DOTTORATO DI RICERCA
in
SCIENZE COMPUTAZIONALI E INFORMATICHE
Ciclo XXIV
Consorzio tra Università di Catania, Università di Napoli Federico II,
Seconda Università di Napoli, Università di Palermo, Università di Salerno

SEDE AMMINISTRATIVA: UNIVERSITÀ DI NAPOLI FEDERICO II

MARIA ROSARIA DEL SORBO

Algoritmi per l'elaborazione di serie temporali:

Compressione con perdite di dati GPS
Verifica di co-regolazione di geni direttamente collegati in network biologici

TESI DI DOTTORATO DI RICERCA

COORDINATORE: Prof.. Ernesto Burattini
TUTOR: Prof.ssa Amelia Giuseppina Nobile

Anno Accademico 2010/2011

Maria Rosaria Del Sorbo

Dipartimento di Matematica e Applicazioni “R. Caccioppoli"
Università degli Studi di Napoli “Federico II”
Complesso Universitario di Monte Sant’Angelo,
via Cintia - 80126 Napoli, Italy
e-mail: marodel@unina.it

La cosa più bella che possiamo sperimentare è il mistero; è la fonte di ogni vera arte e di ogni vera scienza.

Albert Einstein

Non sono i frutti della ricerca scientifica che elevano un uomo ed arricchiscono la sua natura, ma la necessità di capire e il lavoro intellettuale.

Albert Einstein

I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale.

Marie Curie

Indice

Abstract

Tra le tante differenti attività che hanno richiamato la mia attenzione durante il periodo triennale del dottorato di ricerca, è possibile individuare un'area d'interesse comune, l'estrazione di conoscenza da dati ottenuti in contesti diversi ma raccolti sotto forma di serie temporali; nella presente lavoro conclusivo ho scelto di presentare casi di studio di serie temporali su due tematiche particolari, riguardanti argomenti di grande attualità e decisamente avvincenti. All'interesse di base ha fatto seguito un'attività di progettazione, realizzazione e sperimentazione in parte eseguito all'estero. Risvolto concreto dello studio è stata l'implementazione a scopi di simulazione di alcuni algoritmi per l'elaborazione di elementi provenienti da dataset di grandi dimensioni, che ha fruttato, tra l'altro, anche la preparazione di alcuni articoli e la pubblicazione di uno di essi. I due argomenti cui sarà dedicato questo studio sono i seguenti:

a. *Compressione con perdita di dati generati da Global Positioning System (GPS) basata su segmentazione spazio-temporale.*

Il Global Positioning System (GPS) è uno dei sistemi più comunemente utilizzati nelle attuali tecnologie di posizionamento, a causa della sua utilità pratica del suo facile reperimento e della sua versatilità che ne rende possibile l'utilizzo anche per studi scientifici. Alcune tra le tante applicazioni richiedono il salvataggio dei dati da inviare successivamente ad una stazione ricevente a distanza. In questi casi, la quantità di dati da inviare può essere piuttosto cospicua e, quindi, la compressione di questi dati si rende in alcuni casi un'operazione indispensabile. Anzi, il problema della compressione di dati GPS diviene sempre più incalzante con l'aumentare di applicazioni che richiedono, ad esempio, una elaborazione in remoto real time dei dati generati dai dispositivi di posizionamento. Un altro motivo per cui la compressione risulta utile è legato alla indicizzazione dei dati nei dispositivi mobili, in cui si evidenzia la frequente necessità di effettuare ricerche sui dati raccolti. Nello studio cui ho collaborato che è stato recentemente pubblicato [W. Balzano, M.R. Del Sorbo, CoTracks: a lossy compression schema for tracking logs data based on space-time segmentation, CCP1, 1st IEEE International Conference on Data Compression, Communication and Processing, Palinuro, 21-24 June 2011, pagg. 168-171], le serie temporali prese in esame sono dei record di dati prodotti da dispositivi GPS per campionamento ad intervalli di tempo regolari e prefissati e contengono nei loro campi tutte le informazioni relative a parametri istantanei sia di tipo cinematico, come posizione, tempo, velocità, che di controllo della qualità, come i parametri DOP.

Le suddette serie temporali sono state sottoposte a un trattamento preventivo, volto a depurarle dai dati ridondanti, come quelli derivanti da sovra campionamento (oversampling), e dai dati scarsamente affidabili, come quelli che presentano i valori dei parametri di qualità minori di un valore prefissato di soglia. Per realizzare l'obiettivo di una accettabile ed efficace compressione con perdita, poi, la serie temporale, considerata come una sequenza di punti in un iperspazio di dimensione pari a quella dei vettori, è stata sottoposta a una clusterizzazione basata sull'appartenenza dei punti a Minimum Bounding Box, riferiti a parametri dinamici e di qualità, opportunamente generati. In tal modo è stato possibile ottenere rapporti di compressione molto elevati, anche oltre il 90%, senza alcuna significativa perdita dell'informazione contenuta nei segnali.

b. *Verifica di co-regolazione di geni direttamente collegati in network biologici sulla base di serie temporali di espressione genica.*

La seconda parte dello studio sulle serie temporali, maturato durante il recente visiting presso l'Intelligent Systems and Bioinformatics Laboratory della Wayne State University di Detroit, ha preso spunto da argomenti di Bioinformatica e precisamente da analisi di dataset prodotti da esperimenti di tipo microarray per l'indagine sull'espressione genica. Si tratta di dati raccolti ad intervalli di tempo regolari, che costituiscono una serie temporale breve di espressione genica. Questo tipo d'indagine è collegata con i signaling pathway cellulari ed è finalizzata ad accrescere la base di conoscenza su questi schemi di comunicazione tra cellule. In particolare lo studio è basato sull'osservazione che i geni differenzialmente espressi negli esperimenti di genomica ad elevato throughput possono evidenziare e quindi individuare i percorsi di segnalazione cellulare coinvolti nelle reazioni ad un determinato stimolo biologico, ad esempio di tipo farmacologico. L'analisi d'impatto sui percorsi di segnalazione cellulare proposta in [A.L Tarca, S. Draghici et al., A novel signaling pathway impact analysis, Bioinformatics 2009 25(1):75-82] considera i percorsi stessi come grafi orientati e pesati i cui nodi sono occupati dai geni che prendono parte al fenomeno biologico o alla funzione descritta dal percorso. Gli archi del grafo riportano come peso l'intensità dell'interazione tra i geni che collegano. A questo peso si assegna il nome *fattore di efficienza regolatoria* e gli si attribuisce valore +1 o -1 a seconda del carattere di attivazione/induzione o di repressione/inibizione dell'interazione tra i geni.

L'argomento della presente ricerca è stato centrato sulla verifica del fattore di efficienza regolatoria mediante un'analisi sulla similarità di serie temporali biologiche rilevate per geni o proteine che siano direttamente collegate nei signaling pathways cellulari o metabolici. Questa analisi di similarità è fondata su tre differenti strategie di confronto, in cui vengono prese in considerazione alcune caratteristiche significative selezionate da questi brevi segnali, il cui paragone è reso particolarmente difficile dall'esiguità degli elementi stessi della serie temporale. In particolare le caratteristiche sono sia di tipo qualitativo-formale, legate alla forma del segnale, con particolare attenzione al time warping, sia di tipo quantitativo, come una valutazione della correlazione tra componenti spettrali

dominanti. Mediante un modello ibrido, le serie temporali sono state valutate secondo pesi opportunamente scelti sulla base di sperimentazioni condotte indipendentemente Infine è stato elaborato uno score dal quale scaturiscono conclusioni sulla regolazione o inibizione da parte dei geni a monte verso quelli a valle coerenti con conoscenze biologiche già consolidate. Le simulazioni, su serie temporali generate da microarray su un esperimento di somministrazione farmacologica mirata a cellule tumorali, sono state condotte in ambiente R [SIT16], un software opensource specifico per l'analisi statistica dei dati. Come supporto per la topologia dei pathway è stata ampiamente adottata la Kyoto Encyclopedia of Genes and Genomes (KEGG)[SIT24], attualmente uno dei massimi database mondiali nel campo della genomica e la libreria SPIA di R (Signaling Pathways Impact Analysis) [SIT25], in cui sono codificate tutte le informazioni relative all'interazione tra geni in specifici percorsi di segnalazione cellulare.

1. Le serie temporali

1.1 Introduzione

L'analisi di serie temporali, considerate come branca specifica della più vasta *Intelligent Data Analysis,* costituisce un rilevante e consolidato argomento di ricerca applicata, occupando aree di indagine diverse come la statistica, l'econometria, la teoria dei sistemi dinamici, la teoria del caos, la *Knowledge Discovery*. Nel contesto della Knowledge Discovery lo studio delle serie temporali permette di:

- trovare una più efficace rappresentazione dei dati (fitting);
- eseguire misure di similarità tra serie temporali;
- filtrare e analizzare i processi stocastici descritti dai dati;
- classificare i dati, ad esempio per mezzo di algoritmi non supervisionati (clustering).
- prevedere l'evoluzione futura dei fenomeni.

Il rilievo ad intervalli controllati di tempo dei valori registrati da molteplici strumentazioni durante l'evoluzione dinamica di un fenomeno di qualsiasi natura produce una stringa di valori numerici osservati sequenzialmente, eventualmente anche rappresentabile su un grafico per evidenziarne l'andamento globale. L'origine stessa delle serie temporali conduce a ipotizzare che un dato rilevato in un certo istante t sia più simile a quello rilevato all'istante t-1 che ad altri dati rilevati in momenti precedenti; in questo senso, si può ritenere che le serie temporali abbiano "memoria di sé". Questa proprietà è generalmente indicata col nome di *persistenza* e differenzia fortemente i campioni di serie temporali da quelle cross-section, perché nei primi l'ordine dei dati ha un'importanza fondamentale, mentre nei secondi esso è del tutto irrilevante, in quanto si tratta di dati registrati in un singolo istante o periodo di tempo e riferiti a molti differenti individui indipendenti tra loro.

Nelle serie temporali i dati rilevati in sequenza ordinata sono considerati come entità dipendenti, in senso comunque causale. L'analisi di serie temporali è lo studio della natura di questa dipendenza, volto ad individuare modelli stocastici e dinamici che razionalizzino i dati raccolti e diano ad essi un senso compiuto, ne traggano informazione più rilevante e produttiva. Questo insieme di dati, in una visione pratica, può essere considerato come un insieme campionario.

La teoria sviluppata sulle serie temporali è volta alla individuazione di metodi per l'interpretazione dei dati al fine di estrarne informazione significativa sia riguardo allo stato del sistema da cui è generata, sia riguardo alla previsione sulla dinamica temporale del fenomeno stesso, o di serie di durata molto maggiore del periodo di osservazione. I risultati dell'inferenza statistica in particolare, però, si applicano soprattutto a dati che devono essere indipendenti e non connessi temporalmente; per questa ragione sono stati ricercati strumenti diversi, i processi stocastici, adatti all'esecuzione dell'analisi sulle serie temporali; per meglio dire, lo strumento utilizzato per far fronte all'esigenza di trovare una metafora probabilistica per le serie temporali osservate è per l'appunto il processo stocastico.

È bene chiarire la distinzione tra serie temporale, processo e modello [Claps P., 2002]:

- *La serie temporale è una collezione di numeri reali, ordinati secondo la variabile tempo, la quale costituisce una parte finita di una realizzazione di un processo stocastico.*
- *Un processo stocastico a parametro discreto è caratterizzato dalla sua famiglia delle ripartizioni finite, cioè dalla conoscenza di tutte le possibili funzioni di ripartizione di probabilità.*
- *Un modello stocastico è una parametrizzazione di un processo in termini di una funzione esplicita di parametri noti. Mentre un processo stocastico è noto oppure non lo è, un modello può essere stimato a partire dai dati, ovvero dalla serie temporale osservata.*

Generalizzando, tra le tante possibilità, i campi di applicazione più interessanti dell'analisi di serie temporali possono essere considerati i seguenti [Box G. et al, 2008]:

1. la *previsione* di evoluzioni future dei fenomeni, basata su valori della serie temporale rilevati in istanti precedenti;
2. Date due serie temporali distinte, in alcuni casi è possibile la determinazione di una *funzione di trasferimento* di un sistema dinamico che evidenzi la relazione I/O tra le due serie temporali.
3. L'analisi degli *effetti di eventi eccezionali* su un sistema (Box-Tiao intervention model).
4. L'esame delle relazioni tra diverse serie temporali prodotte da un unico processo (serie multivariate), al fine di determinare opportuni modelli dinamici che rappresentino queste relazioni congiunte nel tempo.

5. il progetto di semplici schemi di controllo di processi, volto ad ottenere la convergenza delle serie verso un target predefinito con opportune correzioni sui valori in ingresso.

L'indagine statistica sulle serie temporali può essere suddivisa sostanzialmente in tre fasi: la prima di esse è l'identificazione del modello teorico cui segue la stima dei parametri del modello e l'ultima è il controllo di adattamento del modello ai dati (fitting).

La fase di identificazione del modello necessita non solo di un'approfondita conoscenza dei processi stocastici ma anche di molta perizia: l'opportunità di individuare un modello univoco, che è poi la finalità ultima della ricerca, si verifica solamente in casi particolari, ad esempio quelli in cui si hanno a disposizione alcune informazioni a priori sul fenomeno.

La tecnica impiegata per l'identificazione dei modelli nel metodo statistico classico consente di ridurne l'indeterminazione ad un numero limitato di modelli tra i quali si procede alla selezione di quello definitivo. Con la locuzione approccio statistico classico si vuole intendere un metodo di analisi nel quale sia la serie temporale ad orientare verso il modello e non viceversa (*series speaking for themselves*), evidenziando, così, che eventuali conoscenze a priori sulla serie temporale che si intende esaminare potrebbero portare all'identificazione di modelli non ottimali sotto il profilo della bontà di adattamento.

1.2 Definizioni

Dopo aver presentato nel paragrafo introduttivo i concetti di serie temporale e di processo stocastico, è opportuno fornire in sintesi una definizione più rigorosa degli stessi concetti, come è descritta in letteratura.

Definizione 1.1 Si chiama *serie temporale semplice* o *univariata*, e la si indica con $\{x_t\}$, $t \in \Omega$, un insieme ordinato di misure relative ad una caratteristica di un certo fenomeno. L'indice t ordina gli elementi dell'insieme e se appartiene ad un insieme Ω di numeri interi si dice che la serie è *discreta*, mentre se appartiene ad un insieme continuo di numeri reali si dice che la serie è *continua*.

Definizione 1.2 Una *serie temporale multipla* o *multivariata* (*k-variata*) è una collezione di più serie storiche $\{x_{1t}\}$, $\{x_{2t}\}$, …, $\{x_{rt}\}$, $t \in \Omega$, associate l'una all'altra e si indica con $\{\boldsymbol{x}_t\}$, $t \in \Omega$, dove

$$\boldsymbol{x}(t) = \begin{bmatrix} \boldsymbol{X}_1(t) \\ \boldsymbol{X}_2(t) \\ \vdots \\ \boldsymbol{X}_k(t) \end{bmatrix}$$

è un vettore di k dati tutti relativi al tempo t.

Per poter rendere questa definizione ancora più precisa, è necessario introdurre la teoria dei processi stocastici, da cui è derivata come caso particolare quella delle serie temporali,

Definizione 1.3 Uno *spazio di probabilità* è una terna $(\Omega; \mathcal{F}; \mathcal{P})$, dove Ω è un insieme qualunque, in genere pensato come l'insieme dei risultati possibili di un esperimento casuale, $\mathcal{F}$ è detta una σ-algebra, ovvero un insieme di insiemi (gli eventi) per i quali si può calcolare una probabilità,e $\mathcal{P}()$ è appunto una misura di probabilità su Ω ($P : \Omega \rightarrow [0, 1]$). Per la precisione, una σ-algebra è una famiglia di insiemi tali che:

- $0 \in \mathcal{F}$;
- se $A \in \mathcal{F}$ allora anche il suo complementare $\bar{A}$ è in $\mathcal{F}$;
- unioni numerabili di elementi di $\mathcal{F}$ appartengono ancora ad $\mathcal{F}$

Dato uno spazio di probabilità $(\Omega; \mathcal{F}; \mathcal{P})$, una *serie temporale* può essere considerata come una realizzazione finita di un processo stocastico.

Definizione 1.4 Dati uno spazio di probabilità $(\Omega; \mathcal{F}; \mathcal{P})$ e uno spazio parametrico T, che è un insieme di indici, un *processo stocastico* è definito come una

funzione X(t, ω) finita e a valori reali che, $\forall t \in T$, è una funzione misurabile di $\omega \in \Omega$ (cioè una variabile casuale).

In analogia a quanto riferito alle serie temporali, anche i processi stocastici possono essere univariati o multivariati.

Come già accennato sopra, ogni serie temporale rappresenta una parte finita di una particolare realizzazione di un processo stocastico e il suo andamento è uno solo tra i possibili infiniti tracciati che il processo può generare.

Si può fare inferenza sulle future realizzazioni di un processo solo stabilendo alcune sue proprietà ed ipotizzando una forma particolare di dipendenza temporale, per mezzo della quale dare forma ad un modello statistico per il processo.

Definizione 1.5 Un *processo stocastico a parametro discreto* è caratterizzato dalla sua famiglia delle ripartizioni finite, cioè dalla conoscenza di tutte le possibili funzioni di ripartizione

$$F(x_1, \ldots, x_n) = \Pr(X_{t_1} \leq x_1, \ldots, X_{t_k} \leq x_k)$$

per ogni valore di k e per ogni insieme di indici $\{t_1, \ldots, t_k\}$.

1.3 Varie tipologie di serie temporali

Esistono diversi criteri di classificazione delle varie tipologie di serie temporali. Le serie temporali sono classificabili in base al carattere *continuo* o *discreto* dell'insieme di definizione e dell'insieme dei valori assunti. Ogni serie storica è formata da elementi o valori, di cui quelli conosciuti sono anche detti dati od osservazioni. Le serie discrete, in particolare, sono generate di solito in uno dei seguenti modi:

- *estraendo dati da una serie continua ad intervalli regolari di tempo, come nel caso di temperature minime e massime osservate ogni 24 ore;*
- *come successioni di dati discreti per loro natura, come ad esempio la quotazione giornaliera di chiusura del petrolio;*
- *come successioni di quantità cumulate durante uguali intervalli di tempo, come ad esempio i centimetri di neve caduta in una certa località ed accumulata ogni ora.*

Esistono casi particolari in cui le misure di alcuni fenomeni non sono associate ad un solo indice t ma a più indici $t_1, t_2, \ldots, t_s$; le serie di dati in questo caso si chiamano *spaziali*. Come esempio di serie spaziale si può pensare ad un'onda sismica che si propaga nel tempo e nello spazio e che quindi è associata sia a un indice temporale t_1 che a due coordinate geografiche, ovvero t_2 e t_3.

Nel caso in cui due o più serie temporali siano collegate tra di loro in modo che alcune rappresentino un segnale in *ingresso* (input) in un sistema ed altre costituiscano segnali in *uscita* (output), è possibile ottenere i legami tra serie in ingresso e in uscita utilizzando relazioni funzionali tra le serie e controllare l'evolversi di quelle in uscita attraverso opportuni interventi effettuati sulla serie in ingresso.

L'analisi sistematica dell'evoluzione dinamica di una o più serie temporali e delle loro relazioni parte dalla considerazione di un modello che compendi le caratteristiche più importanti della serie. In particolare, le serie possono essere *deterministiche* o *stocastiche*, e quindi un modello additivo di serie temporale semplice, in cui si sommino la parte deterministica e quella stocastica, è il seguente:

$$\boldsymbol{x}_t = m(t) + \boldsymbol{y}_t \qquad\qquad t \in \Omega$$

Nella precedente espressione $m(t)$ è una funzione deterministica del tempo ed $\boldsymbol{y}_t$ è la parte stocastica.
L'analogo modello moltiplicativo è il seguente:

$$\boldsymbol{x}_t = m(t) \cdot \boldsymbol{y}_t \qquad\qquad t \in \Omega$$

Poiché l'indice t varia nell'insieme Ω le $\boldsymbol{x}_t$ e $\boldsymbol{y}_t$ sono variabili aleatorie appartenenti ai processi stocastici $\{\boldsymbol{x}_t\}$ ed $\{\boldsymbol{y}_t\}$ con $t \in \Omega$, che sono appunto insiemi di variabili aleatorie dipendenti da un parametro t variabile in un dato insieme Ω.

Per quanto fin qui detto, la relazione esistente tra serie temporale $\{x_t\}$, $t \in T$, e processo stocastico $\{\boldsymbol{x}_t\}$, $t \in \Omega$, diviene abbastanza chiara. La serie temporale $\{x_t\}$, $t \in T$ è semplicemente considerata come la parte nota, ovvero una realizzazione, del processo stocastico $\{\boldsymbol{x}_t\}$, $t \in \Omega$, in un certo intervallo di Ω. Costruire il modello a partire dalla serie temporale $\{x_t\}$ significa allora specificare il processo stocastico $\{\boldsymbol{x}_t\}$, $t \in \Omega$, di cui la serie è una parte nota, il che equivale a individuare la parte deterministica $m(t)$ e quella stocastica $\{\boldsymbol{y}_t\}$ con $t \in \Omega$.

Si parla di processo stocastico (e quindi di corrispondenti serie temporali) stazionario in due sensi: *stazionarietà forte* (anche detta *stretta*) e *stazionarietà debole*. Un processo è *stazionario in senso forte* se le caratteristiche di tutte le distribuzioni marginali rimangono costanti al passare del tempo, In altre parole, slittare gli istanti temporali di una quantità h, non altera la distribuzione congiunta. Invece un processo è *stazionario in senso debole* quando questa caratteristica vale solo per i momenti del primo e del secondo ordine per variabili aleatorie doppie.

Nonostante i nomi, la stazionarietà forte non implica quella debole; ad esempio, un processo può essere stazionario in senso forte ma non avere alcun momento; viceversa, la stazionarietà dei momenti non implica che le varie distribuzioni marginali abbiano lo stesso andamento al variare del tempo. Nel caso particolare di processo gaussiano, ossia quando la distribuzione congiunta di un qualunque sottoinsieme di elementi del processo è una normale multivariata, le due definizioni coincidono.

L'*ergodicità*, invece, esprime un concetto legato alla memoria di un processo: un processo non ergodico ha caratteristiche di persistenza così marcate da far sì che un

segmento del processo, per quanto lungo, sia insufficiente a dire alcunché sulle sue caratteristiche. Volendo riassumere quanto detto, la parte necessaria a conoscere un processo non ergodico coincide con tutto il processo.

In un processo ergodico, al contrario, la memoria del processo è debole su lunghi orizzonti e all'aumentare dell'ampiezza del campione aumenta in modo significativo anche l'informazione in nostro possesso.

In senso largo e intuitivo, si può dire che un processo è ergodico se eventi molto lontani fra loro possono essere considerati virtualmente indipendenti; se si osserva un processo ergodico per un intervallo di tempo abbastanza lungo, è possibile individuare quasi tutte le sottosequenze che il processo è in grado di generare. In altri termini, si può dire che, in un sistema ergodico, se qualcosa ha virtualità di accadere, allora prima o poi accadrà. Il fatto che eventi lontani fra loro nel tempo possano essere considerati indipendenti da un punto di vista pratico è poi spesso sintetizzato nella seguente proprietà dei processi ergodici (che a volte viene usata come definizione di processo ergodico):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathrm{Cov}(x_t, x_{t-k}) = 0$$

Di conseguenza, se un processo è ergodico, è possibile, almeno in linea di principio, utilizzare le informazioni del suo andamento nel tempo per inferirne le caratteristiche. Esiste un teorema, chiamato appunto 'teorema ergodico', che afferma che, se un processo è ergodico, l'osservazione di una sua realizzazione "abbastanza" lunga è equivalente, ai fini inferenziali, all'osservazione di un gran numero di realizzazioni.

In definitiva, si può concludere che considerare un'inferenza è ammissibile solo se il processo stocastico oggetto di studio è sia stazionario che ergodico. Va precisato, tra l'altro che, mentre esistono dei metodi per sottoporre a test l'ipotesi di non stazionarietà, l'ipotesi di ergodicità non è testabile se si dispone di una sola realizzazione del processo, quand'anche fosse di ampiezza infinita.

1.4 Esempi

Nella figura 1 inserita di seguito sono presentati alcuni caratteristici esempi di serie temporali ricavate da diverse aree d'interesse pratico. In particolare, di queste serie temporali è possibile eseguire una classificazione macroscopica a seconda del dominio e codominio continuo o discreto e del carattere deterministico o stocastico oppure della relazione esistente tra due o più serie temporali distinte.

- *Serie temporali deterministiche a tempo e valori continui:*

Figura 1.a1 e Figura 1.a2: Evoluzione del transitorio di un circuito elettrico con elementi conservativi. Si possono osservare l'andamento nel tempo della tensione misurata in Volt e della corrente misurata in milliAmpere rilevate ai capi di un condensatore durante il transitorio di chiusura dell'interruttore in un circuito RC.

- *Serie temporali stocastiche a tempo e valori continui:*

Figura 1b: Tracciato di un elettrocardiogramma di cuore durante un evento di fibrillazione atriale. Questo esempio rappresenta una serie stocastica a tempo e valori continui, relativa ad un elettrocardiogramma, in cui sono riportati i segnali bioelettrici continui nel tempo prodotti dal miocardio durante il ciclo cardiaco.

  - In Figura 1f sono rappresentate serie temporali in ingresso (1) e in uscita (2) e (3) da un sistema.

- *Serie temporali stocastiche a tempo e valori discreti:*

  - Figura 1c: Evoluzione per un intervallo di tempo di 50 anni, ad intervalli costanti di 10 anni, della quantità di popolazione suddivisa in base al titolo di studio conseguito, in una regione geografica prescelta.
  - Figura 1g: Evoluzione negli anni dal 1994 al 2011 del numero di studenti laureati quinquennali, di primo livello e della laurea specialistica dopo l'ultima riforma universitaria nell'Università di Salerno.
  - Figura 1h: Evoluzione su base giornaliera, rilevata ad intervalli regolari di due ore, del valore dell'indice FTSE MIB di venerdì 4 novembre 2011.

- *Serie temporale stocastica a tempo discreto e valori continui:*

  - Figura 1d: Evoluzione del prezzo dell'oro in dollari USA, rilevato giornalmente alla chiusura della borsa di New York, nel periodo luglio 1999-luglio2009.

- *Serie temporale stocastica a tempo continuo e valori discreti:*

  - In Figura 1e: Serie temporale stocastica a tempo continuo e valori discreti relativa a numero di particelle emesse da una sostanza radioattiva e registrate su modulo continuo da un contatore Geiger.

- *Relazione ingresso/uscita di serie temporali continue::*

  - Figura 1f: Serie temporali che rappresentano segnali in ingresso (1) e in uscita (2) e (3) da un sistema elettrico.

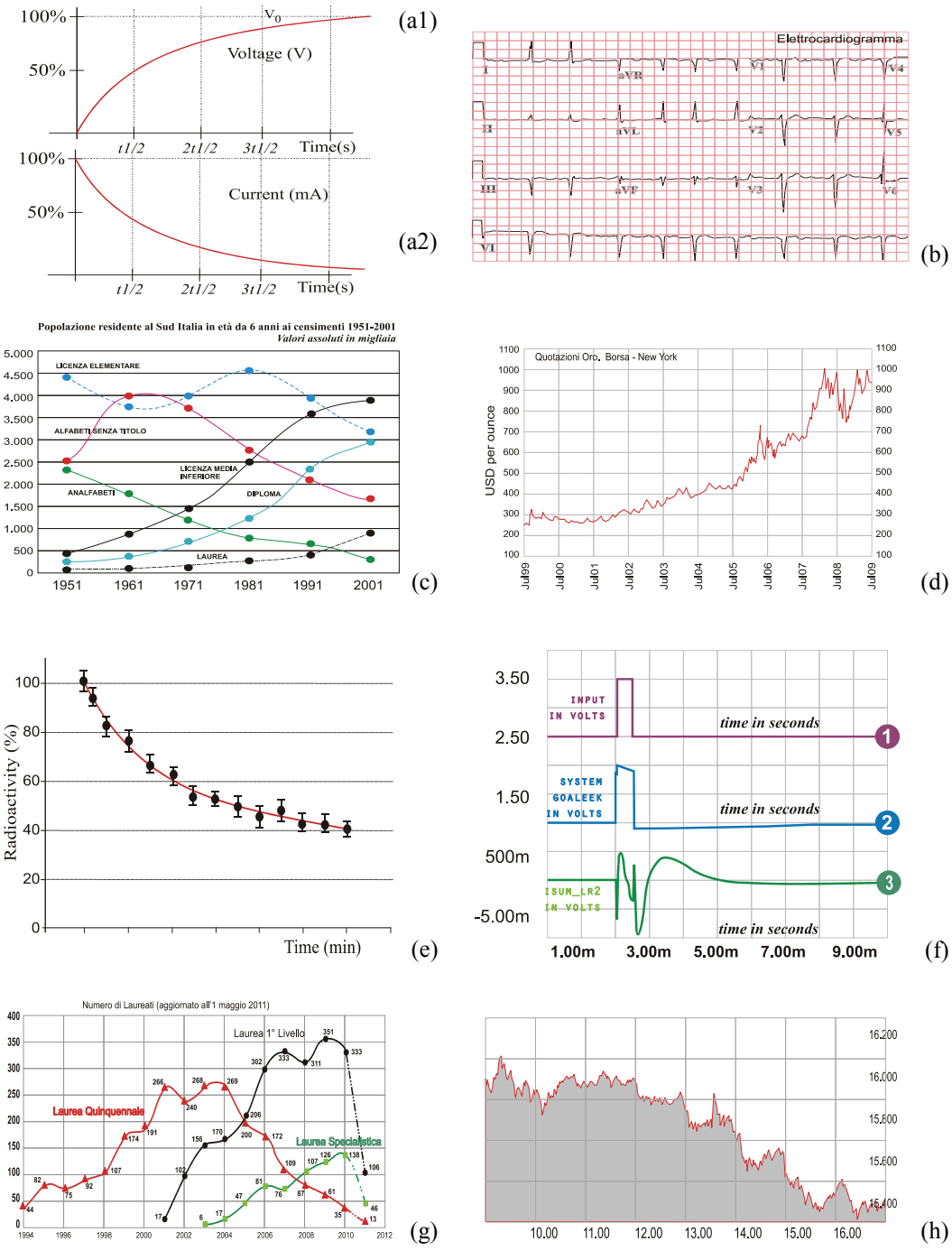


Figura 1: Alcuni rappresentativi esempi di Serie Temporali

1.5 Metodi e Tecniche

Le metodologie di analisi di serie temporali possono essere sinteticamente suddivise in due grandi gruppi: metodi nel *dominio della frequenza* e metodi nel *dominio del tempo*. Al primo gruppo appartiene la tradizionale analisi spettrale, realizzata nediante trasformazione di Fourier e la più recente analisi mediante trasformata wavelet, mentre il secondo gruppo comprende analisi di tipo statistico di auto e mutua correlazione tra le serie.

Quasi tutte le misure sono realizzate nel dominio del tempo, come ad esempio quelle che danno origine ai segnali in tempo continuo e nel nome stesso delle serie temporali è intrinseca la loro natura strettamente legata al dominio del tempo. Tuttavia, sottoponendo le serie temporali a una trasformazione di Fourier si può introdurre un'analisi nel dominio della frequenza, detta anche *analisi spettrale*, mediante la quale si può pervenire all'identificazione della funzione di trasferimento del sistema che ha generato serie temporale in oggetto. Una delle caratteristiche distintive delle tecniche di analisi nel dominio della frequenza consiste nel fatto che la modellizzazione dei sistemi a tempo continuo può avvenire isolando i segnali dal rumore, grosso vantaggio rispetto alle metodologie nel dominio del tempo.

Tuttavia l'analisi spettrale può presentare alcuni drawback: non è adatta a funzioni con scarsa oscillazione; un valore locale della funzione influenza tutti i coefficienti della trasformata o della serie, se sono presenti errori locali nel dominio del tempo essi si spalmano su tutte le frequenze; usando la trasformata di Fourier si perde la visione della funzione nel dominio del tempo, l'analisi spettrale non si adatta allo studio di problemi non lineari poiché piccole variazioni nell'input possono causare grandi variazioni nell'output; nell'analisi di Fourier ogni istante di un segnale è equivalente a qualsiasi altro, perché mancano informazioni temporali sulle frequenze di ciascun segnale e quindi si ignora quando esse si verificano.

Una funzione può essere rappresentata contemporaneamente nei due domini, del tempo e della frequenza, mediante un metodo grafico, la cosiddetta "Windowed Fourier Transform": si dividono gli assi della frequenza e del tempo in segmenti di uguale lunghezza. Nel piano (t,ω) essi formano una griglia di lati $\Delta\omega \cdot \Delta t$. Il rettangolo di lati generici $[\omega_j; \omega_{j+1}]$ e contiene l'approssimazione tra le frequenze ω_j e ω_{j+1} della funzione in $[t_j ; t_{j+1}]$. La dimensione della finestra è fissa, invece il numero di oscillazioni è variabile nel tempo. Quindi una siffatta finestra può essere troppo ridotta per le basse frequenze o troppo estesa per alte frequenze.

Le wavelet, sia nel continuo che nel discreto, sono uno strumento relativamente nuovo di analisi nel dominio della frequenza, che evidenzia le variazioni del segnale, rappresentato dalla serie temporale, intorno a medie locali, servendosi di funzioni elementari, dette appunto wavelet. L'idea originale alla base delle wavelet è quella di un'analisi che viene accordata alla scala dei tempi. Le wavelet sono funzioni che soddisfano alcuni requisiti matematici e sono utilizzate per rappresentare dati o altre

funzioni. Ma non è questa l'idea originale, in quanto si rifà al concetto di base della trasformazione secondo Fourier. Tuttavia, gli algoritmi wavelet processano i dati a diverse scale e risoluzioni.

Osservando un segnale attraverso una finestra ampia si possono cogliere le sue caratteristiche macroscopiche. Allo stesso modo, con una finestra stretta, si notano i particolari, come quando si fa uno zoom. Il risultato dell'analisi di serie temporali mediante wavelet è quello di riuscire ad avere una visione d'insieme e contemporaneamente dettagliata. L'utilizzo delle funzioni wavelet, che sono contenute in domini finiti, risolve il problema della "non-località" delle funzioni goniometriche le cui combinazioni sono utilizzate per approssimare altre funzioni mediante serie e trasformate di Fourier. In questo modo diventa possibile rappresentare segnali che variano molto velocemente nel tempo senza difficoltà. La procedura di analisi mediante wavelet prevede l'adozione di una funzione wavelet prototipo, chiamata wavelet madre. L'analisi temporale è eseguita con una versione contratta ad alta frequenza della wavelet prototipo, mentre l'analisi in frequenza viene effettuata con una versione dilatata, a bassa frequenza della wavelet stessa. In questo modo la serie originale può essere rappresentata mediante una combinazione lineare di wavelet i cui coefficienti si possono considerare come le uniche informazioni indispensabili. Scegliendo per rappresentare i dati la wavelet madre che meglio vi si adatta, oppure troncando opportunamente i coefficienti secondo un criterio stabilito da una soglia, si può ottenere una rappresentazione compressa dei dati. Rispetto all'analisi di Fourier le wavelets offrono i seguenti vantaggi: sono molto più adattabili ai vari tipi di funzione; errori locali nel calcolo della funzione comportano solo errori locali nelle wavelets; all' aumentare della frequenza la finestra t-ω automaticamente rimpicciolisce, catturando i minimi dettagli, mentre al diminuire della frequenza si estende; la caratteristica delle funzioni più semplici, ottenute dalla scomposizione di una funzione con l'analisi wavelet, è quella di avere una scala progressivamente minore. Esse cioè recano un livello di dettaglio progressivamente crescente, da una visione d'insieme ad un effetto zoom, in modo da carpire ogni informazione. Come drawback sostanziale, le wavelet presentano una maggiore complessità della trattazione matematica.

Volendo effettuare un paragone tra le due principali metodologie di analisi nel dominio della frequenza, senza scendere in eccessivo dettaglio e confrontando le wavelet, funzioni su cui sono costruite le trasformate omonime, con le sinusoidi, che sono le funzioni alla base dell'analisi di Fourier, emergono chiaramente le seguenti differenze: la sinusoide è una funzione periodica e il suo dominio si estende su tutto l'asse dei tempi, da meno infinito a più infinito, e pertanto è uniforme e simmetrica; tutte le wavelet di base, invece, sono caratterizzate dall'essere temporalmente finite e tendono ad essere irregolari e asimmetriche. Queste brevi note sulle caratteristiche lasciano intuire la maggiore duttilità delle wavelet nella rappresentazione di funzioni qualsiasi del tempo rispetto alle sinusoidi e la possibilità di conservare informazioni sul tempo.

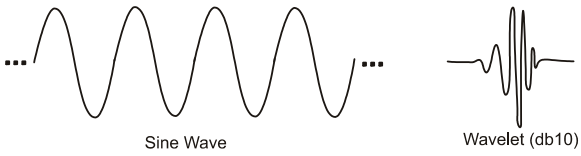


Figura 2: Confronto tra funzioni base della trasformata di Fourier e della trasformata wavelet

L'analisi di Fourier consiste nel decomporre un segnale in sinusoidi di diverse frequenze. Similarmente l'analisi wavelet decompone il segnale in versioni shiftate e scalate della wavelet originale (o onda madre). Osservando la figura si può immediatamente comprendere come i segnali che contengono irregolarità possono essere analizzati meglio facendo ricorso ad una wavelet di forma irregolare piuttosto che ad una sinusoide regolare.

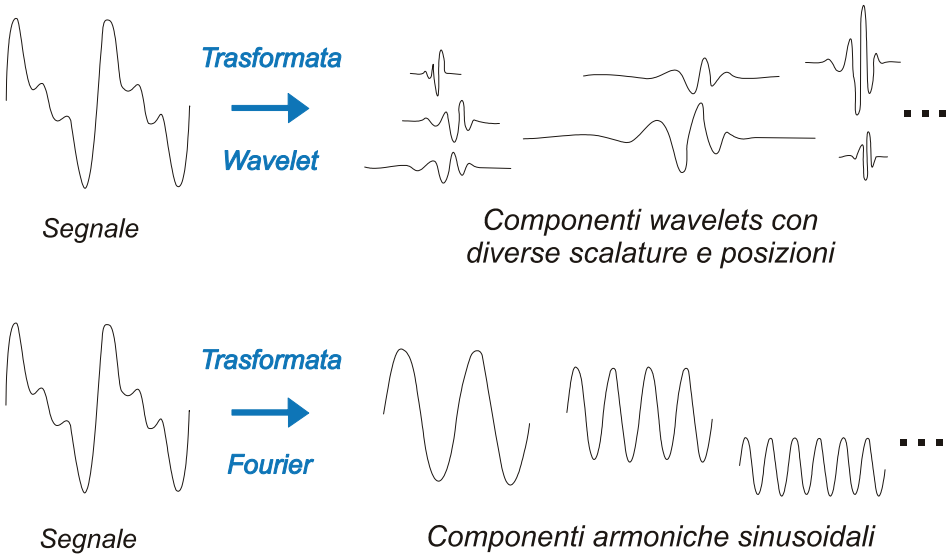


Figura 3: Confronto qualitativo tra componenti wavelet e componenti elementari delle trasformate di Fourier

La funzione di *autocorrelazione*, che deriva direttamente da quella di auto covarianza, fornisce un'indicazione sul legame di dipendenza che esiste tra un dato di una serie temporale, rilevato ad un certo istante t, con un altro dato della stessa serie rilevato in un istante di tempo precedente t-h. In altri termini l'autocorrelazione di una serie temporale discreta è semplicemente la correlazione di questa con una sua versione traslata all'indietro nel tempo. L'analisi dell'andamento della funzione di autocorrelazione al variare di questo h, ovvero lo studio dell'autocorrelazione a (t–1), (t–2), (t–3),... e così via, è di fondamentale importanza per valutare la cosiddetta "memoria" della serie storica, ovvero quanto a lungo permane un'influenza di una particolare realizzazione (osservazione) di detta serie storica sulle realizzazioni seguenti. L'analisi della funzione di autocorrelazione è ampiamente impiegata come strumento decisionale nella costruzione dei modelli, noti dalla teoria dei processi stocastici, autoregressivi e a media mobile (ARMA), i quali pongono la serie temporale nell'ottica di "traiettoria di un processo stocastico". In pratica l'autocorrelazione è utile

per individuare una ripetizione di pattern in una serie temporale, per scoprire una periodicità nascosta sotto il rumore oppure per identificare la frequenza fondamentale di un segnale.

