RETE NAZIONALE DI BIOINFORMATICA ONCOLOGICA (RNBIO)

Paolo Romano (a), Marco Crescenzi (b)
(a) Struttura Complessa Bioinformatica e Proteomica Strutturale, Istituto Nazionale per la Ricerca sul Cancro, Genova
(b) Dipartimento di Biologia Cellulare e Neuroscienze, Istituto Superiore di Sanità, Roma

Base di partenza e razionale

L'esigenza di allestire una rete di bioinformatica per gli Istituti che fanno parte di Alleanza contro il Cancro (ACC) nasce dalla constatazione che la ricerca biomedica dipenderà sempre più dall'analisi delle informazioni disponibili e, quindi, dalla consapevolezza che la bioinformatica diventerà nei prossimi anni il più importante strumento di supporto all'analisi disponibile per i ricercatori. Già ora, la genomica e la proteomica dipendono fortemente dall'analisi automatica delle informazioni per svolgere le ormai classiche elaborazioni di analisi di sequenza, predizione di domini genomici attivi, di struttura, analisi funzionale dell'espressione genica, ecc. In prospettiva, altri ambiti di ricerca, quali l'analisi della variabilità genetica e delle mutazioni e l'analisi del metaboloma, produrranno grandi quantità di dati che potranno essere analizzati esclusivamente in-silico. A titolo d'esempio, si considerino alcuni numeri: le banche dati di sequenze nucleotidiche hanno incrementato la propria dimensione del 40% in media negli ultimi tre anni, una delle principali banche dati di esperimenti di *microarray*, *ArrayExpress*, ha duplicato la propria dimensione in ciascuno degli ultimi due anni, la lista dei siti SRS pubblici comprende un elenco di più di 1.300 distinti database, il supplemento annuale di *Nucleic Acids Research* dedicato alle banche dati di biologia molecolare ha elencato nel 2006 più di 680 database.

Le ovvie problematiche di analisi dati in-silico che derivano da questa ingente mole di informazioni sono ulteriormente complicate dalla distribuzione dei dati sulla rete Internet e dalla eterogeneità dei sistemi informativi. A questa eterogeneità corrispondono diversi software di gestione dati, diversi formati, diverse sintassi e, a volte, diverse semantiche. In questa situazione, anche la gestione dei dati e l'integrazione delle informazioni derivate da diverse sorgenti informative, compiti attualmente svolti dai ricercatori, diventano esse stesse motivazioni sufficienti per un supporto bioinformatico infrastrutturale.

A fianco di queste attività, legate prevalentemente alla ricerca, si sta affermando anche un settore di interesse traslazionale e clinico, la bioinformatica clinica. Appare, infatti, chiaro come sia sempre più necessario integrare le informazioni cliniche dei pazienti oncologici con informazioni genomiche per orientare la pratica diagnostica e terapeutica alla medicina personalizzata. In tale contesto, la pianificazione di studi e l'analisi statistica di dati integrati di tipo clinico e "omico", per la valutazione del contributo diagnostico e prognostico di tecnologie molecolari avanzate, si giova della cooperazione fra ricercatori informatici e statistici biomedici.

Gli IRCCS oncologici non hanno sinora sviluppato competenze, risorse ed esperienze bioinformatiche adeguate a questo contesto, salvo limitati casi. Al contrario, molti Istituti oncologici europei, quali il *Deutsches Krebsforschungszentrum* (DKFZ) di Heidelberg e il *Centro Nacional de Investigaciones Oncológicas* (CNIO) di Madrid, nonché il *National Cancer Institute* (NCI) negli Stati Uniti, hanno da tempo investito cospicue risorse, attivato importanti gruppi di lavoro, e iniziano a ottenere i primi risultati significativi.

Gli Istituti di Alleanza contro il Cancro devono quindi elevare le competenze in questo settore strategico a un livello adeguato alle esigenze dei prossimi anni. La peculiarità di ACC, una federazione di Istituti autonomi e paritetici, nessuno dei quali avrebbe la "massa critica" necessaria, fa sì che una rete di coordinamento e cooperazione sia la struttura più idonea a consentire un efficace confronto tra bioinformatici, biologi e medici, un effettivo trasferimento di competenze tra Istituti, la valorizzazione delle competenze e dei risultati dell'attività bioinformatica svolta e la progettualità necessaria per risolvere efficacemente i problemi che si presenteranno nei prossimi anni.

