HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

Bioinformatiikkaa ja koneoppimista

Samuel Kaski
Tietojenkäsittelytieteen laitos

Bioinformatiikan / laskennallisen biologian UKK

Mitä se on? Modernien laskennallisten ja tilastollisten menetelmien soveltamista ja kehittämistä biologisten järjestelmien ja prosessien ymmärtämiseksi.

Miksi se on tärkeää? Auttaa mm. näissä:

- Elämän ymmärtäminen ym. biologinen perustutkimus
- Lääketieteellinen diagnostiikka
- Lääkkeiden kehitys

Miksi se on kiinnostavaa? Tarjolla on uudenlaisia ongelmia ja sovelluksia uudenlaisille menetelmille ja algoritmeille! ”Biology easily has 500 years of exciting problems to work on” (Donald Knuth)

Mitkä ovat sen haitat? Data-analyysistä *sinänsä* ei onneksi liene haittaa, mutta sen sovelluksista voi olla. Geenitutkimuksen eettiset implikaatiot pitää harkita tapauskohtaisesti!

Miksi se on ajankohtaista? Paljastetaan seuraavalla sivulla.

Missä voin oppia lisää? Bioinformatiikan ja laskennallisen biologian suuntautumisvaihtoehdossa

Recent news

10 technologies that will change the world. Number 4: Bayesian machine learning, 10: personal genomics (Technology Review, 2004)

Opportunities in bioinformatics once abounded for the self-taught and industrially minded, but employers are now turning towards the formally trained and academics (Nature, 2004)

Bioinformatics attracts big guns (Nature Biotechnology, 2004)

Stein gives bioinformatics ten years to live (O’Reilly bioinformatics Technology conference, 2003)

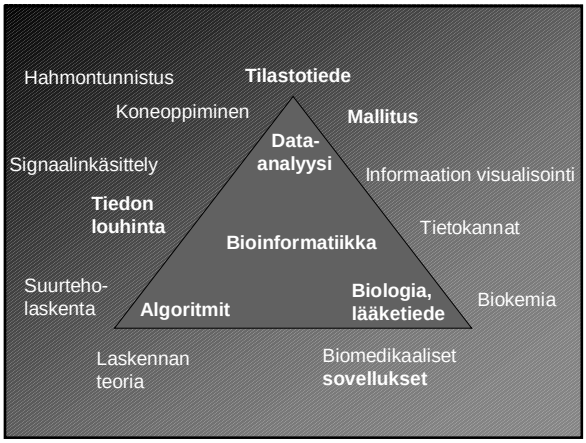
Monenlaista bioinformatiikkaa

Menetelmä-tieteet | Bioinformatiikan menetelmä-kehitys | Biologisten ongelmien ratkaisu valmiilla menetelmillä | Biologia

Miksi bioinformatiikka on vaikeaa?

Tutkimuksen ja haastavan ongelmanratkaisun normaalien ongelmien lisäksi:

- Pitää hallita monta alaa

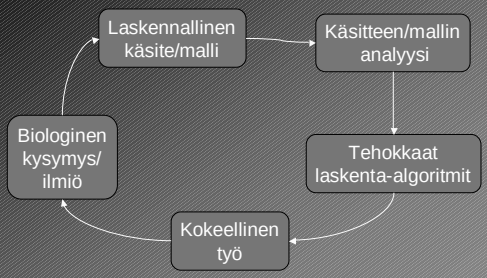


Miksi bioinformatiikka on vaikeaa?

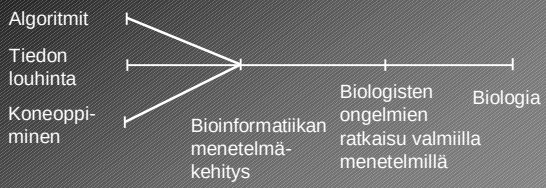
Tutkimuksen ja haastavan ongelmanratkaisun normaalien ongelmien lisäksi:

- Pitää hallita monta alaa
- Ongelmat usein aluksi huonosti määriteltyjä.”Pitäisi ymmärtää tätä ilmiötä.”
- Data-analyysin menetelmäkehityksen yleinen ongelma: Miten kehittää yleisiä menetelmiä ja ratkaista sovelluskohtaisia ongelmia samanaikaisesti?

