Studying gene regulation by data mining approaches

Jaak Vilo
vilo@egeen.ee

EGeen

Estonian Biocentre
Eesti Biokeskus Riia 23b Tartu 51010 Estonia

Tartu Ülikool
University of Tartu

EMBL-EBI
European Bioinformatics Institute

Estonian Computer Science Theory days: Pedase, 3.10.2003

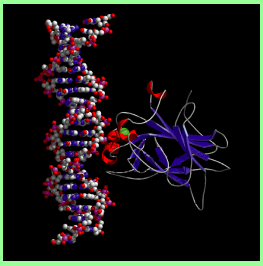
DNA determines function?

DNA → **PROTEIN** → **STRUCTURE**

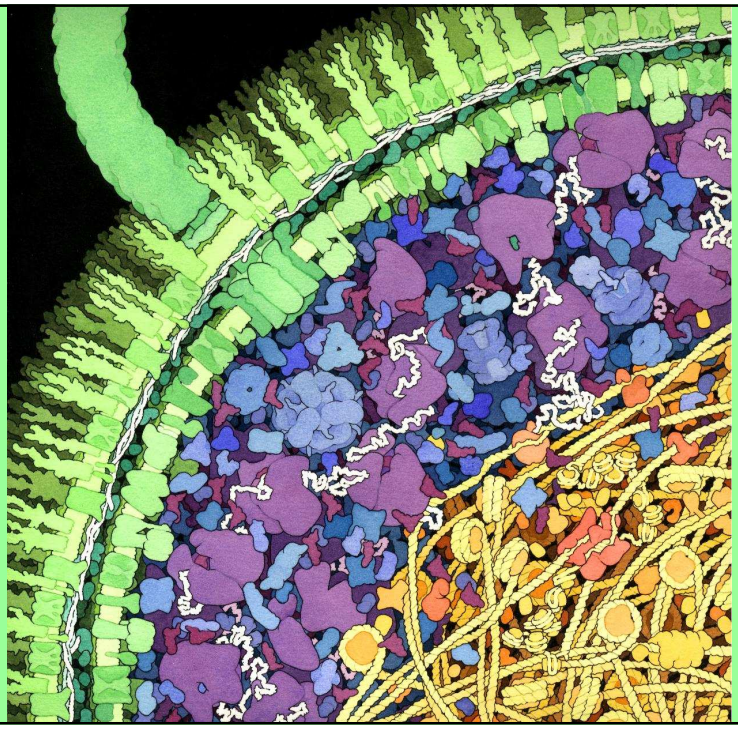
GenBank / EMBL Bank | SwissProt/TrEMBL | PDB/Molecular Structure Database

4 Nucleotides

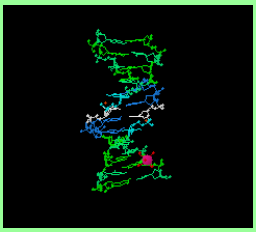
20+ Amino Acids
(3nt → 1 AA)



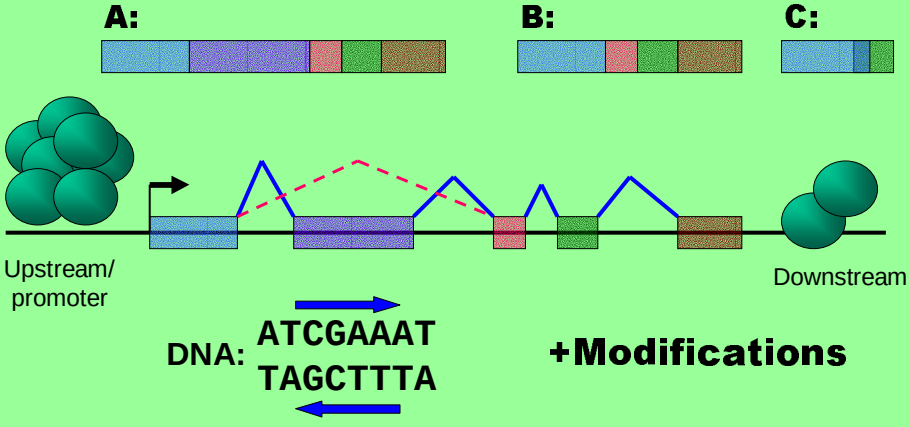
Function?
Dynamics?



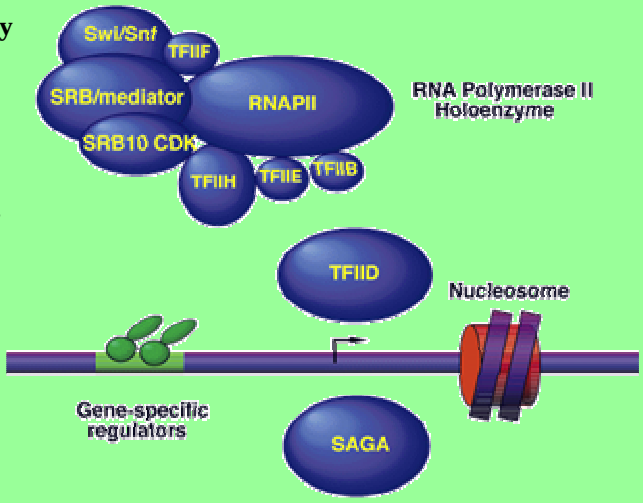
David S. Goodsell
http://www.scripps.edu/pub/goodsell/



A Simple Gene



Model of RNA Polymerase II Transcription Initiation Machinery. **The machinery depicted here encompasses over 85 polypeptides in ten (sub) complexes**: core RNA polymerase II (RNAPII) consists of 12 subunits; TFIIH, 9 subunits; TFIIE, 2 subunits; TFIIF, 3 subunits; TFIIB, 1 subunit, TFIID, 14 subunits; core SRB/mediator, more than 16 subunits; Swi/Snf complex, 11 subunits; Srb10 kinase complex, 4 subunits; and SAGA, 13 subunits.

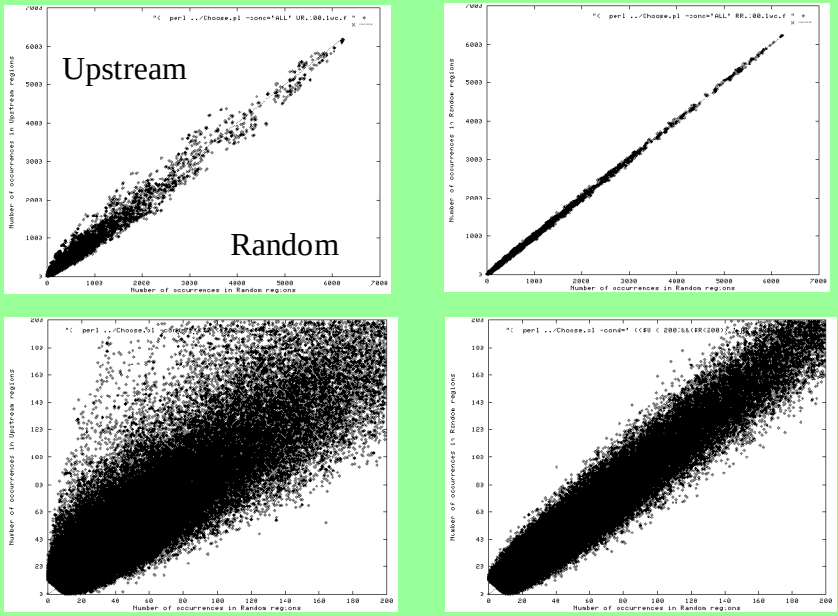


F.C.P. Holstege, E.G. Jenning, J.J. Wyrick, Tong Ihn Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young
Dissecting the Regulatory Circuitry of a Eukaryotic Genome
Cell 95: 717-728 (1998)

TGTTCTTTCTTCTTTCATACATCCTTTTCCTTTTTTTCC
TTCTCCTTTCATTTCCTGACTTTTAATATAGGCTTACCA
TCCTTCTTCTCTTCAATAACCTTCTTACATTGCTTCTTC
TTCGATTGCTTCAAAGTAGTTCGTGAATCATCCTTCAAT
GCCTCAGCACCTTCAGCACTTGCACTTCATTCTCTGGAA
GTGCTGCACCTGCGCTGTCTTGCTAATGGATTTGGAGTT
GGCGTGGCACTGATTTCTTCGACATGGGCGGCGTCTTCT
TCGAATTCCATCAGTCCTCATAGTTCTGTTGGTTCTTTT
CTCTGATGATCGTCATCTTTCACTGATCTGATGTTCCTG
TGCCCTATCTATATCATCTCAAAGTTCACCTTTGCCACT
TTCCAAGATCTCTCATTCATAATGGGCTTAAAGCCGTAC
TTTTTTCACTCGATGAGCTATAAGAGTTTTCCACTTTTA
GATCGTGGCTGGGCTTATATTACGGTGTGATGAGGGCGC
TTGAAAAGATTTTTTCATCTCACAAGCGACGAGGGCCCG
AGTGTTTGAAGCTAGATGCAGTAGGTGCAAGCGTAGAGT
CTTAGAAGATAAAGTAGTGAATTACAATAGATTCGATAC

