Università degli Studi di Padova

Facoltà di Scienze MM. FF. NN.

Laurea di Primo Livello in Biologia Molecolare

Elaborato di Laurea

Ricerca di sequenze regolatorie della trascrizione tramite analisi bioinformatica delle interazioni tra fattori di trascrizione

Tutor: Prof. Giorgio Valle, Dipartimento di Biologia

Co-Tutor: Dott. Nicola Cannata, CRIBI – Dipartimento di Biologia

Laureando: Andrea Telatin

Anno accademico 2003/2004

Indice

1. Abstract

Il progetto Transfactory consiste nello sviluppo di una suite di tools bioinformatici per la ricerca di regioni regolative della trascrizione, ricercando nel genoma gli elementi cui possono legarsi specifici fattori di trascrizione e valutando le interazioni che possano instaurarsi fra questi.

Transfactory ricostruisce un database di fattori di trascrizione e dei corrispondenti siti di legame sul DNA a partire dalle informazioni disponibili in rete dalla banca dati TRANSFAC®, analizza i record dei fattori di trascrizione per determinare quali possano legare il DNA ed elabora le possibili interazioni che possono instaurarsi fra essi, direttamente o mediante fattori intermedi.

La suite ricerca in una sequenza genomica i siti di legame dei fattori catalogati, valuta la densità dei ritrovamenti, calcola il numero di interazioni che possono instaurarsi in ciascuna regione ad alta densità.

Transfactory permette da un lato di far emergere regioni promotoriali putative, dall'altro di verificare quali interazioni si instaurino tra fattori di trascrizione in promotori identificati sperimentalmente.

2. Introduzione

2.1. Fattori di trascrizione

Nella cellula eucariotica sono i fattori di trascrizione, piuttosto che l'RNA polimerasi, ad essere responsabili per il riconoscimento del promotore, fungendo da sito di attracco della RNA polimerasi (che da sola, infatti, ha scarsa affinità per matrici con frammenti di dsDNA).

La definizione di fattore di trascrizione è di tipo funzionale: appartengono a questa classe tutte le proteine necessarie per avere trascrizione ad un determinato promotore (o set di promotori)[1], questo significa che l'ingresso nella categoria è concesso anche a proteine poco caratterizzate.

I geni eucariotici sono sotto il controllo di un promotore organizzato modularmente, contenente diversi siti di legame per altrettanti fattori di trascrizione, nessuno dei quali è sufficiente ad avviare la trascrizione da solo. Il confronto di diverse regioni promotoriali ha mostrato una scarsa omologia, e sembra che il "minimo comun promotore" dei geni codificanti mRNA siano due soli elementi: un *Initiator* (Inr) che contiene il sito di inizio (generalmente una Adenina), ed una TATA-box circa in posizione -30, o in alternativa un "*Downstream Promoter Element*" (DPE) a valle del sito d'inizio. Escludendo questo motivo comune, ogni promotore è costituito da un peculiare set di elementi, spesso ripetuti e disposti secondo i due orientamenti; la modularità dei promotori è una caratteristica fondamentale dei genomi eucariotici.

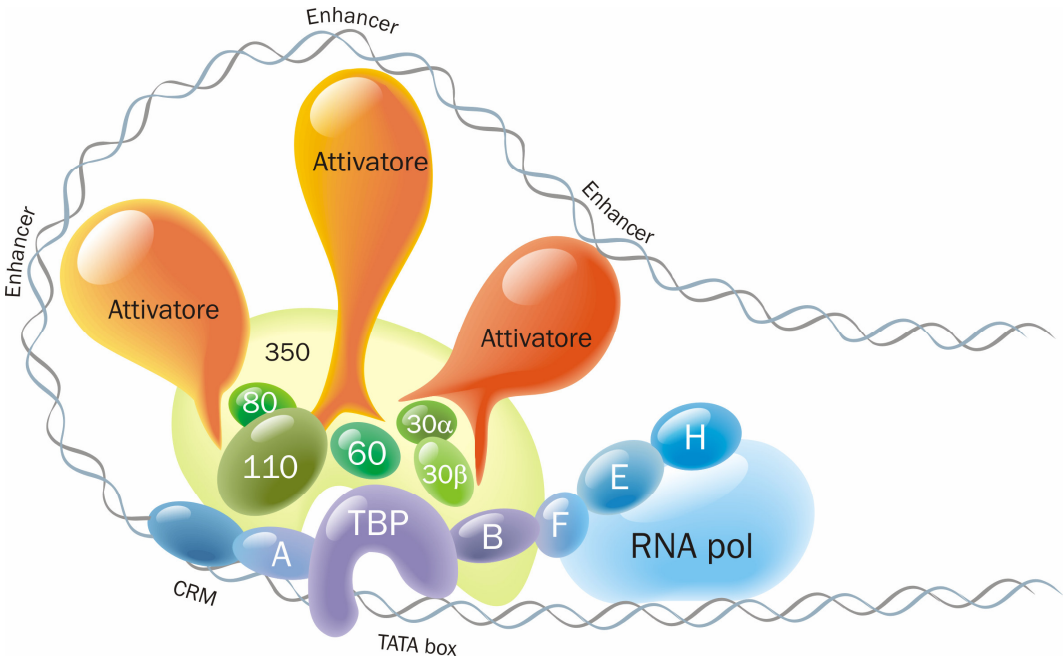

Fig. 2.1 – Rappresentazione schematica di un promotore eucariotico, con legati i fattori di trascrizioni (alcuni leganti il DNA altri interagenti con quest'ultimi) e la RNA polimerasi. Si notano i tre *enhancer* a monte, il CRM (*cis regulatory module*) e la TATA box, e l'organizzazione modulare dello stesso. In azzurro-lilla i fattori di trascrizione dell'apparato basale, in arancione i complessi degli *enhancer* ed in verde-giallo i fattori del *co-activator complex*.

Il promotore recluta i suoi fattori di trascrizione, che formano una piattaforma per ulteriori fattori che non legano direttamente il DNA (collettivamente formanti il "*co-activator complex*"). Quest'ultimo stabilisce contatti con fattori legati a cassette prossimali, ma anche con fattori di altre isole di elementi, dette **enhancer**, poste diverse kilobasi a monte (o a valle) del punto di inizio. Gli enhancer spesso contengono gli stessi elementi che si trovano nel promotore prossimale.

