e-GOV
Mensile
Data 11-2004
Pagina 61/63
Foglio 1 / 3

Mara Tabusso

Bioinformatica: applicare l'information technology alla genetica

La si può definire sotto certi aspetti come gestione e all'analisi dei dati di sequenza. Così Alessandro Guffanti, programmatore Ifom, spiega l'importanza della bioinformatica. Un esempio di come l'Ict possa rivestire un valore sociale.

IFOM is a research institute supported by the Italian foundation for cancer research. Under many respects, it represents a model for scientific research: for example, IFOM is supported jointly by public and private institutions and makes extensive use of information technologies. In fact, today's cancer research is mostly based on genetic researches and particularly on DNA sequencing – a method that has been made possible by dramatic improvements of computing abilities in recent years.
To further improve such abilities, the institute decided to upgrade its machines to Itanium technology.
Since IFOM already had a number of HP servers in use, it asked the company to provide its own Itanium products. The new flexible HP Itanium server allowed IFOM to manage an ever increasing amount of data in an easy way, while making them available to nonprofessional users (e.g. students) and using customized open source applications developed by the researchers themselves.

Nato nel 1998 per iniziativa della **Fondazione italiana per la ricerca sul cancro (Firc), l'Istituto Firc di oncologia molecolare (Ifom)** è un centro di ricerca specializzato nello studio dei meccanismi di formazione e di sviluppo dei tumori. **Ifom** sviluppa la propria attività di analisi e di studio sui meccanismi di formazione e di sviluppo tumorale attraverso il lavoro di ricercatori altamente qualificati, che si avvalgono del supporto di infrastrutture It all'avanguardia. Inaugurato nell'aprile del 2003, l'istituto sorge su un'area ex industriale del milanese che occupa **11.200 mq** ed è ripartita su 6 edifici, 6.200 mq di laboratori, 2.200 mq di uffici e 2.800 mq di spazi adibiti a biblioteca, auditorium, aule per seminari, mensa, foresteria, con una capacità di accoglienza per oltre **300 ricercatori**.

Ifom comprende un core tecnologico che propone metodologie sperimentali avanzate quali nanotecnologie, **bioinformatica**, tecnologie di sequenziamento del Dna, organismi modello, patologia molecolare, colture cellulari, tecniche di imaging, immunologia e biologia strutturale. Grazie al facile accesso alle risorse logistiche e a una politica di condivisione del know-how, l'istituto è un vero e proprio **incubatore di conoscenza** e rappresenta un modello altamente funzionale allo sviluppo moderno della ricerca in oncologia molecolare, attraverso l'integrazione tra **finanziamenti pubblici e privati**, la sinergia tra i gruppi di ricerca, l'ottimizzazione delle risorse e l'attenzione all'applicabilità dei risultati.

Volendo mettere a punto una strategia di prevenzione basata su metodi diagnostici e terapie farmacologiche mirate, il primo passo da compiere è lo studio e l'identificazione delle **anomalie genetiche**, vale a dire la rilevazione degli errori presenti nei geni delle cellule tumorali. Durante gli ultimi dieci anni, la quantità d'informazioni disponibili nel settore della genetica molecolare è letteralmente esplosa, grazie allo sviluppo di **tecniche di**

sequenziamento del Dna. Per avere un'idea della massa critica dei dati, basta pensare che se nel 1999 le sequenze di Dna in termini di singole basi (A, C, G o T) disponibili pubblicamente erano poco più di tre miliardi, nel febbraio del 2004 erano arrivate a **circa 38 miliardi**, divisi in **più di 32 milioni** di record (fonte: *Genetic Sequence Data Bank*).

L'importanza dell'elaborazione e del trattamento dei dati

"La bioinformatica si può definire come l'applicazione dell'Information technology alla gestione e all'analisi dei dati di sequenza – sottolinea **Alessandro Guffanti**, programmatore e bioinformatico Ifom. – *Lo scopo è fornire una risposta a determinati problemi biologici attraverso lo studio dell'espressione genica e delle evidenze biologiche, il che frequentemente si traduce in una base di milioni e milioni di dati. Gli elaboratori elettronici impiegati per queste ricerche devono essere molto potenti e veloci, e utilizzare programmi specializzati".*

