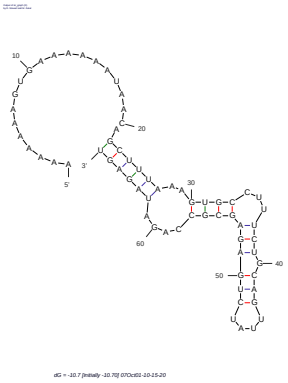


Angewandte Bioinformatik I - Lösungen zu Übungsaufgaben

Strukturbioinformatik: RNA

Manuelle Sekundärstrukturvorhersage

Lösung 1 Die meisten Faltungsprogramme beschränken sich auf die folgenden drei Paarungen:

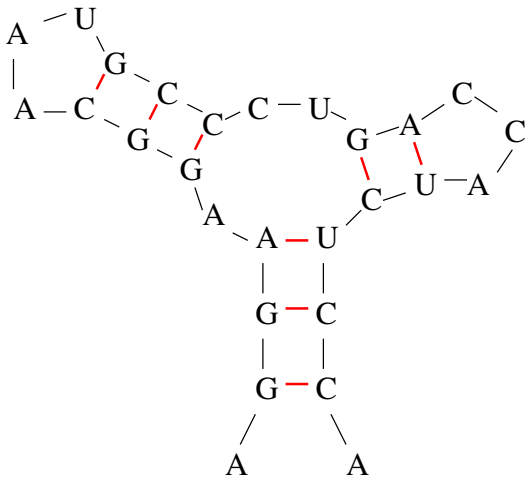
a) kanonischen Watson-Crick-Paarungen: G-C, A-U

b) und die Wobble-Paarung: G-U

Die Sequenz hat folgende Struktur in der Vienna-Darstellung:

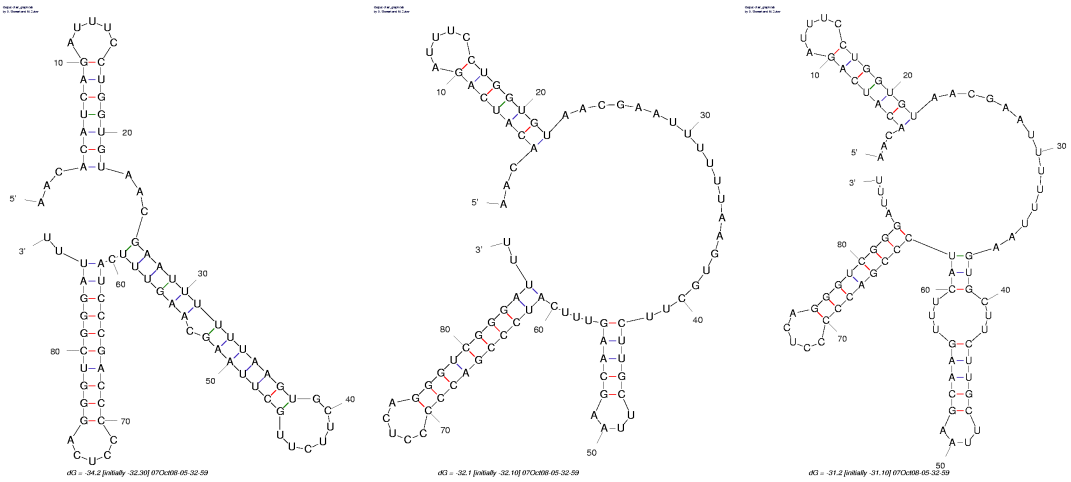
```
AGUAACAGC
.((.)(.))
```

Lösung 2 Die Sequenz/Struktur kann wie folgt dargestellt:



`Mfold`-Webserver zur Vorhersage von RNA Sekundärstrukturen

Lösung 3 Wird die DsrA RNA mit den Standardparametern von `Mfold` gefaltet, so werden die folgenden drei Strukturen ausgegeben:



a) `Mfold` bestimmt nicht nur die Struktur mit der minimalen freien Faltungsenergie, sondern auch alle Strukturen, deren Faltungsenergie sich in einem bestimmten Bereich bewegt. Dieser Bereich wird mit dem Parameter *percent suboptimality* festgelegt. Als Standardparameter sind 5% angesetzt, d.h. es werden alle Strukturen berechnet, deren Faltungsenergie sich maximal um 5% von der MFE unterscheiden.