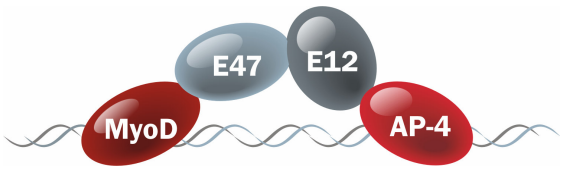

Fig. 2.2 – Schema di una "catena di interazione".

Alcuni particolari set di elementi si trovano in diversi promotori, detti *cis regulatory modules*, **CRM**) nei quali agiscono per ottenere una specifica funzione regolatoria (ad esempio alcuni CRM sono peculiari di un particolare pattern spazio-temporale di espressione)[5].

Ciascun fattore di trascrizione può interagire con un (generalmente ristretto) *pool* di fattori di trascrizione; solo alcuni di essi legano direttamente il DNA in un sito specifico. Tramite queste interazioni, due fattori, leganti il DNA,possono interagire indirettamente, come MyoD e AP-4 rappresentati in rosso in fig. 1.2, tramite dei fattori intermedi (in avio).

2.2. TRANSFAC®

TRANSFAC® è un database contenente informazioni sia sui fattori di trascrizione *trans*-agenti che sugli elementi regolatori di DNA *cis*-agenti di organismi eucariotici, per quanto riguarda geni codificanti proteine[6] (trascritti dall'RNA polimerasi II).

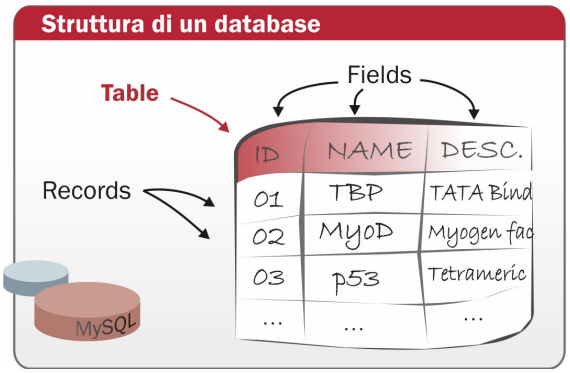

Fig. 2.3 – Schema di una tabella di database.

Un database è un insieme di tabelle (*tables*) correlate. Ogni tabella contiene i dati strutturati secondo campi (fields), ed ogni elemento della popolazione di dati è detto *record* (vedi figura 2.3).

TRANSFAC® contiene sei tabelle, due delle quali di interesse per il progetto: **SITE** (contenente i siti di legame al DNA, od elementi), e **FACTOR** (contenente informazioni sui fattori di trascrizione). Il database viene compilato a partire da pubblicazioni sui fattori di trascrizione, e le informazioni disponibili in ciascun record sono eterogenee, in quanto non per tutti i fattori sono disponibili i dati sperimentali per poter compilare allo stesso modo ciascun campo (es: per diversi fattori il cui campo "*Interagisce-con*" è vuoto, mentre fisiologicamente tutti i fattori interagiscono almeno con una proteina...).

2.3. Ricerca bioinformatica di regioni regolatorie

I fattori di trascrizione di interesse nella ricerca bioinformatica di regioni regolatorie sono primariamente quelli che legano il DNA. A partire da analisi biochimiche si può costruire un *modello* del sito di legame. Il modello più semplice è una sequenza consensus, uno più sofisticato ma flessibile, una matrice posizionale di peso. In TRANSFAC® è disponibile la sequenza consensus di tutti i fattori di trascrizione annotati come leganti il DNA, mentre – come evidenziato dai programmi del progetto – solo per il 5% dei fattori di trascrizione Umani è disponibile una matrice posizionale di peso, motivo per cui Transfactory utilizza le prime.

Ottenuto il modello che descriva il legame al DNA, questo viene utilizzato bioinformaticamente per la ricerca di regioni regolatorie, ma diverse assunzioni arbitrarie tendono ad inficiare la bontà della predizione. Il risultato finale è una eccessiva quantità di **falsi-positivi**, che sfocia nel "*futility theorem*" di Wasserman[4], secondo cui essenzialmente ogni sito di legame predetto non ha un ruolo fisiologico, in quanto spesso l'eccesso di falsi positivi è di oltre 1000:1 (come è facile aspettarsi considerando l'esigua lunghezza media delle sequenze di legame in confronto all'elevata lunghezza delle sequenze genomiche analizzate).

Generalmente si assume che il legame di un fattore sia indipendente da quello degli altri, non considerando quindi i legami cooperativi che si instaurano tra fattori di trascrizione adiacenti.

Inoltre mentre tutte le regioni di una sequenza risultano ugualmente accessibili all'analisi informatica, non è detto che sia altrettanto vero considerando i diversi stati di condensazione che può assumere il DNA in una cellula eucariotica.

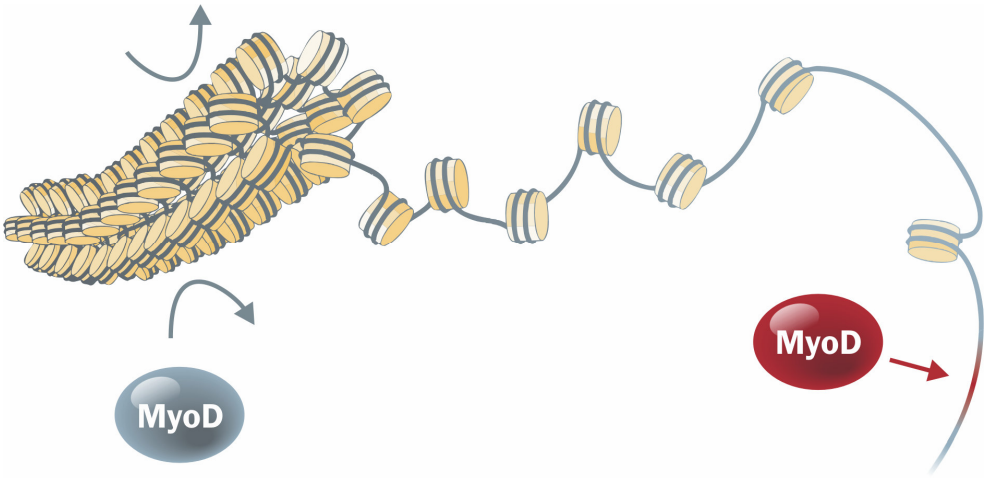

Fig 2.4 – Fra le possibili cause di predizioni "false-positive" vi è lo stato di condensazione che il DNA assume *in vivo*, che può rendere inaccessibili a fattori di trascrizioni sequenze che altrimenti legherebbero prontamente. Se una sequenza (a) contiene tre siti di legame per un fattore di trascrizione (nell'esempio CSBP-1) questi verranno identificati da un sistema bioinformatico, mentre *in vivo* uno solo potrebbe avere un ruolo funzionale.

