Prédiction des structures secondaires d'ARNs: AMSAG et les super-optimaux

Waldispühl Jérôme, Behzadi Behshad,
Steyaert Jean-Marc

École Polytechnique
FRANCE
e-mail: {waldispu,behzadi,steyaert}@lix.polytechnique.fr

Plan

Le problème

- Système permettant de modéliser les structures secondaires d'ARN (sans pseudo-noeud),
- méchanisme exprimant les propriétés locales et globales des macromolecules,
- identification des optimaux (recherche d'une énergie minimale),
- algorithmique efficace.

Le problème

$$\Rightarrow \text{Grammaires S-attribuées} = \left\{ \begin{array}{c} \text{grammaire hors-contexte} \equiv \text{structure} \\ + \\ \text{attributs} \equiv \text{critère d'optimalité} \end{array} \right.$$

Grammaires S-attribuées

Grammaires hors-contexte

Definition $G = \{V_T, V_N, P, S\}$

- V_T *est un ensemble de terminaux,*
- V_N *un ensemble de non-terminaux,*
- P *un ensemble de règles de production,*
- S *l'axiome.*

Grammaires S-attribuees

Definition $G = \{V_T, V_N, P, S, \mathcal{A}, \lambda_{\mathcal{A}}, F_P\}$

- $\mathcal{A}$ *un ensemble d'attributs,*
- $\lambda_{\mathcal{A}}$ *un ensemble de fonctions d'attribution des terminaux,*
- F_P *un ensemble de fonctions de calcul des attributs des non-terminaux.*

= Grammaire hors-contexte + Système d'attributs (synthétizés)

Un exemple : les expressions arithmétiques

Grammaire hors-contexte

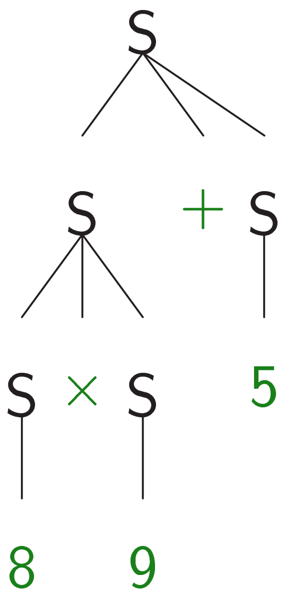
- $V_T = \{0, \cdots, 9, +, \times\}$,
- $V_N = S$,
- $P = \begin{cases} S \to S + S \\ S \to S \times S \\ S \to 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9 \end{cases}$

Système d'attributs

- $\mathcal{A} = \mathbb{N}$
- $\lambda_{\mathcal{A}} = \begin{cases} S_{\mathcal{A}}(0) = 0 \\ \vdots \\ S_{\mathcal{A}}(9) = 9 \\ S_{\mathcal{A}}(+, \times) = 0 \end{cases}$
- $F_P = \left\{ \begin{array}{l} f_{S \to S+S}(xyz) = x + z \\ f_{S \to S \times S}(xyz) = x \times z \\ f_{S \to a \in V_T}(x) = x \end{array} \right\}$

