ОБЩ ПРЕГЛЕД НА МЕТОДИТЕ И АЛГОРИТМИТЕ ЗА МНОЖЕСТВЕНО ПОДРАВНЯВАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

ПРОФ. БОРОВСКА

ФОРМУЛИРАНЕ НА ПРОБЛЕМА

- Дадени са *m* секвенции S_i, всяка с дължина n_i , $i = 1, 2, \ldots, m$:

$$S := \begin{Bmatrix} S_1 = & (S_{11}\, S_{12} \ \ldots & S_{1n_1}) \\ S_2 = & (S_{21}\, S_{22} \ \ldots & S_{2n_2}) \\ S_m = & (S_{m1}\, S_{m2} \ \ldots & S_{mn_m}) \end{Bmatrix}$$

Първата стъпка е да се изравнят дължините на секвенциите.

Ако означим с L дължината на най-дългата секвенция, т.е. $L = MAX\ (n_1, n_2, \ldots, n_m)$, то всяка от секвенциите S_i се запълва с $(L\text{-}n_i)$ на брой празни позиции.

$$S' := \begin{Bmatrix} S'_1 = & (S''_{11}\, S'_{12} \ \ldots & S'_{1n_1}) \\ S'_2 = & (S_{21}\, S_{22} \ \ldots & S_{2n_2}) \\ S'_m = & (S_{m1}\, S_{m2} \ \ldots & S_{mn_m}) \end{Bmatrix}$$

Оптимално глобално подравняване

- *Оптималното глобално подравняване* A* на две секвенции *s* и *t* е такова подравняване A(s,t), при което се получава максимална стойност на общата оценъчна функция *M*(*A*) в сравнение с всички възможни подравнявания.

$$A^* = \max M(A_i)$$

- *Намирането на оптималното глобално подравняване A* е комбинаторен оптимизационен проблем и обхваща следните стъпки:*

1. Генериране на всички възможни подреждания;
2. Изчисляване на качеството на всички подреждания *M*;
3. Селекция на оптималното подравняване *A** с максимално качество *M*;*

избор на оценъчната матрица

- Най-често често използваната оценъчна функция при множественото подравняване на секвенции е сумата по двойки ("Sum of Pairs" - SP).
- При този метод се изчислява за всяка колона сумата от съвпаденията / разликите за всяка двойка секвенции, като за протеини се използва матрицата PAM или BLOSUM, и добавяйки „наказателни точки“ за празните позиции.
- Недостатък на SP методологията е, че относителната разлика между правилното и неправилното подравняване намалява с увеличаването на броя на участващите в подравняването секвенции.

оценъчна функция за качеството на подравняването (*Alignment scoring function*)

- взема се предвид цената на подравняването (alignment cost) на два символа x_i и y_j и се оценява с функцията $\sigma(x_i,y_j)$ по следния начин:

a) Вмъкване на празна позиция в секвенцията $\sigma(-,a) = \sigma(a,-) = -1$

b) Различни символи на секвенциите в една и съща позиция (колона)

$$\sigma(a,b) = -1 \text{ if } a \neq b$$

c) Еднакви символи на секвенциите в една и съща позиция (колона)

$$\sigma(a,b) = 1 \text{ if } a = b$$

Качеството на подравняването на целите секвенции

се оценява посредством сумата от оценките на отделните символи в секвенциите:

$$M = \sum_{i=1}^{c} \sigma(x_i, y_i)$$

Резултатите се подобряват като се използва по-реалистична оценъчна функция, която е инспирирана от биологията - матрица на заместванията.

NP-труден проблем

- Показано е, че проблемът за оптималното глобално множественото подравняване с най-популярната схема за оценяване „сума на двойки“ SP (Sum-of-Pairs) е NP-труден за всяка оценяваща матрица в широк клас M, който включва повечето матрици, които действително се използват в биологичните приложения.
- Проблемът остава NP-труден дори ако секвенциите могат да бъдат изместени една спрямо друга и не се допускат вътрешни празни позиции

Сравнителен анализ на методите и алгоритмите за множествено подравняване на биологични секвенции

- Динамично програмиране;
- Метод на прогресивното подравняване;
- Итеративни методи;
- Консенсусни методи;
- Методи на максималната пестеливост;
- Подравняване по блокове;
- Евристични методи;
- Оптимизационни методи.