Tutkimussykli



Monenlaista bioinformatiikkaa TKTL:ssä



Bioinformatiikan menetelmätutkimusta

Tietojenkäsittelytieteen laitoksella

- Rakennebiologian ja systeemibiologian algoritmit (Ukkonen, Rousu)
- Geenipaikannuksen laskennalliset menetelmät (Toivonen, Mannila)
- Genomien struktuurin algoritmiset kysymykset (Mannila, Ukkonen, Salmenkivi)
- Geeniekspression data-analyysimenetelmät (Kaski)

Muualla Helsingin yliopistossa:

- Matematiikan ja tilastotieteen laitos: Arjas, Gyllenberg
- Viikin Biokeskus: Holm
- Biomedicum Bioinformatics Unit: Saharinen,

Suurin osa TKTL:n bioinformatiikkaryhmistä toimii laitoksen tutkimusyksiköissä:

- Akatemian Datasta tietoon-huippututkimusyksikkö
- Tietotekniikan tutkimuslaitos HIIT

Ryhmillä on laajaa yhteistyötä sekä tieteiden välillä että tietojenkäsittelytieteessä, sekä kotimaassa että kansainvälisesti.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

Algoritmiikkaa

Esimerkki: Sekvenssien rinnastus

```
-GCGC-ATGGATTGAGCGA
TGCGCCATTGAT-GACC-A
```

Tehtävä: Selvitä onko kahdessa sekvenssissä samankaltaisia osia, ja mittaa sekvenssien samankaltaisuus.

Miksi? Rinnastus on perusoperaatio, jota tarvitaan monessa data-analyysitehtävässä, esim:

- tietokantahaut: etsitään tiettyjä toiminnallisia osia
- genomin rakenne: mitkä osat/geenit ovat samanlaisia eri eliöillä

Algoritmi: Dynaaminen ohjelmointi

Valitaan pisteytys/kustannukset:

- Sama symboli: +1
- Eri symboli: -1
- Puuttuva symboli: -2

		A	G	C
	0	-2	-4	-6
A	-2	1	-1	-3
A	-4	-1	0	-2
A	-6	-3	-2	-2
C	-8	-5	-4	-1

AAAC
A-GC

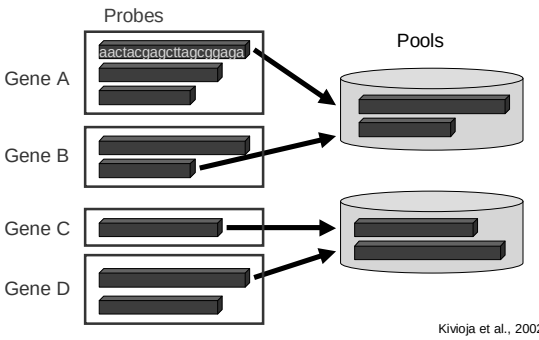
Ongelmia:

- kompleksisuus on $O(n^2)$
- rajoittava: aina ei haluta rinnastaa koko sekvenssejä
- miten valita pisteytykset?

On kehitetty paljon nopeutuskeinoja ja erityistilanteisiin sopivia algoritmeja, mm. lokaalien samakaltaisuuksien hakuun.

Tarvitaan vankkaa näkemystä sekä tehokkaista algoritmeista että siitä, millaisia yksinkertaistuksia voi tehdä.

