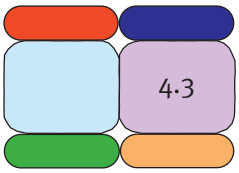
INFORMATICA E BIOLOGIA DEI SISTEMI

Corrado Priami

In questo contributo si esaminano i più recenti sviluppi informatici in campo biologico. Si mostra come l'evoluzione delle scienze biologiche abbia imposto un cambio di paradigma di riferimento nella bioinformatica: ovvero, passando dai progetti di sequenziamento del genoma alla genomica e proteomica funzionale le tecniche informatiche maggiormente adatte si rifanno alla teoria dei linguaggi di programmazione con particolare riferimento alla concorrenza e alla mobilità del codice.



1. INTRODUZIONE

L'uso di tecniche informatiche in domini applicativi legati alla biologia risale ormai agli anni ottanta quando venne coniato il termine **bioinformatica**. Essenzialmente si trattava di memorizzare ed esaminare la grande quantità di dati che veniva prodotta dai biologi. Pertanto, i settori dell'informatica maggiormente interessati furono gli algoritmi, le basi di dati e alcune branche dell'intelligenza artificiale (in particolare, reti neurali) per cercare di estrarre dati significativi e fare predizioni da insiemi spuri di dati prodotti da esperimenti. Le tecniche informatiche maggiormente usate per questi scopi riguardano modelli statici di fenomeni biologici; nessun riferimento viene fatto a possibili evoluzioni funzionali e temporali dei fenomeni. Tuttavia, la definizione di bioinformatica fornita da Hwa Lim che ne coniò il temine è "studio del contenuto informativo e del flusso informativo in sistemi e processi biologici". Nei primi decenni di bioinformatica si è guardato quasi esclusivamente al contenuto informativo (e, quindi, a tecniche statiche) e si è ignorato il flusso informativo.

Per capire le motivazioni che privilegiano la bioinformatica statica a quella dinamica occorre guardare brevemente all'evoluzione che ha avuto la biologia in questi ultimi anni. Il lancio del progetto "genoma umano" ha portato alla scoperta di quantità di dati superiori alle aspettative di chiunque e in breve tempo. A questo punto, conoscendo tutti i geni che compongono il DNA umano si apre una nuova sfida che spesso viene indicata come genoma funzionale (*functional genomics*). Attualmente, si conoscono tutti i "mattoncini" del funzionamento del corpo umano (i geni) e di alcuni si conoscono anche le funzionalità se presi in isolamento, ma ben poco si sa di come i geni e le proteine che questi codificano si comportano in situazioni normali o patologiche. Da qui, la necessità di nuove tecniche per modellare il comportamento di sistemi biologici, e non solo la loro struttura come avveniva fino a pochi anni fa.

La complessità dei sistemi da trattare è tale da non consentire uno studio accurato e

Il termine **"Bioinformatica"** è stato coniato da Hwa Lim alla fine degli anni ottanta per indicare l'applicazione di tecniche informatiche nel dominio applicativo delle scienze della vita. La definizione proposta recita: "lo studio del contenuto informativo e del flusso di informazione nei sistemi e nei processi correlati alla biologia." Tuttavia, questa definizione è unilateralmente legata alla biologia e questo non consente di sviluppare la dignità paritetica che informatica e biologia devono avere in questa area di ricerca al fine di ottenere importanti risultati per entrambe le discipline. A questo proposito una definizione migliore di bioinformatica potrebbe essere "La Bioinformatica è il campo della scienza in cui biologia e informatica si fondono in una unica disciplina per facilitare nuove scoperte biologiche e determinare nuovi paradigmi computazionali sul modello dei sistemi viventi." Questa definizione presa da NCBI al sito www.ncbi.nlm.gov/Education è molto generale, ed evidenzia nettamente la necessità di interazione tra informatici e biologi e, quindi, la necessità di costruire un linguaggio comune alle due discipline per poter interagire e collaborare. Questo può essere fatto solo mediante lo sviluppo di curricula formativi interdisciplinari e mediante l'attivazione di grandi progetti di ricerca. Dal punto di vista dei contenuti, la nuova disciplina deve sicuramente comprendere la definizione di tecniche statistiche e algoritmi necessari a studiare la grande mole di dati che si rende disponibile come risultato degli esperimenti, la definizione di strumenti di scambio e memorizzazione di informazioni accessibili su larga scala, la definizione di metodologie di rappresentazione e simulazione del comportamento di sistemi complessi come le reti geniche o metaboliche o i meccanismi di segnalazione *intra* e *inter*-cellulari. Infine, la bioinformatica dovrebbe assumere il ruolo che ha la matematica per la fisica, e cioè quello di fornire le basi teoriche per i recenti sviluppi biologici nelle aree *omics* (genomics, proteomics, metabolomics ecc.). Questa visione ultima è quella più cara ai biologi teorici che vorrebbero vedere inserita in questa disciplina la loro lunga esperienza sulle teorie evolutive accoppiata con i recenti sviluppi sulla genomica funzionale. Da qui il ruolo primario di modellazione e analisi di sistemi che si vuole affidare alla bioinformatica e che esamineremo più in dettaglio rispetto alle altre possibilità in questo contributo. L'obiettivo ultimo è, quindi, quello di avere teorie predittive del comportamento dei sistemi e anche metodi prescrittivi della loro evoluzione (si veda, a tal proposito il paragrafo ... con le conclusioni del contributo).

La **biologia dei sistemi** è un approccio introdotto recentemente da Leroy Hood e basato sulla teoria dei sistemi per studiare fenomeni biologici. Inizialmente la biologia basava la sua ricerca su un approccio riduzionistico in cui i sistemi venivano scomposti nei loro componenti elementari, si studiavano i singoli componenti per acquisire nuova conoscenza per poi cercare di ricombinarli insieme e avere conoscenza sull'intero sistema. Questo approccio è fallito a causa della enorme complessità dei sistemi biologici e quindi l'incapacità di dominare intellettualmente il processo di ricombinazione. L'idea alla base della biologia dei sistemi è quella di trasformare la biologia in una scienza di scoperta in cui si individua un sistema e se ne studiano le caratteristiche. Anche se il passaggio paradigmatico ai sistemi è interessante e utile di per se, avendo ormai a disposizione l'informazione completa sulle sue potenzialità fornita dai risultati del progetto genoma, l'impatto di un tale approaccio ha portata enorme. Potremmo sintetizzare dicendo che la biologia sta passando dalla produzione della conoscenza all'organizzazione della conoscenza acquisita. Ovviamente dopo aver organizzato il materiale disponibile si dovrà procedere alternando fasi di produzione e fasi di organizzazione come avviene per tutte le scienze sperimentali.
L'obiettivo della biologia dei sistemi si sposa perfettamente con quello della bioinformatica che consideriamo in questo contributo, infatti è prevedere correttamente e modificare il comportamento dei sistemi biologici. Per raggiungere questo obiettivo le strategie della biologia dei sistemi prevedono l'uso di sistematiche perturbazioni genetiche e ambientali dei modelli con un monitoraggio accurato delle risposte globali a questi cambiamenti al livello di geni, proteine, meccanismi di segnalazione e fenotipi. Il monitoraggio deve basarsi non solo su osservazioni qualitative, ma anche quantitative che devono guidare la definizione di strutture dinamiche per la modellazione del comportamento dei sistemi. Su tali modelli si devono poi definire verifiche e controlli iterativi come accade nella definizione di nuovi programmi software per permettere la previsione di nuovi comportamenti.