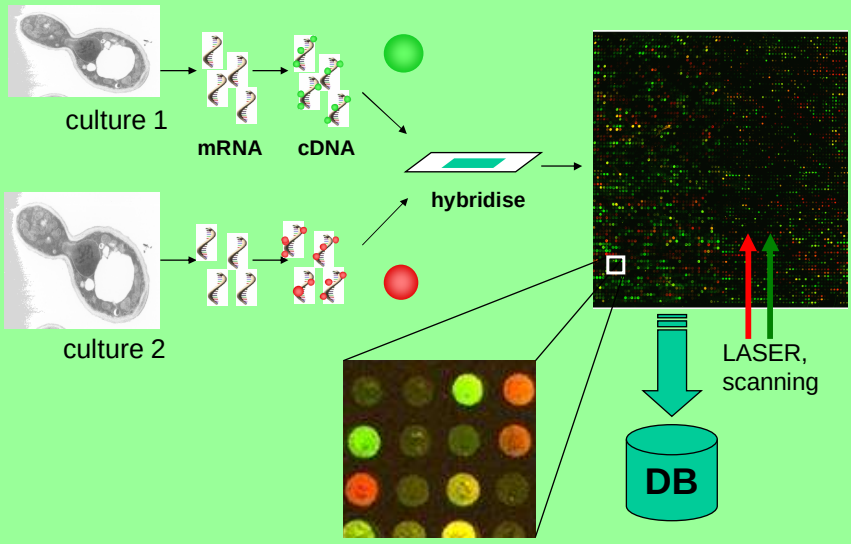
Patterns: AT

Patterns: [AT][ACT]AT (WHAT)



Genome Research, 1998

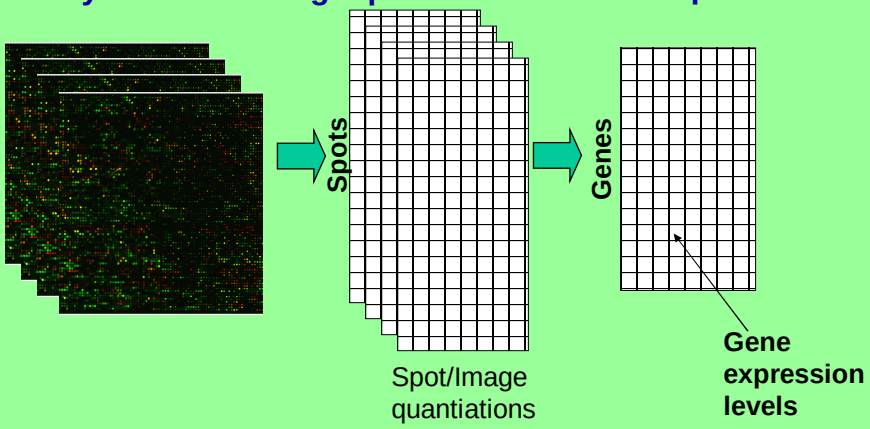
Analysis of biological samples with microarrays



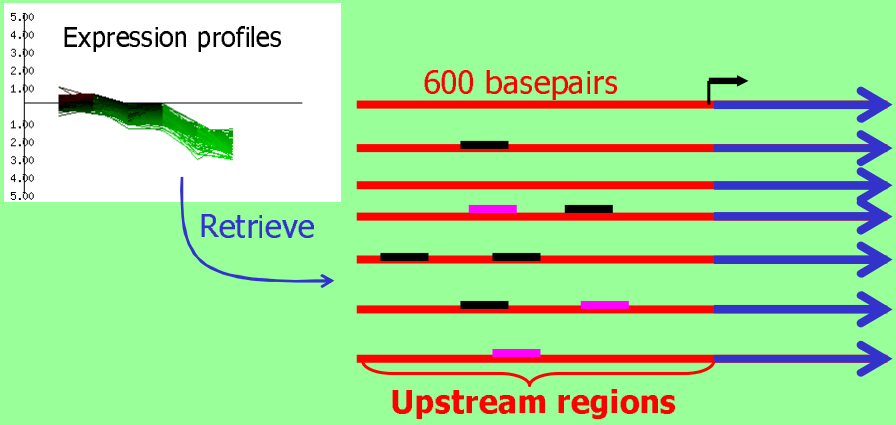
From microarray images to gene expression data

Raw data **Intermediate data** **Final data**

Array scans **Image quantifications** **Samples**

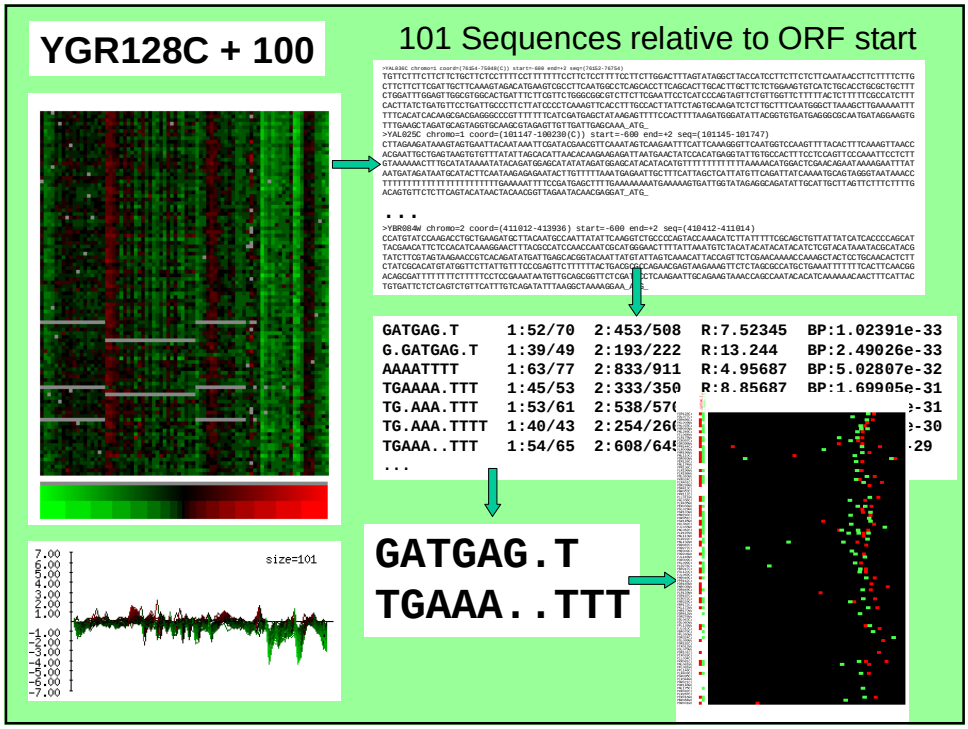


Cluster of co-expressed genes, pattern discovery in regulatory regions



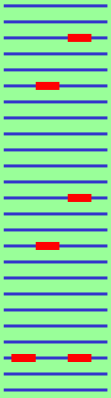
Find patterns over-represented within cluster

Genome Research 1998;
ISMB (Intelligent Systems in Mol. Biol.) 2000



Pattern selection criteria
Binomial distribution

Background - ALL upstream sequences



Cluster:

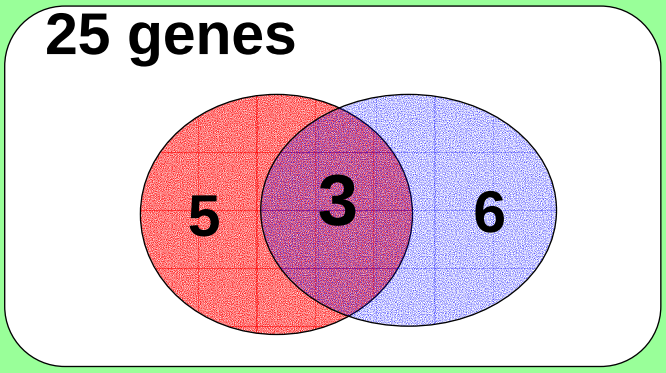
π occurs 3 times

P(3,6,0.2) is probability of having ≥3 matches in 6 sequences

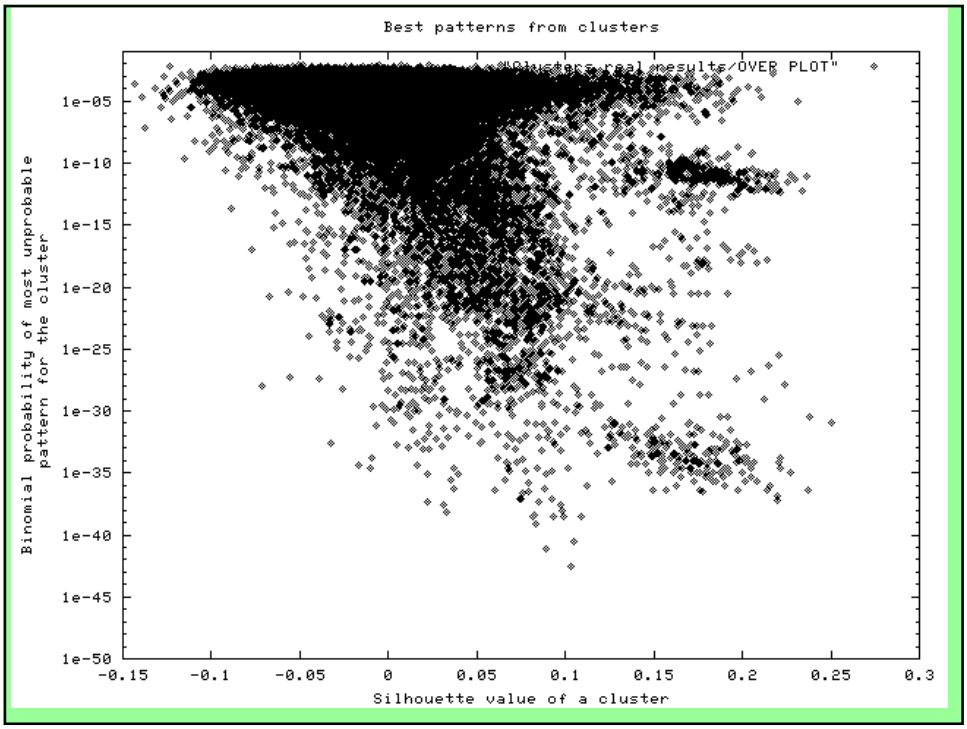
P(π,3,6,0.2) =0.0989

5 out of 25, ***p* = 0.2**

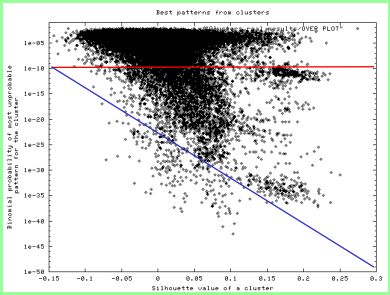
Set overlap



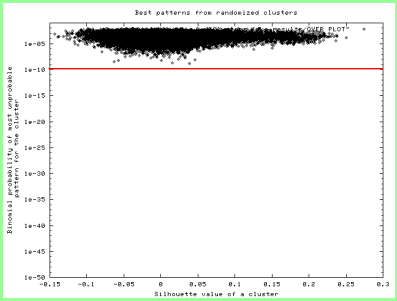
P(choose 6 balls randomly from 25, of which 5 reds, and observe 3 or more red)



Pattern vs cluster “strength”



The pattern probability vs. the average silhouette for the cluster



The same for randomised clusters

Vilo *et.al.* ISMB 2000

Regular patterns (SPEXS)

- Substrings ATCGA
- Add groups ATC[GC][AT]
- Add (unrestricted) wildcards AT*CG
- Add restricted wildcards AT*(2,5)CG
- Combine all above

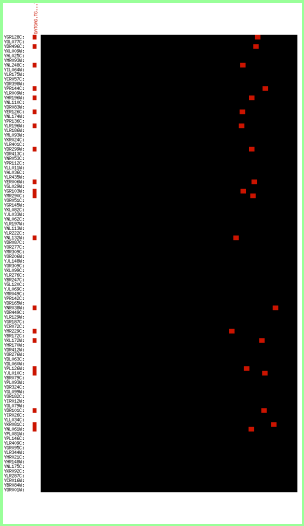
AT[GC]*(1,3)[GT]AC

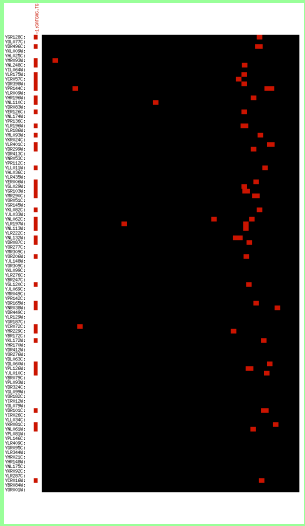
TGC……ACG

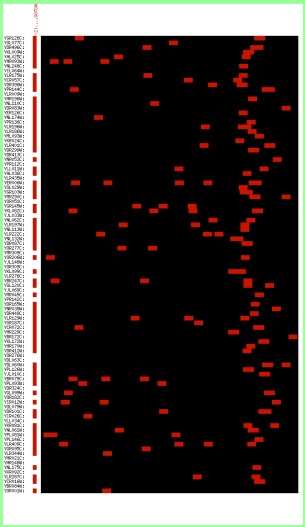
GATGAG.TG...T : 20/98 vs 6/1000

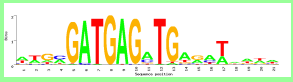
-1: GATGAG.TG...T : 40/98

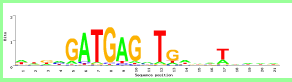
-2: GATGAG.TG...T : 75/98









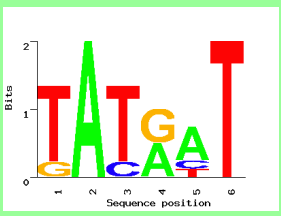


Consensus matrix building

```
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
```

A	0	(6)	0	(3)	(4)	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	(5)	0	(5)	0	1	(6)

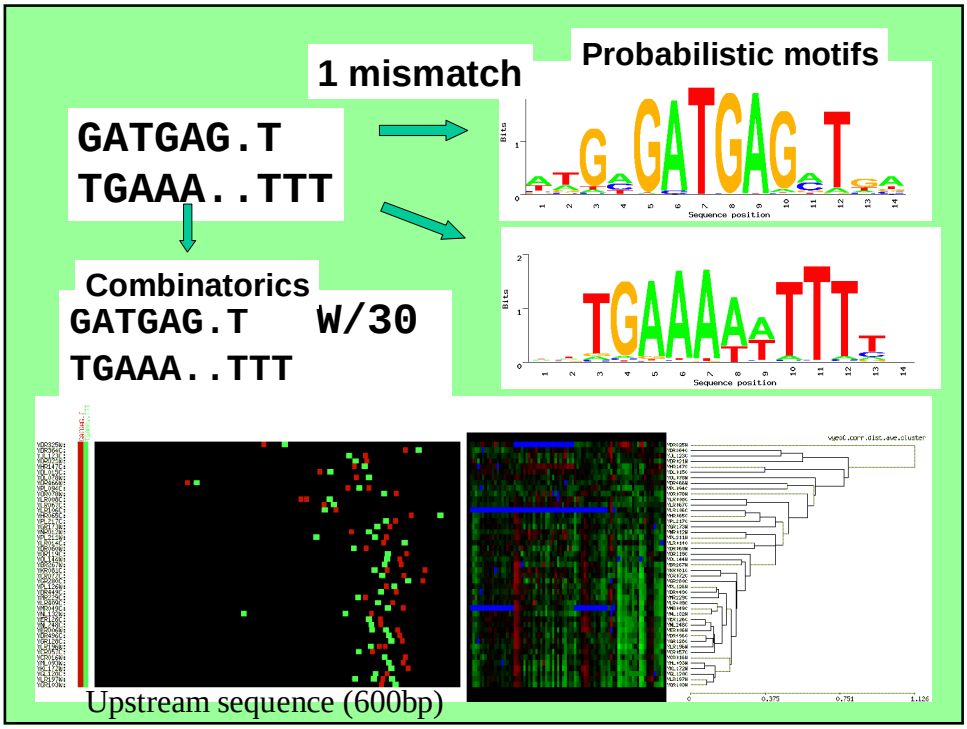
$$I_i = 2 + \sum_{b \in A.C.T,G} f_{b,i} \log_2 f_{b,i}$$

