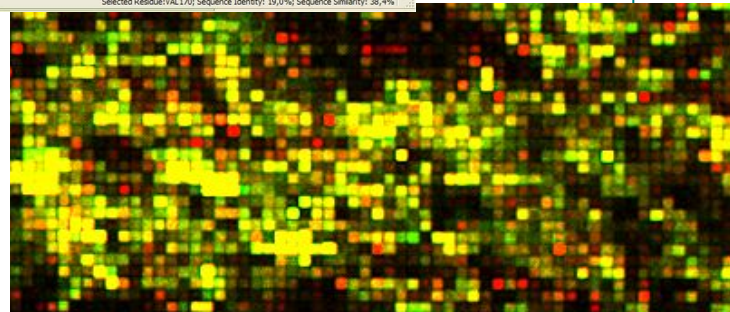
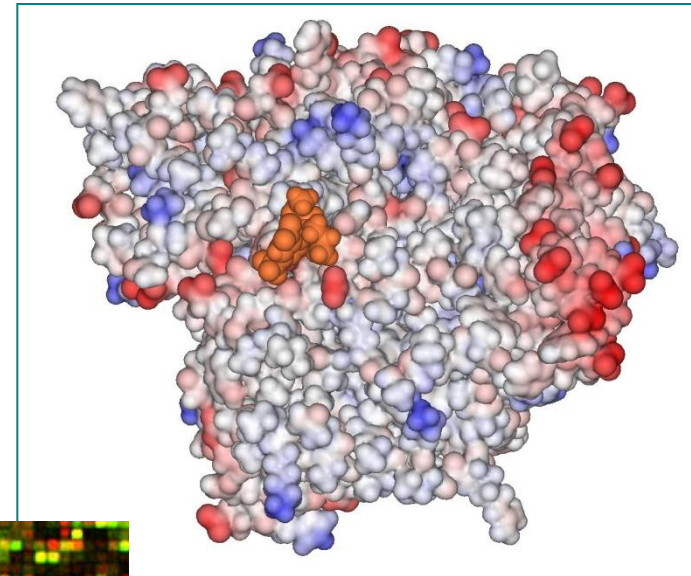
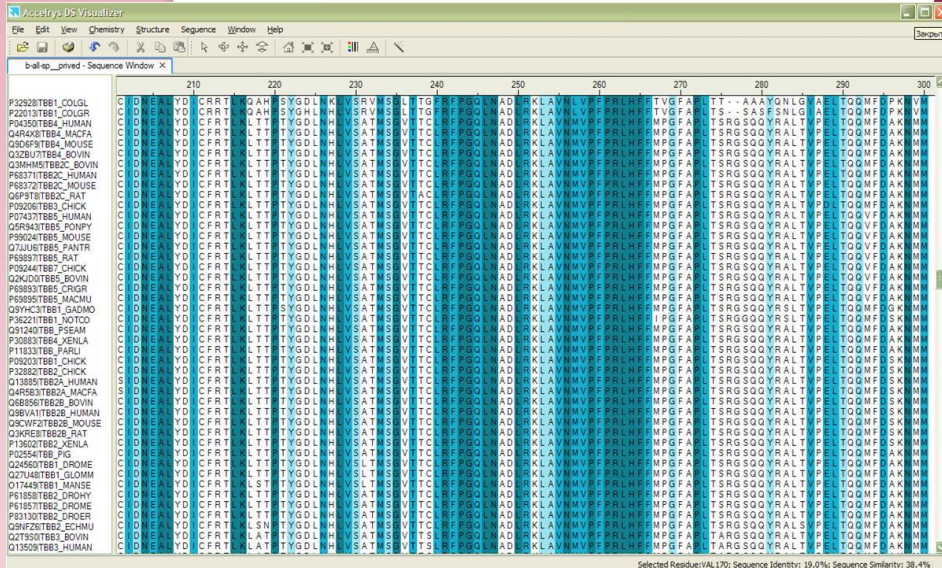


БІОІНФОРМАТИК

к.б.н. Нидорко О.
О.



Основний спосіб визначити схожість двох послідовностей - вирівняти їх

>EC_Tr : MQNRLTI KDI ARLSGVGKSTVSRVNLNNEYR

>EC_Fr : MKLDEI ARLAGVSRRTTASYVI NGKAKQYR

- При аналізі первинних структур процедура вирівнювання виявляє сходство між послідовностями (**sequence similarity**), яке може свідчити про гомологію (**homology**), тобто еволюційну спорідненість макромолекул.

Геп – пропуск в послідовності

>EC_Tr : MQNRLTIKDIARLSGVGKSTVSRVNLNNE---YR
>EC_Fr : ---MKLDEIARLAGVSRRTTASYVINGKAKQYR

**Гомологичные
последовательности –
последовательности, имеющие
общее происхождение (общего
предка).**

**Признаки гомологичности белков
сходная 3D-структура
в той или иной степени похожая
аминокислотная последовательность**

- **разные другие соображения...**

Что изображено?

Номер столбца выравнивания

```
          *                20                *
MTA1_YEAST : ----KSSIS P Q A R A F L E Q V E R R K --- Q S L N S : 24
MAT2_YEAST : K P Y R G H R F T K E N V R I L E S W E A K N I E N P Y L D T : 31
          3 2                L E    F    4                L 1 3
```

```
          40                *                60
MTA1_YEAST : K E K E E V A K K C G I T P L Q V R V W F I N K R M R S K - : 53
MAT2_YEAST : K G L E N I M K N T S L S R I Q I K N W V S N R R R K E K T : 61
          K    E    6    K          6 3    6 Q 6 4    W    N 4 R    4    K
```

Название последовательности

Консервативный остаток

Функционально консервативная позиция

Номер последнего в строке остатка из этой последовательности

«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

домовой
скупидом
водомерка ?

Гэп – пропуск в последовательности

лесовоз
ледоход

? --лесо---воз
лед---оход---

Ортологи и паралоги

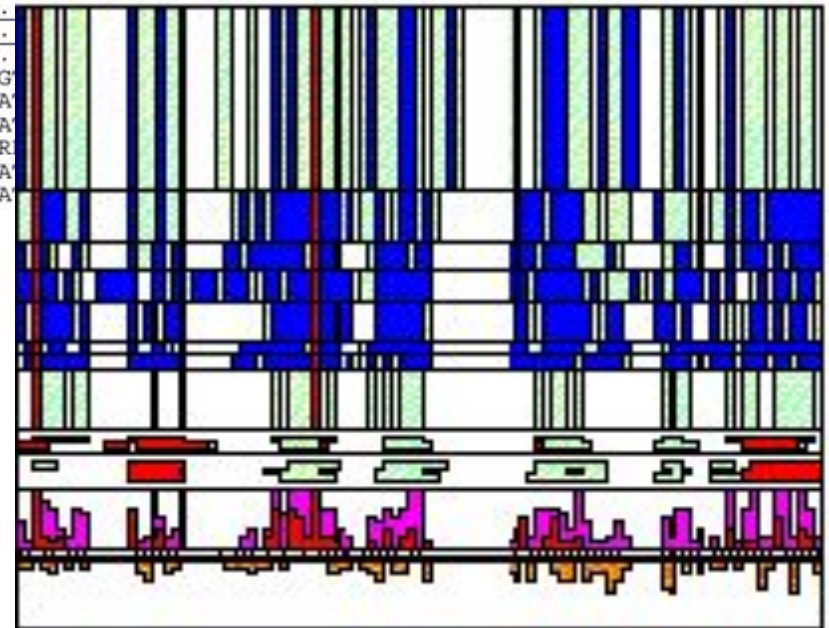
- Ортологи – гени з різних організмів, що розійшлися при видоутворенні.
 - Мається на увазі, що ортологи мають спільного «предка» і однакову функцію (якщо тиск відбору слабкий, то функція может «плисти»).
- Паралоги – гени, що розійшлися при дуплікації («копіюванні»)
 - Копії гена не зазнавали тиска відбора, а значить, могли змінити функцію.

Множественное выравнивание: содержание

- Определение, разновидности, решаемые задачи, общие проблемы
- Глобальное выравнивание
- **Прогрессивное выравнивание**
- **Итерационные методы**
- **Локальные множественные выравнивания**
- **Вероятностно-статистические методы множественного выравнивания**
- Оценка качества выравнивания
- Структурное выравнивание

Множественное выравнивание: Иллюстрации

AGRI_CHICK	154	CVCPAS.....	CS....G	Va.ESI	VCGSDG	KDYR	SECDLN	KHAC.....	DK.....	QENV	FKKFDG	AC	201																																
AGRI_RAT	165	CLCPTT.....	CF....	Gap.	DGT	VCGSDG	VDF	SEC	QLLSHA	C.....	AS.....	QEH	IFKKFNG	FC	212																														
FSA_HUMAN	116	CVCAPD.....	CS....	Nitw	KG	PVCG	LDG	KTYR	NECALL	KARC.....	KE.....	QPE	LEVQYCG	RC	164																														
FSA_PIG	116	CVCAPD.....	CS....	Nitw	KG	PVCG	LDG	KTYR	NECALL	KARC.....	KE.....	QPE	LEVQYCG	RC	164																														
FSA_RAT	116	CVCAPD.....	CS....	Nitw	KG	PVCG	LDG	KTYR	NECALL	KARC.....	KE.....	QPE	LEVQYCG	RC	164																														
FSA_SHEEP	109	CVCAPD.....	CS....	Nitw	KG	PVCG	LDG	KTYR	NECALL	KARC.....	KE.....	QPE	LEVQYCG	RC	157																														
IAC1_BOVIN	14	CKVYTEA.....	CT....	RE..	YN	PI	CD	SA	AKTY	SNE	CTF....	CNE	KM.NN.....	DAD	IHF	NHF	G	CEC	61																										
IAC2_BOVIN	7	CAEFKDP.....	KVY	CT....	RE..	SN	HC	GS	NGE	TY	G	NK	CAF....	G	K	A	M	.K	S.....	G	G	K	I	N	L	K	H	R	G	C	57														
IACA_PIG	7	ONVYRSH.....	LFB	CT....	RQ..	MD	P	I	C	G	N	G	S	Y	A	N	P	C	I	F....	C	S	E	K	G	.L	R.....	N	Q	K	F	D	F	G	H	W	G	H	C	57					
IACS_PIG	12	ODVYRSH.....	LFB	CT....	RE..	MD	P	I	C	G	N	G	S	Y	A	N	P	C	I	F....	C	S	E	K	L	.G	R.....	N	E	K	F	D	F	G	H	W	G	H	C	62					
IAC_MACFA	33	CARYQLPG.....	CF....	RD..	F	N	P	V	C	G	D	M	I	T	P	N	E	C	T	L....	C	M	K	I	R	.E	S.....	G	Q	N	I	K	I	L	R	R	G	F	C	81					
IOV7_CHICK	94	CSPYLQVVRDGNtMVAc	RI..	L	K	P	V	C	G	S	D	S	T	Y	D	N	E	C	G	I....	G	A	Y	N	A	.E	H.....	H	T	N	I	S	K	L	H	D	G	E	C	150				
IOVO_ABUPI	8	CSDHKKP.....	ACL....	QE..	Q	K	P	L	C	G	S	D	N	K	T	Y	D	N	K	C	S	F....	C	N	A	V	.D	S.....	N	G	T	L	T	L	S	H	F	G	K	C	56				
IOVO_ALECH	6	CSEYPKP.....	ACT....	LE..	Y	R	P	L	C	G	S	D	S	K	T	Y	G	N	K	C	N	F....	C	N	A	V	.E	S.....	N	G	T	L	T	L	S	H	F	G	K	C	54				
IPSG_VULVU	68	CTEYSDM.....	CT....	MD..	Y	R	P	L	C	G	S	D	G	N	Y	S	N	K	C	I	F....	C	N	A	V	.R	S.....	R	G	T	I	F	L	A	K	H	G	E	C	115					
IPST_ANGAN	12	CGEMSAMHA.....	CF....	MN..	F	A	P	V	C	G	D	G	N	T	P	N	E	C	S	L....	C	F	Q	R	Q	.N	T.....	K	T	D	I	L	I	T	K	D	D	R	C	61					
IPST_BOVIN	9	CTNEVNG.....	CF....	RI..	Y	N	P	V	C	G	D	G	V	T	Y	S	N	E	C	L	L....	C	M	E	N	K	.E	R.....	Q	T	P	V	L	I	Q	K	S	G	F	C	56				
IPST_PIG	9	CTSEVSG.....	CF....	KI..	Y	N	P	V	C	G	D	G	T	Y	S	N	E	C	V	L....	C	S	E	N	K	.K	R.....	Q	T	P	V	L	I	Q	K	S	G	F	C	56					
IPST_SHEEP	9	CTNEVNG.....	CF....	RI..	Y	N	P	V	C	G	D	G	V	T	Y	S	N	E	C	L	L....	C	M	E	N	K	.E	R.....	Q	T	P	V	L	I	Q	K	S	G	F	C	56				
OATP_HUMAN	439	ONVDCN.....	CFs...	KI..	W	D	P	V	C	G	N	G	I	S	Y	L	S	A	C	L	A...G	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485						
OATP_RAT	439	ONTRCS.....	CS....	Tn	t	W	D	P	V	C	G	N	G	V	A	M	S	A	C	L	A...G	C	K	K	F	V	.G	T.....	G	T	N	M	.V	F	Q	D	C	S	486						
PE60_PIG	37	CEHMTESPD.....	CS....	RI..	Y	D	P	V	C	G	D	G	V	T	Y	S	E	S	E	C	K	L...G	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485					
PGT_RAT	444	CRRDCS.....	CF....	DS	f	F	H	P	V	C	G	D	G	V	E	V	S	F	C	H	A...G	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485						
PSG1_MOUSE	33	CHDAVAG.....	CF....	RI..	Y	D	P	V	C	G	D	G	T	Y	S	N	E	C	V	L...G	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485							
QR1_COTJA	466	CICQDPA.....	ACPs..	t	K	D	Y	K	R	V	C	G	D	N	K	T	Y	D	G	T	C	Q	L	F	G	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485	
SC1_RAT	424	CVCQDPET.....	CFp..	a	K	I	L	D	Q	A	C	G	D	N	C	T	Y	A	S	S	C	H	L	F	A	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485	
SPRC_BOVIN	93	CVCQDP.TS.....	CFap.i	g	E	..	F	E	K	V	C	S	N	D	N	K	T	F	D	S	S	C	H	F	F	A	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485
SPRC_CAEL	74	CECISK.....	CFeld	g	D	P	..	M	D	K	V	C	A	N	N	O	T	F	T	S	L	C	D	L	Y	R	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485
SPRC_MOUSE	92	CVCQDP.TS.....	CFap.i	g	E	..	F	E	K	V	C	S	N	D	N	K	T	F	D	S	S	C	H	F	F	A	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485
SPRC_XENLA	90	CVCQDPST.....	CFpts	.v	G	E	..	F	E	K	I	C	G	D	N	K	T	Y	D	S	S	C	H	F	F	A	C	.E	T	.S	I.....	G	T	G	I	N	M	V	F	Q	N	C	S	485



Множественное выравнивание: определение и проблемы

- **Определение:** найти оптимальное соответствие между несколькими последовательностями, если заданы
 - Матрица соответствия
 - Штраф за делецию
 - Функция веса выравнивания
- **Проблемы:**
 - Множество делеций, замен,...
 - Ограниченное обобщение метода динамического программирования
 - Подсчет суммарного веса замен в колонке
 - Размещение делеций в разных местах и штрафы за них

Множественное выравнивание: проблемы (прод.)

● Проблемы:

○ Локальные минимумы

- накопление первоначальных ошибок в иерархических алгоритмах
- лучшее дерево соответствует лучшему выравниванию

○ Выбор параметров

- один набор параметров не может быть пригодным на все случаи жизни

● Сложности выравнивания нарастают с ростом различий между последовательностями

Множественное выравнивание: решаемые задачи

- Поиск мотивов (блоков) – коротких сигнатур, идентифицируемых в консервативных участках множественного выравнивания
 - отсутствие вставок и делеций
- Построение профилей (матриц весов): оценка частоты встречаемости каждой АК в каждой позиции
- Построение скрытых марковских моделей (НММ) – обобщенных профилей, описываемых строго математически

Выравнивание
хорошо изучен-
ного семейства

Функционально
важные остатки

4-5
консервативных
остатков

Паттерн

Поиск в
UniProt

Если
находим
только «пра-
вильные», то
ОК

Если много
лишнего, то
увеличиваем
паттерн

Паттерн – регулярное выражение UNIX'a:

[AC]-x-V-x(4)-{ED}

Ala или Cys- x-Val- x- x- x - x- (любой, но не Glu и не Asp)

Seq1	F	K	L	L	S	H	C	L	L	V
Seq2	F	K	A	F	G	Q	T	M	F	Q
Seq3	Y	P	I	V	G	Q	E	L	L	G
Seq4	F	P	V	V	K	E	A	I	L	K
Seq5	F	K	V	L	A	A	V	I	A	D
Seq6	L	E	F	I	S	E	C	I	I	Q
Seq7	F	K	L	L	G	N	V	L	V	C

Паттерн:
F-[KP]-x(3)-[EQ]-x(4)

Не найдем!

