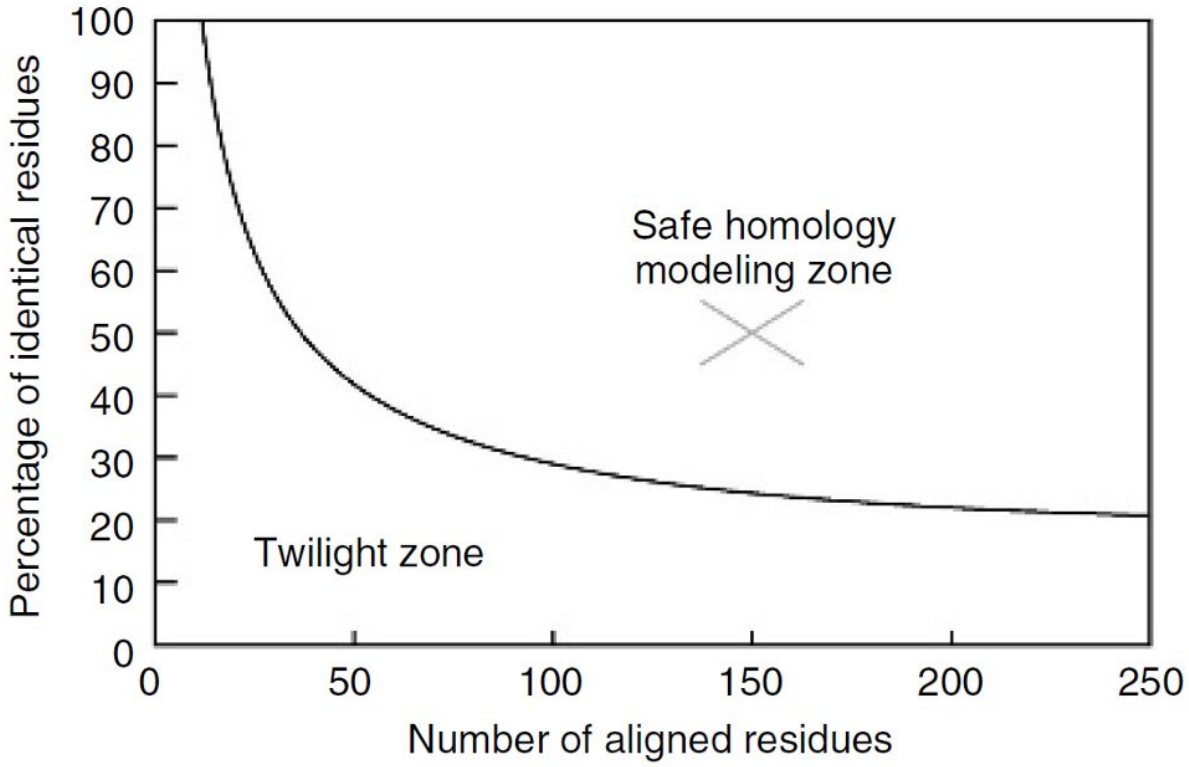
- последовательностей ДНК?
- последовательностей РНК?

2 основних підходи до відтворення просторової структури білка *in silico*

моделювання за гомологією

конформаційний пошук

Утворення тривимірної структури білка in vivo відбувається при біосинтезі або відразу після нього. Чудово, проте, що воно може відбуватися не тільки при біосинтезі: близько 50 років тому Анфінсен показав, що воно може йти і при ренатурації розгорненого білкового ланцюгу in vitro; причому йти абсолютно спонтанно, без допомоги інших макромолекул. Це означає, що амінокислотна послідовність сама (при відповідній температурі і pH води!) визначає просторову структуру білка, тобто білок здатний до самоорганізації.



1999 рік – Rost B. Twilight zone of protein sequence alignment

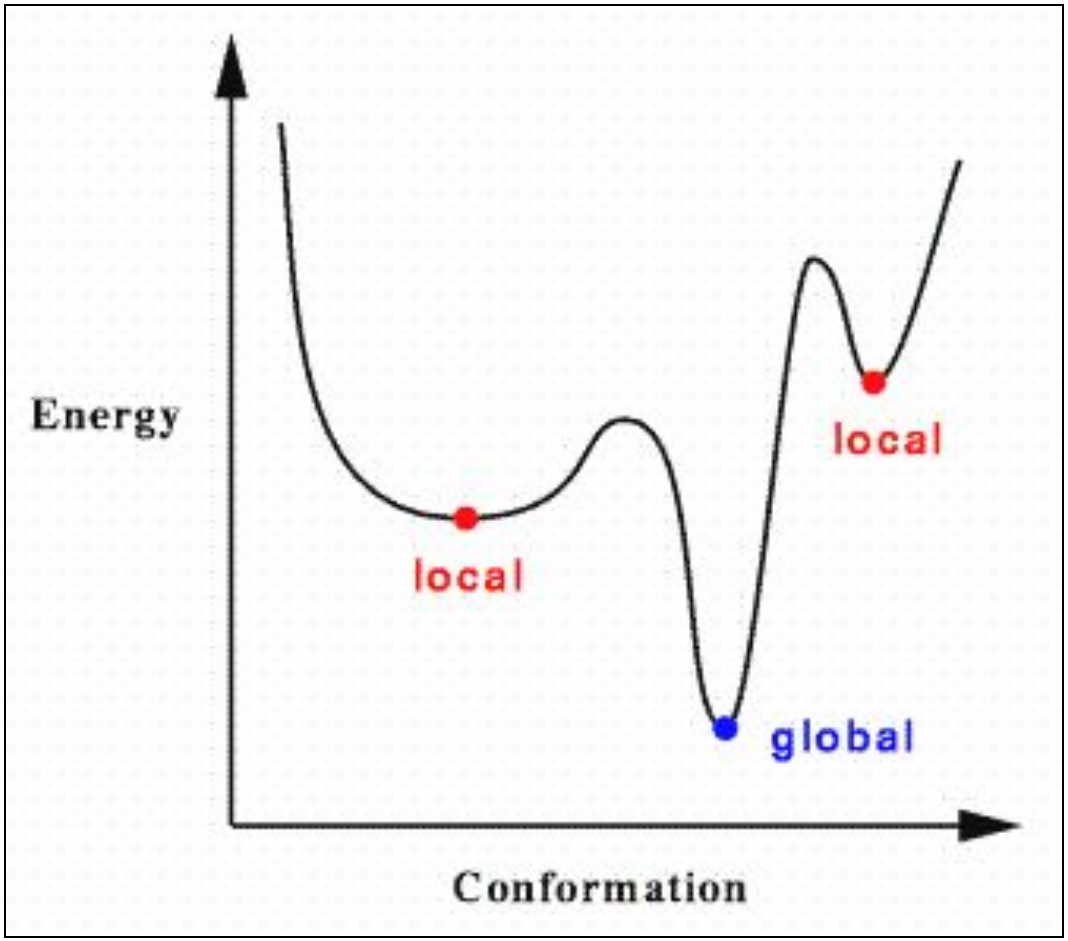


схема залежності енергії молекули від її конформації

Для тубулінів будь-якого походження є характерним явище специфічної взаємодії з низькомолекулярними і не тільки органічними речовинами .

