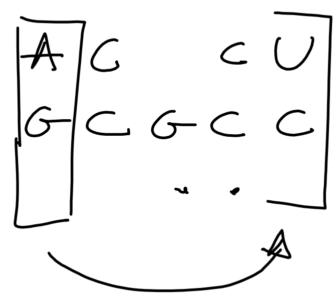
\NA equence Structure alignment

⊖ Notivation

- single ructure rediction is
  imited ~70% base pairs correctly
  icted)
  - ructures predictions are
  ained through parative
  rodeling.
  1. obtain a uence ignment of
     rologous NA
     find inimal set of ovari -ing
     pairs

tation n
.1 → mutation
in col. 2



Ex: 6 rRNA atabase from Gutell

Very accurate ≥ 9 % curate

onclusion: hould e
l. nformation
uest ion . ow? e dont have an
ignment)

Strategies

homologous NA s/s

Plan A
seq. alignment
- coffee etc.
△
aligned sequences
old alignment
fold RNAalifold △

Plan B
multiple
old alignment
Sankoff, 1983)
⋮
▽
aligned & structures

Plan C
fold sequences
▷ NA fold
homologous
structures
structure
alignment
NA forester
△

Pro ns:
n A may fail if uences are almost)
imilarity (0%)

lan C structures ay not e available.
ctions are not accurate

:ion. Don't do things quence by
multaneous folding & alignment ▷
Sankoff ▷ 83)

uestion. how to do SFA?

O equence ignment

ynomial -ime & ace ith eddleman-Wansch

$$D(i,j) = \max \begin{cases})(i,j-1) + \quad (-, w_j') \\ )(i-1,j) + \delta(w_i -) \\ )(i-1,j-1) + \delta(w_i, w_j') \end{cases}$$

$$(i,0) = D\ 0,i) \quad \geq \delta(w_i,-) \quad \sum_i \delta(-, w_i)$$

$$,-) = \delta\ -,x) = - \text{ op penalty}$$

$$(x,y) = \begin{cases} 1 \text{ if } == y \\ 1 \text{ otherwise} \end{cases}$$

- sequences O n^2 time pace
(can be optimized)

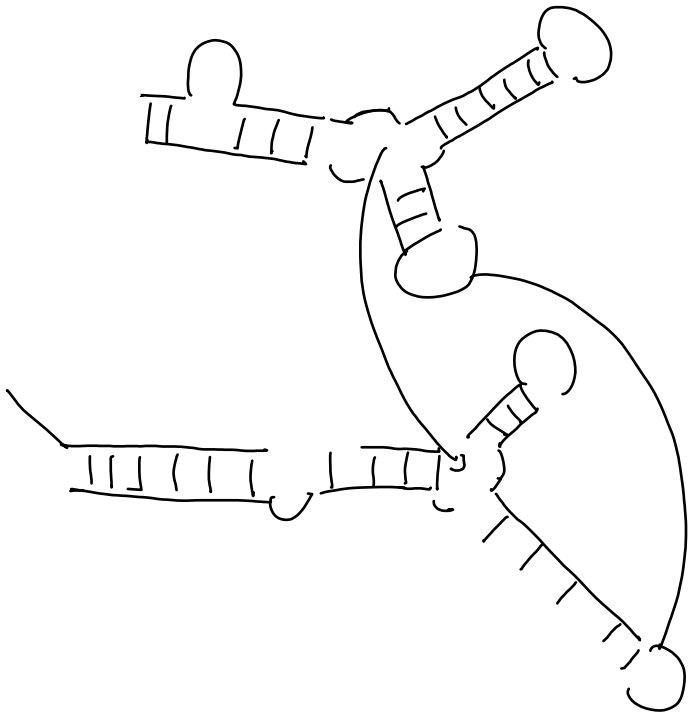
O olding

(n^3) time $O(n^2)$ space NJ, Zuker, etc.)

▷ e hould le ro do both jeratin in lyn ial -ime & ace.

○ Sankoff -lgorithm

B.1 concept: mpatible structures



milar branching
ructures.

ed to constraint
ignments

1 se $(i,j) \in S$ s.t. $i < \pi < j$
$\nexists (i',j')$ s.t. $i < i' < \pi < j' < j$
→ is accessible from (i,j)

an structure alignment is .t.

