

Построение ранжирующей функции для прогнозирования третичной структуры белка

Карасиков Михаил Евгеньевич

Научный руководитель:
д.ф.-м.н. Стрижов В. В.
МФТИ

Консультант:
к.ф.-м.н. Максимов Ю. В.
СКОЛТЕХ

Консультант:
к.ф.-м.н. Грудинин С. В.
INRIA

Московский физико-технический институт

Москва
15 Июня 2017

Белки

Цепочки аминокислот, сворачивающиеся в **пространственные структуры** при определенных условиях

Цель работы

Изучение задачи обратного фолдинга — прогнозирования молекул белка заданной геометрии

Приложения в биологии

Определение молекул, обладающих заданными свойствами:

- лекарств,
- новых ферментов,
- самоорганизующихся белков и пептидов.

Задачи

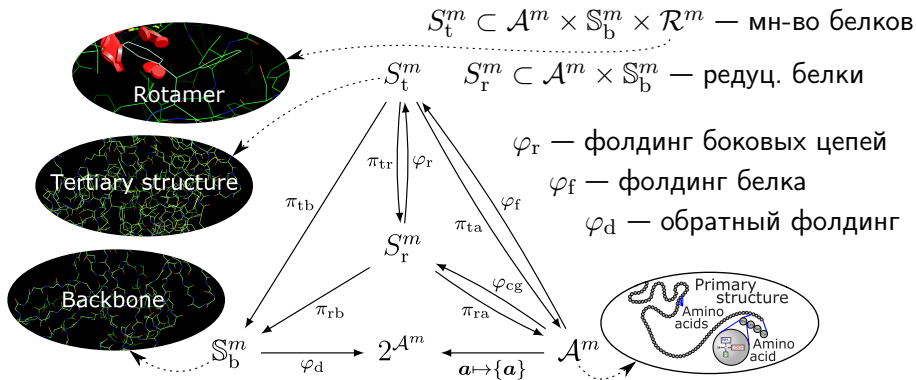
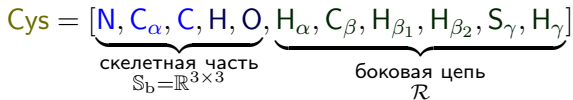
- 1 Постановка оптимизационной задачи
- 2 Введение поправок на априорное распределение аминокислот в прогнозируемых последовательностях
- 3 Решение поставленной оптимизационной задачи




Проблемы

- Целевая функция не определена
- NP-трудная задача дискретной оптимизации
- Огромная размерность
- Оценка качества требует проведения химико-биологических экспериментов

Задачи структурной биологии для белков длины m

$$\mathcal{A} = \{\text{Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, \dots, Trp, Tyr, Val}\}$$



-  Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014).
Protein folding and de novo protein design for biotechnological applications.
Trends in Biotechnology, 32(2), 99–109.
-  Samish, I., Macdermaid, C., Perez-Aguilar, J., and Saven, J. (2011).
Theoretical and computational protein design.
Annual Review of Physical Chemistry, 62(1), 129–149.
-  Liu, Y., Zeng, J., and Gong, H. (2014).
Improving the orientation-dependent statistical potential using a reference state.
Proteins, 82(10), 2383–2393.

$b \in \mathbb{S}_b^m = \mathbb{R}^{m \times 3 \times 3}$ — скелет нативной структуры

$b' \in \mathbb{S}_b^m$ — произвольный скелет белка (модельная структура)

- Среднее квадратическое отклонение

$$\underbrace{\text{RMSD}(b', b)}_{\in [0, \infty)} = \left(\frac{1}{3m} \min_{\substack{t \in \mathbb{R}^3 \\ \mathbf{S} \in \text{SO}(3)}} \sum_{i=1}^m \sum_{k=1}^3 \|b_{ik} - \mathbf{S}b'_{ik} + t\|_2^2 \right)^{1/2}$$

- Template modeling score ($\rho_{\text{TM}} = 1 - \text{TM-score}$)

$$\underbrace{\text{TM-score}(b', b)}_{\in (0, 1]} = \frac{1}{m} \max_{\substack{t \in \mathbb{R}^3 \\ \mathbf{S} \in \text{SO}(3)}} \sum_{i=1}^m \left(1 + \frac{\|b_{i2} - \mathbf{S}b'_{i2} + t\|_2^2}{d_0^2} \right)^{-1}$$