scientifico se non si fa uso di tecniche strutturate e non ambigue di modellazione e di analisi. Da qui la nascita di una nuova branca di biologia chiamata ***biologia dei sistemi*** (***systems biology***) che ripercorre tappe che altre discipline hanno percorso in passato e fa ricorso alla teoria dei sistemi complessi per rappresentare sistemi biologici. Questa area della biologia è estremamente attiva in questi anni e cerca di lanciare un progetto simile al progetto genoma umano, ma con enfasi sulle funzionalità e interazioni dei componenti basilari del corpo umano.
Tornando all'informatica, negli ultimi anni c'è stata molta attenzione ai sistemi mobili e distribuiti e ciò ha portato alla definizione di semplici calcoli (linguaggi formali dotati di sintassi, definizione della simbologia che utilizzano, e semantica, significato attribuito ai simboli, definite rigorosamente) in grado di rappresentare i possibili comportamenti di tali sistemi in modo non ambiguo. Inoltre, tali calcoli sono dotati di strumenti formali di supporto in grado di effettuare analisi e verifiche di proprietà. Anche in questo campo occorre confrontarsi con l'enorme complessità di sistemi formati da milioni di entità geograficamente disperse in grado di comunicare e cooperare senza avere una completa conoscenza dell'ambiente di esecuzione globale e senza avere completa affidabilità e disponibilità di risorse.
A questo punto, unendo gli sforzi fatti nell'area dei linguaggi di programmazione per modellare i sistemi di calcolo globali (*global computing*) e quelli fatti in biologia per passare a un approccio sistemico nello studio dei fenomeni naturali nasce la controparte dinamica della bioinformatica statica e inizia lo studio

del flusso informativo nei sistemi biologici. In questa breve nota verranno trattati questi aspetti dinamici per tre motivi: sono più nuovi e, dunque, meno conosciuti, consentono di aprire nuove frontiere di ricerca che, invece, sono ormai ben chiare nella bioinformatica statica e possono, infine, consentire avanzamenti nello stato dell'arte sia nella modellazione di sistemi informatici complessi e mobili sia nelle conoscenze biologiche fornendo predizioni di comportamento sulla base di simulazioni e modelli analitici.

2. RAPPRESENTAZIONE DI SISTEMI BIOLOGICI

Nella biologia moderna è ormai chiara la necessità di integrare tutti i dati provenienti dalle discipline dette "omics" (*genomics, proteomics, metabolomics* ecc.) per ottenere dei modelli di sistemi complessi che possano essere studiati mediante strumenti automatici. Attualmente, seguendo un approccio guidato da ipostesi i formalismi usati per rappresentare i sistemi sono di due tipi: quelli grafici informali utilizzati solitamente nei *data base* pubblici (come per esempio, EcoCyc, KEGG, aMAZE, TransPath, INETRACT, SPAD si vedano per esempio [5, 8, 13] e le Figure 1 e 2) e quelli rigorosi prevalentemente basati su equazioni differenziali. Se, invece, si utilizza un approccio basato su scoperte, i modelli vengono inferiti mediante tecniche statistiche come le reti bayesiane.

I formalismi grafici spesso hanno il difetto di non essere formali e ancor più spesso di essere ambigui; inoltre, non esiste una notazione standard (già le Figure 1 e 2 differiscono nella simbologia grafica e nella loro semantica – cioè il significato attribuito ai simboli utilizzati). I formalismi matematici non sono composizionali e non si riescono a inferire in modo automatico dalle rappresentazioni grafiche. Queste limitazioni impediscono di ottenere dei modelli che permettano di studiare in modo soddisfacente i sistemi biologici sia