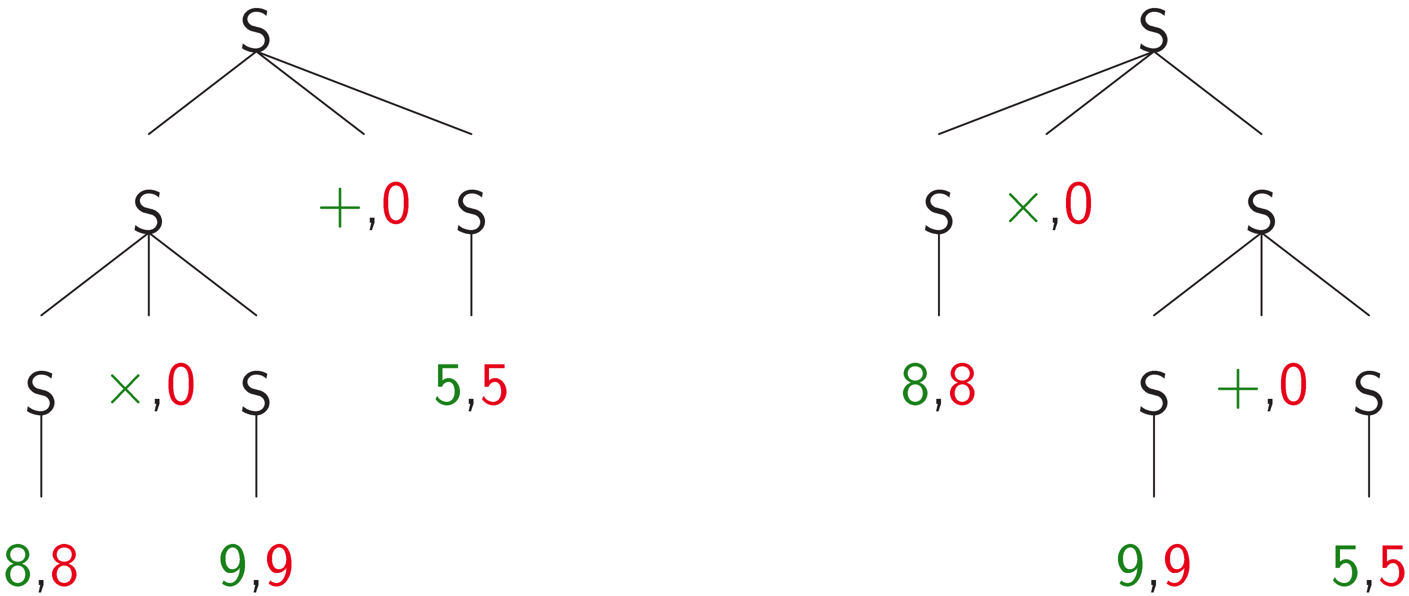
$$8 \times 9 + 5$$

$$8 \times 9 + 5$$

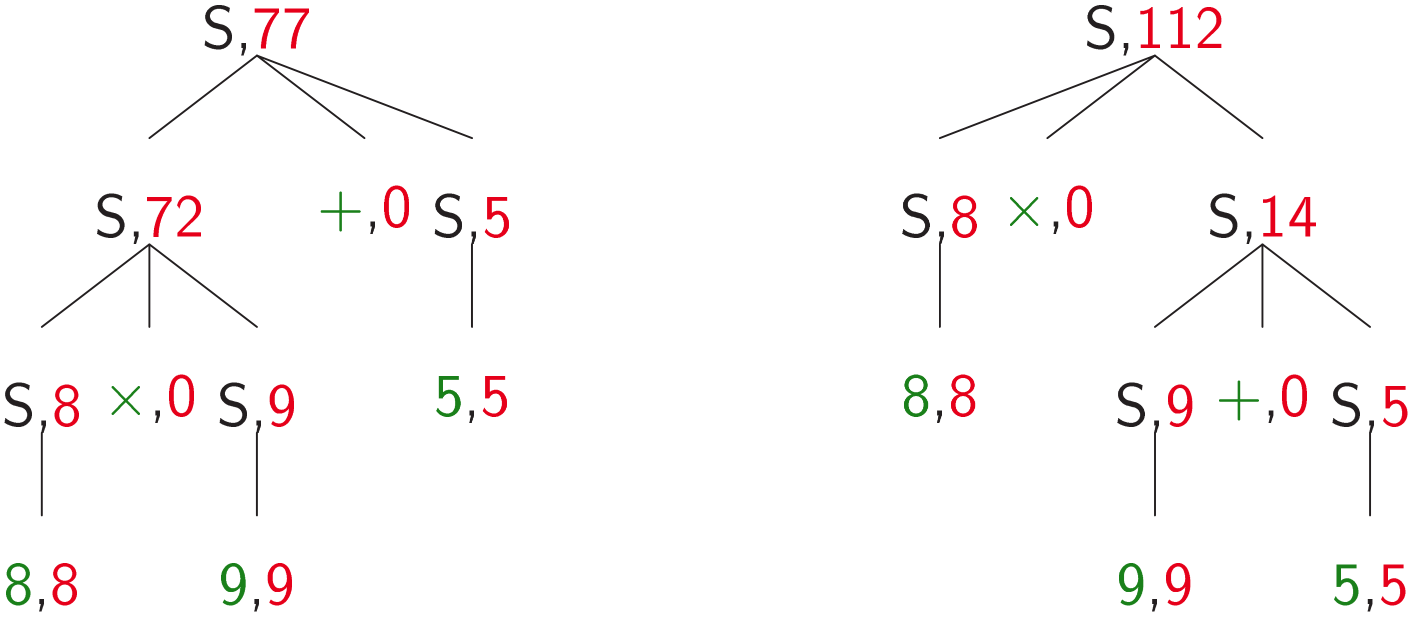


$$8 \times 9 + 5$$

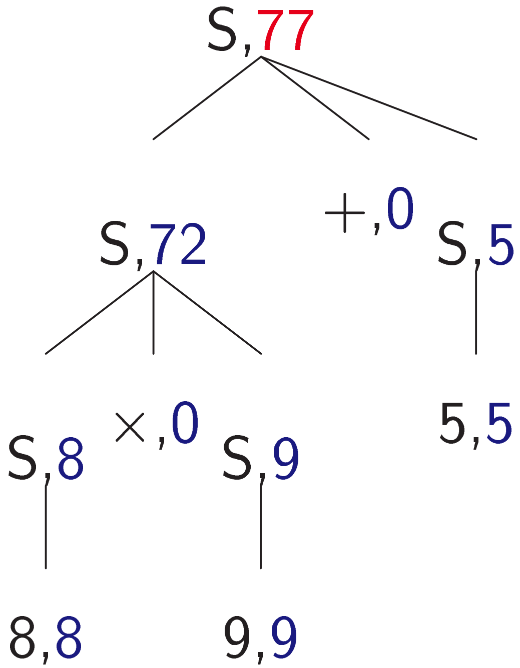
$$8 \times 9 + 5$$



$$8 \times 9 + 5$$



Sélectionner l'arbre optimal



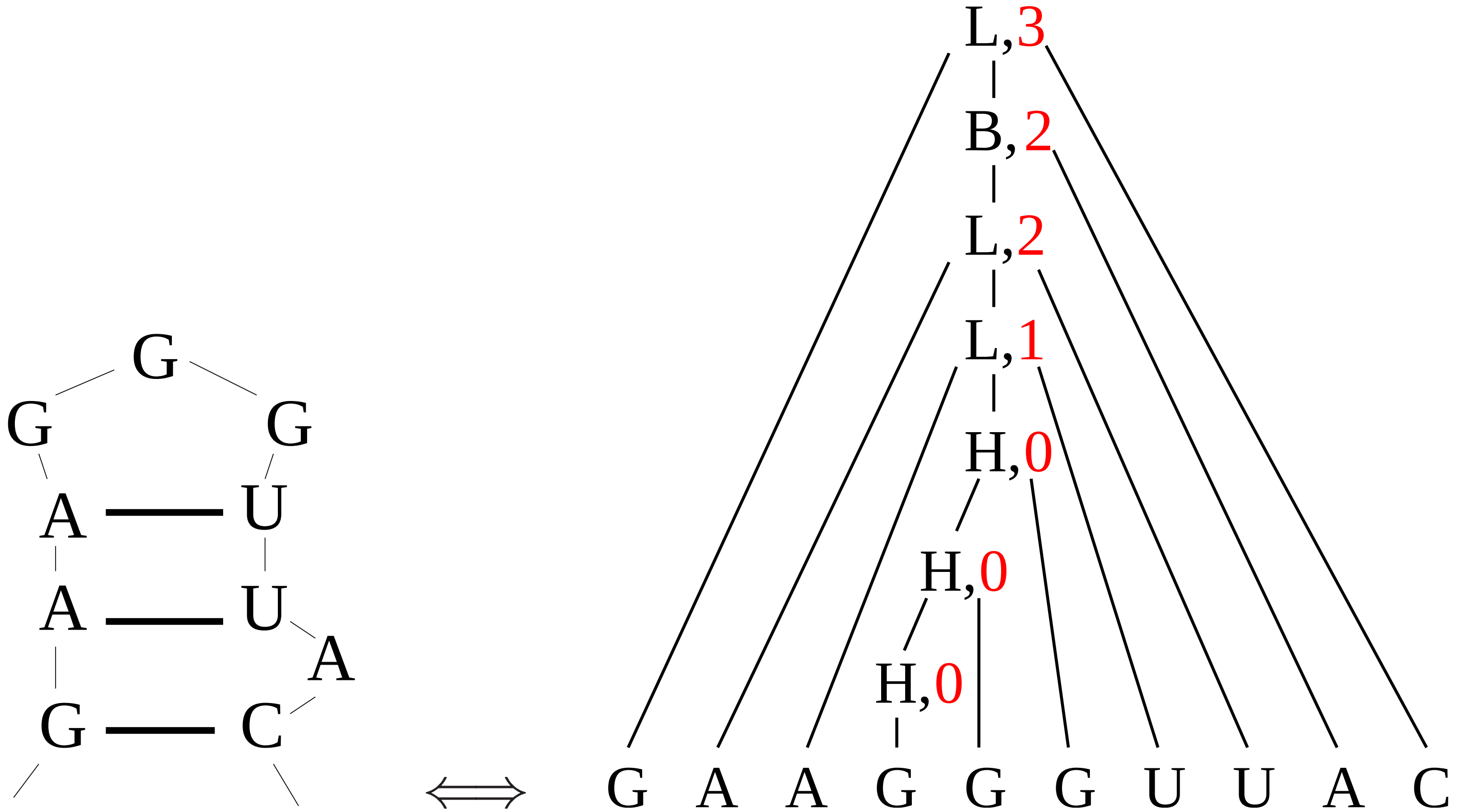
Contrainte de d'optimisation

Definition $\mathcal{C}(x, \lambda_x, y, \lambda_y)$

- $x \ \ et \ y \ \ \in V_T \cup V_N,$
- $\lambda_x \ \ et \ \mathcal{A},$
- *retourne la paire* (z, λ_z) *optimal.*

$\Longrightarrow$ calcul de l'arbre de dérivation optimal.

Structure secondaire des ARNs et Grammaires S-attribuées



Structure secondaire des ARNs et Grammaires S-attribuées

- Structure secondaire $\equiv$ Grammaire hors-contexte:
  un arbre de dérivation
  $\leftrightarrow$ une structure secondaire potentielle,
- énergie de repliement $\equiv$ attributs:
  attribut de la racine de l'arbre de dérivation
  $\leftrightarrow$ énergie de repliement de la structure correspondante,
- structure secondaire d'énergie minimale $\equiv$ arbre de dérivation avec l'attribut optimal à la racine
