CONSORZIO INTERUNIVERSITARIO LOMBARDO
PER L'ELABORAZIONE AUTOMATICA
✉ Via R. Sanzio 4 - 20090 SEGRATE MI Tel. 02 26995.1

CAPI 2006
Biocomputing

Consiglio Nazionale delle Ricerche

# 10° Workshop CAPI2006
# Convegno Calcolo ad Alte Prestazioni nelle scienze della vita (Biocomputing)
# Milano 16-17 Ottobre 2006

## **Titoli ed abstracts degli interventi** (*) relatore

*Keywords*: Supercalcolo, Biocomputing.

**Museo Nazionale della Scienza e della Tecnologia**
*Sala Conte Biancamano, nella scenografica cornice del Padiglione Aeronavale (Via Olona 6 bis)*

http://www.cilea.it/capi2006/

# Lunedì 16 Ottobre 2006

**Progettazione razionale di mimici dei carboidrati**

*Anna Bernardi - Università degli Studi di Milano*

L'interazione tra proteine e carboidrati espressi come glicoconiugati alla superficie cellulare gioca un ruolo determinante in una serie di fenomeni fisiologici e patologici, che coinvolgono il riconoscimento e l'adesione cellulare. Gli eventi di riconoscimento proteina-zucchero sono quindi stati identificati come un potenziale bersaglio per lo sviluppo di nuovi farmaci in svariate aree terapeutiche, che vanno dal cancro alle infezione batteriche e virali, alle malattie immunitarie. Questa area di ricerca costituisce ormai una parte importante delle glicoscienze e della glicomica.

In quest'ambito è molto attiva la ricerca di mimici funzionali dei carboidrati: si tratta di identificare e sintetizzare molecole che siano capaci di riprodurre l'attività biologica del templato naturale (generalmente un oligosaccaride complesso) semplificandone però la struttura, e migliorandone alcune caratteristiche, come la stabilità chimica e biologica, per avvicinarle a quelle tipicamente necessarie per un farmaco. In studi di questo tipo la fase di progettazione si avvale inevitabilmente di modelli computazionali per la riproduzione e previsione della struttura tridimensionale delle molecole di interesse e dei loro complessi con le proteine bersaglio.

Utilizzando queste tecniche, supportate dagli opportuni studi sperimentali, il nostro gruppo ha sviluppato alcuni interessanti glicomimetici, capaci di interferire con alcune proteine coinvolte nelle infezioni batteriche (tossina colerica) e virali (recettore dendritico DC-SIGN). I risultati, e in particolare il ruolo giocato dal supercalcolo nello svolgimento dei progetti, saranno illustrati nella presentazione.

## Computer simulations of biomolecular systems

*Giorgio Colombo – CNR-ICRM, Milano*

Computer simulations of proteins, lipids and nucleic acids at equilibrium have become essentially routine. The challenge for the future is to use such approaches to understand recognition and spontaneous self-organization in biomolecular systems (folding, aggregation and assembly of complexes), processes that cannot be directly observed experimentally.

In this presentation, examples illustrating the extent to which simulations can be used to understand these phenomena in biomolecular systems will be presented along with examples of where the methodology may still be developed further.

The study cases will cover the problems of peptide-receptor recognition and the use of the information obtained for the design of new non-peptidic ligands; the study of the folding mechanism of small proteins and finally the study of the initial stages of peptide self-aggregation.

**BLAST sul Demonstrator: un nuovo portale per lanciare BLAST sfruttando le potenzialità della GRID**

*Giuseppe La Rocca - INFN - Sezione di Catania*

Recent progresses in Grid computing have made it possible to devise platforms able to ensure appropriate support to research activities performed in the area of life sciences. In fact, biological information is increasing at an impressive rate, and it is typical accessed through web interfaces (portals). One of the most relevant problems in bioinformatics is how to manage the increasing amount of data, empirically produced by researchers involved in life sciences. The analyses of this huge data (e.g., when searching the human genome or when carrying out simulations of molecular dynamics for the study of new drugs) need to exploit new and huge resources. The potential of the GRID technology provides a possible solution to this request.

In this paper we propose a Grid portal, based on Enginframe, able to provide the basic tools in order to allow users to run Multi BLAST in a Grid environment. The paper is organized as follow: in sections 2 and 3, we will briefly introduce the tools we used to build all the needed services to run Multi BLAST to the Grid. Then, in section 4, we will describe the implementation of the application exploiting the proposed architecture.