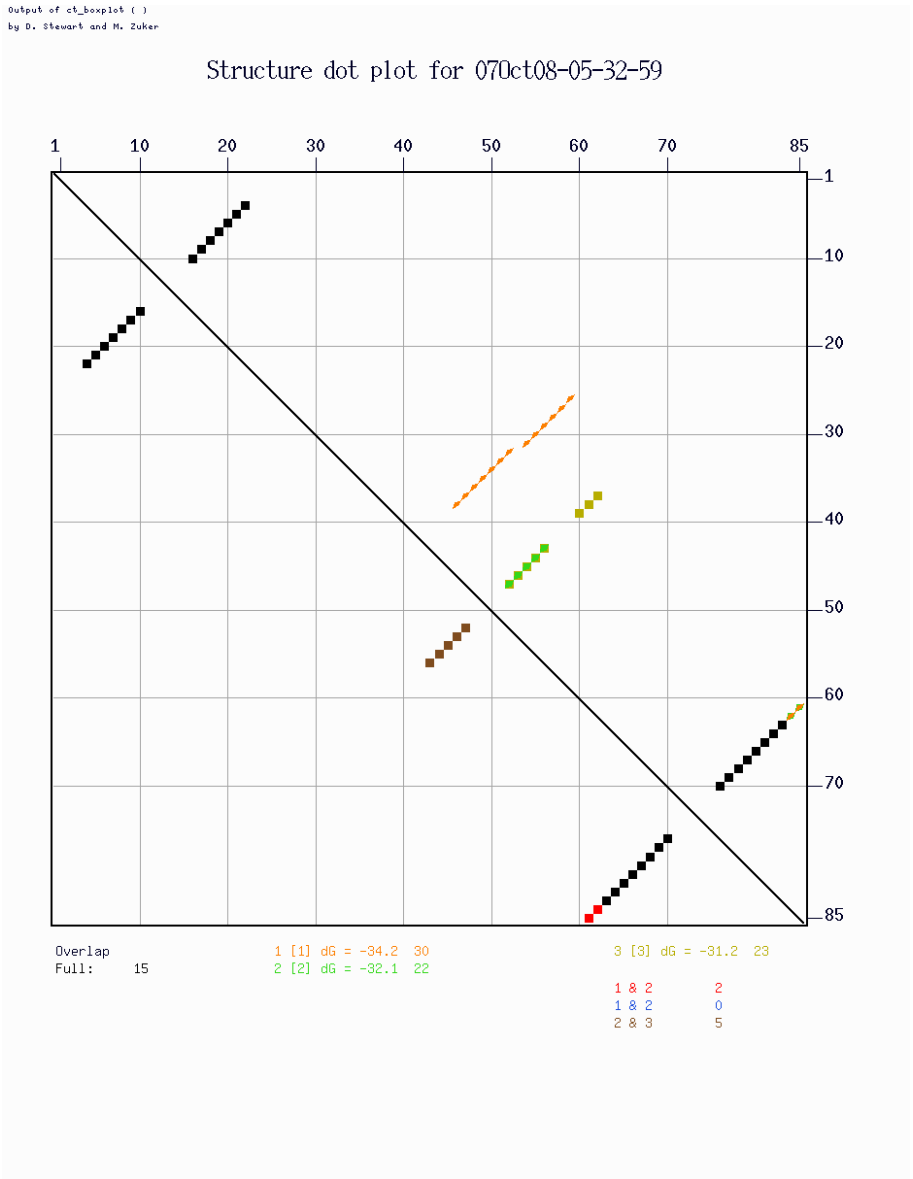
Die Faltungsenergie kann sich bereits beim Hinzufügen oder Entfernen einer einzigen Paarung leicht verändern und damit zu einer enormen Anzahl an leicht variierenden Strukturen führen. Aus diesem Grund wurde ein weiterer Parameter, genannt: *window*, eingeführt. Dieser Parameter legt fest, wie unterschiedlich die angezeigten Strukturen sein sollen. Je grösser der Wert ist, umso unterschiedlicher sind die gezeigten Strukturen und umso weniger von ihnen werden innerhalb eines gegebenen Energiebereichs gefunden. Aus allen möglichen Strukturen im festgelegten Bereich wird also eine Vorauswahl sich unterscheidender Strukturen getroffen, um dem Betrachter mögliche Strukturunterschiede zu zeigen, ohne ihn mit redundaten Einzelheiten zu erschlagen. Da die Anzahl möglicher Faltungen einer Sequenz mit ihrer Länge zunimmt, sind Standardwerte des *window*-Parameters an die Länge der Sequenz gekoppelt.