  $=$ modèle thermodynamique de ZUKER.

État de l'Art

- F. Lefebvre: An optimized parsing algorithm well-suited to RNA folding (1995)
- G. Myers: Approximately matching context-free languages (1995)

Algorithme AMSAG (ECCB 2002)

Generalisation des algorithmes de F. Lefebvre et G. Myers.

Entrée: Une grammaire S-Attributée G, une distance d'édition d, un seuil de recherche $MaxErr$ et une séquence ω.

Sortie: L'arbre de dérivation de $\omega' \in L(G)$ tel que $\lambda_{\omega'}$ soit optimal et $d(\omega, \omega') \leq MaxErr$.

Principe: partitionnment Recursif

Proposition *Soit* $\lambda(A, \omega, e)$ *l'attribut optimal value d'une suite de dérivations* $A \xrightarrow{\star} \omega'$ *pour toute les séquences* ω' *telle que* $d(\omega, \omega') \leq e$. *Alors il éxiste une partition* $\omega_1 \cdots \omega_n = \omega$ *où les sous-chaines* ω_k *sont éventuellement vides, telle que:*

$$\lambda(A, \omega, e) = \max_{A \to \alpha \in P}\{\, f_{A \to \alpha}(\lambda(\alpha_1, \omega_1, e_1) \cdots \lambda(\alpha_n, \omega_n, e_n)) \big| \; e_1 + \cdots + e_n \leq e\}$$