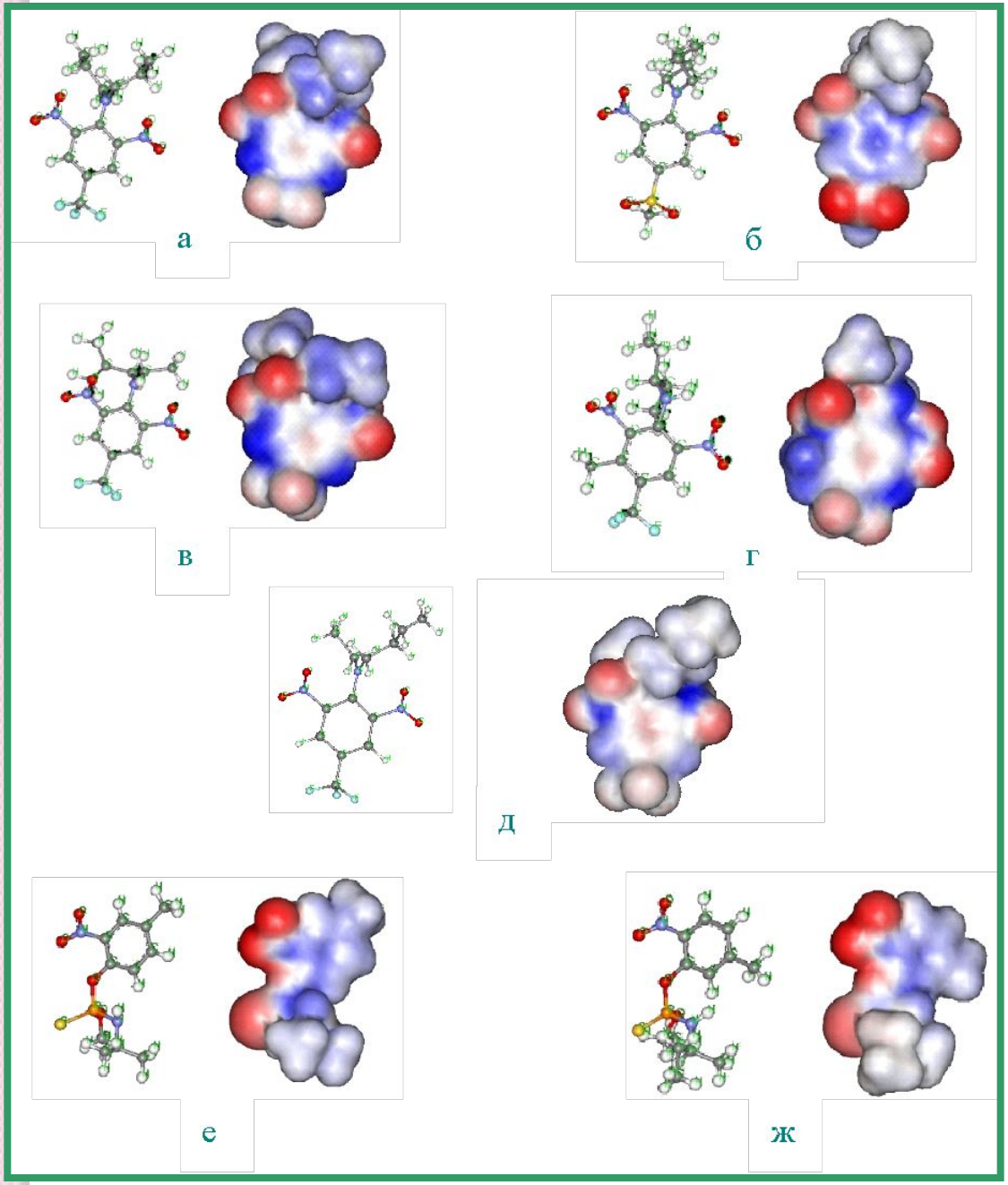
Тубуліни є мішенями для цілого ряду речовин, що характеризуються гербіцидними, протипухлинними, фунгіцидними, протигельмінтними, антипротозойними та іншими видами біологічної активності.

Виникнення стійкості до антимікроторубочкових речовин обумолене точковими мутаціями в молеклулах тубулінів.

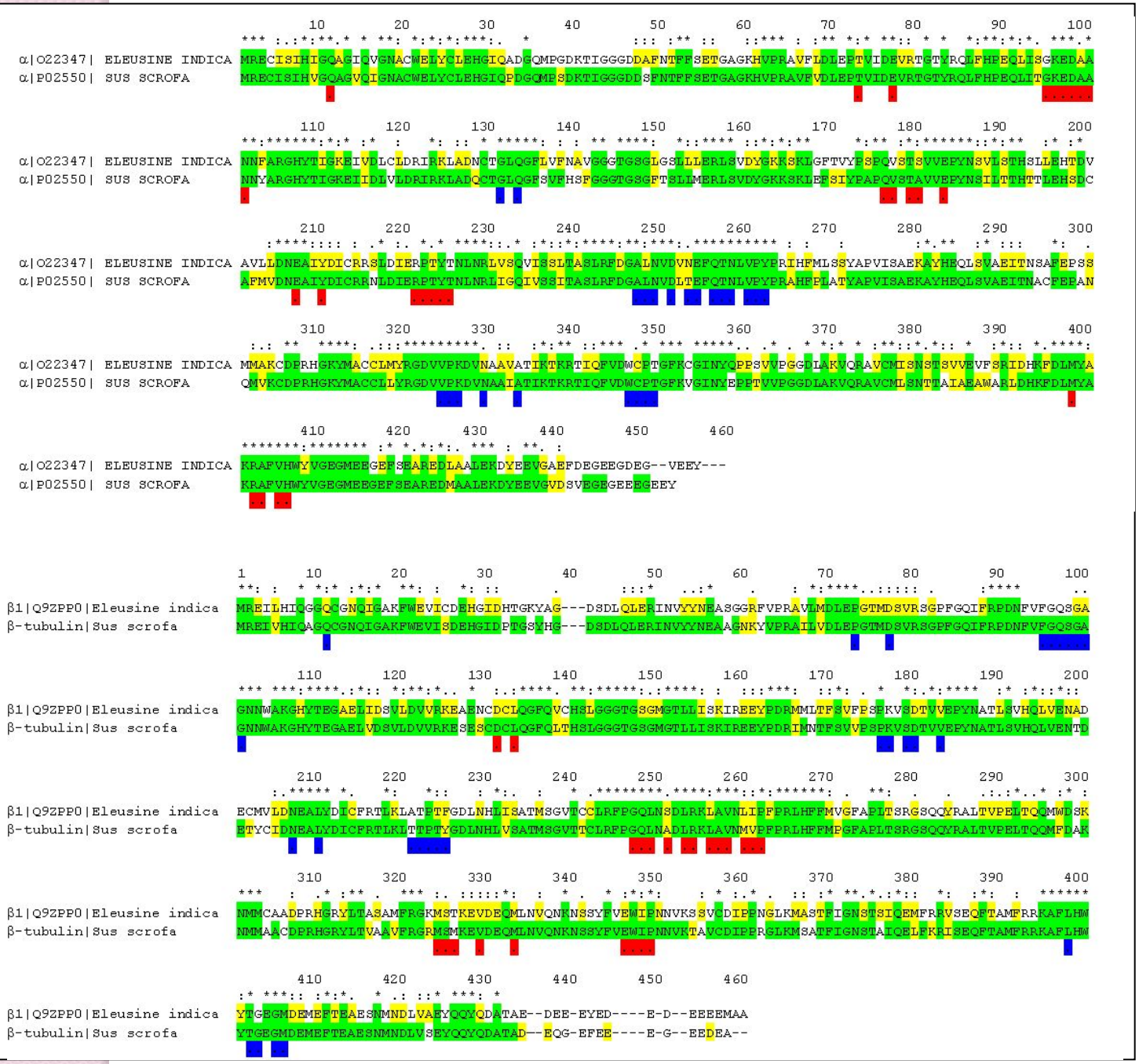
Незважаючи на високу консервативність структури тубулінів різного походження, рослинні тубуліни характеризуються наявністю унікальних властивостей.

