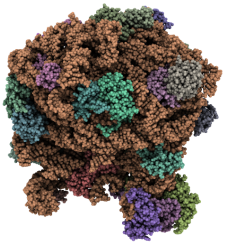
PRÉDICTION ET DESIGN *IN SILICO* EFFICACES D'INTERACTIONS ARN/PROTÉINES

ALAIN DENISE[†], FABRICE LECLERC[‡], AND YANN PONTY[§]



Portion du ribosome
PDBID :1JJ2

Introduction. Initialement, le dogme fondateur de la biologie moléculaire (ADN → ARN → Protéine) n'accordait qu'un rôle de médiateur à l'ARN. Une telle vision, protéocentrique, des mécanismes cellulaires a été considérablement remise en question, au cours des deux dernières décennies, par la découverte de nombreux modes d'actions jusqu'alors insoupçonnés pour l'ARN. A titre d'exemple, on pourra mentionner le phénomène de l'ARN interférence, dans lequel de petits ARN simple-brins inhibent la synthèse de protéines, perturbant ainsi dramatiquement les voies métaboliques concernées. Ce phénomène, considéré comme une alternative crédible à la thérapie génique, est porteur de nombreux espoirs thérapeutiques, et sa découverte aura valu à A. Fire et C. Mello le prix Nobel 2006 de Médecine. Par ailleurs, l'ARN (ribosomal) a encore une fois été a l'honneur avec l'attribution du prix Nobel 2009 de Chimie à V. Ramakrishnan, T. Steitz et A. Yonath pour leurs travaux sur l'étude des structure et fonction du ribosome. Celui-ci, constitué d'un assemblage complexe de protéines et ARN, est un des acteurs cellulaires majeurs.

Cependant, les deux exemples cités (ARN Interference et fonction du ribosome) ont ceci en commun qu'ils sont davantage observés et validés empiriquement, par le biais de données expérimentales (Cristallographie/RMN, expression et/ou cliniques), que compris *ab initio*, i.e. expliqués de façon mécanique. C'est d'ailleurs là une particularité de la biologie et, en un sens, un mal nécessaire, tant la réelle complexité et le grand niveau d'intrication des phénomènes en présence interdisent parfois une modélisation compacte des phénomènes à échelle macroscopique. Cependant, à l'échelle nanométrique, où s'effectuent le repliement et de l'assemblage des molécules, il est possible de s'inspirer de modèles créés et muris par la biochimie et la thermodynamique. Ces phénomènes, dont l'étude est essentielle pour une réelle compréhension des mécanismes du vivant, sont alors modélisés et leur simulation/prédiction *ab initio* associée à des problèmes informatiques *bien définis*, tels les problèmes liés à l'auto-assemblage auxquels on s'intéresse ici.

Enjeux. Récemment, une masse critique de données sur les assemblages protéines/ARN a été atteinte (notamment avec les données structurales sur le ribosome), rendant possible une modélisation plus fine des assemblages ARN/Protéines. Le défi international CAPRI ("Critical Assessment of PRediction of Interactions") teste comparitivement les approches dédiées à la prédiction d'assemblages entre macromolécules. Une interaction ARN/protéine a été dernièrement proposée comme cible lors de l'édition 2008 de CAPRI (round 15, target T34) . Des algorithmes sont donc actuellement élaborés pour prédire les assemblages entre ARN et protéines, notamment à Nancy, au sein des équipes ORPAILLEUR du LORIA (B. Maigret & D. Ritchie) et équipe ARN, RNP, maturation-structure-fonction de AREMS (M. Simoes & F. Leclerc), équipes ayant participé avec succès à l'édition 2008 de CAPRI. Un des challenges actuels dans une optique thérapeutique consiste à prédire la fixation d'un ARN simple-brin, caractérisé par sa séquence, sur une protéine de structure 3D connue.

Une approche possible pour une telle prédiction, la méthode SELEX *in silico*, développée par M. Simoes & F. Leclerc, emprunte à deux grandes familles d'approches utilisées en biologie : D'une part, à la technique expérimentale SELEX [1, 4], qui permet de sélectionner de façon artificielle des ARN ayant une propriété biologique désirée (catalyse, fixation à d'autres biomolécules, etc) ; D'autre part, à des

Key words and phrases. Complexes ARN/protéines ; Design d'ARN ; Algorithmique de graphe ; Génération aléatoire pondérée ; Méthode de Boltzmann.