Implémentation

Calculer recursivement le meilleur attribut possible ete le stocker dans un tableau $Best$ tel que:

$$Best(A,i,j,e) = \max\{\lambda_{\omega'} \mid \text{pour toute suite de dérivations } A \xrightarrow{\star} \omega' \text{ tel que } d(\omega', \omega_{ij}) \leq e \text{ and } \omega_{ij} \text{ sous-mot de } \omega\ \}$$

(programmation dynamique)

Algorithm *AMSAG(G, ω, MaxErr)*

Var $Best$: **Array** $[V, 1..N+1, 0..N, 0..MaxErr]$ **of real**

For $nb\_err \leftarrow 0$ **to** $MaxErr$ **do**

  For $len \leftarrow 0$ **to** N **do**

    For $i \leftarrow 1$ **to** $N - len + 1$ **do**

      $j \leftarrow i + len - 1$

      For $A \in V$ **do**

        $Best[A, i, j, nb\_err] = -\infty$

        For $\Gamma \in P$ *and* $LeftSide(\Gamma) = A$ **do**

          $Cur\_i \leftarrow i, Cur\_j \leftarrow j, Cur\_err \leftarrow nb\_err, Cur\_V \leftarrow A$

          FindAlignment$(\Gamma, RightSide(\Gamma), \epsilon, i, j, nb_err)$

Complexité

- $\mathcal{O}(\|P\|.(MaxErr)^r.n^{r+1})$ en temps,
- $\mathcal{O}((\|V_T\| + \|V_N\|.MaxErr).n^2)$ en espace.

