

Новосибирский государственный университет
Факультет естественных наук
Кафедра информационной биологии

Компьютерный анализ количественной взаимосвязи между структурой Ca^{2+} - регулируемых фотопротеинов и их биолюминесцентными свойствами

Выполнил: Зинченко Д. И., гр. 18410
Научный руководитель: Иванисенко Н. В., к.б.н., н. с., лаб. компьютерной
протеомики

Билюминесценция



Aequorea victoria

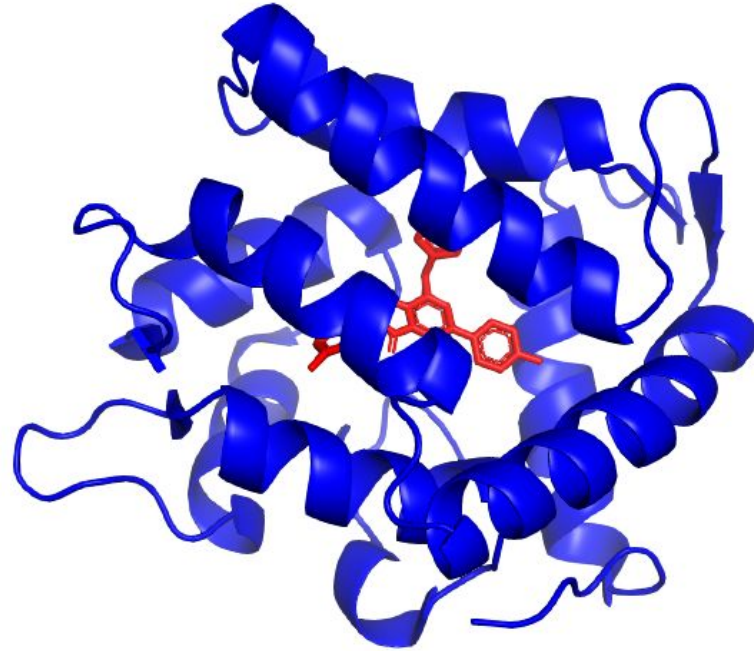


Lampyris noctiluca



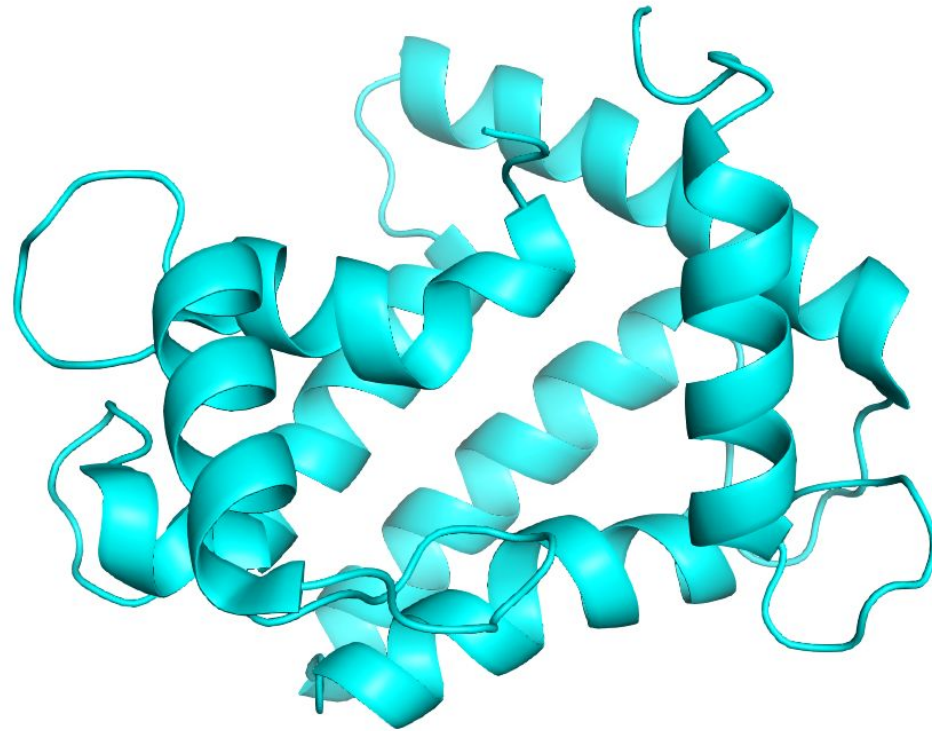
Mycena chlorophos

Ca²⁺-регулируемые фотопротеины



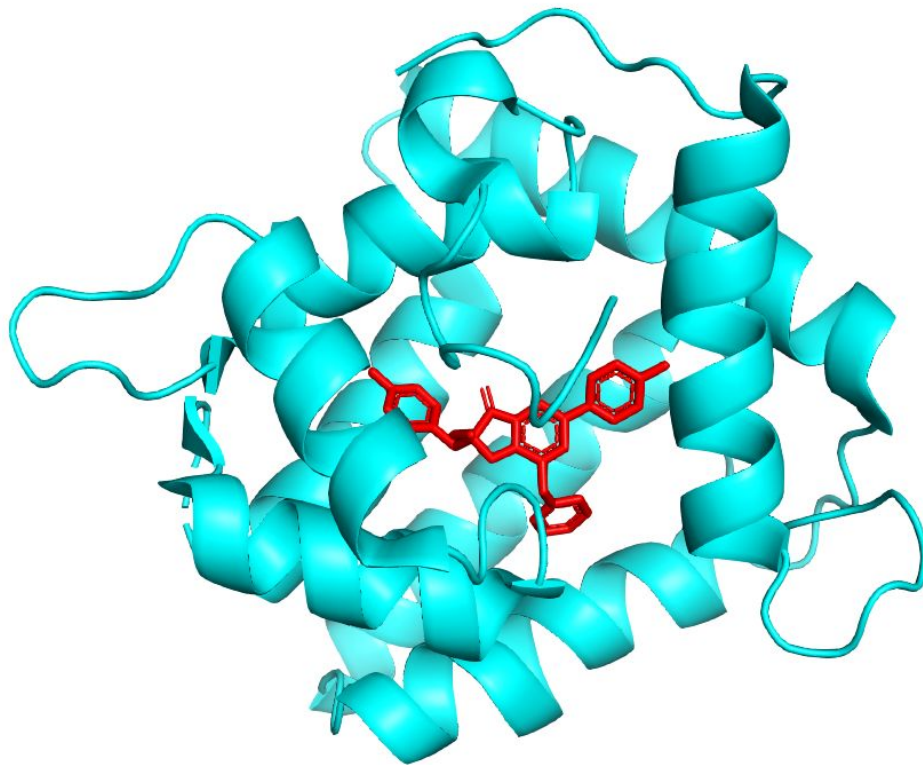
Активный аквапорин (PDB 1EJ3):
синий — апоаквапорин,
красный — 2-гидропероксицелентеразин