**Applicazione a problemi di somministrazione di farmaci delle tecniche di supercalcolo legate alla risoluzione di Equazioni di Hamilton-Jacobi in dimensione alta**

*(*)Piero Lanucara - ISMAC-CNR, Genova*
*Enrico Bersani- CASPUR*
*Alberto Maria Bersani, Marco Rorro - Università La Sapienza Roma*
*Loretta Mastroeni, Università Roma Tre Roma*

La trasduzione del segnale intracellulare è il mezzo attraverso cui la cellula vivente comunica con l'ambiente esterno, risponde e si adatta a esso. La "Systems Biology" si propone di formulare dei modelli matematici che descrivono le complesse reti di trasduzione molecolare, fondate sulla cinetica delle reazioni enzimatiche. Questi modelli sono basati sull'ipotesi di stato quasi stazionario della cinetica di Michaelis-Menten, che a livello di reti di reazioni manifesta forti limitazioni.
Dal punto di vista farmacologico è interessante simulare la risposta della cellula alle perturbazioni delle suddette cinetiche tramite metodi numerici in grado di trattare le condizioni di somministrazione di un farmaco, al fine di ottimizzare la risposta a livello cellulare.
Nel presente lavoro viene utilizzata la tecnica di controllo ottimo, che garantisce solide basi teoriche e una modellistica matematica consolidata, nonché algoritmi efficienti anche per le dimensioni spaziali in cui il problema vive.
Per la risoluzione dell'equazione alle derivate parziali di tipo Hamilton-Jacobi (non lineare) caratterizzante la funzione valore associata alla dinamica del problema vengono utilizzati strumenti di calcolo (hardware e software) ad alte prestazioni.
I risultati relativi alla simulazione di singole reazioni di Michaelis-Menten complete vengono mostrati sia in termini di efficienza computazionale che di efficacia farmacologica.

**Computing infrastructure for drug design within pharmceutical industry**

*Alfonso Pozzan - GlaxoSmithKline S.p.A. Verona*

Nowadays, the success in designing new drugs (drug design) is tightly bound to the computational capabilities available within a research organization. Following the well known Moore's law, during the last few decades, computing power available to computational chemists, cheminformaticians and bioinformaticians has increased dramatically. On the other hand, the technological and scientific improvements in the pharmaceutical field have always created the need for more computing power, network bandwidth and disk space. The current status of a typical computing infrastructure available within a major pharmaceutical industry is reviewed in relationship with the objectives addressed by drug design methods. This includes reviewing and presenting various hardware architectures and how these are integrated with current drug design software and tasks.

**Il supercalcolo paralleo al servizio della farmaceutica: DELOS, una piattaforma bioinformatica per il drug – discovery**

*Piercarlo Fantucci - Università degli Studi di Milano Bicocca*

I tempi sempre più lunghi e gli investimenti sempre più ingenti richiesti dalla ricerca in campo farmaceutico possono essere notevolmente contratti grazie all'uso delle nuove piattaforme bioinformatiche. Fra queste, la piattaforma DELOS (DiscovEry and Lead Optimization System) che integra un software molto avanzato con un cluster di processori in parallelo ha dimostrato di aver raggiunto una buona maturità ed efficienza.
Le applicazioni di supercalcolo nel drug-discovery riguardano la determinazione (e/o raffinamento) della struttura tridimensionale di proteine e l'identificazione/caratterizzazione dei siti recettoriali mediante tecniche di ottimizzazione strutturale quali Meccanica e Dinamica molecolare (MM/MD) in solvente esplicito ed eventualmente in presenza di membrana lipidica. Queste simulazioni dall'elevato numero di parametri e variabili in gioco sono estremamente onerose in tempo macchina e possono essere realizzate in tempi ragionevoli solo con l'utilizzo di macchine parallele.

