Bioinformatik Modul
Transkriptionsregulation

Andrej Aderhold, Jakob Mhmel

Oktober 2005

Zusammenfassung

Im menschlichen Genom kommen etwa 25000 Gene vor, die lediglich 5 % der gesamten DNA Sequenz ausmachen. Etwa 45 % bestehen aus DNA, die funktionaler Natur ist. Darunter fallen die Transkriptionfaktor Bindestellen (TFBS). Das sind cis-wirkende kurze DNA Sequenzen mit i.d.R. 6 - 12 Basenpaare die als Rezeptoren für Transkriptionfaktoren dienen und in dem regulatorischen Bereich eines Gens vorkommen. Die TFBS entwickeln eine zentrale Rolle bei der Regulation von Genen. Vom besonderen Interesse sind die Regulationsrezeptoren an den Hox Genen, die in der Ontogenese eine entscheidende Rolle spielen. In diesem Praktikum untersuchen wir die Regulatorregion von Hox Genen des Zebrafischs, Kugelfisch und Grünen Kugelfisch auf DNA Motive, die Bindungsmotive sein können.

Inhaltsverzeichnis

1 Aufgabenstellung

Für die TFBS der Hoxgene im Zebrafisch, Kugelfisch und Günen Kugelfisch wurde keine Literaturinformation gefunden. Allerdings sind die DNA Sequenzen in Hox Clustern gut konserviert. Die Bindungsmotivdatenbank TRANSFAC und JASPAR beinhalten bekannte TFBS u.a. für Mensch und Maus. Wir werden versuchen Sequenzen ausfindig zu machen, die ein Analog in diesen Datenbanken haben und Motive lokalisieren, die unbekannt aber überpräsent sind. Wir untersuchen die folgenden Hoxcluster. Je Cluster sind etwa 7 - 10 Hoxgene enthalten.

- ▷ Danio rerio: Aa,Ab,Ba,Bb,Ca,Cb,D
- ▷ Fugu rubripes: Aa,Ab,Ba,Bb,Ca,Da,Db,
- ▷ Tetraodon nigroviridis: Aa,Ab,Ba,Bb,Ca,Da,Db

Vorrangiges Ziel ist es verschiedene Tools zur Motiverkennung zu testen und gegenüber zustellen. Bei der Suche nach Motiven wird nach Sequenzmustern gefahndet, die in einer Gruppe von DNA Sequenzen vorkommen. Ein Motiv kann womöglich einer TFBS aus der TRANSFAC Datenbank zugeordnet werden. Allerdings können die Motive auch keinerlei weitere funktionale Bedeutung haben oder sie stellen Sequenzen da deren Funktion noch nicht bekannt ist. Wir werden in den Datenbanken auch direkt mit den DNA Sequenzen nach Motiven suchen die eine TFBS assozieren. Die Tools tfsearch und pwmatch übernehmen diese Aufgabe. Des weiteren werden wir ein lokales Alignment mit dialign über eine Menge von Sequenzen vornehmen um überrepräsentierte DNA Fragmente ausfindig zu machen. Die Ergebnisse all dieser Methoden lassen wir in einem Vergleich einfliessen.

Die Motivsuche in den Sequenzen ermöglicht die Erstellung von 'Position Weight Matrices' (PWM). Diese geben das relative Gewicht für die 4 Nukleotide auf jeder Position des Motivs an. Mit diesen Matrizen kann in der TFBS Datenbank nach bekannten Motiven gesucht werden, die auch als Matrizen vorliegen. Mit dem Resultat können wir einen Verwandtschaftsbaum erstellen, der die Ähnlichkeit von Hox Promotern bezogen auf das vorkommen von Motiven beschreibt. Es folgt eine Liste von Programmen zur Motivsuche:

- ▷ Meme (http://meme.sdsc.edu/meme/meme-intro.html)
- ▷ AlignAce (http://atlas.med.harvard.edu)
- ▷ Bioprospector (http://ai.stanford.edu/ xsliu/BioProspector)
- ▷ rVista (http://rvista.dcode.org)
- ▷ Consensus (http://bifrost.wustl.edu/consensus)
- ▷ YMF (http://wingless.cs.washington.edu/YMF/YMFWeb/YMFInput.pl)

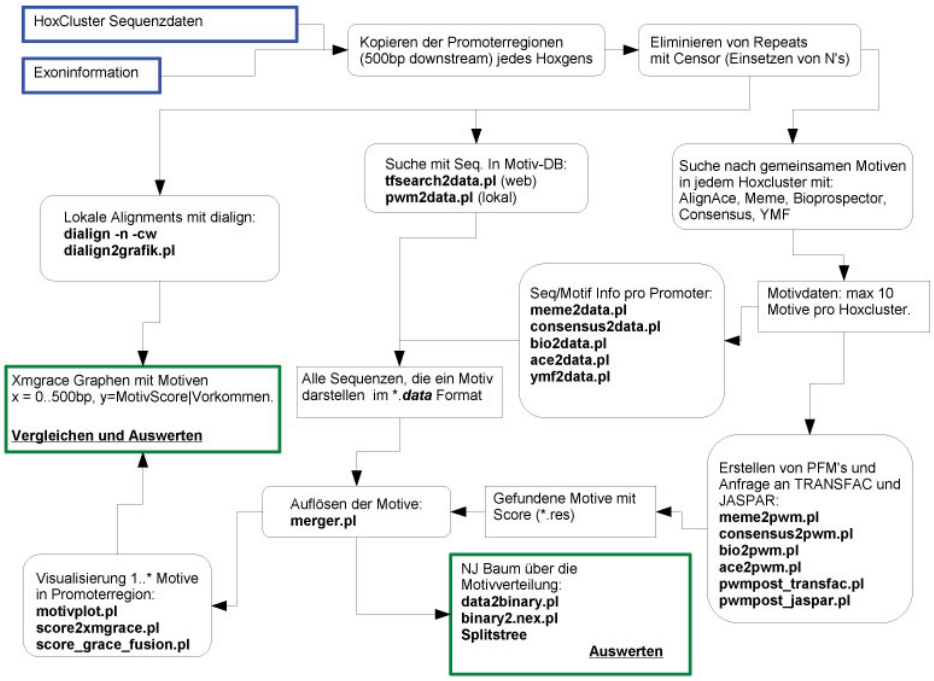

Abbildung 1: Durchführung Übersicht

2 Durchführung

Das folgende Flussdiagramm (Abbildung 1) zeigt den Ablauf der Arbeitsschritte. Die zwei fett gedruckten Ksten links oben beinhalten die Ausgangsdaten (Hox Cluster Sequenz und Exoninformation). Die abgerundeten Kästen sind Aktionen, die Eckigen beinhalten Daten. In die zwei anderen fett gedruckten Ksten fliessen die Endresultate ein.

2.1 Regulationsbereich ausschneiden

Mit Hilfe der Exoninformation aus den .exon Dateien werden aus den Hoxclustern, die als .fasta Dateien vorliegen, 500 Basen der upstream gelegenen regulatorischen Region heraus kopiert. Dazu benutzen wir das Script cutprom.pl und übergeben als Parameter die Fasta Datei mit dem HoxCluster. Die Ergebnisse befinden sich hier.

Wir haben im gesamten Verlauf der Durchführung mit 500 Basen gearbeitet. Allerdings ist dies nur ein Ausschnitt vom Sequenzbereich der Einfluss auf die Regulation des nachfolgenden Gens hat. Besser sind 2500 Basen.

2.2 Repeats eleminieren

Die Sequenzen werden in eine einzelne Datei zusammengefasst und an das Censor Webinterface (http://www.girinst.org/censor) geschickt. Censor sucht nach bekannten/auffälligen Repeats und ersetzt diese durch den Buchstaben N. Das Resultat wird anschliessend so zerschnitten, dass wir für jedes HoxCluster eine .fasta Datei erhalten. Dies geschieht durch das Script splitCensor.pl.