La funzione di autocorrelazione può essere applicata per caratterizzare la struttura di una sequenza di valori nel dominio del tempo. La *mutua correlazione*, invece, è essenzialmente lo stesso processo, ma invece di confrontare una sequenza con una sua versione traslata nel tempo, essa confronta due serie temporali distinte, e in questo modo ne effettua un confronto statistico. La mutua correlazione individua le varie componenti di frequenza che le serie hanno in comune e la relativa differenza di fase. Da un punto di vista pratico, un'analisi di mutua correlazione è particolarmente indicata per evidenziare se una serie sia la versione ritardata dell'altra, perché in tal caso il suo valore è nullo per tutti gli intervalli di ritardo considerabili ed è uguale a uno quando l'intervallo di traslazione tra le serie considerate è uguale al ritardo originariamente esistente tra le serie.

A questa premessa, volutamente sintetica e qualitativa, segue una breve descrizione delle serie temporali che hanno costituito l'oggetto effettivo degli studi esposti nella presente tesi.

1.6 Casi di studio

1.6.1 Serie temporali generate da Global Positioning System (GPS)

Le serie temporali che vengono generate come file di uscita da dispositivi di tipo Global Positioning System sono multivariate, stocastiche e discrete. Esse sono costituite da una sequenza di record, uno per ogni istante di campionamento, raggruppati in file di formato GPX. I singoli record sono formati da una molteplicità di campi che memorizzano sia metadati spazio-temporali, quindi parametri caratteristici dei rilievi cinematici, sia metadati di controllo, ovvero parametri che si riferiscono alla qualità dei dati stessi, come ad esempio la Dilution of precision (DOP) della posizione, della posizione orizzontale e verticale e dell'informazione geometrica globale.

1.6.2 Serie temporali generate da esperimenti microarray

I microarray sono strumenti di una recente tecnologia in grado di evidenziare, tra altri risultati, i livelli di espressione dei geni di un determinato organismo biologico. Anche se si tratta di esperimenti molto complessi e facilmente affetti da rumore, in opportune condizioni di accuratezza essi possono essere applicati ripetutamente, ad intervalli di tempo prefissati, per studiare l'evoluzione nel tempo di fenomeni biologici che si verificano, ad esempio, in risposta alla somministrazione di sostanze farmacologiche o in seguito ad eventi di trasmissione interna di segnali tra cellule o altri processi metabolici spontanei. In questo modo è possibile raccogliere dati, che sono appunto la sequenza di valori di espressione dei geni coinvolti in uno o più signaling pathway cellulari. Ognuna delle serie temporali esaminate è quindi riferita a un singolo gene, è discreta, stocastica e univariata.

2. Serie temporali da GPS

2.1 I dati GPS

Il Global Positioning System (GPS) o più precisamente il sistema NAVSTAR-GPS (NAVigation System for Timing and Ranging –Global Positioning System), è un GNSS (Global Navigation Satellite System) nato negli Stati Uniti negli anni Settanta a scopo militare, ma da alcuni anni è stata concessa un'estensione dell'utilizzo anche per impieghi civili.

Il segmento spaziale del sistema consiste in una costellazione di 28 satelliti orbitanti ad una quota di circa 20.180 Km con un periodo orbitale di 11h 58m 02s, corrispondenti a 12 ore sideree. I satelliti percorrono due orbite complete in poco meno di un giorno solare, in modo da passare per un punto della Terra con quasi quattro minuti di anticipo ogni giorno. I parametri orbitali scelti rendono possibile che, in ogni istante e in ogni luogo, nell'ipotesi di assenza di ostacoli, siano visibili almeno quattro satelliti. I satelliti sono distribuiti su 6 orbite distanti fra loro di un angolo di 60° e formanti un angolo di 55° rispetto al piano equatoriale [Zogg, 2007]. Ogni satellite consente la localizzazione del ricevitore nelle tre coordinate spaziali (latitudine, longitudine e altitudine), la misura del tempo UTC (Universal Coordinated Time) e di altre grandezze quali la velocità, con una copertura globale e continua.

L'interesse per il GPS da parte della comunità scientifica e gli studi condotti hanno permesso agli apparati riceventi di impiegarlo con una precisione considerevolmente maggiore di quella prevista in origine dai progettisti. Grazie alle caratteristiche di precisione ed affidabilità, le applicazioni pratiche nel settore civile sono diventate numerosissime anche al di fuori dei campi tradizionali della navigazione marittima ed aerea. A titolo di esempio si evidenziano:

- *georeferenziazione di automezzi terrestri e di animali o persone*: un ricevitore GPS riesce a determinare con continuità la posizione di veicoli e questo dato può essere trasmesso via radio ad una centrale di controllo in tempo reale.
- *navigazione*: è l'utilizzo più popolare dei GPS. I navigatori satellitari valutano gli itinerari da seguire e sono capaci di memorizzare e riproporre informazioni disparate sui luoghi che si trovano lungo il percorso.
- *misurazioni geodetiche, geofisiche, idrografiche e cartografiche*: l'introduzione del DGPS o GPS differenziale ha consentito di raggiungere precisioni così elevate da poterlo impiegare come strumento di misura per tale tipo di applicazioni.
- *sincronizzazione*: la precisione del segnale di sincronizzazione del GPS si aggira intorno a incertezze di ±10 ns [Lombardi, M. A, 2001] ed è minore solo della precisione degli orologi atomici, su cui il GPS è basato. La sincronizzazione di sistemi di misura remoti può avvenire, ad esempio, tramite GPS.
- *telefonia cellulare*: la sincronizzazione mediante GPS permette una corretta comunicazione tra celle e supporta il rilevamento di posizione ibrido GPS/cellulare per le chiamate di emergenza.
- *geotagging*: alcuni dispositivi digitali dotati di GPS associano a foto o altri documenti informazioni sulle coordinate geodetiche, il che può essere utile in applicazioni diverse.
- *timestamping*: utilizzando dati provenienti da GPS quale riferimento temporale ed eseguendo una correzione della misura per compensare l'errore intrinseco nel GPS , dovuto alla combinazione dell'effetto Doppler, multipath e ritardi d'origine ionosferica ed atmosferica, si può effettuare il time stamping di un evento.
- *misura di fasori*: il GPS rende possibile un timestamping molto accurato dei sistemi di misura di potenza e in questo modo si presta alla delicata misura dei fasori.
- *robotica*: robot autonomi dotati di sensori GPS, che calcolano latitudine, longitudine, tempo, velocità e direzione.
- *tempo libero*: si sono diffusi hobby legati alla possibilità di memorizzare tracciati rilevati mediante GPS come geocaching, geodashing, GPS drawing and waymarking.
- *agrimensura*: l'utilizzo di locazioni assolute è adatto per disegnare mappe e determinare i confini di una proprietà.
- *tettonica*: tramite GPS è possibile effettuare misure dirette di movimenti delle faglie nei terremoti.
- *telematica*: la tecnologia GPS viene integrata con quella dei PC e dei dispositivi mobili di comunicazione in sistemi di navigazione dedicati.

2.1.1 Moduli del GPS

Il GPS è costituito da tre moduli:

- Il *segmento spaziale*, formato dalla costellazione satellitare GPS orbitante intorno alla Terra. Ogni satellite pesa circa 800 kg ed è alimentato con pannelli solari [Kaplan et al 2006]. Ogni satellite si muove su un'orbita non geostazionaria con una velocità circa di 4.000 m/s. La costellazione di satelliti è stata progettata mirando a garantire che qualsiasi utente possa contare sulla presenza sopra l'orizzonte di almeno 4 satelliti che inviano i dati necessari al posizionamento sulla terra e quindi ogni satellite è visibile da un punto per 5 delle 12 ore di ciascun giro intorno alla terra. Ciascun satellite ha a bordo quattro orologi atomici (due al cesio e due al rubidio) che hanno una stabilità media pari a una parte su 1012; ciò implica un ritardo di un secondo ogni 317.000 anni circa. Gli orologi atomici sono il riferimento per la generazione dei segnali in trasmissione, poiché caratterizzati da un'oscillazione di frequenza base di 10,23 MHz. Su ciascun satellite sono alloggiati motori per le correzioni orbitali e sistemi giroscopici per la stabilità.
- Il segmento di *controllo* verifica lo stato di funzionamento dei satelliti e aggiorna le relative orbite ed il funzionamento degli orologi [Kaplan et. al 2006]. Il controllo è realizzato da un insieme di stazioni GPS permanenti poste lungo l'equatore, in modo tale che il segnale di un qualunque satellite sia ricevuto da almeno una di queste stazioni. Le stazioni hanno diversi compiti e sono così suddivise:
  - cinque *Monitor Station* (*MS*) per il controllo dei satelliti, situate a Colorado Springs (Colorado,Stati Uniti), Isole Hawaii e Isola Kwajalein (Oceano Pacifico), Isola di Ascensiòn (Oceano Atlantico) ed Isola Diego Garcia (Oceano Indiano);
  - tre *Ground Antenna* (*GA*) per la trasmissione in banda verso i satelliti dei comandi di controllo e delle informazioni da inserire nei messaggi; esse sono collocate insieme alle Monitor Station delle isole di Ascensiòn, Diego Garcia e Kwajalein;
  - una *Master Control Station* (MCS) situata presso la base aerea Shriever a Colorado Springs (Colorado,Stati Uniti), che è il punto centrale di tutto il segmento di controllo; questa stazione elabora le informazioni giunte da tutti i satelliti attraverso le 5 stazioni di monitoraggio, mette a punto le correzioni necessarie per ogni satellite e comanda la trasmissione attraverso le 3 stazioni di controllo. Per poter effettuare correzioni con grande precisione la MSC è dotata di una serie di orologi atomici estremamente precisi ai quali fanno capo tutti gli altri orologi, sia a terra che a bordo dei satelliti.
  - Il segmento utente, è rappresentato da civili e militari che si avvalgono del servizio GPS; un ricevitore GPS demodula i segnali emessi dai satelliti GPS per stimare in tempo reale la propria posizione.

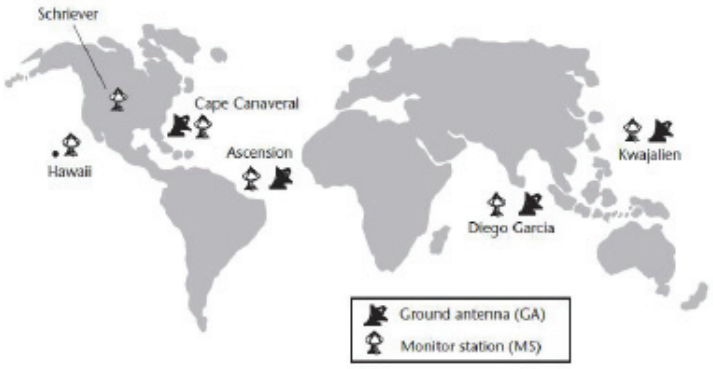


Figura 4: Posizione delle Monitor Station e delle Ground Station sul globo terrestre

2.1.2 Segnale GPS

Tutti i satelliti trasmettono segnali modulati su due diverse portanti nella banda L, entrambe multiple della frequenza fondamentale f_0=10,23 MHz, generata dagli oscillatori installati sui satelliti; indicando con L_1 ed L_2 le 2 portanti e con f_1 ed f_2 le due frequenze corrispondenti, si ha:

$$f_1=154* f_0=1.575,42 \text{ MHz}; \qquad f_2=120* f_0=1227,60\text{MHz}.$$

Le portanti L_1 ed L_2 sono modulate dai seguenti segnali:

1. **P code** (Precision): si tratta di una sequenza di bit di frequenza di 10,23 MHz che si ripete periodicamente dopo circa 38 settimane. L'intera sequenza non è mai trasmessa ma è suddivisa in segmenti lunghi una settimana e rigenerati ogni inizio settimana, comunicati ai satelliti e alle 5 stazioni di controllo. Un segmento rimane inutilizzato [Kaplan et al, 2006]. Il codice P modula entrambe le portanti e consente di raggiungere la massima precisione nella procedura di posizionamento. Su questo codice si basa il PPS (Precise Positioning Service) che fornisce un elevato grado di precisione nel posizionamento assoluto con un accuratezza di 18m. Questo servizio è accessibile solo ai militari, in quanto criptato.
2. **C/A code** (Clear Access or Corse Acquisition): è una sequenza di 1023 bit con frequenza 1,023 Mbps ed un periodo di ripetizione pari a 1 ms; modula soltanto la portante L_1 e risulta di più facile ricezione in quanto è più corto; questo codice è diverso per ciascun satellite, è utilizzato da tutti i ricevitori ma presenta una precisione molto bassa: su di esso si basa la modalità del segnale trasmesso denominata SPS (Standard Positioning Service) accessibile a tutti gli utenti.
3. **D code** (Navigation Data): contiene le informazioni sul posizionamento che il segmento di controllo trasmette all'utenza sfruttando il segnale GPS [Zogg et al., 2001].

I tre codici, come si è appena detto, sono sequenze di bit; il codice D trasporta messaggi informativi mentre i codici P e C/A sono sequenze pseudo casuali di bit [Zogg et al., 2001]. Le sequenze risultano periodiche di periodo 1 ms per il codice C/A e 7 giorni per il codice P. Pertanto i codici P e C/A vengono definiti codici di tipo PesudoRandom Noise (PRN); ogni satellite utilizza una propria sequenza PRN distintiva ed ortogonale alle altre; sono utilizzate 28 sequenze diverse numerate da 0 a 27.

Ogni satellite impiega le stesse frequenze trasmissive: lo schema d'accesso multiplo adottato è la tecnica CDMA (Code Division Multiple Access) che sfrutta l'ortogonalità dei codici sopra evidenziata. Tutti i satelliti utilizzano le stesse frequenze per trasmettere simultaneamente i codici C/A, P e D, per far ciò si utilizza la modulazione BPSK (Binary Phase Shift Keying). In particolare la portante L_1 è modulata dai codici C/A, P e D, invece la portante L_2 è modulata solo dai codici P e D.

2.1.3 Rilevamento della posizione

La tecnica della triangolazione permette di individuare la posizione degli oggetti sul sistema di coordinate. Il satellite invia il suo peculiare C/A code, sincronizzato con il GPS Time e si ripete periodicamente ogni millisecondo. Il ricevitore esegue la correlazione di questo segnale con il C/A code generato localmente, che dovrebbe risultare sincronizzato con il GPS Time, ma tale sincronismo non è perfetto a causa della inferiore qualità degli orologi utilizzati nei terminali.

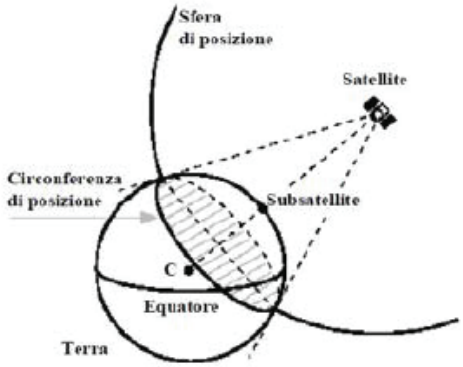


Figura 5: **Intersezione di una superficie sferica con la Terra**

Moltiplicando il ritardo τ ottenuto tramite la correlazione per la velocità di propagazione delle onde elettromagnetiche che è pari a quella della luce c, si ottiene così la distanza dal satellite *d,* detta anche pseudorange. Tramite questa distanza si traccia una sfera con il centro nella posizione occupata dal satellite nell'istante di emissione del segnale ed il raggio pari alla distanza calcolata; tale sfera interseca la superficie terrestre individuando una circonferenza, luogo dei punti in cui può trovarsi il ricevitore (vedi figura 5).

Due misure di distanza, disponibili utilizzando due satelliti, individuano due circonferenze che si intersecano in due punti di cui uno è certamente la posizione dell'osservatore; l'ambiguità fra i due punti può essere risolta con la posizione stimata del ricevitore; considerando come ulteriore incognita anche la quota, sono necessarie tre osservazioni che individuano tre sfere, la cui intersezione individua un volume entro il quale si trova il ricevitore.

I ricevitori terrestri sono dotati di orologi di precisione molto inferiore di quella degli orologi installati nei satelliti. Ciò può comportare una accuratezza grossolana nella misura della posizione, ma l'accuratezza può essere recuperata cercando di determinare l'errore rispetto al segnale di tempo fornito dai satelliti; l'errore dell'orologio

rappresenta pertanto un'ulteriore incognita che può essere determinata con una quarta osservazione (Fig.6); si risolve in definitiva un sistema di quattro equazioni con le quattro incognite: latitudine, longitudine, quota ed offset dell'orologio del ricevitore.

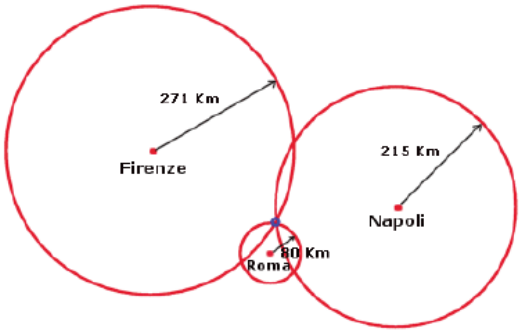


Figura 6: **Intersezione di tre sfere**

2.1.4 Errori del sistema

Il sistema GPS è soggetto a diversi tipi di errori, alcuni naturali, altri dovuti a limitazioni tecnologiche dell'accuratezza della misura; molti di essi possono essere ridotti dopo un'attenta valutazione delle cause e una serie di misure.
I principali errori sono:

Errori dipendenti dai satelliti, fra cui la Selective Availability e l'AntiSpoofing. Si tratta di dispositivi introdotti dal dipartimento della difesa degli USA per proteggere gli utilizzatori militari dal pericolo di essere ingannati da false trasmissioni e impedire ai civili e a potenze straniere ostili di usufruire al completo della capacità di localizzazione del sistema GPS. Dal 1° maggio 2000 il Ministero della Difesa americano ha disattivato tali dispositivi, consentendo ai ricevitori civili un incremento della precisione. Il processo consiste nell'introdurre una piccola alterazione nel funzionamento degli orologi dei satelliti. Inoltre, la posizione del satellite sul percorso riportata nelle effemeridi è trasmessa in modo approssimativo. Ne risulta la degradazione dell'accuratezza della posizione.

Errori di multipath: derivano soprattutto dalla combinazione dei segnali diretti con quelli riflessi dalle superfici circostanti, in particolare dalla superficie marina. Tali errori dipendono dalla natura e dalla localizzazione delle superfici riflettenti, per cui è possibile ridurli con un'opportuna disposizione e progettazione dell'antenna, ma anche adottando adeguate tecniche correttive nei ricevitori. Il segnale ricevuto è formato da una componente diretta e una riflessa, relativa ai percorsi multipli. Un piano di massa choke ring è realizzato mediante diversi anelli conduttori concentrici che circondano il centro di fase dell'antenna, consentendo di attenuare notevolmente la componente riflessa. Le stazioni fisse impiegano antenne di buona qualità, con bassa sensibilità per piccoli angoli di incidenza dei segnali o che abbiano un piano di massa choke ring.

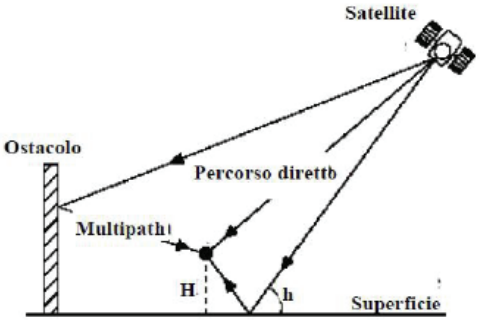


Figura 7: Errore di multipath

Errori dipendenti dal ricevitore: in ogni ricevitore esistono inevitabilmente errori dovuti al rumore interno, alla precisione con cui lavora il correlatore, ai ritardi introdotti dai vari sistemi elettronici e dal software di elaborazione dei dati. Le conseguenze di questi errori possono essere rilevanti nel caso di ricevitori utilizzati su veicoli molto veloci; una progettazione accurata del ricevitore in base all'uso cui è destinato può pertanto minimizzare questo tipo di errori.

Errori prodotti dalla propagazione dei segnali nella ionosfera e nella troposfera: nell'attraversamento di alcuni strati dell'atmosfera, le onde elettromagnetiche subiscono variazioni nella velocità e rifrazioni che creano un allungamento dei percorsi rispetto a quelli rettilinei fra i satelliti e il ricevitore. Per una completa eliminazione di questi errori bisognerebbe ricorrere all'utilizzo alla tecnica di ricezione in diversità oppure, eventualmente, utilizzare il codice P.

Errori introdotti dal Sistema di Controllo: il Manned Spacecraft Center (MSC) può commettere errori nella determinazione delle orbite o nelle correzioni degli orologi. L'errore sulla posizione dipende dagli errori appena detti e da un fattore scalare legato alla distribuzione spaziale dei satelliti utilizzati nelle misure; tale fattore è denominato fattore d'espansione dell'errore o PDOP (Position Dilution Of Precision). Nel caso della navigazione marittima o terrestre, interessa soltanto la posizione e quindi la precisione nel piano orizzontale, per cui il PDOP viene sostituito dall'HDOP (Horizontal DOP), che dipende solo dall'azimut dei satelliti. I ricevitori, utilizzando i dati contenuti nel messaggio di navigazione, riescono a calcolare preventivamente il PDOP o l'HDOP e a scegliere la combinazione migliore tra le diverse combinazioni di satelliti visibili e con il fattore DOP più basso. Dei parametri DOP si parlerà più diffusamente in appendice.

2.2 Compressione di segnali

La riduzione dello spazio di memoria occupato dai segnali digitali senza significative perdite di informazione è un obiettivo cui puntare per molti motivi. Allo stesso modo, per diminuire l'occupazione di banda necessaria in una trasmissione di segnali digitali è necessario applicare delle tecniche idonee a ridurre i bit che devono essere inviati nelle reti di comunicazione.

La compressione di un segnale, come più in generale la compressione di dati, consiste appunto nel ridurre il numero di bit necessari per rappresentare un segnale, rispettando dei vincoli di accuratezza. Molto spesso la compressione è applicata ai segnali digitali in combinazione con la quantizzazione in un passo chiamato codifica sorgente, operazione finalizzata a eliminare le ridondanze e ottenere un'alta efficienza del codice sorgente.

La compressione di segnale è una tecnologia chiave per le applicazioni multimediali sulla National Information Infrastructure, soprattutto in relazione alla trasmissione e allo storage di audio e video.

In pratica, la compressione di un segnale può essere concepita come una corrispondenza della rappresentazione dei dati digitali in esso contenuti da un gruppo di simboli ad un altro gruppo di simboli più conciso.

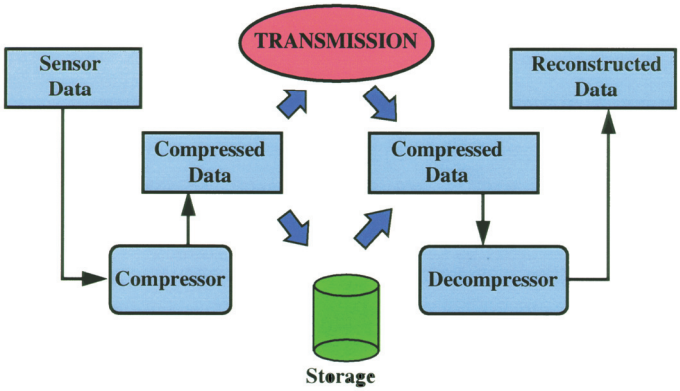


Figura 8: Schema di funzionamento della compressione di segnali

La compressione può essere quantificata attraverso il *quoziente di compressione*, cioè il quoziente tra il numero di bit del file originale e il numero di bit del file compresso. Il *tasso di compressione,* spesso utilizzato, è il reciproco del quoziente di compressione, abitualmente espresso in percentuale. Infine il *guadagno di compressione*, anch'esso espresso in percentuale, è il complemento a 1 del tasso di compressione.

Diversi sono i criteri in base ai quali è possibile classificare gli algoritmi di compressione [Blelloch G. E., 2000]: fisica o logica, con o senza perdite, simmetrica e asimmetrica.

La **compressione fisica** è quella praticata sui segnali in oggetto, sostituendo a una stringa di bit un'altra stringa di lunghezza inferiore trovata in base a certi criteri. La **compressione logica** invece si effettua con un ragionamento logico sostituendo un'informazione con un' altra informazione ad essa equivalente ma diversa.

Nella **compressione senza perdite (o lossless)** la compressione avviene in modo che il file originale possa essere perfettamente ricostruito dal file compresso con un procedimento di decompressione. Esempio classico di compressione senza perdite è l'alfabeto morse, che prevede un codice breve formato da linee e punti per i caratteri più frequenti e un codice più lungo per i caratteri più rari, cosicché in media sono necessari

meno bit che nella rappresentazione standard dei caratteri in codice ASCII, il quale assegna lo stesso numero di bit a tutte le lettere. La compressione senza perdite è l'unica praticabile per i file di testo, i software, i documenti, i database, gli schemi elettrici e così via.

I più noti algoritmi di compressione lossless sono basati sulla codifica aritmetica e in particolare sulla codifica adattiva di Huffman o sui metodi Lempel-Ziv (LZW). Un esempio di algoritmo di compressione senza perdite molto usato nella pratica è quello del formato ZIP, che permette l'archiviazione e la trasmissione di uno o più file contemporaneamente, risparmiando sulle risorse necessarie in termini di spazio su disco e di tempo di trasmissione. Dai file ZIP si ricavano per decompressione file indistinguibili da quelli originali.

Altro esempio è quello delle immagini non fotografiche come schemi, disegni e icone. Per questo tipo di file sono stati progettati formati di compressione come il GIF e il più recente PNG. L'immagine compressa con uno di questi formati presenta un aspetto identico all'originale fin nei dettagli importanti nella grande maggioranza delle applicazioni.

La **compressione con perdite** (o **lossy**) prevede che il file originale non possa essere perfettamente recuperato dal file compresso. La quantizzazione stessa è una forma di compressione con perdite. La maggioranza dei più comuni algoritmi di compressione con perdite effettua una sorta di trasformazione, o codice predittivo, dove il segnale viene preventivamente decomposto e poi quantizzato.
Decomprimendo un file compresso con perdita la copia ottenuta è peggiore dell'originale per livello di precisione delle informazioni che codifica, ma in genere comunque abbastanza simile da non implicare perdita di informazioni indispensabili. Difatti i metodi di compressione lossy in genere tendono a scartare le informazioni poco rilevanti, archiviando solo quelle essenziali: per esempio comprimendo un brano audio secondo la codifica dell'MP3 non vengono memorizzati i suoni non udibili, consentendo di ridurre le dimensioni dei file senza compromettere in modo sostanziale la qualità dell'informazione.

Il principale svantaggio della compressione con perdite è che il rapporto di compressione è limitato nei casi in cui non si riesca ad ottenere una accettabile ricostruzione del segnale d'origine. Una versione dello standard JPEG per la compressione di immagini ferme è un esempio di algoritmo di compressione con perdite: esso utilizza la trasformata discreta del coseno con quantizzazione scalare combinata con una codifica senza perdite. Con un rapporto di compressione di 16:1 si riesce ad ottenere una buona qualità dell'immagine.

Gli algoritmi di compressione con perdite sono adatti a trasmettere velocemente voci ed immagini e sono spesso elementi di algoritmi molto più complessi, come Mosaic, un browser che si proponeva di ricercare file multimediali sul Web. Grazie alla loro flessibilità, gli algoritmi di compressione con perdite consentono agli utenti di trovare un buon compromesso tra la qualità del segnale e la velocità di trasmissione mediante una trasmissione scalabile. Se non fosse presente questa caratteristica di flessibilità non sarebbe possibile neanche lo stesso browsing, e gli archivi di grosse dimensioni non potrebbero essere di nessuna utilità.

Va comunque evidenziata la particolare fragilità dei file compressi impiegati per l'invio di informazioni a distanza, in quanto anche minimi errori nella trasmissione possono causare alterazioni tali da rendere il file completamente inutilizzabile.

In caso di **compressione simmetrica**, si usa lo stesso metodo per comprimere e decomprimere l'informazione, c'è quindi la stessa quantità di lavoro per ciascuna operazione. Questo tipo di compressione si usa normalmente nella trasmissione dei dati.

La **compressione asimmetrica** richiede più lavoro per una delle due operazioni, si cercano algoritmi per cui la compressione sia più lenta rispetto alla decompressione o viceversa. Possono essere necessari degli algoritmi più rapidi in compressione che in decompressione quando si archiviano dati ai quali non si accede spesso, ad esempio per ragioni di sicurezza, dato che così si creano dei file molto compatti.

Gli algoritmi di compressione attualmente in uso sono fondati su idee che sono nate e si sono diffuse sia nella comunità scientifica che nelle aziende negli ultimi 20 anni. L'attenzione della comunità scientifica negli anni più recenti si è spostata dal problema di sviluppare nuove e più sofisticate tecniche a quello dello sviluppo di implementazioni sia hardware che software di tecniche che possano sostenere la sfida con il tempo ma anche al problema dello sviluppo e della determinazione di uno standard che garantisca la interoperabilità degli apparati utilizzati nella rete.

In seguito ai progressi raggiunti, gli standard dovranno essere modificati ed aggiornati. La scelta di standard di compressione avrà un impatto rilevante sui futuri sviluppi e sulla crescita di varie tecniche di compressione per segnali da trasmettere attraverso supporti con una limitazione di banda. Questa tipologia di standard ha già mostrato una profonda influenza sul processo di digitalizzazione nell'ambito dell'elettronica di largo consumo ed ha determinato la affermazione della HDTV che è attualmente molto diffusa. Nella codifica audio di alta qualità, come la compressione audio digitale di qualità CD, i progressi nei banchi di filtri, la codifica percettiva e il mascheramento di frequenze hanno condotto alla fine degli anni '80 fino a standard quale MUSICAM, algoritmo di compressione con perdite, che hanno permesso all'audio digitale in broadcast di soppiantare del tutto le trasmissioni radio in FM.

C'è un limite alla compressione, dato dalla quantità di informazione contenuta nel messaggio, chiamata in gergo tecnico entropia.

Nella teoria dell'informazione, il primo teorema di Shannon (o Teorema della codifica di sorgente), stabilisce dei limiti alla massima compressione possibile di un insieme di dati e definisce il significato operativo dell'entropia. Il teorema stabilisce che, per una serie di variabili aleatorie indipendenti ed identicamente distribuite (i.i.d.) di lunghezza che tende ad infinito, non è possibile comprimere i dati in un messaggio più corto dell'entropia totale, senza rischiare di perdere informazione. Al contrario, compressioni arbitrariamente vicine al valore di entropia sono possibili, con probabilità di perdita di informazione piccola a piacere. Il Teorema della codifica di sorgente per simboli di codice, stabilisce un limite inferiore e superiore alla minima lunghezza attesa di una serie di parole di codice, in funzione dell'entropia della parola in ingresso (vista come una variabile aleatoria) e della dimensione dell'alfabeto in esame.

A pagina seguente è rappresentata una tabella riassuntiva dei principali algoritmi di compressione.

Algorithms	Data encoding	Compression scheme	Compression ratio	Time complexity	Remarks
Static Huffman	Entropy	Static	Achieve good compression ratio in most cases.	N [log (n) + a] + Sn where N is the total number of input symbols, n is the current number of unique symbols, a is the arithmetic to be performed, and S is the time required, if necessary, to maintain internal data structures	· Symbol frequencies must be known in advance. · Symbol frequency table must be stored along with the compressed data. · Require two passes: one pass to compute the Symbol Frequency table and the other for compression. · For large files containing a large number of characters, building
Arithmetic coding	Entropy	Static	Achieve better compression ratio than Huffman.	N [log (n) + a] + Sn where N is the total number of input symbols, n is the current number of unique symbols, a is the arithmetic to be performed, and S is the time required, if necessary, to maintain internal data structures [9]	· Similar to Huffman, symbol frequencies must be known in advance. · Similar to Huffman, symbol frequency table must be stored/transmitted along with the compressed data which produces some overhead. · Complex to implement.
Adaptive Huffman	Dictionary	Adaptive	Achieve relative same compression ratio with Huffman.	N [n+log (2n – 1) + Sn where N is the total number of input symbols, n is the current number of unique symbols, and S is the time required, if necessary, to rebalance the tree [4]	· Require only one pass of the data. · Lose out in terms of speed. · No overhead of Symbol frequency table · Complex in implementing as the tree has to evolve itself
LZW	Dictionary	Adaptive	Relative good compression ratio for Text data. Other type of data varies.	O(n) where n is the number of symbols	· Table can get very large easily. · Require only one pass of the data. · Slow in speed as each time a new character is read in, the algorithm has to search for the new string formed by STRING+CHARACTER. · No overhead of Symbol frequency table.
LZ77	Dictionary	Adaptive	Compression ratio is not as good as others.	O(n) where n is the number of symbols	· Decompression is one of the fastest among all the other algorithms. · Easy to implement.
Lempel-Ziv-Markov chain-Algorithm (LZMA)	LZMA is an algorithm for data compression used in the 7z format of the 7-Zip application. It is an improved version of LZ77, using a dictionary compression scheme similar to LZ77 backed by arithmetic coding. It features a high compression ratio (generally higher than bzip2) and a variable compression-dictionary size (up to 4 GB). Currently, 7-Zip is providing LZMA as an open source SDK and is open to developers to use it to develop applications. The documentation in the SDK is very limited, providing no design information. However, developers can search on third-party discussion, e.g., its user forum for information.				
Deflate	Deflate uses a combination of the LZ77 algorithm and Huffman coding. The deflate compressor provides flexibility on how to compress the data. The programmer design smart algorithms to make compressor make the right choices about how to compress data. There are three modes of compression that the compressor has available: 1. Not compressed at all. This is an intelligent choice for, say, data that's already been compressed. 2. Compression, first with LZ77 and then with Huffman coding. The trees that are used to compress in this mode are defined by the Deflate specification itself, and so no extra space needs to be taken to store those trees. 3. Compression, first with LZ77 and then with Huffman coding with trees that the compressor creates and stores along with the data. The data is broken up in "blocks", and each block uses a single mode of compression. If the compressor wants to switch from uncompressed storage to compression with the trees defined by the specification, or to compression with specified.				
Bzip2	Bzip2 uses several layers of compression techniques such as Run-Length Encoding, Huffman Coding, Move-To-Front Encoding, etc stacked on top of each other, which occur in the following order during compression and the reverse order during decompression. One of the core algorithm used in bzip2 is Burrow-Wheeler Transform (BWT) which permutes the orders of the symbols so that it has runs of the same repeated symbols. The idea is to create more-easily-compressible data before applying the compression algorithms Holger Kruse and Amar Mukherjee [5] have described that by using BWT, text compression can be further increased.				

tabella 1: caratteristiche dei principali algoritmi di compressione (fonte WEB)

2.3 Compressione di serie temporali da GPS

Gli algoritmi di compressione sono orientati alla restrizione dello spazio di archiviazione occupato dai dati in generale e in particolare dei dati geospaziali, ma questa diminuzione di dimensione dei dati non deve incidere minimamente sulla qualità dell'informazione contenuta in essi oppure deve provocare una perdita limitata in un range prefissato.

Gli algoritmi di compressione per dati geospaziali di tipo senza perdite sono adottati necessariamente per applicazioni critiche, che richiedono dettagli di precisione.

Un esempio tipico di questa famiglia di algoritmi è quello di **Hatanaka**, [Hatanaka Y., 2008]. Questo algoritmo effettua una compressione preliminare dei dati, eliminandone le ridondanze; poi registra le variazioni del segnale negli istanti di tempo adiacenti e opera un troncamento dei dati riferendosi alle differenze di ordine n-esimo dei dati stessi. Lo schema di funzionamento del calcolo delle differenze è presentato nella figura seguente:

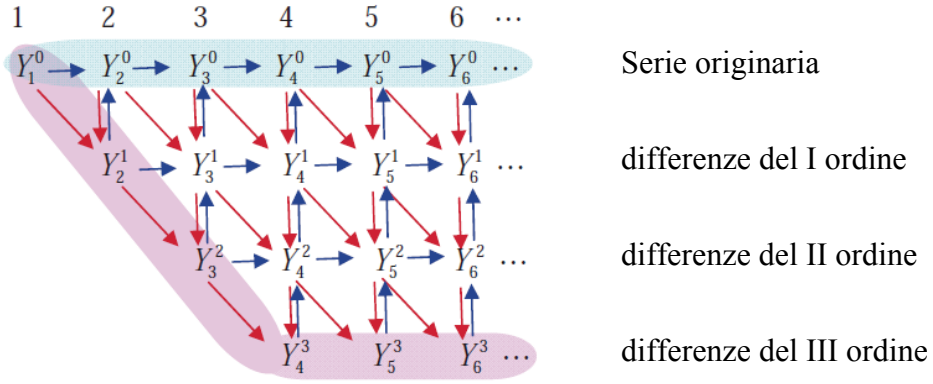


Figura 9: **Schema del processo di calcolo delle differenze (frecce rosse) e del processo di recupero dell'informazione (frecce blu) nel caso di differenze fino al terzo ordine relativo all'algoritmo di Hatanaka. La serie temporale originaria e quella trasformata sono evidenziate rispettivamente in azzurro e in rosa. Le colonne rappresentano i vettori di stato.**

Grazie a questa tecnica la dimensione del file originale può essere ridotta fino a 8 volte ed è su questo algoritmo che è stato basato il formato Rinex, utilizzato per la trasmissione di dati di posizionamento rilevati da satelliti.

Come riportato in [CCSDS, 2006] il CCSDS (Consultative Committee for Space Data Systems) ha messo a punto una raccomandazione riguardo alla compressione senza perdite di dati di posizionamento legati alla ricerca e alle trasmissioni nello spazio. Tra gli algoritmi lossless considerati dalla commissione è stato selezionato

l'algoritmo di Rice per la migliore qualità delle prestazioni offerte. Quello di Rice è un algoritmo di codifica a lunghezza variabile derivato dall'algoritmo di Golomb, con variazioni sul prefisso e sul parametro fissato, che è sempre una potenza di 2.
La codifica prevista dal CCSDS è basata su una strategia a due stadi: uno stadio di pre-processing seguito da una codifica basata sull'entropia informativa, cioè sull'algoritmo di Rice.

Gli algoritmi di tipo lossy sono stati elaborati sull'idea della traiettoria percorsa vista come una linea approssimabile mediante curve di Bézier [Choi J.-W. et al, 2008] o Spline, [Dever C. et al., 2006]. Le curve di Bézier, ideate negli anni '60 per il progetto di carrozzerie di automobili, presentano delle interessanti caratteristiche, come quella di passare sempre per un insieme di punti assegnati; di essere tangenti nel punto iniziale e finale alle linee che connettono il primo con il secondo punto e il penultimo con l'ultimo punto dell'insieme assegnato; di giacere sempre nel convex hull formato dai punti di controllo. Le funzioni spline, anch'esse caratteristiche di applicazioni grafiche, sono una versione adattiva dell'interpolazione polinomiale e hanno le seguenti proprietà: sono la funzione interpolante con curvatura media minima, quindi la funzione interpolante ottenuta con la interpolazione spline è più liscia di quelle ottenute con altri metodi; essa è più facile da valutare dei polinomi di grado elevato; se i dati da interpolare hanno conformazioni particolari, ad esempio formano dei gradini, la spline interpolante può essere soggetta al fenomeno di Gibbs, ampie oscillazioni in vicinanza di un gradino. Per ovviare a questo problema vengono utilizzate le smoothing spline o le tension spline.

L'approccio classico di Douglas-Peucker [Dougles D. H. et al 1973] basato su euristica è frequentemente utilizzato nella grafica per approssimare curve semplicemente mediante segmenti di retta. Questo approccio, però, non è applicabile alla compressione di percorsi GPS, poiché una parte essenziale dell'informazione, ovvero i dati temporali, non sono da esso presi in considerazione ed andrebbe persa. Versioni modificate di questo algoritmo introducono introducono anche l'informazione spazio-temporale mediante metriche centrate sull'errore. Ulteriori particolari sull'algoritmo di Douglas-Peucker sono reperibili in appendice.

