



Bioinformatik

Open Source Software i biologiens tjeneste

Kenneth Geisshirt

`kneth@silex.dk`

Silex Science ApS

Om Silex Science ApS

- Grundlagt maj 2002
- Ejeren er Cortex Holding

Fokusområderne hos Silex Science ApS er:

- High Performance Computing
- Open Source-baseret videnskabeligt og teknisk programmel
- Molekylær visualisering og modellering
- Bioinformatik

Molekylær biologi

Molekylær biologien undersøger levende væsener på det molekylære plan.

- En del af molekylær biologiens data er sekvenser.
- En sekvens er en streng over et endeligt alfabet $\mathcal{A}$

DNA

- DNA er “arvemassen” - de sekvenser som bærer informationen om et levende væsen
- Alfabetet $\mathcal{A}$ består af fire bogstaver (baser): A, C, G, T.
- DNA er dobbelt-strengt
- A–T og G–C sidder overfor hinanden
- Informationen grupperes i gener

Menneskets genom består af 3×10^9 basepar!

Proteiner

Proteiner er kroppens arbejdsheste.

- Gener (DNA) oversættes til proteiner
- Alfabetet består af de 20 aminosyrer
- Proteiner kan modificeres efter oversættelse
- Sekundær struktur er en sekvens over alfabetet H, E, C
- 3-dimensional struktur giver funktionen

Proteiner

- Hormoner er små proteiner (peptider)
- Enzymer katalyserer biokemiske reaktioner
- Membranproteiner kan transportere molekyler over cellemembranen
- Transport af kemiske forbindelser, f.eks. hæmoglobins transport af oxygen

Offentlige databaser

Op gennem 1990'erne blev forskellige organismers arvemasse kortlagt.

GenBank indeholder alle DNA-sekvenser (100 GB)

SwissProt har proteinsekvenser samt annoteringer om funktion

PDB er til 3d-struktur af proteiner

Alle har adgang til disse databaser.

Bioinformatik

Bioinformatik er anvendelsen af informationsteknologi til håndteringen af biologiske data.

Det er en tværfaglig disciplin:

- Molekylær biologi og biokemi
- Datalogi
- Statistik
- Kemi og fysik

Bioinformatik i praksis

En typisk bioinformatik-gruppe i industrien:

- Computerne kører UNIX (og ofte Linux)
- Programmeringssproget er Perl
- Medlemmerne har en ph.d.-grad i biokemi eller molekylær biologi
- Stor åbenhed omkring teknologi

NCBI Toolkit

CoreLib er et generisk bibliotek i ISO C

AsnLib til håndtering af Abstract Syntax Notation
1 (ASN.1)

Vibrant bruges til grafisk brugergrænseflader

SEQ* indeholder funktioner til håndtering af
sekvenser

Licens: public domain

BLAST

Basic Local Alignment Search Tool

- Søgning af sekvenser
- Alignment
- Bygger på NCBI Toolkit

```
A  T  -  C  -  T  G  A  T
-  T  G  C  A  T  -  A  -
```

