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

Задача

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.

Обратный фолдинг как задача оптимизации

Задачи

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

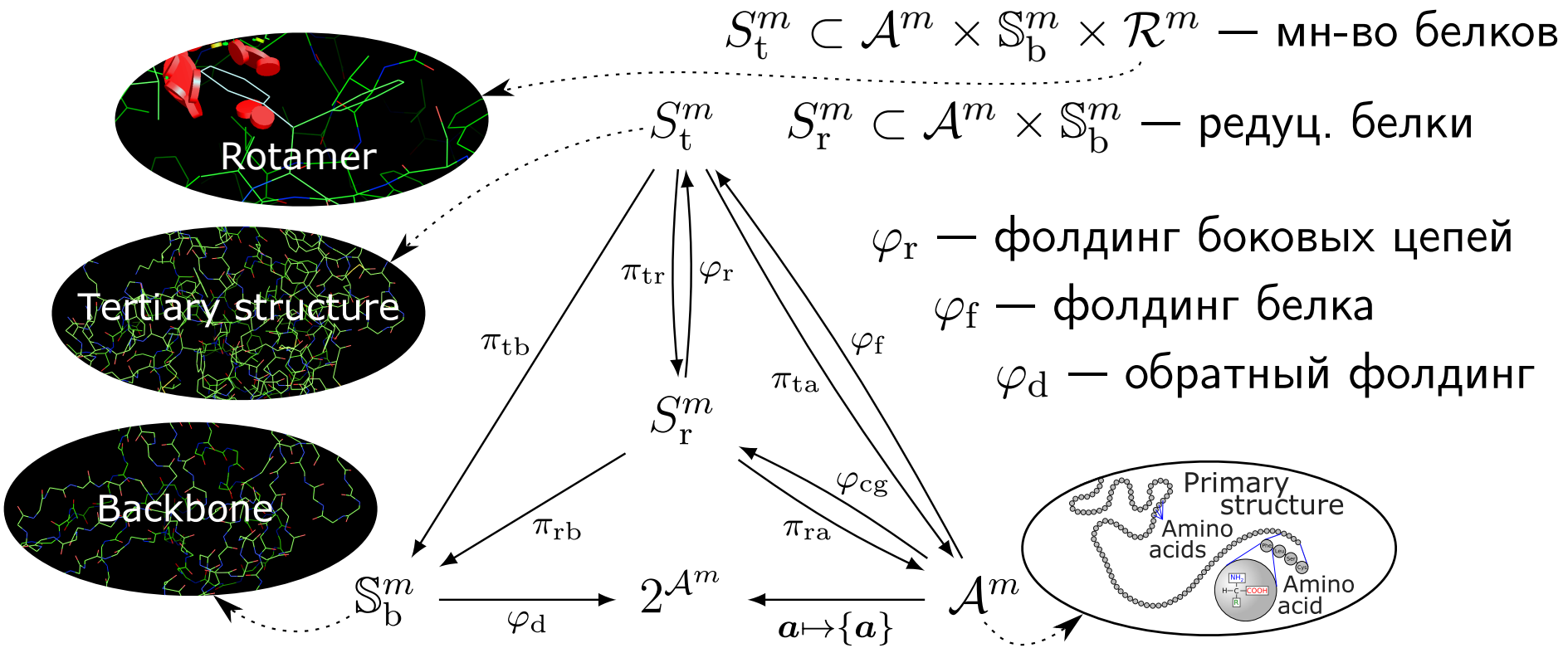
Проблемы

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

Задачи структурной биологии для белков длины m

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

$$\mathsf{Cys} = [\underbrace{\mathsf{N}, \mathsf{C}_\alpha, \mathsf{C}, \mathsf{H}, \mathsf{O}}_{\substack{\text{скелетная часть} \\ \mathbb{S}_\mathrm{b}=\mathbb{R}^{3\times 3}}}, \underbrace{\mathsf{H}_\alpha, \mathsf{C}_\beta, \mathsf{H}_{\beta_1}, \mathsf{H}_{\beta_2}, \mathsf{S}_\gamma, \mathsf{H}_\gamma}_{\substack{\text{боковая цепь} \\ \mathcal{R}}}]$$



Литература


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.

