

Molekularbiologische Datenbanken

Übungen
Einleitung 5
Motive in Proteinsequenzen
Automatisches Update



Silke Trißl
Prof. Ulf Leser



Wissensmanagement in der
Bioinformatik

Motive in Proteinen

- Multiples Alignment von Sequenzen, die alle mit einer Funktion in Verbindung gebracht werden
 - Beispiel: Plant hemoglobins signature

CLUSTAL format alignment of matches to PROSITE PS00208

```
HBL0_PHYPA/78-89      NPPLKkLhAkyVF
HBL1_ARATH/63-74      NPPLcpLhAsvVF
HBL1_ORYSA/68-79      NPPLctLhAsvVF
HBL2_ARATH/60-71      NPPLkaLhAkvVF
HBL2_BRANA/60-71      NPPLkaLhAkvVF
HBL2_GOSHI/57-68      NPPLkaLhAkvVF
HBL2_LYCSB/57-68      NPPLcaLhAkvVF
HBL2_ORYSA/71-82      NPPLctLhAsvVF
HBL3_ORYSA/70-81      NPPLsiLhAkvVF
HBL4_ORYSA/67-78      NPPLckLhAkvVF
HBL_HORVU/64-75      NPPLctLhAsvVF
HBL_MAIZE/67-78       NPPLctLhAsvVF
HBL_TRETO/64-75      NPPLcpLhAtvVF
HBL_ZEAME/67-78      NPPLctLhAsvVF
HBP1_CASGL/56-67     NPPLkaLhAvVF
HBP2_CASGL/63-74     NPPLkaLhAsvVF
HBP4_CASGL/56-67     NPPLkaLhAvVF
HBP4_PARAD/63-74     NPPLcpLhAtvVF
LGB1_LUPLU/57-68     NPPLqaLhAkvVF
LGB1_MEDSA/56-67     SPPLqaLhAkvVF
```

- [SN]-P-x-[LV]-x(2)-H-A-x(3)-F



Motiv-Datenbanken

- Pfam
 - Seeds und dann weitere Detektion mit HMM
- ProDom
 - Findet Proteinfamilien mit PSI-Blast
- Prosite
 - Multiple Alignments von bekannten Proteinfamilien
- Blocks
 - Berechnet Alignments von Pfam, ProDom, ... neu
- Interpro
 - Integrated Resources of Proteins Domains and Functional Sites
 - Integriert Informationen von allen Motiv-Datenbanken

Informationen aus Prosite

Feld	Beschreibung	Kardinalität
ID	Identifikation; Typ	1
AC	Accession Nummer	1
DE	Description	1
PA	Pattern	1
DR	Database reference to Swiss-Prot	n

<http://www.expasy.org/prosite/prosuser.html>

Die Pattern Line

- Der standart IUPAC Einbuchstaben Code für die Aminosäuren wird genutzt
 - Genauso wie bei der Swiss-Prot Sequenz
- Das Symbol 'x' steht für eine Position, wo genau eine Aminosäure akzeptiert wird
 - Entspricht einem '.' bei den regulären Ausdrücken
- Eine der Aminosäuren, die in den eckigen Klammern '[']' steht, kann an dieser Position vorkommen.
 - Beispiel: [ALT] steht für Ala oder Leu oder Thr.
- Aminosäuren, die in geschweiften Klammern '{ }' stehen, können an dieser Position nicht vorkommen, alle anderen schon
 - Beispiel: {AM} bedeutet jede Aminosäure außer Ala und Met.

Die Pattern-Line - cont.

