Vergleichende Genomik nicht-kodierender RNA

Jana Hertel

Professur für Bioinformatik
Fakultät für Mathematik und Informatik
Universität Leipzig
jana@bioinf.uni-leipzig.de

Abstract: Die Idee nichtkodierendender Ribonukleinsäuren als umfangreiche Klasse von regulatorischen Bausteinen in Eukaryotischen Zellen hat im letzten Jahrzent bemerkenswert an Bedeutung gewonnen. Eine Aufgabe der Bioinformatik findet sich hier in der Möglichkeit neue nichtkodierende Ribonukleinsäuren zu identifizieren, deren Funktion zu untersuchen sowie sie in spezielle Klassen einzuordnen. Gemeinsam mit den Ergebnissen experimenteller Analysen lässt sich auf diese Weise das RNA Kompartment einer Spezies beschreiben. Diese Arbeit befasst sich mit der Vorhersage (Detektion) und Annotation neuer RNA Gene sowie deren Klassifizierung anhand funktioneller und struktureller Merkmale. Neben der Entwicklung von neuen Methoden die u.a. auf maschinellem Lernen basieren, wird auch gezeigt, dass Modifikationen von altbewährten Algorithmen, die Detektion neuer RNA Gene, sowie deren Alignierung und Strukturvorhersage signifikant verbessern.

1 Einführung

Organisation des eukaryotischen Genoms

Die Ergebnisse groß angelegter Studien des letzten Jahrzehnts (*Mensch:* (The07), *Maus:* (MKO$^+$06; RSP$^+$06), *Insekten:* (MDS$^+$06) und *Fadenwürmern:* (HWL$^+$07)) haben gezeigt, dass das Genom eukaryotischer Zellen zu einem beachtlichen Teil tatsächlich transkribiert wird. Die Komplexität des eukaryotischen Genoms zeigt sich in einer Menge von kodierenden (proteinbildenden) und nicht-kodierenden (als RNA funktionalen) Genen, die sich gegenseitig überlappen oder bi-direktional verlaufen und somit von ein und demselben Lokus auf der DNA stammen (siehe Abbildung 1). Lange hielt die Theorie, dass die Hauptfunktion von RNA (transkribierter DNA) in der Kodierung von Proteinen liegt, stand. Auch die Annahme, dass die übrigen Transkripte unbrauchbares Material ("junk RNA") seien, gewann an Popularität[1].
Nachdem sich jedoch herrausstellte, dass 97% der transkribierten DNA hätte "junk" sein sollen, kam die Idee auf, dass, gegeben diesem Ungleichgewicht, RNA eine viel wichtigere Rolle im Leben eukaryotischer Zellen übernehmen könnte als bisher angenommen.

Strukturierte nicht-Protein-kodierende RNA Gene

[1] Ausgenommen die gut bekannten Gene für *transfer* und *ribosomalen* RNAs

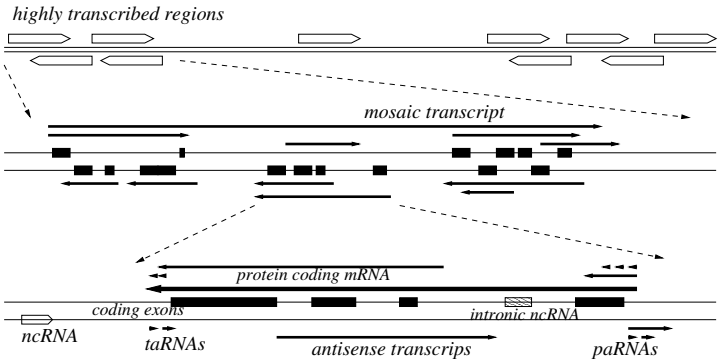


Abbildung 1: Skizze der Sicht auf eukaryotisches Genom anhand der Ergebnisse des ENCODE Projekts. Bild ist extrahiert aus (PS07)

