# Структурная Биоинформатика

## Лекция 3, Предсказание вторичной структуры белка

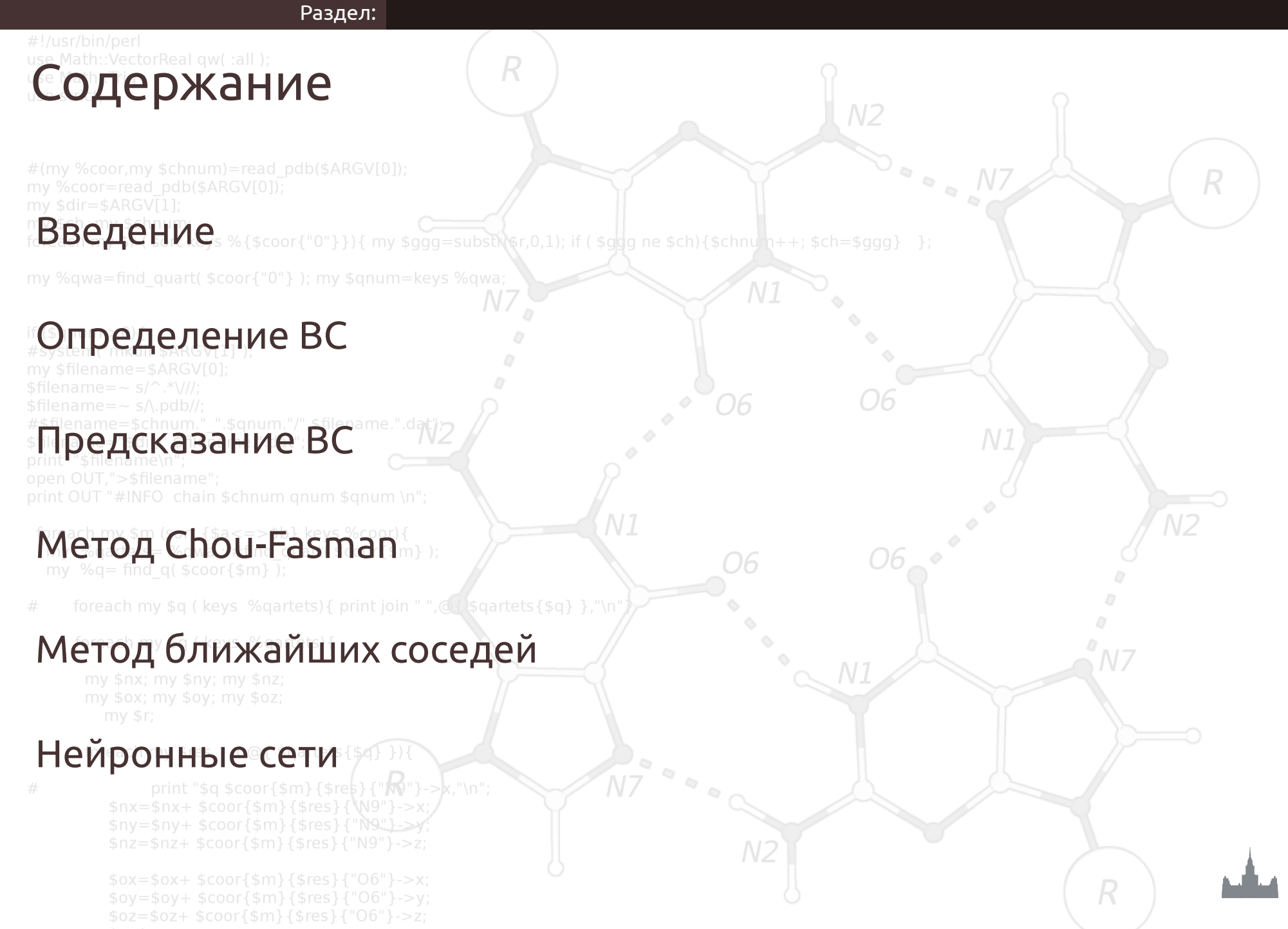
Головин А.В. [1]

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

Москва, 2012

# Содержание





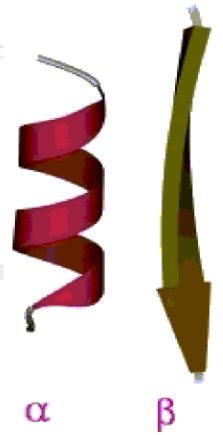
# Первичная структура

Первичная структура – это аминокислотная последовательность:

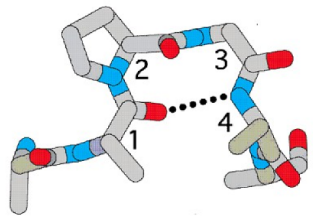
Met-Ala-Gly-Trp-Ala-Val-Asp ...