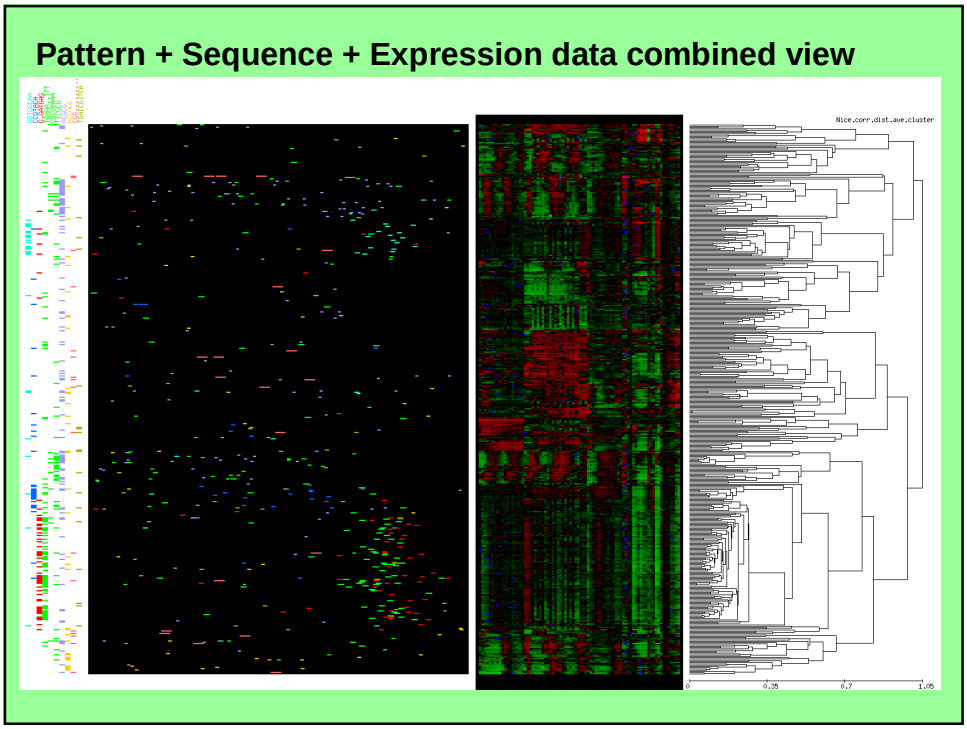


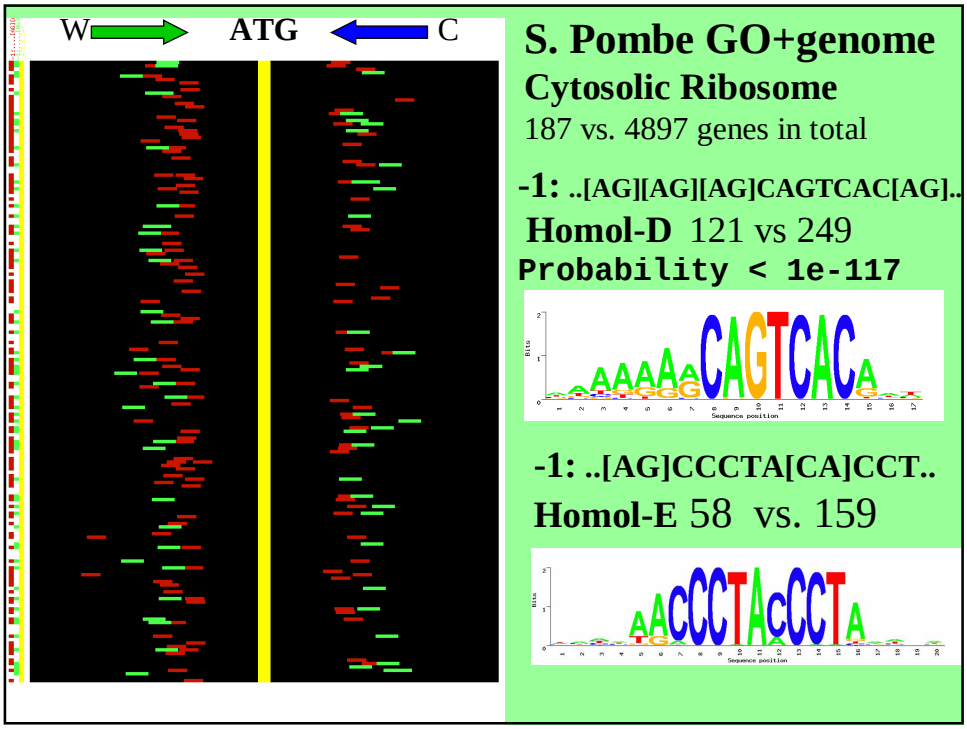
$$- f_{b,i} \log_2 f_{b,i}$$

Consensi:

```
TATAAT
TATRNT
[GT]A[CT][AG][ACT]T
```


1 mismatch
Probabilistic motifs
GATGAG.T
TGAAA..TTT
Combinatorics
GATGAG.T W/30
TGAAA..TTT
Bits
Sequence position
Upstream sequence (600bp)


Pattern + Sequence + Expression data combined view



SPEXS - **S**equence **P**attern **EX**haustive **S**earch

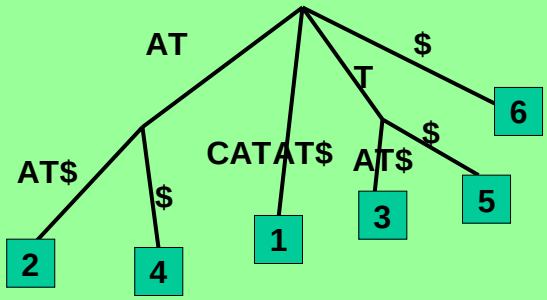
Jaak Vilo, 1998, 2002

- **User-definable pattern language**: substrings, character groups, wildcards, flexible wildcards (c.f. **PROSITE**)
- Fast exhaustive search over pattern language
- “Lazy suffix tree construction”-like algorithm
- **Analyze multiple sets of sequences simultaneously**
- Restrict search to most frequent patterns only (in each set)
- **Report** most frequent patterns, patterns over- or underrepresented in selected subsets, or patterns significant by various statistical criteria, e.g. by binomial distribution

Suffix tree – represent all suffixes

CATAT => suffix tree

```
123456
CATAT$ 1
 ATAT$ 2
  TAT$ 3
   AT$ 4
    T$ 5
     $ 6
```

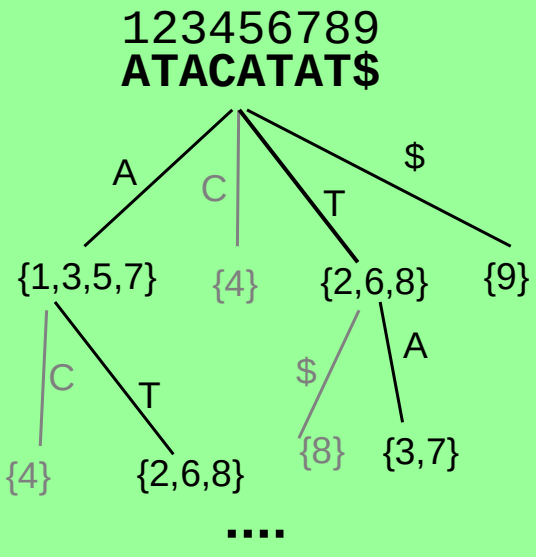


O(n) time and space

"Lazy" construction of trie

ATACATAT$

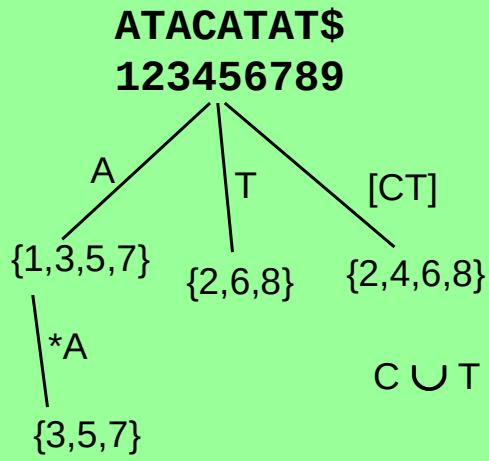
- Suffix trie
- $O(n^2)$
- Kurtz, Giegerich
- Good in practice



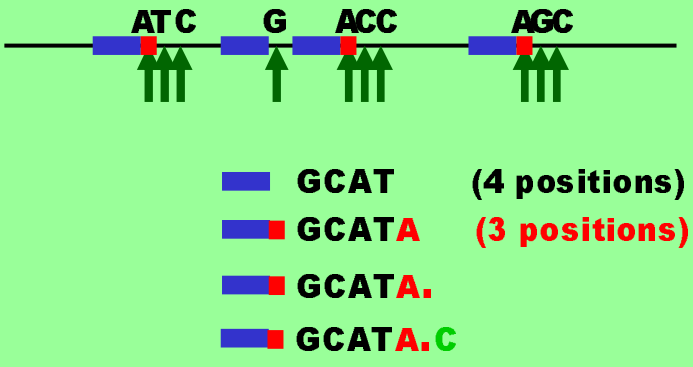
SPEXS: pattern discovery based on pattern trie.