Динамичното програмиране

- ***метод за решаване на изключително сложни проблеми посредством разбиването им на по-прости подпроблеми***.
- Алгоритмите на ДП се използват за оптимизация.
- Прилага се за решаването на проблеми, които съдържат припокриващи се подпроблеми и оптимални субструктури.
- Методът е ефективен в случаите, когато броят на повтарящите се подпроблеми расте експоненциално като функция на размера на входните данни.

прогресивно подравняване

- използва евристично търсене и не гарантира намирането на глобалния оптимум.
- се основава на метода на динамичното програмиране като комбинира двойки подравнявания, започващи с най-сходната двойка и напредва към най-различаващите се двойки.
- Всички методи за прогресивно подравняване изискват два етапа:

(1) изгражда се направляващото дърво, представящо връзките (сходството) между секвенциите на основата на йерархичен метод за клъстериране (напр., UPGMA - Unweighted Pair Group Method with Arithmetic Mean), и

(2) конструиране на подравняването чрез последователно добавяне на секвенциите към нарастващото множество, започвайки от сходните секвенции в посока към най-различните.

- производителността им е ниска за случая на съществено различаващи се секвенции

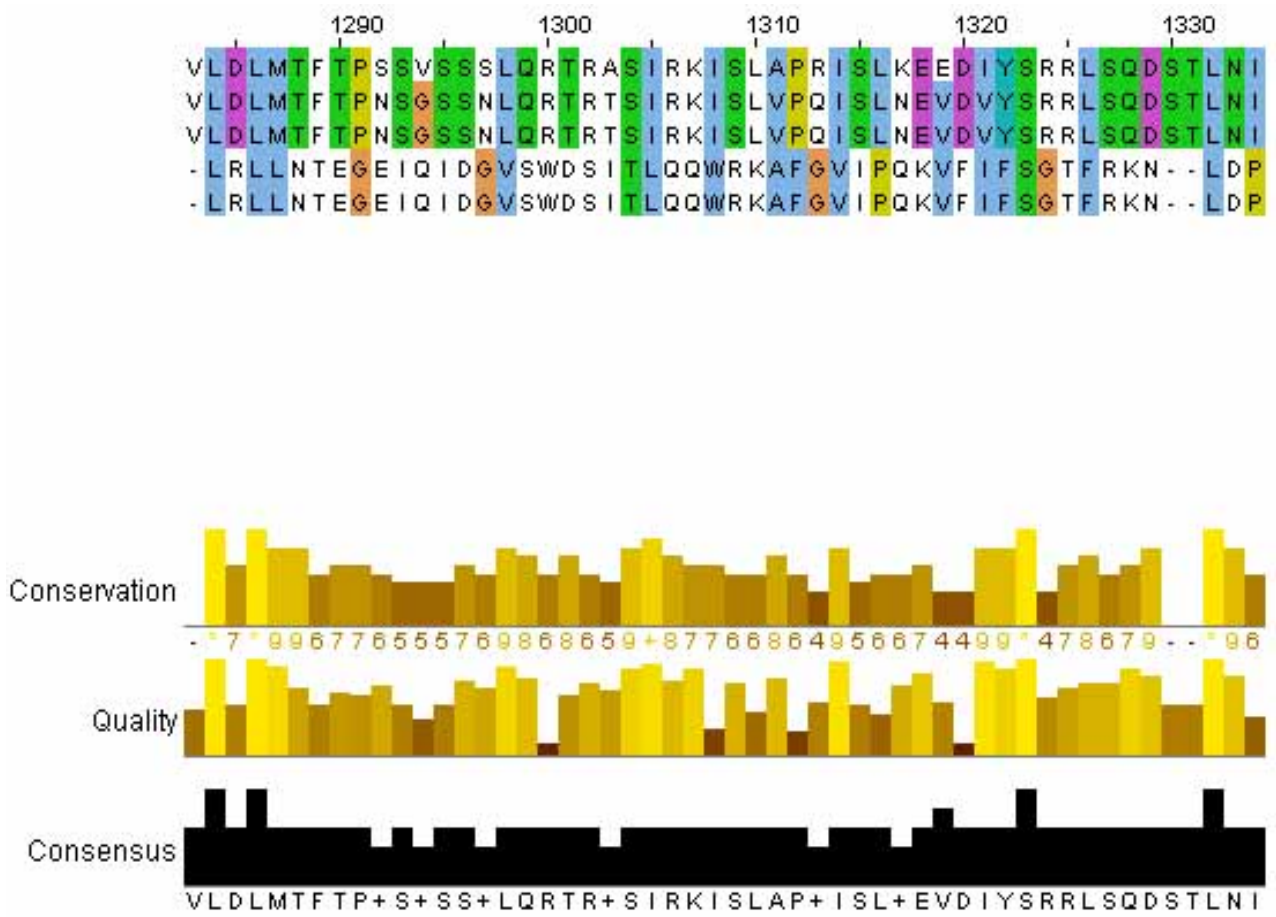
прогресивно подравняване

- Основното забавяне на изпълнението на алгоритъма се дължи на времето за клъстериране на секвенциите в дървовидна или подобна на дендограма структура, т. нар. *направляващо дърво (guide tree)*, което се използва при подравняването на секвенциите и формирането на нарастващо множество от подредени секвенции, следвайки реда на разклоняване в направляващото дърво.
- **Основното преимущество на методите на прогресивното подравняване е, че те са по-бързи и по-ефективни от методите на динамичното програмиране**.