L'algoritmo di Bellman [Bellman R., 1961], basato su programmazione dinamica, è una mera minimizzazione spaziale dell'errore quadratico medio sotto specifiche condizioni. La sua implementazione più elementare è eseguibile in tempo $O(n^3)$ nel caso peggiore, ma, ampliando lo spazio di memoria può essere ridotto a un tempo $O(n^2)$. Questo algoritmo è anche fondato sull'ipotesi che i dati si presentino in ingresso come delle funzioni a un solo valore e la versione di questo algoritmo adattata per le traiettorie esegue l'algoritmo originario ripetutamente sui segmenti dei tracciati prodotti dal GPS. Dimendioni maggiormente ridotte del dataset ovviamente migliorano il tempo di esecuzione, ma ciò ha una ripercussione negativa sulla precisione, poiché non è realizzata una ottimizzazione globale.

L'algoritmo STTrace [Potmias S. et al, 2006] è finalizzato ad archiviare I dati spazio-temporali servendosi di un poligono per eseguire un adattamento del grado di variazione di velocità e direzione. L'appartenenza dei punti della traiettoria al poligono determina la loro memorizzazione. Questo algoritmo può essere soggetto ad un errore di

propagazione e per tale motivo viene introdotto un secondo poligono come area di sicurezza e tale accorgimento elimina il problema della propagazione dell'errore. Il miglioramento della precisione richiede però una più onerosa serie di calcoli e solo i punti che appartengono all'area di sicurezza sono ritenuti affidabili.

I metodi statistici [Gajewski B. J. et al., 2000] di compressione utilizzano i livelli di aggregazione dei dati. Basandosi sulla massima similarità, quindi sulla minima variabilità durante ogni intervallo di tempo, viene individuato un livello ottimale di decomposizione entro un tempo prefissato.

Sulle conclusioni del teorema di Shannon, si può fare ricorso all'analisi con wavelet per identificare le tendenze delle curve, cosicché la decomposizione mediante wavelet è molto spesso conveniente per catturare informazioni importanti nei sistemi di trasporto intelligenti, come riferito in [Yi P. et al, 2004].

In [Cao H. et al, 2006] il problema della compressione dei data point del tipo(x,y,t), spazio-temporali in due dimensioni, è risolto con una tecnica di semplificazione delle linee e l'errore sulle traiettorie approssimate è confinato entro una fascia prefissata. Il conseguente risparmio dello spazio di occupazione di memoria è molto consistente, anche se ci sono dei tipi di query che potrebbero produrre risposte il cui errore risulta illimitato. La sovrapposizione di approssimazioni che causano questo indesiderabile fenomeno dipende dal calcolo e dal tipo di distanza impiegata nel processo di compressione delle traiettorie e dal tipo di query spazio-temporale che viene posta.

2.4 Schema di compressione CoTracks

In alternativa agli algoritmi fino a qui presentati per la compressione di dati provenienti da dispositivi di posizionamento, è stata da me condotta in collaborazione una ricerca al fine di rinvenire una nuova procedura di compressione di dati spazio temporali, basata su una speciale segmentazione dei tracciati rilevati mediante un dispositivo di navigazione satellitare GPS. Questo tipo di compressione è pensata specificamente per una famiglia di dati GPS che si riferiscono a tracciati terrestri ed è pertanto da puntualizzare che questo algoritmo, come tutti gli algoritmi con perdite, non si presta all'applicazione di compressione di dati per impiego spaziale.

Co-Tracks è un algoritmo di compressione con perdite che si propone una semplificazione lineare dei punti costituenti la traiettoria. Le informazioni archiviate nei punti sono parametri caratteristici come una label, la latitudine, la longitudine, l'altitudine, l'angolo formato rispetto a una direzione di riferimento, il tempo, la velocità e la qualità del campione rilevato. Queste "features" sono le componenti di un vettore in uno spazio a F dimensioni (x_1, x_2,…,x_F). Una traiettoria può essere definita come una funzione $T:\{1,\ldots,n\}\rightarrow\mathbf{R}^F$ con $n\in N$:

$T(1)=(x_{11}, x_{12},\ldots, x_{1F})$,

$T(2)=(x_{21}, x_{22},\ldots, x_{2F})$,

$\vdots$

$T(n)=(x_{n1}, x_{n2},\ldots, x_{nF})$;

T(i) è un vertice della traiettoria ed è valida una relazione d'ordine tra vertici in modo che sia $T(i)<T(i+1)\ \forall\ i \in \{1,\ldots,n\}$.

La traiettoria può essere sottoposta a una procedura di semplificazione introducendo degli speciali inviluppi dei dati rappresentativi dei punti del percorso: si sta facendo riferimento a speciali convex-hull, i Minimum Bounding Box (MBBs), i quali altro non sono se non iper-parallelepipedi nello spazio a F dimensioni. In altri termini l'obiettivo è quello di trovare i parallelepipedi di misura minima all'interno dei quali si possano racchiudere sottoinsiemi di punti adiacenti della traiettoria oggetto di studio. La scelta dei Minimum Bounding Box è sembrata la più conveniente perché questi particolari convex-hull possono avere una dimensione molto più sviluppata delle altre, come avviene nei tracciati di tipo GPS. Nella dimensione trasversale può essere rappresentata una fascia entro cui dev'essere contenuto l'errore, realizzata con limitazioni sulla diagonale di base degli MMB in 3 dimensioni. Tutti i punti che cadono all'interno di un MMB sono rappresentabili mediante un segmento di retta i cui estremi sono il primo e l'ultimo dei punti scelti e inseriti nel MBB. L'algoritmo CoTracks è riassunto e rappresentato nel seguente diagramma a blocchi di figura 10:

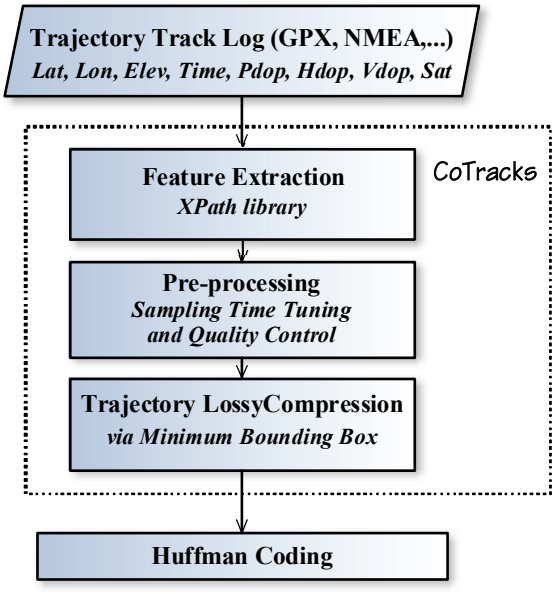


Figura 10: Diagramma a blocchi dell'algoritmo CoTraks

Nei casi ordinari di tracciati registrati sui tragitti delle reti stradali urbane, dato il carattere per lo più rettilineo o quasi rettilineo della maggior parte dei segmenti comunemente percorribili, enormi quantità di punti appartenenti alle traiettorie dei

tracciati possono essere trascurati senza rilevanti perdite di informazione, con risultati di compressione di un certo rilievo dei file che contengono i percorsi.

I dati geodetici sono raccolti da dispositivi di posizionamento e archiviati in file, secondo le indicazioni dello standard della National Marine Electronics Association (NMEA) [NMEA, 2007] o come GPS Exchange Format (GPX) [sito topografix] o anche in altri formati previsti per lo scambio di dati GPS.

Il primo passo consiste nell'estrazione di dati allo stato originario (raw data) direttamente mandando in esecuzione dei parser specifici sui dati in uscita dal dispositivo di tracciamento. A titolo esemplificativo, uno strumento indicato per il trattamento di dati GPX può essere una libreria del linguaggio XPath [SIT6]. Dopo di ciò, i dati sono sottoposti a un pre-processing [Hönle N. et al, 2008] come riassunto schematicamente nel frammento di pseudo-codice presentato nel seguente *Algoritmo 1*:

Algoritmo 1 TRAJECTORYPRE-CONDITIONING.

Input: $T = (T(1), T(2), \ldots T(n))$: Trajectory points sequence.
sampling_time $\rho \geq$ `1`; *dop* $\delta \geq$ `10`; *visib_sat* $\sigma \geq$ `4`.

Output: S pre-conditioned Trajectory of T.

```
 1: S ← T(1); i ← 1; j ← 1;
 2: while i < |T| do                          // |T| = length(T)
 3:     repeat
 4:         j ← j + 1; if j > |T| then return S ; end if
 5:         DiffTime ← TimeStamp(T(j)) - TimeStamp(T(i))
 6:         MDop ← Media(Pdop(T(j)), Vdop(T(j)), Hdop(T(j))
 7:         Sat ← visibles_satellite(T(j))
 8:     until (DiffTime ≥ ρ) ∧ (MDop < δ) ∧ (Sat > σ)
 9:     S ← S ∪ (T(j))                        //S = (s1, s2,...)
10:     i ← j
11: end while
12: return S
```

Questo algoritmo è fondamentalmente orientato a gestire un doppio ordine di circostanze critiche:

- Il tempo di campionamento è molto diverso per i pedoni, le auto e gli aviogetti e sarebbe preferibile poter scegliere i parametri del dispositivo di rilevamento secondo la velocità. tuttavia può capitare che l'intervallo di campionamento sia troppo breve o troppo lungo e che non sia possibile ottenere un campionamento corretto dei punti della traiettoria. Per questo motivo, per intervalli di campionamento troppo brevi è più opportuno effettuare un controllo sui punti prima di mandare in esecuzione l'algoritmo di compressione, al fine di evitare un oversampling. A questo problema sono riferite le istruzioni delle righe 5 e 8 dell'*Algoritmo 1*. Rimane da ammettere che la scelta di intervalli di campionamento troppo lunghi produrrebbe una irreversibile perdita di dati significativi.
- Il numero complessivo di satelliti effettivamente in visibilità e operativi e i parametri di "Dilution of Precision" (DOP) sono indici di qualità del segnale

[Misra P. et al., 1999]. Nella fase preliminare è decisamente opportuno scartare i punti che presentano parametri di qualità inferiori a una soglia prestabilita, come nelle istruzioni che sono scritte nelle righe 6-8 dell'*Algortimo 1*.

Ogni punto è altresì confrontato con i due punti ad esso adiacenti, per rivelare la presenza di valori spuri, generati da cause disparate. La scelta di un livello di soglia deve portare in conto l'errore dovuto alla precisione del sistema di posizionamento. In realtà, lo stadio di pre-filtraggio non mira a un radicale cambiamento dell'insieme di punti originario, ma a una conveniente riduzione della dimensione dei dati ottenuta sfruttando alcune proprietà dei data log dei tracciati.

Dopo la selezione preliminare, il numero di punti da sottoporre alla successiva parte dell'algoritmo è considerevolmente diminuito, con un contributo evidente alla riduzione delle dimensioni del file contenente i dati. A questo punto l'algoritmo CoTracks estende l'idea di compressione dei dati di tipo tracking logs basata sull'uniformità dei parametri 2D ad altri parametri, vale a dire all'altitudine, alla velocità, all'accelerazione e al verso di percorrenza della traiettoria, e calcola le analogie tra punti consecutivi della traiettoria confrontandone le feature, raccogliendoli nei Minimum Bounding Box e riducendo la rappresentazione di tutte le informazioni relative ai punti contenuti nello stesso MBB, ritenuti omogenei, a quelle di una coppia di punti rappresentativi, che si trovano sulle basi del "parallelepipedo".

L'iterazione di questa procedura produce una traiettoria composta da segmenti di retta consecutivi, la cui distanza dalla traiettoria originaria è limitata da una determinata tolleranza. Questa tecnica permette anche di ricostruire i punti mancanti della traiettoria di partenza, supposta la regolarità della distribuzione sia spaziale che temporale dei dati.

L'*Algorimo 2* è una versione breve della rappresentazione in pseudo-codice della procedura che stiamo esaminando.

Algoritmo 2 TRAJECTORYLOSSYCOMPRESSION.

Input: S = ($s_{1,}$ $s_{2,...}$): pre-conditioned Trajectory point seq.
TV: Tolerance Vector.
Output: C: lossy compression of S.

```
1:  C ← s1 ;  PointSet ← s1 ;  i ← 2;
2:  while i ≤ |S|  do                    // |S| = length(S)
3:          PointSet ←  PointSet ∪ si
4:          MBB ← MinBoundingBox(PointSet)
5:          B ← BaseDiagLength(MBB)
6:          V ← VelocityRange(MBB)
7:          A ← AccelerationRange(MBB)
8:          S ← SenseUniformity(MBB)
9:          MC ← (B,V,A,S)       // MBB Characteristics Vector
10:         if  compare(TV,MC)  then              // si  ∉ MBB
11:            C ← C ∪  si-1 ; PointSet ← si-1 ∪ si  // new MBB
12:         end if
13:         if (i == |S|) then C  ← C  ∪  si ; return  C; end if
14:         i ← i + 1;
15: end while
16: return C
```

I dati pre-condizionati sono elaborati sequenzialmente per raggrupparli in sottoinsiemi di elementi con caratteristiche analoghe. Il gruppo di punti così individuati è inviluppato da un Minimum Bounding Box, che può essere considerato come il più piccolo iper-parallelepipedo nell'iperspazio che avvolge il convex hull di un sottoinsieme di dati (si osservi la figura 11).

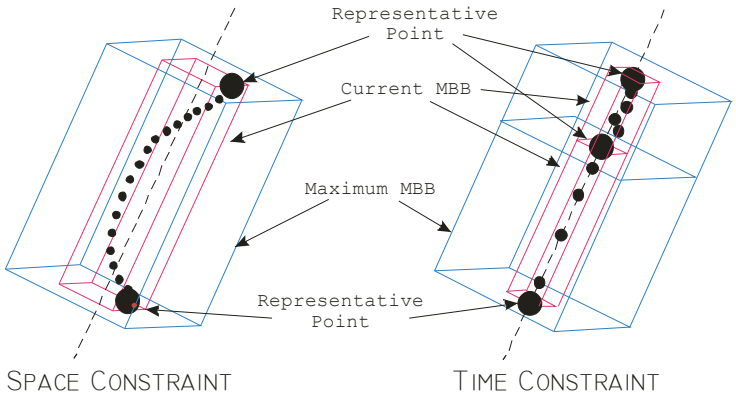


Figura 11: Esempio di Minimum Bounding Box per vincoli spaziali e temporali.

Il vettore TV di tolleranza sull'input memorizza le tolleranze prefissate sulle caratteristiche, mentre alla riga 10, MBB Characteristics Vector contiene i valori dei parametri specifici dell' MBB allo stato corrente, valori che sono aggiornati ad ogni nuovo punto incluso nel Minimum Bounding Box. Le caratteristiche del MBB, fatta eccezione della dimensione longitudinale dominante, sono limitate da una tolleranza con la quale viene determinato il grado di similarità tra la traiettoria campionata e quella approssimata.

D'altra parte la larghezza della soglia ha ripercussioni sul livello di compressione raggiungibile. Valori molto bassi della soglia sono causa di un numero più elevato di MBB per rappresentare una traiettoria, perché i valori delle feature dei punti adiacenti superano molto facilmente la soglia, facendo spesso sorgere l'esigenza della creazione di un nuovo MBB.

La routine contenuta nelle righe comprese tra la 2 e la 10 dell'Algoritmo 2 è ripetuta finché anche uno solo dei limiti di tolleranza sia varcato: in questo caso, a partire dall'ultimo punto processato e che non soddisfaceva i vincoli, è "fondato" un nuovo MBB secondo quanto descritto nelle righe 11-13 dell'Algoritmo 2.

Il rapporto di compressione è una funzione crescente della quantità dei punti inseriti in ogni MBB. La soglia di tolleranza non può essere inferiore di un certo valore di errore sistematico, perennemente presente nei dati.

Il procedimento descritto per il raggruppamento dei punti in una pluralità di MBB consente, una volta che sia stato completamente eseguito, di ridurre tutti i punti appartenenti a ogni MBB a due soli punti, il primo e l'ultimo, a il segmento di retta che unisce gli estremi così individuati è considerata come l'approssimazione della traiettoria reale.

Per motivi di continuità, l'ultimo punto di un MBB è contemporaneamente considerato anche come primo punto dell'MBB successivo e l'intera traiettoria compressa è ottenuta collegando tutti i punti che successivamente si ottengono con questo procedimento.

Come passo terminale, i dati relativi alle traiettorie compresse possono essere sottoposti a una efficiente codifica senza perdite, come la codifica di Huffman, per ridurre al minimo la ridondanza dei dati.

Per ricostruire i punti della traiettoria si ricorre a un processo di decodifica basato sull'approssimazione della traiettoria iniziale con il segmento che unisce i punti estremi rappresentativi, individuati sull' MBB di appartenenza. In questa ricostruzione gioca un ruolo fondamentale l'intervallo di campionamento prescelto. La tolleranza cui ci si è attenuti in fase di compressione garantisce che l'errore sia limitato entro i confini di un intervallo prefissato di valori.

Come è deducibile dai risultati sperimentali raccolti, con l'algoritmo CoTracks è possibile raggiungere rapporti di compressione molto elevati, conseguenza della presenza nei tracciati di numerosissimi punti con caratteristiche cinematiche essenzialmente omogenee.

A fronte di questo successo nella realizzazione della compressione con perdite c'è però da sostenere uno sforzo computazionale notevole, trattandosi di un algoritmo che ha una complessità del'ordine di $O(n^3)$, intrinseca nella procedura della costruzione dei Minimum Bounding Box, che può provocare un allungamento dei tempi di esecuzione dell'algoritmo al crescere del numero dei punti del tracciato da processare. Tale inconveniente è riducibile mediante numerose strategie: ad esempio si può provare a ridurre la cardinalità dell'insieme dei punti con una scelta strategica dell'intervallo di campionamento oppure con una limitazione del massimo numero di punti per MBB e quindi del numero di iterazioni dell'algoritmo. In questo caso, il numero di punti è tenuto basso, ma la compressione finale risulta ugualmente completa.

2.5 Applicazione e valutazione delle prestazioni

Lo schema CoTracks è stato applicato a traiettorie differenti tra loro per lunghezza, pendenza e tortuosità. Per meglio visualizzare le modalità di funzionamento di CoTracks, in questo contesto è stato scelto un tracciato piuttosto breve. In particolare, in figura 12 alla pagina seguente, è riportato come esempio il risultato ottenuto dallo studio del data log di un tracciato menorizzato da un logger GPS in un percorso circolare intorno al campus universitario del complesso di Monte Sant'Angelo, sede del polo scientifico-tecnologico dell'Università di Napoli "Federico II".

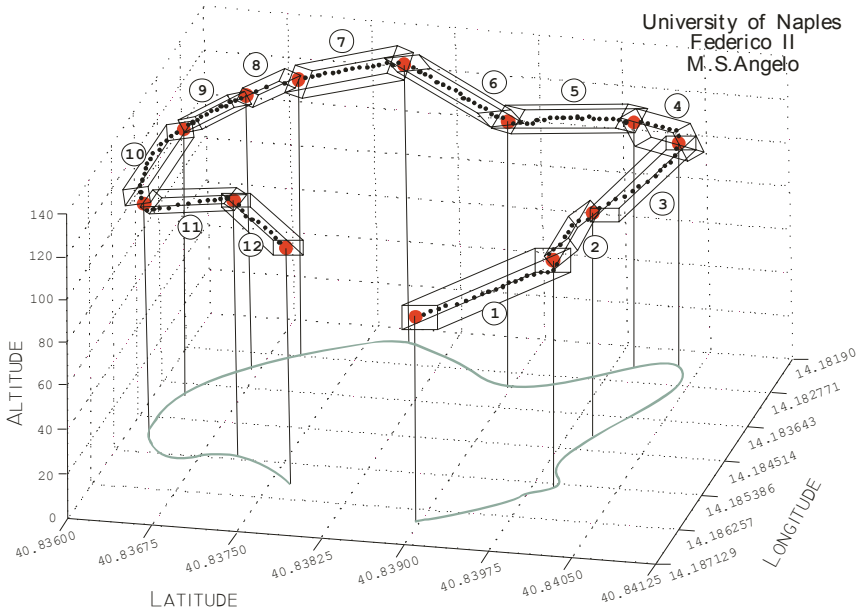


Figura 12: Caso studio, traiettoria formata da punti campionati e relativi MBB. Sono evidenziati i punti rappresentativi, dai quali sarà possibile ricostruire la traiettoria durante il processo di decompressione.

L'insieme di punti è mostrato in una traiettoria in 3 dimensioni nella precedente figura 12. Sono rappresentati anche i Minimum Bounding Box generati dalla procedura dell' Algoritmo 2.

La lunghezza del percorso è di circa 1700 metri, l'altitudine minima è di 93metri e la massima è di 127 metri, il tempo di campionamento è di 1 secondo, il numero di vertici è 171 e le dimensioni del file GPX iniziale sono di 47.7 KB.

I dati sono stati preventivamente sottoposti a parsing mediante una libreria di Xpath per ottenere un database di punti completo di parametri spaziali, temporali, cinematici e di qualità del segnale. Poi i segnali sono stati opportunamente pre-elaborati con la procedura descritta nell'Algoritmo 1 e quindi sono stati introdotti nel successivo stadio del processo, quello di segmentazione MBB. Il kernel dell'algoritmo di compressione con perdite CoTracks è stato implementato nel framework Matlab, con procedure realizzate ad hoc oppure elaborate modificando delle librerie già esistenti [SIT5].

Come inviluppi per la traiettoria di figura 12 sono stati individuati dodici MBB e quindi il data point set d'uscita, ovvero la traiettoria compressa è formata da tredici punti per una dimensione totale del file di 1.18 KB. Questi punti sono stati ulteriormente compressi, mediante codifica di Huffman, giungendo in definitiva ad una occupazione di memoria di 374 byte, intestazioni comprese. Il guadagno di compressione è del 99.2 %. Le variazioni significative di velocità, altitudine e angolo sono efficacemente rappresentate dalla traiettoria approssimata. Ad esempio, l'ottavo MBB nella figura 12 rappresenta un cambiamento di velocità rispetto al precedente MBB, come è rilevabile dall'osservazione dello spazio più ampio tra i punti campionati ad intervalli di tempo costanti.

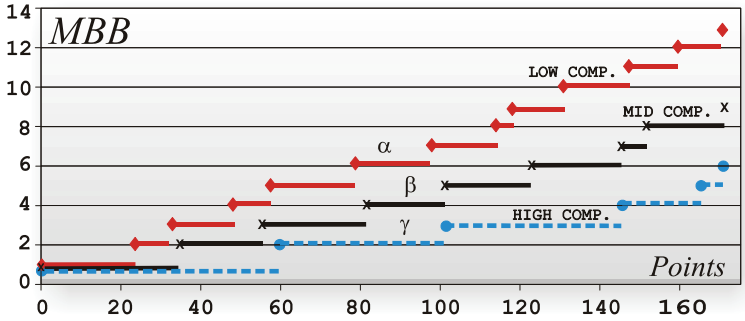


Figura 13: rapporto tra numero di punti campionati e numero di MBB, e quindi di punti della traiettoria approssimata, al variare dei liveli di compressione.

La figura 13 riassume uno studio della dipendenza del livello di compressione raggiungibile per diversi valori della soglia di tolleranza selezionati preventivamente. Il numero di punti originariamente campionati dal dispositivo è rappresentato dall'asse orizzontale e il numero dei punti ottenuti dopo il processo di compressione, uguale al numero di MBB + 1, è rappresentato sull'asse verticale. Il numero di punti della traiettoria approssimata è incrementato di una unità per ogni nuovo MBB introdotto. I risultati del caso di studio sintetizzati in figura 12 corrisponde al valore minimo della soglia di tolleranzae quindi al livello di compressione più basso (funzione a gradini indicata in figura con la lettera α): esiste una corrispondenza tra il numero di punti originali, il numero di MBB e il numero di punti originali inclusi in MBB. Ampliando la soglia di tolleranza, è possibile ottenere un minor numero di punti e quindi un livello di compressione maggiore, come mostrato dalle funzioni contrassegnate in figura 13 con le lettere β e γ.

2.6 Conclusioni

Lo schema di compressione implementato mediante l'algoritmo CoTracks, basato su analogie spazio-temporali tra punti di una traiettoria, è stato applicato a un caso semplice, ma rappresentativo di log data generati da un logger durante il tracciato di un percorso. È emersa una prestazione interessante sia in termini di guadagno di compressione sia nella rappresentazione della traiettoria. Questo algoritmo potrebbe essere impiegato con profitto in applicazioni generiche di trattamento di percorsi terrestri generati da GPS.

Uno sviluppo possibile potrebbe essere un miglioramento della generazione di punti dai Minimum Bounding Box, che tenga conto della regolarità del moto, come ad esempio quella del moto rettilineo uniformemente accelerato.

La forma dei Minimum Bounding Box potrebbe adattarsi meglio ai dati se al posto di parallelepipedi si considerassero dei cilindri.

Nel caso di traiettorie molto contorte o tortuose e limitate ad aree ristrette, con numerosi passaggi per gli stessi punti e con lunghi intervalli in cui i soggetti sono fermi, come avverrebbe ad esempio nel movimento di animali o di pedoni in spazi angusti,

l'algoritmo non farebbe riportare buone prestazioni. Per questo motivo, dovrebbero essere sperimentate delle varianti che non facciano aumentare eccessivamente il numero di MBB a causa dell'eterogeneità tra i punti adiacenti . In tali casi, infatti, l'utilizzo di CoTracks produrrebbe un degrado del guadagno di compressione.
Al fine i limitare l'inconveniente dell'eccessiva complessità computazionale si potrebbero sperimentare dei metodi di riduzione della dimensionalità del problema. Si può infine concepire un approccio multithread, in modo che la traiettoria compressa possa essere ricavata utilizzando processi concorrenti e scegliendo più di un seed point come oggetto iniziale per costruire gli inviluppi MBB fino a trovare la traiettoria finale semplificata.

2.7 Articolo pubblicato

W. Balzano, M.R. Del Sorbo, CoTracks: a lossy compression schema for tracking logs data based on space-time segmentation, CCP1, 1st IEEE International Conference on Data Compression, Communication and Processing, Palinuro, 21-24 June 2011, pagg. 168-171.

3. Serie temporali da Microarray

3.1 Microarray

Come adeguata premessa per un'esauriente esposizione della ricerca condotta, questo paragrafo riassume qualche riferimento sulla tecnica microarray, alla base dell'esperimento che ha generato i dati.

Tranne poche eccezioni, ogni cellula del corpo contiene una serie completa di cromosomi e geni identici. Però solo una piccola parte di questi geni è accesa, ed è il sottoinsieme che si definisce "espresso" e che conferisce proprietà specifiche a ogni tipo di cellula. "Espressione genica" è il termine usato per descrivere la trascrizione delle informazioni contenute nel DNA, il deposito di informazioni genetiche, nelle molecole di mRNA, per una successiva traduzione in proteine che svolgono la maggior parte delle funzioni fondamentali delle cellule. Il tipo e la quantità di mRNA prodotto da una cellula sono utili per sapere quali geni sono espressi, e i geni espressi loro volta forniscono indicazioni su come la cellula reagisce agli stimoli. L'espressione genica è un processo molto complesso e strettamente regolato che permette a una cellula di rispondere dinamicamente sia agli stimoli ambientali che alle cambiare delle proprie caratteristiche. Questo meccanismo agisce sia come un interruttore in cui i geni vengono

espressi in una cellula sia come un controllo di volume che aumenta o diminuisce il livello di espressione di particolari geni, se necessario.

Per lo studio dell'espressione genica, intesa come un processo mediante il quale l'informazione contenuta nel DNA di un gene viene convertita in una macromolecola funzionale, tipicamente una proteina, sono state proposte molte tecniche. Molte di esse richiedono tempi di esecuzione eccessivi e sono limitate dal numero di geni che è possibile studiare in parallelo, contemporaneamente. Al contrario, la tecnica microarray, o DNA chip, permette di trovare conferma o confutazione di molte ipotesi in un unico esperimento, perché è uno strumento di indagine ad elevato parallelismo. Questa tecnologia è fondata sui dati recentemente acquisiti dalla sequenziazione dei genomi e permette di capire come siano regolati i geni, ovvero la loro risposta a determinati input in un particolare istante e in un dato tipo di cellula.

Nella sua forma più generale, un microarray di DNA si presenta come una piastrina di nylon, di vetro o di plastica occupata da uno schieramento regolare di sonde microscopiche, che sono disposte come una sorta di matrice di punti. Su questa piastrina vengono depositate varie sequenze di DNA a singolo filamento (ssDNA). In accordo con la nomenclatura proposta da Duggan (Duggan, D. et al, 1999) indicheremo con il nome di "probe" l'ssDNA stampato sul substrato, la cui natura dipende da quali funzioni geniche debba sondare l'esperimento che si sta conducendo. L'array così predisposto serve per trovare la risposta a un quesito specifico sul DNA mediante un trattamento dell'array con una soluzione contenente ssDNA target, generato da un campione biologico oggetto di studio.

L'idea di base è che il DNA spalmato sulla superficie dell'array ibridizzerà le sequenze complementari contenute nella soluzione contenente il DNA target. La chiave d'interpretazione dell'esperimento microarray è nel DNA utilizzato sull'array per l'ibridazione. Il DNA target è marcato con un colore fluorescente, un elemento radioattivo o altro metodo in modo che il punto in cui avviene l'ibridazione possa essere ben evidenziato e possa fornire anche informazioni quantitative. Nelle applicazioni finalizzate allo studio dell'espressione genica, il DNA target è prodotto per reazione della transcriptasi inversa dall' mRNA estratto da un campione di tessuto (figura14)

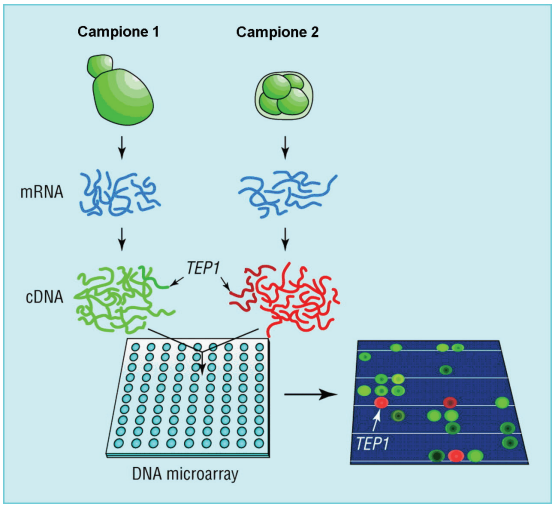


Figura 14: Tecnologia DNA microarray

Questo DNA è marcato con una colorazione fluorescente, in modo che in seguito, irradiando il microarray con una luce apposita, si possa ottenere una specifica immagine dell'array delle caratteristiche, cioè dell'insieme di probe su GeneChips, di punti sugli array di cDNA o di bead di silicio sugli array Illumina. L'intensità luminosa relativa ad ogni punto e la differenza media tra match e mismatch possono rendere nota la quantità di mRNA presente nel tessuto e, di conseguenza, la quantità di proteina prodotta dal gene corrispondente a una determinata caratteristica. I diversi DNA target possono essere marcati con colori diversi ed utilizzati contemporaneamente in un procedimento di ibridazione competitiva nell'ambito di un esperimento multicanale. Successivamente al passo di image processing si ottiene un numero molto elevato di valori di espressione genica. Tipicamente, un unico chip sperimentale può produrre centinaia di migliaia di valori o più.

Chiamiamo profilo di espressione genica i diversi valori di espressione che un gene manifesta in diverse condizioni, sotto determinate sollecitazioni o in diversi tessuti. Le righe della matrice rappresentata nella figura di sopra sono proprio il profilo di espressione genica. Le colonne, invece, rappresentano lo stato di trascrizione, o impronta, di un determinato campione.

Nell'esperimento eseguito nello studio presentato in questa parte della tesi, i microarray sono stati sfruttati per comprendere alcuni meccanismi genetici della cellula vivente mediante l'analisi di una evoluzione temporale dell'espressione genica;

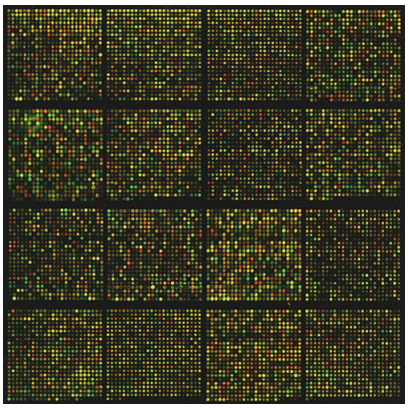


Figura 15: Immagine di un microarray cDNA contenente circa 8.700 sequenze di geni (dall' Incyte GEM1 clone set) elaborata presso NCI Microarray Facility (Advanced Technology Center, Gaithersburg). In essa sono rappresentate le differenze di espressione genica tra due diversi tessuti prelevati da un topo.

oltre che per valutare i livelli di espressione genica, però, i microarray possono essere impiegati con successo in una vasta gamma di applicazioni, come la sequenziazione [Schena M. et al., 2000], l'individuazione di Single Nucleotide Popymorphism [Fan J.-B. et al., 2000], il genotyping (misura di una variazione genetica), associazione di malattie, legami genetici, attenuazione o amplificazione genomica, rivelazione di riadattamenti dei cromosomi e tantissime altre. Al di là delle particolari applicazioni, però, il cospicuo numero di risultati sperimentali ottenuti con i microarray hanno conferito a questa tecnologia di generazione di dati di espressione genica una complessiva e riconosciuta affidabilità [Celis J. E., 200]. I microarray possono essere

anche utilizzati a scopo puramente computazionale, come nel caso del DNA computing [Kari L., 1997]: il microarray può contenere sequenze di DNA che codificano le possibili soluzioni di un problema.

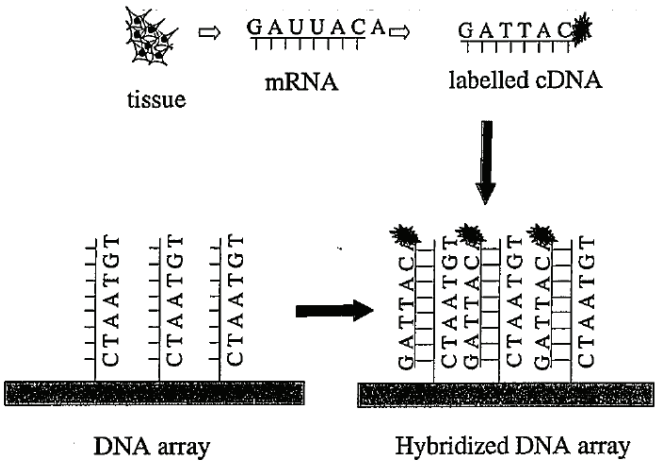


Figura 16: L'mRNA estratto è trasformato in cDNA, ibridizzato con DNA presente sul microarray

In quanto tecnologia relativamente nuova, i microarray comportano una serie di problemi nella loro utilizzazione:

a. Rumorosità: a causa della loro stessa natura, i microarray tendono ad essere molto rumorosi e un esperimento, anche se condotto con gli stessi materiali e preparato nelle stesse identiche condizioni al contorno, produce valori quantitativi diversi per molti dei geni studiati. Il rumore è introdotto ad ogni passo delle varie procedure: la preparazione dell'mRNA (variano tessuti, kit e procedure), la trascrizione (la reazione e gli enzimi non sono sempre uguali), la marcatura (tipo e età del marcatore), l' amplificazione, il tipo di pin (penna, anello, ink-jet), la chimica delle superfici, l'umidità, il volume di destinazione (oscilla anche per uno stesso pin), i parametri di ibridazione (tempo, temperatura, buffering), l'ibridazione aspecifica (cDNA marcato ibridato su superfici che non contengono sequenze perfettamente complementari), l' ibridazione di fondo non specifica, la presenza di polvere, la scansione (settaggio del guadagno, limitazioni del range dinamico, allineamento tra i canali, la segmentazione (separazione caratteristica/background), la quantificazione (media, mediana, percentile dei pixel in un punto) e così via. Quando si confrontano tra loro diversi tessuti o diversi esperimenti su uno stesso tessuto non è sicuro che la variazione riscontrata su un gene particolare sia effettiva oppure che sia una conseguenza del rumore generato da una o più delle cause appena elencate. Inoltre, quanta parte della varianza misurata osservando un gene specifico è dovuta alla regolazione del gene e quanta invece al rumore? Ovviamente il rumore è un fenomeno ineludibile e l'unica arma efficace per combatterlo è la ripetizione degli esperimenti.

b. Normalizzazione: lo scopo della normalizzazione è quello di tenere in conto le differenze sistematiche tra i diversi data ed eliminare i dati spuri, come ad esempio gli effetti non lineari di colorazione dei marcatori. E' opinione comune che la normalizzazione sia indispensabile quando devono essere combinati i risultati raccolti con tecniche sperimentali diverse, ma non tutti concordano sulle specifiche modalità di normalizzazione.

c. Progettazione dell'esperimento: questa fase, cruciale ma spesso ignorata negli esperimenti microarray, consiste nel prevedere quali variabili in input ad un processo sia più opportuno cambiare in modo da ottenere una risposta in output dalla quale sia possibile osservare ed identificare le cause del cambiamento.

d. Grande numero di geni: il successo dei microarray è dovuto, tra l'altro, alla possibilità di accogliere una sperimentazione di migliaia di geni in parallelo, ma questa caratteristica presenta anche degli inconvenienti provocati dalla gestione contemporanea di così tanti dati. Inoltre, il gran numero di geni può modificare la qualità del fenomeno ed i metodi da adottare. L'esempio più calzante è quello del calcolo dei p-value in una situazione di test multiplo, in cui si deve portare in conto la presenza di molti esperimenti paralleli con opportuni fattori correttivi (Draghici, S., 2011).

e. Significatività: per stabilire la significatività di un esperimento, non possono essere applicate semplicemente le tecniche della statistica classica, come il test del chi-quadro, giacché in questo tipo di esperimenti il numero di variabili è molto maggiore del numero di esperimenti.

f. Fattori biologici: nonostante i numerosi vantaggi, per molteplici motivi i microarray non riescono a sostituire del tutto le altre tecniche e strumenti della biologia molecolare.

g. Controllo di qualità dell'array: è utile se l'analisi dei dati non è l'ultimo passo in un processo lineare di esplorazione ma come uno stadio di un loop che produce il feedback necessario alla messa a punto delle procedure atte a produrre il microarray stesso. In tal caso sarebbe assai utile che l'analisi generasse dei valori di qualità dei dati insieme ai valori delle espressioni geniche stesse.

Uno dei vantaggi, ma allo stesso tempo una delle potenziali complicazioni che possono sorgere nell'utilizzo di microarray, come già detto, consiste nella possibilità di realizzare una sperimentazione ad elevatissimo throughput: dopo il passo di ibridazione e di trattamento dell'immagine, ad ogni spot del chip corrisponderà un valore numerico, per un totale che può arrivare fino a decine di migliaia di valori. Data la delicatezza dell'esecuzione delle procedure sperimentali, che possono essere alterate in maniera determinante da molteplici fattori, è necessario ripetere molte volte uno stesso esperimento, affinché possa essere validato inconfutabilmente. In tal modo, la mole dei dati generati da questo tipo di esperimenti è enorme e ben si presta ad applicazioni di molte tecnologie di data mining.

Le più innovative tecniche per i database, come l'indexing multidimensionale, possono rendere più semplice l'elaborazione dei dati microarray, anche se è necessario

stabilire delle regole per fare in modo che i diversi database, come ad esempio quello contenente le serie di valori ricavate sperimentalmente, quello contenente le informazioni funzionali e quello contenente i protocolli di laboratorio possano essere interfacciati nel modo più efficace.

3.2 Signaling Pathways

La massa di informazioni raccolte mediante esperimenti basati su tecnologie ad elevato throughput, come ad esempio i voluminosi dataset di valori di espressione genica, si presta ad essere elaborata per ricavare un'interpretazione dei dati e quindi una conoscenza più dettagliata dei fenomeni biologici su cui si sta effettuando una sperimentazione e in particolare per collocarle nel più ampio quadro d'insieme di un organismo vivente, visto come sistema complesso.