Il secondo punto di applicazione del supercalcolo alla drug discovery riguarda la generazione di librerie virtuali molecolari di grandi dimensioni ($10^5$-$10^6$ composti) a partire da scaffolds predefiniti e da una biblioteca di residui. La generazione delle molecole sfrutta i principi della chimica combinatoriale in termini informatici e computazionali. Le dimensioni molto grandi delle librerie sono richieste dalla necessità di garantire la diversità nello spazio Particolare cura va messa nel garantire la diversità nello spazio molecolare dei composti generati.
La totalità delle molecole generate può essere completamente caratterizzata mediante la valutazione di molte centinaia di descrittori molecolari calcolati con metodi quantomeccanici (QM) semi-empiricio con metodi classici.
Lo screening "in silico" (VHTS) delle librerie di composti verso un target proteico si riconduce alla determinazione per via computazionale di costanti di binding, ed è realizzato con tecniche di docking intermolecolare. Il calcolo esplicito dell'energia di docking in generale rappresenta però il rate determining step del drug design: non è quindi possibile condurre tale calcolo in modo esplicito per l'intera libreria quando il numero dei suoi componenti è $10^5<N_{mol}<10^6$, a meno di non incorrere in serissimi problemi connessi con i tempi di calcolo.
La strategia seguita può procedere attraverso i seguenti steps
- definizione di un sottoinsieme di molecole rappresentative dell'intero spazio molecolare
- calcolo esplicito dell'energia di docking per il sottoinsieme
- sviluppo di un modello statistico (reti neurali e/o classificatore bayesiano) per la stima dell'energia di docking per le molecole della libreria non appartenenti al sottoinsieme.

Nel caso di molecole caratterizzate da grande affinità per il target affini e con caratteristiche chimico-fisiche compatibili con i requisiti della farmacocinetica, si realizza un'ulteriore selezione aggiungendo la flessibilità interna delle molecole stesse, in particolari casi, anche del recettore. Dunque le metodologie di docking si potranno combinare con la dinamica molecolare, pur essendo ben presente che un così formidabile problema computazionale richiede necessariamente potenti mezzi di supercalcolo.
Le ricerche condotte lungo le linee del paradigma sopra esposto, utilizzando la piattaforma DELOS, verranno illustrate per alcuni esempi di target molecolari.

**High performance computing and in silico drug design: quo vadis?**
*Stefano Moro - Università degli Studi di Padova*
Ancora oggi il processo richiesto per lo sviluppo di un nuovo farmaco è lungo in termini temporali (10-12 anni) e costoso in termini di quantità di risorse da investire (600-800 milioni di euro). I recenti progressi compiuti dalla chimica farmaceutica, dalla chimica combinatoria alla chemoinformatica, consentono oggi di poter disporre di diversi milioni di composti chimici da poter immettere nel complesso processo di individuazione di nuovi farmaci. In assenza di una strategia di ottimizzazione delle varie fasi dell'intera filiera che porta alla individuazione e al successivo sviluppo di nuovo farmaco, difficilmente questo rapporto "tempo/risorse" potrà essere migliorato sensibilmente. In questo senso, lo sviluppo e l'applicazione di nuove metodologie computazionali (come per esempio le tecniche di high throughput virtual screening o di farmacologia in silico) in tandem con le nuove frontiere del calcolo ad alte prestazioni (quali per esempio il Grid computing o il calcolo parallelo) possono rappresentare un obbiettivo strategico per ottimizzare al meglio i tempi e le risorse economiche richieste nelle diverse fasi dello sviluppo di un nuovo farmaco. La perfetta sintonia e integrazione di competenze multidisciplinari rappresenta quindi lo strumento essenziale per il raggiungimento del suddetto obbiettivo.

**High performance computing per il drug design cardiovascolare**
*Carmelina Ruggiero - Università degli Studi di Genova*
La disponibilità della sequenza di molti genomi, e in particolare di quello umano, ha effetti estremamente rilevanti per quanto riguarda la messa a punto di nuovi farmaci. La ricerca sui meccanismi patologici a livello molecolare costituisce una nuova base di partenza che ha cambiato profondamente l'approccio alla progettazione di farmaci.