- Сложността на подравняването, след като е конструирано направляващото дърво, е приблизително $\mathcal{O}(N)$ за N секвенции с еднаква дължина. Конструирането на направляващото дърво включва всяка с всяка сравнение на всичките N секвенции, като се генерира матрица на разстоянията, с време $\mathcal{O}(N^2)$. След това се извършва клъстериране със сложност в общия случай $\mathcal{O}(N^2)$ или по-голяма.
- ***При големи стойности на N, конструирането на направляващото дърво е ограничаващ фактор и приложението на тези методи се свежда до обхвата на няколко стотин секвенции.***
- Най-широко използваните имплементации на методите на прогресивното подравняване са **фамилията софтуерни пакети ClustalW**, достъпни на ресурсните web портали за биоинформатика

Софтуерен пакет Clustal Omega

- най-новият софтуер в Clustal family, който замества ClustalW2 при подравняването на стотици секвенции. Clustal Omega прилага „seeded“ направляващи дървета и профилиращи техники базирани на вероятностни статистически модели (скрити модели на Марков - HMM).
- Clustal Omega кластерира секвенциите на основата на малък брой базови (“seed”) секвенции, като при N секвенции броят на seeds е пропорционален на $\log(N)$.
- Кластерирането изисква $\mathcal{O}(NS)$ стъпки, или сложността на алгоритъма е $\mathcal{O}(N\log(N))$.
- В резултат се постига по-високо качество на решенията от софтуерните инструменти на фамилията ClustalW за сметка на повишаване на времето за подравняване.
- Времето за подравняване с верижни направляващи дървета е значително по-голямо и за забавянето значително допринасят и скритите модели на Марков.
- Въпреки този недостатък, времето за подравняване на няколко стотин секвенции се счита за приемливо.