Функции близости скелетов белка $\rho(\boldsymbol{b}', \boldsymbol{b})$

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

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

- Среднее квадратическое отклонение

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

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

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

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

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

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

Постановка задачи обратного фолдинга

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

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

$$\varphi_{\mathrm{d}}(\boldsymbol{b}^0) = \operatorname*{Arg\,min}_{\boldsymbol{a}\in\mathcal{A}^m} \rho(\boldsymbol{b}^0, \underbrace{(\pi_{\mathrm{tb}} \circ \varphi_{\mathrm{f}})(\boldsymbol{a})}_{\text{нат. скелет для } \boldsymbol{a}}).$$

Предлагается решение в два этапа

1. Аппроксимация скоринговой функции

$$S(\boldsymbol{a}, \boldsymbol{b}^0) \approx S^*(\boldsymbol{a}, \boldsymbol{b}^0) := \rho(\boldsymbol{b}^0, (\pi_{\mathrm{tb}} \circ \varphi_{\mathrm{f}})(\boldsymbol{a}))$$

2. Оптимизация

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

Оценка качества скелетов белка — аппроксимация $S^*(a, b)$

Дана функция близости $\rho : \bigcup_{m=1}^{\infty} \mathbb{S}_{\mathrm{b}}^m \times \mathbb{S}_{\mathrm{b}}^m \to \mathbb{R}$
и набор скелетных доменов $\mathcal{D}_1, \ldots, \mathcal{D}_n$:

$$\mathcal{D}_i = \left\{ P_j^i = (\boldsymbol{a}^i, \boldsymbol{b}^{ij}) \mid j = 0, \ldots, t_i \right\} \subset \mathcal{A}^{m_i} \times \mathbb{S}_{\mathrm{b}}^{m_i},$$

где $P_0^i = (\boldsymbol{a}^i, \boldsymbol{b}^{i0}) \in S_{\mathrm{r}}^{m_i}$ — нативный белок с пос-тью $\boldsymbol{a}^i$.

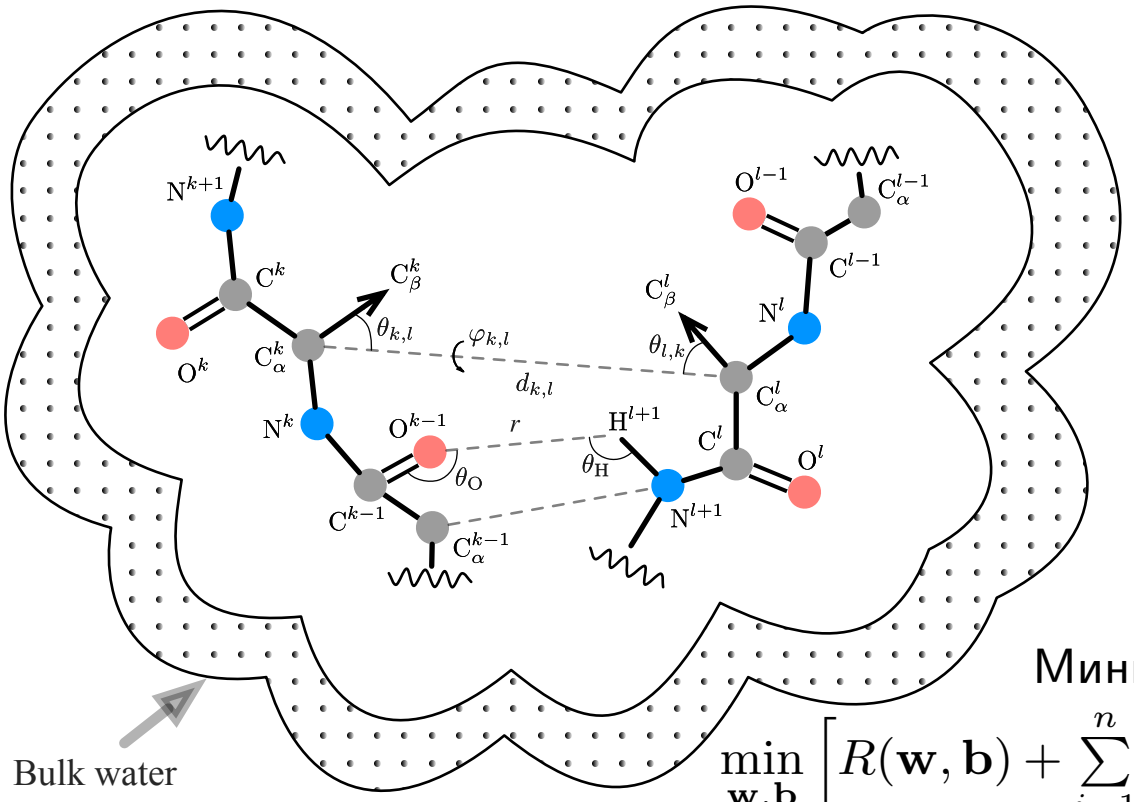
Построить отображение $S : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_{\mathrm{b}}^m \to \mathbb{R}$ —
аппроксимирующее скоринговую функцию S^*:

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

Критерии качества:

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

Модель и признаки



Извлечение признаков:

$$\mathbf{f} : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_{\mathrm{b}}^m \to \mathbb{R}^k$$

Линейная модель:

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

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

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

Минимизация эмпирического риска:

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

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

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

Обратный фолдинг как оптимизация скоринговой функции

Предложенная скоринговая функция парно-сепарабельна:

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

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

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

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

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{x}^{\mathsf{T}} \mathbf{Q} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x} = [\boldsymbol{x}^{1\mathsf{T}}, \ldots, \boldsymbol{x}^{m\mathsf{T}}]^{\mathsf{T}} \\ & \boldsymbol{x}^k \in \{0,1\}^t, \quad k = 1, \ldots, m, \\ & \|\boldsymbol{x}^k\|_0 = 1, \quad k = 1, \ldots, m. \end{aligned}$$

Вычислительный эксперимент

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 $[0.5, 6]\text{A}$ на 100 первых нормальных модах

Зависимость качества скоринговой функции от сглаживания

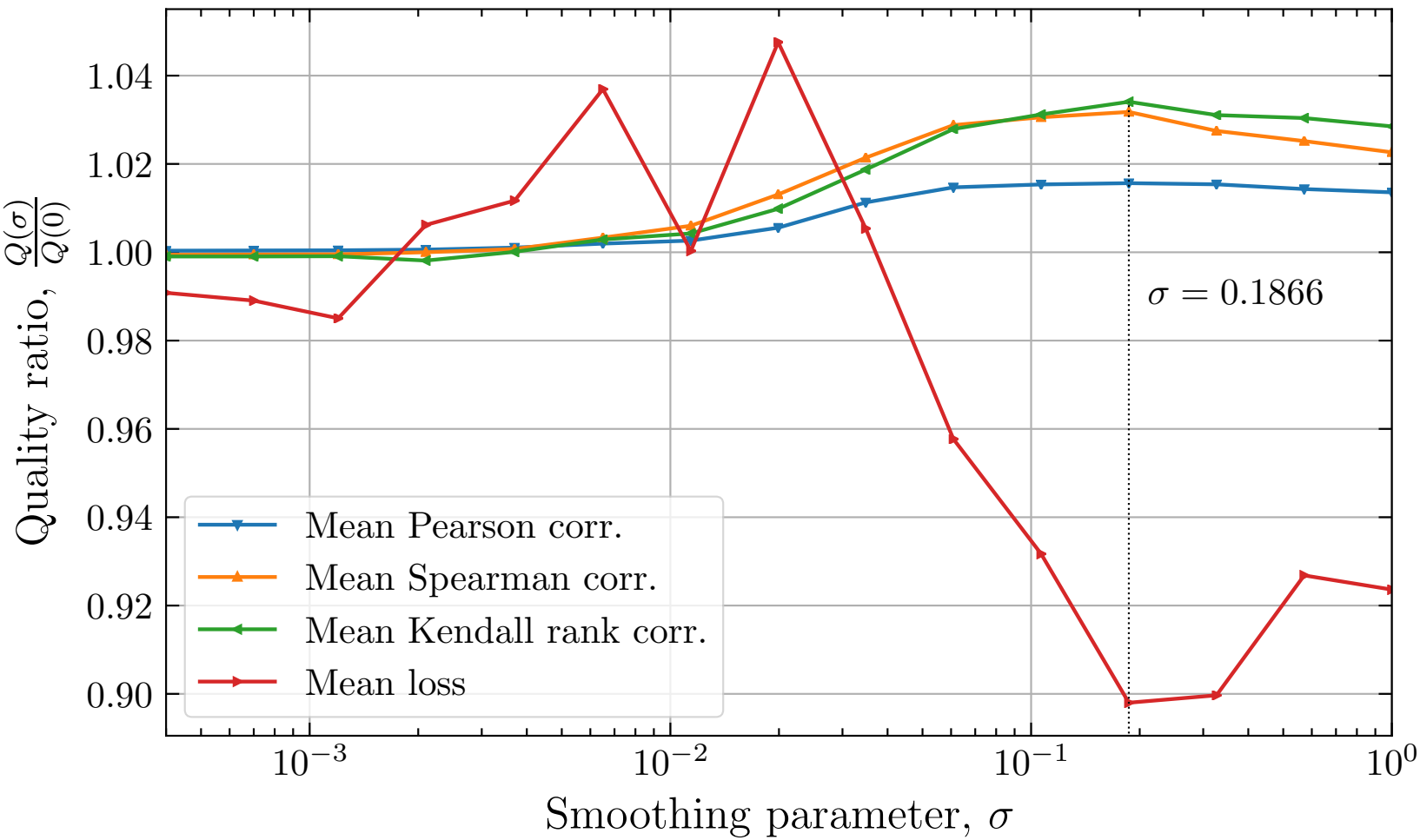


Рис.: Оценка качества структур на выборке CASP10 (stage1 и stage2 вместе) от ширины ядра сглаживания $\sigma^a = \sigma^r = \sigma^h = \sigma^s = \sigma$ при обучении на выборках CASP[5-9] без сглаживания ($\sigma = 0$).

