Fisica teorica e biologia computazionale.

M. Caselle

Napoli, marzo 2005

Indice

- Introduzione alla biologia computazionale:
  - Principali filoni di ricerca
  - Risorse di tipo didattico e preprint databases
  - Congressi
  - FB11
  - Il gruppo di Torino

- Esempio: Identificazione di sequenze regolatrici in lievito ed in uomo.
  - Genomica comparativa: allineamento di sequenze
  - Dati di espressione genica: microarrays
  - Annotazioni funzionali: Gene–Ontology

Filoni di ricerca principali

- Protein folding
- Interazione Proteina–Proteina / Proteina–DNA
- Dinamica del DNA e dell'RNA
- Annotazione del genoma
- Elaborazione di algoritmi (clustering, sequence alignement)
- Studio di problemi legati alla mobilita' cellulare
- ....

Risorse

Didattica

Master

- ”Bioinformatica: Applicazioni Biomediche e Farmaceutiche”, Università di Roma La Sapienza. http://cassandra.bio.uniroma1.it/Master
- ”Bioinformatica”, Università di Torino http://www.masterbioinformatica.it
- ”Bioinformatica”, Università di Milano Bicocca http://www.btbs.unimib.it/ master/bioinformatica2003.htm

Dottorati

”Sistemi complessi applicati alla biologia post-genomica”, Università di Torino

http://www.bioinformatica.unito.it /complex_systems/welcome.html

Preprints

Alla fine del 2003 e' nato (a fianco di hep-th,hep-ph ecc.) un nuovo archivio di "quantitative biology" che si chiama q-bio.

Il link e' http://xxx.lanl.gov/archive/q-bio

Congressi

- Intelligent Systems for Molecular Biology

  ISMB 2004, Glasgow 31 luglio - 4 agosto

  http://www.iscb.org/ismbeccb2004/

  ISMB 2005, Detroit 25-29 giugno

  http://www.iscb.org/ismb2005/

  ECCB 2005, Madrid 27-30 settembre

  http:/www.eccb05.org/

- Research in Computational Biology

  RECOMB 2004, S. Diego 27-31 marzo

  CBrohttp://recomb04.sdsc.edu/

  RECOMB 2005, Boston marzo 2005

  Topics:

  - Genomics
  - Molecular sequence analysis
  - Recognition of genes and regulatory elements
  - Molecular evolution
  - Protein structure
  - Structural genomics
  - Gene Expression
  - Gene Networks
  - Drug Design
  - Combinatorial libraries
  - Computational proteomics
  - Structural and functional genomics

FB11: Applicazione di metodi della fisica teorica a sistemi biologici

Sezioni coinvolte e partecipanti

Sezione	resp. locale	partecipanti
BA	S. Stramaglia	7
BO	A. Bazzani	5
CT	A. Rapisarda	7
FI	S. Bagnoli	11
MI	C. Destri	12
NA	L. Peliti	6
PD	A. Stella	6
Pr	R. Burioni	4
RM2	S. Morante	3
SA	S. Scarpetta	2
TO	M. Caselle	5
	Totale	68

Il gruppo di Torino

- M. Caselle, Dip. di Fisica Teorica
- F. Di Cunto, Dip. di Biologia Molecolare
- I. Pesando, Dip. di Fisica Teorica
- P. Provero, Fondazione per le Biotecnologie
- D. Cora' (Dottorato: Sistemi complessi ...)
- E. Curiotto (Dottorato: Sistemi complessi ...)
- L. Martignetti (Dottorato: Sistemi complessi ...)
- I. Molineris (Dottorato: Sistemi complessi ...)
- A. Re (Dottorato: Sistemi complessi ...)
- G. Sales (Laureando)

Collaborazioni con

- C. Dieterich Max Planck Inst. for Molecular Genetics, Berlin
- C. Herrmann Laboratoire de Genetique et de Physiologie du Developpement (LGPD) Marseille
- I. Sbrana Dipartimento di Biologia dell'Universita' di Pisa

Linee di ricerca

1] Studio della regolazione genica. In particolare:

- identificazione di nuovi fattori di trascrizione in lievito usando Gene Ontology, Microarray e correlazioni.
- uso di metodi di Genomica Comparativa (in particolare il confronto tra topo ed uomo) per l'identificazione di nuovi regolatori nell'uomo.

2] Ricerca di UTR in uomo mediante Genomica comparativa e metodi statistici (catene di Markov)

3] Uso di tecniche di teoria dei grafi per studiare networks di coespressionee di coregolazione.

4] Studio di siti fragili nei cromosomi umani.

5] Studio dell'interazione tra DNA e fattori di trascrizione mediante simulazioni di Dinamica Molecolare.

References

- M. C., F. Di Cunto, P. Provero

  “Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes.”

