

*Предсказание и дизайн  
белковых структур*

## 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

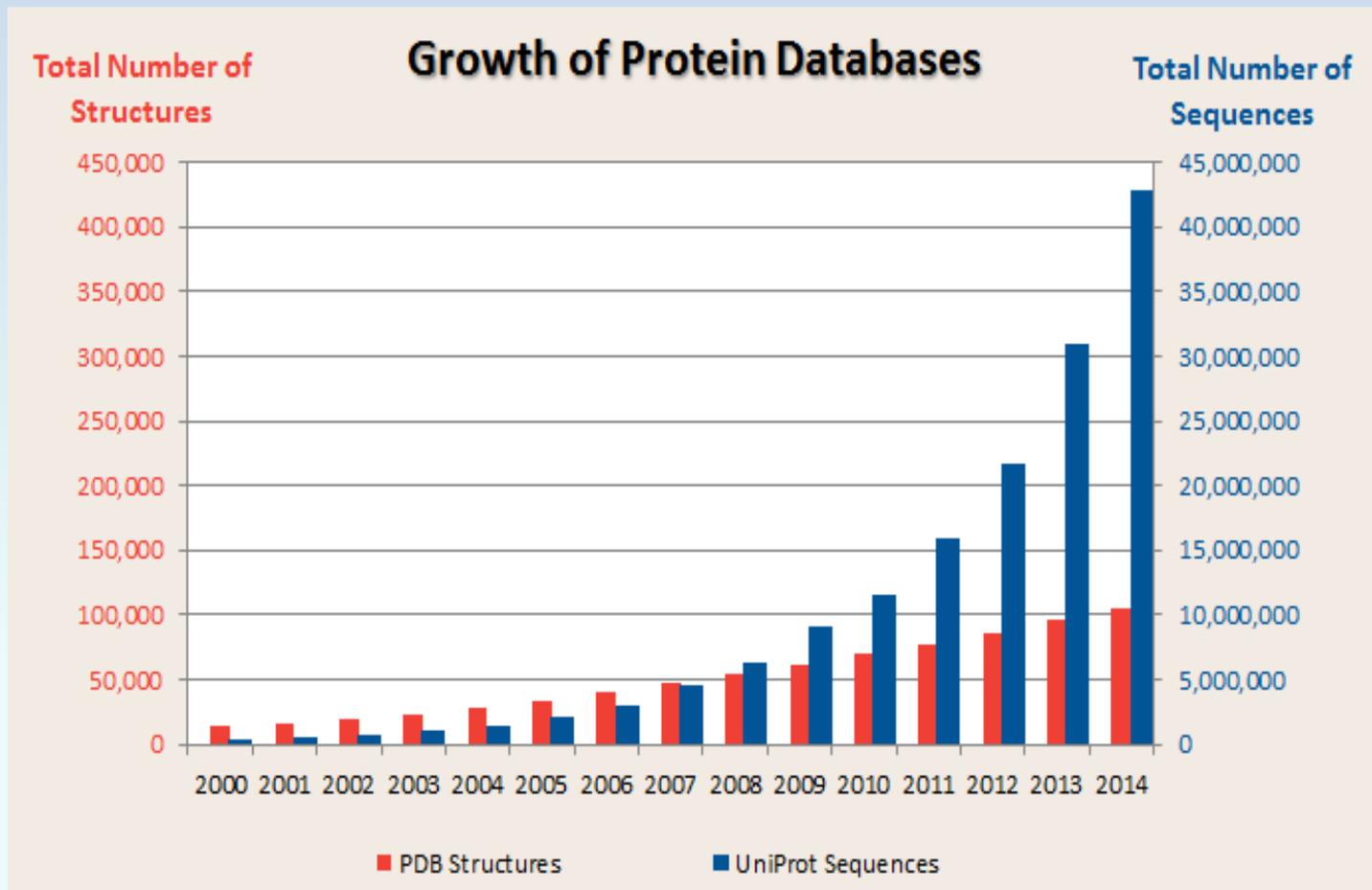
+Монте-Карло

+Марковские цепи - Скрытая марковская модель

[https://ru.wikipedia.org/wiki/Скрытая\\_марковская\\_модель/](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



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

	1	10	20
Human, chimpanzee	<u>GDVEKGKK</u>	<u>IFIMKCSQ</u>	<u>CHTV...</u>
Pig, bovine, sheep	<u>GDVEKGKK</u>	<u>IFVQKCAQ</u>	<u>CHTV...</u>
Chicken, turkey	<u>GDIVEKGKK</u>	<u>IVQKCSQ</u>	<u>CHTV...</u>
Puget sound dogfish	<u>GDVEKGKK</u>	<u>VFVQKCAQ</u>	<u>CHTV...</u>
Screw-worm fly	GVPA	<u>GDVEKGKK</u>	<u>IFVQRCAQ</u>
Rust fungus	GFED	<u>GDAKKGAR</u>	<u>IFKTRCAQ</u>
Rape, cauliflower	ASFDEAPP	<u>GNSKAGEK</u>	<u>IFKTKCAQ</u>

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

	○ ○ ●	◇◇◇ ○	○ ◇	●○ ○ ○ ○
<i>E.coli</i>	MLKQVEIF	<b>TDG</b> SCLGK	<b>---</b> PGPGGY	<b>G</b> AILRYRGRE <b>K</b> TFSAGYTRT <b>TNNRMELMAA</b> IVALEAL
Yeast	YNKSMNYV	<b>CDG</b> SSFGNGTSS	SRAGY <b>G</b> AYFEGAPE <b>EENISPLLSGAQ</b>	<b>TNNRAE</b> IEAVSEALKKI
MMLV	PDADHTWY	<b>TDG</b> SSLLQ	---EGQRKAGAAVTTETE	<b>E</b> VIWAKALPAGTSA <b>QRAEL</b> IALTQ <b>ALK</b> -M
RSV	PVPGPTV	<b>FD</b> ASSSTH	---KGVVWREGPRW	-- <b>E</b> IKEIADL <b>G</b> ---SVQQL <b>E</b> ARAVAMALL-L
HIV	IVGAET	<b>F</b> YVDGAANRE	---TKLGKAGYVTN-KGRQ	<b>K</b> VVPLTNT-- <b>TN</b> QKTELQAIYLALQD-
	=	== :: = :	:	= : : = = :: ==
	← β1 →	← β2 →	← β3 →	← α1 →