- Global distance test scores ($\rho_{\text{GDT-TS}} = 1 - \text{GDT-TS}$)

$$\underbrace{\text{GDT-TS}(b', b)}_{\in [0, 1]} = \frac{1}{4m} \max_{\substack{t \in \mathbb{R}^3 \\ \mathbf{S} \in \text{SO}(3)}} \sum_{i=1}^m \sum_{j=1}^4 \mathbb{1} [\|b_{i2} - \mathbf{S}b'_{i2} + t\|_2 < c_j],$$

$c_{1,2,3,4} = 1, 2, 4, 8\text{\AA}$, $\mathbb{1}[\cdot]$ — индикаторный $\{0, 1\}$ предикат.

Постановка задачи обратного фолдинга

Дан скелет белка $\mathbf{b}^0 \in \mathbb{S}_b^m = \mathbb{R}^{m \times 3 \times 3}$ — координаты троек атомов $[\mathbf{N}, \mathbf{C}_\alpha, \mathbf{C}]$ для m неопределенных аминокислот.

Найти аминокислотные последовательности $\mathbf{a} \in \mathcal{A}^m$, которые сворачиваются в структуры близкие к заданному скелету \mathbf{b}^0 :

$$\varphi_d(\mathbf{b}^0) = \underset{\mathbf{a} \in \mathcal{A}^m}{\text{Arg min}} \rho(\mathbf{b}^0, \underbrace{(\pi_{\text{tb}} \circ \varphi_f)(\mathbf{a})}_{\text{нат. скелет для } \mathbf{a}}).$$

Предлагается решение в два этапа

1 Аппроксимация скоринговой функции

$$S(\mathbf{a}, \mathbf{b}^0) \approx S^*(\mathbf{a}, \mathbf{b}^0) := \rho(\mathbf{b}^0, (\pi_{\text{tb}} \circ \varphi_f)(\mathbf{a}))$$

2 Оптимизация

$$S(\mathbf{a}, \mathbf{b}^0) \rightarrow \min_{\mathbf{a} \in \mathcal{A}^m}$$

Дана функция близости $\rho : \bigcup_{m=1}^{\infty} \mathbb{S}_b^m \times \mathbb{S}_b^m \rightarrow \mathbb{R}$
и набор скелетных доменов $\mathcal{D}_1, \dots, \mathcal{D}_n$:

$$\mathcal{D}_i = \{P_j^i = (\mathbf{a}^i, \mathbf{b}^{ij}) \mid j = 0, \dots, t_i\} \subset \mathcal{A}^{m_i} \times \mathbb{S}_b^{m_i},$$

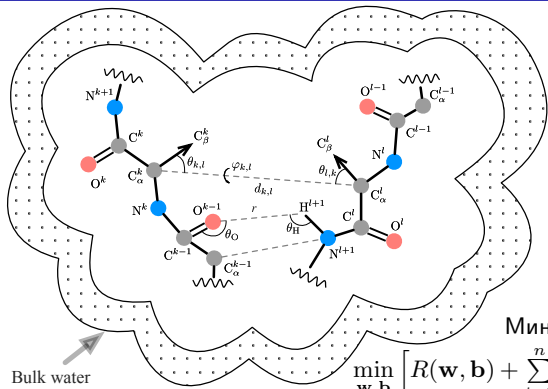
где $P_0^i = (\mathbf{a}^i, \mathbf{b}^{i0}) \in \mathcal{S}_T^{m_i}$ — нативный белок с пост-тью \mathbf{a}^i .