Насамперед це стосується їх здатності специфічним чином зв'язувати низькомолекулярні сполуки динітроанілінового та фосфороамідного рядів, що застосовуються як гербіциди.

Зазначені класи речовин виступають ефекторами для тубулінів рослинного та протозойного походження і взагалі не взаємодіють з тваринними та грибними тубулінами, незважаючи на надзвичайно високий рівень гомології їх амінокислотних послідовностей.



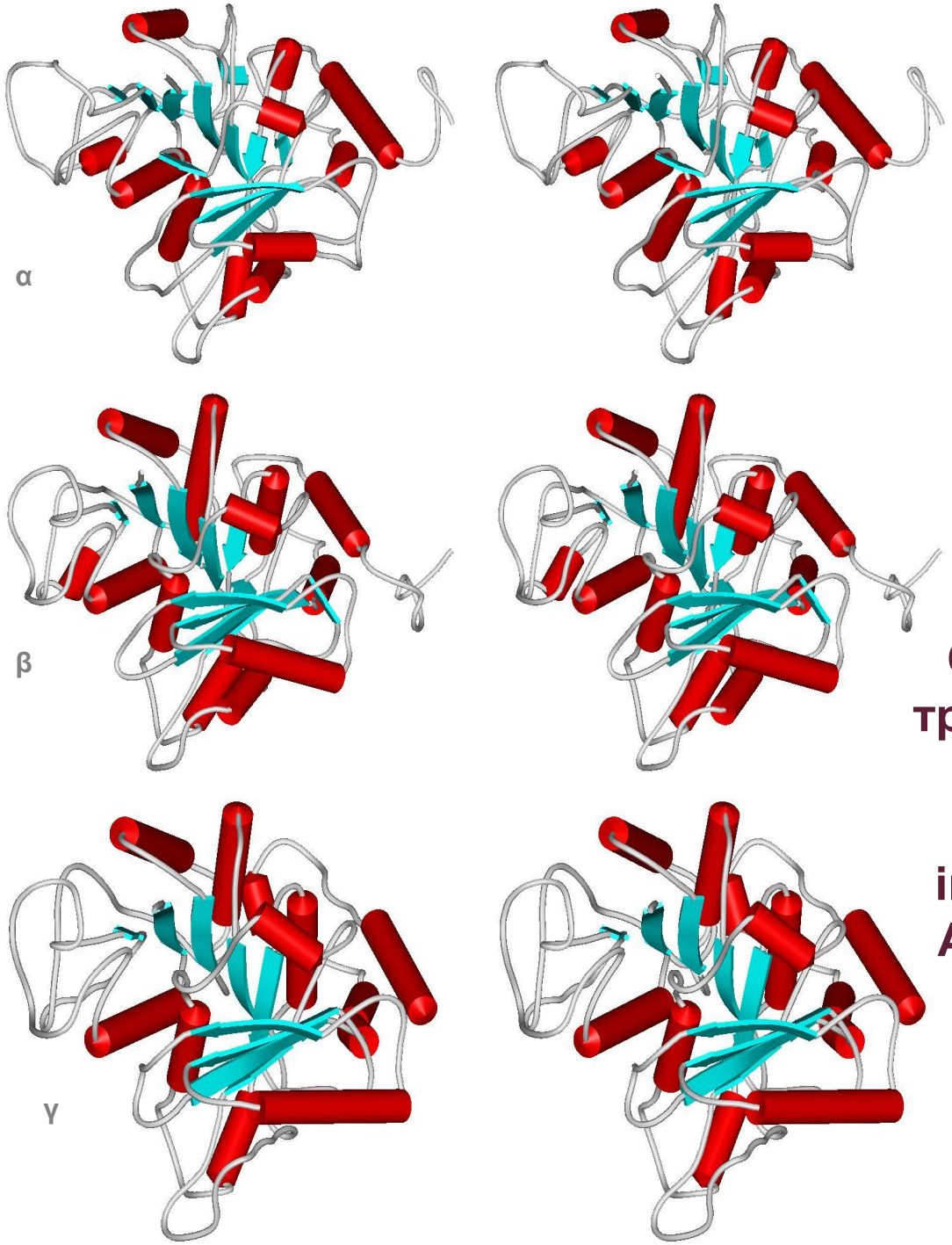
Просторова структура та розподіл електростатичного потенціалу на поверхні представників динітроанілінів а - трифлюралін, б - орізалін, в - еталфлюралін, г - пендіметалін, д - бенефін) та фосфороамідів е-аміпрфосметил, ж-кремарт)



Порівняльне вирівнювання послідовностей тубулінів рослинного (*Eleusine indica*) та тваринного *(Sus scrofa)* походження. Вівень тотожності послідовностей складає 86%

