DNA assembly algorithms

from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874646/
"Assembly Algorithms for Next-Generation Sequencing Data", 2011,
Miller, Koren & Sutton

An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds, sometimes called supercontigs or metacontigs, define the contig order and orientation and the sizes of the gaps between contigs. Scaffold topology may be a simple path or a network. Most assemblers output, in addition, a set of unassembled or partially assembled reads.

DNA sequencing technologies share the fundamental limitation that read lengths are much shorter than even the smallest genomes. Whole Genome Shotgun (WGS) overcomes this limitation by over-sampling the target genome with short reads from random positions. Assembly software reconstructs the target sequence. *Assembly software is challenged by repeat sequences in the target….non-uniform coverage of the target (coverage variation is introduced by chance, by variation in cellular copy number between source DNA molecules, and by compositional bias of sequencing technologies).*

Algorithms (it's NP-Hard, so these are approximations):
*The Overlap/Layout/Consensus (OLC) methods rely on an overlap graph.
*The de Bruijn Graph (DBG) methods use some form of K-mer graph.
*The greedy graph algorithms may use OLC or DBG

OLC and DBG are two robust approaches to assembly.

An overlap graph represents the sequencing reads as nodes and the edges are their overlaps (pre-computed by a series of computationally expensive pair-wise sequence alignments)… Paths through the graph are the potential contigs, and paths can be converted to sequence…There are two ways to force paths to obey the semantics of double-stranded DNA. If the graph has separate nodes for read ends, then paths must exit the opposite end of the read they enter. If the graph has separate edges for the forward and reverse strands, then paths must exit a node on the same strand they enter.

The de Bruijn graph represents strings from a finite alphabet. The nodes represent all possible fixed-length strings. The edges represent suffix-to-prefix perfect overlaps.
A K-mer graph is a form of de Bruijn graph. Its nodes represent all the fixed-length subsequences drawn from a larger sequence. Its edges represent all the fixed-length overlaps between subsequences that were consecutive in the larger sequence…By construction, the graph contains a path corresponding to the original sequence (Figure 1). The path converges on itself at graph elements representing K-mers in the sequence whose multiplicity is greater than one.

The greedy algorithms apply one basic operation, repeated: given any read or contig, add one more read or contig. Each operation uses the next highest-scoring overlap to make the next join. The scoring function measures, for instance, the number of matching bases in the overlap…The greedy algorithms can get stuck at local maxima if the contig at hand takes on reads that would have helped other contigs grow even larger.

https://www.cs.princeton.edu/~wayne/kleinberg-tardos/pdf/06DynamicProgrammingI.pdf

RNA secondary structure

Secondary structure. A set of pairs $S = \{ (b_i, b_j) \}$ that satisfy:

- [Watson–Crick] S is a matching and each pair in S is a Watson–Crick complement: A–U, U–A, C–G, or G–C.
- [No sharp turns] The ends of each pair are separated by at least 4 intervening bases. If $(b_i, b_j) \in S$, then $i < j - 4$.
- [Non-crossing] If (b_i, b_j) and (b_k, b_ℓ) are two pairs in S, then we cannot have $i < k < j < \ell$.

Free-energy hypothesis. RNA molecule will form the secondary structure with the minimum total free energy.

approximate by number of base pairs
(more base pairs ⇒ lower free energy)

Goal. Given an RNA molecule $B = b_1 b_2 \ldots b_n$, find a secondary structure S that maximizes the number of base pairs.

On Finding All Suboptimal Foldings of an RNA Molecule

MICHAEL ZUKER

doi:10.1126/science.2468181
Science, 244, 4900},48–52, 1989

9. R. Nussinov, G. Pieczenik, J. R. Griggs, D. J. Kleitman, *SIAM (Soci. Ind. Appl. Math.) J. Appl. Math.* **35**, 68 (1978).
10. R. Nussinov and A. B. Jacobson, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6309 (1980).
11. M. Zuker and P. Stiegler, *Nucleic Acids Res.* **9**, 133 (1981).