2.3 Lokales Alignment mit Dialign

Mit dem Tool dialign werden über die Menge aller Sequenzen (des kopierten Regulierungsbereichs) eines Clusters lokale multiple Alignments durchgeführt. Die Graphische Darstellung der Resultate geschieht mittels dem Tool dialign2grafik.pl, dessen Ausgabe man sich mit xmgrace (.grace) oder einem Postscript Leser anschauen kann.

```
dialign2-2 -n -cw <.fasta>
dialign2grafik.pl <*.ali>
```

Der entstandene Graph bildet auf der x-Achse die Basenpaare 0..500 ab. Die y-Achse gibt den Grad der Sequenzähnlichkeiten für die entsprechende x-Position an.

2.4 Motivsuche

Über die GUI vom Tool BEST werden folgende Programme ausgeführt: Consensus, Meme, AlignAce, Bioprospector. Für YMF wird das Webinterface benutzt. Als Eingabe dient eine .fasta mit den Sequenzen eines Clusters. Die maximale Anzahl an Motiven, die gefunden werden sollen beträgt 10. Die Programme ermitteln ähnliche Sequenzfragmente über die Menge der gegebenen Sequenzen und errechnen ein Score. Das sind die verschiedenen Scores der Motive: AlignAce ('Map-Score'), Meme (P-Value), Consensus (P-Value), YMF (Z-Score), Bioprospector ('Motif Score'), Tfsearch ('Z-Score').

2.5 Erzeugen von .data Format und PWM

Das .data Format dient dazu die Sequenzinformation aller Tools in eine einheitliche Form zur Weiterverarbeitung zu bringen. Dazu werden aus den Motiv-Daten alle in den Motiven vorkommenen DNA Sequenzen extrahiert und mit zusätzlicher Information in eine tabelarische Form geschrieben. Jeder Promoter besitzt einen Datensatz. Das Format ist

start	end	score	motif	DNA	Seq	tool	fw/rev	hox

Wobei 'start' und 'stop' die Grenzen der DNA Sequenz im Hox Promoterbereich angeben. Die 'score' bezieht sich auf das Motiv, das mittels 'tool' gefunden wurde. Wir erstellen für jedes Motivgenerierende Tool einen seperaten Datensatz. Diese Skripte gibt es hier.

```
meme2data.pl <.meme>
consensus2data.pl <.con_10>
bioprospector2data.pl <.biop_10>
ace2data.pl <.aa_10>
tfsearch2data.pl <.fasta>
pwmatch2data.pl <.fasta>
```

Das Skript tfsearch2data.pl schickt eine .fasta Dateien an die Adresse http://www.cbrc.jp. Aus dem Ergebnis (Motiv aus DB) werden dann die .data Dateien erzeugt. Das Skript pwmatch2data.pl sucht in einer lokalen TRANSFAC Datenbank nach Motiven.

Die PWM (Position Weigth Matrix) werden zu jedem Motiv erstellt. Die Resultate der Programme die unter eingesetzt wurden benötigen jeweils ein spezielles Programm um eine PWM zu generieren, bzw. herauszuextrahieren. Die PWM's werden schliesslich mit den Skripten pwmpost_transfac.pl und pwmpost_jaspar.pl an die TRANSFAC und JASPAR Datenbank von MatCompare (http://rulai.cshl.edu/cgi-bin/MatCompare/home.cgi?process=home) geschickt. Das Resultat sind bekannte Motive aus den Datenbanken, die Rezeptoren darstellen.