Sulla base di conoscenze di genomica è possibile iniziare il processo che porta allo sviluppo di nuovi farmaci partendo dalla sequenza genomica per arrivare all'identificazione dei geni coinvolti in una particolare patologia. Sulla base di queste conoscenze, è possibile affrontare il primo passo critico del processo di sviluppo di un farmaco, cioè l'identificazione e la convalida di un insieme ridotto di molecole bersaglio dei farmaci, quali proteine a carattere enzimatico, recettoriale e di trasduzione del segnale. L'identificazione di queste molecole è la premessa per lo screening e la sintesi di nuove molecole che possono interagire con esse. Dopo aver ottenuto l'insieme di targets molecolari, occorre caratterizzarli mediante analisi strutturale e indentificare un insieme di possibili ligandi che sono in grado di interagire con questi targets.
Le tecniche di sintesi consentono di produrre un gran numero di composti che potrebbero interagire con un target. Ciascuna molecola è caratterizzata da un insieme di proprietà chimico-fisiche in base alle quali è possibile selezionare un numero ristretto di elementi da sottoporre a simulazione di interazione col target (virtual screening). Le risorse informatiche giocano un ruolo rilevante sia nelle selezione dei composti (chemio-informatica) sia per applicazioni di bioinformatica strutturale. E' possibile modellare le relazioni tra le variazioni dei valori di proprietà molecolari e l'attività biologica per un insieme di composti, utilizzano poi queste relazioni come guida per valutare nuovi composti (analisi QSAR). A livello più alto, la fase seguente del processo riguarda la valutazione dell'efficacia del farmaco sviluppato mediante simulazioni dell'assorbimento, diffusione, metabolismo ed escrezione (ADME). Completata questa analisi dei possibili ligandi è possibile sviluppare modello su calcolatore che identificano i percorsi genetici e metabolici nei quali essi sono coinvolti.
Questo approccio è stato adottato nel progetto CARDIOWORKBENCH (finanziato dall'Unione Europea, VI Programma Quadro).
Si prevede di utilizzare la tecnologia GRID per ottimizzare gli aspetti di screening su larga scala, per l'integrazione dei modelli QSAR e ADME, e integrazione delle banche dati contenenti l'informazione necessaria.

*Un momento del Convegno*

# Martedì 17 Ottobre 2006

**Simulazione di processi biologici con i sistemi a membrane**

*(*) Daniela Besozzi - Università degli Studi Milano*
*Paolo Cazzaniga, Enzo Martegani, Giancarlo Mauri - Università degli Studi Milano-Bicocca*

Nell'ambito del Laboratorio di Bioinformatica e Calcolo Naturale del Dipartimento di Informatica dell'Università di Milano-Bicocca è stato recentemente sviluppato il simulatore stocastico di sistemi biologici tau-DPPs, basato sul metodo tau leaping introdotto da Gillespie et al. [1] ed esteso da Cazzaniga et al. [2] a modelli basati sui sistemi a membrane introdotti da George Paun [3].
Tau-DPPs permette di simulare sistemi suddivisi in diversi volumi, tracciando il tempo simulato sia dei singoli compartimenti che dell'intero sistema.
In questo lavoro vengono presentati i risultati della simulazione del pathway Ras/cAMP/PKA del lievito Saccharomyces cerevisiae: ciclo della proteina Ras, attivazione dell'adenilato ciclasi, produzione di AMP ciclico, attivazione di chinasi cAMP-dipendente. I risultati sono stati confrontati con i dati sperimentali [4,5] e danno informazioni sugli elementi regolatori chiave della rete.
Le simulazioni sono state effettuate utilizzando un cluster di computers; la simulazione di sistemi più complessi richiederebbe l'utilizzo di sistemi di calcolo ad alte prestazioni.

**Bio-molecular diagnosis through random subspace ensembles of learning machines**

*(*) Giorgio Valentini, Alberto Bretoni, Raffaella Folgieri - Università degli Studi Milano*

High-throughput bio-technologies (e.g. DNA microarray) generate data characterized by high dimensionality and low cardinality.
The bio-molecular diagnosis of malignancies, based on these biotechnologies, is a difficult learning task, due to the characteristics of these high-dimensional data.
Many supervised machime learning techniques, among them support vector machines (SVMs), have been experimented, using also feature selection methods to reduce the dimensionality of the data.
In this paper we investigate an alternative approach based on random subspace ensemble methods.
The high dimensionality of the data is reduced by randomly sampling subsets of features (gene expression levels), and accuracy is improved by aggregating the resulting base classifiers.
Considering the high computational cost of the proposed technique, we used the High-Performance CILEA. Avogadro cluster of Xeon double processor workstations to perform all our computational experiments.