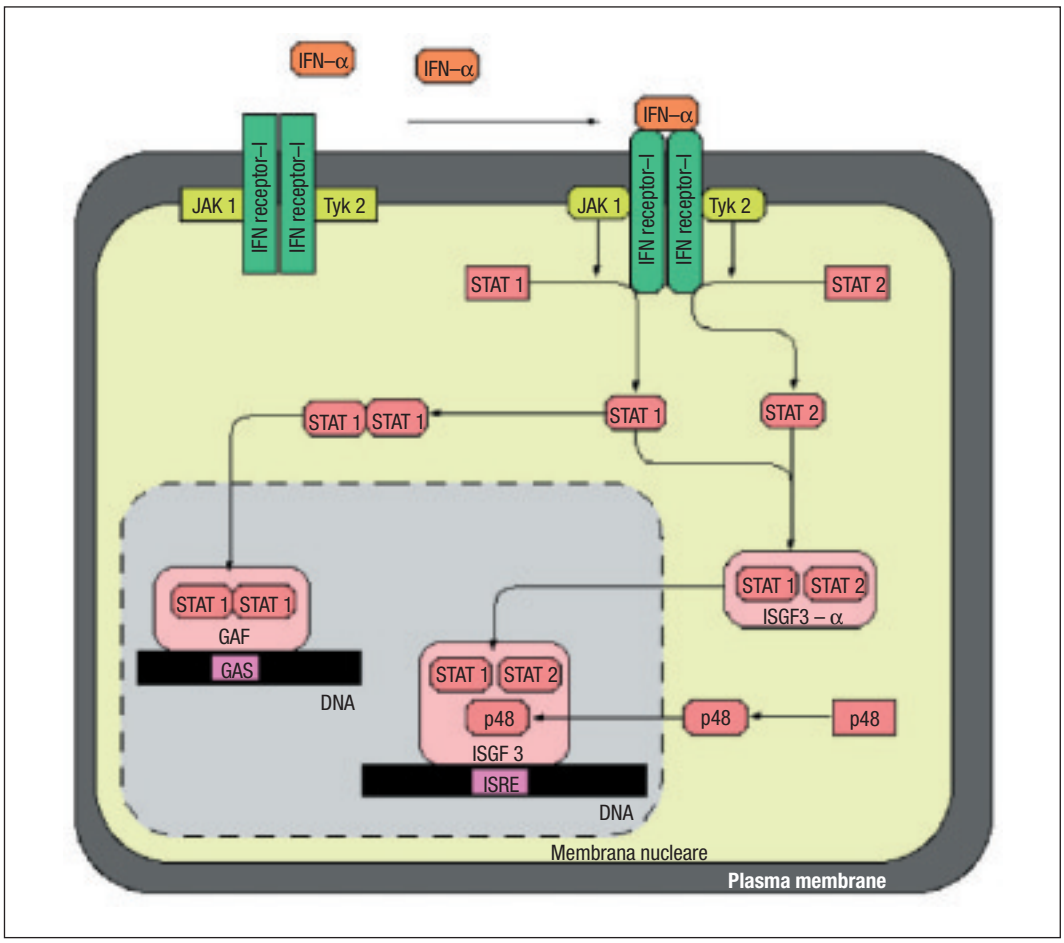


FIGURA 1
Rappresentazione in SPAD

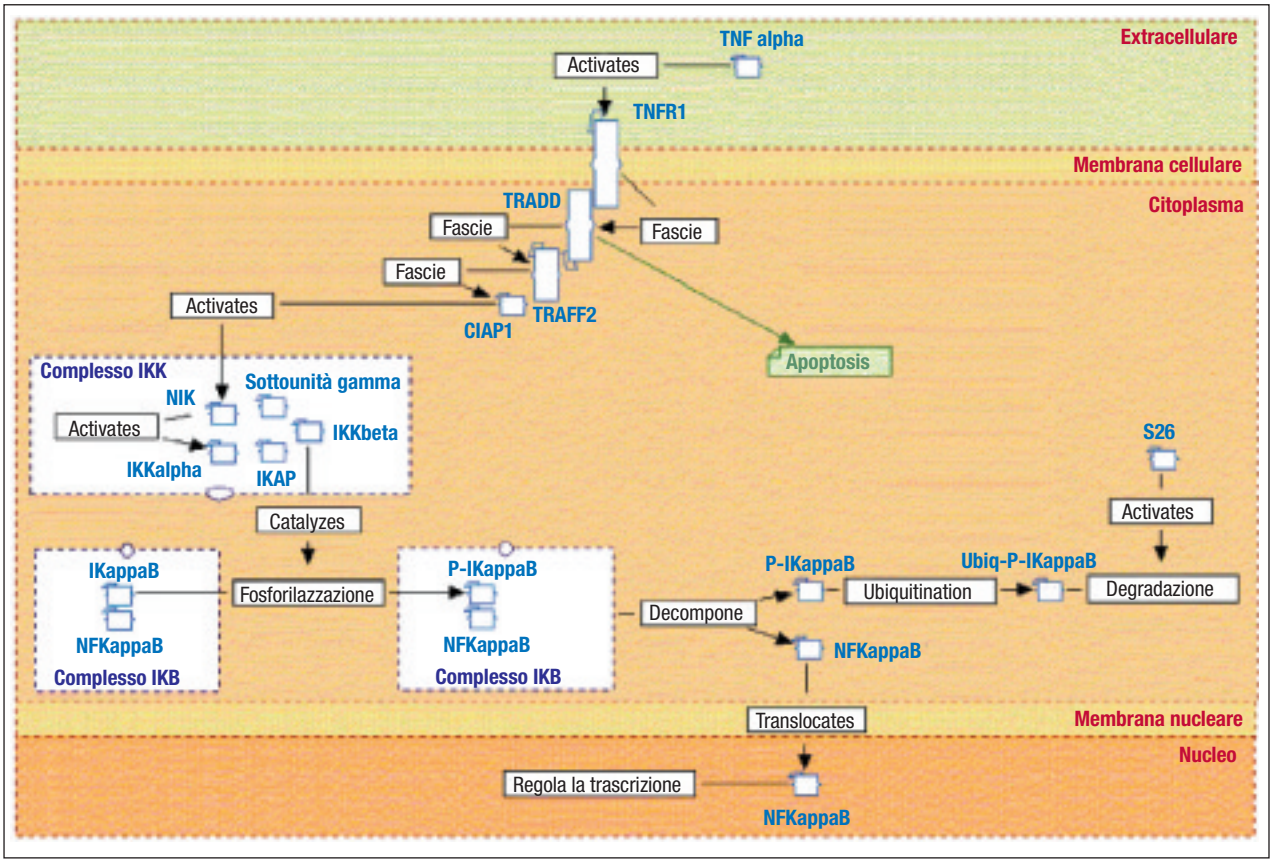


FIGURA 2
Rappresentazione in aMAZE

in termini di struttura che di funzionalità. Recenti sforzi della comunità biologica mostrano che tali limitazioni possono essere superate se si adottano tecniche di modellazione concettuale che sono da molti anni alla base della teoria dei sistemi, dello sviluppo del *software* o della progettazione di grandi basi di dati. Un aspetto cruciale è la possibilità di combinare nello stesso modello sia aspetti statici che dinamici del sistema. Infine, l'importanza della semantica viene evidenziata per determinare dipendenze e relazioni tra le varie parti del modello in modo non ambiguo. La possibilità di definire una struttura (testo di un programma) e un meccanismo (semantica operazionale) per derivare da questa la descrizione di un comportamento dinamico (sistema di transizione) è una peculiarità della definizione formale dei linguaggi di programmazione. Quindi la teoria dei linguaggi, in particolare di quelli concorrenti e mobili, può fornire un notevole supporto alla biologia dei sistemi dove molti eventi concorrenti che modificano l'evoluzione del sistema nel suo complesso sono sempre presenti.

Le problematiche che si devono risolvere per avere rappresentazioni utili e non ambigue dei sistemi sono la definizione di:

- una rappresentazione grafica standard e facilmente comprensibile ai biologi;
- una rappresentazione formale dei sistemi a cui si possano applicare metodi rigorosi di analisi e simulazione;
- un meccanismo di estrazione automatica delle rappresentazioni formali da quelle grafiche nascondendo i dettagli matematici agli utenti;
- un meccanismo di riflessione dei risultati delle analisi nell'interfaccia grafica.

Un approccio recentissimo per rappresentare sistemi biologici in modo grafico, ma semi-formale è basato su **UML** (***Unified Modeling Language***) che fornisce meccanismi naturali per descrivere i componenti di un sistema e le loro interazioni. Inoltre, la modularità dei diagrammi e delle descrizioni UML ha una corrispondenza immediata con la struttura multilivello dei sistemi naturali (per esempio, un organismo è composto di organi che sono a loro volta composti da

L'acronimo **UML** indica un linguaggio grafico e semi-formale usato per la progettazione di sistemi basati su tecnologie ad oggetti. Esso è sicuramente il linguaggio più diffuso anche in ambiente industriale; è supportato da oltre 70 strumenti automatici di progettazione e da oltre 80 libri descrittivi. La sua diffusione è in continuo aumento e sta coprendo un sempre maggior numero di domini applicativi compresi recentemente i sistemi biologici. Una delle maggiori caratteristiche che lo rende estremamente versatile è la sua estendibilità con meccanismi (profili) previsti già nella definizione del linguaggio. L'utilizzo di profili è stato recentemente adottato da alcuni biologi [12] per definire uno standard di modellazione di sistemi biologici. Inoltre, l'esistenza dello standard XML (*eXtensible Markup Language*) per salvare i modelli UML facilita l'integrazione tra strumenti automatici di tipo diverso. In particolare, molti data base biologici pubblici consentono di esportare informazioni in questo formato.

cellule che sono composte da componenti quali proteine, apparati, nucleo, DNA ecc.). Sono stati già definiti anche alcuni ambienti software per modellare sistemi biologici con UML (si veda www.biouml.org). Lo sforzo più significativo in questa direzione è comunque fornito dalla definizione di un profilo SB-UML specifico per la biologia dei sistemi [12] sottoposto al comitato di standardizzazione OMG.

Utilizzare UML e i suoi meccanismi di estensione per la biologia dei sistemi è una scelta strategica per i seguenti motivi. Il formalismo è grafico e non molto distante da quelli solitamente usati dai biologi ai quali si richiede, quindi, un piccolo sforzo di adeguamento. Anche se UML non ha una semantica formale, è sufficientemente strutturato da consentire la definizione di traduttori automatici in calcoli formali. Importanti istituti come il Pasteur di Parigi lo utilizzano come meccanismo di rappresentazione per il circolo delle informazioni interne e questo favorisce la sua diffusione. Inoltre, essendo uno standard molto diffuso nell'area IT (*Information Technology*) corredato da molti strumenti automatici, dovrebbe essere ridotta la fase di *start-up* per la produzione di strumenti mirati al dominino biologico.