Queste pecche si riscontrano in generale in tutti i sistemi predittivi che si basino sulla ricerca di siti di legame. Transfactory tenta di limitare l'imprecisione inserendo nel processo computazionale l'analisi delle interazioni tra fattori di trascrizione, tenta quindi di identificare i sopraccitati *CRM*, piuttosto che i singoli siti di legame dei fattori di trascrizione.

2.4. Precisazioni sul metodo applicato

Transfactory elabora (cfr. paragrafo 3.4) delle "catene di interazioni", come quella presentata in figura 2.2, basandosi sui dati del database TRANSFAC®.
La singola "catena di interazione" elaborata, in sé potrebbe non avere significato fisiologico (ad esempio se un componente della stessa non viene espresso assieme agli altri, se non sono corrette le informazioni sulle interazioni riportate su TRANSFAC®), tuttavia nel loro insieme tentano di associare fattori leganti il DNA anche se non risultano interagire direttamente, ai fini di una ricerca su larga scala.
La ricerca nel genoma permette di trovare sequenze legate dai fattori di trascrizione, tuttavia nel contesto di Transfactory si potrà dire che la ricerca permette di trovare i fattori di trascrizione stessi, data la relazione esistente ogni fattore di trascrizione e la propria sequenza target.

3. Materiali e Metodi

2.4. Panoramica di Transfactory Suite

Transfactory è una collezione di tools che permette la ricerca di regioni contenenti siti di legame a fattori di trascrizione, considerando sia la densità degli stessi che le possibili interazioni che si instaurano tra fattori cooperanti.

Un primo programma, **DataMiner**, permette di ricostruire, in una macchina locale, le tabelle di TRANSFAC®, scaricando dalla rete i singoli record delle tabelle SITE e FACTOR. Il programma **DataProcessor** poi elabora le informazioni scaricate generando (ed aggiornando) un database MySQL in un server locale.

Nei record della tabella SITE sono contenuti dati sulle interazioni conosciute fra fattori di trascrizione. **Interactor** elabora queste informazioni per stabilire quali fattori possano interagire tra loro anche indirettamente (per mezzo di fattori intermedi).

Un insieme di programmi (**GenomeScanner**) ricerca nel DNA i siti di legame dei fattori, determina le regioni a densità inaspettatamente alta, e setaccia quest'ultime vagliando la rete di interazioni che possono instaurarsi in ciascuna, effettuando uno scoring basato sul numero di interazioni.

Transfactory può quindi essere uno strumento per identificare regioni promotoriali putative, o più semplicemente può analizzare una sequenza ottenuta clonando la regione 3' di un gene, potendo quindi identificare i fattori di trascrizione dell'organismo in studio che possono legarvisi, evidenziandone le interazioni.

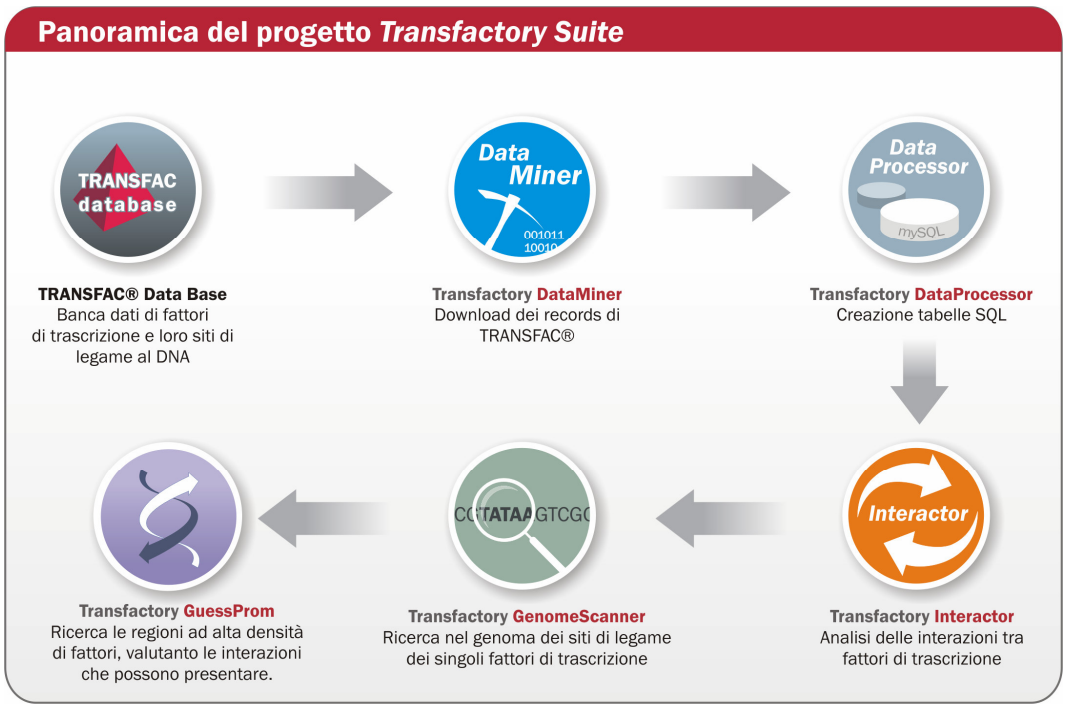

Fig. 3.1 – Rappresentazione schematica dei moduli componenti la suite di programmi "Transfactory".

3.2. DataMiner, download del database TRANSFAC®

Per poter ricostruire in locale una replica delle tabelle di TRANSFAC®, che sono pubblicamente consultabili ma non scaricabili *in toto* (se non a pagamento), è stato sviluppato, in ambiente Borland Delphi™[10], un'applicativo che simula la navigazione all'interno del sito, effettua la log-in, imposta la *query* (ricerca) e scarica uno ad uno i record selezionati dalla *query*. Delphi™ è un ambiente di sviluppo RAD (*Rapid Application Development*) visuale; il programma utilizza per la navigazione in Internet il modulo ActiveX[12] di Microft® Internet Explorer, facilmente riconoscibile in figura 3.1.
Il programma permette di selezionare l'organismo di cui si vogliono ottenere le informazioni, e la tabella da scaricare (FACTORS o SITES). Al termine della sessione (che richiede pochi minuti per scaricare i 960 fattori di trascrizione umani catalogati) tutte le pagine vengono salvate come files HTML in una cartella locale.