Per soddisfare la crescente richiesta di potenza elaborativa, il cuore del sistema attuale di calcolo bioinformatico Ifom si basa su un server **Unix Silicon Graphics Origin2200**, equipaggiato con **8 unità Cpu, 8 Gbyte di memoria** e **1 Tbyte** di spazio disco. Su questo sistema sono installate le principali banche dati pubbliche di acidi nucleici e di proteine, ovvero il prodotto della lettura e traduzione del Dna, con sistemi di aggiornamento giornaliero interamente automatici, assieme ai relativi programmi di ricerca e interrogazione. *"Sfruttiamo quest'architettura, unitamente a un insieme di programmi bioinformatici* – prosegue Guffanti – *per supportare i progetti di ricerca Ifom che richiedono procedure automatiche di lettura, filtraggio e analisi di una considerevole quantità di dati di sequenziamento. Qualche esempio? Collaboriamo a progetti di analisi seriale dell'espressione genica o a progetti di espressione genica basati su microarray. Con i risultati sperimentali creiamo banche dati di facile interrogazione, sviluppando strategie di ricerca automatizzate su database di sequenza per la caratterizzazione e il raggruppamento funzionale dei dati "grezzi" derivati da sequenze prodotte all'interno dell'Istituto. Inoltre, per facilitare la collaborazione tra Ifom e gli istituti partner, i programmi bioinformatici sono disponibili attraverso il nostro sito Internet http://bio.ifom-firc.it/".*

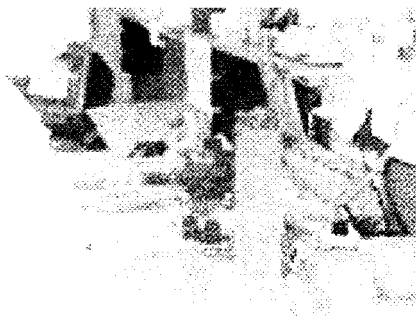
Itanium, chiave del nuovo sviluppo

Grazie alla possibilità di usufruire della **tecnologia Itanium** implementata sul **server Hp Integrity rx2600**, il Dipartimento di bioinformatica ha potuto intraprendere nuovi progetti di ricerca, mantenendo inalterate metodologie e tool di sviluppo grazie alla piena compatibilità della macchina con realtà multipiattaforma e logiche open source, acquistando maggiore potenza su diverse attività computazionali. La divisione di bioinformatica, nel corso del 2003 inizia a vagliare la possibilità di aggiornare i propri strumenti It con macchine più evolute. Lo staff degli operatori Ifom, è curioso di testare la tecnologia Itanium. Tra le piattaforme tecnologiche utilizzate, Ifom ha un nutrito numero di server Hp che supportano una serie di servizi e programmi dedicati alla gestione dell'intera organizzazione. Alla fine dell'estate, l'istituto ottiene in prova una macchina Itanium-based, ovvero il modello **RX2600**. *"Su questa macchina, che abbiamo chiamato 'Leviathan' pensando al possente mostro mitologico* – racconta Guffanti – *abbiamo installato una serie di strumenti bioinformatici complessi. Il nostro interesse principale era applicare alla bioinformatica un server più potente, ricco di funzionalità aggiuntive e, soprattutto, capace di adattarsi a un ambiente multipiattaforma come il nostro".*

"Tutti parlavano di Itanium – precisa **Michael Kahle**, direttore dei sistemi informativi di Ifom – *ma nessuno l'aveva mai visto in azione praticamente. Per la tipologia del lavoro sviluppato, eravamo curiosi di sperimentare il livello di performance possibile grazie a questo nuovo processore. Se era vero quel che si diceva, allora era necessaria una verifica delle sue effettive potenzialità. Per capire la sua validità, infatti, dovevamo poter capire come funzionava 'su strada' e quali problemi ci fossero, per esempio, nel porting da una piattaforma all'altra. Oggi possiamo dire che abbiamo operato una scelta corretta".*

Nuova Hp, un team potente

Le **soluzioni Hp** comprendono infrastrutture It, computer e dispositivi di accesso, servizi globali, sistemi di digitalizzazione delle immagini e stampa per utenti privati, grandi società, piccole e medie aziende, enti pubblici. L'investimento annuo in ricerca e sviluppo, pari a 4 miliardi di dollari, promuove l'invenzione di prodotti, soluzioni e nuove tecnologie per servire meglio i clienti e accedere a nuovi mercati. Hp inventa, progetta e offre soluzioni tecnologiche che forniscono valore aggiunto per il cliente e creano valore sociale per gli utenti.La fusione con Compaq Computer Corporation nel 2002 ha creato un team potente e dinamico di 140 mila persone, con uffici in 178 paesi, operazioni in oltre 40 valute e scambi in più di 10 lingue. Il fatturato complessivo delle società, considerate congiuntamente, già nel 2002 era pari a 72 miliardi di dollari per l'anno fiscale terminato il 31 ottobre. Carly Fiorina, Chairman e Ceo di Hp, guida la società la cui sede centrale si trova a Palo Alto, in California.