# Вторичная структура

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

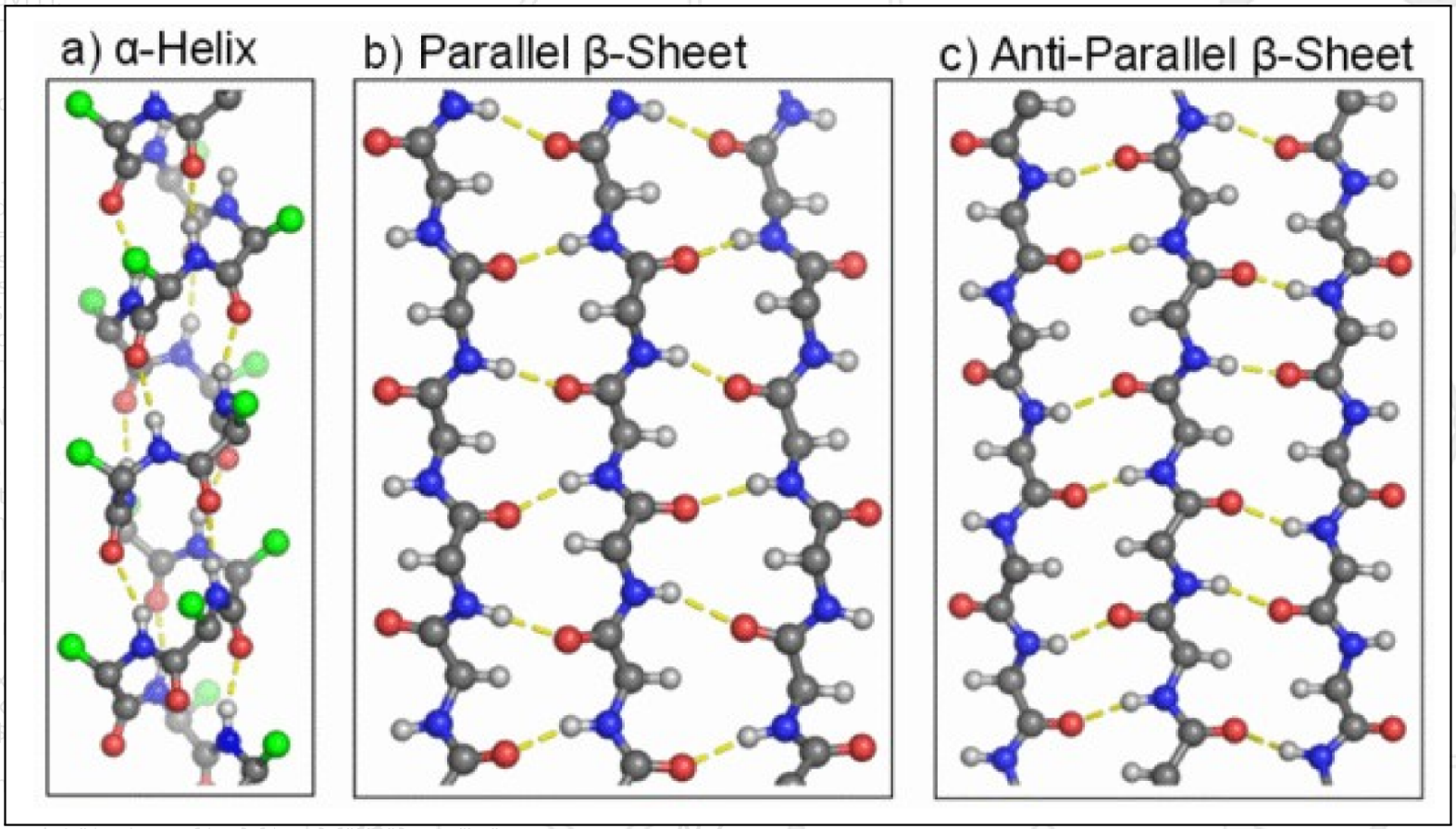



Если упорядоченность такова, что двугранные углы одинаковы у всех остатков, то говорят о регулярной вторичной структуре. Регулярными вторичными структурами являются спирали и $\beta$– структуры.




