ФГБОУ ВО «Волгоградский государственный медицинский университет» МЗ РФ
Кафедра теоретической биохимии с курсом клинической биохимии

Предсказание и дизайн белковых структур

2023

Резниченко Мария Федоровна

План

- Предсказание и дизайн белковых структур.
- Представление о подходах к предсказанию вторичных и пространственных структур белков по их аминокислотным последовательностям.
- "Опознавание" белковых структур по гомологии последовательностей.
- Выделение стабильных структур белковой цепи.
- "Шаблоны" белковых структур.
- Белковая инженерия и дизайн.

Предсказание функции белка — определение биологической роли белка и значения в контексте клетки.

Необходимо для **плохо изученных белков** или для **гипотетических белков**, предсказанных на основе данных геномных последовательностей.

Источник информации –гомология нуклеотидных последовательностей, профили экспрессии генов, доменная структура белков, интеллектуальный анализ текстов публикаций, филогенетические и фенотипические профили, белок-белковые взаимодействия.

Зачем нам нужно предсказывать структуру белка?

✓ Просто интересно с т.з. исследователя

✓ Трудности в экспериментальном определении пространственной структуры

Изучение механизма действия белка, его функций, подбор искусственных ингибиторов или активаторов к нему, невозможно без знаний его пространственной структуры

3 основных подхода к предсказанию пространственной структуры белков.

1. Предсказание структуры методом поиска сходства с известными структурами белков по гомологии (сравнительное моделирование).
2. protein threading
3. Ab initio (или de novo) моделирование.

CASP - конкурс методов предсказания структуры белков

Highly accurate protein structure prediction with AlphaFold.
Nature | Vol 596 | 26 August 2021.
https://www.nature.com/articles/s41586-021-03819-2

Ab initio (или de novo) моделирование.

Молекулярная динамика

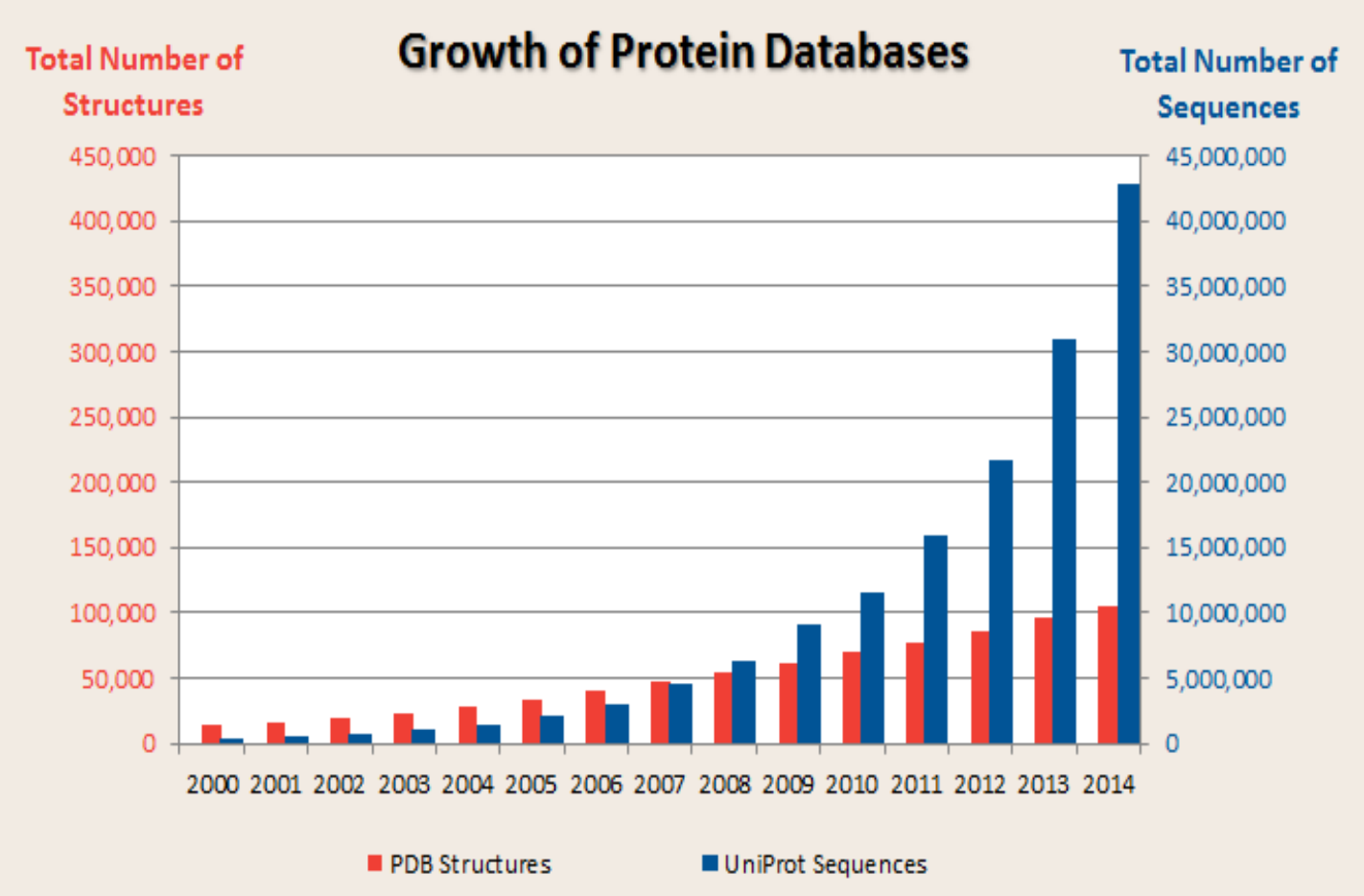
+Монте-Карло

+Марковские цепи - Скрытая марковская модель

https://ru.wikipedia.org/wiki/Скрытая_марковская_модель/

Скрытая марковская модель (СММ) - статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.

Сравнительное моделирование - базис



https://www.dnastar.com/blog/structural-biology/why-structure-prediction-matters/

Гомологичные аминокислотные последовательности N-концевых фрагментов цитохромов с различных митохондрий и хлоропластов эукариотов.

```
                                  1         10        20
Human, chimpanzee                 GDVEKGKKIFIMKCSQCHTV...
Pig, bovine, sheep                GDVEKGKKIFVQKCAQCHTV...
Chicken, turkey                   GDIVEKGKKIVQKCSQCHTV...
Puget sound dogfish               GDVEKGKKVFVQKCAQCHTV...
Screw-worm fly                 GVPAGDVEKGKKIFVQRCAQCHTV...
Rust fungus                    GFEDGDAKKGARIFKTRCAQCHTL...
Rape, cauliflower           ASFDEAPPGNSKAGEKIFKTKCAQCHTV...
```

Аминокислотные последовательности N-концевых фрагментов рибонуклеаз H бактерии (*E.coli*), эукариота (дрожжи, yeast), и трех разных вирусов.

```
                 o  o  •              ◇o◇   o          o   ◇          •o    o  o  o
E.coli  MLKQVEIFTDGSCLGN---PGPGGYGAILRYRGREKTFSAGYTRTTNNRMELMAAIVALEAL
Yeast   YNKSMNYVCDGSSFGNGTSSSRAGYGAYFEGAPEENISPLLSGAQTNNRAEIEAVSEALKKI
MMLV    PDADHTWYTDGSSLLQ---EGQRKAGAAVTTETEVIWAKALPAGTSAQRAELIALTQALK-M
RSV     PVPGPTVFTDASSSTH---KGVVVWREGPRW--EIKEIADLGA---SVQQLEARAVAMALL-L
HIV     IVGAETFYVDGAANRE---TKLGKAGYVTN-KGRQKVVPLTNT--TNQKTELQAIYLALQD-
          =    =:: =    :        :              = :        : =   =  =: :==
            |<── β1 ──>|        |<── β2 ──>|     |<── β3 ──>||<──── α1 ──────>|
```

Зачем нужно выравнивать аминокислотные последовательности?

Данные:

новая последовательность

Биологические задачи:

Предсказание функции, а.к. остатков в «активном центре»

Предсказание 3D-структуры

Реконструкция эволюции

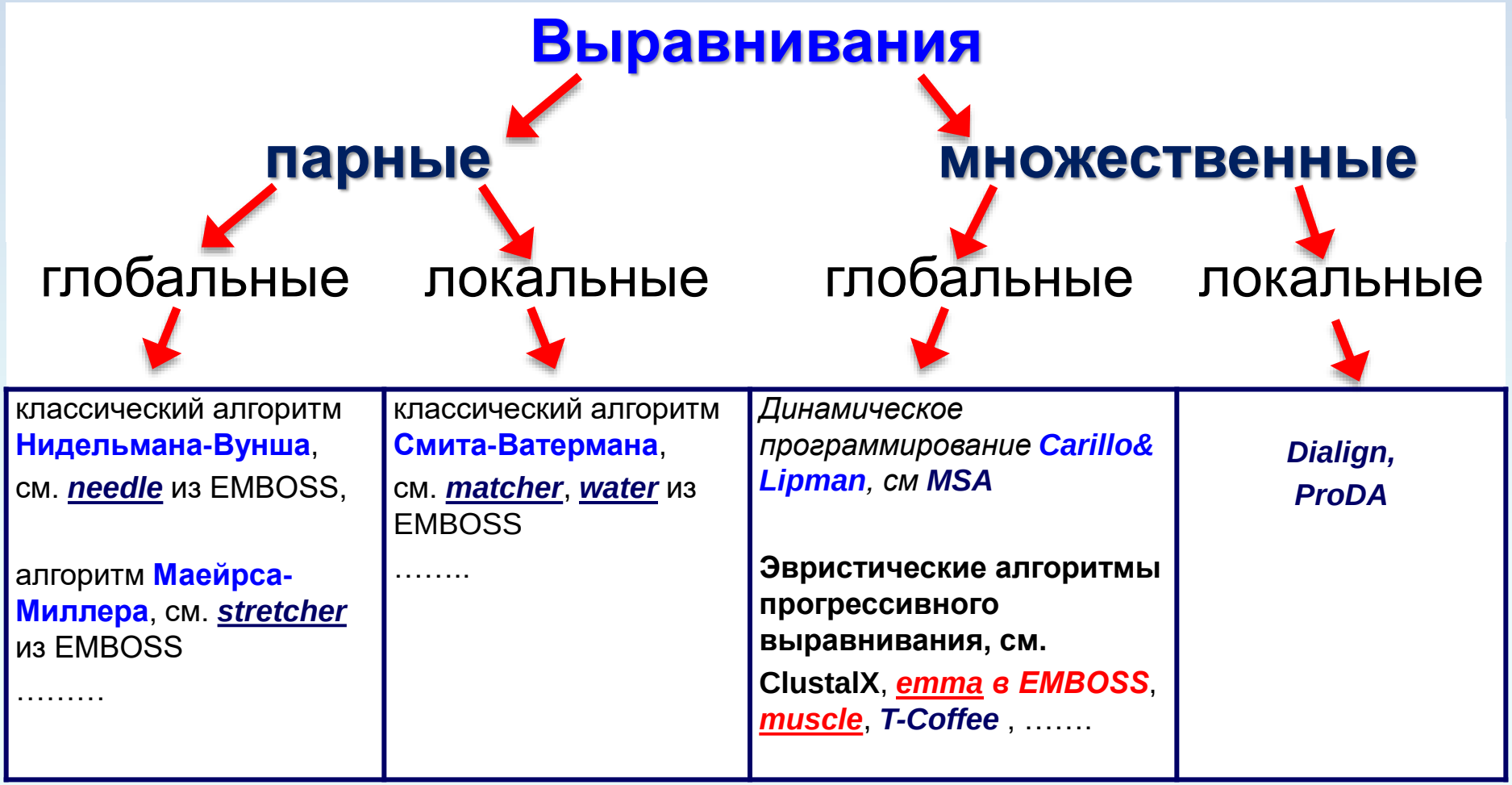
Общий подход к решению — оценка сходства последова-тельностей:

Выбор алгоритма и программы

Построение выравнивания

Анализ выравнивания

Типы выравнивания аминокислотных последовательностей



классический алгоритм **Нидельмана-Вунша**, см. ***needle*** из EMBOSS, алгоритм **Маейрса-Миллера**, см. ***stretcher*** из EMBOSS ……	классический алгоритм **Смита-Ватермана**, см. ***matcher***, ***water*** из EMBOSS ……..	*Динамическое программирование* ***Carillo& Lipman****, см* ***MSA*** **Эвристические алгоритмы прогрессивного выравнивания, см.** **ClustalX**, ***emma в EMBOSS***, ***muscle***, ***T-Coffee*** , …….	***Dialign, ProDA***

Программы для поиска гомологии

BLAST (*Basic **L**ocal **A**lignment **S**earch **T**ool* — средство поиска основного локального выравнивания) — семейство компьютерных программ, служащих для поиска сходных аминокислотных или нуклеотидных последовательностей)

PSI-BLAST,
HMMer,
смита-Ватермана.

Все они строят выравнивание (alignment) последовательностей, добиваясь наибольшего сходства между ними.

Но увеличивая сходства часто приходится «разрывать» последовательности *(обозначено «–» на рис).*

Потом программа оценивает сходство выровненных последовательностей, и сообщает:

(1) гомологичны ли они (т. е. связаны ли они генетическим родством)

(2) как выглядит **наилучшее выравнивание** этих последовательностей.

Правильно установить родство последовательностей можно и при невысоком их сходстве, но — внимание! — при этом часто не удается установить их правильное (вытекающее из сопоставления их пространственных структур)

Критерии качества выравнивания

- ✓ **Количество идентичных (похожих) аминокислот/нуклеотидов**
  - **Для белков – более 25% id при длине > 100 aa**
  - **Для ДНК – более 70% id при длине > 100 nt**
- ✓ **Длина выравнивания**
- ✓ **Вероятность наблюдать такое сходство случайным образом (Зависит от базы данных)**
- ✓ **Score – общая мера сходства (Зависит от программы)**

BLAST

- ✓ **Локальное выравнивание**
- ✓ **Главная задача – поиск похожих последовательностей в базах данных** *(=> главное достоинство – скорость)*
- ✓ **Очень неточно восстанавливает сходство**
- ✓ **Основная программа поиска по БД**
- ✓ **Для специализированных БД часто предлагается на сайте БД**
- ✓ **Для поиска среди известных последовательностей есть специальные сервера**

Родной BLAST – NCBI

(http://www.ncbi.nlm.nih.gov/blast/Blast.cgi)

Что выбрать?

Программа	Query	Тип БД	Сравнивает
Blastn	ДНК	ДНК	ДНК
Blastp	белок	белок	белки
Blastx	ДНК	белок	белки
Tblastn	белок	ДНК	белки
Tblastx	ДНК	ДНК	белки

Дополнительные программы

- ✓ **ДНК:**
  - **megaBLAST – другой алгоритм для сравнения ДНК. Оптимизирован для длинных похожих последовательностей. Оптимален для поиска хитов в родном геноме или очень близких видах**
  - **Discontiguous megaBLAST – аналогично, параметры подобраны для более далеких видов**

- ✓ **Белок:**
  - **PSI-BLAST (Position-Specific Iterated -BLAST) поиск удаленных белковых гомологов с использованием PSSM (position-specific scoring matrix)**
  - **PHI-BLAST (Pattern-Hit Initiated -BLAST) ищет гомологичные белки, удовлетворяющие заданному паттерну**

Средство поиска сходства - выравнивание

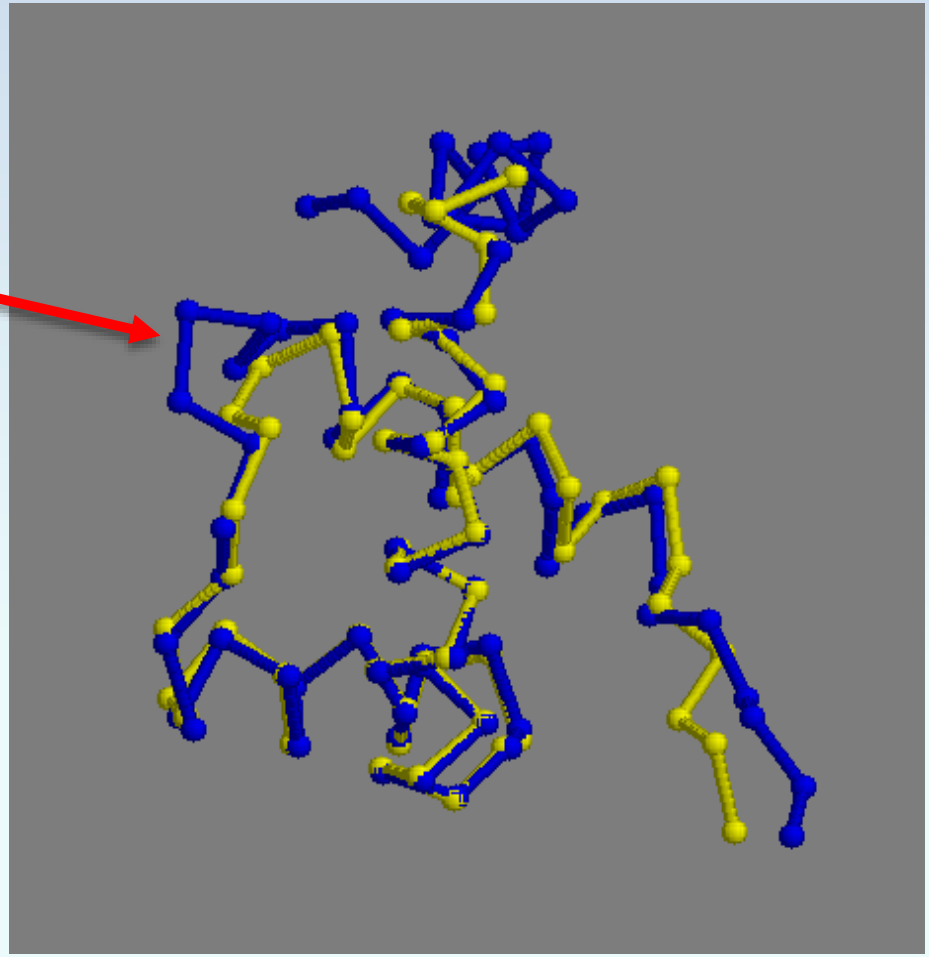
«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

домовой
скупидом
водомерка

лесовоз	---лесо---воз
ледоход	лед---оход---

Схожие 3D структуры

Вставка в «синей» последовательности



Как выровнять 2 последовательности?

Цель - максимальное количество совпадений

- Просто написать их друг под другом
- Двигать друг относительно друга
- Вставлять пробелы
- Что лучше?

Гэп – пропуск в последовательности

лесовоз	---лесо---воз
ледоход	лед---оход---

Матрицы замен

Матрица 20*20 на пересечении 2х аа их уровень сходства (?):

- ***Похожесть по свойствам (объем, гидрофильность, заряд и т.д.)***
- ***Эволюционное родство – частота замен 1ой аа на другую в изученных белках***

2 сорта последних:

PAM (Point Accepted Mutations) – на выравниваниях очень близких белков (PAM20 = PAM^20)

BLOSUM (BLOck Scoring Matrix) – на блоках выравниваний далеких белков (без делеций) (BLOSUM62 – на белках со средним уровнем сходства 62% попарно)

Делеции / инсерции

- ✓ Общий штраф
- ✓ Значительно чаще 1 длинная делеция, чем много коротких => штраф за внесение делеции + штраф за удлинение делеции

