Progetti di integrazione e navigazione web di dati genomici e proteomici

I.1) Realizzazione di un modulo software di download di file da fonti dati biomolecolari
[*Assegnato: Michele Festini*]

L'obiettivo del progetto è l'implementazione in linguaggio Java di un modulo software per il download di file di dati disponibili via HTTP o FTP in banche dati genomiche o proteomiche distribuite e la loro memorizzazione in un file system locale opportunamente strutturato.

A partire da una lista di URL di file di interesse per ogni fonte dati, l'applicazione Java da costruire deve accedere agli URL specificati, scaricare i file di interesse, effettuarne la decompressione (se necessario) e memorizzarne alcune meta informazioni di interesse in un database. Il modulo di download da creare deve poter essere facilmente estendibile, configurabile tramite un file XML di configurazione e consentire di tener traccia nel corso del tempo di possibili variazioni di dimensione e locazione dei file scaricati.

Lo sviluppo del progetto richiede:

- estendere un file XML di configurazione (che verrà fornito) per l'inserimento degli URL dei file di dati da scaricare
- utilizzare la libreria Jakarta Commons Net di Apache (http://commons.apache.org/net/) per l'accesso ai file via FTP e HTTP
- valutare le librerie necessarie per gestire i tipi più comuni di file compressi (zip, gzip, bzip2, tar)
- creare una struttura modulare del file system dove memorizzare i file scaricati
- memorizzare le meta informazioni relative ai file scaricati (es. tempistiche di download, data di accesso, data di ultima modifica del file, …) in un database di supporto (di cui verrà fornito uno schema concettuale e logico).

I.2) Realizzazione di parser per l'importazione di dati genomici e proteomici in un data warehouse
[*Parzialmente assegnato: Luca Goffredi, Mattia Coti Zelati*]

Lo sviluppo di questo progetto richiede la creazione in Java di alcune procedure per l'estrazione di dati genomici e proteomici da file di testo in diversi formati (tabellare, xml, …) e la loro importazione in un data warehouse di dai biomolecolari integrati chiamato Genomic and Proteomic Data Warehouse (GPDW).

Saranno assegnati alcuni file di dati forniti da una o più sorgenti di dati genomici e/o proteomici e verranno forniti gli schemi concettuali e logici relativi a tali dati, o sarà richiesto di generarli a partire dall'analisi dei file di dati assegnati e da esempi analoghi che saranno forniti.

Dovranno essere realizzati dei Parser per l'estrazione dei dati dai file assegnati e dei Loader per l'importazione nel GPDW dei dati estratti dai Parser. La realizzazione dei Parser e dei Loader dovrà seguire un framework esistente, basato su un pattern di tipo produttore-consumatore, che verrà fornito. Tale framework prevede dei file XML di configurazione che dovranno essere compilati per configurare i nuovi Loader realizzati. Le specifiche per la compilazione verranno fornite.

Saranno anche forniti alcuni Parser per i formati di file più semplici che sono già disponibili; tali Parser potranno essere direttamente utilizzati, mentre per formati di file più complessi sarà necessario implementare nuovi Parser estendendo per quanto possibile quelli esistenti.

I.3) Testing di importazione e integrazione di dati genomici e proteomici in un data warehouse

Lo sviluppo di questo progetto richiede la creazione di test per la verifica di un workflow per l'importazione di dati genomici e proteomici disponibili in banche dati distribuite e la loro integrazione in un data warehouse centralizzato.

Lo studente dovrà realizzare dei casi di test, usando il linguaggio Java ed il framework junit4 secondo le seguenti specifiche:

- analizzare il formato dei dati in ingresso e l'output atteso
- scrivere test per verificare:
    - la correttezza (i.e. output del parser = output atteso)
    - la robustezza (i.e. il parser gestisce correttamente input anomali)
    - e la resilienza (i.e. il parser, ove previsto, gestisce correttamente variazioni non significative dell'input)
- definire le condizioni limite dei parser e scrivere i test per esercitare il codice associato a tali condizioni
- realizzare un modulo di sintesi degli input per i casi di test considerati che sia facilmente adattabile/configurabile per fare fronte ad eventuali (non infrequenti) cambi del formato dei dati

Ad ogni studente verrà affidato un numero variabile di parser da testare, in funzione della complessità del parser, dei dati da importare e del numero stimato dei test da realizzare, in modo da modulare il carico di lavoro in base ai requisiti richiesti per i progetti del corso Laboratorio software.