These programs work in two stages. The first part, called the fill algorithm, computes and stores minimum folding energies for all fragments of the sequence. The process begins with all ***pentanucleotides*** and builds up to larger fragments in a recursive fashion. The second algorithm, called the traceback, computes a minimum energy strumre by searching systematically through the matrix of stored energies.

https://www.cs.princeton.edu/~wayne/kleinberg-tardos/pdf/06DynamicProgrammingI.pdf

Theorem. The DP algorithm solves the RNA secondary structure problem in $O(n^3)$ time and $O(n^2)$ space.

Dynamic programming over intervals

Def. $OPT(i, j)$ = maximum number of base pairs in a secondary structure of the substring $b_i b_{i+1} \ldots b_j$.

Case 1. If $i \geq j - 4$.

- $OPT(i, j) = 0$ by no-sharp-turns condition.

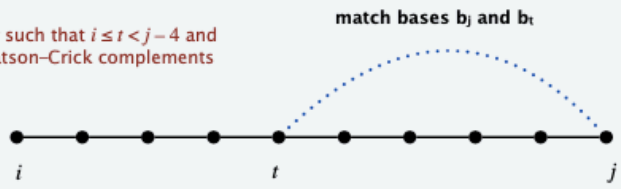
Case 2. Base b_j is not involved in a pair.

- $OPT(i, j) = \text{OPT}(i, j-1)$.

Case 3. Base b_j pairs with b_t for some $i \leq t < j - 4$.

- Non-crossing condition decouples resulting two subproblems.
- $OPT(i, j) = 1 + \max_t \{ OPT(i, t-1) + OPT(t+1, j-1) \}$.

take max over t such that $i \leq t < j - 4$ and b_t and b_j are Watson–Crick complements



52

RNA-SECONDARY-STRUCTURE$(n, b_1, \ldots, b_n)$

FOR $k = 5$ TO $n - 1$

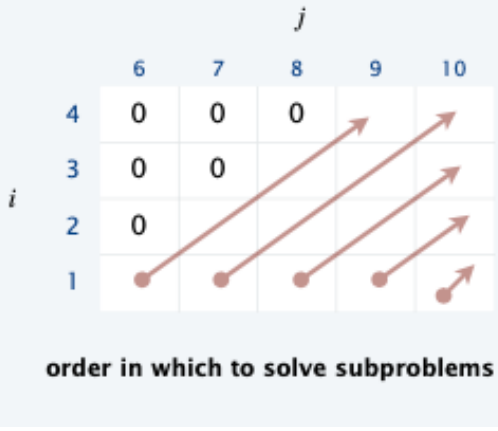
  FOR $i = 1$ TO $n - k$

    $j \leftarrow i + k$.

    Compute $M[i, j]$ using formula.

RETURN $M[1, n]$.

all needed values are already computed



order in which to solve subproblems

RNA Secondary Structure: A Complete Mathematical Analysis

M. S.WATERMAN AND T. F.SMITH
Los Alamos Scientific Laboratory of the Uniwrsiw of Califonia, Los Alamos, New Mexico 87545

Received 9 August 1978; revised 25 August 1978
http://bioinfo.ict.ac.cn/~dbu/AlgorithmCourses/Lectures/Lec6-Advanced-DP-RNA-Waterman1978.pdf

energy to the configurations examined. There has recently been considerable work on the tertiary structure of some nucleic acids, in particular in comparisons with the x-ray data on various tRNAs. However, it should be noted that constraints arising from the most probable secondary structure base pairing are normally imposed on the tertiary structure considerations. This is analogous to the methods of predicting protein tertiary structure by starting with the statistics of forming helical and nonhelical regions.

In the present study the first problem is solved. This is accomplished through an iterative definition of all secondary structures and the extension of the sequence metric algorithms of Sellers [14]. The initial steps are based on the work of Needleman and Wunsch [15] and Tinoco et al. [2]. These ideas lead to the calculation of a minimum "distance" between segments of a RNA sequence, where "distance" is measured in free energy. The most probable secondary structure is then assumed to be the configuration having the minimum sum of all such aligned "distances."