Probe Assignment



Kivioja et al., 2002

- Experiment planning for a novel measurement method developed at VTT Biotechnology.
- The method uses DNA fragments called probes to measure the expression of genes
- Goal: Minimize the resources needed for the experiment.
- Algorithmic problem: Choose one probe for each gene and partition the chosen probes into minimal number of pools so that the probes in each pool have different lengths.
- Results:
  - Negative: The problem is NP-hard.
  - Positive: An algorithm that uses at most twice the optimal number of pools and works well in practice.
  - Software for automatic planning of experiments.

Kivioja et al., 2002

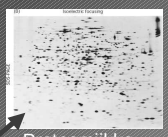
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

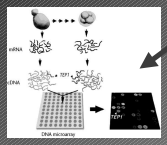
Biologian haaste

Uusilla mittausmenetelmillä suuria geeniaineistoja

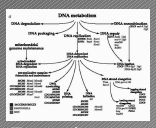
DNA-sekvenssi-tietokannat



Proteomiikka

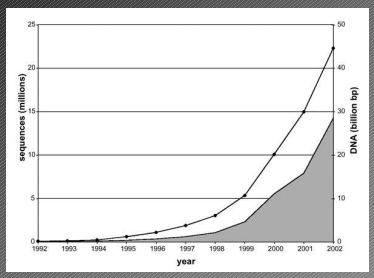


Geenien-ekspressio-tietokannat



Taustatieto: ”geeniontologiat”, artikkelitietokannat

GENBANK-tietokannan kasvu



Venter et al., Nature Genetics 2003

Haaste

Kuinka hyödyntää näitä aineistoja biologisessa ja lääketieteellisessä tutkimuksessa?

Vastaus

Täydentämällä hypoteesilähtöistä tutkimusta datalähtöisemmillä menetelmillä

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

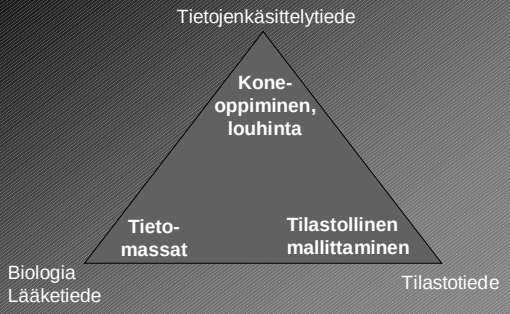
Koneoppivaa bioinformatiikkaa

Tiedon louhintaa oppivilla menetelmillä

Tiedon louhinnan tavoitteena on uusien, kiinnostavien ja hyödyllisten löydösten etsintä tietomassoista.

Oppivat menetelmät: Datalähtöisiä menetelmiä; menetelmiä jotka oppivat datasta.

Monitieteistä tutkimusta

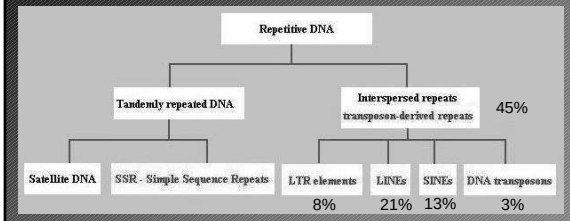


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

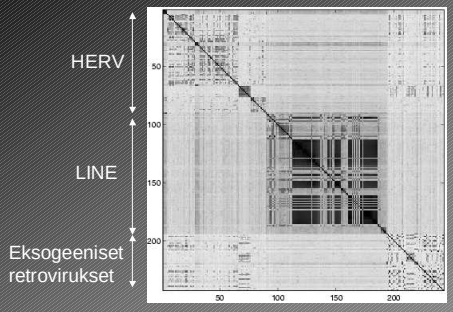
Esimerkki: Retroviruksia ihmisen genomissa

Genomissa on paljon enemmän retrovirusten jäämiä kuin geenejä

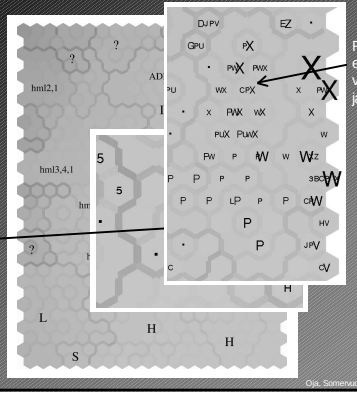


Eriksson et al., 2001

Rakennetta virusnäytejoukossa



Voidaan esim. järjestää samankaltaiset lähekkäin



Oja, Somervuo, Kaski, Kohonen, 2003

Ongelma: Mikä on tärkeää?