b) Der Struktur-Dotplot erlaubt dem Benutzer jede Kombination der angezeigten Strukturen zu vergleichen. In diesem Dotplot wird die Sequenz auf der vertikalen und horizontalen Achse aufgetragen. Paarende Basen an Position i und j in der Sequenz werden durch einen Punkt im Schnittpunkt der i-ten Zeile mit der j-ten Spalte markiert.

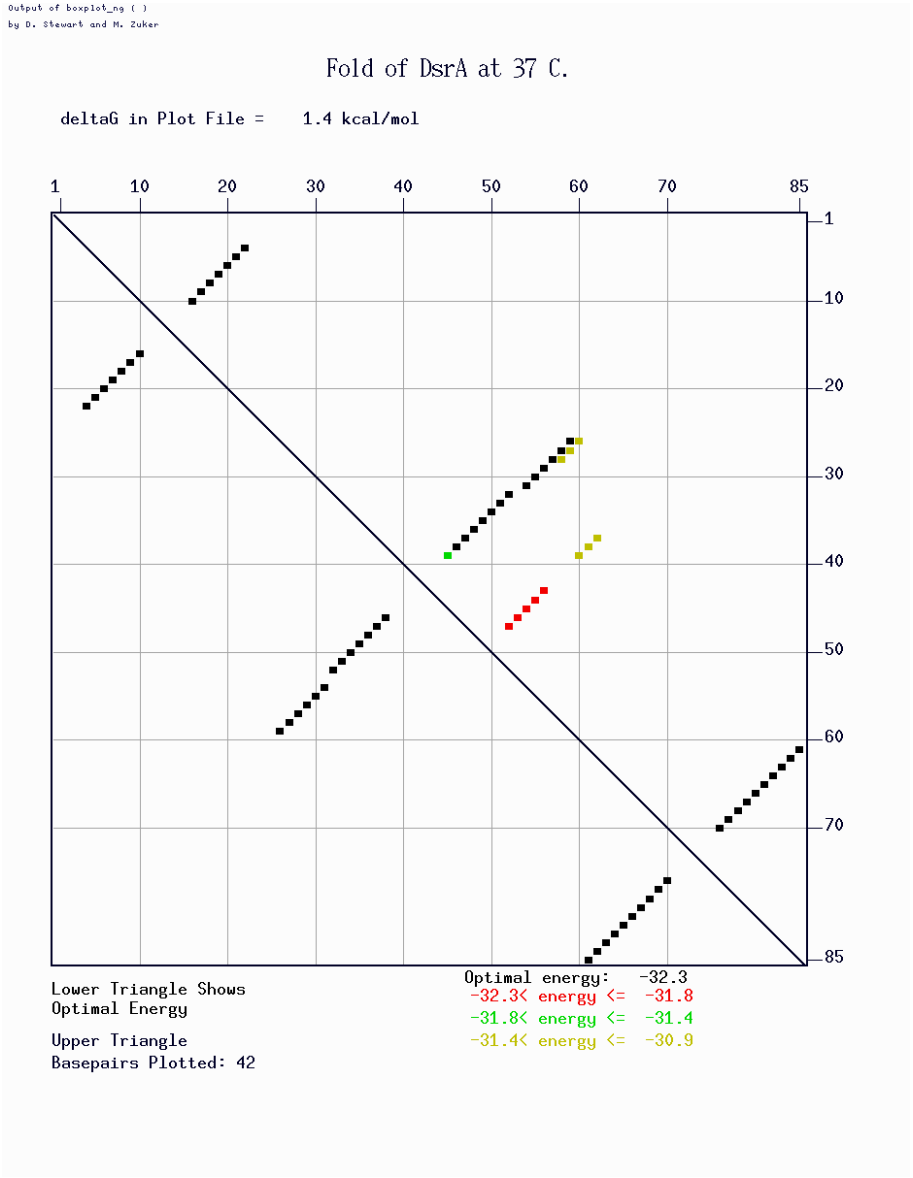
Basenpaare, die in allen Strukturen auftauchen, werden durch einen schwarzen Punkt markiert. Diejenigen, die in zwei oder mehr Strukturen aber nicht in allen vokommen, sind grau. Punkte, die nur zu einer bestimmten Struktur gehören erhalten eine eindeutige Farbe. Das System der eindeutigen Farbzuordnung funktioniert natürlich nur bis zu einer begrenzten Anzahl an Strukturen (hier 15).