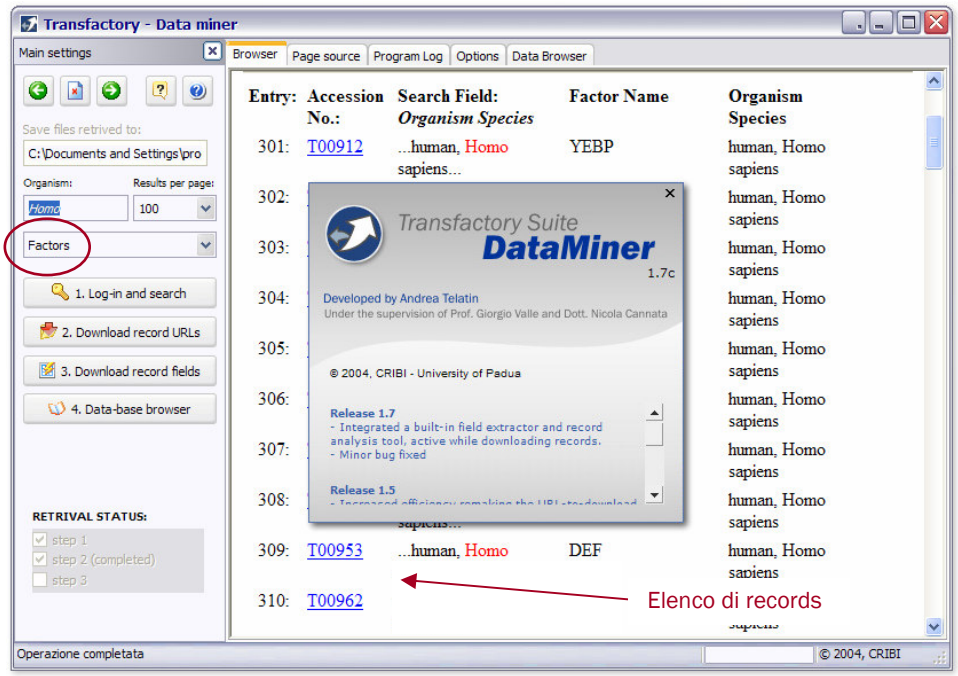

Fig. 3.2 – Schermata di DataMiner, con evidenziate alcune caratteristiche della stessa.

3.3. DataProcessor, estrazione dei record di TRANSFAC®

Scritto in Perl[11], DataProcessor elabora le informazioni scaricate da DataMiner, estraendo da ogni file HTML, tramite *pattern matching*[13], il contenuto di ogni campo, e genera come file di output i file MySQL che permettono di ricreare in locale le tabelle di interesse, ed i file .LST, che vengono interrogati dagli altri script Transfactory.
Da un client MySQL, già a questo step, è possibile ricavare diverse informazioni, ad esempio quali e quanti fattori leghino direttamente il DNA e/o interagiscano con altri fattori di trascrizione. Per ulteriori dettagli si veda il paragrafo 4.1.

3.4. Interactor, interazioni tra fattori di trascrizione

È di interesse sapere se due fattori di trascrizione interagiscano (informazione riposta in TRANSFAC®), non solo direttamente, ma anche indirettamente, come illustrato in fig. 2.2.

Analizzando il genoma, infatti, possiamo trovare gli elementi cui si legano i fattori trascrizione, e Transfactory intende determinare le interazioni fra questi fattori leganti il DNA, anche tramite proteine intermedie.

Interactor, a partire dal pool di fattori che contemporaneamente leghino il DNA e altri fattori (cfr. par. 3.2), effettua una ricerca combinatoria fra i fattori che possono interagire con questo, fino ad approdare ad un altro fattore del medesimo pool. In fig. 3.3 si vede Interactor mentre genera le "catene di interazione".

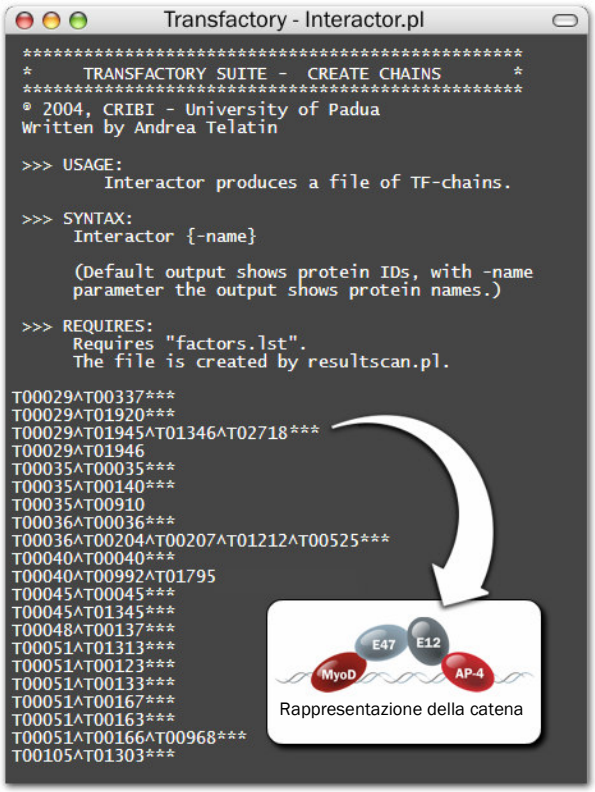

Fig. 3.3 – Screenshot di Interactor.pl, eseguito da riga di comando. Le catene sono rappresentate dal codice identificativo (ID) dei vari fattori.

Le catene terminanti con "***" sono valide: ovvero iniziano e terminano con fattori leganti il DNA; le altre sono catene "abortive" che il programma non è riuscito ad allungare ulteriormente, e verranno scartate.

Ai fini della ricerca in sequenze genomiche i fattori rilevanti sono il primo e l'ultimo di ogni catena, quelli che si legano al DNA. Un programma (*chainsimplifier.pl*) semplifica le catene producendo un file in cui ogni fattore legante il DNA viene associato a tutti i fattori leganti il DNA con cui direttamente od indirettamente può interagire.