**Metodi di apprendimento automatico per l'annotazione strutturale e funzionale di genomi**

*Ivan Rossi - Università degli Studi di Bologna*

Il gruppo di Biocomputing si occupa di fornire soluzioni a problemi specifici di Bioinformatica rilevanti nell'analisi di sequenze proteiche. L'analisi del genoma su larga scala consente infatti una annotazione strutturale e funzionale delle sequenze geniche ottenute da varie specie di procarioti ed eucarioti. In particolare, i nostri metodi sono basati sulla estrazione di informazione da banche dati di sequenze e strutture note per una generalizzazione che consente l'implementazione di predittori, ossia strumenti in grado di rispondere a richieste specifiche in relazione al problema posto. L'integrazione di vari predittori in una suite di programmi consente l'analisi di un genoma procariotico di dimensioni medie in una settimana con due CPU. Al termine viene fornita una annotazione su basi strutturali di ogni gene del genoma in questione. Dato che la procedura può essere parallelizzata i tempi di calcolo potrebbero essere ridotti notevolmente utilizzando risorse opportune. Si ricorda che attualmente i genomi di procarioti e di eucarioti che possono essere oggetto di indagine ammontano a oltre 400.

**Computational dissection of large gene-expression dynamics (by the correlation method)reveals pathways co-regulation changes by conditional transcription factor activation**

(*) *Gastone Castellani – Università degli Studi di Bologna e INFN*
*Claudio Franceschi, Centro L. Galvani Università degli Studi*
*Luciano Milanesi, CNR Milano*
*Mirko Francescani, Daniel Remondini, Università degli Studi di Bologna*

The dynamics of a gene expression time series network is studied. It is showed that the large scale correlation of gene expressions exhibits global dynamic properties that emerge after a cell state perturbation. The main features of the reconstructed network appear to be more robust when compared to those obtained with a network created from a Linear Markov Model (LMM). In particular, the network properties strongly depend on the exact time sequence relationships between genes and are destroyed by random temporal data shuffling.

It is discussed in detail the problem of finding targets of the c-myc proto-oncogene, which encodes a transcriptional regulator whose inappropriate expression has been correlated with a wide array of malignancies.

Two data sets were obtained. The first data set (N data set) contains the gene expression data of the c-myc -/- MycER cell line treated with vehicle (ethanol) only. The second data set (T data set) contains the gene expression data collected after the addition of tamoxifen. Samples were harvested at five time points after the addition of tamoxifen to the culture medium: 1, 2, 4, 8, and 16 h . The entire experiment was repeated on three separate occasions, providing three independent measurements for each gene and each time point.

The integration of this wealth of information into mechanistic models that explain the biological functions of c-Myc. has been greatly complicated not only by the large number of targets, but also by the weak transcriptional effects exerted by c-Myc. Thus, the biologically relevant downstream effectors remain to be comprehensively delineated.

It is showed that the correlation-based model can establish a clear relationship between network structure and the cascade of c-myc activated genes. In comparison with LMM, it is demonstrated that the correlation method is more sensitive to the temporal structure and leads to biologically relevant gene identification that is not found by either Markov modeling or by significance analysis based on ANOVA alone.

A further step toward a better comprehension of such network is the finding that a similar transition is conserved at different scales and is indicative of co-regulation changes. To reduce the dimensionality of the problem and introduce a-priori biological knowledge, the correlation method has been extended by mapping the array onto gene pathways and ontologies. Multiscale correlation shows that the changes in correlation profiles is not only founded at several scales (whole array, gene family and pathways) but it also informative of significant changes induced by c-myc activation and allows pathways synthesis into single funtional forms. This methodology allowed to observe co-regulation between and within several pathways with precise biological functions Finally we show how this method can be applyed to high-resolution time-series microarrays in two situation of great biological interest: caloric restriction and human ageing.

**Modelling Immune System: a challenge for HPC and Grid Computing**

*Santo Motta - Università degli Studi di Catania*

Immune system is a complex adaptive learning system. It has evolved to maintain the healthy state of the organism and to protect against infectious diseases and cancers. Malfunction of immune system may result in autoimmune diseases or allergies. The immune system operates at multiple levels: molecule, cell, tissue, organ, organism, and population. The immune system has immense diversity due to its combinatorial nature. The number of different molecular products of the immune system (such as antibodies or T-cell receptors) is several orders of magnitude larger than the total number of all other proteins produced by the organism. The complexity and diversity of the immune system limit our ability to study the immune system using only experimental approaches. In some cases, e.g. screening of vaccine targets across viral variants, the number of necessary experiments for immunological studies is too large. A significant proportion of experiments that are routinely performed using animal models cannot be done in humans because of ethical considerations.

Computational models are ideal for bridging these gaps. They have been successfully applied, for example, to screening of the immune response targets, analysis of antibody structures, and optimisation of cancer immunotherapies.