# Зачем нужно выравнивать аминокислотные последовательности?

**Данные:**

**Биологические задачи:**

**Общий подход к решению — оценка сходства последовательностей:**

## НОВАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ

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

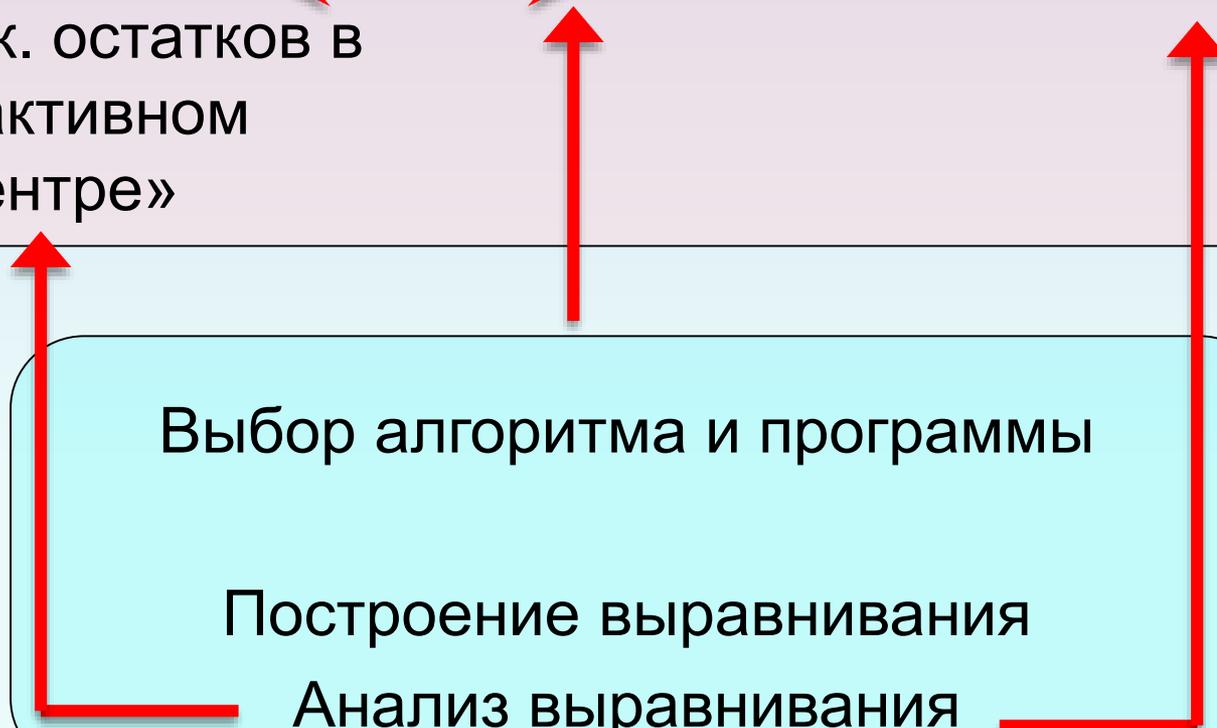
Предсказание 3D-структуры

Реконструкция эволюции

Выбор алгоритма и программы

Построение выравнивания

Анализ выравнивания



# Типы выравнивания аминокислотных последовательностей

## Выравнивания

### парные

### множественные

глобальные

локальные

глобальные

локальные

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*

**BLAST** (*Basic Local Alignment Search Tool* — средство поиска основного локального выравнивания) — семейство компьютерных программ, служащих для поиска сходных аминокислотных или нуклеотидных последовательностей)

**PSI-BLAST,  
HMMer,  
смига-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

### Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

BLAST+ 2.13.0 is here!  
Starting with this release, we are including the blastn\_vdb and tblastn\_vdb executables in the BLAST+ distribution.

Thu, 17 March 2022 [More BLAST news...](#)

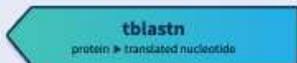
### Web BLAST



**Nucleotide BLAST**  
nucleotide → nucleotide



**blastx**  
translated nucleotide → protein



**tblastn**  
protein → translated nucleotide



**Protein BLAST**  
protein → protein

### BLAST Genomes

Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes

### Standalone and API BLAST



**Download BLAST**  
Get BLAST databases and executables



**Use BLAST API**  
Call BLAST from your application



**Use BLAST in the cloud**  
Start an instance at a cloud provider

### Specialized searches



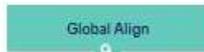
**SmartBLAST**

Find proteins highly similar to your query



**Primer-BLAST**

Design primers specific to your PCR template



**Global Align**

Compare two sequences across their entire span (Needleman-Wursch)



**CD-search**

Find conserved domains in your sequence



**IgBLAST**

Search immunoglobulins and T cell receptor sequences



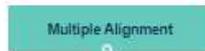
**VecScreen**

Search sequences for vector contamination



**CDART**

Find sequences with similar conserved domain architecture



**Multiple Alignment**

Align sequences using domain and protein constraints



**MOLE-BLAST**

Establish taxonomy for uncultured or environmental sequences

**Родной BLAST – NCBI**

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



## Что выбрать?

Программа	Query	Тип БД	Сравнивает
<b>Blastn</b>	ДНК	ДНК	ДНК
<b>Blastp</b>	белок	белок	белки
<b>Blastx</b>	ДНК	белок	белки
<b>Tblastn</b>	белок	ДНК	белки
<b>Tblastx</b>	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.

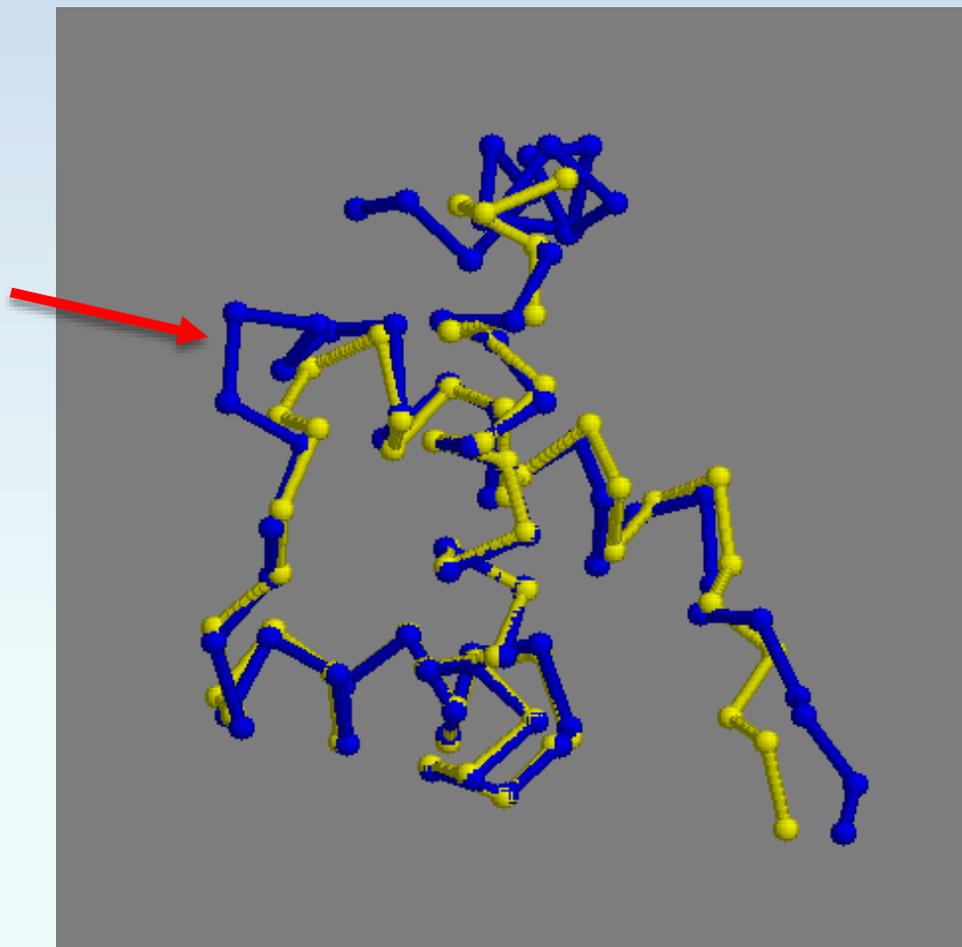
**домовой**  
**скупидом**  
**водомерка**

**лесовоз**  
**ледоход**

**---лесо---воз**  
**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

ледоход

---лесо---воз

лед---оход---

# Матрицы замен

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

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

2 сорта последних:

**PAM** (Point Accepted Mutations) – на выравниваниях очень близких белков (PAM20 = PAM<sup>20</sup>)

**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**

## MAIN NAVIGATION

- Introduction
- Contact
- Downloads & Branding
- Twitter/News
- PSIPRED Team Links
- People
- ProCovar
- Publications
- Vacancies
- PSIPRED Workbench Links
- PSIPRED Workbench
- Workbench Overview
- Workbench Citation
- Help & Tutorials
- REST API
- PSIPRED Github

The PSIPRED Workbench provides a range of protein structure prediction methods. The site can be used interactively via a web browser or programmatically via our REST API. For high-throughput analyses, downloads of all the algorithms are available.