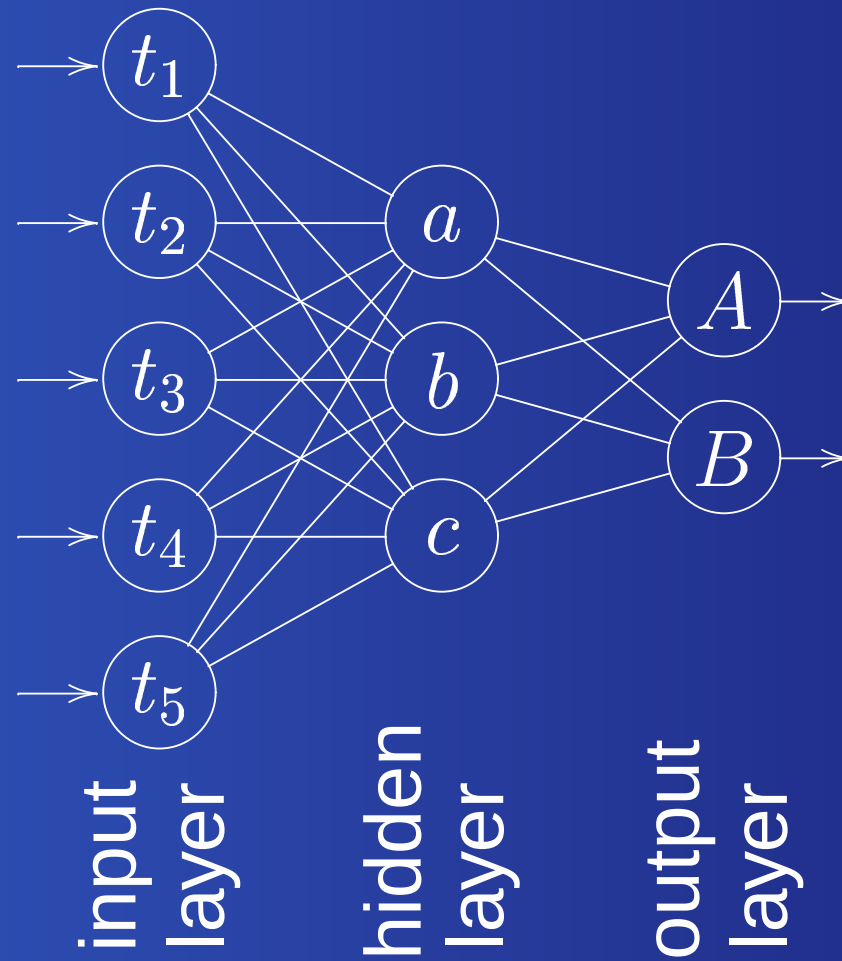
Licens: public domain

Predator

- Forudsigelse af sekundær struktur
- Skrevet i ISO C
- Benytter neurale netværk
- Nøjagtigheden er ca. 68 %

Licens: kræver at en bestemt artikel citeres.

Neurale netværk



EMBOSS

European Molecular Biology Open Software Suite

- Et bibliotek skrevet i ISO C (til UNIX)
- En række (kommando-linje) programmer (100+)
- Undersøtter *Uniform Sequence Addresses*

Licens: GNU LGPL (bibliotek) og GNU GPL (programmer)

EMBOSS - eksempler

Rotte-gener:

```
seqret genbank::nc1102.seq:org:Rattus  
       genbank::rats.seq
```

Gram-positive cocci surface proteins anchoring
hexapeptide:

```
preg ncbi::swissprot 'lp.tg[stgavde]'  
     -outfile swissprot.preg
```

Statistik:

```
pepstats ncbi::swissprot  
         -outfile swissprot.stat
```

BioPerl

BioPerl er et større bibliotek i Perl:

- Håndtering af sekvenser (læse, skrive, intern repræsentation)
- Kald af eksterne programmer (BLAST, Hmmer, etc,)
- Håndtering af “alignment” formater

Licens: Artistic

BioPerl - eksempel

```
use Bio::Seq;
use Bio::SeqIO;
$seq = Bio::Seq->new(-id => "myseq",
    -seq => "CGCCGAAGAAGCATCGTTAAAGTCTCTCTTCACCCTGCC",
    -type => 'Dna');
$seqout = Bio::SeqIO->new('-format' => 'fasta',
    -fh => \*STDOUT);
$trans = $seq->translate(); # make a translation
$seqout->write_seq($trans);
```

Litteratur

- *Developing Bioinformatics Computer Skills.* C. Gibas & P. Jambeck. O'Reilly & Associates, 2001.
- *Introduction to Computational Biology.* M.S. Watermann. Chapman & Hall/CRC, 1995.
- *Bioinformatics: The Machine Learning Approach.* P. Baldi & S. Brunak. MIT Press, 2001.
- *Computational Molecular Biology - An Algorithmic Approach.* P.A. Pevzner. MIT Press, 2000.

Hjemmesider

- Det humane genom-projekt:
<http://www.jgi.doe.gov/programs/hgp.htm>
- National Center for Biotechnology
Information: <http://www.ncbi.nih.gov>
- Open Source Bioinformatics:
<http://bioinformatics.org>
- *How Perl Saved the Human Genome Project*
af L. Stein:
http://bioperl.org/GetStarted/tpj_ls_bio.html