



“La bioinformatica si può definire sotto certi aspetti come l’applicazione dell’Information technology alla gestione e all’analisi dei dati di sequenza.”
Alessandro Guffanti, programmatore e bioinformatico Ifom



Centro di eccellenza tecnologica certificato, 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.

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 multiplatforma e logiche open-source, acquistando maggiore potenza su diverse attività computazionali.

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.

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. per capire meglio le dinamiche sottese all'attività sviluppata è necessario conoscere i presupposti fondamentali. La ricerca, infatti,



ha dimostrato che le diverse manifestazioni patologiche che insieme vanno sotto il nome di cancro sono caratterizzate dalla crescita incontrollata e invasiva di gruppi di cellule geneticamente alterate.

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, immediatamente applicate a interi genomi, e al corrispondente progresso nella tecnologia degli elaboratori elettronici. 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 sotto certi aspetti 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 quello di 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 come chiave del nuovo sviluppo

La divisione di bioinformatica, a fronte di un ritmo evolutivo incalzante, per gestire un'attività di elaborazione sempre più complessa e dettagliata, nel corso del 2003 inizia a vagliare la possibilità di aggiornare i propri strumenti It con macchine più evolute. Lo staff degli operatori Ifom, e in particolare il direttore dei sistemi informativi, 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 multi-piattaforma 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, cioè che a livello di server Itanium dovrà rappresentare il futuro sostituendo le altre tecnologie, 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".

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

Per Ifom il problema non era soltanto di compatibilità hardware e software. 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, tra cui operano attualmente circa 200 ricercatori tra ingegneri, matematici e fisici, 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 attualmente 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; successivamente, una selezione di queste predizioni dovrà essere verificata in laboratorio".

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 a cDNA, 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 10-20mila 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, mediante appositi motori di ricerca, 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".

La reingegnerizzazione viaggia su tecnologie di ultima generazione

È evidente che per indicizzare e archiviare la grandissima mole di dati generata dall'attività del dipartimento di bioinformatica, all'Istituto occorre un server flessibile, robusto, potente e capace di garantire la possibilità di effettuare upgrade progressivi. Il tutto gestito da un'interfaccia che ne permettesse una modalità di utilizzo relativamente 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 -. Trattandosi di una ricerca in banca dati dove si va a cercare con una stringa un'altra stringa, su un database contenente milioni di stringhe, 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, infatti, 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, per nulla scontata, 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".

Per maggiori informazioni sulle soluzioni HP www.hp.com/it

© 2004 Hewlett-Packard Development Company, L.P. Le informazioni contenute in questo documento sono soggette a modifiche senza preavviso. Le garanzie per i prodotti ed i servizi HP sono previste espressamente nella garanzia che accompagna tali prodotti o servizi. Nessuna affermazione contenuta nel presente documento può essere ritenuta una garanzia aggiuntiva. HP non è responsabile per errori tecnici o editoriali od omissioni contenuti nel presente documento.