3.5. GenomeScanner, ricerca degli elementi regolatori

I programmi preliminari hanno ottenuto ed ordinato le informazioni di TRANSFAC®, ottenendo una rete di interazioni – dirette ed indirette – ed associando ad ogni sequenza di legame nel DNA della tabella SITE, uno o più fattori che la leghino specificamente.

Il programma GenomeScanner ricerca nella sequenza di DNA fornitagli in input le posizioni degli elementi di DNA (presi dal file SITES.LST, generato da DataProcessor), e produce un file di output che contiene l'elenco di tutte le posizioni trovate. La ricerca

avviene nella sequenza fornita in input e nell'inversa-complementare che viene calcolata all'avvio. La ricerca avviene tramite il *pattern matching* di Perl, potente sistema di estrazione di stringhe di testo, e riguarda i singoli siti di legame, ignorando per ora le interazioni calcolate da Interactor.

A partire dalla sequenza di DNA da analizzare (sia essa un intero cromosoma o le poche kilobasi dell'inserto di un vettore) viene prodotto un elenco di posizioni.

Fig. 3.4 – Screenshot di GenomeScanner (ELEMENTFINDER.PL). Il programma è stato appena avviato, ha caricato in memoria il cromosoma 22, gli elementi da cercare e sta ora eseguendo la ricerca che richiede circa dieci minuti.

Il file output contiene la posizione del ritrovamento, la sequenza trovata, il fattore od i fattori che la legano, la direzione (*forward* o *reverse*), ed è il punto di partenza per la ricerca delle **regioni promotoriali** da parte di GuessProm. Le ultime righe dell'output generato dalla scansione del Cromosoma Umano 22 sono riportate in fig. 3.4.

Output del programma GenomeScanner

Riga	Posizione	Direzione	Sequenza cercata	Fattori legantisi
1531232	643956	10	REV	ATGAGTCAGA	T00029; T00123; T00133;
1531233	703942	10	FWD	ATGAGTCAGA	T00029; T00123; T00133;
1531234	212476	10	REV	ATGAGTCAGA	T00029; T00123; T00133;
1531235	144606	10	REV	ATGAGTCAGA	T00029; T00123; T00133;

Fig. 3.5 – Ultime righe dell'output generato dal programma nell'analisi del cromosoma 22. L'output consta di oltre un milione e mezzo di elementi trovati. Come si può notare, alcuni elementi di DNA sono bersaglio di diversi fattori di trascrizione.

3.6. GuessProm, ricerca delle regioni regolatorie

GuessProm a partire dall'output di GenomeScanner esegue una ricerca di regioni ad alta densità di siti di legame (utilizzando parametri di soglia e larghezza della finestra definiti dall'utente) considerando non solo il numero di elementi riscontrati, ma anche le possibili interazioni che potrebbero intercorrere tre i fattori leganti gli elementi trovati. Il sistema esegue una scansione con una finestra di larghezza definita dall'utente, e registra il numero di siti di legame presenti in ciascuna finestra. Un parametro di *cutoff* permette di tener conto solo di valori superiori alla soglia. E' possibile specificare il passo della finestra (di quanto distanziare una finestra dalla successiva), in modo da velocizzare la scansione. È prudente scegliere un passo non superiore ad un decimo della finestra, per registrare in maniera affidabile le posizioni di inizio e fine delle regioni promotoriali trovate.

Le finestre adiacenti hanno un alto tasso di sovrapposizione, perciò il programma registra la prima posizione in cui rileva un segnale superiore al *cut-off*, e l'ultima, generando nell'output una serie di regioni ad alta densità, e in queste regioni controlla le interazioni, con un indice che normalizza la densità di interazione rispetto alla densità stessa:

$$score_{r,s} = \frac{(\alpha \cdot f) + (\beta \cdot i)}{(s - r)}$$

In una finestra che nel genoma va dalla posizione (base) ***r*** alla posizione **s**, il programma conta i fattori che si possono legare (**f**) ed il numero di interazioni che si possono instaurare fra gli stessi (**i**). Due coefficienti α e β calibrano il peso che si vuol dare ai due parametri.

Questo indice permette di ricalcolare il peso delle regioni ad alta densità di siti di legame, penalizzando i siti ricchi di siti di legame per fattori che non sembrano interagire tra loro. L'output di Guessprom, in formato CSV (*comma separated values*), può essere aperto direttamente con Microsoft Excel® per ulteriori analisi.

Una considerazione sull'algoritmo

Un primo approccio al problema del calcolo delle interazioni è stato diretto: per ogni fattore trovato in una finestra (in tutto *n*) verificare se interagisca o meno con gli altri (in tutto *n-1*). L'**ordine di complessità** di questo algoritmo è $O(n^2)$, e rallentava pesantemente l'elaborazione dei dati.

Un nuovo approccio adottato è stato di creare una stringa, in ogni finestra, contenente gli ID dei fattori che posso legarsi (es: T00001-T00012-T01844). Dopodiché per ogni fattore di trascrizione si cerca, tramite *pattern matching* di Perl, se i fattori del set con cui può interagire sono contenuti in questa stringa. L'algoritmo diventa di complessità lineare, paragonabile al semplice conteggio della densità di siti di legame.

3.7. Schema delle dipendenze dei programmi

Questo schema (fig. 3.6) presenta le dipendenze dei programmi della suite.

