UNIVERSITÀ DI PADOVA

BIOCOMPUTING GROUP

Capitolato di Appalto

RING

Residue Interaction Network Generator

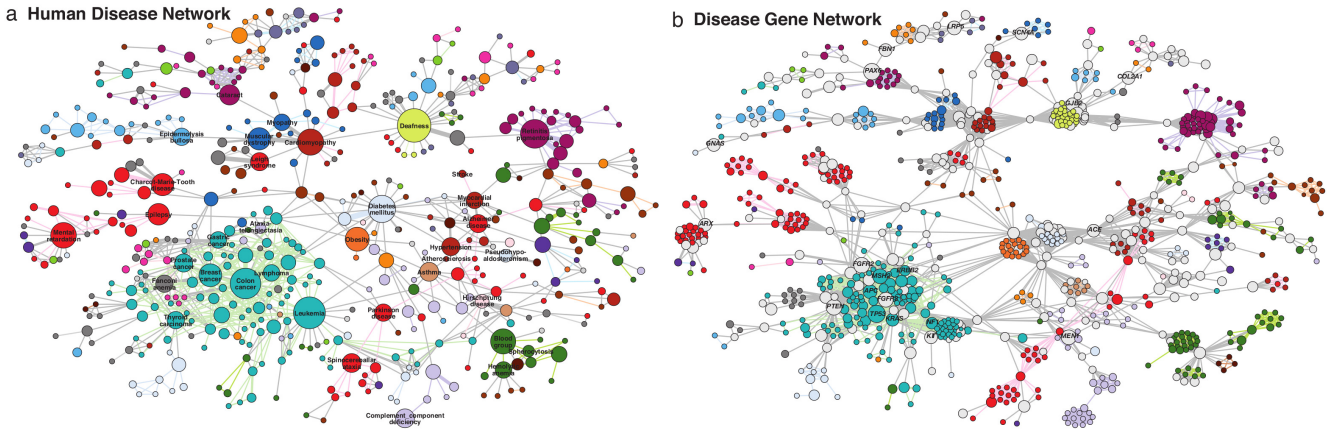
BioComputing lab,
Dipartimento di Scienze Biomedicine,
Università di Padova
E: biocomp@bio.unipd.it
P: (+39) 049 827 6260
F: (+39) 049 827 6269

Introduzione

Networks e bioinformatica

La bioinformatica nonostante sia una disciplina nata recentemente si trova oggi tra le più importanti scienze biologiche. La biologia è diventata a tutti gli effetti una “big science” producendo quantità di dati enormi e in questa prospettiva le nuove idee nascono solo quando una visione olistica è possibile. La bioinformatica oggi è il punto di partenza per il design di moltissimi esperimenti moderni e costituisce l'unico approccio possibile per lo sviluppo della medicina del futuro.

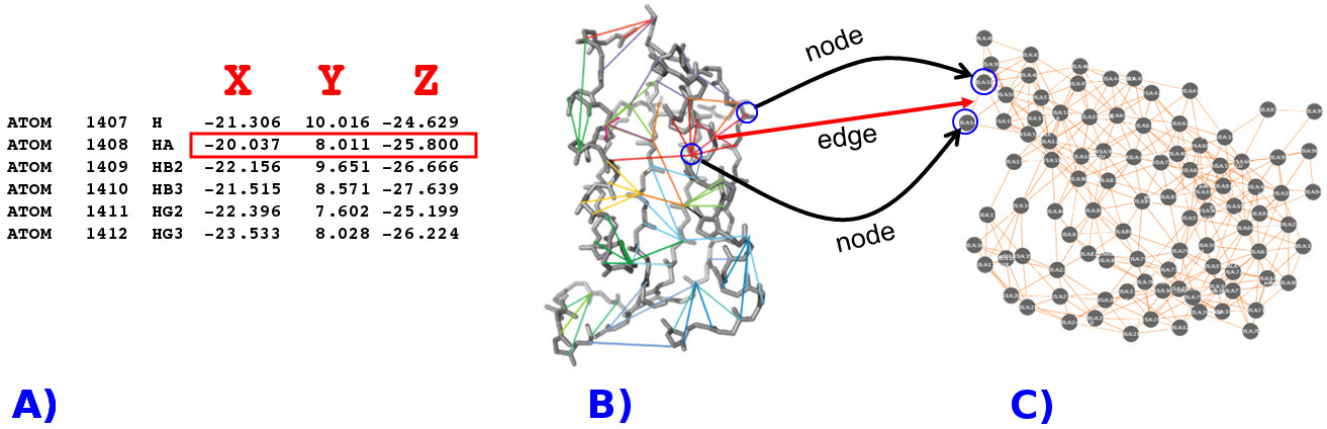
L'utilizzo di reti è diventato uno dei più potenti strumenti per gestire e visualizzare proprietà emergenti di sistemi complessi. Al giorno d'oggi la rappresentazione dei dati tramite reti ha catturato l'attenzione di tutta la comunità scientifica nonostante siano necessari strumenti di analisi più potenti. La gestione di dataset enormi e l'interpretazione dei grafi generati sono ancora problemi aperti. Ne deriva che, a causa della massa di dati prodotti dagli esperimenti della biologia moderna, una nuova classe di tool molto efficienti necessiti di essere sviluppata.