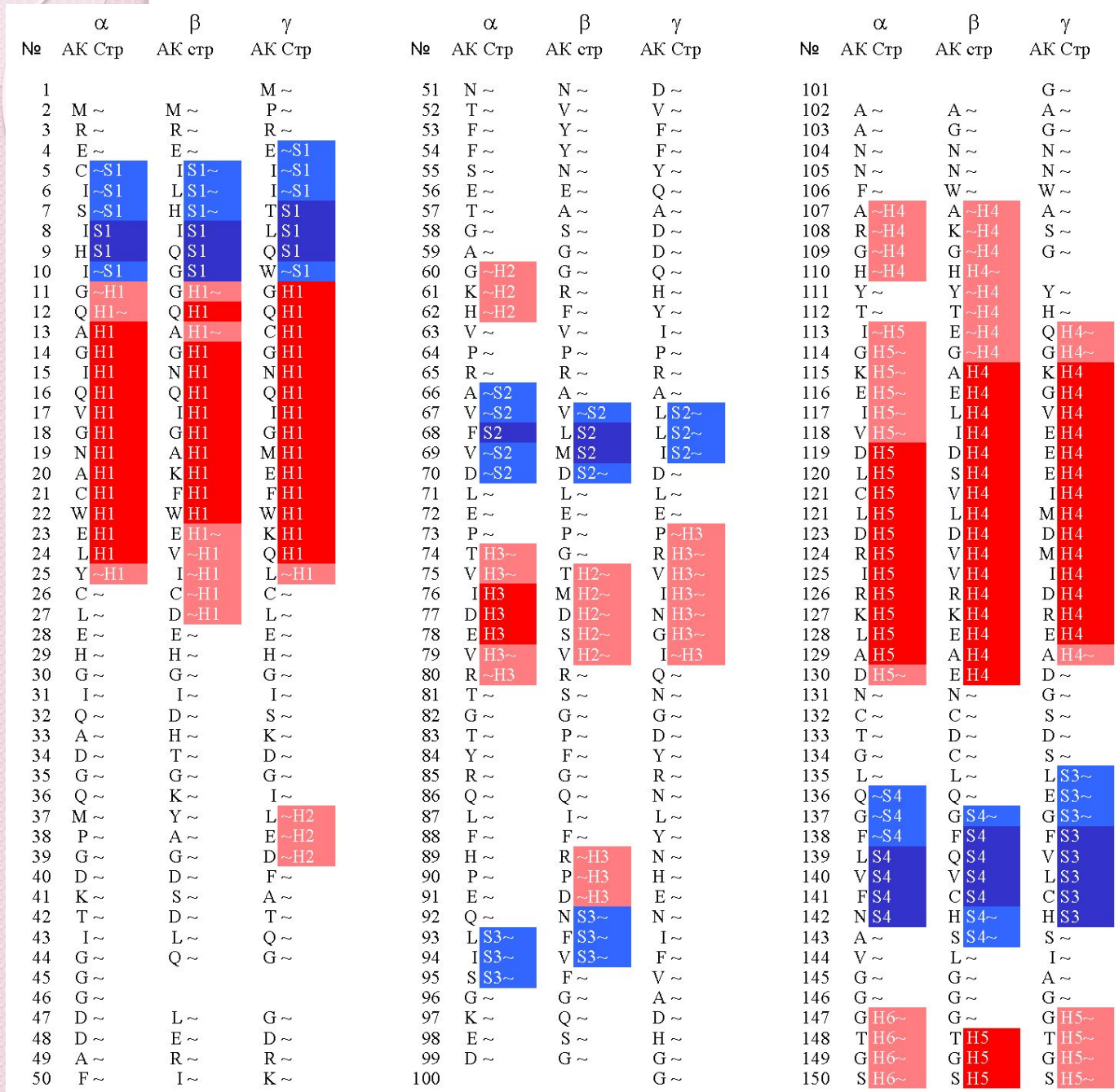
Відсутність досліджень особливостей просторової структури рослинних тубулінів

- труднощі технологічного характеру при отримані рослинних тубулінів із ступенем чистоти, необхідним для їх кристалізації
- обмеження самих кристалографічних методів, що у більшості випадків не дозволяють виявити різниці в просторовій структурі високогомологічних білків.