Modelling the immune system is a formidable challenge because of the large number of components to be considered. For example, there are some $10^{11}$ T-cells and a similar number of B-cells in a human body. There are more than $10^{15}$ possible combinations of Human Leukocyte Antigen molecules which present targets of immune responses to the immune system. In past, models of the immune system were small, idealised, and simplified systems which could not capture the true complexity of the immune system. The revolution in information technology has ensured that available computational resources can deal with the systems of large complexity. The emergence of the Grid computing enables modelling human immune system at a natural scale.

ImmunoGrid is a STREP project funded by European Commission which started o February 1, 2006 (FP6-2004-IST-4, No 028069). The primary aim of ImmunoGrid is the development and implementation of a Grid-based simulator of human immune system for support of clinically relevant applications, such as vaccine development and optimisation of immunotherapies.

Immunogrid partners are: CINECA, Bologna, Italy (Project coordinator); University of Queensland, Australia (Scientific coordinator); CNR, Rome, Italy; CNRS, Montpellier, France; Technical University of Denmark, Lyngby, Denmark; Birkbeck College, University of London, UK; University of Bologna, Italy; University of Catania, Italy.

Main aims of the ImmunoGrid are: (i) Standardisation of immunological concepts and related bioinformatics tools and resources; (ii) Combining data, tools and resources to develop a simulator and create models of the human immune system; (iii) Develop pre-clinical applications of the simulator testing; (iv) Disseminate results and tools to researchers and clinicians.

An example: Finding optimal schedules for cancer immonoprevenction vaccines. Vaccines for cancer treatment is an hot topic in Life Science. An immunoprevenction vaccine for mammary carcinoma has been developed and tested in vivo on HER-2/neu mice at the University of Bologna. In vivo experiments have shown that the vaccine is effective in preventing solid tumor formation if it is administrated with a chronic schedule. Combining biological knowledge, system modeling and HPC computing we suggested a much lighter alternative schedule to test in vivo.

**High performance computing in studies of proteins. A comparative moleculardynamics investigation of psycrophilic and mesophylic elastases**

*Luca De Gioia - Università degli Studi Milano-Bicocca*

Modern bioinformatics and cheminformatics approaches are key approaches to structure-function relationships in biomolecules, and are now widely used in pharmaceutical and biotechnological companies and research institute. Among the computational tools, the molecular dynamics (MD) occupies a central position as it allows to investigate the dynamic properties of a protein and to extract precise information on biologically relevant motions, such as the flexibility of specific regions, the adaptability of the active site or the local structural rearrangement upon ligand binding. However, MD simulations are among the most CPU intensive calculations used to study molecular systems. As a consequence, the use of high-performance computing is crucial to carry out this kind of studies.

As an example of high-performance computing applied to the MD investigation of proteins we present a study in which the molecular basis of cold adaptation inside the specific enzymatic class of pancreatic elastases have been explored by MD simulations.

A comparative MD investigation reveals that specific loop regions are characterized by enhanced flexibility in the cold-adapted enzymes, leading to the conclusions that these differences play a crucial role for catalysis at low temperature. This observation fully supports the hypothesis suggesting that flexibility is the main adaptive character of psychrophilic enzymes. Remarkably, the corresponding mesophilic enzymes are characterized by enhanced flexibility, when compared to the cold-adapted ones, in scattered regions distant from the functional sites. Therefore, our results are also in agreement with a scenario in which local rigidity in regions far from the functional sites can be a positive factor in the adaptation of psychrophilic enzymes.

**Il ruolo di un centro HPC nel supporto delle Scienze della vita**

*Elda Rossi* - CINECA

Cineca è un importante Centro di Calcolo italiano che ha lunga esperienza nel supporto delle attività di ricerca in ambito scientifico.

Negli ultimi anni, oltre a discipline quali la Chimica, la Fisica e l'Astronomia, storicamente legate al calcolo ad altre prestazioni, il Cineca ha aggiunto il supporto verso la Bioinformatica e le Scienze della vita in generale.

Per queste discipline tuttavia, le tecnologie HPC devono intendersi in un senso più ampio. Infatti i ricercatori impegnati in questo ambito non hanno attualmente bisogno di mera potenza di calcolo, o almeno non solo. Sono invece particolarmente interessati a tutta una serie di tecnologie che permettano l'uso più efficace e collaborativo degli strumenti e soprattutto la condivisione dei dati, che sono di centrale importanza per questa disciplina.