Werden nur wenige Strukturen miteinander verglichen und soll sofort erkennbar sein zu welcher Struktur ein Punkt gehört, so muss die Option *Mutiple Overlap*, die sich oberhalb des Strukturdotplots befindet, angeklickt werden. Das führt dazu, dass graue Punkte durch mehrfarbige Punkte ersätzt werden und die Farben eindeutig einer Struktur zuzuordnen sind. Da die relativ kurze DsrA RNA Sequenz nur drei unterschiedliche Strukturen hervorbringt, wird der Struktur-Dotplot automatisch mit dieser Option visualisiert.

Wählt man für den Vergleich das jpg oder png-Format, so können einzelnen Punkte des Dotplots angeklickt werden und man erfährt, in welchen Strukturen diese Paarung auftritt.



c) Im Gegensatz zum Struktur-Dotplot stellt der *Energy Dotplot* nicht nur die ausgewählten (unterschiedlichen) Strukturen, sonder alle Strukturen des festgelegten Energiebereichs dar. Er ist deshalb auch umfassender.



Ein Punkt in der i-ten Zeile und j-ten Spalte steht wieder für eine Paarung der Basen an i-ter und j-ter Position in der Sequenz. Vier Farben kodieren vier Level der Faltungsenergie. Schwarze Punkte bedeuten, dass diese Paarungen zu optimalen Faltungen gehören. Die Level 2-4 beschreiben entsprechend schlechtere Energiebereiche. An dieser Stelle sollten Sie sich daran erinnern, dass je niedriger die Faltungsenergie ist, umso stabiler die resultierende Struktur.

Im unteren Dreieck der Darstellung werden auschliesslich Paarungen, die Strukturen mit minimaler freier Energie gehören, markiert. Es kann übrigens mehr als nur eine optimale Struktur geben. Im oberen Dreieck sind Paarungen, je nach Zugehörigkeit zu einem der Energielevel, mit einer bestimmten Farbe dargestellt.

Lösung 4 Nutzung von *constraint information*.

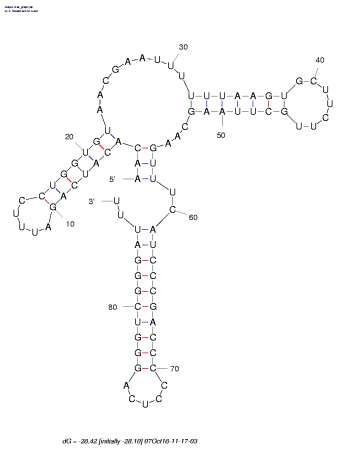


Abbildung zu a)