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Позиционно-специфичная матрица весов аминокислот

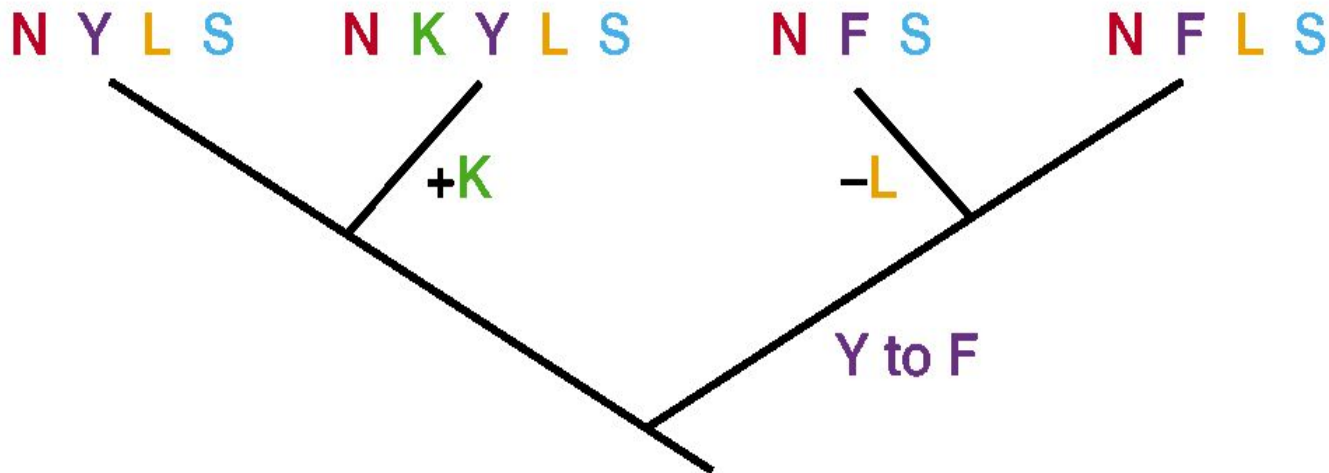
Профиль или весовая матрица (PSSM)

Множественное выравнивание: области применения

- Один из ключевых методов в современной молекулярной биологии
- Сферы применения
 - Филогенетический анализ, **«ЭВОЛЮЦИЯ» ПОС-СТИ**
 - Предсказание вторичной/третичной структуры белков
 - Выявление АК-остатков (**консервативных участков**)
 - экспонированных на поверхности белка
 - формирующих активный центр
 - обеспечивающих субстратную специфичность
 - критичных для стабилизации втор./трет. структуры
 - Выявление характерных **фрагментов** для описания **белковых семейств**
 - Выявление неизвестных ранее **гомологий** между генами и последовательностями
 - Длинные пос-сти из случайных **коротких фрагментов**

Множественное выравнивание и филогенетический анализ

seqA	N	•	F	L	S
seqB	N	•	F	-	S
seqC	N	K	Y	L	S
seqD	N	•	Y	L	S



- ◆ Идея – минимизация числа мутаций
- ◆ Что сначала: выравнивание или дерево?
- ◆ Решение не единственно !

Множественное выравнивание:

консервативные участки во многих пос-стях

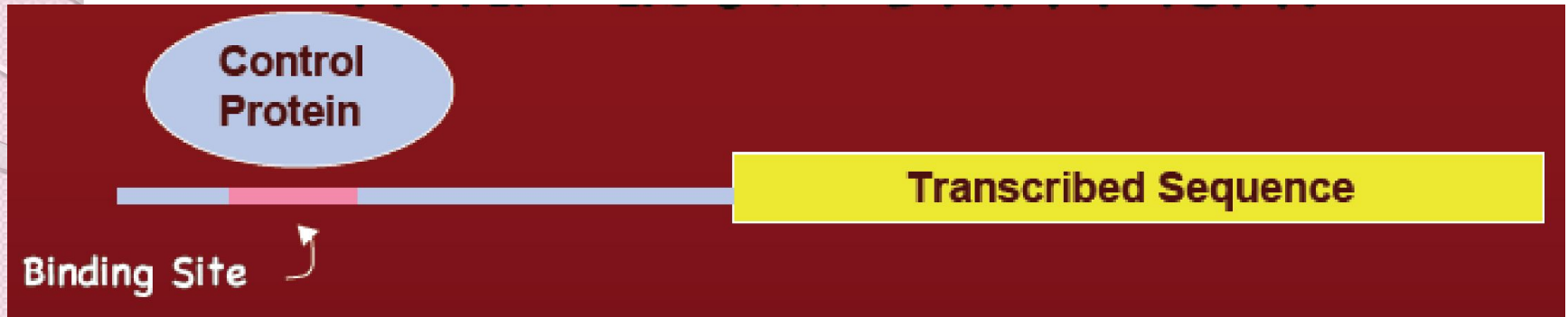
```
DRFKHLKTEAEMKASEDLKKHGVTVLTAALGAILKKGK
PKFAGI-AQADIAGNAAISAHGATVLLKLGELLKAKG
PHF-DLSH-----GSAQVKGHGKQVADALTNAVAHVD
PHF-DLSH-----GSAQVKAHGKKVGDALTLAVGHLD
SHWPDVTP-----GSPHIKAHGKKVMGGIALAVSKID
ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLA
DSFGDLSNPGAVMGNPVKVKAHGKKVLHSGEGVHHLD
PKFKGLTTADELKKSADVRWHAERIINAVDDAVASMD
ADFKGKSVAD-IKASPCLRDSVSRIFTRLNEFVNNA
KRLGNVS---QGMANDKLRGHSITLMYALQNFIDQLD
SFLKGT--SEVPQNNPELQAHAGKVFCLVYEAAIQLE
PQMAGM-SASQLRSSRQMQAHAIRVSSIMSEYVEELD
HKFS-SVPLYGLRSNPAYKAQTLTVINYLDKVVDALG
TQFAG-KDLESIKGTAPFETHANRIVGFFSKIIGELP
GFSGA-----SDPGVAALGAKVLAQIGVAVSHLG
```

Консервативный – не значит «совпадающий» !

Множественное выравнивание: белки vs. ДНК

- Выравнивание белковых семейств
 - В алфавите много «букв»
 - Эволюционная близость белковых молекул, основа для филогенетических деревьев
 - какие события привели к возникновению данного семейства?
 - Идентификация функционально важных областей
 - Данные для предсказания структуры
- Выравнивание некодирующих участков ДНК
 - Консервативные участки, отвечающие за регуляцию экспрессии
 - Установление эволюционной близости
 - Идентификация функционально важных областей

Множественное выравнивание ДНК



- Сайты связывания TFs = мотивы ДНК-последовательностей
- Консерватизм
 - внутривидовой (синергичная регуляция транскрипции нескольких генов)
 - межвидовой (близкие механизмы регуляции транскрипции)
- Дивергенция
 - внутривидовая («специальные» цели, завязанные на метаболизм)
 - межвидовая (эволюционный дрейф)

Множественное выравнивание ДНК: проблемы и варианты решения


- Гораздо сложнее выравнивания белков
 - всего 4 «буквы»
- Отсутствие «золотого стандарта»
- Необходимость оценить
 - способность связывать белки
 - влияние на функцию
- Смысл – тестирование гипотез
 - об общем предке
 - об общих механизмах связывания белков
 - о близости функций

Множественное выравнивание: четыре группы методов

- **Прогрессивное** глобальное выравнивание
 - начать с наиболее близких пос-стей
- **Итерационные** процедуры
 - выравнивание групп пос-стей с последующей оптимизацией
- **Выравнивание по локальным консервативным участкам**
 - построение профилей (разновидности матрицы весов)
 - поиск блоков в пос-стях (выравниваний без делеций)
- **Статистические** методы и вероятностные модели
 - поиск шаблонов (patterns)
 - скрытые марковские модели

Множественное выравнивание: история

- До 1987 г. множественные выравнивания строились вручную
- Sankoff (1975 и 1987) – первый программно реализованный алгоритм
 - основа – филогенетический анализ
- Barton (1990) – оценка качества выравнивания методом рандомизации, **AMPS**
- Russel & Barton (1992) – структурное выравнивание, **STAMP**
- Thomson et al. (1994) – **ClustalW**
- Altshul et al. (1997) – **PSI-BLAST**
- Notredame et al. (2000) – неиерархическое выравнивание, **T-Coffee**
- Clamp (2004) - **JalView**

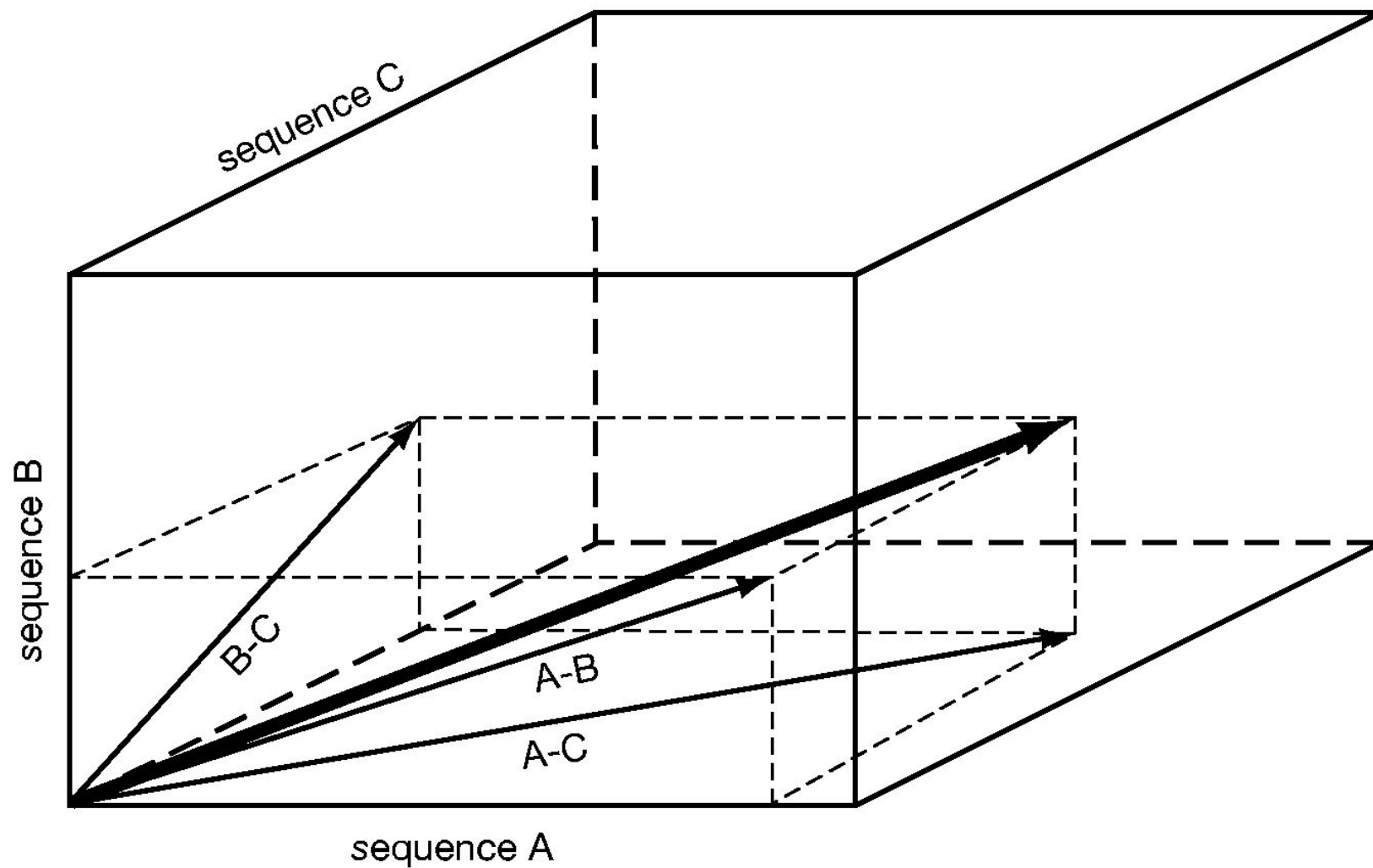


Глобальное выравнивание (обобщение ДП)

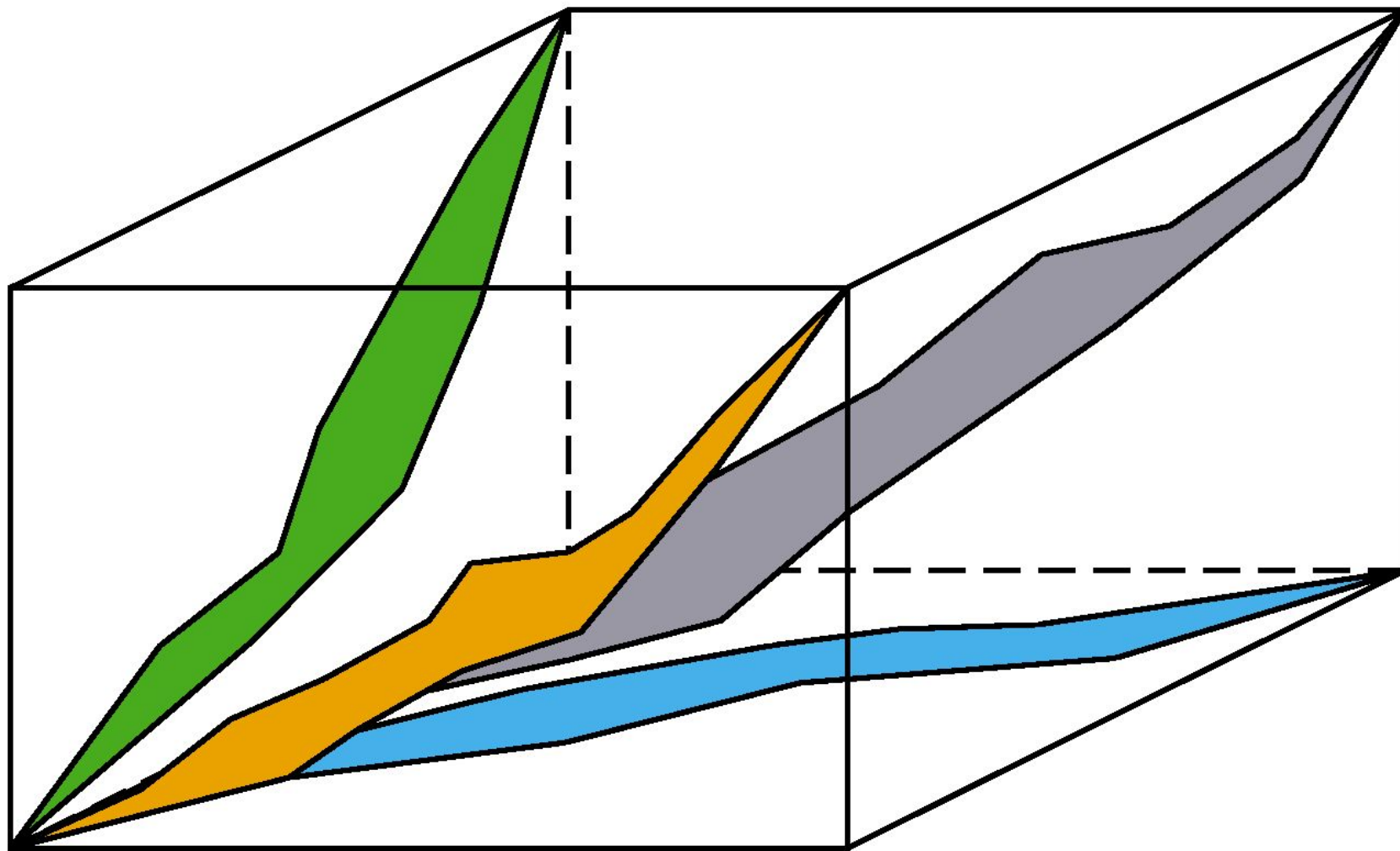
Глобальное выравнивание

- Обобщение метода динамического программирования
 - программа MSA (Lipman et al., 1989)
 - результат далек от оптимального (Gupta et al., 1995)
 - ресурсы: N^m сравнений для m пос-стей длины N
- Развитие MSA
 - метод суммирования пар (sum of pairs, SP) – Carrillo & Lipman (1988)
 - попарные выравнивания
 - филогенетическое дерево
 - выравнивание в ограниченной области куба
 - эвристическое выравнивание \neq оптимальному
 - реализация в ClustalW / ClustalX
 - сокращение необходимых ресурсов – Gupta et al. (1995)

Множественное выравнивание: трехмерное динамическое программирование



Множественное выравнивание: трехмерное динамическое программирование (прод.)



Глобальное выравнивание (прод.)

● Оценка качества

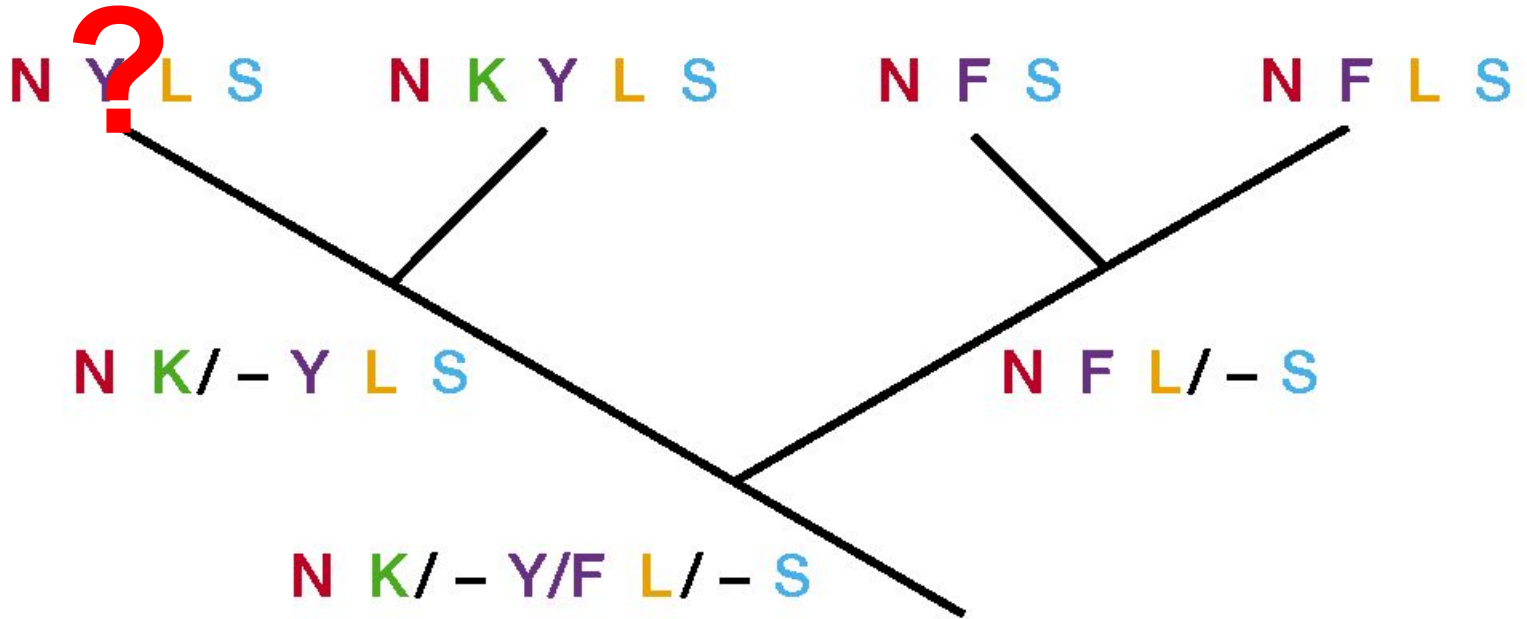
- веса множественных выравниваний (SP score) = сумме весов попарных выравниваний
 - поиск наибольшего суммарного веса
 - взвешивание весов (опционально)
 - по филогенетическому дереву
 - учет эволюционно близких пос-стей
 - «дифференц.» вес ε для каждой пары = (вес пары в MSA) – (вес при оптимальн. парном выравнивании)
 - степень дивергенции пос-стей в выравнивании $\delta = \sum \varepsilon_i$ (чем больше δ , тем сильнее дивергенция)
 - **MSA**: матрица замен PAM250, постоянный штраф за любую делецию
- Возможность применения к большему числу (6-8) коротких последовательностей

