ALGODAN

Sequence analysis

Esko Ukkonen

University of Helsinki
Helsinki University of Technology

Overview

- sequences are everywhere
- combinatorial pattern matching
- pattern discovery in sequences
- dynamic programming, automata theory, advanced data structures, probabilistic modeling

Background

- algorithms on strings and biological sequence analysis studied by the group since 1980
- lots of our earlier results appear in textbooks in the field
  - Gusfield, Cambridge Univ Press 1997
  - Crochemore and Rytter, Oxford Univ Press 1994
  - Navarro and Raffinot, Cambridge Univ Press 2002
  - Smyth, Addison-Wesley 2003
  - Wikipedia

Main tasks in the theme

S.1 String algorithms

- Ukkonen, Mannila, Toivonen

S.2 Finding orders from data

- Mannila, Hyvärinen, Ukkonen

S.3 Sequence segementation and structure

- Mannila, Toivonen, Ukkonen

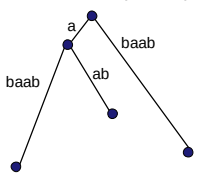
S.4 Kernel algorithms for sequences

- Kivinen, Ukkonen

Example result (suffix-array construction)

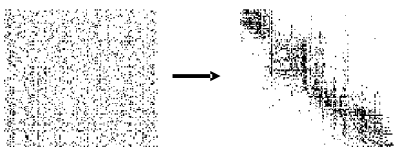
- direct construction of a suffix array in linear time
  - immediately included in teaching materials internationally
  - J. Kärkkäinen, P. Sanders, S. Burkhardt: *Linear work suffix array construction*, J. ACM (in press), ICALP

SuffixTree(abaab)

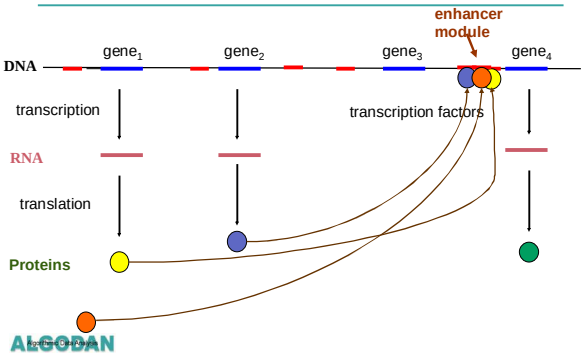


Example (finding orders from data)

- Seriation problem for 0-1 data in paleontology:
  - Find an ordering of the rows such that the 1s are as consecutive as possible
- No errors à polynomial
- Errors à NP-hard
- Spectral techniques; MCMC; finding partial orders (Paleobiology 2006; PLoS Comput Biol 2006; KDD 2006)



Example (finding gene enhancer modules)

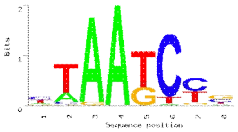


ALGODAN

Binding affinity matrices

- Transcription factor binding sites represented by affinity matrices
- Discovered:
  - Computationally
  - Traditional wet lab
  - Microarrays

9	11	49	51	0	1	1	4
19	3	0	0	0	45	25	16
5	1	2	0	17	0	4	21
18	36	0	0	34	5	21	10

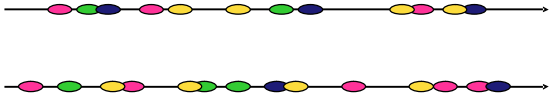


ALGODAN

Good binding sites

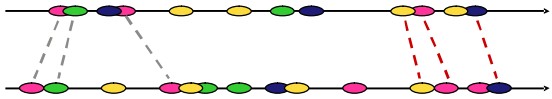
ALGODAN

Clustering and conservation



ALGODAN

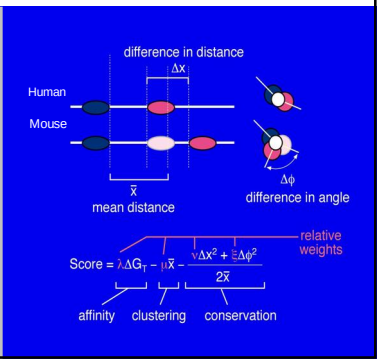
Clustering and conservation



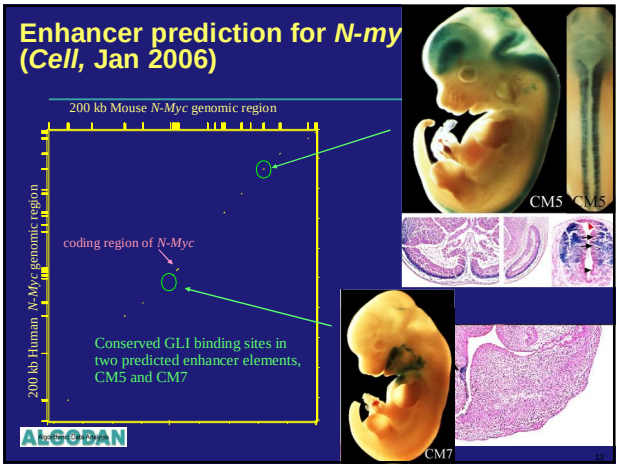
ALGODAN

Computational identification of enhancer elements

- Preserved in evolution:
  - Affinities of functional cis-elements.
  - Spatial arrangement of elements within a module.