Questo è un esempio di come grazie all'utilizzo di reti è stato possibile mappare le interazioni geniche con le malattie. Da questa analisi è stato quindi possibile identificare tutti i geni essenziali (gli hub nella rete), ovvero tutti quei geni che sono determinanti e critici per la maggior parte delle malattie. Tratto da: Goh et al. The human disease network. PNAS, 2007.

Il progetto RING

Con questo bando offriamo l'opportunità di affrontare i problemi classici della biologia strutturale grazie all'utilizzo della teoria dei grafi.
Con il termine "struttura" in biologia molecolare si intende la forma che una data molecola assume quando si trova in un determinato solvente (p.es. acqua) e che si ottiene grazie al fatto che le sue componenti, gli atomi, interagiscono spinti dalle forze atomiche fino a raggiungere un equilibrio termodinamico. Le più interessanti macromolecole nella biologia sono sicuramente le proteine. Queste sono espressione dell'informazione codificata nei geni presenti nel DNA e svolgono tutte le funzioni molecolari all'interno della cellula. La struttura di una proteina è nota quando si conoscono le coordinate tridimensionali degli atomi che la costituiscono (nella figura sottostante è riportato un esempio di come le coordinate spaziali sono solitamente rappresentate). Questo tipo di informazione viene ricavata sperimentalmente e richiede anni di lavoro, e la possibilità di determinare a priori la struttura di una proteina conoscendone solo la sua composizione atomica è un problema che la bioinformatica sta tentando di risolvere già da qualche decennio con risultati non ancora soddisfacenti.
Data una struttura nota, per comprenderne il meccanismo molecolare di funzionamento è necessario conoscere le interazioni che intercorrono tra le sue componenti, gli atomi. Da questa idea è stato sviluppato recentemente un nuovo sistema di rappresentare le strutture proteiche chiamato RIN (Residue Interaction Network). Tecnicamente questo è un grafo in cui i nodi rappresentano gli atomi, e gli archi corrispondono alle interazioni elettrostatiche che intercorrono tra essi. Le interazioni sono determinate sulla base della distanza euclidea calcolata partendo dalle coordinate atomiche e vengono pesate e tipizzate considerando le diverse proprietà chimico-fisiche degli atomi.



(A) Una porzione di un file che contiene le coordinate spaziali degli atomi di una proteina in formato PDB (Protein Data Bank); (B) rappresentazione degli atomi nello spazio tridimensionale; (C) rappresentazione della proteina come RIN dove sono messe in evidenza le interazioni tra atomi.

RING (Residue Interaction Network Generator) è un software sviluppato dal nostro gruppo pochi anni fa ed è stato progettato per calcolare RIN partendo dalle coordinate atomiche delle strutture proteiche. La versione attuale è disponibile online come web service (http://protein.bio.unipd.it/ring/) ed è implementata come una pipeline che include codice fatto in casa più una serie di chiamate a programmi di terze parti per il processing preliminare delle coordinate atomiche e il calcolo di proprietà molecolari.