Modifying the approach of Tinoco et al. [2], define the base pairing matrix $P=(p_{ij})$, for a given RNA sequence $\mathbf{s}=s_1s_2\ldots s_n$ (and the reversed order sequence $\mathbf{s}'=s_ns_{n-1}\ldots s_1$) by $p_{ij}=1$ if s_i and s_j can form a bond and $p_{ij}=0$ otherwise. (the bonds are A—U, G—C, and sometimes G—U.)

A secondary structure for $\mathbf{s}$ is a configuration of the sequence $s_1s_2\ldots s_n$ with two properties: (i) Each point can be bonded to at most one other point. (ii) If s_i and s_j are bonded, then any bonding of s_k $(i<k<j)$ must be with points between i and j. It has been shown [6] that this definition includes all possible substructures (such as hairpins, helicies, bulges, tails, and interior loops). This definition does not include the B_{III} structure[1]

The total number of structures having $i+1$ bonded pairs for a sequence $n+1$ long is given by a recursion relation. Let $N_{l,n}^{i}$ be the number of secondary structures containing exactly i bonded pairs formed on the subsequence $s_ls_{l+1}\ldots s_n$. Then

$$N_{l,n+1}^{i+1}=N_{l,n}^{i+1}+\sum_{j=l}^{n-m}\sum_{k=0}^{i}N_{l,j-1}^{k}N_{j+1,n}^{i-k}p_{j,n+1}, \tag{1}$$

where all hairpin loops have at least m bases. The equation follows from the fact that s_{n+1} is either bonded or not bonded. If s_{n+1} is not bonded, then there are $N_{l,n}^{i+1}$ structures of interest. Otherwise, $n+1$ is bonded to some j, $l<j<n-m$, and if k bonds are formed in $s_l\ldots s_{j-1}$, then $i-k$ must be formed in $s_{j+1}\ldots s_n$. The definition of secondary structure implies that any combination of a k bonded structure with an $i-k$ bonded structure gives a secondary structure. Thus $N_{l,n+1}^{i+1}$ satisfies Eq. (1).

The only sterical constraint in Eq. (1) is that the hairpin loop size must be at least m. It is possible to modify Eq. (1) so that no helices of length one are allowed. This has been done and the recursion applied to real RNA sequences. For many real RNA sequences of length forty there are over 10^6 structures, hundreds of which may have maximal base pairing.

RNA Secondary Structure: A Complete Mathematical Analysis

M. S.WATERMAN AND T. F.SMITH
Los Alamos Scientific Laboratory of the Uniwrsiw of Califonia, Los Alamos, New Mexico 87545

Received 9 August 1978; revised 25 August 1978

The problem is analogous to finding the optimal matching alignment between two evolutionary sequences. The solution to that evolutionary distance problem was proposed by Sellers [14] and generalized by Waterman et al. [19]. To help clarify the relationship between the two problems, it is useful to note that regions of homology between different sequences are analogous to complementary helical regions, nonhomologous regions are analogous to noncomplementary internal loop regions, and deletions/insertions are analogous to bulges. It is also helpful to recall that finding the maximum homology between evolutionary sequences is equivalent to finding the minimum mutational distance between them. As noted above, in this work a minimum "distance" measured in free energy is calculated between all subsequences.

$\alpha_{ij} = \Delta G$ (free energy change) of binding of the ith element of the sequence s with the jth of s′;

$\eta_{ij} = \Delta G$ resulting from nearest neighbor interaction between base pairs $i-1,j-1$ and i,j;

$\beta_j = \Delta G$ of a bulge j bases long;

$\gamma_{ij} = \Delta G$ of an interior loop of lengths i and j;

$\xi_{ij} = \Delta G$ of an end loop $n-i-j$ bases long due to the pairing of bases i and j;

$\tau_i = \Delta G$ of a free end or tail of length i.