Реакция биолюминесценции



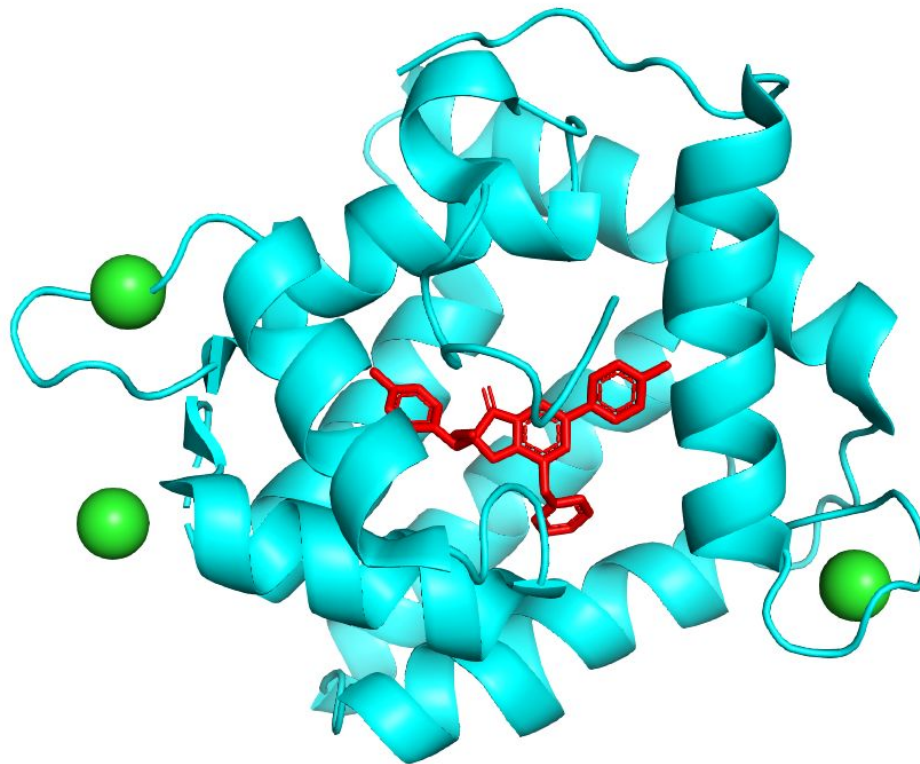
Апообелин (PDB 1SL7)

Реакция биолюминесценции



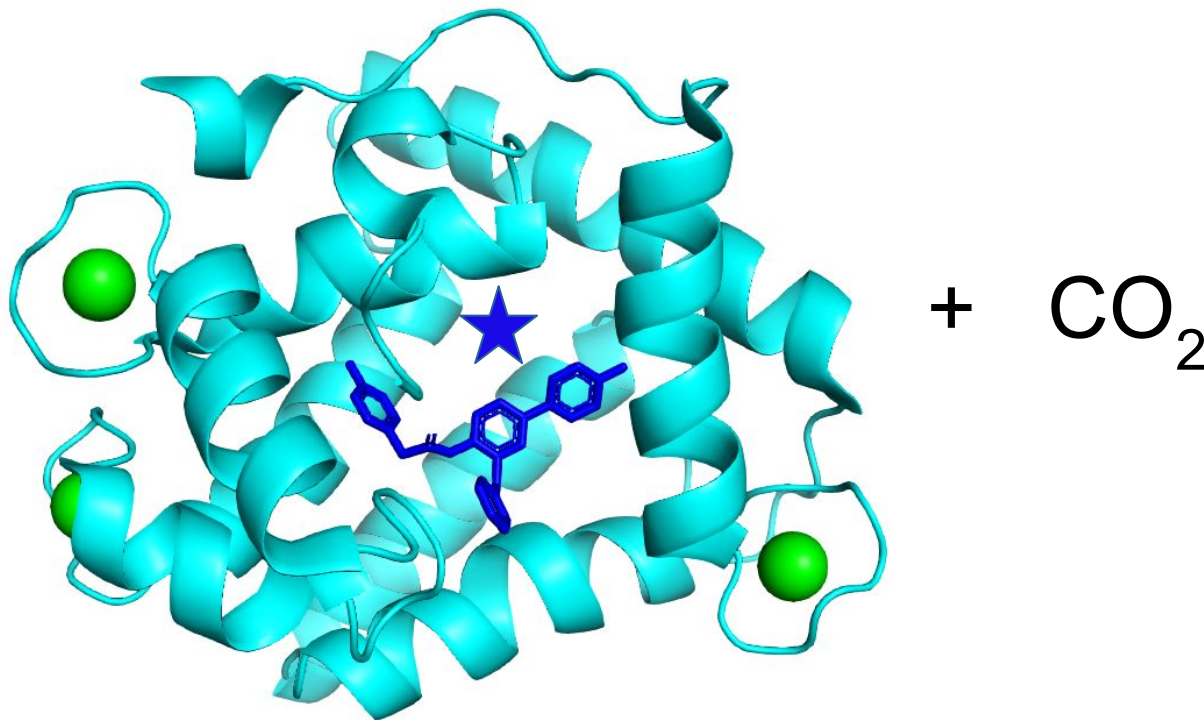
Активный обелин (PDB 1QV0)