- Jedes Element eines Pattern wird durch ein '.' getrennt.
- Wiederholungen eines Elements werden durch einen in Klammern geschriebenen Zahlenwert im Anschluß an das Element ausgedrückt
 - Beispiele: x(3) entspricht x-x-x,
 - x(2,4) entspricht x-x oder x-x-x oder x-x-x-x.
- Wenn ein Pattern auf den Anfang oder das Ende beschränkt ist, dann beginnt es mit dem '<' Symbol oder endet mit dem '>' Symbol.
 - In einigen seltenen Fällen kann das '>' auch in eckigen Klammern stehen.
 - 'F-[GSTV]-P-R-L-[G>]' bedeutet, dass 'F-[GSTV]-P-R-L-G' oder 'F-[GSTV]-P-R-L>'
- Ein Punkt beendet das Pattern

Die Pattern-Line - Beispiele

- PA [AC]-x-V-x(4)-{ED}.
- Dieses Pattern bedeutet:
 - [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
 - [AC].V.{4}[^E^D]
- PA <A-x-[ST](2)-x(0,1)-V.
- Dieses Pattern, das am Anfang einer Kette stehen muß ('<'), wird übersetzt als
 - Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val
 - ^A.[ST]{2}.{0,1}V

Aufgabe 5.1 - Integrieren von Prosite

- Integrieren von Prosite
 - mit vorgegebenen Schema
- Laden der Daten
- SQL Anfragen auf der erweiterten Datenbank stellen
 - finden der Pattern in den Swiss-Prot Sequenzen über reguläre Ausdrücke
 - speichern der ursprünglichen Pattern
 - des daraus resultierenden regulären Ausdrucks

Verbindung zu Swiss-Prot

- Über die in DR angegebenen Links
 - T true positive
 - N false negative; Sequenz, die zur Familie gehört, aber nicht dem Pattern entspricht
 - P partielle Sequenz, das Pattern liegt ausserhalb der Sequenz
 - ? Sequenz gehört möglicherweise zur Familie
 - F Sequenz hat das Pattern, gehört aber nicht zur Familie
- über das eigentliche Pattern
 - PostgreSQL hat regular expressions
 - <http://www.postgresql.org/docs/7.4/interactive/functions-matching.html#FUNCTIONS-POSIX-REGEXP>
 - ähnlich mächtig wie Perl oder Java

In-house Datenbank ↔ Public databases

- Public databases
 - Immer aktuell (Nutzer ändert die Daten dort)
 - stellen
 - große Releases
 - 'weekly' updates
- In-house Databases
 - Anfragen müssen nicht an öffentlich zugängliche Seiten gesandt werden
 - Daten können in-house verändert werden
 - Probleme bei Aktualität

Update prozeß

- Alle Datenbanken stellen
 - ständig
 - aber in unregelmäßigen Abständen neue Daten zur Verfügung
- Daten müssen auch in die In-house Datenbank
 - Manuell: Jede Woche einmal nachsehen
 - unglaublich mühsam
 - Automatisch - cron job:
 - ein script, das erkennt ob eine neue Version verfügbar ist
 - diese herunterlädt
 - und in die Datenbank einspielt

Daten auf ftp-Seiten

- Nutzen des Filetransfers über FTP
 - viele biologische Daten sind von FTP-Seiten zu erhalten
 - werden ständig, aber in unregelmäßigen Abständen aktualisiert
- Automatisieren
 - Je nach Interesse und Aktualisierungszeiten der Datenquelle ein Script ausführen, das eine neue Version herunterlädt und in die Datenbank schreibt.

FTP - Server auf paprika

- Prosite-daten
 - ftp://ca.expasy.org/databases/prosite/release_with_updates/
- <ftp://paprika.hu-berlin.de/prosite>
 - ist nur von Paprika aus zu erreichen
 - arbeitet von 8:00 – 23:00 Uhr
- Daten auf paprika
 - `prosite_new.dat` Neueste Version - bleibt unverändert
 - `prosite.dat` wird alle 30 Minuten durch eine neue Version ersetzt

Aufgabe 5.2 – Präsentation

- 'Live' Präsentation der Lösung
- Zeigen Sie, dass das automatische Herunterladen einer neuen Version funktioniert
 - auf paprika
 - mit Einträgen in die Datenbank
- Daten werden dann alle 5 Minuten erneuert
- Nur die neueste Version muß vorhanden sein!

Fragen?

- Aufgabe 5 auf der Web-page
- Folien auf der Web-page
- Daten auf paprika unter
 - prosite_new.dat unter /daten/mdb/prosite/
 - bzw. über ftp
- Lösung bis 22.06. , 17 Uhr per e-mail oder in RUD25 IV.104