Прогрессивное выравнивание

Прогрессивное выравнивание: идея

- Сначала – эволюционно наиболее близкие пос-сти
- Постепенное добавление новых пос-стей / групп пос-стей
 - Waterman & Perlwitz (1984)
 - Feng & Doolittle (1987, 1996)
 - Higgins *et al.* (1996)
 - ...
- Отображение близости на филогенетическом дереве
 - методы попарного сравнения пос-стей
- Проблема: **неопределенность отдельных замен**

Филогенетический анализ – не панацея



◆ Проблемы

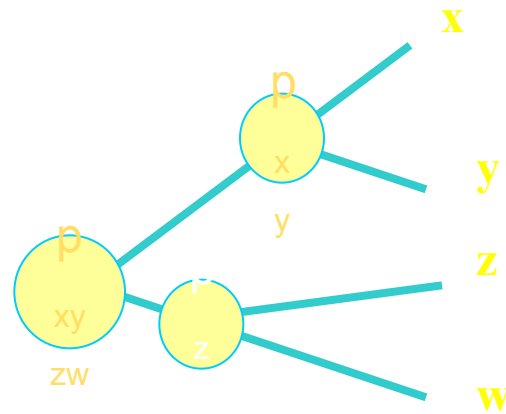
- неопределенность в порядке замен / делеций
- взвешивание ветвей (пос-стей)
- подбор матрицы замен
- назначение штрафов за делеции

◆ Реализация: ClustalW/X и PILEUP

Отражение
эволюции

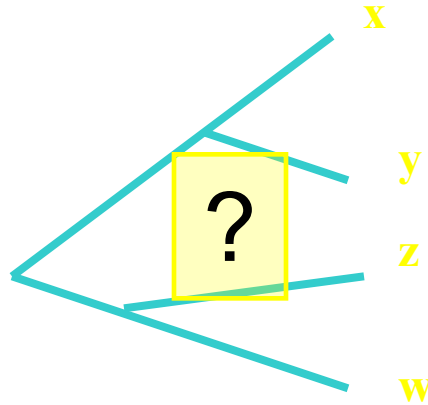


Прогрессивное выравнивание



- Если эволюционное дерево известно
 - сначала выравниваются элементы, самые близкие на эволюционном дереве
 - на каждом шаге выравниваются пос-сти x и y , или профили P_x и P_y для построения нового выравнивания с профилем P_{result}
- Версия со взвешиванием
 - ветви дерева имеют веса, пропорциональные степени расхождения
 - новый профиль – взвешенное среднее двух предыдущих

Прогрессивное выравнивание (прод.)



- Если эволюционное дерево неизвестно:
 - построить всевозможные парные выравнивания
 - определить матрицу расстояний D , элементы которой $D(x, y)$ соответствуют эволюционному расстоянию, определенному по парным выравниваниям
 - реконструировать эволюционное дерево
 - построить выравнивание на основе реконструированного дерева

Прогрессивное выравнивание: детали алгоритма

● Три этапа

- попарные выравнивания «каждая с каждой»
- филогенетическое дерево по весам парных выравниваний (по генетическим расстояниям)
- последовательное построение множественного выравнивания
 - от похожих – к непохожим

● Генетическое расстояние = (число замен)
/ (полное число соответствий)

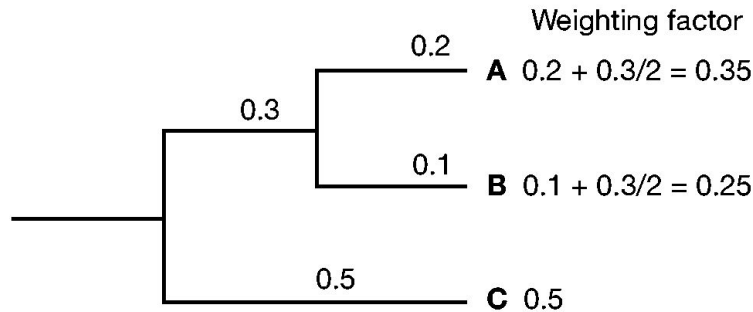
- делеции не учитываются  Дерево

Прогрессивное выравнивание: детали алгоритма (прод.)

- Взвешивание пос-стей (ветвей дерева)
 - мультипликативная модель
- Штрафы за делеции
 - предыдущие делеции влияют на последующие выравнивания
 - местоположение делеций (учет вторичной структуры)
 - таблица встречаемости делеций
 - штраф за открытие делеции и ее продолжение на каждую позицию
 - штрафы во множественном выравнивании модифицируются с учетом матрицы замен, степени сходства и длины пос-стей
- Схема назначения штрафов в Clustal противоположна таковой в MSA
 - чем уникальнее пос-сть, тем больше вес

Прогрессивное выравнивание: взвешивание ветвей дерева

A. Calculation of sequence weights



B. Use of sequence weights

Column in alignment 1

Sequence A (weight a) K.....

Sequence B (weight b) I.....

Column in alignment 2

Sequence C (weight c) L.....

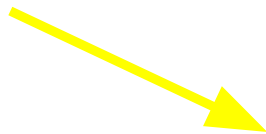
Sequence D (weight d) V.....

Score for matching these two column in an msa =

$$[a \times c \times \text{score (K,L)} + a \times d \times \text{score (K,V)} + b \times c \times \text{score (I,L)} + b \times d \times \text{score (I,V)}] / 4$$

Множественное выравнивание: популярный инструментарий

ClustalX



ClustalX (1.64b)

File Edit Alignment **Trees** Colors Quality Help

Multiple Alignment

- Draw NJ Tree
- Bootstrap NJ Tree
- Exclude positions with gaps
- Correct for multiple substitutions
- Save Log File
- Output Format Options

1 MG088
2 MP606
3 rpsG
4 HIN170
5 HP1196
6 s111097
7 MJ1047
8 YJR123w
9 YJR113c

LAQRILYGAFEIIEKRTNQPIITVFEKAVDNYMFRLELVRRRIA-GSN
LAQRILYGAFDLIEORTKEKPIITVFERAVGNVMPRELELVRRRIA-GSN
TAESIVYSALETLAQRSGKSELEAFEVALENVRFVTVMSRRVVG-CST
VAESIVYGALETLAQRIGKEPFEAFEVALENVRFVTVMSRRVVG-CST
VAEKIIVKAFNKIEEKSGEKGIEVFEKALERVRELVEVMSRRVVG-GAT
VNR---SGKNSLASSIVYNALASVGEKTDGDFLEVFKAIKNITELVEVMARRVVG-GAT
VMRREENTGKMLKALKIVENAFEIEKRTKQNPVQVLVDAIENAGFPREDITRISVVG-GIV
LNMNGRNGKMLKAVRIKHTLDIINVLTDONPIQVVDATITNGFPREDITRVGGG-GAA
INR---HGKNEKACTILSRATVLYCOTRQDFIQALEKSLDELAFLLMHTFNITGVAKA

ruler 130 140 150 160 170 180

File D:\Tree2\single-gene-tree{others}49.ali loaded.

DCSE: total.ali

File Position Markers Primary Secondary Alignment Varia Opti

Saccharomyces cerevisiae D 142 Org:

Aa : -[G G C U C G U G G]-[A U C G]-[A U G A A]-[G - A C C G C]
Dm : -[G S C U C A U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Hm : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Xb : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Ml : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Hs : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Mm : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Rn : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Ce : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Ca : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Sc : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Sp : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Pc : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Cn : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Sj(: -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Mr(: -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Os : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
At : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Bn : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Cl(: -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Fa : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Le(: -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]
Sa : -[G S C U C G U G G]-[G U C G]-[A U G A A]-[G - A C C G C]

Align

Align

Blocksizes 30,15

Type of alignment

1 to many cached 1 to many Many to 1

Algorithm used for alignment

Sellers Myers-Miller Blocksearch only

How to handle unaccommodated inserts

No insert Global Local Push

Go Set Costs Old routines Help Cancel



DCSE

Прогрессивное выравнивание: штрафы за делеции

- Существующие делеции влияют на выравнивание следующих пос-стей
 - их позиции фиксируются
- **ClustalW**: размещение делеций между консервативными доменами
 - Pascarella & Argos (1992): частоты встречаемости делеций после каждой АК в неконсервативных участках структурно близких белков
- Штрафы
 - за открытие делеции
 - за продолжение делеции
 - та же схема за делеции внутри существующих делеций

Прогрессивное выравнивание: штрафы за делеции (прод.)

- Компенсационная модификация штрафов
 - средний вес соответствий по матрице замен
 - уровень гомологии между пос-стями
 - длины пос-стей
- Таблица делеций для каждой группы выравниваемых пос-стей
- Другие варианты модификаций
 - ↓ штрафов для областей с существующими делециями
 - ↑ штрафов для областей, соседствующих с делециями
 - ↑ штрафов для областей с гидрофильными АК

Прогрессивное выравнивание: проблемы

- Результат зависит от начальных парных выравниваний
 - ошибки первых выравниваний накапливаются
 - выравнивание непохожих пос-стейБайесовские методы (e.g. HMM)
- Матрица замен и штрафы за делеции должны отражать специфику всего набора пос-стей



Итерационное выравнивание

Итерационное выравнивание: идея метода

● Задача

- избежать накопления ошибок начальных выравниваний, свойственных прогрессивным методам

● Вариант решения

- многократные итерационные выравнивания **подгрупп последовательностей**
- построение общего глобального выравнивания
- оптимизация общего веса выравнивания (суммы парных весов)

Итерационное выравнивание:

варианты реализации

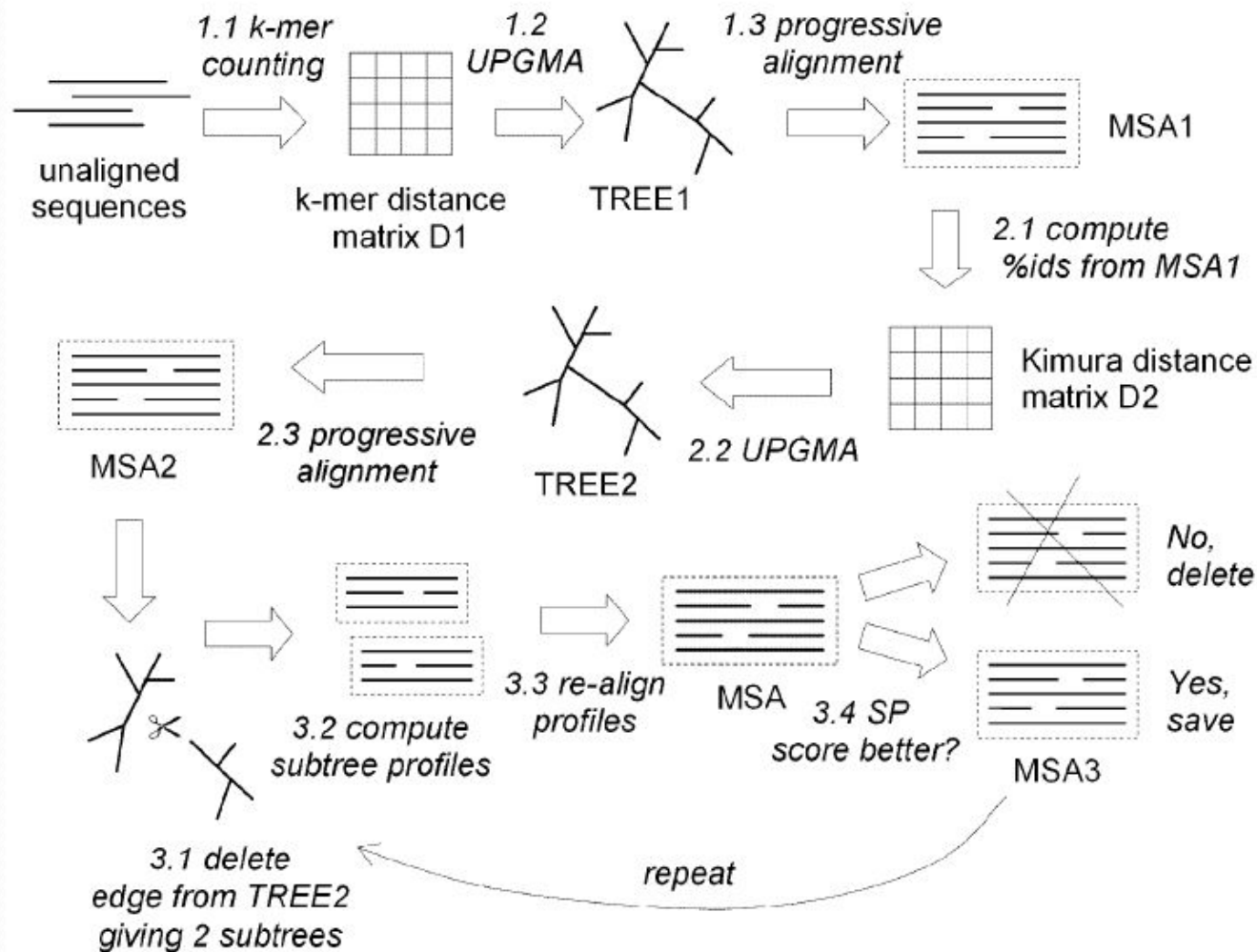
- MultAlin (Corpet, 1998)


- пересчет весов парных выравниваний в прогрессивном алгоритме
- использование весов для пересчета дерева
- улучшение множественного выравнивания

- PRRP (1994)

- построение дерева по начальным парным выравниваниям
- вычисление весов по дереву и построение выравниваний по аналогии с MSA (но: локальные участки вместо глобального выравнивания + возможны делеции)
- итерационный пересчет локально выровненных участков для повышения веса выравнивания
- выравнивание с наибольшим весом □ новое дерево, новые веса и новые выравнивания
- повторение, пока суммарный вес не перестанет меняться

Muscle или как исправить ClustalW





Локальные множественные выравнивания

Локальные множественные выравнивания: виды алгоритмов

- Анализ профилей
- Блочное выравнивание
- Поиск мотивов
- Статистические методы

Анализ профилей: введение

● Идея:

- MSA для группы пос-стей
- Выделение высоко консервативных участков в мини-MSA
- **Профиль - матрица весов для мини-MSA**
- Профиль допускает соответствия, замены, делеции и вставки

● Применения

- поиск соответствий профилю в последовательности-мишени (программа Profilesearch)
- в качестве матрицы замен для построения выравниваний (программа Profilegap)

Анализ профилей: идентификация в семействе белков теплового шока (hsp70)

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len
I	8	3	-2	5	4	5	5	-4	<u>24</u>	0	15	13	1	1	1	-7	2	22	21	-18	-6	4	100	100
T	13	19	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	<u>29</u>	3	-28	-14	15	100	100
L	5	5	-5	3	4	13	4	2	8	-4	<u>14</u>	12	8	-5	0	-10	0	10	10	-1	5	2	22	22
S	17	14	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	<u>34</u>	19	1	-8	-15	4	100	100
T	15	3	22	0	-1	-5	12	-2	7	-3	-8	-6	5	7	-8	-7	16	<u>29</u>	9	-22	6	-4	100	100
T	8	-1	12	-2	0	5	6	-4	19	-4	8	5	-1	2	-8	-8	7	<u>22</u>	19	-15	4	-3	100	100
C	17	0	<u>24</u>	-1	-3	11	8	-1	7	-10	1	-2	1	-3	-8	-14	8	5	9	-5	14	-7	100	100
V	11	0	18	-1	-2	2	14	-10	26	-4	9	7	-3	7	-7	-7	21	10	<u>31</u>	-19	-5	-5	100	100
C	10	-8	<u>15</u>	-11	-11	6	8	-7	11	-10	4	3	-7	0	-11	-4	11	5	15	-22	14	-11	100	100
V	7	7	-3	8	8	-3	11	1	20	-1	14	10	4	2	8	-5	0	5	<u>26</u>	-24	-6	8	100	100

Матрица весов (профиль) содержит вероятности встречаемости АК в разных позициях

Блочное выравнивание: семейство из 34 тубулиновых белков

a MFRRKAF LHWYTGEGMDEMEFTEAESNMNDPVAEYQQY
 MFKRKAF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFKRKAF LHWYTGEGMDEMEFTEVRANMNDLVAEYQQY
 MFKRKAF LHWYTGEGMDELEFSEAESNMNDLVSEYQQY
 MFKRKGF LHWYTGEGMEPVEFSEAQSDLEDL I LEYQQY
 MFRRKAF LHWFTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFRRKAF LHWYTGEGMDEMEFSEAEGNTNDLVSEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNDLMSEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNDLVAEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNDLVHEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNELVSEYQQY
 MFRRKAF LHWYTLGMEELFTEAESNMNDLVVEYQQY
 MFRRKAF LHWYTNÉGM D I TEFAEAESNMNDLVSEYQQY
 MFRRKAF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFRRKRF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFRRNAF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFRRQAF LHWYTGEGMDEMEFTEAESNMNDLVSEYQQY
 MFSRKAF LHWYTGEGMEEGDFAEADNNVSDLLSEYQQY

MF GKRAFVHHYV GEGME ENEFTDARQDLYELEVDYANL
 MF KKRAFVHWYV GEGMEEGEFTEAREN I AVLERDFEEV
 MFVKRAFVHWYV GEGMEEGEFAEARDDL LALEKDYESV
 MYAKRAFVHWYV GEGMEEGEFAEAREDLAALEKDYEEV
 MYAKRAFVHWYV GEGMEEGEFSEARED IAALEKDYEEV
 MYAKRAFVHWYV GEGMEEGEFSEAREDLAALEKDFEEV
 MYAKRAFVHWYV GEGMEEGEFSEAREDLAALEKDYEEV
 MYAKRAFVHWYV GEGMEEGEFSEAREDMAALEKDYEEV
 MYAKRAFVHWYV GEGMEEGEFSEVREDLAALEKDYEEV
 MYAKRAFVHWYV GEGMEEGEFTEAREDLAALEKDYEEV
 MYAKRAFVHWYV GEGMEEVEFSEAREDLAALEKDYEEV
 MYAKRAFVHWYV SEGMEEGEFAEAREDLAALEKDYDEV
 MYSKRAFVHWYV GEGMEEGEFSEAREDLAALEKDYEEV
 MYSKRAFVHWYV GEGMEEGEFSEAREDLAALERDYEEV

b MF . K . . FVH . F . . EGMQ . . QFPQ . . . Q QF . . .
 Y R L Y N N AN N NY
 W I W E E GE E EW
 D D SD D D
 T

c MF . KR . FLHWFT . EGMQ . . QFPE . . . Q . . DLI . DYQQY
 R Y N N A N L

Анализ профилей: ограничения

- Профиль отражает вариабельность в данном MSA
 - смещение в сторону похожих пос-стей
 - вариант коррекции: Gribskov & Veternik, 1996
 - взвешивание пос-стей по удаленности на филогенетическом дереве: чем меньше расстояние, тем меньше вес
- Недостаточное число пос-стей в MSA
 - некоторые АК на некоторых позициях не представлены

Профиль – позиционно-специфическая матрица замен

- 2I столбец и N строк
 - N – длина последовательностей в выравнивании

2hhb Human Alpha Hemoglobin	R	V	D	C	V	A	Y	K	
HAHU	R	V	D	C	V	A	Y	K	100
HADG	R	V	D	C	V	A	Y	K	89
HTOR	R	V	D	C	A	A	Y	Q	76
HBA_CAIMO	R	V	D	P	V	A	Y	K	73
HBA \bar{T} _HORSE	R	V	D	P	A	A	Y	Q	62
1mbd Whale Myoglobin	A	I	C	A	P	A	Y	E	
MYWHP	A	I	C	A	P	A	Y	E	100
MYG_CASFI	R	I	C	A	P	A	Y	E	85
MYH \bar{U}	R	I	C	V	C	A	Y	D	75
MYBAO	R	I	C	V	C	A	Y	D	71

Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	

Consensus = Most Typical A.A.