Où r est la taille maximale d'une partie gauche d'une règle de production dans P et n la taille de la chaine d'entrée.

Remark *En practique nous observons un comportement en* $\mathcal{O}(n^r)$ *au lieu de* $\mathcal{O}(n^{r+1})$.

Améliorations

- Séquence voisine sous-optimale,
- Approche multi-critère.

Séquence voisine sous-optimale (extension)

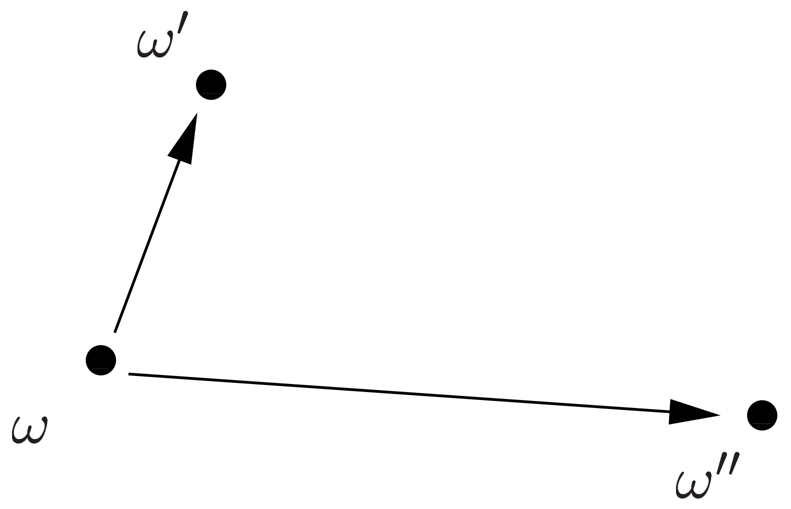
Stocker la liste des k-optimaux attributs dans $Best[V, i, j, NbErr]$.

Complexité: $\mathcal{O}\left(k^r.log(k).\|P\|.(MaxErr)^r.n^{r+1}\right)$ en temps, $\mathcal{O}\left(k.(\|V_T\| + \|V_N\|.MaxErr).n^2\right)$ en espace.

Approche Multi-critère

Sélectionner une séquence plus proche même si l'attribut n'est pas l'optimal. Fonction de valuation $\mathcal{V} : V_T^\star \times V_T^\star \to \mathbb{R}$.
Sortie: la chaine $\omega' \in L(G)$ optimisant $\mathcal{V}(\omega, \omega')$.

Example



$\mathcal{V}(\omega, \omega') = \frac{\lambda_{\omega'}}{d(\omega,\omega')+1}$, $\lambda_{\omega'} \leq \lambda_{\omega''}$ *mais* $\mathcal{V}(\omega, \omega') \geq \mathcal{V}(\omega, \omega'')$.

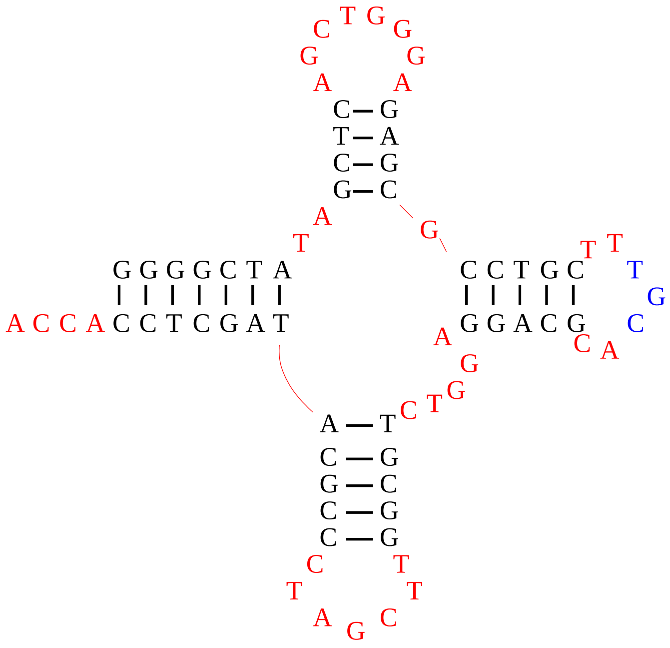
Applications & discussion

Prédiction des structures secondaires d'ARN sans pseudo-noeud (modèle thermodynamique).