  BMC Bioinformatics 2002, 3:7, physics/0203013

- M. C., D. Corá, L. Silengo, F. Di Cunto, P. Provero

  “Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs” BMC Bioinformatics 2004, 5:57, q-bio.GN/0310040

- M. C., F. Di Cunto, P. Provero

  ”A computational approach to regulatory element discovery in eukaryotes”

  Proceedings of the 2002 ECMTB conference, cond-mat/0305279

- M.C., F. Di Cunto, M. Pellegrino and P.Provero

  "Finding regulatory sites from statistical analysis of nucleotide frequencies in the upstream region of eukaryotic genes"

  Proceedings of the International Workshop "Modelling Bio-medical signals", Bari, September 2001, physics/0201033

1. Introduction

Genome Structure

- The density of protein-coding and RNA-coding sequences becomes lower and lower as the complexity of the organism increases. It is rather high in Prokaryotes, low in S. Cerevisiae, very low in the human genome: most of DNA in the human genome is not coding ($\sim 99\%$)

- The biological role of non-coding part of DNA is poorly understood. The common lore is that it should be involved in the regulation of gene expression

Gene regulation

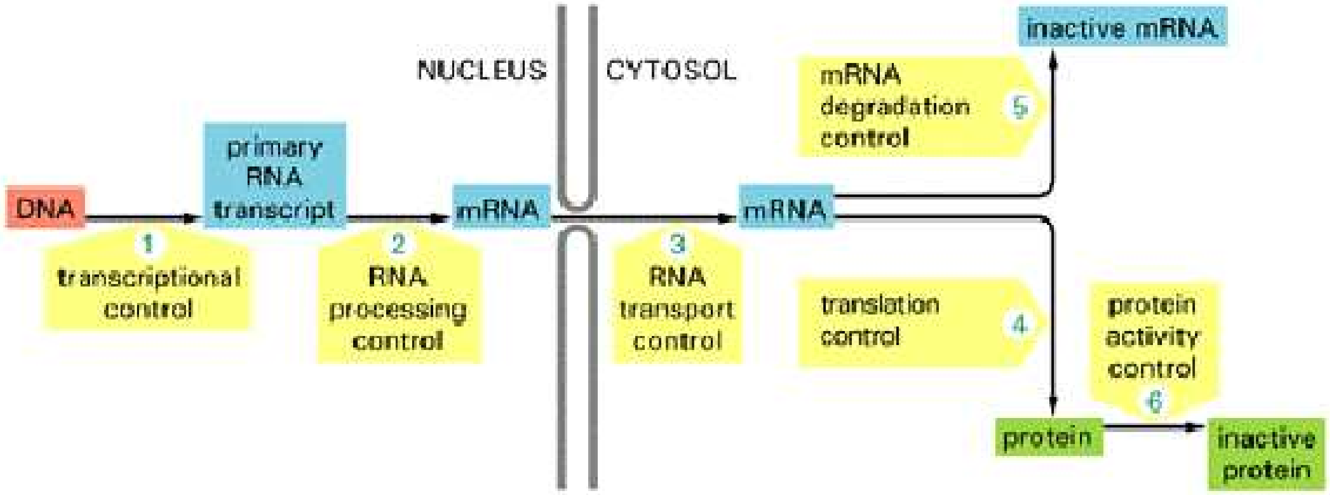
Gene expression is tightly controlled and regulated:

- All cells in the body carry the full set of genes, but only express about 20% of them at any particular time

- Different proteins are expressed in different cells (neurons, muscle cells....) according to the different functions of the cell.

As more and more complete genomes are decoded it is becoming of crucial importance to understand how the gene expression is regulated.

The challenge is now to identify and fully characterize the network of interactions among genes and their products in an organism.



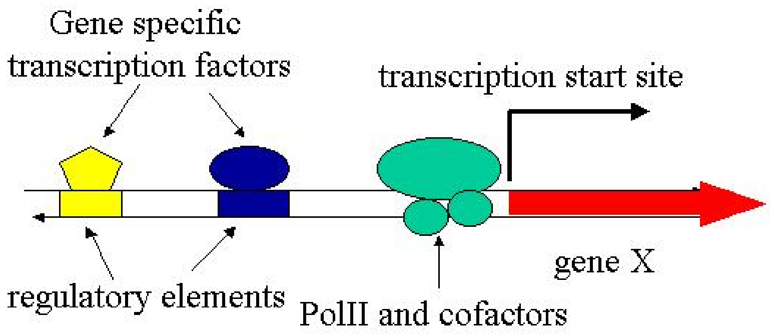
The most important example of such interactions is the transcriptional regulation of protein coding genes. Even if this is not the only regulatory mechanism of gene expression in eukaryotes it is certainly the most widespread one.

The goal of our research project (as of many others in the world) is to reconstruct these interactions by comparing existing biological information (like the coregulation of sets of genes) with the statistical properties of the sequence data.