Per rappresentare le funzionalità biologiche, metaboliche e cellulari degli organismi è possibile ricorrere ad una rappresentazione basata sul comportamento dei fattori di trascrizione, che sono delle proteine che si legano a specifiche sequenze di DNA controllando la trascrizione dell'informazione genica dal DNA al mRNA. I fattori di trascrizione possono facilitare o limitare l'afflusso di RNA-polimerasi, enzima che catalizza la sintesi di un filamento di RNA che è proprio la trascrizione di una informazione genetica dal DNA all'RNA, e così condizionano fortemente le dinamiche cellulari.

A partire da questi concetti è possibile costituire una rete di trascrizione, costituita da nodi e archi orientati e dotati di un attributo. Nella corrispondenza con il sistema della cellula si ha che ogni nodo rappresenta un gene, i segnali d'ingresso portano con sé informazioni relative allo stato del sistema, mentre gli archi codificano le interazioni del sistema.

Un network genico è una rete che rappresenta le interazioni tra gli elementi di un insieme di geni, in dipendenza dalle funzioni che esplicano uno nei confronti dell'altro. I network genici possono quindi essere schematizzati con grafi i cui nodi sono occupati appunto dai trasmettitori e recettori dei segnali e gli archi orientati rappresentano la presenza di una via di comunicazione e il verso della comunicazione stessa. Tenendo conto che i network biologici sono solo la parte hardware della rappresentazione biologica, che deve essere necessariamente completata dall'insieme dei segnali di attivazione o inibizione o tanti altri, si introduce la locuzione "signaling pathway", per significare un'entità che comprende i network e la loro funzionalità di trasmissione di messaggi tra geni.

Con il procedere della sperimentazione sugli organismi, alla luce delle più aggiornate conquiste della genomica e grazie alla collaborazione globale di tutti i laboratori impegnati in questo tipo di ricerche, sono state annotate grandi moli di informazioni su queste reti di segnalazione. In particolare al momento della stesura di questa tesi (ultimo trimestre 2011), da un'indagine condotta su KEGG, il più grande e più consultato database di informazioni genomiche, risulta che il numero dei pathway

attualmente presenti nel database è di 412, di cui 246 appartengono all' homo sapiens. Questo numero è destinato a crescere con la raccolta di nuovi dati, così come la rappresentazione della realtà fornita dai pathway già individuati e catalogati sarà sempre più precisa e dettagliata e le annotazioni dei singoli elementi saranno sempre più circostanziate e ricche di particolari. Si comprende evidentemente anche l'opportunità fornita da questi pathway di rivelare i malfunzionamenti della comunicazione biologica cui conseguono stati patologici. Quindi, rilevando i profili delle espressioni differenziali dei geni, ad esempio mediante esperimenti microarray, è possibile individuare i pathway coinvolti in determinati fenomeni fisiologici o patologici. Questo tipo d'indagine è quella che va sotto il nome di "pathway analysis". L'analisi dei pathway permette ai ricercatori di accrescere significativamente la sensibilità dei metodi di analisi genetica.

Al momento, gli approcci di pathway analysis possono essere distinti usando come criterio le annotazioni funzionali supportate (ad esempio GO, KEGG, Biocarta, Reactome), tipo di analisi impiegata (sovrarappresentazione, funzional class scoring, basata sulla topologia del pathway) e tipo di funzioni statistiche usate (ipergeometrica, binomiale, t-test, correlazione). Adotteremo qui una suddivisione per tipo di analisi impiegata.

Il punto di partenza è una lista di geni differenzialmente espressi in determinate condizioni oppure la corrispondente lista di p-value, o statistiche simili che quantifichino il grado di espressione differenziale per ogni gene considerato nell'esperimento. L'idea è di esaminare i p-value associati con un particolare sottoinsieme di geni, appartenenti tutti alla stessa categoria di GO [SIT22], per vedere se il sottoinsieme mostra una tendenza ad avere p-value più piccoli rispetto a quelli di tutti gli altri geni non appartenenti alla categoria considerata. Per fare ciò, bisogna ricorrere a un test statistico per quantificare la differenza e mettere a punto uno schema di ricampionamento per poter valutare se la differenza riscontrata sia reale o solo attribuibile al caso. Questo procedimento si può ripetere per tutte le categorie di geni che possano essere d'interesse.

Allo scopo, possono essere adottati molti test statistici diversi che sono classificabili come segue:

Analisi di sovrarappresentazione (Over Repesentation analysis, ORA): con questo approccio viene fissata una soglia per i p-value, ad esempio il 5%, e ogni gene di una certa categoria di Gene Ontology (GO) è etichettato come "affidabile" se il suo p-value è maggiore della soglia (>0.05) o "non significativo" se è minore della soglia (>0.05). La proporzione di geni significativi, poi, è confrontata mediante test esatto di Fisher con la corrispondente proporzione della popolazione di geni in esame per verificare se c'è una sovrarappresentazione dei geni significativi nella intera categoria di geni [Hosack, D. A. et al., 2003, Doniger S. W. Et al.,2003, Zeeberg B. W et al, 2003].

Punteggio di classe funzionale (Functional class scoring, FCS): questo approccio calcola una statistica che riassume i p-value per tutti i geni appartenenti a una data categoria di GO [Pavlidis, P. et al., 2004].

Punteggi di distribuzione: questi test confrontano la distribuzione di p-value appartenenti a una categoria di GO con la distribuzione di tutti i p-value della lista iniziale [Mootha V. K. et al., 2003].

Nel 2002 Draghici e Khatri [Khatri P. et al., 2002, Draghici S. et al., 2003] hanno proposto un approccio automatico di analisi fondato sulla GO. Tale metodologia prevede l'applicazione di un'analisi statistica alle liste di geni differenzialmente espressi per individuare le categorie di GO, ad esempio la categoria "componente cellulare", "processo biologico" e così via, che sono sovrarappresentate o sottorappresentate. Dato un insieme di geni differenzialmente espressi, questo approccio confronta il numero di geni differenzialmente espressi trovati in ciascuna categoria d'interesse con il numero di geni che dovrebbero essere differenzialmente espressi per puro caso. Nella circostanza in cui le quantità di geni differenzialmente espressi risultino molto diverse tra loro, la corrispondente categoria di GO è indicata come significativa. Per calcolare la probabilità di osservare il vero numero di geni differenzialmente espressi per puro caso, ovvero il p-value, si può utilizzare un modello statistico, come ad esempio quello ipergeometrico.
Nell'articolo [Doniger S. W. et al., 2003] sono individuati insiemi di geni associati alla divisione e alla crescita cellulare con il criterio di sovrarappresentazione.

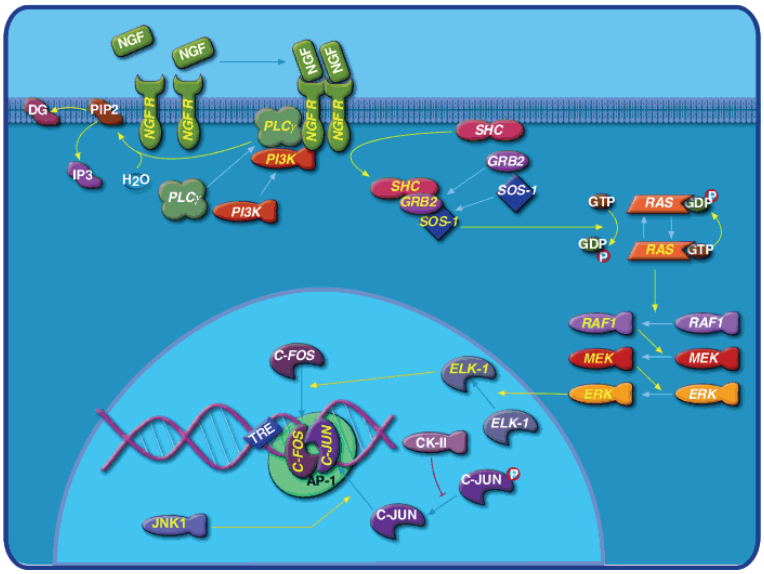


Figura 17 –Pathway del fattore di crescita nervoso (da Biocarta)

Attualmente esistono più di 50 tool tra quelli che adottano in vari modi un criterio di sovrarappresentazione (Over Representation Approach, ORA) [Khatri P. et al., 2005 e 2010), come è possibile rilevare su [SIT22]. Generalmente per ogni gene della lista in input sono recuperate le annotazioni del pathway corrispondente. Dopo vengono conteggiati tutti i geni della lista di input che si trovano in ciascun pathway e si ripete il processo per tutti i geni oggetto di studio. Infine sono individuati i pathway che sono sovra o sottorappresentati nella lista dei geni differenzialmente espressi, in relazione alla lista di riferimento dei geni, calcolando la significatività statistica per ogni pathway,

ad esempio mediante test esatto di Fischer, distribuzione ipergeometrica, distribuzione binomiale o distribuzione chi-quadro. Nonostante la discreta diffusione, questo approccio è molto limitato da tipo, qualità e struttura delle annotazioni disponibili: i diversi strumenti statistici utilizzati in ORA sono indifferenti ai cambiamenti nelle espressioni geniche misurate. In ogni caso i geni sono espressi con misure diverse in ogni data condizione. I dati che forniscono informazioni sulla misura della regolazione genica, come i "fold-change" o la significatività di un cambiamento nell'espressione genica, possono servire per assegnare un peso diverso ai geni in input e ai pathway cui appartengono. In secondo luogo ORA assume che i geni siano tra loro indipendenti, mentre la realtà biologica conferma l'esistenza di strette interazioni tra questi e i loro prodotti tramite i pathway di segnalazione finalizzate all'espletamento di svariate funzioni. Anzi, l'analisi dell'espressione genica è principalmente finalizzata alla conquista di più approfondita conoscenza di come le interazioni tra i prodotti dei geni si manifestino sotto forma di cambiamenti dell'espressione genica e riescano a spiegare i fenomeni biologici sottostanti.

Un approccio alternativo considera la distribuzione dei geni che compaiono nel pathway sull'intera lista di tutti i geni e prevede un esame della distribuzione statistica dei punteggi dei singoli geni tra tutti i geni contenuti nella stessa classe di GO e non richiede una fase iniziale di selezione di geni. In questo modo è eseguita un'elaborazione di un punteggio di classe funzionale (Functional Class Scoring, FCS) che permette anche correzioni per le correlazioni tra i geni [Goeman J.I. et al., 2004, Pavlidis, P., 2004]. A questo modello è improntata la Gene Set Enrichment Analysis (GSEA), che è un metodo computazionale che determina se un set di geni individuato a priori mostri differenze significative e concordi tra due stati biologici. [Mootha V. K. et al., 2003, Subramanian A. et al., 2005, Tian L. et al.,2005]. Questo tipo di analisi valuta tutti i geni a seconda della loro espressione e dei fenotipi dati ed attribuisce un punteggio che riflette il livello di rappresentazione di un determinato pathway P. Più precisamente la valutazione del punteggio avviene scorrendo la lista dei geni, ordinati per cambiamento di espressione decrescente. Il punteggio viene incrementato per ciascun gene appartenente al pathway P e viene decrementato per ogni gene che non appartiene allo stesso pathway. La significatività statistica è valutata rispetto a una distribuzione nulla costruita mediante permutazioni.

Le tecniche attualmente più utilizzate sia di tipo ORA che di tipo FCS, sono limitate dal fatto che ciascuna categoria funzionale è analizzata indipendentemente da tutte le altre, senza un'analisi globale a livello di pathway o di sistema. Si comprende come questo non possa essere ritenuto un approccio ideale per ottenere una visione di insieme delle dipendenze ed interazioni tra entità biologiche oppure per identificare perturbazioni e cambiamenti a livello dei pathway o dell'organismo [Stelling, J., 2004].
I più importanti database di pathway, come quello contenuto in Kyoto Encyclopedia of Genes and Genomes (KEGG) [Ogata, H. S., 1999], in Biocarta e in Reactome, descrivono normalmente in modo assai semplificato i pathway metabolici e le reti di

segnalazione tra geni, ma questa descrizione è soltanto una base per analisi più complesse, dettagliate ed utili.

La tecnica ScorePage, realizzata da un gruppo di ricercatori del Max Plank Institut informatik [Rahnenführer J. et al., 2004], segue un approccio statistico per valutare i cambiamenti nell'attività dei pathway metabolici deducibili dai dati di espressione genica. Questo metodo identifica i pathway metabolici biologicamente rilevanti e la corrispondente significatività statistica e il peso dato alla topologia del pathway nella valutazione migliora ulteriormente la sensibilità di questo metodo. Tuttavia questa tecnica o altre affini ad essa non è stata ancora applicata nel campo dell'analisi delle reti di segnalazione genica. Tutti i tool di analisi dei pathway attualmente in uso sfruttano un approccio di tipo ORA e non traggono alcun vantaggio dall'abbondanza di informazione insita in questo tipo di dati. Ne diamo qui sotto un elenco dei più rappresentativi:

- GenMAPP/MAPPFinder e GeneSifter utilizzano un generico Z-score [Doniger S. W. et al., 2003, Dahlquist, K. Et al., 2002].
- Pathway processor [Grosu, P. et al., 2002]
- PathMAPA [Pan D. et al., 2003]
- Cytoscape [Shannon P. et al., 2003] and PathwayMiner [Pandey R. et al., 2004] eseguono un test esatto di Fisher, MetaCore sfrutta un modello ipergeometrico
- ArrayXPath [Chung H.-J. Et al. 2004] mette a disposizione sia un test esatto di Fisher che il False Discovery Rate (FDR), metodo statistico per le correzioni su confronto multiplo nel multiple hypothesis testing, che controlla la proporzione di errori di tipo I in una lista di ipotesi respinte.
- VitaPad [Holford M. et al., 2004] and Pathway Studio [Nikitin A. et al., 2003] sono strumenti di visualizzazione e non di analisi di pathway.

La totalità degli strumenti attualmente disponibili e in uso per l'analisi delle reti di segnalazione genica presentano carenze rimarchevoli. Come primo punto della lista si deve evidenziare l'assoluta mancanza di qualsiasi riferimento alla topologia delle reti in cui sono inseriti i geni: questi strumenti, infatti, si limitano a considerare un insieme di geni del pathway, ma ne trascurano la posizione nella rete costituente il pathway stesso.

Questa semplificazione non trova riscontro nella realtà biologica dei geni, che sono strettamente interattivi tra loro. La sfida della ricerca nell'ambito genetico, attualmente, è proprio quella di ricavare una conoscenza approfondita del funzionamento dei pathway di segnalazione e dell'interazione tra i geni, che costituiscono i nodi di queste reti di comunicazione. Inoltre, se anche si verificasse che un pathway sia innescato da un singolo prodotto genico o attivato attraverso un singolo recettore, ma una certa particolare proteina non è prodotta, allora il pathway ne riceverebbe un grosso impatto, e nonostante l'attivazione dei recettori forse sarebbe completamente spento, come avviene nel pathway dell'insulina qualora manchi il recettore INSR. D'altra parte, se accade che un numero cospicuo di geni di un pathway

manifesti un cambiamento del livello di espressione, ma questi geni sono in una parte recondita della rete di collegamento e non in punti cruciali, allora il pathway è solo scarsamente coinvolto nella manifestazione di cambiamento.

In secondo luogo, alcuni geni rivestono ruoli diversi in vari pathway distinti. Un esempio di questo caso è il recettore INSR. Questo recettore è presente nel diagramma che rappresenta il pathway dell'insulina e si trova contemporaneamente anche in altri pathway, tra cui quello delle proteine di giunzione cellulare, ma il suo ruolo in quest'ultimo pathway è meno importante che nel primo. Come mostrato nella figura 18, nel pathway dell'insulina il posto occupato da INSR è centrale, mentre nell'altro pathway (Adherens junction) è decisamente più periferico.

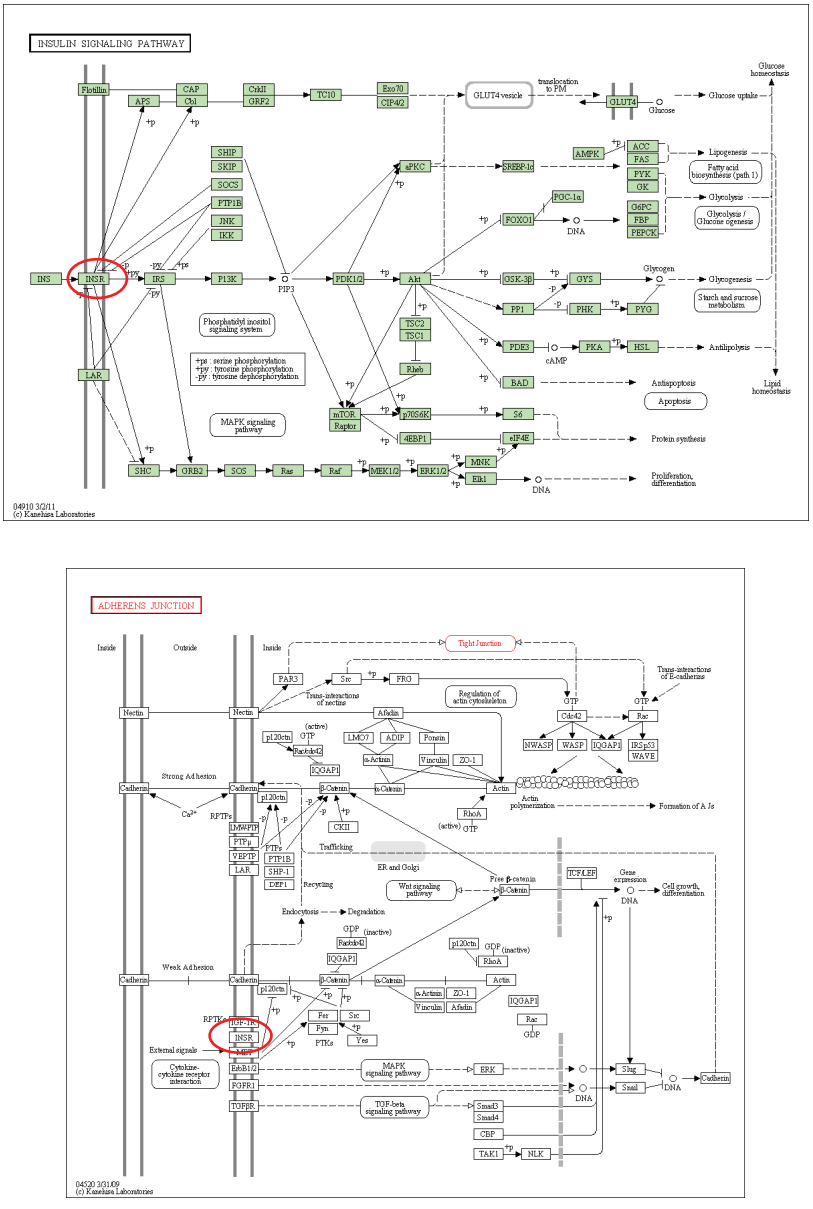


Figura 18: Pathway diversi in cui ISNR ricopre ruoli più o meno rilevanti. a) pathway dell'insulina in cui occupa un nodo centrale in ingresso e b) pathway delle proteine di giunzione cellulare in cui è relegato in un nodo decisamente più periferico (immagini tratte da KEGG).

Il vantaggio davvero considerevole della rappresentazione grafica offerta dai diagrammi dei pathway è proprio quello di visualizzare distintamente le interazioni tra i geni coinvolti nel pathway e anche il modo in cui avvengono le reciproche regolazioni tra i geni. Risulta ora chiaro il motivo per il quale, al fine di ottenere una osservazione fattiva dei fenomeni che si manifestano tra i geni, è quindi essenziale abbandonare gli approcci che prescindono dalla topologia dei pathway e dalla posizione che i geni occupano negli stessi. La prospettiva futura, alla luce dell'attuale proliferare di sperimentazioni sul tema, è sicuramente quella di ottenere un progressivo approfondimento delle conoscenze dei vari pathway con l'aumentare dei dati sperimentali raccolti: i pathway saranno di certo soggetti a numerose modifiche dei nodi, che potranno essere aggiunti o eliminati, e degli archi orientati del diagramma che potranno essere creati, soppressi oppure subire cambiamenti di verso. Nonostante queste trasformazioni, però, la maggior parte delle tecniche esistenti sarebbe incapace di cogliere mutamenti, non si accorgerebbe di nulla e produrrebbe esattamente gli stessi risultati, a patto che si ottemperi all'unica condizione che i geni del diagramma rappresentativo del pathway rimangano gli stessi, anche se le interazioni tra di essi siano state completamente stravolte.

3.3 Analisi d'impatto

Allo stato attuale delle metodologie di analisi dei pathway, i cambiamenti di espressione genica, deducibili dai dati che possono essere raccolti in esperimenti ad elevato throughput come i microarray, sono stati considerati al semplice scopo di individuare ognuno dei geni differenzialmente espressi, limitatamente agli approcci di tipo ORA, oppure per attribuire un punteggio e classificare i geni nei metodi FCS. Non è invece presa in nessuna considerazione la relazione di tali cambiamenti di espressione genica su pathway specifici: in particolare, le tecniche di sovrarappresentazione non discriminano tra la situazione in cui un sottoinsieme di geni presenta un'espressione che supera di poco la soglia prefissata come limite minimo per poterlo considerare di "differenzialmente espresso", ad esempio se l'espressione ha un valore doppio rispetto alla soglia, e la situazione in cui lo stesso sottoinsieme di geni è differenzialmente espresso e la differenza rispetto alla soglia prefissata è di alcuni ordini di grandezza, ad esempio se l'espressione ha un valore pari a cento volte la suddetta soglia. Le tecniche di attribuzione di un punteggio, allo stesso modo, nel caso in cui le correlazioni tra geni e fenotipi rimangano simili, producono una stessa classifica dei geni nonostante i valori dell'espressione genica varino in un intervallo anche piuttosto esteso. In definitiva quindi, per i metodi che seguono un criterio di tipo "enrichment analysis", come ORA, e per quelli di tipo FCS i pathway sono semplicemente degli insiemi non strutturati di geni e non delle reti di interazione in cui si manifesti una intensa e complessa attività di segnalazione tra le varie entità biologiche.

Per riuscire a dare il giusto risalto agli aspetti topologici di cui a questo punto si è ripetutamente evidenziata l'importanza, è stato presentato in [Tarca A.D et al., 2009] un

approccio completamente diverso dai precedenti per effettuare l'analisi dei pathway, chiamato Signaling Pathway Impact Analysis (SPIA).

La principale novità di questo metodo è la definizione di un impact factor (IF), calcolato per ciascun pathway, che ingloba parametri come il fold change (rapporto di cambiamento) dei geni differenzialmente espressi, la significatività statistica dell'insieme dei geni appartenenti al pathway e la topologia del signaling pathway stesso. Lo scopo è quello di sviluppare un modello di analisi che tenga in conto sia di un numero statisticamente significativo di geni differenzialmente espressi sia dei cambiamenti significativi nell'espressione genica su un dato pathway.
Relativamente a questo modello, definiamo preliminarmente un ***fattore di perturbazione di un gene*** *g* nel modo seguente:

$$\boldsymbol{PF}(\boldsymbol{g}) = \Delta \boldsymbol{E}(\boldsymbol{g}) + \sum_{\boldsymbol{u} \in \boldsymbol{US}_{\boldsymbol{g}}} \beta_{\boldsymbol{ug}} \cdot \frac{\boldsymbol{PF}(\boldsymbol{u})}{\boldsymbol{N}_{\boldsymbol{ds}}(\boldsymbol{u})} \qquad (3.1)$$

In questa equazione il primo termine compendia l'informazione quantitativa misurata in un esperimento di espressione genica: il fattore $\Delta \boldsymbol{E}(\boldsymbol{g})$ è rappresentativo della misura, espressa come numero relativo normalizzato, del cambiamento di espressione genica di *g*, determinata mediante uno dei metodi noti [Churchill, G. A. et al., 2002, Draghici S., 2002, Quackenbush J., 2001].
Il secondo termine è una somma di tutti i fattori di perturbazione dei geni *u*, che sono collegati direttamente a monte del gene *g*, normalizzati per il numero, $N_{ds}(u)$, di geni che si trovano a valle (downstream) di *u* e pesato per un fattore β_{ug}, detto di ***efficienza regolatoria***, che dipende dal tipo di interazione tra *u* e *g*.

Ad esempio $\beta_{ug} = 1$ per una attivazione e $\beta_{ug} = -1$ per una repressione. In KEGG l'informazione sul tipo di interazione è disponibile nella descrizione topologica del pathway per tutti i link tra coppie di geni, come illustrato sotto in figura 19.

US_g è l'insieme di tutti i geni che si trovano a monte (upstream) del gene *g* nel pathway considerato. In qualche modo questo secondo termine è simile all'indice PageRank utilizzato da Google [Page, L. et al., 1998], con la differenza che in questo caso hanno importanza le connessioni a valle invece che a monte di un gene: un gene è tanto più importante quando più riesce a far sentire la sua regolazione al più elevato numero possibile di altri geni cui invia i suoi messaggi biologici, mentre una pagina web è tanto più rilevante quante più pagine puntano ad essa.

A questo punto è possibile definire l'***impact factor*** *IF* dell'*i*-esimo pathway tra quelli presi in considerazione in un determinato contesto, chiamiamolo *P*. In breve, $IF(P_i)$ è calcolato secondo la seguente formula:

$$\boldsymbol{IF}(\boldsymbol{P_i}) = \log\left(\frac{1}{\boldsymbol{p_i}}\right) + \frac{\sum_{\boldsymbol{g} \in \boldsymbol{P_i}} \left|\boldsymbol{PF}(\boldsymbol{g})\right|}{\left|\Delta \boldsymbol{E}\right| \cdot \boldsymbol{N}_{\boldsymbol{de}}(\boldsymbol{P_i})} \qquad (3.2)$$

che è la somma di due termini:

il primo termine è una funzione probabilistica logaritmica in cui la variabile indipendente è la significatività dell'i-esimo pathway P_i dal punto di vista dell'insieme dei geni in esso contenuti. Questo termine contiene l'informazione fornita dagli approcci statistici classici più comunemente utilizzati e può essere calcolato sia con test

di tipo ORA (come z-test [Doniger S. W. et al., 2003], tabelle di contingenza [Pan D. et al., 2003], FCS come GSEA [Shi J. et al., 2007] o tanti altri. Specificamente, il valore p_i rappresenta la probabilità di ottenere con la statistica adottata un valore almeno tanto estremo quanto quello osservato quando è vera l'ipotesi nulla. Utilizzando il modello ipergeometrico, p_i è la probabilità di ottenere almeno il numero osservato di geni differenzialmente espressi, N_{de}, solo per caso.

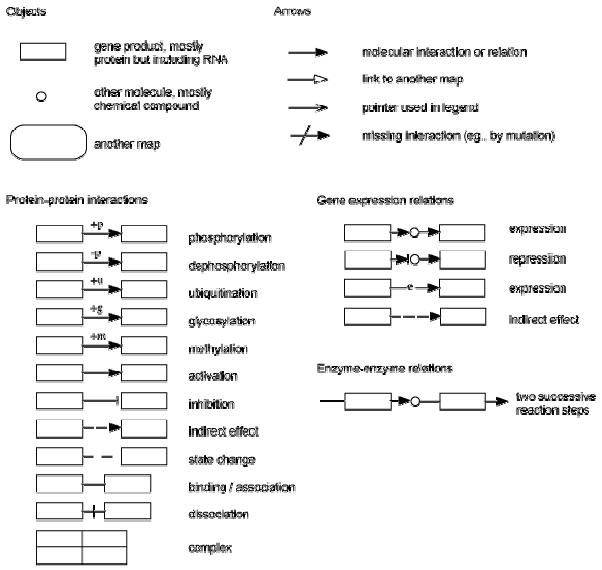


Figura 19: Legenda della rappresentazione dei diversi tipi di interazione tra entità geniche nei pathway (da KEGG)

Il secondo termine della formula (3.2) è una funzione sia delle caratteristiche specifiche dei particolari geni che si evidenziano come differenzialmente espressi che delle interazioni rappresentate nel pathway, vale a dire della topologia del pathway stesso. Al numeratore di questo termine compare la somma dei valori assoluti dei fattori di perturbazione (*PF*) per tutti i geni appartenenti a un dato pathway P_i. Accettando l'ipotesi nulla, consistente nel fatto che la lista di geni differenzialmente espressi contiene soltanto geni casuali, la probabilità che un pathway abbia un valore elevato dell'impact factor è proporzionale al numero dei geni differenzialmente espressi che appartengono al pathway, che a sua volta è proporzionale alla grandezza del pathway. Quindi è necessario che in questo secondo termine il fattore di perturbazione sia normalizzato rispetto alla grandezza del pathway, dividendo il fattore di perturbazione per il numero totale di geni differenzialmente espressi, $N_{de}(P_i)$, appartenenti al pathway in esame.

Le varie tecnologie impiegate possono inoltre produrre risultati diversi della stima dei fold change: i fold change riportati dai microarray tendono ad essere più bassi rispetto a quelli riportati dalla RT-PCR, altra tecnologia utilizzata in laboratorio per amplificare e quantificare un DNA oggetto di studio [Draghici S. et al, 2006]. Per rendere l'impact factor quanto più indipendente possibile dalla tecnologia e anche confrontabile tra le diverse applicazioni, il secondo termine dell'equazione 3.2 viene diviso per il valore medio assoluto dei fold change $|\Delta E|$, calcolato per ogni gene

differenzialmente espresso. Assumendo che esistano alcuni geni differenzialmente espressi da qualche parte nel dataset, sia $N_{de}(P_i)$ che $|\Delta E|$ sono diversi da zero e quindi il secondo termine assume un valore convenientemente definito. Con un semplice passaggio matematico, basato sulle proprietà della funzione logaritmo, si potrebbe mostrare che gli impact factor corrispondono all'opposto del logaritmo della probabilità totale di avere un numero statisticamente significativo di geni differenzialmente espressi e una elevata perturbazione nel pathway in oggetto. È stato dimostrato che i valori di impact factor, *IF*, seguono una distribuzione $\Gamma(2, 1)$ da cui possono essere calcolati i p-value come

$$p = (IF + 1) \cdot e^{-if} \qquad (3.3)$$

Di conseguenza l'accumulo di perturbazione nella rete è calcolato come risultante del contributo dei singoli geni.
Si definisce dunque l'accumulo del gene g_i come la differenza tra il fattore di perturbazione di un gene ed il corrispondente log fold-change:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \qquad (3.4)$$

Questa sottrazione è necessaria per assicurare che i geni differenzialmente espressi non connessi con nessun altro gene non forniscano alcun contributo all'accumulazione, perché questi geni sono già portati in conto nella sovrarappresentazione.
Si può dimostrare che il vettore delle accumulazioni di perturbazioni può essere ottenuto dalla relazione:

$$Acc = B \cdot (I - B)^{-1} \cdot \Delta E \qquad (3.5)$$

dove B rappresenta la matrice delle adiacenze normalizzate e pesate del grafo che descrive la rete di segnalazione genica, tenendo anche in conto i versi degli archi orientati:

$$\boldsymbol{B} = \begin{pmatrix} \frac{\beta_{11}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_1)} & \frac{\beta_{12}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_2)} & \cdots & \frac{\beta_{1\boldsymbol{n}}}{\boldsymbol{N_{ds}}(\boldsymbol{g_n})} \\ \frac{\beta_{21}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_1)} & \frac{\beta_{22}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_1)} & \cdots & \frac{\beta_{2\boldsymbol{n}}}{\boldsymbol{N_{ds}}(\boldsymbol{g_n})} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{\boldsymbol{n}1}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_1)} & \frac{\beta_{\boldsymbol{n}2}}{\boldsymbol{N_{ds}}(\boldsymbol{g}_2)} & \cdots & \frac{\beta_{\boldsymbol{nn}}}{\boldsymbol{N_{ds}}(\boldsymbol{g_n})} \end{pmatrix} \qquad (3.6)$$

I è la matrice identità e:

$$\Delta \boldsymbol{E} = \begin{pmatrix} \Delta \boldsymbol{E}(\boldsymbol{g}_1) \\ \Delta \boldsymbol{E}(\boldsymbol{g}_2) \\ \cdots \\ \Delta \boldsymbol{E}(\boldsymbol{g_n}) \end{pmatrix} \qquad (3.7)$$

Per l'analisi sono utili solo i pathway che presentano il determinante della matrice $I-B$ diverso da zero e questa condizione di non singolarità è ottenibile abbastanza semplicemente operando delle opportune trasformazioni sulla matrice B.
Dei 246 pathway di homo sapiens attualmente disponibili in KEGG, una grande maggioranza possiede questo requisito senza alcuna altra trasformazione. Le situazioni in cui i pathway hanno matrici simgolari non sono oggetto d'interesse per questa trattazione.
La perturbazione totale netta accumulata nel pathway è calcolata come:

$$t_A = \sum_i Acc(g_i) \qquad (3.8)$$

L'analisi di impatto estende ed arricchisce gli altri approcci statistici preesistenti, introducendo gli aspetti innovativi che sono stati sopra descritti. Ad esempio, il secondo termine del fattore di perturbazione di un gene, nell'equazione 3.1, fa aumentare il valore di PF dei geni che sono connessi da link diretti ad altri geni differenzialmente espressi. Come conseguenza aumentano i valori del fattore di perturbazione totale dei pathway nei quali i geni differenzialmente espressi sono localizzabili in un sottografo completamente connesso. È interessante notare che, se fossero imposte le limitazioni degli approcci preesistenti, come ad esempio quella di trascurare l'entità dei cambiamenti nelle espressioni geniche rilevate o di non tener conto delle interazioni regolative tra geni, allora l'analisi d'impatto si ridurrebbe ai metodi analitici classici e produrrebbe risultati identici. Ad esempio, se non ci fossero perturbazioni direttamente a monte di un dato gene, allora il secondo termine dell'equazione 3.1 sarebbe nullo e il fattore di perturbazione *PF* si ridurrebbe al valore del cambiamento rilevato nell'espressione genica.

L'accertamento della validità di qualsiasi metodo di analisi di pathway non è semplice, giacché non esiste un benchmark unico ed universalmente accettato per questo tipo di valutazioni. In mancanza di uno standard perfetto, le possibilità sono due: la valutazione dei risultati del metodo di analisi d'impatto alla luce delle conoscenze biologiche, ottenute per altre vie oppure il confronto dell'analisi d'impatto con gli altri metodi preesistenti sempre nell'ambito della conoscenza biologica. Ovviamente in queste condizioni risulterà impossibile calcolare dei valori esatti per la sensibilità, la specificità, oppure poter tracciare curve ROC. Tuttavia i metodi possono essere valutati e confrontati sulla base dei pathway che si rivelano significativi in una data condizione e considerando l'adattamento dei pathway significativi coinvolti in una data condizione a quanto previsto dalle conoscenze biologiche. Questo tipo di valutazione è ritenuta attualmente la migliore prassi in questo settore [Subramanian et al., 2005].

Sulle basi di un'ampia sperimentazione su molti dataset indipendenti ampiamente documentata in letteratura [Draghici S., 2011], è possibile affermare che l'analisi d'impatto è un metodo che riesce a fornire risultati più completi e utili sia dell'Over Representation Approach sia del metodo FCS per evidenziare i signaling pathway coinvolti in determinati fenomeni biologici e quindi per dedurre una conoscenza più approfondita dei fenomeni legati al metabolismo, alla farmacocinetica e all'innesco di patologie. Pathway Express, un tool freeware e web-based implementa il metodo SPIA

ed è una libreria di Onto-Tools [SIT22]. La fonte dei dati utilizzati in Pathway Express è da ricercarsi in KEGG. Pathway Express implementa anche un metodo di tipo ORA e permette così un confronto tra i risultati ricavati con i due metodi diversi. Questo strumento consente anche di eseguire delle query rapide di geni e pathway e la visualizzazione di pathway interi.

3.4 Caratteristiche delle serie temporali da microarray

Scopo principale di questo studio è l'analisi di uno specifico tipo di esperimenti microarray, in cui l'espressione genica sia misurata ripetutamente in istanti successivi di tempo in diverse condizioni biologiche. Le motivazioni allo studio dinamico dell'espressione genica possono essere fondamentalmente di tre tipi:

- comprendere i fondamenti dei fenomeni fisiologici e dei processi di sviluppo, come avviene ad esempio nello studio del ciclo cellulare.
- studiare la risposta degli organismi viventi a determinati stimoli o trattamenti.
- analizzare le reti di interazione tra geni, come i signaling pathways, per approfondire la conoscenza di molti aspetti della complessa interdipendenza tra le singole entità biologiche.

Dal punto di vista della progettazione sperimentale, e serie temporali possono essere classificate in base a:

- numero di rilievi, pari al *numero di punti* da cui la serie è formata. Si parla di serie temporale molto breve quando i punti della serie sono in numero compreso tra 3 e 6. Per un numero di punti compreso tra 6 e 12 la serie è considerata breve. Se i punti sono da 10 a 20 la serie è considerata media. Per valori maggiori la serie è considerata lunga: le serie temporali da microarray di questo tipo sono molto rare. Naturalmente questa distinzione vale nell'ambito dell'insieme delle serie temporali da microarray, in quanto tutte le serie appartenenti a questo insieme sono da ritenersi brevi se confrontate alle serie temporali comunemente studiate negli altri ambiti d'interesse scientifico.
- *numero di condizioni biologiche diverse* in cui avvengono i rilievi. Le serie temporali possono essere singole, se è studiato un solo gruppo sperimentale o multiple, se i gruppi oggetto di studio sono in numero maggiore di 1.
- *dipendenza o indipendenza* dei punti della serie. Nel primo caso si parla di serie longitudinale, che si rileva quando i rilievi sono fatti sullo stesso individuo ma in istanti di tempo successivi; nel secondo caso si dice che i dati sono rilevati trasversalmente, poiché si riferiscono ad istanti di tempo diversi e a individui diversi ed indipendenti tra loro.

Gli esperimenti che hanno per oggetto lo studio di processi biologici periodici, come il ciclo cellulare e i ritmi circadiani, sono rappresentati da singole serie temporali, lunghe e longitudinali.

Al contrario, gli esperimenti rivolti ad individuare le risposte a determinate sollecitazioni utilizzano serie temporali multiple, brevi e indipendenti. Lo scopo degli

esperimenti che fanno ricorso a serie temporali multiple è quello di analizzare le differenza tra le espressioni geniche di diversi gruppi sperimentali, che possono essere stati sottoposti a trattamenti diversi o possono essere stati prelevati da tessuti diversi. Le risposte agli stimoli sono tipicamente attese entro intervalli di tempo predefiniti ed è per tale motivo che le serie in oggetto possono essere brevi.

Infine, i dati relativi alle reti geniche sono di solito associati a serie temporali brevi o medie al fine di ricavare evidenze di co-espressione e co-regolazione in riferimento a un'ampia varietà di condizioni biologiche.

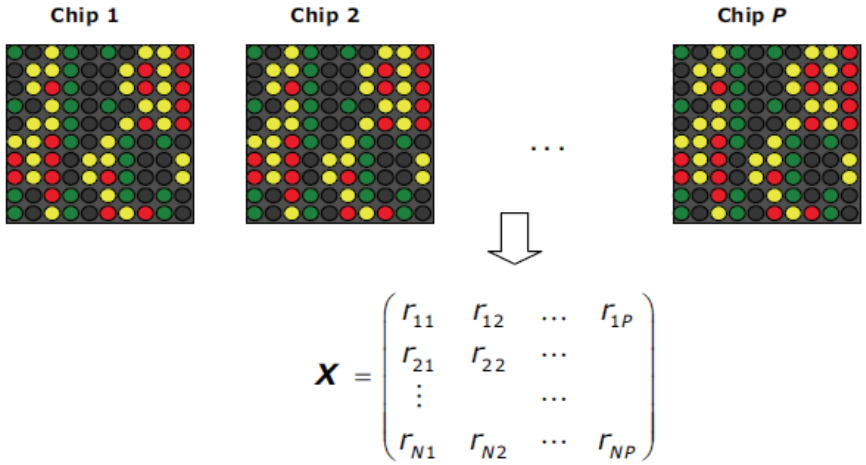


Figura 20: Microarray dopo l'ibridazione. Profili di espressione genica per N geni in P microarray. I punti colorati sono l'immagine di geni diversi che sono stati opportunamente marcati, l'intensità luminosa è legata all'espressione genica. In particolare in corrispondenza ad un rapporto di espressione genica <1 è visibile un colore verde, mentre se questo rapporto è >1 è visibile un colore rosso. Il colore giallo esprime una condizione in cui il rapporto è unitario. I valori di espressione genica sono rappresentati dalla matrice X, in cui le righe sono i geni e le colonne rappresentano la sequenza di chip microarray.

