Université de Montréal

Conception de microARNs pour atténuer l'expression de gènes

par
Maxime Caron

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue d'obtention du grade de maître
en bio-informatique

Septembre 2008

Université de Montréal
Faculté des études supérieurs
Ce mémoire intitulé :

Conception de microARNs pour atténuer l'expression de gènes

présenté par :
Maxime Caron

a été évalué par un jury composé des personnes suivantes :

Jean-Yves Potvin, Ph.D.
président-rapporteur

François Major, Ph.D.
directeur de recherche

Gerardo Ferbeyre, M.D. Ph.D.
codirecteur

Jean-Claude Labbé, Ph.D.
membre du jury

Abstract

MicroRNAs belong to the family of small non-coding RNAs and act as down regulators of messenger RNAs and/or their protein products. microRNAs differ from siRNAs by downregulating instead of shutting down. In recent years, numerous microRNAs and their targets have been found in mammals and plants. Bioinformatics plays a big role in this field, as software has emerged to find new microRNA targets. Each individual microRNA can regulate hundreds of genes, and it has been shown that microRNA expression profiles can classify human cancers. The need for artificially created microRNAs is then justified, as one could target overexpressed oncogenes and promote healthy cell proliferation. MultiTar V1.0, a tool for creating artificial microRNAs, has been implemented and is available as a web application. The tool relies on structural and biological properties of microRNAs and uses a Tabusearch metaheuristic. A typical biological problem is presented and it is shown that an in-silico microRNA has in-vitro effects. The 3'UTR sequences of E2F1, E2F2 and E2F3 were given as input to the tool, and predicted microRNAs were then tested using luciferase essays, western blots and growth curves. At least one microRNA is able to regulate the three genes with luciferase essays and all of the created microRNAs were able to regulate the expression of E2F1 and E2F2 with western blots. Growth curves were also studied in order to investigate overall biological effects, and reduction in growth was observed for all solutions. Results obtained with the predicted microRNAs and the target genes open a new door into therapeutic possibilities.

Key words : microRNA, E2F, tabusearch, downregulation, mir20

Résumé

Les microARNs appartiennent à la famille des petits ARNs non-codants et agissent comme inhibiteurs des ARN messagers et/ou de leurs produits protéiques. Les microARNs sont différents des petits ARNs interférants (siARN) car ils atténuent l'expression au lieu de l'éliminer. Dans les dernières années, de nombreux microARNs et leurs cibles ont été découverts chez les mammifères et les plantes. La bioinformatique joue un rôle important dans ce domaine, et des programmes informatiques de découvertes de cibles ont été mis à la disposition de la communauté scientifique. Les microARNs peuvent réguler chacun des centaines de gènes, et les profils d'expression de ces derniers peuvent servir comme classificateurs de certains cancers. La modélisation des microARNs artificiels est donc justifiable, où l'un pourrait cibler des oncogènes surexprimés et promouvoir une prolifération de cellules en santé. Un outil pour créer des microARNs artificiels, nommé MultiTar V1.0, a été créé et est disponible comme application web. L'outil se base sur des propriétés structurelles et biochimiques des microARNs et utilise la recherche tabou, une métaheuristique. Il est démontré que des microARNs conçus in-silico peuvent avoir des effets lorsque testés in-vitro. Les séquences 3'UTR des gènes E2F1, E2F2 et E2F3 ont été soumises en entrée au programme MultiTar, et les microARNs prédits ont ensuite été testés avec des essais luciférases, des western blots et des courbes de croissance cellulaire. Au moins un microARN artificiel est capable de réguler les trois gènes par essais luciférases, et chacun des microARNs a pu réguler l'expression de E2F1 et E2F2 dans les western blots. Les courbes de croissance démontrent que chacun des microARNs interfère avec la croissance cellulaire. Ces résultats ouvrent de nouvelles portes vers des possibilités thérapeutiques.

Mots clés : microARN, E2F, recherche tabou, régulation de l'expression, mir20

Table des matières

Liste des figures

Liste des abréviations

ARN : **A**cide **D**ésoxyribo**N**ucléique

ARN : **A**cide **R**ibo**N**ucléique

poly-A : **p**oly **A**dénosine

RISC : **R**NA **I**nduced **S**ilencing **C**omplex

RMSD : **R**oot **M**ean **S**quare **D**eviation

Å : **Å**rmstrong

NP : **N**on-deterministic **P**olynomial-time

IP : **I**nternet **P**rotocol

Remerciements

Je tiens à remercier tous ceux et celles qui ont participé au sujet de recherche sur lequel j'ai travaillé, dont mon directeur François Major, mon co-directeur Gerardo Ferbeyre ainsi que mes collègues de laboratoire, Vincent De Guire et Marie-France Gaumont-Leclerc.

Je remercie aussi mes coéquipiers de maîtrise en bioinformatique qui ont su apporter bonheur, enthousiasme et réflexions tout au long de cette aventure.

Je n'oublie pas le soutien moral et émotionnel de ma famille et de ma blonde.

Première partie

Introduction

1. Biologie

L'ARN (acide ribonucléique) est une molécule composée de nucléotides qui se trouve dans les cellules des organismes. Les nucléotides sont constitués de trois molécules principales, soit d'une base azotée, d'un sucre ribose et d'un phosphate. Dès 1939, l'ARN était soupçonné de jouer un rôle dans la synthèse des protéines, mais ce n'est qu'en 1959 que le gagnant du prix Nobel de Médecine, Severo Ochua, a découvert comment il était synthétisé. L'ARN diffère de l'ADN sur trois points principaux :

1. l'ARN est habituellement sous forme simple brin tandis que l'ADN est sous forme double brins ;
2. l'ARN est constitué d'un sucre ribose ayant un atome d'oxygène de moins que celui de l'ADN ;
3. l'ARN est constitué de la base azotée uracile qui est une forme non méthylée de la thymine, présente chez l'ADN.

L'ARN est transcrit à partir de l'ADN à l'aide de l'enzyme ARN polymérase. Plusieurs sortes d'ARN existent, dont certains vont donner un produit protéique tandis que d'autres vont effectuer des activités enzymatiques sous forme native.

1.1 ARN messager

L'ARN messager (ARNm) est le précurseur des produits protéiques et est composé principalement de régions codantes (exons) et de régions non codantes (introns). Son cycle commence lorsqu'il est transcrit de l'ADN dans le noyau de la cellule par l'ARN

polymérase II. Ce produit s'appelle le pré-ARNm, une version précaire qui doit subir plusieurs transformations avant de devenir l'ARNm mature, la molécule qui sera traduite en protéines. Les transformations du pré-ARNm incluent la greffe d'une coiffe 7-méthylguanosine à l'extrémité 5', l'excision des introns, une édition des nucléotides (dans certains cas) ainsi que l'addition d'une queue poly-A (plusieurs adénosines). Un des rôles de la coiffe et de la queue poly-A est d'apporter une résistance à la dégradation par les enzymes exonucléases. Une fois modifié, l' ARNm mature (figure 1.1) est transporté à l'extérieur du noyau et est traduit en protéines par le ribosome.

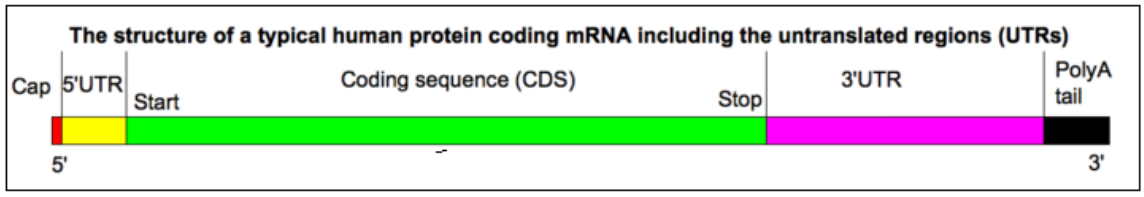

Figure 1.1: Parties principales d'un ARN messager. L'ARN messager (ARNm) contient une coiffe en 5', une région codante (exons) ainsi qu'une queue poly-A. Photo tirée de wikipedia.

1.2 ARN de transfert

L'ARN de transfert (ARNt) est la molécule responsable du recrutement des acides aminés lors de la traduction de l'ARN messager. Sa taille est d'environ 80 nucléotides et il a une structure typique en forme de trèfle (Figure 1.2). L'ARN de transfert est composé de quatre parties principales dont une tige recrutrice d'acides aminés à son extrémité 3', une boucle anti codon qui se lie à la séquence de l'ARNm, un bras 'D' et un bras 'T'.

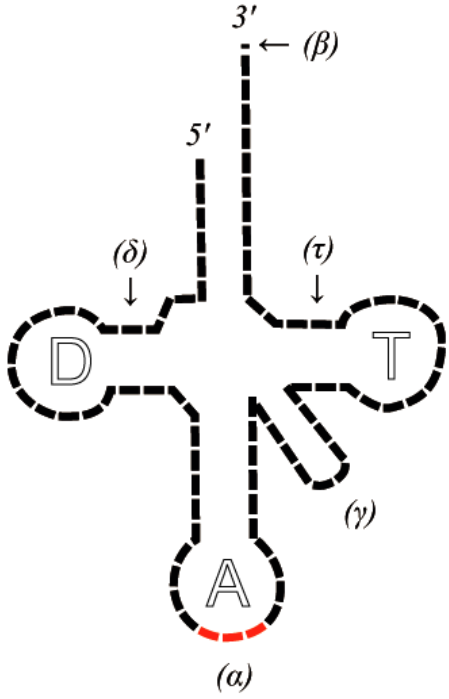

Figure 1.2: Parties principales d'un ARN de transfert. L'ARN de transfert (ARNt) est composé d'une tige recrutrice d'acides aminés (α), d'une boucle anti codon (β), d'un bras variable (γ) d'un bras 'T' (τ) et d'un bras 'D' (δ). Photo tirée de wikipedia.

1.3 ARN ribosomal

L'ARN ribosomal (ARNr) est le noyau catalytique du ribosome, complexe responsable du recrutement des ARNt et de la synthèse de peptides. Il est composé de deux parties principales, soit (chez les eucaryotes) la grande sous-unité 60S et la petite sous-unité 40S. Les chiffres correspondent à la vitesse à laquelle les sous-unités sédimentent lors de la centrifugation. L'ARN ribosomal contient trois sites de liaisons, le site A, P et E, tous responsables d'interagir avec les ARNs de transfert. Notons que l'ARN ribosomal est le gène le plus conservé chez tous les types cellulaires [1].

1.4 Petit ARN d'interférence

Le petit ARN d'interférence (siARN) est une molécule simple brin ayant une longueur de 20 à 25 nucléotides et interférant avec l'expression des gènes. Le brin mature du siARN, apparié au complexe protéique RISC (RNA-induced Silencing Complex), active la dégradation de l'ARNm auquel il se lie lorsque l'appariement des bases est presque parfait. Depuis sa découverte en 1999 comme mécanisme pouvant réguler l'expression des gènes dans les plantes [2], il ne cesse de générer de l'enthousiasme dans la communauté scientifique, principalement pour ses implications thérapeutiques. En 2001, une équipe a démontré que le siARN est capable de réguler l'expression des gènes non seulement chez les plantes, mais aussi chez les mammifères [3].

1.5 microARN

Les microARNs (miARN) sont des molécules simple brin d'environ 22 nucléotides qui ressemblent aux petits ARNs d'interférence, mais dont l'appariement aux gènes n'est que partiel et la régulation moins sévère. Ils se retrouvent aussi bien dans les cellules des plantes que dans les cellules des mammifères. Ils ont été découverts en 1993 dans l'organisme C. elegans [4] mais n'ont gagné en popularité qu'en 2001, lorsqu'ils ont été formellement appelés microARNs [5]. Les microARNs se lient aux régions 3' non traduites des ARNm et agissent de plusieurs façons, par exemple en accélérant la dégradation des ARNm ou en bloquant leur traduction [6]. Ils sont impliqués dans une panoplie de fonctions cellulaires dont le développement [7], l'augmentation de la mort cellulaire [8], le développement du cerveau [9], les cancers [10] [6] et les infections virales [11]. Des centaines de microARNs ont été répertoriés chez les eucaryotes [12] et chacun d'entre eux peut réguler des centaines de gènes, ce qui suggère que l'importance de leur régulation se mesure à l'importance de la régulation des facteurs de transcription. Environ 60% des microARNs sont exprimés seuls, 15% sont exprimés en groupe

et 25% se retrouvent à l'intérieur d'introns.

1.5.1 Cheminement biologique du microARN

Les microARNs débutent en étant transcrits dans le noyau des cellules par l'enzyme ARN polymérase II. Les transcrits comprennent une coiffe en 5', une queue poly-A et ont la forme d'une tige boucle dont la structure est propice au clivage par des nucléases [13] [14]. Les microARNs transcrits sont nommés pri-microARNs [15] et peuvent être transcrits seuls ou en groupe.

Une fois que le microARN est transcrit, il est digéré par la nucléase Drosha et le produit résultant est le pré-microARN [15] , une molécule d'environ 80 nucléotides. Le pré-microARN a certaines caractéristiques typiques dont une séquence simple brin de 1 à 4 nucléotides à son extrémité 3', une double hélice d'environ 30 nucléotides et des cycles internes de petite dimension. Le pré-microARN est par la suite transporté dans le cytoplasme par la molécule Exportine-5, molécule qui se lie seulement aux pré-microARNs qui ont été correctement clivés.

Dans le cytoplasme, le pré-microARN est reconnu par l'enzyme Dicer, une enzyme de la famille des RNases III. Cette dernière clive le pré-microARN à 19 nucléotides du site de clivage de l'enzyme Drosha [14] et produit une molécule double brins ayant une région flanquante simple brin de 1 à 4 nucléotides à l'une ou l'autre de ses extrémités. Seulement un des deux brins est sélectionné comme étant le microARN mature qui effectuera sa fonction de régulation d'expression de gènes. La sélection du brin mature est basée sur l'énergie libre de liaison des nucléotides à leurs extrémités 5' [16] [17] et ce brin mature sera incorporé au complexe protéique RISC, complexe responsable d'amener le microARN à ses cibles. La figure 1.3 résume les étapes principales du cheminement biologique des microARNs.

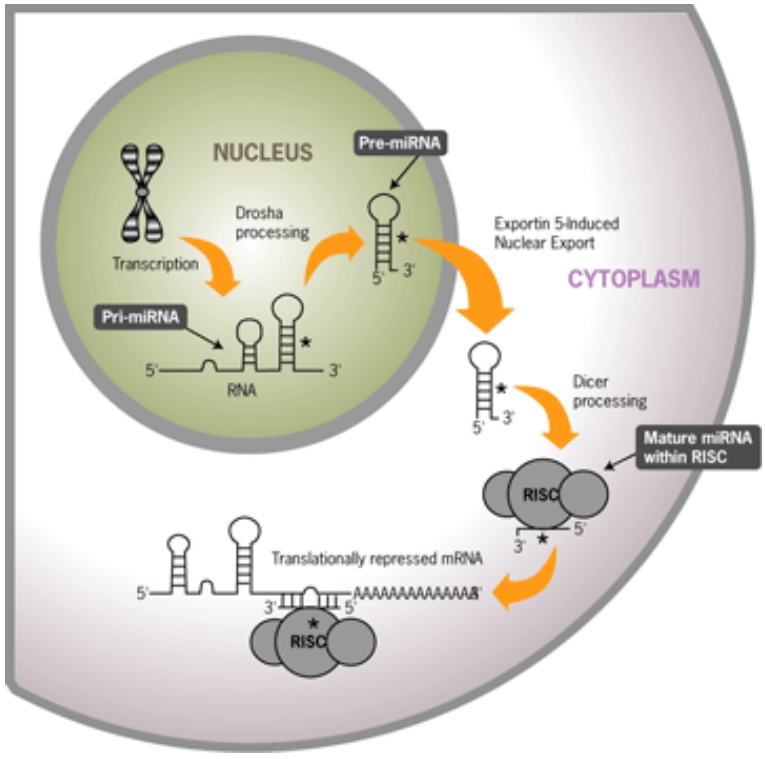

Figure 1.3: Étapes principales du cheminement biologique des microARNs. Le microARN est transcrit dans le noyau de la cellule et transporté dans le cytoplasme pour être ensuite clivé par l'enzyme Drosha. Le pré-microARN résultant est de nouveau clivé par l'enzyme Dicer et un des deux brins est intégré au complexe protéique RISC. Photo tirée de http ://www.ambion.com/techlib/resources/).

1.5.2 Bases de données des microARNs

Le nombre de microARNs endogènes chez les organismes eucaryotes supérieurs (dont les humains et les souris) se situe autour de quelques centaines et leur nombre de cibles est de plusieurs centaines, voire même quelques milliers. Les bases de données des microARNs sont d'une utilité fondamentale, car elles permettent de répertorier toute cette information. mirBase [18] est une base de données créée en 2004 qui regroupe plus de 5000 microARNs répertoriés chez plus de 58 organismes. Dans les deux dernières années, la base de données a reçu plus de 2000 nouvelles soumissions de microARNs. En plus de contenir les microARNs, mirBase contient les cibles potentielles de ces derniers, découverts à l'aide du programme informatique miRanda [19]. Or, le besoin d'avoir des

cibles validées expérimentalement est important, car les programmes informatiques ne sont pas précis à 100%. C'est pourquoi TarBase [20] a vu le jour, une base de données regroupant les cibles de microARNs vérifiées expérimentalement. TarBase contient les cibles des microARNs chez 8 organismes et décrit en détail les types d'inhibitions, les alignements de la liaison des microARNs à leurs cibles ainsi que le type d'expérience qui a été effectuée afin de valider l'inhibition de ces microARNs.

1.6 Gènes E2F1, E2F2 et E2F3

Les gènes E2Fs sont des facteurs de transcription qui interviennent dans la prolifération cellulaire, l'apoptose (mort cellulaire) et la réparation de l'ADN [21]. Au total, 8 protéines font partie de la famille des E2Fs, dont E2F1 E2F2 et E2F3 qui peuvent induire l'activation de la phase S de la division cellulaire chez les cellules dormantes. Il est connu que les gènes E2F1, E2F2 et E2F3 ont des effets redondants au niveau du cycle cellulaire, mais chacun de ces derniers a aussi des effets spécifiques. Une diminution de l'expression de E2F1, E2F2 et E2F3 diminue la prolifération cellulaire [22]. De plus, une suractivation de E2F1 semble être liée à l'apoptose ; plus précisément, un niveau anormalement bas de E2F1 a été découvert dans les cellules cancéreuses humaines comparativement aux niveaux dans les tissus normaux [23].

1.7 mir-20

mir-20 est un microARN faisant parti du groupe de microARNs mir-17-92, groupe impliqué dans le développement du coeur, des poumons, du système immunitaire et des cancers [24], [25], [26]. Le groupe de microARNs mir-17-92 est un gène polycistronique s'étendant sur 800 bases et qui regroupe 6 microARNs dont mir-17, mir-18, mir-19a, mir-19b, mir-20a et mir-92. Ce groupement est hautement conservé chez les vertébrés et une mutation ancestrale a résulté au maintien de deux paralogues. Ces microARNs

sont activés par le facteur de transcription c-Myc, qui est habituellement surexprimé dans les cellules cancéreuses. Les premières cibles des microARNs mir-17-92 qui ont été découvertes sont les gènes E2F1, E2F2 et E2F3 [22]. Il semble que lorsque ces microARNs sont surexprimés, l'activité des gènes E2Fs est inhibée et ceux-ci ne peuvent plus activer le développement cellulaire de façon contrôlée. Les gènes E2Fs peuvent agir contre l'expression du groupe mir-17-92, produisant un cycle d'autorégulation négatif. Il fut démontré que dans certaines cellules, une surexpression ou une atténuation du groupe mir-17-92 est suffisante pour induire ou prévenir l'entrée dans la phase S (phase de réplication de l'ADN) des cellules dormantes [27].