- Substrings
- Group characters
- Wildcard positions
- Variable length wildcards
- Restrictions on the number on each separately
- **At least k occurrences**
- Exact occurrences locations for each pattern

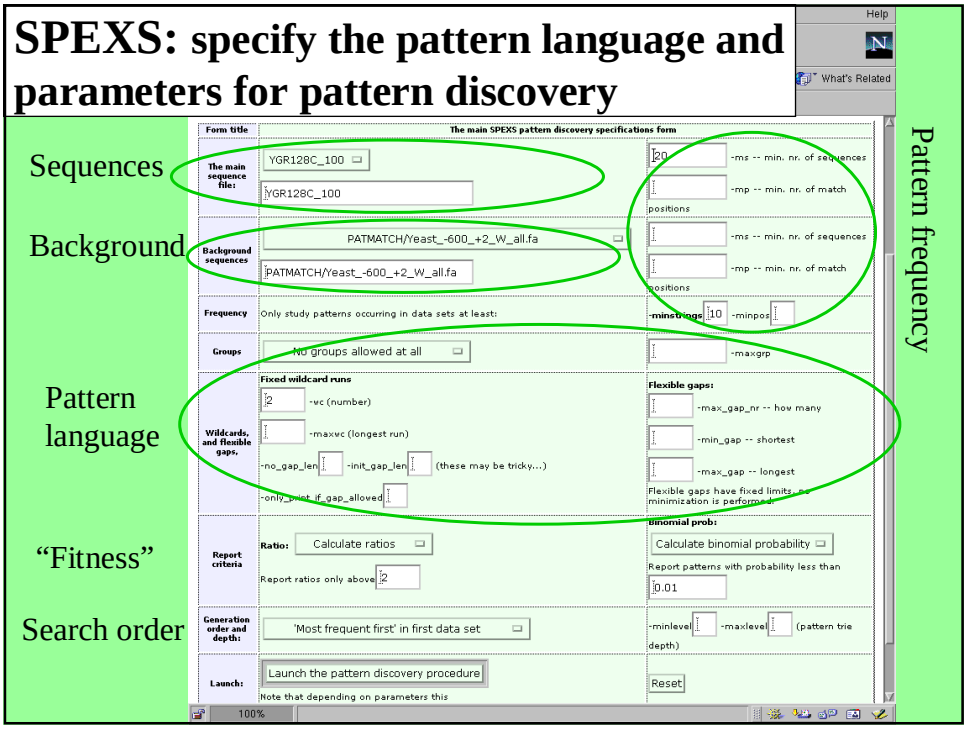
Vilo 1998, 2002



Sequence patterns: the basis of the SPEXS



SPEXS: specify the pattern language and parameters for pattern discovery



Combinatorics of sites

- Which binding sites tend to co-occur frequently together in upstreams
- Association rules data mining
  - **#(A,B) = 200 , #(A,B,C) = 180**
  - **A,B => C (90%)**

- ***Alvis Brazma, Jaak Vilo, Esko Ukkonen and Kimmo Valtonen***
  Data Mining for Regulatory Elements in Yeast Genome.
  Fifth International Conference on Intelligent Systems for Molecular Biology, ISMB-97 (pp. 65-74) June, 1997. *AAAI Press.*

Research goals

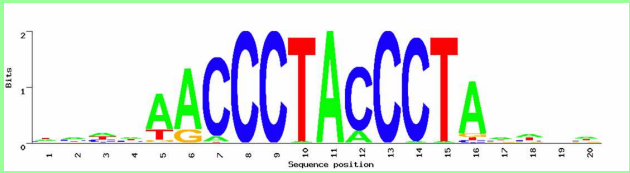
- **Generate a (full) list of hypothetical regulatory signals (for each and every gene)**
- **Maintain a DB of all known or predicted inf.**
- How does this correlate to known information and/or experimental data (e.g. ChIP on chip)
- Predict from unannotated DNA where are the promoters (and genes)
- Predict from DNA how the gene is expressed given concentrations of all TF-s in cell
- Predict the alternative splicing isoforms
- Evolution & comparative genomics approaches

How to know what is known?

- After in silico predictions the first question should be
  - **How does that compare to current knowledge?**
- But what if databases do not allow to answer such questions easily?

Similarity?
Fast (approximate) search?

