Лекция 10. Предсказание вторичной структуры белка; SCOP, CATH; соревнование CASP

Курс: Структурная Биоинформатика и моделирование лекарств (ВШЭ)

Головин А.В. [1]

[1] МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Москва, 2017

Содержание

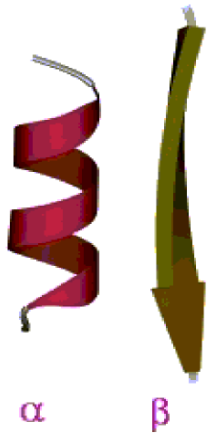
Первичная структура

Первичная структура – это аминокислотная последовательность:

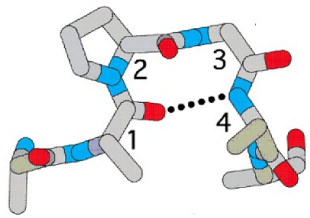
Met-Ala-Gly-Trp-Ala-Val-Asp …

Вторичная структура

Вторичная структура
белка - это упорядоченные расположения атомов основной цепи полипептида, безотносительно к типам боковых цепей (групп) и их конформациям.

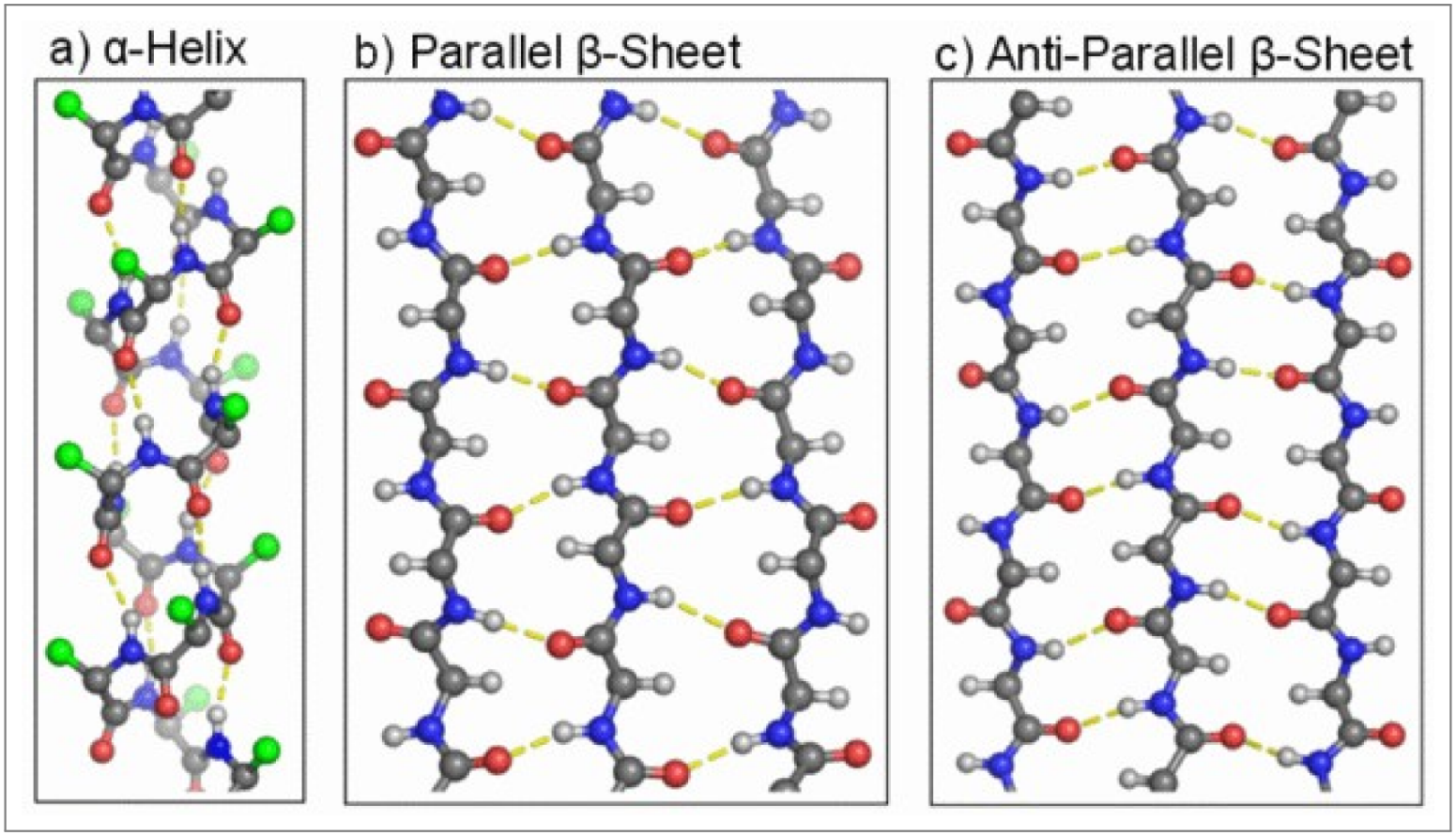


Если упорядоченность такова, что двугранные углы одинаковы у всех остатков, то говорят о регулярной вторичной структуре. Регулярными вторичными структурами являются спирали и β– структуры.



Пример нерегулярной вторичной структуры β–поворот (β–изгиб, реверсивный поворот).

Регулярные вторичные структуры



Постулаты

- Современные методы на позволяют в целом предсказывать структуру белка
- В целом структура белка соответствует минимуму потенциальной энергии белка
- Неточность методов предсказания структуры связанно с недостаточной мощностью компьютеров.

Классификация

Особенности:

- Паттерны водородных связей
- Двугранные углы

Программы для определения

- DSSP
- STRIDE
- Continuum

Обозначения

Восемь состояний аминокислоты в DSSP:

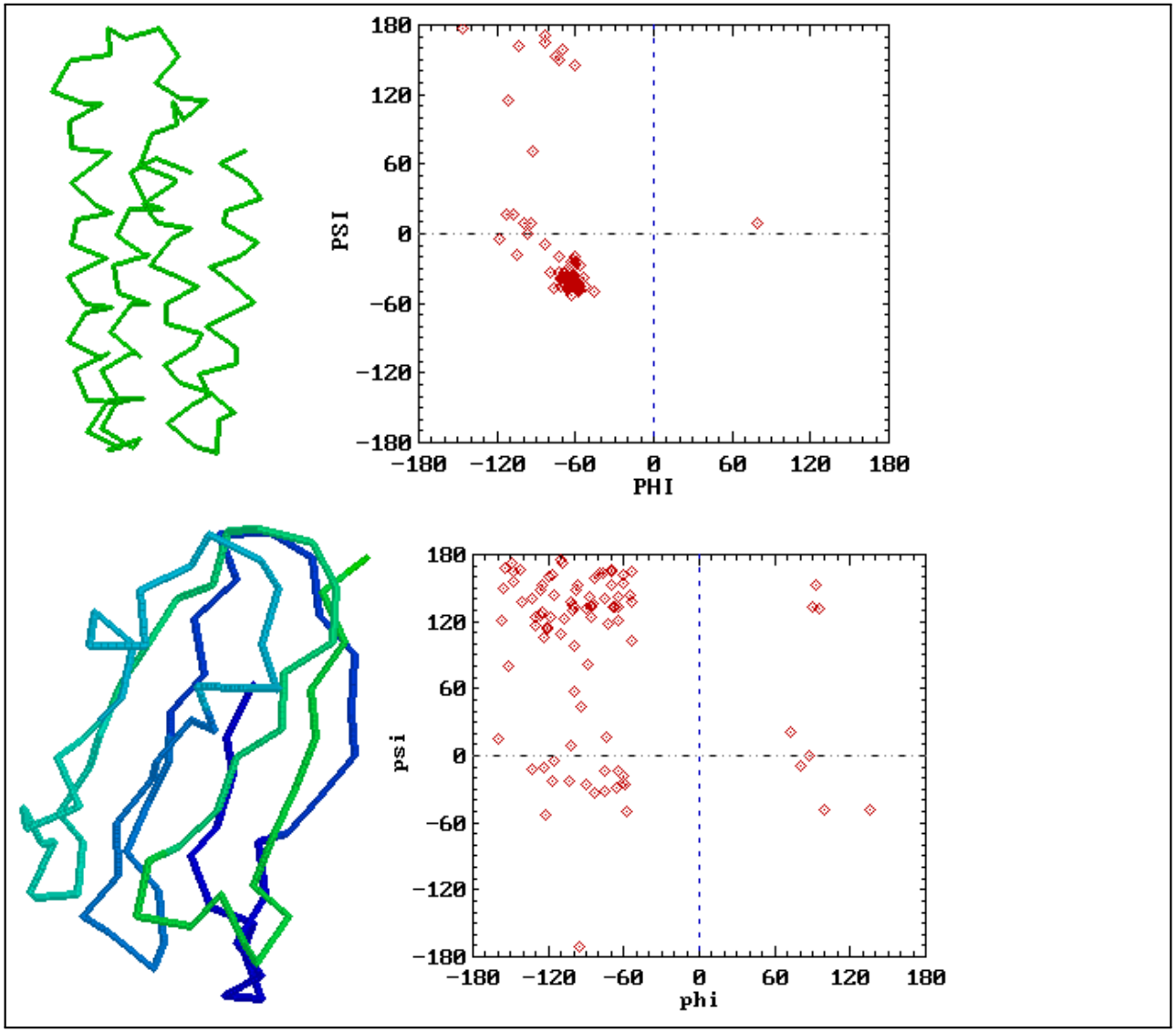
- H: α-helix
- G: 3_{10} -helix
- I: π-helix
- E: β-strand
- B: bridge
- T: β-turn
- S: bend
- C: coil

```
24    26    E   H  <  S+      0    0   132
25    27    R   H  <  S+      0    0   125
26    28    N      <          0    0    41
27    29    K                 0    0   197
28          !                 0    0     0
29    34    C                 0    0    73
30    35    I   E      -cd    58  89B     9
31    36    L   E      -cd    59  90B     2
32    37    V   E      -cd    60  91B     0
33    38    G   E      -cd    61  92B     0
```