Реакция биолюминесценции



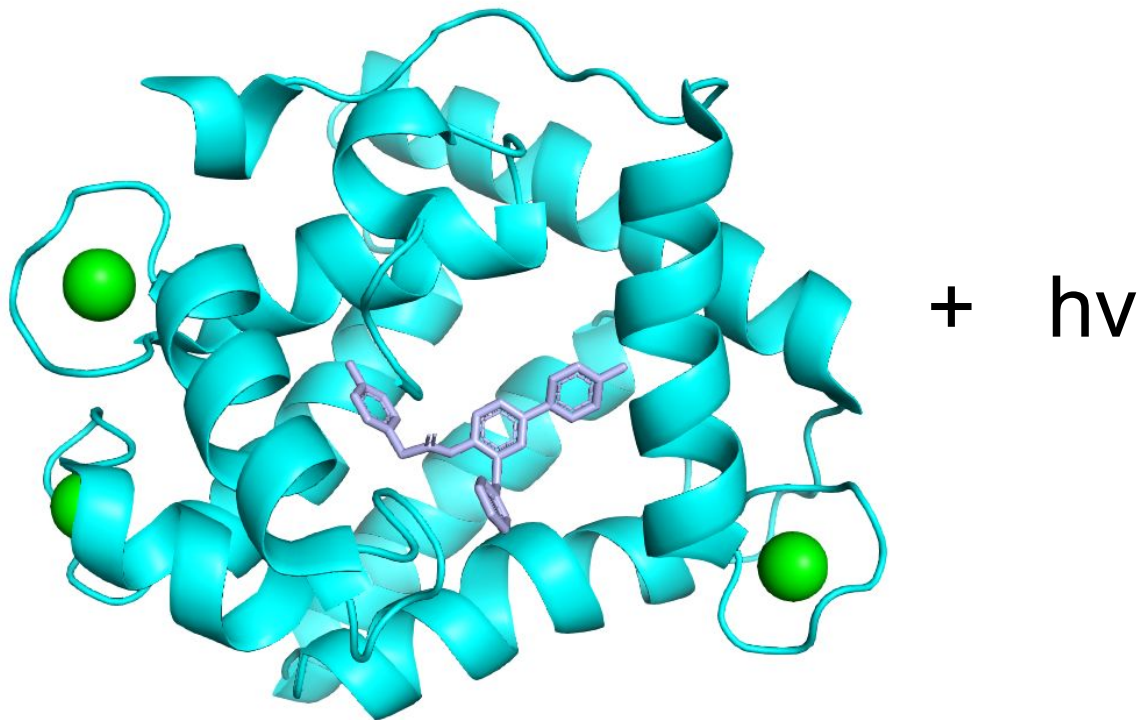
Активный обелин с кальцием (PDB 1QV0)

Реакция биолюминесценции



Обелин с целентерамидом в возбуждённом состоянии (PDB 2F8P)

Реакция биолюминесценции



Обелин с целентерамидом в основном состоянии (PDB 2F8P)

Спектр билюминесценции

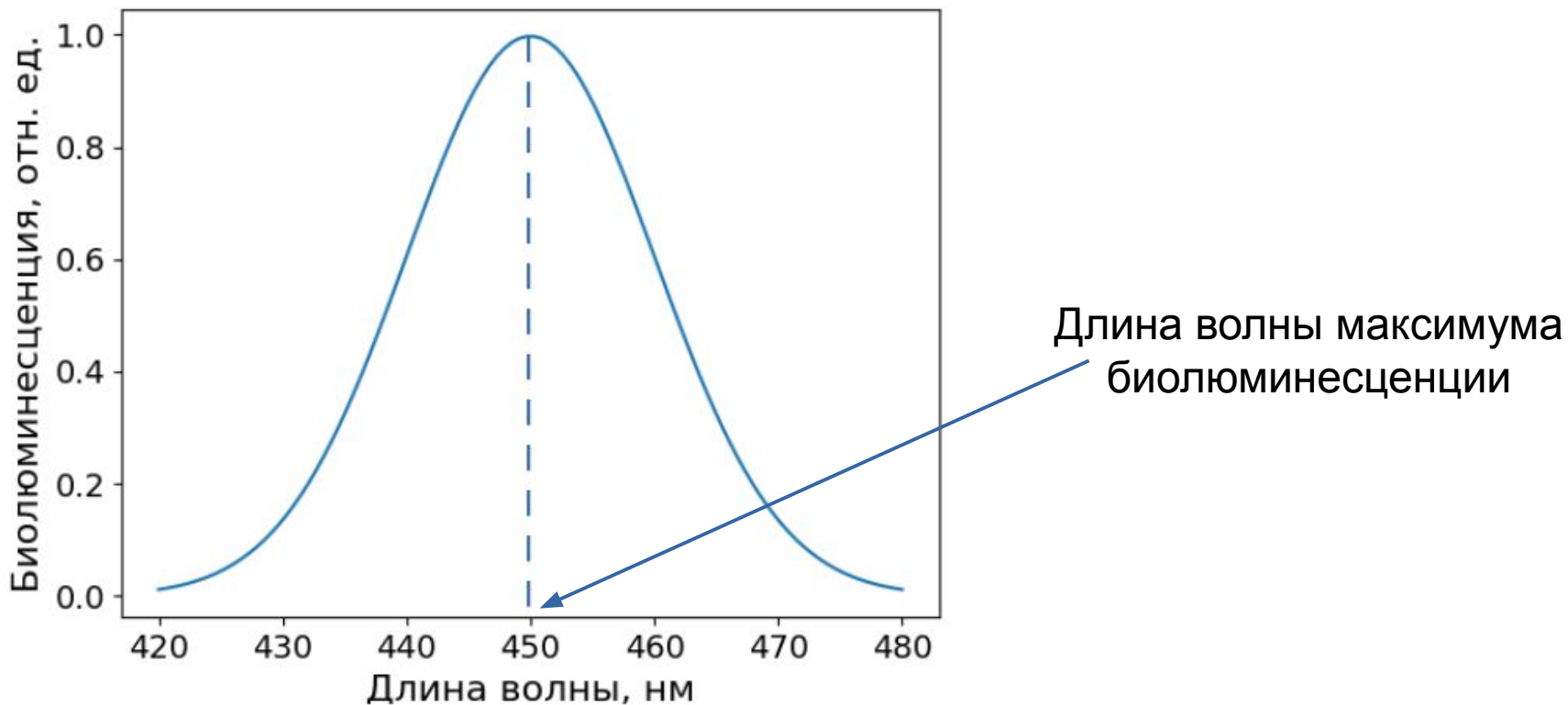
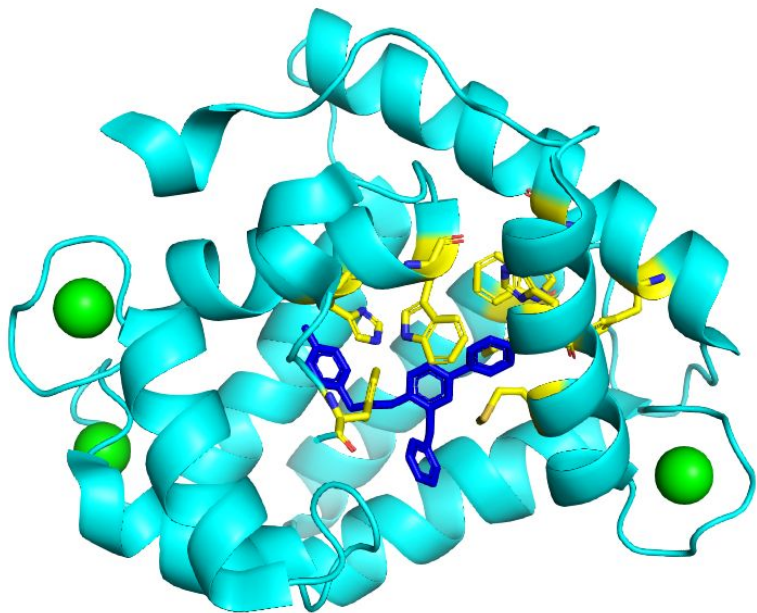


