Computational Methods for High-Throughput Small RNA Analysis in Plants

Monteiro Morgado, Lionel

SUMMARY

SAMENVATTING (DUTCH SUMMARY)

SUMÁRIO (PORTUGUESE SUMMARY)

SUMMARY

Over the last years, the scientific community has centered efforts to unravel the complex world of RNA molecules that are not translated into a protein, but that rather have a regulatory function in the cell. The small RNA (sRNA) is a very important subclass of such molecules and is central in determining the concentration of genic and non-genic messenger RNA through negative regulation. In recent times, sRNA has evolved into an important tool for genetic engineering and functional genomics, but many aspects of sRNA biology remain unexplored. Understanding the biological fundaments of the new emerging sRNA pathways is therefore imperative.

In plants, millions of uncharacterized sRNA sequences can be found for which experimental validation pose an impractical laborious task. On the other hand, it is nowadays relatively easy and cheap to capture sRNA sequences at a genome-wide scale using deep sequencing approaches. Computational methods have been devised to perform preliminary studies of populations of sRNAs and guide downstream experiments. Nonetheless, sRNA biology is complex and demands a battalion of independent computational methods which currently cannot be found in one unifying framework necessary for a thorough examination, while many critical algorithms are inaccurate or remain to be devised. In special, sRNAs that guide epigenetic mechanisms such as DNA methylation are estimated to be among the most abundant in plants, but despite the importance of this category, there were no tools available in the public domain for their identification by the time the work described in this thesis started.

The existing computational solutions employed to identify and categorize sRNAs acquired through deep sequencing are overviewed in the beginning of this thesis. This sets the base for understanding the “why” and the “how” of the new tools that are further introduced. The novel tools, developed for high-throughput sRNA analysis, were applied to real problems in biology bringing new insights to sRNA-mediated epigenetics.

Chapter 1 introduces basic concepts for understanding the work done in this thesis. The categories of sRNA known to exist in plants are explained, their relationship with epigenetic inheritance and the computational challenges currently faced in the field.

Chapter 2 brings an overview of computational methods to aid sRNA categorization in plants. An extended list of software publicly available is presented, including tools that encompass the analysis of features related to sRNA biogenesis and function.

Chapter 3 describes “SAILS” a novel computational approach for Argonaute-sRNA affinity prediction. The inference algorithm was developed using machine learning methodologies adapted to deal with large datasets, and explores features of sRNA primary structure to infer the capacity of a sequence to load into specific plant Argonaute (AGO) proteins. This tool facilitates the classification of sRNA into putatively functional and non-functional sequences, and further distinguishes sRNA according to their role in transcriptional and post-transcriptional silencing.

Chapter 4 describes “hibeRNAte” a computational framework for the identification and characterization of sRNA from high-throughput sequencing libraries. This is the first computational framework to facilitate comprehensive analyses of all sRNA categories known to exist in plants. The framework features a total of ten modules that include functionalities for sRNA-seq library preprocessing, sRNA categorization, downstream analysis and for *in silico* simulations. The first public program for the detection of heterochromatic-sRNA is also presented.

Chapter 5 shows a link between DNA methylation, sRNA and the production of secondary metabolites in plants. The impact of epigenetic variation on the metabolic composition is studied in epigenetic recombinant inbred lines (epiRILs) from Arabidopsis, and evidence is provided for the presence of sRNA-mediated gene regulation over long genomic distances.

Chapter 6 demonstrates that the inheritance of stress-induced sRNA changes can persist for two generations in apomictic dandelions, independently of the stressor.

SAMENVATTING

De wetenschappelijke gemeenschap heeft de laatste jaren haar krachten gebundeld om de complexe wereld van RNA moleculen die niet worden omgezet in eiwitten, maar die een regulerende functie in de cel hebben, te ontrafelen. Een belangrijke subklasse van dit soort RNA moleculen, genaamd small RNA (sRNA), reguleert de hoeveelheid van messenger RNA (mRNA) in de cel, dat afkomstig is van coderend en niet-coderend DNA, door middel van negatieve regulatie. Small RNA speelt sinds kort een belangrijke rol binnen de gentechnologie en de functionele genomica (Engels: functional genomics), maar veel aspecten van de sRNA biologie zijn nog niet onderzocht. Begrip van de biologische grondslagen van onlangs ontdekte sRNA-regulerende netwerken is daarom onmisbaar.

