



Dinas Forskerskole:
Workshop den 6.-7. april 2000

Koldkærgård Landboskole, Skejby.

Redigeret af Anders Ringgaard Kristensen

Dina Notat No. 87 ! April 2000

Dette dokument er også tilgængeligt på www under URL:
<ftp://ftp.dina.kvl.dk/pub/Staff/Anders.R.Kristensen/dina-reports/notat87/notat87.pdf>

Dinas Forskerskole
Institut for Husdyrbrug og Husdyrsundhed
Den Kongelige Veterinær- og Landbohøjskole
Grønnegårdsvej 3, DK-1870 Frederiksberg C

Indhold

Program	3
Torsdag den 6. april	3
Fredag den 7. april	4
Anbefalet læsning	5
Mundtlige indlæg	6
Søgning i databaser over DNA- og proteinsekvenser	7
<i>Kenneth Smith</i>	
Algoritmer for alignment af biosekvenser: hvordan virker de?	8
<i>Peter Sestoft</i>	
Genkort, koblingsanalyse, quantitative trait loci	9
<i>Claus Ekstrøm</i>	
Fylogeni	10
<i>Henrik Christensen & Henrik Stryhn</i>	
Posters	12
Protein analysis and identification by mass spectroscopy and database searching	13
<i>Jeppe Emmersen</i>	
Genetic mapping of Spinal Dysmyelination in cross-bred American Brown Swiss cattle	14
<i>Peter H. Nissen^{1,2}, Naseer M. Shukri¹, Jørgen S. Agerholm³, Merete Fredholm² and Christian Bendixen¹</i>	
High Through-Put Sequencing	15
<i>Mads Lundsgaard, Karen G. Welinder and Kåre Lehmann Nielsen</i>	
Genetic diversity of the Hemagglutinin (H) Glycoprotein gene of Canine Distemper Virus infecting Greenlandic sledge dogs and Danish mink	16
<i>Lotte Dahl, B. Bolt, L. Nielsen, E. Gottschalck, N.L. Jansen & M. Blixenkron-Møller</i>	
Genetic diversity in barley assessed using microsatellite markers	17
<i>Runa Ulsøe Johansen</i>	
QTL-Mapping in Danish Dairy Cattle	18
<i>Bernt Guldbrandtsen</i>	
Deltagerliste	19

Program

Torsdag den 6. april

11.00:

Ankomst og indkvartering

12.00:

Frokost

13.00:

Indledning, præsentation af deltagere.

Anders Ringgaard Kristensen, Dinas Forskerskole

13.15:

Søgning i databaser over DNA- og proteinsekvenser

Kenneth Smith, Applied Bioinformatics Center, Lyngby

13.45:

Algoritmer for alignment af biosekvenser: hvordan virker de?

Peter Sestoft, Dinas Forskerskole, Institut for Matematisk og Fysik, KVL

14.30:

Computerøvelser

15.00:

Eftermiddagskaffe

15.30:

Genkort, koblingsanalyse, quantitative trait loci

Claus Ekstrøm, Biostatistisk Afdeling, Sundhedsvidenskabelige Fakultet, Københavns Universitet

16.30:

Computerøvelser

18.00:

Middag

19.00:

Fylogeni

*Henrik Christensen, Institut for Veterinær Mikrobiologi, KVL;
Henrik Stryhn, Dinas Forskerskole, Statens Veterinære Serumlaboratorium*

20.00:

Computerøvelser

22.00:

Natmad

Fredag den 7. april

7.45:

Morgenmad

8.30:

Opfølgning på øvelser

9.15:

Postersession med plenumpresentation, besigtigelse og kaffe

- *Jeppé Emmersen*: **Protein analysis and identification by mass spectroscopy and database searching**

- *Peter H. Nissen*: **Genetic mapping of Spinal Dysmyelination in cross-bred American Brown Swiss cattle**

- *Mads Lundsgaard*: **High Through-Put Sequencing**

- *Lotte Dahl*: **Genetic diversity of the Hemagglutinin (H) Glycoprotein gene of Canine Distemper Virus infecting Greenlandic sledge dogs and Danish mink**

- *Runa Ulsøe Johansen*: **Genetic diversity in barley assessed using microsatellite markers**

- *Bernt Guldbrandtsen*: **QTL-Mapping in Danish Dairy Cattle**

10.30:

Præsentation og diskussion af perspektiver for brug af biosekvensanalyse i biologisk forskning og anvendelser.