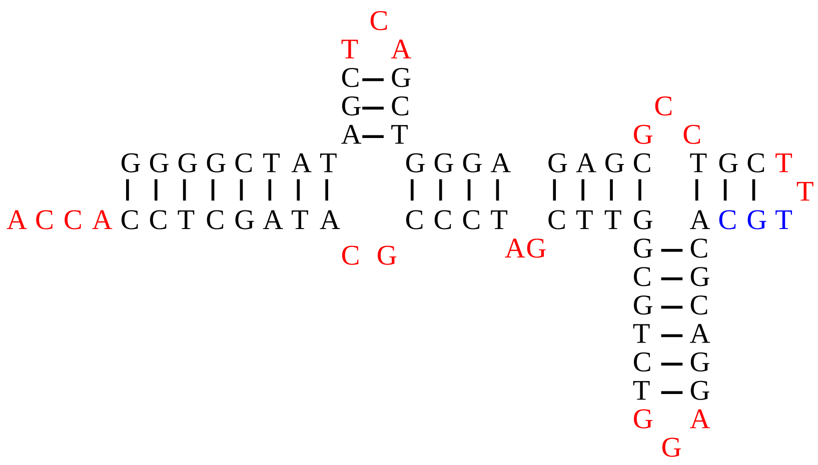
Distance d'édition:

- mutation cost $= 1$,
- deletion cost $= \infty$,
- insertion cost $= \infty$.

tRNA from E. Coli (anti-codon TGC)

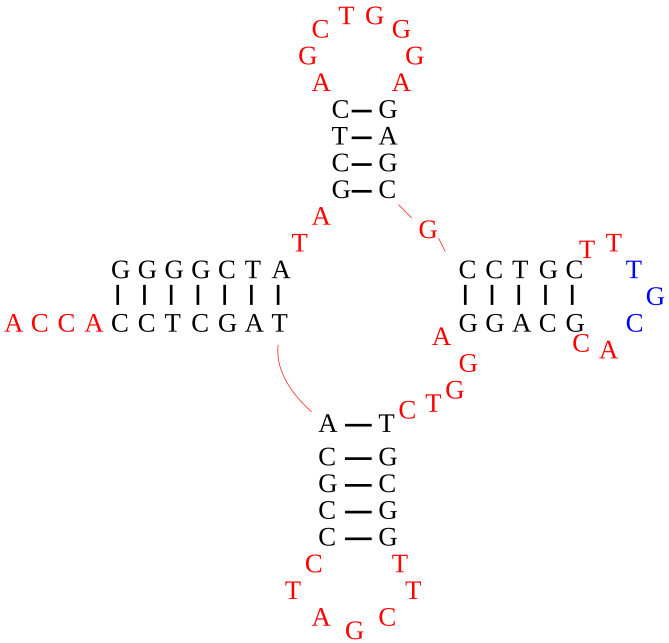


real structure

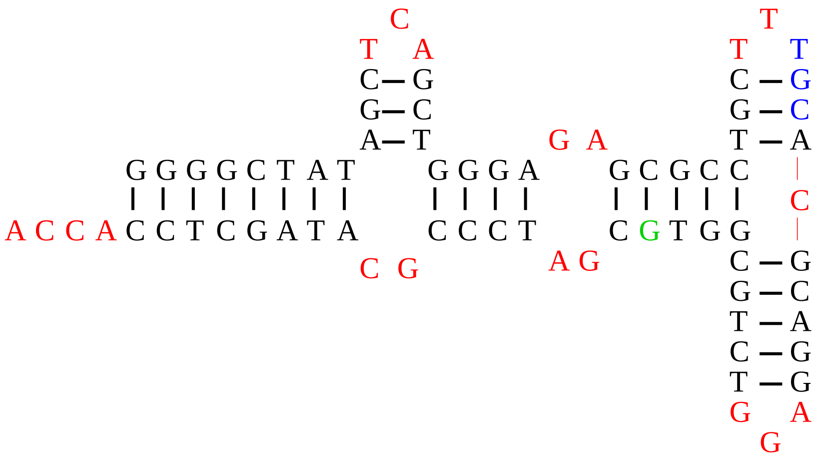


threshold = 0

tRNA from E. Coli (anti-codon TGC)