Un problema particolarmente critico per l'analisi di serie temporali da microarray è la selezione dei geni differenzialmente espressi: a causa della variazione nel tempo, infatti, non è possibile ritenere valide le considerazioni sulle espressioni differenziali relative al caso statico. A tale scopo sono stati proposti alcuni modelli statistici per l'individuazione dei geni differenzialmente espressi in serie temporali, la cui descrizione e sperimentazione è possibile reperire in [Storey J.D. et al., 2005, Ma P. et al., 2009, Kalaitzis A. A. et al. 2011; Gillespie C.S. et al., 2011]. Attualmente nelle librerie di Bioconductor [STI33] è possibile trovare due pacchetti in grado di evidenziare i geni differenzialmente espressi in serie temporali: `timecourse` e `limma`.

Il pacchetto `timecourse` fa un accertamento delle differenze di trattamento confrontando i profili medi delle serie temporali sia all'interno che tra punti temporali. L'algoritmo è basato sul modello di Bayes multivariato empirico proposto da [Tai Y.C et al., 2006] e, dopo aver calcolato la statistica di Hotelling, elabora un elenco dei geni da cui si possono estrarre i geni che hanno un'espressione preminente rispetto a quella degli altri.

Il pacchetto `limma` utilizza un approccio basato su t-test, come descritto in [Smyth, G. K., 2005] e su un approssimazione lineare del comportamento dei singoli geni, in cui i coefficienti dei modelli adattati descrivono le differenze degli RNA che sono rappresentati nel microarray. Ci sono poi molti modi per valutare i geni differenzialmente espressi, ad esempio considerando il loro "log-fold change", analogamente a quanto avviene nel caso statico.

3.5 Similarità ed algoritmi di clustering

Avendo premesso che la misura del grado di co-espressione dei geni è un passo fondamentale per l'analisi dei dati, va aggiunto che oramai da più di un decennio la ricerca si è attivata, mediante teorie, tecniche e strumenti diversi, per proporre, sperimentare e confrontare con varie misure di similarità i profili di espressione, intesi come delle particolari serie temporali [Wit E. el al., 2004].

D'altra parte le serie temporali di origine bioinformatica rappresentano attualmente intorno al 40% dei dataset di espressione genica disponibili e, a causa dell'elevatissimo numero di geni presenti in ogni rilievo sperimentale, si è presentata la necessità di sistematizzare i dati raggruppandoli secondo un criterio di clustering. Riguardo al particolare metodo di clustering da scegliere affinché dall'analisi possano scaturire informazioni quanto più numerose e di buona qualità, sono state avanzate proposte molto disparate e sono state sviluppate centinaia di algoritmi, che possono essere catalogati in gruppi, a seconda del criterio di similarità su cui sono fondati. Una prima grande distinzione è quella tra algoritmi discriminativi e algoritmi generativi.

Gli *algoritmi discriminativi* sono basati sulla definizione preliminare di funzioni di similarità per coppie di elementi. Applicando al dataset queste funzioni si riescono ad individuare dei raggruppamenti di punti. Nel caso specifico delle serie temporali di dati di tipo microarray, gli oggetti da inserire nei cluster sono i profili di espressione dei singoli geni.

Gli algoritmi *discriminativi basati su similarità puntuale*, considerano i profili di espressione genica come successioni di *n* punti temporali visti come dei vettori in uno spazio ad *n* dimensioni e definiscono una metrica, applicando particolari funzioni di distanza o di correlazione a questi vettori per quantificare la distanza tra due profili. Tra le metriche più comunemente utilizzate sono da menzionare la distanza Euclidea, la distanza di Manhattan e i coefficienti di correlazione lineare di Pearson [D'Haeseleer P. et al., 2005]. Alcuni algoritmi, come la maggior parte di quelli gerarchici [Eisen,M. B., 1998] seguono un procedimento aggregativo, poiché costruiscono cluster a partire dai singoli punti e poi, via via, da gruppi più piccoli ad agglomerati sempre maggiori; altri algoritmi, di cui il k-means e le self-organizing maps sono i più rappresentativi, al contrario, lavorano per divisioni successive, spezzando l'insieme di tutti i data point in cluster sempre più ristretti [Bergkvist A., 2010]. Tutti questi algoritmi forniscono buoni risultati nell'applicazione a dati di microarray statici [Boutros P., 2005] e sono generalmente poco complicati e di semplice implementazione e sono stati i primi ad essere implementati per le serie temporali da microarray meno recenti [Eisen M. B.,

1998]. Tuttavia nonostante queste positive caratteristiche, gli algoritmi discriminativi basati sulla similarità punto a punto non si prestano ad un'analisi di dati come le serie temporali, in quanto sono stati elaborati nell'ipotesi che le misure di espressione eseguite su un determinato gene siano indipendenti ed identicamente distribuite [Moller-Levet C. S., 2003]. Come già detto in precedenza, invece, quest'ipotesi non sussiste per i dati di serie temporali, in cui i campioni che si susseguono sono fortemente correlati tra loro [Bar-Joseph Z., 2004]. Quindi, in altri termini, gli algoritmi discriminativi puntuali ignorano il significato dinamico delle serie temporali, giacché l'ordine dei dati non è tenuto in nessun conto.

Gli algoritmi *discriminativi basati su similarità delle caratteristiche (feature)* non processano i raw data dei profili di espressione ma un insieme di feature estratte dai dati stessi. Pertanto prima avviene una trasformazione dei vettori di espressione genica in vettori di caratteristiche (feature vector) e poi a questi ultimi vettori sono applicate delle strategie di clustering simili a quelle discusse per la similarità puntuale. Idealmente il feature vector deve inglobare gli aspetti più rilevanti di un profilo di espressione e quindi è particolarmente interessante e delicata la scelta delle caratteristiche più rappresentative da estrarre. Alcuni hanno proposto semplicemente di indicare con −1, +1, or 0 i geni a seconda della loro sottoespressione, sovraespressione o espressione non differenziale [Di Camillo et al., 2005]. Altri hanno seguito un approccio molto diverso, trasformando ogni vettore di espressione genica in un vettore traiettoria in cui ogni termine indichi i tre possibili cambiamenti, crescenza, decrescenza o costanza, tra due punti consecutivi dell'espressione: per una serie temporale formata da N punti, ci sono N – 1 cambiamenti e 3^{N-1} possibili traiettorie [Phang et al., 2003]. Altri ancora aggiungono a questa sequenza di differenze del primo ordine anche una sequenza di differenze del secondo ordine (concavità o convessità o nessuna curvatura) e combinano entrambe le seuqenze in un solo vettore di caratteristiche [Kim et al., 2007].

È stato poi elaborato un metodo che combina gli approcci di Di Camillo e di Phang: in questo metodo i feture vector contengono sia i livelli di espressione relativa rispetto a un valore di riferimento prefissato (sottoespressione, sovraespressione o espressione non differenziale) sia le differenze di coppie di punti successivi (salita rapida, salita lenta, nessun cambiamento, discesa lenta, discesa veloce), [Phan S. et al. 2007].

Gli algoritmi StepMiner [Sahoo D. et al., 2007] and SlopeMiner [McCormick K. et al., 2008] sono basati sull'ipotesi che un profilo di espressione è completamente definito dalle transizioni dei livelli di espressione. Nel primo algoritmo, ogni vettore di espressione è ridotto ad un semplice pattern binario (salita, discesa, salita e discesa, discesa e salita o nessun cambiamento) e, secondo il pattern assegnato, a uno o due istanti di tempo in cui avvengono le transizioni. Il secondo algoritmo consente anche di considerare transizioni progressive e non soltanto di tipo a gradino. Agli algoritmi discriminativi basati su similarità delle caratteristiche sono associati molti e attraenti vantaggi: sono più veloci di quelli della categoria di clustering basato sui dati raw, perché il vero passo di clustering è effettuato su un vettore semplificato in luogo di quello originario; il passo di estrazione delle feature solitamente riduce il rumore presente nei dati originari e rende quindi più robusti questi algoritmi [Wang et al.,

2008]; ancor più importante, questi metodi sono notevolmente flessibili e quindi permettono di ridurre complessi profili di espressione solo alle caratteristiche ritenute essenziali. Di conseguenza, i geni sono confrontati sulla sola base degli aspetti davvero interessanti dei loro profili e così si riescono a generare dei cluster significativi.

A fronte di questi notevoli vantaggi, però gli algoritmi basati su estrazione di caratteristiche sono sicuramente soggetti a perdite di informazione e anche al pericolo di falsare l'analisi. Infatti, nella selezione delle caratteristiche che sintetizzano i vettori si prevedono già in qualche modo i pattern cui si vuol giungere e quindi si perde la possibilità di osservare similarità e pattern inaspettati. Per evitare questi rischi si possono utilizzare diversi set di caratteristiche nello stesso tempo: questa soluzione è praticabile in quanto questo tipo di algoritmi è relativamente veloce. La soluzione ideale sarebbe quella di sviluppare un algoritmo automatico di feature extraction che selezioni le caratteristiche ottimali da un insieme di possibili caratteristiche. Nell'ambito del Machine Learning questi algoritmi sono conosciuti come filtri o wrapper e nelle applicazioni al clustering di microarray i filtri, più che i wrapper, troppo impegnativi da un punto di vista computazionale, potrebbero agire sulle annotazioni provenienti dai database di Gene Ontology per scegliere il sottoinsieme di caratteristiche che meglio riescano a raggruppare geni conosciuti come co-regolati [Xing, E. P. et al., 2001] .

Una terza categoria è quella degli algoritmi *discriminativi basati su similarità delle forme.* Si pensi ad esempio ad un algoritmo che identifica ed associa tra loro i profili che hanno una data forma, indipendentemente dal fatto che la forma possa essere invertita o traslata nel tempo [Quian et al., 2001]. La procedura di questi algoritmi è basata sull'algoritmo di Smith-Waterman per l'allineamento locale di sequenze ed assegna ad ogni coppia di profili di espressione un punteggio ed una relazione: simultanei, ritardati, invertiti o invertiti e ritardati. Il punteggio, eventualmente pesato per la relazione, può essere considerato una misura di similarità e può quindi generare i cluster di geni. Molte sono state le proposte per migliorare questo algoritmo: ad esempio sviluppare un algoritmo simile a BLAST per velocizzare il processo di individuazione del massimo allineamento locale delle forme [Balasubramaniyan R., 2005]; oppure servirsi di feature vector che rappresentano i cambiamenti nei livelli di espressione genica e confrontare le forme di questi ultimi vettori [He et al., 2006].

Il vantaggio maggiore di questi algoritmi basati sulla similarità delle forme consiste nella loro capacità di identificare come affini due profili di espressione anche se sono traslati e/o invertiti. Da un punto di vista biologico la relazione di traslazione è corrispondente alla regolazione di un gene da parte di un altro gene o a un gene che presenta un ritardo nella risposta allo stesso fattore di trascrizione, mentre la relazione di inversione è significativa di un meccanismo regolatorio che attiva un gene e ne inibisce un altro. Questi algoritmi hanno il vantaggio di rivelare connessioni sconosciute tra geni, perché sono in grado di evidenziare similarità che sfuggono ad altri metodi di clustering. Tuttavia potrebbero essere migliorati per quanto riguarda la presenza di gap nelle serie temporali, causate da campionamento non uniforme e ricorrere al Dinamic Time Warping (DTW) per cogliere le similitudini anche nelle situazioni in cui i profili possano presentare delle deformazioni, compressioni e dilatazioni. Data il collegamento

del DTW con l'algoritmo sviluppato in questa tesi, questo argomento sarà trattato più diffusamente nell'appendice.

Il maggior inconveniente di questo tipo di algoritmi è la loro complessità computazionale e la loro lentezza, giacché il processo per individuare il miglior allineamento locale deve essere ripetuto più volte per creare i cluster. Quindi è necessario mettere a punto un approccio euristico che renda più veloce la ricerca senza troppo compromettere l'accuratezza [Balasubramaniyan R., 2005].

Gli *algoritmi generativi*, invece, partono dal presupposto che i dati sono generati da un set finito di modelli. Questo tipo di algoritmi utilizza i dati come training set per individuare i parametri che caratterizzano i modelli e in seguito crea raggruppamenti di punti generati dallo stesso modello. Essi possono essere suddivisi in due gruppi: basati sul template e basati sul modello.

Negli *algoritmi generativi basati sul template* ogni vettore di espressione genica è messo nel cluster corrispondente a un particolare template, o profilo candidato, che meglio lo descrive. Sebbene questo tipo di algoritmi siano molto simili ai metodi basati sulla similarità delle caratteristiche, tuttavia se ne discostano, perché questi non lavorano misurando la similarità tra una coppia di espressioni, ma assegnando ad ogni espressione al template più adatto.

Gli algoritmi basati sul template si differenziano tra loro prima di tutto per il pre-processing e per la metodologia di scelta dei profili candidati. Uno dei primi e più comuni algoritmi appartenenti a questa categoria è stato presentato in [Peddada S. D. et al. 2003] ed è particolarmente adatto ai casi in cui si presuppone di voler mostrare a quale gruppo appartenga un certo profilo. Il primo passo della procedura consiste nella definizione dei profili di diseguaglianza candidati come funzioni monotone decrescenti o cicliche. Il secondo passo utilizza tecniche statistiche per trovare la somiglianza di ogni profilo di espressione con uno dei candidati o con nessuno. Alcune varianti di questo algoritmo sono più rapide e prevedono anche una misura della significatività dei cluster oppure considerano dei vettori d'informazione più sintetici di quelli originali per ottenere un miglioramento delle prestazioni sui dati ad elevata variabilità. In molti casi, però, il compito del clustering e dell'analisi dei dati è quello di individuare nuove relazioni tra geni basate su pattern non noti e similarità e non avrebbe senso utilizzare profili predefiniti. Per questo sono stati proposti algoritmi basati sul template in cui al primo passo ogni vettore di espressione genica è trasformato in un vettore di pattern, che ne indica le crescenze e decrescenze.

La particolarità di questo algoritmo è che considera come profilo di template ogni possibile vettore di pattern e ogni gene è assegnato al cluster definito dal proprio vettore. Al crescere della lunghezza della serie temporale il numero di profili di template e quindi di cluster aumenta rispetto al numero di geni, fino a rendere inutile il lavoro di clusterizzazione. Applicando questo algoritmo a una serie temporale di 12 punti per 2000 geni risulterebbero $2^{12-1} = 2048$ cluster, vale a dire un numero maggiore di quello dei geni! Per superare queste incongruenze si potrebbe pensare a un algoritmo che selezioni un sottoinsieme rappresentativo ma di cardinalità contenuta dei profili di espressione da utilizzare come template. Ciò comporterebbe che per trovare il migliore di questi sottoinsiemi di cardinalità assegnata, cioè quello contenente i profili a coppie

più diversi uno dall'altro, bisogna risolvere un problema NP-hard, che potrebbe essere aggirato con un algoritmo greedy finalizzato a trovare un insieme buono ma non necessariamente ottimale. Questo algoritmo sarebbe però comunque enormemente lento.
In definitiva anche questo tipo di algoritmi è robusto al rumore e funziona particolarmente bene quando applicato alle serie temporali brevi, che, come visto, per il caso dei microarray sono sicuramente la grande maggioranza.

Gli *algoritmi generativi basati sul modello* sono i più rappresentativi della categoria. I dati in questo caso sono generati da un determinato modello parametrico i cui parametri sono stimati con tecniche statistiche, quasi sempre del tipo expectation-maximization. Ogni modello rappresenta un singolo cluster e un determinato gene è associato al cluster corrispondente al modello che genera profili di espressione più simili al suo.
Recentemente è stato proposto un certo numero di algoritmi di clustering progettati espressamente per serie temporali di espressioni geniche. Un esempio ne è l'approccio di clustering basato sulle dinamiche dei pattern di espressione [Ramoni et al., 2002], quello che utilizza rappresentazioni continue del profilo e [Bar-Joseph et al., 2003] e quello che utilizza un hidden Markov model (HMM) [Schliep et al., 2003]. Tali algoritmi funzionano bene solo per serie temporali maggiori di 10 punti perché, nel caso di serie con un numero di punti più esiguo, opererebbero un overfitting e potrebbero non riuscire a gestire la presenza del rumore.

CATEGORIA DI ALGORITMI	CRITERIO DI SIMILARITÀ	ALGORITMI TIPICI	PRO	CONTRO
Discriminativi	Distanza puntuale	• Gerarchici • K-means • Self-organizing maps	• Semplicità • Diffusione • Test positivo su dati statici	Non dinamicità
	Distanza tra feature	• StepMiner • SlopeMiner	• Veloci • Robusti • Adattabili	• Perdite • Bias
	Forma del profilo	DTW	Riconosce repliche deformate	Complessità computazionale
Generativi	Template-based	Algoritmi a due step	• Robusti • Adatti a serie brevi	• Complessità computazionale • Perdite
	Model-based	• HMM • Bayesiani	• Visione probabilistica • Robusti	• Lentezza • Bias • Complessità computazionale

Tabella 2: Quadro riassuntivo della classificazione degli algoritmi per l'analisi di serie temporali geniche

3.6 Ipotesi e finalità della sperimentazione

3.6.1 Corrispondenza tra sistemi biologici e grafi

La principale ipotesi su cui si fonda questa ricerca è che esistano meccanismi biologici che possono essere rappresentati mediante dei network biologici, nel nostro caso si tratterà di signaling pathway cellulari di homo sapiens (hsa), che sono coinvolti in un processo di reazione a una sostanza somministrata all'istante di tempo iniziale della sperimentazione.

Grafo sparso

I network biologici sono tipicamente alquanto sparsi e i link tra nodi sono significativi delle relazioni esistenti tra le diverse componenti cellulari e organiche.

Funzionalità di regolazione cellulare

La cellula può essere schematizzata come un sistema complesso le cui componenti, alcune migliaia di proteine, interagiscono assolvendo ognuna alla sua particolare funzione. In questo meccanismo d'interazione giocano un ruolo primario determinate proteine, dette trascrittori, che possono regolare, cioè sia attivare che inibire, la funzionalità di altre entità analoghe cui sono collegate nei pathway. In caso di funzione di attivazione, si parla di fattore di trascrizione promotore, in caso di funzione di inibizione, si parla di fattore di trascrizione repressore.

Fattore di efficienza regolatoria

Nella rappresentazione dei pathway di segnalazione cellulare ogni nodo rappresenta un gene, i segnali contengono l'informazione riguardante lo stato del sistema e gli archi sono corrispondenti alle interazioni tra geni stessi, come schematizzato nella seguente figura 21:

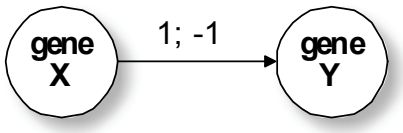


Figura 21: Grafo di un collegamento diretto tra due geni con archi pesati

Questo grafo elementare significa che la produzione del gene X è un fattore di trascrizione che influenza positivamente il grado di produzione del gene Y. Il valore 1 o – 1, attribuito agli archi, si riferisce alla funzione regolatoria di promotori o repressori. Questo valore è esattamente il fattore di efficienza regolatoria, che compare nell'espressione dell'Impact Factor descritto nel paragrafo 3.3 e su cui ci apprestiamo ad indagare.

Geni condivisi da più pathway

Uno stesso gene può essere coinvolto in più funzioni biologiche diverse e quindi lo stesso gene può comparire in due o più pathway distinti in cui ricopre un ruolo diverso. Nella nostra sperimentazione questa ipotesi è causa di sovrapposizione di segnali che si propagano attraverso i pathway e sarà oggetto di una trattazione specifica.

Numerosità degli archi uscenti e interazioni multiple tra geni
Un'altra particolarità, che emerge dall'analisi consiste nel fatto che i sistemi biologici reali spesso contengono nodi con un numero elevato di archi uscenti, individuati con il nome di "hub". Si può dedurre con varie considerazioni che gli archi uscenti da uno stesso nodo tendono ad avere lo stesso segno di regolazione. Al contrario un nodo può essere correlato contemporaneamente da un fattore di trascrizione positivo e da uno negativo.

3.6.2 Casualità delle serie temporali biologiche

Una seconda ipotesi va precisata relativamente alla caratteristica delle serie temporali rappresentative di profili genetici: esse sono di tipo stocastico e soggette pertanto ad errori casuali di due diverse origini:
a) Errori intrinseci al processo biologico, da cui le proteine sono generate;
b) Errori causati dal processo di estrazione dei dati con tecniche di qualsiasi tipo e in particolare di tipo microarray.
Gli errori derivanti dal processo biologico si fondano sulla notevole e imprevedibile variabilità della concentrazione di una singola proteina all'interno di un gruppo di cellule identiche. Conseguentemente il profilo di espressione genica è soggetto a variazioni stocastiche. La distribuzione di una proteina in una cellula ha un andamento spesso simile alla distribuzione gaussiana, risultante dal prodotto di variabili stocastiche. Ciò si verifica anche per la produzione delle proteine, che risulta dal prodotto dei processi di trascrizione e traslazione. La tipologia dei collegamenti tra nodi concorre ancora di più a questa variabilità. Nel caso di una struttura di geni, costituita da un ciclo con feedback negativo, le fluttuazioni vengono diminuite, mentre un'autoregolazione positiva fa aumentare queste fluttuazioni.

3.6.3 Segnali in ingresso ai pathway

Si ipotizza poi che i segnali d'ingresso siano funzioni a gradino, cioè monotone crescenti o decrescenti secondo il segno della regolazione. Nella regolazione rappresentata nella figura alla pagina precedente, ad esempio, se non c'è alcun segnale in ingresso X è inattivo e Y non è prodotto. In seguito a una sollecitazione, X commuta rapidamente nella sua forma attiva, e subito inizia anche la produzione del gene Y in rapporto costante con lo stimolo che applicato su X. Interrompendo repentinamente il segnale d'ingresso, il processo di produzione del gene Y presenta un decadimento di ordine esponenziale dallo stato di equilibrio precedentemente raggiunto. La situazione in caso di un repressore si può facilmente ricavare per analogia.

3.6.4 Costanti di tempo dei sistemi biologici

Un'altra ipotesi che deriva dall'analisi della dinamica delle regolazioni tra geni riguarda il tempo necessario, nei sistemi reali, a raggiungere lo stato d'equilibrio: si nota che si possono distinguere alcune dinamiche molto lente e altre molto più rapide.

3.6.5 Serie temporali brevi

L'ipotesi che si accetta riguardo alle serie temporali che saranno analizzate è che esse siano estremamente brevi, al massimo contenenti poche decine di elementi, che ogni loro elemento sia correlato con quelli lo precedono e lo seguono direttamente e infine che siano affette come già detto da rumore.

L'obiettivo principale di questo studio è alquanto più modesto di quello di voler ricostruire un'intera rete di relazioni tra geni: ci limiteremo ad andare ad isolare i geni che nell'esperimento si siano presentati con un p-value sempre al di sopra di una certa soglia per tutti gli elementi della serie temporale, poi a trovare quelli direttamente connessi in tutti i pathway che fanno parte della attuale conoscenza biologica su hsa e a controllare, con un modello che sarà esposto nel paragrafo 3.9, che il fattore di efficienza regolatoria si accordi in maniera accettabile alla similarità tra gli andamenti delle serie temporali che sarà valutata opportunamente mediante delle caratteristiche collegate sia alla forma che alla eventualità di fenomeni di time warping, naturalmente curando che i vincoli di causalità non siano in nessun caso violati.

Riassumendo in definitiva quanto fin qui detto, sarà eseguita un'analisi di similarità dei profili temporali di espressione genica di geni direttamente collegati tra loro, ritenendo vero, come affermato in [Farina et al., 2008] che la co-espressione sia un valido indicatore per la co-regolazione. In altri termini, quando una pluralità di geni mostra un profilo di espressione simile è molto probabile che essi siano bersaglio dello stesso tipo di fattore di trascrizione.

3.7 Dataset biologico e strumenti

Gli esperimenti di tipo microarray finalizzati alla generazione di serie temporali di espressione genica costituiscono da circa un decennio un metodo comunemente adottato per lo studio di una ingente quantità di processi biologici. La maggior parte di questi esperimenti può essere suddivisa in quattro grandi categorie:

a. gli esperimenti che si propongono di scoprire le dinamiche che si celano al di sotto di alcuni fenomeni biologici, come il ciclo cellulare e il ritmo circadiano [Androulakis I. P., 2007; Bar-Joseph Z., 2004].

b. gli esperimenti che si propongono di scoprire cambiamenti genetici che sottendono ai sintomi osservabili [Androulakis I. P., 2007; Bar-Joseph Z., 2004; Arbeitman et al., 2002].

c. la ricerca ha sfruttato i microarray per studiare patologie come l'Alzheimer [Ginsberg S. D., et al, 2000], l'HIV [Ross J. M., 2006] e il cancro [Whitfield M. L., 2002].

d. La quarta ed ultima include gli esperimenti di serie temporali microarray finalizzati alla determinazione delle risposte dei geni a vari stimoli, come ad esempio i "knockouts", condizioni di stress e somministrazione di farmaci [Stoughton R. B., 2005; Gasch, A.P. 2000]

Le serie temporali da microarray hanno la particolarità di essere molto brevi: esistono pochissime serie che superino i venti elementi, anche se raramente possono arrivare anche fino a qualche centinaio di elementi. La figura sottostante si riferisce al caso dello Stanford Microarray Database (SMD) nel 2004. SMD è attualmente la più ricca repository di risultati provenienti da microarray al mondo, con i suoi 82.542 set sperimentali di cui 1.461 di tipo serie temporale.

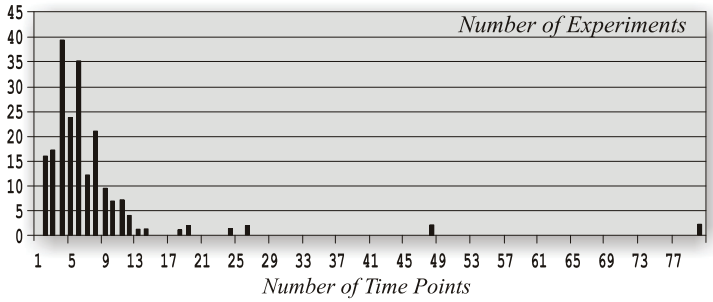


Figura 22: Distribuzione delle lunghezze delle serie temporali contenute nella raccolta della Stanford Microarray Database

Questo database contiene soltanto esperimenti eseguiti presso i laboratori di Stanford, ma è comunque abbastanza rappresentativo della realtà globale dei microarray e la situazione fotografata alcuni anni fa può ancora essere ritenuta un valido riferimento, in quanto la stragrande maggioranza dei dati genomici con evoluzione temporale è sempre formata da un numero di elementi molto contenuto e minore di 20.

Come indicato nella figura, più dell'80% delle serie temporali contiene un numero di punti minore o uguale di 8 ed esistono numerose ragioni per le quali le serie temporali sono formate da così pochi elementi: gli esperimenti che le producono richiedono molti microarray e in molti casi ogni punto è ripetuto almeno una volta e quindi i costi associati all'esecuzione della procedura sono molto elevati.

Anche se questi costi stanno diminuendo, la difficoltà di generare serie molto lunghe permane, a causa dell'oggettiva impossibilità di avere a disposizione determinati materiali biologici da campionare ad intervalli di tempo regolari. Ad esempio, si pensi all'eventualità di effettuare prelievi ematici a pazienti, affetti da una patologia, oltre un certo numero di volte.

Il particolare data set sul quale è stata eseguita la ricerca è il risultato di una sperimentazione del laboratorio del Cancer Institute, Karmanos Cancer Inst., Wayne State University, curato da Aliccia Bollig-Fischer. Attualmente questo dataset è reperibile pubblicamente all'URL [sit].

Si tratta di cellule della linea SUM-225 (Human Breast Cancer) trattate con HER-2-specific inhibitor CP 724,714 per 45 ore. Questo data set raccoglie i risultati del blocco della funzione HER-2 oncogeno chinasi nelle cellule SUM-225mediante trattamento con CP724,714 e misurando l'espressione genica in funzione del tempo. In questo modo è possibile ottenere informazioni circa i geni regolati da HER-2 in questa linea di cellule di cancro al seno.

L'RNA totale è stato raccolto ad intervalli regolari di tre ore da colture parallele durante l'arco di 45 ore di trattamento ed è stata eseguita sull'RNA un'analisi di

espressione dell'intero genoma in ogni rilevo, per un totale di 16 punti temporali, a iniziare dall'ora 0 e finendo all'ora 45.

Protocollo di crescita: le cellule SUM-225 allevamento su mezzo Ham's F-12 con aggiunta del 5% di siero bovino fetale, insulina (5 µg/ml)e idrocortisone (1 µg/mL).

Protocollo di trattamento: le cellule sono state trattate con l'aggiunta di 1uM CP724,714 (Pfizer Inc, Groton, CT) alla media di cultura per 45 ore totali.

Protocollo di estrazione: L'RNA è stato estratto ogni tre ore dalle colture parallele mediante QIAGEN Rneasy Plus kit secondo le indicazioni prescritte dal protocollo allegato al kit. Il controllo di qualità è stato effettuato con Agilent Bioanylizer e Agilent RNA 6000 Nano Kit.

Protocollo di etichetta: il cRNA biotinilato è stato preparato con il kit TotalPrep-96 RNA Amplification.

Protocollo di ibridizzazione: Illumina Standard.

Protocollo di scansione: Illumina Standard.

Pre-processing dei dati: i dati sono stati normalizzati mediante la funzione di normalizzazione quantile contenuta Bioinformatics Toolbox di Matlab.

Definizione di valore: quantile normalizzato.

I dati grezzi sono contenuti in un file in formato tabellare. Essi sono essenzialmente dei campioni provenienti da 24.527 probe dei microarray Illumina, ricavati ad intervalli costanti di tempo di tre ore per un totale di 45 ore. Ne è risultato un set di 24.527 serie temporali formate ognuna da 16 punti, il che qualifica le presenti come serie temporali geniche di lunghezza al di sopra della media. A ciascuno dei valori campionati è associato il corrispondente p-value, per consentire una valutazione della qualità dei dati. In particolare il p-value indica quanto probabile (valori alti) o improbabile (valori bassi) è l'eventualità di osservare esattamente un certo valore s_n della statistica test S_n, sotto l'ipotesi nulla. Nella figura 23 sono rappresentate le colonne iniziali della tabella in cui sono memorizzati i dati grezzi di alcune serie. Nella prima colonna compare l'identificativo del probe di provenienza, nella seconda colonna il simbolo del gene, nella terza, quinta e settima colonna i valori di espressione genica rilevati all'istante iniziale, dopo tre e dopo 6 ore; nella quarta, sesta e ottava colonna compaiono i p-value delle rispettive espressioni.

PROBE_ID	SYMBOL	SAMPLE 1	SAMPLE 1 Det Pval	SAMPLE 2	SAMPLE 2 Det Pval	SAMPLE 3	SAMPLE 3 Det. Pval
ILMN_1809034	15E1.2	673.3592	0	543.5964		498.2716	0
ILMN_1660305	2'-PDE	1145.298	0	942.9919		791.9109	0
ILMN_1762337	7A5	292.0536	0	206.3636	0.07142857	195.5109	0.125
ILMN_2055271	A1BG	237.4949	0.02197802	214.1104	0.03809524	234.5176	0
ILMN_1814316	A2BP1	186.5772	0.3469388	182.9526	0.2333333	156.0063	0.789557

Figura 23: Veduta parziale della tabella in cui sono raccolte le serie temporali geniche. Si osservi che i valori riportati sono quelli ancora non normalizzati.

In questa tabella sono riportate le espressioni geniche originarie e non ancora sottoposte al processo di normalizzazione. I vettori riga della matrice individuano i singoli geni e i valori dei dati sono le serie temporali da analizzare.

Nei seguenti grafici sono visualizzati andamenti nel tempo tipici delle serie temporali corrispondenti ai profili di espressione di geni differenzialmente espressi sia nella versione originale che dopo il processo di denoising:

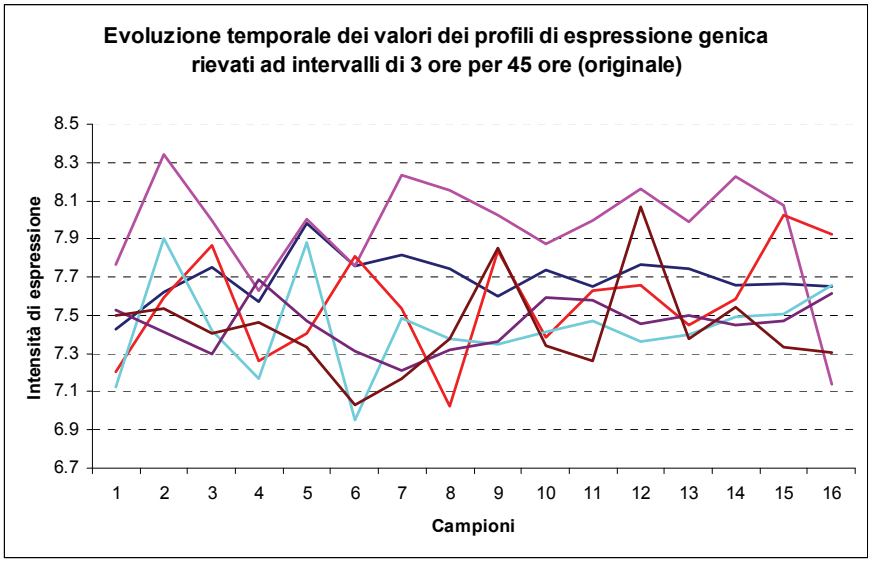


Figura 24: Rappresentazione di alcune serie temporali di espressione genica (versione originale)

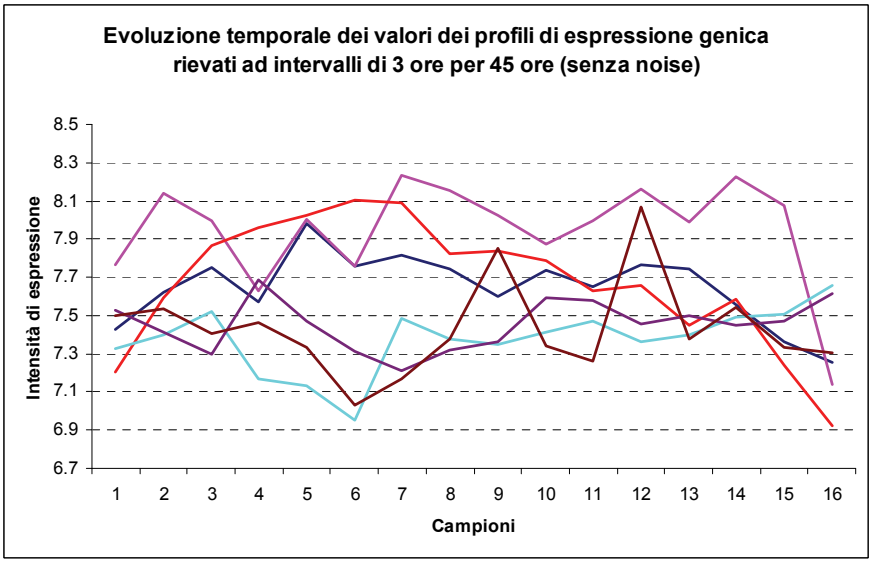


Figura 25: Rappresentazione di alcune serie temporali di espressione genica (versione senza noise)

L'elaborazione dei dati è stata eseguita in linguaggio R, per le specifiche proprietà nel calcolo statistico e perché interfacciabile con la libreria SPIA (Signaling Pathways Impact Analysis), che implementa l'omonimo algoritmo, descritto in [Tarca et al., 2009, Khatri et al., 2007 e Draghici et al., 2007]. Ulteriori dettagli su questo algoritmo sono reperibili in Appendice.

3.8 Calcolo delle relazioni tra coppie di geni

La motivazione della ricerca, già in parte anticipata nel corso dei paragrafi precedenti, è quella di confrontare coppie di geni direttamente collegati nei signaling pathway, per verificare che la co-regolazione rilevata corrisponda al coefficiente di efficienza regolatoria proposto nel modello di "impact analysis". Diamo di seguito una sintetica descrizione dello schema sperimentale che ha consentito di raggiungere l'obiettivo principale di questa ricerca.

Le successive elaborazioni richiedono che le serie temporali normalizzate siano opportunamente pre-processate secondo lo schema della figura 26.

Per quanto riguarda il p-value, al fine di rendere affidabile e significativo il procedimento di estrazione delle informazioni, è realizzato un primo passo, volto a eliminare i probe che presentino in più della metà dei valori della serie temporale un p-value <0.1. Questo criterio è stato suggerito dalla esigenza di conciliare un rifiuto erroneo dell'ipotesi nulla nel 10% dei casi e in non più della metà dei campioni registrati per una serie temporale con la circostanza di selezionare un numero di probe grande abbastanza da conferire a questo esperimento un livello elevato di significatività.

Figura 26: Schema a blocchi della fase di pre-trattamento dei dati.
Con *M* è indicata la tabella o matrice d'ingresso mentre $s_c(t)$ è la matrice di serie temporali condizionate, dalla quale sono state escluse le serie temporali non soddisfacenti i vincoli sulla soglia di significatività e le serie che rappresentano geni non differenzialmente espressi. Alla fase di selezione segue anche uno stadio di filtraggio in cui è attenuato il rumore.

Il passo successivo consiste nell'applicazione dell'algoritmo *limma*, per selezionare le serie temporali corrispondenti a geni differenzialmente espressi, come descritto alla fine del paragrafo 3.4. Dopo l'esecuzione di questo passo, nel dataset saranno contenuti soltanto i geni che abbiano reagito al trattamento farmacologico in modo più netto ed evidente.

Tra la procedura di selezione dei geni differenzialmente espressi e il trattamento di denoising è prevista una selezione dei geni che possano essere effettivamente interessanti ai fini dell'analisi della regolazione tra coppie di geni, cioè dei geni che

subiscono l'influenza di un solo gene a monte oppure sono soggetti a regolazioni che abbiano effetti concordi. A tale scopo è eseguita un'ulteriore procedura selettiva il cui schema è riportato qui sotto:

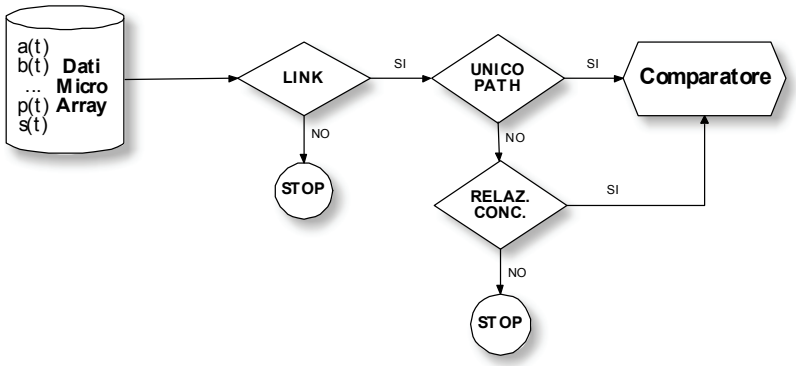


Figura 27: Schema di selezione dei geni interessanti per l'analisi di similarità.

Il primo controllo è eseguito sull'intero insieme di geni differenzialmente espressi che sono stati precedentemente selezionati. Se un gene non risultasse collegato a nessun altro gene, allora sarebbe ritenuto non interessante per una ricerca sulle relazioni e quindi sarebbe scartato.