Adesso si descriverà brevemente come le **algebre di processo** possono rappresentare i sistemi biologici. I processi biomolecolari sono reti di proteine che interagiscono, ciascuna composta da molte parti strutturali distinte e indipendenti chiamate "domini". Le interazioni binarie tra domini dipendono dalla complementarietà strutturale e chimica di particolari porzioni delle proteine. L'interazione tra proteine causa a sua volta variazioni biochimiche dei domini che influenzano le future interazioni dei componenti coinvolti. Inoltre, l'interazione tra proteine guida direttamente il funzionamento delle cellule, e le modifiche delle proprietà biochimiche delle proteine sono, quindi, il meccanismo principale che guida molte funzionalità cellulari. Queste caratteristiche corrispondono piuttosto strettamente a quelle dei sistemi distribuiti in cui la topologia di interconnessione delle varie componenti può variare dinamicamente cambiando così le potenziali interazioni future.

Per avere un parallelo più dettagliato tra sistemi biologici e algebre di processo si possono considerare le molecole che interagiscono come processi concorrenti e la complementarietà delle caratteristiche biochimiche come coppie di operazioni complementari (*send* e *receive*) sullo stesso canale di comunicazione. La modifica successiva all'interazione biologica è modellata consentendo la comunicazione di canali che, quindi, alterano la struttura topologica della rete di interconnessione. Infatti, se un certo processo riceve un nuovo nome di canale, da quel momento in poi lo può utilizzare per comunicare con tutti gli altri processi che lo conoscono. Al contrario, se un certo processo consuma il nome di un canale per effettuare su di lui una comunicazione, non potrà poi più comunicare con i processi che conoscono quel canale fino a che non acquisisce nuovamente il nome. Tecnicamente, il comportamento dinamico dei sistemi biologici viene formalmente definito dalla semantica operazionale dei calcoli. In letteratura sono stati proposti recentemente numerosi calcoli per rappresentare sistemi biologici (si ricordano tra questi Biochemical π-calculus [10], BioAmbients [11], Core Molecular Biology [2], Brane calculus [1]). La descrizione accurata degli aspetti quantitativi che guidano i processi molecolari viene inglobata nel parallelo sopra riportato utilizzando **algebre di processo stocastiche** in cui le transizioni sono governate da distri-

Le **algebre di processo** sono dei semplici calcoli introdotti alla fine degli anni settanta da Tony Hoare e Robin Milner per modellare le peculiarità dei sistemi concorrenti in modo rigoroso. Esse comprendono pochi operatori che compongono azioni elementari indicate con lettere minuscole nel seguito e processi indicati invece con lettere maiuscole: sequenzializzazione di azioni e processi ($a.P$), composizione parallela di processi ($P|Q$), composizione non deterministica di processi ($P + Q$), dichiarazione di nomi nuovi (*new a*), operatore di scelta [$x = y$], ricorsione (rec $X. P$). Le azioni sequenziali possono essere di tre tipi: $a!b$ per spedire il nome b sul canale a, $a?x$ per ricevere un dato che rimpiazzerà la variabile targa x sul canale a, oppure t per rappresentare un'azione interna del sistema non visibile a un osservatore esterno. Lo scopo principale è quello di definire l'interazione e la cooperazione tra processi concorrenti e mobili.
La semantica intuitiva degli operatori elencati sopra è la seguente. L'azione a è la prima azione atomica che il processo $a.P$ può compiere. La ricezione $a?x$ lega le occorrenze della variabile x nel processo prefisso P. In altre parole, un dato sarà ricevuto sul canale a e sostituirà tutte le occorrenze libere della variabile targa x in P. Il prefisso di invio $a!x$ invia il nome x sul canale a senza legare le occorrenze di x in P. Nel processo (new x)P, l'operatore di restrizione new crea un nuovo (unico) nome x il cui raggio di azione è P. L'operatore di scelta [$x = y$] è soddisfatto se i due nomi sono uguali e consente l'esecuzione del processo che prefigge. Se la scelta non è soddisfatta l'esecuzione si ferma. Nella composizione parallela $P|Q$ i due processi sono eseguiti indipendentemente e possono comunicare se condividono uno stesso nome di canale. In altre parole $a!x.P|a?y.Q$ può comunicare inviando dalla parte sinistra alla parte destra della composizione il nome x sul canale a. Il processo risultante dopo la comunicazione sarà $P|Q\{x/y\}$, dove $\{x/y\}$ rappresenta l'operazione di sostituzione del nome x alle occorrenze libere di y nel processo cui è applicata. La somma rappresenta una scelta non deterministica: $P + Q$ si comporterà in modo mutuamente esclusivo o come P o come Q. Infine, rec $X.P$ rappresenta la definizione ricorsiva del processo P, cioè la possibilità di ripetere l'esecuzione del processo P tante volte quante si vuole.
La semantica formale di questi calcoli è solitamente fornita in modo operazionale sfruttando l'approccio operazionale introdotto da Gordon Plotkin e basato su assiomi e regole di inferenza. Il rigore formale che ne deriva consente di dimostrare proprietà dei programmi senza perdere in intuizione. Il comportamento dinamico dei sistemi rappresentati viene espresso mediante sistemi di transizione che sono essenzialmente dei grafi etichettati orientati. Gli stati rappresentano le configurazioni del sistema e le transizioni le azioni che il sistema può compiere per cambiare configurazione. Le etichette delle transizioni forniscono informazioni sul tipo di azione che esse rappresentano.

Algebre di processo stocastiche
Inizialmente le algebre di processo sono state utilizzate solo per descrivere e studiare aspetti qualitativi di sistemi concorrenti e mobili. L'evoluzione della teoria e le prime applicazioni a casi di studio reali hanno subito mostrato il limite di un approccio qualitativo. Per esempio se si vuole progettare un sistema distribuito di una qualche complessità non si può prescindere dalle prestazioni del sistema sin dai primi passi di progettazione. Questo ha fornito la spinta per estendere la teoria delle algebre di processo con informazioni quantitative vedendo la comparsa in letteratura sia di algebre di processo temporali che probabilistiche. Questo primo passo nel quantitativo non è però sufficiente a risolvere tutti i problemi posti dalla progettazione avanzata di sistemi. Il passo decisivo viene fatto da Jane Hillston quando introduce una variante stocastica di una semplice algebra di processo. L'idea di base è quella di arricchire i prefissi sequenziali delle algebre di processo (si veda il riquadro su algebre di processo) con una distribuzione probabilistica: i nuovi prefissi hanno, quindi, la forma $(a,F).P$ dove a è l'azione standard delle algebre di processo e F è la distribuzione probabilistica continua. A questo punto, il supporto a tempo di esecuzione del calcolo viene reso probabilistico introducendo il concetto di gara tra tutte le azioni che sono abilitate per essere eseguite in una data configurazione. L'idea è che tutte le azioni abilitate tentano di eseguire il loro compito, ma solo la più veloce riesce. Un teorema fondamentale delle distribuzioni continue assicura che la probabilità che due azioni abilitate terminino simultaneamente è zero. Ciò rende non ambiguo il meccanismo di scelta della azioni abilitate. A questo punto, il sistema di transizione viene a coincidere, con piccoli aggiustamenti tecnici, con un processo stocastico che può essere studiato per avere misure quantitative del sistema che rappresenta facendo riferimento a tecniche standard. Il vantaggio di questo approccio rispetto, ad esempio, a reti di code è che il passaggio dalla specifica al processo stocastico avviene automaticamente attraverso la semantica del calcolo e quindi può essere dimostrato corretto una volta per tutte.