Изходни данни от множествено подравняване на секвенции с ClustalW



T-coffee (Tree-based Consistency Objective Function For alignment Evaluation)

- използва метода за прогресивно подравняване
- Генерира множествено подравняване с най-високо ниво на консистентност като използва библиотека от предварително обработени глобални и локални подравнявания по двойки.
- Може да комбинира предишни подравнения,
- да оценява нивото на консистентност на подравняванията,
- да извлича серия от мотиви за създаването на локално подравняване.

Консенсусните методи

- опитват се да намерят оптималното консенсусно подравняване на множество секвенции, сравнявайки множество различни подравнения на един и същ набор от секвенции, създадени по различни модели.
- Има два често използвани консенсусни метода, **M-COFFEE и MergeAlign**.
- T-Coffee има специален „Мета“ режим на опериране, наречен M-Coffee.
- **M-Coffee** използва множествени подреждания на секвенции, генерирани по 8 различни метода, за да генерира консенсусно подравняване посредство комбиниране на най-добрите части от тях.
- **MergeAlign** генерира консенсусни подреждания от произволен брой подреждания, генерирани с използване на до 91 различни модели на еволюцията или различни методи за множествено подравняване.

Методи на максималната пестеливост

- Методът на максималната пестеливост е базиран на символи, който дава филогенетично дърво с минимум общ брой на еволюционните стъпки, необходими за обяснение на даден набор от данни, присвоени на листата.
- Търси се филогенетично дърво / мрежа, която, когато реконструираме еволюционните събития, водещи до данните за листата, минимизира сумата от теглата по дъгите.
- Критерият за пестеливост в мрежа се дефинира като общата сума от оценките на субституциите на дъгите на дадено дърво (представено от под-граф в мрежата), която минимизира оценката на пестеливост на сайта.
- Оценката на пестеливост за мрежите е NP - труден изчислителен проблем.
- **MEGA (Molecular Evolutionary Genetics Analysis)** е интегриран пакет от софтуерни инструменти за статистически анализи на данни за ДНК и протеинови секвенции от еволюционна гледна точка
- MetaPIGA е робустна имплементация на няколко стохастични евристики за широкомащабни филогенетични анализи (maximum likelihood)

Евристични методи

- ***Методи за търсене по думи (k-tuple methods).***
- Това са евристични методи, които не гарантират оптимално подравняване, но са значително по-ефективни от алгоритъма на SmithWaterman.
- Особено са полезни при търсене в големи бази биологични данни.
- Методите за търсене по думи са най-известни с тяхното прилагане в популярните софтуерни инструменти за търсене в бази биологични данни **FASTA** и фамилия **BLAST**.
- Алгоритъмът FASTA най-често се използва за търсене в бази данни на биологични секвенции, като осигурява метод, алтернативен на този на динамичното програмиране (ДП) за подравняване на биологични секвенции.
- PLFASTA в състава на пакета FASTA създава диаграма на участъците с най-висок процент на съвпадение, аналогично на метода на точковата матрица, и по този начин осигурява възможност за генериране на алтернативни подреждания.

Евристични методи

- *FASTA suite* се поддържа от Genestream на http://vega.igh.cnrs.fr/
- Програмите включват *ALIGN (global, Needleman-Wunsch alignment), LALIGN (local, Smith-Waterman alignment), LALIGNO (Smith-Waterman alignment, no end gap penalty), FASTA (local alignment, FASTA method), and PRSS* (локално подравняване с разбъркани копия (scrambled copies) на втората секвенция за статистически анализ).
- Практически, най-популярен е алгоритъмът BLAST, който се базира на евристични техники, и осигурява откриването на добро (субоптимално) подравняване за приемливо време.
- BLAST е акроним на "Basic Local Alignment Search Tool".
- Програмите за търсене в биологични бази данни в рамките на фамилията софтуерни пакети BLAST приемат и изпълняват заявки за търсене на подобия на нуклеотидни или протеинови секвенции, които биха могли да покажат хомология.
- Тези програми имплементират варианти на алгоритъма BLAST, който се основава на евристичен метод за бързо намиране на локални подравняване с оценки, достатъчно високи, за да бъдат статистически значими.