Destinazione

La descrizione delle funzionalità del software prodotto e le sue applicazioni verranno utilizzate per la ricerca svolta dal gruppo di BioComputing UP. Se possibile, saranno pubblicate in una rivista internazionale di bioinformatica citando il contributo degli sviluppatori.
Il software verrà distribuito con licenza Creative Commons *Attribution 3.0 Unported* (vedi: http://creativecommons.org/licenses/by/3.0/deed.en_US)

Requisiti

Il programma deve generare un grafo (RIN) dove i nodi corrispondono agli atomi o gruppi di atomi della proteina (a seconda del livello di astrazione) e gli archi rappresentano le interazioni tra questi atomi. Gli archi (le interazioni) vengono generati quando una coppia di atomi si trova ad una distanza spaziale al di sotto di una soglia specifica. La distanza tra atomi viene calcolata partendo dalle loro coordinate x,y,z che vengono fornite in input.
La gestione di file generati da dinamiche molecolari è invece opzionale. Tutti i requisiti opzionali sono denotati con la parola "opzionale" nel testo che segue.

Design pattern

Il programma deve essere progettato come un software "stand-alone" che può essere utilizzato da linea di comando, attraverso un web service e come plugin (app). Inoltre deve essere un sistema client-server basato su three-tier architecture che separi interfaccia, logica e dati come da definizione. La logica che consiste nell'insieme di funzioni che operano il calcolo deve essere la stessa per tutte e 3 le versioni del software (command line, web service, plugin). Questo è necessario per garantire la futura manutenzione del codice.

Input

1. Coordinate atomiche.

L'utente può fornire direttamente uno o più file (da gestire come fossero una molecola sola) con le coordinate atomiche nel formato PDB[1] o PDBML/XML (vedi "Riferimenti utili").

2. Codice identificativo.

Se l'utente non dispone di un file di input può semplicemente fornire un codice identificativo PDB[1] e il programma provvederà a scaricare automaticamente le coordinate online.

3. Coordinate atomiche multiple (opzionale).

L'utente può fornire più file di coordinate indipendenti l'uno dall'altro (ad esempio quelli generati da

una "dinamica molecolare"[2]). In questo caso il programma dovrà gestire ogni file separatamente distribuendo il calcolo in parallelo su richiesta dell'utente.

4. Parametri.
   a) Distanza soglia per la determinazione degli archi per ogni tipo di interazione.
   b) Livello di astrazione. I nodi possono rappresentare tutti gli atomi o gruppi di atomi (identificati dai residui della proteina).
   c) Catene da processare. Ogni catena corrisponde ad un gene, tuttavia spesso una singola proteina è formata da più catene che si assemblano e ne determinano la funzionalità molecolare. L'utente deve poter scegliere se generare il grafo su tutte le catene o un sottoinsieme di esse.
   d) Tipo di output (vedi paragrafo output).
   e) Numero di thread per il calcolo parallelo (opzionale).

Output

1. Il grafo.

Come lista e/o matrice delle adiacenze.

2. Gli attributi dei nodi e degli archi.

Per ogni nodo vengono forniti attributi chimico-fisici e per ogni arco il tipo di interazione e la forza di questa interazione (il "peso" dell'arco).

3. Coordinate atomiche corrette.

Le coordinate strutturali, anche se determinate sperimentalmente, spesso contengono errori o sono incomplete. Possono mancare atomi la cui presenza e posizione può essere dedotta, e ci possono essere atomi posizionati non correttamente che creano collisioni ovvero che si sovrappongono. Il programma si occupa di risolvere questi problemi e l'utente deve poter ottenere il file delle coordinate corretto.

Il formato di output per i punti 1 e 2 deve essere:

1. Testuale, per una visualizzazione diretta.

2. CSV, per l'analisi con fogli di calcolo.

3. JSON, per facilitare la trasmissione via internet.

4. Matriciale, per agevolare l'utilizzo di tool statistici.

Quando viene fornito un file di traiettorie derivato da una dinamica molecolare l'output è solo di tipo matriciale.

Il formato per quanto riguarda il punto 3 deve essere lo stesso adottato dal PDB[1].

Inoltre deve essere prodotto un file unico contenente sia il grafo che gli attributi di archi e nodi con un formato supportato dai principali tool di visualizzazione di reti come Cytoscape, Cytoscape-Web e Gephi.

*PDB denota la "Protein Data Bank" http://www.rcsb.org/. La repository principale di strutture di macromolecole biologiche.