buzioni probabilistiche. Il calcolo delle distribuzioni da associare alle transizioni si basa sull'osservazione che l'interazione tra due proteine è governata da una costante di interazione determinata empiricamente in base alle affinità biochimiche dei reagenti e dalla concentrazione. L'unico dei calcoli menzionati che può gestire aspetti stocastici è il π-calcolo nella sua variante stocastica [9, 10] (si veda, a tal proposito il paragrafo 4 sulla simulazione). Per esempio, il meccanismo di trascrizione regolato da un ciclo a *feedback* positivo astrattamente rappresentato dal diagramma biologico in figura 3 è tradotto dal programma in Biochemical π-calcolo stocastico di tabella 1.
Si conclude questo paragrafo discutendo brevemente i meccanismi di estrazione e riflessione, ipotizzando di avere una rappresentazione grafica UML-like. In informatica sono stati fatti notevoli sforzi per cercare di ottenere modelli formali basati su algebre di processo a partire da rappresentazioni UML [5]. Queste tecniche possono essere usate per ottenere descrizioni formali di sistemi biologici quando questi sono rappresentati mediante UML e, quindi, applicare le tecniche di analisi tipiche delle algebre di processo.

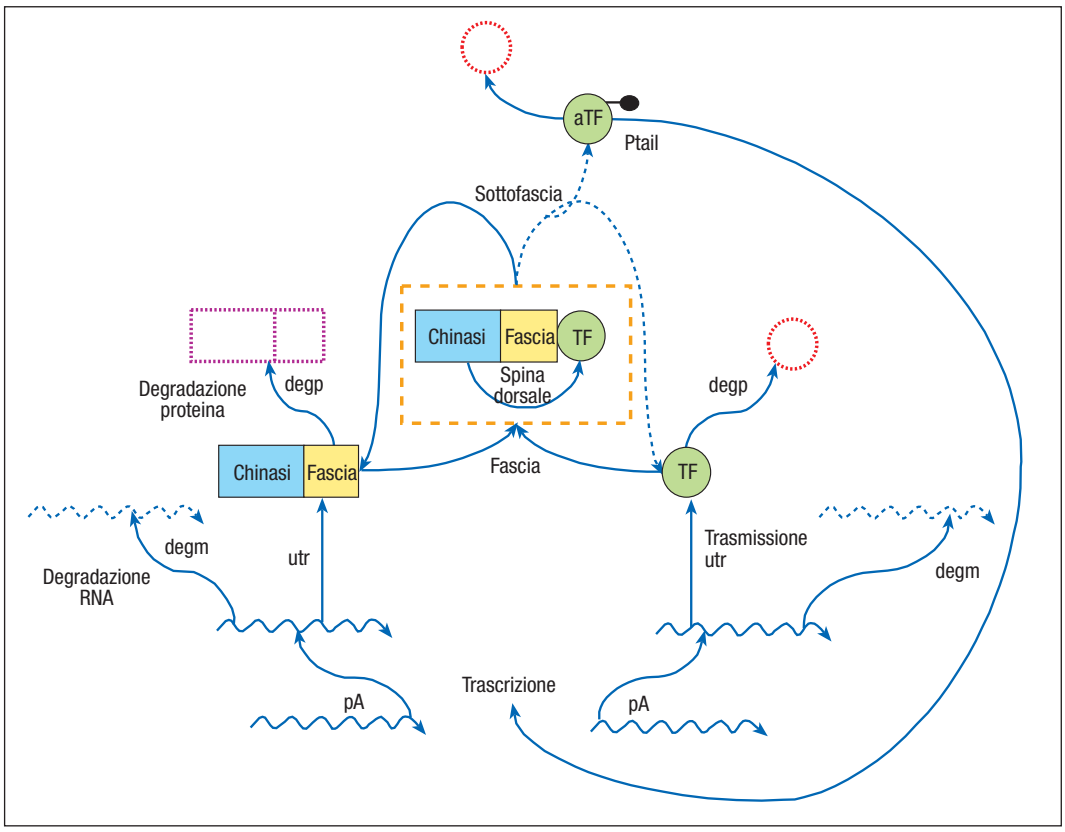


FIGURA 3

Rappresentazione di un meccanismo di trascrizione regolato da un ciclo a feedback positivo

```
Sys = Gene_A|Gene_TF|Transer|Transl|RNA_Deg|Protein_Deg
Gene_A = (basal(), 4).(Gene_A|RNA_A) + (pA(), 40).(Gene_A|RNA_A)
RNA_A = (utr(), 1).(RNA_A|Protein_A) + (degm(), 1)
Protein_A = (νbb1, bb2, bb3)(Binding_Site|Kinase)
Binding_Site = (bind ⟨bb1, bb2, bb3⟩, 0.1).Bound_Site + (degp(), 0.1).(bb3, ∞)
Bound_Site = (bb1, 10).Binding_Site + (degp(), 0.1).(bb3. ∞).(bb3, ∞)
Kinase = (bb2 ⟨ptail⟩, 10).Kinase + (bb3(), ∞)
Gene_TF = (basal(), 4).(Gene_TF|RNA_TF) + (pA(), 40).(Gene_TF|RNA_TF)
RNA_TF = (utr(), 1).(RNA_TF|Protein_TF) + (degm(), 1)
Protein_TF = (bind (c_bb1, c_bb2, c_bb3), 0.1).Bound_TF + (degp(), 0.1)
Bound_TF = (c_bb1(), 10).Protein_TF + (c_bb3(), ∞) + (c_bb2(tail), 10).
           ((c_bb1(), 10).Active_TF(tail) + (c_bb3(), ∞))
Active_TF(tail) = ( tail, 100).Active_TF(tail) + (degp(), 0.1)
Transer = (basal, 4).Transer + (ptail(), 100).(pA, 40).Transer
Transl = ( utr, 1).Transl
RNA_Deg = (degm, 1).RNA_Deg
Protein_Deg = (degp, 0.1).Protein_Deg
```

TABELLA 1

Rappresentazione in BioSPI del sistema di figura 3

3. ANALISI

Le tecniche di analisi del comportamento dei sistemi specificati mediante algebre di processo si dividono in statiche e dinamiche. Quelle più usate fino ad oggi in ambito biologico sono quelle dinamiche che prevedono la costruzione di un modello del comportamento a partire dalla descrizione (per esempio un sistema di transizione, cioè un grafo orientato in cui i nodi rappresentano gli stati del sistema e le transizioni gli eventi che causano il passaggio di stato – si veda tabella su algebre di processo). Le proprietà che si riescono a studiare con queste tecniche possono essere sia di tipo qualitativo che di tipo quantitativo. Tra le prime si ricordano la causalità tra transizioni o eventi, la località in cui certe transizioni avvengono, la concorrenza di transizioni [3].