Mogens Lund, IT-koordinator i Dina, Danmarks Jordbrugsforskning;

Günter Backes, IT-koordinator i Dina, Risø;

Peder Worning, Center for Biological Sequence Analysis, Lyngby.

11.30:

Diskussion, opsummering, evaluering

12.00:

Frokost og afrejse

Anbefalet læsning

Deltagende PhD studerende vil få udleveret Trends Guide to Bioinformatics 1998, Elsevier Science.

En letlæst men informativ introduktion til genetik for ikke-genetikere er Larry Gonick og Mark Wheelis: The Cartoon Guide to Genetics, HarperPerennial 1991. Kan f.eks. købes gennem www.bokus.dk for 120 kroner.

Nyttig forberedelse til Claus Ekstrøms foredrag for deltagende ikke-genetikere er Mary Sara McPeck: An introduction to recombination and linkage analysis, i Genetic Mapping and DNA Sequencing, redigeret af T. Speed og M.S. Waterman, Springer-Verlag 1996.

Mundtlige indlæg

Søgning i databaser over DNA- og proteinsekvenser

Kenneth Smith

Overblik over de molekylære biologiske databaser tilgængelige på nettet. Derefter en kort gennemgang af nogle af de vigtigste databaser såsom Genbank, Swiss-Prot, PDB mm. SRStm fra Lion Biosciences vil blive brugt som sekvenssøgningsprogram.

Algoritmer for alignment af biosekvenser: hvordan virker de?

Peter Sestoft

Der findes mange programmer som giver mulighed for alignment af biosekvenser, først og fremmest Fasta og Blast. Men hvordan virker programmerne egentlig? Hvordan kan de effektivt finde det bedste alignment mellem to biosekvenser?

Teknisk set benyttes der såkaldt dynamisk programmering, som løser et optimeringsproblem ved at tabellere optimale deløsninger.

I forelæsningen skal vi se hvordan denne teknik udnyttes til at løse forskellige slags alignment-problemer:

- Globalt alignment af to sekvenser med simpel lineær gab-omkostning (Needleman-Wunsch algoritmen)
- Lokalt alignment af to sekvenser med simpel lineær gab-omkostning (Smith-Waterman algoritmen)
- Globalt alignment af to sekvenser med affine gab-omkostning (udvidelse af Needleman-Wunsch algoritmen)
- Lokalt alignment af to sekvenser med affine gab-omkostninger (udvidelse af Smith-Waterman algoritmen)

Det meste kan forklares på det begrebsmæssige plan, men vi skal også se på dele af en model-implementation i programmeringssproget Java.

Forelæsningens eksempelprogrammer er tilgængelige fra

<http://www.dina.kvl.dk/~sestoft/bsa.html>

og kan køres direkte fra en nyere browser

Genkort, koblingsanalyse, quantitative trait loci

Claus Ekstrøm

Koblingsanalyser danner baggrund for mange af de metoder, der benyttes i genetisk epidemiologi, og er en af hjørnestenene i statistiske analyser af genetiske markørdata. Ved at benytte koblingsanalyser sammen med ideen bag "allele sharing", er man i stand til at finde kromosomale områder, der med størst sandsynlighed indeholder kodende regioner for både kvalitative og kvantitative træk.

I forelæsningsen gennemgås hovedideerne bag koblingsanalyse og hvorledes man ud fra disse analyser kan konstruere genkort (koblingskort). Ligeledes gennemgås to metoder til single- og multipointanalyse af kvantitative træk (QTL):

- Haseman-Elston regression
- Normalfordelingsmodeller med varianskomponenter

Begge metoder er baseret på allele sharing, og kan benyttes til analyser af både simple designs (så som F1, F2 og BC) så vel som mere komplekse.