**Amino acid** sequences enable: secondary structure prediction, including regions of disorder and transmembrane helix packing; contact analysis; fold recognition; structure modelling; and prediction of domains and function. In addition **PDB Structure files** allow prediction of protein-metal ion contacts, protein-protein hotspot residues, and membrane protein orientation.

## Data Input

Select Input data type

 Sequence Data  PDB Structure Data

Choose prediction methods (hover for short description)

## Popular Analyses

- PSIPRED 4.0 (Predict Secondary Structure)
- MEMSAT-SVM (Membrane Helix Prediction)
- DISOPRED3 (Disopred Prediction)
- pGenTHREADER (Profile Based Fold Recognition)

## Contact Analysis

- DeepMetaPSICOV 1.0 (Structural Contact Prediction)
- MEMPACK (TM Topology and Helix Packing)

## Fold Recognition

- GenTHREADER (Rapid Fold Recognition)
- pDomTHREADER (Protein Domain Fold Recognition)

## Structure Modelling

- Bioserf 2.0 (Automated Homology Modelling)
- DMPfold 1.0 Fast Mode (Protein Structure Prediction)
- Domserf 2.1 (Automated Domain Homology Modelling)

## Domain Prediction

- DomPred (Protein Domain Prediction)

## Function Prediction

- FFPred 3 (Eukaryotic Function Prediction)
- Help...

**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: 987768742244664265311588652002212788898785421444666432233763

Pred: CCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCCHHHHHCCCCCCCCCCCCCCCC

AA: MTERRVFPFSLLRGPSWDFFRDWYPHSRLFDAQAFGLPRLPEEWSQWLGSSWPGYVRPLPP  
10 20 30 40 50 60

Conf: 000373112633342100220898312588795399999827998756599998998999

Pred: CCCCCCCCCCCCCCCCCCHHHCCCCCEEEEECCCCCEEEEECCCCCHHEEEEECEEEEE

AA: AAIESPAVAAPAYSRALSRQLSSGVSEIRHTADRWRVSLDVNHFAPDELTVKTKDGVVEI  
70 80 90 100 110 120

Conf: 998010000188758889998762889885762876657985999971578875447427

Pred: EEEEECHHHCCCCCEEEEEEEEECCCCCCHHEEEEECCCCCEEEEECCCCCCCCCEE

AA: TGKHEERQDEHGYSISRCFTRKYTLPPGVDPTQVSSLSPEGTLTVEAPMPKLATQSNEIT  
130 140 150 160 170 180

Conf: 7574148754467664434312039

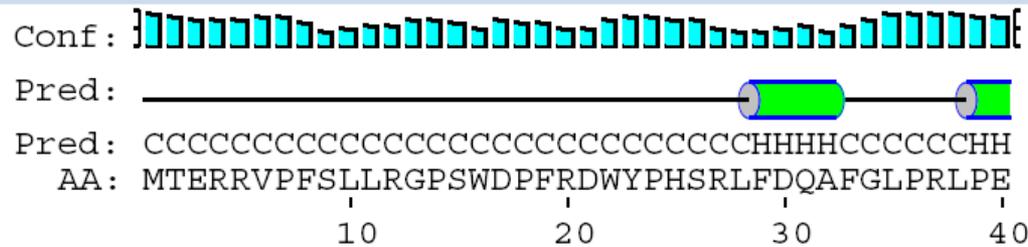
Pred: EEEEECCCCCCCCCCCCCHHHCC

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



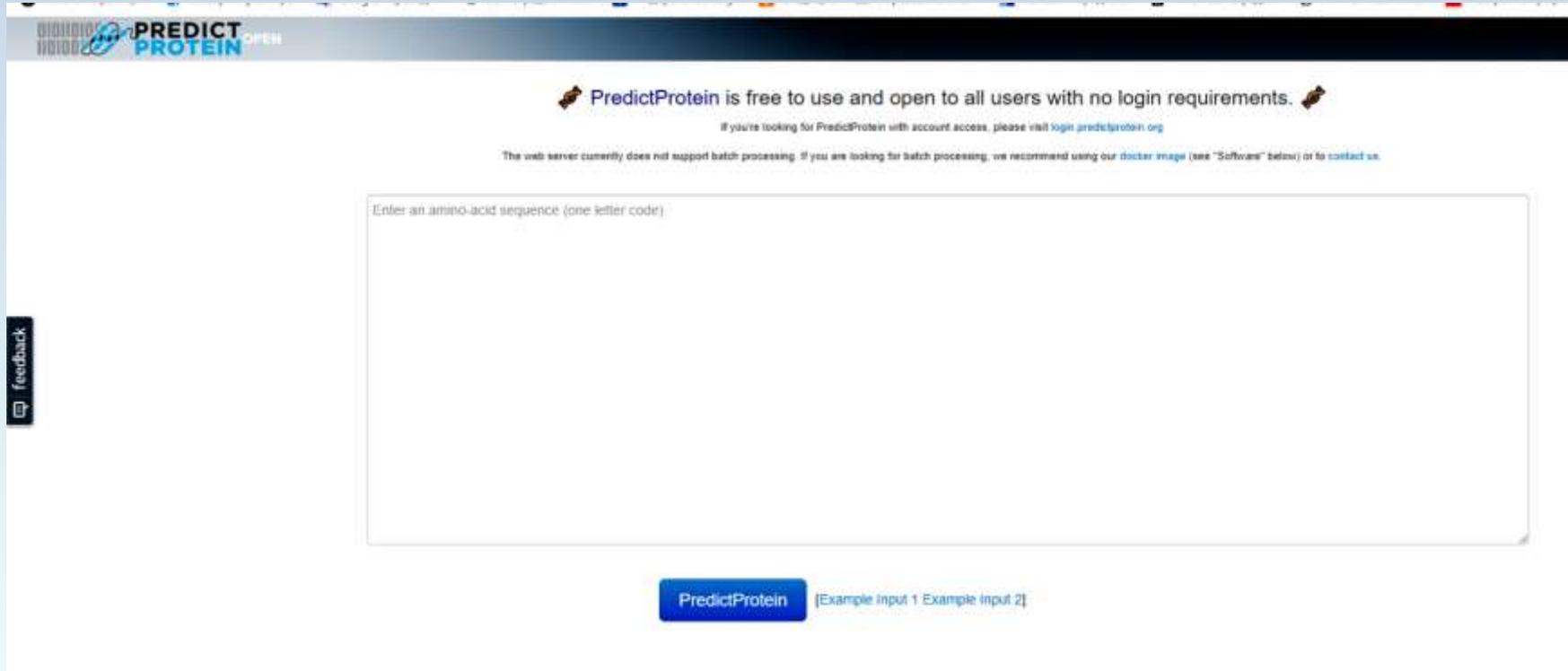
Legend:

-  = helix
-  = strand
-  = coil
- Conf:  = confidence of prediction
- Pred:  = predicted secondary structure
- AA: target sequence

# 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
- Требуется регистрации

# PredictProtein



The screenshot shows the PredictProtein website interface. At the top left is the logo for PredictProtein, which includes a stylized protein structure and the text "PREDICT PROTEIN OPEN". Below the logo, there is a message: "PredictProtein is free to use and open to all users with no login requirements." followed by a small icon of a key. Below this message, there is a link: "If you're looking for PredictProtein with account access, please visit [login.predictprotein.org](http://login.predictprotein.org)". Further down, there is a note: "The web server currently does not support batch processing. If you are looking for batch processing, we recommend using our [docker image](#) (see "Software" below) or to [contact us](#)." In the center of the page is a large text input field with the placeholder text "Enter an amino-acid sequence (one letter code)". Below the input field is a blue button labeled "PredictProtein" and two links: "[Example input 1]" and "[Example input 2]". On the left side of the page, there is a vertical "feedback" button with a speech bubble icon.