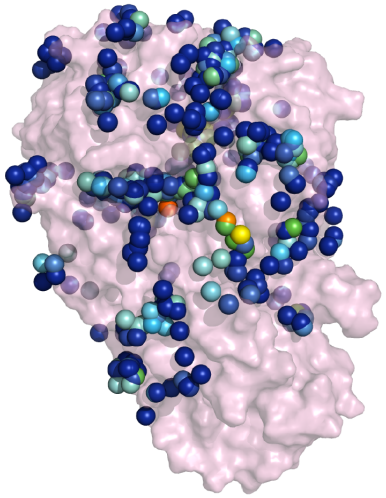


FIG. 1. Distribution de sites de fixation possibles pour le nucléotide A à la surface de la toxine diphtérique (PDBID 1F0L). Chaque nucléotide est représenté par une sphère : 500 sites potentiels de fixation sont représentés et colorés en fonction de la force d'interaction avec la protéine (du rouge pour les plus forts, à l'orange, au jaune, au vert et au bleu pour les moins forts). La distribution des nucléotides A permet alors la recherche de séquences A^n, et la donnée de ces distributions pour les 4 types de nucléotides celle de n'importe quel ARN simple-brin.

méthodes de conception de "drogues" dites "par fragment" [3]. L'approche SELEX *in silico* commence donc par déterminer un ensemble de sites de fixation potentiels côté protéine pour chaque type de nucléotide A, C, G ou U (Voir Figure 1). Ces points sont ensuite utilisés pour initier une reconstruction du site de fixation, illustrée en Figure 2, par élongations successives de la séquence d'ARN. On prend alors soin, au cours de l'élongation, de respecter des critères locaux (proximité de nucléotides adjacents dans la séquence) ou globaux (utilisation unique d'une position donnée, énergie décrits en Figure 1) . Un des avantages majeurs de cette approche réside alors dans son extension naturelle au design d'ARN simple-brin interagissant avec une protéine donnée, extension obtenue en *anonymisant* les nucléotides de la séquence d'ARN en entrée. Cependant, la complexité (nombre d'élongations) d'une implémentation naive de ce principe croit exponentiellement sur la taille de l'ARN considéré, même dans le cas où aucune fixation n'est admissible pour le couple ARN/protéine considéré, et constitue en conséquence un facteur limitant de ce type d'approche.

Formalisation. Nous souhaitons donc développer une approche différente, fondée sur une analogie avec la génération aléatoire de chemins auto-évitants dans un graphe dirigé. Le problème algorithmique sous-jacent peut ainsi être isolé et reformulé de la façon suivante :

> Étant donné un ensemble Σ d'étiquettes, une séquence $\omega \in \Sigma^*$ et un graphe dirigé $G = (V, E)$ étiqueté aux sommets (Par une fonction $\psi : V \to \Sigma$), existe-t-il un chemin dans G, auto-évitant (Au plus une occurrence de chaque noeud dans V), et dont la séquence d'étiquettes associée soit ω ?

Par ailleurs, on pourra souhaiter pondérer ce problème, en adjoignant aux arêtes E du graphe des poids, offrant ainsi la possibilité de modéliser l'affinité du site de fixation avec la surface de la protéine, ou encore des contraintes de chiralité. On s'intéressera enfin au cas d'une séquence *anonyme*, c'est à dire compatible avec tout chemin, auquel cas la résolution du problème ci-dessus permet la conception (Design) d'une séquence d'ARN interagissant avec une protéine donnée.

Aspect théoriques et approche possible. Le problème formulé ci dessus, est NP-complet en général. En effet, on peut facilement vérifier que, dans une restriction à une unique étiquette ($\Sigma = \{x\}$) et une sequence $\omega = x^{|V|}$, la recherche d'un chemin solution est équivalent à celle d'un chemin hamiltonien