Fylogeni

Henrik Christensen & Henrik Stryhn

Med fylogeni kan forstås to ting: den fylogenetiske mekanisme (1.) og resultatet af denne mekanisme (2.).

1. Fylogeni er et biologisk koncept. Grundideen er, at alle grupper af organismer er opstået ved evolution og udgør et beslægtet kontinuum (Lamarckism). Det fylogenetiske koncept indeholder Darwins udviklingslære (den bedst egnede variant udvælges).

2. Fylogeni kan også være resultatet af den fylogenetiske proces, hvilket ofte vises som et "træ". Træets "blade" er artslignende grupper eller højere kategorier af organismer, "grenene" er genotypiske linier, forgreningspunkterne (på engelsk node) udtrykker artsopsplitninger. Hvis "grenenes" længde relateres til tiden og oprindelsen for opsplittningen af arterne, er det et fylogenetisk træ (fylogram). Hvis grenenes længde ikke fastsættes er det et kladogram. Der anvendes kun tveforgreninger, hvilket har en logisk baggrund (Aristoteles) og ikke en biologisk. Ved hvert forgreningspunkt kan en art enten være blevet spaltet i to nye arter og selv være uddø, eller arten kan have dannet én ny art, men fortsat sin eksistens. I forgreningspunkter med mere end tvegreninger, sættes de yderligere grenlængder til 0. Træer kan have "rod", dvs. en gren som ikke udspringer i en art, men har et forgreningspunkt, eller være uden rod, dvs. der ikke findes en sådan gren i træet. Det samme træ kan visualiseres på mange måder, da der er fri rotation om alle forgreningspunkterne.

Den fylogenetiske analyse har til formål at opstille en model for fylogenen (forstået som 2). Her kommer bioinformatikken ind. Hvis der arbejdes direkte på genotypeniveau, hvilket sker med DNA-sekvensanalyse, vil der være en klar sammenhæng til slægtskabet mellem organismer. Før disse informationer blev tilgængelige var den fylogenetiske analyse vanskelig metodisk og resultatet kunne være usikkert, idet de fænotypiske (morfologiske) karakterer kunne være svære at relatere (homologisere) mellem forskellige grupper af organismer. Homologi (udspring i fælles stamform) af DNA er mere entydig, men der kan være faldgruber som rekombination og gendublikation.

Den fylogenetiske analyse har to hovedanvendelser: til rene evolutionære studier og som en hjælpemetode til taksonomiske undersøgelser. Ved de taksonomiske undersøgelser er formålet at afgrænse såkaldte monofyletiske grupper, dvs. grupper af organismer som kan påvises at have en fælles stamform. Klassifikation kan ikke alene baseres på fylogenetisk analyse, da denne ikke fastlægger grænser for hvor store grupperne skal være eller om disse er praktisk anvendelige i taksonomi.

Den fylogenetiske analyse af molekulære sekvensdata sker ofte i seks trin:

1. DNA- eller aminosyresekvenser af homologe gener "alignes", dvs. de arrangeres lodret under hinanden så de positioner, der mest sandsynligt har haft en fælles stamform står under hinanden. Relevante edbprogrammer er Clustal (frisoftware), Pileup/lineup (GCG), Bestfit (GCG, bygger på Smith-Waterman). Se endvidere i PAUP og MacClade.
2. Fylogenetisk analyse efter maksimum likelihood princippet, bygger groft sagt på at vælge

et træ inden for en klasse af statistiske modeller af træer, således at de observerede sekvenser får størst mulig sandsynlighed. I modellerne indgår både træets struktur og grenlængder. Da søgningen efter maksimumsværdien er iterativ ud fra et "starttræ", køres den flere gange for at kontrollere træet med den højeste likelihoodværdi. Relevante programmer er fastDNAMl (frisoftware) og DNAML (phylip-pakken).