Lo studio della relazione di causalità tra transizioni (la prima causa la seconda se è condizione necessaria per la seconda e ne influenza l'esecuzione) consente di determinare su un modello dinamico di una malattia quali sono gli eventi scatenanti e consente anche di tracciare in modo preciso il comportamento di un dato farmaco sui meccanismi di segnalazione della malattia. Da qui il concetto di modello predittivo se attraverso queste analisi si riescono a prevedere nuovi comportamenti biologici validabili attraverso esperimenti di laboratorio.

Anche la località gioca un ruolo essenziale nella modellazione e nell'analisi di sistemi biologici. Infatti, è essenziale sapere la localizzazione di certi componenti per determinare la probabilità o semplicemente la possibilità di una loro interazione. Dato un generico modello comportamentale di un sistema biologico, la località può essere utilizzata per ridurne la dimensione eliminando i comportamenti derivanti da interazioni tra componenti che non sono sufficientemente vicini o che non possono proprio entrare in contatto.

Altro esempio importante di proprietà da considerare è la concorrenza, ossia la possibilità per due o più transizioni di avvenire contemporaneamente. Questo consente di studiare fenomeni come ad esempio il *rolling* dei globuli bianchi in corrispondenza di tessuti infiammati nel suo complesso e non semplicemente studiando il comportamento di un singolo globulo. I risultati che si ottengono nei due casi sono abbastanza difformi e hanno ricadute diverse sull'evolvere dell'infiammazione.

Passando alle proprietà quantitative si ricorda come i sistemi di transizione possano, con piccole manipolazioni, essere interpretati come processi stocastici quando gli archi sono etichettati mediante distribuzioni probabilistiche (tipicamente esponenziali in tempo continuo). Queste tecniche sono state adottate da molti anni nel campo della valutazione delle prestazioni dei sistemi distribuiti, originando quelle che sono chiamate algebre di processo stocastiche.

Analizzare le proprietà di sistemi individuali non è abbastanza. Ulteriori conoscenze sulle funzionalità e possibili evoluzioni di reti molecolari possono essere acquisite confrontando i sistemi relativamente alle condizioni biologiche in cui operano, ai tipi di cellule e organismi che li compongono. La biologia computazionale (disciplina che sviluppa algoritmi efficienti per manipolare grandi quantità di dati, ad esempio al fine di confrontare due o più o sequenze di DNA, per ricostruire sequenze di nucleotidi data una conoscenza frammentaria delle sequenze o per generare alberi evolutivi a partire da un insieme di genotipi) ha ottenuto importanti risultati confrontando le sequenze e le strutture di singole molecole. In modo analogo, si possono usare strumenti messi a disposizione dalla teoria della concorrenza come le equivalenze basate sul concetto di bisimulazione per confrontare il comportamento dinamico di intere reti molecolari. Quello che è possibile, quindi, definire è una misura di omologia dei processi molecolari derivata dallo studio dei modelli. Le ricadute di una tale applicazione sono significative sia in campo informatico che biologico-medico. Dal punto di vista computazionale si può trarre ispirazione per nuove nozioni di equivalenza in quanto la nozione di omologia biologica è molto più complessa di quella di bisimulazione. Sul lato biologico lo studio comparativo di condizioni patologiche (per esempio, confronto del comportamento di un tessuto normale e di un tessuto tumorale) può con-

sentire di tracciare a ritroso importanti passi che stanno alla base dell'attivazione della malattia. Questo è tanto più possibile quanto più si riescono a compenetrare nella definizione delle equivalenze anche le nozioni di causalità e località.
Il problema principale delle tecniche dinamiche sopra descritte è dato dalla dimensione del sistema di transizione che è esponenziale rispetto alla descrizione testuale in algebre di processo. La conseguenza immediata è che data la grande dimensione dei sistemi biologici è impensabile avere algoritmi che possano esaminare in modo esaustivo lo spazio degli stati. Le soluzioni proposte sono prevalentemente orientate a ridurre la complessità computazionale del problema informatico a scapito della precisione dei risultati che si possono ottenere. Si descriverà qui un possibile utilizzo di tecniche di analisi statica e si rimanda al prossimo paragrafo la discussione delle tecniche di simulazione.
Le tecniche di analisi statica sono state introdotte originariamente per effettuare ottimizzazioni di compilatori, ma oggi le loro aree di applicazione sono molto più ampie. L'idea alla base di queste tecniche è la possibilità di estrarre informazioni complesse sul comportamento dinamico di sistemi semplicemente guardando alla loro descrizione testuale: il programma. Questo vuol dire che non è necessario costruire il modello del comportamento dinamico (il sistema di transizione) e, quindi, viene meno il vincolo dato dall'esponenzialità della rappresentazione. Al contrario dell'analisi dinamica, ogni analisi statica deve essere definita in relazione a una particolare proprietà che si vuole studiare e al particolare linguaggio di specifica che si intende usare. Da qui la necessità di avere un formalismo unico per descrivere molti aspetti diversi dei sistemi biologici al fine di limitare il numero di analisi che si devono definire.
Le informazioni che si estraggono mediante analisi statica del testo del programma sono corrette rispetto al comportamento dinamico del programma, ma non è possibile ottenere informazioni esatte. Il compromesso dell'analisi statica per avere algoritmi efficienti è a scapito della precisione delle informazioni ricavate. Da qui il concetto di approssimazione. Ci possono essere sia approssimazioni per eccesso che approssimazioni per difetto. Nel primo caso si ottiene uno spazio delle soluzioni del problema che contiene strettamente le soluzioni esatte e, quindi, si può affermare con certezza solo ciò che non potrà mai accadere. Nel caso di approssimazioni per difetto si ottiene uno spazio delle soluzioni del problema che è strettamente contenuto nello spazio delle soluzioni esatte. In questo caso si può dire con certezza solo ciò che accadrà sicuramente. La situazione inaccettabile per una approssimazione è quando lo spazio delle soluzioni calcolato contiene solo un sottoinsieme delle soluzioni esatte perché in questo caso non si ha alcun controllo sulla correttezza dei risultati ottenuti.
Anche se l'analisi statica è stata introdotta per studiare proprietà completamente diverse, può contribuire in modo significativo allo sviluppo della bioinformatica. Infatti le tecniche di approssimazione individuate consentono di studiare sistemi almeno un ordine di grandezza più grandi di quelli studiati mediante tecniche dinamiche. Le proprietà che possono essere esaminate riguardano la localizzazione di componenti all'interno di strutture più complesse, le loro possibili interazioni e migrazioni (per esempio la traslocazione nel nucleo di una cellula e la conseguente trascrizione), la determinazione di cicli a *feedback* positivo o negativo all'interno di grandi reti di segnalazione [7].