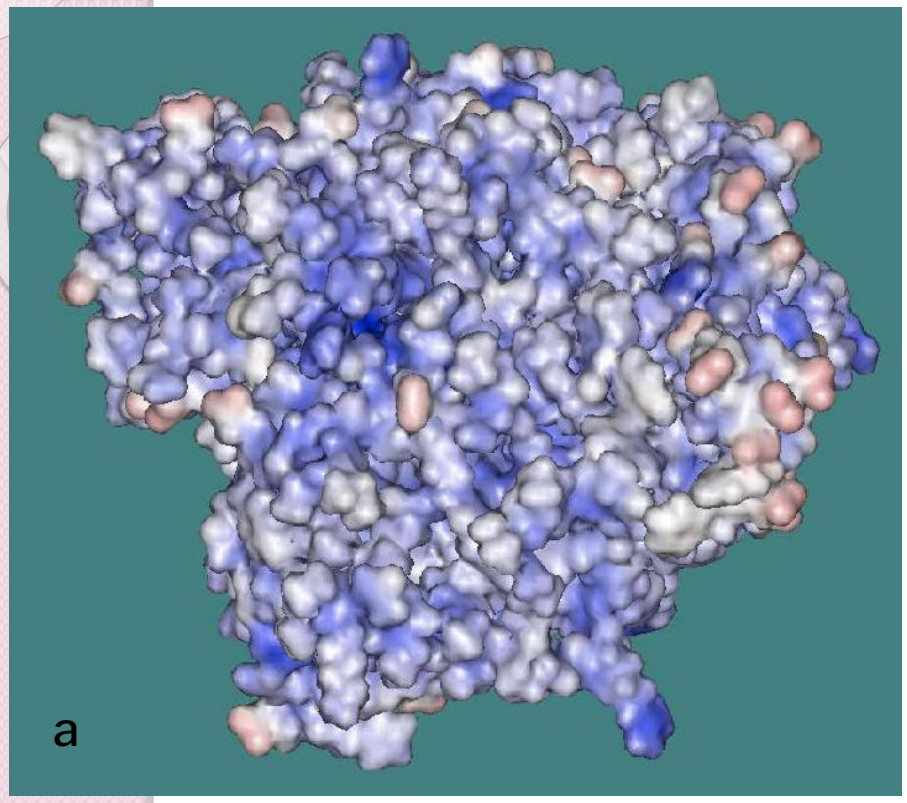


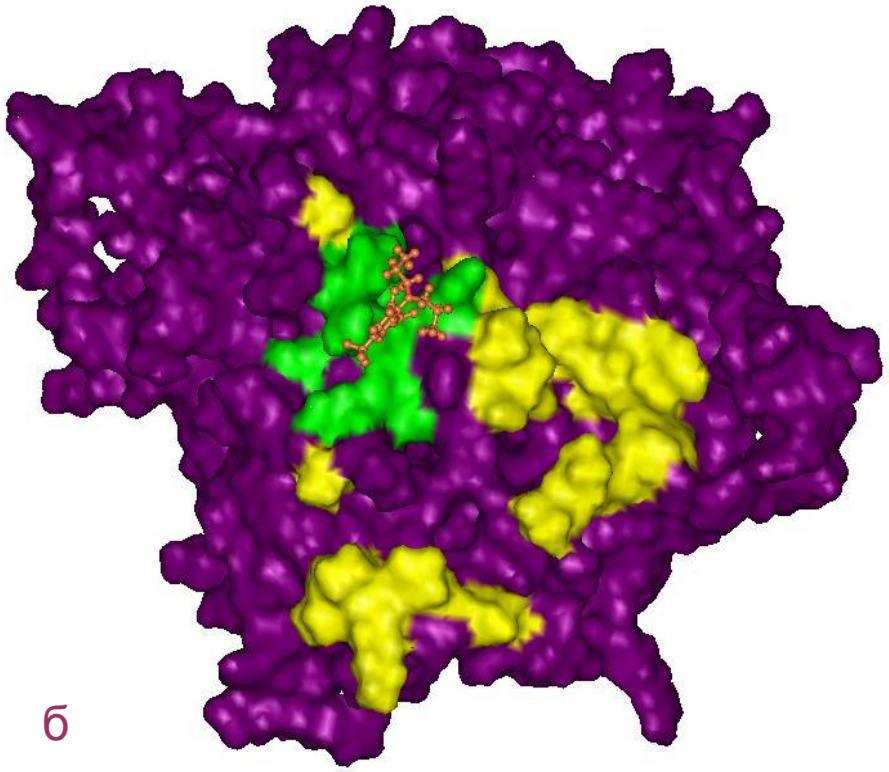
Стереозображення тривимірної упаковки молекул α-і β-тубулінів Eleusine indica та γ -тубуліну Arabidopsis thaliana

- Фундаментальною особливістю тубулінів є явно виражена метастабільність елементів вторинної структури у часі – явище, яке характеризується наявністю переходів цілого ряду амінокислотних залишків, що входять до β-складок і α-спіралей, у неупорядковані структури і назад.

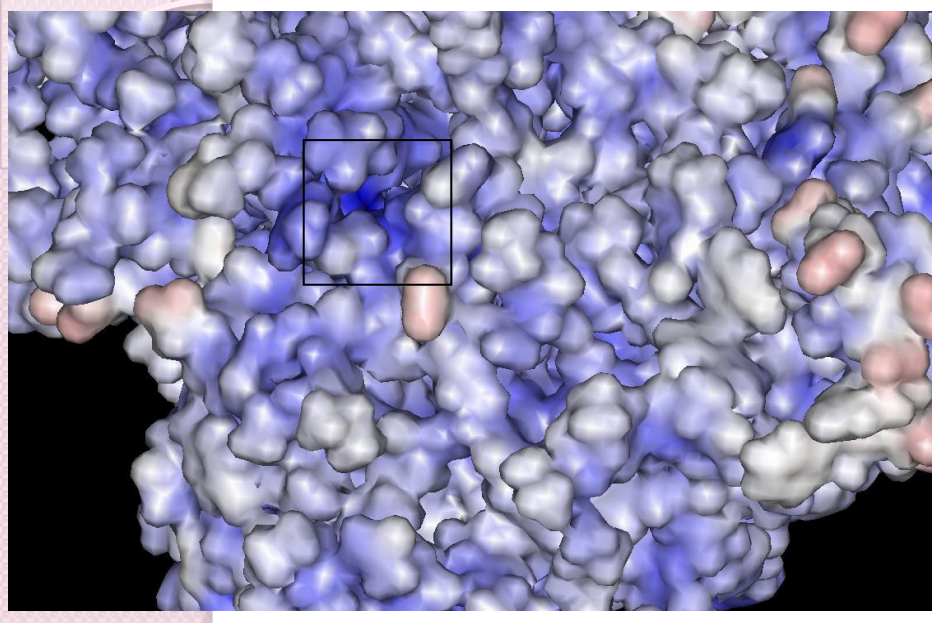


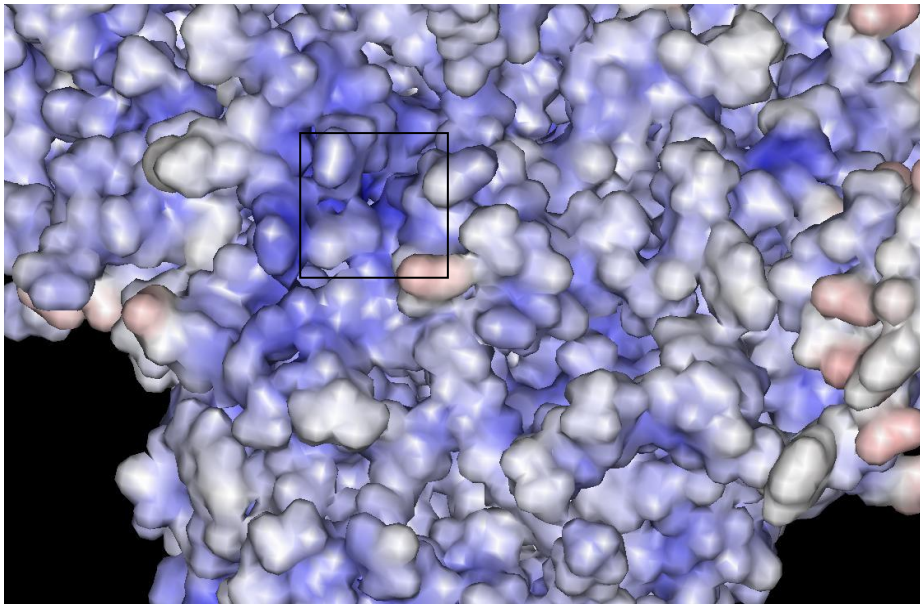
Діаграма розташування елементів вторинної структури в молекулах α, β та γ-тубуліну рослин на ділянці з 1 по 150 амінокислотний залишок