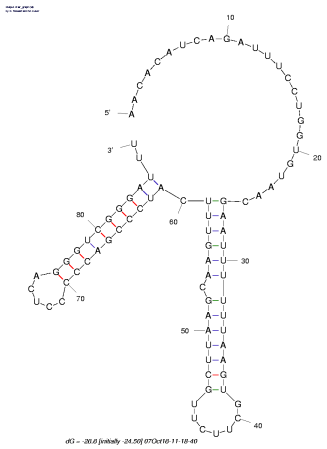


Abbildung zu b)

a) Möchten Sie erreichen, dass ein bestimmter Abschnitt einer Sequenz wenn möglich gepaart wird, dann muss im Eingabefeld neben *constraint information* die Option:

F 1 0 3

eingegeben werden.

b) Um die Paarung der Basen 4 bis 10 zu verhindern, muss *constraint information*

P 4 0 7

gesetzt werden.

c) Weiterhin gibt es die Möglichkeiten:

- zwei Abschnitte miteinander paaren zu lassen, wei die folgenden Paare entstehen sollen: (i,j),(i+1,j-1),...(i+n-1,j-n+1)

  F i j n

- das Paaren zweier Abschnitte zu verhindern, wobei die folgenden Paarungen nicht zustande kommen dürfen: (i,j), (i+1, j-1),...(i+n-1,j-n+1)

  P i j n

- Paarungen zwischen zwei Abschnitten zu verhindern, wobei die Abschnitte auch nicht versetzt miteinander paaren dürfen

  P i-j k-l

Der Buchstabe F steht für Force, also Zwang, wobei P für Prohibit steht und für ein Verbot bedeutet. Die Indizes i, j, k und l geben jeweils eine Position in der Sequenz an und n die Länge des betrachteten Bereichs.

`RNAalifold` : Vorhersage einer gemeinsamen Sekundärstruktur für eine Schar von Sequenzen

RNAalifold berechnet eine evolutionär konservierte Struktur einer RNA Sequenz, indem es die konservierten Strukturinformationen in einem Alignment ausnutzt, um die beste gemeinsame Struktur zu bestimmen. Um die Ausgabe von RNAalifold zu verdeutlichen, sehen Sie sie zum Vergleich direkt unter dem Alignment. In den ersten vier Zeilen stehen die mit ClustalW alignierten Sequenzen. Darunter kommt die mit RNAalifold berechnete Konsensussequenz und die gemeinsame Struktur im Vienne-Format. (Eine Konsensussequenz wird aus einem Alignment bestimmt, indem aus jeder Spalte, der am häufigsten vorkommende Buchstabe genommen wird.)

```
AUCCGAAGCGAAAGCGUCGGGAUAAUAAUAACGAUGA----AAUUCCUCUUUGACGGGCCAAUAGCGAUAUUGG----------
AUCCGAAGCGAAAGCGUCGGGAUAAUAAUAACGAUGA----AAUUCCUCUUUGACGGGCCAAUAGCGAUAUUGGCCAUUUUUUA
AUCCGAAGCGAAAGCGUCGGGAUAAUAAUAACGAUGA----AAUUCCUCUUUGACGGGCCAAUAGCGAUAUUGGCCAUUUUUUU
AUCCGCAAGGAGUGUGAGUGCUUAAUAACAAAAAUAAUGAAAAUUCCUCUUUGACUGGCCGGUAGUGAUAUCGGCCAUUUUUUU

AUCCGAAGCGAAAGCGUCGGGAUAAUAAUAACGAUGA____AAUUCCUCUUUGACGGGCCAAUAGCGAUAUUGGCCAUUUUUUU
.((((((.((....)).)))))).....................................(((((((((....)))))))))........
```

Im Gegensatz zu rein thermodynaischen Methoden, welche die Struktur nur auf der Grundlage der Faltungsenergie vorhersagen können, nutzt RNAalifold zusätzliche, im Alignment konservierte Informationen. Die thermodynamische Strukturvorhersage birgt ein Problem: Die thermodynamisch optimale Struktur muss nicht die reale Struktur einer Sequenz sein. Obwohl sich die Faltungsenergie der realen Struktur eventuell nur gering von der optimalen Faltungsenergie unterscheidet, macht es eine eindeutige Strukturzuordnung schwierig. Zusätliche Informationen zur Identifikation der richtigen Struktur sind höchst willkommen und können aus der Sequenzkonservierung im Alignment entnommen werden.