Схема спектра билюминесценции
фотопротеина

Спектр биолюминесценции



Жёлтый — аминокислотные остатки,
оказывающие влияние на биолюминесценцию

Длина волны максимума
биолюминесценции зависит
от аминокислотных остатков,
оказывающих влияние на биолюминесценцию

Спектр билюминесценции

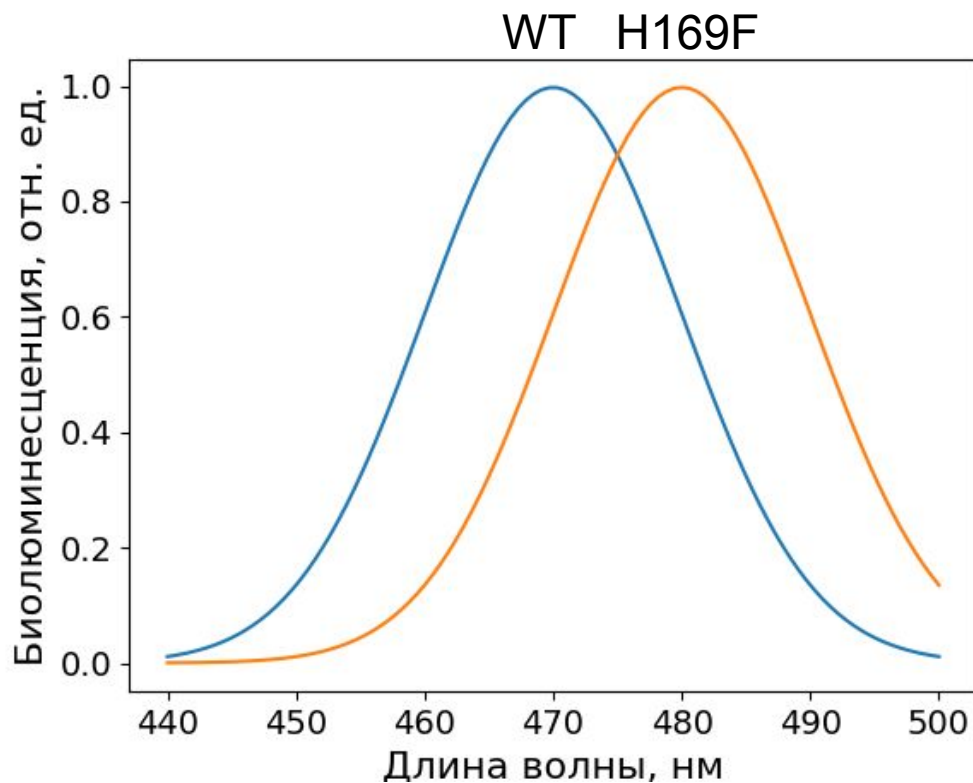


Схема спектра билюминесценции
природного акворина и его мутанта H169F

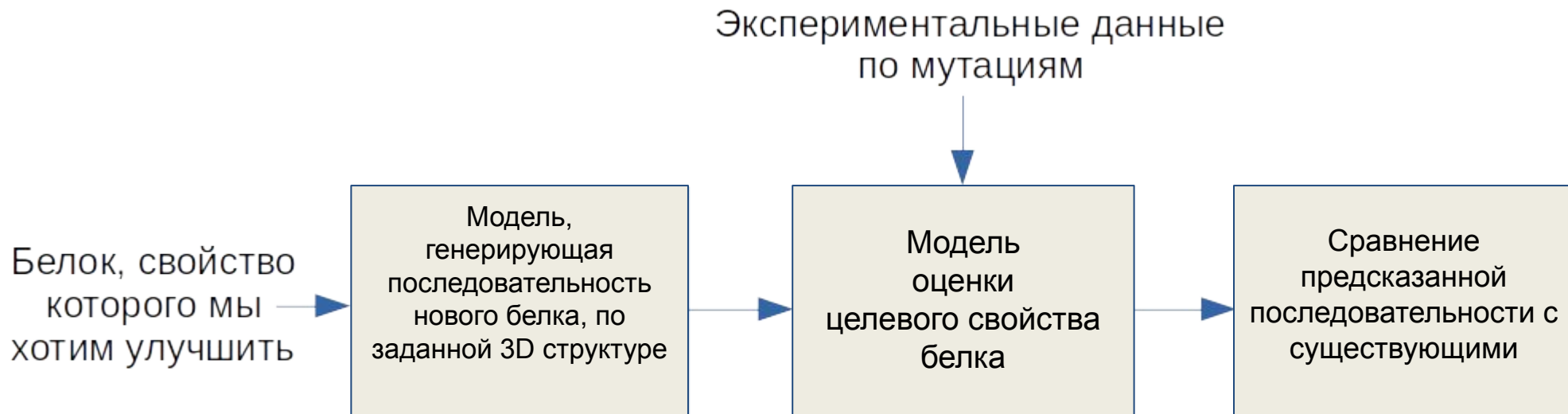
Цель

- *In silico* конструирование синтетических аналогов Ca^{2+} -регулируемых фотопротеинов с использованием методов молекулярного моделирования и машинного обучения.

Задачи

1. Построить логистическую регрессионную модель связи длины волны максимума биолюминесценции фотопротеинов с физико-химическими и структурными свойствами остатков активного сайта Ca^{2+} -регулируемых фотопротеинов
2. Провести предсказание последовательностей синтетических фотопротеинов с использованием генеративных белковых языковых моделей
3. Провести сравнительный анализ синтетических и природных Ca^{2+} -регулируемых фотопротеинов

Протокол компьютерного дизайна белковых последовательностей



Материалы и методы

Исследуемые фотопротиены

Акворин, 465 нм



Aequorea victoria

Беровин, 490 нм



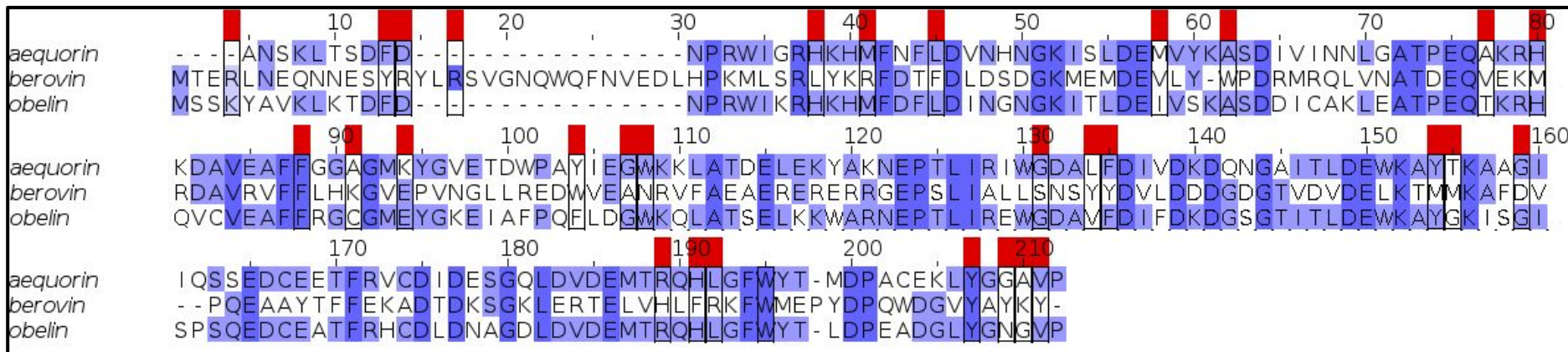
Beroe abyssicola