Построить отображение $S : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_b^m \rightarrow \mathbb{R}$ —
аппроксимирующее скоринговую функцию S^* :

$$S(P_j^i) \approx S^*(\mathbf{a}^i, \mathbf{b}^{ij}) = \rho(\mathbf{b}^{ij}, \underbrace{(\pi_{\text{tb}} \circ \varphi_{\mathbf{f}})(\mathbf{a}^i)}_{\mathbf{b}^{i0}}).$$

Критерии качества:

- $\text{Loss}(S; P_0, \mathcal{D}) = \left| \max_{P' \in \mathcal{D} \setminus \{P_0\}} S^*(P') - S^*(\arg \max_{P' \in \mathcal{D} \setminus \{P_0\}} S(P')) \right|,$
- $\text{Z-score}(S; P_0, \mathcal{D}) = \frac{S^*\left(\arg \max_{P' \in \mathcal{D} \setminus \{P_0\}} S(P')\right) - \mathbb{E}_{P \sim \mathcal{D} \setminus \{P_0\}} S^*(P)}{\sqrt{\mathbb{D}_{P \sim \mathcal{D} \setminus \{P_0\}} S^*(P)}},$
- Корреляции Пирсона, Спирмана, ранг Кенделля и др.



Извлечение признаков:

$$\mathbf{f} : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_b^m \rightarrow \mathbb{R}^k$$

Линейная модель:

$$S(P) = \langle \mathbf{w}, \mathbf{f}(P) \rangle$$

$$\tilde{\mathbf{f}}(P_j^i) := \begin{bmatrix} \mathbf{f}(P_j^i) \\ \beta \mathbf{e}_i \end{bmatrix} \in \mathbb{R}^{k+n}$$

$$\tilde{\mathbf{w}} := \begin{bmatrix} \mathbf{w} \\ \mathbf{b} \end{bmatrix} \in \mathbb{R}^{k+n}$$

Минимизация эмпирического риска:

$$\min_{\mathbf{w}, \mathbf{b}} \left[R(\mathbf{w}, \mathbf{b}) + \sum_{i=1}^n \sum_{j=0}^{t_i} L(S(P_j^i) + \mathbf{b}_i, S^*(P_j^i, P_0^i)) \right]$$

$$\alpha \left(\|\mathbf{w}\|_2^2 + \frac{1}{\beta^2} \|\mathbf{b}\|_2^2 \right) + \sum_{i=1}^n \sum_{j=0}^{t_i} (S(P_j^i) + \mathbf{b}_i - S^*(P_j^i, P_0^i))^2 \rightarrow \min_{\mathbf{w}, \mathbf{b}}$$

$$\alpha \|\tilde{\mathbf{w}}\|_2^2 + \sum_{i=1}^n \sum_{j=0}^{t_i} \left(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{f}}(P_j^i) \rangle - S^*(P_j^i, P_0^i) \right)^2 \rightarrow \min_{\tilde{\mathbf{w}}} \text{ — ридж регрессия}$$

Предложенная скоринговая функция парно-сепарабельна:

$$S(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^m \sum_{l=1}^m E_{kl}^{\mathbf{b}}(a_k, a_l) \rightarrow \min_{\mathbf{a} \in \mathcal{A}^m}.$$

Пусть $\mathcal{A} = \{a^1, \dots, a^t\}$. Сведение к BQP:

$$\sum_{k,l=1}^m E_{kl}^{\mathbf{b}}(a_k, a_l) = \sum_{k,l=1}^m \sum_{i,j=1}^t E_{kl}^{\mathbf{b}}(a^i, a^j) \underbrace{\mathbb{1}[a_k = a^i]}_{x_i^k} \underbrace{\mathbb{1}[a_l = a^j]}_{x_j^l}.$$

Положив $\mathbf{Q} = \left[[E_{kl}^{\mathbf{b}}(a^i, a^j)]_{i,j=1}^t \right]_{k,l=1}^m$, получим задачу BQP

$$\underset{\mathbf{x}}{\text{minimize}} \quad \mathbf{x}^{\top} \mathbf{Q} \mathbf{x}$$

$$\text{subject to} \quad \mathbf{x} = [\mathbf{x}^{1\top}, \dots, \mathbf{x}^{m\top}]^{\top}$$

$$\mathbf{x}^k \in \{0, 1\}^t, \quad k = 1, \dots, m,$$

$$\|\mathbf{x}^k\|_0 = 1, \quad k = 1, \dots, m.$$

Цели:

- 1 Изучение зависимости качества скоринга от объема обучающей выборки и от ядра сглаживания гистограм признаков
- 2 Сравнение качества скоринговой функции с лучшими существующими методами

Данные:

- Модельные структуры с соревнований CASP[5-11]
- По 300 NMA моделей белков для каждой нативной из CASP в RMSD диапазоне $[0.5, 6]$ Å на 100 первых нормальных модах

