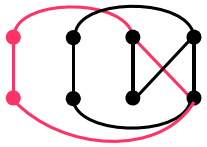
Structural Alignment of two RNA Sequences with Lagrangian Relaxation

RNA ist in der Regel ein einzelsträngiges Molekül, wobei die verschiedenen Basen innerhalb eines Moleküls sog. Wasserstoffbrueckenbindungen ausbilden können, so dass letztendlich verschiedene Sekundärstrukturen entstehen. Will man nun funktionelle Motife (konservierte Strukturen) finden, die biologisch relevant sind, dann muss man in einem Alignment eben auch die Interaktionen innerhalb eines RNA-Moleküls berücksichtigen, da funktionell verwandte RNAs häufig eine geringe Sequenzähnlichkeit dafür aber eine sehr ähnliche Struktur aufweisen.

Ein strukturelles Alignment für 2 RNA-Sequenzen (S_1 , S_2) kann in Form eines Graphen $G=(V, A \cup I, w)$ dargestellt werden:



- Knoten (V): A, C, G, U, - in den Sequenzen S_1 , und S_2
- Alignmentkanten (A): zwischen den beiden Sequenzen
- Interaktionskanten (I): Basenpaarungen innerhalb einer RNA Sequenz (rot)

Allerdings muss gewährleistet werden, dass die Interaktionskanten nur zwischen 2 Basen verlaufen und nicht etwa zwischen einer Base und einem Gap oder umgekehrt. Außerdem muessen die paarweisen Interaktionskanten in S_1 und S_2 durch Alignmentkanten realisiert werden, d.h. in dem gültigen Alignment müssen die Alignmentkanten l und m, die zwei Interaktionskanten verbinden auch tatsächlich enthalten sein. Desweiteren haben sowohl die Alignmentkanten als auch die Interaktionskanten ein positives Gewicht w.

Das strukturelle Alignment lässt sich recht einfach als ILP formulieren:

Die Zielfunktion des ILP beschreibt den zu maximierenden Score. Er berechnet sich aus der Summe der Alignmentkanten x_l und den realisierten Interaktionen y_{ml} :

$$\max \sum_{m \in A} \sum_{l \in A} w_{lm} y_{lm} + \sum_{m \in A} w_m x_m$$

$$x \in \{0,1\}^A, y \in \{0,1\}^{A \times A}$$

Die hier verwendeten Variablen x_l und y_{ml} sind 1 wenn die entsprechende Alignment- bzw. Interaktionskante in dem aktuellen Alignment vorhanden ist und ansonsten 0.

Für die Zielfunktion müssen nun die folgenden 3 Bedingungen gelten:

$$\sum_{l \in I} x_l \leq 1 \qquad \forall I \in \mathcal{I} \qquad (1)$$

$\mathcal{I}$ enhält dabei Teilmengen von Alignmentkanten, die sich paarweise ausschließen, und damit nicht gemeinsam in einem gültigen Alignment vorkommen können. Das betrifft Kanten, die von der gleichen Base ausgehen oder sich kreuzen. Die Bedingung gewährleistet, dass höchstens eine Kante aus solch einer Menge $\mathcal{I}$ in einem Alignment enthalten ist.

$$y_{lm} = y_{ml} \qquad \forall l, m \in A, l < m \qquad (2)$$

Die Interaktionen müssen durch Alignmentkanten (x_l , x_m) realisiert werden, wobei letztendlich für beide Alignmentkanten die gleichen Interaktionen ausgewählt werden sollen.

$$\sum_{l \in A} y_{lm} \leqslant y_m \qquad \forall m \in A \qquad (3)$$

Ist die Alignmentkante x_m im aktuellen Alignment vertreten, so kann eine oder eben auch keine Interaktion mit der entsprechenden Base bestehen. Für alle potentiellen Alignmentkanten, die eben nicht im besten Alignment vorhanden sind, ist dann natürlich keine Interaktion möglich.

Das ILP versucht man erst mal zu lösen, indem das Problem etwas reduziert wird. Hier wurde zunächst die Bedingung für die Interaktionskanten (2) entfernt und ein herkömmliches Alignment berechnet:

Und zwar wird vorab für jede Alignmentkante der maximale Profit aus dem ursprünglichen Gewicht w plus dem Gewicht für die jeweils beste Interaktion bestimmt. Das zu berechnende Alignment liefert dann die Alignmentkanten $\bar{x}$. Die dazugehörigen Interaktionskanten $\bar{y}_{ml}$ ergeben sich dann aus $\bar{x}_l * \hat{y}_{ml}$ wobei $\hat{y}_{ml}$ die Interaktionskanten sind, die vorher für die Berechnung des maximalen Profits der jeweiligen Alignmentkanten herangezogen wurden. Der auf diese Art berechnete Score ist eine mögliche obere Grenze (UB) fuer das ILP.