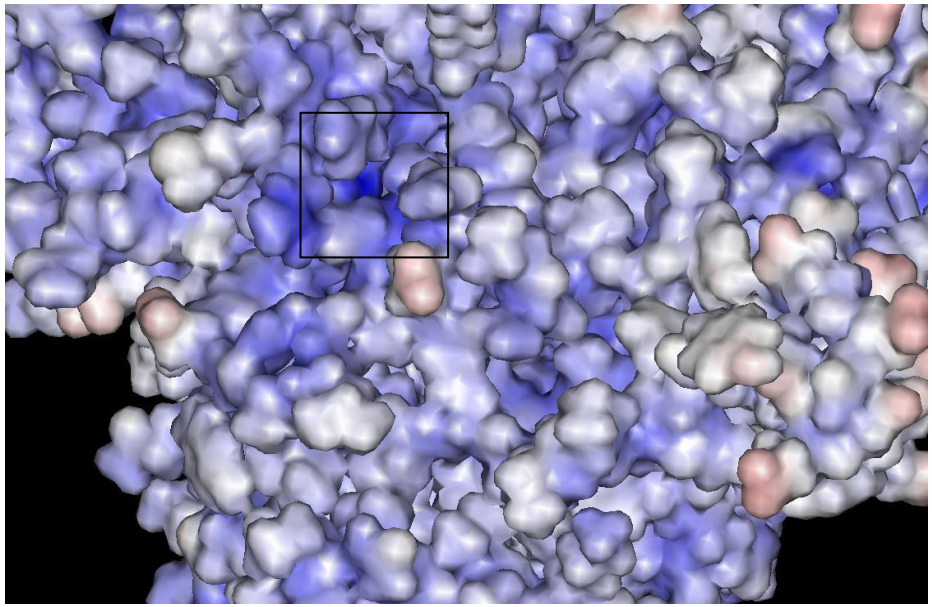




Вид молекулярної поверхні α-тубуліну з боку інтердимерного контакту: а – розподіл електростатичного потенціалу на молекулярній поверхні, б – розташування контактних амінокислотних залишків (жовтий колір) та залишків, що утворюють сайт взаємодії з динітроаніліновими та фосфороамідними сполуками (зелений колір). В сайті розташована молекула трифлюраліну.