The total free energy change of a secondary structure is defined to be the sum of the ΔG's associated with these substructures. This can be accomplished [7] by constructing an f matrix such that each element f_{ij} represents the free energy of formation of the i,j bound pair plus the free energy of that secondary structure having the minimum free energy among all substructures formed from the $i-1$ subsequence of s and the $j-1$ subsequence of s′. The elements of f_{ij} are undefined (plus infinity) for all i,j such that the ith base in s cannot form a Watson-Crick pair with the jth base in s′ ($p_{ij}=0$).

For the case when $p_{ij}=1$, f_{ij} is defined as

$$f_{ij} = \alpha_{ij} + \min\Big\{ f_{i-1,j-1} + \eta_{ij},\ \min_{k>0}\{f_{i-k-1,j-1} + \beta_k\},\ \min_{k>0}\{f_{i-1,j-k-1} + \beta_k\},\ \min_{\substack{k>0\\ l>0}}\{f_{i-k-1,j-l-1} + \gamma_{k,l}\},\ 0\Big\}. \tag{2}$$

The free energy change of the best single loop secondary structure is calculated by

$$F_{1,n} = \min_{\substack{1<j<n\\ 1<j<n}}\{f_{ij} + \xi_{ij}\}, \tag{3}$$

RNA Secondary Structure: A Complete Mathematical Analysis

M. S.WATERMAN AND T. F.SMITH
Los Alamos Scientific Laboratory of the Uniwrsiw of Califonia, Los Alamos, New Mexico 87545

Received 9 August 1978; revised 25 August 1978

which includes the additional free energy associated with the end loops. Figure 1 shows the values of f_{ij} for a simple illustrative example using the component ΔG's given in Table 1, Column A. The insert in Fig. 1 shows the spatial relationship between previous elements of f_{ij} and a given element for a finite value of α. A complete, mathematical proof that this procedure obtains the minimum is given by Waterman [7].

To calculate more complex minimum free energy secondary structures, the single loop F_{ij} must be obtained for all viable subsequences $1 < i < j < n$. Then the bulges, interior loops, and tails must be examined for the possibility that these subsequences may form bonded single loop structures. It is not entirely easy to calculate the proper free energy changes for the addition of these structures. This is because f_{ij} gives tails weight zero, when they could become bulges or joins in the new composite structures. It is useful to calculate F_{ij} only for substructures such that s_i and s'_j are bonded. Waterman [7], using these restricted F_{ij}, was able to iterate and calculate minimum free energy structures of arbitrary complexity.

The secondary structure having the calculated minimum free energy change is obtained from a traceback procedure. Having found the f_{ij} which gives the single loop minimum F_{1n} in Eq. (3), one must trace back to find which terms in Eq. (2) and thus which structural component contributed at each step. There is no guarantee that the minimum free energy structure is unique, but a traceback procedure can locate all such structures.

TABLE 1
Substructural Component Free Energies
in kcal at 23°C

A[a]	B[b]
$\alpha_{ij} = -1.0$[c]	$\alpha_{AU} = -0.25$
$\eta_{ij} = -1.0$[c]	$\alpha_{GC} = -1.4$
	$\alpha_{GU} = +1.9$[d]
$\beta_l = 1.0 + 0.5l$	$\eta_{ij} = -1.0$[c]
$\gamma_{kl} = 1.0 + 0.5(l+k)$	$\beta_l = 1.0 + 0.3l$
$\xi_{kl} = 1.0 + 0.5(n-k-l)$	$\gamma_{lk} = 1.5 + 0.2(l+k)$
$\Delta G_{join} = 0.0$	$\xi_{lk} = 3.05 + 0.1(n-k-l)$
	$\Delta G_{join} = 0.5 + 0.3l$

[a]Values for investigative use only, in the construction of Fig. 1. Such values allow the illustration all the major properties of the proposed algorithm.

[b]Values extrapolated from experimental values. The values for α_{AU}, α_{GC}, and η were chosen to given the standard values of -1.25 and -2.4 kcal in the limit of long bound chains.

[c]For all $i,j = A, C, G, U$.

[d]The value of α_{GU} was set equal to the interior loop value. This results in a ΔG for a G—U pair in the interior of a helical region of only -0.1 kcal. The values for the bulge, interior loop, and end loop contributions are linearizations of those due to DeLisi and Crothers [3], and as such have a limited argument range.