I.4) Realizzazione di un'interfaccia di configurazione per l'importazione e integrazione di dati genomici in un data warehouse

Il progetto prevede la progettazione e realizzazione di un'applicazione software con interfaccia grafica per la configurazione di procedure automatiche di importazione e integrazione di dati genomici in un data warehouse, nell'ambito del progetto "Virtual BioInformatics Lab".

La procedura di importazione preleva i dati di input dalle locazioni indicate, lancia i corrispondenti parser e inserisce i dati nel data warehouse, usando ad ogni passo le informazioni contenute in un file XML di configurazione; al processo di importazione seguono un numero variabile (anche zero) di passi di post-processing dei dati, definiti nel file di configurazione.

La procedura di integrazione utilizza i dati importati per creare delle entità (di entità biomolecolari o caratteristiche (feature) biomediche) di alto livello; la procedura di integrazione è guidata dalle informazioni di configurazione contenute in un file XML.

I parametri di configurazione si dividono in due categorie principali:

1. Registrazione delle singole fonti dati: dati di input, parser da utilizzare, struttura dei dati di input, eventuali relazioni con altri dati, eventuali post-processing da applicare ai dati importati

2. Definizioni delle "features" di alto livello e loro composizione in termini di sorgenti dati

L'applicazione software con interfaccia grafica di configurazione da progettare e realizzare dovrà essere sviluppata in linguaggio Java e fornire una GUI che consenta di configurare in modo visuale le procedure descritte per il popolamento del data warehouse, salvandone i dati di configurazione nei file di configurazione in formato XML; le specifiche e la documentazione del formato sono disponibili e verranno fornite.

I.5) Analisi quantitativa di dati genomici e proteomici in un data warehouse integrato

L'obiettivo di questo progetto riguarda la creazione un framework software per supportare l'analisi quantitativa di dati genomici e proteomici presenti nel data warehouse integrato Genomic and Proteomic Data Warehouse (GPDW), implementato nell'ambito del progetto "Virtual BioInformatics Lab". Tale analisi dovrà avere una doppia finalità: in primo luogo calcolare il numero totale dei dati importati e integrati nel data warehouse ad ogni suo aggiornamento, in secondo luogo individuare eventuali incongruenze ed errori di importazione o di integrazione di tali dati.

Allo studente verrà fornita una descrizione di tutti i dati presenti nel GPDW, sia dei metadati che descrivono il data warehouse e i dati in esso contenuti, sia dei dati di entità biomolecolari e feature biomediche, delle loro annotazioni con vocabolari controllati e ontologie e dei vari altri dati contenuti nel GPDW. Verrà fornita una copia campione di tale data warehouse per DBMS PostgreSQL, comprensiva di schemi concettuali e logici e del dizionario dei dati principali.

Le principali query di analisi e quantificazione dei dati verranno suggerite allo studente dal team di sviluppo, lo studente in particolare avrà il compito di:

- implementare una procedura configurabile per l'esecuzione automatica di analisi e quantificazioni di controllo che forniscano una descrizione dettagliata quantitativa dei dati presenti nel GPDW; tale procedura dovrà integrarsi con le attuali procedure configurabili di implementazione e integrazione delle sorgenti dati di cui verrà illustrato e fornito il codice e la documentazione
- progettare e implementare la struttura dello schema del database per la memorizzazione dei dati quantitativi calcolati
- definire e implementare ulteriori operazioni di analisi quantitativa sui dati.