In planten kunnen miljoenen nog niet gekarakteriseerde sRNA sequenties worden gevonden waarvoor experimentele validatie een onbegonnen arbeidsintensieve taak is. Daarentegen is het tegenwoordig relatief gemakkelijk en goedkoop om door middel van deep sequencing de sRNA sequenties te bepalen die gegenereerd worden door het hele genoom. Computationele methodes zijn ontwikkeld om vooronderzoek uit te voeren aan sRNA populaties zodat richting gegeven kan worden aan vervolg experimenten op het laboratorium. Niettemin is de sRNA biologie complex en vraagt het om een enorme hoeveelheid onafhankelijke, maar elkaar aanvullende, computationele methoden die op dit moment niet kunnen worden gevonden voor de uitvoering van een grondig onderzoek. Daarnaast zijn veel bestaande computerprogramma's onnauwkeurig en moeten vele andere computerprogramma's nog ontwikkeld worden. In het bijzonder schat men dat sRNA moleculen die epigenetische mechanismen sturen, zoals het aanbrengen van DNA methylatie, het meest overvloedig aanwezig zijn in planten. Aan het begin van het onderzoek dat beschreven is in deze thesis waren er echter nog geen programma's beschikbaar voor de identificatie van deze categorie van sRNA moleculen.

Het eerste deel van dit proefschrift bestaat uit een overzicht van bestaande computationele oplossingen die worden gebruikt om sRNA moleculen, die zijn verkregen door deep sequencing, te identificeren en te categoriseren. Dit geeft de context waarbinnen de programma's die verderop worden geïntroduceerd zijn ontwikkeld en beschrijft waarom en hoe de programma's zijn bedacht. De nieuwe applicaties, die zijn ontwikkeld voor high-throughput sRNA analyse, zijn toegepast op bestaande problemen in de biologie en geven nieuwe inzichten over sRNA-gemedieerde epigenetica.

Hoofdstuk 1 introduceert de basis concepten zodat het werk dat in dit proefschrift is gedaan beter begrepen kan worden. De categorieën van sRNA die bestaan in planten, de relatie met epigenetische

overerving en de computationele uitdagingen waarmee men in het veld wordt geconfronteerd worden beschreven.

Hoofdstuk 2 biedt een overzicht van computationele methoden die gebruikt kunnen worden voor de sRNA categorisatie in planten. Een uitgebreide lijst van openbaar verkrijgbare software wordt gepresenteerd. Dit is inclusief applicaties die de analyse omvatten van kenmerken die gerelateerd zijn aan de biogenese en functie van sRNA moleculen.

Hoofdstuk 3 beschrijft 'SAILS', een nieuwe computationele benadering voor de voorspelling van Argonaute-sRNA verwantschap. Het inferentie-algoritme is ontwikkeld door gebruik te maken van machine learning methodologieën, maar wel op zo'n manier dat het aangepast is voor het verwerken van grote databestanden. Verder exploreert het algoritme kenmerken van de primaire sRNA structuur om zo de affiniteit tussen een sequentie en een plant-specifieke Argonaute (AGO) – eiwit te bepalen. Deze applicatie faciliteert de classificatie van sRNA moleculen in vermeende functionele- en niet-functionele sequenties en onderscheidt daarnaast sRNA aan de hand van hun rol in transcriptionele- en post-transcriptionele regulatie.

Hoofdstuk 4 beschrijft 'hibeRNAte', een computationeel raamwerk voor de identificatie en karakterisering van sRNA moleculen die afkomstig zijn van high-throughput sequencing libraries. Dit is het eerste computationele raamwerk waarmee met een uitgebreide analyse sRNA moleculen ondergebracht kunnen worden in alle bekende sRNA categorieën in planten. Het raamwerk bevat een tiental modules die functionaliteiten bevatten voor sRNA-seq library preprocessing, sRNA categorisering, vervolg-analyse en *in silico* simulaties. Daarnaast wordt ook het eerste openbaar verkrijgbare programma voor de detectie van heterochromatisch sRNA gepresenteerd.

Hoofdstuk 5 laat een verband zien tussen DNA methylatie, de aanwezigheid van sRNA moleculen en de productie van secundaire metabolieten in planten. De impact van epigenetische variatie op de metabolische compositie is bestudeerd in zogenoemde epigenetische recombinante inteelt lijnen (epiRILs) van *Arabidopsis thaliana*. Aan de hand van dit onderzoek is er bewijs geleverd voor de aanwezigheid van sRNA-gemedieerde regulatie over lange afstanden op het genoom.

Hoofdstuk 6 toont aan dat de overerving van stress-geïnduceerde sRNA veranderingen voor twee generaties stabiel kunnen blijven bestaan in ongeslachtelijk voortgeplante paardenbloemen, onafhankelijk van de aanwezigheid van een stressfactor.

