IL PROGETTO

Combattere gravi malattie con la genomica computazionale guidata dai dati

di Stefano Ceri, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano 13 Giu 2016

TAG: analisi banche accesso costi Bologna cineca corso dati genomica data informazione medici milano processo telematico polimi open regioni scienza servizi studio UE stefano ceri

genomica computazionale dal punto di vista dei dati, tramite nuovi modelli, linguaggi e strumenti per la loro analisi e gestione. Solidi dal punto di vista dei concetti utilizzati e capaci di operare in modo super-efficiente. Il gruppo di ricerca del Politecnico di Milano è tra i primi e pochissimi al mondo a mettere al centro della ricerca la integrazione di dati genomici

La Genomica computazionale è la scienza che, partendo dal

sequenziamento del genoma e grazie all'uso di analisi statistiche e computazionali, decifra la funzione delle regioni del genoma e costituisce pertanto il presupposto per le future scoperte nel campo della biologia e della medicina. Le tecniche di sequenziamento del genoma di nuova generazione (NGS) consentono oggi la produzione dell'intera seguenza del genoma umano a costi molto bassi (circa 1000 dollari). Parallelamente sono stati sviluppati algoritmi specializzati per estrarre le caratteristiche salienti del genoma ed evidenziare le sue caratteristiche (i "segnali" che ci invia), ad esempio le mutazioni o l'espressione dei geni, cioà la loro attività di trascrizione. C'è però una grande lacuna da colmare, e cioè costruire un sistema che sia capace di integrare i dati genomici estratti da tali algoritmi, ottenendo un "senso biologico" interpretabile dai medici per comprendere meglio, ad esempio, lo sviluppo di tumori o la loro dipendenza da fattori ambientali. Il progetto GeCo (Data-Driven Genomic Computing, ERC Advanced Grant,

contratto attualmente in corso di finalizzazione con la UE) ha l'obiettivo di rivisitare la genomica computazionale dal punto di vista dei dati, tramite nuovi modelli, linguaggi e strumenti per la loro analisi e gestione, solidi dal punto di vista dei concetti utilizzati e capaci di operare in modo superefficiente. Il contratto avrà una durata di 5 anni e un finanziamento di 2.5M di Euro. Il gruppo di ricerca del Politecnico di Milano è tra i primi e pochissimi al mondo a mettere al centro della ricerca la integrazione di dati genomici. Partendo da un modello di dati astratto che garantisce interoperabilità fra i

vari formati dei dati, abbiamo già sviluppato un sistema (denominato

GenData 2020) per interrogare dati genomici; il sistema utilizza il nuovo linguaggio di interrrogazione GMQL (GenoMetric Query Language) – il cui nome è dovuto alla presenza di operazioni che calcolano la distanza tra regioni geniche, ad esempio per estrarre i geni sovra- o sotto-espressi che sono a specifiche distanze da regioni che fungono da attivatori della trascrizione. Il linguaggio offre vari operatori di alto livello, in parte standard e in parte specificamente dedicati alla manipolazione di regioni genomiche. GenData 2020 consente anche di tracciare i "metadati" che caratterizzano ciascun dato sperimentale e di trasferire questa informazione dai dati di ingresso ai risultati di una interrogazione, consentendo così di conoscere in modo dettagliato la "provenienza" dei dati (interpretare i risultati a partire dai dati di ingresso.) Il sistema è stato realizzato traducendo le interrogazioni GMQL negli

operatori di vari framework per la gestione dei dati disponibili sulle piattaforme "cloud", inizialmente (Versione 1) abbiamo usato un framework denominato "Pig", poi (Versione 2) siamo passati a "Spark" e a "Flink", e stiamo anche realizzando una traduzione verso "SciDB", un framework per la gestione di dati per applicazioni scientifiche. La disponibilità di varie traduzioni consente di utilizzare la versione di sistema più adatta allo specifico contesto e consente anche un confronto prestazionale tra i vari framework, che sono abbastanza diversi fra loro. Il sistema GenData 2020 in Versione 2 è già pubblicamente utilizzabile

oppure scaricabile in entrambe le versioni dai server del Politecnico. Nel corso del progetto ERC, il sistema GenData 20202 evolverà verso un nuovo sistema, denominato GeCo, che verrà arricchito di strumenti per l'analisi dei dati e verrà reso sempre più efficiente, utilizzando nuove tecniche di traduzione e di ottimizzazione verso i vari framework. Nel progetto GeCo offriremo un accesso integrato ai dati di tipo "processato" offerti da vari consorzi internazionali, tra cui ENCODE,

presso il Cineca (Consorzio Interuniversitario per il Calcolo Automatico),

Roadmap Epigenomics, TCGA e 1000 Genomes. In aggiunta all'accesso ai dati, cercheremo di dare uniformità ai metadati, sia tramite conversioni, sia soprattutto tramite la connessione a ontologie di dati biologici e clinici; abbiamo ad esempio già utilizzato UMLS (una ontologia universale di termini clinici) per inferire equivalenze tra i termini presenti in ENCODE. L'integrazione di metadati provenienti da consorzi differenti sarà ovviamente più difficile.

> La gestione del dato come cardine dei nuovi percorsi di trasformazione digitale - Iscriviti al

L'obiettivo più ambizioso del progetto è la realizzazione, in un prossimo futuro, di un "Internet per la genomica", cioè di un modo di raccogliere dati genomici pubblicati da consorzi internazionali e dai ricercatori, e di un "Google per la genomica", cioè un sistema di indicizzazione e ricerca su grandi raccolte di dati genomici pubblici.