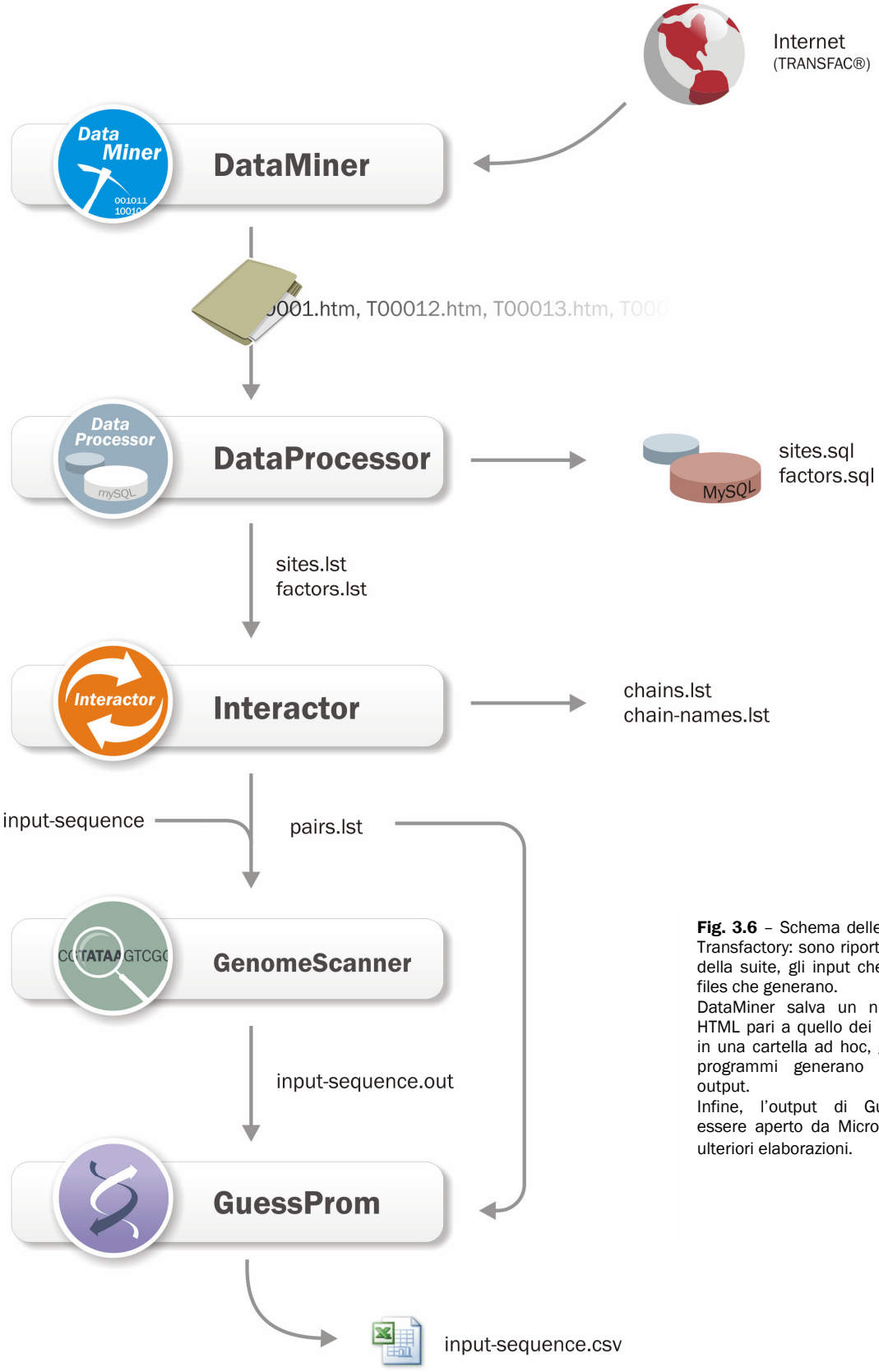

Fig. 3.6 – Schema delle dipendenze di Transfactory: sono riportati i programmi della suite, gli input che utlizzano ed i files che generano.
DataMiner salva un numero di files HTML pari a quello dei record scaricati in una cartella ad hoc, gli altri, invece, programmi generano singoli file di output.
Infine, l'output di GuessProm può essere aperto da Microsoft Excel® per ulteriori elaborazioni.

4. Risultati e Discussione

4.1. Download dei record di TRANSFAC®

Utilizzando DataMiner (vedi paragrafo 2.1) per l'organismo *Homo sapiens* sono stati scaricati 960 files dalla tabella FACTOR e 1281 dalla tabella SITE. Dopo aver ricostruito il database MySQL con DataProcessor è stato possibile ottenere il numero effettivo di fattori di trascrizione: di alcuni (482) infatti non si conoscono le interazioni con altre proteine o con il DNA e sono stati scartati.

Tramite semplici query SQL è stata individuata la struttura dei record significativi. In particolare per 98 si sa che legano il DNA ed interagiscono con altri fattori di trascrizione, sono il pool di partenza e di arrivo per la ricerca di "catene di interazione" (vedi fig. 2.2) da parte del programma Interactor (cfr. paragrafo 3.3).

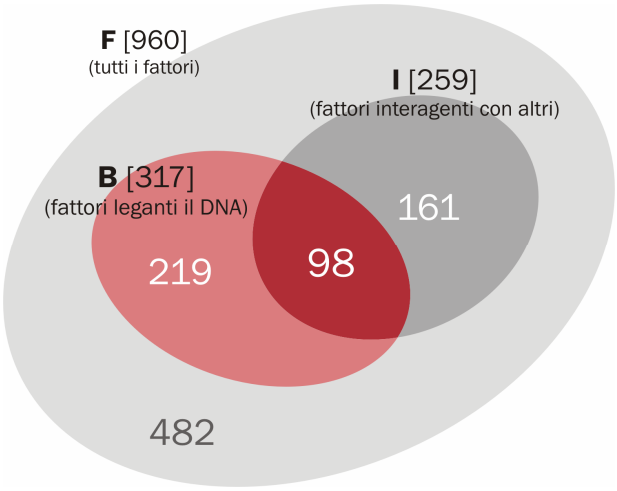

Fig. 4.1 – Distribuzione dei record della tabella FACTOR. Il sottoinsieme **B** (Binding DNA) ne contiene 317 ed il sottoinsieme **I** (Interacting with other factors) 161. L'insieme **I∩B** ne contiene 98. Queste informazioni sono recuperate tramite *query* SQL, ad esempio per l'insieme IB :

```
select count(id) from factors where (not isnull(bindsdna)) and (not isnull(interacts));
```

4.2. Catene di interazione elaborate da Interactor

Il programma Interactor ha ricostruito 169 catene di interazione compiute, delle quali ne vengono qui riportate quattro a titolo esemplificativo:

ID	FATTORI DI TRASCRIZIONE DELLA CATENA						
01	TBP	TAF(II)100	TAF(II)18	TAF(II)28	TAF(II)20	TAF(II)30	ER-α
02	c-Rel	NF-κB1 *p*	IκB-α	NF-κB			
03	E2F-1	P53					
04	c-Rel	NF-κB					

TRANSFAC® spesso include come record distinti di fattori di trascrizione, sia un complesso, che i suoi componenti. Questa è una duplicazione di informazione che non agevola le elaborazioni automatiche e può portare ad artefatti. Inoltre isoforme, varianti e precursori possono appesantire l'elenco: c-Rel interagisce direttamente con

NF-κB, ma anche tramite NF-κB1 *p*, che ne è il precursore (secondo esempio della tabella).

