Metodi di fisica teorica applicati alla biologia computazionale

Davide Corà

Università degli studi di Torino – INFN

Torino, 27 Febbraio 2006

Outline

- **Introduzione alla biologia computazionale**
  - Filoni di ricerca (cosa è la "genomica" ?)
  - Risorse / riviste / congressi
- **Il gruppo TO61 di Torino**
  - Il problema della regolazione genica
  - Ricerca di elementi regolativi in human
    - approccio statistico
    - genomica comparativa
    - dati di espressione
    - dati di annotazione funzionale (Gene Ontology)
  - Altri lavori in corso …..

La biologia computazionale (1)

Sotto il nome di “biologia computazionale” è incluso tutto ciò che riguarda l’analisi di dati biologici con metodi provenienti dalla matematica / fisica / statistica / computer-science.

Alcune linee più “classiche”, ad es:

- **biofisica**
- **reti neurali**
- **analisi di sistemi dinamici**

La biologia computazionale (2)

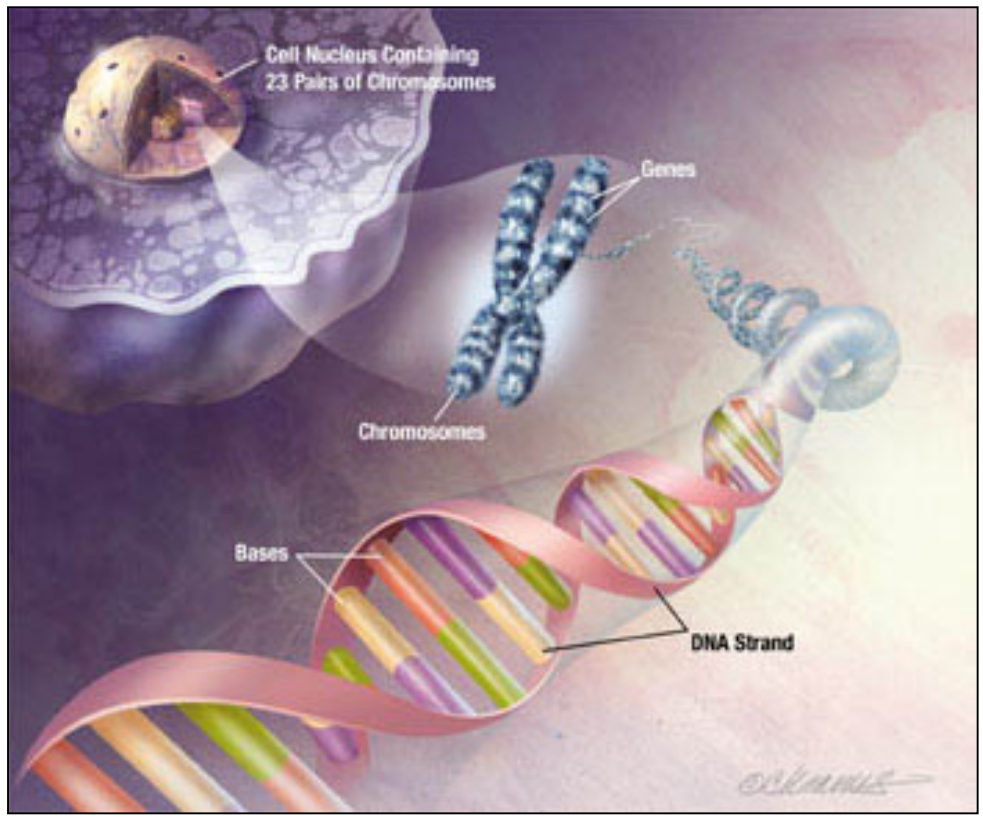
Alla fine degli anni ’90 → nasce “era genomica”

La biologia diventa sempre più quantitativa:

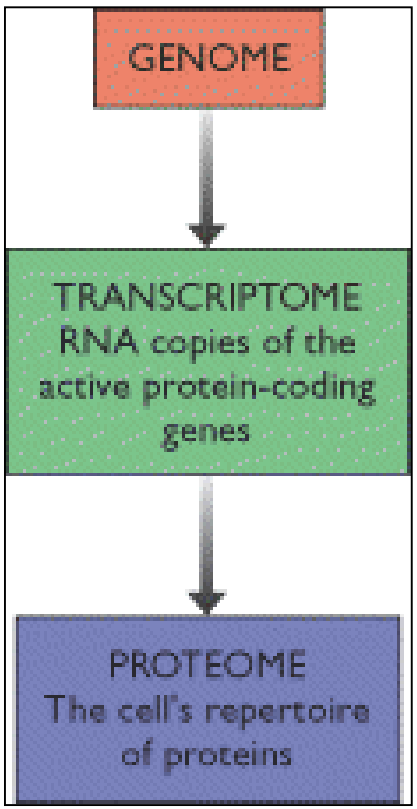
- **sequenziamento dei genomi di interi organismi**
- **microarray**
- **dati proteomici**