La relativa novità, per gli Istituti ACC, della tematica impone lo svolgimento di uno studio di fattibilità che, nel corso del primo anno di progetto, consenta di definire precisamente le aree scientifiche di ricerca e cliniche di interesse per la rete, avendo cura di coinvolgere il maggior numero possibile di IRCCS oncologici e altri istituti di ricerca interessati.

Obiettivo principale e obiettivi secondari del progetto

L'obiettivo principale della Rete Nazionale di Bioinformatica Oncologica (RNBIO) è la creazione di un efficace coordinamento delle attività bioinformatiche degli Istituti partecipanti ad ACC al fine di integrare ed elevare le attuali competenze e poter quindi ottimizzare e innovare le attività di ricerca e cliniche in oncologia basate sull'analisi in-silico e sull'automazione delle procedure e dei processi, nonché di partecipare a progetti di livello internazionale e, più in generale, competere ai massimi livelli della ricerca in questo settore.

Gli obiettivi secondari del progetto fanno riferimento ad aspetti relativi ad attività di supporto alla ricerca e alla clinica, alla formazione del personale, alla collaborazione interistituzionale e all'identificazione e definizione di nuovi progetti di ricerca di base e traslazionale.

Si ritiene in particolare di poter identificare i seguenti obiettivi concreti:

- promozione dell'uso di strumenti bioinformatici e dello sviluppo degli stessi, tramite le tecnologie informatiche e telematiche più innovative con l'obiettivo di migliorare l'efficienza e la qualità dell'analisi in-silico;
- coordinamento delle attività di ricerca e sviluppo su tematiche specifiche di ricerca e cliniche che possano portare all'ideazione di nuovi strumenti bioinformatici, la cui realizzazione può essere portata a termine in progetti finalizzati, finanziati su bandi nazionali e internazionali distinti;
- avvio e sviluppo di collaborazioni tra la rete di bioinformatica di ACC e Istituti oncologici europei e internazionali d'eccellenza, sulla base sia di progetti comunitari che bilaterali, nonché con altre reti e infrastrutture di ricerca, anch'esse da concretizzare con finanziamenti distinti;
- valorizzazione delle competenze e degli strumenti/servizi sviluppati e mantenuti dagli IRCCS in supporto all'oncologia clinica e sperimentale, anche nell'ottica di favorirne lo sviluppo secondo modalità informatiche di buon livello e di migliorarne le prestazioni;
- avvio e sviluppo di collaborazioni con gestori di servizi di *High Performance Computing* e infrastrutture di rete avanzate (Grid), nazionali e internazionali, per favorire l'utilizzo di software di elevata complessità e di grandi esigenze computazionali.

In particolare, il progetto:

- si propone di fornire strumenti e infrastrutture bioinformatiche e telematiche che facilitano il lavoro personale e collaborativo dei membri di ACC tramite l'implementazione di un opportuno sito di riferimento;

- prevede la partecipazione della maggioranza dei membri di ACC in quanto non si pone come una rete riservata ai bioinformatici, ma aperta a tutti i ricercatori e i clinici, ponendosi come un luogo di incontro e confronto tra le diverse professionalità tramite il quale sia possibile identificare e affrontare le esigenze e gli interessi di tutti; si intende allargare al massimo la partecipazione, soprattutto all'attività formativa, sfruttando le collaborazioni esistenti dei partner con altri enti/ricercatori.
- favorisce la realizzazione e l'ampliamento di reti regionali e interregionali che possono essere propedeutiche a uno sviluppo in ambito europeo in quanto i partner si rendono disponibili a sostenere e promuovere le attività della rete nei loro rispettivi ambiti regionali;
- ha numerosi agganci con progettualità europee per la partecipazione a progetti ERA-NET (*European Research Area NET*: registro tumori) ed ESFRI (*European Strategy Forum on Research Infrastructures*: biobanche, biologia strutturale, bioinformatica), nonché ai bandi del VII Programma Quadro HEALTH e IST (Challenges 1.2 'Service and Software Architectures, Infrastructures and Engineering' and 5.3 'Towards Sustainable and Personalised Healthcare - Virtual Physiological Human').