CASP:

- H = (H, G, I), E = (E, B), C = (C, T, S)

Двугранные углы



Предсказание вторичной структуры

- Для последовательности:
  GHWIATRGQLIREAYEDYRHFSSECPFI
- Предсказать состояние каждой аминокислоты

GHWIATRGQLIREAYEDYRHFSSECPFI
CEEEEECHHHHHHHHHHHCCCHHCCCCC

Цели предсказания

- Предсказание вторичной структуры проще чем предсказание третичной
- Аккуратное предсказание может упростить предсказание третичной структуры
- На основе вторичной структуры можно предположить функцию белка
- Классификация белков
- Предсказание изменения структуры при функционировании белка

Основные методы

- Статистические:
  Chou-Fasman method, GOR I-IV
- Методы ближайших соседей
  NNSSP, SSPAL
- Нейронные сети
  PHD, Psi-Pred, J-Pred
- HMM

Аккуратность

- Three-state prediction accuracy: Q_3

$$Q_3 = \frac{N_{\text{Correctly predicted residues}}}{N_{\text{Number of residues}}}$$

- Для всех петель $Q_3 \sim 40\%$

Развитие аккуратности

1974	Chou, Fasman	∼50-53%
1978	Garnier	63%
1987	Zvelebil	66%
1988	Qian, Sejnowski	64.3%
1993	Rost, Sander	∼70.8-72.0%
1997	Frishman, Argos	<75%
1999	Cuff, Barton	72.9%
1999	Jones	76.5%
2000	Petersen et al.	77.9%

Основные предположения

- Последовательность содержит достаточно информации для предсказания
- Боковые группы определяют структуру
- Окно в 13-17 остатков достаточно для предсказания
- Основы для выбора размера окна:
  - α-спирали 5-40 остатков
  - β-тяжи 5-10 остатков

Метод Chou-Fasman

- Из банка данных PDB установим вероятность нахождения остатка в той или иной структуре:

$$P_\alpha^i = \frac{P(\alpha|aa_i)}{p(\alpha)} = \frac{p(\alpha, aa_i)}{p(\alpha)p(aa_i)}$$

- Пример:
  #Ala=2000, #residues=20000, #helix=4000, #Ala in helix=500
  P(α,aa$_i$) = 500/20000
  p(α)=4000/20000
  p(aa$_i$) = 2000/20000
  P = 500 / (4000/10) = 1.25

Результат:

Chou-Fasman Parameters

Pα		Pβ		Pt	
Glu	1.51	Val	1.70	Asn	1.56
Met	1.45	Ile	1.60	Gly	1.56
Ala	1.42	Tyr	1.47	Pro	1.52
Leu	1.21	Phe	1.38	Asp	1.46
Lys	1.16	Trp	1.37	Ser	1.43
Phe	1.13	Leu	1.30	Cys	1.19
Gln	1.11	Cys	1.19	Tyr	1.14
Trp	1.08	Thr	1.19	Lys	1.01
Ile	1.08	Gln	1.10	Gln	0.98
Val	1.06	Met	1.05	Thr	0.96
Asp	1.01	Arg	0.93	Trp	0.96
His	1.00	Asn	0.89	Arg	0.95
Arg	0.98	His	0.87	His	0.95
Thr	0.83	Ala	0.83	Glu	0.74
Ser	0.77	Ser	0.75	Ala	0.66
Cys	0.70	Gly	0.75	Met	0.60
Tyr	0.69	Lys	0.74	Phe	0.60
Asn	0.67	Pro	0.55	Leu	0.59
Pro	0.57	Asp	0.54	Val	0.50
Gly	0.57	Glu	0.37	Ile	0.47

Chou-Fasman, Алгоритм

- Для элементов Спираль, Тяж
  - Ищем окно с 6 остатками где суммарный score > 1
  - Расширяем окно до score < 1
  - Двигаемся вперёд и повторяем
- Конфликт: Если получается, что в одном месте и спираль и тяж, то сравниваем P(h) и P(b)
- Точность: до $\sim 60\%$

Метод ближайших соседей

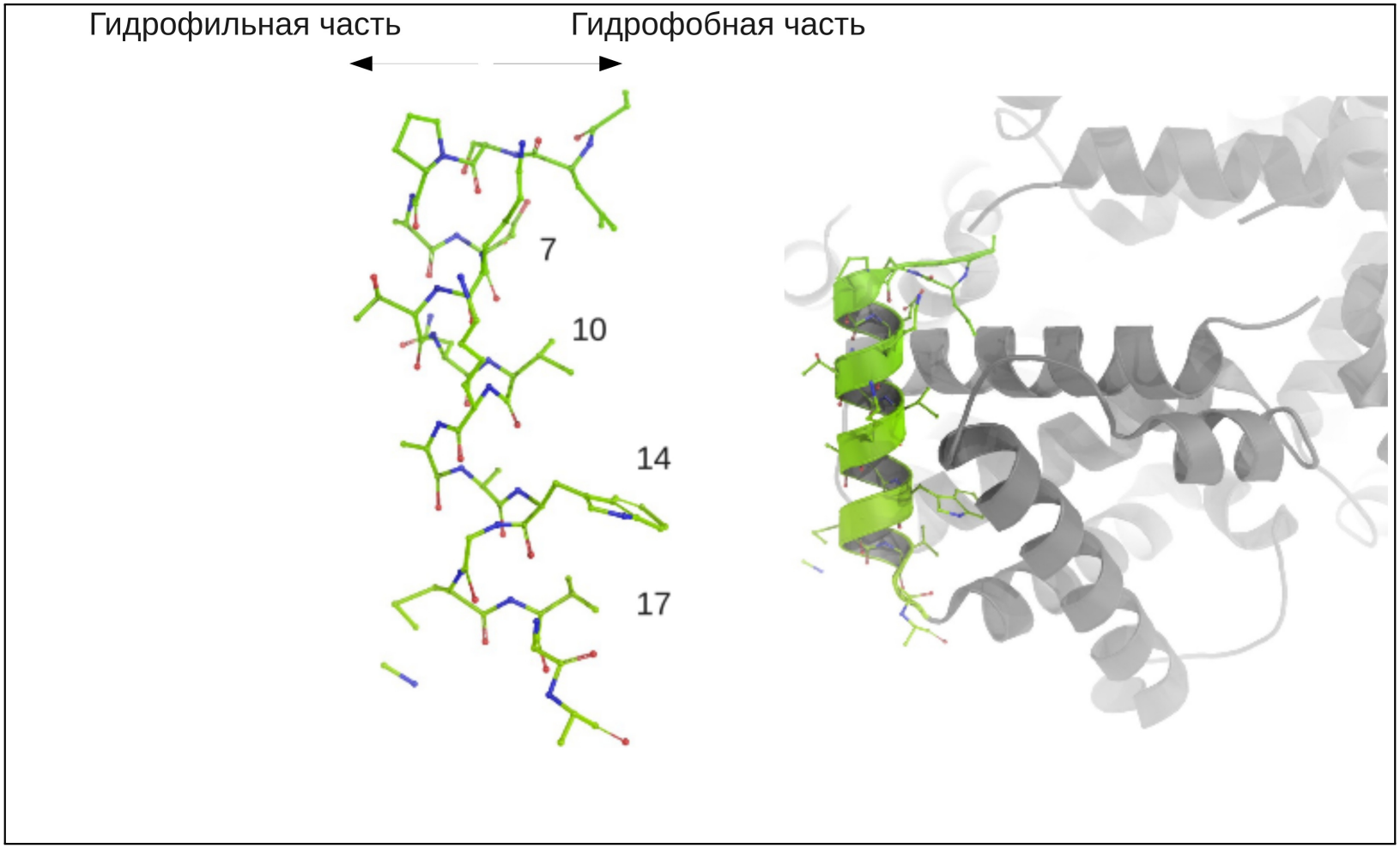
- Предсказываем BC центрального остатка для выбранного сегмента, для которого известен ближайший гомолог
  - Из базы данных находим близкую последовательность
  - Или строим N лучших локальных выравнивания входной последовательности со всеми последовательностями базы
- Используем max(n_α, n_β, n_c) для ближайшего соседа или max(s_α, s_β, s_c) для консенсуса из выравнивания