I.6) Realizzazione di un'applicazione web per la navigazione di dati genomici e proteomici
[*Assegnato: Fabio Pozzi*]

L'obiettivo del progetto è l'implementazione in linguaggio PHP di un'applicazione web che permetta ad utenti registrati di interrogare ed esplorare le informazioni genomiche e proteomiche raccolte nel Genomic and Proteomic Data Warehouse (GPDW) sviluppato dal gruppo di Bioinformatica del Politecnico di Milano.

Verrà fornita una copia campione di tale data warehouse per DBMS PostgeSQL, comprensiva di schemi concettuali e logici, e del dizionario dei dati principali. Lo scopo del lavoro è realizzare un'interfaccia web dinamica che consenta di realizzare interrogazioni per parole chiave sulle varie informazioni contenute nel GPDW e di visualizzare i risultati estratti, in modo simile a quanto usualmente disponibile nelle interfacce web delle banche dati biomolecolari quali Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) o UniProt (http://www.uniprot.org/).

L'applicazione dovrà consentire:

- La registrazione di un utente (senza il controllo dell'abilitazione di accesso)
- Il login di un utente registrato
- Ad un utente registrato:
    - di realizzare query per la navigazione dei dati, che permettano di interrogare per codici di geni e/o per parole chiave le varie informazioni contenute nel GPDW. Tali query dovranno poter essere realizzare senza dover scrivere le query complete in linguaggio SQL, ma inserendo in un form i parametri della query. I risultati di tali query dovranno essere adeguatamente visualizzati nell'interfaccia web
    - di salvare in un proprio spazio personale l'"history" delle query realizzate che ritenga più significative, in modo tale da poterle rieseguire successivamente in modo immediato.

I.7) Realizzazione di un'applicazione web per la visualizzazione e l'interrogazione grafica di una "resource network"

Il progetto prevede la realizzazione di un'applicazione web che consenta la visualizzazione grafica dei principali metadati (tipi di dati, sorgenti che li forniscono, …) che descrivono i dati genomici e proteomici integrati nel Genomic and Proteomic Data Warehouse (GPDW), implementato nel DBMS relazionale PostgeSQL nell'ambito del progetto "Virtual BioInformatics Lab". Inoltre, l'applicazione da implementare dovrà consentire all'utente di comporre e realizzare query sui dati contenuti nel GPDW a partire dalla selezione grafica di tali metadati e dall'inserimento di keyword da usare come parametri restrittivi della query.

Lo scopo principale dell'applicazione è rappresentare il grafico a nodi dei metadati descrittivi del GPDW, il quale comprende le informazioni relative a entità biomolecolari e feature biomediche, alle loro sorgenti dati, alle relazioni tra entità biomolecolari e feature biomediche, ai dati delle ontologie usate per rappresentare le feature biomediche, ecc. La rete realizzata dovrà essere interattiva in modo da consentire agli utenti di navigare ed estrarre ulteriori informazioni dalla selezione di un nodo.

I dati in ingresso sono costituiti da tutti i dati presenti nel GPDW, sia i metadati descrittivi del data warehouse, sia i dati di entità biomolecolari e feature biomediche, di loro annotazioni, informazioni ontologiche e vari altri dati contenuti nel GPDW. Verrà fornita una copia campione di tale data warehouse per DBMS PostgeSQL, comprensiva di schemi concettuali e logici e del dizionario dei dati principali.

L'ambiente di sviluppo da utilizzare per realizzare l'applicazione dovrà essere PHP, sfruttando una libreria per la visualizzazione di grafici come JpGraph (http://www.aditus.nu/jpgraph/) o mxGraph (http://www.jgraph.com/mxgraph.html), o il web service client Cytoscape (http://www.cytoscape.org/) adatto per visualizzare dati estratti da database genomici, o direttamente Java sfruttando la libreria JUNG (http://jung.sourceforge.net/) ed un modulo che consenta di eseguire via web l'applicazione Java realizzata (es. in JSP), oppure in Flash usando il framework Flex (http://www.adobe.com/products/flex/).