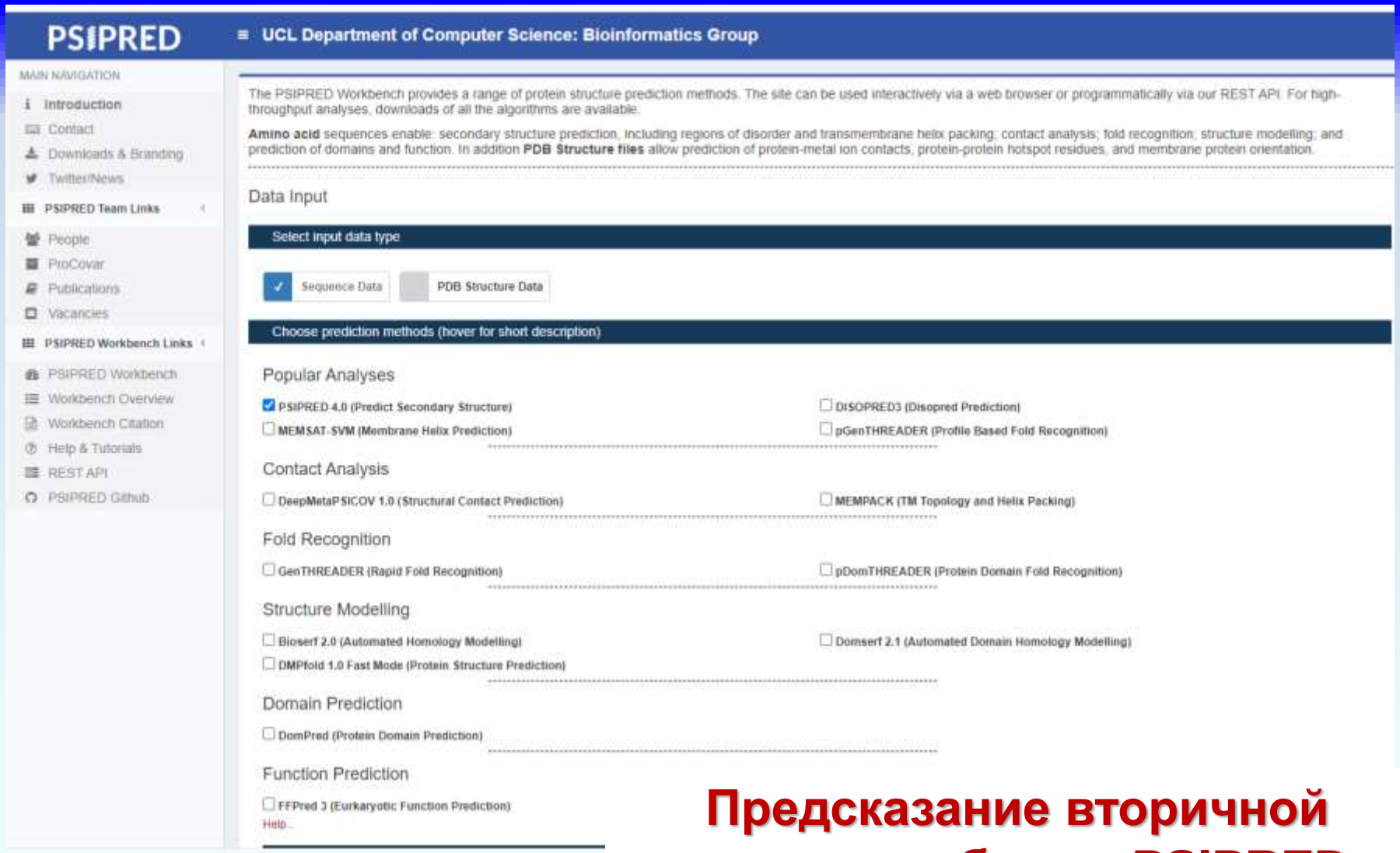
Предсказание пространственной структуры белка

- *Ab initio* - моделирование укладки “из первых принципов” - без использования дополнительной информации о структурах схожих белков.

- Предсказание на основе гомологии (homology modeling) - моделирование на основе известных структур схожих белков.

- Тридинг (Threading) - моделирование на основе слабой гомологии.

Алгоритмы предсказания вторичной структуры белка можно условно разделить на группы, основываясь на принципах, лежащих в их основе.
Эти группы включают в себя статистические методы, методы ближайших соседей, методы, использующие нейронные сети, методы опорных векторов и методы, основанные на скрытых марковских моделях



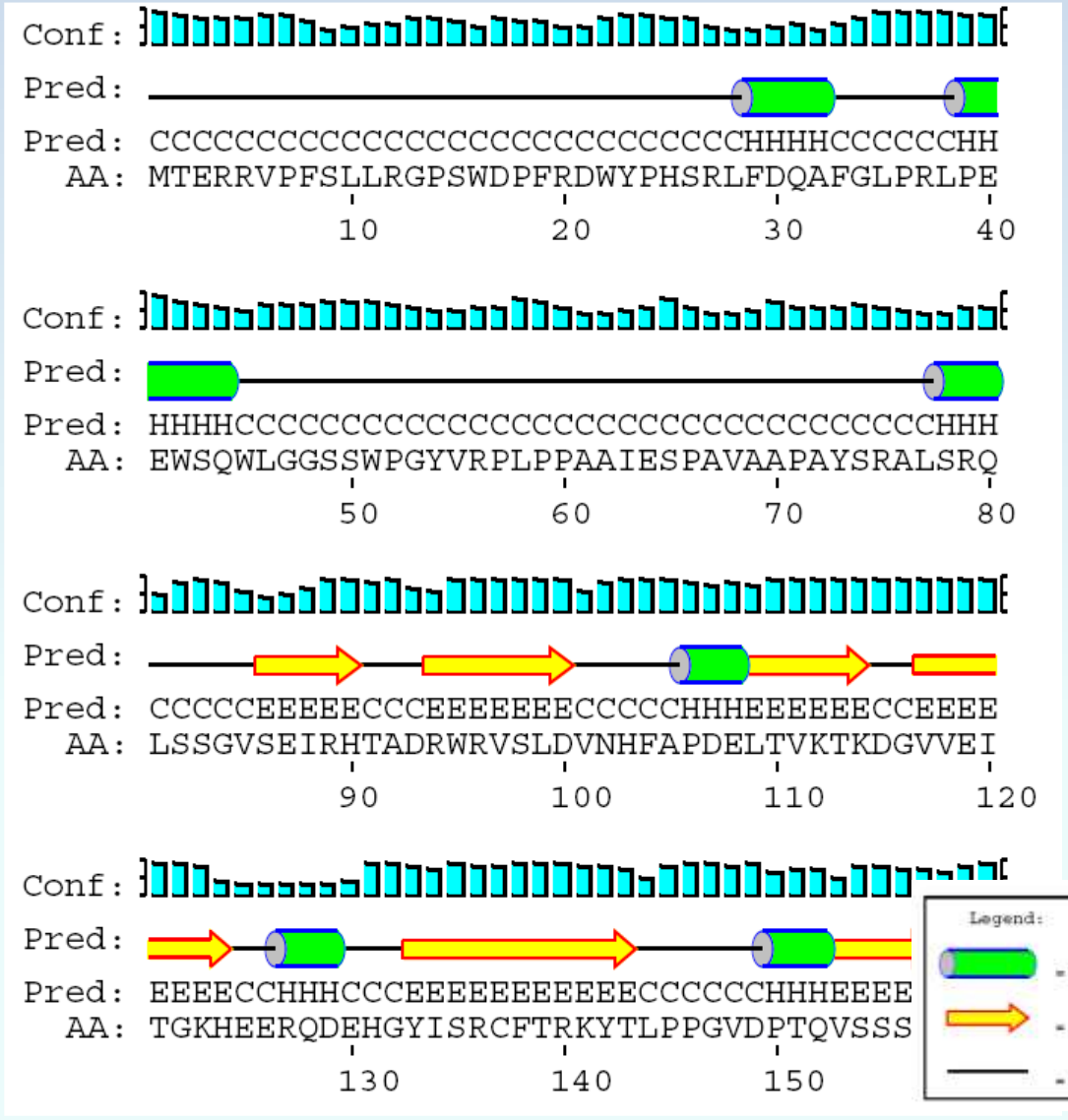
Предсказание вторичной структуры белка: PSIPRED

http://bioinf.cs.ucl.ac.uk/psipred/

PSIPRED - output

```
PSIPRED PREDICTION RESULTS

Key

Conf: Confidence (0=low, 9=high)
Pred: Predicted secondary structure (H=helix, E=strand, C=coil)
  AA: Target sequence


# PSIPRED HFORMAT (PSIPRED V2.5 by David Jones)

Conf: 9877687422446642653115886520022127888987854214446664322337 63
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCCHHHHHHCCCCCCCCCCCCCCCC
  AA: MTERRVPFSLLRGPSWDPFRDWYPHSRLFDQAFGLPRLPEEWSQWLGGSSWPGYVRPLPP
              10        20        30        40        50        60


Conf: 000373112633342100220898312588795399999827998756599998998999
Pred: CCCCCCCCCCCCCCCCCHHHCCCCCEEEEECCCEEEEEEECCCCCHHHEEEEEECCEEEE
  AA: AAIESPAVAAPAYSRALSRQLSSGVSEIRHTADRWRVSLDVNHFAPDELTVKTKDGVVEI
              70        80        90       100       110       120


Conf: 998010000188758889998762889885762876657985999971578875447427
Pred: EEEECCHHHCCCEEEEEEEEEEECCCCCCCHHHEEEEECCCCEEEEEEECCCCCCCCCCEE
  AA: TGKHEERQDEHGYISRCFTRKYTLPPGVDPTQVSSSLSPEGTLTVEAPMPKLATQSNEIT
             130       140       150       160       170       180


Conf: 75741487544676644343120 39
Pred: EEEEECCCCCCCCCCCCCCCHHHCC
  AA: IPVTFESRAQLGGPEAAKSDETAAK
             190       200


Calculate PostScript, PDF and JPEG graphical output for this result using:
http://bioinf2.cs.ucl.ac.uk/cgi-bin/psipred/gra/nph-view2.cgi?id=05471106883edf08.psi
```

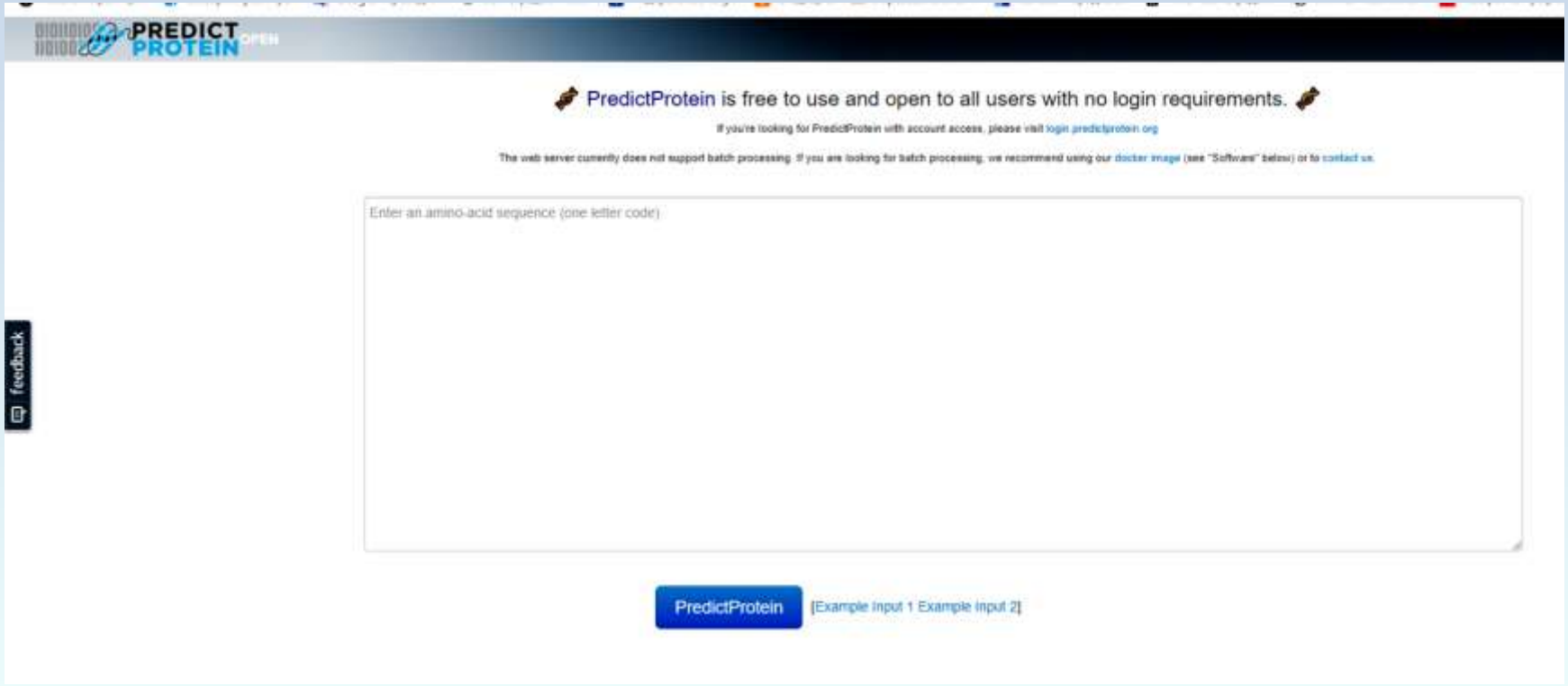
PSIPRED – графический выход



PredictProtein

- Предсказывает:
- ✓ Вторичную структуру (H, E, C)
- ✓ Для каждого остатка – доступность для растворителей
- ✓ Трансмембранные сегменты и их топологию
- ✓ Глобулярные участки белка
- ✓ Coiled coil участки
- ✓ PROSITE мотивы в белке
- ✓ Prodom домены
- ✓ Дисульфидные связи
- ✓ Участки с неравномерным а.к.-составом
- Запускает META server для исследуемого белка
- Требует регистрации