Environment preference score

- Предположим, что каждая а.к. имеет предпочтение для специфического структурного окружения
- Структурные переменные: вторичная структура, доступность растворителю
- Для базы данных уникальных белков FSSP

$$S(i,j) = log\frac{p(aa_i|E_j)}{p(aa_i)} = log\frac{p(aa_i, Ej)}{p(aa_i)p(E_j)}$$

Environment preference score



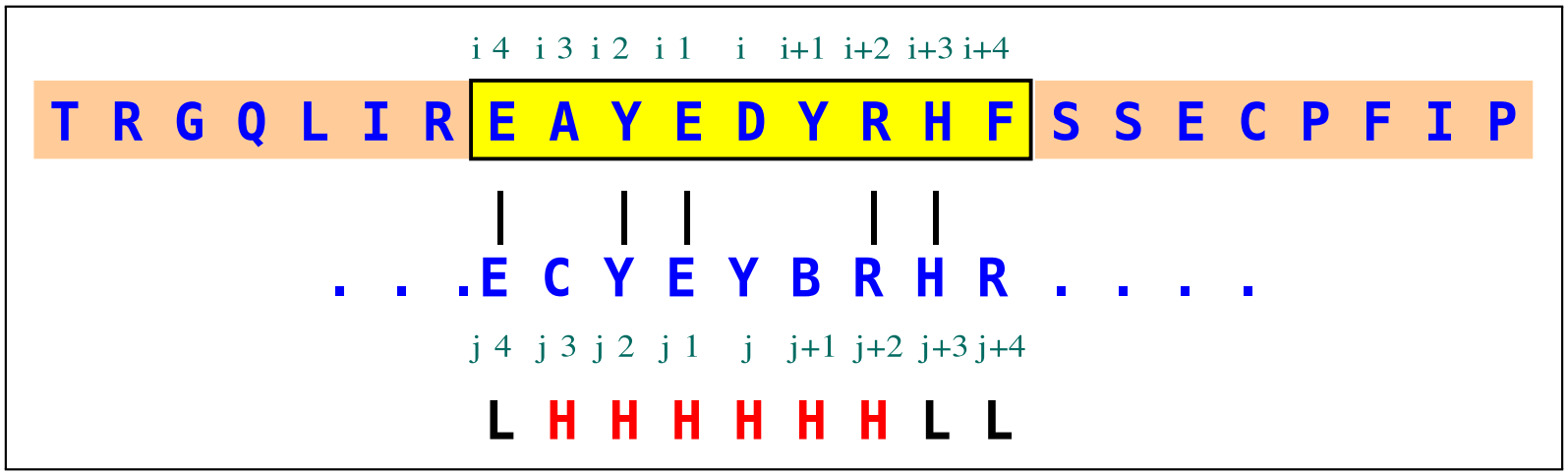
Матрица “Singleton”

	Helix			Sheet			Loop		
	Buried	Inter	Exposed	Buried	Inter	Exposed	Buried	Inter	Exposed
ALA	-0.578	-0.119	-0.160	0.010	0.583	0.921	0.023	0.218	0.368
ARG	0.997	-0.507	-0.488	1.267	-0.345	-0.580	0.930	-0.005	-0.032
ASN	0.819	0.090	-0.007	0.844	0.221	0.046	0.030	-0.322	-0.487
ASP	1.050	0.172	-0.426	1.145	0.322	0.061	0.308	-0.224	-0.541
CYS	-0.360	0.333	1.831	-0.671	0.003	1.216	-0.690	-0.225	1.216
GLN	1.047	-0.294	-0.939	1.452	0.139	-0.555	1.326	0.486	-0.244
GLU	0.670	-0.313	-0.721	0.999	0.031	-0.494	0.845	0.248	-0.144
GLY	0.414	0.932	0.969	0.177	0.565	0.989	-0.562	-0.299	-0.601
HIS	0.479	-0.223	0.136	0.306	-0.343	-0.014	0.019	-0.285	0.051
ILE	-0.551	0.087	1.248	-0.875	-0.182	0.500	-0.166	0.384	1.336
LEU	-0.744	-0.218	0.940	-0.411	0.179	0.900	-0.205	0.169	1.217
LYS	1.863	-0.045	-0.865	2.109	-0.017	-0.901	1.925	0.474	-0.498
MET	-0.641	-0.183	0.779	-0.269	0.197	0.658	-0.228	0.113	0.714
PHE	-0.491	0.057	1.364	-0.649	-0.200	0.776	-0.375	-0.001	1.251
PRO	1.090	0.705	0.236	1.249	0.695	0.145	-0.412	-0.491	-0.641
SER	0.350	0.260	-0.020	0.303	0.058	-0.075	-0.173	-0.210	-0.228
THR	0.291	0.215	0.304	0.156	-0.382	-0.584	-0.012	-0.103	-0.125
TRP	-0.379	-0.363	1.178	-0.270	-0.477	0.682	-0.220	-0.099	1.267
TYR	-0.111	-0.292	0.942	-0.267	-0.691	0.292	-0.015	-0.176	0.946
VAL	-0.374	0.236	1.144	-0.912	-0.334	0.089	-0.030	0.309	0.998

Total score

- Общее значение это сумма для окна l

$$Score(i,j) = \sum_{k=-l/2}^{l/2} [M(i+k, j+k) + cS(i+k, j+k)]$$



"Соседи"

1	**-**	**L**	**H**	**H**	**H**	**H**	**H**	**H**	**L**	**L**	**-**	S_1
2	**-**	**L**	**L**	**H**	**H**	**H**	**H**	**H**	**L**	**L**	**-**	S_2
3	**-**	**L**	**E**	**E**	**E**	**E**	**E**	**E**	**L**	**L**	**-**	S_3
4	**-**	**L**	**E**	**E**	**E**	**E**	**E**	**E**	**L**	**L**	**-**	S_4
n	**-**	**L**	**L**	**L**	**L**	**E**	**E**	**E**	**E**	**E**	**-**	S_n
***n*+1**	**-**	**H**	**H**	**H**	**L**	**L**	**L**	**E**	**E**	**E**	**-**	S_{n+1}
						⋮						

$$max(n_\alpha, n_\beta, n_L) \quad OR \quad max(\sum S_\alpha, \sum S_\beta, \sum S_L)$$