Ein transkribiertes Gen, welches nicht für ein Protein kodiert sondern selbst eine Funktion im Lebenszyklus einer Zelle hat, bezeichnet man als nicht-(Protein-)kodierend (*non-(protein-)coding*, im Folgenden als ncRNA bezeichnet). Die spezielle Funktion einer solchen RNA wird hauptsächlich durch die Struktur welche das Molekül im Raum einnimmt definiert. Wie sich herausgestellt hat, zeigen bereits die Sekundärstrukturen starke charakteristische Merkmale anhand derer man viele ncRNAs ihrer Funktion zuordnen kann.
Wie auch DNA und Proteine, so unterliegen natürlich auch RNA Gene evolutionären Veränderungen. Der Selektionsdruck arbeitet hier hauptsächlich auf der Erhaltung der strukturierten Elemente um die Funktion nicht zu beeinträchtigen. Aus diesem Grund können wir Variation in der Sequenz homologer RNA Gene beobachten während die Struktur erstaunlich gut konserviert ist.
Auf Basis der Funktion/Struktur werden die RNA Gene in verschiedene Klassen eingeteilt. Diese ncRNA Klassen werden bezüglich ihrer Basensequenz weiter unterteilt in ncRNA Familien. Neben den sogenannten "housekeeping RNAs" *ribosomal* RNAs (rRNAs) und *transfer* RNAs (tRNAs) existieren eine Menge anderer Familien. In dieser Arbeit beschäftigen wir uns hauptsächlich mit *micro* RNAs und *small nucleolar RNAs*.

Micro RNAs

Die Klasse der *micro* RNAs (miRNAs) umfasst eine sehr umfangreiche Menge an kurzen ncRNAs, welche in den Prozess der post-transkriptionellen Regulation der Genexpression involviert sind. Sie werden als ca. 90nt langer Precursor transkribiert, der eine charakteristische Stem-loop Struktur einnimmt. Mindestens eine der beiden Stem-Seiten enthält die reife miRNA, welche durch ein Enzym ausgeschnitten und zu ihrem Ziel – dem untranslatierten Ende einer mRNA (3'UTR) geleitet wird.

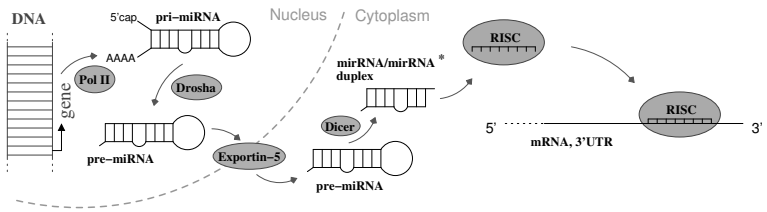


Abbildung 2: Prozessierung und Funktion von miRNAs.

Small nucleolar RNAs

Small nucleolar RNAs (snoRNAs) leiten wichtige chemische Modifikationen anderer RNA Moleküle ein. Die RNA Targets und andere notwendige Proteinkomponenten werden an charakteristischen Sequenz-Strukturmotifen der snoRNA spezifisch gebunden.
Zwei verschiedene Arten der Modifizierung werden von zwei verschiedenen Arten snoRNAs katalysiert. Die Klasse der H/ACA snoRNAs ist verantwortlich für die Pseudouridylation bestimmter U-Residuen der ribosomalen RNAs. Diese snoRNAs weißen neben einer doppelten Stem-loop Struktur auch 2 charakteristische Sequenzmotife $ANANNA$ (H Box) und ACA auf.