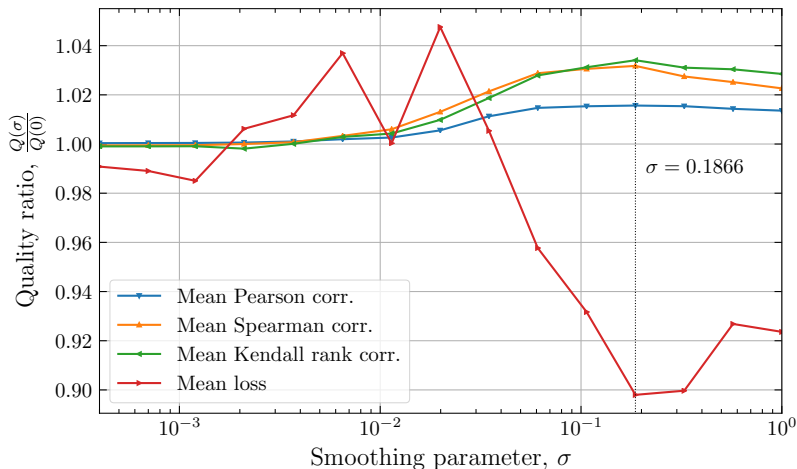


Рис.: Оценка качества структур на выборке CASP10 (stage1 и stage2 вместе) от ширины ядра сглаживания $\sigma^a = \sigma^r = \sigma^h = \sigma^s = \sigma$ при обучении на выборках CASP[5-9] без сглаживания ($\sigma = 0$).

Исследование скоринговой функции

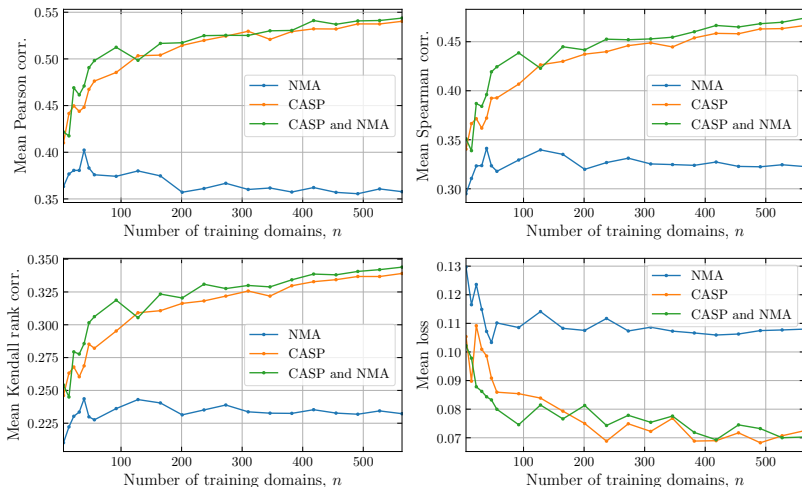


Рис.: Зависимость качества скоринговой структур от объема обучающей выборки. Обучение: случайные подвыборки CASP[5-10]. Контроль: CASP11 (stage1 и stage2 вместе).

QA Method	CASP11 Stage1			CASP11 Stage2		
	Loss	PCC	SCC	Loss	PCC	SCC
This study	0.083	0.645	0.522	0.057	0.441	0.426
ProQ2	0.090	0.643	0.506	0.058	0.372	0.366
VoroMQA	0.108	0.561	0.426	0.069	0.401	0.386
Wang-SVM	0.109	0.655	0.535	0.085	0.362	0.351
Dope	0.111	0.542	0.416	0.077	0.304	0.324
RWplus	0.135	0.536	0.433	0.084	0.295	0.314

Таблица: Качество ранжирования структур выборки CASP11. Метрики качества: Mean metric loss (Loss), коэффициент корреляции Пирсона и Спирмана (PCC и SCC) между оценками качества структур разными методами и функцией близости $\rho_{\text{GDT-TS}}$. Обучение: CASP[5-10].