I.8) Integrazione e utilizzo di servizi web bioinformatici pubblicamente disponibili

Molteplici servizi web bioinformatici sono pubblicamente disponibili per l'interrogazione di dati genomici e proteomici biomedici eterogenei, quali ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) per l'interrogazione di dati di espressione genica, BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi http://www.ebi.ac.uk/Tools/blast2/index.html) per la ricerca di sequenze nucleotidiche e aminoacidiche in banche dati di geni e proteine, PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) per la ricerca di pubblicazioni scientifiche, etc. Questi servizi permettono di rispondere a semplici interrogazioni dell'utente, fornendo risultati che spesso includono un valore di ranking di adeguatezza all'interrogazione realizzata. L'obiettivo del progetto è analizzare alcuni di tali servizi e sviluppare un framework software che, utilizzando le API disponibili per il loro utilizzo, integri i risultati forniti da tali servizi; lo scopo è di supportare interrogazioni più articolate e multi dominio, con valore aggiunto per l'utente, non direttamente eseguibili sui singoli servizi.

Dato un insieme selezionato di servizi (quali quelli precedentemente citati), lo studente dovrà:

- analizzare i servizi selezionati, identificando le loro diverse modalità di interrogazione
- individuare e "mappare" sui singoli servizi alcuni esempi di possibili interrogazioni articolate di interesse
- individuare eventuali raffinamenti o espansioni dell'interrogazione iniziale che possono essere eseguiti dall'utente a partire dai risultati dell'interrogazione iniziale stessa, al fine di navigare e usufruire meglio dei dati esposti dai servizi selezionati
- realizzare un semplice applicativo web che supporti l'esecuzione di tali interrogazioni articolate.

L'attività di progetto ha una spiccata natura esplorativa e pertanto richiede una frequente interazione con il corpo docente. Appositi strumenti software saranno forniti per facilitare la realizzazione dell'applicativo web integrato. Il software dovrà essere realizzato in Java, Javascript e HTML.

Riferimenti:

- ArrayExpress:
  - http://www.ebi.ac.uk/microarray/doc/help/programmatic_access.html
  - http://www.ebi.ac.uk/microarray/doc/atlas/api.html
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686529/
- BLAST:
  - http://www.ebi.ac.uk/Tools/webservices/services/wublast
  - http://blast.ncbi.nlm.nih.gov/blast_overview.shtml#access
- PubMed:
  - http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html
  - http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html

Progetti di ambito analisi informazioni genomiche e proteomiche

A.1) Predizione di annotazioni funzionali in basi di dati biomolecolari utilizzando un approccio pLSA

Lo sviluppo del progetto richiede l'implementazione in C++ di un algoritmo per la predizione di annotazioni in basi di dati genomiche e proteomiche basato sull'utilizzo della pLSA (probabilistic Latent Semantic Analysis), estensione dell'approccio basato sulla decomposizione SVD (Singular Value Decomposition) di matrici alla base di metodi di "Latent Semantic Indexing" utilizzati nell'ambito dell'interrogazione di basi di dati contenenti documenti testuali.

I dati in ingresso consistono nell'insieme delle annotazioni funzionali attualmente disponibili per vari organismi. Per ciascun gene e/o proteina di ognuno di tali organismi, è disponibile l'elenco dei termini che lo caratterizzano dal punto di vista funzionale, utilizzando l'insieme delle sue annotazioni descritte mediante la "Gene Ontology" e altre ontologie e vocabolari controllati. La "Gene Ontology" è un'ontologia che definisce un numero finito di termini funzionali e strutturali ammissibili e, per ciascuno, ne fornisce un identificativo univoco. Inoltre la "Gene Ontology" definisce relazioni semantiche tra i termini, sotto forma di un grafo direzionale aciclico.

L'algoritmo di predizione ha come obiettivo quello di individuare possibili annotazioni (ovvero coppie gene-termine) che non sono ancora presenti nella base di dati considerata.

Esiste un'implementazione in linguaggio C++ dell'algoritmo SVD, che riceve i dati in ingresso estratti da una base di dati relazionale PostgreSQL e produce in uscita il risultato della predizione per ogni coppia gene-termine.