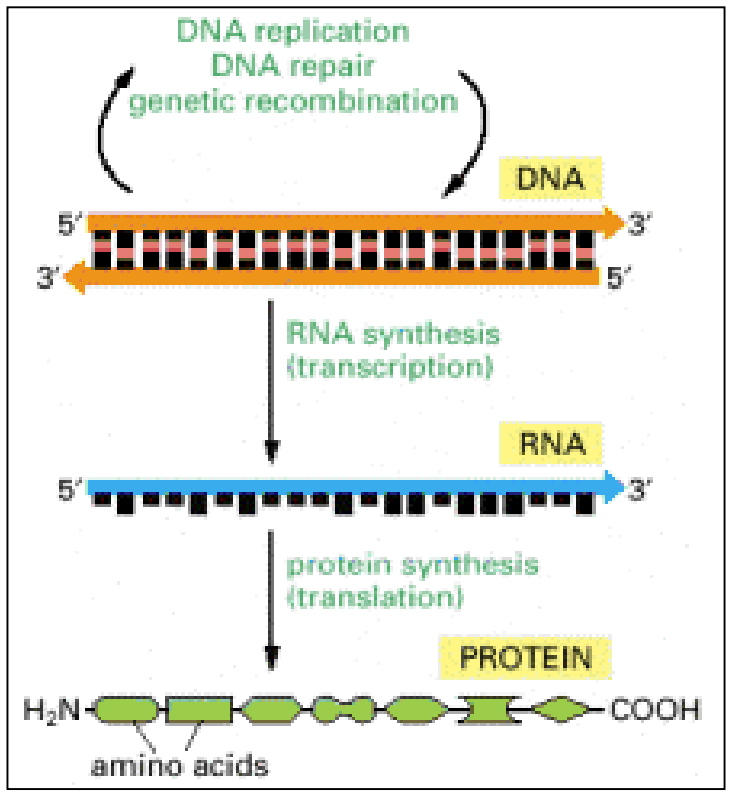
La cellula

- Ogni organismo vivente è composto da una o più cellule. Ogni cellula può essere vista come una macchina complessa che esegue delle istruzioni scritte e memorizzate nel proprio **genoma**.



L'informazione dentro la cellula





La genomica moderna: sequenze

- il dato primario del materiale genetico di un organismo, la sua sequenza di DNA, è risolvibile in maniera sistematica.

> *homo_sapiens*
ACTTTTTTACCCTCGTGTGTTGC
AGACTTTTTGCCACTTTTAAAAC
GCTGACAATTCGACCCTTTCCAA
GTGCAAAAAGTGCCAAGATTTA
CGATAAAATTCCCCCGAGAGAC
GTGTGCA……

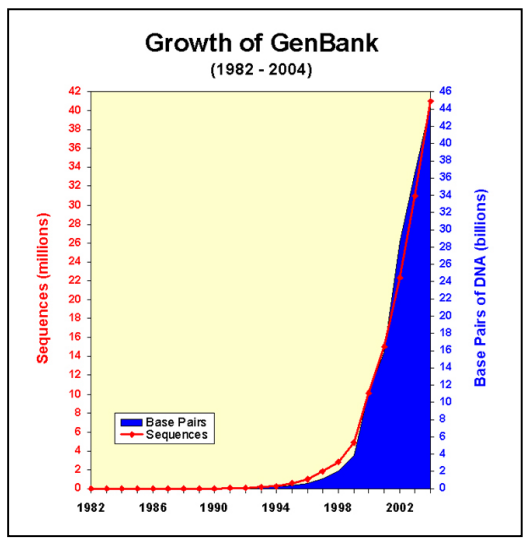
articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

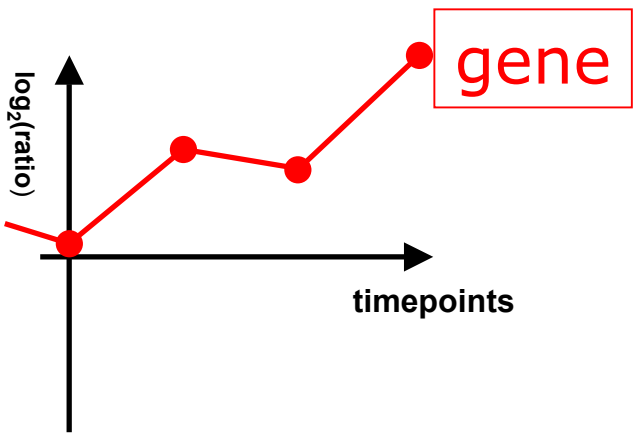
* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

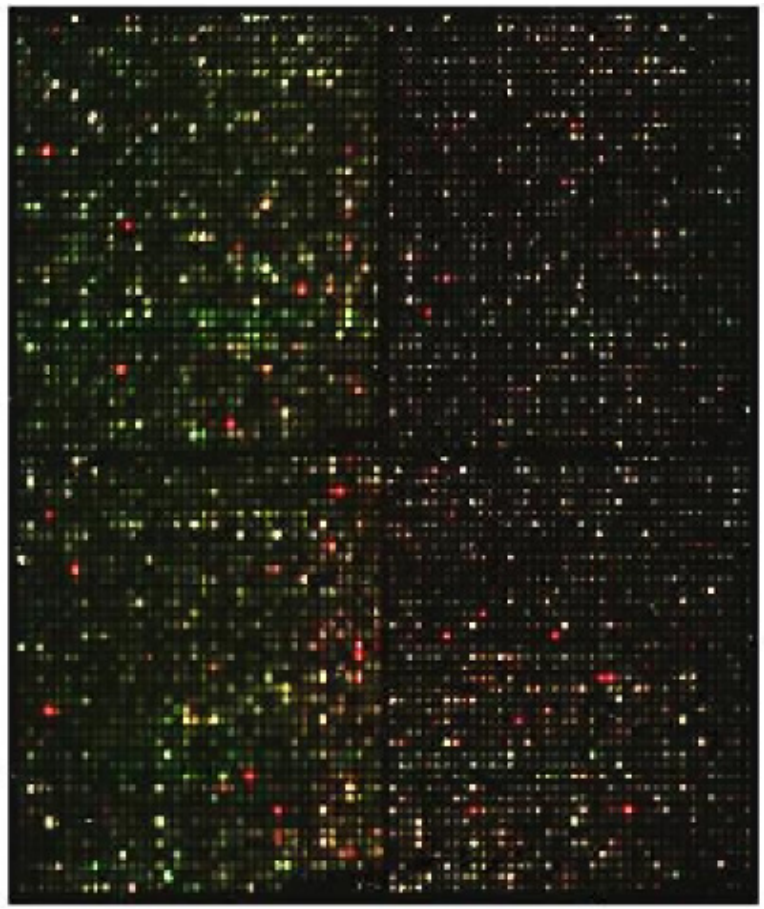
The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.



