

Repliement avec pseudonoeuds simples

Introduction

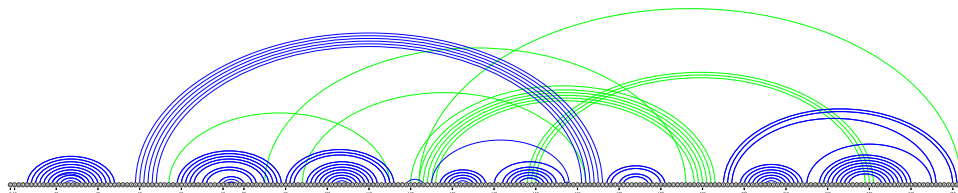


FIGURE 1 – Structure secondaire **avec pseudo-noeuds** d'un ribozyme (Intron du groupe I – Staphylococcus phage twort – PDBID 1Y0Q :A)

Dans de nombreux exemples, l'hypothèse d'une structure secondaire native sans pseudo-noeuds, c'est à dire associée à une représentation linéaire sans croisement, est mise en défaut (Voir Figure 1 ou l'ensemble des structures de la PseudoBase [9, 8]). Cette hypothèse, sous-jacente aux algorithmes de Nussinov et MFold, est cependant essentielle d'un point de vue algorithmique, car l'extension du problème du repliement à des pseudo-noeuds arbitraires rend ce problème NP-complet, même dans des modèles d'énergie simples [4]. Il est cependant possible de limiter les pseudo-noeuds explorés par les algorithmes à ceux rencontrés le plus fréquemment, et d'obtenir ainsi des algorithmes efficaces (polynomiaux) pour des espaces de conformations limités [7, 3, 1, 5, 2, 6, 10].

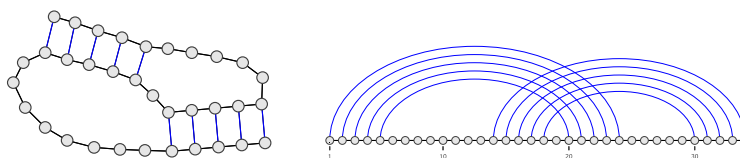


FIGURE 2 – Un pseudonoeud de type H (Gauche) et sa représentation linéaire (Droite)

Dans ce projet, on s'intéressera à une classe simple (mais très peuplée) de pseudo-noeuds, celle considérée par Akutsu [1], qui capture les pseudo-noeuds de type H (Voir Figure 2).

Travail demandé

Implémenter l'algorithme d'Akutsu en un programme `python` :

- **Entrée** : Une séquence d'ARN r
- **Sortie** : La structure secondaire avec pseudonoeuds simples \mathcal{S} , d'énergie minimale pour un modèle de Nussinov.

Une application naive des équations de programmation dynamique (Voir ci-dessous) donne un algorithme en temps $\mathcal{O}(n^5)$. Vous expliquerez pourquoi dans le rapport, ainsi qu'une façon de calculer ces mêmes équations peuvent en temps $\mathcal{O}(n^4)$.

Appariements

On ne considérera que les appariements canoniques : A/U, U/A, C/G, G/C (Watson/Crick), G/U et U/G (Wobble), pondérés par leur nombre de liaison hydrogène. On introduit donc une fonction $v : \{A, C, G, U\} \times \{A, C, G, U\} \rightarrow \mathbb{R}$, qui renvoie la contribution

énergétique associée à une paire de base. Cette fonction prendra typiquement comme valeurs $v(A, U) = v(U, A) = -2$, $v(G, C) = v(C, G) = -3$, $v(G, U) = v(U, G) = -1$, et $v(-, -) = \infty$ sinon.

Par ailleurs, les appariements devront impliquer des bases séparées par au moins θ bases.

Équations de programmation dynamique

L'algorithme d'Akutsu est principalement basée sur une exploration (redondante) de l'ensemble des structures secondaires, au sein desquelles sont insérés des pseudo-noeuds simples. Ces pseudonoeuds simples correspondent à des pseudonoeuds de type H.