Obiettivo del progetto è implementare in linguaggio C++ l'algoritmo pLSA e realizzare una versione equivalente dal punto di vista funzionale dell'implementazione che utilizza l'algoritmo SVD. I dati di input da utilizzare sono disponibili sotto forma tabellare in una base di dati relazionale PostgreSQL.

Riferimenti:

- Khatri P, Done B, Rao A, Done A, Draghici S. A semantic analysis of the annotations of the human genome. *Bioinformatics* 2005 Aug 15; 21(16): 3416-3421.
- Berry MW, Dumais ST, O'Brien GW. Using linear algebra for intelligent information retrieval. *SIAM Review* 1995 Dec; 37(4): 573-595.
- Hofmann T. Probabilistic Latent Semantic Analysis. In *Proc. Uncertainty in artificial intelligence, UAI'99*, Stockholm, 1999, pp. 289-296.

A.2) Clustering funzionale di geni o proteine basato su annotazioni in basi di dati biomolecolari

Lo sviluppo del progetto richiede l'implementazione in C++ di un algoritmo per il clustering di geni o proteine in base al loro profilo di annotazione funzionale, utilizzando approcci simili a quelli utilizzati per il clustering di dati sperimentali. In particolare sarà necessario implementare sia l'algoritmo per la scelta del numero k di clusters in cui suddividere i dati di ingresso, sia un algoritmo per definire la suddivisione e appartenenza dei dati a tali k clusters; relativamente a tale secondo tipo di algoritmo, esiste già un'implementazione in C++ dell'approccio k-Nearest Neighbours (k-NN) che può essere usata per il confronto dei risultati di altri algoritmi analoghi che verranno implementati).

I dati di ingresso da utilizzare sono disponibili sotto forma tabellare in una base di dati relazionale PostgreSQL e consistono nell'insieme delle annotazioni funzionali attualmente disponibili per uno specifico organismo. Per ciascun gene e/o proteina di tale organismo, è disponibile l'elenco dei termini che lo caratterizzano dal punto di vista funzionale, utilizzando l'insieme delle sue annotazioni descritte mediante bio-terminologie e/o bio-ontologie, tra cui la "Gene Ontology".

Riferimenti:

- Jonnalagadda S, Srinivasan R. NIFTI: an evolutionary approach for finding number of clusters in microarray data. *BMC Bioinformatics* 2009 Jan 30; 10: 40.
- Drineas P, Frieze A, Kannan R, Vempala S, Vinay V. Clustering Large Graphs via the Singular Value Decomposition, *Machine Learning* 2004; 56(1-3): 9-33.

A.3) Analisi combinatoriale di risultati proteomici sperimentali per la purificazione di proteine

Il progetto richiede di implementare in Java (o eventualmente altro linguaggio da concordare) un software in grado di compiere calcoli insiemistici su gruppi di proteine selezionate mediante esperimenti di analisi proteomica. In funzione dei risultati dei calcoli realizzati e in base a metriche di costo variabili a seconda dello specifico problema, il software dovrà "suggerire" di volta in volta una strategia sperimentale ottima da compiere per purificare, a partire dai gruppi di proteine considerati, una singola o più proteine scelte dall'utente. Il software dovrà anche calcolare la percentuale di purificazione raggiungibile alla fine del processo e di ogni suo passo, nel caso la strategia sperimentale ottima richieda più passi, fornendo la lista di proteine estratte ad ogni passo e finale. Il software sviluppato dovrà essere corredato da adeguata documentazione. I dati di input dovranno essere caricati da file (excel) o da database (che verrà fornito).

Riferimenti:

- Bachi A, Simó C, Restuccia U, Guerrier L, Fortis F, Boschetti E, Masseroli M, Righetti PG. Performance of combinatorial peptide libraries in capturing the low-abundance proteome of red blood cells. 2. Behavior of resins containing individual amino acids. *Anal Chem*. 2008 May 15; 80(10): 3557-3565.