R	V	D	C	V	A	Y	E
---	---	---	---	---	---	---	---

Better Consensus = Freq. Pattern (PCA)

R	iv	cd	š	š	A	Y	μ
---	----	----	---	---	---	---	---

š = (A, 2V, C, P); μ = (4K, 2Q, 3E, 2D)

Entropy ⇒ Sequence Variability

3	7	7	14	14	0	0	14
----------	----------	----------	-----------	-----------	----------	----------	-----------

Множественное выравнивание на базе вероятностно-статистических методов

- Максимизация математического ожидания
- Сэмплирование Гиббса
- Скрытые марковские модели
 - see Russ Altman, Lecture 4-27-06, pp. 8-20

Наиболее известные программы множественного выравнивания:

1. **MSA => оптимальное выравнивание, если дождаться результата**
2. **ClustalW (реализации – ClustalX, emma из EMBOSS) – до сих пор самый популярный алгоритм, в сложных случаях может ошибиться.**
3. **Muscle – итеративный прогрессивный алгоритм, точнее и быстрее ClustalW**
4. **T-COFFEE – немного точнее, но существенно медленнее**
5. **HMMER – часто ошибается, но хорошо строит профили**
6. **.....**



Структурное выравнивание

Правильно ли выровнены последовательности?

```

                *                20                *
MTA1_YEAST : ----KSSIS P Q A R A F L E Q V E R R K --- Q S L N S : 24
MAT2_YEAST : K P Y R G H R F T K E N V R I L E S W E A K N I E N P Y L D T : 31
                3 2                L E    F    4                L13

                40                *                60
MTA1_YEAST : K E K E E V A K K C G I T P L Q V R V W F I N K R M R S K - : 53
MAT2_YEAST : K G L E N I M K N T S L S R I Q I K N W V S N R R R K E K T : 61
                K    E    6    K                63    6Q64    W    N4R    4    K

```

В чем биологический смысл выравнивания?

- *Буквы в одной колонке определяют сопоставление аминокислотных остатков двух белков*
- Сопоставленные остатки, по идее, должны иметь что-то общее в молекулах белка; что???

Предложение: биологический смысл имеет сопоставление одинаковых или функционально сходных остатков белка.

Эти остатки играют сходную роль.

Сопоставление непохожих остатков не имеет смысла.

Какое выравнивание “правильнее”?

```

          *           20           *
MTA1_YEAST : ----KSSISPCARAFLEQVFRK---QSLNS : 24
MAT2_YEAST : KPYRGHRFTKENVRILESWFAKNIENPYLDT : 31
          3 2           LE  F  4           L13
    
```

```

          40           *           60
MTA1_YEAST : KEKEEVAKKCGITPLQVRVWFINKRMRSK- : 53
MAT2_YEAST : KGLNLMKNTSLSRIQIKNWVSNRRRKEKT : 61
          K  E  6  K           63  6Q64  W  N4R  4  K
    
```

12 консервативных остатков


```

          *           20           *           4
MTA1_YEAST : K----SSISPCA-R-----A-----F-----LEQVFR : 17
MAT2_YEAST : KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLNLMK : 39
          K           3 2  R           A           5           LE  6  4
    
```

```

          0           *           60           *
MTA1_YEAST : RKQSLNSKKEKEEVAKKCGITPLQVRVWFINKRMRSK- : 53
MAT2_YEAST : NT-SL-SR-----IQIKNWVSNRRRKEKT : 61
          SL  S4           6Q64  W  N4R  4  K
    
```

13 “консервативных” остатков



Чтобы понять смысл
выравнивания, вернемся к тому,
что такое последовательность
аминокислотных остатков и что
такое белок

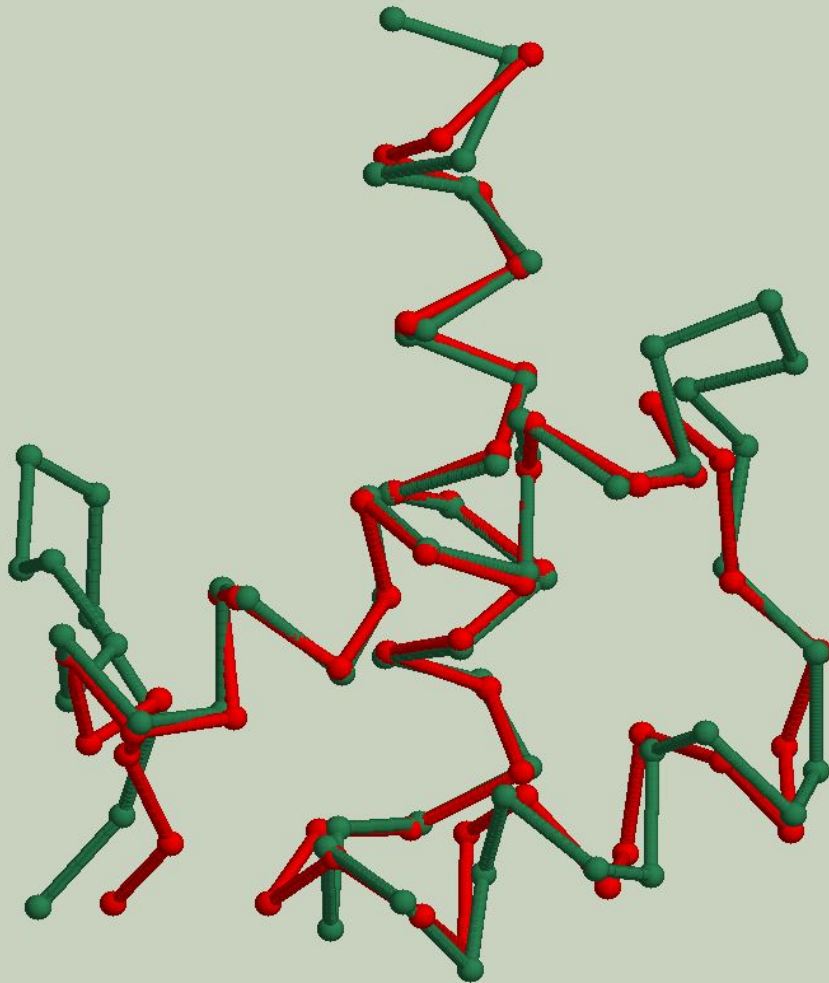
(i) Последовательность - удобный способ закодировать структурную (химическую) формулу молекулы белка (до посттрансляционных модификаций)

(ii) Белок - это большая молекула, сохраняющая в живой клетке постоянную пространственную структуру, т.е.- взаимное расположение ковалентно связанных атомов (конформацию)

(iii) Последовательность однозначно определяет в какую пространственную **структуру** свернется белок в клетке

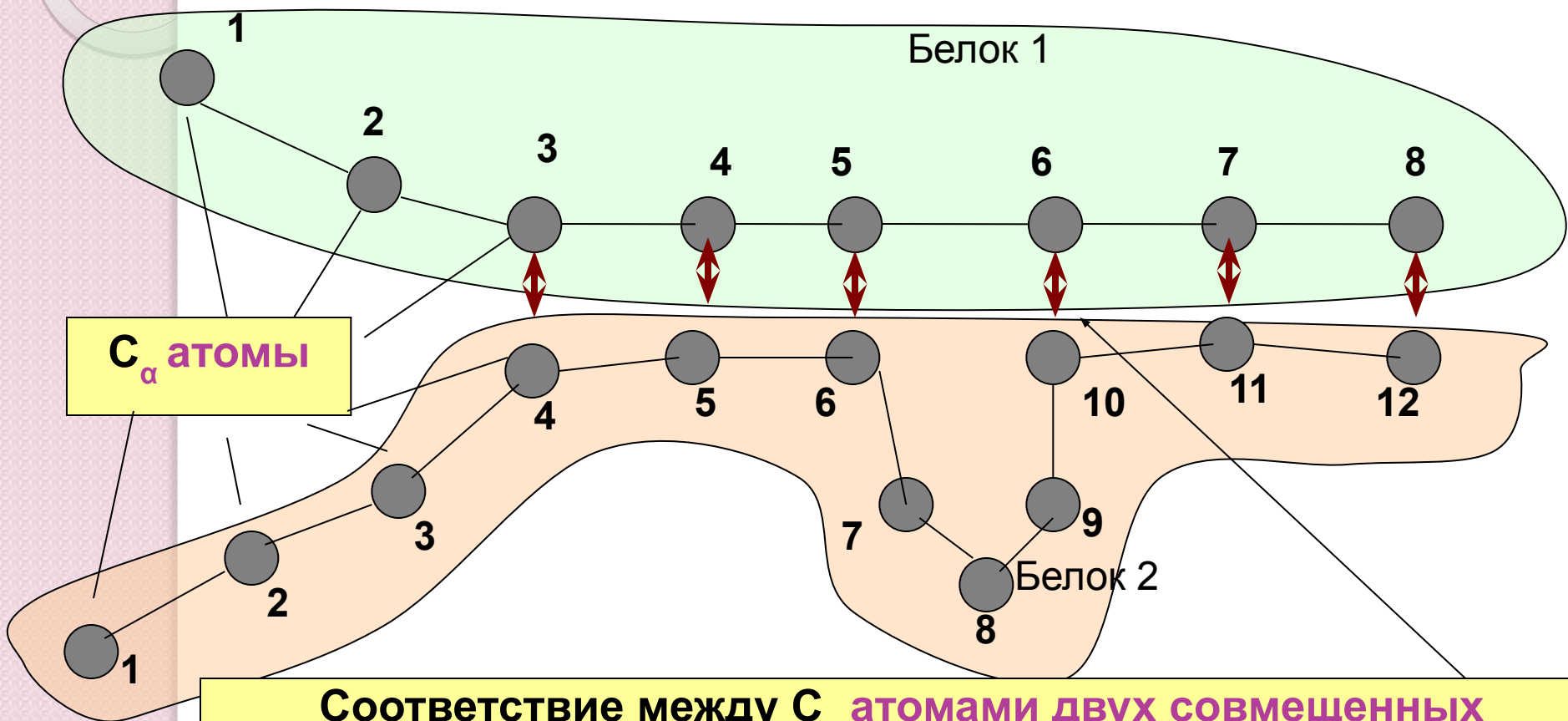
(iv) Функция белка в клетке **проявляется** только при сохранении **уникальной пространственной структуры**

Пространственное совмещение полипептидных цепей белков mta1_yeast и mat2_yeast



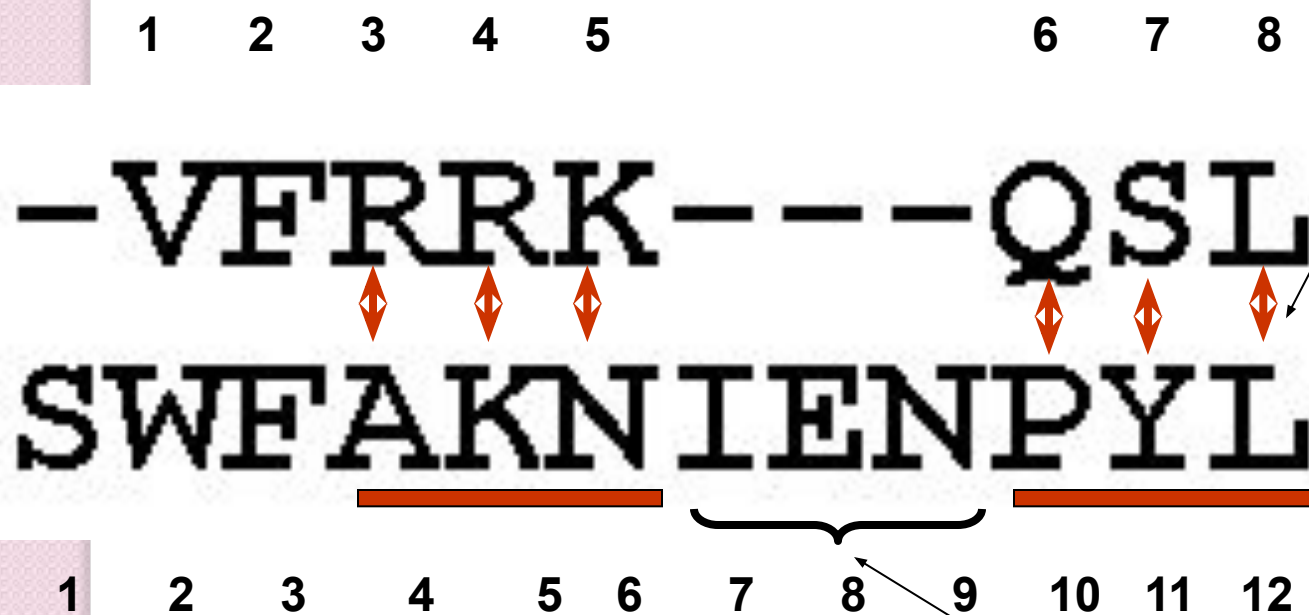
На плоской картинке
видно плохо 😞

Схематическое изображение совмещенных структур

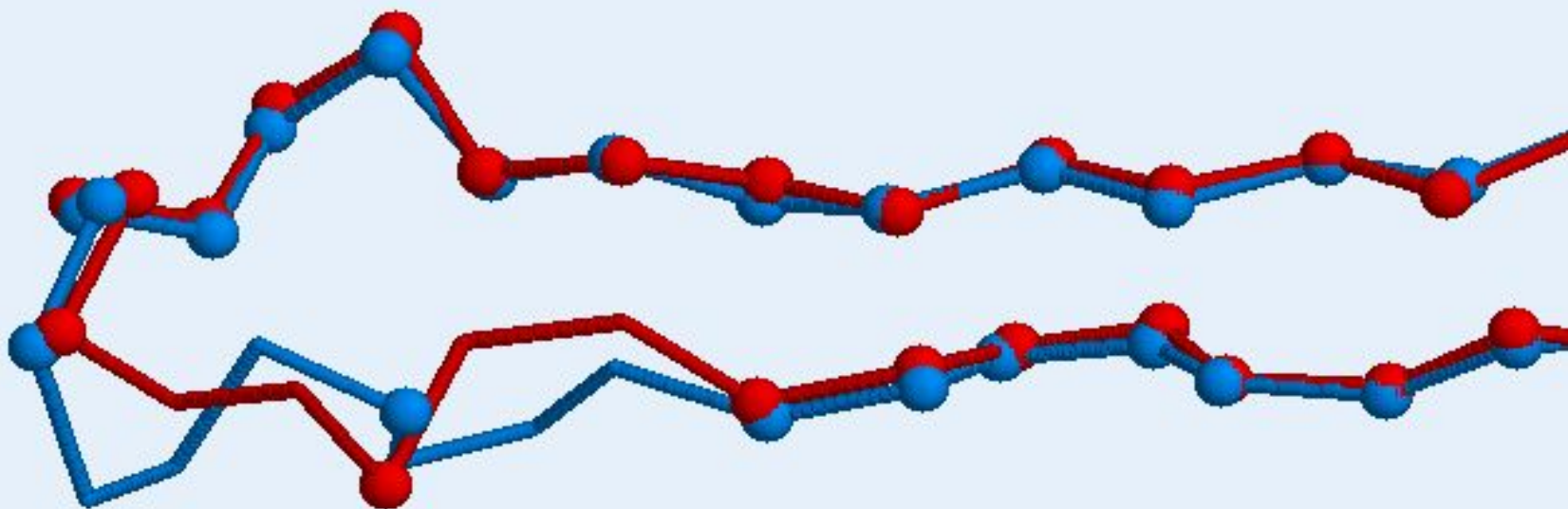


Соответствие между C_α атомами двух совмещенных структур,
основанное на близости в пространстве

Другой способ отобразить совмещение полипептидных цепей называется структурным выравниванием последовательностей



Совмещение структур и выравнивание последовательностей



		*		20		*	
Seq_A	:	LTGYGRWEAEFagnkae	--	sdt	aqg	KTrlAFAGLK	: 33
Seq_B	:	LTGYGQWEYNFqgn	se	gada	qtgn	KTrlAFAGLK	: 35
Aligned	:	AAAAAAAAAAAAAAAA	----			AAAAAAAAAAAA	: 27

Еще раз: разметка по совмещенным структурам

```
                *                20                *
Seq_A      : LTGYGRWEAEFagnkae--sdtaqqKTrlAFAGLK : 33
Seq_B      : LTGYGQWEYNFqgnnsegadaqtgnKTrlAFAGLK : 35
Aligned    : AAAAAAAAAAAAAAAAAA-----AAAAAAAAAA : 27
```