Пример нерегулярной вторичной структуры $\beta$–поворот ($\beta$–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:

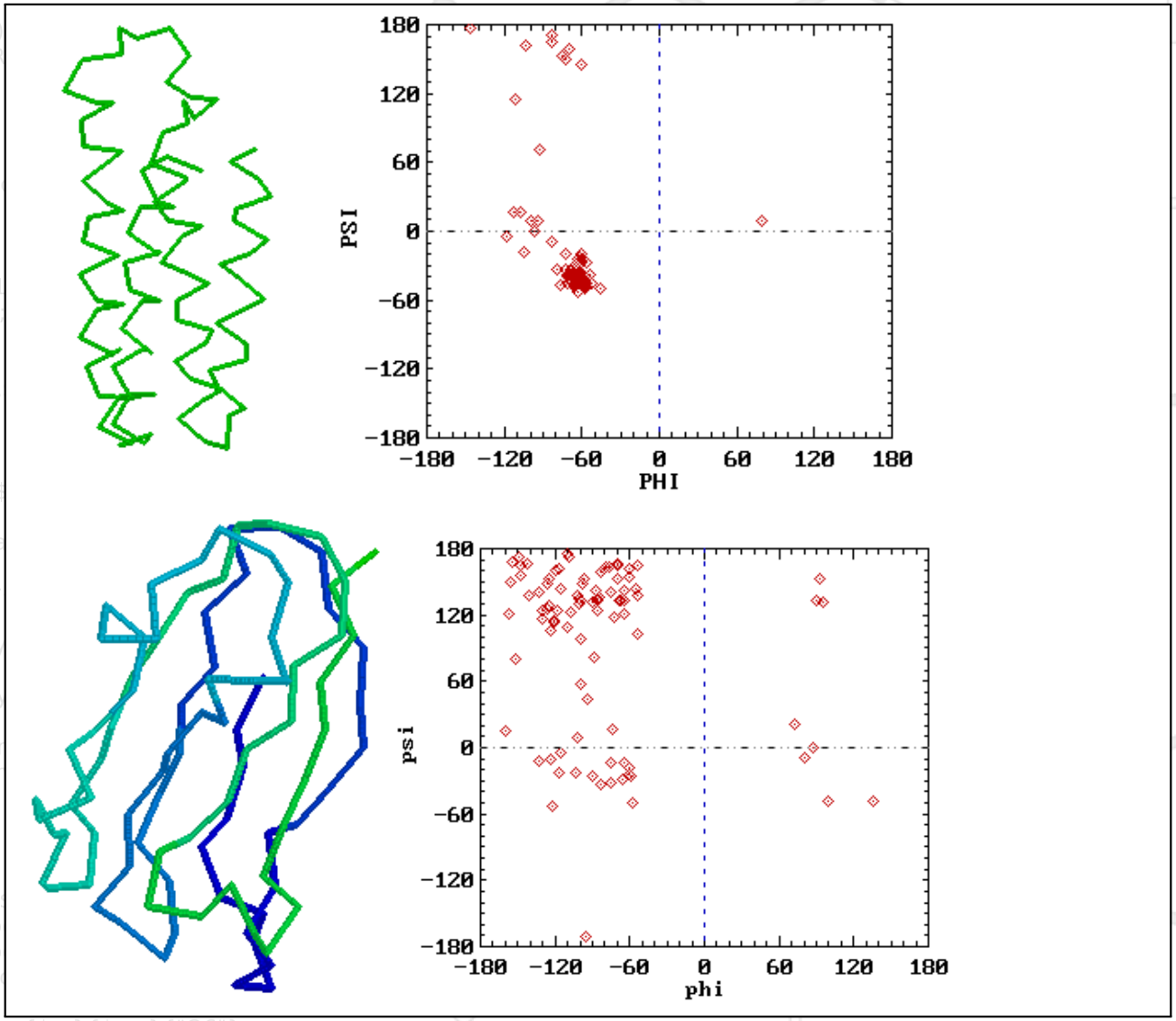
- H: $\alpha$-helix
- G: $3_{10}$ -helix
- I: $\pi$-helix
- E: $\beta$-strand
- B: bridge
- T: $\beta$-turn
- S: bend
- C: coil

CASP:

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

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 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 |

# 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: $Q_3$

$$Q_3 = \frac{correctlypredictedresidues}{numberofresidues}$$

- 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:
  - $\alpha$-спирали 5-40 остатков
  - $\beta$-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($\alpha$,aa$_i$) = 500/20000
  p($\alpha$)=4000/20000
  p(aa$_i$) = 2000/20000
  P = 500 / (4000/10) = 1.25