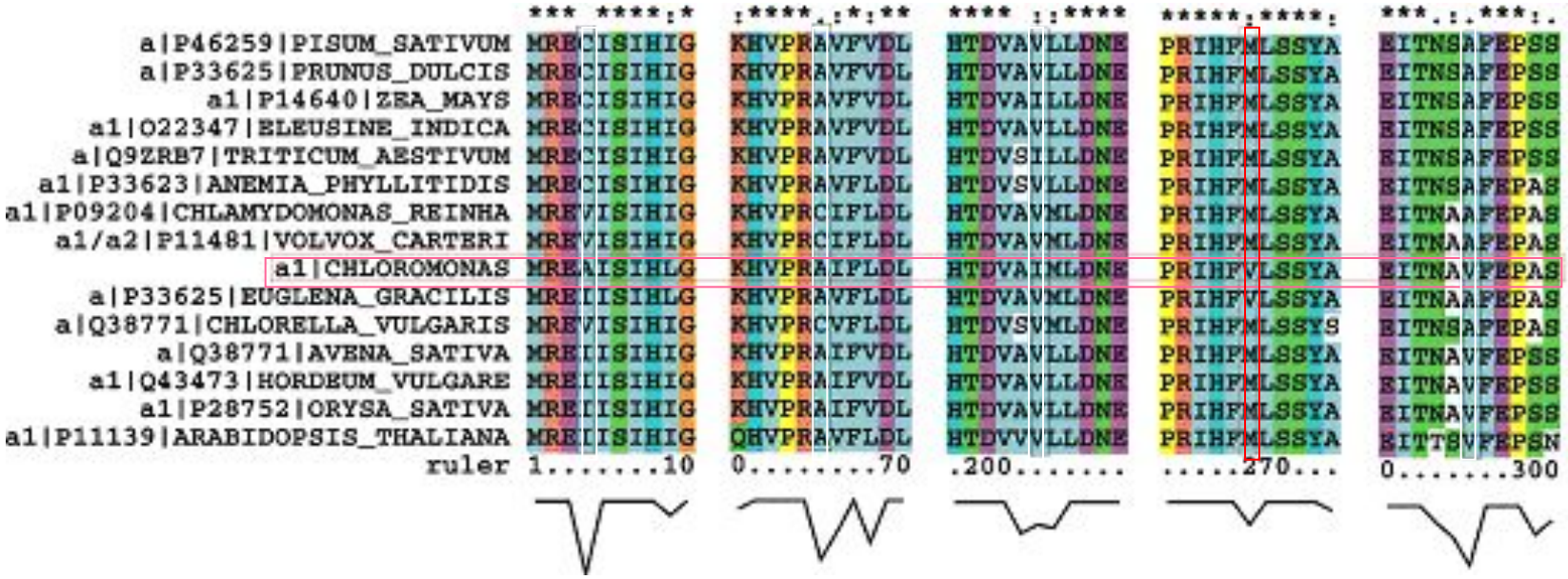






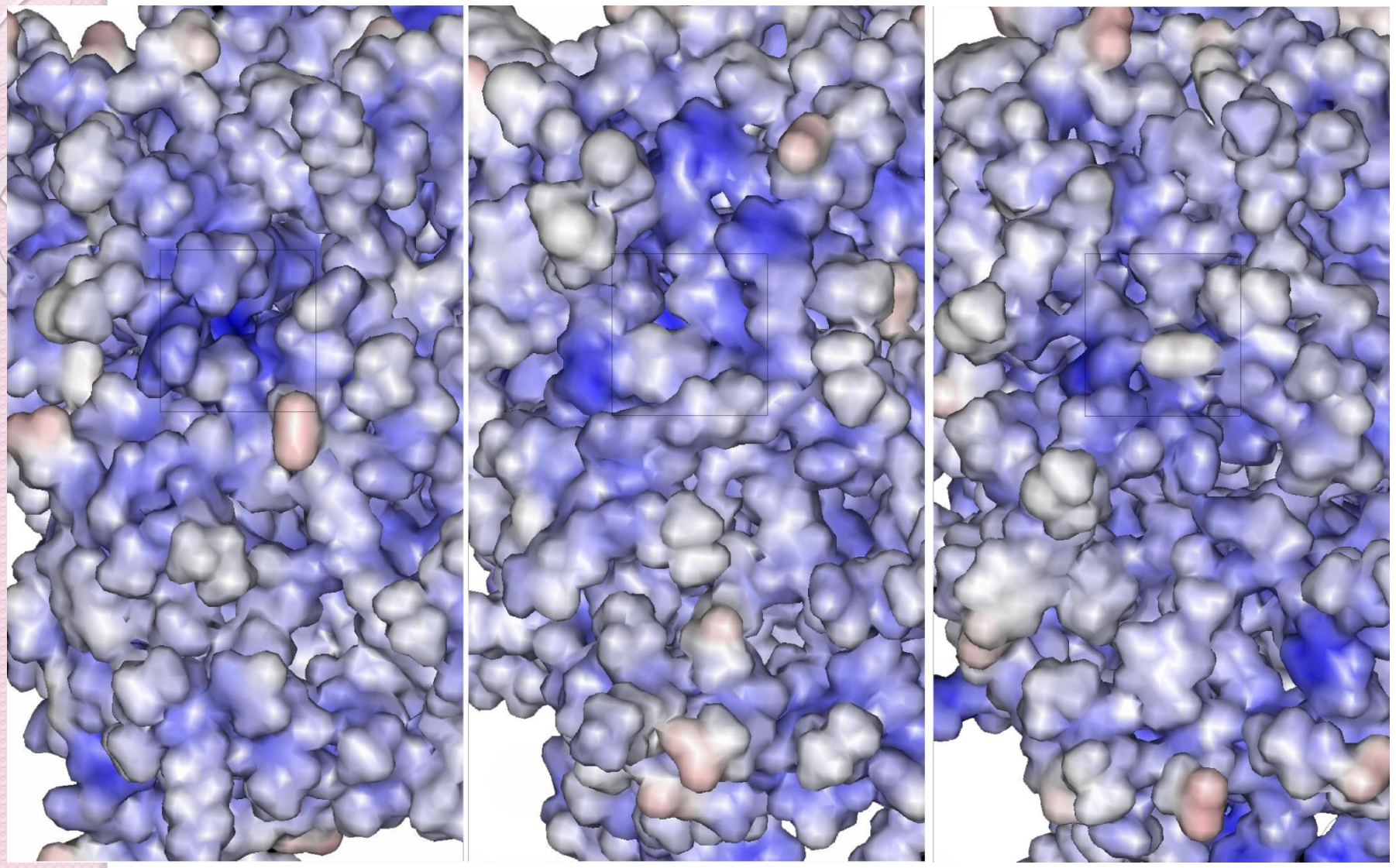
Особливості рельєфу поверхні та розподілу електростатичного потенціалу в області сайту взаємодії α-тубуліну *E. indica* з динітроаніліновими та фосфороамідними сполуками

- Мутація Met–>Thr в позиції 268 рослинного α-тубуліну, яка викликає виникнення проміжної стійкості до динітроанілінових гербіцидів, співпадає з позицією заміни Met–>Val, яка спричиняє підвищення рівня холодостійкості і, в свою чергу, приводить до перебудов поверхні інтердимерного контакту.



Порівняльний аналіз послідовностей рослинних α-тубулінів

Представлено ділянки послідовностей, що безпосередньо прилягають до амінокислотних залишків, для яких виявлені заміни в α-тубуліні хлоромонаса. Місця розташування цих залишків виділені рамкою



Карти молекулярної поверхні рослинних тубулінів в області, що відповідає сайту зв’язування на поверхні α-тубуліна. α-тубулін – зліва, β-тубулін – посередині, γ-тубулін – справа

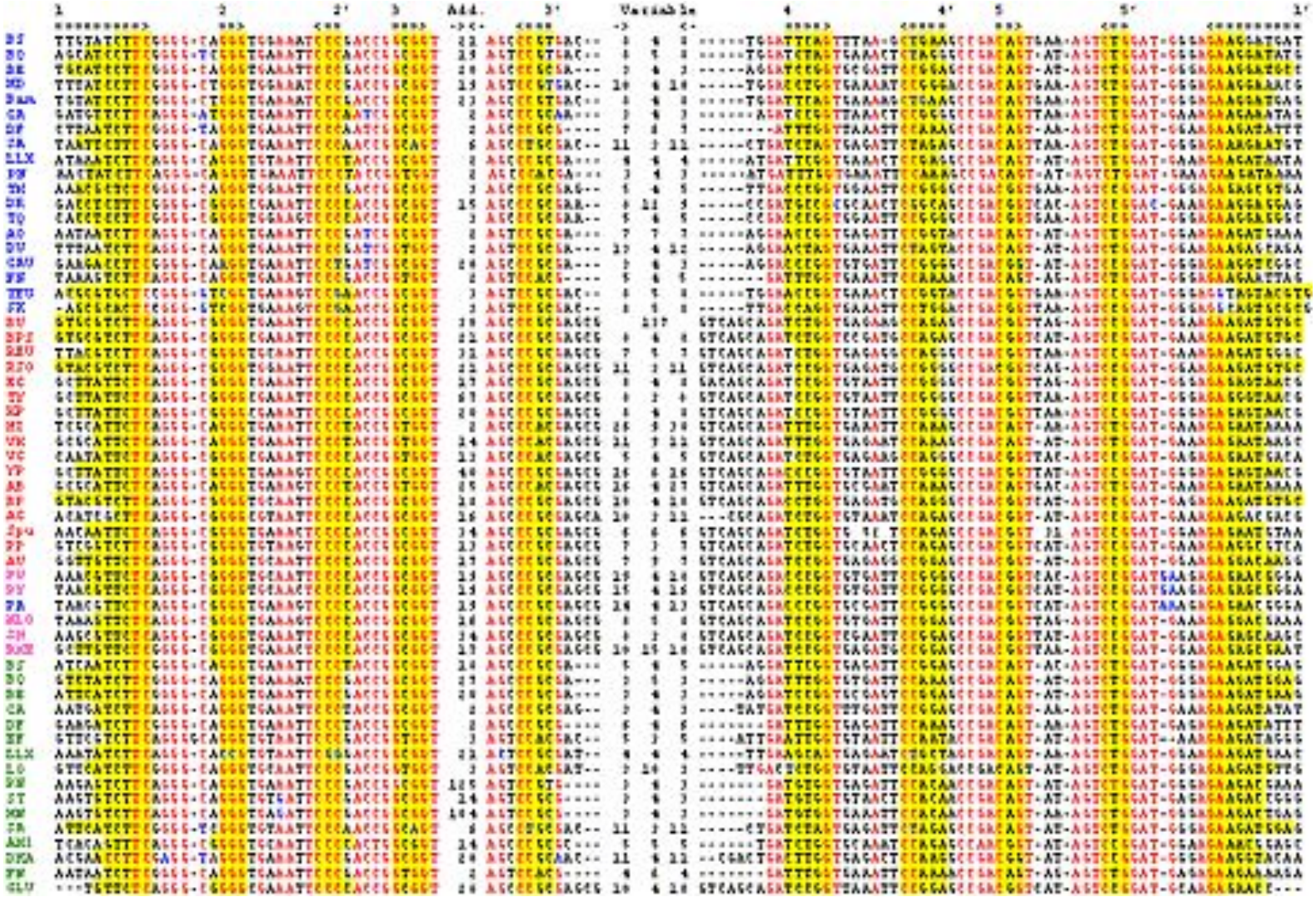
Распознавание генов

- Поиск открытых рамок считывания
- Использование статистики (отличия белок-кодирующих и некодирующих областей)
- Идентификация начал генов – участки связывания рибосом (прокариоты)
- Экзон-интронная структура (эукариоты)
- Сравнения с известными генами
- Геномные сравнения