Ils sont parcourus par l'algorithme en examinant la configuration de trois bases distinguées i , j et k . Trois cas apparaissent alors : i peut être apparié avec j (Matrice S_L), j avec k (Matrice S_R) ou aucune interaction (Matrice S_M). On poursuit la récurrence en examinant les bases laissées libres par les appariements (ou un sous-ensemble strict de celles-ci si aucune interaction locale).

On obtient alors les décompositions/équations de programmation dynamique résumées par la Figure 3. Ces trois matrices sont alors combinées dans une récurrence générale sur l'énergie $S_P(i_0, k_0)$ d'une région $[i_0, k_0]$ peuplée avec un pseudonoeud simple :

$$S_P(i_0, k_0) = \min_{i_0 \leq i < j < k \leq k_0} (S_L^{i_0}(i, j, k), S_R^{i_0}(i, j, k), S_M^{i_0}(i, j, k))$$

On insère enfin les pseudonoeuds simples dans des structures secondaires normales grâce à l'équation suivante, variante de l'équation historique de Nussinov :

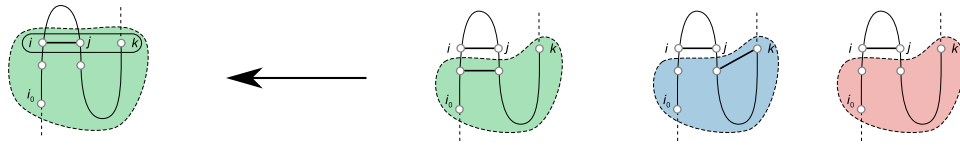
$$S(i, j) = \min \left(S_P(i, j), S(i+1, j-1) + v(r_i, r_j), \min_{i < k \leq j} (S(i, k-1) + S(k, j)) \right)$$

Contact

N'hésitez pas à me contacter (yann.ponty[AT]lix.polytechnique.fr) si un point vous semble obscur ...

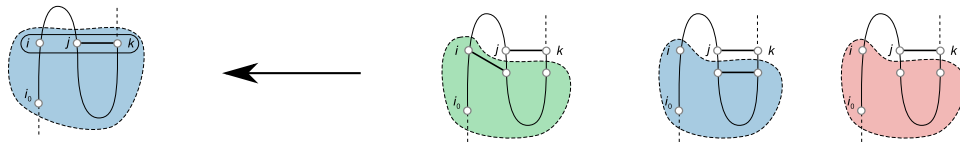
Références

- [1] Tatsuya Akutsu, *Dynamic programming algorithms for rna secondary structure prediction with pseudoknots*, Discrete Appl. Math. **104** (2000), no. 1-3, 45–62.
- [2] S. Cao and S.J. Chen, *Predicting RNA pseudoknot folding thermodynamics*, Nucleic Acids Research **34** (2006), no. 9, 2634–2652.
- [3] Robert M Dirks and Niles A Pierce, *A partition function algorithm for nucleic acid secondary structure including pseudoknots.*, J Comput Chem **24** (2003), no. 13, 1664–1677.
- [4] R. B. Lyngsø and C. N. S. Pedersen, *RNA pseudoknot prediction in energy-based models*, Journal of Computational Biology **7** (2000), no. 3-4, 409–427.
- [5] J. Reeder and R. Giegerich, *Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics*, BMC Bioinformatics **5** (2004), 104.
- [6] Jens Reeder, Peter Steffen, and Robert Giegerich, *pknotsrg : Rna pseudoknot folding including near-optimal structures and sliding windows.*, Nucleic Acids Res **35** (2007), no. Web Server issue, W320–W324.