2. Bioinformatique de l'ARN

La bioinformatique, peu après ses débuts, s'est concentrée sur l'ARN, dont la prédiction de la structure 2D et 3D, ses caractéristiques structurelles, les interactions ARN-ARN et ARN-ADN, la découverte de microARNs et leurs cibles, et plus encore. Les approches développées se basent sur de grandes quantités de données biologiques recueillies en laboratoire et infèrent leurs résultats en utilisant des méthodes mathématiques et informatiques.

2.1 Structure 2D de l'ARN

La prédiction de la structure 2D de l'ARN remonte à 1978 avec une forme simple d'algorithme proposée par Nussivo [28]. Par la suite, les algorithmes de minimisation de l'énergie libre (MFE) ont vu le jour en 1981 [29]. Ces algorithmes se basent sur la programmation dynamique appliquée à des données calorimétriques de sous-structures d'ARN. Ces données ont été mises à jour en 2004 [30]. Les deux implantations les plus populaires de la prédiction par MFE sont MFOLD [31] et RNAFold [32]. Le point faible de ces approches est qu'elles ne considèrent pas les pseudonoeuds, une sous-structure contenant deux hélices dont une fait partie de l'autre. Pourtant, ces structures existent dans les formes endogènes tridimensionnelles des ARNs. En 1999, un algorithme traitant ces pseudonoeuds a vu le jour [33]. Or, la complexité de ce dernier ne rend pas son utilisation pratique pour des problèmes courants, donc certaines restrictions qui sont émises par les utilisateurs sur le type de pseudonoeuds à considérer. De plus, ces approches ne considèrent que les appariements de bases canoniques et ignorent les

appariements non canoniques. Une étude sur l'efficacité de prédiction des différentes approches de prédiction de la structure 2D de l'ARN a été faite en 2005 [34]. Cette étude considère deux approches, soit la prédiction d'une seule séquence et la prédiction d'une séquence utilisant l'information de séquences homologues. MFOLD et RNAFold sont des algorithmes de prédiction n'utilisant qu'une seule séquence, et leur meilleure valeur de prédiction est chiffrée à environ 70%.

2.2 Structure 3D de l'ARN

La prédiction 2D de l'ARN offre une approximation sur la structure, mais ne peut se rapprocher de l'efficacité que pourrait obtenir des algorithmes qui tiennent compte des aspects tridimensionnels de l'ARN. Ce problème, beaucoup plus compliqué, a connu des percées intéressantes au début des années 1990 avec la création de MC-SYM, un programme de modélisation par contraintes [35]. MC-SYM a permis de reproduire la structure tridimensionnelle d'un ARN de transfert de la levure. Près de 15 ans plus tard, une méthode de prédiction de structure basée sur des motifs cycliques de l'ARN et MC-SYM est apparue [36], permettant d'obtenir des représentations 3D de structures d'ARN à partir d'une séquence primaire (figure 2.1). Un aspect intéressant de cette méthode est que la structure secondaire de l'ARN est déterminée en considérant non seulement les appariements de bases canoniques, mais aussi les non canoniques. Ces appariements se trouvent à l'intérieur des structures endogènes d'ARN. La précision de l'approche est déterminée par la RMSD (Root Mean Square Deviation) des structure 3D de plusieurs petits ARNs et se situe entre 1 et 3 Å. Une autre approche différente est apparue peu après, nommée iFoldRNA [37]. Cette méthode se base sur des simulations moléculaires dynamiques et peut aussi faire des prédictions à partir de la séquence primaire. La RMSD de la méthode est, en moyenne, de 2 à 5 Å lorsque testée sur de petits ARNs.

Figure 2.1: Prédiction de la structure 3D de l'ARN. La structure secondaire de l'ARN est prédite à partir de la séquence primaire, et la structure tertiaire est obtenue à partir de la structure secondaire. Image tirée de MC-Pipeline.

2.3 Prédiction des cibles des microARNs

La découverte des microARNs a engendré un intérêt en bioinformatique et des approches pour détecter les cibles de nouveaux microARNs ont émergé. Quelques unes des premières approches ont utilisé l'information sur la conservation des séquences chez plusieurs organismes et la prédiction de structure 2D de l'ARN pour trouver des cibles chez la Drosophile [38] [39]. Par la suite, d'autres programmes ont fait surface dont TargetScan [40], miRanda [19], PicTar [19] et RNAhybrid [41], des programmes de prédiction des cibles des microARNs chez les mammifères. Ces méthodes utilisent de nombreuses propriétés biologiques des microARNs et de leurs cibles (séquence, énergie du duplexe microARN/ARNm, etc.) ainsi que la conservation des séquences chez plusieurs espèces. D'autres méthodes plus récentes utilisant l'apprentissage machine [42] et le classificateur naïf de Bayes [43] ont vu le jour. La plupart des méthodes énumérées trouvent la majorité des cibles vérifiées expérimentalement (jusqu'a 94% de vrais positifs) mais ont l'inconvénient de trouver énormément de faux positifs (certains jusqu'à 50%). Malgré tout, ces approches permettent de guider les biologistes vers des pistes

prometteuses. Grâce à ces méthodes, il est estimé que 1% des gènes qui codent pour les microARNs dans le génome régulent l'expression de 10% des produits protéiques.

3. Métaheuristiques

3.1 Description

Les métaheuristiques, par définition, sont des heuristiques qui gouvernent d'autres heuristiques. Elles sont utilisées pour trouver des solutions quasi optimales à des problèmes dont l'espace des solutions est immense. Plusieurs exemples sont dans cette catégorie, dont le problème combinatoire du voyageur de commerce. Ce problème est NP-dur, c'est-à-dire qu'il est au moins aussi difficile que le problème NP le plus exigeant. Les problèmes NP ne pouvent être résolus en temps polynomial. Plusieurs métaheuristiques sont donc grandement utilisées aujourd'hui, dont la recherche locale, les algorithmes génétiques [44], la recherche tabou [45] et les colonies de fourmis [46]. Les métaheuristiques utilisent une fonction d'évaluation définie par l'utilisateur (normalement ajustée au problème étudié) qui permet de distinguer les bonnes solutions des mauvaises ; il faut donc minimiser ou maximiser cette fonction. La recherche de la solution optimale débute avec une solution initiale prometteuse (à déterminer selon le problème) et est guidée vers d'autres solutions avoisinantes. Une classe de transformations adaptée au problème permet de générer des solutions voisines à partir de celle qui est obtenue à l'étape courante de la recherche. Ceci est répété jusqu'à ce qu'une condition d'arrêt, spécifiée par l'utilisateur, soit vérifiée (généralement un nombre X d'itérations ou un seuil Y pour la fonction d'évaluation). Il n'est pas garanti que les métaheuristiques vont trouver une solution optimale, et elles ont parfois tendance à rester coincées dans un minimum ou maximum local, c'est-à-dire dans un sous-espace de solutions qui ne contient pas de solution ayant une meilleure évaluation que la so-

lution courante, qui elle n'est pas optimale.

3.2 Minimum local versus minimum global

Lorsqu'on essaie de minimiser une fonction d'évaluation, on essaie de trouver le minimum global, c'est-à-dire la solution qui a la meilleure valeur dans l'espace de toutes les solutions. La figure 3.1 montre un exemple de minimum local, minimum global, maximum local et maximum global pour une fonction d'évaluation.

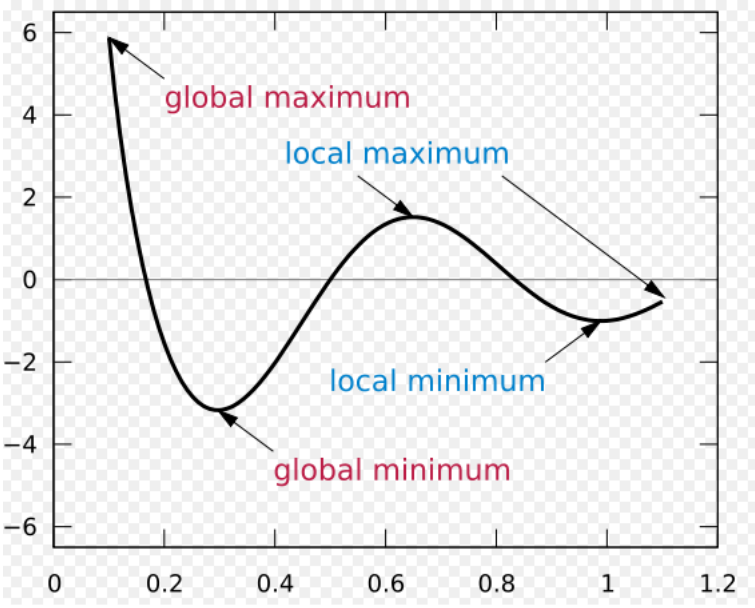

Figure 3.1: Exemple de valeurs locales et globales. La figure illustre un minimum local, un minimum global, un maximum local et un maximum global pour une fonction d'évaluation. Image tirée de wikipedia.

Les métaheuristiques ont certains mécanismes qui leur permettent de sortir de ces minima/maxima locaux. Par exemple, la recherche tabou possède une liste tabou qui empêche d'explorer des solutions précédemment visitées (mémoire à court terme). Elle possède aussi une procédure de diversification qui permet de repartir la recherche à un autre endroit dans l'espace des solutions.

3.3 Applications en bioinformatique

Plusieurs applications en bioinformatique font usage des métaheuristiques. Les algorithmes génétiques sont utilisés pour l'alignement de séquences [47], le design de séquences initiatrices [48] et la prédiction des données d'expression de gènes [49]. La recherche tabou a été utilisée pour la prédiction de structures d'ARN en construisant des réseaux biologiques [50] et les colonies de fourmis pour la sélection des sommets dans les spectres MALDI-TOF [51]. Les métaheuristiques ont leur place en bioinformatique, et plusieurs méthodes innovatrices les utilisent afin de résoudre des problèmes complexes.

Deuxième partie

Méthodologie

4. Problématique, hypothèses et données à l'étude

Les microARNs endogènes ciblent chacun des centaines de gènes à l'intérieur des cellules. Serait-il possible de concevoir un microARN artificiel ciblant plusieurs gènes spécifiques ? Telle est la motivation derrière la création du projet de ce mémoire. Étant donné les séquences 3'UTR de plusieurs gènes, pourrait-on trouver le microARN le plus efficace qui ciblerait tous ces gènes en même temps. Contrairement aux siARNs, les microARNs ne diminuent pas entièrement l'expression d'un gène, mais atténuent son effet et ramènent un équilibre du niveau d'expression au sein de la cellule. Une application concrète serait de cibler les gènes surexprimés de souches cellulaires cancéreuses et de s'attaquer aux gènes surexprimés afin de ramener leurs niveaux d'expression à un seuil approprié, sans avoir à injecter de nombreux siARNs limitant ainsi les effets non spécifiques de ces derniers.

Les données utilisées pour ce projet proviennent de DIANA Tarbase [20], une base de données contenant les microARNs et leurs cibles qui ont été vérifiés expérimentalement. La base de données contient présentement des séquences pour 8 organismes, mais seuls les microARNs de l'humain furent considérés, car il s'agit de l'organisme cible du projet. Au moment de l'obtention des données, 124 microARNs et cibles différentes furent utilisées, que le microARN agisse au niveau de la répression de la traduction ou au niveau de la dégradation de l'ARN messager. Après nettoyage des données, 110 microARNs furent retenus. Le nettoyage a éliminé les alignements non réalistes (boucles trop grandes (> 10 nucléotides)) ou les alignements avec information manquante sur les

nucléotides (alignement contenant de multiples 'N'). La liste finale a été utilisée principalement pour obtenir des données de base, par exemple la longueur moyenne des microARNs, le nombre de mésappariements dans le 'seed', le nombre de paires G/U, etc. Outre cette base de données, plusieurs caractéristiques de l'algorithme furent obtenues de divers articles qui ont étudié les aspects biologiques et bio-informatiques de l'interaction des microARNs avec leurs cibles.

5. Propriétés des microARNs et de leurs cibles

Afin de pouvoir bien modéliser les microARNs, il faut se baser sur plusieurs aspects biologiques et structuraux. Certains aspects sont très rudimentaires (longueur du microARN) et d'autres plus complexes (cycles structuraux 2D, énergie libre). Cette section décrit en profondeur toutes les caractéristiques utilisées pour concevoir les microARNs artificiels.

5.1 Longueur du microARN

Il est souvent rapporté dans la littérature qu'un microARN a une longueur d'environ 22 nucléotides, variant généralement de 19 à 23 nucléotides. Lors de la conception d'un microARN, il faut se décider sur une longueur précise, car la séquence générée doit avoir une longueur fixe. Après avoir analysé la longueur de 110 microARNs pour leur longueur (Figure 5.1), 57% des microARNs ont une taille de 22 nucléotides, donc c'est la taille qui a été retenue.

5.2 Énergie libre du seed

Un microARN est composé de trois composantes principales (Figure 5.2). En allant de 5' à 3', la première composante est le premier nucléotide qui fut déterminé comme étant une adénosine dans la grande majorité des cas [52]. Chacun des microARNs générés par l'algorithme contient une adénosine en première position.

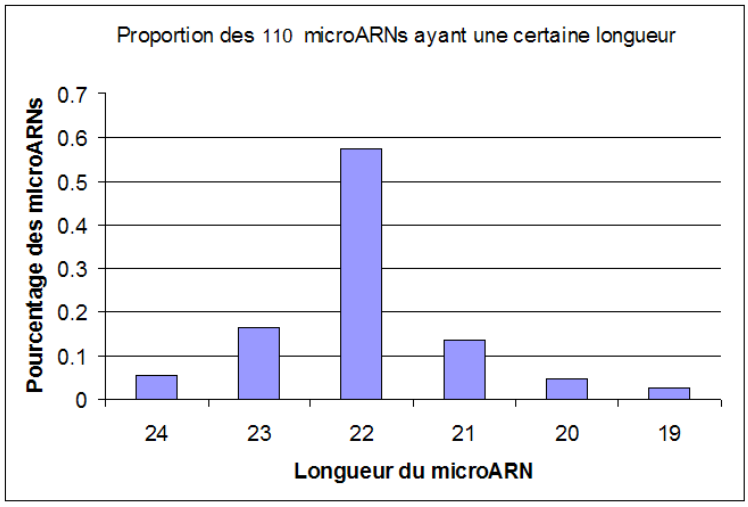

Figure 5.1: Longueur des microARNs. La longueur des microARNs qui s'échelonne sur 19 à 24 nucléotides. La majorité a une longueur de 22 nucléotides.

Ensuite il y a le 'seed' (semence), séquence partant de la deuxième à la huitième position. Finalement, il y a la partie 3' UTR qui commence à la neuvième position et s'échelonne jusqu'à la fin du microARN. Il est reconnu que le seed joue en grande partie le rôle d'ancrage pour les ARN messagers, mais qu'un équilibre existe entre la force de liaison du seed et de la partie 3'. Ainsi, quand la force de liaison du seed est insuffisante, une force de liaison supplémentaire au niveau du 3' est requise.

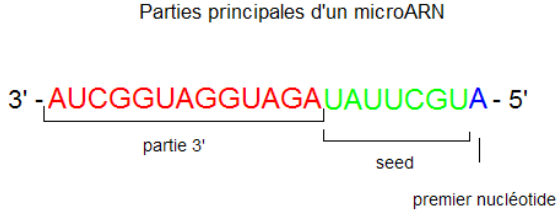

Figure 5.2: Composantes principales du microARN. Le microARN comporte trois composantes principales. À partir de 5', le premier nucléotide (position 1), le seed (positions 2-8) et la partie 3' (positions 9 et +).

En analysant l'énergie libre de la liaison du seed aux ARN messagers, nous pouvons déterminer le seuil minimal acceptable. Lorsque l'on regarde la distribution de l'énergie de liaison du seed sur les 110 complexes microARNs/mARN (Figure 5.3) et en tenant compte d'une étude sur l'énergie libre du seed des microARNs [53], le seuil minimal

acceptable a été fixé à -5 kcal/mol.

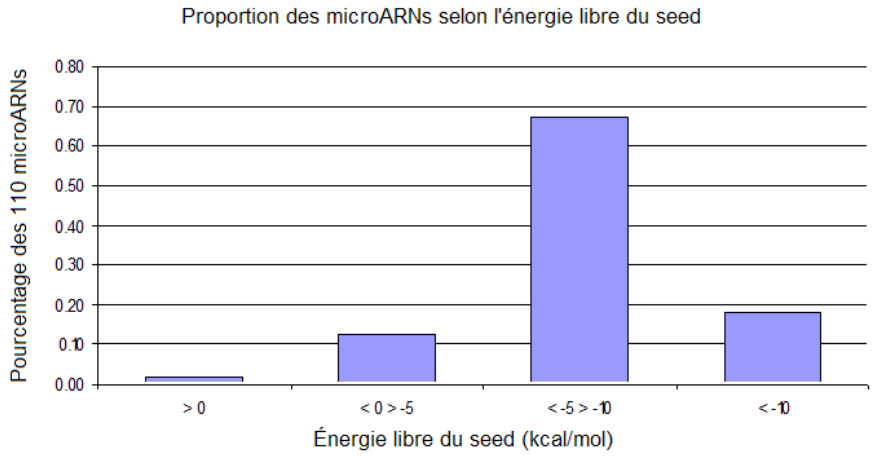

Figure 5.3: Énergie libre de liaison des seeds. La majorité des seeds ont une énergie libre de liaison inférieure à -5 kcal/mol.

5.3 Propriétés structurelles et génomiques du seed

Outre l'énergie libre de liaison du seed, il faut déterminer d'autres facteurs structuraux et génomiques qui peuvent influencer l'efficacité de l'appariement. En se basant sur les 110 microARNs de notre ensemble de données, la majorité de ces derniers ont un appariement parfait (7/7 nucléotides) (Figure 5.4). D'autres ont soit un ou deux mésappariements. Intuitivement, il semble être souhaitable dans la majorité des cas de forcer un appariement parfait, mais ceci ne convient pas pour deux raisons principales. Premièrement, un appariement parfait contenant plusieurs nucléotides A/U sera moins favorable énergétiquement qu'un mésappariement contenant plusieurs nucléotides G/C, étant donné le nombre de ponts hydrogènes formés. De plus, forcer un appariement parfait réduit grandement le nombre de solutions possibles, surtout si le nombre de gènes ciblés est grand. Pour ces raisons, l'algorithme accepte des appariements parfaits ou des mésappariements de un ou deux nucléotides.

Dans les programmes d'alignement 2D de l'ARN, il existe la paire de bases G/U, appelée paire 'wobble'. Plusieurs données contradictoires existent quant à l'inclusion de cette paire à l'intérieur de l'alignement d'un microARN. Une étude sur l'efficacité de la répression des microARNs en présence de paires G/U a démontré que l'incorporation d'une seule paire à l'intérieur du seed a un effet néfaste, même si cette paire est énergétiquement favorable [53]. L'incorporation de 3 paires a complètement annulé l'effet du microARN. D'un autre côté, un programme bio-informatique de recherche de cibles potentielles de microARNs sur le génome humain [40] autorise les paires G/U et obtient des résultats intéressants. En tenant compte de ces deux études, et de la distribution de ces paires parmi les 110 microARNs (Figure 5.5), l'incorporation de paires G/U a été considérée comme étant un facteur risqué à introduire dans l'algorithme, donc les paires G/U ne sont pas permises.

