Ministero della Salute – Direzione Generale della Ricerca e dell'Innovazione in Sanità

Fondi 5 per mille ANNO 2016
Abstract ed elenco pubblicazioni scientifiche

Ente della Ricerca Sanitaria
Denominazione Ente:
Associazione La Nostra Famiglia – IRCCS "Eugenio Medea"
Codice fiscale: 00307430132
Sede legale: Via Don Luigi Monza n. 1 – Ponte Lambro (Co)
Indirizzo di posta elettronica dell'ente: segreteria.scientifica@pec.emedea.it
Dati del rappresentante legale: Luisa Minoli nata il 14.01.1968 a Busto Arsizio (Va) – CF: MNLLSU68A54B300V

Titolo del progetto:
Personalized Rehabilitation: nuovi approcci analitici nell'identificazione di pathway e processi biologici di risposta ai trattamenti riabilitativi

Abstract dei risultati ottenuti:

Obiettivo generale di questo progetto era dotare il nostro istituto di una serie di strumenti in grado di estrarre da un singolo genoma informazioni il più possibile utili in campo riabilitativo attraverso l'implementazione di tools di analisi che siano in grado di:
A. Identificare le varianti di un genoma in modo omogeneo tra diversi esperimenti: verranno presi in considerazione i tools disponibili scegliendo quello più adatto al nostro scopo configurandolo in modo da garantire un'adeguata sensibilità limitando al tempo stesso il numero di falsi positivi.
B. Annotare ciascuna variante in relazione a parametri di interesse quali il possibile meccanismo influenzato (splicing, trascrizione), la patogenicità, il livello di conservazione evolutiva.
C. Analizzare un intero genoma sia in termini di singole varianti che in termini di effetto combinato di insiemi varianti associabili a specifiche caratteristiche biologiche/fenotipiche dell'individuo.

Sono stati inoltre indicati i seguenti obiettivi secondari:
Implementare metodiche di machine learning in grado di predire l'effetto di varianti a livello genomico su diversi meccanismi, in particolare, dato il loro impatto, quelli che coinvolgono interazioni tra DNA/RNA e proteine.
Applicare i metodi implementati o configurati (quando si tratta di tools già esistenti) a gruppi di soggetti di particolare interesse nell'ambito riabilitativo.
Durante lo svolgimento del progetto sono state messe a punto pipelines formalizzate per l'analisi di esomi e genomi in modo da rendere uniforme il trattamento analitico dei dati sperimentali ed ottenere dati il più possibile confrontabili tra diversi esperimenti tenendo conto delle caratteristiche specifiche delle apparecchiature (sequenziatori NGS) impiegate che, necessariamente, cambiano nel tempo, tra diversi studi e per quanto riguarda studi multicentrici tra i diversi istituti coinvolti.
Per l'annotazione delle varianti identificate dal punto di vista della patogenicità e della conservazione evolutiva vengono utilizzati database pubblici; sono state messe a punto automazioni software (pipelines) per aggiornare periodicamente i dati scaricandoli localmente e per

analizzare un set di varianti (vcf files).
Per quanto riguarda invece l'annotazione funzionale abbiamo sviluppato una serie di metodi per l'identificazione di motivi di binding per proteine DNA/RNA binding. Questi motivi vengono poi utilizzati per predire, mediante approcci machine learning, il potenziale di binding intorno alle varianti identificate. Misurare la variazione di questo potenziale indotta dalla presenza di una variante consente di associare ad essa la probabilità che abbia un effetto (sia positivo che negativo) sul binding di ciascuna proteina.
In particolare è stata sviluppata una libreria software che consente le seguenti operazioni:
Identificazione di motivi di binding per specifiche proteine (sia RNA che DNA binding) a partire da esperimenti quali ChIP-seq SELEX o simili.
Classificazione e confronto dei motivi identificati rispetto a collezioni di motivi di dominio pubblico
Identificazione di set di motivi non ridondanti e specifici
Predizione di binding sites a partire dai motivi
Caratteristiche di questa libreria sono l'estrema scalabilità un base alle dotazioni hardware e la qualità delle predizioni che risulta superiore a quella dia altri metodi. È possibile ottenere i motivi e i modelli predittivi per una TF in poche decine di secondi.
Sono quindi stati analizzati 2037 diversi esperimenti ChIP-seq derivati dal progetto ENCODE ottenendo modelli predittivi per più di 844 fattori di trascrizione in 70 tipi cellulari diversi e un totale di 941 motivi specifici. Questo set consente quindi di predire l'effetto regolatorio di una variante in modo preciso e specifico per TF e tipo cellulare.

L'identificazione delle varianti funzionali è ancora una sfida chiave, in particolare per le varianti regolatorie. Infatti la conoscenza dei meccanismi regolatori è incompleta e spesso l'impatto di una singola variante è limitato. Tuttavia possiamo ipotizzare che più varianti (in particolare quelle regolatorie) abbiano un effetto combinato sulla funzionalità dei geni loro associati. Il nostro obiettivo è quello di fornire uno strumento computazionale per valutare l'arricchimento di varianti (rispetto a una frequenza di background) nei domini funzionali associati a gruppi di geni annotati per specifiche proprietà biologiche. Le regioni target potrebbero essere identificate come quelle ospitanti varianti regolatorie in un genoma, nonché attraverso esperimenti come ChIP-seq e approcci simili. Le proprietà biologiche dei geni possono includere processi biologici, funzioni e percorsi molecolari, come definito su database, come l'ontologia genica, il fenotipo umano e i percorsi di Kegg.
Un approccio analogo è stato proposto per investigare "termini biologici" associati a geni target attraverso Ontologie. GREAT è disponibile esclusivamente tramite un'interfaccia web che non è adatta per l'analisi di campioni multipli. L'analisi viene eseguita contro domini regolatori la cui definizione è limitata a tre approcci basati sulla distanza dai geni. Inoltre l'utente può scegliere solo tra un piccolo numero di database predefiniti senza alcuna indicazione su quali set di geni, definizioni di ontologie e fonti di associazioni geniche sono state integrate. Il nostro obiettivo è fornire uno strumento autonomo veloce e flessibile per superare queste limitazioni integrando ed estendendo GRANDI funzionalità mantenendo l'input il più semplice possibile.