# Результат:

## Chou-Fasman Parameters

| $P\alpha$ | | $P\beta$ | | $Pt$ | |
|---|---|---|---|---|---|
| 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

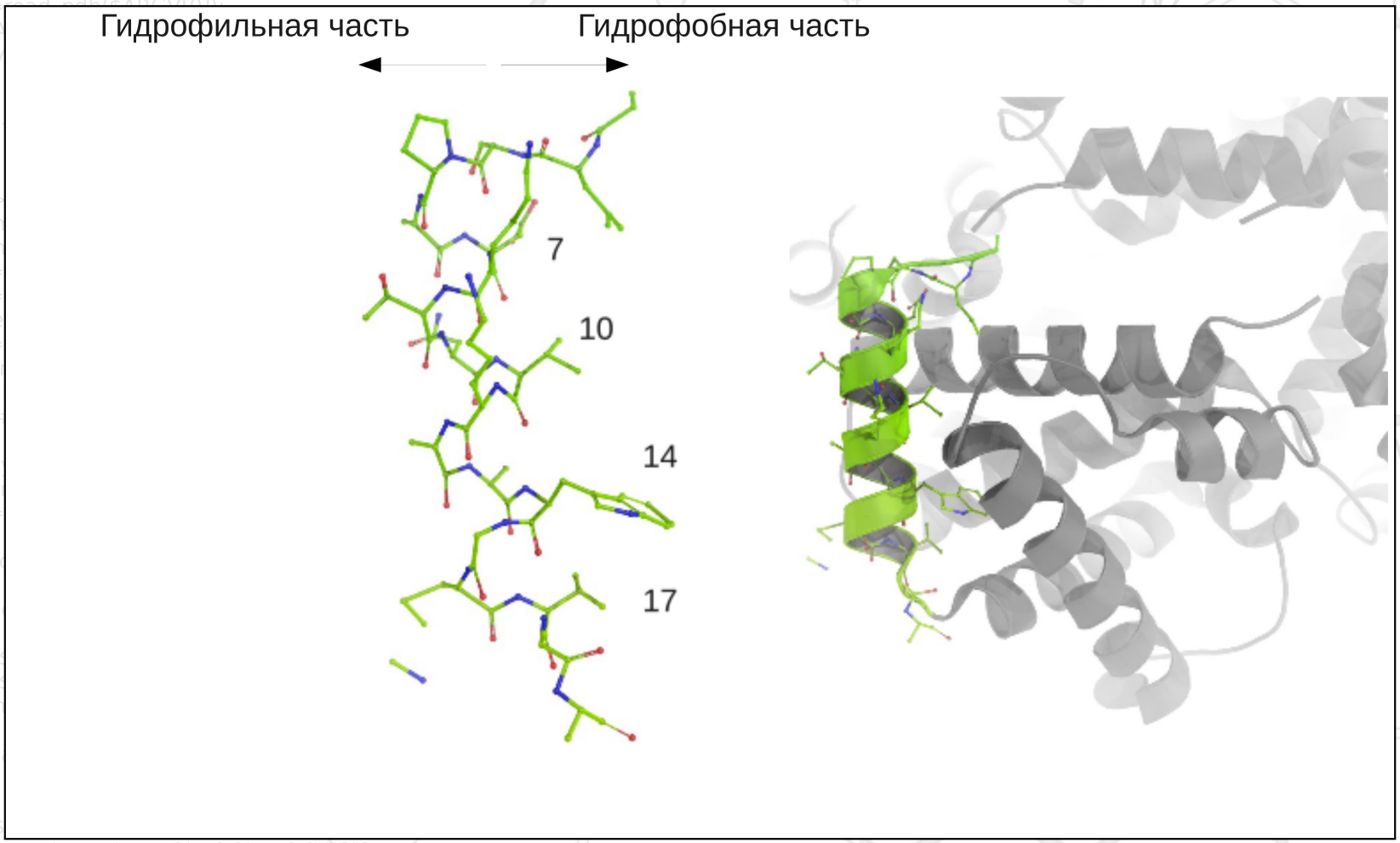
- Предсказываем ВС центрального остатка для выбранного сегмента, для которого известен ближайший гомолог
  - Из базы данных находим близкую последовательность
  - Или строим N лучших локальных выравнивания входной последовательности со всеми последовательностями базы
- Используем max($n_\alpha$, $n_\beta$, $n_c$) для ближайшего соседа или max($s_\alpha$, $s_\beta$, $s_c$) 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