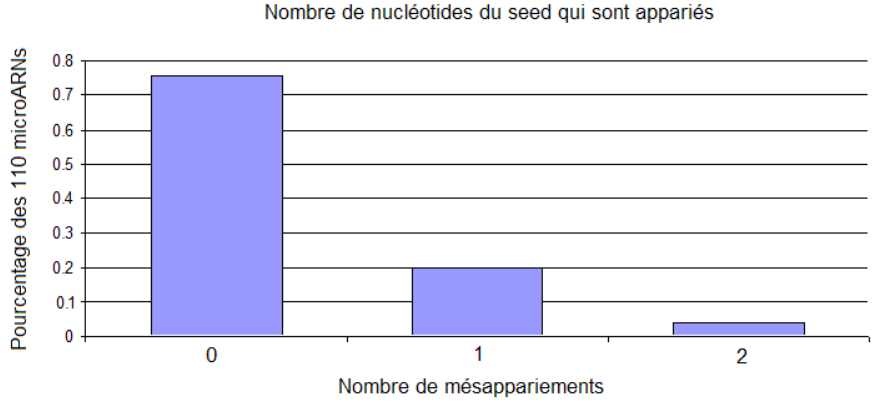

Figure 5.4: Nombre de mésappariements dans les seeds des microARNs. Les seeds des microARNs ont majoritairement un appariement parfait, mais peuvent contenir un ou deux mésappariements.

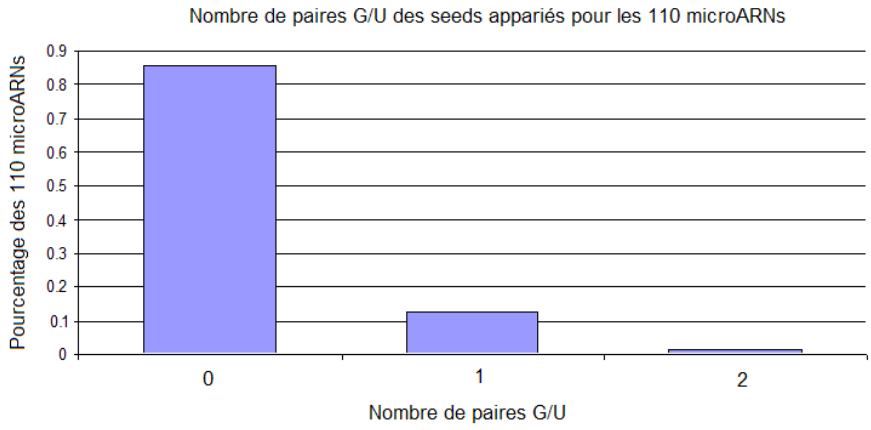

Figure 5.5: Nombre de paires G/U dans les seeds des microARNs. Les seeds des microARNs ne contiennent majoritairement aucune paire G/U, mais peuvent rarement en contenir une ou deux.

5.4 Incorporation des brins dans le complexe protéique RISC

Comme présenté à la section 1.5, l'incorporation du brin mature du microARN n'est pas aléatoire. Cette incorporation dépend de l'énergie libre de liaison des deux extrémités (5' et 3'). Afin de favoriser l'incorporation de notre brin dans le complexe RISC, chaque seed doit avoir 2 nucléotides A ou U dans les 3 premiers nucléotides de l'extrémité 5'. Ceci va de concert avec la contrainte d'avoir au moins 2 nucléotides G ou C sur 4 dans les 4 derniers nucléotides de l'extrémité 3' (voir section 5.7). De cette manière, on s'assure que l'extrémité 5' est liée avec moins de force que l'extrémité 3', favorisant ainsi la sélection de notre brin artificiel.

5.5 Cycles structuraux 2D - Decompose.java

Les cycles structuraux observés lors de la liaison entre un microARN et sa cible contiennent de l'information sur les préférences structurelles d'une telle liaison. Une étude sur ce sujet rapporte, entres autres, que la présence d'un cycle unilatéral de

taille variant entre 2 et 6 nucléotides ou d'un cycle bilatéral de 2 à 3 nucléotides se retrouve majoritairement au milieu des complexes microARN/ARNm [54]. Un programme développé au cours du projet (Decompose.java) décompose et quantifie la nature des cycles structuraux 2D de ces complexes. La même conclusion n'a pu être observée dans notre base de données de 110 microARNs. Ne pas utiliser l'information structurelle des cycles pourrait nuire à la modélisation d'un microARN artificiel, donc une approche alternative a été adoptée. L'algorithme tient compte de l'information structurelle des cycles, mais indépendamment de leur position. Les cycles structuraux et leurs fréquences obtenues à partir des 110 microARNs sont illustrés dans la figure 5.6.

Cycle	Fréquence	Cycle	Fréquence
0-4	0.016	0-3	0.038
0-2	0.099	4-0	0.020
1-2	0.018	5-0	0.016
1-0	0.204	3-0	0.063
0-6	0.014	2-0	0.097
2-1	0.016	1-1	0.038
0-5	0.012	0-1	0.274

Figure 5.6: Cycles structuraux 2D des complexes microARN/ARNm et leurs fréquences. Les cycles répertoriés ont leurs fréquences déterminées par le nombre d'occurrences divisé par le nombre total de cycles.

5.6 Énergie du complexe microARN/ARN messager

L'énergie libre du complexe microARN/ARN messager doit être suffisante pour qu'il y ait une liaison stable permettant au duplexe de pouvoir agir selon ses mécanismes d'activation. En se basant sur plusieurs études [38, 39, 55] ainsi qu'un programme bio-

informatique populaire de prédiction de cibles de microARNs [56], un consensus sur un seuil de 30% est rapporté. Ce seuil est calculé en divisant l'énergie libre du duplexe microARN/ARNm par l'énergie libre du duplexe microARN/miARN-complémentaire-inversé (figure 5.7). Ce seuil est implanté dans l'algorithme et toute solution sous le seuil est rejetée.

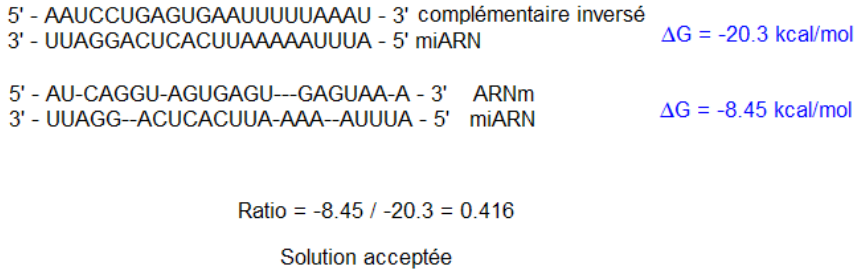

Figure 5.7: Énergie libre du complexe microARN/ARN messager. L'énergie libre du duplexe miARN/ARNm est divisée par l'énergie libre du duplexe miARN/miARN-complémentaire-inversé. Si ce ratio est supérieur à 0.30, la solution est acceptée.

5.7 Appariement de la partie 3' du microARN

Un équilibre existe entre la force de liaison du seed et celle de la partie 3' du microARN. Un microARN ayant un seed avec liaison faible aura besoin d'une compensation de liaison dans sa partie 3', et vice-versa. Une étude démontre une association entre l'intensité de la répression du microARN et le nombre de nucléotides pairés dans la région 3' [57]. Les positions des nucléotides visés par cette étude sont les 4-mer partant de la position 11 et allant jusqu'à la position 15 à partir de la partie 5' du microARN (figure 5.8). Cette contrainte est stricte et élimine énormément de solutions, donc l'algorithme permet de forcer un appariement aux positions 12, 13 et 14 ou de forcer 5 appariements parmi les 10 derniers nucléotides du microARN.

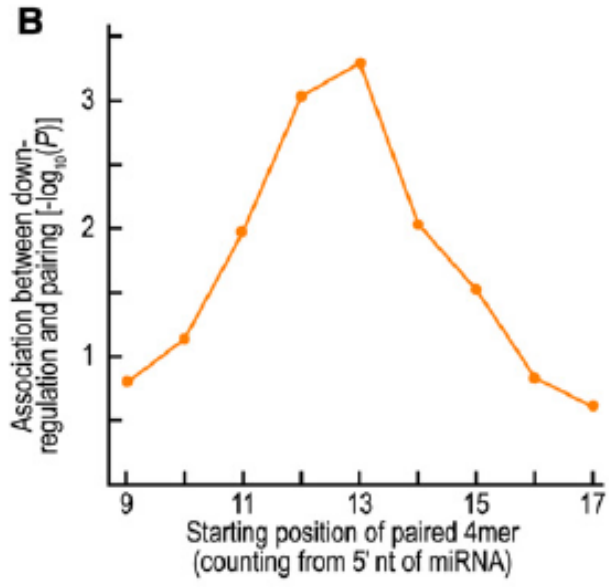

Figure 5.8: Association entre l'appariement de bases dans la partie 3' du microARN et la régulation de l'expression. Les 4-mer débutant à la position 11 du microARN et s'échelonnant jusqu'à la position 15 ont la plus forte association avec la régulation de l'expression.

5.8 Structure des sites ciblés par les microARNs

Les microARNs sont recrutés à leur site d'intérêt à l'aide du complexe protéique RISC, une molécule composée de nombreuses protéines actives. Il est intuitif de penser que ce complexe doit avoir le champ libre dans la région entourant les sites ciblés afin de bien pouvoir transporter les microARNs. Une étude a démontré que les régions entourant les sites ciblés doivent avoir une structure ouverte afin de favoriser l'introduction du complexe protéique RISC [58]. Cette même étude affirme que la structure locale sur laquelle se lie les microARNs (17 nucléotides en amont et 13 en aval, ainsi que le site de liaison) doit être accessible afin de promouvoir l'hybridation des microARNs aux ARNs messager. Il semble que le désappariement de la région cible suivi de l'appariement du microARN requiert plus d'énergie que de simplement apparier le microARN. Certains affirment que l'accessibilité joue un rôle aussi important que la reconnaissance de la séquence dans l'activité des microARNs [58]. .

5.8.1 Structure en amont et en aval des sites ciblés par les microARNs

Des régions ayant une structure ouverte plutôt que refermée vont favoriser l'accès au complexe protéique RISC. Deux approches peuvent être utilisées : obtenir une structure 2D globale de l'ARNm et analyser la section d'intérêt séparément, ou obtenir une structure 2D locale. En général, plus la séquence de l'ARN est grande, plus la précision des algorithmes de repliement de séquences diminue. Or, la représentation globale de la structure peut se rapprocher davantage de celle retrouvée dans les cellules, comparativement à une représentation locale. D'un autre côté, un repliement local est plus rapide qu'un repliement global. Afin de pouvoir décider de l'approche à utiliser, les deux méthodes ont été évaluées par rapport aux sites endogènes de liaison de mir-20 sur les gènes E2F1, E2F2 et E2F3 en utilisant RNAfold. L'accessibilité (nombre de nucléotides non-pairés) d'une fenêtre de 70 nucléotides a été étudiée pour l'approche globale et locale (repliement du ARN messager entier versus repliement de la partie 3' UTR seulement ; voir figure 5.9). Si l'on considère que ces sites endogènes doivent avoir des structures favorables au recrutement du complexe RISC, alors l'accessibilité devrait être haute. Notons que les 70 nucléotides étudiés dans la figure 5.9 recoupent avec les sites ciblés par mir-20.

Étant donné que le repliement local obtient une majorité de bons résultats, et que ce repliement est moins coûteux à exécuter, il a été choisi dans l'algorithme. L'implantation finale de l'évaluation de la structure en amont et en aval du site ciblé consiste, pour chaque site potentiel, à obtenir l'énergie libre des 70 nucléotides chevauchant à droite et à gauche le site cible (figure 5.10).

Ainsi, une énergie libre proche de 0 sera privilégiée à une énergie hautement négative, car l'on recherche des structures locales en désordre (non appariées).

Accessibilité des sites de liaisons de mir-20 sur les gènes E2F1
E2F2 et E2F3

	Index	Global	3'UTR	Local
E2F1	393	0.27	0.51	0.46
	986	0.67	0.49	0.77
E2F2	904	0.43	0.32	0.37
	1518	0.19	0.18	0.30
E2F3	1822	0.30	0.45	0.51

Figure 5.9: Accessibilité des sites endogènes de mir-20 des gènes E2F1, E2F2 et E2F3. L'accessibilité est calculée comme étant le nombre de nucléotides non pairés sur un total de 70. Les régions recoupent le site de liaison de mir-20. L'index commence à partir de la première position des séquences 3'UTR. Plus le résultat se rapproche de 1, plus l'accesibilité est grande. L'approche locale obtient trois résultats optimaux sur cinq, tandis que les approches globales pour la séquence complète et pour la séquence 3'UTR obtiennent chacune un résultat optimal.

5.8.2 Structure des sites ciblés par les microARNs

En plus de considérer les structures avoisinantes aux sites ciblés par les microARNs, il faut aussi porter attention à la structure du site lui-même. L'énergie requise pour briser l'hybridation d'un duplexe ARNm/ARNm pour ensuite y lier un microARN est supérieure à l'énergie requise pour lier un microARN. Les structures des sites cibles ayant le nombre le plus élevé de nucléotides non appariés seront privilégiées. Un site cible est défini comme étant les 17 nucléotides en amont du site, le site lui-même (32

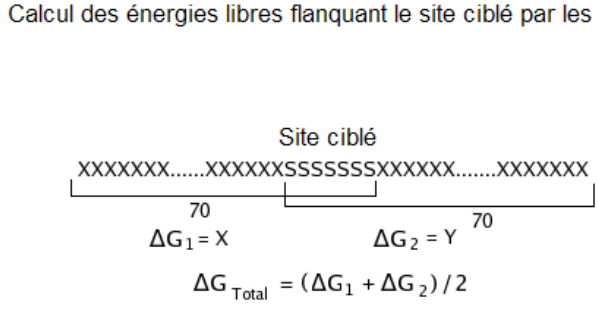

Figure 5.10: Calcul de l'énergie libre des 70 nucléotides en amont et en aval qui chevauchent le site ciblé par les microARNs. L'énergie libre totale est la moyenne des énergies libres.

nucléotides) et les 13 nucléotides en aval du site [58]. Afin de bien représenter la structure de cette région, un sous-ensemble de structures est généré par RNAsubopt, un programme de la suite ViennaRNA qui retournent des structures 2D sous-optimales. Ce sous-ensemble contient la structure optimale ainsi qu'un nombre prédéterminé de structures sous-optimales. Le résultat de l'accessibilité du site cible est donc la moyenne des résultats pour chaque structure obtenue (Figure 5.11).

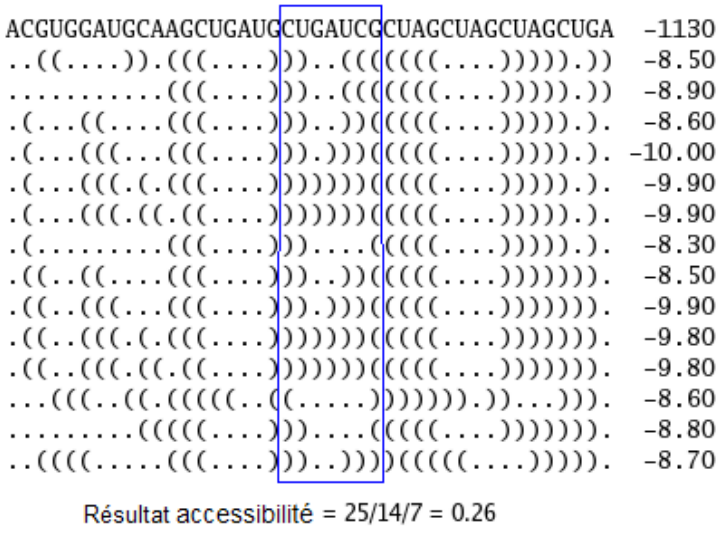

Figure 5.11: Exemple de calcul de l'accessibilité pour un site ciblé par les microARNs. Représentation d'un ensemble de 14 structures ayant un site cible de 7 nucléotides. L'accessibilité est calculée en fonction du nombre de nucléotides non appariés (25) divisé par le nombre de structures (14) divisé par la longueur du site cible (7). Le nombre réel de structures varie entre 100 et 300 pour une séquence de 62 nucléotides (le site cible) et la longueur du site cible est la structure entière (62 nucléotides).

6. Implantation : MultiTar V1.0

Les propriétés des microARNs et leurs cibles déterminées à la section 5 servent de base à l'implantation de l'algorithme qui a pour but de générer un microARN artificiel ciblant un nombre X de gènes. MultiTar V1.0 est l'implantation de cet algorithme à travers une interface web qui permet à quiconque de pouvoir lancer ses propres analyses. Ce programme comprend quatre parties principales dont l'interface usager, l'algorithme de recherche du seed optimal, l'algorithme de génération de la séquence 3' du microARN et le formatage des résultats. Le programme est actuellement hébergé sur un ordinateur de l'IRIC (Institut de recherche en immunologie et cancérologie) à l'adresse suivante : www.maj03.iric.ca/c̃aronmax/cgi-bin/launch.cgi.

Multitar V1.0 utilise principalement les langages de programmation Java, Perl et CGI ainsi que certains programmes externes de la suite Vienna RNA package codés en C. Le serveur web est Apache roulant sous Linux. Chaque exécution est stockée sur le disque dur en prenant soin de répertorier la date et l'heure d'exécution ainsi que l'adresse IP de l'exécuteur. Ceci permet de retrouver facilement les fichiers générés par le programme en cas d'analyses ultérieures. Exemple : runs/Wed-Aug-20-2008-heure-24.203.208.131.

6.1 Pipeline de MultiTar

La figure 6.1 décrit les grandes étapes de l'exécution de MultiTar, les principales fonctions de chacune ainsi que les programmes associés. Le tout débute avec l'interface

usager qui prend en entrée les gènes d'intérêt, suivi de l'algorithme itératif pour trouver le seed le plus prometteur, suivi de l'heuristique recherche tabou qui s'occupe de générer la partie 3' restante du microARN artificiel et le tout se termine avec le formatage des résultats. La sortie des résultats comprend le microARN le plus probable selon nos critères ainsi que les alignements optimaux de ce microARN avec chacun des gènes.

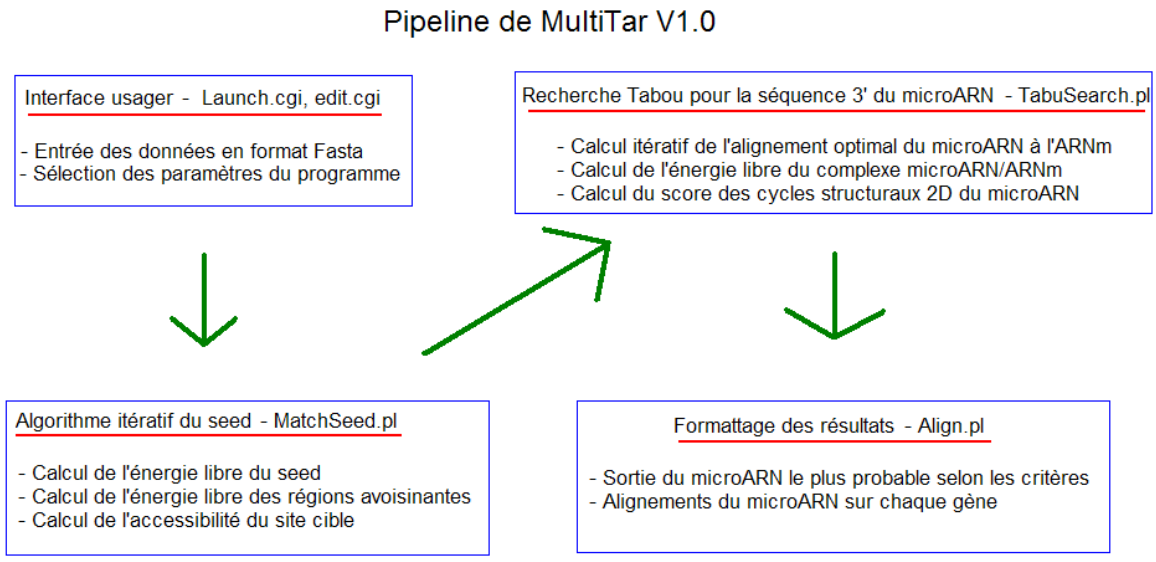

Figure 6.1: Étapes principales du programme MultiTar V1.0 L'illustration des étapes principales de MultiTar : l'interface usager, algorithme itératif du seed, génération de la séquence 3' du microARN artificiel et formatage des résultats.