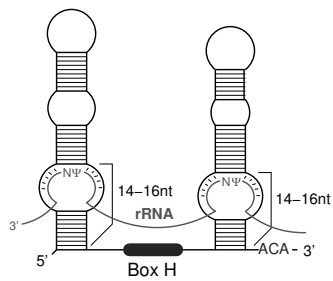


Abbildung 3: H/ACA snoRNA

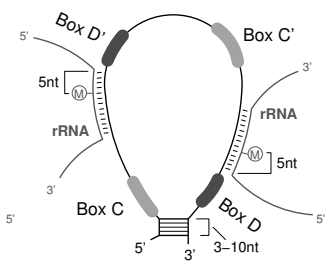


Abbildung 4: C/D snoRNA

Die zweite Klasse der snoRNAs wird durch die C/D snoRNAs beschrieben. Sie sind verantwortlich für die Methylierung des Sauerstoffatoms am C-3 Atom aller vier Arten von Nukleotiden anderer RNAs. SnoRNAs sind weitaus weniger und unregelmäßiger strukturiert. Die meisten C/D snoRNAs bilden einen *kurzen* (4-10nt) Stem, der den Anfang des Moleküls mit dessen Ende verbindet. Eingeschlossen von diesen gepaarten Basen finden sich die charakteristischen Sequenzmotife $UGAUGA$ (C Box) und $CUGA$ (D Box). Oft existiert ein zusätzliches Paar dieser Boxen sodass sich eine Reihenfolge von C-D'-C'-D ergibt. Die Bindungsstelle an der die snoRNA mit der Target RNA interagiert befindet sich unmittelbar stromaufwärts der D bzw. D' Box.

Obwohl diese Gene charakteristische Sekundärstrukturen und kurze spezifische Sequenzmotife aufweisen, lassen sie sich oft nur schwer in großen Mengen transkribierter RNA finden. Ein Grund dafür ist die Kürze der Sequenzen von nur 50 bis 120nt und die schlechte Sequenzkonservierung. Man bedenke, dass, abgesehen von den Sequenzmotifen, die restliche Sequenz (evolutionär) nicht interessant ist, solange die charakteristische Struktur ausgebildet wird. Im Fall der C/D snoRNAs kommt hinzu, dass neben der schlechten Sequenzkonservierung, die Struktur allein zu instabil ist, als dass sie mit den üblichen Strukturvorhersagemechanismen gefunden werden kann.

2 RNA Struktur and Alignment

Die Struktur einer RNA Sequenz lassen sich auf der Ebene von Basenpaarungen als spezielle Typen von gewichteten Matching Problemen modellieren. Die Gewichtung beschreibt die durch die Strukturbildung freigesetzte Energie. Diese Energiebeiträge werden im einfachsten Fall als Kantengewichte modelliert (NPGK78) in realistischeren Modellen jedoch wird die Struktur in "loops" zerlegt (HFS$^+$94). Das Problem der RNA Faltung ist ähnlich zum Sequenzalignment mit dem die Homologie zweier Sequenzen bestimmt wird. Beide Optimierungsprobleme lassen sich effizient mittels Dynamischer Programmierung lösen (DEKM99).

`RNAsalsa`

NcRNAs zeichnen sich durch bestimmte Sequenz- und Strukturmerkmale aus, welche ihre Funktion definieren. Das Alignieren von RNA Sequenzen auf Basis ihrer Sequenz *und* ihrer Strukturinformation ist ein bekanntes Problem in der Bioinformatik. Insbesondere bei langen Sequenzen unterscheiden sich die Ergebnisse verschiedener Programme stark. *Ribosomale* RNAs (rRNAs) sind die meist genutzte Quelle phylogenetischer Information. Allerdings sind diese RNAs sehr lang und weißen in verschiedenen Spezien große Variationen in Substitutionsraten auf. Daher ist es unmöglich sinnvolle Alignments nur auf Sequenzbasis zu erstellen. Obwohl rRNAs stark strukturierte konservierte Muster aufweisen, ist es nicht trivial eine naturnahe Sekundärstruktur vorherzusagen.
Implementiert im Program `RNAsalsa`(SLH$^+$09) haben wir einen Weg gefunden, ein evolutionär sinnvolles multiples Alignment mehrerer verwandter langer RNA Sequenzen zu konstruieren. Unter Beachtung der Sequenzinformation werden die strukturierten Bereiche angemessen aligniert. Außerdem werden in einem Zwischenschritt vernünftige Sekundärstrukturen einer jeden einzelnen Sequenz berechnet.
Angewendet auf zwei naheliegende Beispiele, den Säugetieren und Stachelhäutern (Seeigel, -sterne, etc.), konnte das `RNAsalsa` Alignment die phylogenetische Rekonstruktion erheblich verbessern.