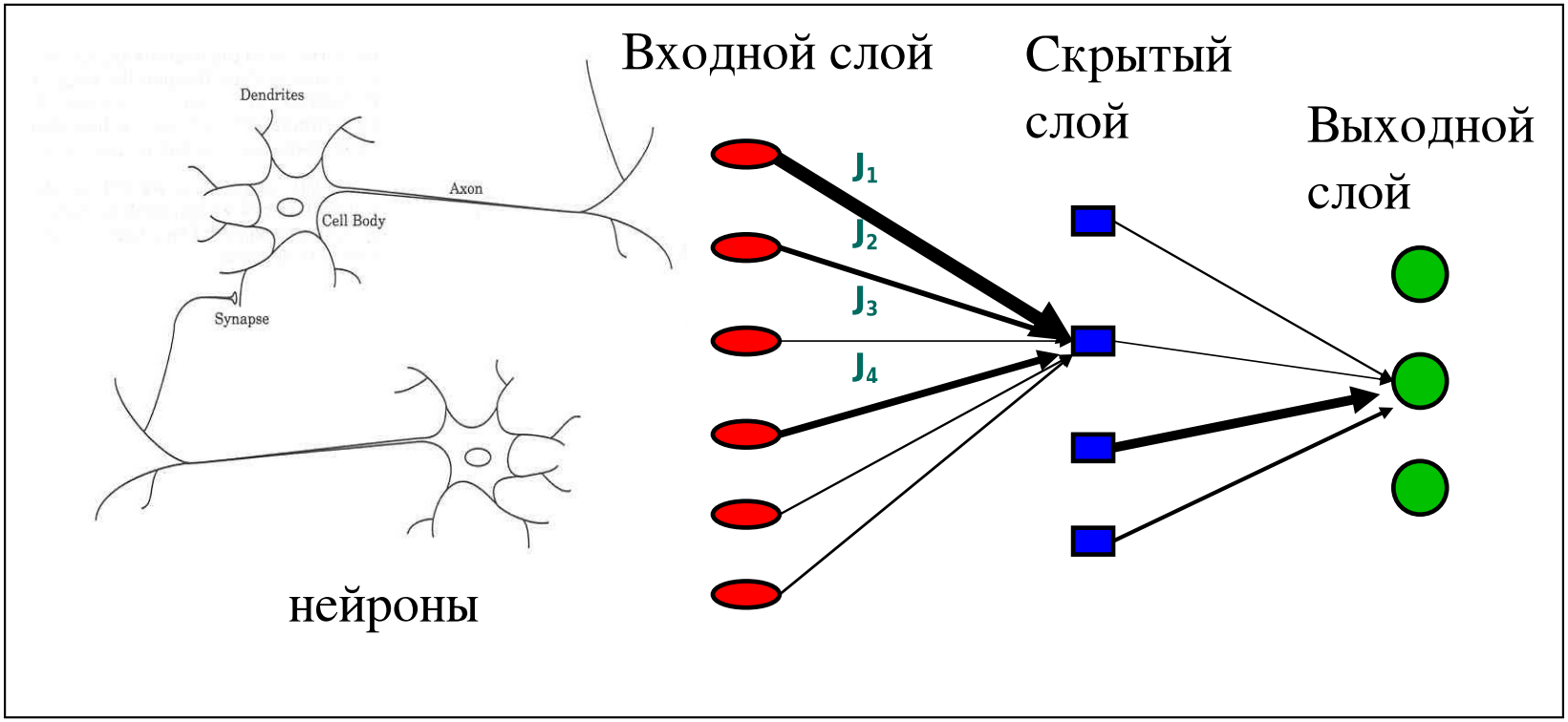
Эволюционная информация

- ”Все белки, которые эволюционировали в природе, при совпадении более 35% позиций при длине белка боле 100 а.к. имеют подобные структуры”
- Структура белка остаётся стабильной если изменилось не более 12% остатков
- PPM содержит много эволюционной информации, которую можно утилизировать.
- ”Гэпы”редко встречаются в тяжах и спиралях
- 1.4% рост Q_3 только благодаря росту содержания баз данных

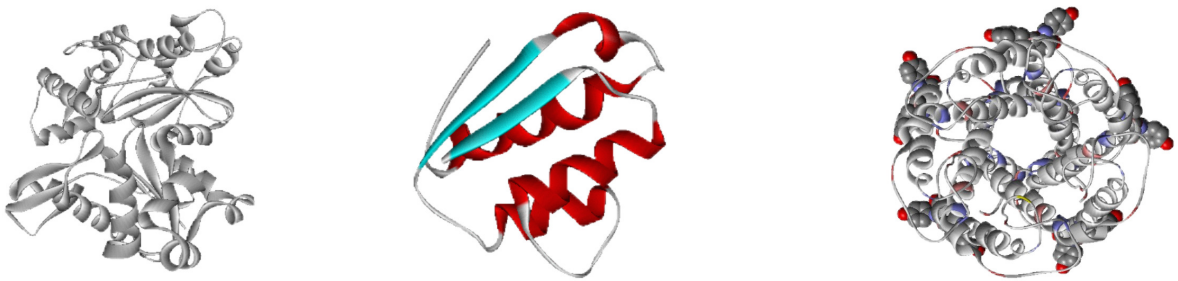
Как использовать эту информацию:

- Выравнивание последовательность-профиль
- Сравнение последовательности с белковыми семействами
- BLAST против PSI-BLAST.
- Использование PSSM вместо матриц PAM или BLOSUM.

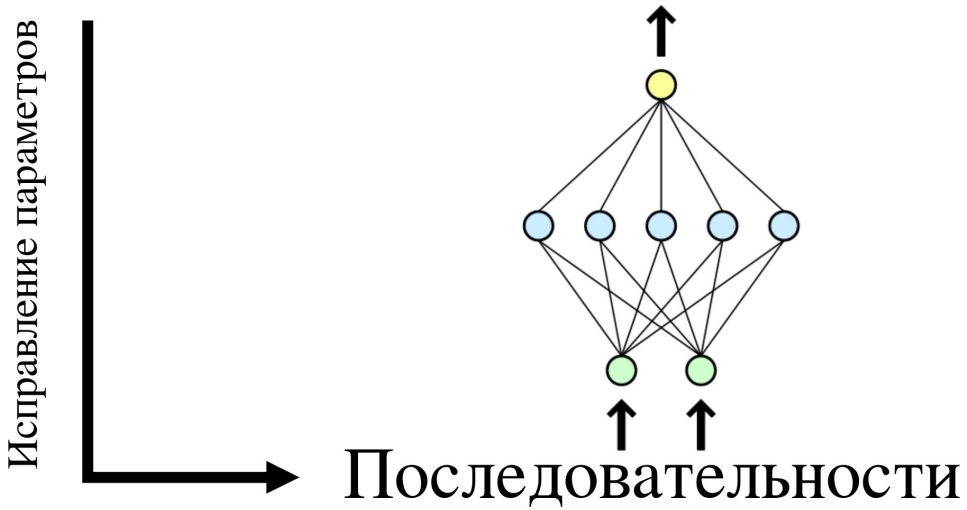
Нейронные сети



Обучение сети



Сравнение предсказания с реальностью



Стратегия разумного предсказания

- Построить выравнивание с гомологами, чем больше тем лучше
- Применить как можно больше разных современных методов предсказания
- Внимательно следить за паттернами и консервативными остатками
- Построить консенсус

Реализация

Серверы:

- JPRED : `http://www.compbio.dundee.ac.uk/~www-jpred/`
- PHD: `http://cubic.bioc.columbia.edu/predictprotein/`
- PSIPRED: `http://bioinf.cs.ucl.ac.uk/psipred/`
- NNPREDICT: `http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html`
- Chou and Fassman: `http://fasta.bioch.virginia.edu/fasta_www/chofas.htm`

Интересная статья:
Rost and Eyrich. EVA: Large-scale analysis of secondary structure prediction. Proteins 5:192-199 (2001)

Вопросы?