- <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



[EVA mirrors](#) - [Secondary structure](#) [Comparative modelling](#) [Threading](#) [Contacts](#) - [FTP](#) - [search](#)

Version  
May 25, 2001

[email](#)

## OBJECTIVES:

EVA continuously 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](#)

## INFORMATION:

- EVA [flow chart](#)
- EVA [concept](#)
- Structure prediction [servers](#) participating
- Related [resources](#) ( [CAFASP results](#).)

## CONTACT:

- EVA [eva@cubic.bioc.columbia.edu](mailto:eva@cubic.bioc.columbia.edu)
- EVA [team](#)

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

[EVA mirrors](#) - [Secondary structure](#) [Comparative modelling](#) [Threading](#) [Contacts](#) - [FTP](#) - [search](#)

# PDB – универсальный репозиторий данных по пространственной структуре белка

**RCSB PDB**  
PROTEIN DATA BANK

An Information Portal to Biological Macromolecules  
As of Tuesday Dec 04, 2007 there are 47625 Structures

CONTACT US | HELP | PRINT PAGE

PDB ID or keyword  Author

Advanced Search

Home Search

- Home
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- General Education
- Site Tutorials
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions
- Report Bugs/Comments

*Quick Tips:*

The top bar "site search" will search structure files as well as web site pages.

**Are you missing data updates? The PDB archive has moved to <ftp://ftp.wwpdb.org>. For more information click [here](#).**

## Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the [wwPDB](#) whose mission is to ensure that the PDB archive remains an international resource with uniform data.

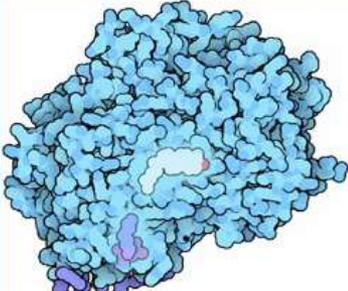
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found [here](#).

A [narrated tutorial](#) illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the [Macromedia Flash player download](#).]

Comments? [info@rcsb.org](mailto:info@rcsb.org)

### Molecule of the Month: Oxidosqualene Cyclase



Cholesterol has gained a bad reputation in recent years. It is absolutely essential in our lives: it is needed to keep our membranes fluid and it is the raw material used to build a host of important molecules such as vitamin D and steroid hormones. However, elevated levels of cholesterol (for instance from a fat-rich diet) have been linked to the formation of atherosclerosis and heart disease. Today, doctors suggest that a combination of a healthy low-fat diet and exercise will keep these two faces of cholesterol in balance.

- More ...
- Previous Features

# PDB – стандартная запись

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

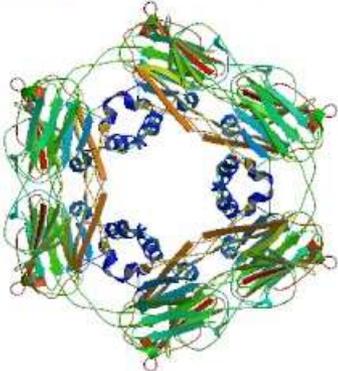
**1gme**    DOI 10.2210/pdb1gme/pdb

Red - Derived Information

<b>Title</b>	CRYSTAL STRUCTURE AND ASSEMBLY OF AN EUKARYOTIC SMALL HEAT SHOCK PROTEIN						
<b>Authors</b>	Van Montfort, R.L.M., Basha, E., Friedrich, K.L., Slingsby, C., Vierling, E.						
<b>Primary Citation</b>	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. <i>Nat.Struct.Biol.</i> 8: 1025-1030 [Abstract] 						
<b>History</b>	Deposition 2001-09-13 Release 2001-11-29						
<b>Experimental Method</b>	Type: X-RAY DIFFRACTION Data  [ EDS ]						
<b>Parameters</b>	Resolution(Å) 	R-Value	R-Free	Space Group			
	2.70	0.231 (obs.)	0.286	H 3 2			
<b>Unit Cell</b>	Length [Å]	a	171.65	b	171.65	c	124.16
	Angles [°]	alpha	90.00	beta	90.00	gamma	120.00
<b>Molecular Description Asymmetric Unit</b>	Polymer: 1 Molecule: HEAT SHOCK PROTEIN 16.9B Chains: A,B,C,D						
<b>Classification</b>	Chaperone						
<b>Source</b>	Polymer: 1 Scientific Name: <b>Triticum aestivum</b>  Common Name: <b>Wheat</b> Expression system: <b>Escherichia coli</b>						
<b>SCOP</b>	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

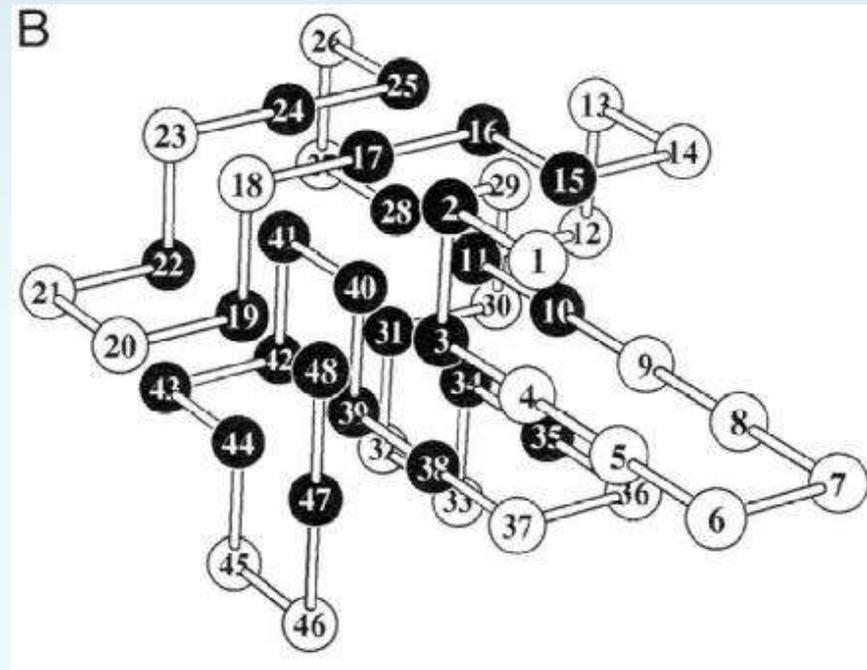
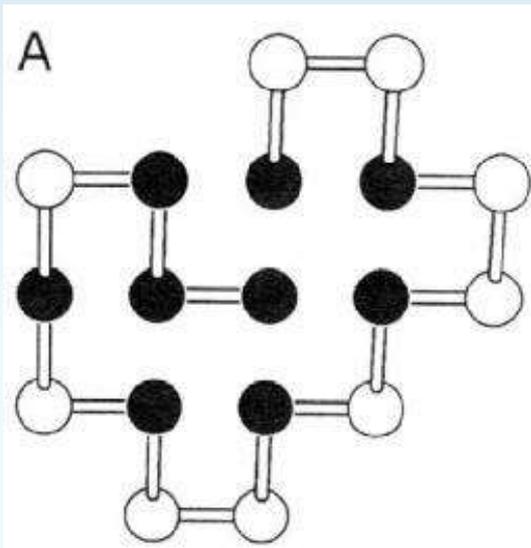
\* Capable of displaying biological molecules.