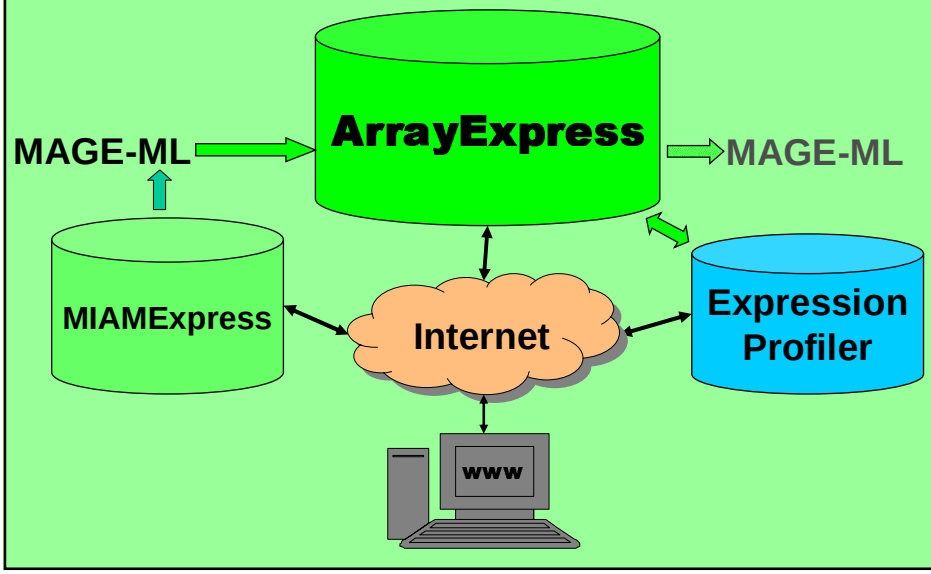
CCTAGTAG

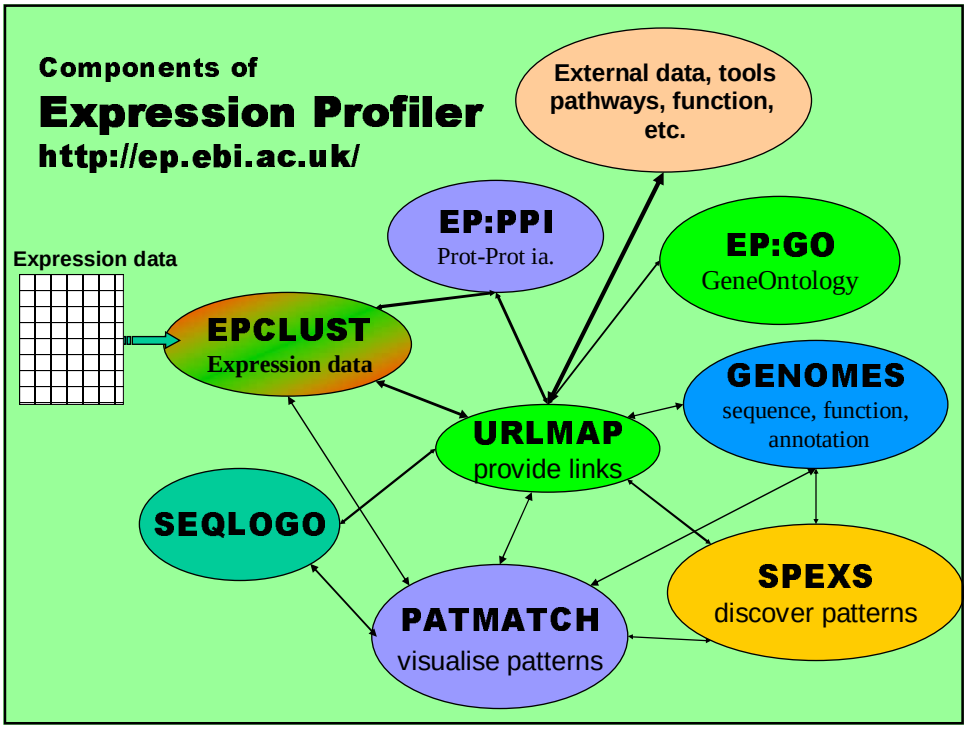


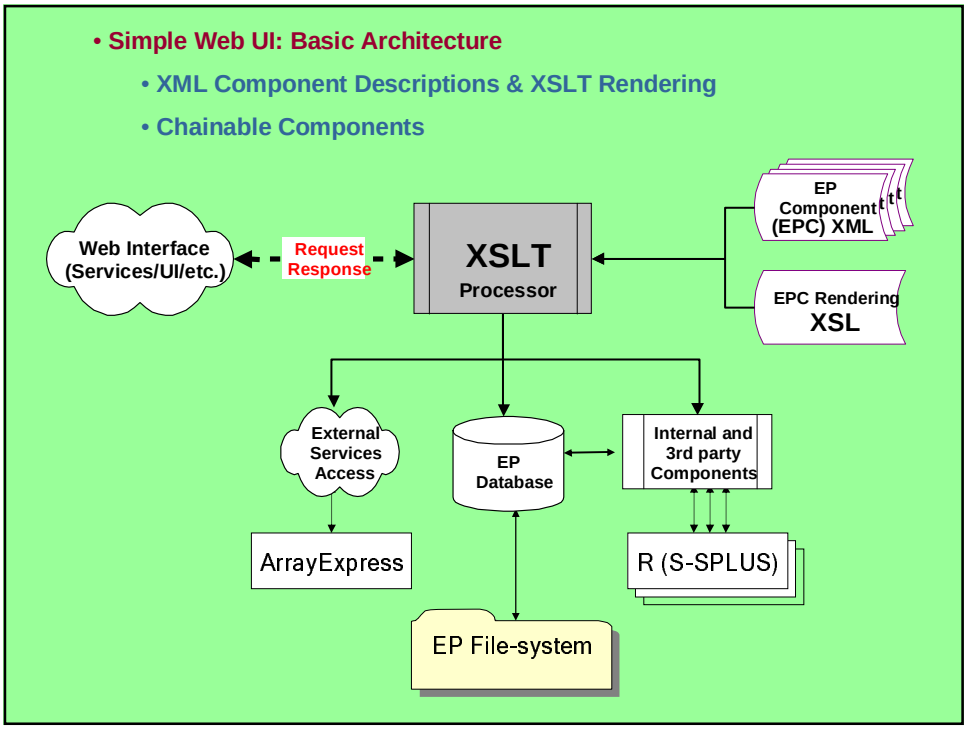
GTAA..CCT..CCT

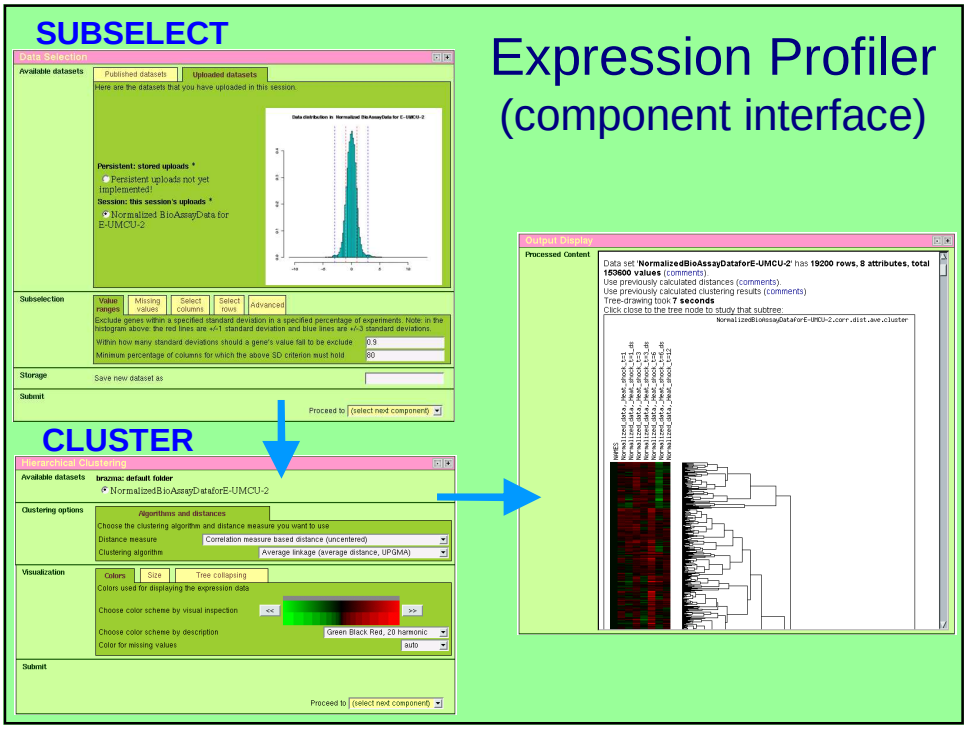
ArrayExpress (EBI)

a public repository for microarray gene expression data




Components of
Expression Profiler
http://ep.ebi.ac.uk/
External data, tools pathways, function, etc.
EP:PPI
Prot-Prot ia.
EP:GO
GeneOntology
Expression data
EPCLUST
Expression data
GENOMES
sequence, function, annotation
URLMAP
provide links
SEQLOGO
SPEXS
discover patterns
PATMATCH
visualise patterns


• Simple Web UI: Basic Architecture
• XML Component Descriptions & XSLT Rendering
• Chainable Components
Web Interface (Services/UI/etc.)
Request
Response
XSLT
Processor
EP Component (EPC) XML
EPC Rendering XSL
External Services Access
EP Database
Internal and 3rd party Components
ArrayExpress
R (S-SPLUS)
EP File-system



Projects started at Tartu

- **Database for gene regulation information**
- **Tools for using that database**
- **Pattern matching and discovery**
- **Alternative splicing regulation (3yr EU project)**
- **(Fast) gene expression data clustering**
- **Data mining seminar series;**
- **other DM and ML methods**
- **...**

Gene regulation

- Promoter analysis, also in higher eukaryotes
- Alternative Splicing data analysis
- Genetic networks
- Gene expression data analysis
- Integration of many different data types
  - Protein-protein interactions
  - Phenotypes
- Metabolic pathways
- Signaling pathways

Pattern discovery

- Pattern discovery and pattern matching in sequences; sequence algorithms

- Regulatory sequence analysis
- GPCR receptor bioinformatics

Data mining

- **Data mining methods development for bioinformatics**
  - **Fast clustering methods**
  - **Gene networks and regularities**
  - **Machine learning methods**
- **Text Mining**
  - **Information extraction**
  - **categorization**
- **Information retrieval, dictionaries**
- **Medical and clinical data handling and storage, population and statistical genetics, pharmacogenetics.**

Software engineering

- **Database development, software engineering**
  - **UML based development and code generation**
  - **XML based UI-s**
- **Expression Profiler**
- **Farm and GRID computing**

Acknowledgements

Alvis Brazma

Misha Kapushesky

+ the EBI microarray team

Frank Holstege, and **Patrick Kemmeren**, UMC Utrecht

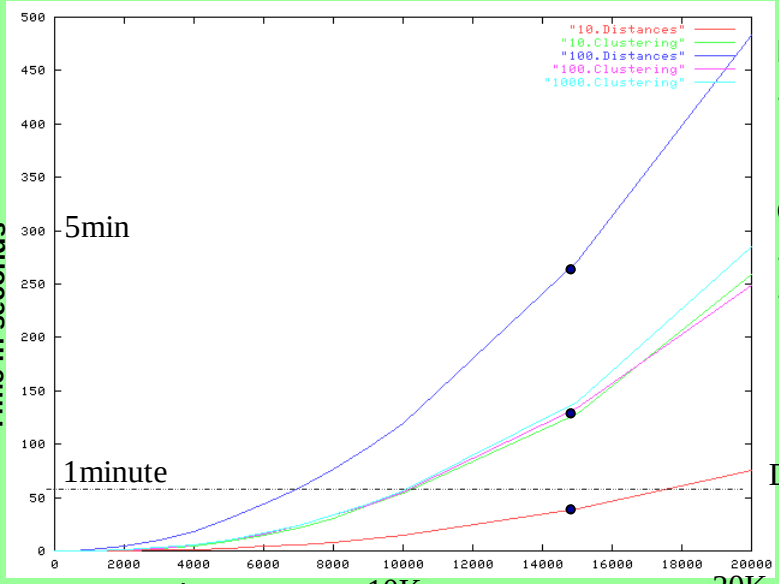
Mike Croning, Steffen Möller (ex EBI)

Esko Ukkonen, U. of Helsinki

Inge Jonassen, Bergen U.

Meelis Kull
Hedi Peterson
Ireen Meho
…
+ ~ 10 new faces

Running time for hierarchical clustering



Distances 100 attrib

Clustering 10,100, 1000 dim

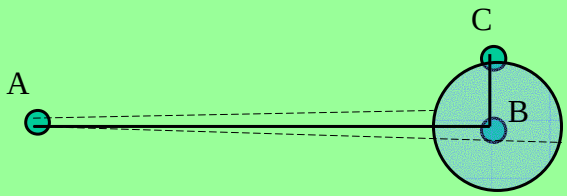
Distances 10 attrib.

