Bioinformatisches Praktikum WS 2012/13

begleitend zur Vorlesung Sequenzanalyse und Genomik und Fortgeschrittene Methoden der Bioinformatik

1 Allgemeines

Small nucleolar RNAs (snoRNAs) sind kleine nicht-kodierende RNA Gene mit der Aufgabe ribosomale (rRNAs) und small nuclear (snRNAs) RNAs zu modifizieren, bzw. deren Modifikation einzuleiten. Box C/D snoRNA Moleküle sind notwendig um die Positionen in der target Sequenz aufzuspüren, welche methyliert werden müssen, während box H/ACA snoRNAs die Uridine lokalisieren, welche durch Pseudouridylation modfiziert werden sollen. Small cajal body RNAs (scaRNAs) sind snoRNAs die entweder box C/D, box H/ACA oder aber Charakteristika beider Typen aufweisen und somit auch entweder Methylierungen, Pseudourydilierungen oder beide Funktionen ausüben. Diese speziellen snoRNAs sind in den Cajal Bodies im Nukleous zu finden.

Anhand eines Basis-Datensatzes von experimentell verifizierten snoRNAs in Invertebraten sollen deren Homologe gefunden werden. Ziel des Praktikums ist eine möglichst vollständige Annotation des snoRNA Komplements in Invertebraten.

Das Praktikum der Module 10-202-2207 *(Sequenzanalyse und Genomik, Gruppe I)* und 10-202-2206 *(Fortgeschrittene Methoden der Bioinformatik, Gruppe II)* findet überlappend statt:

Gruppe I, A:	07.01.2012 - 18.01.2012
Gruppe II,B:	14.01.2012 - 25.01.2012
Gruppe I, C:	21.01.2012 - 01.02.2012

Wo? Härterlstr. 16-18, Raum 309

2 Ablauf

Gruppe A

Mo 07.01 bis Fr 11.01.2013

2.1 Erstellung des Basis-Datensatzes - Mo bis Mi

2.1.1 Small Nucleolar RNAs

Bereits bekannte, experimentell verifizierte snoRNAs von Invertebraten werden gesammelt und zur weiteren Verarbeitung aufbereitet. Zur Verfügung stehen Annotationen aus *D. melanogaster*, *C. elegans* und *B. mori*, sowie evtl. einzelne snoRNAs aus anderen Organismen, siehe Literaturliste.

- finde zugehörige Sequenzen im Supplement, NCBI-Datenbank, o. anderen Quellen, siehe jeweilige Publikation
- erstelle Fasta-file mit eindeutigem header (AC o. Name laut Publikation,etc.)
- annotiere Boxen manuell, o. entspr. Publikation
- extrahiere Koordinaten, Chromosom, Leserichtung
- formatiere Header jeder Sequenz entsprechend der Vorlage

2.1.2 Target RNAs

Target RNAs: snRNAs und rRNAs werden gesammelt. Hier kann es vorkommen, dass das rRNA Operon nur zum Teil annotiert ist. In solchen Fällen wird versucht, es manuelll zusammenzubauen. Sollten rRNAs bestimmter Spezien nicht annotiert sein, wird versucht diese anhand von bereits gefundenen rRNA Genen verwandter Spezien zu finden. SnRNAs einiger Invertebraten finden sich hier http://www.bioinf.uni-leipzig.de/publications/supplements/08-001.

2.2 Homologiesuche - Do

2.2.1 Entferne Homologien im Basis-Datensatz

Bevor in den verfügbaren Invertebraten nach Homologen snoRNAs gesucht werden kann, müssen die bereits gesammelten Sequenzen selbst auf Homologe untersucht werden. Häufig kommt es vor, dass bei der Neuannotation von Genen nicht geschaut wird, ob es möglicherweise eine verwandte (bereits annotierte) Sequenz in einer anderen Spezies gibt. Dies führt zu Inkonsistenzen in der Namensgebung und in statischen Analysen. Hier soll ein Mapping der Ausgangsnamen erstellt und homologe snoRNAs zusammengefasst werden (in 1 Fasta file).

2.2.2 Verfeinerungen

Mit snoRNAs aus Würmern und aus Insekten wird mittels der automatischen Pipeline `snoStrip` in den jeweiligen Linien gesucht (ohne Targetsuche, ohne Infernal). Evtl. müssen Box Positionen korrigiert werden. Zum Testen der Korrektur wieder `snoStrip` verwenden.