Semi-globales Alignment von ncRNAs – `GotohScan`

Im Allgemeinen werden Sequenzalignments auch dazu benutzt homologe Gene zu finden. Das NCBI-`Blast` Program (AGM$^+$90) ist wohl das bekannteste und am häufigsten benutzte Program für diesen Job. Basierend auf der Methode des *localen paarweisen Alignierens* bekommt man dabei den besten partiellen Match zwischen Query Sequenz und Datenbank.
Für viele ncRNAs benötigen wir allerdings den besten kompletten Match der Query- und Datenbanksequenz. Da viele ncRNAs lange Insertionen/Deletionen/Variationen aufweisen (z.Bsp. von unstrukturierten Bereichen, die zwischen funktionalen (strukturierten) Bereichen liegen) würde der `Blast`-Ansatz das Problem in Teilprobleme brechen. Es kann dann auch passieren, dass es für diese Teilprobleme keine vernünftige Lösung gibt.
Die Berechnung eines *semi-globalen Alignments* mit einem affinen Gap-Kostenmodell scheint hier der natürlichere Weg zu sein um homologe ncRNAs in groß Datenmengen zu finden. `GotohScan`(HdJM$^+$09) implementiert einen solchen Algorithmus nach (Got82). Zeit und Speicher des ursprünglichen Algorithmus sind $\mathcal{O}(n \times m)$. Da wir jedoch nicht

jedes Alignment ausgeben wollen brauchen wir in der Vorwärtsrekursion für zwei der drei Matrizen nur lineare Arrays zu speichern:

$$\begin{aligned} D_{ij} &= \max\{S_{i-1,j} + \gamma_o, D_{i-1,j} + \gamma_e\} & S_{00} &= 0, \\ F_{ij} &= \max\{S_{i,j-1} + \gamma_o, F_{i,j-1} + \gamma_e\} & D_{0j} &= -\infty,\ S_{0j} = F_{0,j} = \gamma_o + (j-1)\gamma_e, \\ S_{ij} &= \max\{D_{ij}, F_{ij}, S_{i-1,j-1} + \sigma(p_i, q_j)\} & F_{i0} &= -\infty,\ S_{i0} = D_{i,0} = \gamma_o + (i-1)\gamma_e. \end{aligned}$$

Nur eine geringe Zahl signifikanter Hits soll ausgegeben werden, darum muss nur für diese Kandidaten die Rückwärtsrekursion durchgeführt werden. Die Signifikanz aller Kandidaten wird mit eine E-Wert Berechnung bestimmt.

Empirisch zeigt sich, dass unser Score Histogram der Alignments einer Gamma-Verteilung folgt (siehe Abbildungen 5)

$$f(s; k, \theta) = \frac{1}{\theta\Gamma(k)} \left(\frac{s}{\theta}\right)^{k-1} e^{-s/\theta}$$

Die Parameter θ und k der Gamma-Verteilung werden dann mit einem Least-square Fitting von $log(f)$ geschätzt. Die eigentliche E-Wert Berechnung erfolgt mit der asymptotischen Erweiterung der unvollständigen Gammafunktion:

$$\log E = (k-1)(\log s - \log\theta) - \log\Gamma(k) + U_k(s/\theta) - \frac{x}{\theta}$$

`GotohScan` erwies sich als ziemlich hilfreich in den Annotationsprojekten des RNA Komplements von *Trichoplax adhaerens* (HdJM$^+$09) und *Aspergillus fumigatus* (JRH$^+$08). Viele ncRNA Gene, deren Vorhandensein erwartet wurde konnten mit den üblichen Methoden wie z.Bsp. `Blast` nicht gefunden werden während `GotohScan` die richtigen Ergebnisse lieferte.