La genomica moderna: microarray

- In un esperimento di DNA microarray si misura il livello di espressione (mRNA) di migliaia di geni contemporaneamente

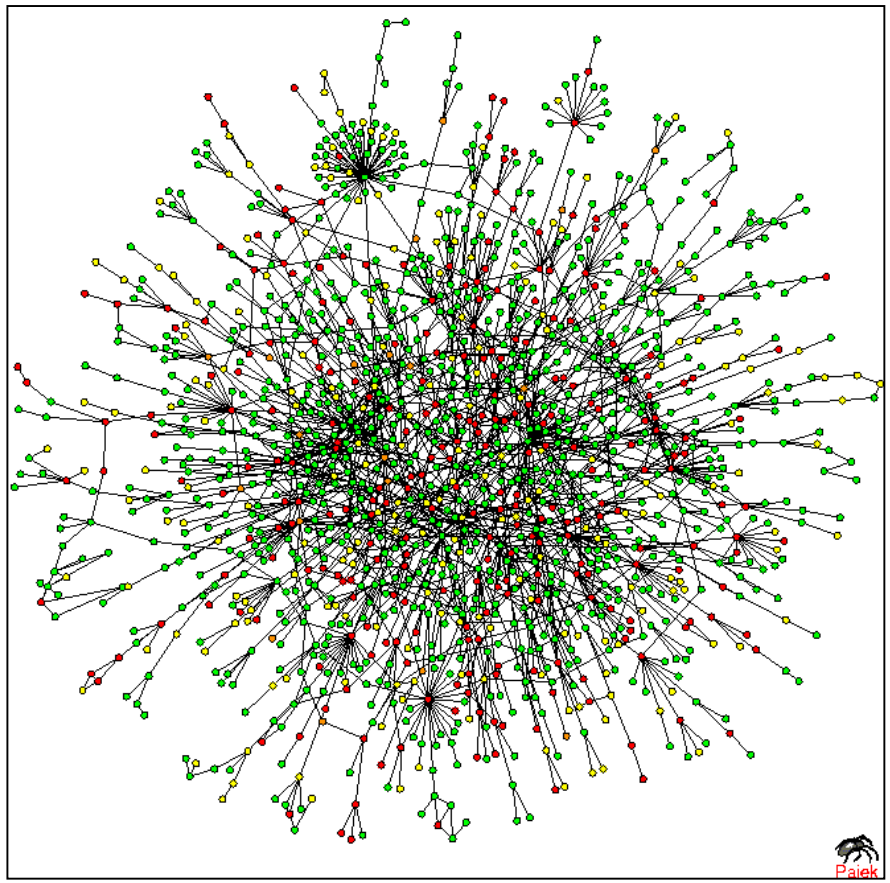




La genomica moderna: networks di interazione

- Le proteine (geni) dentro una cellula formano un network.

- La risposta di una cellula ad un certo stimolo è una risposta "globale", non di singole unità separate



La genomica moderna: perché interessa ad un fisico ?

- La biologia moderna è quindi caratterizzata da una enorme mole di dati

- Questi dati sono:

  - quantitativi
  - interessanti & nuovi
  - complessi (c'è molta informazione, ma è "nascosta")

- L'analisi di questi dati richiedi l'uso di tecniche matematiche, statistiche ed informatiche avanzate che un biologo non conosce.

Risorse / Congressi

- Master in bioinformatica – Università di Torino & Fondazione per le Biotecnologie
  http://www.masterbioinformatica.it

- Dottorato in "Sistemi complessi" – Università di Torino
  http://www.bioinformatica.unito.it/complex_systems/welcome.html

- letteratura:
  - archivio preprints:
    http://xxx.lanl.gov/archive/q-bio
  - NCBI pubmed:
    http://www.ncbi.nlm.nih.gov/

Risorse / Congressi

- ECCB 2006 (Eilat, Israele – 10 / 13 Settembre 2006)
  http://www.eccb06.org

- RECOMB 2006 (Venezia – 2 / 5 Aprile 2006)
  http://recomb06.dei.unipd.it

- BITS 2006 (Bologna – 28 / 29 Aprile 2006)
  http://www.biocomp.unibo.it/bits2006/home.html

Il gruppo TO61 di Torino

- Caselle Michele
- Pesando Igor
  - Corà Davide
  - Martignetti Loredana
  - Molineris Ivan
  - Re Angela
  - Sales Gabriele

Collaborazioni ----->

Paolo Provero	***@ Dip. Genetica***
F. Di Cunto	***@ Dip. Genetica***
Lorenzo Silengo	***@ Dip. Genetica***
Bussolino et al.	***@ IRCC***
Isabella Sbrana	***@ Univ. di Pisa***
C. Herrmann	***@ Marsiglia***
C. Dieterich	***@ MPI Tubinga***

Il gruppo TO61 di Torino

Il mio lavoro di ricerca:

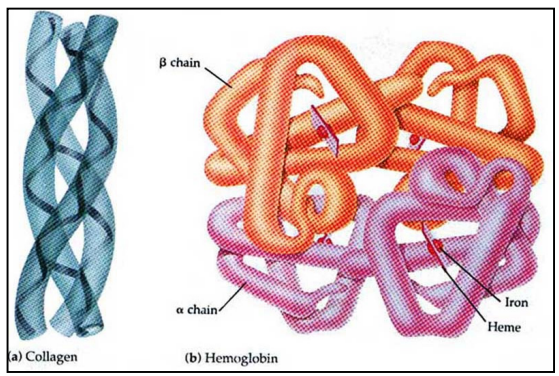
Sviluppo di metodi integrati per la identificazione di elementi regolativi in eucarioti superiori

Motivazione biologica

- Sequenza del genoma umano
  (2001 draft, 2004 finished)

~ 25000 "genes": proteins

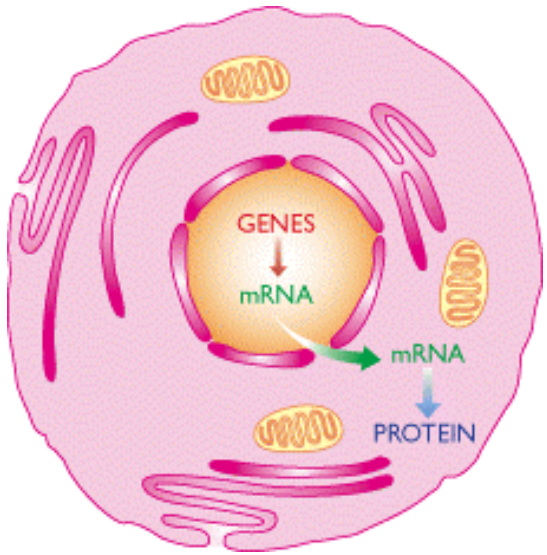
  - **3.2 x 10^9 bp di DNA**
  - **~ 3 % is coding: my "machinery"**
  - **~ 97 % è <u>non coding</u>**



Si pensa che formi la parte di regolazione
"regulation of my machinery"

Motivazione biologica

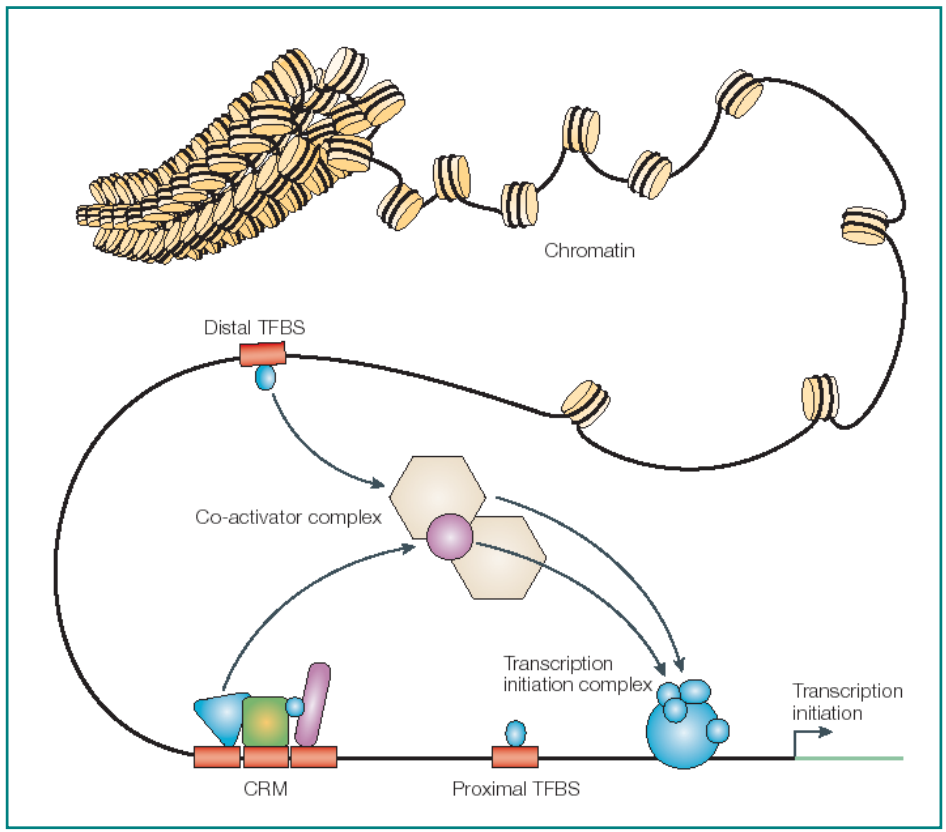
- La maggioranza dei geni specifica uno o più proteine: "**espressi**".
- L'espressione dei geni coinvolge un intermediario detto **messaggero** or **mRNA**.
- Il processo di espressione inizia con una fase detta "**trascrizione**" che è accuratamente controllata in ogni tipo cellulare.