Структурный домен (биоинформатика)

Обособленная в пространстве часть белка, его структурная единица, имеющая:

- сравнительно мало контактов с другими частями белка
- собственное гидрофобное ядро

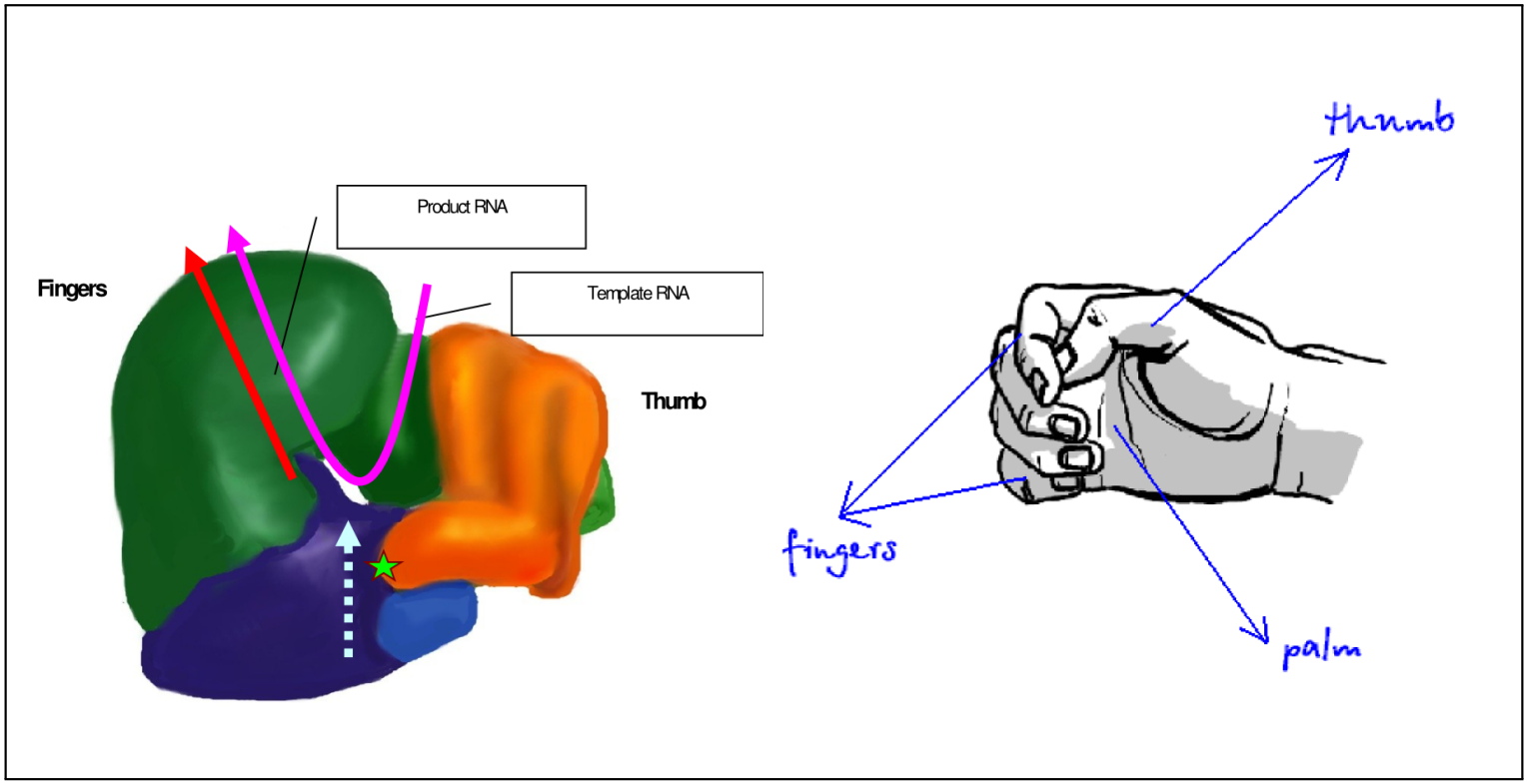
Домен белка XXX (жизнь)

Часть белка, названная доменом:

- Субъективизм
- Образность
- Традиция

Пример

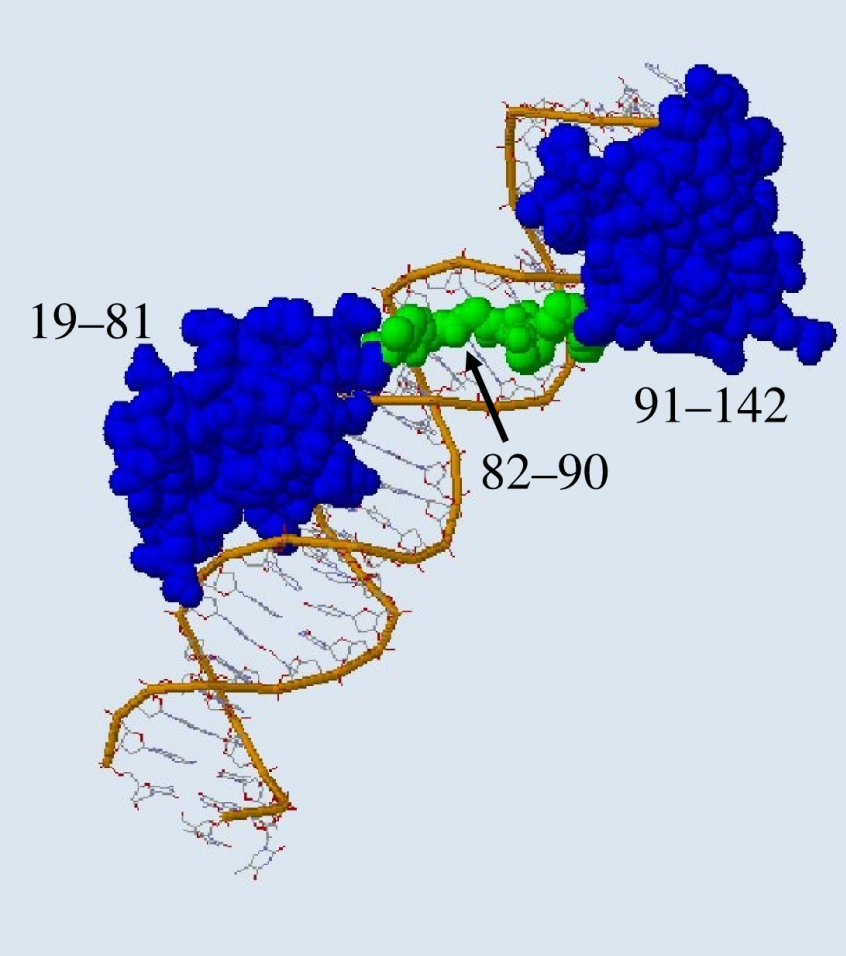
В полимеразах обычно выделяют три домена: fingers, palm, thumb



Итог

Три определения доменов часто дают похожие результаты!
Но не всегда

Пример



“Парный” (“Paired”) домен из транскрипционного фактора PAX5 человека (PDB 1K78) – очевидно, два структурных домена
Эволюционный домен (PAX в Pfam) включает оба структурных домена (126 а.о.)

Пример

Забавно, что полипептидные цепи обоих структурных доменов имеют общую топологию

- одинаковое число спиралей,
- одинаковые межспиральные взаимодействия,
- одинаковый порядок следования спиралей вдоль цепи;
- * минорные элементы вторичной структуры не в счет!

Структурные домены, Алгоритмы

Основы методов

- Домен имеет собственное гидрофобное ядро (пример: алгоритм DETECTIVE Swindells, 1995)
- Домен – это часть белка, внутри которой много контактов аминокислотных остатков, а между доменами – мало контактов (пример: алгоритм DOMAK, Siddiqui&Barton, 1995)

Siddiqui &Barton, 1995: DOMAK

Сверху – вниз, от целого – к части!

- Предпосылки: домен состоит из одного или двух непрерывных участков полипептидной цепи
- Число контактов между остатками внутри домена больше, чем число междоменных контактов

Формализация

- Два остатка контактируют, если расстояние между ними меньше 5Å
- Если белок разбит на две части, A и B, то определяется индекс разделенности:

$$SplitValue = (\frac{int_A}{ext_{AB}})(\frac{int_B}{ext_{AB}})$$

  - int_A число пар контактирующих остатков из A;
  - int_B число пар контактирующих остатков из B;
  - int_{AB} число пар контактирующих остатков, один из A, а другой – из B

Пример

Структура 1CD4. Часть A: N-конец полипептидной цепи до остатка i; часть B – от (i+1) до C-конца