6.2 Interface usager - Launch.cgi, edit.cgi

L'interface usager est la première étape du programme MultiTar (figure 6.2). L'usager doit entrer en format fasta les séquences 3'UTR des gènes d'intérêt. L'interface comprend de nombreuses options allant des options pour le seed, des options pour la séquence 3' du microARN ainsi que des options générales. Les options du seed comprennent l'énergie libre, le nombre de mésappariements, et la distribution des résultats pour la fonction de pondération (section 6.3.1). Les options de la partie 3' du microARN comprennent le choix de forcer un appariement aux positions 12,13,14 ou de forcer un

appariement sur 5 des 10 derniers nucléotides. Finalement, les options générales permettent de spécifier le nombre de sites par gènes, c'est-à-dire que l'on peut forcer un microARN à se lier à exactement 1, 2 ou 3 cibles par gènes. Il est aussi possible de spécifier le nombre de solutions, donc si l'on spécifie plus qu'une solution, celles qui vont sortir seront des solutions sous-optimales (les options complètes sont illustrées en Annexe).

Figure 6.2: Interface web de MultiTar V1.0 Interface usager de MultiTar V1.0. L'utilisateur doit entrer les séquences 3' UTR des gènes d'intérêt en format FASTA.

6.3 Algorithme itératif de recherche du seed optimal - MatchSeed.pl

La première étape de l'algorithme consiste à trouver le seed optimal pour chacune des séquences 3'UTR de gènes. Si l'utilisateur requiert d'avoir un seed avec un match parfait, chacun des 7-mer de la séquence la plus petite sera comparé itérativement à chacune des autres séquences. Si l'on permet un ou deux mésappariements, chacun des 7-mer possibles (16384) sera comparé à chacune des séquences. Lorsqu'un seed

satisfait la contrainte d'appariement ainsi que la contrainte de sélection de brin, il est itérativement recalculé pour chaque séquence en tenant compte de son énergie libre, de l'énergie libre de ses régions avoisinantes ainsi que l'accessibilité de son site cible. Ceci est répété pour chacun des seeds potentiels et les meilleurs seeds sont retournés comme solutions finales.

6.3.1 Fonction de résultat des seeds

La fonction de résultat est déterminée de la façon suivante :

Résultat total = A x [Résultat de l'énergie libre du seed] + B x [Résultat de l'énergie libre des régions avoisinantes] + C x [Résultat de l'accessibilité du site cible]

où A est le poids accordé à l'importance de l'énergie libre du seed, B est le poids accordé à l'importance de l'énergie libre des régions avoisinantes et C est le poids accordé à l'importance du résultat d'accessibilité. Il faut que A+B+C = 1. Étant donné que l'importance de chacune des composantes est difficile à déterminer a priori, les constantes suivantes sont implantées par défaut dans l'algorithme : A = 0.30, B = 0.35 et C = 0.35.

Le résultat de l'énergie libre du seed est obtenu en exécutant RNAcofold (un programme de la suite ViennaRNA qui retourne l'énergie libre d'un duplexe ARN :ARN) sur le duplexe d'ARN qui comprend la séquence du seed et la région de l'ARN messager à l'itération courante. Ce résultat est divisé par l'énergie libre minimale possible d'un seed, déterminée empiriquement (-16 kcal/mol). Plus le résultat se rapproche de 1, plus le seed a une énergie libre favorable.

Le résultat de l'énergie libre des régions avoisinantes est obtenu en exécutant RNAfold (un programme de la suite ViennaRNA qui retourne la structure 2D optimale

d'une séquence d'ARN) sur les 70 nucléotides en amont et en aval chevauchant le site cible. Chacune des deux énergies libres est divisée par l'énergie libre minimale déterminée empiriquement (-40kcal/mol). Ensuite, la moyenne des deux énergies est retenue et l'inverse de cette dernière détermine le résultat final de l'énergie libre des régions avoisinantes. Dans ce cas-ci, contrairement à l'énergie libre du seed, nous voulons obtenir des énergies libres proches de 0 afin de privilégier des régions ayant des structures secondaires en désordre (ouvertes). Plus le résultat de rapproche de 1, plus les régions avoisinantes sont désordonnées.

En dernier lieu, le résultat de l'accessibilité est obtenu en exécutant RNAsubopt (un programme de la suite ViennaRNA qui retournent des structures 2D sous-optimales) sur une région de 62 nucléotides englobant le site cible (17 nucléotides en amont, le site cible de 32 nucléotides ainsi que 13 nucléotides en aval). RNAsubopt est exécuté avec l'option -e 3, c'est à dire en obtenant toutes les solutions sous-optimales jusqu'à 3 kcal/mol de la solution optimale. Le résultat est obtenu en divisant le nombre de nucléotides non pairés par 62 (longueur de la région) et ensuite divisé par le nombre de solutions retournées. Plus le résultat se rapproche de 1, plus l'accessibilité du site cible est favorable.

Pour conclure, le résultat final du seed est intéressant lorsqu'il s'approche de 1.

La figure 6.3 montre le résultat du seed optimal sans aucun mésappariement des séquences 3'UTR des gènes E2F1, E2F2 et E2F3. Il se peut qu'un résultat supérieur soit retourné si l'utilisateur permet des mésappariements dans le seed. Dans la section suivante, nous présentons le pseudo-code de cet algorithme itératif de recherche du seed optimal.

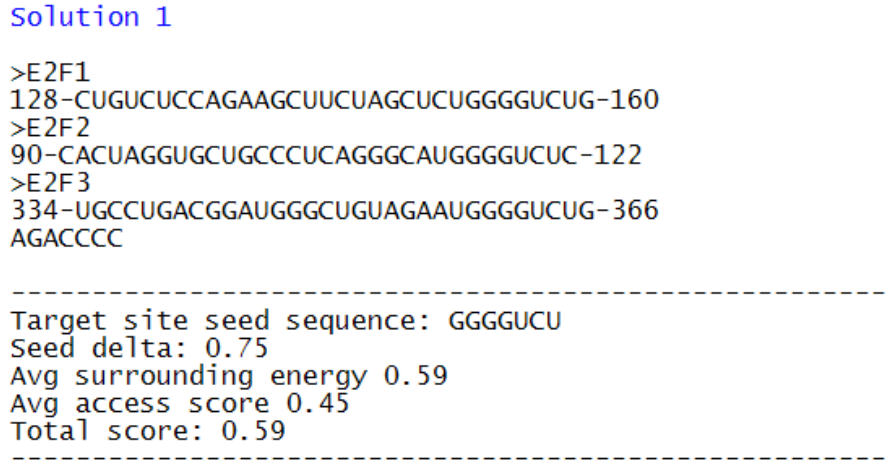

Figure 6.3: Exemple d'un résultat retourné pour un seed avec aucun mésappariement sur les séquences 3'UTR des gènes E2 F1, E2F2 et E2F3. Le seed avec match parfait a la séquence 'GGGGUCU'. Le résultat de l'énergie libre du seed est de 0.75, le résultat de l'énergie libre des régions avoisinantes est de 0.59 et le résultat de l'accessibilité du site ciblé est de 0.45. Le résultat total est de 0.59.

6.3.2 Pseudocode

```
Procedure Algorithme_iteratif(sequences, contraintes,
ponderation_seed, ponderation_envir, ponderation_acces) {

FAIRE

 Initialiser solution_optimale
 Initialiser liste_seeds <- generer_7mer(A,C,G,T)
 resultat_solution_optimale <- 0
 y <- 0

 TANT QUE y < retourner_longueur_liste(liste_seeds)

  FAIRE
   Initialiser liste_resultat
   p <- 0

   TANT QUE p < retourner_longueur_liste(sequences)

    FAIRE
     meilleur_resultat_sequence <- 0
     i <- 0

     TANT QUE i < retourner_longueur(sequences[p])

      FAIRE
       continuer <- 0
       continuer <- passer_contraintes(liste_seeds[y], sequences[p],
       i, contraintes)
       SI continuer == 1
        FAIRE
         a <- calcul_energie(liste_seeds[y], sequences[p], i, ponderation_seed)
         b <- calcul_environs(liste_seeds[y], sequence[p], i, ponderation_envir)
         c <- calcul_acces(liste_seeds[y], sequence[p], i, ponderation_acces)
         SI a + b + c > meilleur_resultat_sequence
          FAIRE
           meilleur_resultat_sequence <- a + b + c
           liste_resultat[p] <- meilleur_resultat_sequence
          FIN FAIRE
         FIN SI
        FIN FAIRE
       FIN SI
      FIN FAIRE
      i <- i + 1
```

```
     FIN TANT QUE

    FIN FAIRE
    p <- p + 1

   FIN TANT QUE

   meilleure_solution_seed <- 0
   m <- 0

   TANT QUE m < retourner_longueur_liste(sequences)

    meilleure_solution_seed <- meilleure_solution_seed + liste_resultat[m]

   FIN TANT QUE

   SI meilleure_solution_seed > resultat_solution_optimale
    FAIRE
     resultat_solution_optimale <- meilleure_solution_seed
     solution_optimale <- liste_seed[y]
    FIN FAIRE
   FIN SI
  FIN FAIRE
  y <- y + 1

 FIN TANT QUE

 RETOURNER solution_optimale, resultat_solution_optimale

FIN FAIRE


}
```

La procédure Algorithme_iteratif utilise les procédures suivantes :

1. Procédure generer_7mer : procédure qui génère toutes les séquences de 7 nucléotides possibles ;
2. Procédure retourner_longueur_liste : procédure qui retourne la taille d'une liste ;
3. Procédure retourner_longueur : procédure qui retourne la longueur d'une séquence ;
4. Procédure passer_contraintes : procédure qui retourne une valeur booléene ; 1 si le seed satisfait les contraintes de l'algorithme, 0 sinon ;
5. Procédure calcul_energie : procédure qui retourne le résultat de l'énergie du seed ;

6. Procédure calcul_environs : procédure qui retourne le résultat des régions avoisinantes au seed ;
7. Procédure calcul_acces : procédure qui retourne le résultat de l'accessibilité du seed.

6.3.3 Complexité de l'algorithme itératif

La complexité de l'algorithme itératif de recherche du seed optimal doit tenir compte du nombre de séquences 3'UTR (données en entrée), de la longueur de celles-ci, du nombre d'itérations de l'algorithme (nombre de seeds) et du temps de calcul de chaque seed. Soit M le nombre de séquences 3'UTR, N la longueur de la plus grande séquence, K le nombre de seeds (pour simplification, considérons toujours un mésappariement, donc 16384 seeds) et C le temps de calcul pour chaque seed (calcul de l'énergie libre du seed et des régions avoisinantes, ainsi que le score d'accessibilité). La complexité totale dans le pire cas est de O(MNKC).

6.4 Détermination de la région 3' du microARN - TabuSearch.pl

Une fois que le seed optimal est trouvé, il reste 14 nucléotides à générer afin de former le microARN entier. Ceci peut sembler trivial, mais le nombre possible d'alignements d'un microARN artificiel avec un ensemble de séquences de gènes fait en sorte qu'un nombre préalable de solutions ne peut être déterminé facilement.

6.4.1 Espace de solutions

On peut être tenté de déterminer a priori certaines séquences de 14 nucléotides qui semblent plus favorables que d'autres, mais le nombre d'alignements possibles rend cette tâche difficile. Considérons une séquence cible et un microARN artificiel. Ce der-

nier a 22 nucléotides et la séquence cible peut avoir une longueur indéterminée. Dans cet algorithme, la longueur de la séquence cible varie de 22 nucléotides à 32 nucléotides, donc 11 alignements possibles (Figure 6.3). L'élimination de certaines solutions sans observer l'effet de celles-ci sur les alignements est une tâche non triviale, et ceci est d'autant plus vrai lorsque le nombre de séquences cibles est grand. Il faut donc considérer toutes les solutions possibles, ce qui se chiffre à environ 268 millions de solutions, un nombre élevé. Dans un problème comme celui-ci, une métaheuristique est une méthode permettant de survoler un sous-espace de solutions de façon efficace. La métaheuristique choisie est la recherche tabou. Il est suggéré de partir avec une bonne solution initiale afin de bien entamer la recherche de solutions. La solution initiale est la séquence consensus complémentaire des 14 nucléotides situés après le seed (régions retournées par la recherche itérative du seed). Chaque nucléotide retenu pour chacune des 14 positions sera celui qui apparaît le plus souvent pour la position donnée. Afin de simuler l'appariement d'un duplexe d'ARN, le complémentaire de ces nucléotides sera utilisé.

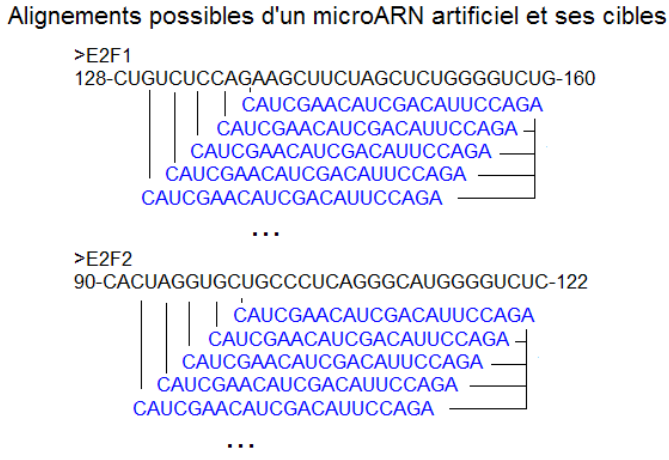

Figure 6.4: Alignements d'un microARN artificiel avec ses cibles. Le microARN artificiel s'aligne sur 22 à 32 nucléotides par cible.

6.4.2 Fonction d'évaluation des solutions

Une métaheuristique a besoin d'une fonction d'évaluation des solutions afin de pouvoir différencier les bonnes solutions des mauvaises et ainsi guider son choix dans l'exploration des solutions. La fonction d'évaluation consiste à analyser le résultat du meilleur alignement parmi les 11, pour chaque séquence de gènes. Le meilleur alignement comporte deux aspects, soit le ratio de l'énergie libre de l'alignement et le résultat des cycles structuraux 2D. Plus précisément :

résultat meilleur alignement = A x [Ratio de l'énergie libre de l'alignement] + B x [résultat des cycles structuraux 2D]

où A est le poids accordé à l'importance du ratio de l'énergie libre de l'alignement et B est le poids accordé à l'importance du résultat des cycles structuraux 2D, avec A+B = 1. Les valeurs de A = 0.7 et B = 0.3 ont été déterminées par inspection visuelle des alignements retournés par l'algorithme.

Le résultat de l'énergie libre de l'alignement versus l'énergie libre du complémentaire inversé est calculé comme décrit à la section 5.6

Le résultat des cycles structuraux est calculé en prenant la moyenne des résultats des cycles contenus dans l'alignement divisé par le résultat maximal du cycle le plus fréquent (0.274).

Deux contraintes supplémentaires doivent être satisfaites afin d'obtenir une solution valide. Selon l'option sélectionnée par l'utilisateur, les nucléotides 12,13 et 14 ou 5 des 10 derniers nucléotides du microARN doivent être pairés. De plus, seuls les duplexes classiques sont acceptés, c'est-à-dire aucune boucle tige parmi les alignements n'est tolérée.

De façon similaire au résultat de la recherche itérative du seed, nous voulons que le résultat se rapproche de 1.

6.4.3 Solutions avoisinantes

Afin d'explorer d'autres solutions potentielles, une classe de transformation permettant de générer des solutions avoisinantes doit être établie. Les solutions avoisinantes sont générées en mutant un des 14 nucléotides de la partie 3' du microARN par un nucléotide différent.

6.4.4 Liste tabou

Une des caractéristiques de la recherche taboue est l'implantation d'une liste taboue. Cette liste permet de créer des règles visant à ne pas considérer des solutions avoisinantes qui ont été préalablement explorées (mémoire court terme). Basée sur [45], la taille de la liste tabou est de 12. Une solution est donc dite tabou si celle-ci a été préalablement visitée dans les 12 itérations précédentes. Une transformation générant un voisin est entrée dans la liste tabou sous forme de l'inverse de la mutation, c'est-à-dire si la position 8 a été mutée de C à G, alors la mutation de la position 8 de G à C sera tabou.

6.4.5 Intensification

L'intensification est une procédure dans la recherche tabou qui permet d'augmenter le nombre de voisins lorsque l'on se trouve dans une région de solutions prometteuses. L'algorithme considère que lorsque 3 itérations consécutives donnent une meilleure solution que celle trouvée à l'itération précédente, le nombre de voisins est augmenté de 15

à 30. Ceci permet d'explorer davantage les régions contenant des solutions intéressantes.

6.4.6 Diversification

La diversification permet de s'échapper d'un minimum ou maximum local lorsqu'aucune nouvelle meilleure solution n'est trouvée pour un nombre prédéterminé d'itérations. Après 5 itérations non fructueuses, une procédure de diversification est enclenchée par l'algorithme, qui consiste à générer une solution complètement aléatoire et de repartir la recherche à partir de cette dernière.

6.4.7 Critère d'aspiration

Le critère d'aspiration est un aspect de la recherche tabou qui permet de lever l'interdiction de la liste tabou si la solution avoisante répond à un critère bien précis. Dans le cas de notre algorithme, si la solution avoisinante obtient un résultat supérieur au résultat optimal de la recherche, l'interdiction de la liste taboue est levée et la recherche peut continuer avec cette nouvelle solution.

6.4.8 Paramètres de la Recherche Tabou

La recherche tabou contient plusieurs paramètres, dont le nombre d'itérations de la recherche, le nombre de voisins, le nombre d'itérations avant l'intensification, le nombre d'itérations avant la diversification, la taille de la liste tabou ainsi que le nombre de voisins en mode intensification. La figure 6.5 présente les valeurs des paramètres utilisées.

Dans la section suivante, nous présentons le pseudo-code de l'algorithme de la recherche tabou.

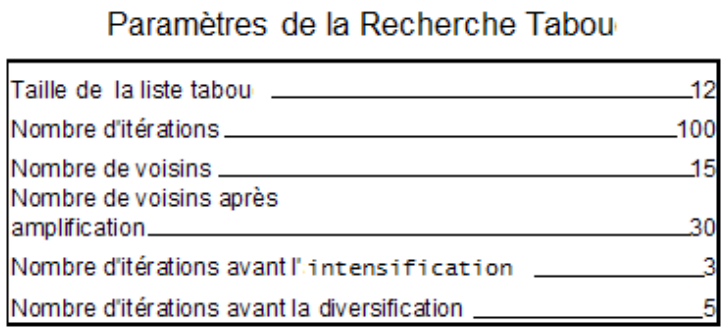

Figure 6.5: Les paramètres de la recherche tabou. La recherche tabou contient de nombreux paramètres, dont le nombre d'itérations, le nombre de voisins, etc.