Le tecnologie più indicate afferiscono all'ambito delle Grid Computazionali e più precisamente alla accezione dei "servizi", sia Web Services che Grid Services. Per l'utilizzo di questi servizi gli strumenti di orchestrazione e sottomissione, chiamati "Workflow", sono altrettanto importanti poiché costituiscono l'interfaccia utente all'ambiente computazionale e ne permettono una efficace fruizione.

Le attività del Cineca in questo ambito sono soprattutto legate ad alcuni progetti attualmente in corso che verranno brevemente discussi:

- LIBI: un progetto nazionale per la progettazione e realizzazione di un ambiente distribuito per la bioinformatica
- EMBRACE: un progetto europeo per la realizzazione di un ambiente distribuito per la bioinformatica
- ImmunoGrid: un progetto europeo per la realizzazione di un simulatore del sistema immunitario umano.

**Applicazione GRID per il confronto esaustivo delle regioni genomiche conservate tra uomo e topo**

*(*)Flavio Mignone 'Università degli Studi di Milano*
*Graziano Pesole - Università degli Studi di Bari*
*Giacinto Donvito, Giorgio Maggi - INFN, Università di Bari*
*Giorgio Grillo, Vito Flavio Licciulli, Sabino Liuni, Istituto Tecnologie Biomediche, CNR, Bari*

Il sequenziamento del genoma di numerosi organismi rappresenta importante punto di partenza per il successivo processo di annotazione funzionale, che comprende la corretta identificazione dei geni e delle regioni che ne regolano l'espressione. Tale processo è indispensabile per comprendere i meccanismi molecolari che guidano lo sviluppo e consentono la vita di un organismo.

Un approccio che ha già mostrato la sua efficacia è lo studio dei genomi mediante analisi comparata e, in particolare, mediante lo studio delle regioni genomiche evolutivamente conservate.

E' infatti noto che le regioni del genoma più rilevanti dal punto di vista funzionale sono sottoposte a maggiori pressioni selettive che comportano un minore tasso di mutazione e quindi una maggiore conservazione nel corso dell'evoluzione.

Un algoritmo sviluppato dal nostro gruppo, implementato nel software denominato CSTminer, è in grado di identificare efficacemente regioni conservate e di discriminare tra regioni codificanti (che fanno parte dei geni) e regioni non-codificanti (che possono svolgere attività strutturale o regolatoria) a partire dal confronto di sequenze omologhe.

Mediante l'utilizzo di tale software ci siamo proposti di condurre il confronto esaustivo dei genomi di uomo e topo per identificare tutte le regioni conservate (da noi chiamate Conserved Sequence Tag - CST) e per creare una collezione di tutti i CST codificanti e non codificanti identificati tra tali organismi.

Considerate le dimensioni dei genomi di uomo e topo - 3 e 2.6 miliardi di basi rispettivamente - e che mediante CSTminer è possibile confrontare sequenze di lunghezza non superiore a 10 Kbp, è evidente che il tempo di elaborazione necessario per compiere tale analisi su una comune workstation o un server di una certa potenza non è "umanamente" accettabile.

Per effettuare i circa 800 milioni di confronti necessari a condurre tale analisi è stato quindi adottato un approccio che utilizza la tecnologia GRID, in quanto questa permette l'accesso e l'elaborazione di processi su un gran numero di elaboratori. L'infrastruttura GRID e le risorse di calcolo utilizzate sono quelle di INFN-GRID, grid di produzione dell'INFN, la componente italiana dell'infrastruttura europea di Grid realizzata dal progetto EGEE, accessibili mediante la Virtual Organization "bio".

Un ulteriore problema affrontato - oltre alla riduzione del costo computazionale dell'operazione - è stata la gestione di un tale numero di confronti. A questo scopo è stato implementato un sistema di gestione dei processi (confronti) utilizzando un database a supporto. In questo modo è stato possibile gestire efficacemente i confronti sottomettendo alle CPU libere i confronti ancora da eseguire, monitorando lo stato di quelli già eseguiti e sottomettendo nuovamente quelli eventualmente falliti.

Questa implementazione permette inoltre di dare la priorità ad alcuni confronti (ad esempio effettuando i confronti relativi a un particolare cromosoma) potendo quindi iniziare ad analizzare i risultati senza dover attendere il completamento dell'intera procedura.