Regolazione trascrizionale

Motivazione biologica

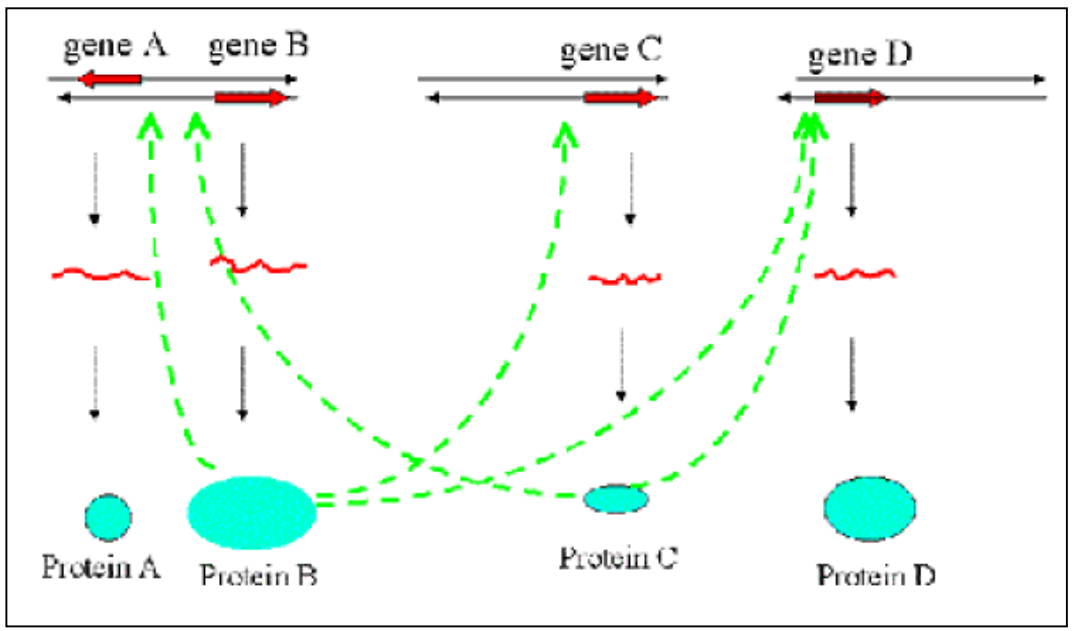
- Negli eucarioti superiori (es: uomo) l'evento di trascrizione è molto complesso



Wassermann, Nat. Rev. Genetics 2004 Apr;5(4):276-87

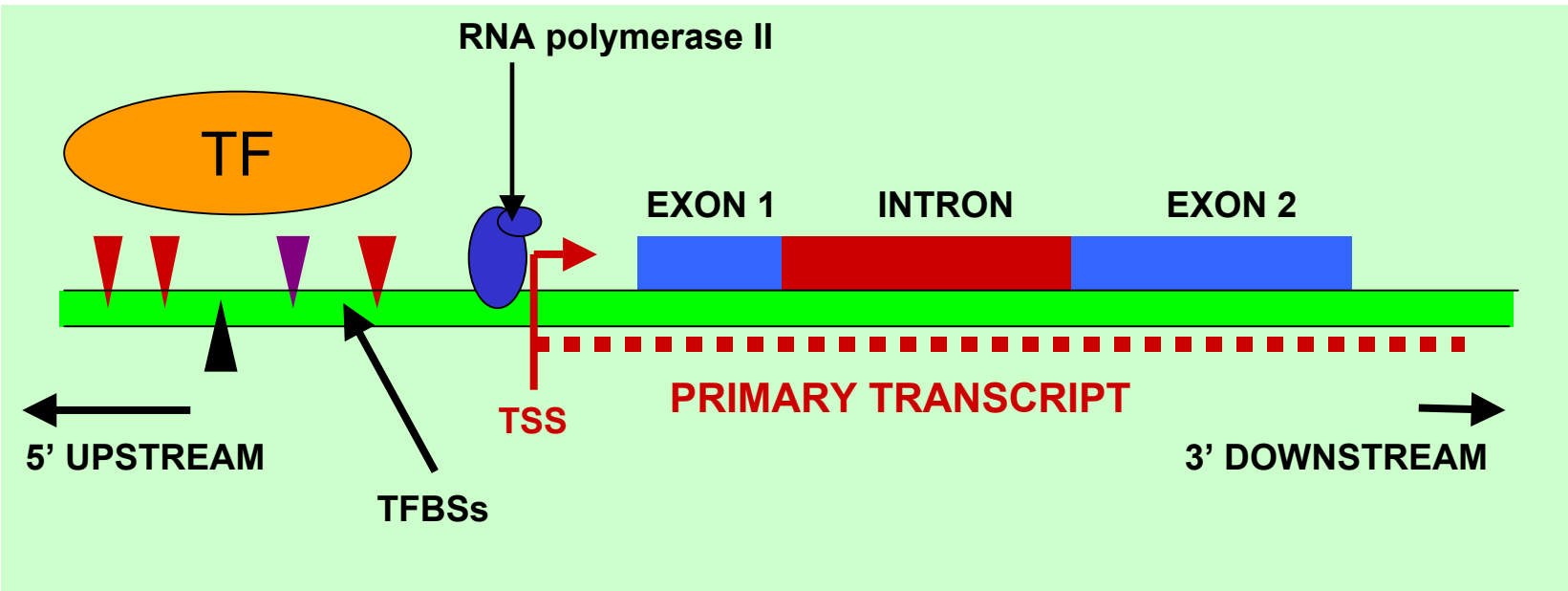
Motivazione biologica

- Negli eucarioti superiori la risposta trascrizionale è organizzata in un network.



Motivazione biologica

- Regolazione trascrizionale:
  fattori di trascrizione (TF) si accoppiano a particolari DNA motifs (TFBS) localizzati upstream del gene regolato.



Scopo del nostro lavoro

- Lo scopo del nostro lavoro è di identificare, a partire dalla sola sequenza genomica, dei candidati TFBS *ovvero* identificare il "vocabolario" di DNA motifs che regolano l'espressione dei geni.

➢ TFBS sono di solito corti (5-20 bp di DNA).

➢ TFBS sono di solito variabili.

➢ TFBS sono di solito dispersi su lunghe distanze(≥ 15000 bp nel caso umano).

➢ TFBS sono di solito attivi in entrambe le orientazioni.

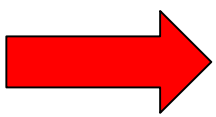
Il rapporo segnale / rumore è molto basso

Metodologia proposta

- TFBS sono di solito overrepresented.