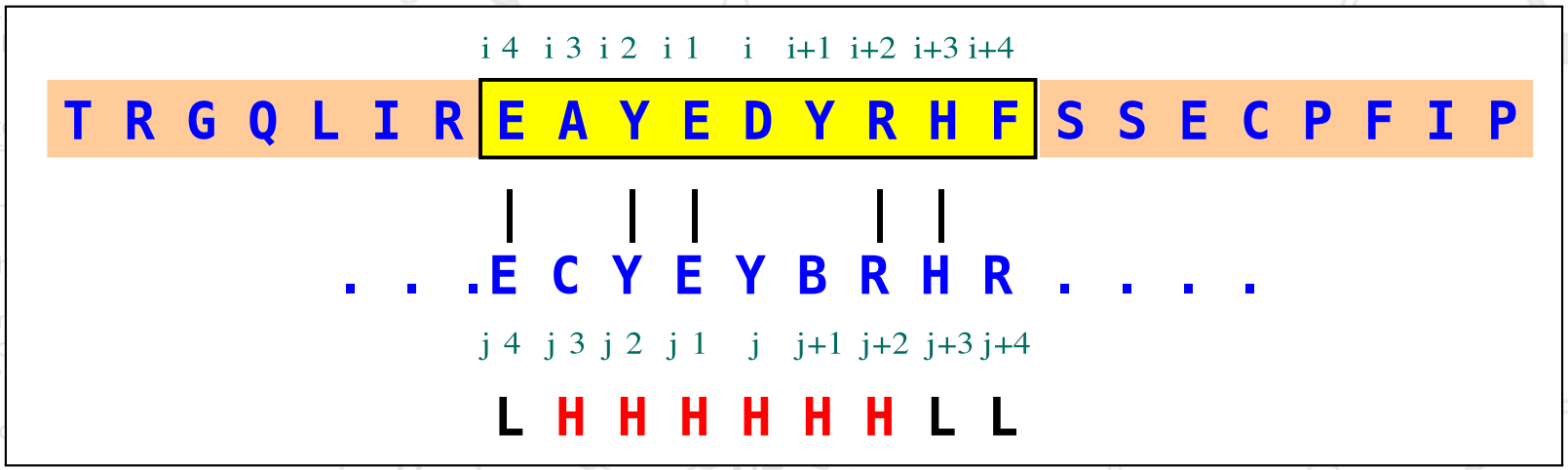
# Матрица “Singleton”

| | 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** | **-** | $S_{n+1}$ |
| | | | | | | ⋮ | | | | | | |