Mediante questo approccio - tuttora in corso - è stato possibile eseguire più della metà dei confronti in circa un mese.

Ringraziamenti. Questo progetto è finanziato dal MIUR progetto FIRB-LIBI, "Laboratorio Internazionale di Bioinformatica".

**Applicazioni del calcolo avanzato nell'ambito delle biotecnologie presso il CEINGE**
*Giovanni Paolella - Università degli Studi di Napoli "Federico II"*
L'attività di ricerca svolta presso il CEINGE, Napoli, e incentrata sulle problematiche relative alla identificazione di nuovi elementi funzionali attraverso un approccio genomico comparativo. Il lavoro in quest'area di ricerca svolto in questi anni ha portato alla generazione di due database che contengono un numero elevato di sequenze conservate (CST) identificate nell'ambito del genoma umano e delle loro controparti identificate nel genoma di altri vertebrati, dal topo al pesce. Occasionalmente questi elementi si possono trovare anche in organismi filogeneticamente più distanti, come gli insetti e gli invertebrati.
CST sono state identificate nelle regioni genomiche che includono e circondano geni coinvolti in malattie trasmesse geneticamente (DG-CST) e geni che codificano per tutte le proteine che hanno come dominio funzionale un'attività tipo tirosina chinasi (KinWeb). I due database sono mantenuti in costante aggiornamento e disponibili "online" per la comunità scientifica.
Un goal importante dell'analisi genomica è l'identificazione di elementi che corrispondono a regioni non codificanti dell'RNA. Per questo scopo, 50 genomi batterici sono stati usati come sistema modello, e analizzati per cercare famiglie di sequenze ripetute caratterizzate dalla presenza di strutture di RNA tipo stem-loop. Quest'approccio ha portato all'identificazione di tutte le famiglie già conosciute e a numerose nuove, che sono state raggruppate in un database annotato e che presto sarà reso disponibile "online".
Le analisi sopra descritte richiedono elevate risorse di calcolo, e sono state rese possibili dalla disponibilità di un sistema di calcolo parallelo, costituito da un cluster di 56 nodi biprocessore, attualmente disponibile presso il centro.

**e-bioscienze@CASPUR: sviluppo e consolidamento di una soluzione dedicata alla biologia**
*Tiziana Castrignanò – CASPUR, Roma*
Il CASPUR da oltre cinque anni ha intrapreso una iniziativa per offrire agli utenti dell'area bio(chimico)-fisica un ambiente dedicato, basato su tecnologie e risorse umane adatte a rispondere alle sempre maggiori richieste computazionali in questo settore scientifico. Il progetto BioGrid, proposto dal gruppo di chimica e biologia computazionale del CASPUR, ha seguito l'evoluzione tecnologica delle architetture di microprocessori (alpha, ia64, Opteron) per il calcolo e l'elaborazione dati in parallelo per offrire le migliori prestazioni alle applicazioni utente nei settori della dinamica molecolare, della bioinformatica, del datawarehousing in ambito biologico e dei DNA microarray. Grazie alla formazione di personale specializzato, BioGrid raccoglie oggi decine di collaborazioni a progetto con la controparte scientifica finanziate a livello nazionale ed europeo con la possibilità di estendere il raggio di interesse verso nuove soluzioni computazionali. Nell'intervento verranno presentate le esperienze in corso e i risultati conseguiti negli ultimi anni nel settore della bioinformatica, con una particolare attenzione rivolta alle nuove applicazioni basate su tecnologia a griglia e ad alcuni nuovi progetti in fase di attuazione nell'ambito dello sviluppo di applicazioni di interesse biomedico su grid orientate ai servizi.

**CILEA e LITBIO: una sinergia di successo**
*Claudio Arlandini - CILEA, Segrate Milano*
Il Progetto LITBIO (Laboratorio Interdisciplinare di Tecnologie BIOinformatiche) è una infrastruttura, finanziata dal Ministero dell'Istruzione, dell'Università e della Ricerca con un grant FIRB 2003 per il periodo 2005 - 2010, capace di fornire a progetti di ricerca internazionali nuove strategie di analisi di dati biomedici e biotecnologici. Tra gli scopi del progetto vi è quello di stabilire una collaborazione tra pubblico e privato nel campo così come di stimolare la crescita di nuove imprese nel settore della Bioinformatica.
CILEA è partner fondamentale del progetto, incaricato di sviluppare e gestire le infrastrutture informatiche avanzate del Laboratorio.