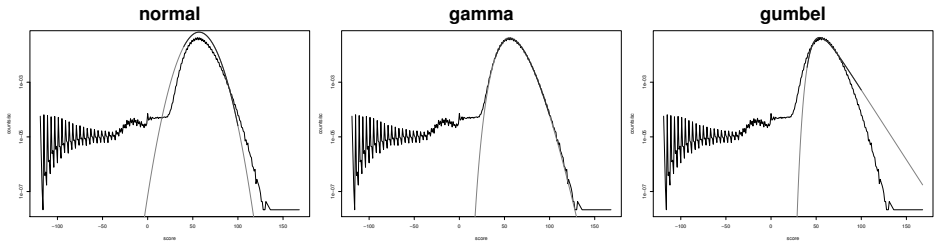


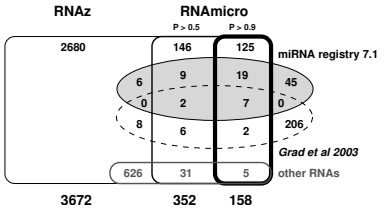
Abbildung 5: Empirischer Test der Scoreverteilung von `GotohScan`.

3 Spezifische RNA Annotation

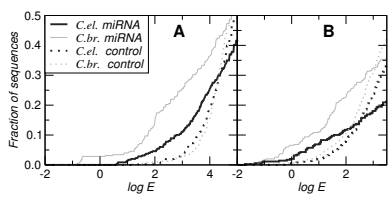
Mögliche Kandidaten für neue ncRNAs können bioinformatisch auf verschiedene Weise gefunden werden. Die meisten Ansätze basieren auf der hohen Konservierung und thermodynamischen Stabilität der Sekundärstrukturen von ncRNAs, z.Bsp. `RNAz` (WH04).

Auch mit Hilfe von Strukturvorhersagen in Kombination mit phylogenetischem Modelling, implementiert im Program `Evofold` (PBS$^+$06) konnten erfolgreich neue ncRNA Gene entdeckt werden. Groß angelegte Anwendungen dieser Programme auf ganze Organismen und natürlich auch weitere experimentelle Analysen lieferten uns eine riesige Menge möglicher neuer ncRNAs u.a. in Wirbeltieren und Fadenwürmern (Nematoden).
Die ncRNAs bilden jedoch keine homogene Klasse sondern zerfallen in eine Vielzahl von spezialisierten Molekülklassen, die sich durch spezifische, durch ihre Funktion bestimmte, Sequenz- und Strukturmerkmale auszeichen. Die Erkennung und Klassifikation von ncRNA Genen kann damit als typische Anwendung fuer Techniken des Maschinellen Lernens verstanden werden.
Der nächste Schritt zur Annotation der RNA eines Organismus ist die Zuordnung der ncRNA Kandidaten zu den einzelnen ncRNA Klassen. Auch hierbei gibt es viele Ansätze, die man komibinieren sollte um möglichst viele falsch positive Annotierungen zu eliminieren. Zunächst vergleicht man die Nukleotidsequenz der Kandidaten mit ncRNA Datenbanken unter Benutzung von Alignmentmechanismen (z.Bsp. `Blast` oder `GotohScan`). Die tRNAs lassen sich recht verlässlich mit Hilfe des Programs `tRNAscan-SE` (LE97) identifizieren. Die korrekte Zuordnung von miRNAs und snoRNAs war jedoch bisher ein größeres Problem. Unter Verwendung von maschinellem Lernen, genauer gesagt einer von uns speziel trainierten *Support Vector Machine* (SVM) konnten wir die zwei Programme `RNAmicro` (HS06) und `snoReport` (HHS08) erstellen, welche dieses Problem überraschend gut lösen. Die speziellen ncRNAs werden hierbei anhand charakteristischer Merkmale der Struktur, deren Stabilität, Sequenzmerkmale und auch Konservierungsinformation von Sequenz und Struktur als miRNA oder snoRNA (beide Arten) identifiziert.