Dei geni rimanenti, si fa una distinzione tra sorgenti e destinazioni della relazione di regolazione. Il test seguente viene eseguito sui geni destinazione, ovvero su quelli che sono condizionati da altri geni che occupano una posizione immediatamente a monte di essi. Questa considerazione è legata alla rilevanza topologica della collocazione dei nodi nel grafo rappresentativo dei signaling pathway. Orbene, se un gene risulta condizionato da più di un gene a monte, allora si ispezionano i tipi di regolazione che lo influenzano, ad esempio "activation", "expression", "phosphorylation", "repression" ed altri, rilevabili dai pathway riportati nella raccolta di KEGG [SIT24] e perfettamente schematizzati in matrici di adiacenza nella libreria SPIA di R. Se le relazioni sono concordi, come ad esempio "activation" e "expression" oppure come "repression" e "inhibition" allora il gene è conservato e destinato all'analisi di similitudine. Se invece le relazioni con i geni immediatamente a monte sono di segno opposto, come "activation" e "inhibition", allora quel gene è scartato, perché si vuole evitare di dover discutere e quantificare l'effetto di due o più meccanismi di regolazione contrastanti.

Il blocco finale di denoising è inserito per tener conto del fatto che i diversi fattori di disturbo, che sono stati descritti nei paragrafi precedenti, si sovrappongono al segnale genico in modo da determinarne una contaminazione casuale. La serie temporale vista come funzione dinamica è la sovrapposizione di una componente di segnale, quella che si vuole effettivamente analizzare, che varia lentamente nel tempo, e una componente stocastica che presenta una dinamica molto più rapida.

Per meglio cogliere la differenza delle velocità di transizione tra segnale e rumore, risulta più agevole trasporre il ragionamento nel dominio della frequenza, tramite trasformate di Fourier dei segnali. In generale, da quanto si conosce sui processi biologici, ci si aspetta che il segnale sia limitato in banda, mentre il rumore ha banda praticamente infinita. Inserendo quindi nella pipeline di processo un filtro passa basso

con frequenza adeguata a quella del segnale, si può avere la possibilità di tagliare fuori tutte le frequenze non contenute nella banda del segnale, isolando quasi completamente l'effettiva porzione significativa di segnale. La scelta più adeguata per il denoising di serie temporali, basandosi su considerazioni di confronto specifiche annotate in [Zanini, 2011], è quella di un filtro di Savitzky-Golay. È stato dimostrato che l'errore, nel caso di processo della serie temporale con questo filtro, è di un ordine di grandezza inferiore, rispetto a quello raggiunto dai altri tipi di filtri messi a confronto, come quello di Butterworth, di Chebyshev, di Cauer, con fitting lineare o quadratico e a media di 5 valori con finestra scorrevole. I particolari in merito al funzionamento del filtro di Savitzky-Golay sono reperibili in appendice.

Il confronto tra coppie di geni opportunamente selezionate, da cui dovrà risultare verificata la relazione di co-regolazione, come attivazione o inibizione, è uno degli argomenti di più ampia discussione sulle serie temporali di origine biologica.

Nel paragrafo 3.5 è stata condotta una breve rassegna degli algoritmi per il confronto di serie temporali geniche e il clustering o la classificazione che ne risultano.

L'idea più immediata ed elementare per quantificare la relazione tra due dinamiche casuali è quella di calcolare il coefficiente di correlazione tra le due serie temporali coinvolte nel confronto. Un'analisi di correlazione è stata appunto la prima esperienza di questa ricerca.

Applicando al set di geni opportunamente pre-condizionati un calcolo di coefficiente di correlazione lineare si sperava di riuscire ad evidenziare il risultato previsto. Per tener conto dei tempi di propagazione dei segnali biologici dal gene a monte al gene direttamente collegato a valle, è stata contemplata anche la correlazione della serie temporale relativa al primo gene con versioni traslate fino a 9 ore della serie temporale relativa al secondo gene. Le prove eseguite, però, hanno rivelato l'inadeguatezza dello strumento scelto. Con questo metodo non si riescono in nessun caso a distinguere apprezzabilmente le azioni di attivazione da quelle di inibizione, come deducibile dai grafici presentati di seguito.

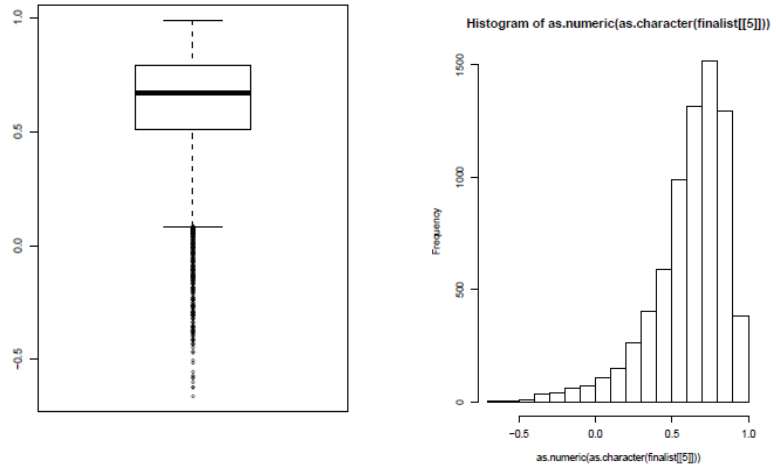


Figura 28: Boxplot e istogramma della distribuzione delle correlazioni tra geni direttamente collegati da cui ci si aspetta una relazione di attivazione.

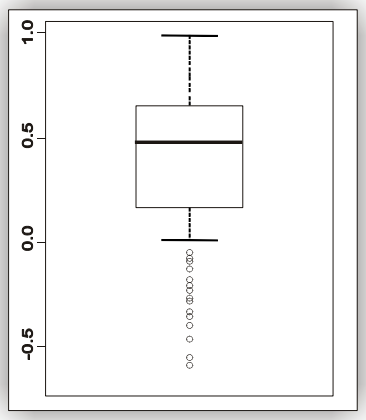


Figura 29: Boxplot della distribuzione delle correlazioni tra geni direttamente collegati da cui ci si aspetta una relazione di inibizione, non si osservano significative variazioni rispetto al caso precedente.

La spiegazione di questo risultato risiede nella natura stessa dell'indice di correlazione. Per la precisione, indicatori di correlazione lineare tra serie, come l'indice di correlazione di Pearson e simili, non garantiscono buone prestazioni nella valutazione della co-regolazione genica in quanto sono adatti a rivelare una dipendenza lineare e monotonica, ma non riescono a cogliere legami di dipendenza non lineare, come quelli esistenti tra le serie temporali in oggetto.

Notevole è la quantità di letteratura dedicata nell'ultimo decennio al confronto tra serie temporali prodotte da esperimenti di tipo microarray, come sinteticamente riferito nel paragrafo 3.5. Tuttora nuovi modelli sull'argomento continuano ad essere frequentemente messi a punto e pubblicati, sebbene stia diffondendosi rapidamente l'idea che non è più necessario elaborare ulteriori modelli per riscontrare similarità nei geni, ma basta cercare di sfruttare al meglio quelli già esistenti. L'obiettivo diventa quello di adattare ad ogni diverso tipo di analisi il modello di confronto più adeguato.

Nel caso della verifica del fattore di efficienza regolatoria l'analisi è riducibile al confronto diretto di due sole serie temporali (pairwise comparison) e alla ricerca della similarità tra le due serie.

Sono stati vagliati in particolare tre approcci distinti, al fine di valutare quale sia il più efficace nello studio di similarità in questione. Poi, sulla considerazione dei particolari risultati ottenuti con le tre metodologie, sono stati proposti e testati alcuni schemi di combinazione dei risultati stessi, volti a cogliere le possibili sinergie e a conseguire delle più ampie ed approfondite valutazioni di similarità.

3.8.1 Metodo differenziale

Data la natura dei fenomeni biologici, una serie potrà essere una versione traslata, attenuata e deformata dell'altra, ma ne potrebbe ricordare l'andamento, ben sintetizzato dalle "crescenze" e dalle "concavità" della funzione discreta che la rappresenta. Ciò è da

riferire anche a versioni traslate di due o tre intervalli temporali, lunghi 3 ore, delle forme d'onda che rappresentano la serie temporale.

Per questi motivi è parsa naturale l'idea di privilegiare il confronto dell'andamento qualitativo della serie temporale inserendo nei suoi parametri rappresentativi un' informazione differenziale, che consenta di portare in conto crescenze, decrescenze, concavità e convessità. Per la precisione, per ogni tratto che unisce due punti temporali è memorizzata una coppia di bit, che è 00 se la differenza tra i valori estremi di quel tratto è positiva, 11, se è negativa e 01 se è nulla. Lo stesso ragionamento si ripete sulle differenze delle differenze, o differenze del secondo ordine. In tal modo è possibile sintetizzare un'informazione sulla concavità o convessità della forma d'onda.

Il risultato di questa sintesi è una stringa di 58 bit, 30 per le differenze del primo e 28 per quelle del secondo ordine. Il confronto delle serie temporali della coppia di geni direttamente connessi si riduce quindi a un confronto tra due stringhe di bit, con al più uno shift per tenere conto dell'eventuale ritardo di regolazione dovuto alle costanti di tempo delle reti biologiche. Il risultato di questo confronto e quindi la bontà della divisione operata sulla base dei differenziali della funzione che rappresenta la serie temporale può essere valutata con criteri propri dell'Information Retrieval. Il confronto tra due geni, infatti può dar luogo a quattro esiti diversi:

1. correlazione positiva in corrispondenza di una effettiva regolazione positiva (Vero positivo VP);
2. correlazione positiva in corrispondenza di una effettiva correlazione negativa (Falso Positivo FP);
3. correlazione negativa in corrispondenza di una effettiva regolazione negativa (Vero Negativo VN);
4. correlazione negativa in corrispondenza di una effettiva regolazione positiva (Falso Negativo FN).

Sono ritenute rilevanti tutte le coppie di geni individuate dal dataset dopo il pre-processing.

Nel nostro caso la classe delle relazioni positive è molto più vasta di quella delle relazioni negative, quindi, invece di ricorrere a indici come Precison e Recall e Accuratezza si preferisce calcolare il coefficiente di correlazione di Mattew (Matthew Correlation Coefficient, MCC), particolarmente indicato nel caso di classificazione binaria asimmetrica, con notevole differenza di grandezza delle classi. Il coefficiente di correlazione di Matthews indica essenzialmente il rapporto tra la classificazione predetta e quella osservata e può assumere valori compresi nell'intervallo [−1;1]. Assume valore −1 in corrispondenza di una predizione opposta al valore osservato, +1 in corrispondenza di una predizione corretta e 0 nel caso in cui il classificatore si comporti come se fosse avvenuta una scelta casuale della classe di appartenenza. Questo coefficiente può essere calcolato direttamente dalla matrice di confusione mediante la formula:

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

Se qualche fattore al denominatore è nullo, la frazione perderebbe di significato. In tal caso il denominatore è posto arbitrariamente uguale a uno.

A conferma di quanto fin qui detto, si presentano in forma di istogramma i valori della Accuratezza, Precisone, Recall e coefficiente di correlazione di Matthews nel caso del dataset in esame e dell'utilizzo della procedura di selezione con approccio qualitativo appena descritta.

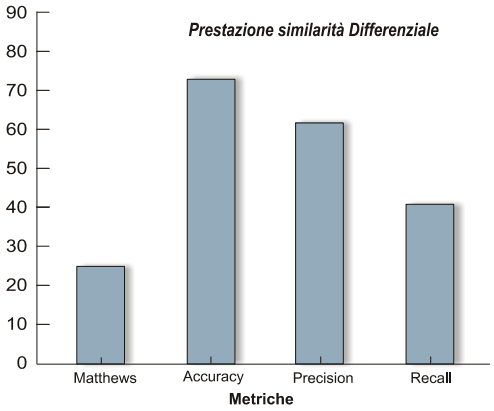


Figura 30: Rappresentazione della qualità del classificatore con approccio qualitativo relativa agli indici Accuratezza, Precision, Recall e indice di Matthews.

3.8.2 Metodo Dinamic Time Warping

Dinamic Time Warping, uno degli algoritmi discriminativi basati su similarità delle forme o di *pattern matching*, menzionato nella disamina del paragrafo 3.5 e diffusamente trattato in appendice. Mutuato dalle pratiche di riconoscimento vocale e motorio, questo algoritmo permette l'allineamento tra due sequenze e fornisce una misura di distanza tra le due sequenze una volta allineate. E' adatto specificamente per trattare confronti di forme d'onda con caratteristiche tempovarianti e per le quali la semplice espansione o compressione lineare non porterebbe conclusioni accettabili.

In generale, DTW è un metodo che riesce a scoprire una corrispondenza ottima tra due serie temporali, attraverso una distorsione non lineare rispetto alla variabile indipendente (tipicamente il tempo). Devono però essere soddisfatti alcuni vincoli per il calcolo della corrispondenza tra serie temporali: deve essere garantita la monotonicità nelle corrispondenze, ed il limite massimo di possibili corrispondenze tra elementi contigui della sequenza.

In particolare, l'idea di sfruttare questo tipo di algoritmo per le serie temporali di espressione genica è presentata in [Aach et al., 2001], con risultati che pongono in risalto i miglioramenti apportati da esso nel trovare comportamenti similari delle serie temporali rispetto ad altri tipi di clustering.

Ulteriori dettagli sull'algoritmo e un frammento di macrocodice relativo all'implementazione di questo algoritmo sono reperibili nell'appendice. Qui di seguito sono riportati gli istogrammi significativi della qualità della classificazione basata su DTW.

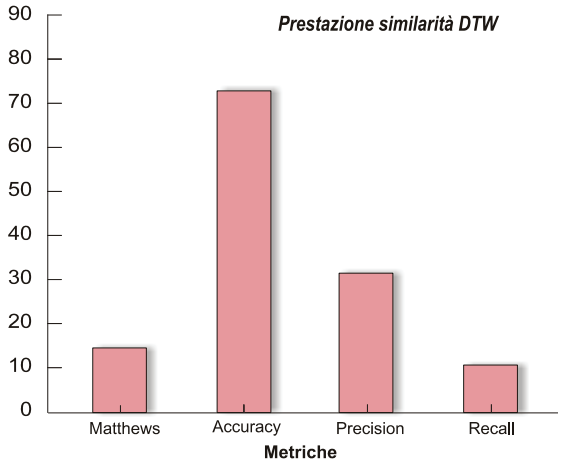


Figura 31: Rappresentazione della qualità del classificatore con approccio Dinamic Time Warping relativa agli indici Accuratezza, Precision, Recall e indice di Matthews

3.8.3 Metodo della Componente Spettrale Dominante

Il terzo metodo di studio della similarità si discosta alquanto dai due precedenti. Si tratta di un'analisi basata sulla Componente Spettrale Dominante (CSD). Come deducibile dal modello presentato in [Yeung et al., 2004], dalla decomposizione spettrale dei profili di espressione può scaturire un confronto di correlazione tra componenti in frequenza, capace di cogliere legami che sfuggono completamente alla correlazione lineare tradizionale. Inoltre, nel caso di geni che subiscono influenze contemporanee da diversi geni a monte, è possibile anche distinguere l'influenza dovuta a geni diversi. L'idea di base di questa tecnica è quella di decomporre la serie temporale $x(n)$, $n \in N$, in un insieme di sinusoidi ad ampiezza variabile e aventi varie frequenze:

$$x[n] = \sum_{i=1}^{M} x_i[n] = \sum_{i=1}^{M} \alpha_i \exp(\sigma_i n)\cos(\omega_i n + \varphi_i) \qquad (3.9)$$

I parametri α_i, σ_i, ω_i, and φ_i (i = 1, 2, …, M), sono l'ampiezza, il fattore di smorzamento, la pulsazione e lo sfasamento della i-esima componente; essi possono essere calcolati con il metodo di autoregressione [Yan, 2005] e definiscono completamente lo spettro dell'espressione genica. La correlazione di $x[n]$ con un'altra sequenza $y[n]$ può essere riformulata come somma di componenti di correlazione parziali pesate:

$$x[n] \circ y[n] = \sum_i \sum_j \sqrt{\frac{E_{x_i} E_{y_i}}{E_x E_y}}\, x_i[n] \circ y_i[n] \qquad (3.10)$$

il simbolo "$\circ$" rappresenta l'operazione di correlazione e i termini E_{xi}, E_{yi}, E_x e E_y rappresentano l'energia totale di una sequenza o l'energia di una sua particolare componente. Questa equazione spiega il modo in cui una correlazione tra due sequenze possa essere separata in un insieme di componenti di correlazione parziali pesate, che

può fornire dettagli più approfonditi circa la relazione che intercorre tra una coppia di geni, in quanto si suppone che le componenti spettrali di ampiezza di un gene generate dall'azione di un altro gene abbiano con questo una maggiore correlazione.

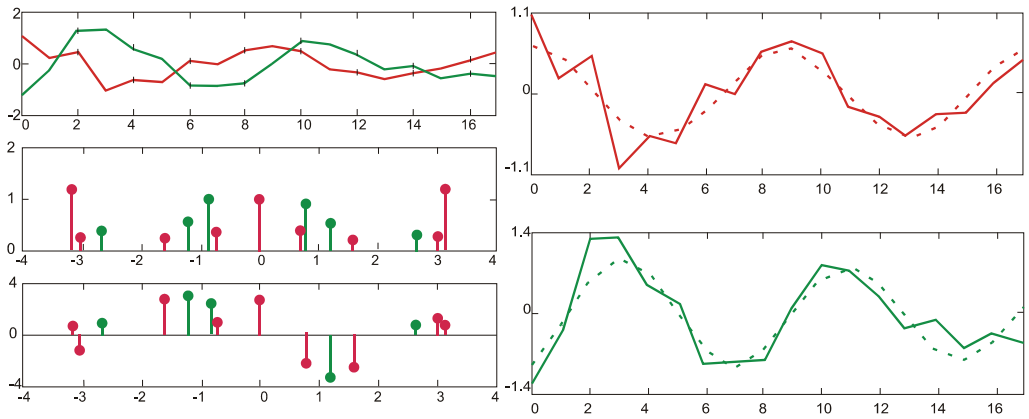


Figura 32: Rappresentazione delle serie temporali, degli spettri di ampiezza e di fase e delle sinusoidi ad ampiezza variabile corrispondenti alla componente spettrale dominante di due espressioni geniche legate da regolazione di tipo "attivazione".

In questa valutazione di correlazione possono essere trascurate le informazioni relative allo spettro di fase e in questo modo è possibile cogliere le similarità tra segnali traslati nel tempo senza ulteriori complicazioni.

Spesso si rivela utile trascurare componenti irrilevanti e azioni di disturbo, come il rumore, che potenzialmente costituiscono un ostacolo alla rivelazione di similarità ed è anche piuttosto frequente riportare bassi valori di correlazione in serie temporali a causa di componenti rumorose.

Proprio per questo motivo, le componenti di correlazione parziali pesate possono essere considerate come una più affidabile misura della relazione tra geni. In particolare si può fare riferimento alla massima componente pesata risultante dalla scomposizione con allineamento di fase come metrica per misurare le relazioni tra geni. Il corrispondente valore non pesato è chiamato coefficiente di correlazione relativo alla componente.

Nella figura 32 è rappresentato l'andamento di due serie temporali espressione di due geni legati da relazione di attivazione, la corrispondente scomposizione spettrale in ampiezza e fase e le sinusoidi corrispondenti alle componenti spettrali dominanti.

Il metodo di decomposizione spettrale trova impiego non solo per espressioni geniche cicliche, ma anche per i casi non periodici ed è quindi applicabile in tutti i casi di interesse per le indagini su profili di evoluzione dinamica di espressioni geniche.

La valutazione del metodo della Componente Spettrale Dominante, realizzata come nei due casi precedenti con gli indici di Accuratezza, Precision, Recall e di Matthews fanno rilevare una discreta qualità di questo algoritmo, che riesce ad avere prestazioni migliori di tutti gli altri fin qui elencati. Il motivo della percentuale di successi nell'individuare il tipo di relazioni esistenti tra i geni presi in esame va ricercato nelle caratteristiche di questo metodo, tutto concentrato sul contenuto in frequenza del segnale.

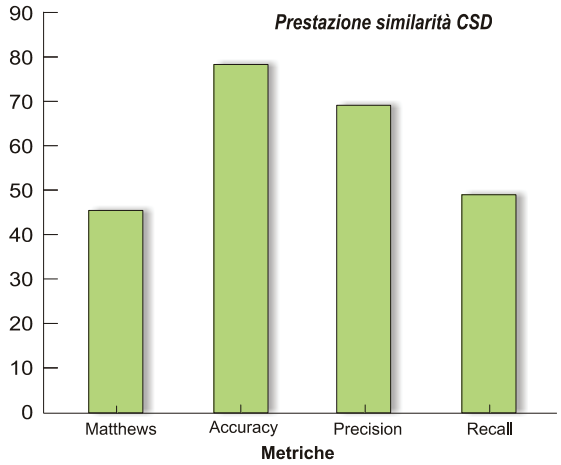


Figura 33: Rappresentazione della qualità del classificatore con approccio CSD relativa agli indici Accuratezza, Precision, Recall e indice di Matthews

3.8.4 Metodo di aggregazione dei risultati

Dopo aver applicato separatamente i tre metodi di valutazione delle relazioni tra coppie di geni è stata eseguita una procedura ibrida, per elaborare un punteggio complessivo di similarità che riassuma gli aspetti relativi a ciascuna delle metodologie. L'idea di combinare le informazioni ottenute è stata suggerita da un'evidenza sperimentale. Al di là del dato sintetico sulle prestazioni degli algoritmi differenziale, DTW e CSD, è stato possibile osservare che le particolari coppie riconosciute come simili da un metodo spesso non sono riconosciute come simili con un altro metodo. E allora, se si potessero aggregare le potenzialità dei vari metodi, che sono basati su parametri diversi, sarebbe possibile ottenere un criterio di similarità migliore?

Per rispondere a questo interrogativo è sembrato conveniente approfondire un'indagine sui metodi di combinazione dei risultati ottenuti da studi di similarità, in particolare su un lavoro che ha rappresentato una significativa milestone in questo ambito [Belkin et al., 1995]. In esso, con particolare riferimento alla similarità di testi e alle prestazioni ottenute su query in un insieme di documenti, sono argomentate diverse considerazioni circa la correttezza analitica e la convenienza nel ricercare la combinazione o la fusione dei risultati ottenuti con diverse metodologie sullo stesso sistema o con la stessa metodologia su sistemi diversi. L'attestazione di validità di un certo schema di combinazione dei dati risiede primariamente nella ricerca empirica. Esistono tuttavia diversi motivi in virtù dei quali ci si attende un miglioramento della funzione obiettivo dalla combinazione di risultati ricavati per vie diverse:

- l'arbitrarietà nella scelta di una soglia al di sopra della quale si decide sulla positività o negatività di una regolazione genica implica una certa probabilità che si verifichino sia dei falsi positivi che dei falsi negativi. La probabilità di una combinazione mediante funzione logica AND di risultati indipendenti conduce a una probabilità data dal prodotto delle singole probabilità che, nel caso di errori è quindi sicuramente minore che in ognuno dei risultati componenti.
- nell'ottica di una massimizzazione di una funzione obiettivo, espandere lo spazio dei metodi con cui effettuare una misura, e quindi lo spazio dei parametri

considerati, conduce almeno ad ottenere un risultato uguale e al più a raggiungere un risultato migliore. Si potrebbe addirittura verificare che ignorare uno dei due metodi sia una scelta migliorativa.

Gli algoritmi di combinazione sono sostanzialmente euristici. Nel caso in esame, se l'obiettivo ultimo è quello di minimizzare i falsi negativi, si potrebbe pensare di operare una scelta del massimo tra i tre punteggi di similarità ottenuti con i tre metodi precedenti, ovvero di effettuare una OR logica tra le appartenenze a classi di regolazione di attivazione o repressione (vedi figura 34).

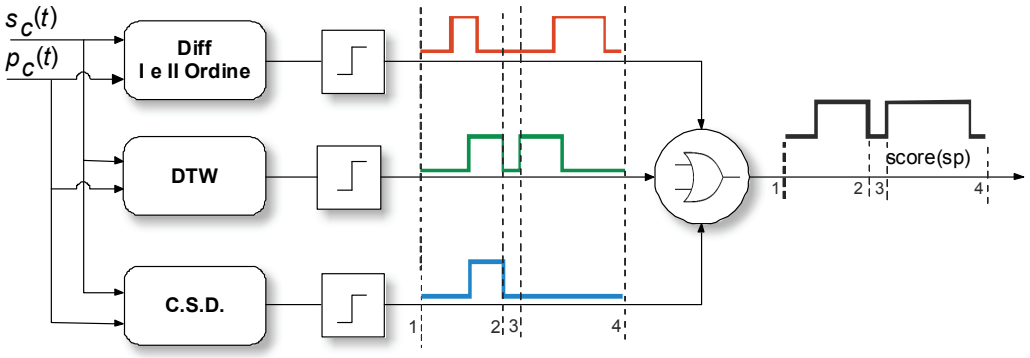


Figura 34: schema a blocchi dell'algoritmo ibrido di generazione dello score di co-regolazione

Se si vuole invece minimizzare la probabilità di falsi positivi si opererà allo stesso modo, con una funzione logica AND. Per avere un controllo più ampio sulla stima del tipo di regolazione di un gene sull'altro si può anche pensare a una combinazione dei risultati mediante coefficienti che vanno opportunamente "sintonizzati" (figura 35).

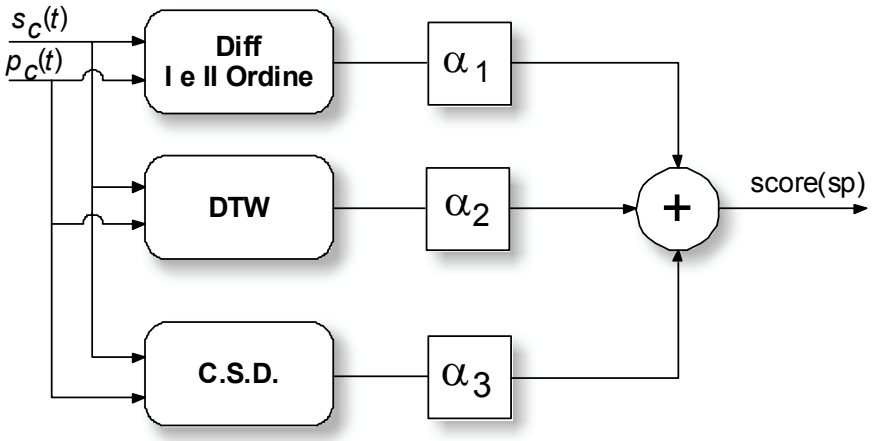


Figura 35: schema a blocchi dell'algoritmo ibrido di generazione dello score di co-regolazione

Riportiamo di seguito l'istogramma che valuta le prestazioni di uno questi di algoritmi ibridi.

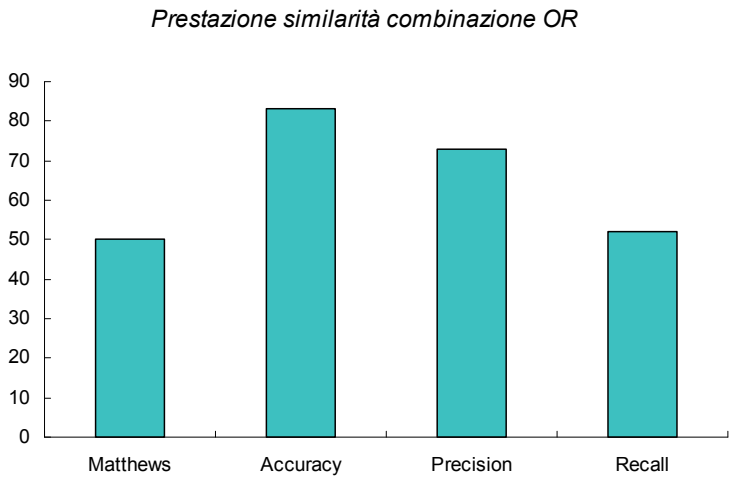


Figura 36: Rappresentazione della qualità del classificatore con metodo di combinazione dei risultati con funzione OR relativa agli indici Accuratezza, Precision, Recall e indice di Matthews

3.9 Valutazione dei risultati

Come passo finale del processo di confronto tra serie temporali alla ricerca di evidenze sulle tracce che la regolazione tra geni dissemina nelle dinamiche dei suoi profili di espressione, annotiamo qualche considerazione conclusiva, alla luce di tutti i risultati ottenuti.

In primis, la caratteristica rumorosità delle serie temporali di origine biologica, come già ripetutamente sottolineato in precedenza, deve essere opportunamente limitata per poter ottenere buoni risultati da qualsiasi algoritmo. Volendo quantificare, in questi casi, da valutazioni rilevate dall'elaborazione di elementi del dataset, si hanno rapporti segnale rumore che possono arrivare fino a limiti inferiori di pochi dB.

L'analisi di similarità condotta sulla base del coefficiente di correlazione di Pearson fornisce risultati affatto deludenti, riuscendo a individuare abbastanza correttamente le relazioni di attivazione, nel 45% dei casi, ma confondendo la relazione con la sua opposta nel 50% dei casi in caso di inibizione.

Degli altri tre metodi applicati per l'analisi di similarità, l'approccio nel dominio della frequenza ha presentato le migliori performance. Grazie alle sue peculiarità, è possibile processare con esso tutte le coppie di geni direttamente collegate, indipendentemente dal fatto che possano esserci delle influenze sovrapposte su uno stesso gene. E' inoltre stata discussa la sua robustezza rispetto al rumore.

Tuttavia, per sfruttare le informazioni contenute in alcuni parametri associati alla forma d'onda che rappresenta la serie temporale, è sembrata praticabile l'idea di un metodo combinato, in cui all'analisi nel dominio della frequenza fossero associate

anche informazioni sintetiche dell'andamento differenziale, opportunamente codificato in stringhe di bit, e della deformazione dell'onda.

I risultati ottenuti mostrano un miglioramento di quasi il 5% in termini di coefficiente di correlazione di Matthews del migliore approccio ibrido rispetto al migliore approccio assoluto, quello della Componente Spettrale Dominante, che era già abbastanza accettabile, essendosi attestato intorno al 45%.

In figura è rappresentato un istogramma riassuntivo dei casi precedenti.

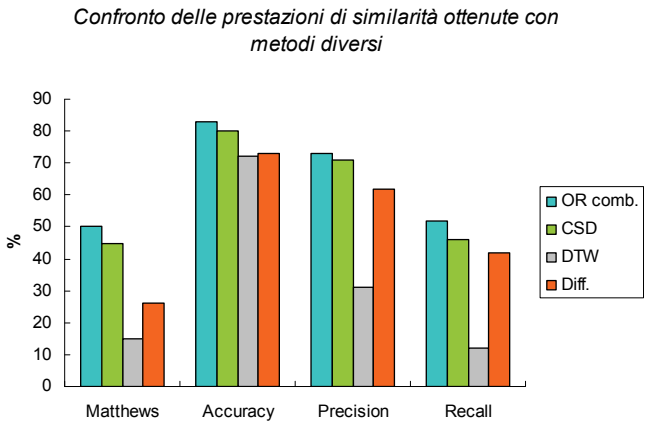


Figura 37: Comparazione delle qualità dei classificatori assoluti con quello ibrido

3.10 Conclusioni

Il confronto di serie temporali generate da esperimenti di tipo microarray, al fine di ricostruire il tipo di regolazione esistente tra coppie di geni, è stato condotto secondo diverse metodologie di trattamento dei dati.

Il risultato della sperimentazione ha mostrato una buona qualità dei classificatori combinati messi a punto in questa ricerca e confermato dal confronto di questo con altri tipi di classificatori che ne costituiscono le parti. I miglioramenti rispetto agli approcci assoluti conseguono da una maggiore ricchezza di informazioni a disposizione del classificatore e da una possibilità di sfruttare sinergie nascoste nella possibilità di valutare la similarità su una più ampia base di parametri. I coefficienti di peso o le funzioni logiche inserite nel computo dello score finale conferiscono maggior rilievo all'approccio che ha mostrato miglior attitudine alla classificazione per questo tipo di dati, riuscendo a modulare opportunamente i valori dei singoli componenti, in modo da recuperare alcune similarità significative e che erano emerse in uno solo degli indici assoluti .

A scopo di verifica, si prevede in tempi brevi di ampliare la sperimentazione ad altre serie temporali biologiche. Ad esempio i dati disponibili di serie temporali più

numerose, fino a 30 punti, oppure raccolte da studi su organismi diversi da Homo Sapiens, come Saccaromyces Cerevisiae, Drosofila e così via.

Sarebbe anche un'interessante spunto di ricerca quello di trovare una interpretazione biologica delle componenti spettrali, ricavate mediante l'approccio del passaggio nel dominio della frequenza e inoltre, obiettivo ancora più ambizioso, riuscire a collegare ogni componente spettrale a un gene specifico oppure a un determinato signaling pathway. In tale ottica i singoli geni o i percorsi di segnalazione genica potrebbero essere visti come delle sorgenti di segnali ad una frequenza caratteristica.

Molto pertinente appare anche la possibilità di studiare e quantificare i tempi di propagazione dei segnali biologici e quindi di individuare delle costanti di tempo dei circuiti biologici coinvolti in specifiche funzionalità. Con questi elementi lo studio dei network regolatori potrebbe essere facilitato dall'adozione delle pratiche oramai consolidate della teoria dei circuiti elettrici e da valutazioni dei tempi di propagazione dei segnali nei pathway potrebbero essere rilevate delle anomalie nella fisiologia dei geni.

In definitiva l'enorme massa di dati ancora non trattati e la mancanza di modelli stabili e non abbastanza universali nel campo dei network biologici è una implicita istanza di intensificazione delle ricerche in campo Bioinformatico. Anche la nascita di sempre nuove branche della cosiddetta "omica" presuppone quanto ancora siano vasti i margini di sviluppo di questa branca delle Scienze computazionali, quanto ancora sia lungo il cammino verso l'integrazione delle competenze e quanto urgente si dimostri la necessità di realizzare percorsi multidisciplinari che consentiranno un più produttivo e rapido sviluppo della conoscenza sui sistemi viventi.

3.11 Articoli sviluppati sul tema:

In questa pagina è riportato l'elenco degli articoli, in vista di sottomissione e in preparazione, riguardanti studi connessi con quelli presentati in questo capitolo.

1. Hybrid models for gene time series comparison (with Sorin Draghici and Walter Balzano). To be submitted.
2. Similarity metrics in gene time series. (with Sorin Draghici e Walter Balzano). To be submitted.
3. Time constants in gene regulatory networks. In preparation.

4. Appendice

A1. Illumina BeadArray

La tecnologia BeadArray per la fabbricazione di microarray è stata recentemente sviluppata da Illumina Inc., una delle aziende leader del settore delle scienze biologiche e dei sistemi integrati per analisi su larga scala di funzioni e variazioni genetiche. Questa tecnologia utilizza delle perline di silicio di 3 μm di diametro inserite in piccoli buchi scavati in uno o due substrati, fasci di fibre ottiche o piastrine di silicio planare. Queste perline sono auto-assemblate casualmente su uno di questi substrati in uno schieramento uniforme in cui le perline sono una a 5,7 μm di distanza una dall'altra.

Ogni perlina è coperta da centinaia di migliaia di copie di una data sequenza nucleotidica, formando il "probe" specifico per un determinato test. Ognuno di questi nucleotidi è lungo 50 bp ed è concatenato con un'altra sequenza fatta su misura che può essere utilizzata come un indirizzo associato in modo univoco ad ogni sito di destinazione specifica. Poiché lo spazio necessario per il probe che corrisponde a un solo gene è così piccolo, sull'array si può raggiungere un'elevata densità. Inoltre, gli array possono essere multiplexati per testare più campioni in parallelo.

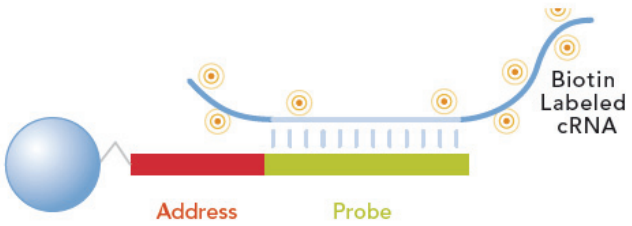


Figura A1: un "probe" di Illumina ad ibridizzazione diretta è simile a quelli che sono adoperati nelle altre tecnologie di microarray, ma è attaccato a una perlina di silicio e non si trova su una superficie piatta. Le perline sono fissate su fibre ottiche o distribuite su una piastrina di silicio, come mostrato nella figura A2. Immagine tratta da http://www.Illumina.com

Attualmente Illumina sta producendo un array a matrice di puntini con 96 campioni, che consente ai ricercatori di testare fino a 96 campioni contemporaneamente. Può essere adoperato lo stesso formato in applicazioni che vanno dalla tipizzazione all'espressione genica. Nelle applicazioni di espressione genica, Illumina produce array dell'intero genoma, così come array più mirati, che includono probe per un sottoinsieme di geni relativi a una condizione specifica. Ci sono uno o due probe per gene, a seconsa del tipo di array (mirato a un sottoinsieme o all'intero genoma). Gli array Illumina per l'espressione genica sono prodotti in due versioni, corrispondenti a due diversi approcci. Il test a Ibridizzazione diretta usa una singola sequenza di DNA per perlina, molto simile alle altre tecnologie. Questa sequenza a singola elica è finalizzata all'ibridizzazione con una sequenza bersaglio presente nel saggio. La quantità di fluorescenza fornisce una misura della quantità del gene bersaglio contenuta nel saggio.

L'altro approccio alla misura del livello di espressione è chiamato DASL, che significa appaiamento mediato da cDNA, Selezione, Estensione e Legatura. I questo test ogni gene è rappresentato scegliendo da 3 a 10 siti bersaglio. Una coppia di nucleotidi è associata a ogni sito bersaglio e possono essere multiplexati fino a 1536 coppie di oligonucleotidi in una singola reazione. Viene raggiunta un'elevata specificità richiedendo che entrambi i membri di una coppia di oligonucleotidi debba ibridizzare in prossimità affinché il test possa generare un segnale intenso.

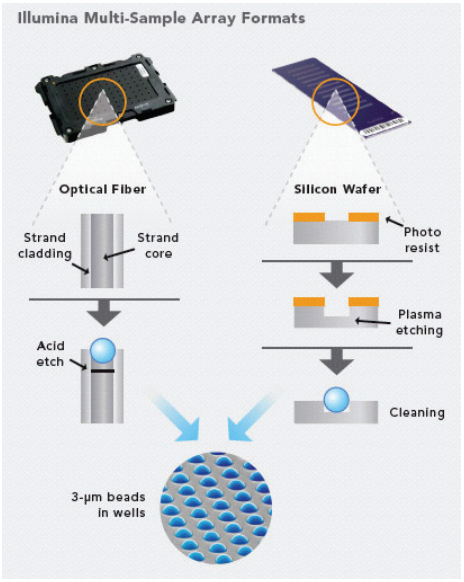


Figura A2: tecnologia Illumina BeadArray. Perliine di silicio del diametro medio di 3 µm sono messe in piccoli buchi ricavati dalle fibre ottiche o da wafer di silicio. Le perline sono tenute ferme da forze di Van der Waals e da interazioni idrostatiche con le pareti del buco. la superficie di ogni perlina è ricoperta con centinaia di migliaia di copie della sequenza scelta per rappresentare un gene. Immagine tratta da http://www.Illumina.com

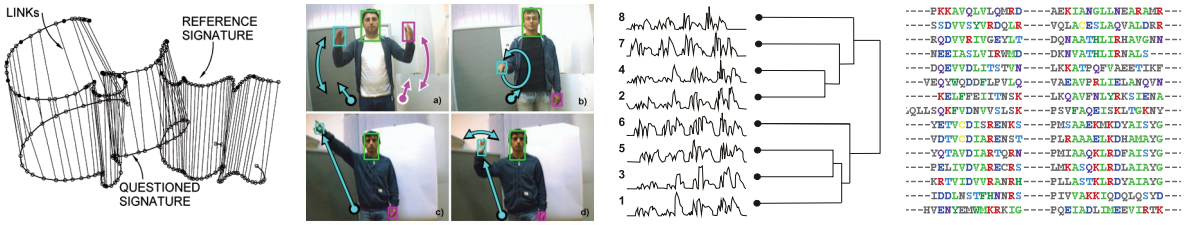
Il principale vantaggio dell'approccio DASL rispetto all'ibridizzazione diretta riguarda la qualità del mRNA che può essere testato. Poiché il test DASL utilizza due sequenze brevi che nel gene sono separate da un gap, è consentita notevole flessibilità

nella scelta delle sequenze. Inoltre, dato che questi probe sono calibrati solo per 50 basi, può essere usato RNA parzialmente degradato. Al contrario, gli array di cDNA utilizzano sequenze molto lunghe e richiedono RNA di buona qualità, che può essere ottenuto solo da tessuti freschi o tessuti congelati subito dopo il prelievo.