Transcription factors.

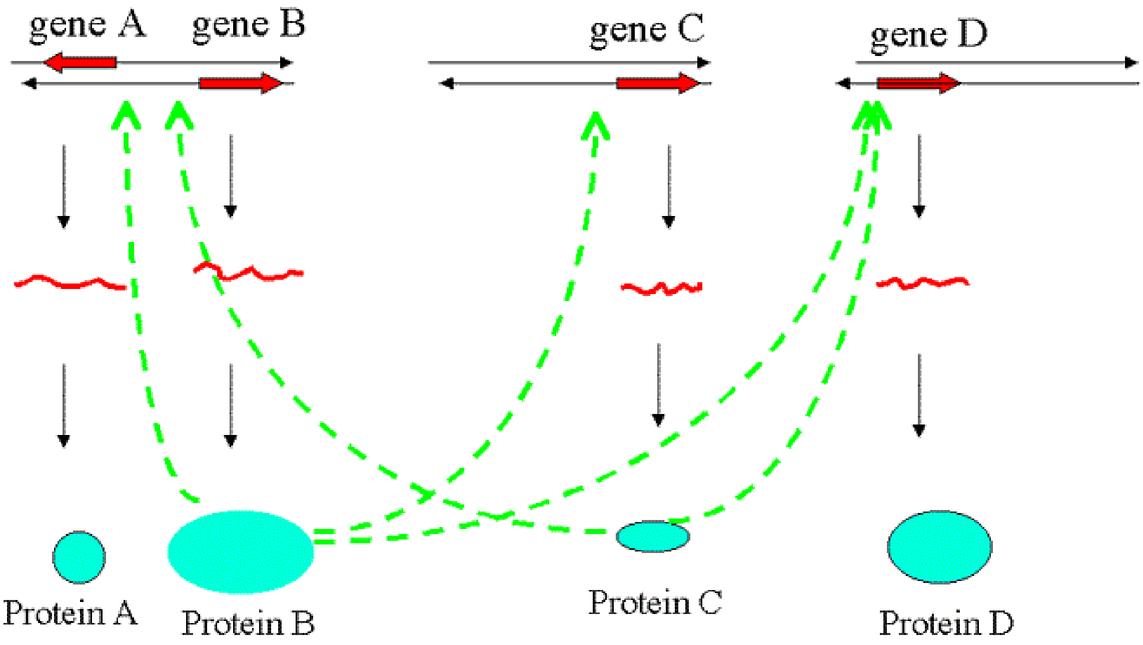
TFs act by binding to specific, often short (5-10 bp) DNA sequences in the upstream noncoding region of genes.

Transcriptional regulation



Regulatory network

T.F.'s themselves are proteins produced by other genes.



The Genome is a complex network of interactions between genes and their products This network pattern is ubiquitous in Postgenomic biology

The problem.

However, computational detection of regulatory sites is a difficult task, specially in eukaryotes:

- the consensus sequences recognized by transcriptional factors are generally rather short (5-20 bp)
- they can be quite variable
- they are in general dispersed over large distances
- they are generally active in both orientations

A simple study of relative frequencies of sequences can be meaningless

2. Our Strategy.

We have a few tools to attack the problem:

- Binding sites are often overrepresented. One can use this to separate the signal (binding site) from the noise (background upstream sequence)

- Binding sites are often evolutionary conserved. One can use comparative genomics to recognize them.

- Genes which share the same functions may also share the same regulatory mechanisms. One may use microarray experiments or functional annotations to identify binding sites.

Overrepresented words in the upstream regions

Many binding sites are effective only when repeated many times in the upstream region of the gene they regulate.

Example: the word GATAAG—CTTATC is a known binding factor for nitrogen-regulated genes: Examine the 500 bp's upstream of two of them.

>YPR138C upstream sequence, from -500 to -1

```
TCCACCTTATCTCGGCGCCAAATCCTTATC
TCTCGTAGCTGGTTTGCCCGCGATAAGGCG
GGCGAGTTATTTTGAAGTTTTCCATAAACT
GGTTTTCCATCTCGAGGTTTTTCCTCGCTT
TCCACGCTATGACCCTTTTTAGTTAAGGTA
CCCGATGGCATACTTTATATATTATATATA
TATGTTAAGTTAATATGTTTTAGCAGATTT
GATATGCTGATATGCAGCACGGACTTTCCC
TCTCCTTGTCTTATCGCATCTTATCGCAAC
AATTTGATAGATATCTTCTCCCTTTCCTAT
CTTGTAGAATAAGGTTGTGTGCTTTGAGTC
TGATAGCCGTCTTCTTTCGGTCGCTTCTTC
TCTCTTTTGGTTCTTTGATTGTCTATTACA
ATCAATGCAGGCTAGTTAAGGGTCCAATCA
CTTTTGAAATTGTTTTGTAAAAAGCGAAGG
CATTTTTTTTTTAGAAGATACAATTGAAAA
CATATAGATTTAGAGTTCAC
```