Исследование скоринговой функции

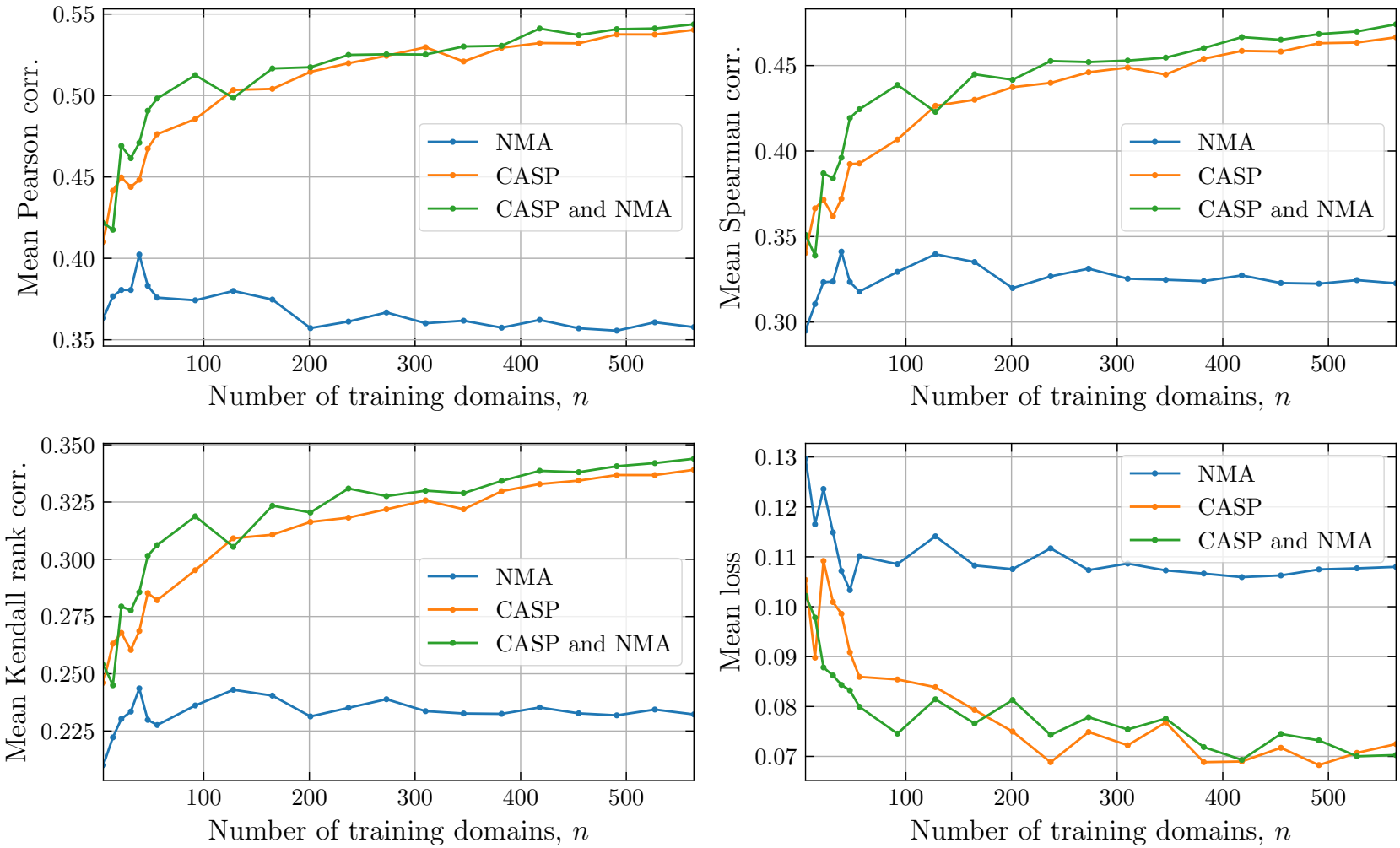


Рис.: Зависимость качества скоринговой структур от объема обучающей выборки. Обучение: случайные подвыборки CASP[5-10]. Контроль: CASP11 (stage1 и stage2 вместе).

Сравнение

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].

Заключение. Результаты, выносимые на защиту

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