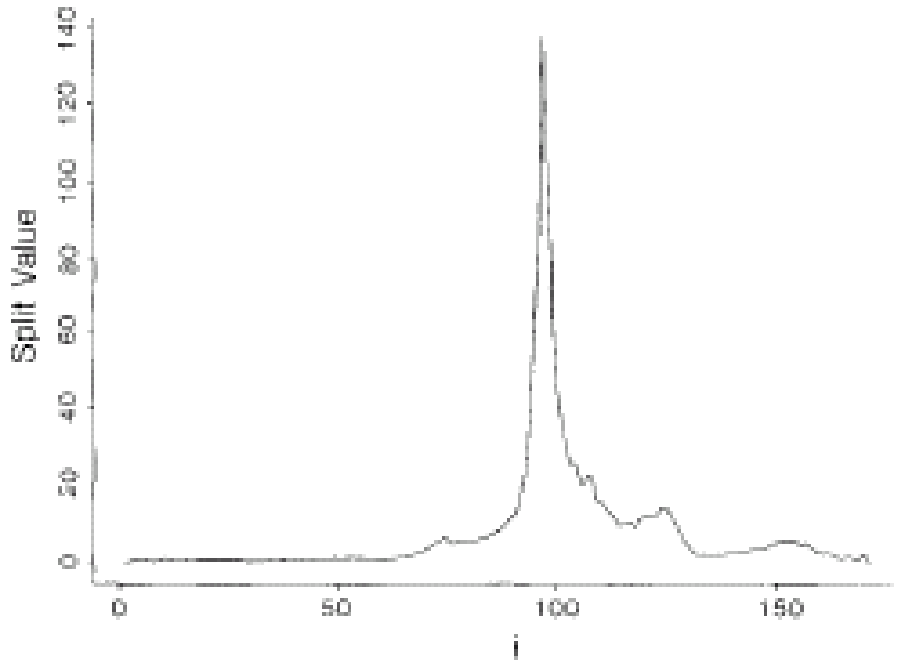


График зависимости индекса разделенности от номера граничного остатка

Алгоритм

- К полной цепи применяются 2 метода (целый или сегментный домен). Выбирается разделение с лучшим индексом
- К полученным двум доменам применяется та же процедура. В случае, когда домен состоит из двух сегментов, применяется также дополнительный метод.
- Алгоритм останавливается в зависимости от пороговых значений:
  - MDS – минимальный размер домена (в числе остатков)
  - MSS – минимальный размер сегмента
- Отдельная процедура предусмотрена для сегментов, длина которых между MDS и MSS
- Найденные домены проверяются на “компактность”; некомпактные – сливаются в один

Swindells, 1995 DETECTIVE

Снизу – вверх, наращивание частей!

Предпосылка: каждый домен имеет свое гидрофобное ядро.

Этапы:

- выявление гидрофобных ядер в структуре
- “натягивание” доменов на гидрофобные ядра

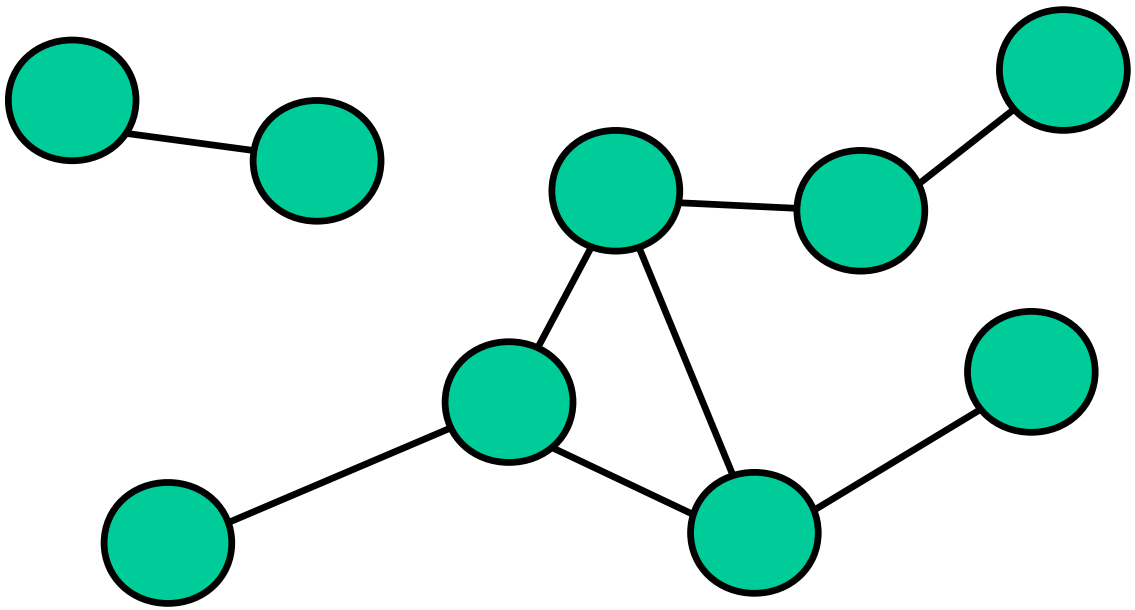
Подробности

- Отбираются остатки, которые
  - Слабо экспонированы (<7%)
  - Принадлежат спиралям или тяжам
  - Более 75% контактов их атомов с другими атомами классифицируются как гидрофобные
  - Контактом считается сближение “тяжелых” атомов на сумму vdW радиусов + 1Å
  - Гидрофобным контактом считается контакт углеродов
- Два остатка из отобранных считаются взаимодействующими гидрофобно, если число гидрофобных межатомных контактов превосходит число негидрофобных межатомных контактов

Подробности

Строится граф:

- Вершина – отобранный остаток
- Ребро соединяет вершины, если соответствующие остатки гидрофобно взаимодействуют
- Связные компоненты графа, содержащие 5 или более остатков, называются гидрофобными ядрами



Всё сложнее

Гидрофобные ядра – еще не домены!

Для получения доменов применяется много ходовая процедура чистки-слияния

Топология

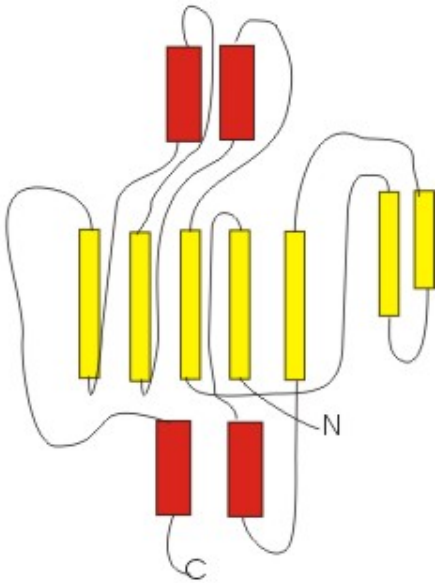
Это описание последовательности элементов вторичной структуры и их взаимного расположения в пространстве.

Белки имеют одинаковую топологию если **основные** элементы вторичной структуры расположены в последовательности в одном и том же порядке и взаимное расположение этих элементов в пространстве сходно.

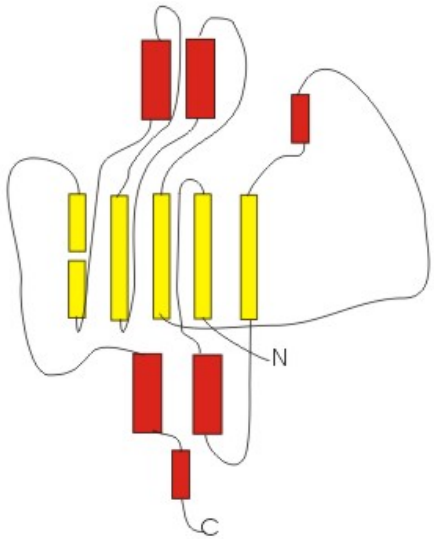
Термины “укладка” (folding) и “топология” (topology) обычно трактуются как синонимы

Пример

Каталаза (С-концевой домен)



Флаводоксин



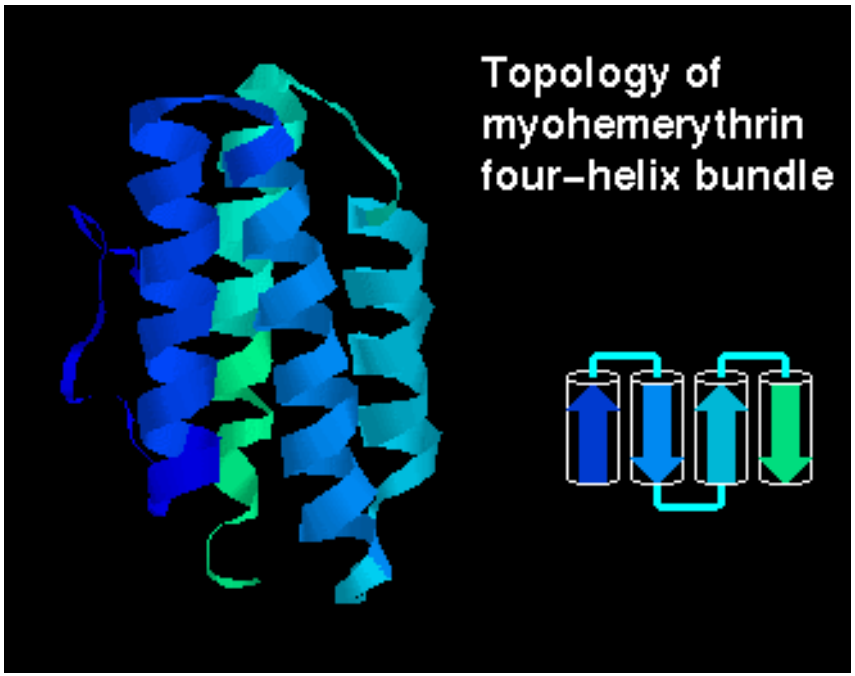
Архитектура

это описание взаимного расположения в пространстве элементов вторичной структуры без учета их последовательности в полипептидной цепи