- ○ cessible tion of extern & multi loop are
  iped
- ○ ing pairs of ipins are aligned
- ○ ops index & ali ed, inserted
  leted

① $i_1 < \dots < i_k$ positions of an internal ~~loop~~ pair
accessible pair positions of a multi-loop in w

$j_1 < \dots < j_l$ in w'

→ the algo produces an alignment such that

- $k = l$
- $(i_f, i_g) \in -$ iff $(j_f, j_g) \in B$

are branching figuration.

recursive equations

$$D(i,j,i',j') = \max \begin{cases} D(i,j-1,i',j-1) \\ D(i,j-1,i',j') \\ D(i-1,j,i',j-1) \\ D(i-1,j,i',j') \\ D(i-1,j-1,i',j-1) + \delta(w_i, w'_{j'}) \end{cases}$$

$D(i,j,i',j') = (w_i, w'_{j'})$ et $\delta(x,y) = \begin{cases} \text{if } x == y \\ \infty \text{ otherwise} \end{cases}$

→ pairwise alignment of loop regions.

& Sankoff Algorithm

$\mathcal{D}\left[\binom{i_1}{i_2},\binom{j_1}{j_2}\right]$ ← best alignment score (sequence only)
↑ $-_{i_1 j_1}$ & $B_{i_2 j_2}$

$\mathcal{V}\left[\binom{i_1}{i_2},\binom{j_1}{j_2}\right]$ ← min - alignment (seq + struct.)
- $-_{i_1 j_1}$ & $B_{i_2 j_2}$

$\mathcal{I}$ ← same with multi-loop
nality for internal bases

$\left[\binom{i_1}{i_2},\binom{j_1}{j_2}\right]$ — min cost alignment seq & struct)
of $A_{i_1 j_1}$ & $B_{i_2 j_2}$ s.t. (i_1, j_1) & (i_2, j_2)
are base pairs

Equations:

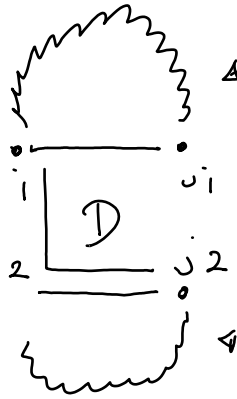
D can be filled using Needleman-Wunsch

$\mathcal{V}\left[\binom{i_1}{i_2},\binom{j_1}{j_2}\right]$ // base pair at extremities

$\cdot\ \mathcal{W}\left[\binom{i_1}{i_2}\binom{j_1}{j_2}\right] = \min \begin{cases} \min\limits_{i_1,k_2} \mathcal{V}\left[\binom{i_1}{i_2}\binom{k_1}{k_2}\right] + \mathcal{WL}\left[\binom{k_1+1}{k_2+1}\binom{j_1}{j_2}\right] & \text{// all decomp.} \\ \mathcal{D}\left[\binom{i_1}{i_2}\binom{j_1}{j_2}\right] & \text{// no structure} \end{cases}$

$$V\left[\begin{pmatrix} i_1 \\ i_2 \end{pmatrix}, \begin{pmatrix} j_1 \\ j_2 \end{pmatrix}\right] = \min \begin{cases} D\left[\begin{pmatrix} i_1 \\ i_2 \end{pmatrix}, \begin{pmatrix} j_1 \\ j_2 \end{pmatrix}\right] + E_H^A(i_1, j_1) + E_H^B(i_2, j_2) & \text{①} \\ \min_{k_1 k_2 l_1 l_2} = E_L^A(i_1, k_1, l_1, j_1) + E_L^B(i_2, k_2, l_2, j_2) & \text{②} \\ \quad + D\left[\begin{pmatrix} i_1 \\ i_2 \end{pmatrix}\begin{pmatrix} k_1 - 1 \\ k_2 - 1 \end{pmatrix}\right] + D\left[\begin{pmatrix} l_1 + 1 \\ l_2 + 1 \end{pmatrix}\begin{pmatrix} j_1 \\ j_2 \end{pmatrix}\right] \\ \quad + V\left[\begin{pmatrix} k_1 \\ k_2 \end{pmatrix}\begin{pmatrix} l_1 \\ l_2 \end{pmatrix}\right] \\ = E^{\text{close}}(i_1, j_1) + E^{\text{close}}(i_2, j_2) & \text{③} \\ \min_{k_1, k_2} V\left[\begin{pmatrix} i_1 \\ i_2 + 1 \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \end{pmatrix}\right] + W\left[\begin{pmatrix} k_1 + 1 \\ k_2 + 1 \end{pmatrix}\begin{pmatrix} j_1 \\ j_2 - 1 \end{pmatrix}\right] & \text{④} \end{cases}$$