2.6 Auflösen (Mergen) der Motive

Die m.H. der PWM Matrizen gefundenen Motive müssen mit den schon bestehenden Datenbestand der Motive abgeglichen werden. Dabei werden die Motivnummern aller Sequenzfragmente, die ein bekanntes Motiv darstellen durch die Nummern aus den Datenbanken ersetzt. Es kommt auch vor, dass für eine PWM mehr als eine Motivnummer in den Datenbanken gefunden wurden. In diesem Falle werden die Motivdatensätze mit zusätzlichen Einträgen erweitert. Bei der späteren graphischen Auswertung wird es u.a. darum gehen Motive die in der DB gefunden wurde und Motive, die nicht gefunden wurden farblich zu unterscheiden.

Das Tool merger.pl nimmt eine Datei im Format .res (MatCompare Rückgabe) entgegen und liest alle dazu passenden .data Dateien aus dem Arbeitsverzeichnis raus. Mit dem Paramter -m kann man bestimmen ob nicht gefundene Motive in den neuen Datensatz übernommen werden. Das Ergebnis sind Dateien mit der Endung .merged.

```
merger.pl -m <.res>
```

2.7 NJ Baum erstellen

Über die Motivverteilung in allen Clustern wird ein NJ Baum m.H. von SplitsTree erstellt. Das Programm data2.binary.pl erstellt einen fasta konformen Eintrag für jeden Promoter. Alle in den Clustern vorkommenden Motive werden gesammelt und sortiert. Die Sequenz unter dem fasta header ist ein String aus Nullen und Einsen. Jede Position gibt an ob ein Motiv aus der Sammlung aller bekannten Motive in diesem Promoter vorkommt oder nicht. Beispiel:

>DrAa_hoxa1a_80301-81442
0110011100101011100000111001

Die Motive wurden in diesem Fall vorher mit AlignAce ermittelt. Ein ähnliches Bindungsmotiv wurde in der TRANSFAC gefunden. Die ersten 4 Motive die global gefunden wurden waren M00010, M00014, M00018 und M00019. In Danio rerio, HoxCluster Aa, Gen 1 sind M00014 und M008 vorgekommen.

Die Erstellung von Bäumen über wenige Promoter bringt i.d.R. Bäume mit sehr geringer Verzweigung, da dann wenig Motive vorkommen.

2.8 Motivplots erstellen

Die .merged Daten werden mit xmgrace visualisiert. Es werden Graphen erstellt, die auf der x-Achse die Basenpositionen 0..500 abbilden und auf der y-Achse den Score eines Motivs (Scoregraph) oder ein Plot seines Vorkommens (Motivplot). Beide Varianten müssen mit jeweils einem Skript erstellt und anschliessend zusammen gefügt werden. Für den Scoregraph wird das Skript score2xmgrace.pl und für den Motivplot motivplot.pl eingesetzt. Die Skripte haben folgende Parameter:

```
motivplot.pl [-m] [-p] list-of-files
score2motiv.pl [-m] [-p] list-of-files
score_grace_fusion.pl
```

Wobei -t angibt ob die Stufen von Align-Ace gelöscht werden sollten. Der Parameter -p gibt an ob der Hox Gen Name oder das Programm, das zum ermitteln der Motive benutzt wurde, in die Legende gezeichnet werden soll. Das Script score_grace_fusion.pl erstellt aus den Daten, die von den anderen Skripten in den Verzeichnissen graphplot und motivplot abgelegt wurden, einen zusammengesetzten Graphen.

Die Graphen werden mit den Ergebnissen von dialign verglichen. Es werden auch Graphen erzeugt, die die Ergebnisse von den verschiedenen Tools darstellen.

3 Ergebnisse

3.1 Motivbäume

Wir generieren die NJ Bäume über die Menge aller gefundenen Motive für das jeweilige Tool. Die Bäume für Tfsearch und Pwmatch sind sich in der Art der Verzweigung und der Verteilung der Motive änlich (Abbildung 2, 3). Das war auch anzunehmen, da sie beide Bindungsstellen in der TRANSFAC Datenbank über die fasta Sequenzen suchen. Charakteristisch sind hier das eng beieinander auftretende Vorkommen von gleichen Hoxpromotern in den beiden Kugelfischen (z.B. Hox Ab1 in Fr und Tn). Promoter aus dem Zebrafisch kommen aber auch in einigen Fällen zusammen mit den Kugelfischen vor (Hox Ca9 in Fr,Tn und Dr). Diese werden wir im Abschnitt 3.2 näher betrachten. Allerdings ist Aufgrund der starken, sehr gleichmässigen Verzweigung und der auffälligen Zentralisierung eine Interpretation der Daten schwierig.