# Предсказание структуры *ab initio*

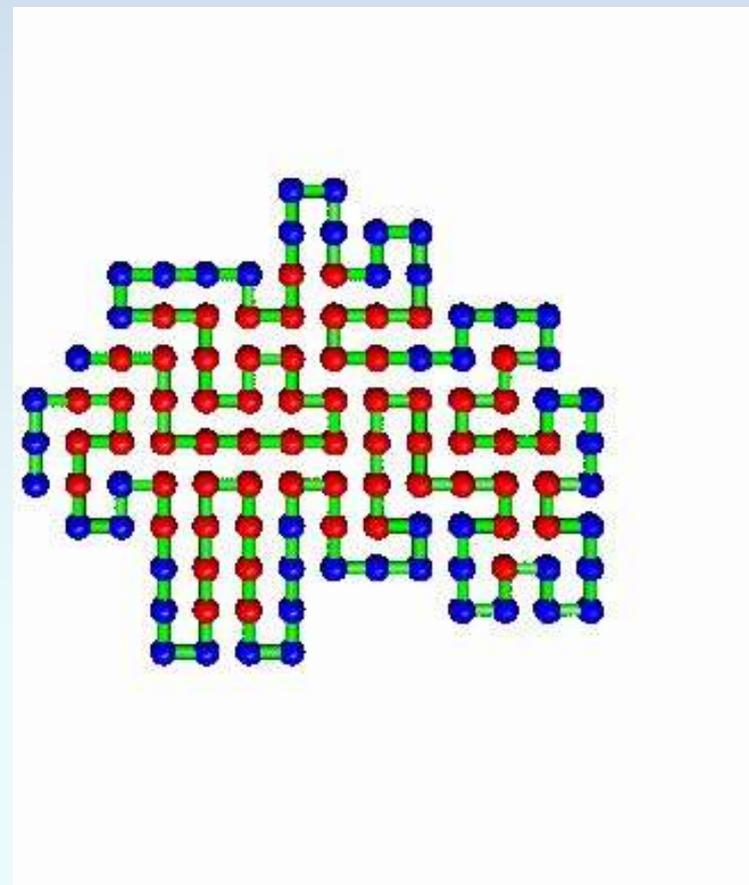
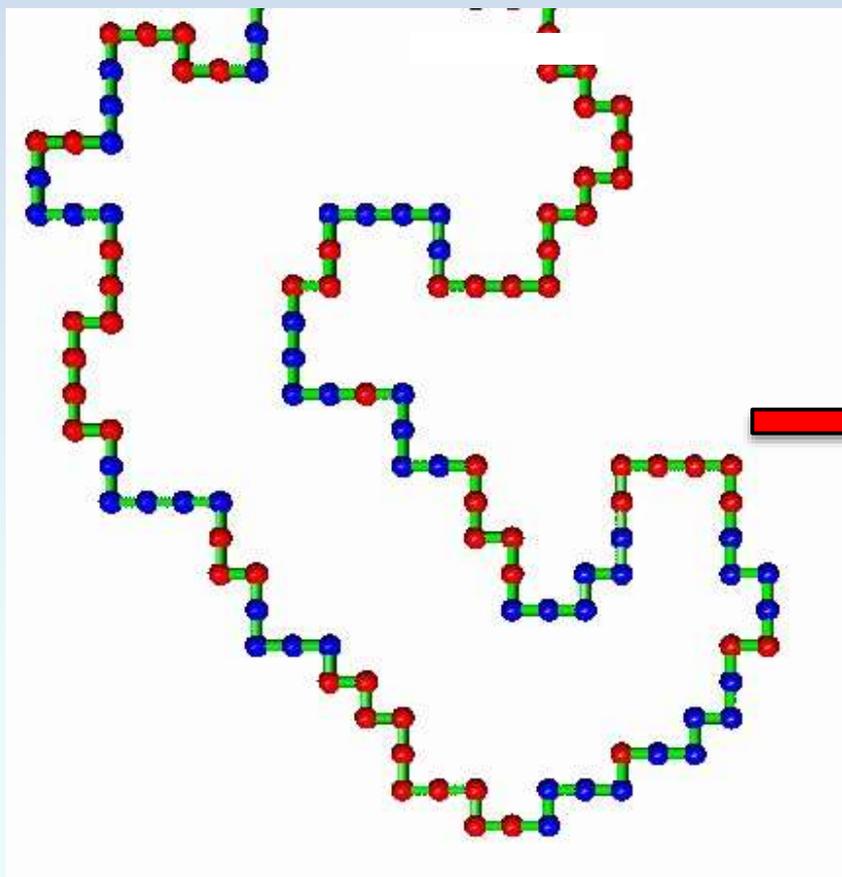
- Функция потенциальной энергии
  - Модель водного раствора
  - Оценка попарного взаимодействия между аминокислотами
- Поиск в пространстве всевозможных конформаций
  - Модель на основе “решетки”
  - Молекулярная динамика
  - Использование библиотек известных 3D фрагментов
- Предсказание вторичной структуры