Enthält ein ILP Bedingungen, die es schwer lösbar machen, dann kann man eben solche Bedingungen in die zu optimierende Zielfunktion integrieren. Für das strukturelle RNA Alignment wurde die Bedingung für die Interaktionskanten (2) in die Zielfunktion eingefügt:

$$\max \sum_{m \in A} \sum_{l \in A} w_{lm} y_{lm} + \sum_{m \in A} w_m x_m + \sum_{l \in A} \sum_{m \in A, l < m} \lambda_{lm} (y_{lm} - y_{ml}) \qquad (4)$$

Das Problem besteht nun darin für jede gültige Alignmentkante einen geeigneten Faktor λ zu finden:

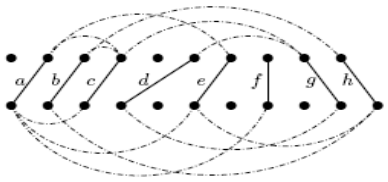
$$\lambda_{lm}^{i+1} = \begin{cases} \lambda_{lm}^i & \text{if } s_{lm}^i = 0 \\ \max(\lambda_{lm}^i - \gamma_i, -w_{lm}) & \text{if } s_{lm}^i = 1 \\ \min(\lambda_{lm}^i + \gamma_i, w_{lm}) & \text{if } s_{lm}^i = -1 \end{cases}$$

$$s_{lm}^i = \bar{y}_{lm} - \bar{y}_{ml} \text{ for all } l, m \in A, l < m$$

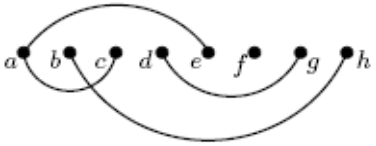
Im ersten Durchlauf wird der maximale Score fuer $\lambda_{lm}^0 = 0 \; \forall m, l \in A$ bestimmt. In jeder Iteration wird dann für jede Alignmentkante, die im aktuellen Alignment vorhanden ist der Faktor λ neu bestimmt, je nach dem ob für diese Kante die zweite Bedingung verletzt wurde oder nicht (s_{lm}^i). Die neuen λ_i hängen neben den bekannten Gewichten (w) auch von der Schrittgröße γ_i ab, die in jedem Durchgang neu bestimmt wird:

$$\gamma_i = \mu \frac{\mathrm{UB} - \mathrm{LB}}{\sum_{m,l \in A} (s_{lm}^i)^2}$$

Mit den aktuellen λ_i wird das Maximum für die Zielfunktion erneut berechnet. Der Score für das strukturelle Alignment sollte dann recht schnell gegen ein Optimum konvergieren.
Die in der Iteration verwendete Schrittgröße benötigt u.a. eine untere Grenze (LB) für das ILP. Diese lässt sich recht einfach bestimmen, indem man sich den Interaktionsgraphen anschaut:



Alignmentgraph



Interaktionsgraph

Der Interaktionsgraph ergibt sich aus dem Alignmentgraphen wobei alle Alignmentkanten als Knoten dargestellt werden und dann auch nur jeweils eine Interaktionskante übernommen wird wenn die Bedingung zu den Interaktionskaten (2) tatsächlich erfüllt ist. Anschließend muss noch eine gültige Menge von Interaktionskanten bestimmt werden für die der Score maximal ist.

Die hier vorgestellte Methode ermöglicht es also einen Alignmentscore für zwei RNA Sequenzen mit bekannter Sekundärstruktur zu berechnen, d.h. man bekommt eine Aussage darüber ob bzw. wie ähnlich die RNA Strukturen sind. Außerdem kann man eben auch ein Alignment für RNA Sequenzen mit unbekannter Sekundärstruktur (Funktion) bestimmen und so bei einem hohen Alignmentscore, der ja für eine ähnliche bzw. konservierte Struktur spricht, auf eine mögliche Funktion schließen, da das Verfahren letztendlich eine gemeinsame Sekundärstruktur für zwei RNA Sequenzen bestimmt.

Mit diesem Verfahren können jetzt auch sog. Pseudoknoten gefunden werden und wie die Ergebnisse zeigen, können die Alignments deutlich schneller als mit bisherigen Verfahren berechnet werden. Außerdem kann das Verfahren wohl recht gut für ein multiples Alignment erweitert werden. Das Problem hierbei ist allerdings, dass der berechnete Score nur eine Annäherung an das tatsächliche Optimum darstellt. Verbesserungen lassen sich ggf. erzielen indem der hier verwendete Ansatz mit anderen Methoden wie „branch and bound“ kombiniert wird.

Anhang:

Lagrangefunktion:

Eines der wichtigsten Werkzeute der Optimierung ist die Lagrangefunktion, die dazu dient, eine gewisses „Gleichgewicht“ zwischen der Zielfunktion und den Nebenbedingungen zu beschreiben.

Sei nun das folgende Optimierungsproblem mit nur einer Bedingung gegeben

$$\inf\{f_0(x) \mid f_1(x) \leq 0\} \qquad (5)$$

Dann führt man zu jedem Parameter y $y \geq 0, y \in \mathbb{R}$, Hilfsprobleme ein, die von y abhängen:

$$\inf\{f_0(x) + y f_1(x) \mid x \in \mathbb{R}\} \qquad (6)$$

Der Parameter y beschreibt das Gewicht, das man der Erfüllung der Nebenbedingung $f_1(x) \leq 0$ beimisst. Wir nehmen an, dass (6) für jedes festes $y \geq 0$ eine Optimallösung $x^*(y)$ besitzt.

Für $y = 0$ wird der Optimalpunkt $x^*(0)$ die Nebenbedingung $f_1(x) \leq 0$ im Allgemeinen verletzten, es sei denn, diese Nebenbedingung war „überflüssig“.

Wenn man y aber sehr groß wählt, wird das Hauptgewicht des Problems (5) bei der Minimierung von f_1 liegen; In der Regel wird dann $f_1(x^*(y)) \leq 0$ gelten und $x^*(y)$ wird für (5) nicht optimal sein.

Lässt man nun, beginnend bei $y = 0$, den Wert von y langsam wachsen und verfolgt die zugehöringen Lösungen $x^*(y)$, so wird es einen Zwischenwert bzw. „Gleichgewichtspunkt“ $\bar{y} > 0$ geben, für den $f_1(x^*(\bar{y})) = 0$ gilt. Dann löst $x^*(\bar{y})$ auch (5).