>YIR028W upstream sequence, from -500 to -1

```
ATTCTCGGGTCTAATGTGGCTCGAGGGTAT
CTCTTATCGGTATTACTTTCTTATCAATGA
AAAATTTCTGCCAGGGAAAATGCGCCCGCT
TTTTTTCCGGCCATCCTTACTCGCTGTCGC
ATACAAAATAGCGCCTCTAATCTAGTTGCG
ATAAGGAATGTGTATGTGTAATTGAAGATC
CAGGATGTTTTCCTTTTCAGGGAGATGAGA
AGGAATAATAGGATGGATTGACCGCTTTGC
TGTCACGTCGATAAGGTTCCTTTAAAAATT
GTGTCCAATGATTAGCATAGAGAGGTAGAG
TATCAGAGAAACAAGTTTGTAATCGAGAAA
CTTGATCTGCTAGTGTTGAGCATAGAAGGC
TAGGAAAACATGGGGAAGAAAAAAAAAGTA
TAAATAATTAGCTTGATGAGTAGTTTGAAT
ATATATGTTACTTTAGTTTCCCTTTTTGAC
CTTTTATATTCATCTACATCTTGTGATATA
AAACATCAACAAAGACGAGA
```

Our Proposal

first step Grouping of genes based on the motifs that are overrepresented in their upstream regions. To each possible word w we associate the set S_w of all the genes in whose upstream region the word w is overrepresented

second step Select those sets which show some kind of functional characterization using microarray experiments or Gene Ontology annotations.

- Microarray: For each set S_w we compare the expression distribution within the set with the genome wide one (using for example Kolmogorv-Smirnov test).

- Gene Ontology: For each set S_w we compute the prevalence of all GO terms among the annotated genes in the set, and the probability that such prevalence would occur in a randomly chosen set of the same size:
  - hypergeometric distribution to assess the significance of the intersection
  - evaluation of false discovery rate through comparison with randomly generated gene sets (using only the best p-value for each set as criterion for the comparison)

The words which survive this analysis are candidates to be binding sites.

The Gene Ontology Consortium "Gene Ontology: tool for the unification of Biology" Nature Genetics **25** (2000) 25.

The sets $S(\text{word})$

- For each word (5 to 8 bp's) compute the frequency in the upstream sequences of the whole genome considered as a single sample: these will be our reference frequencies.

- Then count occurrences of the word in the upstream region of each gene separately.

- If the number of occurencies of the word in the upstream region of gene G is statistically significant (compared to a binomial distribution based on the above reference frequencies), then the gene G belongs to the set $S(\text{word})$.

Choices in our study on yeast:

- *upstream sequences length: 500 bp*

- *probability cutoff* $P = 0.01$

The Gene–Ontology filter.

For each set $S(m)$ we computed the prevalence of all Gene Ontology (GO) terms among the annotated genes in the set, and the probability that such prevalence would occur in a randomly chosen set of genes of the same size.

For a given GO term t let $K(t)$ be the total number of ORFs annotated to it in the genome, and $k(m,t)$ the number of ORFs annotated to it in the set $S(m)$. If J and $j(m)$ denote the number of ORFs in the genome and in $S(m)$ respectively, such probability is given by the right tail of the appropriate hypergeometric distribution:

$$P(J,K(t),j(m),k(m,t)) = \sum_{h=k(m,t)}^{\min(j(m),K(t))} F(J,K(t),j(m),h)$$

where

$$F(M,m,N,n) = \frac{\binom{m}{n}\binom{M-m}{N-n}}{\binom{M}{N}}$$

In this way a P-value can be associated to each pair made of a motif and a Gene Ontology term.

False discovery rate

Problem:

Given the huge number of P-values that we compute (in principle equal to the number of GO terms multiplied by the number of words analysed) it is clear that very low P-values could appear simply by chance.

The usual way of dealing with this issue, that is the Bonferroni correction, is not appropriate, because due to the hierarchical nature of the Gene Ontology annotation scheme, the P-values we compute are very far from being independent from each other.

Our proposal

We randomly generated a large number N_R of sets of genes comparable in size to the typical size of the sets associated to the motifs and ranked the random sets based on their **best** P-values.