6.4.9 Pseudocode

```
Procedure Recherche_tabou(solution_initiale,
iterations_tabou, nombre_voisins_standard,
nombre_voisins_intensification, parametre_intensification,
parametre_diversification, taille_liste_tabou,
fonction_evaluation, classe_transformation) {

FAIRE

 Initialiser solution_optimale
 Initialiser liste_tabou
 resultat_solution_optimale <-0
 intensification <- 0
 diversification <- 0
 nombre_voisins <- nombre_voisins_standard
 i <- 0

 TANT QUE i < iterations_tabou

  Initialiser liste_voisins
  Initialiser liste_resultats_voisins

  FAIRE
   SI intensification == parametre_amplification
    FAIRE
     nombre_voisins <- nombre_voisins_intensification
     intensification <- 0
    FIN FAIRE
   FIN SI

   SI diversification == parametre_diversification
    FAIRE
```

```
  solution_initiale <- generer_diversification(solution_initiale)
  diversification <- 0
 FIN FAIRE
FIN SI

liste_voisins = generer_voisins(solution_initiale,
nombre_voisins, liste_tabou, classe_transformation)

liste_resultats_voisins = evaluer_voisins(liste_voisins,
fonction_evaluation)

trouver_optimale <- 0
y <- 0

TANT QUE y < nombre_voisins

 SI liste_resultats_voisins[y] < resultat_solution_optimale
  FAIRE

   solution_initiale <- liste_voisins[y]
   solution_optimale <- liste_voisins[y]
   resultat_solution_optimale <- liste_resultats_voisins[y]
   intensification <- intensification + 1
   diversification <- 0
   actualiser_liste_tabou(liste_tabou, taille_liste_tabou,
   liste_voisins[y])
   trouver_optimale <- 1

  FIN FAIRE
 FIN SI
 y <- y + 1

FIN TANT QUE

 SI trouver_optimale == 0
  FAIRE
   resultat_solutions_voisins <- 0
   y <- 0

   TANT QUE y < nombre_voisins

    SI liste_resultats_voisins[y] < resultat_solutions_voisins
     FAIRE
      solution_initiale <- liste_voisins[y]
      diversification <- diversification + 1
      actualiser_liste_tabou(liste_tabou, taille_liste_tabou,
      liste_voisins[y])
```

```
        FIN FAIRE
       FIN SI
       y <- y + 1

      FIN TANT QUE

     FIN FAIRE
    FIN SI

    nombre_voisins <- nombre_voisins_standard

   FIN FAIRE

   i <- i + 1

  FIN TANT QUE

 RETOURNER solution_optimale, resultat_solution_optimale

FIN FAIRE

}
```

La procédure Recherche_tabou utilise les procédures suivantes :

1. Procédure generer_diversification : procédure qui génère une solution diversifiée ;
2. Procédure generer_voisins : procédure qui génère les solutions voisines ;
3. Procédure evaluer_voisins : procédure qui retourne les résultats de l'évaluation des voisins ;
4. Procédure actualiser_liste_tabou : procédure qui met à jour la liste tabou.

6.4.10 Complexité de la Recherche Tabou

La complexité de la recherche tabou dépend du nombre de séquences 3'UTR, du nombre de cibles par séquence, du nombre d'itérations de l'algorithme, du nombre de voisins par itération, du nombre d'alignements par séquence (11) et du temps d'exécution de chaque alignement (calcul du ratio de l'énergie libre de l'alignement et calcul du score des cycles structuraux 2D). Soit M le nombre de séquences, K le nombre de cibles par séquences (1,2 ou 3), T le nombre d'itérations de la recherche (100), V le

nombre de voisins (afin de simplifier le calcul, considérer le nombre de voisins en mode intensification : 30), H le nombre d'alignements par séquence (11) et R le temps d'exécution de chaque alignement, la complexité totale dans le pire cas est de O(MKTVHR). La complexité est multiplicative car pour chaque séquence, il faut évaluer chacune des cibles, pour chacune des iterations de l'algorithme, ainsi de suite.

6.5 Sortie des résultats - Align.pl

La dernière étape de MultiTar consiste à produire en sortie la séquence du microARN artificiel le plus probable, le résultat de la recherche tabou ainsi que l'alignement du microARN avec chacune des séquences 3'UTR des gènes spécifiés en entrée. Les alignements sont obtenus par RNAcofold et sont transformés pour pouvoir être facilement interprétés. La figure 6.6 montre le résultat et les alignements du microARN obtenus après avoir fait une recherche tabou sur le seed trouvé à la figure 6.3.

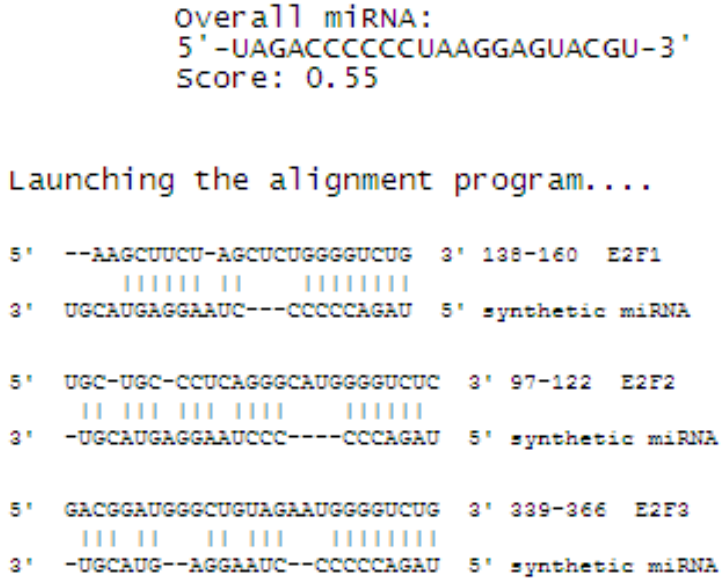

Figure 6.6: Exemple de sortie d'une solution de MultiTar. Sortie d'une recherche tabou exécutée sur le seed 'GGGGUCU'. Le résultat de la recherche tabou est de 0.55. Les trois alignements du microARN sur les régions 3'UTR des gènes E2F1,E2F2 et E2F3 sont illustrés.

6.6 Évaluation empirique du nombre de solutions pour un seed

Tel que mentionné précédemment, un seed est une séquence de 7 nucléotides qui peut être parfaitement appariée à sa séquence complémentaire, ou avoir un à deux mésappariements. Nous voulons étudier le nombre théorique d'appariements possibles pour un nombre variable de séquences, afin d'estimer l'utilité de l'approche pour des problèmes réels. Des séquences de longueur 500, 1000, 2000 et 3000 nucléotides ont été générées avec une fréquence uniforme (0.25 pour chaque nucléotide). Par la suite, les 16374 7-mer possibles ont été générés et comparés à un nombre de séquences variant entre 2 et 10. L'expérience a été répétée 10 fois et la moyenne des résultats a été calculée. Le nombre d'appariements parfaits ainsi que le nombre de mésappariements de un nucléotide retrouvés dans les séquences générées sont rapportés aux figures 6.7 et 6.8.

Lorsqu'on requiert un appariement parfait, le nombre de solutions possibles diminue rapidement lorsque le nombre de séquences augmente. À partir de 4 séquences, il y a très peu de seeds qui ont un appariement parfait. La longueur des séquences joue aussi un rôle dans le nombre de solutions possibles ; plus les séquences sont longues, plus elles permettent de trouver un appariement parfait. Lorsqu'on autorise un mésappariement, le nombre de solutions possibles est beaucoup plus grand et moins sensible au nombre et à la longueur des séquences. Pour 10 séquences ayant 1000 nucléotides, on peut espérer avoir environ 1000 solutions potentielles.

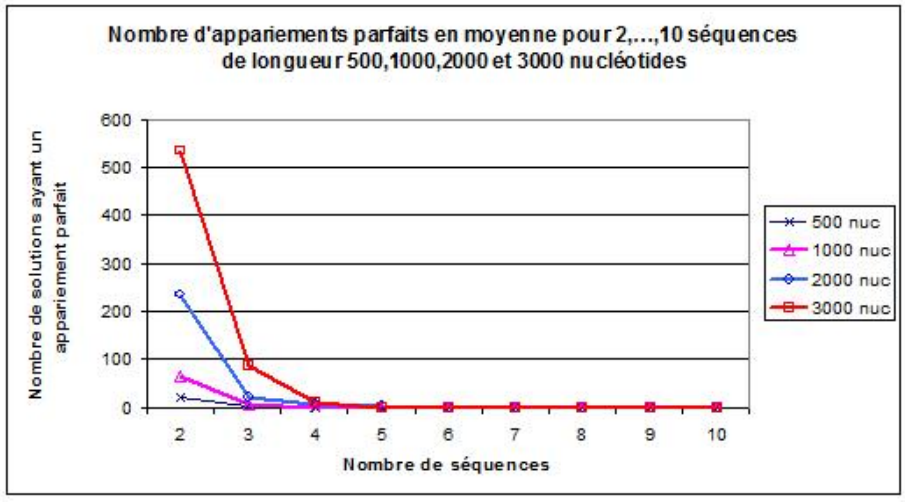

Figure 6.7: Nombre de solutions de seeds avec un appariement parfait. Nombre de solutions pour un seed lorsqu'un appariement parfait est requis. Le nombre de séquences varie de 2 à 10 et la taille des séquences est de longueur 500, 1000, 2000 et 3000 nucléotides.

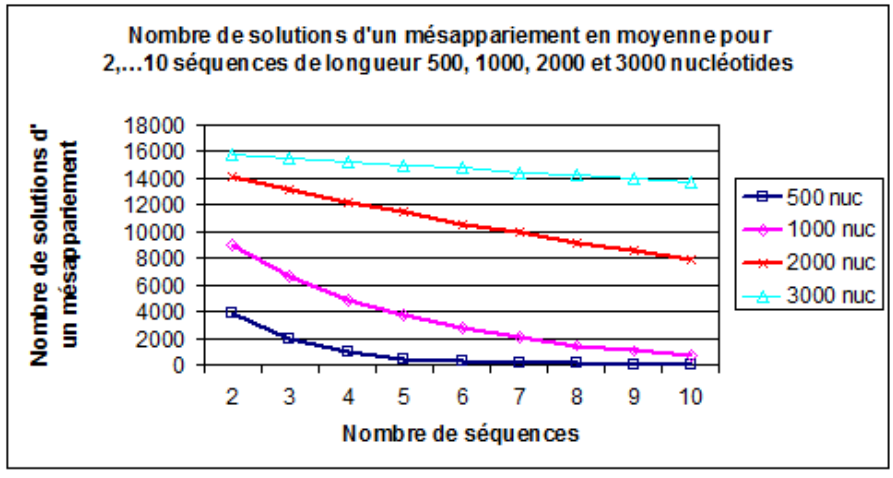

Figure 6.8: Nombre de solutions de seeds avec un mésappariement. Nombre de solutions pour un seed lorsqu'un mésappariement est alloué. Le nombre de séquences varie de 2 à 10 et la taille des séquences est de longueur 500, 1000, 2000 et 3000 nucléotides.

6.7 Validation in silico de miR-20 et miR-206

La validation in silico de microARNs endogènes permet de vérifier si les règles de l'algorithme ainsi que leurs pondérations représentent bien les observations réelles. Deux microARNs ont été choisis, soit miR-20 et miR-206. La figure 6.9 illustre le résultat des six sites de liaisons de miR-20, soit 0.52. Le résultat des meilleurs alignements de ces sites avec le microARN est de 0.4. Notons que la contrainte d'alignement dans la partie 3' du microARN fut levée pour obtenir ces solutions, une contrainte considérée sévère.

La figure 6.10 présente le résultat obtenu pour les trois sites de liaison de miR-206, soit 0.6. Le meilleur résultat pour l'alignement de ces sites avec le microARN est de 0.375. Pour ces deux microARNs endogènes, notre algorithme a pu trouver des solutions valides selon nos règles préétablies. Cela n'est pas surprenant étant donné que les règles sont basées sur des observations de microARNs endogènes. Chaque solution retournée par l'algorithme satisfait les contraintes de base et les résultats permettent de différencier les solutions ordinaires des solutions exceptionnelles. Le résultat maximal d'une solution est de 1. Ainsi, plus le résultat se rapproche de 1, plus la solution est exceptionnelle. Le résultat du seed et le résultat de la recherche tabou sont indépendants ; une interprétation subjective est donc requise. Nous pouvons accorder une importance égale à ces deux résultats, ou privilégier un ou l'autre.

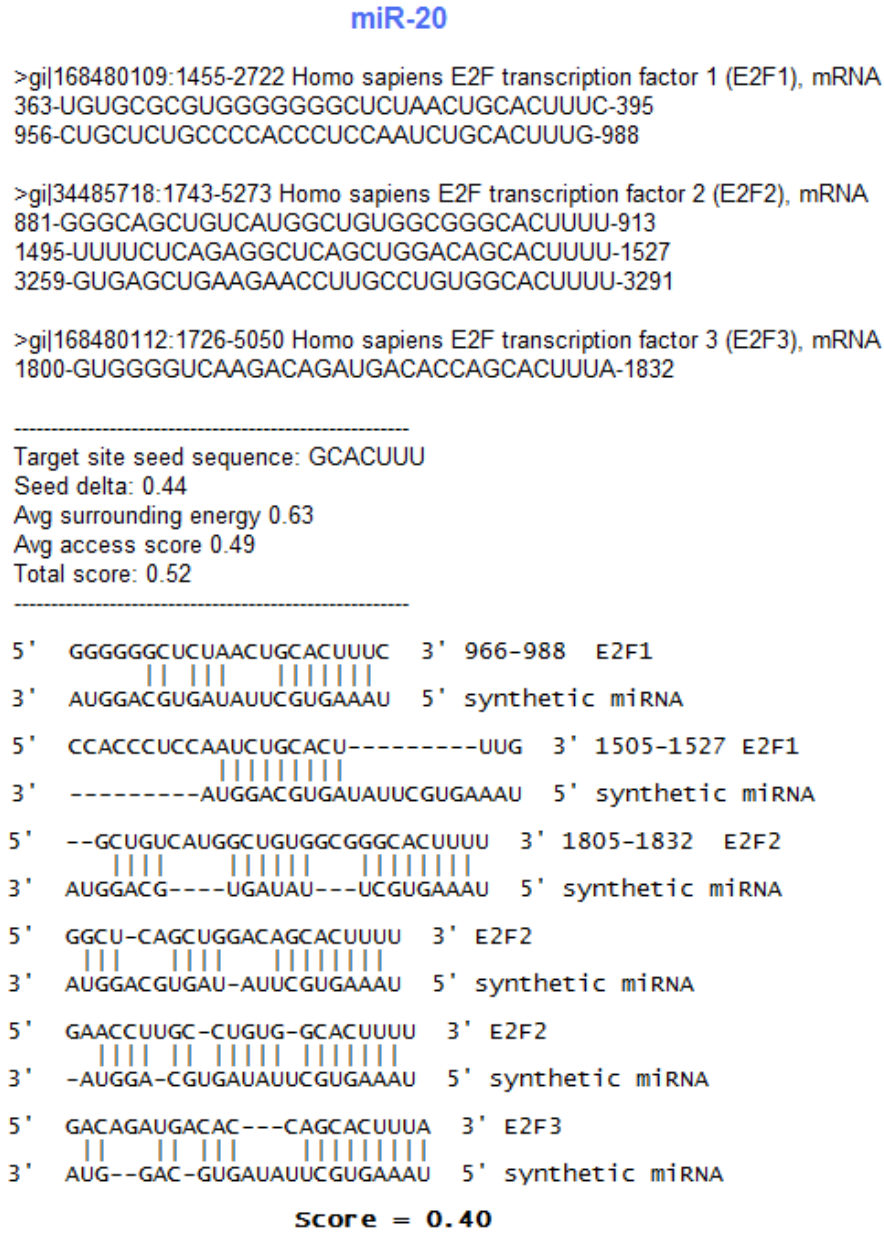

Figure 6.9: Résultat MultiTar des sites endogènes de miR-20. Le résultat de l'algorithme itératif du seed est de 0.52 et le résultat des meilleurs alignements est de 0.4.

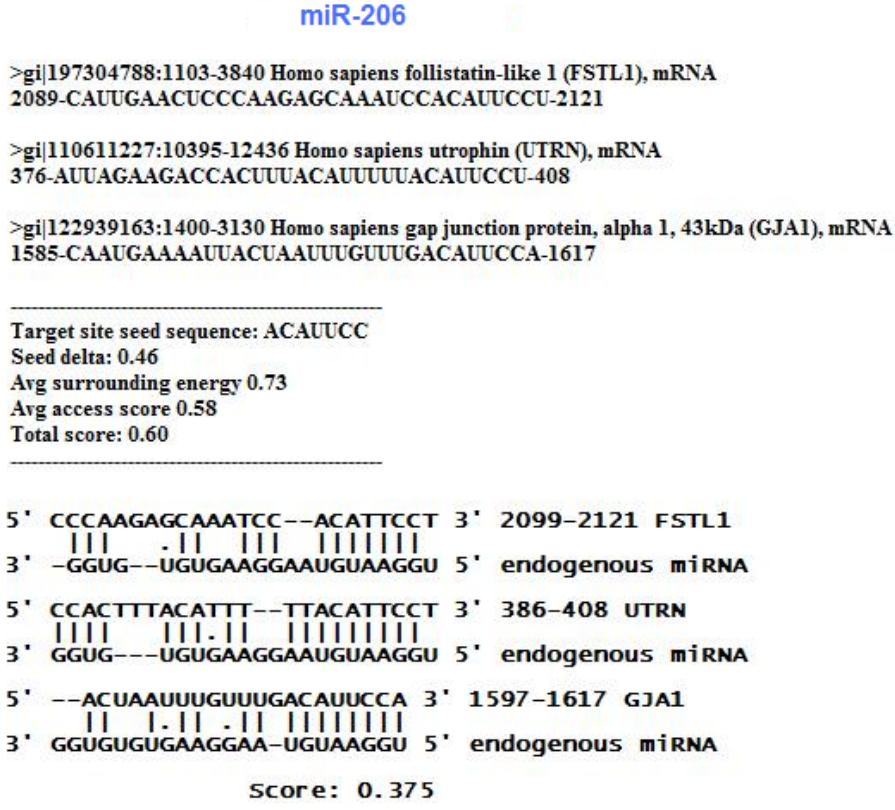

Figure 6.10: Résultat MultiTar des sites endogènes de miR-206. Le résultat de l'algorithme itératif du seed est de 0.6 et le résultat des meilleurs alignements est de 0.375.

7. Protocoles biochimiques

Les expériences biochimiques qui ont été faites pour investiguer l'activité des microARNs artificiels sont grandement utilisées dans le domaine, mais certaines particularités peuvent être propres au problème étudié.

7.1 Essais luciférase

Les séquences 3' UTR des gènes E2F* ont été greffées sur le gène de la luciférase et transfectées dans des cellules HeLa pour ensuite quantifier son expression. Les conditions de l'expérience sont les suivantes : 24 heures avant la transfection, les cellules HeLa ont été déposées dans des plaques de 25 puits. Le plasmide pGL3 contrôle contenant les 3'UTRs des différents E2Fs (5ng) a été transfecté avec pRL-globin (5ng) et avec la lipofectamine LTX (invitrogen) en suivant les directives de la compagnie. L'activité luciférase a été mesurée 24h après la transfection.

7.2 Western blots

Les fibroblastes normaux IMR90 (ATCC) ont été cultivés dans du milieu DMEM enrichi avec 10% de sérum de foetus de boeuf (FBS, Hyclone,USA) et avec 1% de pénicilline/streptomycine (GIBCO). Le transfert de gène par rétrovirus a été effectué comme rapporté précédemment [59]. Les IMR90 ont été infectés avec les rétrovirus exprimant les différentes constructions et incubés avec de la puromycine à une concentration de 2.5ug/ml pendant 48h afin d'éliminer les cellules non infectées. Ensuite, les cellules IMR90 exprimant les différentes constructions ont été trypsinisées et lavées une

fois au PBS pour être ensuite lysées avec 100ul de LAEMLI 6X, incubées sur glace pendant 5 minutes et chauffées pendant 5 minutes à 100 degrés celsius. 20 ug de protéines ont été déposées sur un gel SDS-PAGE avec une concentration d'acrylamide de 7.5% et transférées à une membrane Immobilion P (Millipore). Les anticorps suivants ont été utilisés pour la détection : anti E2F1 KH-95 (Santa Cruz 1 :1000), anti-alpha-tubuline (B-5-1-2 1 :5000) (Sigma) et anti MCM6. Le signal a été révélé après une incubation avec un anticorps secondaire anti-souris (1 :1500) couplé à la péroxidase (Amersham Biosciences) en utilisant du ECL (Amersham Biosciences).