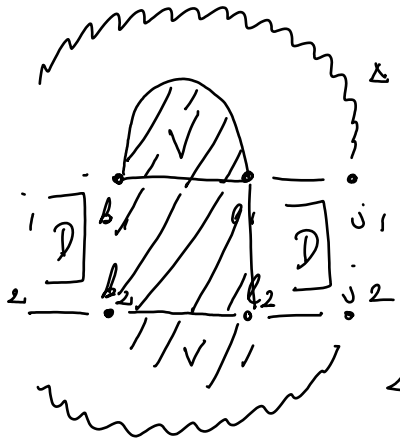
① Align ends



$= E_H^A(i_1, j_1)$ energy of ending i_1, j_1

$= E_H^B(i_2, j_2)$

② Align loops of index 2

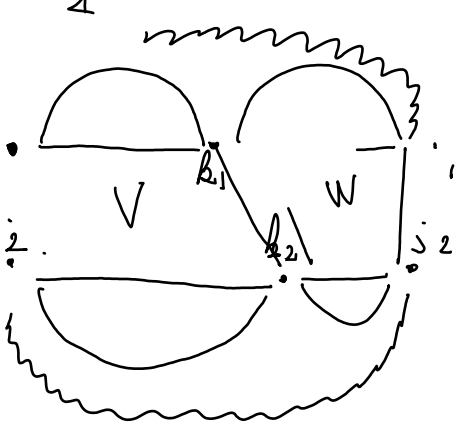


$= E_L^A(i_1, k_1, l_1, j_1)$ copy of the top.

$= E_L^B(i_2, k_2, l_2, j_2)$

③ Align branchings

Rq: we need to recalculate $2 \times W$ - or ensure ℓ rems



· olution of the problem is found in
$N[\binom{1}{1}\binom{|A|}{|B|}]$ & backtrack

mplexity $O(|A|^3|B|^3)$ -ime
$(|A|\,|B|^2)$ ace

▷ too high for mediate application
especially n 1985!)

3.4. scussion

hy does it work? FE of ifle seq is not
ccurate, but if a structure has a low
ergy & homologous sequences imultaneously
-hen his is m - likely a ctly prediction

○ ankoff-style algo in practice

first -ical implementations peared n
-he 2000's

Dynalign in RNA structure) Mathews Turner 2 02)

me complexity but ficient implementation
if restricted. (See next page.)

Dynalign complexity $O((\min(|A|,|B|))^3 K^3)$

where K is the maximum distance between aligned nucleotides.
⇒ allows to restrict search depth

plication: ection of ncRNA
Uzilov - al., 200)

rotocal:
.u - d - abase of ncRNA
· align query RNA with ps in DB
· generate nrol ps for this in of RNA & calculate Z-score
· accept as cRNA if Z-score is significant

M np Mofacher et al., 20 4)

Dynalign or another implementation - Sankoff lgo (foldalign, orodkin et al. 1997)
o not implement the full model

idea: se ase pair proba calculated y .c Caskill, 1990) -o ped up Sankoff.

ou? · n in the algo only base pair with proba higher than . threshold.

efits:

- deals with a number of potential
  ase pair linear with the length of RNAs
  deals with base pairs (i.e. not stacks) → simpler

methodology can be used to align stochastic
contact matrices. re informative)
(can be tuned with convex opt.)

plication. LocaRNA (Will et al. 200)
re implementation)

- lustering · ncRNA.
  1 take a t of ncRNA
  2 align all is of ncRNA & compute
  a score for distance)
  3 luster RNA used on this distance matrix
  here weighted pair method algo)
  istances d(i,j) = 0, -score)
  here q is the 9%-quantile , all
  inwise scores.

is method ied on Rfam ds)
- discover some classes , ncRNA
- found classes , ncRNA in
  ona intestinalis genome