Enhancer prediction for *N-myc* (*Cell*, Jan 2006)



ALGODAN

Regulatory modules in gene regulatory networks

- Still today, organ specific growth control remains poorly understood and difficult to study with traditional genetics
- Now we can start to investigate the "program of development" in top-down fashion instead of bottom-up
  - Soon (within 10 years) the binding specificities of all TFs will be known (?)
  - Tools like EEL can be used to build genome-wide predictions of transcriptional regulation
- Central idea: Developmental regulatory circuits are embedded in the genome-wide transcriptional network but extracting this knowledge requires new computational tools

ALGODAN

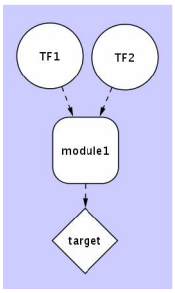
Model

- The control system for growth consists mainly of
  - Transcription factors (TFs)
    - proteins that bind to DNA
  - Regulatory modules
    - clusters of TF binding sites in the regulatory areas of genes that specify the interactions of TFs with the regulated genes.
- A module acts as a **logic gate** which
  - input is the expression levels of the TFs and
  - output is the expression of the regulated gene
- The regulatory interactions form a **complex logic circuit**
  - **One TF usually interacts with many modules and the genes encoding TFs are themselves controlled using the same mechanism**