- 1 Построена функция, ранжирующая 3D структуры белка
 - Является **парно-сепарабельной** скоринговой функцией
 - Использует интерпретируемую **физическую модель**
 - Использует только структуру **скелета**
 - **Робастна** к ошибкам в расстановке боковых цепей
 - Сохраняет **гладкость** скоринговой функции
 - Достигает **state-of-the-art** качества
- 2 Проведено экспериментальное сравнение выпуклых релаксаций между собой и с методами дискретной оптимизации при решении задачи обратного фолдинга и фолдинга боковых цепей
- 3 Предложены энергетические поправки для контроля частоты встречаемости различных аминокислот в предсказанных последовательностях

Максимизация апостериорного распределения

$$p(\mathbf{a}|\mathbf{b}^0) \rightarrow \max_{\mathbf{a} \in \mathcal{A}^m},$$

где правдоподобие определяется распределением Больцмана

$$p(\mathbf{b}^0|\mathbf{a}) \propto \exp\left(-\frac{\sum_{k=1}^m \sum_{l=1}^m E_{kl}^{\mathbf{b}}(a_k, a_l)}{T}\right).$$

При этом

$$p(\mathbf{b}^0|\mathbf{a}) \rightarrow \min_{\mathbf{a} \in \mathcal{A}^m} \iff \sum_{k=1}^m \sum_{l=1}^m E_{kl}^{\mathbf{b}}(a_k, a_l) \rightarrow \min_{\mathbf{a} \in \mathcal{A}^m}$$

Введем априорное распределение

$$p(a_1, \dots, a_m) = C \prod_{a \in \mathcal{A}} \mathcal{N}(m_a | m p_a, m \sigma_a^2),$$

где $\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $m_a = \sum_{i=1}^m \mathbb{1}[a_i = a]$, $a \in \mathcal{A}$,

p_a и σ_a для каждого $a \in \mathcal{A}$ — параметры распределения,
 C — константа нормализации.

Лемма (Энергетические поправки, Карасиков 2016)

Пусть априорное распределение $p(a_1, \dots, a_m)$ задается формулой $p(a_1, \dots, a_m) = C \prod_{a \in \mathcal{A}} \mathcal{N}(m_a | mp_a, m\sigma_a^2)$.

Тогда, задача максимизации апостериорного распределения $p(\mathbf{a} | \mathbf{b}^0) \rightarrow \max_{\mathbf{a} \in \mathcal{A}^m}$ эквивалентна минимизации полной энергии

$$\sum_{k=1}^m \sum_{l=1}^m [E_{kl}(a_k, a_l) + E'_{kl}(a_k, a_l)] \rightarrow \min_{a_1, \dots, a_m},$$

где энергетические поправки введены следующими образом:

$$E'_{kl}(a_k, a_l) := \begin{cases} \frac{T}{2m} \cdot \frac{1-2p_{a_k}}{\sigma_{a_k}^2}, & a_k = a_l; \\ -\frac{T}{2m} \cdot \left(\frac{p_{a_k}}{\sigma_{a_k}^2} + \frac{p_{a_l}}{\sigma_{a_l}^2} \right), & a_k \neq a_l. \end{cases}$$

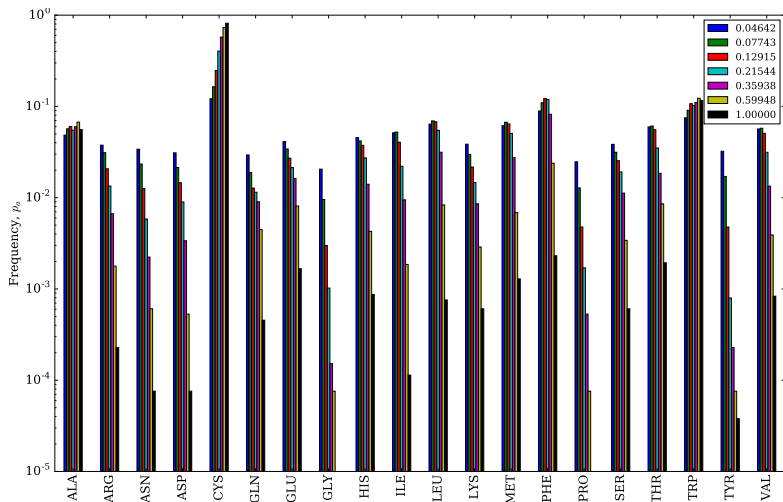


Рис.: Встречаемость кислотных остатков в решениях при различных параметрах $\beta = 1/T$. Усреднение по выборке SCWRL4.