Die Bäume von Meme und AlignAce haben eine enge Anordung von Regulationsbereichen der gleichen Organismen und Hox Clustern. Sie zeigen keine Ähnlichkeit zu denen von Tfsearch und Pwmatch. Das liegt an der Durchführung der Motivsuche. Wir haben den beiden Programmen jeweils alle Sequenzen eines Hox Clusters übergeben. Aus dieser Menge von Sequenzen wurden die Motive ermittelt. Die Abbildungen 4 und 5 zeigen letztenendes nur unsere Vorgehensweise, also wie Meme und AlignAce gearbeitet haben. Die Cluster, die wir den Programmen übergeben haben, bilden Motiv Cluster in den Bäumen. Damit scheinen die Ergebnisse nicht weiter von Bedeutung zu sein. Die Abbildung 6 zeigt den Baum von Consensus. Dort wurden nur Motive im Zebrafisch gefunden. Generell hat Consensus

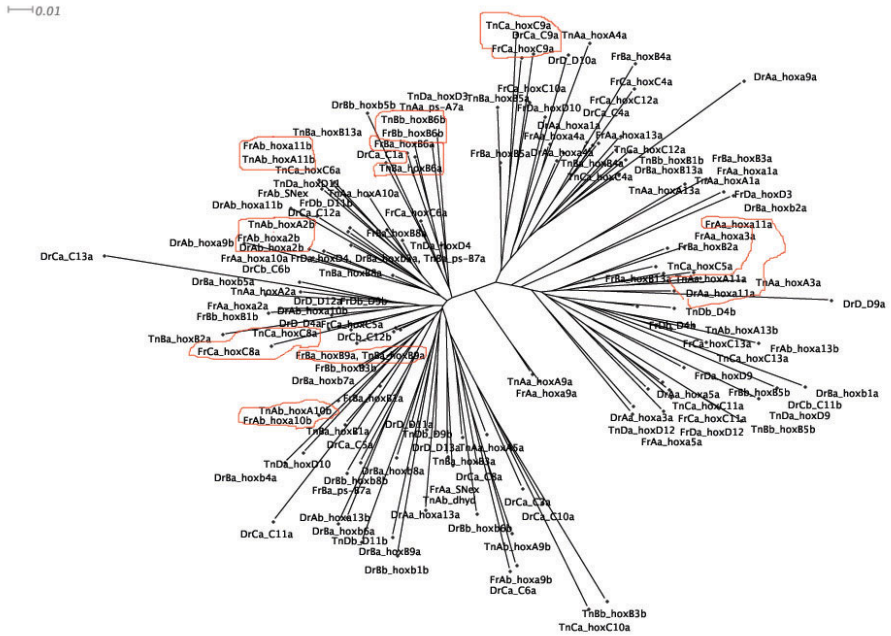

Abbildung 2: Tfsearch

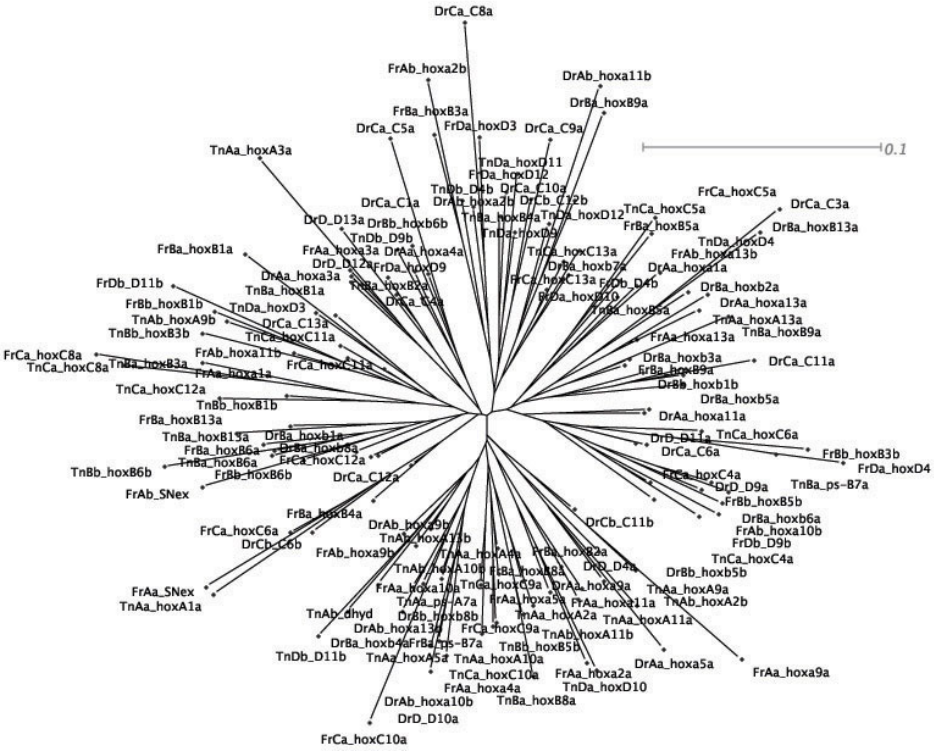

Abbildung 3: PWMatch

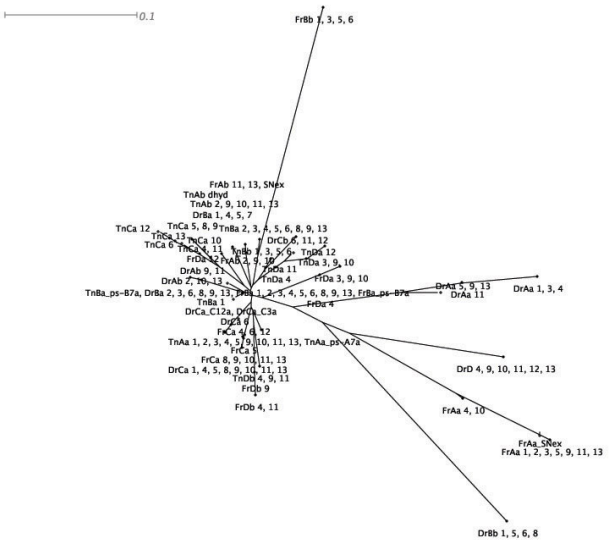

Abbildung 4: Meme

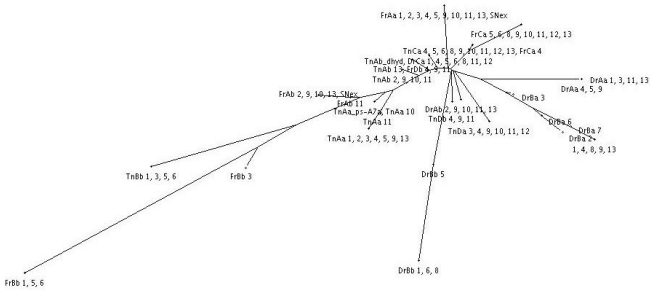

Abbildung 5: AlignAce

nur sehr auffällige (u.a. USF) Motive gefunden.