7.3 Courbes de croissance

Les cellules préalablement infectées pour les western blots ont été collectées le lendemain de la fin de la sélection et déposées dans des plaques de 24 puits (Costat, Corning Inc., NY USA). Les cellules ont été comptées par incorporation de violet de cristal immédiatement après l'attachement (jour 0) et les jours 2, 5, 8 et 9. La sénescence a été évaluée en déterminant le pourcentage de la population démontrant une activité Béta-galactosidase comme décrit précédemment [60].

Troisième partie

Résultats

8. Résultats des microARNs artificiels

Afin d'évaluer l'efficacité des solutions produites par l'algorithme MultiTar, une validation biochimique a été choisie au lieu d'une validation informatique pour deux raisons principales. Premièrement, le but ultime du développement d'une méthode in-silico de conception de microARNs est de tenter d'observer des effets immédiats sur des modèles biologiques pré établis, et donc de supposer que ces effets pourraient s'appliquer à des situations médicales réelles. Le projet a été entrepris en tandem avec un laboratoire biochimique de l'Université de Montréal, et il a été possible de tester in vitro des microARNs artificiels. Deuxièmement, une validation informatique est difficile à interpréter ; les scores produits par l'algorithme se basent sur des règles biochimiques et des observations statistiques, mais il n'y a aucun moyen clair de vérifier l'efficacité des microARNs artificiels. La validation biochimique est composée de quatre volets : des essais luciférases, des western blots, des études sur la croissance cellulaire et des études sur la sénescence.

8.1 Données

L'algorithme a été exécuté sur les séquences 3'UTR des gènes E2F1 NM_005225, E2F2 NM_004091 et E2F3 NM_001949 de l'humain. La longueur de ces séquences est de 1268, 3531 et 3325 nucléotides respectivement. Ces trois gènes ont été choisis car il fut démontré préalablement que mir-20 est un microARN endogène qui cible

simultanément ces trois gènes, donc il constitue un contrôle positif parfait. Quatre solutions ont été choisies, soit deux solutions avec seulement un site cible par gènes, une solution avec deux sites cibles par gènes et une solution avec trois sites cibles par gènes. Il est sous-entendu dans la littérature que plusieurs sites de liaison peuvent augmenter l'inhibition de l'expression, donc nous avons choisi des microARNs qui ciblent plusieurs sites à la fois. Les options de MultiTar utilisées sont les options par défaut et en permettant deux mésappariements pour le seed. Les solutions retenues sont répertoriées dans la figure 8.1.

Abbréviation	Mode d'action	Séquence
MTE2F4	1 site cible	5' - UUUCCCUUUUCGCCCGGCCCU - 3'
MTE2F5	1 site cible	5' - UAUCUGACUUACGUGACUGCUU - 3'
MTMS1	2 sites cibles	5' - UAGUGGGGAGGGGGUUUCCGGU - 3'
MTMS2	3 sites cibles	5' - UAGUGGGGAUGUUUUUGGCAGG - 3'

Figure 8.1: Les solutions retenues ainsi que leurs séquences. MTE2F4 et MTE2F5 ont un site cible tandis que MTMS1 a deux sites cibles et MTMS2 trois sites cibles.

8.2 Essais luciférase

Les essais luciférase sont grandement utilisés dans le domaine des microARNs car ils constituent une méthode rapide et efficace de vérifier la régulation potentielle d'un microARN sur un ARNm. Puisque les microARNs inhibent l'expression de leurs cibles en se liant dans la région 3'UTR, les séquences 3'UTR de E2F1, E2F2 et E2F3 ont été clonés en aval du gène de la luciférase afin d'observer leurs niveaux d'expression. La figure 8.2 contient les résultats pour chacun des microARNs pour les gènes E2F1 et E2F3. Les essais luciférases pour E2F2 ne démontraient pas de diminution pour le contrôle mir-20 donc ils ne sont pas inclus.

Afin de vérifier si un microARN peut réguler de manière efficace chacun des gènes E2F1, E2F2 et E2F3, l'expérience a été répétée avec MTE2F5 (un site cible) et les résultats sont présentés aux figures 8.3, 8.4 et 8.5.

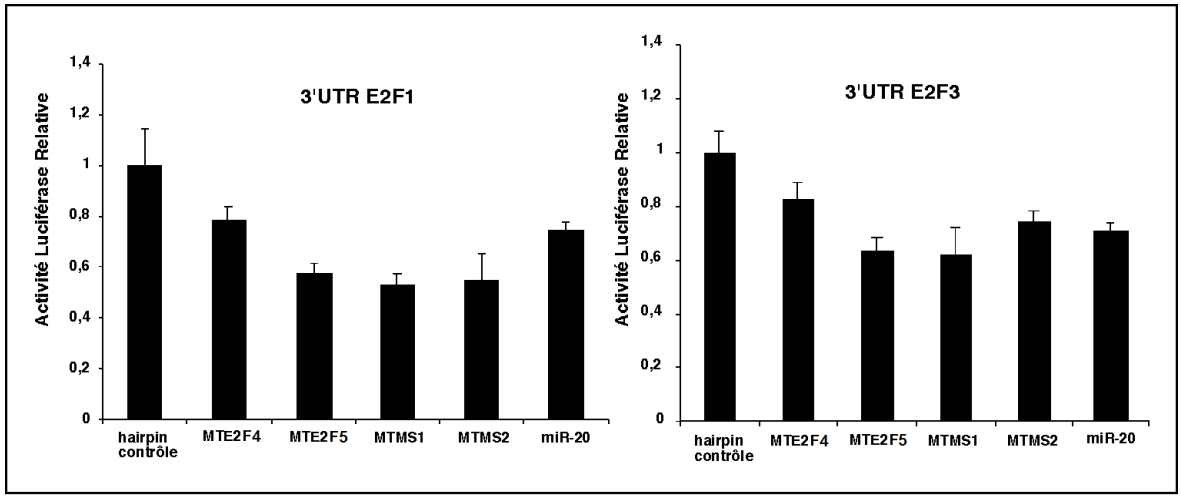

Figure 8.2: Les différents microARNs artificiels régulent l'expression de E2F1 et E2F3. Les constructions possédant les 3'UTRs de E2F1 ou E2F3 ont été co-transfectés dans les cellules HeLa avec les différents microARNs (MTE2F4, MTE2F5, MTMS1 et MTMS2), mir-20, le harpin contrôle ainsi que le contrôle de la transfection de la Renilla. La référence de l'expression de E2F1 et E2F3 est illustrée sous la barre hairpin contrôle.

Sur la figure 8.2, on aperçoit les taux d'expression des gènes E2F1 et E2F3 pour les quatre microARNs artificiels. Parmi eux, notons MTE2F4 et MTE2F5, deux microARNs à un site d'action sur le 3'UTR des gènes et MTMS1, MTMS2, deux microARNs à deux et trois sites d'action respectivement. Pour les deux figures, l'expression avec le hairpin contrôle est ramené à 1 et, comme référence, l'expression du microARN mir-20 est illustrée. Pour le gène E2F1, mir-20 diminue l'expression à environ 75% et chacun des microARNs artificiels diminue aussi l'expression. Le microARN le moins efficace pour l'inhibition de E2F1 est MTE2F4 à 80%, et le plus efficace est MTMS1 à 50%. On peut remarquer que MTMS2, même s'il contient trois sites cibles sur le 3'UTR, ne diminue pas mieux l'expression que MTMS1, qui a deux sites cibles. Ces deux microARNs n'ont pas la même séquence et n'agissent pas au même emplacement sur la séquence 3'UTR, ce qui peut expliquer ce résultat. Notons aussi que l'inhibition de l'expression par MTMS1 n'est pas tellement plus prononcée que celle de MTE2F5 qui n'a qu'un seul site cible. Trois des quatre microARNs artificiels régulent davantage E2F1 que mir-20. Lorsque l'on regarde les niveaux d'expression pour E2F3, la tendance générale est semblable à celle de E2F1, mais globalement les niveaux d'ex-

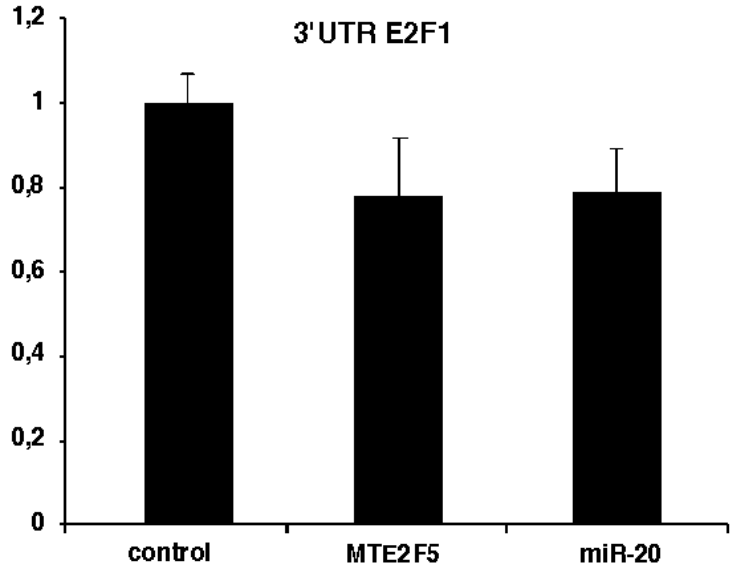

Figure 8.3: Niveaux d'expression de E2F1 avec MTE2F5 et mir-20. MTE2F5 et mir-20 régulent l'expression de E2F1. La référence de l'expression de E2F1 est illustrée sous la barre control.

pression sont plus hauts. Le microARN le moins efficace réduit l'expression à 85% et le plus efficace à 60%, tandis que mir-20 diminue l'expression à 70%. Contrairement à E2F1, MTMS2 est moins actif sur E2F3 et se compare à l'efficacité endogène de mir-20. Deux des quatre microARNs artificiels régulent davantage E2F3 que mir-20. Au moment de l'écriture de ce mémoire, les expériences faites sur E2F2 n'étaient pas concluantes pour mir-20 (aucune diminution observée). Ceci peut être dû au manque de sensibilité de la technique ; il faut trouver un juste équilibre entre l'expression du microARN et l'expression du rapporteur luciférase.

D'autres expériences d'essais luciférases ont été exécutées et sont rapportées aux figures 8.3 8.4 et 8.5. Le microARN choisi est MTE2F5, qui a bien performé dans les expériences ultérieures. Pour les 3 gènes E2Fs, MTE2F5 et mir-20 montrent des diminutions d'expression. Les niveaux d'inhibition de ces deux microARNs sont quasi identiques et sont de 75% pour E2F1, de 70% pour E2F2 et de 70% pour E2F3. Les niveaux d'inhibition de mir-20 pour les gènes E2F1 et E2F3 sont à quelques pourcentages près de ceux reportés dans l'expérience précédente. Par contre, le niveau d'inhibition

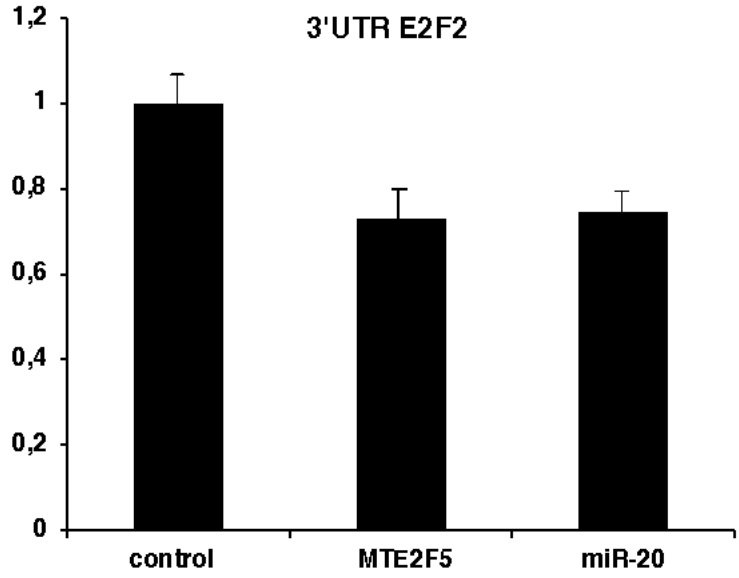

Figure 8.4: Niveaux d'expression de E2F2 avec MTE2F5 et mir-20. MTE2F5 et mir-20 régulent l'expression de E2F2. La référence de l'expression de E2F2 est illustrée sous la barre control.

de MTE2F5 est moins important chez E2F1 et presque identique chez E2F3. Ceci peut être dû aux variations expérimentales. Il est intéressant de remarquer qu'un microARN synthétique peut égaler les niveaux d'expression d'un microARN endogène sur 3 gènes.

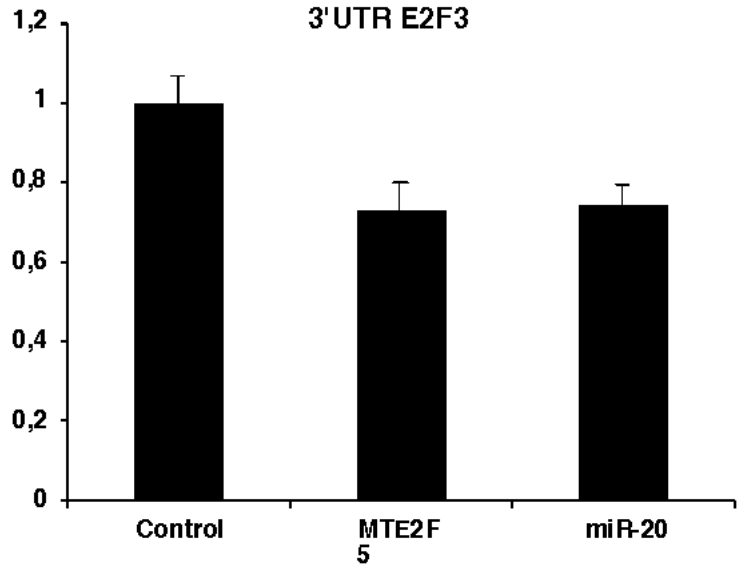

Figure 8.5: Niveaux d'expression de E2F3 avec MTE2F5 et mir-20. MTE2F5 et mir-20 régulent l'expression de E2F3. La référence de l'expression de E2F3 est illustrée sous la barre control.

8.3 Western Blots

L'essai luciférase est un outil efficace, mais assez artificiel, car il ne permet pas d'observer l'effet d'un microARN sur l'expression de la protéine endogène ciblée. Pour remédier à cette situation, des western blots ont été entrepris pour vérifier l'effet des microARNs sur les niveaux endogènes des gènes E2Fs.

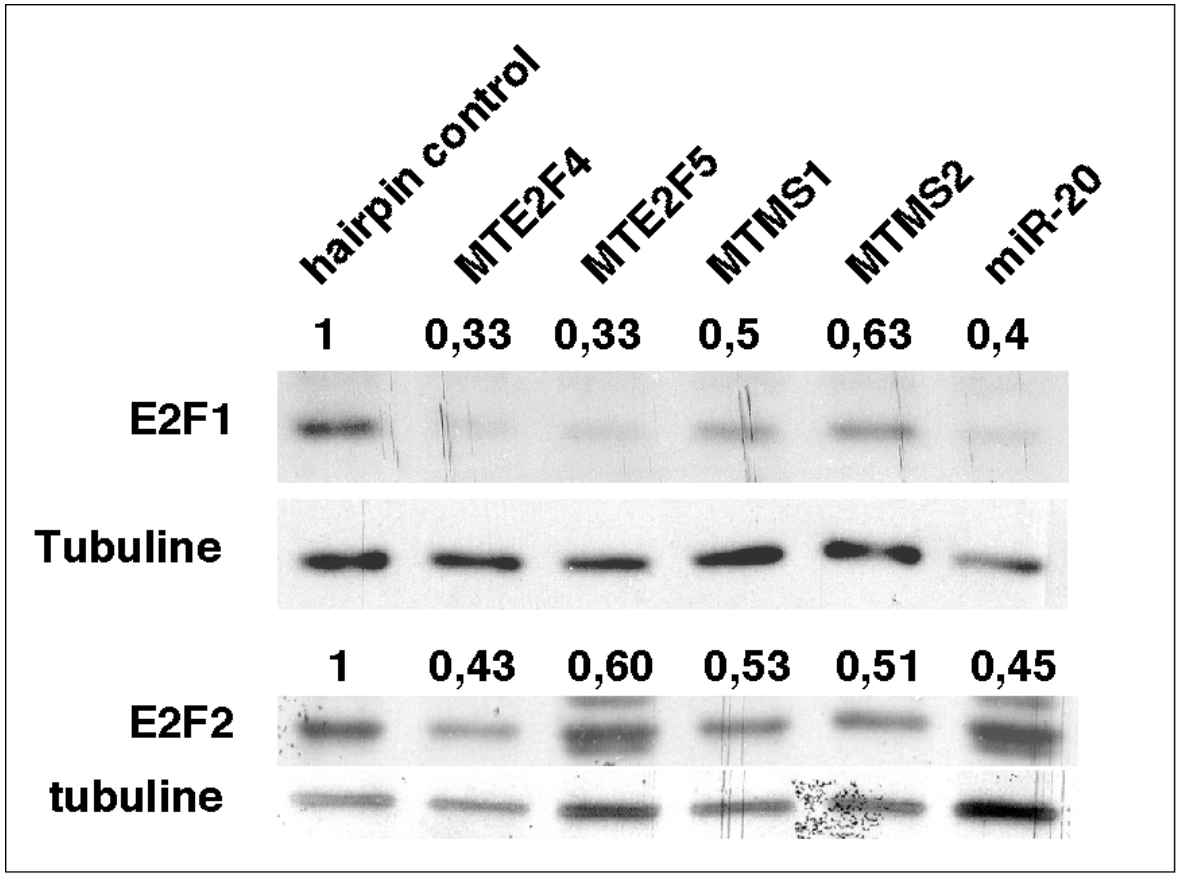

Figure 8.6: Les différents microARNs artificiels régulent l'expression endogène de E2F1 et E2F2. Détection des niveaux endogènes de E2F1 et E2F2 par western blot dans les cellules IMR90 infectées avec les différents microARNs artificiels, mir-20 et le hairpin contrôle.

L'effet des quatre microARNs artificiels sur les gènes E2F1 et E2F2 est démontré avec des western blots à la figure 8.6. Les niveaux d'expressions de ces deux gènes sont ramenés à 1 avec un hairpin contrôle, et les niveaux de la protéine Tubuline sont illustrés. Nous pouvons apercevoir que pour E2F1 et E2F2, tous les microARNs et mir-20 réduisent l'expression. Pour E2F1, les microARNs les plus efficaces sont MTE2F4 et MTE2F5 avec un niveau d'expression de 33% et le microARN le moins efficace est

MTMS2 avec 63%, tandis que mir-20 obtient 40%. Deux des quatre microARNs artificiels ont une efficacité supérieure au microARN endogène. Pour le gène E2F2, le microARN le plus actif est MTE2F4 avec un niveau d'expression de 43%, le moins efficace est MTE2F5 avec 60% et mir-20 obtient 45%. Même si le niveau d'inhibition de MTE2F5 est identique à celui de MTE2F4 pour E2F1, ce n'est pas le cas pour E2F2. Seulement un des quatre microARNs artificiels a une efficacité supérieure à mir-20. Les westerns blots de E2F3 étaient sales et donc n'ont pu être considérés au moment de l'écriture 10.2. Le nombre excessif de bandes révélées suggère que l'anticorps E2F3 n'est pas assez spécifique pour détecter la bande correspondant à E2F3.

On remarque une tendance générale à l'effet que le nombre de sites cibles par microARN ne semble pas être relié à l'efficacité de l'atténuation de l'expression. Ceci peut être dû au fait que les effets ne sont pas additifs, que les séquences du microARN sont moins efficaces, ou que les sites de liaison sur le 3'UTR sont moins accessibles).