A2. Dinamic Time Warping

L'algoritmo Dynamic Time Warping algorithm (DTW) è un noto algoritmo introdotto negli anni'60 ed ampiamente esplorato negli anni '70 in molte aree (Figura X1) tra cui:

- Riconoscimento del parlato (speech recognition)
- Riconoscimento della scrittura (handwriting and online signature matching)
- Riconoscimento dei gesti (gesture recognition)
- Datamining e Time series clustering
- Computer Vision and Animation
- Allineamento delle sequenza di proteine
- Musica ed elaborazione di segnali



(a) (b) (c) (d)

Figura A3: (a) Riconoscimento di una firma basata sull'abbinamento e sul confronto di punti significativi della firma di riferimento e quella da verificare (b) Riconoscimento dei gesti mediante il confronto di traiettorie di punti osservati con traiettorie di riferimento; (c) Time series clustering con l'ausilio di dendogrammi (d); Allineamento delle sequenze di due campioni di proteine mediante la ricerca di sottopattern comuni.

L'algoritmo DTW ha ottenuto grande popolarità per l'efficiente calcolo della misura di similitudine di serie temporali mediante deformazioni (Warping) e trasformazioni 'elastiche' che minimizzano gli effetti di sfasamento e distorsioni presenti nei modelli confrontati.

DTW Classico

L'obiettivo primario del DTW consiste nel confrontare 2 sequenze temporali $X=(x_1, x_2,\ldots,x_N)$ ed $Y=(y_1, y_2,\ldots,y_M)$ con N e M $\in \mathbb{N}$ e le cui componenti, o Features, appartengono ad un prefissato 'spazio delle caratteristiche' $\mathbb{F}$.
Per confrontare una coppia di elementi di $\mathbb{F}$ occorre una funzione c: $\mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R} \geq 0$ detta metrica locale dei costi nota anche come misura locale della distanza. Si introduce quindi la matrice dei costi C_{NxM} con $c(x_i,y_j)$ $(i,j) \in [1:N] \times [1:M]$ che definisce le misure locali di tutte le coppie (x_i,y_j). L'obiettivo della DTW è quello di ricercare il più basso

'costo complessivo' di un allineamento tra X ed Y composto da una serie scelta di L coppie (x_i, y_j). La complessità per la gestione del DTW è O(NxM); di recente è stata proposta una metodologia, nota con il nome di FastDTW che ha invece una complessità lineare.
Dal punto di vista intuitivo, un allineamento ottimale tra le sequenze X ed Y percorre un sentiero di avvallamenti più profondi possibile (valori bassi) all'interno della matrice C:

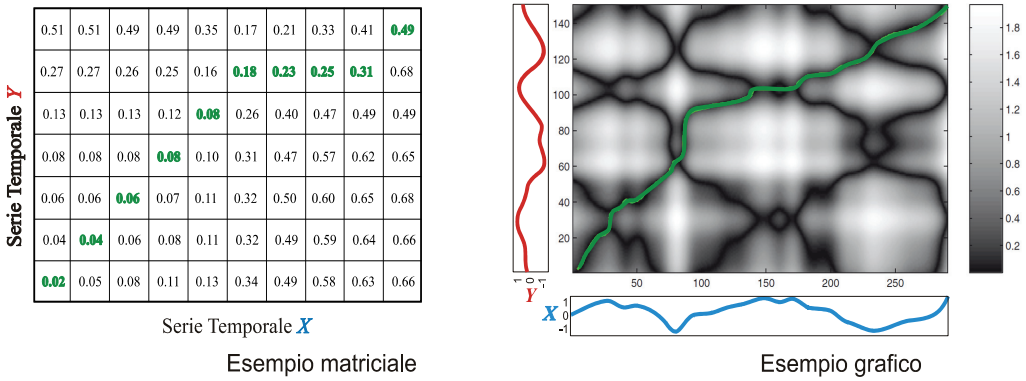


Figura A4: Rappresentazione tabellare e grafica della matrice dei costi C. Nell'esempio grafico in cui i valori bassi delle distanze euclidee c(x_i,y_j) sono rappresentati da tonalità più scure ed i valori alti da tonalità più chiare. La DTW, mediante una opportuna scelta di coppie di punti (x_i,y_j), individua il miglior percorso, cioè quello a costo complessivo più basso, da (x_1,y_1) a (x_N,y_M) così come rappresentato dai valori verdi nella tabella e dal percorso verde dell'esempio grafico.

Una premessa di base dell'algoritmo DTW presume che i punti delle due serie temporali X, Y siano state campionate con gli stessi intervalli di tempo altrimenti si effettua un ricampionamento. Una metrica tradizionale di confronto come quella della distanza Euclidea (fig. A5 a) relaziona in modo biunivoco l'i-esimo punto di X con l'i-esimo di Y; invece, nella metrica DTW invece tale corrispondenza non è biunivoca (fig. A5 b) e ciò comporta un confronto 'elastico' più intuitivo.

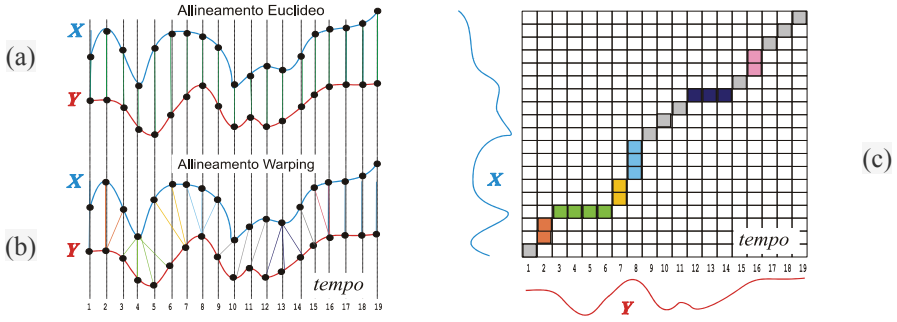


Figura A5: Confronto di metriche di serie temporali: la metrica Euclidea (a) e la metrica DTW (b). Dall'osservazione del percorso ricavato sulla matrice C_{NxM} delle distanze locali risulta semplice notare che gli allineamenti 'migliori' sono composti da un numero più ampio possibile di percorrenze oblique: se tale percorso su C coincide con la diagonale allora le due sequenze temporali X ed Y sono perfettamente allineate.

L'individuazione di una scelta ottimale di tali coppie di punti (x_i ,y_i) corrisponde alla individuazione di un percorso p da (x_1, y_1) a (x_N, y_M) nella matrice C dei costi (Figura A5 c), con costo minimo. Un generico percorso su C può essere rappresentato con una sequenza p di elementi

$$p = (p_1,\ldots, p_L) \text{ in cui } p_\ell = (n_\ell , m_\ell) \in [1{:}N] \text{ x } [1{:}M] \text{ con } \ell \in [1{:}L]$$

Di fatto un (N, M)-warping path definisce un allineamento tra 2 sequenze X=(x_1, x_2,…, x_n) ed Y=(y_1,y_2,…, y_n) assegnando elementi di X con elementi di Y. Poiché ciò che si intende realizzare è un allineamento totale tra le 2 serie temporali X ed Y allora il primo abbinamento sarà identificato dalla coppia (x_1, y_1) e l'ultimo abbinamento dalla coppia (x_N, y_M) (si veda la Figura X3-c).
Definiamo ora il costo totale $c_p(X,Y)$ di un warping path p tra le serie X ed Y basato sulla misura di costo locale c, la somma:

$$c_p(X,Y) := \sum_{l=1}^{L} c(x_{nl}, y_{ml}).$$

Un Warping path ottimale tra X ed Y è un warping path p* che ha il costo minimo tra tutti i possibili warping path; Da ciò definiamo la distanza DTW(X,Y) tra X ed Y uguale al costo totale di p*.

$$\text{DTW}(X,Y) = c_{p^*}(X,Y) = \min\{\text{cp}(X,Y) \mid p \text{ è un (N, M)-warping path}\}$$

Questa nuova metrica soddisfa molte proprietà come quella della simmetria (a patto che sia simmetrico lo stesso costo locale c su cui la metrica DTW è basata). Si noti esplicitamente che i percorsi descrivibili in C sono numerosi ed il percorso che soddisfa il requisito di costo minimo potrebbe non essere unico; Definiamo ara dei vincoli, alcuni di base ed altri opzionali, che facilitano il calcolo di questa nuova metrica. Un (N, M)-warping path è caratterizzato dalle seguenti 3 condizioni di base:

- **i.** Punti estremi (boundary): $p_1 = (1,1)$ e $p_L = (N, M)$.
- **ii.** Monotonia (monotonicity): $n_1 \le n_2 \le\ldots \le n_L$ ed $m_1 \le m_2 \le\ldots \le m_L$.
- **iii.** Incremento (step size): $p_{\ell+1} - p_\ell \in \{(1,0),(0,1),(1,1)\}$ per $\ell \in [1{:}L\text{-}1]$.

Per migliorare la complessità di calcolo del (N, M)-warping path è possibile limitare lo spazio di ricerca nella matrice dei costi C aggiungendo, ad esempio, queste ulteriori condizioni:

- **iv.** Intervallo di deformazione (warping window): $|n_\ell - m_\ell| \le r$, con $r > 0$
- **v.** Intervallo di pendenza (slope constraint): $(m_\ell - m_z) / (n_\ell - n_z) \le p$ con $p \ge 0$
  ed $(n_\ell - n_z) / (m_\ell - m_z) \le q$ con $q \ge 0$ ed $\ell, z \in [1{:}L]$.

Per interpretare il significato di tali 5 condizioni si ricorre ad un sistema di assi cartesiano ***n*0*m*** per la rappresentazione grafica della matrice C (Figura A5-c e Figura A6): per uniformare il senso di crescenza degli assi coordinati del sistema di riferimento ***n*0*m*** al tradizionale senso di crescenza di indicizzazione riga/colonna della matrice C, si considera CR che corrisponde alla rotazione antioraria di 90° di C. In tal modo l'elemento C(1,1) corrisponderà all'elemento (1,1) in basso a sinistra di ***n*0*m*** e C(N,M) corrisponde all'elemento (M,N) in alto a destra di ***n*0*m***.

La condizione **i.** richiede che l'algoritmo di costruzione dell'(N, M)-warping path determini un percorso *p* nella matrice C in cui il primo punto è l'elemento in basso a sinistra della matrice C e l'ultimo punto sia l'elemento di C in alto a destra (parte superiore della Figura A6-*i.*); tale condizione, riferita alle due serie temporali *X* ed *Y*, implica una scelta di coppie (n_i ,m_i) che, contrariamente a quanto mostrato nella parte inferiore della Figura A6-**i**., garantisce l'inclusione dei punti di estremità di *X* ed *Y*.

La condizione **ii.** di monotonia serve a garantire che non ci siano ripetizioni di allineamenti (Figura A6-**ii**.). In altre parole, se una sequenza di elementi all'interno della matrice dei costi C definisce un percorso che decresce allora ciò implica che uno medesimo sottoinsieme della sequenza temporale *Y* viene fatto corrispondere a due sottoinsiemi diversi della sequenza temporale *X*.

La condizione **iii.**, nota anche come condizione di continuità, garantisce che nel (N, M)-warping path *p* non siano presenti salti temporali generando discontinuità in *p*. Un possibile effetto della violazione di questa condizione potrebbe indurre a non considerare importanti caratteristiche delle serie temporali (Figura A6-**iii**.).

La condizione **iv.**, detta condizione di 'warping window', limita eccessive deformazioni e garantisce una limitata differenza di velocità di scorrimento dei punti da confrontare (Figura A6-**iv**.).

Infine, la condizione **v.** che riguarda le pendenze della curva (troppo accentuata o troppo piatta), impedisce che una serie di pochi punti consecutivi di una serie temporale che costituiscono cioè un breve tratto di essa, siano abbinati ad una serie troppo numerosa di punti consecutivi, cioè un lungo tratto, dell'altra serie temporale (Figura A6-**v**.); viceversa per il caso di curva troppo piatta.

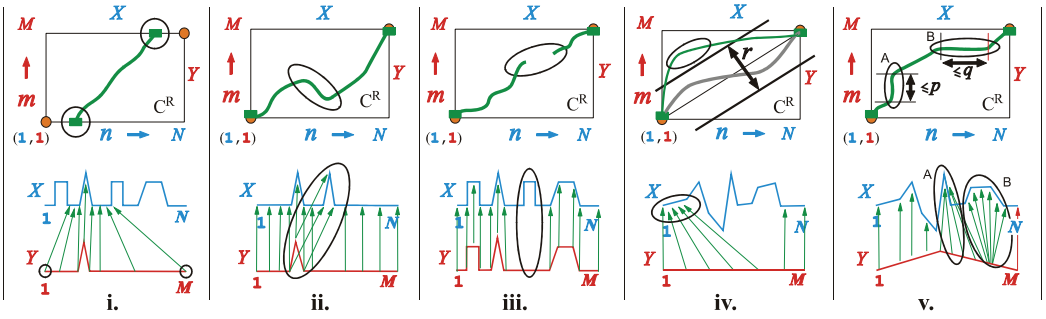


Figura A6: significato grafico (matriciale e serie temporali) corrispondente alle condizioni i. ii. iii. iv. ed v. relative al confronto di due serie temporali *X* ed *Y* con metrica DTW.

Dtw-normalizzato

Come precedentemente detto, per poter iniziare a confrontare le due sequenze temporali *X* ed *Y* occorre fare alcune ipotesi fondamentali riguardanti proprio la variabile 'tempo'; è necessario cioè considerare in quale modo il tempo di campionamento possa influenzare la misura DTW.

Alcuni importanti fatti che riguardano il tempo sono i seguenti:

Si presume che le serie temporali considerate *X* ed *Y* siano state campionate con gli stessi intervalli di tempo altrimenti è necessario un ricampionamento.

I campionamenti possono essere stati effettuati sia in modalità statica, cioè ad intervalli di tempo costante, sia in modalità dinamica, cioè ad intervalli di tempo variabile.

I campionamenti, in dipendenza da quale istante di tempo si riferiscono, possono essere caratterizzati da diversi gradi di rilevanza o attendibilità. L'abbinamento dei punti di *X* ed *Y* dovrebbe essere pertanto normalizzato considerando anche aspetti qualitativi del campionamento: le serie temporali sono di sovente corredate etichette che definiscono la qualità dei dati ai quali esse sono riferite. Ad esempio, nel caso di una serie temporale da Microarray si considerano i 'pvalue' che determinano l'attendibilità del dato rilevato e nel caso di una serie temporale da GPS si considerano i valori di *DOP* (Diluition Of Precision) che, anche in questa circostanza, determinano la qualità del dato considerato.

Tralasciando aspetti di dipendenza come quelli riferiti alla qualità del dato menzionata al precedente punto 3., in questa trattazione ci si occuperà di valutare un modo per normalizzare la misurazione DTW rispetto al tempo di campionamento. Anzitutto è semplice verificare la dipendenza della misurazione rispetto al campionamento perché se ad esempio si considerano le serie temporali *X* ed *Y* entrambe campionate ad intervalli regolari di tempo e se si considerano altre due serie temporali *X'* ed *Y'* definite come copie sotto-campionate rispettivamente di *X* ed *Y* allora risulta evidente che l'applicazione della precedente definizione di DTW implicherà che DTW(*X*,*Y*) > DTW(*X'*,*Y'*). Ciò scaturisce dal fatto che il numero di abbinamenti tra i punti delle due serie temporali *X* ed *Y* è sicuramente maggiore del numero di abbinamenti tra i punti delle due serie temporali *X'* ed *Y'*: di conseguenza il percorso di allineamento nella matrice dei costi C risulterà diverso.

Un primo tentativo di normalizzazione del DTW può essere suggerito dal formato del seguente schema:

Distanza DTW Time-normalized tra le due serie temporali *X* ed *Y*:

$$\mathrm{D}(X,Y) = \min_{p}\left[\frac{\sum_{\ell=1}^{L} c(p_\ell)\cdot w_\ell}{\sum_{\ell=1}^{L} w_\ell}\right]$$

in cui:
c(p_ℓ) è il costo locale associato alla coppia (n_ℓ, m_ℓ) ∈ [1:N] x [1:M] con l ∈ [1:L]
$w_\ell > 0$ rappresenta un coefficiente di peso

$\mathrm{D}(X, Y)$ definisce un DTW normalizzato come il miglior percorso di allineamento quantitativo tra le due serie temporali *X* ed *Y* indipendentemente dagli aspetti qualitativi dei campionamenti. Occorre ora definire i pesi w_ℓ. Il significato di $w_\ell = 1$ può corrispondere al caso in cui i campionamenti di *X* ed *Y* siano costanti ed effettuati ogni unità di tempo: ciò renderebbe più semplice il calcolo del $\mathrm{D}(X, Y)$. In letteratura è possibile trovare diversi per la definizioni dei pesi; tuttavia è facile rendersi conto che da essi dipende parte della complessità totale della $\mathrm{D}(X, Y)$. Per questo motivo è ragionevole fissare qualche vincolo costruttivo che semplifichi tale calcolo. Se, ad esempio, si suppone di fissare una costante *K* per la quale valga la relazione:

$$\sum_{\ell=1}^{L} w_\ell = K$$

Allora la precedente formulazione di $\mathfrak{D}(X, Y)$ si sostituisce con la più semplice

$$\mathfrak{D}(X, Y) = \frac{1}{K} \min_{p} \left[\sum_{\ell=1}^{L} c(p_\ell) \cdot w_\ell \right]$$

che garantisce così una misurazione tra le due serie temporali X ed Y senz'altro più accurata del DTW classico precedentemente descritto. Il problema successivo resta quindi legato alla numerosità dei percorsi identificabili nella matrice C dei costi locali.

Calcolo del DTW mediante Programmazione Dinamica

Il problema principale nel calcolo del DTW è connesso alla numerosità dei percorsi costruibili su C e l'approccio risolutivo impiega i modelli di Programmazione Dinamica.
L'algoritmo generale per il calcolo del DTW è costituito da due distinte fasi principali:

FASE 1: Calcolo di una matrice G detta 'Matrice Cumulativa dei Costi'
(si veda in seguito ACCUMULATEDCOSTMATRIX del frammento di codice 1)

FASE 2: Individuazione del percorso p
(si veda in seguito OPTIMALWARPINGPATH del frammento di codice 2)

Partendo dunque dai costi locali definiti dagli elementi $c(x_i,y_j)$ della matrice C, è possibile costruire una nuova matrice G (Accumulated Cost Matrix) definita dalle seguenti 4 condizioni di base:

$G(1,1) = 1$

$G(n,1) = \sum_{z=1}^{n} c(x_z, y_1)$ per $n \in [1:N]$

$G(1,m) = \sum_{z=1}^{m} c(x_1, y_z)$ per $m \in [1:M]$

$G(n,m) = \min \{G(n\text{-}1, m\text{-}1), G(n\text{-}1,m), G(n,m\text{-}1)\} + \mathbf{c(x_n, y_m)}$
per $1 < n \leq N$ e $1 < m \leq M$

La prima semplice condizione imposta ad 1 il valore di G(1,1). Le condizioni 2. e 3. occorrono per il calcolo dei valori di G rispettivamente della prima colonna e della prima riga di G (gli step 2. e 3. possono anche essere invertiti). Infine, la condizione espressa in 4. esprime la regola costruttiva, di tipo ricorsivo, per il calcolo di tutti gli altri elementi di G. Dalle condizioni 1.-4. deriva che

$$DTW(X, Y) = G(N,M)$$

risolubile a complessità $O(NM)$ con l'ausilio di tecniche di Programmazione Dinamica.
Occorre inoltre notare che le 4 definizioni di base espresse in 1-4 trovano in letteratura molteplici varianti di particolare rilievo basate sopratutto su diversi '*fattori di incremento aggiuntivi*'. (il $c(x_n, y_m)$ della precedente condizione 4.)
Tali fattori di incremento identificano il peso w_ℓ associato all'ℓ-esimo step di avanzamento p_ℓ sulla matrice G durante la costruzione ricorsiva del percorso p da G(N,M) a G(1,1). In particolare i modelli maggiormente impiegati per la definizione di tali pesi sono:

Modello Simmetrico: $w_\ell = (n_\ell - n_{\ell-1}) + (m_\ell - m_{\ell-1}) \Rightarrow K = N + M$

Modello Asimmetrico: $w_\ell = (n_\ell - n_{\ell-1}) \Rightarrow K = \mathrm{N}$

[o, equivalentemente, $w_\ell = (m_\ell - m_{\ell-1}) \Rightarrow K = \mathrm{M}$]

Modello Quasi-Simmetrico, che differisce dal modello Simmetrico solo per la definizione del peso da considerare, 1 anziché 2, nel caso di avanzamento obliquo sulla matrice dei costi C (si veda Figura A7 e Tabella A1).

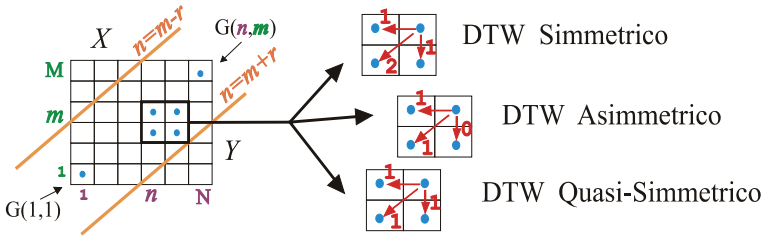


Figura A7: Calcolo dei pesi cumulativi della matrice G per i modelli DTW Simmetrico, Asimmetrico e Quasi-Simmetrico.

I valori di *K* del modello Simmetrico, Asimmetrico e Quasi-Simmetrico ottenuti come varianti del modello di base, permettono di semplificarne l'implementazione dell'algoritmo e di realizzare la normalizzazione della misurazione. Il valore finale della misurazione normalizzata è così definito:

$$\mathfrak{D}(X, Y) = \mathrm{G(N,M)} / K$$

Una prima approssimazione per la realizzazione di algoritmi di Programmazione Dinamica, comprensivi di vincolo 'warping window', con i tre sistemi di pesatura DTW Simmetrico, Asimmetrico e Quasi Simmetrico possono essere schematizzati nella tabella A1:

	DTW Simmetrico	DTW Asimmetrico	DTW Quasi Simmetrico
Condiz. limite G(1,1)	2c(1,1).	c(1,1).	c(1,1).
Condiz. limite prima colonna	$G(n,1) = \sum_{z=1}^{n} c(x_z, y_1)$		
Condiz. limite prima riga	$G(1,m) = \sum_{z=1}^{m} c(x_1, y_z)$		
Eq. di ricorsione G(n,m)	$\min\begin{Bmatrix} G(n,m-1)+c(n,m) \\ G(n-1,m-1)+2c(n,m) \\ G(n-1,m)+c(n,m) \end{Bmatrix}$	$\min\begin{Bmatrix} G(n,m-1) \\ G(n-1,m-1)+c(n,m) \\ G(n-1,m)+c(n,m) \end{Bmatrix}$	$\min\begin{Bmatrix} G(n,m-1)+c(n,m) \\ G(n-1,m-1)+2c(n,m) \\ G(n-1,m)+c(n,m) \end{Bmatrix}$
Condiz. di Warping window	$m - r \le n \le m + r.$		
K=somma costi	N + M	N	N + M
distanza normalizz $\mathfrak{D}$(X, Y)	G(N,M) / *K*		

Tabella A1: Tre varianti risolutive dell'algoritmo di Programmazione Dinamica. Si noti espressamente che l'algoritmo, per ogni step di scelta descritto nell'equazione di ricorsione, ha 3 alternative possibili. In altre parole, così come illustrato anche in Figura A7, ad ogni step di calcolo del percorso in G, le 3 direzioni possibili sono (a) spostamento a sinistra (b) spostamento in basso (c): spostamento obliquo in basso a sinistra. La scelta tra questi tre possibili movimenti dipende da quale di essi abbia il costo più piccolo. Infine, la memorizzazione di tutte le scelte effettuate in questo passo di ricorsione individueranno univocamente il percorso ottimo nella matrice G.

In sintesi, applicando uno degli schemi di Programmazione Dinamica sopra proposti, otterremo nell'ultima posizione di G, cioè G(N,M) il valore della misurazione richiesta della distanza tra le 2 serie temporali confrontate *X* ed *Y* da cui, dividendo per *K* otterremo il valore normalizzato finale.
La FASE 2 sopra citata sarà eseguita per disegnare il percorso *p* sulla matrice cumulativa G dei costi. Per la costruzione di *p* sarà sufficiente, durante la FASE 1, aver individuato e memorizzato tutti i singoli spostamenti ricavati al momento della scelta del valore minimo descritto nell'equazione di ricorsione G(*n*,*m*) di Tabella A1 e come illustrato in Figura A8 (si vedano i corrispondenti diversi singoli step di spostamento a destra di Figura A8).

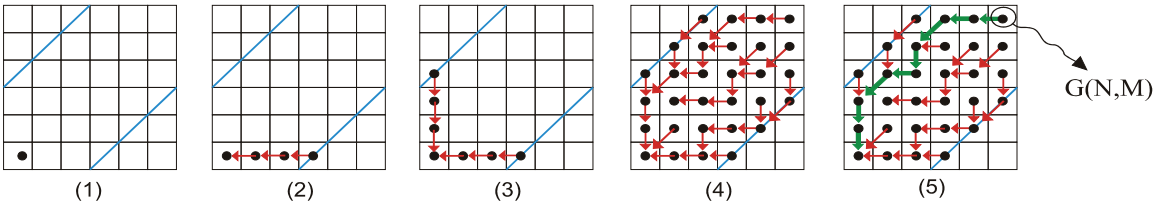


Figura A8: Progressione dell'algoritmo di programmazione dinamica, comprensivo del vincolo di 'warping window'. (1) impostazione del primo valore (2) calcolo della prima riga (3) calcolo della prima colonna (4) calcolo di tutti gli altri elementi di G (5) calcolo del percorso *p* con produzione del valore di G(N,M).

Derivative Dynamic Time Warping (DDTW)

Prima di concludere questa trattazione sul DTW, si ritiene opportuno citare a titolo di esempio, un'altra variante del DTW, il Derivative Dynamic Time Warping (Keogh – Pazzani) nota anche con il nome di DDTW. Tale variante che integra i modelli precedentemente descritti, basa la sua caratteristica predominante sul fatto che sostituisce la matrice C dei costi 'euclidei' tra i campioni delle due serie temporali *X* ed *Y* in cui $c(x_i, y_i) = | x_i - y_i |$, con una nuova matrice C' costruita partendo dalla solita matrice euclidea C_{NxM}. C' è costruita usando la differenza tra le stime della derivata. In concreto, il valore $c(x_i, y_i)$ di C corrisponde in C' al valore

$$c'(x_i, y_i) = | E(x_i) - E(y_i) | 2$$

in cui:

$$E(x_i) = \{ (x_i - x_{i-1}) + [(x_{i+1} - x_{i-1}) / 2] \} / 2$$
$$E(y_i) = \{ (y_i - y_{i-1}) + [(y_{i+1} - y_{i-1}) / 2] \} / 2$$

Questo algoritmo risulta essere più robusto del precedente nell'allineamento dei due segnali lungo l'asse delle ordinate (cioè variazioni in ampiezza), ed ha un comportamento migliore in alcuni punti di singolarità (Figura A9).

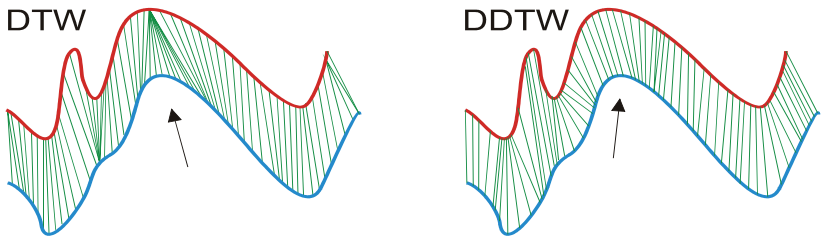


Figura A9: Confronto tra DTW e DDTW

Fast Dynamic Time Warping (FastDTW)

Come visto in precedenza l'implementazione del DTW comporta una complessità spaziale e temporale assai elevata che ne limita l'utilizzo. Tra i molteplici approcci e le molteplici derivazioni che agiscono suoi vincoli, va citata la metodologia nota come FastDTW (Stan Salvador e Philip Chan in "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space") basata su approccio multilivello che agisce in modo ricorsivo e progressivamente migliora una soluzione di base iniziale di bassa risoluzione. La peculiarità di questo approccio innovativo risiede proprio nella migliore complessità computazionale che è di tipo lineare in tempo e spazio.
Il nucleo dell'approccio FastDTW è basato sulla ricerca della soluzione su spazi a risoluzione sempre più elevata. In altre parole si esegue una DTW su una matrice di distanza molto 'grezza' composta cioè da un numero molto ridotto di righe e di colonne. L'algoritmo successivamente procede ricercando una soluzione solo nella parte di percorso selezionata al passo precedente. Le iterazioni successive considerano matrici con numero di righe e colonne sempre maggiori.

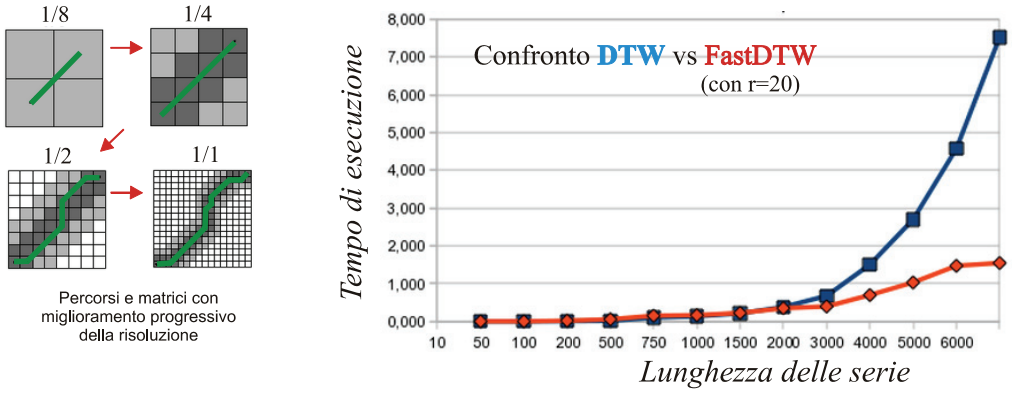


Figura A10: Progressione dell'algoritmo FastDTW e confronto con DTW

Ulteriore peculiarità dell'algoritmo FastDTW risiede nel fatto che esso è applicabile, oltre al DTW tradizionale, anche al Derivative DTW (DDTW).

Descrizione algoritmica del DTW

Di seguito vengono illustrate le due procedure principale dell'algoritmo DWT. Come precedentemente detto nella sezione sull'approccio al calcolo del DTW mediante Programmazione Dinamica, l'algoritmo completo è composto da due fasi distinte: nella prima fase viene calcolata la matrice cumulativa dei costi di cui l'ultimo elemento (N,M) della matrice è proprio il valore ricercato della distanza tra le due serie temporali *X* ed *Y*. La seconda fase, mediante una strategia greedy di backtracking, partendo dal punto finale del percorso = (N,M) e giungendo al punto iniziale del percorso = (1,1) ricostruisce il percorso di warping path.

Algorithm ACCUMULATEDCOSTMATRIX(X,Y,C)

Input: serie Temporali X, Y.
Matrice di costi (o distanza locale) C
Output: G matrice cumulativa della distanza X, Y

```
 1: n ← | X |; m ← | Y |; G[]← new[n x m]
 2: for i = 1; i ≤ n; i++ do
 3:     G(i,1) ← G(i – 1,1) + c(i, 1)
 4: end for
 5: for j = 1; j ≤ m; j++ do
 3:     G(1, j) ← G(1, j – 1) + c(1, j)
 4: end for
 5: for i = 1; i ≤ n; i++ do
 6:   for j = 1; j ≤ m; j++ do
 7:     G(i,j) ← c(i, j) + min {G(i – 1, j); G(i, j – 1); G(i – 1, j – 1)}
 8:   end for
 9: end for
10: return G
```

Frammento di codice 1: Calcolo della matrice *G* = Matrice Cumulativa dei Costi

Algorithm OPTIMALWARPINGPATH(G)

Input: G matrice cumulativa della distanza X, Y
Output: path = percorso nella matrice cumulativa *dtw*

```
 1: path ← new array; i = rows(G); j = columns(G)
 2: while ( i > 1) & (j > 1) do
 3:    if i == 1 then j = j – 1 else
 4:       if j == 1 then i = i – 1 else
 5:          if G(i – 1, j) == min {G(i – 1, j); G(i, j – 1); G(i – 1, j – 1)}
 6:             then i = i – 1 else
 7:                if G(i, j – 1) == min {G(i – 1, j); G(i, j – 1); G(i – 1, j – 1)}
 8:                   then j = j – 1 else i = i – 1 ; j = j – 1
 9:                endif
10:          path.add((i, j))
11:          endif
12: end while
13 return path
```

Frammento di codice 2: Calcolo del percorso path a costo minimo nella matrice *G*

A3. Parametri DOP

L'accuratezza con la quale è determinata la posizione è strettamente correlata all'erroredella misura mediante il fattore DOP (Diluition of Precision) che può essere espresso come rapporto tra la precisione nella posizione e quella della misura da: $\sigma = DOP \cdot \sigma_0$ con σ_0 deviazione standard dell'errore di misura e σ deviazione standard dell'errore di posizione. Il DOP è appunto uno scalare che "quantifica" il contributo geometrico della configurazione alla precisione della posizione. I vari tipi di DOP esistenti dipendono unicamente dalla particolare coordinata o combinazioni di coordinate di cui si vuole considerare la precisione. Indicate con σ_x^2, σ_y^2,σ_z^2, σ_T^2 le varianze sugli errori nelle tre dimensioni e nel tempo, si avranno i seguenti parametri DOP:

Posizionamento tridimensionale: $PDOP = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}$

Posizionamento bidimensionale: $HDOP = \sqrt{\sigma_x^2 + \sigma_y^2}$

Posizionamento verticale: $VDOP = \sqrt{\sigma_z^2}$

Determinazione tempo: $TDOP = \sqrt{\sigma_T^2}$

Diluizione geometrica di precisione: $GDOP = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + \sigma_T^2}$

Per un ricevitore capace di inseguire 4 satelliti simultaneamente è stato mostrato che il *PDOP* (e quindi *GDOP*) è inversamente proporzionale al volume V della figura spaziale formata dai vettori unitari dal ricevitore dai 4 satelliti.
La configurazione geometrica a massimo volume con 4 satelliti coincide con quella ottaedrica e nel caso del point positioning statico o cinematico in 4 dimensioni la scelta ottimale coincide con il minimo GDOP.

Qualunque sia la modalità di rilievo GPS impiegata, sono consigliabili valori di GDOP e di PDOP minori di 6 anche se, in una lunga acquisizione, possono essere tollerati valori superiori per brevi intervalli di tempo. Tanto minore è il tempo di stazionamento sui punti, tanto più importante è una buona configurazione satellitare e conseguentemente l'influenza del valore degli indici di DOP sull'affidabilità dei risultati del calcolo. Gli indici sopra descritti, naturalmente variano continuamente a causa del moto dei satelliti; quindi, si comprende come l'analisi delle orbite satellitari consente di individuare, per la zona del rilievo, possibili configurazioni "deboli" rispetto ad alcuni parametri e quindi definire, da un punto di vista puramente geometrico, i periodi ottimali per lo stazionamento che sono sintetizzati nella seguente tabella:

Finestra	**Satelliti**	**DOP**
Buona	5 o più	≤ 5
Utilizzabile ma non raccomandata	4	≤ 7
Inutilizzabile	4	≥ 7
Assolutamente inutilizzabile	3 o meno	

A4. Codifica di Huffman

La Codifica di Huffman [Huffman D.A., 1952] è un algoritmo greedy di codifica ideato da David A. Huffman, dottorando presso il MIT. Questo algoritmo è ispirato ad alcuni concetti base dell'algoritmo di Shannon-Fano, ma a differenza di questo produce un codice ottimale. Alla base dell'algoritmo di compressione c'è la costruzione di un albero binario che serve per ricavare il codice stesso. Lo schema dell'algoritmo è molto semplice:

1. dopo aver computato la frequenza di ogni simbolo di un messaggio originale M, ciascun simbolo è visto come un albero formato dalla sola radice ed è disposto in ordine non decrescente.
2. i due alberi meno frequenti sono uniti in un nuovo albero, che può essere considerato come un unico albero di frequenza pari alla somma delle frequenze dei due alberi.
3. si continua in maniera iterativa unendo sempre i due alberi aventi frequenza minore.

L'albero binario contenente tutti i simboli è chiamato albero di Huffman. In particolare un albero binario pesato è un albero di Huffman se soddisfa le seguenti proprietà:
le foglie hanno peso non negativo e il peso di ogni nodo interno è la somma dei pesi dei suoi figli;
l'albero gode della proprietà del sibling, cioè: i nodi possono essere elencati, tramite una lista L, in ordine non decrescente di peso; ogni nodo appare adiacente a suo fratello nella lista L e i figli precedono sempre il loro padre.

Si può osservare che in tale albero i simboli più frequenti sono più vicini alla radice. Assegnando 1 a tutti i rami verso destra e 0 a tutti quelli verso sinistra (o viceversa), ad ogni simbolo del messaggio originale sarà associato il codice ottenuto unendo in sequenza i bit che si incontrano nel cammino dalla radice fino al simbolo considerato. Nel caso in cui l'albero fosse costituito da un solo nodo, il codice associato al simbolo sarebbe indifferentemente 1 o 0.
Si produce in output il nuovo testo compresso, sostituendo ad ogni simbolo il codice ottenuto.
Una dimostrazione di come funziona l'algoritmo è illustrato nella figura A11. Se la parola da codificare è per esempio "HUFFMAN", avremo che :
• I simboli utilizzati sono {A,F,H,M,N,U};
• il numero di simboli totali è 7;
• Frequenza del simbolo 'A'=1/7;
• freq. 'F'=2/7;
• freq. 'H'=1/7;
• freq. 'M'=1/7;
• freq. 'N'=1/7;

• freq. 'U'=1/7;

A partire dal codice ottenuto, la parola "HUFFMAN" sarà codificata dalla seguente stringa binaria: "001 10 11 11 010 000 011".

La fase di decodifica avviene alla stessa maniera di quella di Shannon-Fano: al decodificatore insieme al messaggio codificato viene passata anche la tabella delle corrispondenze creata nella fase di codifica, la parola di codice iniziale del messaggio codificato viene identificata senza alcuna ambiguità dato che si tratta di un codice prefisso; viene quindi trasformata nel simbolo originario, rimossa dal file codificato e viene ripetuto il procedimento di decodifica sulla parte rimanente del file codificato.