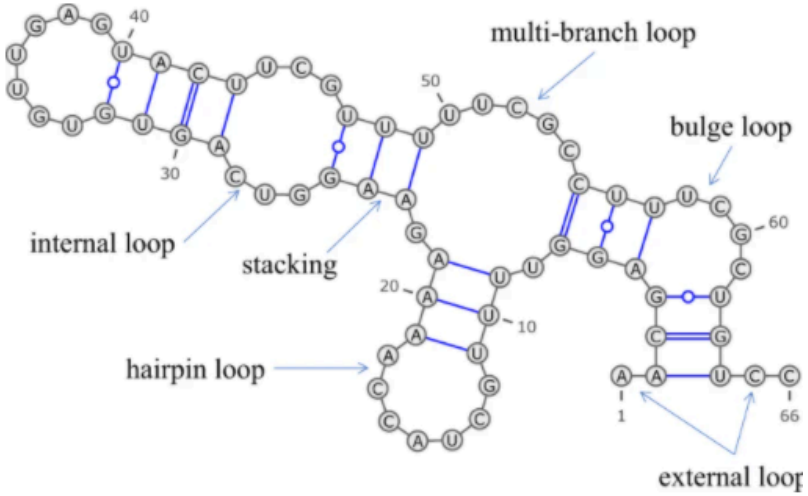
https://en.wikipedia.org/wiki/Nucleic_acid_secondary_structure

RNA secondary structure can be determined from atomic coordinates (tertiary structure) obtained by X-ray crystallography, often deposited in the Protein Data Bank. Current methods include 3DNA/DSSR[16] and MC-annotate.[17]

“RNA secondary structure prediction using deep learning with thermodynamic integration.“
Sato, K., Akiyama, M. & Sakakibara, Y.
Nat Commun 12, 941 (2021).
https://doi.org/10.1038/s41467-021-21194-4

Accurate predictions of RNA secondary structures can help uncover the roles of functional non-coding RNAs. Although machine learning-based models have achieved high performance in terms of prediction accuracy, overfitting is a common risk for such highly parameterized models. Here we show that overfitting can be minimized when RNA folding scores learnt using a deep neural network are integrated together with Turner’s nearest-neighbor free energy parameters. Training the model with thermodynamic regularization ensures that folding scores and the calculated free energy are as close as possible. In computational experiments designed for newly discovered non-coding RNAs, our algorithm (MXfold2) achieves the most robust and accurate predictions of RNA secondary structures without sacrificing computational efficiency compared to several other algorithms. The results suggest that integrating thermodynamic information could help improve the robustness of deep learning-based predictions of RNA secondary structure.
…
there are major difficulties in determining RNA tertiary structures through experimental assays such as nuclear magnetic resonance and X-ray crystal structure analysis, because of the high experimental costs and resolution limits on measurements of RNA. Although considerable advances in cryo-electron microscopy research on RNA tertiary structure determination have been achieved in recent years2, these limitations have not yet been completely overcome. Therefore, instead of conducting such experimental assays, we frequently perform computational prediction of RNA secondary structures, defined as sets of base-pairs with hydrogen bonds between the nucleotides.
…
We can efficiently calculate an optimal secondary structure that has the minimum free energy using a dynamic programming (DP) technique, the well-known Zuker algorithm

Fig. 1: Decomposition of an RNA secondary structure into nearest-neighbor loops.



An RNA secondary structure can be decomposed into several types of nearest-neighbor loops, including hairpin loops (e.g., bases 11–19), internal loops (e.g., bases 25–29 and 43–47), bulge loops (e.g., bases 4–5 and 57–62), base-pair stackings (e.g., bases 23–24 and 48–49), multi-branch loops (e.g., bases 7–9, 21–23, and 49–55), and external loops (e.g., bases 1–2 and 64–66). This diagram was drawn using VARNA[45].

https://rna.urmc.rochester.edu/NNDB/turner04/index.html
NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure.
Turner & Matthews