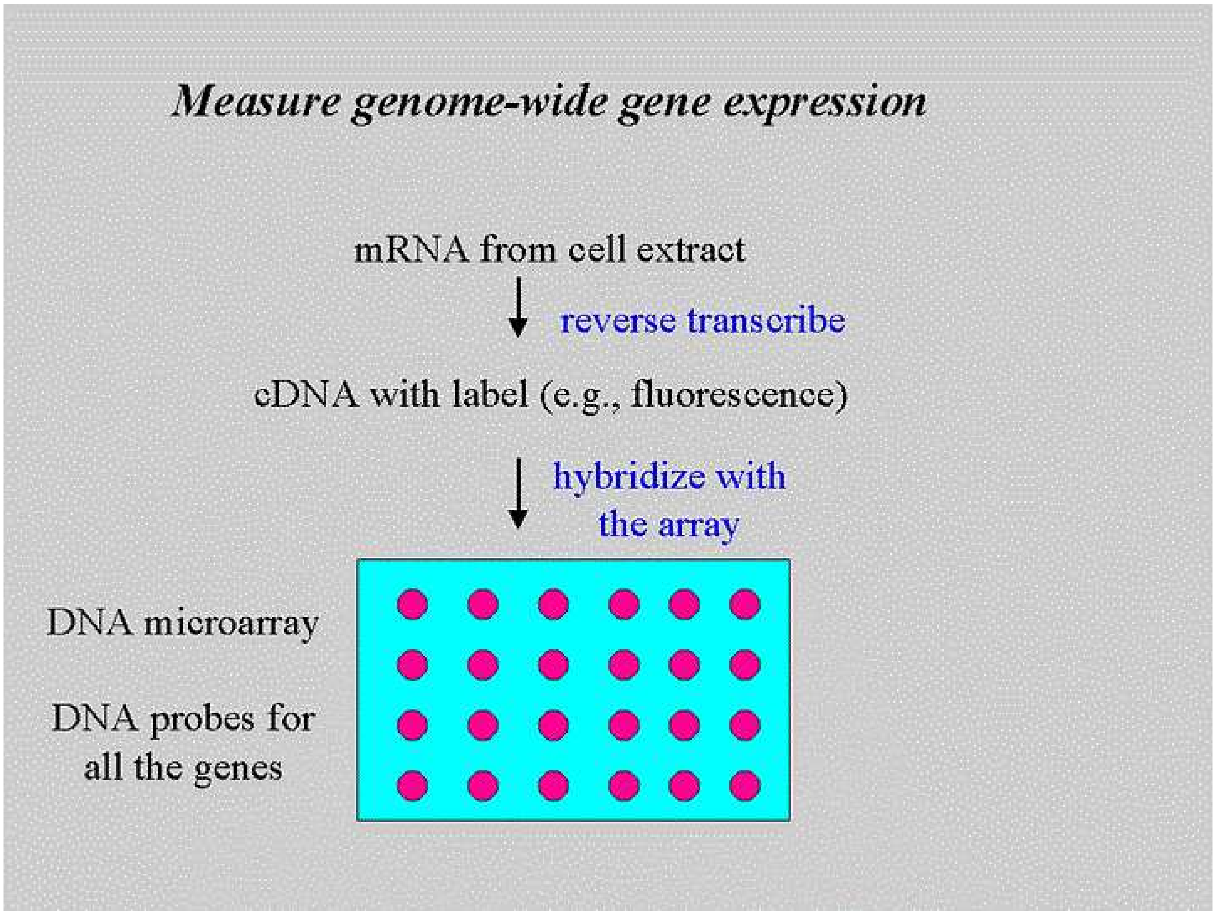
In this way we can determine a false discovery probability $p_f(C)$ as a function of the cutoff on P-values C

Warning:

The lower is the FDR required, the higher is the precision required in determining the function $p_f(C)$ and hence the number N_R of sets to be generated randomly. For instance a FDR of 0.01 requires the generation of 3.5×10^6 randomly chosen sets.

The microarray filter

DNA microarrays can estimate genome-wide gene expression levels by measuring the amount of mRNA levels in the cell. Thousands of genes can be simultaenously studied in a single microchip.



Measure genome-wide gene expression

mRNA from cell extract

reverse transcribe

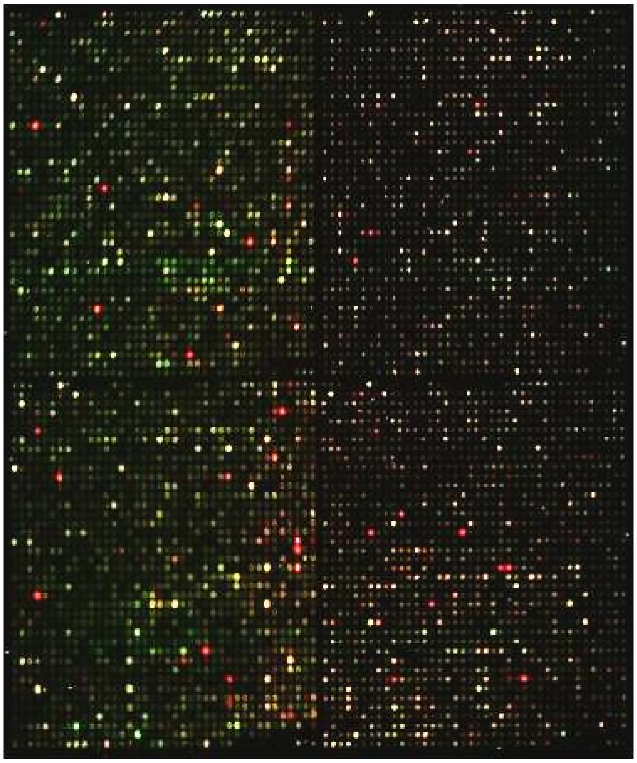
cDNA with label (e.g., fluorescence)