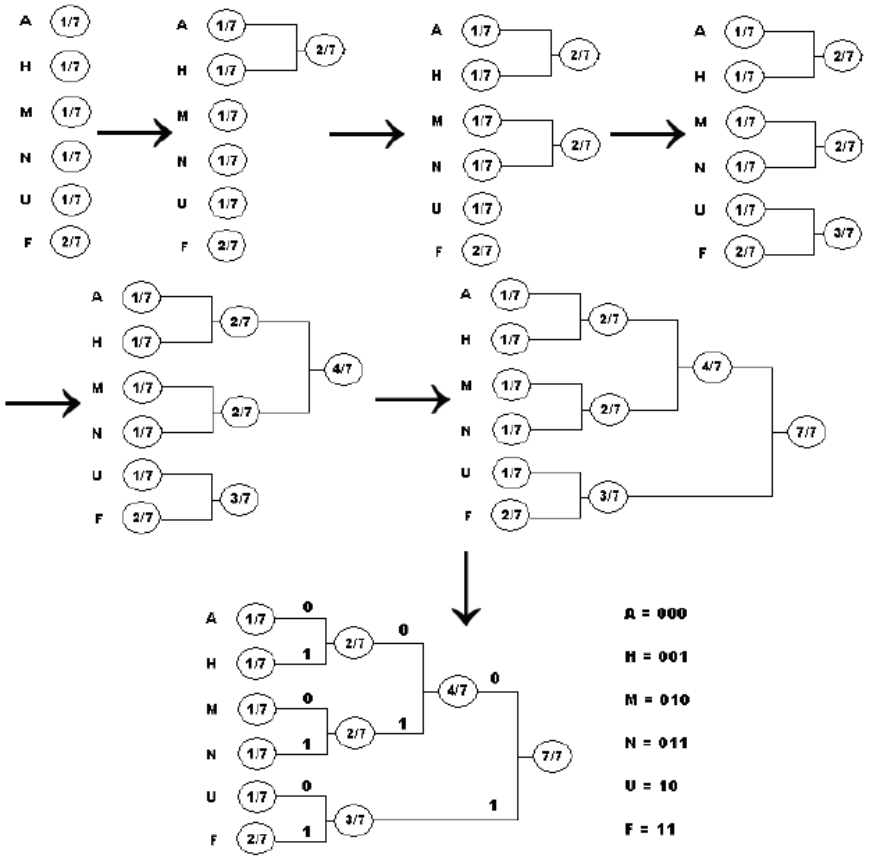


Figura A11: Schema della procedura di codifica di Huffman

Nel seguente pseudocodice supponiamo che C sia un insieme di simboli che rappresentano l'alfabeto su cui è costruito *M*, e che ogni elemento $c \in C$sia un oggetto con una frequenza definita $f[c]$. L'algoritmo costruisce l'albero T che corrisponde al codice ottimo nel modo seguente: inizia con un insieme di |C| foglie ed esegue una sequenza di |C|−1 fusioni per creare l'albero finale. Una coda di min-priorità Q, con chiave f, è utilizzata per identificare i due oggetti con frequenze minime da fondere insieme. Il risultato della fusione dei due oggetti è un nuovo oggetto la cui frequenza è la somma delle due frequenze dei due oggetti originali.

```
HUFFMAN-CREATECODE(C)
1 n ← |C|
2 Q ← C
3 for i ← 1 to n – 1
4 do alloca un nuovo nodo z
5 left[z] ← x ← Exstract –Min(Q)
6 right[z] ← y ← Exstract –Min(Q)
7 f[z] ← f[x] + f[y]
8 Insert(Q, z)
9 return Exstract –Min(Q)
```

L'algoritmo di Huffman è molto più efficace su distribuzioni di frequenza molto disomogenee. Se la distribuzione degli elementi è troppo uniforme può diventare addirittura più lungo dell'originale. La codifica di Huffman è ottimale quando la probabilità di ciascun simbolo in input è una potenza negativa di due, mentre il caso limite è collegato alla sequenza di Fibonacci. Infatti se le frequenze dei simboli coincidono con i termini della successione di Fibonacci, l'albero risultante appare bilanciato tutto da una parte, quindi la lunghezza dei codici per ogni simbolo, a partire da quello più frequente che sarà codificato con un bit, aumenterà di un bit per ogni simbolo man mano che si scende nell'albero. Solo i due simboli meno frequenti avranno un codice della stessa lunghezza, differente soltanto per l'ultimo bit.

Codifica di Huffman adattiva

Un inconveniente della codifica di Huffman, o meglio dei codici a lunghezza variabile (e quindi anche della codifica di Shannon-Fano), è che per calcolare le frequenze dei singoli simboli bisogna preliminarmente esaminare tutto il file, e ciò può essere dispendioso in termini di tempo. Proprio per ovviare a questo problema esiste una variante della codifica di Huffman standard, chiamata codifica di Huffman adattiva, che comprime il messaggio man mano che viene letto, in modo da ottenere dei risultati in forma immediata.
L'algoritmo di codifica si basa principalmente su due fasi:
1. inizializzazione dell'albero
2. aggiornamento dell'albero, che consiste in: incrementare il numero di occorrenze di un carattere nell'albero; ogni volta che viene aggiornato un nodo, bisogna controllare che non venga violata la proprietà del sibling ed in caso contrario essa deve essere ripristinata.

Parallelamente alla fase di costruzione dell'albero, viene creato il testo codificato in output. Da tale codifica è possibile risalire all'albero originario e quindi effettuare la decodifica. In definitiva è possibile concludere affermando che l'algoritmo di Huffman adattativo risulta essere più veloce di quello standard in quanto esegue una sola scansione del testo in input al posto delle due effettuate dall'altro algoritmo. Come conseguenza, la complessità di tempo diminuisce. Un altro vantaggio è che non è

necessaria la memorizzazione dell'albero in quanto la decompressione avviene in modo analogo alla compressione.

A5. Filtro Savitzky-Golay

Il filtro di Savitzky-Golay deriva direttamente da una particolare formulazione del problema di smoothing nel dominio del tempo. Nella pratica, il filtro Savitzky-Golay è utilizzato per rimuovere il rumore sovrapposto a segnali limitati in banda. L'idea su cui si basa questo filtro è la seguente: si prendono N campioni del segnale originale, si calcola il miglior interpolatore polinomiale di ordine prefissato, si sostituisce al campione centrale il corrispondente valore della funzione polinomiale approssimante. Si può dimostrare che questa operazione può essere effettuata usando un filtro FIR con opportuni coefficienti. Il filtro poi è applicato ad una serie di campioni equispaziati $f_i \equiv f(t_i)$ per cui vale $t_i \equiv t_0 + i\Delta$, con Δ costante spaziale e $i = \ldots, -2, -1, 0, 1, 2, \ldots$.
Il più semplice filtro digitale possibile, non ricorsivo e con risposta finita all'impulso, sostituisce ogni campione f_i con una combinazione lineare g_i di un sottoinsieme di campioni vicini:

$$g_i = \sum_{n=n_s}^{n_D} c_n f_{i+n}$$

con c_n opportuni coefficienti.
Si ha che n_S e n_D sono rispettivamente il numero di campioni di sinistra e di destra, da utilizzare per la combinazione lineare. Il filtro si dice causale se risulta $n_D = 0$, per cui la sommatoria non riguarda campioni successivi a quello preso in esame.
Per capire il modo di funzionamento del filtro Savitzky-Golay si può considerare la più elementare procedura per il calcolo della media. Fissato n_S=n_D si calcola ogni g_i come l'effettiva media dei campioni tra f_{i-nS} e $f_{i\,+\,nD}$. Tale procedura è chiamata "moving window averaging" e corrisponde all'equazione precedente, con:

$$c_n = \frac{1}{n_S + n_D + 1}$$

Nel caso in cui la funzione principale del segnale sia costante oppure monotona crescente o decrescente, non viene aggiunta ai risultati nessuna distorsione.
Una distorsione è comunque aggiunta se la funzione da analizzare ha una derivata seconda nulla. In corrispondenza di un massimo locale, ad esempio, la media a finestra mobile riduce sempre il valore risultante della funzione. Nel campo spettrografico questa riduzione corrisponde ad una diminuzione della larghezza e dell'altezza di una linea dello spettro. Dato che questo effetto riduce alcuni parametri caratteristici del segnale, bisogna cercare di evitare questo tipo di distorsione. Si sostituisce alla finestra di tipo rettangolare, in cui la pesatura dei coefficienti è costante, con una finestra di tipo polinomiale o di ordine superiore, ad esempio quadratica oppure del quarto ordine. Per ogni campione f_i si effettua un fitting polinomiale ai minimi quadrati dei n_S+n_D+1

campioni della finestra scorrevole e si prende come coefficiente g_i l'elemento che si trova in posizione *i*-esima. Il resto dei punti viene ignorato nel polinomio calcolato. Nel passaggio ad un campione seguente nella successione, si effettua un nuovo calcolo completo di fitting ai minimi quadrati, ottenuto trasponendo la finestra.
Dal punto di vista computazionale questo procedimento risulta molto dispendioso, in quanto richiede per ogni campione un numero di operazioni decisamente cospicuo. Per fortuna si può notare che il processo di fitting richiede solo l'inversione di una matrice. E' possibile effettuare a priori tutti fitting per tutti i campioni, in quanto ognuno di essi è ottenibile come combinazione lineare di fitting semplici, in cui la matrice è composta di tutti zeri ed un unico valore uguale ad uno. Questa è l'idea chiave attorno alla quale si sviluppa la costruzione del filtro. In particolare esistono dei set specifici di coefficienti c_n, per cui risulta che l'equazione base del filtro soddisfa automaticamente il processo di approssimazione ai minimi quadrati, all'interno della finestra scorrevole. Per derivare tali coefficienti basta pensare a come può essere ottenuto g_0. Si intende approssimare con un polinomio di grado *M* nella variabile *i*, del tipo con $a_0 + a_1 i + a_2 i^2 \cdot \cdot \cdot + a_M i^M$ i valori $f_{-nS}, \ldots, f_{nD}$.
Allora g_0 è il valore del polinomio in corrispondenza di $i = 0$, a cui diamo il nome a_0. La matrice relativa a questo problema diventa
$A_{i,j} = i^j$ con $i = -n_S, \ldots, n_D$ e $j = 0, \ldots, M$.
Le equazioni normalizzate per il vettore degli a_j , espresso in termini del vettore delle f_i, utilizzando la notazione matriciale, risultano le seguenti: $\left(\boldsymbol{A^T \cdot A}\right) \cdot \boldsymbol{a} = \boldsymbol{A^T \cdot f}$ oppure
$\boldsymbol{a} = \left(\boldsymbol{A^T \cdot A}\right)^{-1} \cdot \boldsymbol{A^T \cdot f}$
Queste equazioni possono essere espresse anche nella forma:

$$\left\{\boldsymbol{A^T \cdot A}\right\}_{\boldsymbol{i,j}} = \sum_{\boldsymbol{k=n_S}}^{\boldsymbol{n_D}} \boldsymbol{A_{k,i} A_{k,j}} = \sum_{\boldsymbol{k=n_S}}^{\boldsymbol{n_D}} \boldsymbol{k^{i+j}}$$

e

$$\left\{\boldsymbol{A^T \cdot f}\right\}_{\boldsymbol{j}} = \sum_{\boldsymbol{k=n_S}}^{\boldsymbol{n_D}} \boldsymbol{A_{k,i} f_k} = \sum_{\boldsymbol{k=n_S}}^{\boldsymbol{n_D}} \boldsymbol{k^j f_k}$$

Dato che i coefficienti c_n sono i componenti a_0 in cui *f* è sostituita dal vettore unità e_n, $-n_S \leq n \leq n_D$ si ottiene:

$$\boldsymbol{c_n} = \left\{\left(\boldsymbol{A^T \cdot A}\right)^{-1} \cdot \left(\boldsymbol{A^T \cdot e_n}\right)\right\}_0 = \sum_{\boldsymbol{m=0}}^{\boldsymbol{M}} \left\{\left(\boldsymbol{A^T \cdot A}\right)^{-1}\right\}_{\boldsymbol{0,m}} \cdot \boldsymbol{n^m}$$

Quest'equazione dimostra che è necessaria un'unica riga della matrice inversa, che può a sua volta essere ottenuta con un'unica riduzione, mediante decomposizione LU.
Il filtro Savitzky-Golay è contraddistinto da tre parametri, n_S, n_D e *M*, che determinano completamente il suo comportamento.
Una proprietà, che deriva direttamente dalla struttura propria del filtro Savitzky-Golay, è quella che lega il livello dello smoothing con la grandezza dei parametri n_S e n_D: maggiori sono questi parametri, maggiore sarà lo smoothig, in quanto per ogni campione risulta maggiore l'effetto dell'operazione di media e quindi di riduzione della componente stocastica.

5. Ringraziamenti

Mai come alla fine di un percorso si riesce a percepire la profondità dello spazio e la distanza che separa dal punto di partenza. Si ritrovano, raccolte e vicine, tracce insospettabili di tratti di cammino, di idee e di parole messe in comune, di occasioni colte e mancate, di aspettative confermate o deluse, di persone che hanno reso unica questa breve esperienza. Al di sopra di ogni altra, il prof. Luigi Maria Ricciardi, coordinatore del nostro dottorato, che pochi mesi fa ha lasciato la sua vita come la più alta e fulgida lezione dalla quale apprendere. A lui la mia gratitudine, che trascende la sua dimensione di illustre uomo di scienza per giungere alla impareggiabile umanità e nobiltà d'animo che lo hanno reso stimato e caro a tutti.

Un pensiero di smisurata riconoscenza a Walter Balzano, che mi ha accompagnata in molte attività, con cui ho incessantemente collaborato e discusso, con cui ho condiviso mille circostanze, progetti, difficoltà, successi, cambiamenti e conferme. Senza la sua abilità didattica, la sua generosa e costante presenza e il suo genuino interesse per la ricerca mi sarebbe mancato l'interlocutore ideale per lo sviluppo del mio lavoro.

Grazie infinite alla prof.ssa Amelia Giuseppina Nobile, che con continua disponibilità mi ha indirizzata e consigliata, con equilibrio e lungimiranza, dimostrandomi stima e calore spontanei e gratificanti.

Uno speciale ringraziamento al prof. Sorin Draghici per la fiducia riposta nelle mie capacità, per la splendida ospitalità offertami presso il suo laboratorio alla Wayne State University di Detroit e per avermi concesso l'opportunità di avvicinarmi a tematiche di ricerca avanzate e con notevoli prospettive di sviluppo. Un'esperienza di studio, di cultura e di umanità che hanno qualificato enormemente questi tre anni di studio, che non sarebbe stata così perfetta e divertente senza la compagnia simpatica dei compagni si studio. Ringrazio in particolare Michele Donato e Sonia Haiduc, due giovani ricercatori e ragazzi di una affabilità e accoglienza rare, che hanno condiviso con me molto del loro tempo, facendomi sentire a casa.

Una riconoscente menzione al prof. Leopoldo Milano, prezioso custode di esperienza e generatore d'idee, con il quale ho intenzione di riprendere una ricerca interrotta, che non abbandonerò fino ad aver raggiunto la meta e al prof. Scalia Tomba, che ha saputo abilmente farmi appassionare alle tematiche del mondo dei probabilisti.

Grazie specialmente a mio figlio Stefano e a mia madre Iole, che, dotati di pazienza e spirito di sacrificio unici, spesso privati della mia presenza e condividendomi con disagio con i miei impegni di lavoro, non hanno mai smesso di incoraggiarmi e di riempirmi di attenzioni e ottimismo.

Grazie infine ai miei colleghi dottorandi che hanno condiviso con me mille momenti di incertezze e glorie.

Napoli, 30 novembre 2011

6. Lista delle figure

7. Riferimenti Bigliografici

- TIME SERIES

Box G., Jenkins G. M., Reinse G., Time Series Analysis: Forecasting & Control (4th Edition), J. Wiley ans sons, 2008.

Bradley E., Time-series analysis, in M. Berthold and D. Hand, editors, Intelligent Data Analysis: An Introduction, Springer Verlag, 1999.

Brillinger D. R., Time Series: Data Analysis and Theory (Classics in Applied Mathematics), SIAM, 2001.

Bremaud P., Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis, Springer, May 2002.

Brockwell P. J., Davis, R. A., Introduction to time series and forecasting, Springer, 2002.

Burrus C. S., Gopinath, R. A., Guo, H., Introduction To Wavelets and Wavelet Transforms, a primer, Upper Saddle River, NJ (USA): Prentice Hall, 1998.

Celis J. E., Kruhoffer M., Gromova I., Frederiksen C., Ostergaard M., Thykjaer T., et al. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics, Federàtion oj Europeàn Biochemical Societies Letters, (23892):1-15, 2000.

Chatfield C., The future of the time-series forecasting, International Journal of Forecasting Volume 4, Issue 3, 1988, Pages 411-419.

Chatfield C., The analysis of time series: an introduction, CRC press, 2004.

Claps P., Serie storiche, processi e modelli stocastici per l' idrologia e la gestione delle risorse idriche, Lezioni per il corso di III livello, Politecnico di Torino, 20-22 Novembre 2002.

Diggle P. J., Time series: a biostatistical introduction, Oxford University Press, 2004.

Grossmann A., Kronland-Martinet R. and Morlet J., Reading and understanding continuous wavelet transforms, in J.M. Combes, A. Grossman and P. Tchamitchian (eds.), Wavelets: Time-Frequency Methods and Phase Space, pp. 2-20, Springer-Verlag, Berlin, 1989.

Hamilton, J. D., Time Series Analysis, Princeton University Press, 1994.

Hannan E. J., Multiple Time Series, Wiley and sons, 1970

Kirchgässner, G., Wolters, J., Introduction to modern time series analysis, Springer, 2007.

Mc Calvey, T., Relations between time domain and frequency domain prediction error methods, Control Systems Robotics, vol. V.

Percival, D. B., Walden, A. T., Wavelet Methods for Time Series Analysis, Cambridge Series in Statistical and Probabilistic Mathematics.

Proakis, J. G., Manolakis, D. G., Digital signal processing, Pearson Prentice Hall, 2007.

Shumway, R. H., Stoffer, D. S., Time series analysis and its applications, Springer, 2010.

Smith, S. W., Digital signal processing: a practical guide for engineers and scientists, Newnes, 2003.

- SITOGRAFIA TIME SERIES

[SIT1] http://seriestoriche.istat.it/

[SIT2] http://www.pitt.edu/~super4/lecture/lec5131/index.htm

[SIT3] http://dep.eco.uniroma1.it/~carlucci/docs/Modulo06-01.pdf

[SIT4] http://www.meteorologia.it/wavelet.htm

- GPS

Bellman R. E., On the approximation of curves by line segments using dynamic programming, CACM, 4(6):284, 1961.

Blelloch G. E., Introduction to Data Compression, Course notes for: Algorithms for the real world, 2000.

Cao H., Wolfson O., and Trajcevski G., Spatio-temporal data reduction with deterministic error bounds, The VLDB Journal, vol. 15, no. 3, pp. 211–228, September 2006,

CCSDS, Lossless data compression, informational report, green book, December 2006.

Choi J.-W. and Elkaim G. H., Bézier curves for trajectory guidance, in WCECS ’08: Proceedings of the World Congress on Engineering and Computer Science, Oct. 2008, pp. 625–630.

Douglas D. H. and Peucker T. K., Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, Canadian Cartographer, vol. 10, no. 2, pp. 112–122, December 1973.

Dever C., Mettler B., Feron E., Popovic J. and Mcconley M., Nonlinear trajectory generation for autonomous vehicles via parameterized maneuver classes, Journal of Guidance, Control and Dynamics, vol. 29, pp. 289–302, 2006.

Gajewski B. J., Turner S. M., Eisele W. L., and Spiegelman C. H. Intelligent transportation system data archiving: Statistical techniques for determining optimal aggregation widths for inductive loop detector speed data, Transportation Research Record, 1719:85, 2000.

Galati S. R., Geographic Information Systems Demystified, Artech House, 2006.

Hatanaka Y., A Compression Format and Tools for GNSS Observation Data, Bulletin of the Geographical Survey Institute, Vol.55 March, 2008.

Hönle N., Grossmann M., Nicklas D., Mitschang B., Preprocessing Position Data of Mobile Objects, Proceedings of the The Ninth International Conference on Mobile Data Management MDM '08, IEEE Computer Society Washington, DC, USA, 2008.

Huffman D.A., A method for the construction of minimum-redundancy codes, Proceedings of the I.R.E., pagg. 1098-1102, September 1952.

Kaplan E. D., Hegarty C. J., Understanding GPS: principles and applications, Artech House, 2006.

Langley R.B., Dilution of precision, GPS World, May 2001.

Lever R., Hinze A. and Buchanan G., Compressing GPS Data on Mobile Devices, On the move to meaningful internet systems 2006: OTM 2006 WORKSHOPS, Lecture Notes in Computer Science, Volume 4278, 2006.

Lombardi, M. A., Nelson, L. M., Novick, A. N., Zhang, V. S., Time and frequency Measurements Using the Global Positioning System, Cal. Lab. Int. J. Metrology, pp. 26-33, July-September 2001.

Misra P., Enge P.K., Global Positioning System: Signals, measurements, and performance, second edition, Ganga-Jamuna, 2006

Muckell J., Hwang J.-H., Lawson C.T., Ravi S. S., Algorithms for compressing GPS trajectory data: an empirical evaluation, ACMGIS'10, November 2-5, 2010, San Jose, CA, USA.

M. Potamias, K. Patroumpas, and T. Sellis. Sampling trajectory streams with spatio-temporal criteria, 18th International Conference on Scientific and Statistical Database Management (SSDBM'06), pages 275-284, 2006.

NMEA Reference Manual. SiRF Technology, Inc.2007.

Sayood K., Lossless Compression Handbook. Academic Press, 2003.

Tan S.-L., Lee C.-H., P. Q.-H., A comparative study of compression techniques for GPS Data for tectonic and volcanic observations, Proceedings of the 2009 IEEE 9th Malaysia International Conference on Communications, Kuala Lumpur Malaysia,15 -17 December 2009.

Talbot N.C. , Compact Data Transmission Standard for High-Precision GPS, Proceedings of ION-GPS-96; Kansas City, Missouri, 861-871, USA, 1996.

Xu G., GPS Theory, Algorithms and Applications, Second Edition, Springer, 2007.

Yi P., Sheng L., and Yu L., Wavelet transform for feature extraction to improve volume adjustment factors for rural roads, Transportation Research Record, 1879:24-29, 2004.

Zogg J-M., GPS Basic, Introduction to the system Application overview, u-blox AG, GPS-X-02007, 2002.

- SITOGRAFIA GPS

[SIT5] http://www.mathworks.com/matlabcentral/fileexchange/18264
[SIT6] http://www.w3.org/TR/xpath
[SIT7] http://www.topografix.com/gpx.asp
[SIT8] http://www.dmoz.org/Science/Earth_Sciences/Geomatics/Global_Positioning_System/

[SIT9] http://waas.stanford.edu/
[SIT10] http://www.kowoma.de/en/gps/errors.htm
[SIT11] http://www.pnt.gov/public/docs/2008/spsps2008.pdf
[SIT12] http://www.data-compression.com/index.shtml
[SIT13] http://www.cs.princeton.edu/~rs/AlgsDS07/20Compression.pdf
[SIT14] http://mattmahoney.net/dc/dce.html

- PATHWAYS

Aach, J. and Church, G. M., Aligning gene expression time series with time warping algorithms, Bioinformatics, Vol. 17 no. 6, Pages 495–508, 2001.

Androulakis I. P., Yang E., and Almon R. R.. Analysis of time-series gene expression data: Methods, challenges, and opportunities. Annu Rev Biomed Eng, 9:205–228, 2007.

Alon, U., Network motifs: theory and experimental approaches, Nature Reviews Genetics 8, 450-461, June 2007.

Anderle P., Duval M., Draghici S., Kuklin A., Littlejohn T. G., Medrano J. F., Vilanova D. and Roberts M. A., Gene expression databases and data Mining. BioTechniques, Microarrays and Cancer: Research and Applications, Suppl: 36-44, March 2003.

Arbeitman M.N., Furlong E.E., Imam F.J., Johnson E., Null B.H., Baker B.S., Krasnow M.A., Scott M.P., Davis R.W. and White K.P. Gene expression during the life cycle of Drosophila melanogaster. Science, 298, 2270–2275, 2002.

Balasubramaniyan R., Hullermeier E., Weskamp N. and Kamper J., Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics, 21:1069–1077, 2005.

Bar-Joseph Z., Analyzing time series gene expression data. Bioinformatics, 20(16):2493–2503, 2004.

Belkin N.J., Kantor P., Combining the Evidence of Multiple Query Representations for Information Retrieval, Information Processing & Management, Vol. 31, No. 3, pp. 431-448, 1995.

Bergkvist A., Rusnakova V., Sindelka R., Garda J. M., Sjogreen B., Lindh D., Forootan, A. and Kubista M., Gene expression profiling–Clusters of possibilities, Methods, 50:323–335, 2010.

Boutros P. and Okey A., Unsupervised pattern recognitions: An introduction to the whys and wherefores of clustering microarray data, Briefings in Bioinformatics, 6(4):331– 43, 2005.

Canales R. D., Luo Y., Willey J. C., Austermiller B., Barbacioru C. C., Boysen C., Hunkapiller K., Jensen R. V., Knight C. R., Lee K. Y., Ma Y., Maqsodi B., Papallo A., Peters E. H., Poulter K., Ruppel P. L., Samaha R. R., Shi L., Yang W., Zhang L., and Goodsaid F. M., EvaIuation of DNA microarray results with quantitative gene expression platfonns. Nature Biotechnology, 24(9): 1115-22, September 2006.

Churchill G. A., Fundamentals of experimental design for cDNA microarrays. Nature Genetics Supplement, vol. 32, Dec. 2002, pp.490-495.

Chung H.-J., M. Kim, C. H. Park, J. Kim, andJ. H. Kim. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using scalable vector graphics. *Nucleic Acids Research,* 32(Web Server issue):W460-W464, Jul 2004.

Dahlquist K., SaIomonis N., Vranizan K., Lawlor S., and Conklin B., GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics,* 31(1):19-20, May 2002.

Di Camillo B., Sanchez-Cabo F., Toffolo G., Nair S. K., Trajanoski Z., and Cobelli C., A quantization method based on threshold optimization for microarray short time series, BMC Bioinformatics, 6 Suppl 4:S11, 2005.

D'Haeseleer P., How does gene expression clustering work? Nature Biotechnology 23, 1499 – 1501, 2005.

Dimmer E., Huntley R., Barrell D., Binns D, Draghici S., Camon E., Hubank M., Talmud P., Apweiler R., Lovering R., The Gene Ontology: providing a Functional role in Proteomic studies. Practical Proteomics, 17 Jul 2008.

Doniger S. W., Salomonis N., Dahlquist K. D., Vranizan K., Lawlor S. C., and Conklin B. R.. MAPPFinder: using gene ontology and genmapp to create a global gene expression profile from microarray data. *Genome biology,*4(1):R7,2003.

Draghici S., Statistical and Data Analysis for Microarrays using R, CRC Press, 2011.

Draghici S., Khatri P., Tarca A. L., Amin K., Done A., Voichita C., Georgescu C., and Romero R., A systems biology approach for pathway level analysis. Genome Research, 17:1537-1545, 2007.

Draghici S., Khatri P., Eklund C. A., and Szallasi Z.. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet, 22(2):101-9,2006.

Draghici S., Khatri P., Martins R. P., Ostermeier G. C., and Krawetz S. A., Global functional profiling of gene expression. Genomics, 81(2):98-104, February 2003.

Draghic, S., Statistical intelligence: effective analysis of high-density microarray data, Drug Discovery Today, 7(11):S55-S63, 2002.

Duggan D., Bittner M., Chen Y., , Meltzer P., and Trent J.. Expression profiling using cDNA microarrays. Nature Genet., 21(1 Suppl):1O-14, 1999.

Eisen M. B., Spellman P. T., Brown P. O. and Botstein D., Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, USA, 95:14863–14868, 1998.

Ernst J. et al., (2005) Clustering short time series gene expression data. Bioinformatics, 21 (Suppl 1), I159–I168.

Faloutsos C., Ranganathan M., Manolopoulos Y., Fast Subsequence Matching in Time Series Databases, Proceedings of the 1994 ACM SIGMOD international conference on Management of data, Proceedings of the 1994 ACM SIGMOD international conference on Management of data, ACM New York, NY, USA ©1994.

Fan J.-B., X. Chen, M. Halushka, A. Berno, X. Huang, T. Ryder, R. Lipshutz, D. Lockhart, and A. Chakravarti. Parallel genotyping of human SNPs using generic high density oligonucletide tag arrays. Genome Research, (10):853-860, 2000.

Farina L., De Santis, A., Salvucci, S., Morelli, G., Ruberti, I., Embedding mRNA Stability in Correlation Analysis of Time-Series Gene Expression Data. PLoS Computational Biology, 2008, 4(8): e1000141.

Gasch A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O., Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell, 11, 4241–4257, 2000.

Geier F., Timmer J., and Fleck C., Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge, BMC Systems Biology 2007, 1:11.

Gillespie C.S., Lei G., Boys R. J., Greenall A. and Wilkinson D.J., Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays, BMC Research Notes, 3:81, 2011.

Ginsberg S. D., Hemby S. E., S. E. Lee, V. M. Lee, and J. H. Eberwine, Expression profiling of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. Annals of Neurology, 48:77–87, 2000.

Goeman J.I., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20(1):93-99, 2004.

Goldin D. Q. and Kanellakis P. C., On similarity queries for time series data: Constraint specification and implementation. In Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming (CP'95), Cassis, France, 1995. Springer Verlag.

Grosu P., Townsend J. P., Hartl D. L., and Cavalieri D., Pathway processor: a tool for integrating whole-genome expression results into metabolic networks, *Genome Research,* 12:1121-1126,2002.

Guillemin K., Salama N., Tompkins L. and Falkow S., Cag pathogenicity island-specific responses of gastric epithelial cells to Helicobacter pylori infection. Proc. Natl Acad. Sci. USA, 99, 15136–15141, 2002.

He F. and Zeng A. P., In search of functional association from time-series microarray data based on the change trend and level of gene expression, BMC Bioinformatics, 7:69, 2006.

Holford M., Li N., Nadkarni P., and Zhao H., VitaPad: visualization tools for the analysis of pathway data. *Bioinformatics,* 21(8):1596-1602, Apr 2004.

Hosack D. A., Dennis G., Sherman B. T., Lane H. C. and Lempicki R. A., Identifying biological themes within lists of genes with EASE, Genome Biology 2003, 4:R70.

Hui-Huang Hsu, Andy C. Yang, Ming-Da Lu, KNN-DTW Based Missing Value Imputation for Microarray Time Series Data Journal of Computers, Vol. 6, No. 3, March 2011.

Kalaitzis A. A. and Lawrence N. D., A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression, BMC Bioinformatics, 12:180, 2011.

Kari L., DNA Computing: arrival of biological mathematics, The Mathematical Intelligencer, 19(2):9-22, 1997

Khatri P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A., Profiling gene expression using Onto-Express. Genomics, 79(2):266-270, February 2002.

Khatri P. and Draghici S., Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21 (18):3587-3595, 2005.

Khatri P. and Sarwal M. M., Functional Pathway Analysis for Understanding Immunologic Signature of Rejection, in F. M. Marinicola and E. Wang, Immunologic Signatures of Rejection, Springer, 2010, pagg. 239-257.

Kin-Pong Chan and Ada Wai-Chee Fu, Efficient time series matching by wavelets, International Conference on Data Engineering, pages 126-133, 1999.

Kim J. and Kim J. H.. Difference-based clustering of short time-course microarray data with replicates. BMC Bioinformatics, 8:253, 2007.

Kuenzel L., Gene clustering methods for time series microarray data, Biochemistry 218, June 6, 2010.

Ma P., Zhong W., Liu J. S., Identifying Differentially Expressed Genes in Time Course Microarray Data, Statistics in Biosciences, 1: 144–159, 2009.

McCormick K., Shrivastava R. and Liao L., Slopeminer: An improved method for mining subtle signals in time course microarray data. In F. P. Preparata, X. Wu, and J. Yin, editors, Frontiers in Algorithmics, volume 5059 of Lecture Notes in Computer Science, pages 28–34. Springer, 2008.

Moller-Levet C. S., Cho K.-H., Yin H. and Wolkenhauer O., Clustering of gene expression time-series data. Technical report, University of Rostock, 2003.

Mootha V. K., Lindgren C. M., Eriksson K.-F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., M. Ridderstrrue, E. Laurila, N. Houstis, M. 1. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. A1tshuler, and L. C. Groop. PGC-l alpha-responsive genes involved in oxidative phosphorylation are co-ordinately downregulated in human diabetes. Nature Genetics, 34(3):267-273, July 2003.

Mörchen F., Ultsch A., Mining Hierarchical Temporal Patterns in Multivariate Time Series, Lecture Notes in Computer Science, Volume: 3238, Springer, 127-140, 2004.

Mörchen F., Time series feature extraction for data mining using DWT and DFT, Technical Report No. 33, Departement of Mathematics and Computer Science Philipps-University Marburg, 2003.

Nikitin A., Egorov S., Daraselia N., and Mazo I., Pathway Studio - the analysis and navigation of molecular networks. *Bioinformatics, 19(16):2155-2157,*2003.

Ogata H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, et al., KEGG: Kyoto Encyclopedia of Genes and genomes. Nucleic Acids Research, 27(1):29-34, 1999.

Page, L., Brin, S., Motwani R., and Winograd T.. The PageRank citation ranking: Bringing order to the web. Technical report, 1998.

Pan D., N. Sun, K.-H. Cheung, Z. Guan, L. Ma, M. Holford, X. Deng, and H. Zhao. PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arbidopsis. BMC Bioinformatics, 4(1):56, Nov 2003.

Pandey, R., Guru, R. K. and D. W. Mount. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. Bioinformatics, 20(13):*2156*- 2158, Sep 2004.

Pavlidis P., Qin, J., Arango, V., J. J. Mann, and E. Sibille. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. Neurachemical Research, 29(6): 1213 -1222, June 2004.

Peddada S. D., Lobenhofer E. K., Li L., Afshari C. A., Weinberg C. R., and Umbach D. M.. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics, 19:834–841, May 2003.

Phan S., Famili F., Tang Z., Pan Y., Liu, Z., Ouyang, J., Lenferink, A., and O'Connor, M. M.-C., A novel pattern based clustering methodology for time series microarray data, International Journal of Computer Mathematics, 84(5):585–597, 2007.

Phang T. L., Neville, M. C., Rudolph, M., and Hunter, L.. Trajectory clustering: a non-parametric method for grouping gene expression time courses, with applications to mammary development. Pacific Symposium On Biocomputing, 8:351–362, 2003.

Popivanov, I., and Miller, R. J.,Similarity search over time-series data using wavelets, International Conference on Data Engineering, page 0212, 2002.

Quackenbush J., Computational analysis of microarray data. Nature Reviews Genetics, 2:418-427, 2001.

Qian J., Dolled-Filhart M., Lin J., Yu H. and Gerstein M.. Beyond synexpression relationships: Local clustering of time shifted and inverted gene expression profiles identifies new, biologically relevant interactions. Journal of Molecular Biology, 314:1053–1066, 2001.

Raghavan N., Amaratunga, D., Cabrera, J., Nie, A., Qin, J. and McMillian, M., On Methods for Gene Function Scoring as a Means of Facilitating the Interpretation of Microarray Results, Journal of Computational Biology Volume 13, Number 3, Mary Ann Liebert, Inc., Pp. 798–809, 2006.

Rahnenführer J., F. S. Domingues, J. Maydt, and T. Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. Statistical Applications in Genetics and Molecular Biology, 3(1), 2004.

Ramoni M. F., Sebastiani P. and Kohane I. S., Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences, USA, 99:9121–9126, 2002.

Ross J. M., C. Fan, M. D. Ross, T.-H. Chu, Y. Shi, L. Kaufman, W. Zhang, M. E. Klotman, and P. E. Klotman, HIV- 1 infection initiatites an inflammatory cascade in human renal tubular epithelial cells, Journal of Acquired Immune Deficiency Syndromes, 42(1):1–11, 2006.

Sahoo D., Dill D. L., Tibshirani, R. and S. K. Plevritis, Extracting binary signals from microarray time-course data. Nucleic Acids Research, 35:3705–3712, 2007.

Schena M., Microarray Biochip Technology, EatonPublishing, SunnyvaIe, CA, 2000.

Schliep A., Schonhuth A., and Steinhoff C., Using hidden Markov models to analyze gene expression time course data. Bioinformatics, 19 Suppl 1:i255– 263, 2003.

Shannon P., Markiel A., Ozier O., BaIiga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., and Ideker T., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research,* 13:2498-504, 2003.

Shi, J., Walker, M. G., Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles, Current Bioinformatics, 2007, 2, 000-000.

Smyth G. K., Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397–420, 2005.

Spellman P.T., Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273–3297,1998.

Stelling J., Mathematical models in microbiaI systems biology. *Current opinion in microbiology,* 7(5):513-8, 2004.

Storey J.D., Xiao, W., Leek J.T., Tompkins, R., Davis, G., Significance of time course microarray experiments. Proceedings of National Academy of Science 102:12837–12842, 2005.

Stoughton R. B., Applications of DNA microarrays in biology. Annual Review of Biochemistry, 74:53–82, 2005.

Subramanian A., Tamayo P., Mootha V. K., Mukherjee S., Ebert B. L., Gillette M. A., Paulovich A., Pomeroy S. L., Golub T. R., Lander E. S., and Mesirov J. P.. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding oj The National Academy oj Sciences oj the USA,* 102(43):15545-15550, 2005.

Tai Y.C., Speed, T.P., A multivariate empirical Bayes statistic for replicated microarray time course data. The Annals of Statistics 34:2387, 2006.

Tarca A. L., Draghici S., Khatri P., Hassan S., Mittal P., Kim J. S., Kim C. J., Kusanovic J. P., Romero R., A Novel Signaling Pathway Impact Analysis (SPIA). Bioinformatics 2009 25(1):75-82.

Tavazoie S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics,*22:281-285,1999.

Tian L., Greenberg S. A., Kong S. W., Altschuler I., Kohane I. S., and Park P. J., Discovering statistically significant pathways in expression profiling studies, Proceedings of the National Academy of Sciences of USA, 102(38):13544-13549,2005.

Tim O., Firoiu L. and Cohen P., Clustering time series with Hidden Markov Models and dynamic time warping, 1999.

Wang X., Wu M., Li Z. and Chan C., Short time-series microarray analysis: Methods and challenges. BMC Systems Biology, 2:58–63, 2008.

Wit E., McClure, J., Statistics for Microarrays Design, Analysis and Inference. Chichester, UK: John Wiley & Sons Ltd, 2004.

Whitfield M. L., Sherlock G., Saldanha A. J., Murray J. I., Ball C. A., Alexander K. E., Matese J. C., Perou C. M., Hurt M. M., Brown P. O., and Bosteon D., Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular Biology of the Cell, 13(6):1977–2000, 2002.

Xing E. P., Jordan M. I., and Karp R. M., Feature selection for high-dimensional genomic microarray data. In ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning, pages 601–608, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2001.

Yan H., Efficient matching and retrieval of gene expression time series data based on spectral information, Computational Science and its Applications – ICCSA 2005, Lecture Notes in Computer Science, 2005, Volume 3482/2005, 251-291.

Yang-Sae Moon, Kyu-Young Whang, and Woong-Kee Loh. Efficient time-series subsequence matching using duality in constructing windows. Information Systems, 26(4):279{293, 2001.

Yeung L. K., Yan H., Liew A. W., Szeto L. K., Yang M. and Kong R., Measuring Correlation between Microarray Time-series DataUsing Dominant Spectral Component, 2nd Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand. Conferences in Research and Practice in Information Technology, Vol. 29.

Zeeberg, B., W. Feng, G. Wang, M. Wang, A. Fojo, M. Sunshine, S. Narasimhan, D. Kane, W. Reinhold, S. Lababidi, K. Bussey, J. Riss, J. Barrett, J. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biology, Vol. 4, No. 4. (2003), R28.

• SITOGRAFIA PATHWAYS

[SIT15] http://www.illumina.com/documents/products/datasheets/datasheet_whole_genome_dasl.pdf

[SIT16] http://www.r-project.org/

[SIT17] http://www.bio.davidson.edu/courses/genomics/chip/chip.html

[SIT18] http://www.eurekalert.org/pub_releases/2011-06/wifb-spp060611.php

[SIT19] http://vcell.ndsu.nodak.edu/animations/

[SIT20] http://amigo.geneontology.org/

[SIT21] http://www.broadinstitute.org/gsea/index.jsp

[SIT22] geneontology.org/GO.tools.shtml#micro

[SIT23] www.biocarta.com

[SIT24] http://www.genome.jp/kegg/

[SIT25] http://www.bioconductor.org/packages/2.6/bioc/html/SPIA.html

[SIT26] http://www.reactome.org/ReactomeGWT/entrypoint.html

[SIT27] http://www.mpi-inf.mpg.de/departments/d3/areas/scorepage.html

[SIT28] http://www.genmapp.org/

[SIT29] http://www.geospiza.com/Products/AnalysisEdition.shtml

[SIT30] http://vortex.cs.wayne.edu/ontoexpress/

[SIT31] http://smd.stanford.edu/

[SIT32] http://www.ncbi.nlm.nih.gov/geo/

[SIT33] http://www.bioconductor.org/

[SIT34] http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE22955