Биологически обоснованное выравнивание гомеодоменов

```

                *           20           *
MTA1_YEAST(1LE8:A) : ----KSSISSPQARAFLEQVFRRK---QSLNS : 24
MAT2_YEAST(1MNM:C) : KPYRGHRFTKENVRILESWFAKNIENPYLDT : 31
Aligned           : -----AAAAAAAAAAAAAAAAAAAA---AAAAA : 21
    
```

```

                40           *           60
MTA1_YEAST(1LE8:A) : KEKEEVAKKCGITPLQVRVWFINKRMRSK- : 53
MAT2_YEAST(1MNM:C) : KGLENLMKNTSLSRIQIKNWVSNRRRRKEKT : 61
Aligned           : AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA- : 50
    
```


Совмещение 5-и гомеодоменов



Множественное выравнивание гомеодоменов

```

          *           20           *           40           *           60
MTA1_YEAST(1LE8) : ----KSSISPQARAFLEQVFRRK---QSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK- : 53
HMP1_MOUSE(1AU7) : --KRRTTISIAAKDAERHFGEH---SKPSSQEIMRMAEELNLEKEVVRVWFCNRRQREKR : 56
VND_DROME(1NK2)  : KRKRRVLFITKAQTYELERRFRQQ---RYLSAPEREHLASLIRLPTQVKIWFQNHRYKTKR : 58
MAT2_YEAST(1MNM) : KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLNLMKNTSLSRIQIKNWSNRRRKEKT : 61
PBX1_HUMAN(1PUF) : ARRKRRNFNKQATEILNEYFYSHLSNPYPSEEAKEELAKKCGITVSQVSNWFGNKRIRYKK : 61
          Le F           e a           qv Wf N R K

```

Красным выделены консервативные (одинаковые у всех) остатки;

желтым – на 80% консервативные (одинаковые почти у всех)

остатки

```

          *           20           *           40           *           60
MTA1_YEAST(1LE8) : ----KSSISPQARAFLEQVFRRK---QSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK- : 53
HMP1_MOUSE(1AU7) : --KRRTTISIAAKDAERHFGEH---SKPSSQEIMRMAEELNLEKEVVRVWFCNRRQREKR : 56
VND_DROME(1NK2)  : KRKRRVLFITKAQTYELERRFRQQ---RYLSAPEREHLASLIRLPTQVKIWFQNHRYKTKR : 58
MAT2_YEAST(1MNM) : KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLNLMKNTSLSRIQIKNWSNRRRKEKT : 61
PBX1_HUMAN(1PUF) : ARRKRRNFNKQATEILNEYFYSHLSNPYPSEEAKEELAKKCGITVSQVSNWFGNKRIRYKK : 61
          Le F           e 6a 6 q6 Wf N R 4 K

```

Красным выделены консервативные и функционально консервативные остатки

Размеченное множественное выравнивание

```

                *           20           *
MTA1_YEAST : ----KSSISPQARAFLEQVERRK---QSLNSKEK : 27
HMP1_MOUSE : --KRRTTISIAAKDALERHFEHGEH---SKPSSQEI : 29
VND_DROME  : KRKRRVLFRTKAQTYELERRERQQ---RYLSAPER : 31
MAT2_YEAST : KPYRGHRFTKENVRILESWFAKNIENPYLDTKGL : 34
PBX1_HUMAN : ARKRRNFENKQATEILNEYFYSHLSNPYPSEEAK : 34
Aligned    : -----AAAAAAAAAAAAAAAAAA---AAAAAAA : 24

```

```

                40           *           60
MTA1_YEAST : EEVAKKCGITPLQVRVWFINKRMRSK- : 53
HMP1_MOUSE : MRMAEELNLEKEVVRVWFCNRRQREKR : 56
VND_DROME  : EHLASLIRLTPTQVKIWFQNHRYKTKR : 58
MAT2_YEAST : ENLMKNTSLSRIQIKNWVSNRRRKEKT : 61
PBX1_HUMAN : EELAKKCGITVSVSNWFGNKRIRYKK : 61
Aligned    : AAAAAAAAAAAAAAAAAAAAAAAAAAA??- : 47

```

Функции аминокислотных остатков

Leu16

```

                *           20           *
MTA1_YEAST : ----KSSISPQARAFLEQVERRK---QSLNSKEK : 27
HMP1_MOUSE : --KRRTTISIAAKDALERHFGEH---SKPSSQEI : 29
VND_DROME  : KRKRRVLF'TKAQTYELERREFRQQ---RYLSAPER : 31
MAT2_YEAST : KPYRGHRFTKENVRILESWFAKNIENPYLDTKGL : 34
PBX1_HUMAN : ARRKRRNFENKQATEILNEYFYSHLSNPYPSEEAK : 34
Aligned    : -----AAAAAAAAAAAAAAAAAAAA---AAAAA : 24
  
```


Pro442/
Lys442

```

                40           *           60
MTA1_YEAST : EEVAKKCGITPLQVRVWFINKRMRSK- : 53
HMP1_MOUSE : MRMAEELNLEKEVVRVWFCNRRQREKR : 56
VND_DROME  : EHLASLIRITPTQVKIWFQNHRYKTKR : 58
MAT2_YEAST : ENLMKNTSLSRIQIKNWVSNRRRKEKT : 61
PBX1_HUMAN : EELAKKCGITVSQVSNWFGNKRIRYKK : 61
Aligned    : AAAAAAAAAAAAAAAAAAAAAAAAAAAA??- : 47
  
```

Arg53

Trp48



В “правильном” выравнивании много консервативных аминокислотных остатков и функционально консервативных позиций

Выравнивание и эволюция

		*		20		*		4	
POLG_CXB4J	:	GAQVSTQKTGAHETSLSASGNSIIHYTNINYYKDAASNS						:	39
POLG_CXB4E	:	GAQVSTQKTGAHETSLSATGNSIIHYTNINYYKDAASNS						:	39
		0		*		60			
POLG_CXB4J	:	ANRQDFTQDPSKFTEPVKDVMIKSLPALN						:	68
POLG_CXB4E	:	ANRQDFTQDPSKFTEPVKDVMIKSLPALN						:	68

Последовательности белка оболочки из двух штаммов вируса Коксаки

..

		*		20	*		4	
POLG_CXB4J	:	GAQVSTQKTGAHETSL	SASGNSIIHYTN	INYYKDAASNS	:	39		
POLG_CXB4E	:	GAQVSTQKTGAHETSL	SATGNSIIHYTN	INYYKDAASNS	:	39		
POLG_HE71B	:	GSQVSTQRS	GSHENSNSATEG	STINYYTTINYYKDSYAAT	:	39		

		0	*		60	
POLG_CXB4J	:	ANRQDFTQDPSK	FTEPVKDVMIKSLPALN	:	68	
POLG_CXB4E	:	ANRQDFTQDPSK	FTEPVKDVMIKSLPALN	:	68	
POLG_HE71B	:	AGKQSLKQDP	KEANPVKDI	FTEMAAPLK	:	68

Последовательности белка оболочки из двух штаммов вируса Коксаки и энтеровируса человека

Аминокислотные остатки в одной колонке биологически обоснованного выравнивания, как правило, "произошли" из одного и того же остатка - их общего предка

Алгоритмические решения проблемы воплощены в программах

Программы выравнивания
последовательностей тестируются путем
сравнения с биологически обоснованными –
построенными по совмещению структур –
выравниваниями

Существуют базы данных структурных
выравниваний последовательностей
(BAlivAse и др.)

Предположим, известны структуры
родственных белков и, значит,

биологически обоснованное

выравнивание последовательностей

- При $> 60\%$ совпадающих букв любая современная программа даст (почти) правильный результат
- При $< 20\%$ совпадающих букв (такие примеры существуют) ни одна программа не даст правильного выравнивания
- Между 20% и 60% , обычно, результат программы частично правилен

● Применения

- «**ЗОЛОТОЙ СТАНДАРТ**» для выравнивания высоко гомологичных белков – выявление общего предка
- идентификация общих **значимых элементов** структуры для негомологичных белков
- кластеризация белков (разбиение на **белковые семейства**) на основе структурной близости

● Выравнивание должно отражать сходство структур

- совпадение общих структурных и функциональных элементов

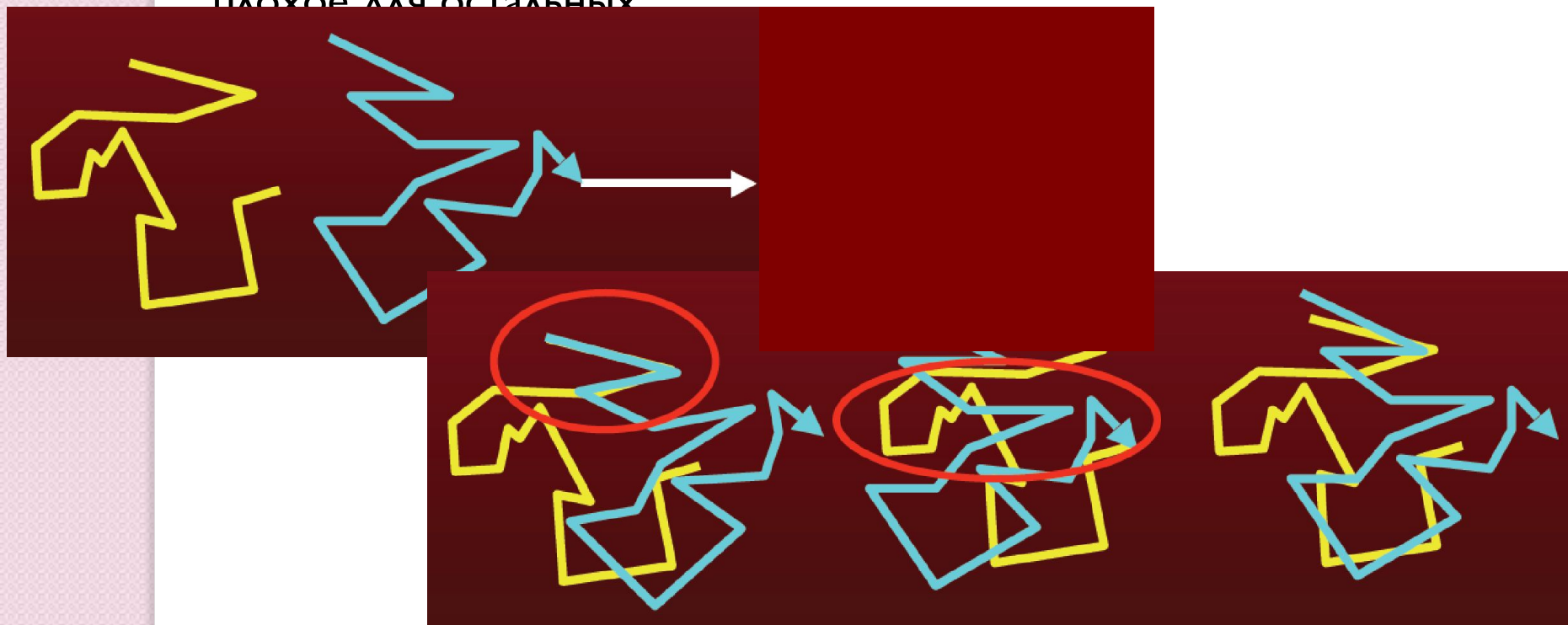
● Проблема: **ОПТИМУМ В ВЫЧИСЛЕНИЯХ**

≠

ОПТИМУМУ В БИОЛОГИИ

Структурное выравнивание: постановка задачи

- Для двух пространственных структур найти соответствие между атомами, обеспечивающее наилучшее «выравнивание»
 - для большинства атомов достигается минимум с.к.о.
 - **проблема:** «идеальное» выравнивание для нескольких атомов и плохое для остальных



Структурное выравнивание:

оценка результата

● Критерии

- число соответствий между АК
- суммарное евклидово расстояние между выровненными АК
- доля идентичных АК среди выровненных
- число введенных делеций
- размер сравниваемых белков
- консерватизм окружения известных активных центров

● Универсальных критериев не существует

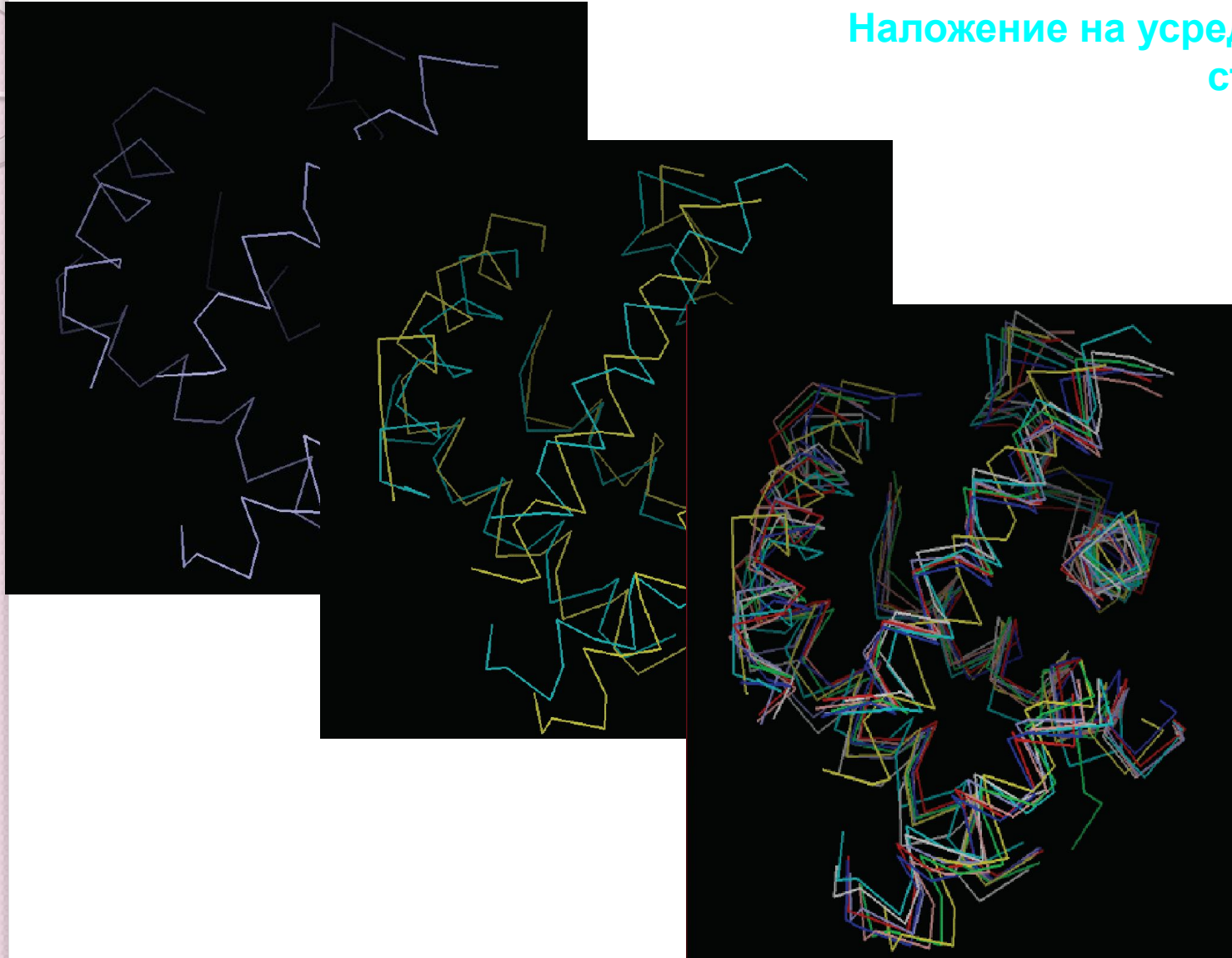


Замечание

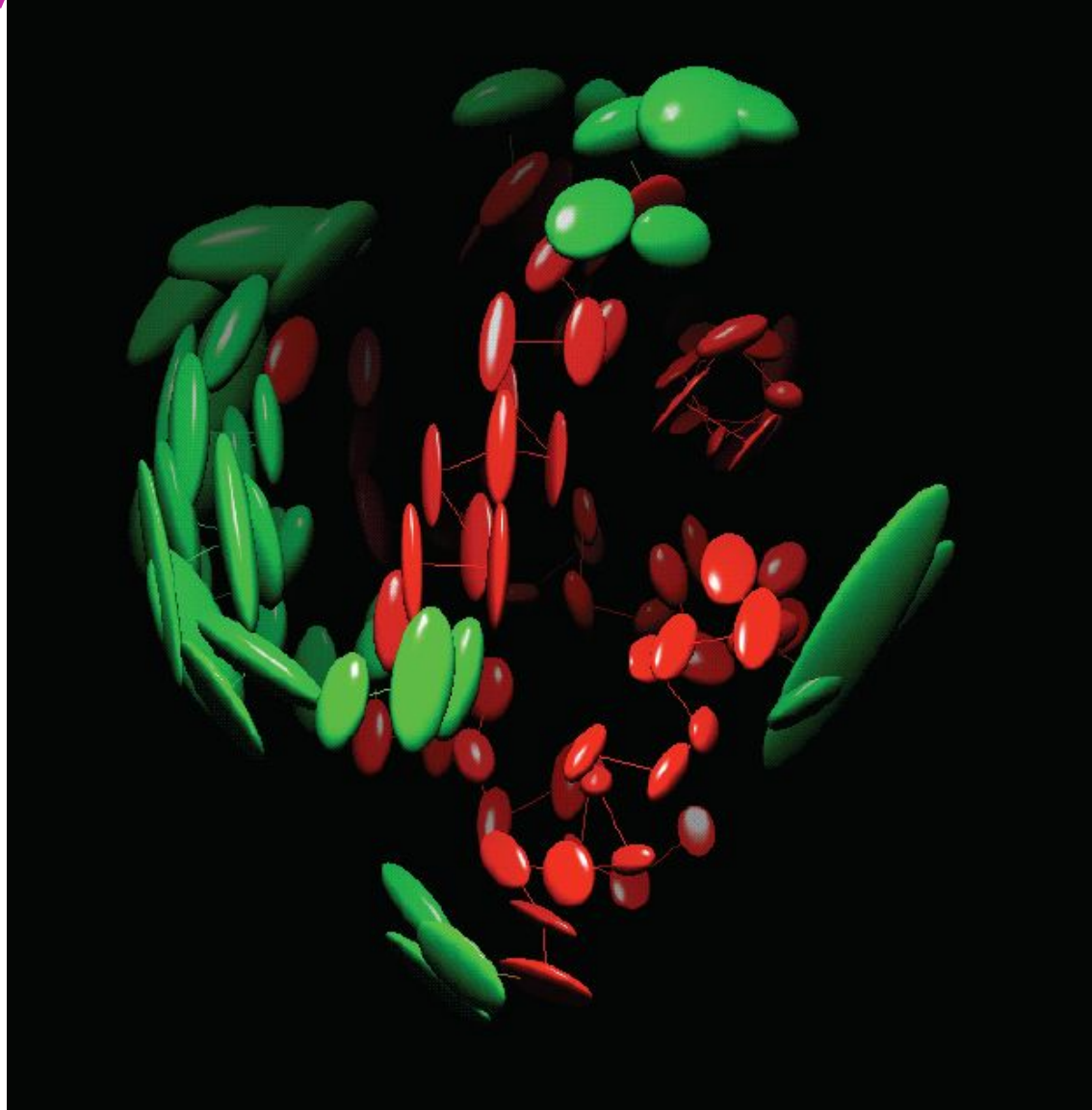
- отличие от поиска минимума евклидова расстояния при известном соответствии атомов
- с к о используется только в качестве метрики