Le catene più lunghe (7-8 fattori) includono diversi TAFs (*TBP associated factors*), come nel primo esempio della tabella. Si nota inoltre che il 67% delle catene è lunga due unità: si tratta quindi di fattori che interagiscono direttamente. La distribuzione delle lunghezze delle catene è presentata in figura 4.2.

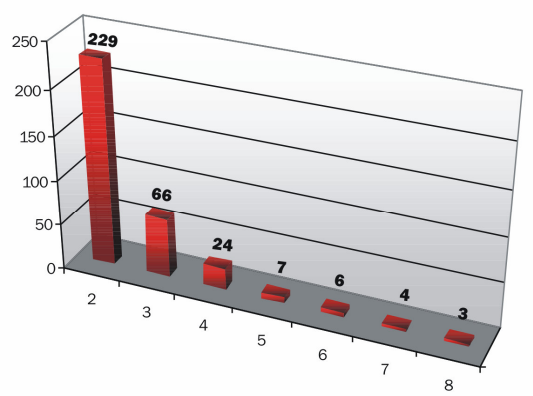

Fig. 4.2 – Distribuzione delle lunghezze delle catene di interazione elaborate da Interactor.

4.3. Ricerca di siti di legame

La ricerca di siti di legame (elencati nel file sites.lst, generato da DataProcessor) viene eseguita da GenomeScanner. La ricerca di test è stata effettuata nel Cromosoma Umano 22 (lungo circa 35 kb).

SEQUENZA UTILIZZATA	ELEMENTI TROVATI
22N - Cromosoma 22 (Sequenza completa)	1 531 233
22M - Cromosoma 22 (Repeat Masker)	847 624

Il numero di picchi trovati viene quasi dimezzato se si fornisce a GenomeScanner una la sequenza genomica processata da *Repeat Masker*[14], un tool che sostituisce le basi di regioni a bassa complessità con delle "N", che vengono quindi ignorate dai software che effettuano ricerche nella sequenza.

4.4. Ricerca di regioni regolatorie

L'analisi dell'output di GenomeScanner effettuata da GuessProm è volta a rilevare le regioni ad alto segnale (densità di siti di legame e densità di possibili interazioni, cfr. paragrafo 3.6).

Il numero di picchi rilevati dipende dai parametri fornito per la ricerca (larghezza finestra, passo della finestra, *cut-off*).

L'analisi delle sequenze precedenti è avvenuta con due set di questi valori:

	Set A	Set B
Finestra	200 bp	500 bp
Passo	20 bp	50 bp
Cut-off	25 (fattori/finestra)	50 (fattori/finestra)
Picchi trovati (22N)	1 556	703
Picchi trovati (22M)	894	449

Il file di output di GuessProm può essere visualizzato direttamente con Microsoft Excel®, foglio di calcolo che permette facilmente di visualizzare in forma grafica i valori della tabella.

Per la sequenza 22M (Cromosoma 22 elaborato da *Repeat Masker*) si è ottenuto un grafico in cui sono rappresentati, in rosso, i picchi delle interazioni, ed in grigio, i picchi. I valori dei picchi (espressi in percentuale) sono relativi al massimo valore (il picco di interazioni pari al 100%).

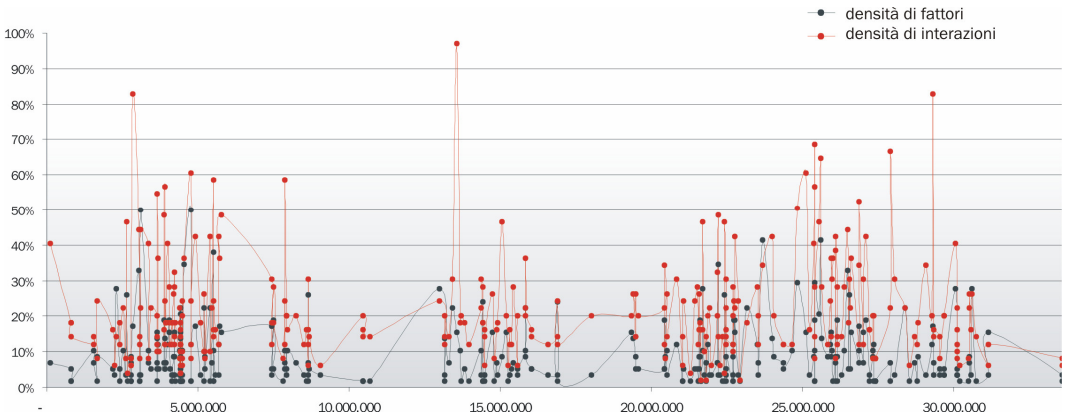

Fig. 4.3 – Elaborazione della scansione del Cromosoma Umano 22 in Microsoft Excel®.

Aprendo una finestra di 880.000 basi di questo cromosoma con il GenomeBrowser[8], e confrontandola con la predizione, si nota certa corrispondenza fra gli mRNA identificati nella stessa e le predizioni di GuessProm, come si può notare in fig. 4.4.