hybridize with the array

DNA microarray

DNA probes for all the genes

The result of the experiment is a slide of this type:



The fluorescence level is proportional to the amount of mRNA produced in the experimental condition under study (usually one studies the ratio with respect to the expression level in some “reference” state of the cell).

Example : Microarray samples in S. Cerevisiae

The diauxic shift

DeRisi et al., Science 278 (1997) 680

- a yeast culture is inoculated into a glucose-rich medium
- rapid anaerobic growth fueled by fermentation, with production of ethanol, insues
- upon glucose depletion, the yest cells turn to ethanol as a carbon source for aerobic growth (respiration)

Expression data from DNA microarrays

- samples of cells are harvested at seven time-points during the diauxic shift

- using DNA microarray techniques mRNA levels for all the genes can be measured and compared to their initial values

- therefore the experiment answers the question: which genes are switched on, and which are switched off, as the available glucose becomes progressively scarcer?

The output of the experiment is, for each gene, the ratio between initial expression level and expression level at each of the seven timepoints during the diauxic shift.

The idea is to look for statistical correlation between these numerical data and the presence of binding sites in the upstream region of each gene.

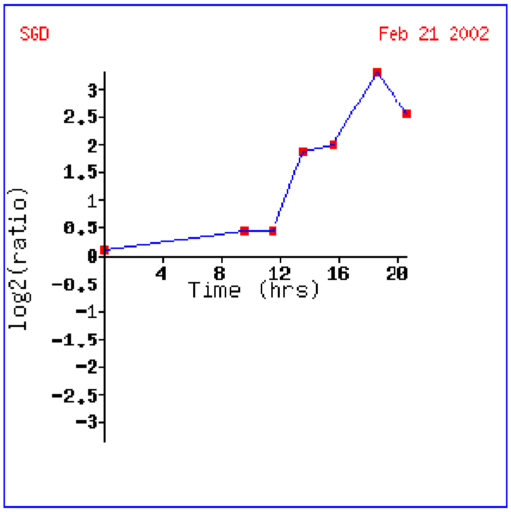
Studying expression level for each set

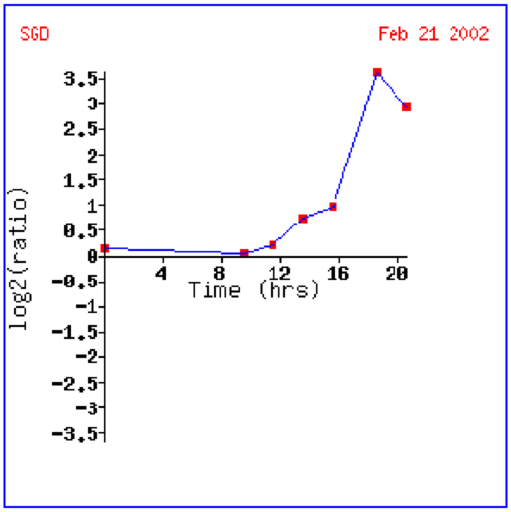
For each set $S(\text{word})$ we compute the average expression level of the genes in the set at the seven timepoints of the diauxic shift experiment.

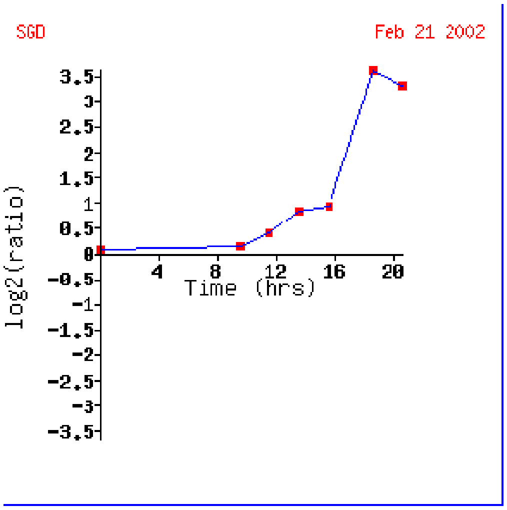
More precisely, the average $\log_2$ of the ratio between measured mRNA at each timepoint and measured initial mRNA.