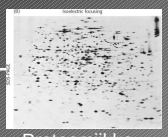
Mikä löydöksissä on uutta, kiinnostavaa ja hyödyllistä?

Mikä osa löydöksistä ei ole kohinaa: mittauskohinaa tai biologista kohinaa?

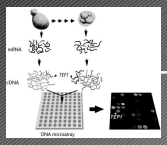
Kuinka päätellä mikä on relevanttia?

DNA-sekvenssi-tietokannat

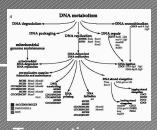
Oletus: Useille aineistoille yhteinen on kiinnostavaa



Proteomiikka

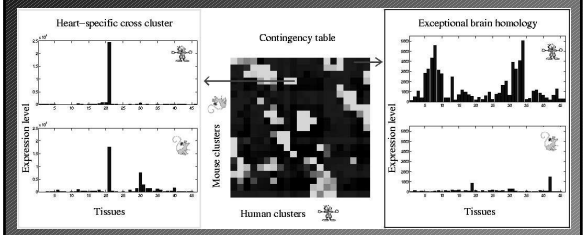


Geeniekspressio-tietokannat



Taustatieto: ”geeniontologiat”, artikkelitietokannat

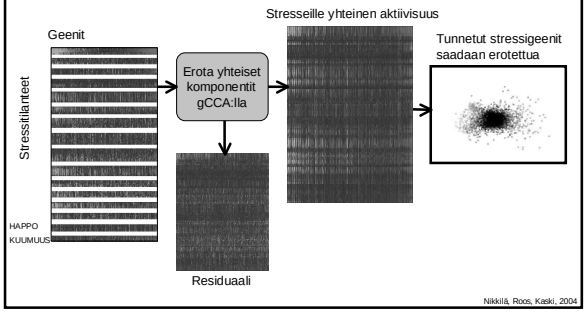
Hiiriä ja ihmisiä



Riippuvuuksien louhinta: Etsitään geenien toiminnasta säännönmukaisuuksia ja **poikkeuksia** hiiren ja ihmisen välillä.

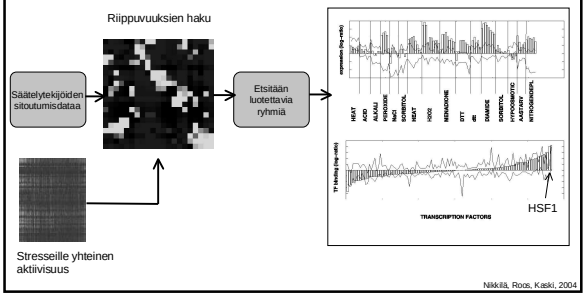
Nikkilä, Sinkkonen, Lahti, Knuutila, Roos, Kaski, 2004

Millaista on stressi?



Nikkilä, Roos, Kaski, 2004

Mikä säätelee stressiä?



Nikkilä, Roos, Kaski, 2004

Millaista tutkimus on käytännössä?

Hyvin monentyyppistä riippuen suuntautumisesta:

Yksin tai pienessä porukassa tehtävästä teoreettisesta työstä (esim. algoritmitutkimuksesta)

Ryhmässä biologien ja lääkärien kanssa kiinteässä yhteistyössä tehtävään ongelmanratkaisuun.