Обелин, 485 нм



Obelia longissima

Мутации фотопротеинов



	A	B	C	D	E	F
1	Aequorin		Berovin		Obelin	
2	mutation	wave_length	mutation	wave_length	mutation	wave_length
3	H16D	468	R4A	492	H22D	395
4	H16E	395	R4K	490	H22E	393
5	H16N	393	Y13F	490	H22N	395
6	H16Q	390	Y13Q	490	H22Q	395
7	H16T	388	Y13A	492	H22T	393
8	H16K	452	Y13R	490	H22K	0
9	H16I	388	R14A	0	H22I	390

33 мутации для акворина,

73 - для беровина,

41 - для обелина

Линейная регрессия

Линейные регрессионные модели строились с помощью библиотеки scikit-learn для Python. В качестве зависимой переменной выступала длина волны максимума биолюминесценции (y), а в качестве независимых переменных значения некоторого признака аминокислот в определённых позициях последовательности белка (x).

$$b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_jx_{i,j} + \dots + b_nx_{i,n} = y_i,$$

где $x_{i,j}$ - значение признака для i -ой мутации в j -ой позиции (из всех n позиций, для которых известны мутации), y_i - длина волны максимуму биолюминесценции при i -ой мутации, b_0, \dots, b_n - коэффициенты регрессии.

Логистическая регрессия

Логистические регрессионные модели строились аналогично линейным, только в качестве зависимой переменной выступало наличие (значение 1) или отсутствие биолюминесценции (0).