Funzioni core

1. Processing delle coordinate atomiche.
   a) Calcolare le coordinate degli atomi di idrogeno mancanti.
   b) Correggere le coordinate che creano conflitti (orientamento sbagliato, collisioni di atomi).

2. Calcolo RIN.
   a) Calcolo della lista delle adiacenze in $O(n^2)$.
   b) Determinazione del tipo di interazione.
   c) Calcolo degli attributi degli archi.
   d) Calcolo attributi dei nodi.
   e) Gestione del calcolo in parallelo. Ad esempio per gestire l'input derivante dalle "dinamiche molecolari"[2] (opzionale).

Il punto 2.d (Calcolo attributi dei nodi) quando richiede l'utilizzo di programmi esterni può servirsi dell'utilizzo di tool installati in locale (se possibile) oppure appoggiarsi ad appositi web service di terze parti.

Interfaccia (wrapper)

1. Command-line.

Il software deve rispettare gli standard POSIX e deve prevedere di poter richiamare un "help" che descriva la sintassi del programma.

2. Web service.

Deve essere implementare un sistema REST e/o SOAP per la comunicazione con il server.

3. Pagina Web.

Va sviluppata una pagina web che permetta di utilizzare lo strumento e di visualizzare le reti calcolate su un browser moderno (iEplorer 8+, Firefox, Chrome).

4. Plugin.

Deve essere implementato un wrapper che permetta di utilizzare le funzionalità del programma all'interno di Cytoscape e/o Gephi.

Documentazione

Devono essere prodotti un documento con la descrizione della logica e gli algoritmi usati, un manuale utente e il codice sorgente deve essere attentamente commentato utilizzando il sistema Doxygen e le relative convenzioni.

Linguaggio di implementazione

Suggeriamo di utilizzare il linguaggio C++ e/o Java sia per la disponibilità di librerie avanzate (in C++) che per facilitare l'integrazione di RING in Cytoscape e Gephi (implementati in Java).

Variazione dei requisiti

Non sono ammesse variazioni se non a evidente miglioramento di quanto richiesto dal committente. Non è esclusa la comunicazione, da parte del committente, di variazioni ai requisiti sia precedentemente alla consegna delle offerte che durante la realizzazione del sistema.

Riferimenti utili e undertaking del proponente

Un'ampia documentazione per quanto riguarda l'utilizzo e la descrizione delle funzionalità della prima versione di RING è disponibile all'indirizzo http://protein.bio.unipd.it/ring/ nelle sezioni "Quick Help" e "Method". Allo stesso indirizzo è possibile inoltre reperire le referenze per quanto riguarda la letteratura scientifica (sezione "References").
La documentazione riguardante il formato "PDB" e "PDBML/XML" delle coordinate atomiche delle strutture proteiche è disponibile in http://www.wwpdb.org/docs.html#format.
Fondamentale inoltre per quanto riguarda l'annotazione degli attributi dei nodi (vedi "Requisiti" -> "Funzioni core") sono le sezioni "Methods" e "XML schema" della banca dati SIFTS (http://www.ebi.ac.uk/pdbe/docs/sifts/index.html).

Per quanto riguarda il supporto sistematico e qualsiasi tipo di chiarimento tecnico/scientifico riguardante lo sviluppo delle nuove funzionalità, per accedere al vecchio codice, casi di prova e materiale integrativo riguardante l'approfondimento della bioinformatica strutturale si deve fare riferimento al *Dr. Damiano Piovesan*. Sarà disponibile per interazioni orali (su appuntamento) e per corrispondenza via mail ai seguenti recapiti:

- Indirizzo: *BioComputing UP, Dipartimento di Scienze Biomediche, Edificio Vallisneri, aula 19 – 4° piano sud, Viale G. Colombo 3, PD*
- E-mail: *damianopiovesan@gmail.com*
- Tel: *(+39) 049 8276269*

Glossario

[1]PDB denota la "Protein Data Bank" http://www.rcsb.org/. La repository principale di strutture di macromolecole biologiche.
[2]Dinamica Molecolare denota una tecnica computazionale attraverso la quale si simula una perturbazione della struttura di una molecola e si calcolano le variazioni delle coordinate spaziali in piccoli intervalli di tempo.