2.2.3 Finde zusätzliche Homologe snoRNAs in Invertebraten - Fr

Finde Homologe der Datenbasis in *allen* Invertebraten mittels `snoStrip` (mit Targetsuche, ohne Infernal). *Dies soll übers Wochenende rechnen!*

Gruppe A + B

Mo 14.01 bis Fr 18.01.2013

Die Alignments werden angesehen und auf Fehler in der Boxannotation u. falsch alignierte Positionen untersucht u. ggfs. korrigiert. Sie werden um Sekundärstrukturannotation erweitert.

2.2.4 Auswertung Datenbasis - Gruppe A

Alle mit Hilfe von `snoStrip` gefunden Targets werden mit evtl. Target-Annotationen der Publikation (dme, cel, bmo) verglichen.

- Haben alle snoRNAs einer Familie das gleiche Target?
- Gibt es Targets die von versch. snoRNA Familien modifiziert werden können?
- Sonstige Sonderfälle?

Wurden snoRNAs gefunden, die bereits in anderen Publikationen "gefunden" (predicted) worden sind (z.Bsp. *S. mansoni* und *S. japonicum*)?

- # snoRNAs in Input (wieviele snoRNAs, wieviele Familien)
- # snoRNAs in Output (wieviele snoRNAs, wieviele Familien)
- # snoRNAs mit/ohne Target

→ Bilder, Statistiken, hübsche u. schlechte Beispiele

2.3 Homologiesuche auf Distanz - Gruppe B

Basierend auf den resultierenden Alignments von snoRNA Familien in Invertebraten der Gruppe A, sollen nun distante Homologien aufgedeckt werden und neue snoRNA Familien gefunden werden.

Alle Alignments werden in das .stk Format überführt. Dies erfordert eine zusätzliche Strukturannotation der Alignments, d.h. es muss eine Consensus-Struktur berechnet werden, die die constraints des snoRNA Typs berücksichtigt.

Zusätzlich werden `Infernal`modelle aus den Alignments erstellt. Diese werden benutzt um entfernte Verwandte bzw. gleichartige snoRNAs in Invertebraten zu finden. ***`Infernal`*** *Lauf am Freitag starten, übers Wochenende rechnen!*

Im nächsten Schritt werden alle snoRNA Familien selbst auf distante Homologien untersucht. Dazu wird das Programm `cmcompare` benutzt, welches alle Sequenzen über Kreuz vergleicht und das Vergleichsergebniss bezüglich eines randomisierten Backgrounds bewertet. Ggfs. werden die Familien zusammengefasst. ***`cmcompare`*** *Lauf am Freitag starten, übers Wochenende rechnen!*

Gruppe B

Mo 21.01 bis Fr 25.01.2013

2.3.1 Gesamtauswertung - Gruppe B

Alignments und deren Analyse zusammenfassend auswerten:

- # snoRNAs in Input (wieviele snoRNAs, wieviele Familien/Alignments)
- # Alignments mit guter/richtiger u. schlechter/falscher Sekundärstruktur
- # zusätzlicher mit `Infernal` gefundener Alignments
- # zusammengeführter Familien mittels `cmcompare`

→ Bilder, Statistiken, hübsche u. schlechte Beispiele

Gruppe C

Mo 21.01 bis Fr 01.02.2013

2.4 Genomische Umgebung und Host Gene von snoRNAs

2.4.1 Motifsuche

Mit `MEME` wird in der näheren Umgebung gefundener snoRNAs nach Sequenzmotifen, Promotoren gesucht. Gefunden Motife werden innerhalb der Familie auf Konservierung verglichen.

Mit Blick auf alle gefundnen Motife in allen snoRNA Familien:

- Sind Motife innerhalb einer Familie konserviert?
- Gibt es Motife die generell häufig vorkommen?
  Welche, wie oft und wo?

2.4.2 Host Gen und Cluster Analyse

In den meisten Fällen kommen snoRNAs in Introns protein-kodierender Gene, seltener zwischen solchen Genen vor. Auch wurde beobachtet, dass sich mehrere snoRNAs in genomischen Clustern befinden können. Es sollen die Host-Gene, bzw. umliegende Gene der snoRNAs in einer Familie verglichen werden.

- Sind Host Gene in einer Familie Konserviert
- "Springt" die snoRNA eventuell im Genom
- In welchem Ausmaß sind snoRNA Cluster konserviert

Verteilung der snoRNAs in Invertebraten