PredictProtein



- http://www.predictprotein.org/

Evaluation of secondary structure prediction

EVA: http://pdg.cnb.uam.es/eva/

- ✓ сравнивает различные серверы по предсказанию вторичной структуры
- ✓ часто обновляемый список действующих серверов

Evaluation of secondary structure prediction

Version
May 25, 2001

email

OBJECTIVES:

EVA continously and automatically analyses protein structure prediction servers in 'real time' (more details)

RESULTS:

- PDB statistics

- secondary structure
- comparative modelling
- inter-residue distances and contacts
- threading

- FTP archives

http://pdg.cnb.uam.es/eva/

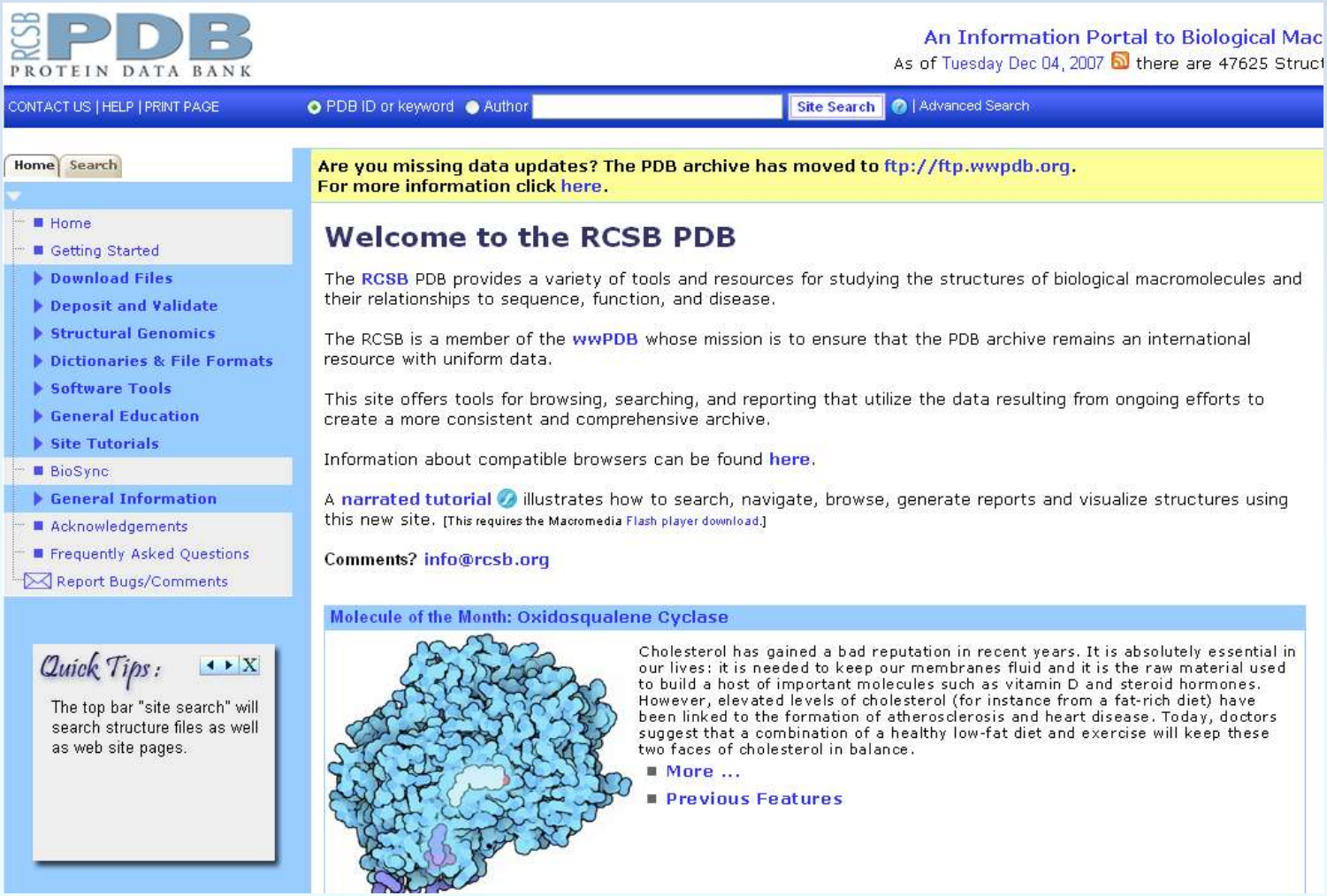
INFORMATION:

- EVA **flow chart**
- EVA **concept**
- Structure prediction **servers** participating
- Related resources (CAFASP results)

CONTACT:

- EVA eva@cubic.bioc.columbia.edu
- EVA team

PDB – универсальный репозиторий данных по пространственной структуре белка



PDB – стандартная запись

Help | Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

1gme

DOI 10.2210/pdb1gme/pdb

Red - Derived Information

Title	CRYSTAL STRUCTURE AND ASSEMBLY OF AN EUKARYOTIC SMALL HEAT SHOCK PROTEIN
Authors	Van Montfort, R.L.M., Basha, E., Friedrich, K.L., Slingsby, C., Vierling, E.
Primary Citation	**van Montfort, R.L., Basha, E., Friedrich, K.L., Slingsby, C., Vierling, E.** (2001) Crystal structure and assembly of a eukaryotic small heat shock protein. *Nat.Struct.Biol.* **8:** *1025-1030* [Abstract]
History	Deposition 2001-09-13 Release 2001-11-29
Experimental Method	Type X-RAY DIFFRACTION Data [EDS]

Parameters	Resolution[Å]	R-Value	R-Free	Space Group
	2.70	0.231 (obs.)	0.286	H 3 2

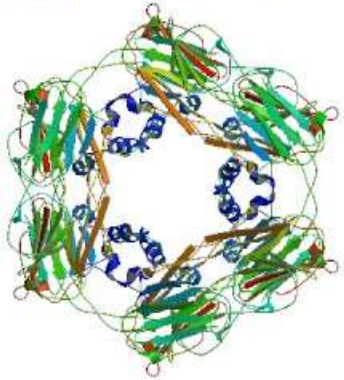
Unit Cell						
Length [Å]	a	171.65	b	171.65	c	124.16
Angles [°]	alpha	90.00	beta	90.00	gamma	120.00

Molecular Description Asymmetric Unit	Polymer: 1 Molecule: HEAT SHOCK PROTEIN 16.9B Chains: A,B,C,D
Classification	**Chaperone**
Source	Polymer: 1 Scientific Name: **Triticum aestivum** Common Name: **Wheat** Expression system: **Escherichia coli**

SCOP Domain Info Class Fold Superfamily Family Domain Species

Images and Visualization

<< Biological Molecule >>



Display Options

KiNG
Jmol
WebMol
MBT SimpleViewer*
MBT Protein Workshop
QuickPDB
All Images

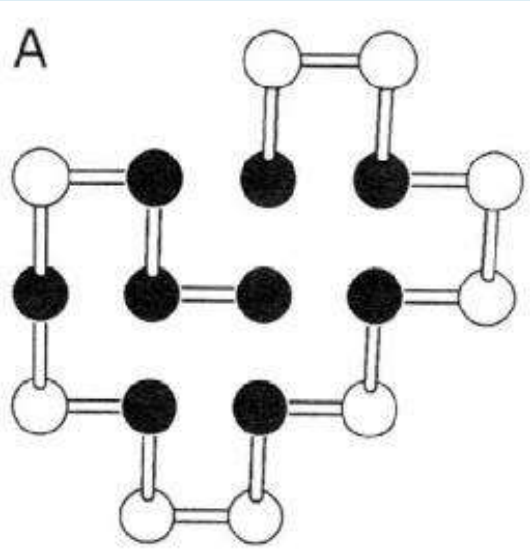
* Capable of displaying biological molecules.