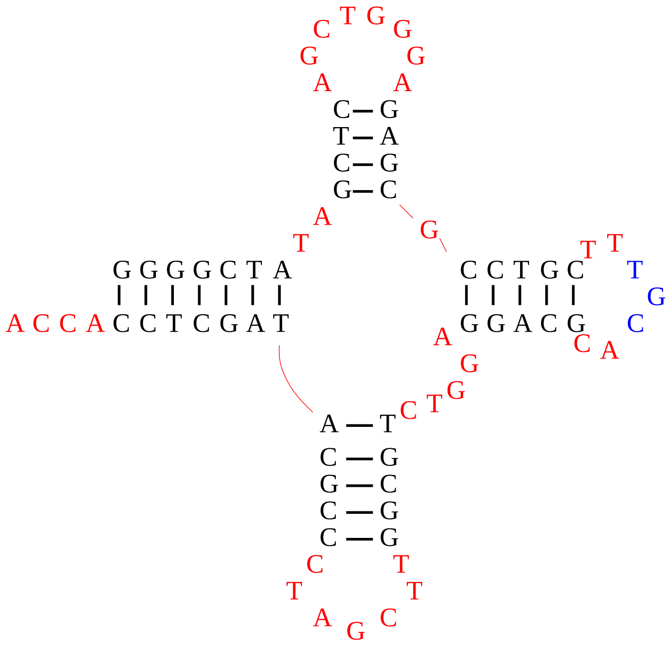


real structure

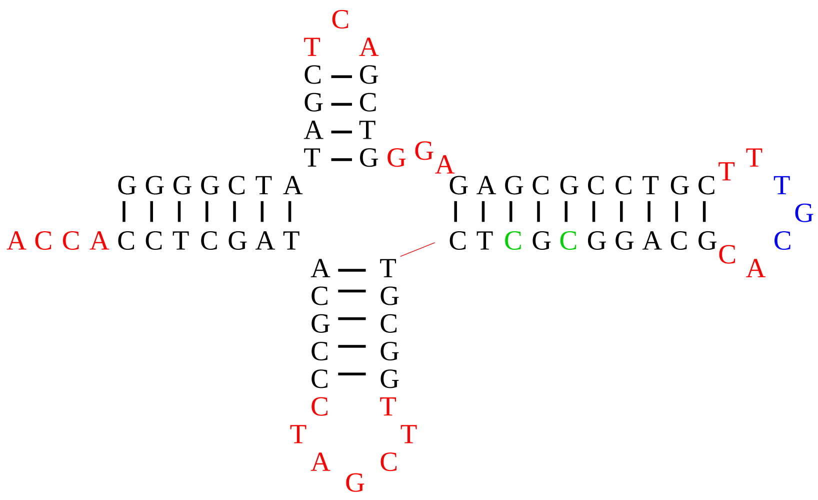


threshold = 1

tRNA from E. Coli (anti-codon TGC)