# 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

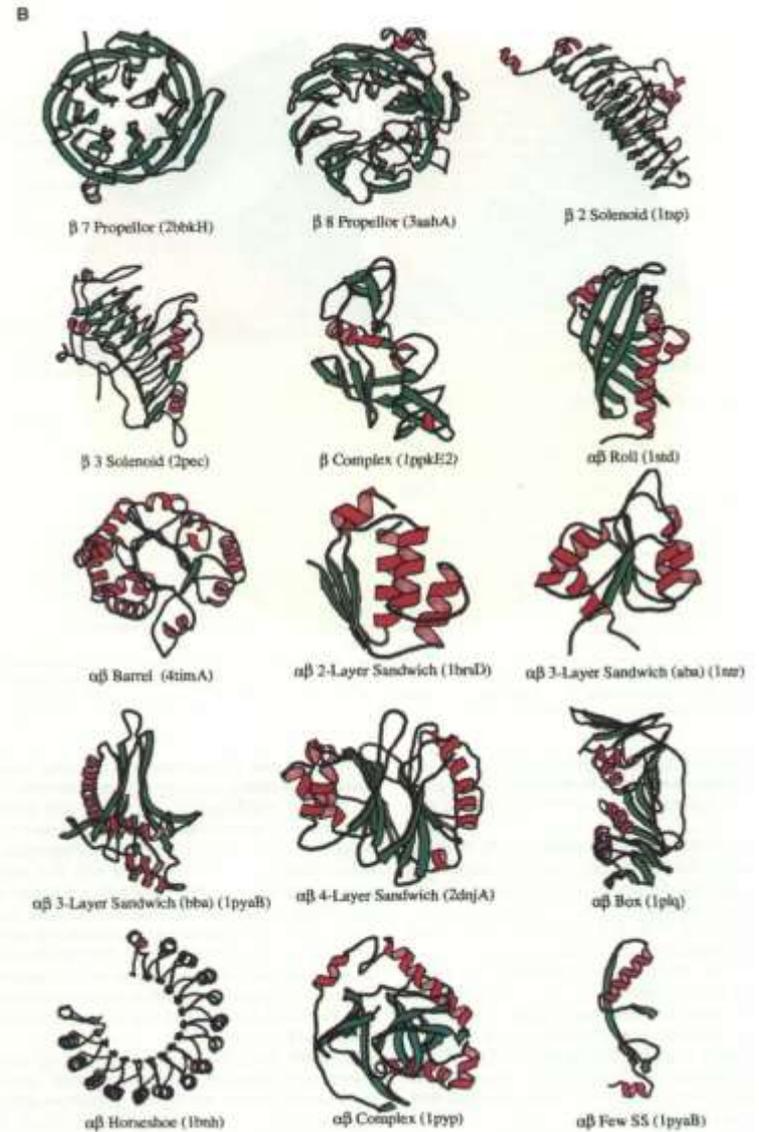
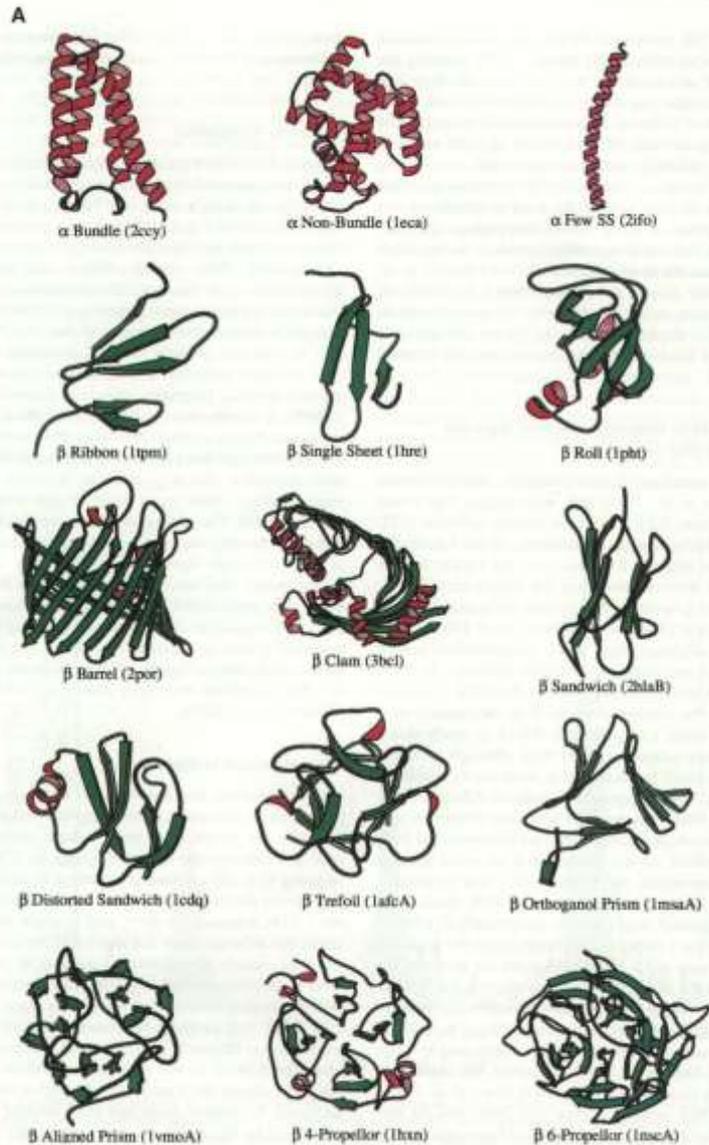
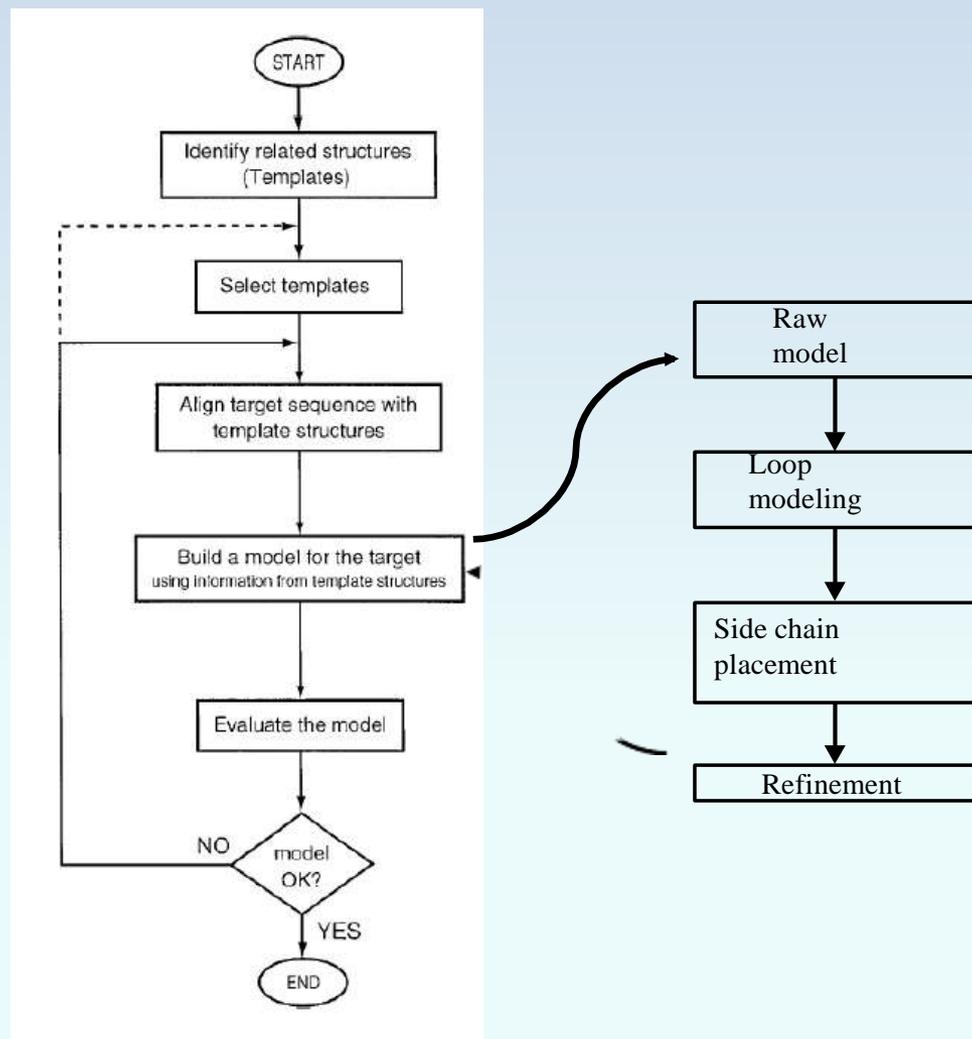
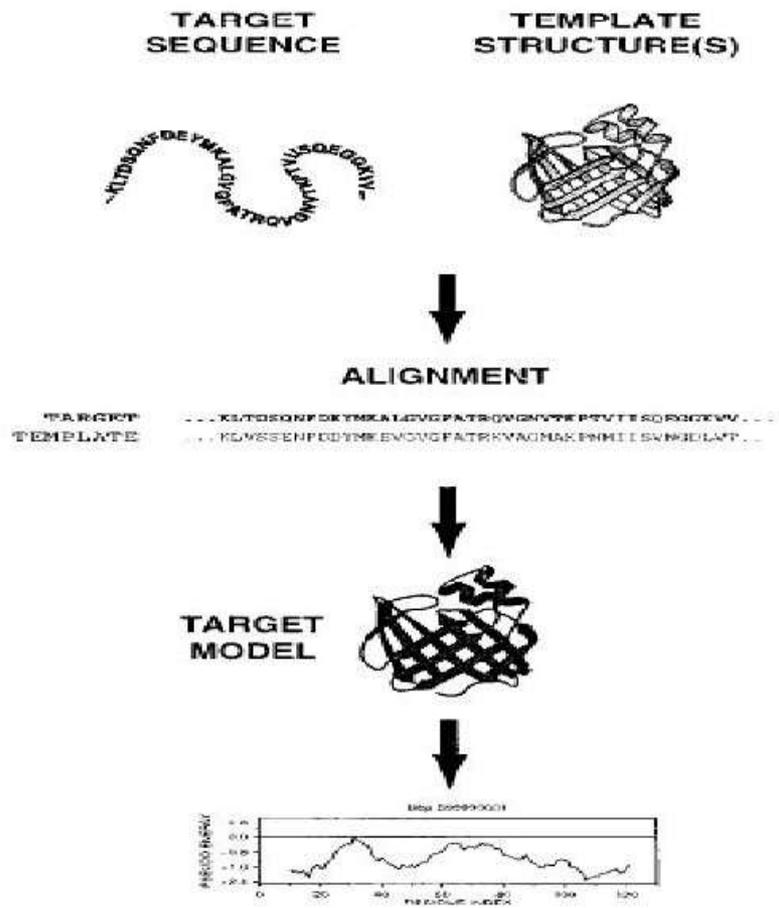


Fig. 4. Structures from the PDB representing the 35 different protein architectures in the CATH protein structure classification system.