4. SIMULAZIONE

Come accennato nella precedente sezione, anche le tecniche di simulazione consentono di evitare la costruzione di un intero modello del comportamento dinamico, eliminando il problema dell'esponenzialità delle rappresentazioni. L'idea alla base della simulazione è quella di eseguire il programma che rappresenta il sistema biologico scegliendo una tra tutte le possibili esecuzioni (in termini di modello dinamico vuol dire scegliere un cammino sul sistema di transizione). Ripetendo un numero molto elevato di volte l'esecuzione si ottiene una descrizione "media" del comportamento dinamico del sistema considerato.
I principali tentativi di modellare il comportamento dinamico dei sistemi biologici so-

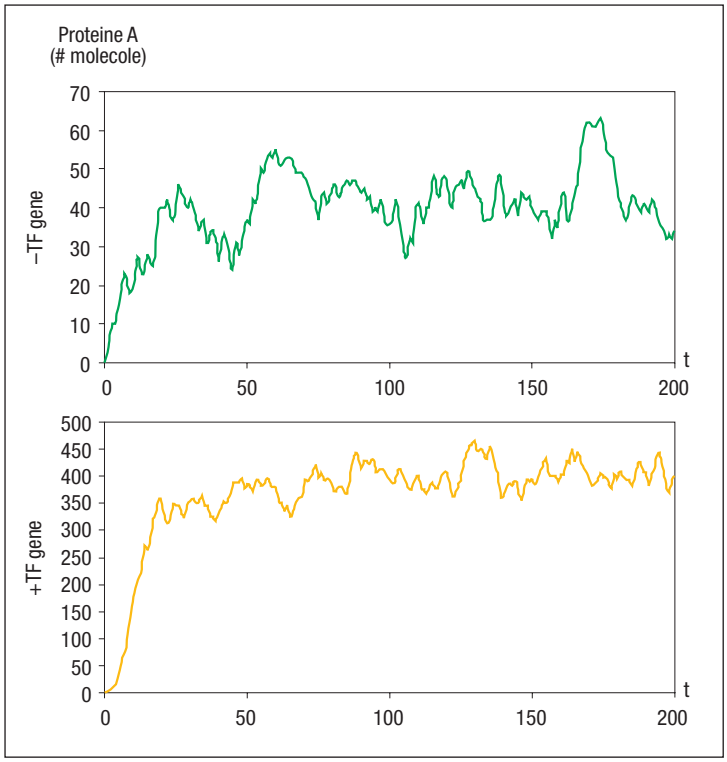


FIGURA 4
Risultati della simulazione mediante BioSPI del sistema rappresentato graficamente in figura 3 e in stocastico p-calcolo in tabella 1

no basati su equazioni differenziali ordinarie o stocastiche, su metodi di simulazione discreta che possono rifarsi alle tecniche Monte-Carlo, reti bayesiane [4]. Ciascuno degli approcci menzionati è in grado di catturare alcuni degli aspetti specifici dei meccanismi di segnalazione cellulare, ma nessuno è in grado di integrare la dinamica con gli aspetti molecolari e biochimici. Queste limitazioni possono essere superate mediante l'utilizzo di algebre di processo come si è già visto nelle precedenti sezioni.

Il primo ambiente di simulazione basato sulla realizzazione di un supporto a tempo di esecuzione probabilistico per il π-calcolo, implementando quindi una variante dello stocastico πi-calcolo [9] è BioSPI [10]. La realizzazione è basata su Flat Concurrent Prolog e supporta completamente comunicazioni e scelte non deterministiche (rese poi probabilistiche dal supporto a tempo di esecuzione che implementa l'algoritmo di Gillespie).

Il sistema realizzato consente di specificare le quantità iniziali dei vari componenti di cui si vogliono studiare le potenziali interazioni e i tassi probabilistici con cui avvengono le comunicazioni sui vari canali che compaiono nella specifica del sistema (si veda per esempio Tabella 1). Il meccanismo di simulazione permette poi di monitorare come le concentrazioni e i prodotti delle interazioni variano al variare del tempo (anche la scala temporale può essere variata scegliendo quella più adeguata al fenomeno considerato). Per esempio il risultato della simulazione del programma di tabella 1 fornisce il risultato riportato in figura 4.

5. CONCLUSIONI E SVILUPPI FUTURI

In questo contributo sono stati esaminati gli aspetti della bioinformatica principalmente legati alla descrizione dei comportamenti dinamici dei sistemi biologici complessi. Gli obiettivi scientifici principali che sono stati considerati riguardano:

- la rappresentazione dei sistemi che sia amichevole per i biologi, ma che consenta al tempo stesso di derivare in modo automatico modelli formali per lo studio rigoroso del comportamento biologico;
- l'analisi qualitativa e quantitativa di proprietà dei sistemi e possibile definizione di una omologia di processi molecolari basata su nozioni di equivalenze comportamentali definite i teoria della concorrenza;
- la simulazione del comportamento di sistemi basata su implementazioni di supporti a tempo di esecuzione probabilistici per le algebre di processo;
- la necessità di costituire un linguaggio comune tra informatici e biologi per la forte interdisciplinarietà della bioinformatica.

Le tecniche di analisi e di simulazione non sono in alternativa, ma servono entrambe per avere una visione complessiva del sistema studiato quanto più precisa possibile. Infatti, mentre con le simulazioni si riescono a studiare quantitativamente le variazioni percentuali delle sostanze coinvolte in certi fenomeni biologici esse sono inerentemente limitate nella capacità di fornire risposte generali su proprietà intrinseche di sistemi. Le tecniche di analisi analitica basate sui metodi formali sono, invece, adatte a fornire risposte generali a domande del tipo "un segnale può passare attraverso una certa

molecola sotto certe condizioni? Sotto ogni condizione?" oppure " se modifichiamo il comportamento di una certa molecola, cosa accadrà a un'altra data molecola? O al sistema nel suo complesso?"
Uno dei motivi principali del graduale abbandono dell'approccio riduzionista (si veda il riquadro sulla biologia dei sistemi) nella biologia attuale è dovuto all'impossibilità di coniugare la conoscenza acquisita sulle componenti minimali dei sistemi per ottenere un modello dell'insieme. Ciò è dovuto alla enorme complessità dei sistemi biologici. Un vantaggio che deriva dall'uso delle algebre di processo (e dalla teoria dei linguaggi in generale) è la loro **natura** inerentemente **composizionale** che fornisce regole precise e non ambigue di composizione di oggetti elementari per costruire oggetti più complessi.
In conclusione di questo paragrafo si riportano alcune considerazioni sulle possibili evoluzioni della bioinformatica. L'obiettivo a lungo termine della bioinformatica (o almeno degli aspetti di questa disciplina maggiormente discussi in questo contributo) sono quelli di fornire tecniche sia *predittive* che *prescrittive* del comportamento dei sistemi biologici complessi. Le tecniche predittive sono in grado di prevedere i comportamenti di un sistema e, quindi, sono semplicemente descrittive, mentre le tecniche prescrittive dovrebbero essere in grado di imporre un determinato comportamento (o classe di comportamenti) ai sistemi, essendo, quindi, invasive. Si discutono adesso brevemente le due tipologie di tecniche.
Una tecnica predittiva si basa sulla bontà del modello del sistema biologico e mediante analisi delle proprietà del modello determina possibili evoluzioni. Questa strategia può essere usata dai ricercatori per dimostrare in laboratorio certi comportamenti ancora non noti oppure per prevedere le reazioni di un farmaco in presenza di determinate malattie. Essenziale per ottenere buoni risultati è la fase di validazione dei modelli che si utilizzano e, da qui, la grande attività attuale nella modellazione di sistemi biologici reali mediante algebre di processo al fine di ottenere dal modello almeno tutti i comportamenti noti del sistema considerato mediante applicazione di tecniche di analisi.