3.2 Motivplots

Ausgehend aus NJ Bäumen von Tfsearch aus 3.1 untersuchen wir die Motivverteilung für das Hox Aa11 in Zebrafisch und den beiden Kugelfischen. In Abbildung 7 kann man einen Motivplot über die 500 Basen stromaufwärts erkennen (Position 501 entspricht dem Transkriptionsstart des Hoxgens). Auffällig ist die Ansammlung von Motiven in den Bereich um Basenpaar 325 (8,10 Motive) und 350 (2 Motive). Die folgenden Motive kommen bei Tn und Fr gleich vor: 121 (fw/rev), 122(fw/rev), 217(fw/rev), 236 (fw/rev), 101 (fw), 75 (fw), 76 (fw). Bei Danio rerio fehlen 236, 101, 75 und 76. Die Motive 236 und 101 schliessen sich unmittelbar hintereinander an und sind als etwas längerer Strich dargestellt. 75 und 76 sind die 2 Striche auf der Höhe von Basenpaar 350.

Die Motive 121,122 und 217 binden den stimulierenden Faktor USF (Upstream stimulatory factor), der u.a. in Ratte, Maus und Mensch identifiziert wurde. Informationen zu den einzelnen Motiven befinden sich in diesem TRANSFAC Dokument.

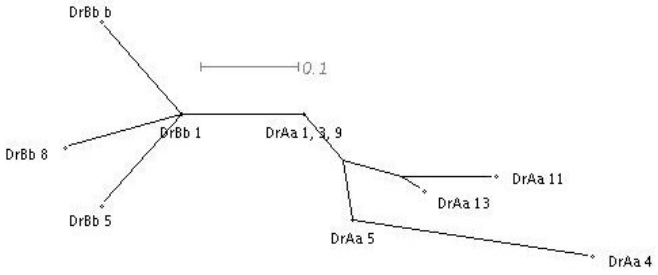

Abbildung 6: Consensus

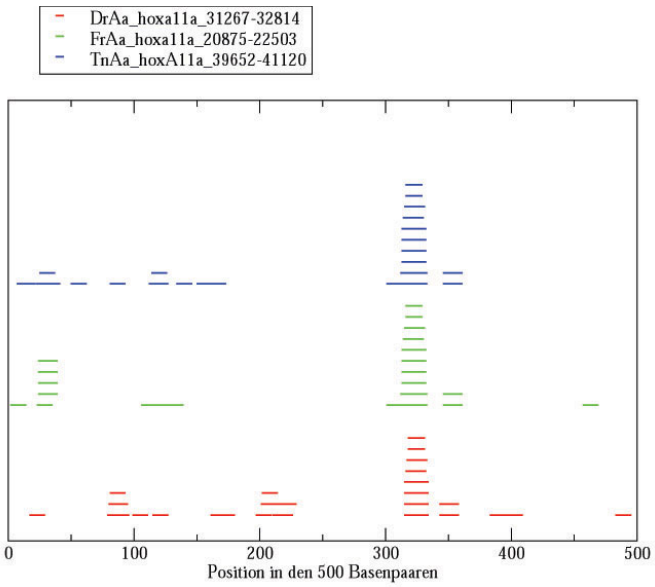

Abbildung 7: 500 bp upstream von HoxAa11 Transkriptionstart

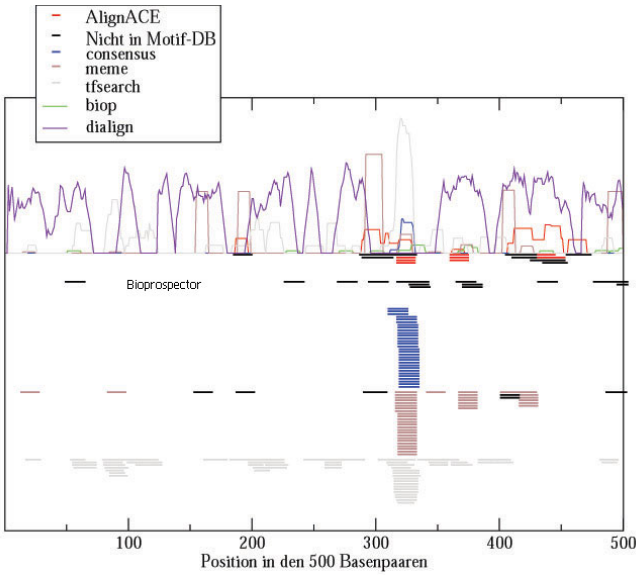

Abbildung 8: Vergleich der gefundenen Motive

3.3 Programmvergleich

In Abbildung 8 werden Motive wiedergegeben, die durch verschiedenene Programme gefunden wurden. Die Regulatorische Region des Hoxgens Aa11 vom Zebrafisch wurde untersucht. Jedes Programm besitzt eine Farbe und sind vertikal in Banden angeordnet. Die schwarzen Striche sind Motive, die durch die Programme gefunden wurden aber nicht in der TRANSFAC Datenbank in MatComapre. Farblich hervorgehoben sind Motive, die auch in der Datenbank erkannt wurden.

Auffallend ist, dass Meme, Consensus und Tfsearch die gleiche Tendenz bei der Anzahl der gefundenen Motive im Bereich der Basenpaare 325 zeigen. Allerdings hat Consensus nur dort Ergebnisse und diese beschränken sich auch nur auf den Zebrafisch (siehe Abbildung 6). AlignAce hat ebenfalls eine geringe Ansammlung von gefundenen Motiven und Bioprospector hat in diesem, sowie in weiteren Plots keine, bzw. kaum