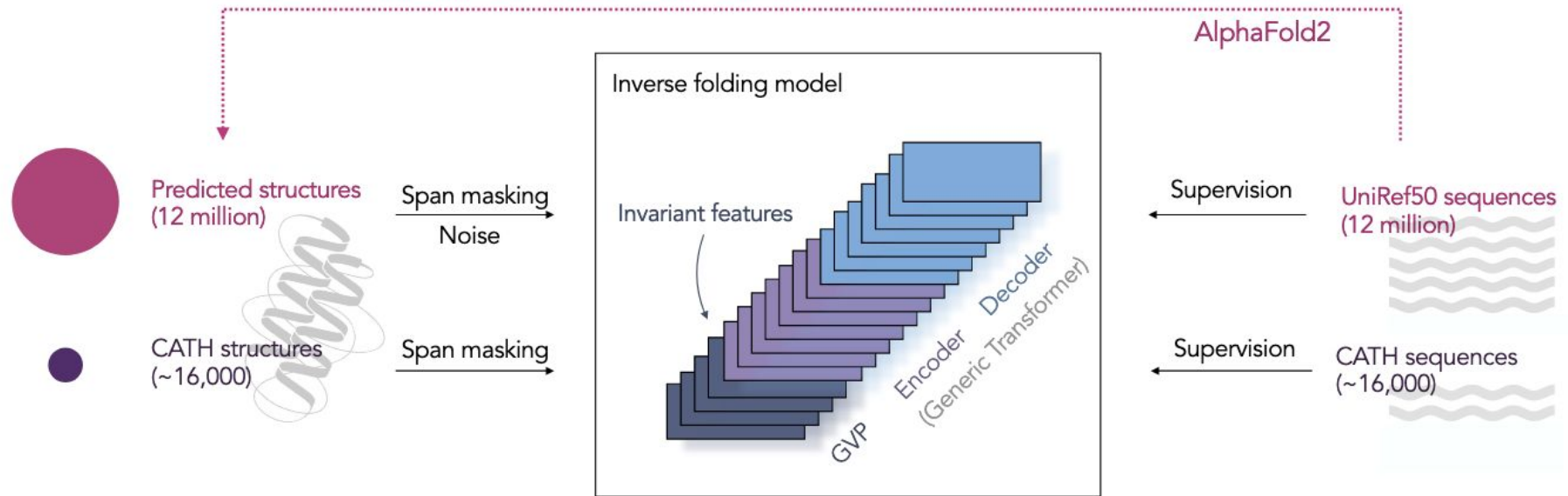
$$b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_jx_{i,j} + \dots + b_mx_{i,m} = z_i,$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

где $x_{i,j}$ - значение признака для i -ой мутации в j -ой позиции (из всех m позиций, для которых известны мутации), $f(z_i)$ - вероятность того, что сохранится биолюминесценция при i -ой мутации, b_0, \dots, b_m - коэффициенты регрессии.

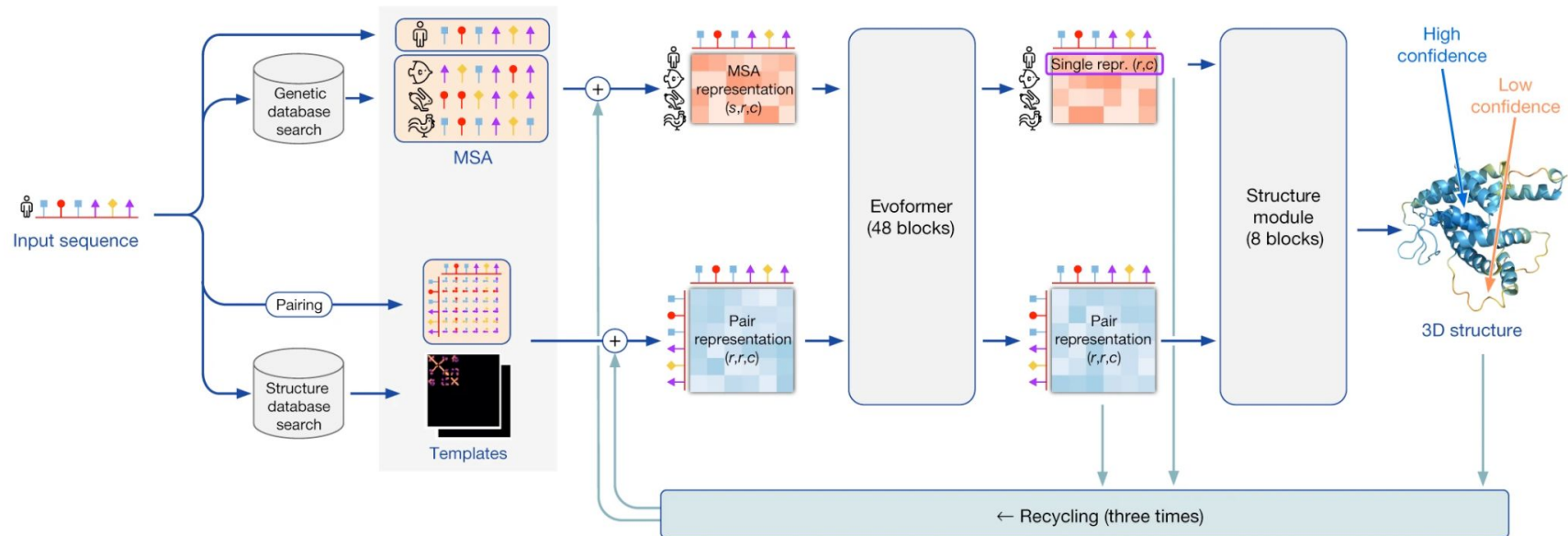
ESM-IF1

Модель ESM-IF1 для генерации белковых последовательностей по структуре



AlphaFold

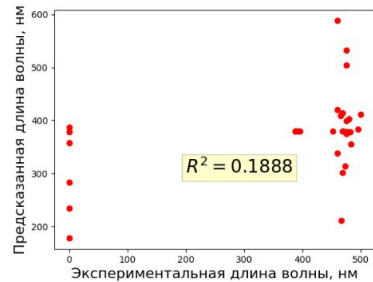
AlphaFold2 - программа на основе нейронных сетей по предсказанию пространственных структур белков по их аминокислотной последовательности.



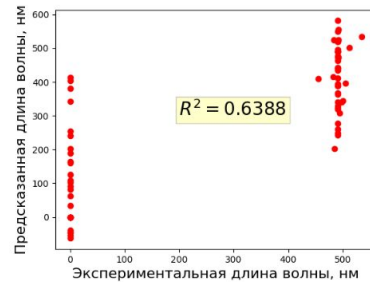
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.