Tra gli obiettivi del progetto vi è anche la costruzione di un ambiente aperto

(open source) messo a disposizione dei ricercatori biologici e clinici, che

webinar del 24 maggio

potranno usare servizi offerti dal sistema oppure scaricare e istallare il sistema presso i loro centri; mentre i servizi realizzati dal Politecnico useranno esclusivamente dati pubblici, messi a disposizione per "uso secondario" (cioè per fare attività di ricerca), l'istallazione protetta del sistema in un contesto clinico potrà favorire la cosiddetta "medicina personalizzata", cioè l'adattamento delle terapie ai dati genomici di specifici pazienti. Tramite l'uso estensivo di banche dati pubbliche, sarà possibile dare risposta a problemi biologici fondamentali come lo studio di gravi malattie.

Gran parte delle patologie hanno una componente genetica, e quindi le ricadute di un sistema capace di integrare "big data" sono potenzialmente importantissime: è un nuovo approccio alla biologia e alla medicina. Già adesso abbiamo in corso studi molto intessanti, che riguardano la classificazione dei tumori andando a caratterizzare dati di espressione genica in relazione con i "domini topologici funzionali", cioè regioni del genoma identificate tramite tecniche di segmentazione tridimensionale del genoma. Altri studi biologici riguardano l'interazione tra fattori di trascrizione, cioè proteine che, legandosi al DNA, abilitano il processo di trascrizione operato dai geni. Tra gli studi clinici, vorremmo intraprendere a breve uno studio sul microbioma polmonare dei pazienti di fibrosi cistica che sono colonizzati da micobattere. Questi studi sono resi possibili da collaborazioni con biologi e clinici di molti diversi centri di ricerca, tra cui lo IEO (Istituto Europeo di Oncologia), il

Policlinico Milano, l' Università di Singapore, il Broad Institute di Cambridge (US); altre collaborazioni informatiche coinvolgono alcune università italiane (tra cui Sapienza, Bologna e UniRoma3, che hanno partecipato al progetto PRIN GenData 2020) e europee (tra cui TU Berlin, che ha un griuppo molto attivo allo sviluppo del framework Flink).

Commenta per primo.



0 Commenti

Condividi Inizia la discussione...

Agendadigitale

Ordina dal migliore -

Accedi

SEMPRE SU AGENDADIGITALE

https://www.voutube.com/wat...

2 commenti • 22 giorni fa•

Maurizio Lunghi - esercizio o meglio 'sfida' molto interessante e molto promettente ... potrebbe rivisitare e ridisegnare compeltamente l'interazione tra cittadino e PA ... mi sembra ovvia che si dovrebbe partire da 3 passi fodnamentlai:1- definire le entità nel ... Perché l'Italia digitale è frenata da deficit di norme e cultura

Ada — Consiglio all'autore dell'articolo di guardarsi questa intervista di un paio di anni fa

Le ontologie nella PA: ecco perché anche ai dati serve la carta d'identità

giuseppe carnemolla - Pienamente d'accordo

Il 5G non sarà come vogliono farci credere. Ecco perché

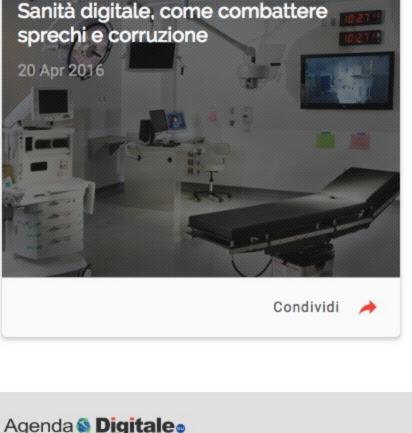
Le tre novità che cambieranno la cyber security nazionale, con il nuovo decreto 1 commento o 7 giorni fao

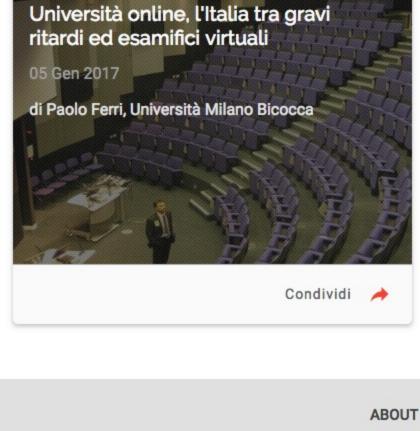
1 commento • 9 giorni fa•

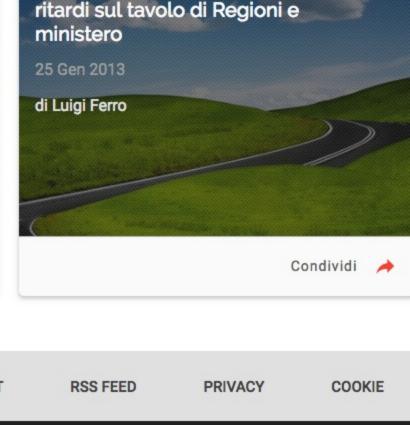
Articoli correlati

Marco Di Muzio — https://www.cvpherius.com ----> CYBER SECURITY -

L'ANALISI







Banda larga in agricoltura: gravi

NETWORK DIGITAL L.

DIGITAL 360 Group

Testate orizzontali

ECONOMYUP

STARTUP BUSINESS

UNIVERSITY2BUSINESS

FORUM PA

ZEROUNO

About

Digital360 aiuta imprese e pubbliche amministrazioni nella

Via Copernico, 38

Milano - Italia

CAP 20125

Indirizzo

Digital360 SRL - Codice fiscale 08053820968 - P.IVA: 08053820968 - © 2016 DIGITAL 360. ALL RIGHTS RESERVED - Mappa del sito