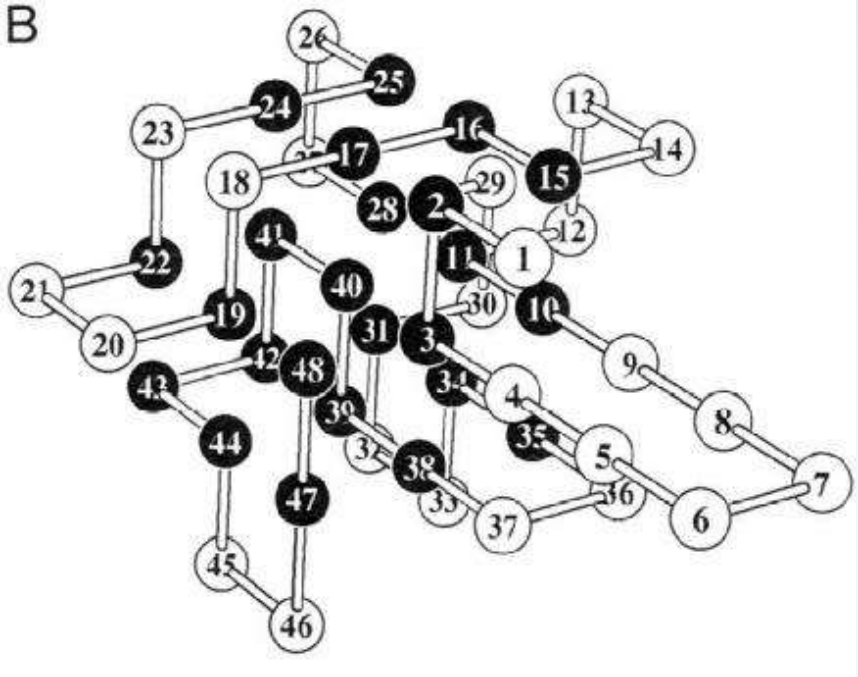
Предсказание структуры *ab initio*

- Функция потенциальной энергии
  - Модель водного раствора
  - Оценка попарного взаимодействия между аминокислотами
- Поиск в пространстве всевозможных конформаций
  - Модель на основе “решетки”
  - Молекулярная динамика
  - Использование библиотек известных 3D фрагментов
- Предсказание вторичной структуры

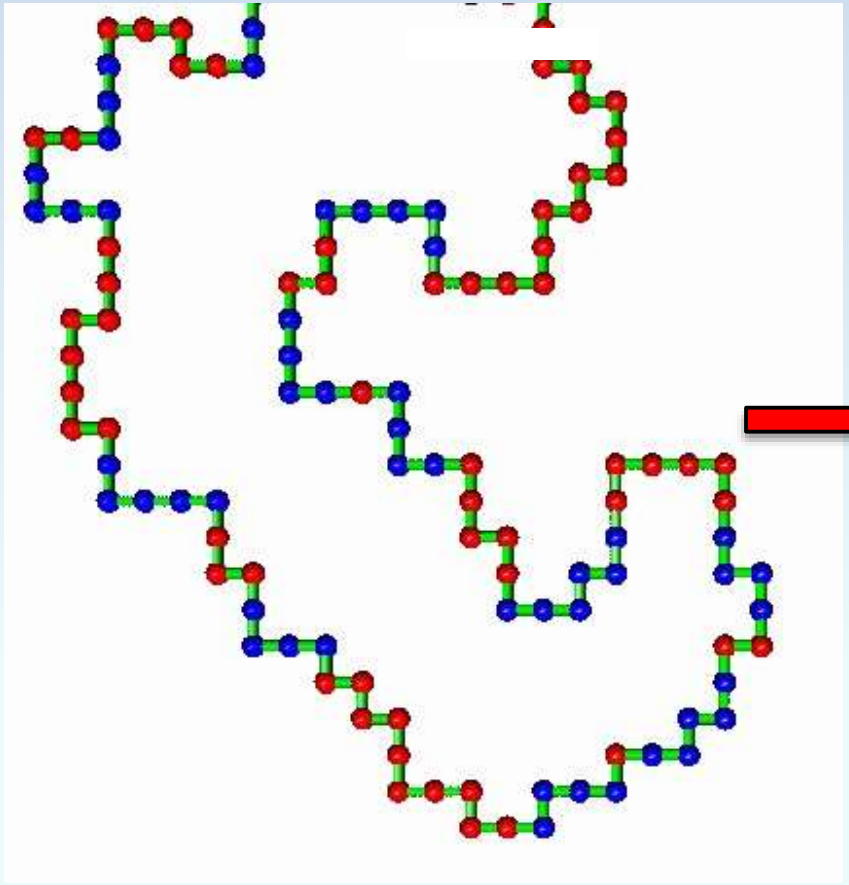
Предсказание структуры с использованием решетки

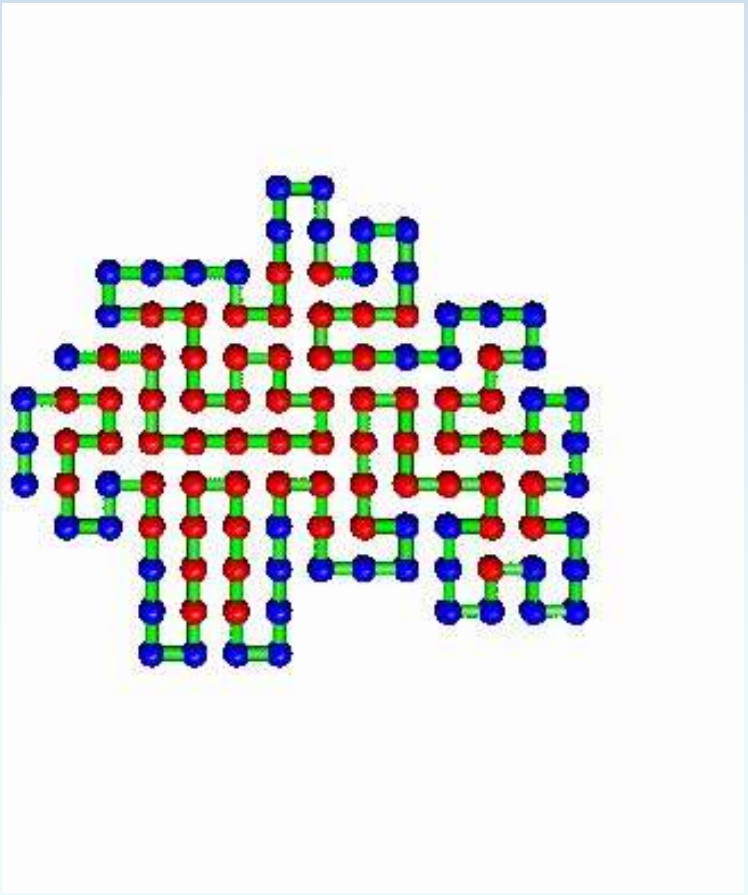
- **HP-модель (Hydrophobic-Polar)** - рассматривает гидрофобные взаимодействия как наиболее важные.
  - Не существует эффективных алгоритмов
  - Плохо отражает реальность





Предсказание структуры с использованием решетки





ROSETTA

- Используются структурно консервативные фрагменты длиной 4-10 аминокислот
- Поиск в пространстве конформаций осуществляется методом Монте Карло
- Полученные структуры кластеризуются и в качестве результата выдаются наилучшие структуры для каждого кластера

Предсказание структуры на основе гомологии

- Выравнивание рассматриваемой последовательности с последовательностями белков с известной 3D структурой (обычно >30% сходства)
- Наложение моделируемой последовательности на известную структуру согласно выравниванию
- Локальное улучшение полученной пространственной структуры

- Число уникальных укладок (фолдов), наблюдающихся в белках, ограничено (несколько тысяч)
- 90% помещаемых в PDB структур имеют уже известные укладки (фолды)