Il termine **"composizionalità"** in teoria dei linguaggi indica la possibilità di definire la semantica di un costrutto in termini della semantica dei suoi componenti. Questa è chiaramente una proprietà fondamentale per poter definire in modo finito e chiaro la semantica dei linguaggi di programmazione. Infatti se la definizione della semantica non fosse composizionale dovremmo elencare la semantica di tutti i possibili programmi esprimibili in un dato linguaggio e per ogni linguaggio interessante questi sono infiniti. Parlando di sistemi (biologici) la composizionalità è la possibilità di definire un modello mediante integrazione (composizione) dei modelli dei sotto-sistemi che lo costituiscono. Quando le regole di composizione sono chiare per un certo formalismo e dominio applicativo, la composizionalità è anche una metodologia di progettazione e sviluppo che consente di esaminare e determinare soluzioni per problemi semplici che poi verranno composte per risolvere problemi più complessi.

Le tecniche prescrittive definiscono, invece, algoritmi che vengono eseguiti su *hardware* vivente. Considerando che ogni singola cellula ha approssimativamente 1 MIPS di potenza di calcolo e 1 MEGA di memoria, le potenzialità dei computer viventi sono estremamente interessanti. La possibilità di usare le cellule per eseguire algoritmi avrebbe una ricaduta immensa anche in campo biologico-medico. Per esempio, riprogrammando le cellule che esibiscono comportamenti anomali si potrebbero trovare cure efficaci per tutta la classe delle malattie auto-immuni come la sclerosi multipla oppure per i tumori. Su questa strada si stanno muovendo numerosi importanti gruppi di ricerca cercando di definire un modello completo del funzionamento della cellula. Questo è sicuramente il primo passo per ipotizzare poi metodologie in grado di controllare ed eventualmente modificare il comportamento delle cellule.
Concludendo si può certamente affermare che gli aspetti dinamici dei sistemi biologici e le tecniche informatiche per dominarne la complessità sono un campo di ricerca agli esordi e che probabilmente dominerà la scena bioinformatica dei prossimi anni con attività altamente interdisciplinari. Infatti, per validare i modelli di comportamento che portino a tecniche predittive è assolutamente necessario interagire con biologi e per poter comprendere completamente i risultati delle analisi i biologi devono poter comprendere ciò che gli informatici hanno fatto. Da qui la necessità di creare una co-

munità scientifica internazionale che sia a cavallo tra le discipline delle scienze della vita e quelle del settore dell'informazione. Recentemente la branca della biologia dei sistemi sta cercando di percorrere questa strada reclutando fisici, matematici e informatici all'interno dei loro istituti di ricerca. Analogo sforzo dovrebbe essere fatto nei centri di ricerca bioinformatici che nascono dai dipartimenti di informatica. Occorre tener presente che come sempre quando si creano nuove comunità scientifiche da aggregazioni di persone provenienti da aree disciplinari diverse è essenziale stabilire immediatamente un criterio di pariteticità tra le varie anime del gruppo. Per far questo è necessario individuare con chiarezza quelle che sono le necessità e le aspettative dei biologi e quali quelle degli informatici assumendo che nessuna delle due scienze è al servizio dell'altra. Su questa base paritaria sarà poi possibile attivare progetti di grande respiro e creare una robusta comunità scientifica.

Bibliografia

[1] Cardelli L.: *Brane calculus*. Rapporto tecnico Microsoft Research Cambridge, 2003.

[2] Danos V., Laneve C.: *Graphs for Core molecular biology*. Proceedings of CMSB03, LNCS 2602, Springer-Verlag, 2003.

[3] Degano P., Priami C.: Noninterleaving semantics of mobile processes. *Theoretical Computer Science*, Vol. 216, n. 1-2, 1999, p. 237-270.

[4] De Jong H.: Modeling and Simulation of Genetic Regulatory Systems. *A Literature Review. Journal of Computational Biology*, Vol. 9, n. 1, 2002, p. 67-103.

[5] Karp P.D., Krummenacker M., Paley S., Wagg J.: Integrated pathway/genome databases and their role in drug discovery. *Trends in Biotechnology*, Vol. 17, n. 7, 1999, p. 275-281.

[6] Koremblat K., Priami C.: *Towards extracting _-calculus from UML sequence and state diagrams*. In Compositional verification of UML models 03, apparirà su ENTCS 2003.

[7] Nielson F., Nielson H.R., Priami C., Scuch da Rosa D.: *Control Flow Analysis of BioAmbients*. Bioconcur 2003, Apparirà su ENTCS.

[8] Ogata H., Goto S., K. Sato, Fujibuchi W., Bono H., Kanehisa Kegg M.: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, Vol. 27, n. 1, 2000, p. 29–34.

[9] Priami C.: Stochastic π-calculus. *The Computer Journal*, Vol. 38, n. 6, 1995, p. 578-589.

[10] Priami C., Regev, Shapiro E., Silverman W.: Application of a stochastic passing-name calculus to representation and simulation of molecular processes. *Information Processing Letters*, Vol. 80, 2001, p. 25-31.

[11] Regev A., Panina E.M., Silverman W., Cardelli L., Shapiro E.: *BioAmbients: an abstraction for biological compartments*. Apparirà su Theoretical Computer Science.

[12] Roux-Rouquie M., Caritey N., Gaubert L., Rosenthal-Sabroux C.: *Using the Unified Modeling Language (UML) to guide Systemic description of Biological Processes and Systems*. Apparirà su BioSystems.

[13] Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhauser R., Pruss M., Schacherer F., Thiele S., Urbach S.: The transfac system on gene expression regulation. *Nucleic Acids Research*, Vol. 29, n. 1, 2001, p. 281–283.

Corrado Priami ha ricevuto laurea e dottorato di ricerca in informatica all'Università di Pisa, è stato ricercatore all'Ecole Normale Superieur di Parigi e poi all'Università di Verona dove è diventato professore associato di Informatica. Attualmente è professore straordinario di informatica all'Università di Trento, Dipartimento di Informatica e Tlc, dove guida il gruppo di bioinformatica ed è responsabile dei corsi di laurea e laurea specialistica in Informatica. Rappresenta l'Università di Trento nel consiglio di amministrazione della Trento School of Management ed è delegato di ateneo per i progetti europei nel sesto programma quadro. I suoi interessi di ricerca coprono, oltre la bioinformatica, i sistemi distribuiti e mobili e i linguaggi di programmazione in senso lato. Coordina un progetto europeo nell'area del global computing e un progetto nazionale sulle tematiche riportate nell'articolo. Partecipa, inoltre, a numerosi altri progetti sia nazionali che internazionali e ha numerose collaborazioni con industrie nel settore delle comunicazioni mobili e delle biotecnologie. Opera come revisore di progetti europei e nazionali nelle sue aree di interesse. Ha pubblicato oltre 80 articoli su riviste e convegni internazionali ed è autore e curatore di alcuni libri sulle tematiche sopra esposte. Ha partecipato e partecipa, anche come presidente, a numerosi comitati di programma di eventi internazionali prevalentemente nell'area della bioinformatica. Co-dirige una scuola di dottorato aperta a giovani ricercatori informatici e biologi per la costruzione di un gruppo scientifico interdisciplinare.
priami@dit.unitn.it