Limits of standard clustering

- Hierarchical clustering is (very) good for visualization (first impression) and browsing
- Speed for modern data sets remains relatively slow (minutes or even hours)
- ArrayExpress database needs some faster analytical tools
- Hard to predict number of clusters (=>Unsupervised)

Approximate distances

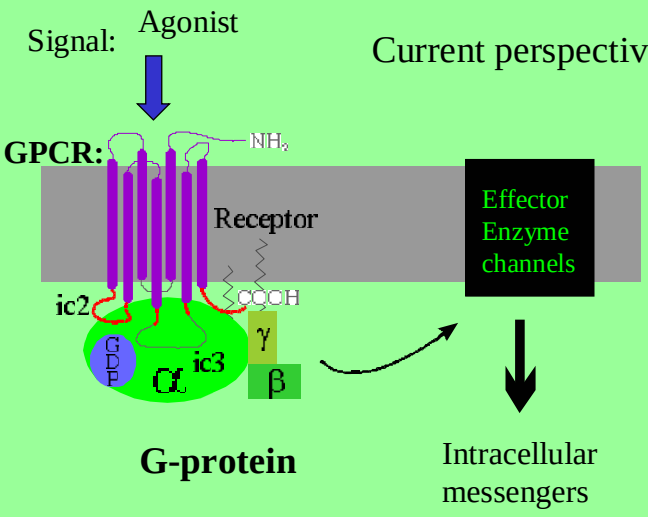
- Triangle inequality for metrics



d(A,B) and d(B,C) allow us to estimate d(A,C) within certain limits

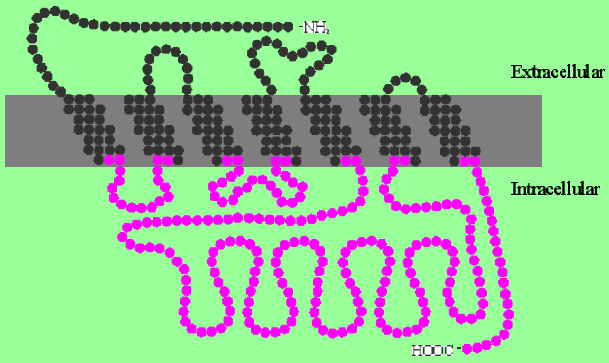
|d(A,B) – d(B,C)| <= d(A,C) <= d(A,B) + d(B,C)

GPCR coupling

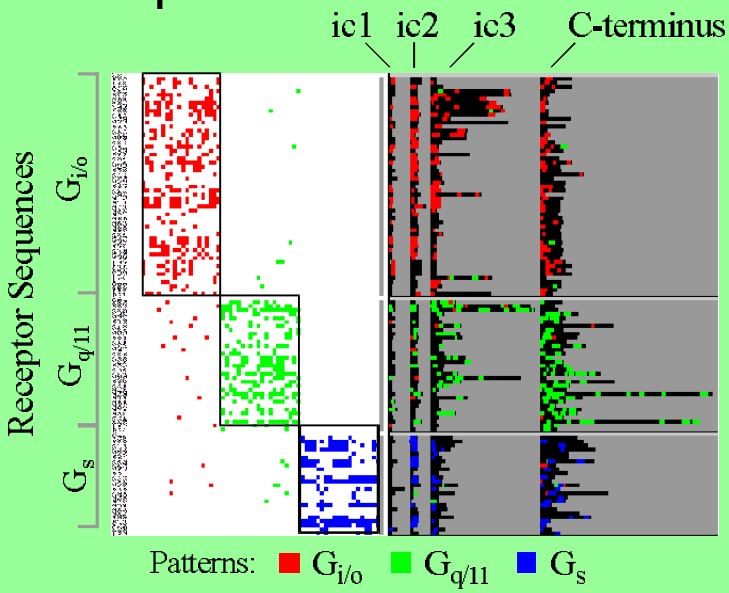


Our Computational Approach

- Using a new membrane topology prediction algorithm (designed specifically for GPCRs), we constrained our pattern search to the intracellular domains of ≈ 100 receptor sequences with well-characterised, and non-promiscuous coupling (split into G_s, $G_{i/o}$ and $G_{q/11}$)



Receptor Match Positions

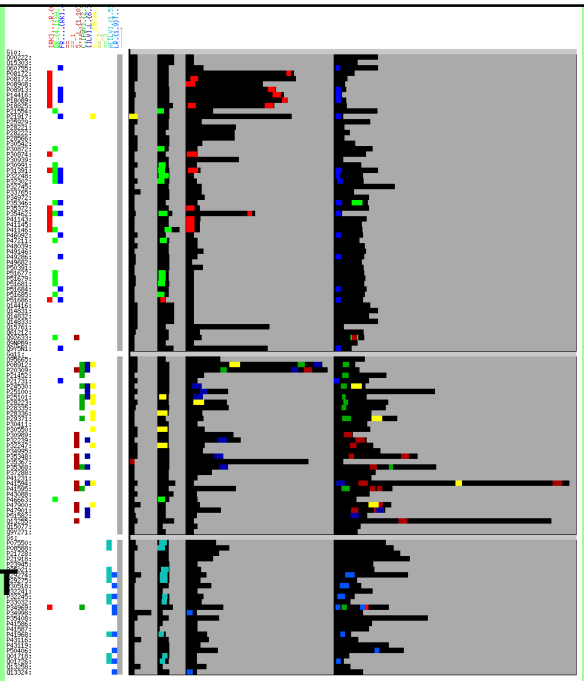


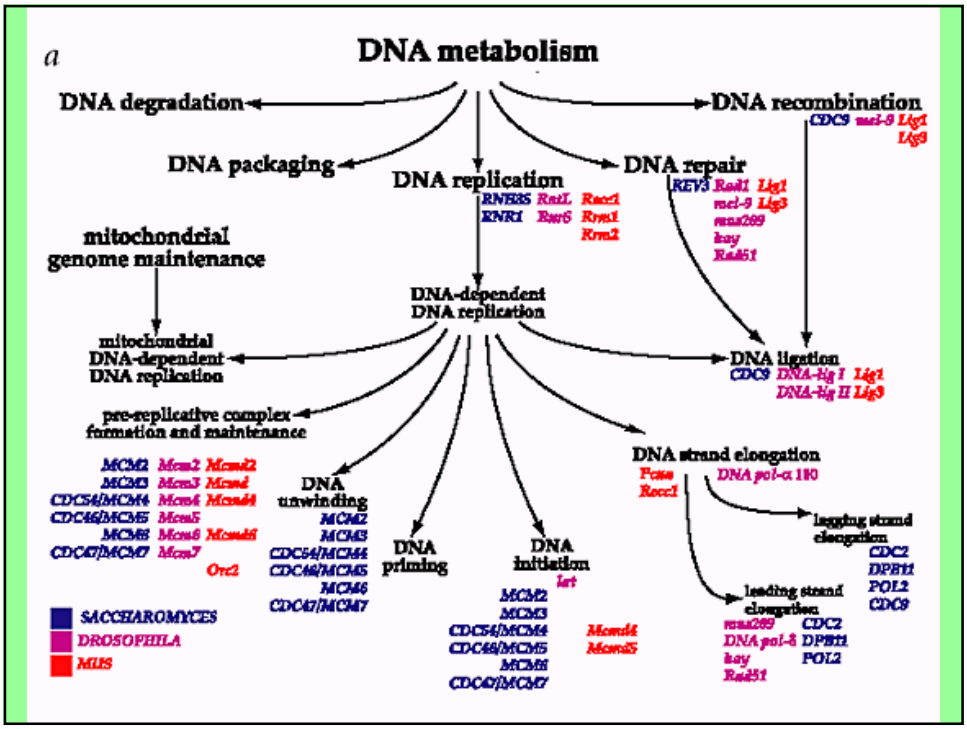
Croning, Vilo, Möller, *ISMB 2001*

[RK]....R.{0,9}EK
DR.{4,11}H...[AGS]
FR....[RK].{0,3}L

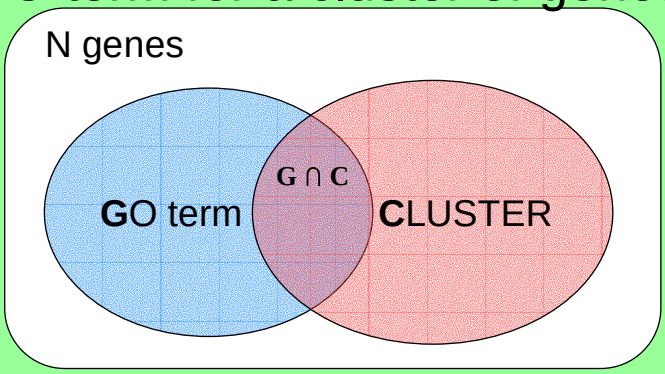
S...L.{1,10}T[ILV]
C.[FWY].{2,11}K
[ILV].L.{6,10}A.T
S....[RK]A.{3,10}S

A[ILV].{1,5}Y..[ILV].T
LR.{1,9}T...[ILV]





Determine the significance of GO term for a cluster of genes



A: |G ⋂ C| / min(|G|, |C|)

B: **P**(choose |C| from N with |G|, observe |G ⋂ C|+)

Annotation of clusters

GO:0042254 <U:L> **Process: ribosome biogenesis and assembly (+2:15) (depth=7) [sgd:2:187]**
GO:0042254: 47 from cluster (size 98) vs 187 in this class (including subclasses)

GO:0006364 <U:L> **Process: rRNA processing (+3:3) (depth=8) [sgd:50:126]**
GO:0006364: 35 from cluster (size 98) vs 126 in this class (including subclasses)