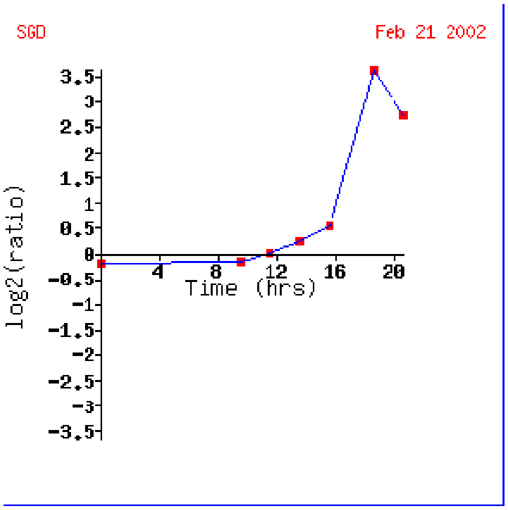
This value is then compared to the average expression taken over the whole genome at each timepoint.

If the difference is larger than six standard deviations the word defining the set is a candidate binding site for the regulations of the genes in the set.


SGD
Feb 21 2002
log2(ratio)
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
4
8
12
16
20
Time (hrs)


SGD
Feb 21 2002
log2(ratio)
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
4
8
12
16
20
Time (hrs)


SGD
Feb 21 2002
log2(ratio)
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
4
8
12
16
20
Time (hrs)


SGD
Feb 21 2002
log2(ratio)
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
4
8
12
16
20
Time (hrs)

3. Example: Yeast

Identification of TF binding sites in yeast using Gene–Ontology

Output of the analysis:

- With the false discovery rate set at 0.01 we find a total of 108 associations between 80 different words (of 5-8 letters) and 41 Gene Ontology terms.

- The words can be organized in 12 different groups. Within each group the motifs are very similar to each other and are associated to the same or to very similar Gene Ontology terms. For each group we construct a consensus sequence ("motifs") by aligning the words.

motif	C	F	P
AGGGTGC	-	-	siderophore transport
AGGGTGCA	-	-	siderophore transport
TGGGTGCA	-	-	siderophore transport
GGGTGCA	-	-	siderophore transport
GGGTGC	-	-	siderophore transport
GGTGCA	-	heavy metal ion porter	siderophore transport
GGTGC	cell wall (sensu Fungi)	- -	- -
AGGGTGCACC			
CGGCGCC	-	-	tricarboxylic acid cycle
CGGCGCCG	-	-	tricarboxylic acid cycle
GGCGCCGA	-	-	tricarboxylic acid cycle
GCGCCGAG	-	-	tricarboxylic acid cycle
CGGCGCCGAG			

Table 1: *Two examples of motifs.*

Validation:

- Comparison with known TF's and binding sites (Transfac + literature survey)

- Comparison with the genome wide ChIP experiment of: T.I. Lee et al., Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science 298, (2002) 799.*

motif	C	F	P	TF
TGAAAC	-	-	sexual reproduction	DIG1 STE12
TGAAACA	-	-	sexual reproduction	DIG1 STE12
TGAAACA				
ACTGTG	-	-	sulfur amino acid transport	MET4
TGTGGC	-	-	sulfur metabolism	MET4 MET31
ACTGTGGC				

Table 2: *Two examples of motifs with significant intersection with ChIP data*

Results:

- All the motifs we find correspond to known binding sites. (No false positive!)

- For some of the motifs we are able to

  - refine the putative binding sequences.
  - identify candidates for combinatorial regulation (example: PAC and RRPE))
  - Refine the functional annotation of already known TF's
  - identify new potential targets of known TF's (example: Hcm1p)

MPS1	(YDL028C)
CIN8	(YEL061C)
PDS1	(YDR113C)
SPC98	(YNL126W)
VIK1	(YPL253C)
SPC25	(YER018C)
ESP1	(YGR098C)
STU2	(YLR045C)
SLI15	(YBR156C)

Table 3: *Candidate targets of regulation by the Hcm1p transcription factor*

motif	C	F	P
GATGAGA	nucleolus	-	ribosome biogenesis
GATGAGAT	nucleolus	-	ribosome biogenesis
ATGAGAT	nucleolus	-	ribosome biogenesis
ATGAGATG	-	-	ribosome biogenesis
TGAGATG	-	-	ribosome biogenesis and assembly
TGAGATGA	-	-	ribosome biogenesis and assembly
GAGATG	-	-	ribosome biogenesis and assembly
GAGATGAG	nucleolus	-	ribosome biogenesis and assembly
GAGATGA	nucleolus	-	ribosome biogenesis and assembly
AGATGAG	nucleolus	-	ribosome biogenesis
GATGAG	nucleolus	-	ribosome biogenesis
GATGA	-	-	ribosome biogenesis
ATGAGCT	nucleolus	-	ribosome biogenesis
TGAGCT	nucleolus	-	rRNA processing
GATGAGATGAGCT			
AAAAATT	nucleolus	-	ribosome biogenesis
AAAAATTT	nucleolus complex	-	transcription from Pol I promoter
AAAATT	nucleolus	-	ribosome biogenesis
AAAATTT	nucleolus	-	ribosome biogenesis
AAAATTTT	nucleolus	-	ribosome biogenesis
AAATT	nucleolus	-	35S primary transcript processing
AAATTTTC	small nucleolar ribonucleoprotein complex	-	35S primary transcript processing
AAAAATTTTC			