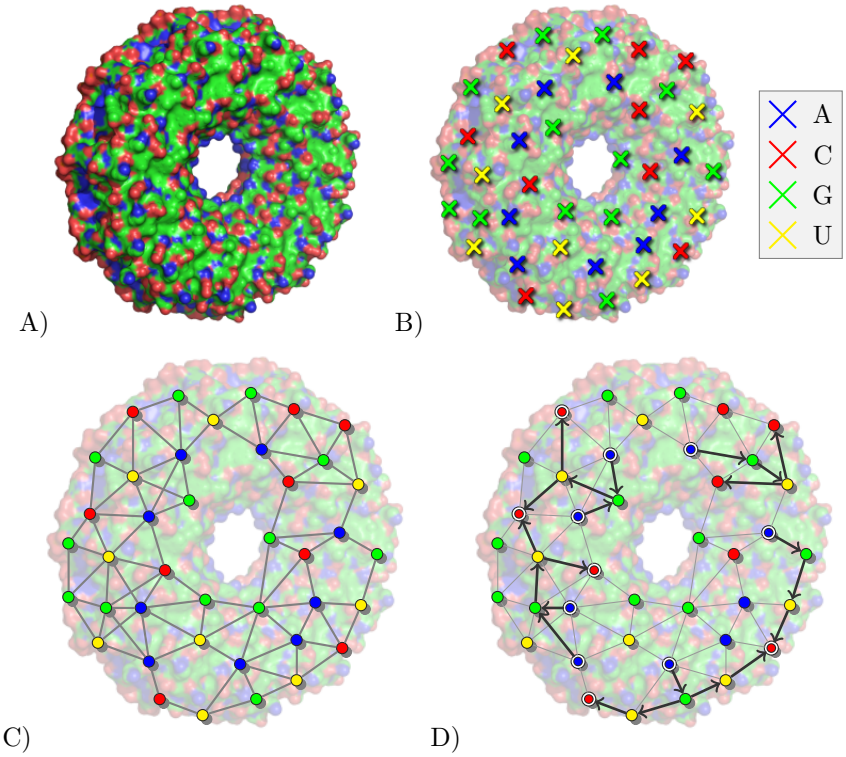


FIG. 2. Méthodologie générale : Partant d'une structure 3D de protéine connue (A, PDBID 1C9S :L-V), la méthode SELEX construit un ensemble de site de fixation pour des nucléotides individuels (B). Des contraintes géométriques, stériques et énergétiques permettent alors de construire un graphe compatibilité locale (C), dans lequel nous rechercherons un ensemble de chemins correspondant à une séquence d'ARN donnée (D, chemins correspondant aux 13 fixations possibles pour la séquence A–G–U–C).

dans G. Ce problème, NP-complet dans le cas d'un graphe général, demeure NP-complet même dans des graphes offrant de fortes régularités, comme par exemple les sous-graphes de grilles carrées [2].

Cependant, la difficulté algorithmique tient ici uniquement à la contrainte d'auto-évitement. En effet, il est possible de résoudre ce problème de façon exacte en temps polynomial en relâchant la contrainte d'auto-évitement, i.e. en autorisant les collisions. Il est alors même possible de pratiquer une génération aléatoire uniforme et/ou exhaustive des candidats par une application de la méthode dite *récursive* [5]. On essaiera donc d'exploiter cette *faiblesse* du problème pour proposer des approches algorithmiques efficaces, en pratique par le biais d'heuristiques ou garanties dans le cadre plus formel de la complexité paramétrée.

Objectifs. Le but principal de ce stage consiste à concevoir et implémenter une stratégie algorithmique d'exploration efficace pour la prédiction et le design d'interactions ARN/protéine. Les moyens à mettre en oeuvre sont multiples :

- Biologie/Biochimie : Affiner les objectifs d'une génération efficace et pertinente biochimiquement dans un dialogue avec le partenaire bioinformaticien/biochimiste du projet (F. Leclerc, Nancy)
- Informatique : Poursuivre une étude des aspects algorithmiques du problème. En particulier, on essaiera d'identifier des contraintes pesant sur les objets du fait de leur nature biologique afin de déduire des simplifications algorithmiques et/ou des heuristiques efficaces.
- Combinatoire : On pourra quantifier (empiriquement, voir analytiquement) la proportion de séquences auto-évitantes dans un cheminement imposé par la séquence, et ainsi évaluer le gain de temps comparé à l'approche naive (backtracking).

Parallèlement à une étude des aspects algorithmiques du projet, un effort d'implémentation devra être fourni afin de tester régulièrement les hypothèses et approches développées sur les données disponibles. L'approche SELEX *in silico* sera appliquée à des cibles protéiques d'intérêt (associées à certaines dystrophies humaines) et pourra faire l'objet d'une validation expérimentale à Nancy à l'issue du stage.

Déroulement du stage. Le stage se déroulera au LIX et sera co-encadré par Y. Ponty (CNRS LIX, Polytechnique) et A. Denise (Professeur LRI/IGM, Paris XI). Des contacts réguliers avec le partenaire co-développeur de la méthode SELEX (F. Leclerc, CNRS AREMS, Nancy I) sont également prévus.

Prérequis. Le candidat devra être doté d'un gout réel pour l'algorithmique et d'une capacité à l'implémentation. Il devra idéalement posséder une bonne maîtrise d'un langage de script (Python, Perl, ...) et d'un langage "classique" (C/C++, Caml, Java, ...).
Des connaissances poussées en biologie ne sont pas nécessaires, mais une curiosité envers la biologie moléculaire/biochimie est souhaitable.

Références

1. A D Ellington and J W Szostak, *In vitro selection of rna molecules that bind specific ligands*, Nature **346** (1990), no. 6287, 818–22 (eng).
2. A. Itai, C. H. Papadimitriou, and J. L. Szwarcfiter, *Hamilton paths in grid graphs*, SICOMP **11** (1982), no. 4, 676–686.
3. W Jahnke, D Erlanson, and 2006, *Fragment-based approaches in drug discovery*, books.google.com (2006), 1–369.
4. Regina Stoltenburg, Christine Reinemann, and Beate Strehlitz, *Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands*, Biomol Eng **24** (2007), no. 4, 381–403 (eng).
5. H. S. Wilf, *A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects*, Advances in Mathematics **24** (1977), 281–291.

† LRI/IGM, Université Paris XI/Paris-Sud – `Alain.Denise@lri.fr`

‡ AREMS, Université Henri Poincaré-Nancy 1 – `Fabrice.Leclerc@uhp-nancy.fr`

§ LIX, École Polytechnique (**Encadrant LIX**) – `Yann.Ponty@lix.polytechnique.fr`