Результаты

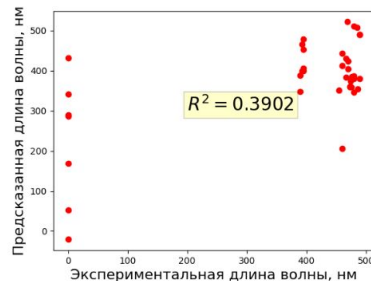
Построение регрессионных моделей



(a) Акворин



(b) Беровин



(c) Обелин

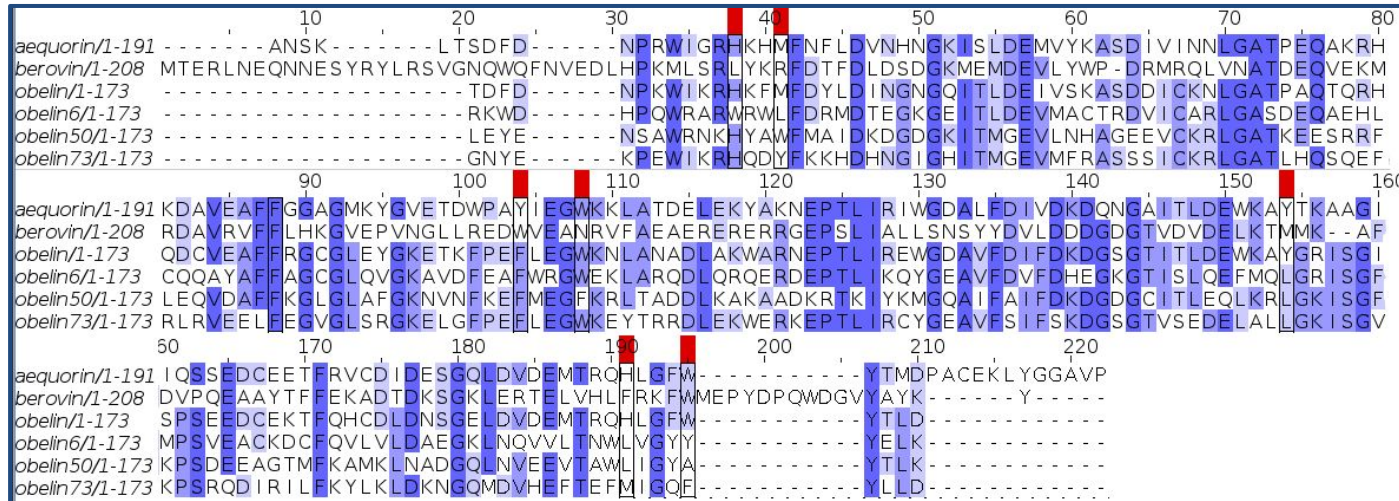
Результаты построения линейных регрессионных моделей.

Результат построения логистических регрессионных моделей.
Точность предсказания исчезновения биолюминисценции в результате мутаций:

- Акворин - 85%,
- Обелин - 93%,
- Беровин - 89%

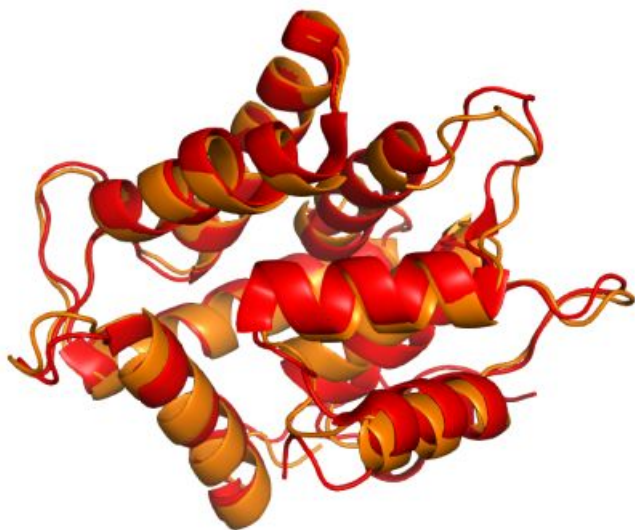
Дизайн фотопротеинов с помощью ESM-IF1

- Получили 100 предсказанных последовательностей для одного из фотопротеинов (обелина из *Obelia geniculata*)
- Отобрали 55 стабильных последовательностей по оценке ESM-IF1
- Все полученные фотопротеины обладают биолуминесценцией согласно построенной ранее регрессионной модели



Идентичность с
исходной
последовательностью
40-50%.

Верификация пространственных структур синтетических последовательностей с помощью AlphaFold2



(a) Синтетическая последовательность №6,
RMSD = 1.387 Å



(b) Синтетическая последовательность №50,
RMSD = 2.234 Å

Сравнение пространственных структур синтетических последовательностей с исходным фотопротеином (обелин, PDB 1JF0, красный)