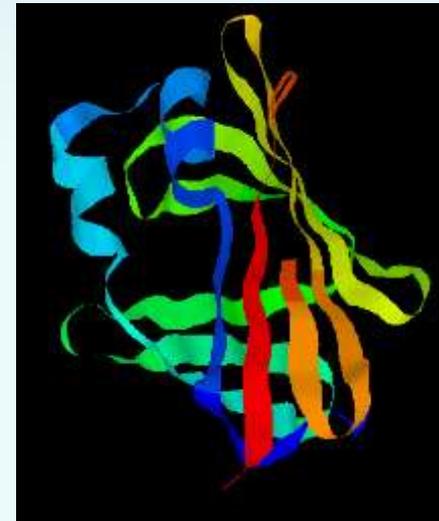
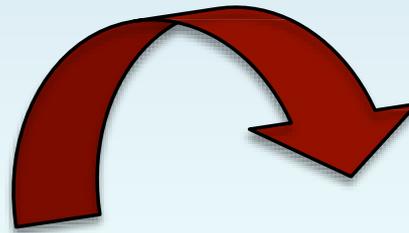
# Предсказание структуры на основе ГОМОЛОГИИ



# 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

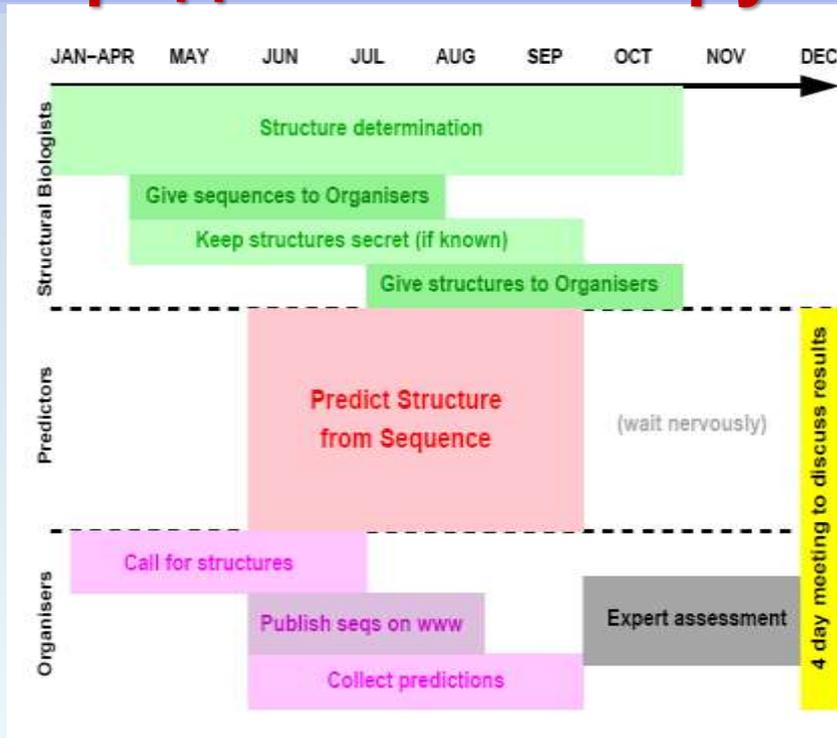


# Основные компоненты тридинга

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

# CASP - конкурс методов предсказания структуры белков

[FoldIt](#)



What's New

New Paper out of the Baker Lab

## Critical Assessment of protein Structure Prediction, CASP

**Baker lab site** and click on second item listed (Global analysis of protein folding using massively parallel design, synthesis, and testing).

Traditionally, if you wanted to measure a protein's structural stability, you'd first have to purify the protein (which is expensive and time-intensive) and then run some kind of protein unfolding experiment with your purified protein sample.

The authors of this paper developed a protease assay that can approximate the stability of thousands of proteins in parallel, without having to purify and test each individual protein.

The idea is that unfolded or unstable proteins will be readily chewed up by proteases, but stable, well-folded proteins will resist protease degradation and remain intact. We can then sort out which proteins survived the proteases.

From this large dataset of protein stability, the authors were also able to make some interesting conclusions about protein design. For example, they found that ASP and GLU are especially stable in the first turn of a helix; ARG and LYS are preferred in the last turn.

GET STARTED: DOWNLOAD



Are you new to FoldIt? Click here.

Are you a student? Click here.

Are you an educator? Click here.

SEARCH

Google Search Only search beta

RECOMMEND FOLDIT

Send

USER LOGIN

Username \*

Password \*

Log in

Create new account  
Request new password

SOLVERS EVOLVERS GROUPS TOPICS

PLAYER	PUZZLE	SCORE
Neaven W...	75 1426	Beginner Puzzle...in 7,143
Enzyme 31	25	1426: Revolut...y 3 10,071
Enzyme 31	25	1424: Extra Lac...ign 11,034
TwoEdge 75	1197	Beginner Puzzle...yle 8,581
downward 75	200	Beginner Puzzle...ste 8,722
downward 75	200	Beginner Puzzle...ign 10,008
downward 75	200	Beginner Puzzle...ign 8,949
downward 75	200	Beginner Puzzle...ity 14,744

## 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 учётом 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

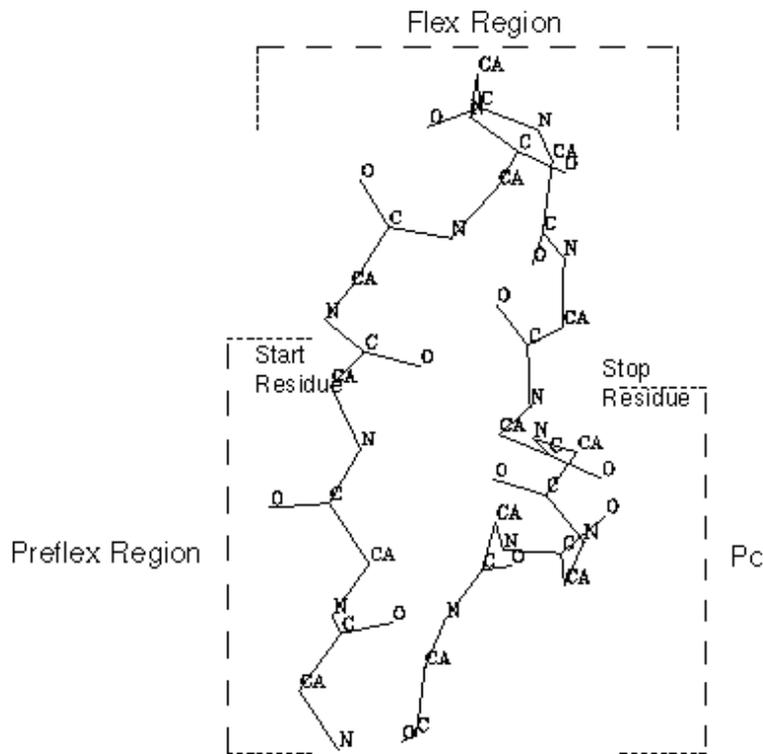
В первую очередь присваиваются координаты атомам полипептидной цепи.

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  $\chi$  одинаковы).