Структурное выравнивание: наложение пространственных структур

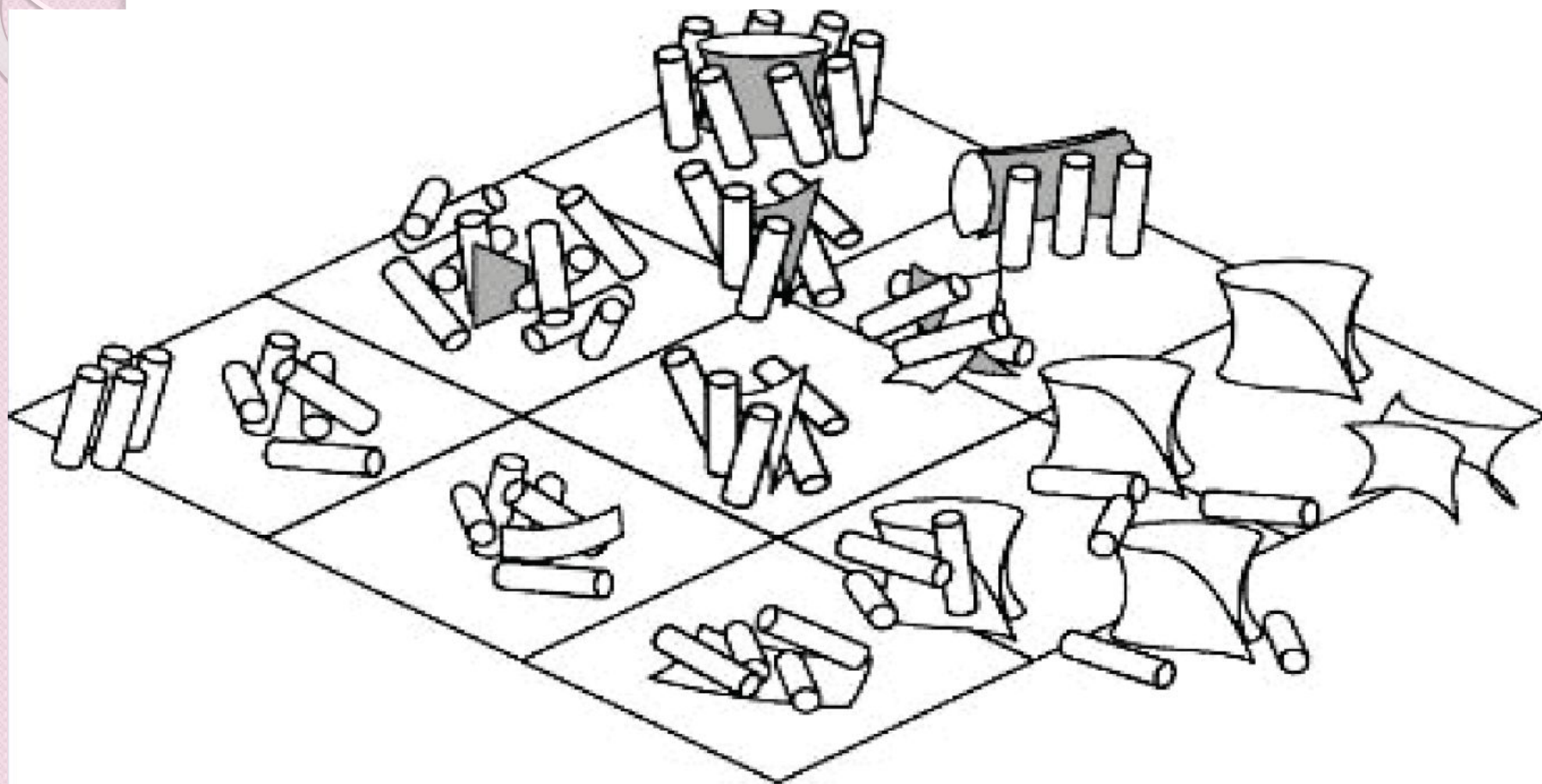
Наложение на усредненную
структуру



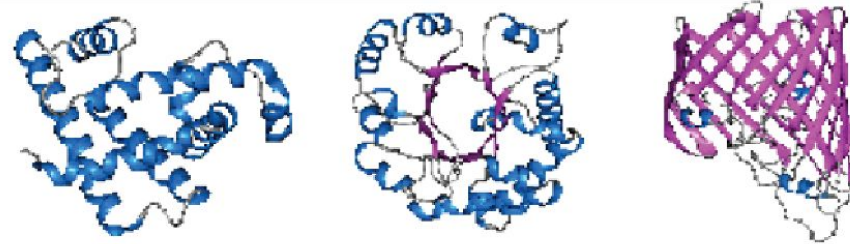
Структурное выравнивание: наложение пространственных структур



Структурное выравнивание: различные классы белковых структур (I)



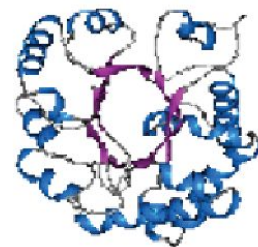
Структурное выравнивание: различные классы белковых структур (2)



α

$\alpha\&\beta$

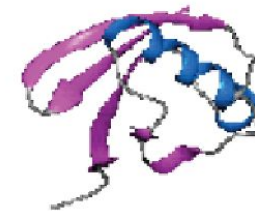
β



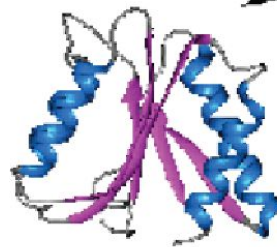
TIM barrel



Sandwich



Roll



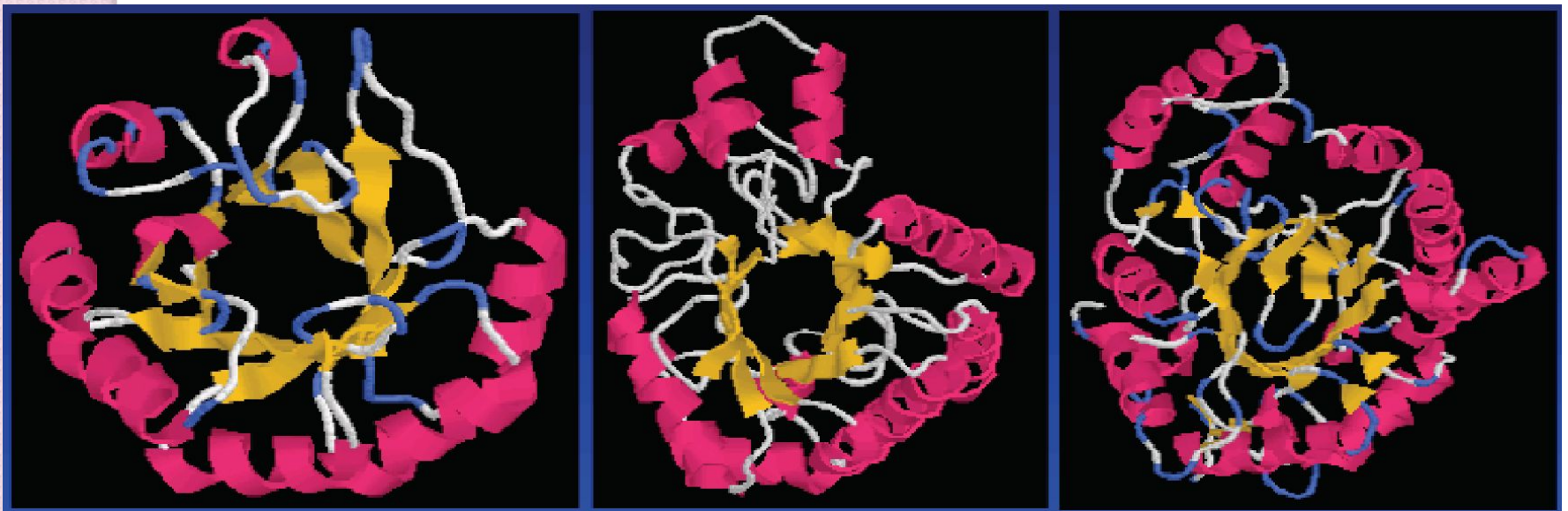
flavodoxin
(4fxn)



β -lactamase
(1 mblA1)

Структурное выравнивание: различные классы белковых структур (3)

Разные суперсемейства «бочонков»



Поиск структурного выравнивания

«вручную»

◆ Класс

- похожие вторич. структуры
- все α , все β , $\alpha + \beta$, α/β

◆ Слой (fold)

- значительное структурное сходство
- сходная организация вторичной структуры

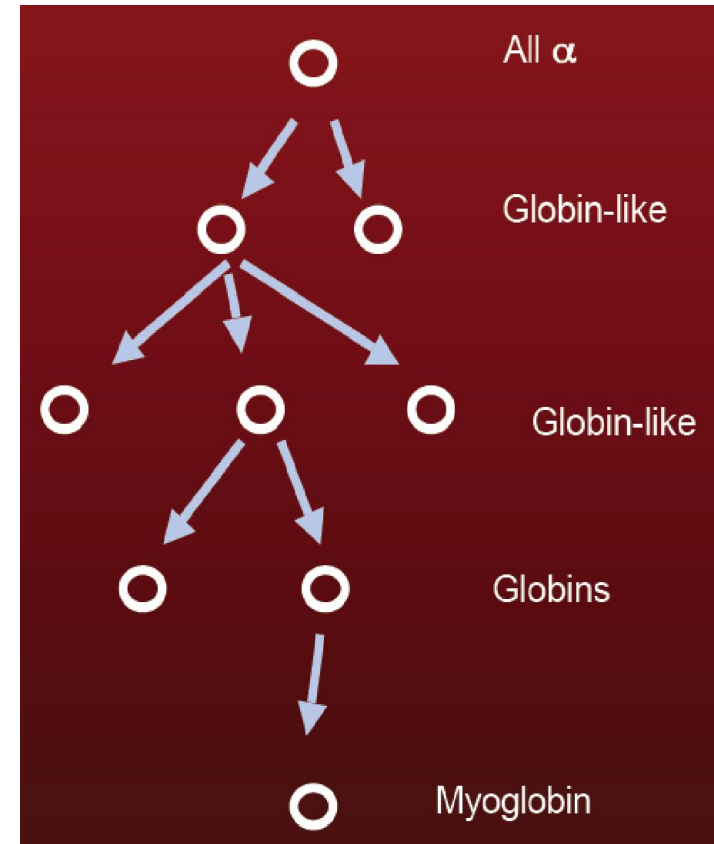
◆ Суперсемейство (топология)

- предположительный общий предок

◆ Семейство

- очевидные эволюционные отношения
- гомологичность последовательностей $> 25\%$

◆ Конкретный белок













Пример инструментария: Structural Classification Of Proteins (SCOP)

SCOP: Root: scop - Netscape

File Edit View Go Communicator Help

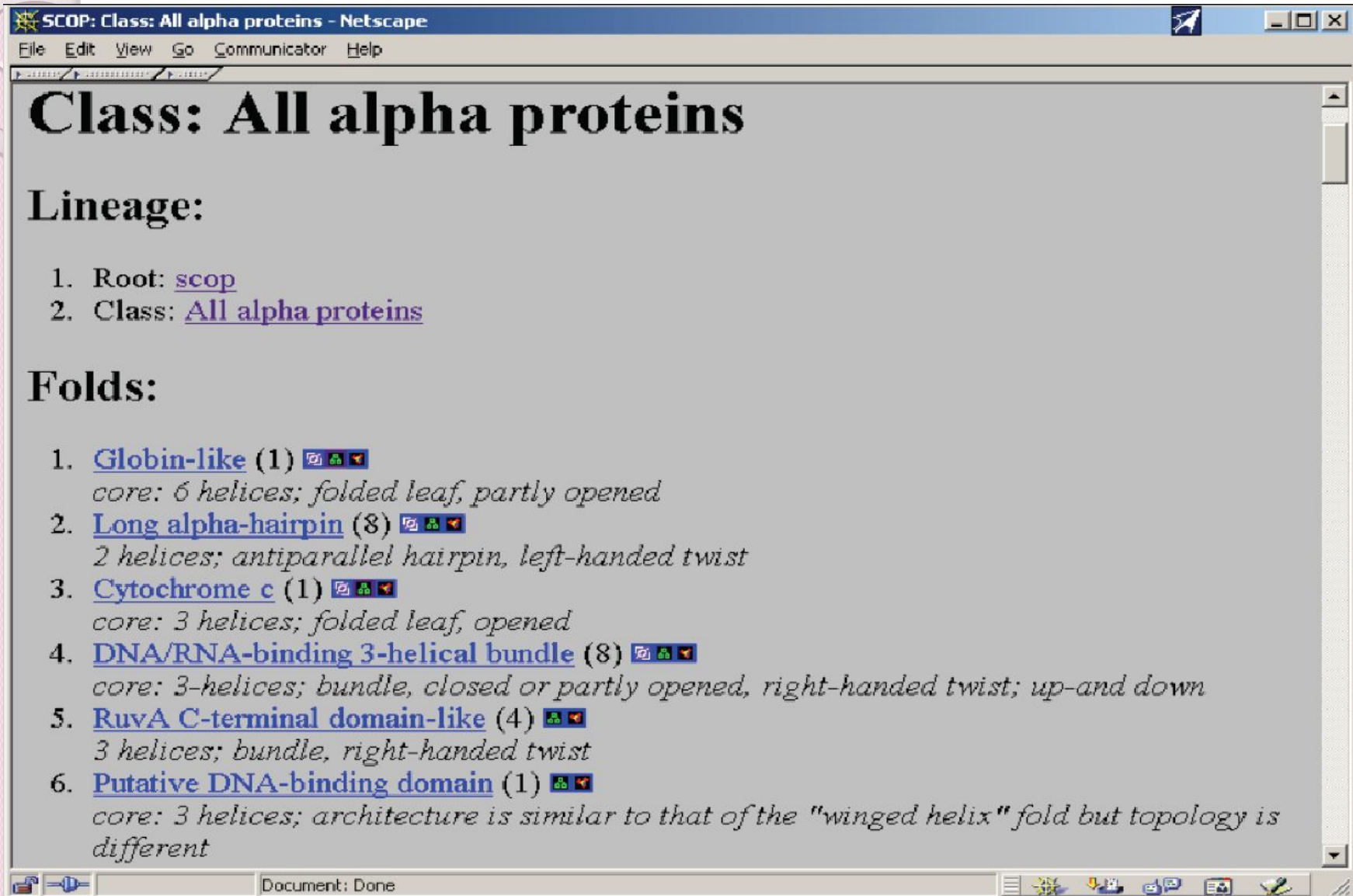
Root: scop

Classes:

1. [All alpha proteins](#) (128)   
2. [All beta proteins](#) (87)   
3. [Alpha and beta proteins \(a/b\)](#) (93)   
Mainly parallel beta sheets (beta-alpha-beta units)
4. [Alpha and beta proteins \(a+b\)](#) (168)   
Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. [Multi-domain proteins \(alpha and beta\)](#) (25)   
Folds consisting of two or more domains belonging to different classes
6. [Membrane and cell surface proteins and peptides](#) (11)   
Does not include proteins in the immune system
7. [Small proteins](#) (52)   
Usually dominated by metal ligand, heme, and/or disulfide bridges
8. [Coiled coil proteins](#) (5)   
9. [Low resolution protein structures](#) (10)  
10. [Peptides](#) (65)   
Peptides and fragments
11. [De](#)
Exp

<http://scop.stanford.edu>
<http://scop.mrc-lmb.cam.ac.uk/scop/>

Пример инструментария: SCOP (прод.)