$$\begin{aligned} & \underset{\mathbf{x} \in \{0,1\}^n}{\text{minimize}} && \mathbf{x}^\top \mathbf{Q} \mathbf{x} \\ & \text{subject to} && \mathbf{A} \mathbf{x} = \mathbf{1}_m, \end{aligned}$$

1 Непрерывная

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{x}^\top (\mathbf{Q} - \lambda_{\min} \mathbf{I}_n) \mathbf{x} + \lambda_{\min} \mathbf{1}_n^\top \mathbf{x} \mid \mathbf{A} \mathbf{x} = \mathbf{1}_m, \mathbf{0}_n \leq \mathbf{x} \leq \mathbf{1}_n \right\}$$

2 Лагранжева (двойственная)

$$\min_{\lambda \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m} \left\{ \gamma - r(\mathbf{u}) \mid \gamma \leq 0, \begin{bmatrix} \mathbf{P}(\lambda) & \frac{1}{2} \mathbf{q}(\lambda, \mathbf{u}) \\ \frac{1}{2} \mathbf{q}^\top(\lambda, \mathbf{u}) & -\gamma \end{bmatrix} \in \mathcal{S}_+^{n+1} \right\}$$

3 Положительно полуопределенная

$$\min_{\mathbf{x}, \mathbf{X}} \left\{ \text{Tr}(\mathbf{Q} \mathbf{X}) \mid \begin{array}{l} \mathbf{A} \mathbf{x} = \mathbf{1}_m, \\ \mathbf{A} \mathbf{X} = \mathbf{1}_m \mathbf{x}^\top, \\ X_{ij} \in [0, 1], \\ X_{ii} = x_i, \\ i, j = 1, \dots, n, \end{array} \begin{bmatrix} \mathbf{X} & \mathbf{x} \\ \mathbf{x}^\top & 1 \end{bmatrix} \in \mathcal{S}_+^{n+1} \right\}$$

Семплирование для полуопределенной релаксации

$$\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \underbrace{\mathbf{X} - \mathbf{x}\mathbf{x}^\top}_{\in \mathcal{S}_+^n})$$

Округление

Проекция $\hat{\mathbf{x}} \in \text{Proj}_V \mathbf{x}$, где $V = \{\mathbf{x} \in \{0, 1\}^n \mid \mathbf{A}\mathbf{x} = \mathbf{1}_m\}$ вычисляется по формуле:

$$\hat{x}_i^k := \begin{cases} 1, & i = \arg \max_{j=1, \dots, t} \hat{x}_j^k, \\ 0, & \text{otherwise,} \end{cases}$$

где $k = 1, \dots, m$, $\arg \max_j \hat{x}_j^k = \min(\text{Arg max}_j \hat{x}_j^k)$.

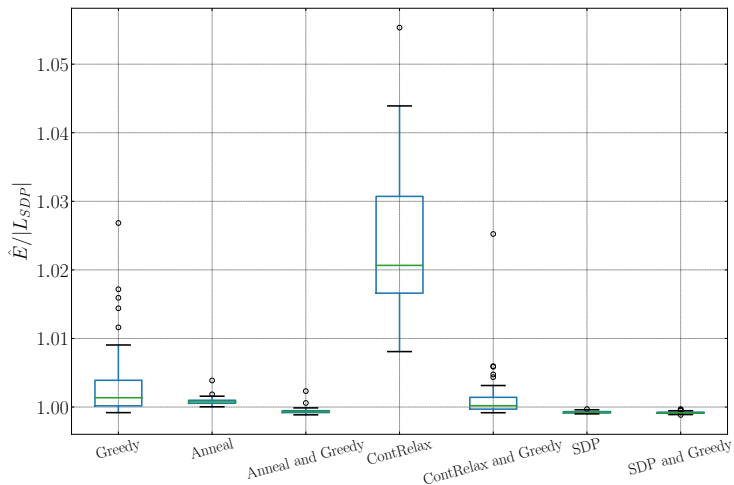


Рис.: Box plots for normalized approximate optimal values obtained by different optimization methods. Averaged over first 40 structures from the SCWRL4 dataset.