Выводы

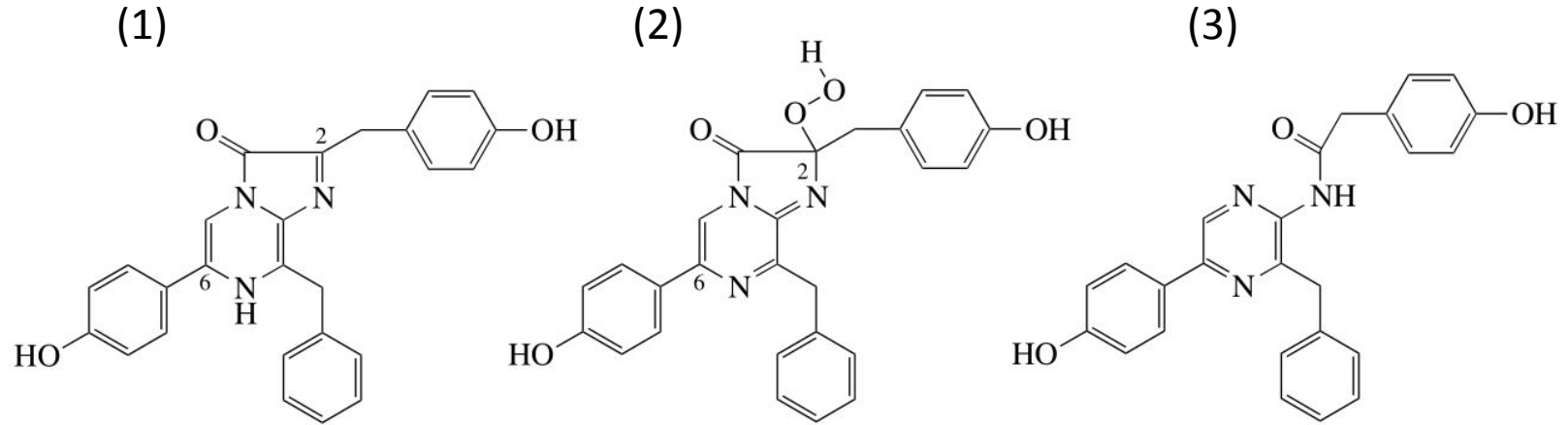
- . 1. Была построена логистическая регрессионная модель, способная предсказывать исчезновение сигнала биolumинесценции фотопротеинов при внесении мутаций в активном сайте.
- . 2. Были предсказаны структуры синтетических фотопротеинов с использованием генеративной модели ESM-IF1.
- . 3. С использованием построенной регрессионной модели была предсказана активность синтетических белков.
- . 4. С использованием AlphaFold2 была проведена валидация способности предсказанных последовательностей иметь третичный тип укладки, соответствующий известным фотопротеинам.

Спасибо за внимание!

Использование фотопротеинов

- Индикаторы внутриклеточного кальция
- Иммуноанализ
- Гибридизационный анализ
- Биоимиджинг
- Фотопротеины + флуоресцентные белки (например, GFP)

Субстрат

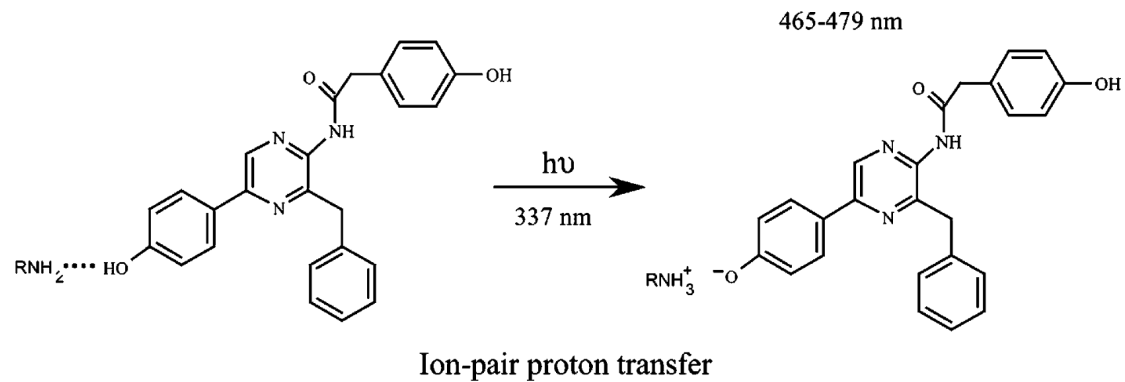
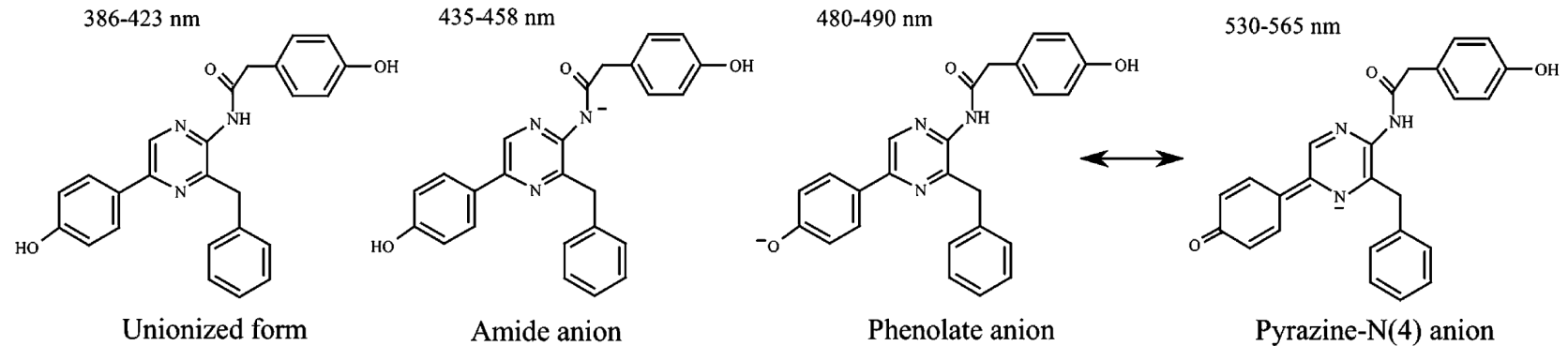


(1) Субстрат – целентеразин

(2) Преактивированная форма – 2-гидропероксицелентеразин

(3) Продукт реакции – целентерамид

Флуоресцентные формы целентерамида



Линейная регрессия

$$y = Xb + \varepsilon$$

$$\hat{y} = Xb, \quad e = y - \hat{y} = y - Xb$$

$$RSS = e^T e = (y - Xb)^T (y - Xb)$$

$$(X^T X)b = X^T y$$

Анализ консервативности активного сайта фотопротеинов

С помощью BLAST и построения филогенетических деревьев были найдены группы гомологов.

Для акворина:

Акворин из *Aequorea coerulescens*, *Aequorea parva*, *Aequorea macrodactyla* (80-85 % идентичность с исходной последовательностью)

Митрокомин из *Mitrocoma cellularia* (≈ 65 %);

Клитин-I, II, III из *Clytia gregaria*; клитин 1, 2, 3 из *Clytia hemisphaerica*; (58-62 %);

Обелин из *Obelia longissima* и *Obelia geniculata* (60-65 %);