Ein Anwendungsbeispiel beider Programme sei hier kurz vorgestellt. In den Nematoden konnten bekannte miRNAs mit einer Sensitivität von 80% und Spezifität von 96% identifiziert werden (Abbildung 6a). Außerdem konnte die Überrepräsentation eines bekanntes Sequenzmotif stromaufwärts bekannter miRNAs in Nematoden für die `RNAmicro` Kandidaten nachgewiesen werden (verglichen mit den ncRNA Kandidaten, welche von `RNAmicro` abgewiesen worden, siehe Abbildung 6b). Auch die Tatsache, dass die Mehrheit der Vorhersagen in intergenischen bzw. intronischen Regionen zu finden ist, bestätigt die Signifikanz der Hits. Da viele bekannte miRNAs in Clustern auf der DNA kodiert



(a) Venn Diagramm des Ergebnisses der Anwendung von `RNAmicro` auf mögliche ncRNA Kandidaten in Nematoden.



(b) Nachweis der Existenz eines stromaufwärts von bekannten miRNAs gelegenen Sequenzmotife (A und B) in Nematoden für RNAmicro Vorhersagen.

Abbildung 6: Ergebnisse von `RNAmicro` angewendet auf ncRNA mögliche Kandidaten.

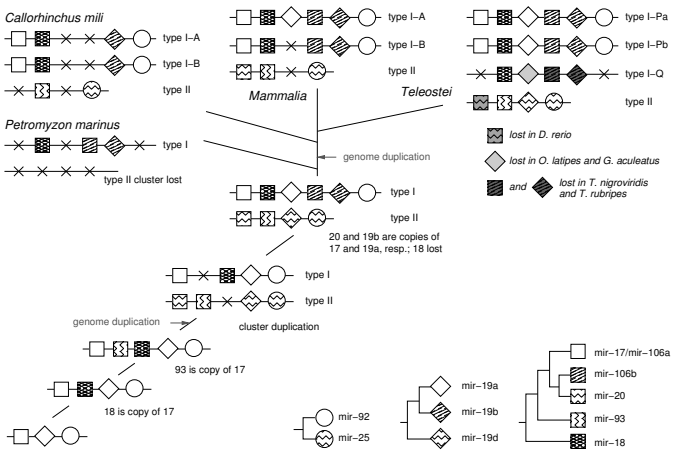


Abbildung 7: Evolution des *mir-17* Clusters, siehe (TS04) und (HLM$^+$06)

sind, wurde auch dieser Aspekt für die vorhergesagten Kandidaten untersucht und in vielen Fällen konnte auch dies nachgewiesen werden.
Auch die Vorhersage von snoRNAs in Nematoden erzielte beeindruckende Ergebnisse. Verglichen mit einer Menge Ergebnisse anderer Studien, zeigte sich `SnoReport` sehr sensitiv in seinen Klassifikationen. Auch die Tatsache, dass das Program viele ncRNA Kandidaten, welche zu anderen Klassen gehören auch nicht als snoRNA klassifizierte ist ein großer Erfolg. Gerade die Gene für H/ACA snoRNAs werden von anderen Programmen oft als miRNA annotiert und umgedreht. Auch nachfolgende Targetanalysen für die snoRNA Kandiaten fruchteten in vielen sinnvollen Annotation, z.Bsp. im Projekt der RNA Annotation im Schimmelpilz *Aspergillus fumigatus* (JRH$^+$08).