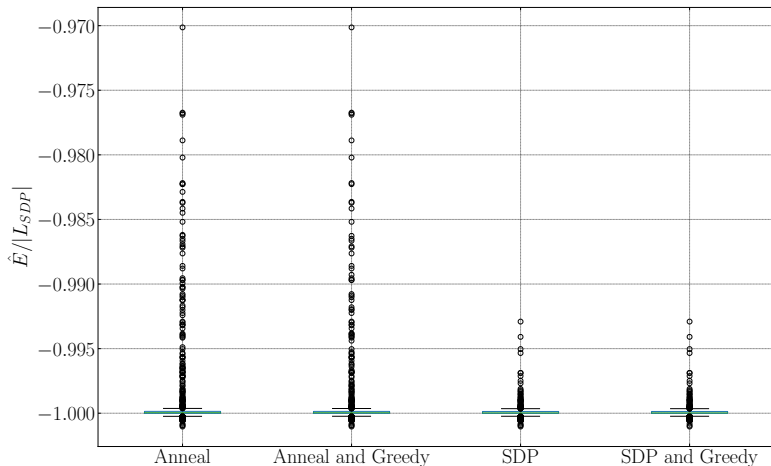


Рис.: Box plots for normalized approximate optimal values obtained by different optimization methods. Averaged over structures from the SCWRL4 dataset and sequence lengths $m = 5, 10, 15, 20, 25, 30$.

Резерв: оптимизация в задаче обратного фолдинга

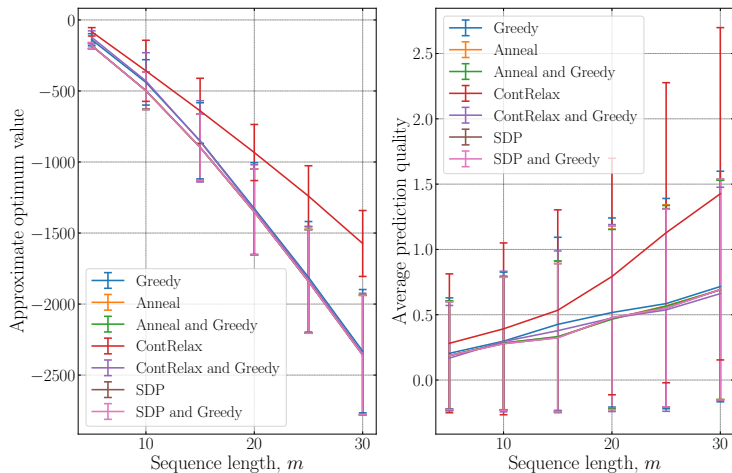
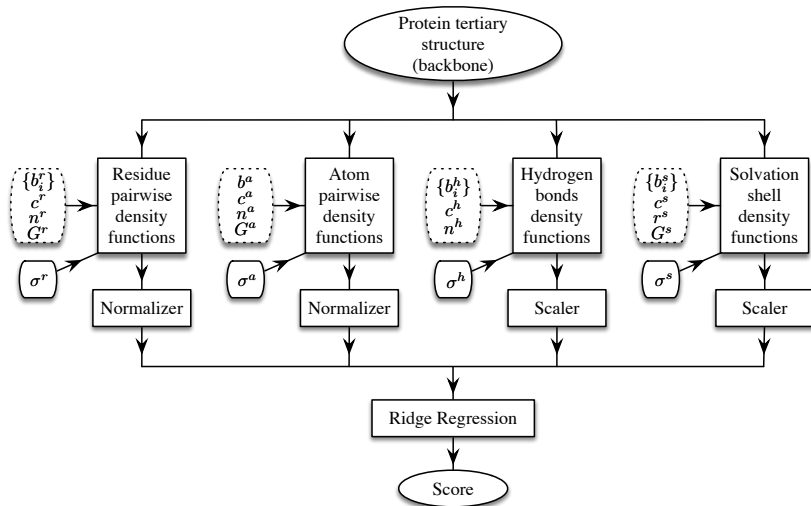


Рис.: Upper bounds on the optimal value and Average ratio of correctly predicted amino acids. Averaged over 352 protein structures in the SCWRL4 dataset. DFIRE- C_α (Zhang et al., 2004) energy function.



Резерв: распределения для функций близости