Примеры укладок

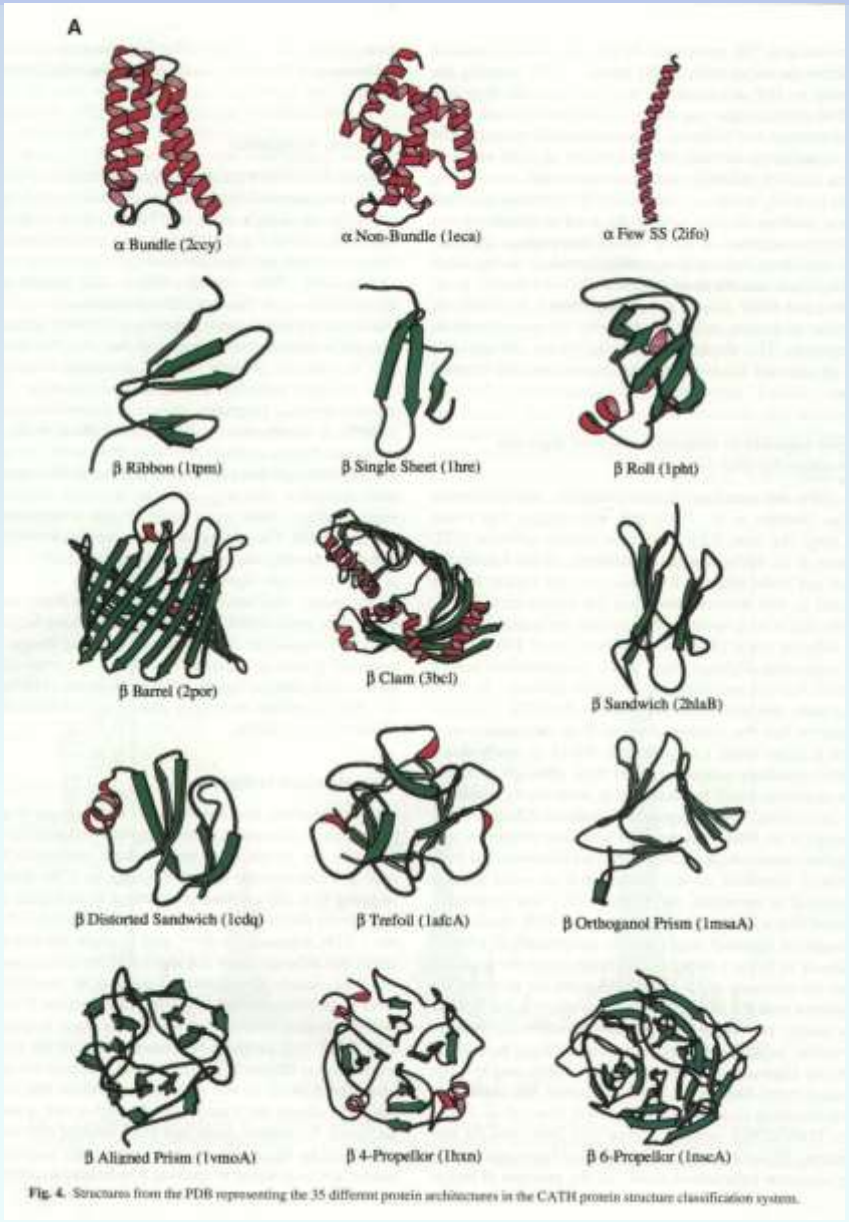
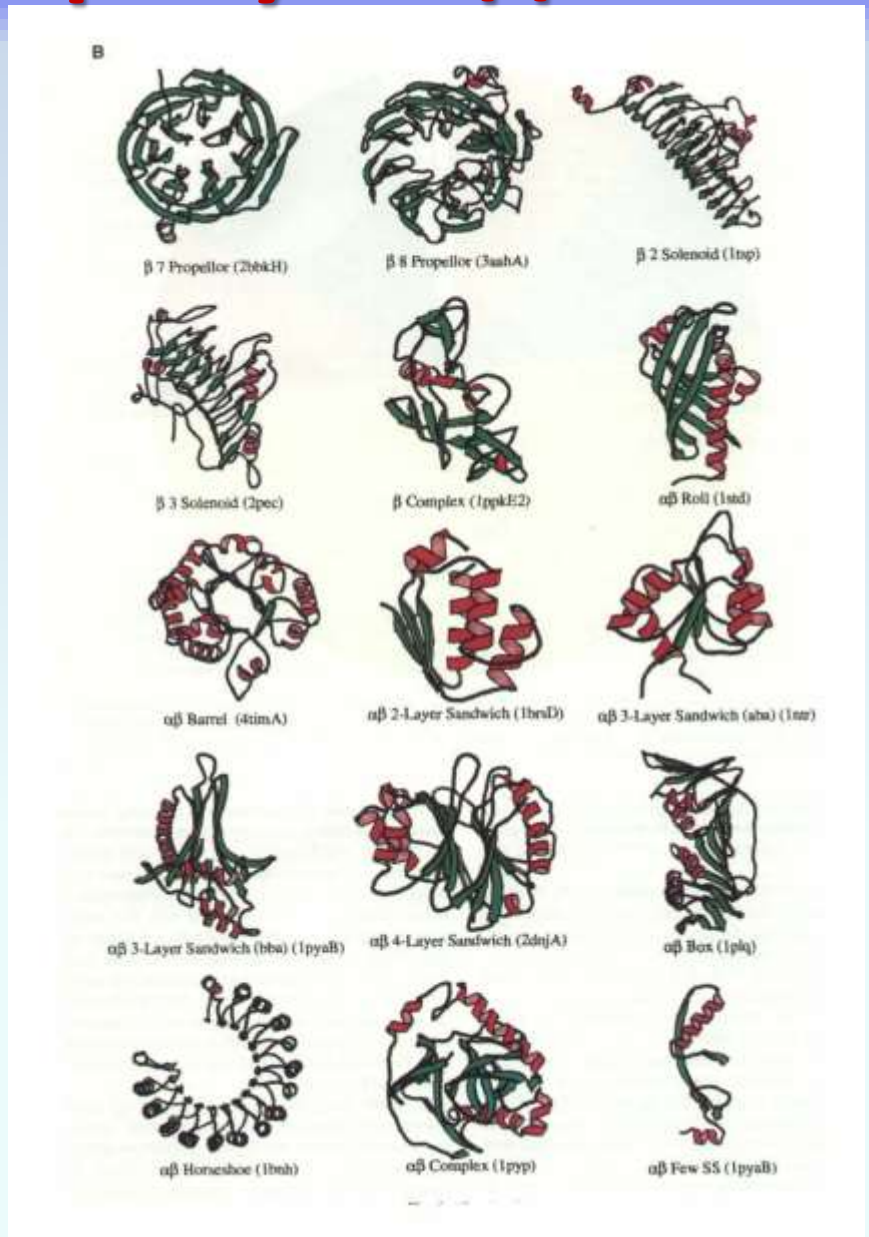
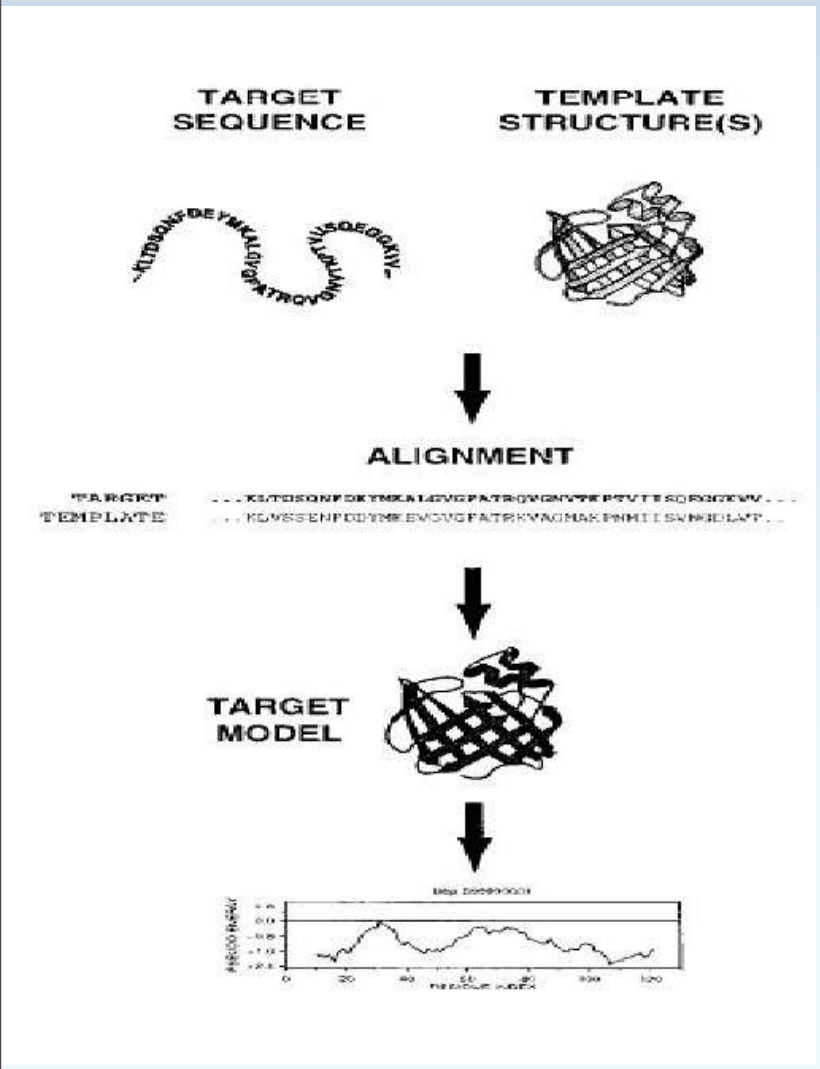
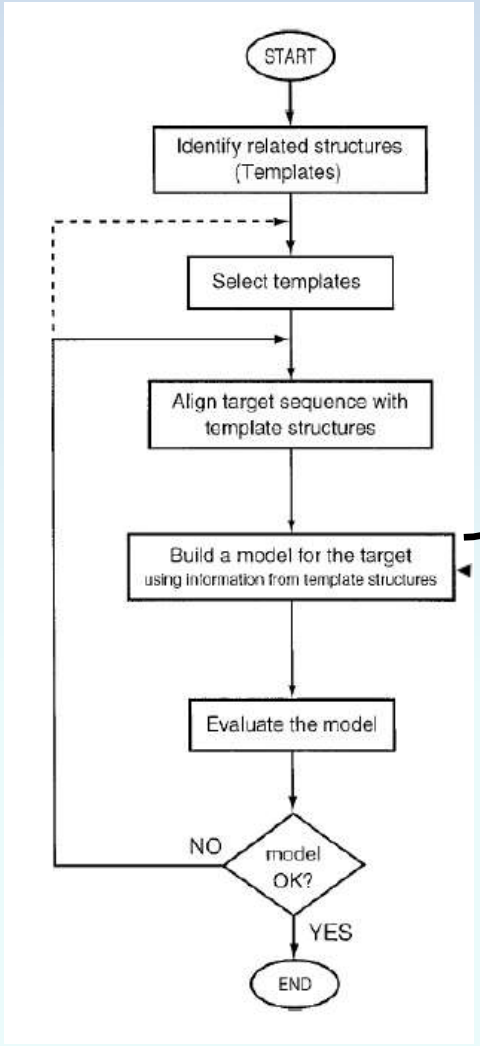


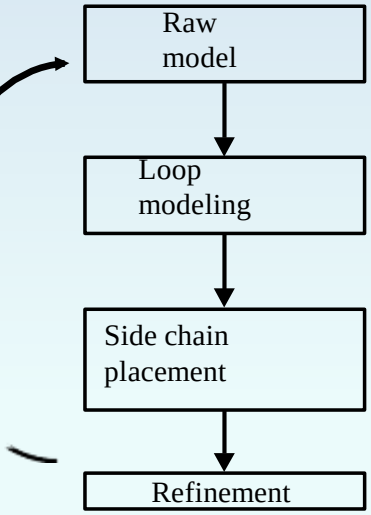
Fig. 4. Structures from the PDB representing the 35 different protein architectures in the CATH protein structure classification system.



Предсказание структуры на основе гомологии

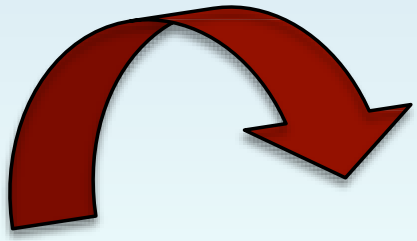




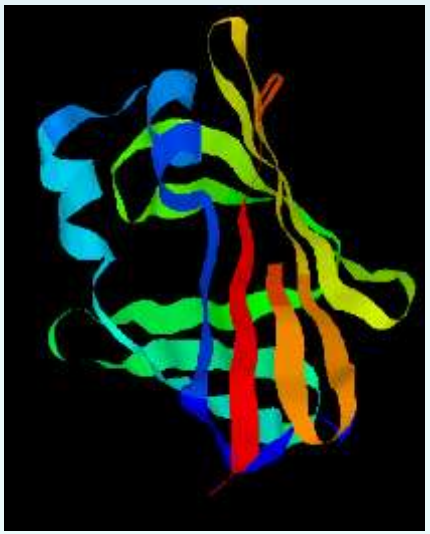


Тридинг (Threading) - предсказание структуры на основе слабой гомологии

- Главное отличие от моделирования по гомологии - поиск наилучшей структуры осуществляется с помощью выравнивания последовательности со структурой, а не с последовательностью. При этом используется специальным образом определенная весовая функция.



MTYKLILN ….
NGVDGEWTYTE

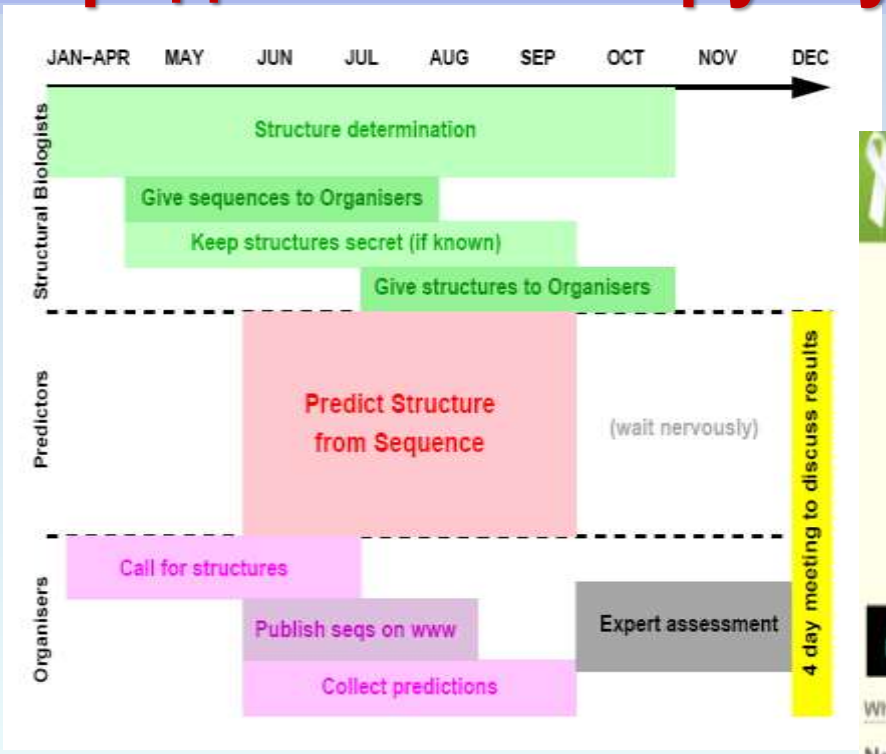


Основные компоненты тридинга

- библиотека уникальных укладок (фолдов)
- функция, определяющая вес выравнивания последовательности со структурой
- алгоритм нахождения наилучшего выравнивания

CASP - конкурс методов предсказания структуры белков

FoldIt



Critical Assessment of protein Structure Prediction, CASP

Предсказание геометрии боковых радикалов

Точное предсказание расположения боковых аминокислотных радикалов в структуре представляет собой отдельную проблему в прогнозировании структуры белка.