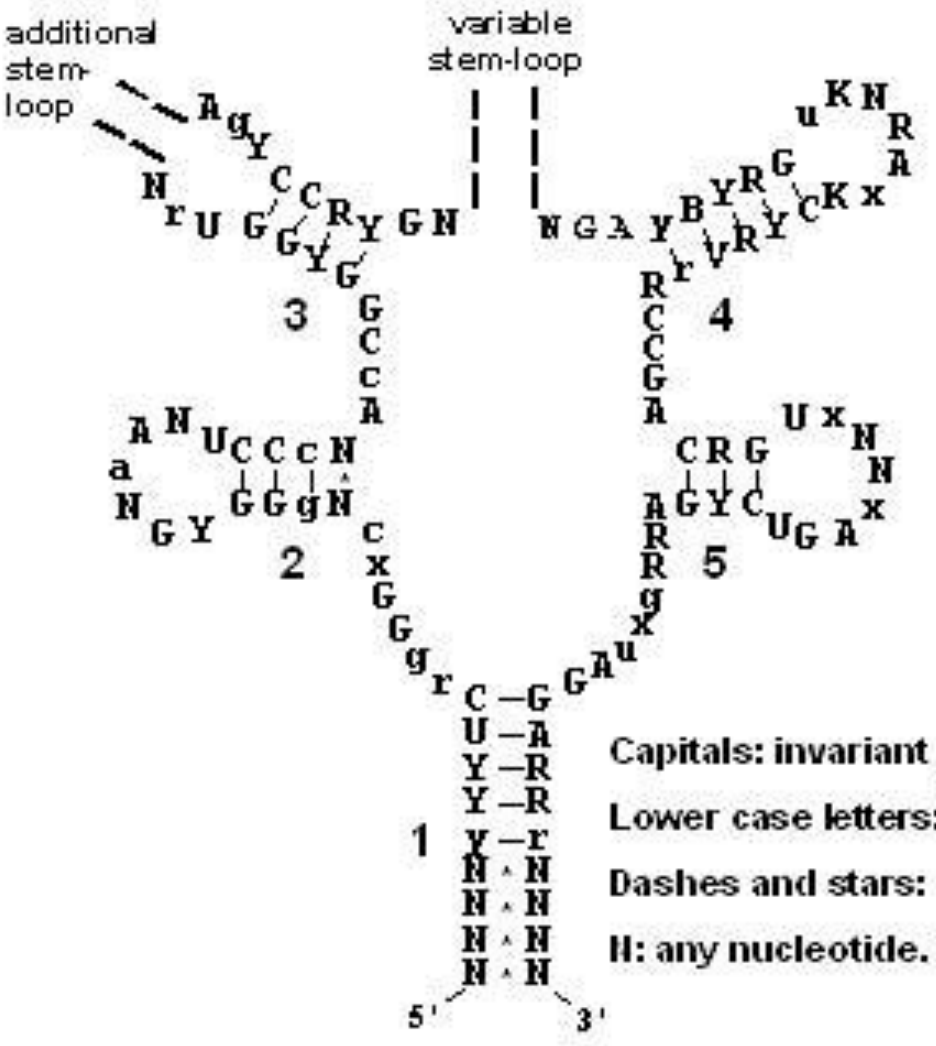
Ортологи и паралоги

- Ортологи – гени з різних організмів, що розійшлися при видоутворенні.
  - Мається на увазі, що ортологи мають спільного «предка» і однакову функцію (якщо тиск відбора слабкий, то функція может «плисти»).
- Паралоги – гени, що розійшлися при дуплікації («копіюванні»).
  - Копії гена не зазнавали тиска відбора, а значить, могли змінити функцію.

Регуляторні послідовності в геномі бактерій



Регуляторні послідовності в геномі бактерій



Capitals: invariant (absolutely conserved) positions.

Lower case letters: strongly conserved positions.

Dashes and stars: obligatory and facultative base pairs

H: any nucleotide. X: any nucleotide or deletion

Цель (глобальная)

Предсказать свойства организма путем (компьютерного) анализа его генома

(возможно, с использованием дополнительной информации: эпигенетика, белок-белковые взаимодействия и т.п.)

сейчас: метаболическая реконструкция, транспортные системы, ответ на стресс и т.д.

“Понять” эволюцию геномов/организмов

«Неприкладная» биоинформатика

- Молекулярная эволюция
  - филогения генов
  - таксономия организмов
  - горизонтальные переносы и т.п.
  - положительный и отрицательный отбор
    - что сделало нас людьми?
    - лекарственная устойчивость
  - эволюция геномов
- Системная биология
  - строение геномов
  - сети взаимодействий
    - белок-белковые
    - регуляция транскрипции
    - сигнальные пути

Задачи

- С проверяемым ответом
  - предсказание функции, регуляции, структуры и т.п.:
    - ставим эксперимент
- С непроверяемым ответом
  - эволюционные деревья
    - но если бы знать все геномы всех (в том числе *очень* давно умерших) существ, то задача станет тривиальной
- С принципиально непроверяемым ответом (который зависит от операциональных определений)
  - идентификация повторов, консервативных областей, островов метилирования и т.п.
    - (так ли он непроверяем?)
- Без ответа (общеописательные)
  - статистика геномов (изохоры и т.п.)
  - описание регуляторных и пр. сетей (hubs, мотивы и т.п.)

«В принципе не проверяемые ответы» (зависящие от определений)

Так ли они непроверяемы?

- Повторы
  - если иметь ***все*** геномы, то можно описывать вставки/замены фрагментов генома и их последующее расхождение
- Консервативные области
  - если иметь ***все*** геномы, то можно просто оценивать локальную скорость эволюции (но это будет функцией времени)
- Статистика ДНК (локальный нуклеотидный состав)
  - это следствие локального паттерна замен, так и надо описывать
- Микросателлиты
  - можно ли «функционально» (а не операционально) определить микросателлит, исходя из динамики вставок/замен/дупликаций?
- CpG-острова
  - можно ли «функционально» (а не операционально) определить CpG-остров, исходя из паттерна мутаций, состояния метилирования и т.п.? (тут уже эволюция + эксперимент)

Цель (недостижимая?)

откуда оно все взялось?

первое приближение – реконструкция генома/свойств

реально ли заглянуть глубже?

реально ли смоделировать? (времена)

реально ли смоделировать «по частям»?

Дякую за увагу
Благодарю за внимание
Thank you for your attention