TRANSFAC Datenbank hits. Alle Motive sind deswegen schwarz gekennzeichnet. Meme und AlignAce haben ausserhalb der starken Motivansammlung auch Treffer, die in der Datenbank Analoge zu haben scheinen. Wir fanden die Ergebnisse von Meme und AlignAce von der Trefferverteilung und der Anzahl der in der Datenbank gefundenen Motive am aufschlussreichsten.

Die Programme rVista und YMF wurden nicht weiter benutzt. RVista nimmt zur Motiverkennung lediglich 2 .fasta Dateien entgegen.

4 Bemerkungen

Bei der Auswertung der Ergebnisse fiel auf, dass die Art und Weise der Durchführung nicht immer optimal war. Ein paar Faktoren könnten starken Einfluss auf die Endresultate genommen haben. Hier eine Aufzählung.

- ▷ Die Motivsuche wurde lediglich über die vereinzelten Cluster durchgeführt. Das hat sich direkt auf die Motivverteilung in den Organismen

ausgewirkt. Es haben sich wiederum die gleichen Cluster gebildet. Eine Motiverkennung über alle Cluster hätte die Menge der möglichen Motive stark beeinflusst. So versucht z.B. AlignAce ein Motiv über alle Sequenzen zu finden, also einen Konsens. Die Ergebnismenge wäre geringer aber dafür die Treffer umso deutlicher. Andere Methoden hätten sicherlich stark abweichende Resultate erbracht.

- ▷ Bei den Programmen wurden immer die ersten 10 Motive genommen. Daraus kann sich ein Verlust oder Redundanz von Motiven ergeben. Wir sollten für jede Score abwegen, wie gross der cut-off ausfallen soll. Zwar wird durch die spätere Suche in den Bindungsmotivdatenbanken der Score bereinigt (>0.9) aber es ist nicht auszuschliessen, dass durch oberflächliche Auswahl der Ausgangmotive, der Datenbestand nachhaltig von einem Optimum abweicht.

- ▷ Es wurden lediglich 500 Basen upstream untersucht. Es wurde zum Ende festgestellt, dass der Regulationsbereich darüber hinausgeht und 2500 Basen angemessen sind. Wir haben einen Motivplot (Abbildung ...) mit Meme, Ace und TfSearch mit 2500 Basen durchgeführt aber sind wegen Zeitmangels nicht weiter darauf eingegangen.

Literatur

[1] Hoffmann, R., Valencia, A. A gene network for navigating the literature. Nature Genetics 36, 664 (2004)