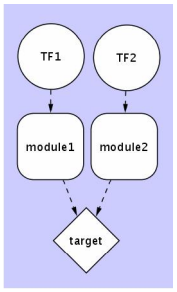
ALGODAN

Model: Logical Building Blocks

- AND/OR

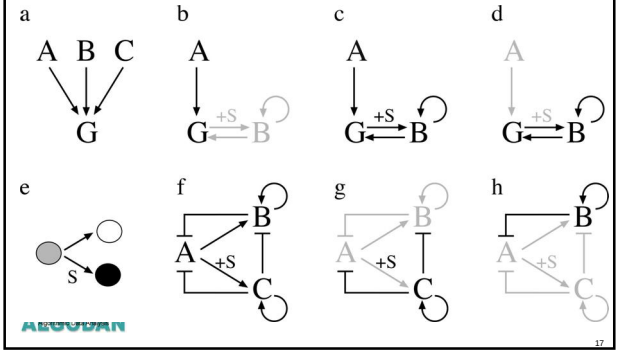


- OR

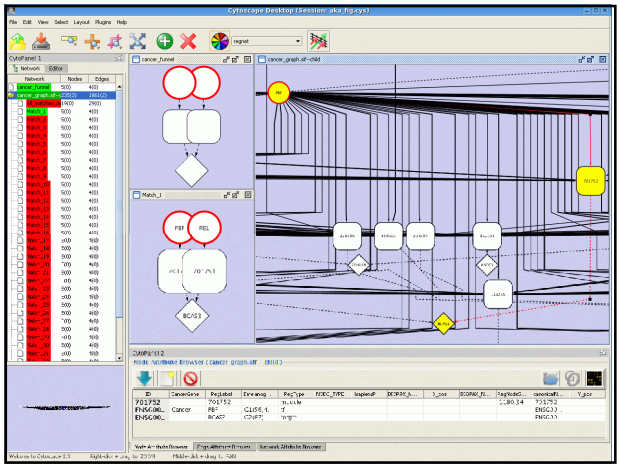


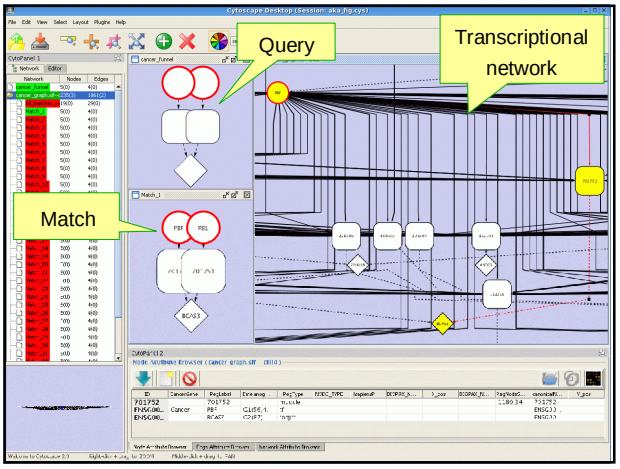
ALGODAN

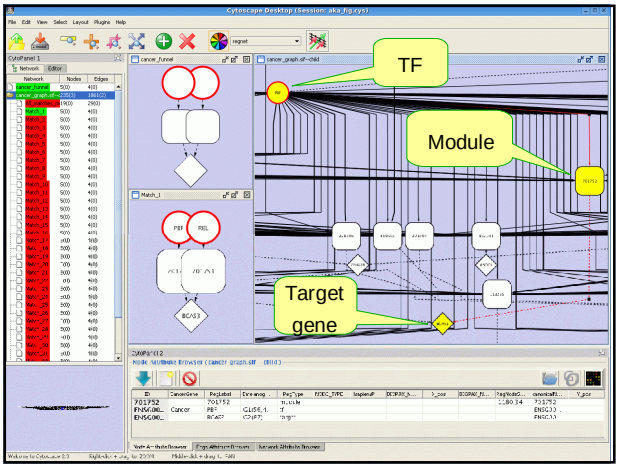
Biological hypothesis

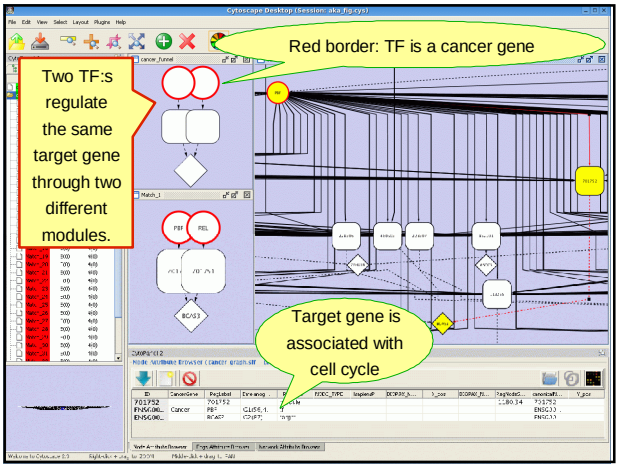


ALGODAN



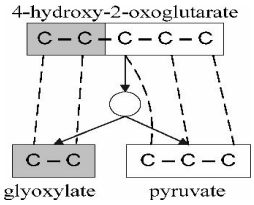






^{13}C Metabolic Flux Analysis

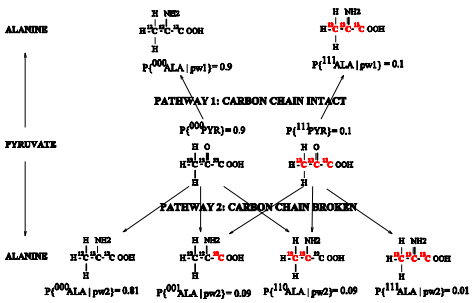
- Metabolites (chemical compounds)
- Metabolic reactions
- Enzymes
- labelings by ^{13}C



ALGODAN

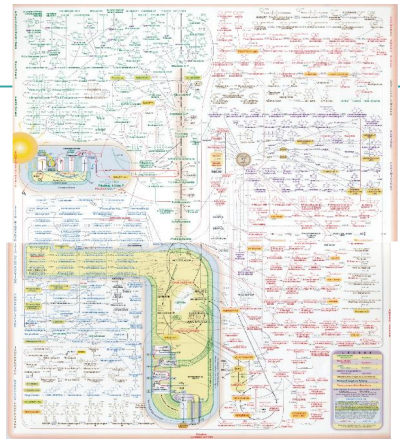
22

^{13}C metabolic flux analysis



ALGODAN

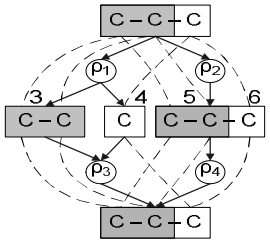
23



ALGODAN

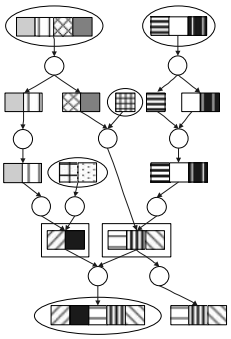
24

^{13}C Metabolic Flux Analysis



ALGODAN

^{13}C Metabolic Flux Analysis



ALGODAN

^{13}C Metabolic Flux Analysis

- Problem: Estimate the fluxes of metabolites
- Analogous problem: How much gin and how much tonic in gin-tonic?
- Measure the alcohol contents => solution of the problem
  - **Alcohol content ≈ isotopomer content**
  - **gin: 40 %, tonic: 0 % , gin-tonic: 10 %**

    Vg + Vt = Vs

    0,4*Vg + 0*Vt = 0.1*Vs

    ⟶

    gin-tonic: ¼ gin, ¾ tonic

ALGODAN

Future goals

- Indexing sequential data for approximate searches?
- Distance functions for sequences?
  - General theoretical framework, efficient evaluation algorithms, complexity bounds, relations between distances
  - Application specific distances
  - Generalizations to different generalized sequences (XML, images, music, …)
- Finding structure and signals in sequences
  - Supervised and unsupervised learning of signals
  - Statistical significance of the findings

ALGODAN

Future goals (cont.)

- How does the program encoded in genomes work?
- Integrated analysis of genomic, proteomic, and metabolic data
  - metabolic modeling (new important area in bioinformatics)
  - metabolic fluxes
- Challenge problems
  - Learning network structures
  - Finding structure and knowledge in natural language data
  - The vocabulary, grammar and history of genomes

ALGODAN