SUMÁRIO

Ao longo dos últimos anos, a comunidade científica tem centrado esforços para desvendar o complexo mundo das moléculas de ácido ribonucleico (em inglês “ribocucleic acid”, abreviado RNA) que não são traduzidas para proteína, mas que contudo têm uma função reguladora na célula. O pequeno RNA (em inglês “small RNA”, abreviado sRNA) constitui uma importante subclasse de tais moléculas, central para determinar a concentração de RNA mensageiro génico e não-génico através de um mecanismo de regulação negativa. Mais recentemente, o sRNA tem evoluído como uma importante ferramenta para a engenharia genética e genética funcional, mas muitos aspectos da biologia do sRNA continuam por explorar. Perceber os fundamentos biológicos dos novos mecanismos que envolvem sRNA é por isso imperativo.

Nas plantas, há milhões de sequências de sRNA não caracterizadas, cuja validação experimental impõe um desafio laborioso face às metodologias actualmente disponíveis. Por outro lado, é hoje em dia relativamente fácil e barato ler as sequências de sRNA cobrindo a totalidade do genoma de uma célula através de sequenciação profunda (em inglês “deep sequencing”). Diversos métodos computacionais têm sido desenvolvidos para realizar estudos preliminares de populações de sRNA e para guiar a experimentação a jusante. Contudo, a biologia do sRNA é complexa e exige um batalhão de métodos computacionais independentes que correntemente não podem ser encontrados numa plataforma unificadora necessária para análises detalhadas, sendo que ao mesmo tempo muitos algorítmos críticos apresentam baixa precisão e outros aguardam ainda ser desenvolvidos. Em especial, estima-se que os sRNAs que guiam mecanismos epigenéticos, tais como a metilação do DNA, são dos mais abundantes nas plantas, mas apesar da importância desta categoria, não havia ferramentas em domínio público para sua identificação aquando do início do trabalho descrito nesta tese.

As soluções computacionais empregues para identificar e categorizar sRNAs obtidos por sequenciação profunda são revistos no início desta tese. Isto forma a base para a compreensão do “porquê” e do “como” das novas ferramentas que são mais tarde introduzidas. Estas novas ferramentas, desenvolvidas para análise em alto débito de sRNA, foram aplicadas a problemas reais em biologia trazendo novo conhecimento à epigenética mediada por sRNAs.

O **capítulo 1** introduz conceitos básicos para a compreensão do trabalho feito no âmbito desta tese. As categorias de sRNA conhecidas em plantas são explicadas, a sua relação com a temática da herança epigenética e os desafios computacionais enfrentados presentemente na área.

O **capítulo 2** traz uma revisão de métodos computacionais para a categorização de sRNA em plantas. Uma lista extensiva do software disponível publicamente é apresentada, incluíndo ferramentas para a análise de aspectos relacionados com a génese e função dos sRNAs.

O **capítulo 3** descreve "SAILS", um novo método computacional para predição da afinidade entre proteínas Argonautas e sRNA. O algoritmo de inferência foi desenvolvido usando metodologias de aprendizagem máquina adaptadas para processar grandes quantidades de dados, e explorar propriedades da estrutura primária dos sRNAs, que possam ser usadas para inferir a capacidade de uma sequência de se associar a proteinas Argonautas específicas presentes em plantas. Esta ferramenta facilita a classificação de sRNAs em sequências funcionais e não-funcionais, e ainda distingue sRNAs de acordo com a sua intervenção em mecanismos de silenciamento transcricional e pós-transcricional.

O **capítulo 4** descreve "hibeRNAte", uma plataforma computacional para a identificação e caracterização de sRNA em livrarias de sequenciação de alto débito. Esta é a primeira plataforma computacional que permite uma análise compreensiva de todas as categorias de sRNA conhecidas em plantas. A plataforma contém um total de dez módulos que incluem funcionalidades para o preprocessamento de livrarias de sRNA-seq, categorização de sRNA, análise a jusante e para simulações *in silico*. O primeiro programa público para a deteção de sRNA heterocromático é também apresentado.

O **capítulo 5** mostra uma ligação entre a metilação do DNA e a produção de metabolitos secondários em plantas. O impacto da variação epigenética na composição dos metabolitos é estudada em linhas recombinantes epigenéticas (em inglês "epigenetic recombinant inbred lines", abreviado epiRILs) de Arabidopsis, e evidência é fornecida relativa à presença de mecanismos de co-regulação génica mediada por sRNA entre localizações genómicas distantes.

O **capítulo 6** demonstra que a herança de alterações no sRNA devido a stresse pode persistir por duas gerações em dentes-de-leão apomicticos, independentemente do stressor.