L'importanza di un'architettura potente, flessibile e scalabile

Il lavoro di ricerca svolto all'interno dell'istituto è molto articolato: la maggior parte dei tool applicativi utilizzati sono sviluppati dagli stessi scienziati Ifom mediante una customizzazione di **soluzioni open source** o sviluppate ex novo. Il profilo delle competenze della maggior parte degli operatori, è costituito da un robusto know-how informatico soprattutto in relazione all'attività di analisi e programmazione. In pratica, si tratta di un **ambiente altamente verticale** dove l'It, in termini computazionali, rappresenta uno dei bracci armati della ricerca. La richiesta ad Hp di un server dotato di processore Itanium rispondeva alla necessità di supportare una nuova linea di ricerca dell'istituto, applicando una procedura di analisi bioinformatica su un'architettura più potente rispetto a quella già disponibile. Lo scopo era applicare ed estendere una procedura di **data mining** preesistente ma web-based, al fine di supportare la ricerca di "geni antisenso" nell'intero genoma umano.

"Scoperte recenti hanno evidenziato come sia possibile analizzare la sequenza genomica anche in un'altra prospettiva direzionale – precisa Guffanti – *da cui è possibile rilevare una nuova serie d'informazioni molto importanti. Alla luce delle nuove relazioni descritte dalle sequenze 'antisenso', stiamo producendo una serie di dati di predizione innovativi rispetto a queste strutture geniche, lavorando su una dimensione molto elevata di dati e con potenziali contributi interessanti per la ricerca sulla genetica molecolare del cancro".*

Un **secondo progetto** su cui sono state testate le prestazioni dell'Rx2600 riguarda la gestione di una consistente mole di dati, generati nel laboratorio di **array acDNA**, che consentono di monitorare contemporaneamente l'espressione genica di migliaia di geni, in diverse condizioni sperimentali e patologie tumorali. *"Grazie alle prestazioni di Rx2600* – aggiunge **James F. Reid**, ricercatore bioinformatico Int/Ifom – *oggi possiamo analizzare e gestire una campionatura di oltre 10-20 mila geni effettuandone senza problemi lo stoccaggio, l'archiviazione e il recupero. In una logica di integrazione e condivisione delle informazioni, ho sviluppato un tool web-based in cui è possibile inserire dei dati grezzi e identificare, tutte le informazioni necessarie al fine di capitalizzare la ricerca senza ridondanze e in un'ottica di condivisione delle informazioni. Grazie alle funzionalità di report del software bioinformatico che abbiamo installato sull'Rx2600, in ogni momento è possibile ricostruire tutti gli step relativi all'analisi che hanno portato a un determinato risultato, in una chiave di knowledge management evoluto".*

Tecnologie di ultima generazione

È evidente che per indicizzare e archiviare la **grandissima mole di dati** generata dall'attività dal dipartimento di bioinformatica, all'istituto occorreva un server capace di garantire la possibilità di effettuare upgrade progressivi. Il tutto gestito da un'interfaccia che ne permettesse una modalità di utilizzo semplice, dal momento che nel novero degli utilizzatori Ifom sono compresi anche molti studenti. Attraverso il sito, infatti, parte delle informazioni vengono offerte sotto forma di servizio e rese accessibili all'esterno. Nell'ottica della massima condivisione delle risorse, dunque, Leviathan viene ubicato nel centro stella e, attraverso una **Lan aziendale Ethernet based**, collegato ai vari dipartimenti che necessitavano di una capacità di elaborazione più evoluta.

Per supportare l'attività computazionale della nuova ricerca genomica, il gruppo di bioinformatici ha replicato sulla macchina Rx2600 una serie di software specialistici, tra cui **Spidey, AntiHunter, RepeatMasker e BlastN**. *"Una delle incognite riguardava BlastN, un programma che risultava difficile riuscire a far funzionare su diverse architetture con i parametri necessari alla nostra ricerca...* – precisa Guffanti –. *Con questo tipo di impostazioni le criticità sono molte ma la macchina Hp è riuscita a risolvere l'impasse"..* Reid aggiunge che *"...il server Hp Integrity Rx2600 ha risposto molto bene su più livelli di compatibilità. Grazie a Leviathan, abbiamo potuto definire al meglio le nostre esigenze in ambito bioinformatico e oggi siamo finalmente in grado di esprimere una prospettiva di reingegnerizzazione su un'architettura che consideriamo ben definita".*

"È indubbio che quest'esperienza ci ha aiutato moltissimo nell'analisi delle nostre reali necessità – conclude Kahle – *Prima di tutto abbiamo potuto vagliare nel concreto la sua compatibilità anche sul versante Linux, e questo su tutti i moduli applicativi di cui fanno uso i ricercatori e programmatori bioinformatici Ifom, per un arco di tempo molto ampio. In conclusione, riteniamo che l'esperienza sia stata davvero valida al punto che stiamo pensando a una sua estensione applicativa".*

www.ifom-firc.it