8.4 Courbes de croissance

Les courbes de croissance permettent de quantifier la croissance des cellules sur un intervalle de temps donné. Étant donné que les gènes E2Fs sont impliqués au niveau de la progression cellulaire, il est intéressant d'observer l'impact de l'inhibition de ces facteurs de transcription par les microARNs artificiels. Les gènes E2Fs ont une fonction redondante au niveau de la transition G1/S du cycle cellulaire et une diminution simultanée de leur niveau d'expression pourrait refléter une diminution de la croissance cellulaire. Les courbes de croissance ont été séparées en trois graphiques pour faciliter leur étude, et se trouvent aux figures 8.7, 8.8 et 8.9.

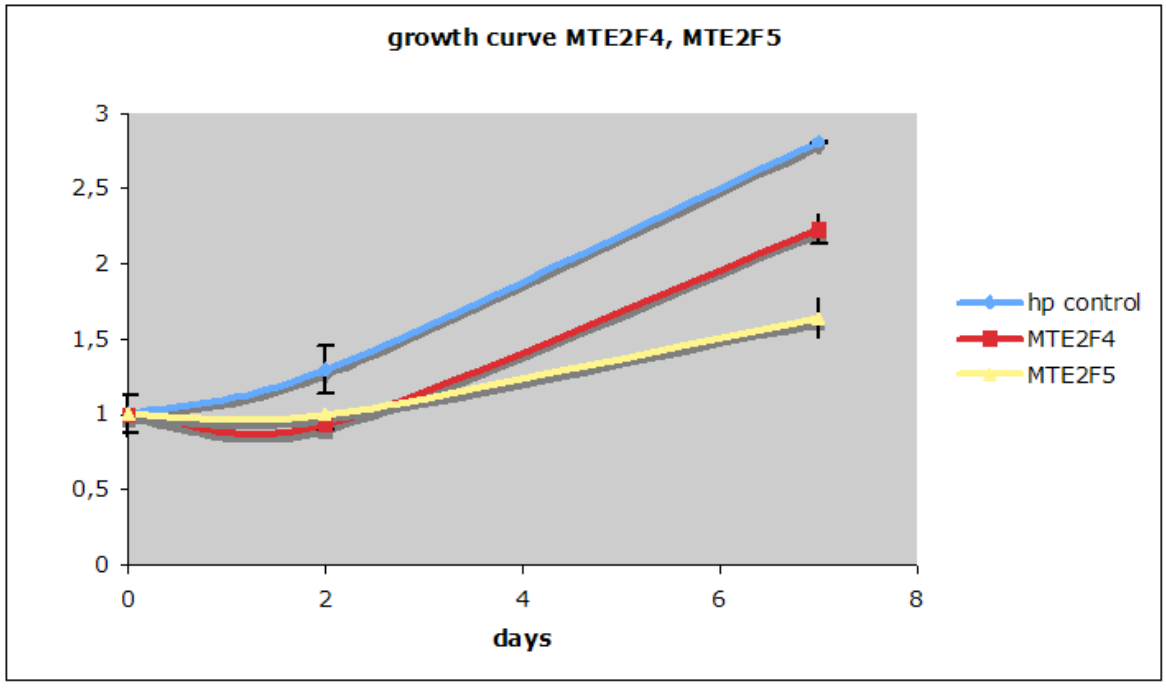

Figure 8.7: Courbe de croissance des cellules IMR90. Quantification du niveau de croissance des cellules IMR90 en présence de deux microARNs artificiels à un site (MTE2F4, MTE2F5) et le contrôle hairpin.

Les courbes de croissance sont un moyen simple de détecter des effets biologiques globaux. L'atténuation de l'expression de trois gènes responsables de la croissance cellulaire devrait en principe résulter en une diminution de la croissance. Les gènes E2Fs ne sont pas activés durant la totalité de la croissance cellulaire, mais sont activés par

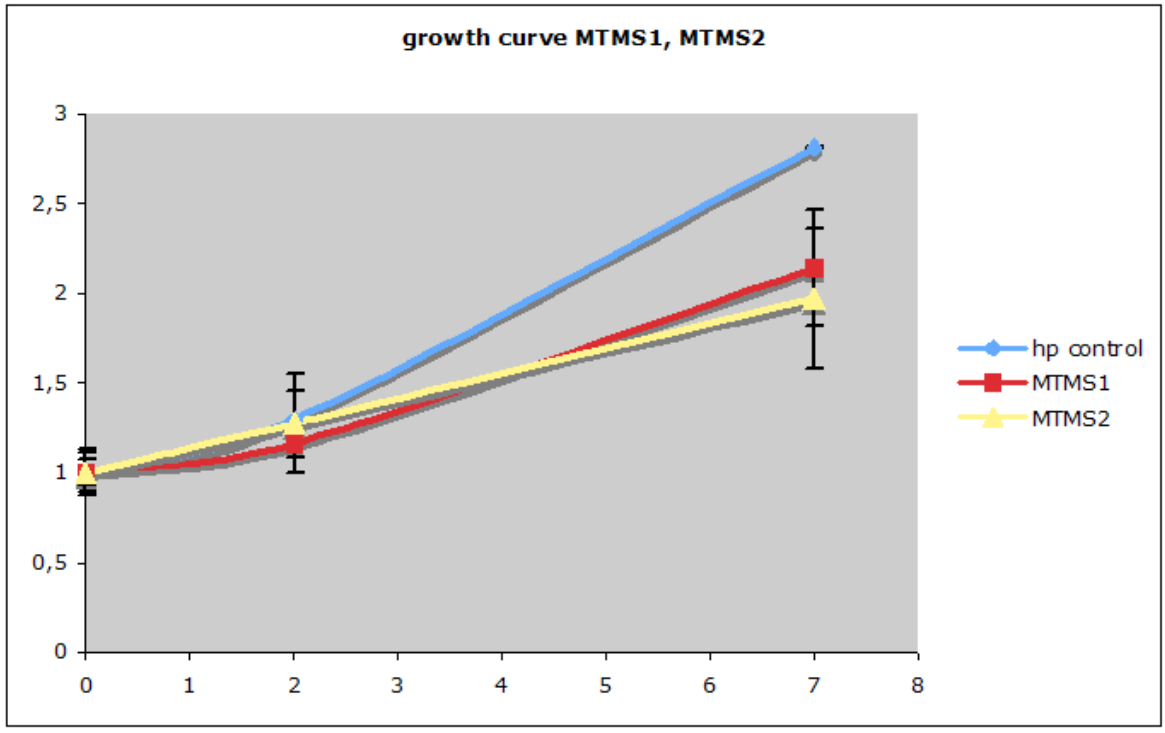

Figure 8.8: Courbe de croissance des cellules IMR90. Quantification du niveau de croissance des cellules IMR90 en présence de deux microARNs artificiels à deux et trois sites (MTMS1, MTMS2) et le contrôle hairpin.

périodes. C'est pourquoi l'étude des courbes de croissance s'échelonne sur plusieurs jours. Les figures 8.7 8.8 et 8.9 montrent les niveaux de croissance pour les quatre microARNs artificiels, le hairpin contrôle, mir-20 et Ras, une protéine reconnue pour interférer de façon agressive avec la croissance cellulaire. Les courbes sont divisées en trois graphiques pour faciliter leur étude.

Premièrement, le hairpin contrôle démontre un compte cellulaire de 2.75 qui est la référence. Tel qu'attendu, Ras obtient un compte de 1.2, ce qui est très bas. Le microARN artificiel ayant le compte le plus bas est MTE2F5 avec 1.6, celui ayant le compte le plus haut est MTMS1 avec 2.1 et mir-20 obtient 1.75. Seul MTE2F5 obtient un compte plus bas que le microARN endogène. Le début de la croissance cellulaire commence 2 jours après le début de l'attachement et augmente de tendance linéaire par la suite. MTE2F5 obtient une bonne inhibition de la croissance, et cela reflète les bonnes inhibitions de l'expression des gènes E2Fs obtenus avec les essais luciférases et

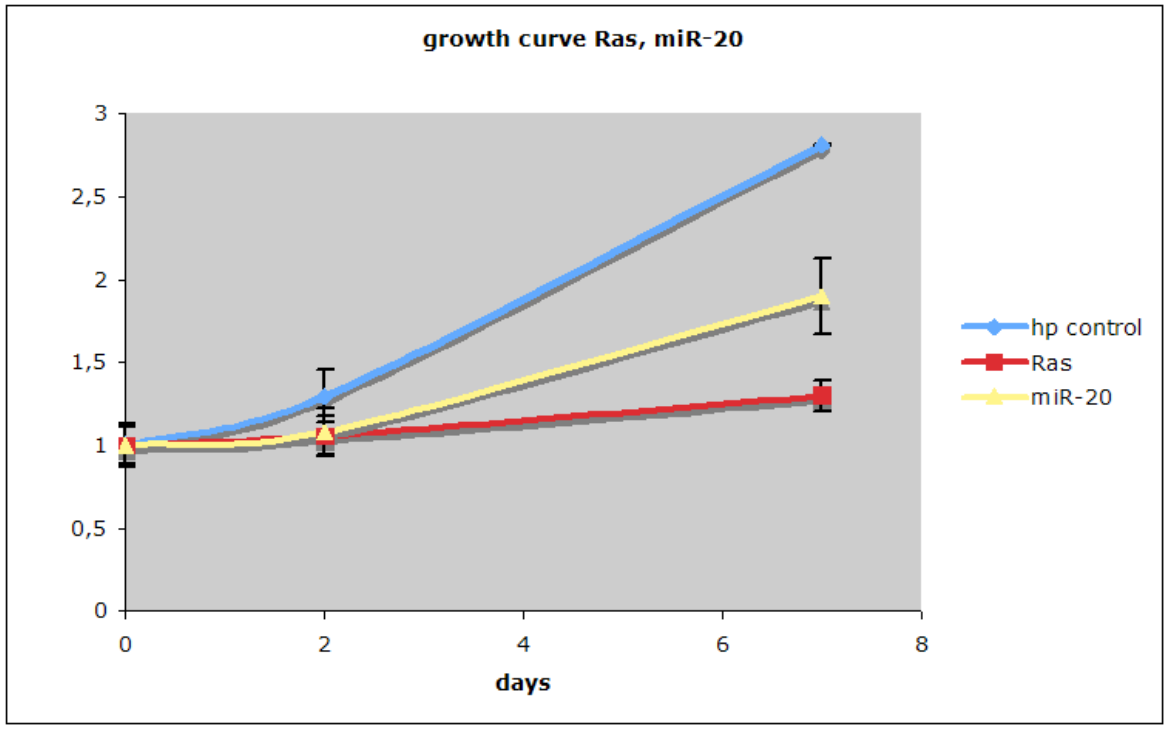

Figure 8.9: Courbe de croissance des cellules IMR90. Quantification du niveau de croissance des cellules IMR90 en présence de mir-20, du contrôle hairpin et de RAS.

les western blots. Il est réaliste de penser que l'inhibition de la croissance cellulaire puisse être dûe à l'inactivation de gènes autres que ceux ciblés par les microARNs. Or, quatre microARNs différents ciblent ces mêmes gènes et l'effet est observable pour chacun de ceux-ci, permettant de déduire que c'est bel et bien l'inactivation des gènes E2Fs qui inhibent la croissance cellulaire.

8.5 Étude de la sénescence

La sénescence cellulaire est un modèle qui permet de vérifier si l'activité redondante des gènes E2Fs est bel et bien atténuée simultanément. Il fut démontré qu'un knockdown siARN de la protéine DP1, impliquée dans l'activation des E2Fs, induit une sénescence prématurée et que cette sénescence est due à la diminution simultanée des trois gènes E2Fs, et non seulement de deux sur trois [61]. Pour vérifier l'efficacité sénescente globale des microARNs artificiels, l'activité B-Galactosidase des cellules infectées a été mesurée après coloration avec le substrat X-gal. Les résultats sont rapportés à la figure 8.10.

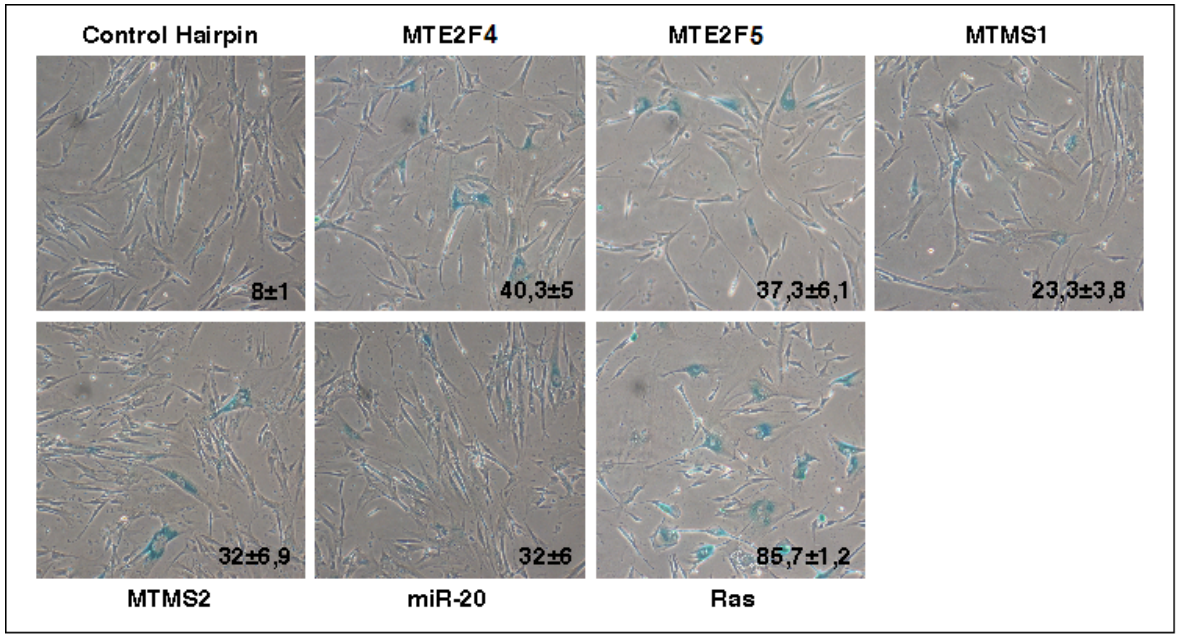

Figure 8.10: Sénescence des cellules IMR90 après incorporation des microARNs artificiels. Sénescence du contrôle hairpin, mir-20, Ras et les quatre microARNs artificiels. Les microARNs induisent une sénescence prématurée (coloration bleuâtre présente chez les cellules contenant les microARNs versus les cellules contenant le contrôle hairpin).

Aussi, une expérience sur la formation de colonies a été performée sur les cellules cancéreuses de la prostate (PC3) en incorporant les microARNs artificiels MTE2F4 et MTE2F5. La formation de colonies est une caractéristique des cellules cancéreuses qui est basée sur la perte du mécanisme de défense de contact d'inhibition. Une diminution de la formation de colonies et de leur taille pourrait donc démontrer une diminution

du potentiel oncogénique de ces cellules, ce qui peut être prometteur au niveau de l'utilisation des microARNs artificiels à des fins thérapeutiques. Une diminution de la formation de colonies a été observée pour ces deux microARNs comparativement au contrôle (38 et 45 colonies pour les cellules contenant MTE2F4 et MTE2F5, versus 64 colonies pour les cellules contrôles). Ceci est illustré à la figure 8.11.

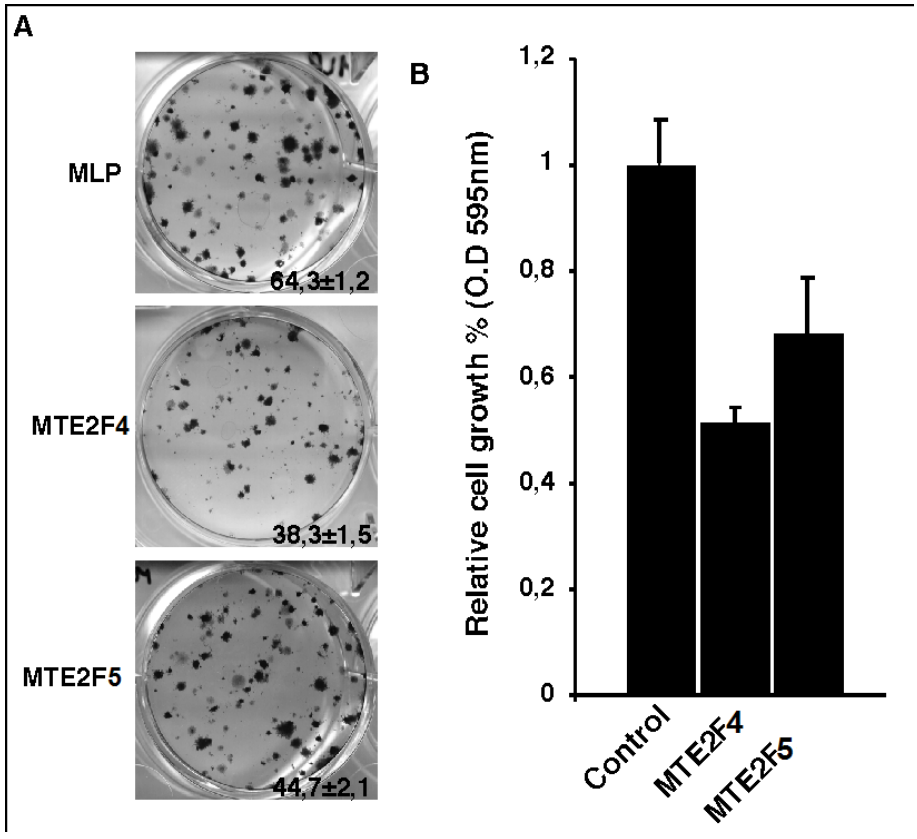

Figure 8.11: Formation de colonies de cellules cancéreuses de la prostate PC3. Les microARNs MTE2F4 et MTE2F5 diminuent la formation de colonies par rapport au contrôle. A. Visualisation des colonies. B. Niveaux de prolifération des colonies versus le contrôle.

La sénescence prématurée des cellules après incorporation des microARNs artificiels est rapportée à la figure 8.10. Le niveau de sénescence du contrôle est de 1 et la protéine Ras induit 85% de sénescence. Par ailleurs, mir-20 obtient 32%, le meilleur microARN 40% et le pire 23%. Les niveaux de sénescence des microARNs artificiels sont proches de celui de mir-20 et les résultats obtenus sont corrélés avec ceux obtenus pour les courbes de croissance. Encore une fois, il ne semble pas que le nombre de sites cibles par microARN améliore la sénescence. Les résultats confirment que l'action des

microARNs artificiels agit globalement sur les trois gènes E2Fs, et non pas seulement sur un ou deux.

La formation de colonies est rapportée à la figure 8.11. Le nombre de colonies pour le contrôle est défini à 1 et MTE2F4 obtient une formation de colonies d'environ 50% et MTE2F5 d'environ 70%. Il est intéressant de voir que l'introduction de microARNs artificiels dans les cellules cancéreuses peut diminuer la formation de colonies. On peut donc penser à des approches thérapeutiques en ciblant dynamiquement les oncogènes les plus exprimés dans différentes souches cellulaires cancéreuses.

Quatrième partie

Discussion

9. Discussion

9.1 Effets non spécifiques

Les microARNs ont chacun des centaines de cibles dans le génome humain. Il faut donc se soucier du problème d'effets non spécifiques qui correspond à la régulation de l'expression des gènes en dehors de ceux ciblés initialement. A priori, on peut estimer le nombre de cibles en se basant sur des programmes bio-informatiques qui se spécialisent dans ce domaine. Or, ces programmes n'ont jamais une précision parfaite. Des expériences supplémentaires, comme des études de toxicité cellulaire, devront être entreprises afin d'étudier les conséquences des effets non spécifiques.