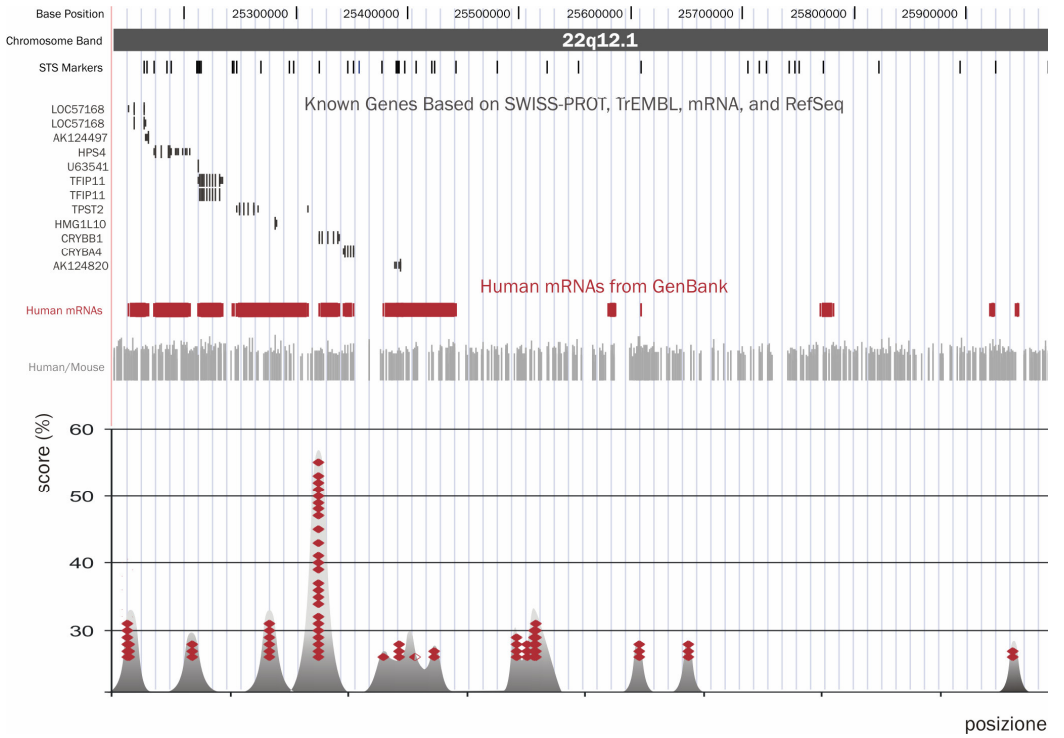

Fig. 4.4 – Confronto di una regione (Chr22:25.136.565-26.250.000) del Cromosoma 22 da 880.000 basi visualizzata con GenomeBrowser (in alto) e i picchi individuati da GuessProm (in basso). Si nota una certa corrispondenza di quest'ultime con la densità degli mRNA.

In figura 4.5, infine, si confronta la predizione effettuata in una regione contenente un gene (*TNNC1*), con finestra centrata sul gene stesso in GenomeBrowser.
GuessProm in questo caso individua due picchi, uno a monte ed uno a valle del sito.

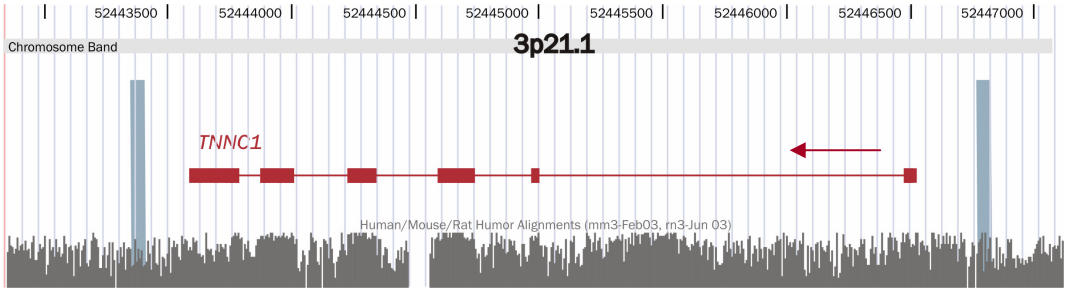

Fig. 4.5 – Confronto analogo al precedente applicato ad una regione ristretta (4.400 basi del terzo Cromosoma Umano che includono il gene *TNNC1*). L'orientamento del gene è indicato dalla freccia. I picchi di GuessProm sono rappresentati dalle due barre avio verticali, la seconda potrebbe rappresentare il promotore del gene.

4.5. Prospettive future

Un programma predittivo deve essere validato con un *training set* di sequenze note per poterne calcolare l'**affidabilità**[2]. Dopo aver eseguito la predizione, si possono ottenere i seguenti parametri:

- Specificità ($\frac{VN}{VN+FP}$);
- Sensitività ($\frac{VP}{VP+FN}$);
- Selettività ($\frac{VP}{VP+FP}$);

VP = Veri Positivi;
VN = Veri Negativi;
FP = Falsi Positivi;
FN = Falsi Negativi;

L'affidabilità del sistema può essere espressa anche con un unico indice di correlazione, che vale 1 in caso di perfetta corrispondenza predizione-realtà e -1 in caso di predizione totalmente imperfetta:

$$Correlazione = \frac{(VP)(VN)-(FN)(FP)}{\sqrt{(VN+FN)(VN+FP)(VP+FN)(VP+FP)}}$$

Nel caso di Transfactory si deve dunque disporre di un set di promotori ben annotati e porzioni sicuramente prive di funzione regolatoria.
Data la difficoltà di reperire un *training set* di dimensioni adeguate da essere statisticamente significativo, queste analisi possono essere affiancate a confronti dei risultati ottenuti per la medesima regione con altri sistemi predittivi.
Questa verifica permetterà anche di **ottimizzare** il sistema impostando, in GuessProm, i parametri (cfr. paragrafo 3.6) ed i coefficienti che forniscono mediamente la miglior correlazione.

5. Bibliografia

Testi di riferimento

1. **Lewin B.**, *Genes VIII,* 2003, Pearson – Prentice Hall
2. **Valle G. et al,** *Introduzione alla Bioinformatica*, 2003, Zanichelli (Bologna)
3. **Siever E., Spainhour S., Patwardhan N.**, *Perl in a nutshell*, 1999, O'Reilly

Articoli

4. **Wasserman W. W., Sandelin A.,** Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*. 2004 Apr;5(4):276-87. Review.
5. **Fickett, J. W. & Hatzigeorgiou, A. G.** Eukaryotic promoter recognition. *Genome Res*. 7, 861–878 (1997).
6. **Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. and Urbach, S.**, The TRANSFAC system on gene expression regulation, *Nucleic Acids Res*. 2001 Jan 1;29(1):281-3.

Web links

7. Aggiornamenti su Transfactory, http://transfactory.neofox.net/
8. GenomeBrowser (italian mirror), http://pandora.cribi.unipd.it/
9. MySQL®, http://www.mysql.com/
10. Borland Delphi™, http://www.borland.com/delphi/
11. Perl, www.perl.com, www.cpan.org
12. ActiveX™ technology by Microsoft®, http://www.microsoft.com/com/tech/ActiveX.asp
13. Perl – Regular Expressions, http://www.perldoc.com/perl5.8.0/pod/perlrequick.html
14. Repeat Masker, http://www.repeatmasker.org/