4. Binding site identification in human.

The extension of our algorithm to the human genome is not straightforward. At least 15.000 bp long upstream regions must be taken into account leading to a very small signal to noise ratio.

It is mandatory to perform a comparative analysis selecting only those parts of the upstream regions which are conserved between men and mouse.

This can be done using the CORG database:

C. Dieterich et al., CORG: a database for comparative regulatory genomics. *Nucleic Acid Res.*, **31**, *(2003) 374.*

The CORG database.

CORG is a collection of conserved sequence blocks in the non-coding, upstream regions of orthologous genes from man and mouse.

These blocks are obtained by searching statistically significant local suboptimal alignments of 15kb regions upstream of the translation start site.

The database contains more than 10,000 pairs of orthologous genes. The alignments were obtained using the Waterman-Eggert algorithm. We used two different choices of the PAM matrix: PAM1 and PAM10 to test the robustness of the results.

An important role in the following analysis is played by the fact that more than half of the genes in the database are annotated in the GO database.

The two releases are very different:

- PAM1
  - number of genes in the database: 10999
  - mean number of conserved blocks for gene: ~ 20
  - mean length of the union of conserved blocks: ~ 500
  - number of genes with a GO annotation 6187
- PAM10
  - number of genes in the database: 12943
  - mean number of conserved blocks for gene: ~ 40
  - mean length of the union of conserved blocks: ~ 900
  - number of genes with a GO annotation 7260

Results.

In the PAM10 case, out of the 43250 possible words of 5,6,7 and 8 letters

- 154 different words survive the G–O filter
- 331 words survive the Microarray filter
- the intersection between the two sets is 109 words which corresponds to a p–value e^{-201}
- similar results are obtained with PAM1. Despite the fact that the PAM1 and PAM10 CORG databases are very different our results seems to be very robust: most of the words are present in both releases.

Clustering of words.

Due to the larger amount of words and to the higher motif's variability, clustering of words is more delicate than in the yeast case. To decide if two words belong to the same motif we make a two steps analysis.

- First step: we check if at least one of the following conditions is met:
  - at least one GO term is significant for both motifs
  - there is at least one time point in the cell cycle MA experiment in which both motifs are simultaneously significant.
  - the intersection of the two sets of genes (labeled by the two words that we are testing) is statistically significant.
- Second step: we check if an alignment can be found between the two words with no gaps, at least 4 bases correctly aligned and at most 1 mismatch.

Validation.

Comparing our finding with the data collected in the Transfac database we were able to recognize some well known TF's.

Example: NF–kB

motif	C	F	P
GGAAATTC	-	chemoattractant	-
GGRAAKTCCC	Transfac consensus		

Table 4: *The putative NF–kB motif.*

Example: E2F

motif	C	F	P
TTTCGCGC	-	-	DNA replication initiation
TTTSGCGC	Transfac consensus		

Table 5: *The putative E2F motif.*

Example: A putative new motif

motif	C	F	P
AATGTTG	Golgi lumen	-	-
TGTTGA	Golgi lumen	-	-
ATGTTGA	Golgi lumen	-	-
TTATGTA	Golgi lumen	-	-
TWATGTTGA			

Table 6: *A putative motif with no reference in Transfac.*

Conclusions.

We propose a new method to extract relevant biological information on the Transcription Factors (and more generally on the mutual interactions among genes) from the statistical distribution of oligonucleotides in the upstream region of the genes.

- The method requires a complete knowledge of the upstream oligonucleotide sequences and thus it can be applied for the moment only to those organisms for which the complete genome has been sequenced.
- It does not require any external bias. The significance criterion only depends on the statistical distribution of oligonucleotides in the upstream region
- It can be easily implemented and could be used as a standard preliminary test, to guide more refined analysis

- It makes use of G–O annotations and/or Microarray data to assess the significance of the results. Both these tools are becoming more and more precise. This should lead to improved performances of future releases of our analysis.

We studied its performances in two cases: yeast and human. In both cases we found some already known TFs which we used as a validation test of the method. In the human case we also found some previuosly unknown candidates binding sites, which we expect to be of biological relevance.