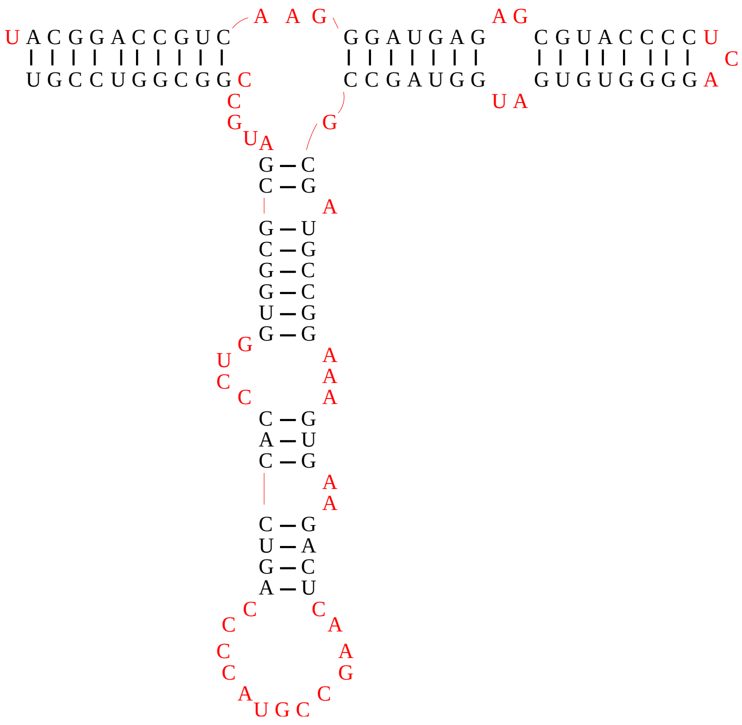


real structure

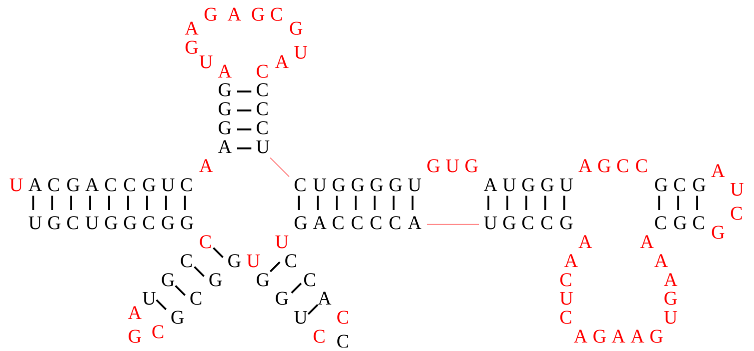


threshold = 2

RNA from E. Coli

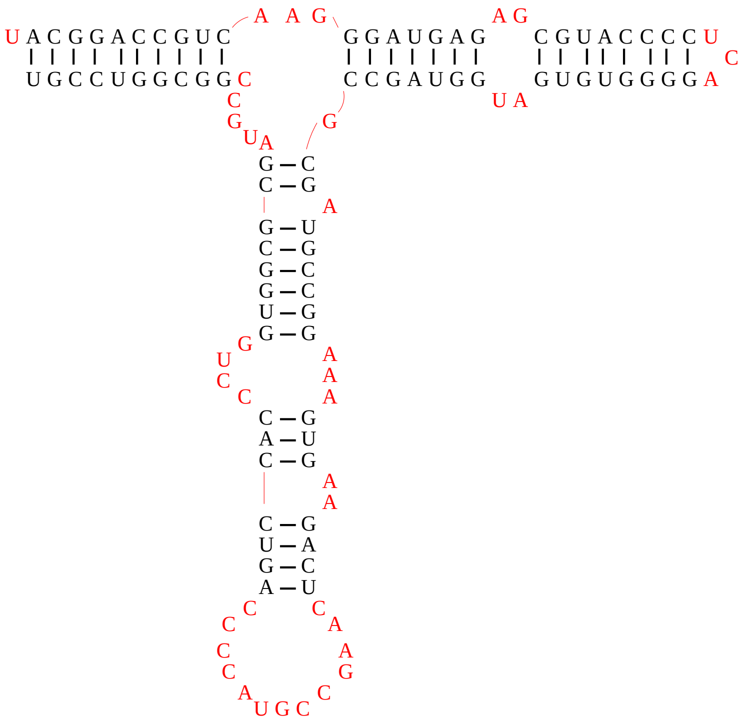


real structure

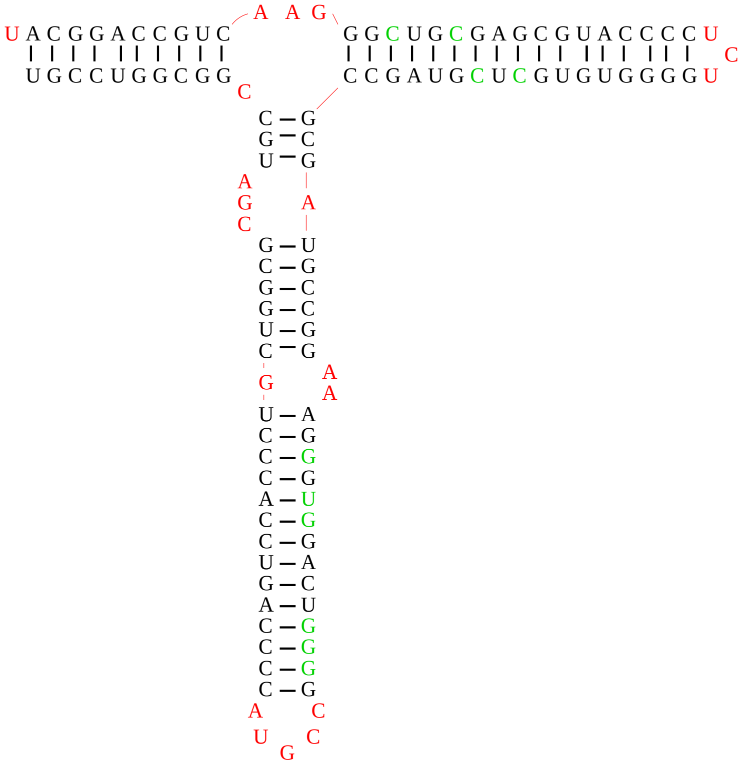


threshold = 0

RNA from E. Coli



real structure



threshold $= 10$