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

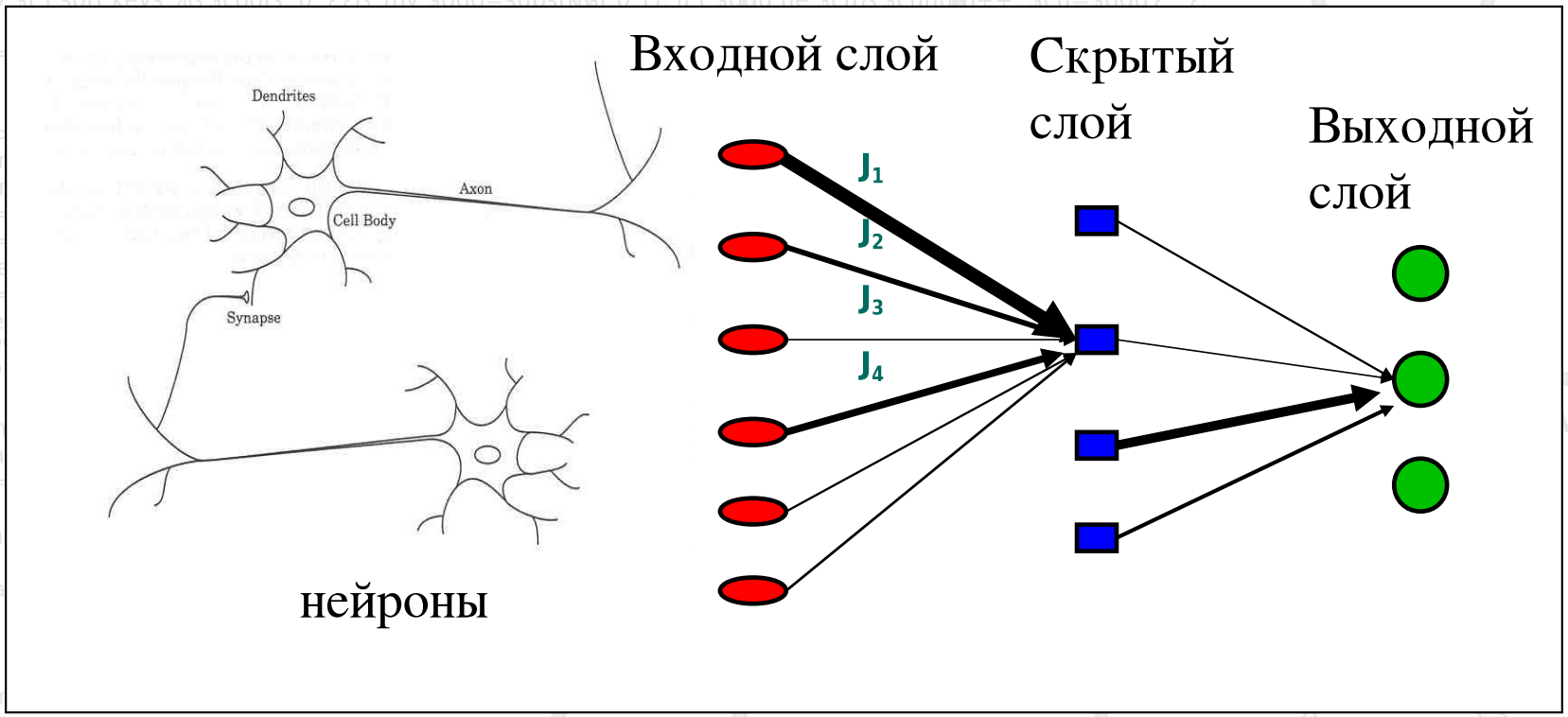
# Эволюционная информация

- "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

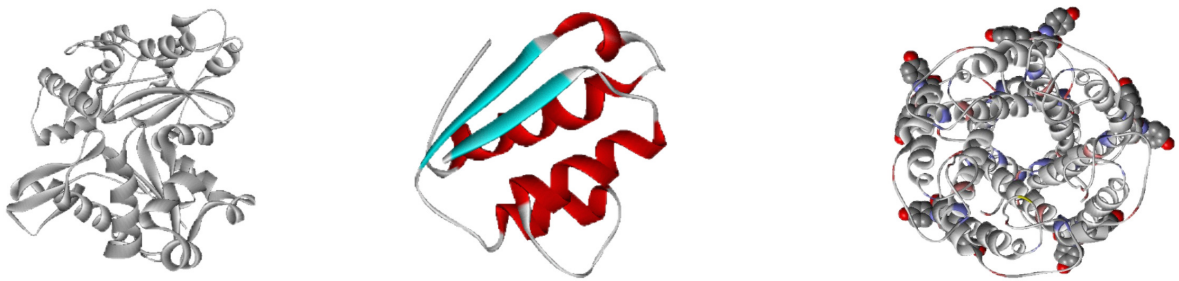
# Как использовать эту информацию:

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

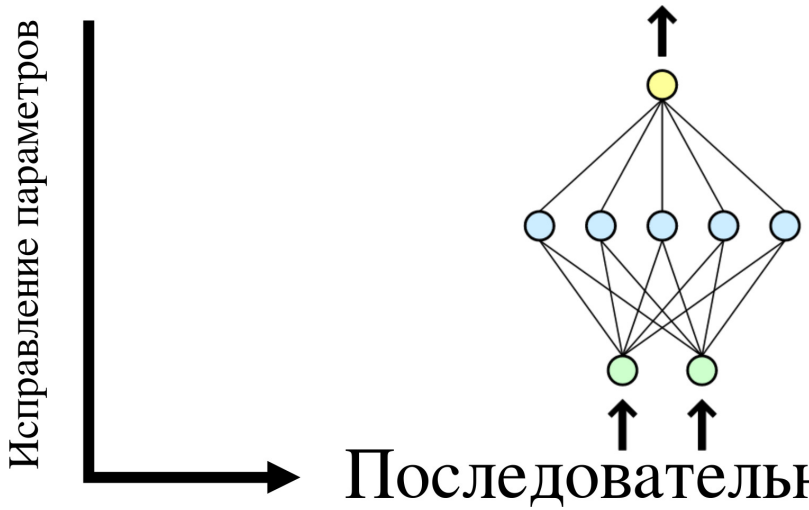
# Нейронные сети



# 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)