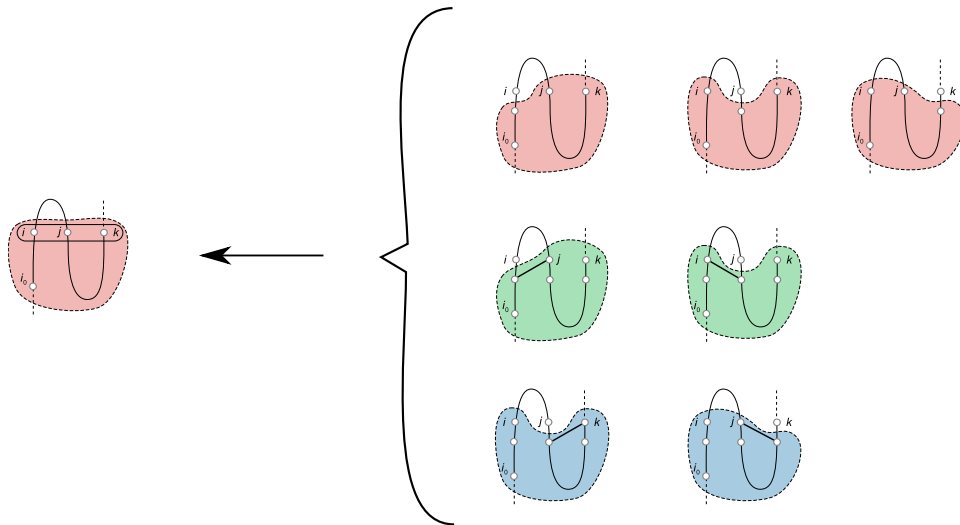
$$S_L^{i_0}(i, j, k) = v(r_i, r_j) + \min \left\{ \begin{array}{l} S_L^{i_0}(i-1, j+1, k), \\ S_M^{i_0}(i-1, j+1, k), \\ S_R^{i_0}(i-1, j+1, k) \end{array} \right\},$$

$$S_L^{i_0}(i, j, j) = v(r_i, r_j), \quad \forall i < j$$



$$S_R^{i_0}(i, j, k) = v(r_j, r_k) + \min \left\{ \begin{array}{l} S_L^{i_0}(i, j+1, k-1), \\ S_M^{i_0}(i, j+1, k-1), \\ S_R^{i_0}(i, j+1, k-1) \end{array} \right\},$$

$$S_R^{i_0}(i_0-1, j, j+\theta+1) = v(r_i, r_{j+\theta+1}), \quad \forall j$$



$$S_M^{i_0}(i, j, k) = \min \left\{ \begin{array}{l} S_M^{i_0}(i-1, j, k), S_M^{i_0}(i, j+1, k), S_M^{i_0}(i, j, k-1), \\ S_L^{i_0}(i-1, j, k), S_L^{i_0}(i, j+1, k) \\ S_R^{i_0}(i, j+1, k), S_R^{i_0}(i, j, k-1) \end{array} \right\}$$

$$S_L^{i_0}(i_0-1, j, k) = S_R^{i_0}(i_0-1, j, k) = S_M^{i_0}(i_0-1, j, k) = 0, \quad \forall j, k \text{ tels que } k-j \leq \theta$$

FIGURE 3 – Décompositions et équations correspondant à des appariement à gauche, droite et absence d'appariement (De haut en Bas) au sein du triplet (i, j, k) .

-
- [7] E. Rivas and S. R. Eddy, *A dynamic programming algorithm for rna structure prediction including pseudoknots.*, J Mol Biol **285** (1999), no. 5, 2053–2068.
 - [8] Michela Taufer, Abel Licon, Roberto Araiza, David Mireles, F. H D van Batenburg, Alexander P Gulyaev, and Ming-Ying Leung, *Pseudobase++ : an extension of pseudobase for easy searching, formatting and visualization of pseudoknots.*, Nucleic Acids Res **37** (2009), no. Database issue, D127–D135.
 - [9] F. H. van Batenburg, A. P. Gulyaev, and C. W. Pleij, *Pseudobase : structural information on rna pseudoknots.*, Nucleic Acids Res **29** (2001), no. 1, 194–195.
 - [10] Jizhen Zhao, Russell L Malmberg, and Liming Cai, *Rapid ab initio prediction of rna pseudoknots via graph tree decomposition.*, J Math Biol **56** (2008), no. 1-2, 145–159.