3. Fylogenetisk analyse efter "Neighbour Joining" sker i to trin. Først opbygges en similaritetsmatrice (DNADIST, PROTDIST i phylip-pakken), derefter dannes et træ efter en clustering-algoritme der successivt forbinder sekvenserne med den mindste indbyrdes afstand som naboer (NEIGHBOR i phylip-pakken). Der findes beslægtede programmer der også opstiller træer på grundlag af en similaritetsmatrice.

4. Fylogenetisk analyse efter parsimonianalyse. Det træ findes som har færrest trin, dvs. færrest substitutioner mellem nukleinsyre eller aminosyrer. Grenlængder estimeres ikke og træet er derfor et kladogram. PAUP-pakken er "født" til denne metode, men analysen kan også ske med DNAPARS+ CONSENSE og andre programmer i phylip pakken.

5. Bootstrapanalyse foretages for at dokumentere monofyletiske grupper. Kolonner i alignment byttes om ved tilfældig udtagning med tilbagelægning. Et stort antal (fx 100) sådanne tilfældige alignments analyseres med maksimum likelihood (eller et af de andre fylogenetimetoder). Med CONSENSE i phylip-pakken bestemmes hvor mange gange et givet forgreningspunkt forekommer ud af 100 træer. Jo højere tal, desto større sandsynlighed er der for at gruppen er monofyletisk. Proceduren er glimrende forklaret i "Trends Guide ..".

6. På grundlag af informationer i punkt 2. - 5. opstilles det bedste fylogenetiske træ. Træet kan tegnes manuelt, men genereres og manipuleres bedst med programmer i MacClade-pakken.

Ved meget store datasæt eller hvor mindre akkurathed kræves, kan analysen reduceres til et eller flere af trinene 2. - 3.

Posters

Protein analysis and identification by mass spectroscopy and database searching

Jeppe Emmersen

Dept. of Life Science, Aalborg University, Sohngaardsholmsvej 49, 9000 Aalborg (+ 45) 96358461 i5je@civil.auc.dk

As the information gained by sequencing of DNA has increased to an amount unthinkable 10 years ago, the need (and ability) to correlate this information with knowledge about protein structure and function has become more important.

Such information can be obtained quickly from a purified protein by mass spectroscopy performed on proteolytic digests by for example trypsin, which cleaves a protein specifically at arginine and lysine, thus giving well defined peptides. The peptides are then separated by RP-HPLC and detected both by UV and Electrospray Ionization Mass Spectroscopy (ESI-MS).

This information can then be used to search databases to obtain identification.

De Novo sequence information is also possible with ESI-MS, as the instrument is able to fragment peptides by MS/MS and perhaps give 2-4 aminoacids of sequence per peptide per digest.

Genetic mapping of Spinal Dysmyelination in cross-bred American Brown Swiss cattle

Peter H. Nissen^{1,2}, Naseer M. Shukri¹, Jørgen S. Agerholm³, Merete Fredholm² and Christian Bendixen¹

¹Danish Institute of Agricultural Science, Department of Animal Breeding and Genetics, Tjele, Denmark; ²Royal Veterinary and Agricultural University, Department of Animal Genetics and Breeding, Copenhagen, Denmark and ³Danish Veterinary Laboratory, Department of Pathology and Epidemiology, Copenhagen, Denmark