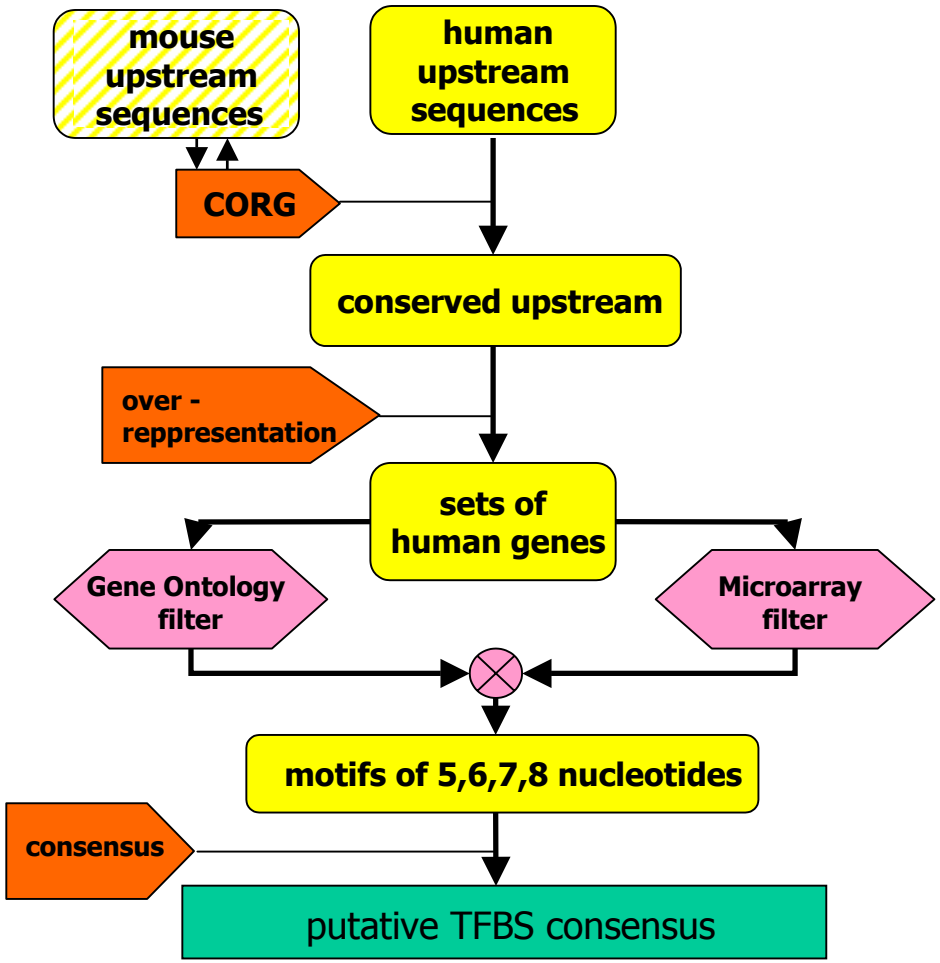
- TFBS sono di solito conservati evolutivamente.
  Usare genomica comparativa (phylogenetic footprinting).

- Geni con la stessa funzione di solito condividono anche gli stessi meccanismi di regolazione -> coregolati.
  Si possono usare esperimenti di microarray o annotazioni funzionali per indagare questo fatto.



focus sul caso umano

Metodologia proposta

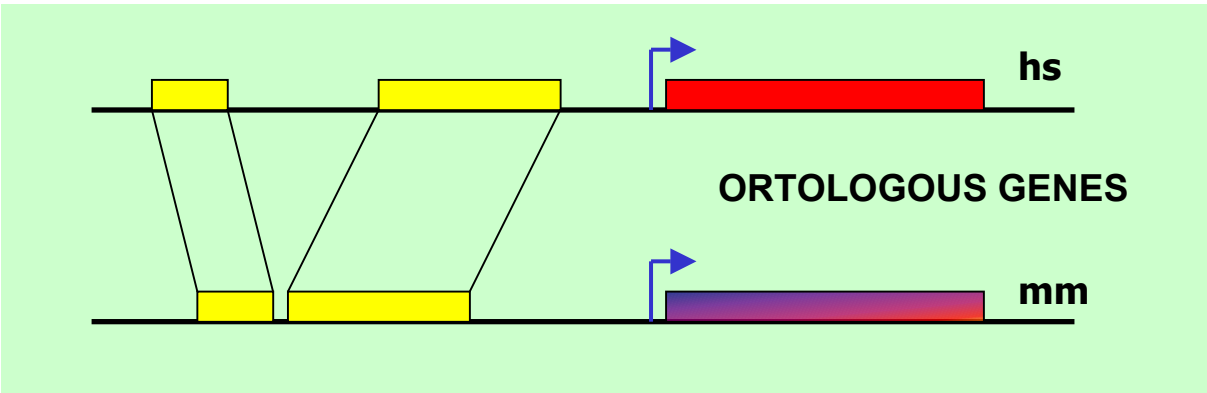


Metodologia di lavoro

- E’ un lavoro al calcolatore, es: perl, C/C++, mySLQ, applicativi specifici (Blast, ClustalW, R …).
- E’ un lavoro di analisi dati
- E’ un lavoro che richiede l’uso e l’implementazione di metodi e modelli statistici, analisi numerica, metodi Montecarlo.
- E’ un lavoro che richiede stretta collaborazione con i biologi.

CORG: COmparative Regulatory Genomics

- In human: almeno 15.000 bp upstream di tutti i geni.
- Esecuzione di una analisi comparativa per selezionare solo i tratti del genoma simili tra uomo e topo.