L'organizzazione dell'algoritmo è riportata nella Figura 1. L'analisi di arricchimento si basa sia su test ipergeometrici che su test binomiali. I risultati vengono corretti per più test tramite False Discovery Rate. GREATER è progettato in C++ ed è stato testato in ambiente Linux. Le dipendenze includono le librerie BOOST C++ per l'analisi dei parametri, l'i/o dei dati e le distribuzioni statistiche, GeCo++ [2] per la rappresentazione e l'analisi delle regioni genomiche. Ove appropriato, il codice include la parallelizzazione automatica utilizzando openMP.

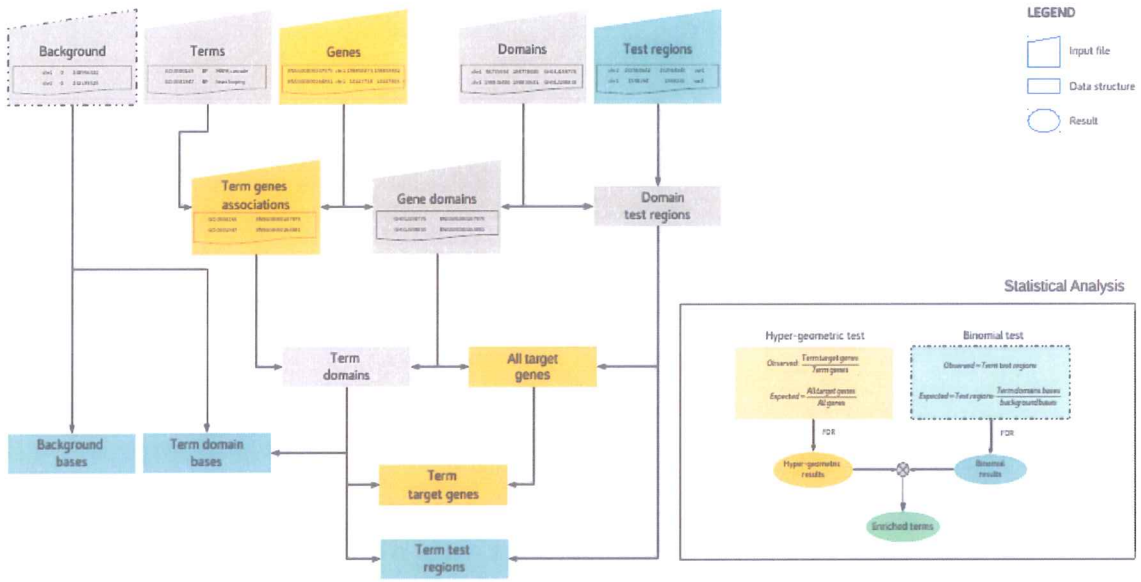

Figura 1. Organizzazione dell'algoritmo. Sulla destra il riquadro Analisi statistica descrive l'analisi di arricchimento. Nel flusso di lavoro, i contenitori blu e arancione rappresentano rispettivamente gli input necessari per il test binomiale o ipergeometrico. Le linee tratteggiate sono per input o analisi facoltativi. Le caselle di input contengono esempi dei formati di file.

Gli input tipici includono il set di geni, un insieme di regioni di dominio, l'insieme di termini di ontologia, le associazioni tra geni e domini e termini e il set di regioni target. È possibile fornire facoltativamente un set di regioni di background. A partire da questi input la costruzione delle strutture dati necessarie per l'analisi è un compito che richiede tempo. Pertanto il codice GREATER è stato organizzato per mantenere facoltativamente in memoria queste strutture al fine di analizzare più insiemi di regioni target senza sovraccarico di tempo.
GREATER è uno strumento computazionale di bioinformatica che consente all'utente di analizzare grandi set di dati (ad esempio dati di sequenziamento di nuova generazione) con tracce di annotazione del genoma sia pubbliche che personalizzate. Il tempo di esecuzione dipende fortemente dal numero di regioni di dominio da analizzare per la preparazione del set di dati e dal numero di termini da analizzare per l'arricchimento del campione. Una tipica durata di esecuzione, inclusa la creazione di set di dati, varia nell'ordine di decine di secondi. È necessario circa un minuto per creare il set di dati e dieci secondi per eseguire gli arricchimenti.
GREATER è facile da usare perché richiede formati di file di input semplificati e consente molteplici scelte di analisi (incluso l'arricchimento del set di geni): è uno strumento a riga di comando veloce e affidabile per valutare l'arricchimento delle regioni target in domini genomici selezionati, facilitando le attività genomiche di routine.

Prodotti della Ricerca (correlati al progetto):
Pipelines
Pacchetto R contenente le funzionalità per l'identificazione dei motivi di binding
Pacchetto R contenente le funzionalità di GREATER

Tre pubblicazioni sono attualmente in preparazione relativamente alla presentazione dei pacchetti di R e una relativa all'applicazione ad un gruppo di Pazienti ASD

Data, 24/02/2023

Il Responsabile del Progetto
Ing. Uberto Pozzoli

Il Legale Rappresentante
Dr.ssa Luisa Minoli

Si autorizza al trattamento dei dati ai sensi del d.lgs. 196/2003

Il Legale Rappresentante
Dr.ssa Luisa Minoli