The congenital neurological disorder spinal dysmyelination has been diagnosed in several national cattle breeds upgraded with American Brown Swiss (ABS). This is also the case in the Red Danish Dairy breed where all diagnosed cases of spinal dysmyelination have been genetically related to one single ABS bull. The disease is characterised by congenital recumbency, opisthotonus and extension of the limbs. Recently spinal dysmyelination have been shown to be inherited as an autosomal recessive disease. From an experiment where a number of cows descending from a known carrier were inseminated with semen from a known carrier bull, all having the same grand-grand sire (Agerholm & Andersen 1995, J. Vet. Med. A, 42, 9-12), we used thirteen affected calves and their mothers to conduct a full genomes can with approximately 200 microsatellite markers. Based on this investigation we were able to locate the spinal dysmyelination locus to a 17 cM interval. Assuming that the imported American Brown Swiss bull serves as single founder for the disease in the Danish population, the Spinal Dysmyelination gene must be located in a chromosomal region of homozygosity. Using newly found cases of spinal dysmyelination from the general population we can limit the most likely region to an interval of approximately 5 cM. Currently we are evaluating a comparative candidate gene picked from a region of conserved synteny in the human genome.

High Through-Put Sequencing

Mads Lundsgaard, Karen G. Welinder and Kåre Lehmann Nielsen

Department of Life Science at Aalborg University

Aalborg University has recently purchased the hardware needed to conduct high throughput sequencing, for the purpose of building Expressed Sequence Tag (EST) databases. The work presented here has mainly been addressing the setting up of the hardware and establishment of procedures. This have been done giving special attention to efficiency both regarding hands-on time and money, producing an affordable high through-put sequencing facility. These demands have called for an alternative method of sample preparation. The system has therefore been set-up to sequence PCR products instead of the traditional plasmid DNA preparations.

A sequencing facility of this capacity, however, presents a serious work load in terms of data processing. The procedures for sequence evaluation and homology searches have not yet been automated, and we are currently investigating different software tools for this purpose.

Genetic diversity of the Hemagglutinin (H) Glycoprotein gene of Canine Distemper Virus infecting Greenlandic sledge dogs and Danish mink

Lotte Dahl, B. Bolt, L. Nielsen, E. Gottschalck, N.L. Jansen & M. Blixenkron-Møller

Laboratory of Virology and Immunology, Department for Veterinary Microbiology, The Royal Veterinary and Agricultural University, Bülowsvej 13, DK-1870 Frederiksberg, Denmark.

The introduction of live attenuated canine distemper virus (CDV) vaccines in the 1950's has drastically reduced the incidence of CDV in dogs. However outbreaks, in which previously immunised dogs become infected have been observed. This raises the questions of whether the vaccines currently used efficiently protect against present-day circulating wild-types.

In 1998-1999 we experienced distemper outbreaks among Danish farmed mink. In contrast to earlier observations, these new outbreaks had protracted clinical and epidemiological course despite the claimed use of distemper vaccination.

To study the molecular biological characteristics, we sequenced the H genes of these viruses and of viruses involved in outbreaks among partly vaccinated sledge dogs in arctic Greenland. These sequences were compared to previously determined H gene sequences from earlier outbreaks in Europe, Greenland and North America. Phylogenetic analyses revealed that the viruses involved in the recent outbreaks in Danish mink had the characteristics of American field isolates. It appears that these viruses were introduced in Denmark by import of subclinically infected mink. The recent arctic viruses were very closely related to CDV from a serious outbreak among Greenlandic sledge dogs in 1988. Thus, the genotype appear to be remarkably conserved for more than 10 years, probably persisting among dogs and wildlife carnivores in this remote arctic region. We have found CDV in polar foxes. The new wild-type viruses exhibited a substantial genetic distance from the vaccine strains, as shown for other field isolates.

Studies on the pathogenesis and immunogenicity of the new wild-types are in progress.

Genetic diversity in barley assessed using microsatellite markers

Runa Ulsøe Johansen

Introduction

To be able to breed for durable disease resistance it is important to use genetically diverse resistance sources. As a prerequisite to select suitable parents for crosses, it is necessary to evaluate and analyse the available germplasm. The aim of the present study is to: 1) Investigate how the diversity within barley (*Hordeum*) is structured, and how big a part of the diversity is included in *Hordeum vulgare*, for instance modern varieties, landraces or exotic varieties, versus *Hordeum bulbosum* and wild barley species. 2) Elucidate the phylogeny of barley using microsatellite DNA fingerprints; does it correlate with traditional phylogeny and/or with pedigree information? 3) Study the frequency of different microsatellite alleles in barley.