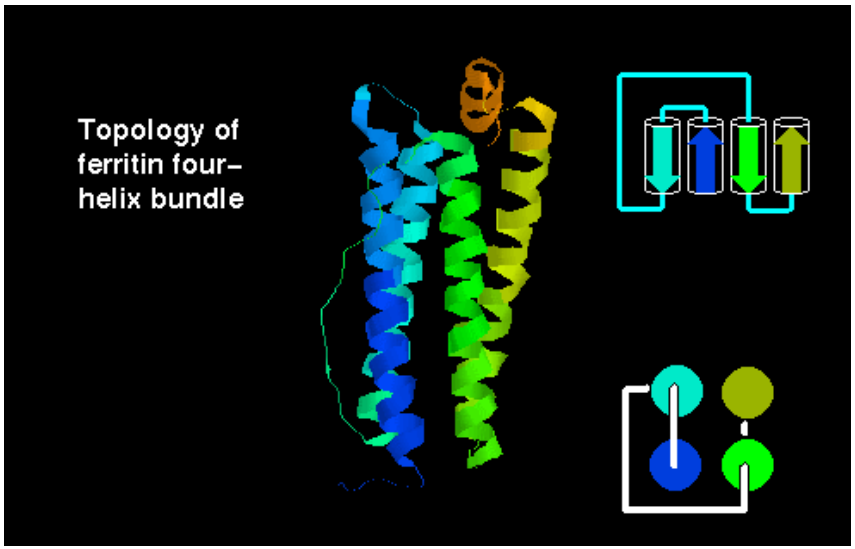
Т.о., белки с одинаковой топологией (укладкой) имеют одинаковую архитектуру, по определению.
Обратное не верно!

Пример

Архитектура – пучок 4х параллельных спиралей



Топология пучка параллельных спиралей может быть отражена диаграммой TOPS

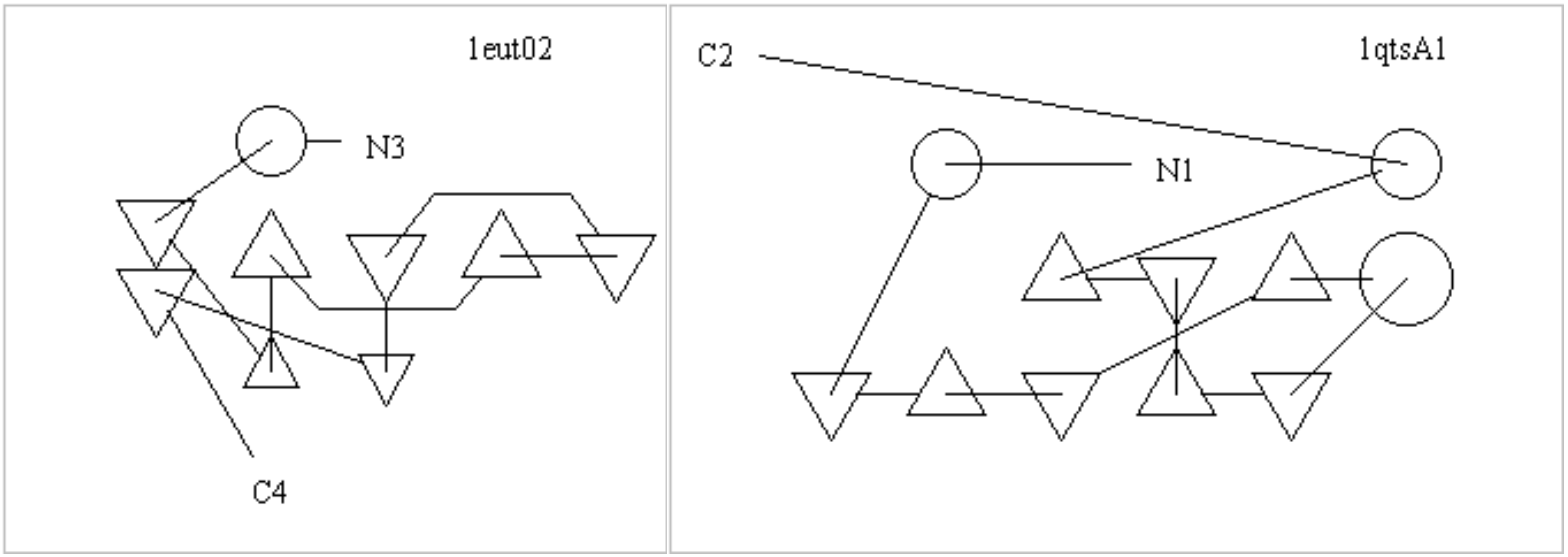


Недостаток

К сожалению, на сегодня отсутствуют адекватные универсальные способы описания топологии … Поэтому остается большой произвол в трактовке того, какие топологии встречаются в белках, сколько разных топологий и т.п.

В отличие от сравнительной однозначности в трактовке элементов вторичной структуры разными авторами и программами!

Примеры описания топологии. TOPS



На самом деле, у этих белков сходная топология

Примеры, частые в глобулярных белках

- Бета-сэндвич (sandwich)
- Бета-баррель (barrel)
- Рулет (jelly roll) – по существу, сэндвич, содержащий мотив “рулет”
- Бета-спираль (beta-helix)
- 3х (4х) спиральный узел (3helical bundle)
- Пучек спиралей (parrallel helical bundles)
- Спирализованная спираль
- Бета-цилиндр (TIM barrel)
- Укладка Россмана (Rossman fold)

Классификации структурных доменов

- SCOP (Murzin, Benner, Hubbard, Chotia, 1995)
- CATH (Orengo et al., 1993, 1997)
- FSSP (Holm&Sander, 1993)
- другие

Structural Classification of Proteins, SCOP

- Экспертное выделение доменов
- Экспертная классификация

Уровни классификации в SCOP

- Класс
- Укладка (fold) – сходная топология
- Суперсемейство – структурная гомология (?)
- Семейство – сходство последовательностей и/или хорошее пространственной выравнивание цепей
- Белок – б.м. ортологичные белковые домены
- Вид – конкретный белок

Классы

Основные

- Альфа-спиральные домены (218 укладок)
- Бета-структурные домены (144)
- Альфа/бета структурные домены (a/b) (136)
- (бета-альфа-бета структурные единицы)
- Альфа+бета домены (a+b) (279)
  - (разделенные альфа спиральные и бета-структурные области)

Классы

Специфические

- Многодоменные белки (46)
  - (сложные домены)
- Мембранные (47)
  - (кроме белков иммунной системы)
- Маленькие (75)

Формально собранные классы

- Спирализованные спирали (6)
- Низкое разрешение (24)
- Пептиды, фрагменты (116)
- Искусственные белки (42)

Scop Classification Statistics. 1.73 release

34 494 PDB Entries (Sep 2007). 97 178 Domains

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	448	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202

Class Architecture Topology Homologous superfamily, CATH

- Белок делится на домены автоматически при согласованных результатах трех алгоритмов:
  - DETECTIVE (Swindells, 1995),
  - PUU (Holm & Sander, 1994)
  - DOMAK (Siddiqui and Barton, 1995).
- При несовпадении результатов алгоритмов – решение о доменах за экспертом

CATH: уровни классификации

- Класс: основные all-alpha, all-beta, alpha-beta
- Архитектура: сходное пространственное расположение элементов вторичной структуры без учета их
- последовательности
- Топология (укладка): сходное взаимное расположение вдоль цепи и в пространстве элементов вторичной структуры
- Суперсемейство: предположительно или несомненно гомологичные домены
- Семейство: сходные последовательности (>35% identity и выровненные участки покрывают >60% длины)

Внимание

Классификации SCOP и CATH часто не совпадают

CASP

CASP - Critical Assessment of Techniques for Protein Structure Prediction

или соревнование между различными группами по предсказанию структуры белка, это происходит раз в два года начиная с 1994 (http://predictioncenter.org/).