GO:0006360 <U:L> **Process: transcription from Pol I promoter (+6:14) (depth=8) [sgd:23:155]**
GO:0006360: 38 from cluster (size 98) vs 155 in this class (including subclasses)

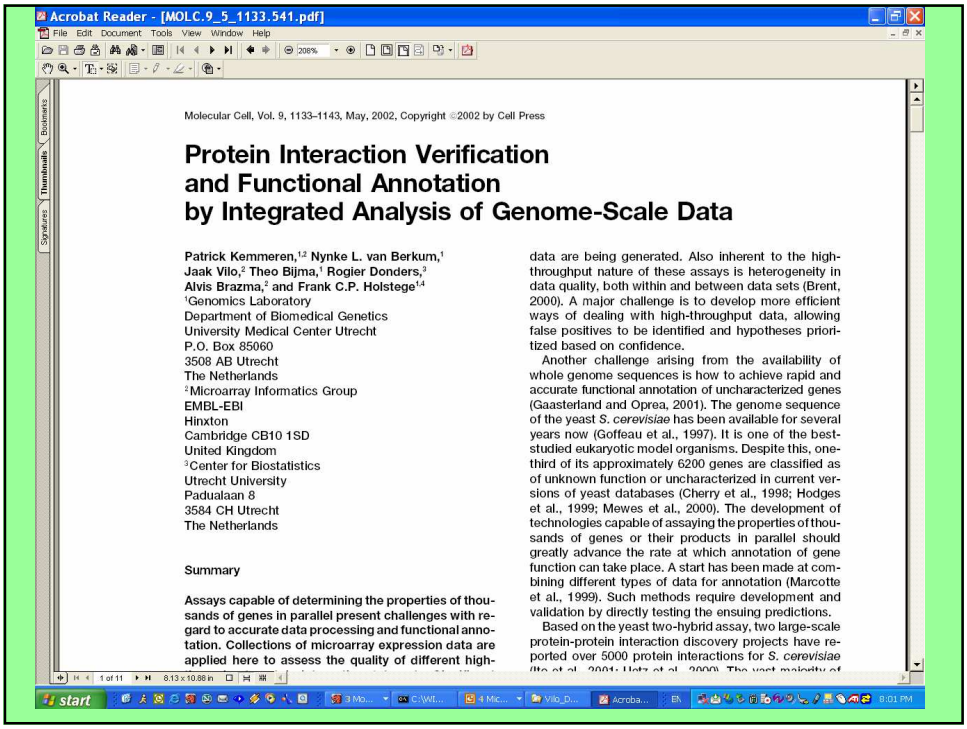
GO:0005730 <U:L> **Component: nucleolus (+10:17) (depth=6) [sgd:154:210]**
GO:0005730: 45 from cluster (size 98) vs 210 in this class (including subclasses)

GO:0030515 <U:L> **Function: snoRNA binding (depth=6) [sgd:23:23]**
GO:0030515: 17 from cluster (size 98) vs 23 in this class (including subclasses)

GO:0030490 <U:L> **Process: processing of 20S pre-rRNA (depth=9) [sgd:33:33]**
GO:0030490: 18 from cluster (size 98) vs 33 in this class (including subclasses)

GO:0005732 <U:L> **Component: small nucleolar ribonucleoprotein complex (depth=6) [sgd:30:30]**
GO:0005732: 16 from cluster (size 98) vs 30 in this class (including subclasses)

GO:0006396 <U:L> **Process: RNA processing (+7:52) (depth=7) [sgd:7:370]**
GO:0006396: 40 from cluster (size 98) vs 370 in this class (including subclasses)



Acrobat Reader - [MOLC.9_5_1133.541.pdf]

File Edit Document Tools View Window Help

Molecular Cell, Vol. 9, 1133–1143, May, 2002, Copyright ©2002 by Cell Press

Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data

Patrick Kemmeren,[1,2] Nynke L. van Berkum,[1] Jaak Vilo,[2] Theo Bijma,[1] Rogier Donders,[3] Alvis Brazma,[2] and Frank C.P. Holstege[1,4]
[1]Genomics Laboratory
Department of Biomedical Genetics
University Medical Center Utrecht
P.O. Box 85060
3508 AB Utrecht
The Netherlands
[2]Microarray Informatics Group
EMBL-EBI
Hinxton
Cambridge CB10 1SD
United Kingdom
[3]Center for Biostatistics
Utrecht University
Padualaan 8
3584 CH Utrecht
The Netherlands

Summary

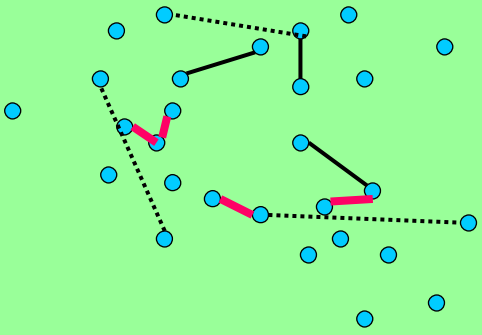
Assays capable of determining the properties of thousands of genes in parallel present challenges with regard to accurate data processing and functional annotation. Collections of microarray expression data are applied here to assess the quality of different high-

data are being generated. Also inherent to the high-throughput nature of these assays is heterogeneity in data quality, both within and between data sets (Brent, 2000). A major challenge is to develop more efficient ways of dealing with high-throughput data, allowing false positives to be identified and hypotheses prioritized based on confidence.

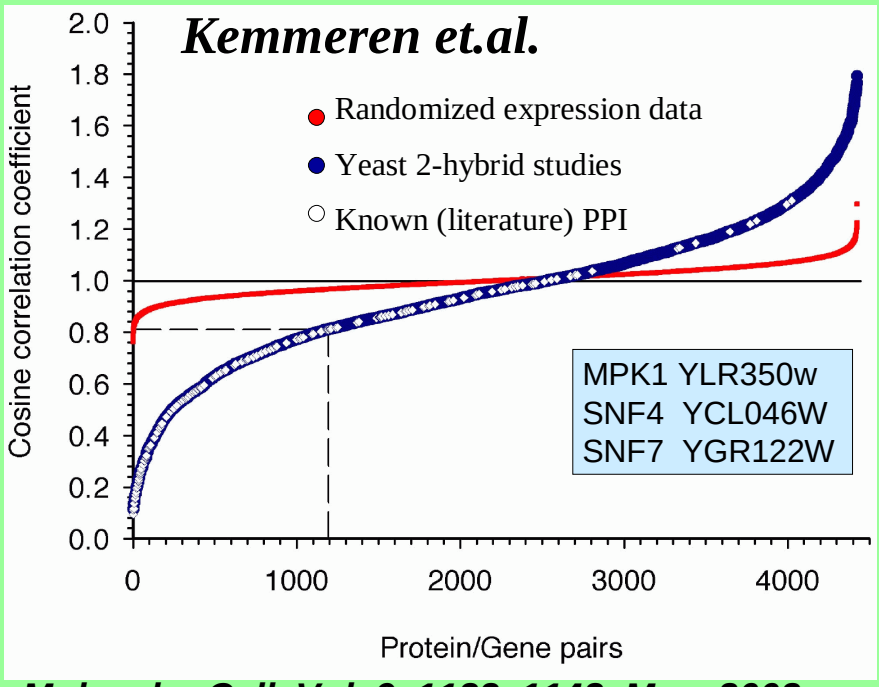
Another challenge arising from the availability of whole genome sequences is how to achieve rapid and accurate functional annotation of uncharacterized genes (Gaasterland and Oprea, 2001). The genome sequence of the yeast *S. cerevisiae* has been available for several years now (Goffeau et al., 1997). It is one of the best-studied eukaryotic model organisms. Despite this, one-third of its approximately 6200 genes are classified as of unknown function or uncharacterized in current versions of yeast databases (Cherry et al., 1998; Hodges et al., 1999; Mewes et al., 2000). The development of technologies capable of assaying the properties of thousands of genes or their products in parallel should greatly advance the rate at which annotation of gene function can take place. A start has been made at combining different types of data for annotation (Marcotte et al., 1999). Such methods require development and validation by directly testing the ensuing predictions.

Based on the yeast two-hybrid assay, two large-scale protein-protein interaction discovery projects have reported over 5000 protein interactions for *S. cerevisiae*

Protein-protein interactions: which to trust more?



Answer: Use the distance measure alone



Molecular Cell, Vol. 9, 1133–1143, May, 2002

Results from PPI & expression

- **Confidence in 973 out of 5342 putative two-hybrid interactions from *S. cerevisiae* is increased.**
- **Besides verification, integration of expression and interaction data is employed to provide functional annotation for over 300 previously uncharacterized genes.**
- **The robustness of these approaches is demonstrated by experiments that test the in silico predictions made.**
- **This study shows how integration improves the utility of different types of functional genomic data and how well this contributes to functional annotation.**

Start summarizing…