Materials and Methods

Three sets of materials comprising approximately 180 different barley lines from various barley growing regions will be analysed. Wild species have been provided by the nordic gene bank, *H. bulbosum* were provided by The Abed Foundation, and some modern European and exotic varieties is included as well to give a reasonable coverage of the whole genus.

Genomic DNA is extracted from green leaves using the CTAB method, and used as template in a PCR to amplify selected microsatellites. Microsatellites are thus used as molecular markers to produce a DNA fingerprint for each accession. A number of published microsatellite primers have been tested and the most polymorphic chosen to give a unique fingerprint for each accession. Following the PCR the product is run on a polyacrylamide gel, and the DNA is visualized by silverstaining, or by using a labeled primer in the PCR and following running the gel in an ABI sequencer where the DNA is detected through fluorescence.

Data is analysed using "Statistica", "Sigma Plot" and "microsat".

QTL-Mapping in Danish Dairy Cattle***Bernt Guldbrandtsen***

Deltagerliste

Name	Email
Gunter Backes (kun fredag)	gunter.backes@risoe.dk
Morten Buch-Pedersen	mob@kvl.dk
Charlotte Buchwaldt	cbu@kvl.dk
Henrik Christensen	Henrik.Christensen@vetmi.kvl.dk
Lotte Dahl	lod@kvl.dk
Claus Dethlefsen (fjern-overnatning)	dethlef@math.auc.dk
Claus Ekstrøm	C.Ekstroem@biostat.ku.dk
Jeppe Emmersen (fjern-overnatning)	i5je@civil.auc.dk
Susanne Gammelgaard Bøttcher (fjern-overnatning)	alma@math.auc.dk
Britta Gavnholt	britta.gavnholt@risoe.dk
Bernt Guldbrandtsen (ingen overnatning)	Bernt.Guldbrandtsen@agrsci.dk
Kirsten Bagge Holm	kirsten.bagge.holm@risoe.dk
Michael Karsten Höhle	hoehle@dina.kvl.dk
Malene Højbjerg	malene@math.auc.dk
Maria Johansson	maria.johansson@vpat.slu.se
Nina Jøhnk	nij@kvl.dk
Steen Randers Knudsen (ingen overnatning)	steenrknudsen@hotmail.com
Anders Ringgaard Kristensen	ark@dina.kvl.dk
Lene Lind Mikkelsen	lenel.mikkelsen@agrsci.dk
Mogens Sandø Lund (ingen overnatning)	mogens.lund@agrsci.dk
Mads Lundsgaard (fjern-overnatning)	I5ml@civil.auc.dk
Helle Frank Mortensen (ingen overnatning)	hfm@svs.dk
Kirstine Klitgaard Nielsen	kni@svs.dk
Peter H. Nissen	PeterH.Nissen@agrsci.dk
Britta Nylin	bny@svs.dk
Hanne Gro Olsen	hanne-gro.olsen@ihf.nlh.no

Ulf Olsson	Ulf.Olsson@sdi.slu.se
Karen-Margrethe Pedersen	kmh@kvl.dk
Hannu Rita	rita@silvia.helsinki.fi
Peter Sestoft	sestoft@dina.kvl.dk
Gerhard Skagestein	gerhard.skagestein@imf.nlh.no
Kenneth Smith	ksm@uni-c.dk
Henrik Stryhn	hes@svs.dk
Kirsten Kørup Sørensen (ingen overnatning)	KirstenK.Sorensen@agrsci.dk
Peter Sørensen (ingen overnatning)	psø@genetics.agrsci.dk
Nils Toft	nt@dina.kvl.dk
Mia Torpdahl	mto@svs.dk
Runa Ulsøe Johansen (fra torsdag eftermiddag)	runa.ulsoe@risoe.dk
Jan Vintov	jvi@svs.dk
Peter Worning	
Jukka Öfversten	JUKKA.OFVERSTEN@MTT.FI