Stato generale di sviluppo del progetto e conseguimento dei risultati

Durante il primo anno di progetto, l'attività della RNBIO si è sviluppata secondo le linee previste, con la partecipazione effettiva e continua dei partner. Complessivamente, le attività svolte sono in leggero ritardo rispetto alle previsioni, come dettagliato in seguito, il ritardo essendo principalmente legato all'incertezza iniziale sulla data reale di inizio del progetto e alla successiva difficoltà a reclutare rapidamente il personale a contratto previsto per le attività di progetto. Il progetto si è ora avviato pienamente e si ritiene che i ritardi possano essere recuperati.

Nei primi dodici mesi, tutti i filoni di attività sono stati avviati. Il coordinamento si è concretizzato attraverso teleconferenze tramite software skype, posta elettronica, incontri generali e limitati a un numero ridotto di partner. In particolare, sono state attivate tre *mailing list*: la lista members@rnbio.it è un forum aperto, non moderato, dedicato alla comunicazione tra i partner, la lista news@rnbio.it è dedicata alla diffusione di annunci dalla rete, e la lista newsletter@rnbio.it è destinata agli annunci relativi alla Newsletter.

Il sito web di progetto è stato implementato all'indirizzo http://www.rnbio.it/, utilizzando il software Plone, un *Contents Management System* con utili caratteristiche per la gestione collaborativa di siti web. Il sito comprende un'area pubblica e una riservata ai partner. Nell'area pubblica sono presenti sezioni su news e eventi scientifici, i partner, documenti prodotti dalla rete, software e formazione. Nella sezione riservata sono invece disponibili sotto-sezioni dedicate alla formazione (progettazione dei corsi), ai gruppi di lavoro, ai meeting di progetto e alle teleconferenze. Tutti i partner sono registrati con un proprio account e abilitati all'inserimento e pubblicazione di news, eventi scientifici e documenti di varia natura.

Per quanto riguarda la formazione e l'aggiornamento, sono stati definiti due filoni principali relativi rispettivamente alla formazione sulla programmazione e sullo sviluppo di strumenti bioinformatici e al miglior utilizzo di software di particolare interesse e utilità. Il primo filone mira anche a sottolineare l'importanza della *good programming practice*, del *software reuse* e dell'interoperabilità dei tool sviluppati, mentre il secondo si pone l'obiettivo di definire metodi comuni di analisi, promuovere l'uso dei migliori software e, in definitiva, migliorare l'efficacia

dell'analisi. Il primo corso ha riguardato l'introduzione alla programmazione con il linguaggio R (*Introduction to R* – I2R) e si è svolto a Casalecchio di Reno, presso il CINECA, che collabora strettamente con l'unità operativa IOR, dal 5 al 6 giugno 2008. Il corso, che ha compreso anche una parte pratica, è stato destinato in primis ai partner della rete, ma allargato ai dipendenti di tutti gli IRCCS, con precedenza per quelli di ACC. La partecipazione è stata elevata, tanto che non è stato possibile accettare tutte le richieste pervenute, e il gradimento notevole. Attualmente sono in discussione alcune ipotesi per corsi da tenersi tra la fine del 2008 e i primi mesi del 2009. La discussione avviene all'interno dell'attività dei gruppi di lavoro.

Per quanto riguarda i gruppi di lavoro, si è proceduto alla loro definizione tramite un iter concordato che ha privilegiato la massima partecipazione dei partner. Inizialmente, si è chiesto ai partner di avanzare proposte. Queste sono state vagliate da tutti i partecipanti e si è quindi avviata una discussione generale su ciascuna proposta. Ciascuna delle proposte accettate ha quindi definito un programma di massima per la durata del progetto. La gestione dei gruppi avverrà sulla base di meeting periodici, e via apposite mailing list e/o teleconferenze. Il sito web è destinato a ospitare tutta la relativa attività. I gruppi di lavoro approvati sono relativi alle seguenti tematiche:

i) automazione dei processi d'analisi dei dati online,
ii) oncoproteomica,
iii) oncogenomica,
iv) metodi statistici,
v) analisi dati di *deep sequencing*.