Методы, которые решают проблему прогнозирования геометрии боковых радикалов, включают в себя устранение тупиков и методы самосогласованного поля.

Конформации боковых радикалов с низкой энергией обычно определяются на жёстком полипептидном остове и используют набор дискретных конформаций боковой цепи, «ротамеров». Принцип работы таких методов заключается в поиске набора ротамеров, минимизирующего общую энергию модели.

Гомологическое моделирование третичной структуры белка на основе первичной структуры

Стратегия построения пространственной структуры белков методом моделирования по гомологиям:

- Определения круга гомологичных белков;
- Нахождение структурно-консервативных элементов в структуре гомологов (SCRs);
- Выравнивание последовательности модельного белка с последовательностями гомологов, с учѐтом наличия SCR;
- Присвоение координат атомов остатков, входящих в SCR, соответствующим атомам модельного белка согласно выравниванию;
- Предсказание конформации петель, соединяющих SCR, а также N- и C-концов пептидной цепи белка;
- Поиск оптимальной конформации боковых остатков аминокислот модельного

  белка, отличающихся от остатков опорного белка;
- Использование методов регуляризации структуры (энергетическая минимизация и молекулярная динамика) для уточнения молекулярной структуры с целью устранения стерических напряжений созданных при построении моделей.

Присвоение координат атомов

В первую очередь присваиваются координаты атомам полипептидной цепи.

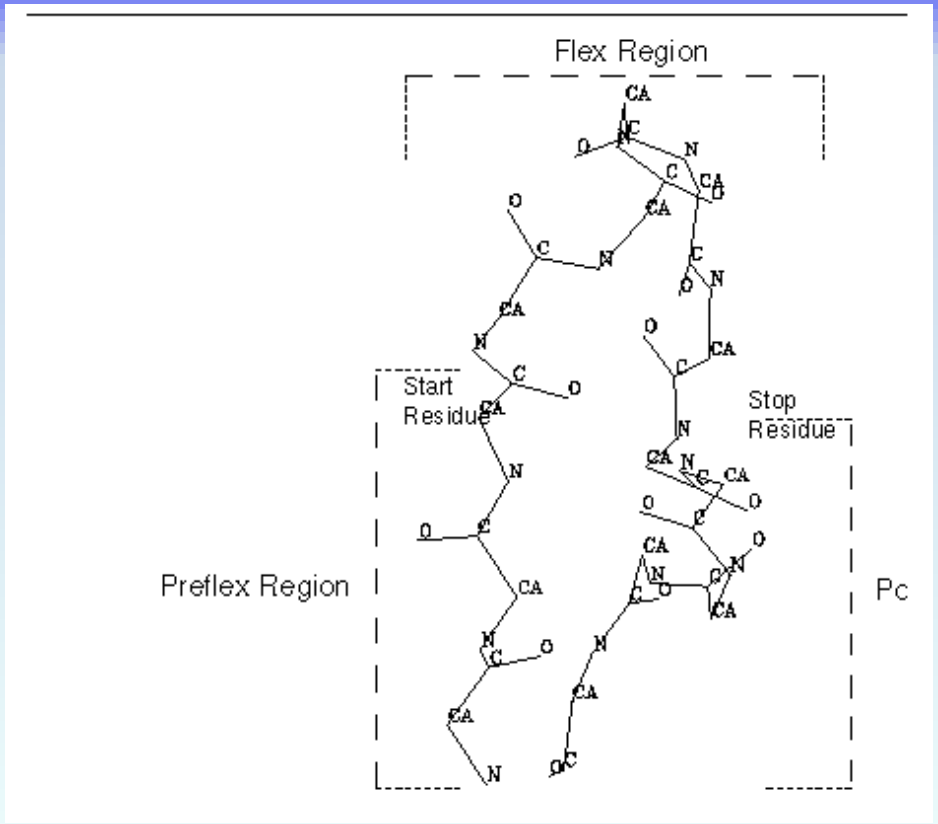
Затем присваиваются координаты атомам боковых цепей.

Благоприятный случай, когда аминокислота модельного белка совпадает с соответствующей кислотой белка- гомолога. В этом случае конформация боковой цепи остаётся неизменной. Если боковая цепь аминокислоты модельного белка короче, чем соответствующая цепь аминокислоты гомолога, более короткая цепь повторяет насколько это возможно более длинную

(торсионные углы χ одинаковы).

Если же аминокислота модельного белка более длинная, то начальный ход повторяет ход боковой цепи в белке-гомологе, а последующие атомы цепи помещаются в развѐрнутую (extended) конформацию, вероятно вызывая сильные напряжения в структуре модельного белка.

Поиск конформации соединяющих петель



После того, как присвоены координаты атомам, составляющим петли, мы имеем модельную структуру, которая нуждается в приведении её в соответствие со следующими требованиями:

- ***Геометрия пептидной цепи*** модельной структуры должна быть регулярной (транс- конформация пептидных групп, близкие к равновесным значения валентных углов и дли связей);
- ***Атомы не должны перекрываться***, т.е. расстояния между несвязанными атомами не должны быть существенно меньше, чем сумма их ван-дер-ваальсовских радиусов;
- Боковые цепи аминокислот должны находиться в равновесной конфигурации;

-Если в молекуле имеются дисульфидные мостики (Cys-Cys связи), то расстояния между соответствующими атомами серы должны быть приведены в соответствие с геометрией;

- В структуру должны быть помещены необходимые простетические группы.

Построение пространственной структуры D-amino-acid oxidase из *Trigonopsis variabilis (Yeast)*

В качестве опорного белка была использована пространственная структура D-Amino Acid Oxidase из *Rhodotorula gracilis* (PDB идентификатор 1C0L)

Типичная процедура регуляризации модельной структуры белка

1. Энергетическая минимизация участков сочленения SCR и петель с упором на восстановление нормальной пептидных связей;
2. Энергетическая минимизация пептидной цепи и боковых остатков петель;
3. Энергетическая минимизация боковых цепей аминокислот, принадлежащих SCR, подвергшихся замене при присваивании координат;
4. Энергетическая минимизация всех боковых остатков белка;
5. Энергетическая минимизация (500-1000 шагов) всей структуры модельного белка;
6. Молекулярная динамика модельного белка в вакууме на протяжении 20-50 пикосекунд;
7. Финальная энергетическая минимизация структуры белка (200-500 шагов).

Типичная процедура регуляризации модельной структуры белка

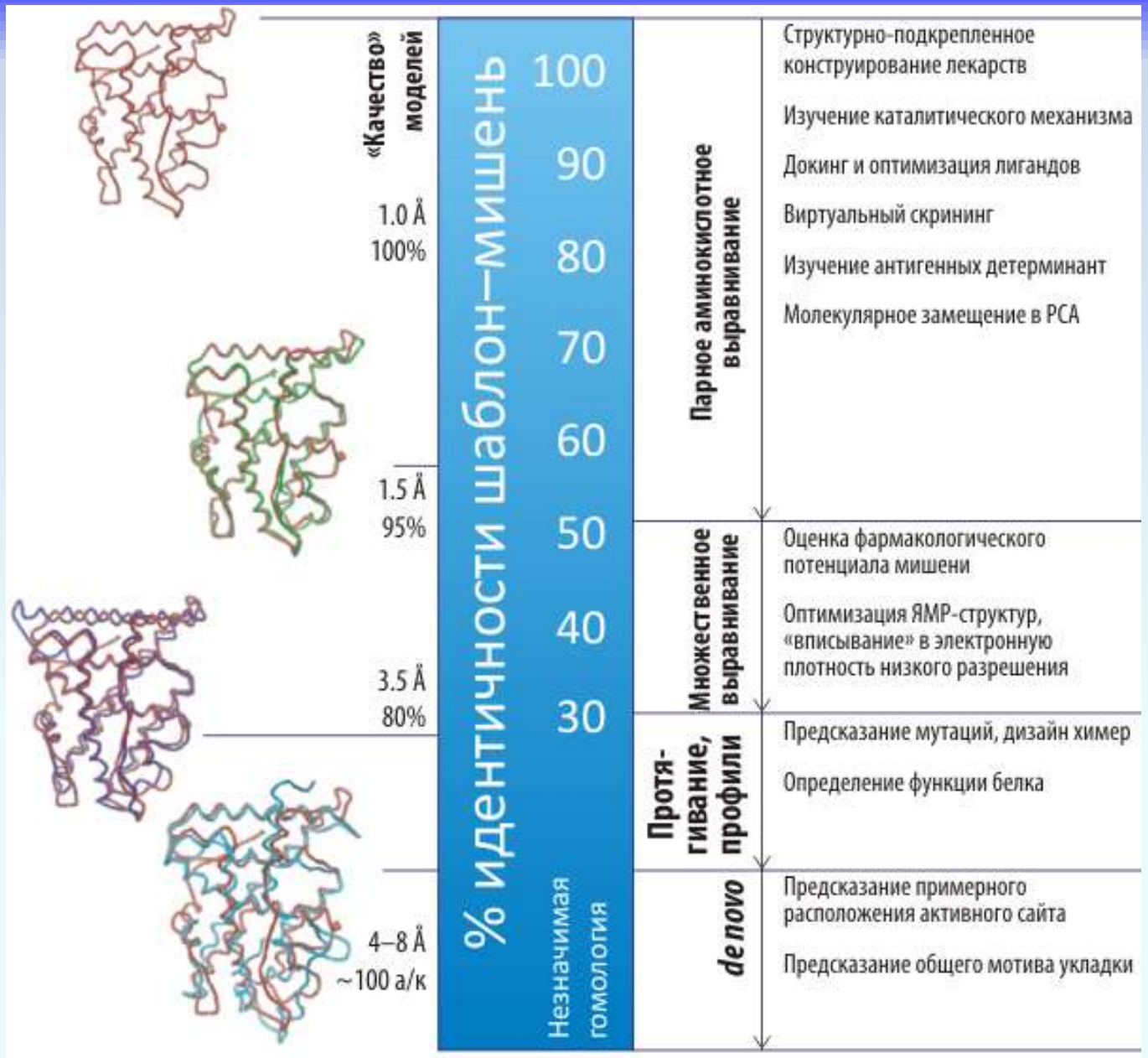
Результат этой процедуры - белковая структура с правильной стереохимией (длины валентных связей и значения валентных углов не будут существенно отличаться от равновесных значений), с отрицательной энергией **несвязанных** взаимодействий (свидетельство того, что не наблюдается перекрытие ван-дер-ваальсовских радиусов атомов), с отрицательной энергией **электростатических** взаимодействий (произошло сближение противоположно заряженных атомов) и с **ненулевой энергией водородных** связей (в молекуле установились водородные связи).