CORG è una collezione di blocchi upstream conservati tra uomo e topo (*Dieterich et al, NAR. 2003 Jan 1;31(1):55-7*)

Costruzione di set di geni

- Per ogni parola *w* di 5, 6, 7 e 8 nucleotidi costruiamo il set di geni in cui la parola *w* è <u>overrepresented</u>.

  - Ipotesi nulla:
    distribuzione random binomiale
    (parola *w* + il suo reverse complement)

  - Otteniamo per ogni parola *w* un insieme di geni *S(w)* da esaminare per evidenze di coregolazione.

> *gene name*
ACTTTTTTACCCTCGTGTGTT
GCAGACTTTTTGCCACTTTTA
AAACGCTGACAATTCGACCC
TTTCCAATCTCTCAAAAGTTT
CGACGAGCTGTACAACCCCC
CCCCC ……..

$$b_g(w) = \sum_{n=n_g(w)}^{L_g(w)} \binom{L_g(w)}{n} p(w)^n \left[1 - p(w)\right]^{L_g(w)-n}$$

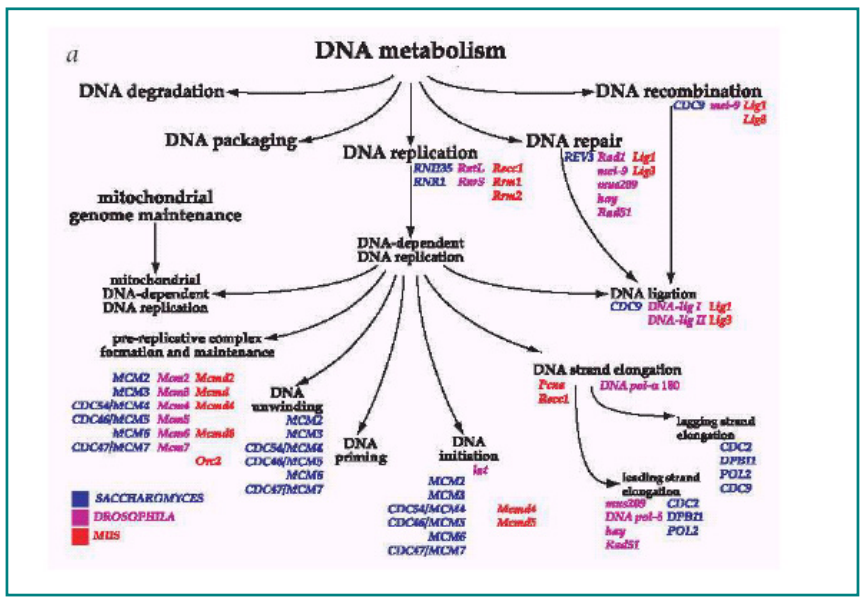
Van Helden et al,
Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.
J Mol Biol. 1998 Sep 4;281(5):827-42

Gene Ontology

- **Gene Ontology è un modello per l'unificazione di dati biologici.**

- **Lo scopo di GO è di costruire un vocabolario controllato per la descrizione di:**
  - Molecular function
  - Biological process
  - Cellular component
  di un certo gene.

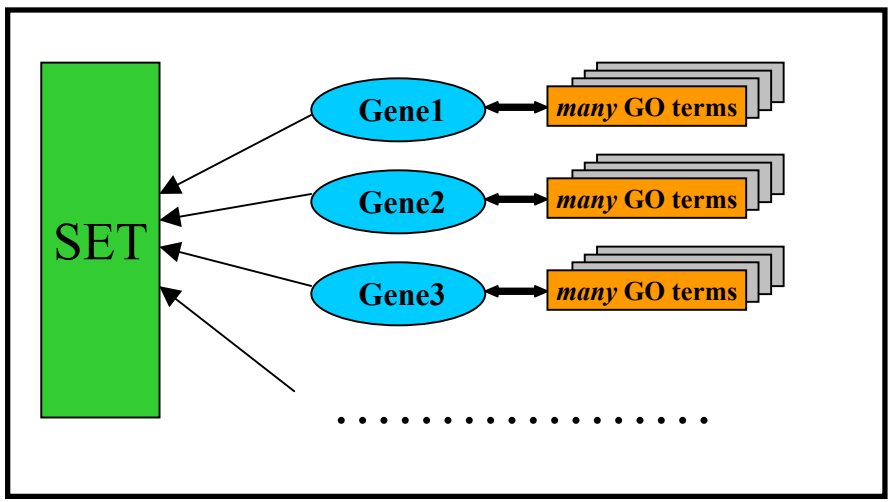
- **I vocabolari sono organizzati in un network gerarchico.**



Gene Ontology Consortium, Gene Ontology: tool for the unification of biology.
Nat Genet. 2000 May;25(1):25-9

Filtro Gene Ontology

- Per ogni set *S(w)* calcoliamo la prevalenza di tutti i termini di Gene Ontology (GO) relativi ai geni presenti nei set e paragoniamo i risultati ai numeri che ci aspetteremmo per caso in base ad un modello ipergeometrico.