The screenshot shows a Netscape browser window with the title "SCOP: Class: All alpha proteins - Netscape". The address bar is empty. The main content area displays the following information:

Class: All alpha proteins

Lineage:

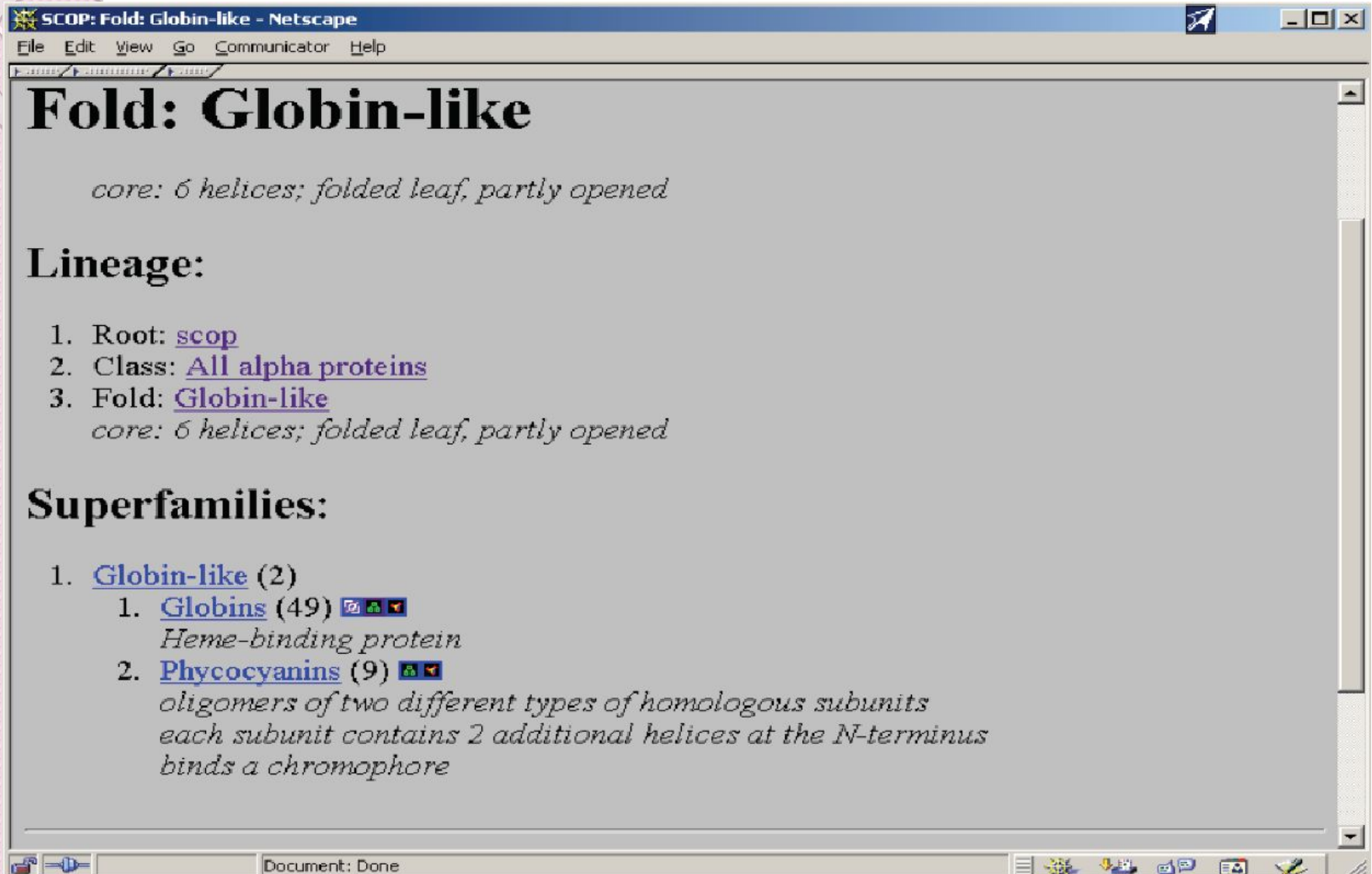
1. Root: [scop](#)
2. Class: [All alpha proteins](#)

Folds:

1. [Globin-like](#) (1) 
core: 6 helices; folded leaf, partly opened
2. [Long alpha-hairpin](#) (8) 
2 helices; antiparallel hairpin, left-handed twist
3. [Cytochrome c](#) (1) 
core: 3 helices; folded leaf, opened
4. [DNA/RNA-binding 3-helical bundle](#) (8) 
core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down
5. [RuvA C-terminal domain-like](#) (4) 
3 helices; bundle, right-handed twist
6. [Putative DNA-binding domain](#) (1) 
core: 3 helices; architecture is similar to that of the "winged helix" fold but topology is different

The browser's status bar at the bottom shows "Document: Done" and various system icons.

Пример инструментария: SCOP (прод.)



The screenshot shows a Netscape browser window with the title "SCOP: Fold: Globin-like - Netscape". The address bar is empty. The main content area displays the following information:






Fold: Globin-like

core: 6 helices; folded leaf, partly opened

Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#)
3. Fold: [Globin-like](#)
core: 6 helices; folded leaf, partly opened

Superfamilies:

1. [Globin-like](#) (2)
 1. [Globins](#) (49)   
Heme-binding protein
 2. [Phycocyanins](#) (9)  
*oligomers of two different types of homologous subunits
each subunit contains 2 additional helices at the N-terminus
binds a chromophore*

The browser's status bar at the bottom shows "Document: Done" and a taskbar with various system icons.

Пример инструментария: SCOP (прод.)



The screenshot shows a Netscape browser window with the title "SCOP: Family: Globins - Netscape". The address bar is empty. The main content area displays the following information:

Family: Globins

Heme-binding protein

Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#)
3. Fold: [Globin-like](#)
core: 6 helices; folded leaf, partly opened
4. Superfamily: [Globin-like](#)
5. Family: [Globins](#)
Heme-binding protein

Protein Domains:

1. Hemoglobin I
 1. [Ark clam \(*Scapharca inaequivalvis*\)](#) (10) 
 2. [Clam \(*Lucina pectinata*\)](#) (4) 
2. Glycera globin
 1. [Marine bloodworm \(*Glycera dibranchiata*\)](#) (4) 
3. Myoglobin

The browser's status bar at the bottom shows "Document: Done" and various system icons.

<http://scop.stanford.edu>

<http://scop.mrc-lmb.cam.ac.uk/scop/>

Как распознать близость структур?

- На глаз

- Алгоритмически

- точечные методы: установление соответствий по точечным свойствам (расстояниям)
- анализ вторичной структуры: установление соответствий по векторам, изображающим элементы вторичной структуры

- Четыре метода, оперирующих прототипами

- **STRUCTAL** (Levitt, Subbiah, Gerstein)
- **DALI** (Holm, Sander)
- **LOCK** (Singh, Brutlag)

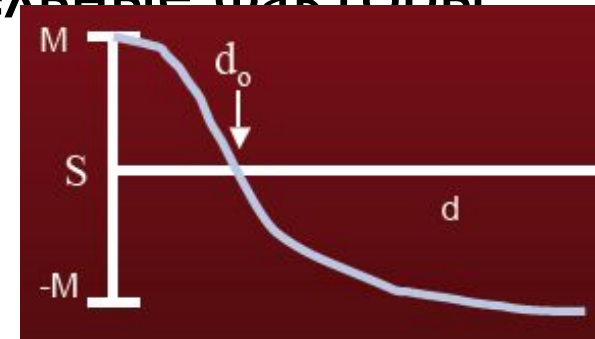
Структурное выравнивание при помощи прототипов: STRUSTAL

- Итерационное динамическое программирование для улучшения случайно выбранного начального выравнивания
- Шаги алгоритма
 - 1) начать с произвольного **набора соответствий** между двумя структурами (выравнивание пос-стей, вторичных структур, на глаз, случайное)
 - 2) **выровнять две структуры**, исходя из текущего набора соответствий
 - 3) построить **матрицу весов** (Нидлмана-Вунша), исходя из расстояний между всевозможными парами точек
 - 4) **ДП**: обратное движение по матрице весов для нахождения выравнивания с наибольшим суммарным весом
 - 5) повторение шагов 2-4, пока суммарный вес не перестанет меняться
- **Метод эвристический, не гарантирует результата, зависит от выбора начального выравнивания**

Структурное выравнивание при помощи прототипов: STRUSTAL (прод.)

- Оценка выравнивания: чем лучше выравнивание, тем выше суммарный вес
 - возможность учесть дополнительные факторы

● Вес



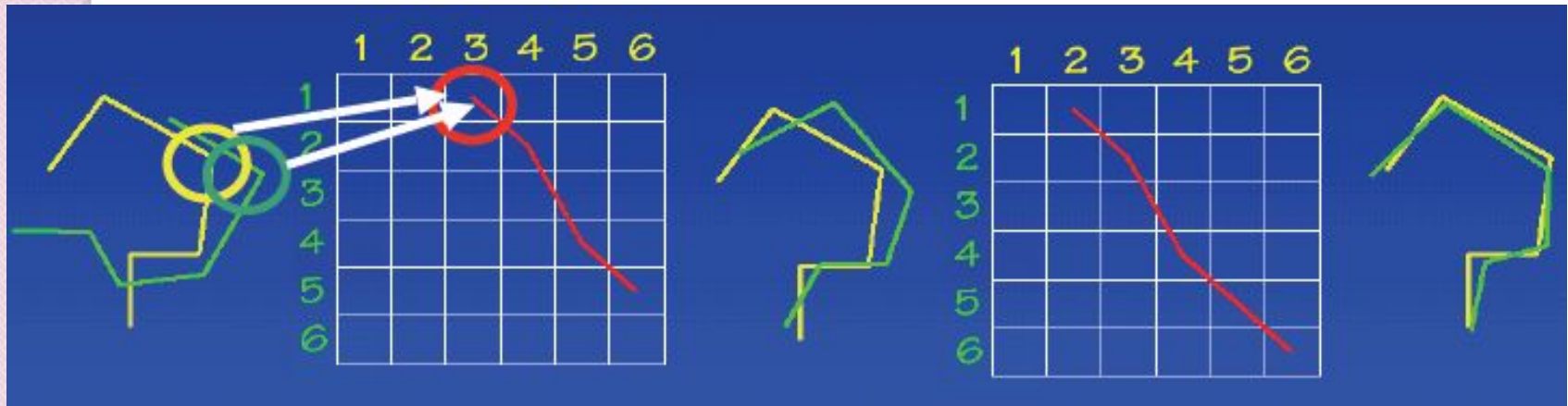
$$S(d) = M \left\{ \frac{2}{1 + (d/d_0)^2} - 1 \right\}$$

где M – максимальный ожидаемый вес, d – измеряемая величина (e.g. расстояние между точками), d_0 – значение d , соответствующее $M = 0$

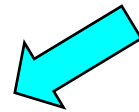
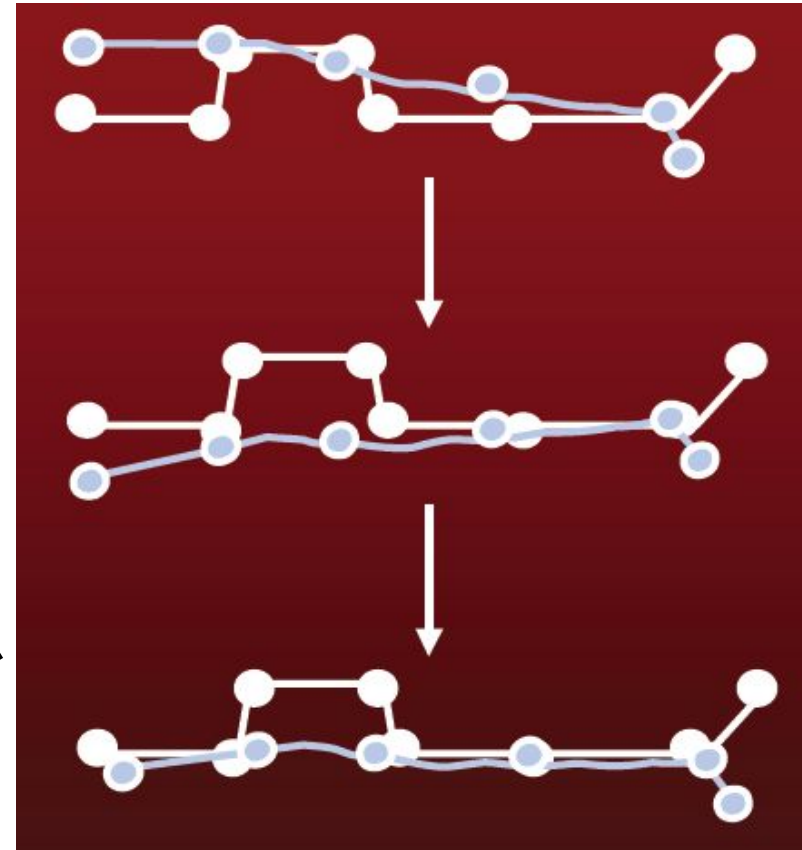
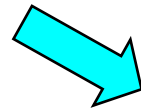
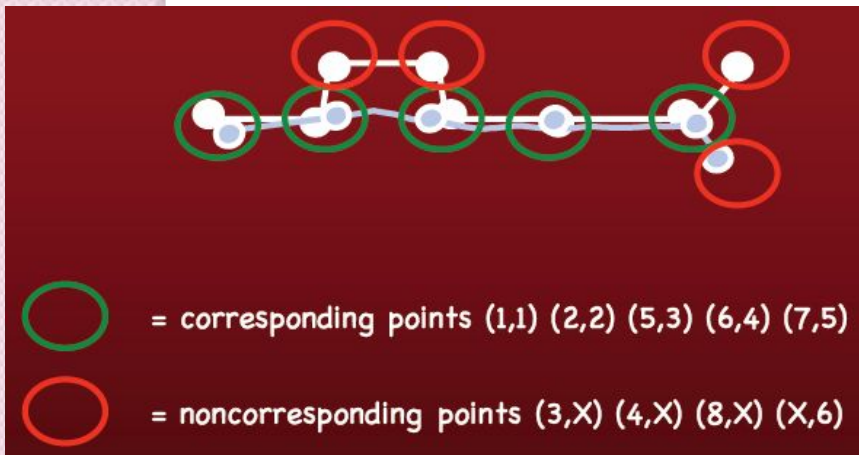
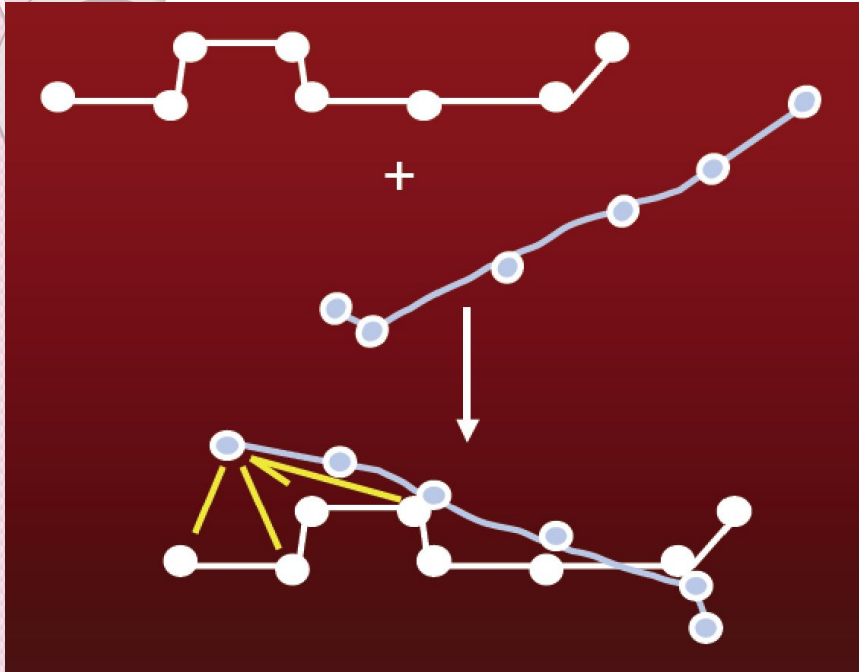
□ $0 < d < d_0$... $d > d_0$...

Структурное выравнивание при помощи прототипов: STRUCSTAL (прод.)

Итерационное динамическое программирование

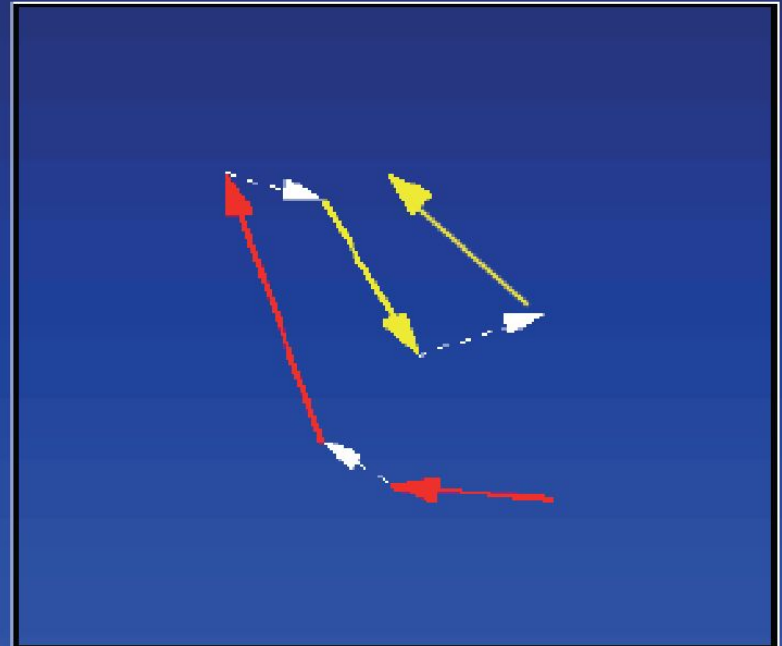


Структурное выравнивание при помощи прототипов: STRUSTAL (прод.)



Структурное выравнивание при помощи прототипов: LOCK

- Основная идея:
 - элементы вторичной структуры представляются при помощи векторов
 - быстрый поиск похожих структур



Структурное выравнивание при помощи прототипов: LOCK (прод.)

Сравнение «векторов вторичной структуры»



Orientation Independent Scores:

$$S = S(|\text{angle}(i,k) - \text{angle}(p,r)|)$$

$$S = S(|\text{distance}(i,k) - \text{distance}(p,r)|)$$

$$S = S(|\text{length}(k) - \text{length}(r)|)$$

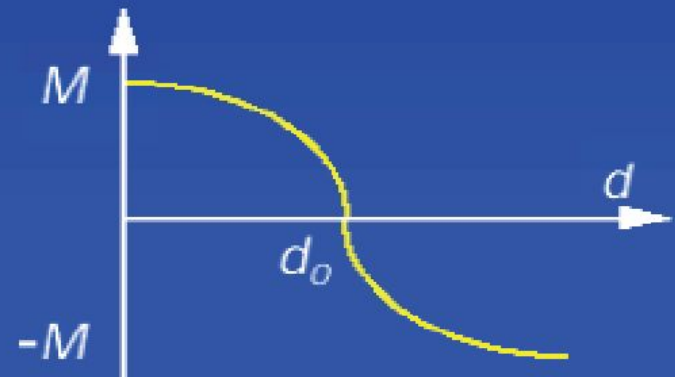


Orientation Dependent Scores:

$$S = S(\text{angle}(k,r))$$

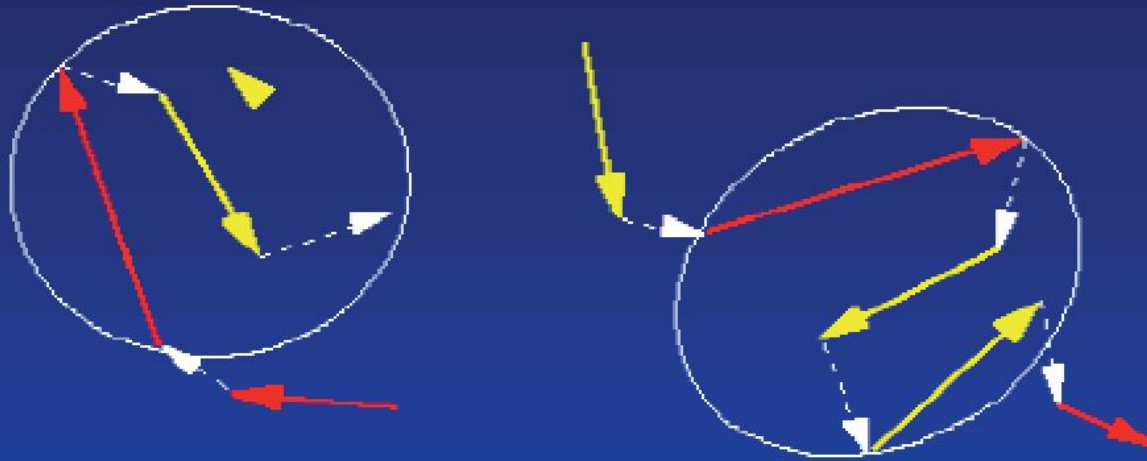
$$S = S(\text{distance}(k,r))$$

$$S(d) = \left[\frac{2M}{1 + \left[\frac{d}{d_0} \right]^2} - M \right]$$



Структурное выравнивание при помощи прототипов: LOCK (прод.)