4 Micro RNA Evolution

Der evolutionäre selektive Druck liegt bei miRNAs, den kleinen entscheidenden Regulatoren in eukaryotischen Zellen, in der Tatsache, dass das Target stets erkannt werden muss. Außerdem erfordert der Prozessierungsweg der miRNAs die Stem-loop Struktur. Demzufolge, muss beim Auftreten von Mutationen, die Targetsequenz mitevolvieren, und strukturzerstörende Mutationen müssen kompensiert werden.
Die Evolution von miRNAs ist deshalb so interessant, da sie starke phylogenetische Signale aufweisen, obwohl sie so kurz sind (ca. 90nt). In dem meistzitierten Artikel über miRNA Evolution derzeit (HLM$^+$06) haben wir alle erwarteten Ereignisse innherhalb der Metazoen [2] anhand einer riesigen Auswertung der Vorkommen und Familienzugehörigkeiten der miRNAs, nachvollziehen können. In Abbildung 7 ist dies an dem Cluster der miRNA Familie *mir-17* graphisch dargestellt.

[2] lokale/nicht-lokale Genduplikationen, Genverluste, einzelne und kompensatorische Mutationen, sowie auch spontane Innovationen

5 Ausblick

Während dieser Arbeit habe ich viel Erfahrung im Umgang mit Methoden der Bioinformatik nicht nur in Bezug auf die Auswertung von RNA Sequenzen gewonnen. Dadurch gibt es bereits neue Ideen die entwickelten Programme zu verbesser bzw. um weitere Bestandteile zu erweitern. Außerdem schreit die Fülle neuer (biologischer) Daten geradezu danach, insbesondere die Programme welche maschinelles Lernen benutzen neu zu trainieren und eventuell bestimmte Parameter neu zu adjustieren. Mittlerweile wurden auch snoRNAs gefunden, welche sich in bestimmten Situationen wie miRNAs verhalten (EKF$^+$08; SW08). Hier stellt sich die Frage, ob eine Kombination meiner beiden Programme `RNAmicro` und `SnoReport` auch solche Gene identifizieren könnte. Nicht zuletzt verbessert natürlich die Kombination experimenteller Validierung und bioinformatischer Vorhersage die Ergebnisse beträchtlich.

An dieser Stelle möchte ich mich für die großartige Unterstützung meines Doktorvaters Peter F. Stadler und natürlich all meiner Kollegen in den Bioinformatik-Gruppen der Universitäten Leipzig, Wien und Freiburg bedanken.

Literatur

[AGM$^+$90] S F Altschul, W Gish, W Miller, E W Myers und D J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215:403–10, 1990.

[DEKM99] R Durbin, S R Eddy, A Krogh und G Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

[EKF$^+$08] C Ender, A Krek, M R Friedländer, M Beitzinger, L Weinmann, W Chen, S Pfeffer, N Rajewsky und G Meister. A Human snoRNA with MicroRNA-Like Functions. *Mol Cell*, 32(4):519–28, 2008.

[Got82] O Gotoh. An improved algorithm for matching biological sequences. *J Mol Biol*, 162(3):705–708, 1982.

[HdJM$^+$09] J Hertel, D de Jong, M Marz, D Rose, H Tafer, A Tanzer, B Schierwater und P F Stadler. Non-Coding RNA Annotation of the Genome of *Trichoplax adhaerens*. *Nucleic Acids Res*, 37:1602–1615, 2009.

[HFS$^+$94] I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker und P Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh Chem*, 125:167–188, 1994.

[HHS08] J Hertel, I L Hofacker und P F Stadler. `snoReport`: Computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24:158–164, 2008.

[HLM$^+$06] J Hertel, M Lindemeyer, K Missal, C Fried, A Tanzer, C Flamm, I L Hofacker, P F Stadler und Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:25, 2006.

[HS06] J Hertel und P F Stadler. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):e197–202, 2006.

[HWL$^+$07] H He, J Wang, T Liu, X S Liu, T Li, Y Wang, Z Qian, H Zheng, X Zhu, T Wu, B Shi, W Deng, W Zhou, G Skogerbø und R Chen. Mapping the *C elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res*, 17:1471–1477, 2007.

[JRH$^+$08] C Jöchl, M Rederstorff, J Hertel, P F Stadler, I L Hofacker, M Schrettl, H Haas und A Hüttenhofer. Small ncRNA transcriptome analysis from Aspergillus fumigatus suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Res*, 36(8):2677–89, 2008.