Если же аминокислота модельного белка более длинная, то начальный ход повторяет ход боковой цепи в белке-гомологе, а последующие атомы цепи помещаются в развёрнутую (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)

```
TRIVR ( 1) MAKIVVIGAGVAGLTTALQLLRKGHEVTIVSEFTPGDLSI g-yTSPWAGANWLTfydgg (58 )
COL ( A999) LMMHSQKRVVVLGSGVIGLSSALILARKGYSVHILARDLPEDVSSQTFASPWAGANWTF----- ( gap )

TRIVR ( 59) klADYDavsypILRELARSSPEAGIRLISqrsHVLKRDLPKLEVAMSAICQrnpWFKNTVDSFEII (124 )
COL ( gap ) -MTLTDG---PRQAKWEEESTFKKVELVPT-GHAMWLKGTTRRFAQNEGLIG-HWYKDITPNYRPL (A1118 )

TRIVR ( 125) EdrsRIVHDDVaylvEFRSVCIHTEGVYLNWLMSQCLSLGATVVKRRVNHIKDanllhssgsrpDVI (190 )
COL ( A1119) PS-SECPPGAI G--V TYDTLSVHAPKYCQYLARELQKLGATFERRTVTSLEQA--FD--G--ADLV (A1175 )

TRIVR ( 191) VNCGLFARFLGGVEDKKMYPPIRGQVVLVRNSLPFMASFSSTPEKenedealyIMTRFDGTSIIGG (256 )
COL ( A1176) VNATGLGAKSIAGIDDQAAEPIRGQTVLVKSPCKRCTMDSSDPASP-----A-YIIPRPGGEVICGG (A1236 )

TRIVR ( 257) CFQPNNWSSEPDPSLTHRILSRALDRFPELTKDGPLd---iVRECVGHRPGREGGPRVELEKi--p (317 )
COL ( A1237) TYGVGDWDLVSNPETVQRILKHCLRLDPTISSDGTIEGIEVLRHNVGLRPARRGGPRVEAERIVLP (A1302 )

TRIVR ( gap ) -----gvg-----fVVHNYGAAGAGYQSSYGMADAVSYVERALTRPNI (356 )
COL ( A1303) LDRTKSPLSLGRGSARAAKEKEVTLVHAYGFSSAGYQOSWGAAEDVAQLVDEAFORYHG (A1361 )
```

## 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

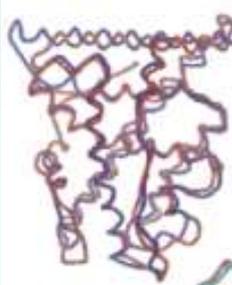
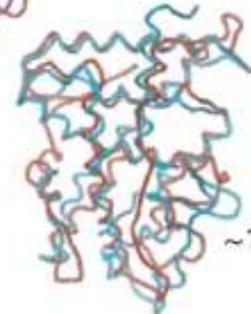
Результат этой процедуры - белковая структура с правильной стереохимией (длины валентных связей и значения валентных углов не будут существенно отличаться от равновесных значений), с отрицательной энергией **несвязанных** взаимодействий (свидетельство того, что не наблюдается перекрытие ван-дер-ваальсовских 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 приведёт к её улучшению с точки зрения стереохимии, но при этом возрастут искажения структуры активного центра (центра связывания) вашей структуры.

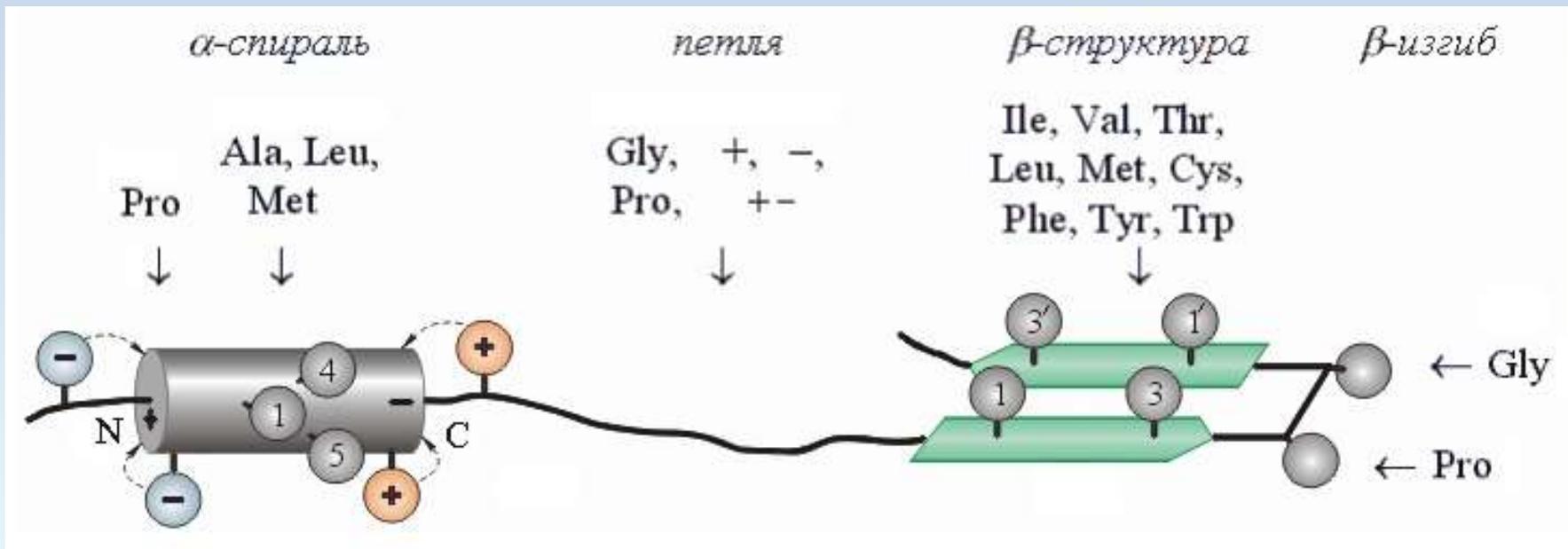
Модельная структура построена и отрелаксирована. Она обладает участками структурно-консервативных областей, унаследованных от белков гомологов, правильной стереохимией (результат регуляризации). Дальнейшие манипуляции с этой структурой (подгонка геометрии активного центра, точечные мутации) зависят от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

 <p>«Качество» моделей</p> <p>1.0 Å 100%</p>	100	<b>% идентичности шаблон–мишень</b>  Незначимая гомология	<b>Парное аминокислотное выравнивание</b>	Структурно-подкрепленное конструирование лекарств Изучение каталитического механизма Докинг и оптимизация лигандов Виртуальный скрининг Изучение антигенных детерминант Молекулярное замещение в PCA		
 <p>1.5 Å 95%</p>	90			<b>Множественное выравнивание</b>	Оценка фармакологического потенциала мишени Оптимизация ЯМР-структур, «вписывание» в электронную плотность низкого разрешения	
 <p>3.5 Å 80%</p>	80				<b>Протягивание, профили</b>	Предсказание мутаций, дизайн химер Определение функции белка
 <p>4–8 Å ~100 а/к</p>	70					<b>de novo</b>
	60					
	50					
	40					
	30					

# «Шаблоны» для опознавания $\alpha$ -спирали, петли, $\beta$ -структуры и $\beta$ -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 — C-конец, мы получаем нечто вроде **«шаблона»** спирали.

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

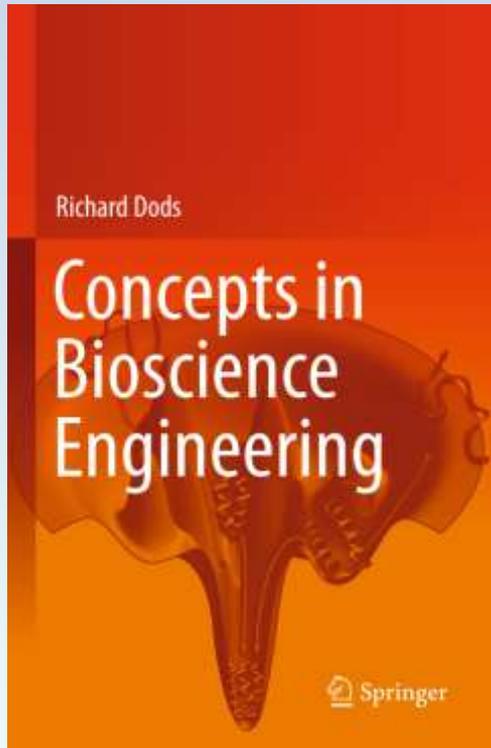
такое же описание для образования  $\beta$ -структуры, отдельных участков цепи или образующих более сложные структуры  $\beta$ - $\alpha$ - $\beta$ -суперспиралей,

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  $\phi \approx 60^\circ$ , 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!*