9.2 Autres utilisations

L'effet des microARNs artificiels a été démontré sur 3 gènes impliqués dans la croissance cellulaire. L'approche est applicable à des problèmes thérapeutiques, comme le traitement du cancer. Une utilisation potentielle de MultiTar est de détecter, à l'aide de micro arrays, les niveaux d'ARNs dans les cellules cancéreuses et les comparer à ceux de cellules saines. Les oncogènes ayant des niveaux d'expression plus élevés que la normale peuvent être soumis en entrée au programme afin de générer des microARNs artificiels. Il est alors possible de vérifier l'effet biologique de ces microARNs en regardant les courbes de croissances des cellules cancéreuses et regarder s'il y a un effet d'atténuation de croissance. Toute autre utilisation nécessitant une atténuation de gènes peut profiter de cette approche.

Aussi, il est connu dans la littérature que les microARNs se lient aux séquences 3'UTR des gènes, mais il n'est pas exclu que ceux-ci puissent aussi se lier à l'intérieur d'introns et d'exons. Le programme peut supporter des séquences entières et donc pourrait possiblement trouver des microARNs potentiels à l'extérieur des régions 3'UTR.

9.3 Limitations

La limitation la plus importante de l'algorithme est qu'il ne considère que des aspects structuraux en deux dimensions alors que l'on sait que la structure tridimensionnelle joue un rôle important dans la fonction des molécules. Les algorithmes de repliement 3D d'ARN pourraient apporter de l'information structurelle et donc améliorer l'efficacité du programme.

Un autre aspect à considérer est le nombre de séquences que l'on peut soumettre en entrée. Théoriquement, l'étape de recherche du seed optimal peut produire énormément de solutions lorsque l'on permet des mésappariements. Or, l'étape limitante est la recherche tabou, car avec un nombre de séquences qui augmente, ou avec un nombre de sites cibles par séquence qui augmente, il faut trouver une seule séquence nucléotidique capable de s'aligner efficacement avec tous les sites possibles. Après inspection visuelle, lorsque le nombre de séquences dépasse 10 et le nombre de sites cibles est de 3, l'algorithme a de la difficulté à produire des solutions qui satisfont les seuils préétablis. On pourrait diminuer ces seuils, mais les solutions seraient moins bonnes selon nos critères.

Finalement, le temps d'exécution peut aussi devenir un problème lorsque l'on a beaucoup de séquences et que l'on permet des mésappariements dans le seed. Pour trois séquences allouant deux mésappariements, le temps d'exécution peut prendre quelques heures.

Cinquième partie

Conclusion

10. Conclusions & perspectives

10.1 Conclusions

La découverte des microARNs a engendré un intérêt dans la communauté de recherche quant à leur fonctionnement et leurs cibles. Chacun d'entre eux agit sur des centaines de gènes et ensemble ils sont responsables de la régulation d'environ 10% des produits protéiques chez l'humain. Les microARNs diffèrent des siARNs, car ils amènent un équilibre au niveau de l'expression des gènes au lieu de l'atténuer complètement. Plusieurs outils bio-informatiques ont vu le jour afin de prédire de nouvelles cibles pour les microARNs découverts en laboratoire. Un algorithme nommé MultiTar V1.0 a été présenté dans ce mémoire qui permet de créer des microARNs artificiels qui ciblent les régions 3'UTR des gènes. Quatre microARNs ont été testés sur les gènes E2F1, E2F2 et E2F3, des facteurs de transcription participant à la croissance cellulaire. À l'aide d'essais luciférase, il fut démontré que ces quatre microARNs atténuent l'expression de E2F1 et E2F3. De plus, MTE2F5, un microARN à un site cible, est capable de réguler simultanément l'expression de E2F1, E2F2 et E2F3. Subséquemment, des western blots ont été entrepris afin de regarder les niveaux endogènes des gènes E2Fs en compagnie des microARNs artificiels, et les quatre microARNs atténuent E2F1 et E2F2. Par la suite, une expérience sur la croissance cellulaire a été faite afin de voir si un effet biologique global est observé lorsque l'on incorpore les microARNs aux gènes E2Fs. Pour chacun des microARNs, une diminution de la croissance cellulaire fut observée. Finalement, une expérience sur la formation de colonies chez les cellules cancéreuses de la prostate PC3 a démontré que les deux microARNs à un site cible,

MTE2F4 et MTE2F5, réduisent la formation de colonies. Ceci ouvre une piste vers d'autres expériences similaires, qui ont pour but d'atténuer l'expression d'oncogènes chez les cellules cancéreuses.

Pour chacune des expériences effectuées, les niveaux d'inhibition de l'expression des microARNs artificiels sont proches de ceux de mir-20, un microARN endogène reconnu pour cibler les gènes E2Fs simultanément. De plus, il ne semble pas que la présence de plusieurs sites cibles sur la région 3' UTR des gènes soit liée à une diminution plus accentuée de l'expression, car l'effet des microARNs ayant plus d'un site cible ressemble à celui des microARNs avec seulement un site cible. En se basant sur les résultats obtenus, il semble que la création in-silico d'un microARN artificiel puisse induire l'expression de gènes in-vitro.

10.2 Perspectives

Les résultats obtenus sont intéressants à première vue, mais il faut pouvoir répéter ces expériences au moins trois fois (travaux en cours au laboratoire de Dr. Ferbeyre au moment de l'écriture), en obtenant les mêmes résultats, avant de pouvoir arriver à une conclusion définitive. Les versions futures de MultiTar devront tenir compte d'aspects structuraux tridimensionnels afin de mieux refléter les mécanismes d'action des microARNs. Ceci permettrait de perfectionner la modélisation afin de pouvoir assurer une inhibition simultanée et complète à tout coup. D'autres améliorations possibles peuvent inclure une augmentation de la performance de l'algorithme afin de réduire le temps d'exécution et une solution au problème du nombre de séquences qui est actuellement limité à environ 10. Au moment de l'écriture, six oncogènes surexprimés dans une autre lignée cellulaire cancéreuse ont été soumis à MultiTar afin d'analyser l'effet biologique des microARNs. Cette approche permet d'ouvrir une voie vers la thérapie cancéreuse et offre une alternative pratique au problème de surexpression génique.

Bibliographie

[1] S. Smit, J. Widmann, and R. Knight. Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Research*, 35(10) :3339, 2007.

[2] A.J. Hamilton and D.C. Baulcombe. A Species of Small Antisense RNA in Post-transcriptional Gene Silencing in Plants. *Science*, 286(5441) :950, 1999.

[3] S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature-London-*, pages 494–497, 2001.

[4] R.C. Lee, R.L. Feinbaum, V. Ambros, et al. The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complimentarity to lin-14. *Cell-Cambridge MA-*, 75 :843–843, 1993.

[5] G. Ruvkun. Molecular Biology : Glimpses of a Tiny RNA World. *Science*, 294(5543) :797–799, 2001.

[6] M.Z. Michael, S.M. O Connor, N.G. van Holst Pellekaan, G.P. Young, and R.J. James. Reduced Accumulation of Specific MicroRNAs in Colorectal Neoplasia. *Molecular Cancer Research*, 1 :882–891, 2003.

[7] BJ Reinhart, FJ Slack, M. Basson, AE Pasquinelli, JC Bettinger, AE Rougvie, HR Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature(London)*, 403(6772) :901–906, 2000.

[8] J. Brennecke, D.R. Hipfner, A. Stark, R.B. Russell, and S.M. Cohen. bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell*, 113(1) :25–36, 2003.

[9] A.M. Krichevsky, K.S. King, C.P. Donahue, K. Khrapko, and K.S. Kosik. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10) :1274–1281, 2003.

[10] G.A. Calin, C. Sevignani, C.D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences*, 101(9) :2999–3004, 2004.

[11] S. Pfeffer, M. Zavolan, FA Grasser, M. Chien, JJ Russo, J. Ju, B. John, AJ Enright, D. Marks, and C. Sander. Virology : Identification of Virus-Encoded MicroRNAs. *Science-New York Then Washington-*, pages 734–735, 2004.

[12] L.P. Lim, M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. Vertebrate MicroRNA Genes. *Science*, 299(5612) :1540–1540, 2003.

[13] N.R. Smalheiser. EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. *Genome Biol*, 4(7) :403, 2003.

[14] X. Cai, C.H. Hagedorn, and B.R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12) :1957–1966, 2004.

[15] Y. Lee, K. Jeon, J.T. Lee, S. Kim, and V.N. Kim. MicroRNA maturation : stepwise processing and subcellular localization. *The EMBO Journal*, 21 :4663–4670, 2002.

[16] D.S. Schwarz, G. Hutvágner, T. Du, Z. Xu, N. Aronin, and P.D. Zamore. Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell*, 115(2) :199–208, 2003.

[17] A. Khvorova, A. Reynolds, and S.D. Jayasena. Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell*, 115(2) :209–216, 2003.

[18] S. Griffiths-Jones, H.K. Saini, S. Dongen, and A.J. Enright. miRBase : tools for microRNA genomics. *Nucleic Acids Research*, 2007.

[19] B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander, and D.S. Marks. Human MicroRNA targets. *PLoS Biol*, 2(11) :e363, 2004.

[20] P. Sethupathy, B. Corda, and A.G. Hatzigeorgiouh. TarBase : A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2) :192–197, 2006.

[21] PK Tsantoulis and VG Gorgoulis. Involvement of E2F transcription factor family in cancer. *European Journal of Cancer*, 41(16) :2403–2414, 2005.

[22] Y. Sylvestre, V. De Guire, E. Querido, U.K. Mukhopadhyay, V. Bourdeau, F. Major, G. Ferbeyre, and P. Chartrand. An E2F/miR-20a Autoregulatory Feedback Loop. *Journal of Biological Chemistry*, 282(4) :2135, 2007.

[23] C. Press and P. Reminder. miRiad Roles for the miR-17-92 Cluster in Development and Disease. *Cell*, 133 :217–222, 2008.

[24] A. Ventura, AG Young, MM Winslow, L. Lintault, A. Meissner, SJ Erkeland, J. Newman, RT Bronson, D. Crowley, JR Stone, et al. Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell*, 132(5) :875, 2008.

[25] S.B. Koralov, S.A. Muljo, G.R. Galler, A. Krek, T. Chakraborty, C. Kanellopoulou, K. Jensen, B.S. Cobb, M. Merkenschlager, N. Rajewsky, et al. Dicer Ablation Affects Antibody Diversity and Cell Survival in the B Lymphocyte Lineage. *Cell*, 132(5) :860–874, 2008.

[26] C. Xiao, L. Srinivasan, D.P. Calado, H.C. Patterson, B. Zhang, J. Wang, J.M. Henderson, J.L. Kutok, and K. Rajewsky. Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes. *Nature Immunology*, 9 :405–414, 2008.

[27] I. Ivanovska, A.S. Ball, R.L. Diaz, J.F. Magnus, M. Kibukawa, J.M. Schelter, S.V. Kobayashi, L. Lim, J. Burchard, A.L. Jackson, et al. MicroRNAs in the miR-106b Family Regulate p21/CDKN1A and Promote Cell Cycle Progression. *Molecular and Cellular Biology*, 28(7) :2167–2174, 2008.

[28] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matchings. *SIAM J. Appl. Math*, 35(1) :68–82, 1978.

[29] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and aux iliaryinformation. *Nucleic Acids Res*, 9(1) :133–148, 1981.

[30] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19) :7287–7292, 2004.

[31] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13) :3406, 2003.

[32] I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13) :3429, 2003.

[33] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285(5) :2053–2068, 1999.

[34] P.P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *feedback*, 2005.

[35] F. Major, D. Gautheret, and R. Cedergren. Reproducing the Three-Dimensional Structure of a tRNA Molecule from Structural Constraints. *Proceedings of the National Academy of Sciences*, 90(20) :9408–9412, 1993.

[36] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452 :51–55, 2008.

[37] S. Sharma, F. Ding, and N.V. Dokholyan. iFoldRNA : Three-dimensional RNA Structure Prediction and Folding. *Bioinformatics*, 2008.

[38] A. Stark, J. Brennecke, RB Russell, and SM Cohen. Identification of Drosophila MicroRNA Targets. *PLoS Biol*, 1(3) :e60, 2003.

[39] N. Rajewsky and N.D. Socci. Computational identification of microRNA targets. *Developmental Biology*, 267(2) :529–535, 2004.

[40] B.P. Lewis, I. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7) :787–798, 2003.

[41] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10) :1507–1517, 2004.

[42] X. Yan, T. Chao, K. Tu, Y. Zhang, L. Xie, Y. Gong, J. Yuan, B. Qiang, and X. Peng. Improving the prediction of human microRNA target genes by using ensemble algorithm. *FEBS Letters*, 581(8) :1587–1593, 2007.

[43] M. Yousef, S. Jung, A.V. Kossenkov, L.C. Showe, and M.K. Showe. Naive Bayes for microRNA target predictions machine learning for microRNA targets. *Bioinformatics*, 23(22) :2987, 2007.

[44] A. Fraser and D.G. Burnell. *Computer models in genetics*. McGraw-Hill New York, 1970.

[45] F. Glover. Tabu Search-Part I. *ORSA Journal on Computing*, 1(3) :190–206, 1989.

[46] M. Dorigo. Optimization, Learning and Natural Algorithms. *PhDthesis, PolitecnicodiMilano*, 1992.

[47] C. Zhang and A.K.C. Wong. A genetic algorithm for multiple molecular sequence alignment. *Bioinformatics*, 13(6) :565–581, 1997.

[48] J.S. Wu, C. Lee, C.C. Wu, and Y.L. Shiue. Primer design using genetic algorithm. *Bioinformatics*, 20(11) :1710–1717, 2004.

[49] CH Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1) :37–44, 2003.

[50] J. Blazewicz, M. Szachniuk, and A. Wojtowicz. RNA tertiary structure determination : NOE pathways construction by tabu search. *Bioinformatics*, 21(10) :2356–2361, 2005.

[51] HW Ressom, RS Varghese, SK Drake, GL Hortin, M. Abdel-Hamid, CA Loffredo, and R. Goldman. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, 23(5) :619, 2007.

[52] B.P. Lewis, C.B. Burge, and D.P. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1) :15–20, 2005.

[53] J.G. Doench and P.A. Sharp. Specificity of microRNA target selection in translational repression. *Genes & Development*, 18(5) :504, 2004.

[54] M. Kiriakidou, P.T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes & Development*, 18(10) :1165–1178, 2004.

[55] A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D.S. Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1) :1–1, 2004.

[56] A. Krek, D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. da Piedade, K.C. Gunsalus, M. Stoffel, et al. Combinatorial microRNA target predictions. *Nature Genetics*, 37 :495–500, 2005.

[57] A. Grimson, K.K.H. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel. MicroRNA Targeting Specificity in Mammals : Determinants beyond Seed Pairing. *Molecular Cell*, 27(1) :91–105, 2007.

[58] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10) :1278, 2007.

[59] G. Ferbeyre, E. de Stanchina, E. Querido, N. Baptiste, C. Prives, and S.W. Lowe. PML is induced by oncogenic ras and promotes premature senescence. *Genes & Development*, 14(16) :2015, 2000.

[60] GP Dimri, X. Lee, G. Basile, M. Acosta, G. Scott, C. Roskelley, EE Medrano, M. Linskens, I. Rubelj, O. Pereira-Smith, et al. A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proc Natl Acad Sci US A*, 92(20) :9363–9367, 1995.

[61] E. Hara et al. Reduction of total E2F/DP activity induces senescence-like cell cycle arrest in cancer cellslacking functional pRB and p53. *Journal of Cell Biology*, 168(4) :553–560, 2005.

Annexe I
Options de MultiTar V1.0

MultiTar Options

Seed options

Seed minimum free energy

◉ <= -5.0 kcal/mol ○ <= -7.0 kcal/mol ○ <= -10.0 kcal/mol

Seed mismatches

◉ Perfect seed match ○ 1 mismatch ○ 2 mismatches

Score function - Total must equal 1

Seed score 0.30

Surroundings score 0.35

Accessibility score 0.35

3'UTR options

Base pairing at the 3' end of the microRNA

◉ Force 5 of the last 10 nucleotides to be paired ○ Force nucleotides 12,13 and 14 to be paired

Base pairing at the 5' end of the microRNA

☐ Force 3 out of the first 5 nucleotides to be paired in the 5' part of the microRNA

General options

Multiplicity

◉ 1 ○ 2 ○ 3

Number of solutions

◉ 1 ○ 2 ○ 3 ○ 5 ○ 10

Figure 10.1: Options de MultiTar V1.0 Le programme comporte de nombreuses options, dont des options pour le seed, pour la partie 3' du microARN ainsi que des options pour le nombre de sites par gènes et le nombre de solutions à obtenir.

Annexe 2
Western blot de E2F3

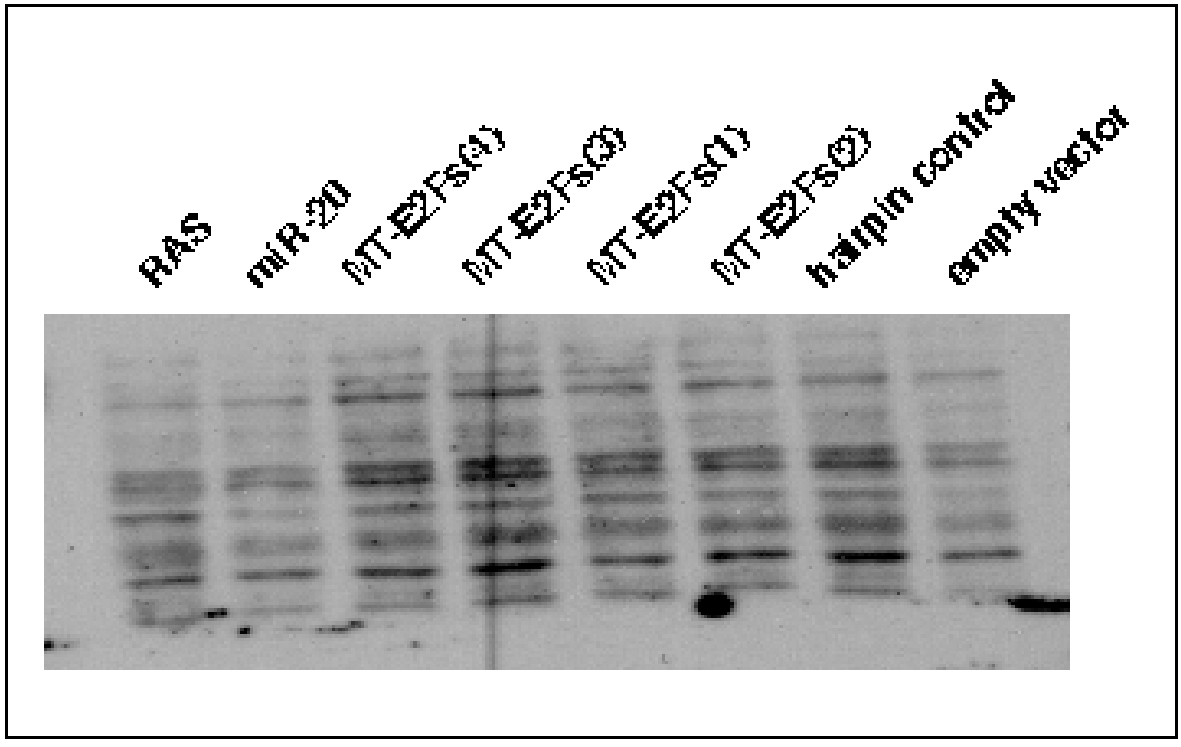

Figure 10.2: Western blot de E2F3. Western blot effectué pour E2F3 en présence des différents microARNs. Le nombre excessif de bandes détectées suggère que l'anticorps E2F3 n'est pas assez spécifique.