Выравнивание «векторов вторичной структуры»



	H	H	S	S
S				
H				
S				
S				
H				

Best local alignment : **HHSS**
SHSSH

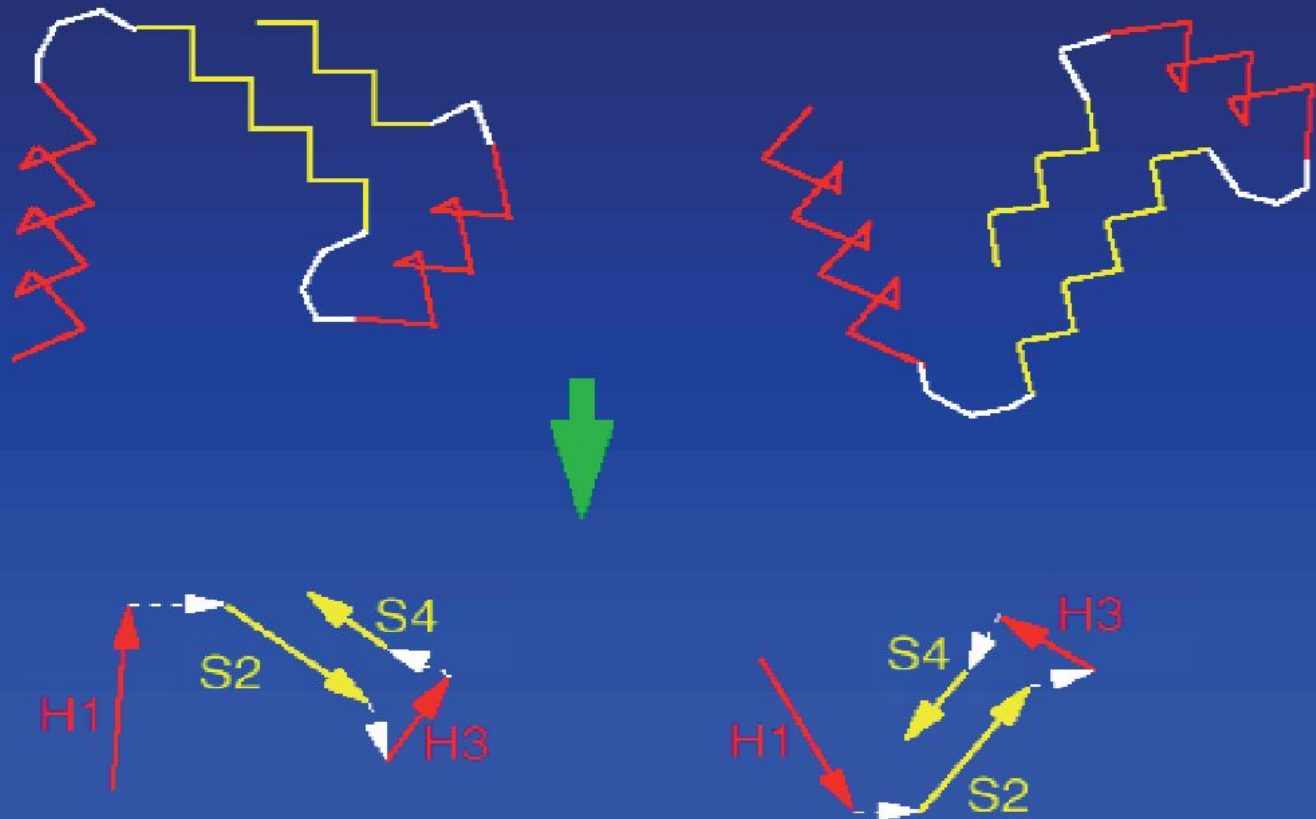
Структурное выравнивание при помощи прототипов: LOCK (прод.)

Шаги алгоритма

- 1) определить локальные элементы вторичной структуры
- 2) построить начальное наложение структур методом ДП, используя
 - выбранную функцию веса
 - векторное представление элементов вторичной структуры
- 3) определить ближайших соседей, минимизируя евклидовы расстояния
- 4) удалить лишние атомы, чтобы получить минимальное с.к.о.

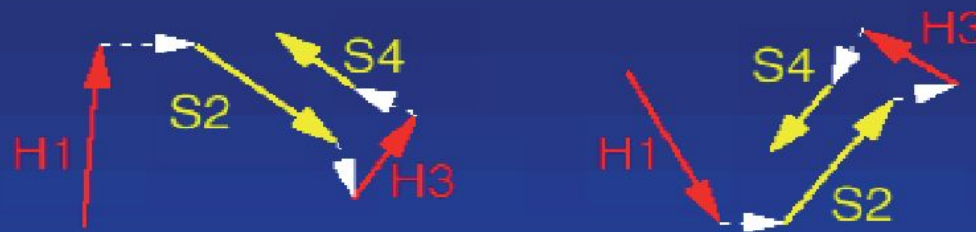
Структурное выравнивание при помощи прототипов: шаги алгоритма LOCK (I)

Step 1: Local Secondary Structure Superposition



Структурное выравнивание при помощи прототипов: шаги алгоритма LOCK (1a)

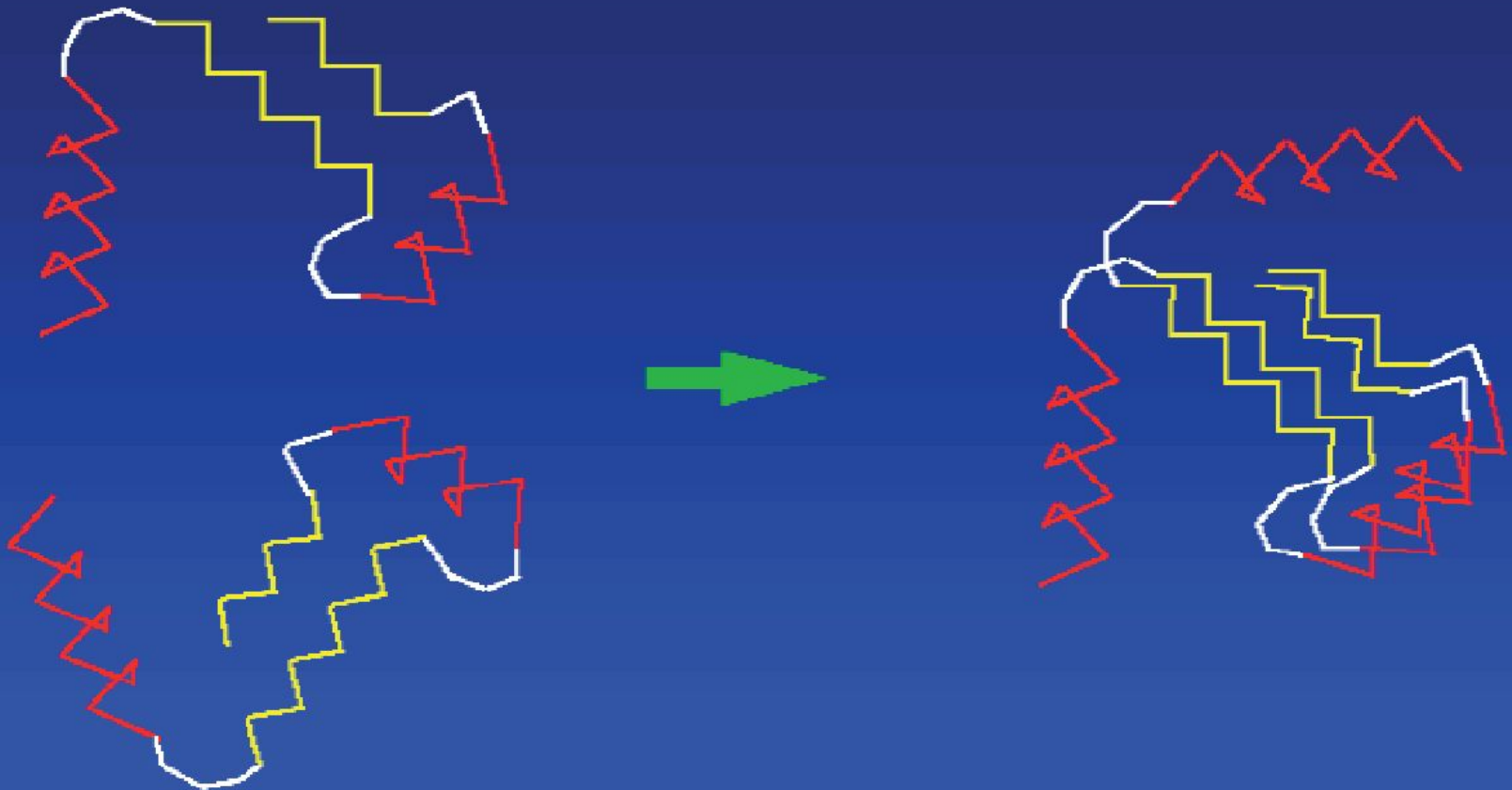
Step 1: Local Secondary Structure Superposition



pair	# of aligned vectors	total alignment score
H1,S2	2	27
S2,H3	3	65
H3,S4	3	<u>71</u>
S2,S4	3	68

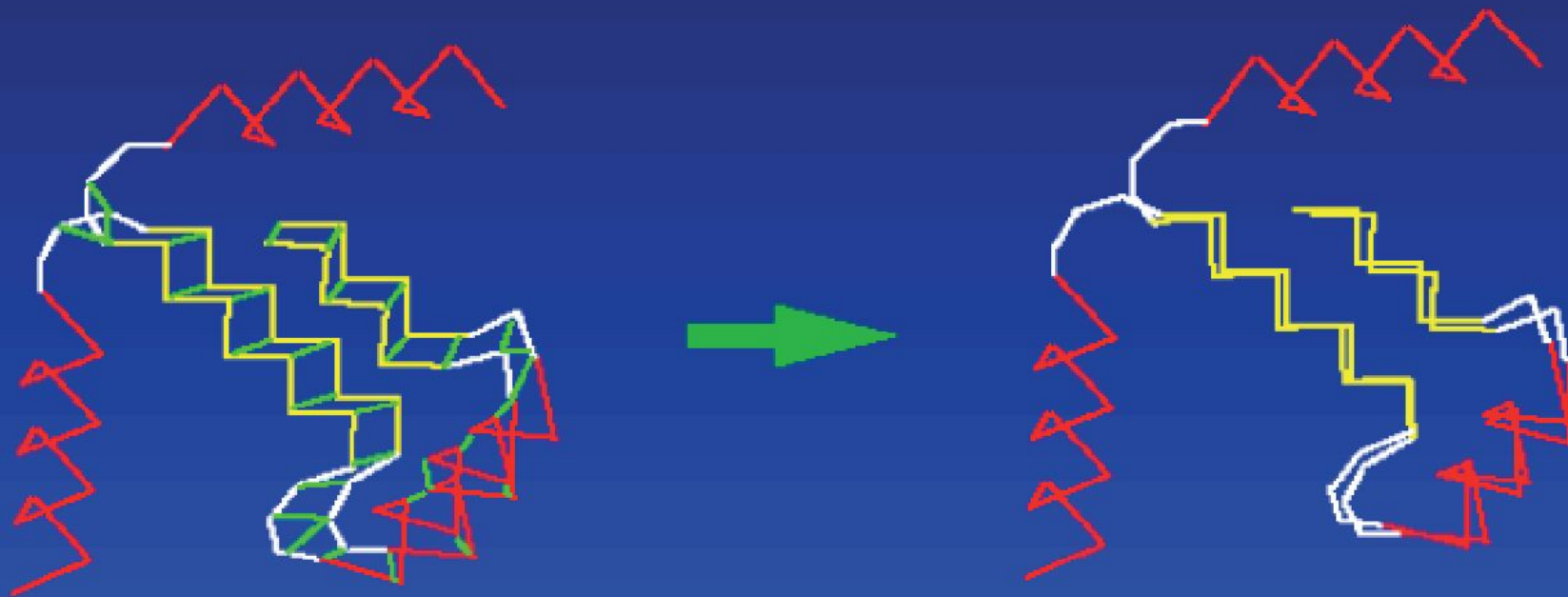
Структурное выравнивание при помощи прототипов: шаги алгоритма LOCK (Ib)

Step 1: Local Secondary Structure Superposition



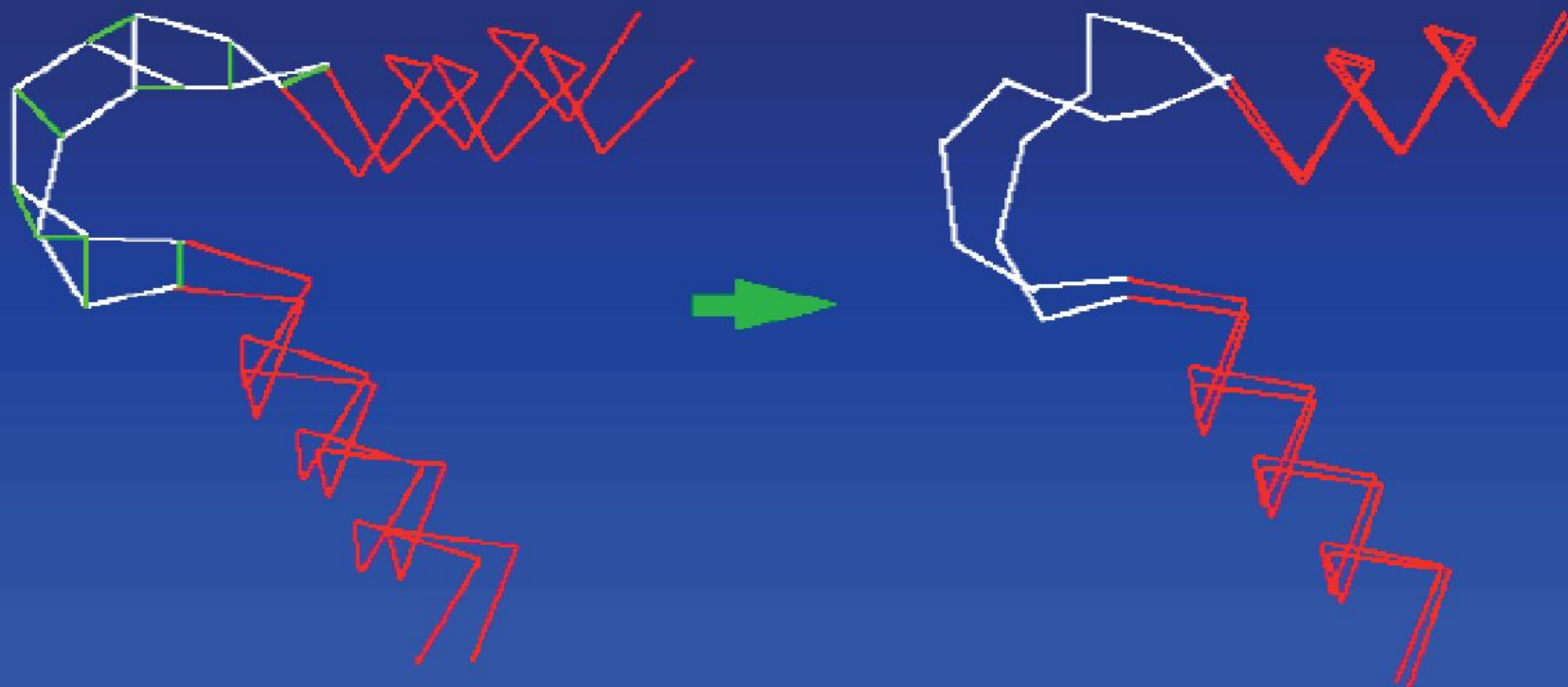
Структурное выравнивание при помощи прототипов: шаги алгоритма LOCK (2)

Step 2: Atomic Superposition



Структурное выравнивание при помощи прототипов: шаги алгоритма LOCK (3)

Step 3: Core Superposition




Структурное выравнивание:

«за» и «против»



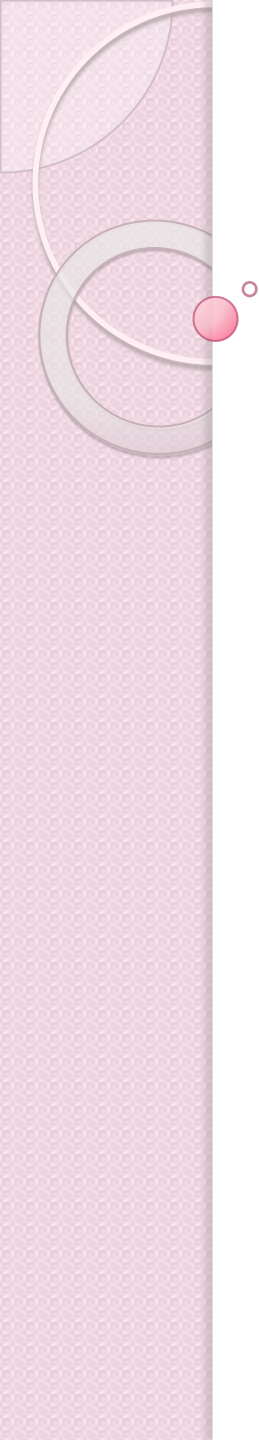
- ◆ «Золотой» стандарт для выравнивания пос-стей
- ◆ Трехмерная структура часто неизвестна
- ◆ Структурное выравнивание не всегда отражает ход эволюции
 - точная последовательность вставок/замен/делеций неизвестна



ПРОБЛЕМА: как построить
“правильное” выравнивание
последовательностей белков если
структуры белков неизвестны?

На сегодня известны:

- более 10 млн(!!!) последовательностей белков (включая фрагменты и трансляты)
- пространственные структуры около 70 тыс. белков



Дякую за увагу
Благодарю за внимание
Thank you for your attention