08.10.2000 - Was ist Bioinformatik

Heidelberger Life Science Lab (HDLSL)
AG BioInformatik
Protokoll vom 8.10.2000

Am Vortag war beim Brainstorming folgende Liste von möglichen Themen entstanden:
- Neuronale Netze
- Kompression von DNA
- optische Erkennung
- Membranen / Zellfkt. simulieren
- Einführung in BioInformatik
- Klassenvererbung
- Evolutionsvorgänge simulieren
- Programmieren lernen
- Proteine...
- Sequenzierung

Zu Beginn des Treffens gab es eine Diskussion über die in der AG zu verwendenen Programmiersprachen. Dabei kam heraus, dass wir C, C++ und / oder Java verwenden wollen. Für C / C++ haben wir uns auf gcc als Compiler geeignet, weil dieser frei erhältlich ist, sowohl für Unix / Linux alsauch für Windows. Als Betriebssystem legten wir generell Unix/Linux fest, wobei man natürlich auch Windows einsetzen kann, wenn die Programme auch unter Linux laufen.

Um allen Grundlagen in C zu vermitteln, wurde ein C-Kurs geplant, der bei einem regulären Treffen Anfang November gehalten werden soll. Oliver Kahl und ?? haben sich bereiterklärt, den Kurs zusammen vorzubereiten und zu halten.

Für die Teilnehmer an der AG stellte Sven (oder Jens?) eine Liste zusammen, was für Themen wir uns zur Einführung in das Thema Bioinformatik ansehen sollten:

- DNS
- Proteine
z.B. sollte man die Namen der 20 verwendeten Aminosäuren wiedererkennen können.
- Datenbanken im WWW
-> eine Liste mit Bio-Datenbanken gibt es unter
http://www.cse.ucsc.edu/classes/cmps243/Fall99/resources.html
z.B. könnte man nach den Protein p53 (des Menschen) suchen
- Humangenomprojekt
- grundlegende Wahrscheinlichkeitsrechnung
- Algorithmus

Es wurden zur Grundlagenbildung Referate verteilt, die jeweils zu zweit ausgearbeitet werden sollen. Dabei stehen kurze Artikel auf Englisch aus "Trends Guide to Bioinformatics" als Grundlage zur Verfügung, die bei Bedarf durch andere Quellen ergänzt werden können. Die Referate sollten mündlich in maximal 10 Minuten von einem der Beiden vorgetragen werden. Zusätzlich sollte in Zusammenarbeit eine einseitige (DIN A4) Zusammenfassung des Referats (auf deutsch) erstellt werden, die an alle

Teilnehmer ausgeteilt wird. Bei dem Referat sollte man sich auf die 5 wichtigsten Punkte der Quelle beschränken, nicht die gesamte Information verarbeiten wollen. Der Zeitaufwand wurde grob auf "zwei Nachmittage" geschätzt. Die Referaten sollen auf zwei Treffen verteilt vorgetragen werden, am 20.10.2000 und der Rest zwei Wochen danach.

Als weitere Massnahme, um alle Teilnehmer auf eine gemeinsame Wissensbasis zu bekommen wurde weiterhin vorgeschlagen, dass jeder selbstständig die Treffen nachbereiten sollte, falls er etwas nicht verstanden hat.

Um das leibliche Wohl sollte man sich möglichst vor oder nach dem Treffen kümmern, um unseren Tagungsraum nicht unnötig zu verschmutzen. z.B. gibt es hinter dem Kommunikationszentrum ein Bistro

Gegen Ende hielt Jens Stoye einen Vortrag als Einführung in die Bioinformatik. Nachfolgend meine Mitschrift:

Grundlagen der Bioinformatik:
- Biologie
- Molekularbiologie
- Biochemie
u.a., aber nicht: Botanik, Zoologie
- Mathematik
- Statistik
- Stochastik
- Informatik
- Kombinatorik
- Datenbanken

Anwendungen der Bioinformatik:
1. Sequenzanalyse:
- DNA
- RNA
- Proteine
1.1 Beispiele Sequenzanalyse bei DNA
- exaktes matching:
Kommt eine kurze Sequenz q in einer langen Sequenz t vor?
- approximatives matching: Kommt eine kurze Sequenz q in einer Variation in einer langen Sequenz t vor?
Die lange Sequenz t kann sehr lang sein, z.B. die 3,3 bis 3,5 Mrd. Buchstaben, aus denen das menschliche Erbgut besteht.
1.2 Beispiel Sequenzanalyse bei RNA
- Vorhersage der räumlichen Faltungsstruktur anhand der Basensequenz
1.3 Beispiel Sequenzanalyse bei Proteinen
- Vorhersage der räumlichen Faltungsstruktur anhand der Basensequenz
-> diese Aufgabe ist bei Proteinen weitgehend ungelöst

Neben dem normalen Sequenzvergleich von zwei Sequenzen gibt es noch den multiplen Frequenzvergleich, bei dem mehr als zwei Proteine auf Ähnlichkeiten untersucht werden.

2. Evolution und Genomanalyse
Rekonstruktion von Stammbäumen durch Vergleich von DNA-Sequenzen bestimmter Gene verschiedener Lebewesen, z.B. Vergleich

MenschDrosophila

3. Sequenzierung und Kartierung (z.B. Human Genome Project)
Ein zu sequenzierender DNA-Strang wird vervielfältigt und danach an zufälligen Stellen zerschnitten, so dass jede Kopie des DNA-Strangs an unterschiedlichen Stellen aufgetrennt wurde. Die Reihenfolge dieser Schnipsel, die jeweils etwa 50000 Basen lang sind, ist nicht bekannt. Nun wird jeder Schnipsel einzeln sequenziert. Danach wird mittels Computern die einzelnen Sequenzen untersucht. Da verschiedene Kopien des gleichen Strangs an unterschiedlichen Stellen zerschnitten wurden, gibt es notwendigerweise Überlappungen. Der Computer muss nun die einzelnen Schnipsel durch Auffinden der Überlappungen zu einem durchgehenden Strang zusammensetzen. Dabei kann er sich natürlich irren, falls es Wiederholungen in der Gesamtsequenz oder zu geringe Überschneidungen gibt. Je mehr verschiede überlappende Schnipsel der Computer untersucht und in sein Puzzle einpassen kann, desto sicherer ist es, dass sein Ergebnis richtig ist.

4. Datenbanken
Hier geht es um die Fragestellung, wie man Daten sinnvollerweise in Datenbanken speichern kann, wenn es um riesige Datenmengen wie beim HGP geht. Verschienene Möglichkeiten sind z.B.
- reine Textdateien
- XML (Extended Markup Language)
- Oracle

5. Expression: welche Gene sind aktiv?
(exprimierte Gene = aktive Gene)
Hier geht es darum, die Expressionslevel von Genen zu messen.
Bei Krebszellen sind oft die "falschen" Gene aktiv.