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

Gene Ontology Consortium, Gene Ontology: tool for the unification of biology.
Nat Genet. 2000 May;25(1):25-9

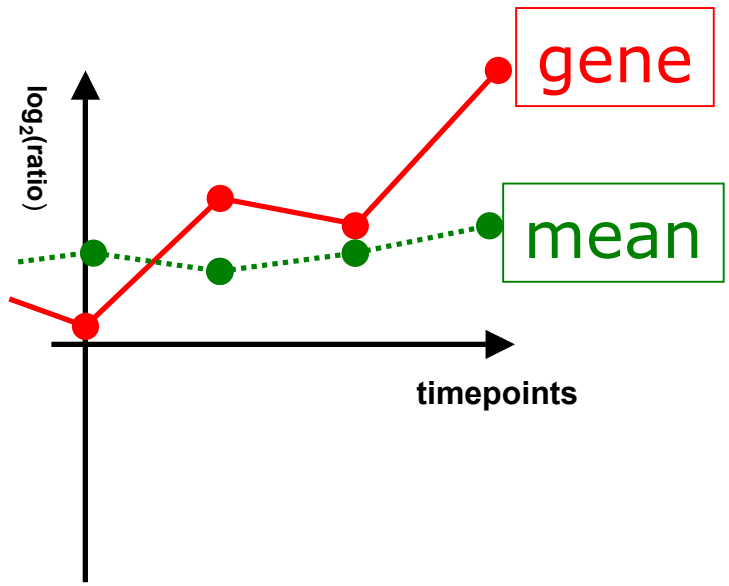
- Valutazione del False Discovery Rate attraverso una simulazione Montecarlo.

Filtro Microarray

- Per ogni set *S(w)* consideriamo il livello di espressione dei geni in 114 timepoints di un esperimento di cell cycle (Stanford human cell-cycle experiment).

Confronto con il livello medio calcolato su tutto il genoma per ogni timepoint usando un test di Kolmogorov-Smirnov.

(Jensen et al, Bioinformatics. 2000 16(4):326-33)



Costruzione di un consensus

Consideriamo le parole che sopravvivono ai tests precedenti i putative TFBS per i geni nel corrispondente set.

In molti casi alcune parole, simili tra di loro, risultano essere associate allo stesso Gene Ontology term, o allo stesso esperimento di microarray.

```
--ATCACG------
--- TCACGG----
CGTCACG------
--- TCACGGAA-
-----CACGGAAT
```

Si assemblano queste parole in una sequenza consensus.

→ ***WCONSENSUS package***

(Stormo et al, PNAS. 1989 86:1183-1187)

Risultati

Alla fine si ottiene un "dizionario" di putative TFBSs.

	1	2	3	4	6
1	actin filament-based process	P	GGGATTA	1	
2	ATP metabolism	P	CCGTCCC	1	
3	biosynthesis	P	CGCACG	1	
4	cell adhesion	P	AGGGC	1	*
5	cell communication	P	CGCGA	2	*
6	cell growth and/or maintenance	P	CTTCA	1	
7	cell motility	P	AGGGG	3	*
8	defense response	P	AGGAA	1	*
9	development	P	CCCC	16	*
10	DNA metabolism	P	GCGGGAA	7	
11	DNA replication	P	GCGCCAA	1	
12	DNA replication and chromosome cycle	P	TTCCCGCG	4	
13	DNA replication initiation	P	GCGCGAAA	1	
14	embryogenesis and morphogenesis	P	CGGGC	1	*
15	enzyme linked receptor protein signaling pathway	P	AGGGGG	1	*
16	epidermal differentiation	P	AGGCA	1	*
17	fluid secretion	P	TATAA	1	*
18	frizzled-2 signaling pathway	P	GCTGGAGA	1	
19	glycoprotein catabolism	P	CTGACCTA	1	
20	G-protein coupled receptor protein signaling pathway	P	AATTTA	4	*
21	G-protein signaling\, adenylate cyclase inhibiting pathway	P	ATCAA	1	
22	heterophilic cell adhesion	P	CTAAACTC	1	
23	immune response	P	GAAAC	1	

Esempi di risultati (robustezza)

- Di tutte le parole di 5,6,7 e 8 lettere (≈ 40000 in totale)
  - 93 differenti parole sopravvivono al filtro di Gene Ontology
  - 181 differenti parole sopravvivono al filtro di Microarray
  - L'intersezione tra le due classi è di 55 parole il che corrisponde ad un p–value ~ e^{-107}

Esempi di risultati (validazione)

- Confronto con databases di TFBS noti sperimentalmente (es: *TRANSFAC*) …..

Es:

TTGGCGC	Dna replication related terms	**E2F**
AGCAGG	Muscle / Muscle development / Sarcomere	**MyoD** **Myf5**
ACGTG CCACGTCA GACGTGGC	Endoplasmatic reticulum / protein transport	**ATF6**

Esempi di risultati (predizioni)

GGGGGGGT	Organogenesis	?
AGAGCCTT	Dna replication related terms / nucleosome assembly	?

..... esperimenti !!!!

Working group + references

- M. Caselle and D. Corà (Dip. Fisica Teorica - Torino)
- F. Di Cunto (Dip. Genetica, Biologia e Biochimica - Torino)
- P. Provero (Dip. Genetica, Biologia e Biochimica - Torino)
- C. Dieterich (Max Plank Institute - Tubinga)
- C. Herrmann (L.G.P.D. - Marsiglia)

- M. Caselle, F. Di Cunto and P. Provero
  BMC Bioinformatics. 2002;3:7.

- D. Corà, F. Di Cunto, P. Provero, L. Silengo and M. Caselle
  BMC Bioinformatics (2004); 5(1):57

- D.Corà, C.Herrmann, C, Dieterich, F. Di Cunto, P. Provero, and M. Caselle
  BMC Bioinformatics (2005); 6(1):110

Altri lavori in corso

- **Martignetti Loredana**
  → **Analisi di zone UTR, genomica comparativa**

- **Molineris Ivan & Sales Gabriele**
  → **Allineamenti genome–wide**

- **Re Angela**
  → **Analisi dei siti fragili**

- **Pesando Igor**
  → **Correlazioni tra DNA motifs**

The end

- **Special thanks**

  →INFN TO61 per l'ineguagliabile supporto

- **Questions and complains**

  → caselle@to.infn.it