Дальнейшая регуляризация структуры приведѐт к еѐ улучшению с точки зрения стереохимии, но при этом возрастут искажения структуры активного центра (центра связывания) вашей структуры.

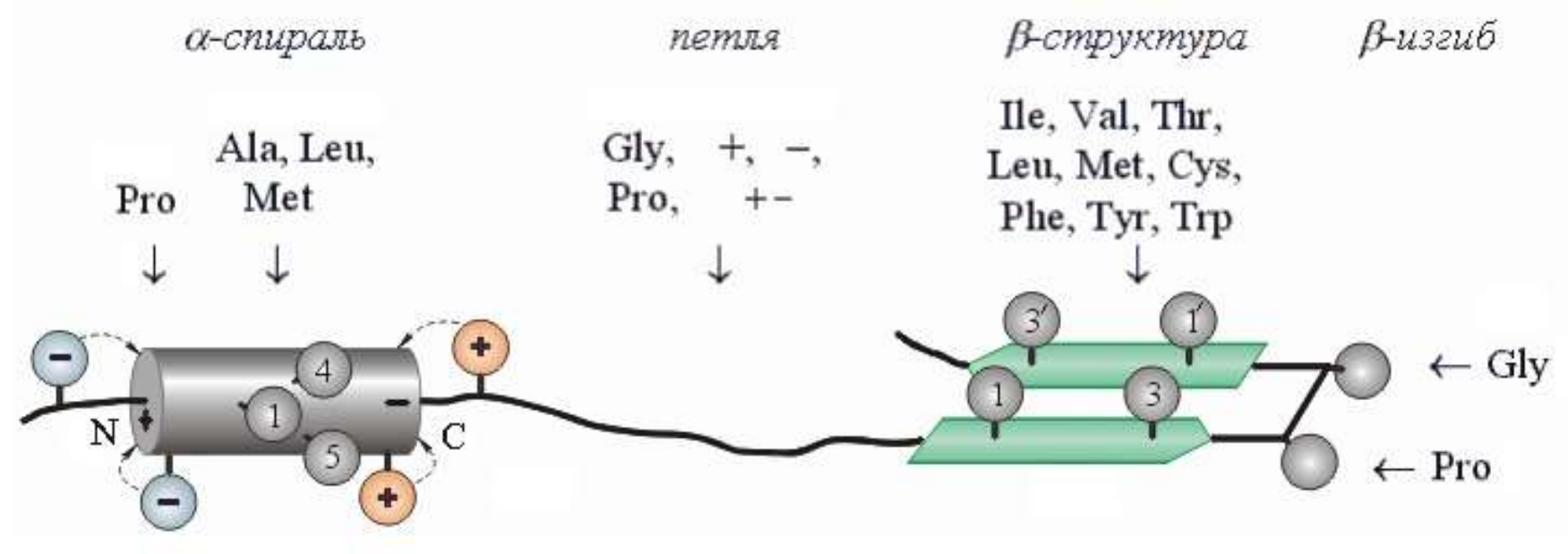
Модельная структура построена и отрелаксирована. Она обладает участками структурно- консервативных областей, унаследованных от белков гомологов, правильной стереохимией (результат регуляризации). Дальнейшие манипуляции с этой структурой (подгонка геометрии активного центра, точечные мутации) зависят от цели исследований.

Полученную структуру надо рассматривать как средство иллюстрации результатов вашей работы (объяснения экспериментальных фактов, гипотезы).

Качество и сфера пригодности компьютерных моделей белков, основанных на различной степени гомологии



«Шаблоны» для опознавания α-спирали, петли, β-структуры и β-изгиба



Выделены остатки, которые стабилизируют
«+» означает все положительно заряженные аминокислоты,
«–» все отрицательно заряженные,
«+ –» все аминокислоты с диполем в боковой цепи.

Зная, какие аминокислотные остатки стабилизируют середину спирали, какие — ее N-конец, а какие — С-конец, мы получаем нечто вроде «шаблона» спирали.

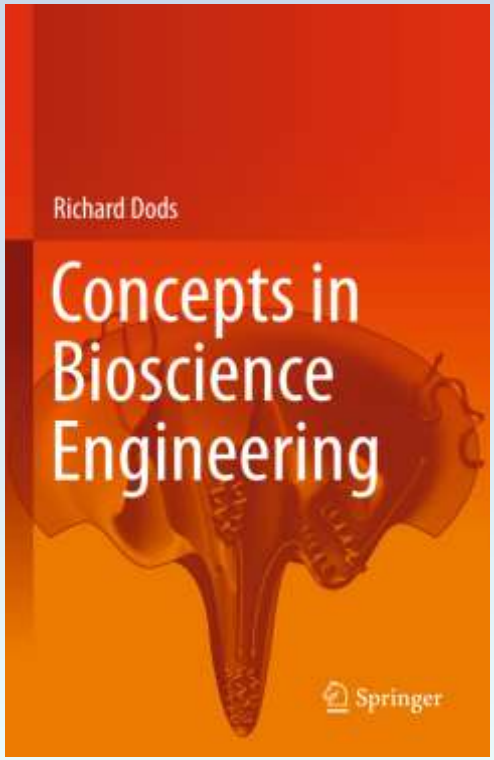
«шаблон» качественно описывает аминокислотную последовательность, подходящую для образования α-спирали. Чем лучше аминокислотная последовательность удовлетворяет этому шаблону, тем вероятнее спираль в данном месте цепи.

такое же описание для образования β-структуры, отдельных участков цепи или образующих более сложные структуры β-α-β-суперспиралей,

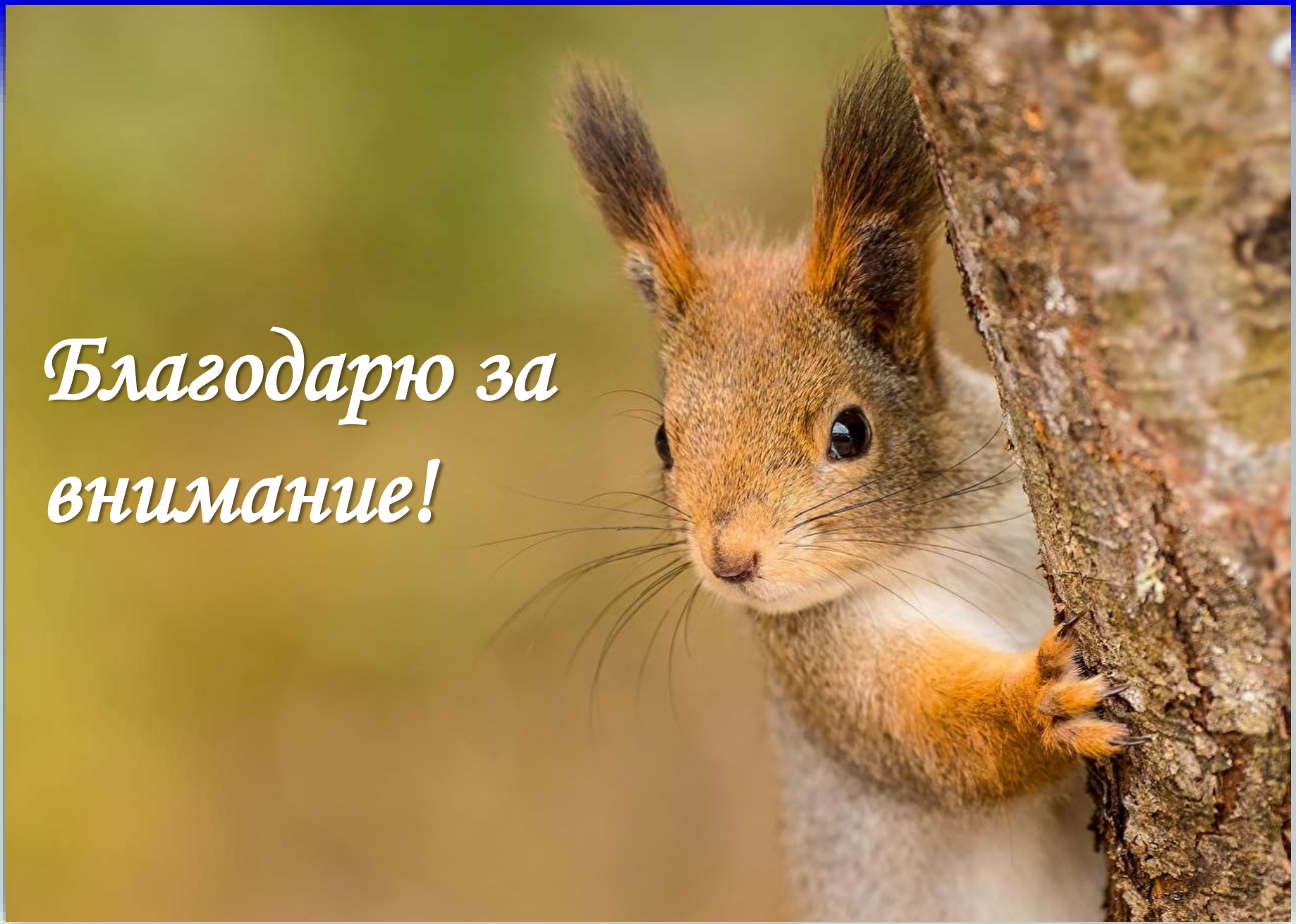
Особую роль в «шаблоне» играют **«ключевые позиции»,** которые могут быть заняты только **строго определенными аминокислотными** остатками,
например, Gly: только этот остаток может находиться в конформации с $\phi \approx 60°$, недоступной всем остальным остаткам.

«шаблоны» могут содержать, помимо структурной, и функциональную информацию

Рекомендуем – 2020 г.



https://doi.org/10.1007/978-3-030-28303-2


Благодарю за
внимание!