Основная задача этого явления это выявление возможностей современных методов предсказания структуры. Для соревнования выдаётся последовательность белка, для которого известна структура, но не опубликована.

Цели CASP

В CASP7 решались следующие вопросы:

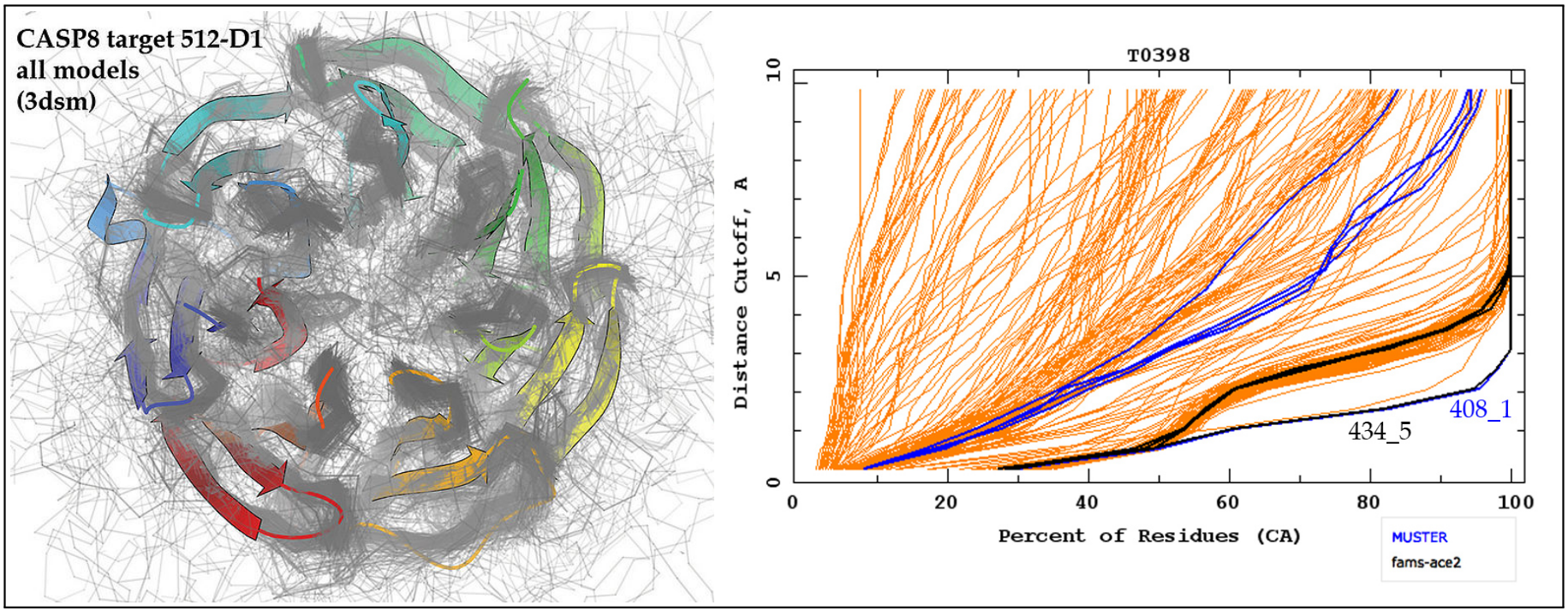
- Насколько модели похожи на экспериментальную структуру?
- Насколько верно выравнивание?
- Были ли ранее решены подобные структуры?
- Насколько гомологичное моделирование улучшает структуру, по сравнению с копированием?
- Есть ли прогресс по сравнению с предыдущими соревнованиями?
- Какие методы самые эффективные?

Категории CASP

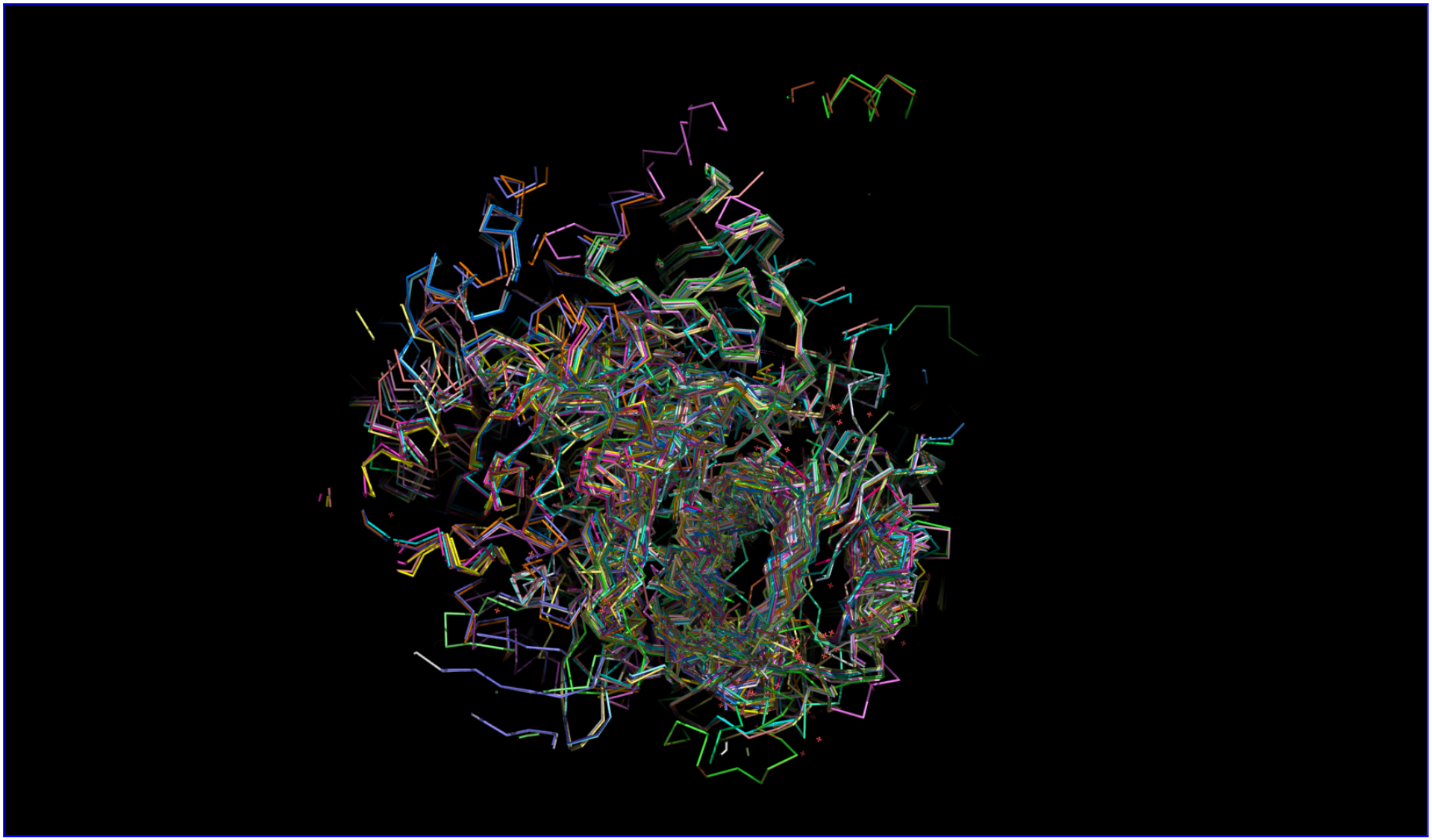
Результаты сравнивались в следующих категориях:

- 3D структура (все CASPs)
  - Гомологичное моделирование
  - Распознавание укладки.
  - de novo предсказание.
- 2D структура (прекратили после CASP5)
- Предсказание структуры комплексов (CAPRI)
- Предсказание контактов (с CASP4)
- Предсказание неструктурированных областей (с CASP5)
- Предсказание границ домена (с CASP6)
- Предсказание функции (с CASP6)
- Определение качества модели (с CASP7)
- Улучшение модели (с CASP7)

Визуализация CASP



Поиск ядра метилаз



Вопросы?