Non è per ora stato attivato nessun gruppo in collegamento con il progetto ESFRI INSTRUCT, ma la discussione sulle modalità di collaborazione con esso è tuttora in corso.

Una massa critica e attività sinergiche sono importanti per la ricerca e lo sviluppo bioinformatici. Per questo motivo, sono state incentivate nuove collaborazioni con altre reti oncologiche e bioinformatiche e con gestori di sistemi di supercalcolo. In particolare, si è partecipato alla proposta di un'infrastruttura italiana per la bioinformatica con la Società Italiana di Bioinformatica (*Bioinformatics ITalian Society*, BITS) e con il progetto interdipartimentale "Bioinformatica" del CNR. Si sono anche stabiliti contatti per future collaborazioni con i responsabili del progetto ESFRI ELIXIR e della *Informatics Initiative del National Cancer Research Institute* inglese.

La rete è stata presentata in vari workshop e convegni, quali quello della Società Italiana di Cancerologia del 2007 (SIC2007), e della Società Italiana di Bioinformatica, congiunto alla *European Conference of Computational Biology* (ECCB/BITS 2008).

Inoltre, la rete ha organizzato una sessione di bioinformatica oncologica nell'ambito del workshop NETTAB 2008 dedicato a "Bioinformatics Methods for biomedical complex system applications", Varenna (LC), 19-21 maggio 2008. Alla sessione sono stati anche presentati diversi contributi dei partner della rete.

Come ulteriore strumento di promozione, sarà pubblicata una *newsletter* in formato elettronico. Un primo numero, di prova, denominato numero 0, sarà presentato nei prossimi giorni. L'obiettivo della Newsletter è quello di informare su attività e risultati ottenuti, contribuendo così anche allo sviluppo di ulteriori collaborazioni. La *newsletter* conterrà, tra l'altro, una rassegna dei risultati, l'annuncio di corsi ed eventi scientifici, abstract di pubblicazioni recenti dei partner e *research news* (sia interne che esterne alla rete), *training note* su software, metodi e algoritmi specifici.

A queste attività, prettamente di rete e di coordinamento, vanno aggiunte quelle svolte singolarmente dai diversi partner e opportunamente dettagliate nelle relazioni scientifiche individuali.

Articolazione del progetto

L'articolazione del progetto è descritta nella Tabella 1.

Tabella 1. Articolazione della Rete Nazionale Bioinformatica in oncologia (RNBIO)

Proponente *(Coodinatori della rete)*	**Ente di appartenenza dell'UO**	**Responsabile scientifico delle UO**
ISTGE *(Paolo Romano)* ISS *(Marco Crescenzi)*	ISTGE	Paolo Romano
	IEO	Francesca Ciccarelli
	INT	Adriano De Carli
	IRE	Giulia Piaggio
	CRO	Valter Gattei
	ITB	Stefania Tommasi
	IOR	Luca Sangiorgi
	Humanitas	Massimo Locati
	HSR	Giovanni Lavorgna
	Istituto di Scienze dell'Alimentazione (CNR)	Angelo Facchiano
	IDI	Giandomenico Russo
	ISS	Paolo Roazzi

Pubblicazioni conseguite nell'ambito del progetto

Il presente progetto ha prodotto in questo primo anno di attività le seguenti pubblicazioni:

1. Benedetti D, Bomben R, Dal-Bo M, Marconi D, Zucchetto A, Degan M, Forconi F, Del-Poeta G, Gaidano G, Gattei V. Are surrogates of IGHV gene mutational status useful in B-cell chronic lymphocytic leukemia? The example of Septin-10. *Leukemia* 2008;4(3):355-8.
2. Bevilacqua V, Chiarappa P, Mastronardi G, Menolascina F, Paradiso A, Tommasi S. Identification of tumour evolution patterns by means of inductive logic programmino. *Genomics, Proteomics and Bioinformatics* 2008;6(2):91-7.
3. Bevilacqua V, Pannarale P, Mastronardi G, Azzariti A, Tommasi S, Menolascina F, Iorio F, Di Bernardo D, Paradiso A, Colabufo NA, Berardi F, Perrone R, Tagliaferri R. High-throughput analysis of the drug mode of action of PB28, MC18 and MC70, three cyclohexylpiperazine derivative new molecole. In: Huang D-S, *et al.* (Ed.). *Advanced intelligent computing theories and applications with aspects of contemporary intelligent computing techniques: 4th International Conference on Intelligent Computing, ICIC 2008. Shangai (China); September 2008. Proceedings.* Berlin, Heidelberg: Springer-Verlag; 2008. p. 1085-92
4. Costantini S, Colonna G, Facchiano AM. FASMA: a service to format and analyze sequences in multiple alignments. *Genomics Proteomics Bioinformatics* 2007;5(3-4):253-5.
5. Costantini S, Paladino A, Facchiano AM. CALCOM: A software for calculating the center of mass of proteins. *Bioinformation* 2008;2(7):271-2.
6. Facchiano A, Facchiano F. Transglutaminases and their substrates in biology and human diseases: 50 years of growing. *Amino Acids* 2008. [Epub ahead of print]
7. Gattei V, Bulian P, Del Principe MI, Zucchetto A, Maurillo L, Buccisano F, Bomben R, Dal-Bo M, Luciano F, Rossi FM, Degan M, Amadori S, Del PG. Relevance of CD49d protein expression as overall survival and progressive disease prognosticator in chronic lymphocytic leukemia. *Blood* 2008;111(2):865-73.

8. Marabotti A. Modeling the conformation of side chians in proteins: approaches, problems and possible developments. *Current Chemical Biology* 2008;2:200-14.
9. Marabotti A, Spyrakis F, Facchiano A, Cozzini P, Alberti S, Kellogg GE, Mozzarelli A. Energy-based prediction of amino acid-nucleotide base recognition. *J Comput Chem* 2008;29(12):1955-69.
10. Menolascina F, Alves RT, Tommasi S, Chiarappa P, Delgado M, Bevilacqua V, Mastronardi G, Freitas A, Paradiso A. Fuzzy rule induction and artificial immune systems in female breast cancer familiarity profiling. *The International Journal of Hybrid Intelligent Systems* 2008;5(3):161-5.
11. Menolascina F, Bevilacqua V, Zarrilli M, Mastronardi G. Induction of fuzzy rules by means of artificial immune systems in bioinformatics. In: Jin Y, Wang L (Ed.). *Fuzzy systems in bioinformatics, bioengineering and computational biology.* Berlin, Heidelberg: Springer-Verlag; 2009. p. 1-18.
12. Monti L, Cinquetti R, Guffanti A, Cremona M, Lavorgna G, Cazzola M, Vignati F, Cittaro D, Taramelli R, Acquati F. In silico prediction and experimental validation of natural antisense transcripts in two cancer-associated regions of human chromosome 6. *International Journal of Oncology* 2008 (in corso di stampa).
13. Mutarelli M, Cicatiello L, Ferraro L, Grober OM, Ravo M, Facchiano AM, Angelini C, Weisz A. Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. *BMC Bioinformatics* 2008;9(Suppl 2):S12.
14. Orfanelli U, Wenke AK, Doglioni C, Russo V, Bosserhoff AK, Lavorgna G. Identification of novel sense and antisense transcription at the TRPM2 locus in cancer. *Cell Res* 2008;18(11):1128-40.
15. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. *Trends Genet* 2008;24(9):427-30.
16. Romano P, Marra D. SWS: accessing SRS sites contents through Web Services. *BMC Bioinformatics* 2008;9(Suppl 2):S15.
17. Romano P. Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics* 2008 9(1):57-68.
18. Rossi D, Zucchetto A, Rossi FM, Capello D, Cerri M, Deambrogi C, Cresta S, Rasi S, De PL, Lobetti BC, Bulian P, Del PG, Ladetto M, Gattei V, Gaidano G. CD49d expression is an independent risk factor of progressive disease in early stage chronic lymphocytic leukemia. *Haematologica* 2008;93(10):1575-9.