Tyypillisesti erityyppiset vaiheet vuorottelevat.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

Bioinformatiikan opetusta TKTL:ssä

Koulutusfilosofia

Vankka menetelmätieteinen perusta
+ perustiedot biologiasta ja bioinformatiikasta
+ syventyminen johonkin menetelmäalueeseen
+ paljon valinnanvaraa
+ keskitytään periaatteisiin jotka eivät vanhene

= laskentamenetelmien asiantuntijan tutkinto, joka on sopiva pohja hyvin monenlaisille tehtäville

Painopisteen voi valita joustavasti bioinformatiikan ja laskennallisen data-analyysin välillä

Bioinformatiikan opetus TKTL:ssä

- Bioinformatiikan perusteet, 3 ov
- Data analysis for gene expression, 3-5cu
- Geenisekvenssit. Tänä vuonna tarjolla ainakin Merkkijonomenetelmät (4ov) ja matematiikan ja tilastotieteen laitoksen kurssi Statistical methods in bioinformatics
- Computational systems biology, 3cu
- Bioinformatiikkaan liittyviä seminaareja ja erikoiskursseja

Menetelmällisiä syventymisalueita

Algoritmit:
- Algoritmien suunnittelu ja analyysi, 5ov
- Merkkijonomenetelmät, 4ov
- Kombinatorinen optimointi, 5ov
- Approximation algorithms, 4ov

Tiedon louhinta ja laskennallinen data-analyysi
- Tutkimustiedonhallinnan peruskurssi, 3ov (jollei jo luettu cumussa)
- Tiedon louhinnan menetelmät, 3ov
- Special course on data mining, 3ov
- Algorithms for segmentation problems, 2ov
- Paikkatiedon hallinta ja analyysi, 3ov
- Linear algebra methods for data mining, 2ov
- Information visualization, 2ov

Koneoppiminen
- Koneoppiminen, 4ov
- Tekoäly, 4ov
- Kolme käsitettä -sarja (tänä vuonna Informaatio)
- Kernel methods for pattern analysis, 2ov
- Classification, 4ov

Informaatiojärjestelmät
- Tutkimustiedonhallinnan peruskurssi, 3ov (jollei jo luettu cumussa)
- Tietokannan mallinnus, 2ov
- Tietokantarakenteet ja algoritmit, 4ov
- Tietovarastot, 2ov
- Tiedonhakumenetelmät, 3ov

Mitä isona?

Laskentamenetelmien ja tiedonhallinnan erikoisasiantuntijoiksi tutkimuslaitoksiin ja yrityksiin

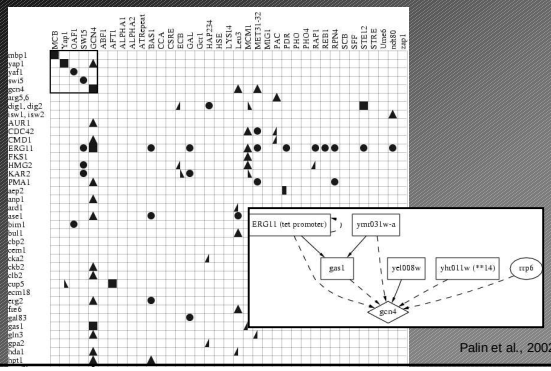
Ala muuttuu koko ajan — tarvitaan ihmisiä jotka määrittelevät alan ja työtehtävät tulevaisuudessa!

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tietojenkäsittely-tieteen esittely
29.9. 2004

Muita bioinformatiikkatöitä

Correlating promoters and expression



Palin et al., 2002

Lisätietoa

Tietojenkäsittelytieteen laitos:
http://www.cs.helsinki.fi

Bioinformatiikan linja tietojenkäsittelytieteen laitoksella:
http://www.cs.helsinki.fi/bioinformatiikka

Helsingin yliopiston bioinformatiikkasivut:
http://www.helsinki.fi/bioinfo