[LE97] T M Lowe und S R Eddy. `tRNAscan-SE`: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25:955–864, 1997.

[MDS+06] J R Manak, S Dike, V Sementchenko, P Kapranov, F Biemar, J Long, J Cheng, I Bell, S. Ghosh, A Piccolboni und T R Gingeras. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet*, 38:1151–1158, 2006.

[MKO+06] N Maeda, T Kasukawa, R Oyama, J Gough, M Frith, P G Engström, B Lenhard, R N Aturaliya, S Batalov, K W Beisel, C J Bult, C F Fletcher, A R Forrest, M Furuno, D Hill, M Itoh, M Kanamori-Katayama, S Katayama, M Katoh, T Kawashima, J Quackenbush, T Ravasi, B Z Ring, K Shibata, K Sugiura, Y Takenaka, R D Teasdale, C A Wells, Y Zhu, C Kai, J Kawai, D A Hume, P Carninci und Y Hayashizaki. Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. *PLoS Genet*, 2:e62, 2006.

[NPGK78] Ruth Nussinov, George Piecznik, Jerrold R Griggs und Daniel J Kleitman. Algorithms for Loop Matching. *SIAM J Appl Math*, 35(1):68–82, 1978.

[PBS+06] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S Lander, Jim Kent, Webb Miller und David Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, 2(4):e33, 2006.

[PS07] Sonja J. Prohaska und Peter F. Stadler. A story of Growing Confusion: Genes and Their Regulation. In Rubem P. Mondaini und Rui Dilão, Hrsg., *BIOMAT-2007: International Symposium on Mathematical and Computational Biology*, Seiten 325–345, Singapore, 2007. World Scientific. Armação dos Búzios, RJ, Brazil, 24-29 November 2008.

[RSP+06] Timothy Ravasi, Harukazu Suzuki, Ken C Pang, Shintaro Katayama, Masaaki Furuno, Rie Okunishi, Shiro Fukuda, Kelin Ru, Martin C Frith, M Milena Gongora, Sean M Grimmond, David A Hume, Yoshihide Hayashizaki und John S Mattick. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res*, 16:11–19, 2006.

[SLH+09] R R Stocsits, H Letsch, J Hertel, B Misof und P F Stadler. Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res*, 37(18):6184–93, 2009.

[SW08] A A Saraiya und C C Wang. snoRNA, a Novel Precursor of microRNA in Giardia lamblia. *PLoS Pathog*, 2008. e1000224.

[The07] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[TS04] A Tanzer und P F Stadler. Molecular evolution of a microRNA cluster. *J Mol Biol*, 339(2):327–35, 2004.

[WH04] S Washietl und I L Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, 342(1):19–30, 2004.

Jana Hertel geboren am 11. September 1981 in Leipzig, absolvierte das naturwissenschaftlich orientierte Gymnasium in Brandis nahe Leipzig. Im Jahr 2000 schließt sie dies erfolgreich mit dem Abitur ab und beginnt im gleichen Jahr das Studium der Bioinformatik an der Universität Leipzig. Im Dezember 2005 erhält sie ihr Diplom und die Möglichkeit der weiterführenden Forschung im Gebiet der nichtkodierenden Nukleinsäuremoleküle in der Arbeitsgruppe von Prof. Peter F. Stadler an der Professur für Bioinformatik der Universität Leipzig. Um weitere Erfahrungen und neue Kontakte auf dem Forschungsgebiet zu knüpfen, führt sie ihre Forschungen im Jahr 2007 am Institut für theoretische Chemie an der Universität Wien in Österreich fort. Zurück in Leipzig rundet sie ihre Forschungsergebnisse mit ihrer Doktorarbeit im Dezember 2008 ab. Im April 2009 verteidigt sie sehr erfolgreich ihre Arbeit und erhält den Titel *Dr. rer. nat.* mit Auszeichnung. Seit August 2009 ist sie stolze Mutter eines Sohnes und befindet sich im Mutterschaftsurlaub.