

CHIMICA & FARMACEUTICA



Stefano Alcaro, Anna Artese, Francesco Ortuso
Dipartimento di Scienze Farmacobiologiche
Università di Catanzaro "Magna Graecia"
alcaro@unicz.it

INDAGINE STATISTICA SULLA DIFFUSIONE DELLE TECNOLOGIE COMPUTAZIONALI AVANZATE PER LA PROGETTAZIONE DI NUOVI FARMACI

Un'indagine sull'uso di tecnologie computazionali è stata sottoposta ad un campione di industrie farmaceutiche italiane. I risultati indicano l'interesse per specifiche tecniche e aiutano di tracciare l'identikit del chimico farmaceutico computazionale da formare per le attuali esigenze industriali.

La scoperta di nuovi farmaci (*drug discovery*) rappresenta uno dei più ambiziosi e costosi processi dell'industria chimica. Essa comprende una serie di passaggi che oggi parte dallo studio approfondito del target macromolecolare, procede con l'identificazione di molecole bioattive (*lead compounds*), prevede prove biologiche, farmacologiche e cliniche per approdare, dopo diversi anni, all'immissione sul mercato.

Il successo nello sviluppo di nuovi farmaci è piuttosto raro, se si considera il gran numero di composti che entrano nel processo del *drug discovery*. I tempi ed i costi correlati a tale straordinario successo risultano di norma elevati, rispettivamente compresi tra i 10-12 anni e i 200-400 milioni di dollari.

Pertanto ogni mezzo per ridurre tempo e costi risulta utile per chi si occupa a livello industriale e accademico di *drug discovery*. Negli ultimi decenni un forte impulso in tal senso si deve allo sviluppo e alla diffusione di tecnologie computazionali avanzate orientate alla progettazione di nuovi farmaci. Tali metodi [1-3] hanno consentito di aumentare l'efficienza del processo, come nel caso di alcuni recenti esempi. Uno di quelli più citati riguarda la lotta contro il virus dell'Aids in cui le tecniche computazionali hanno molto contribuito, ad esempio, per il disegno razionale di inibitori della proteasi [4]. Anche nell'industria farmaceutica italiana tali tecniche hanno avuto una rapida diffusione, pertanto, è stata condotta un'indagine con lo scopo di monitorarne l'utilizzo. Tale studio è stato realizzato su un

campione di 10 industrie farmaceutiche italiane. Le aziende coinvolte, la cui identità non sarà rivelata per motivi di riservatezza, sono state scelte sulla base della loro disponibilità a collaborare in questo progetto e considerando la loro attività in vari settori farmaceutici, quali immunologia, oncologia, ginecologia, dermatologia e malattie neurodegenerative.

L'indagine, condotta in accordo con il Direttivo della Divisione di Chimica Farmaceutica della Società Chimica Italiana, ha avuto lo scopo di derivare informazioni utili per meglio indirizzare la formazione sia teorica che sperimentale dei giovani, studenti e dottorandi, interessati ad acquisire il background culturale riguardante le metodologie computazionali attualmente indispensabili per gli studi di progettazione razionale di un farmaco e dare loro maggiori opportunità professionali sia nell'industria che nell'accademia.

Lo studio condotto ha permesso, infatti, di delineare l'identikit del chimico computazionale attualmente impegnato nell'industria farmaceutica italiana e ha permesso di determinare le informazioni necessarie alla formazione di figure professionali adatte a soddisfare le esigenze della ricerca in questo settore.

Attraverso un apposito questionario, redatto in stretta collaborazione con esperti che operano nelle aziende farmaceutiche italiane, nel settembre del 2007 è stata inviata un'informativa atta ad identificare 10 aziende, sulle quali condurre l'indagine statistica, con caratteristiche differenti in termini di ambito operativo, dimensioni ed esperienza nelle tecniche computazionali.

Strutturazione del questionario

Il questionario proposto si articola su 26 domande a risposta multipla suddivise in 4 sezioni:

1. sezione conoscitiva circa l'impiego di tecniche computazionali, incluse le risorse dedicate a questo ambito di ricerca all'interno dell'azienda;
2. sezione monografica per ciascuna tecnica computazionale e sommario;
3. sezione software;
4. sezione hardware;

La prima sezione consta di 5 domande:

- 1.1 *Presso la Sua azienda si fa uso di tecniche computazionali nel settore di ricerca e sviluppo chimico-farmaceutico?*
- 1.2 *Se sì, da quanto tempo?*
- 1.3 *Quante persone strutturate sono dedicate allo studio computazionale?*
- 1.4 *Quante persone non strutturate (dottorandi, postdoct, tirocinanti) sono dedicate allo studio computazionale?*
- 1.5 *Indicare la percentuale di personale coinvolto nei progetti di ricerca chimico-farmaceutici dell'azienda.*

Nel corso dell'indagine 10 aziende hanno dato una risposta positiva alla domanda 1.1.

Su queste si è quindi concentrato il resto dell'indagine statistica.

Alla domanda 1.2 il questionario offriva la possibilità di rispondere su 4 fasce temporali comprese tra i >10 e <2 anni.

Alla domanda 1.3 e 1.4 erano previste 6 risposte comprese tra 0 ed oltre 15 unità dedicate allo studio computazionale, sia strutturate che non strutturate nell'azienda.

Alla domanda 1.5 era invece previsto di esplicitare la percentuale del personale coinvolto in attività *in silico* ed in altre attività di ricerca e sviluppo dell'azienda.

Alla seconda sezione, la più corposa del questionario, alle 15 domande su interesse ed impatto applicativo di ciascuna delle tecniche computazionali previste, si potevano dare 5 diverse tipologie di risposta, da interesse nullo a molto elevato. Le 15 tecniche alle quali è rivolta la seconda sezione del questionario sono di seguito elencate:

- 2.1 *Ab initio;*
- 2.2 *Semiempirici;*
- 2.3 *Meccanica Molecolare;*
- 2.4 *Dinamica Molecolare;*
- 2.5 *Docking;*
- 2.6 *Virtual screening bidimensionale;*
- 2.7 *Virtual screening tridimensionale;*
- 2.8 *QSAR;*
- 2.9 *3D-QSAR;*
- 2.10 *Multiallineamento proteico;*
- 2.11 *QSPP;*
- 2.12 *Homology Modelling;*
- 2.13 *ADME prediction;*
- 2.14 *De novo design;*
- 2.15 *Librerie chimiche.*

La sezione si conclude con un sommario che indica in percentuale l'uso di ognuna delle tecniche nell'ambito del comparto di R&S di ciascuna azienda farmaceutica.

La terza sezione è dedicata al software impiegato in ambito industriale. L'attenzione non è stata focalizzata sui singoli prodotti commerciali ma, piuttosto, su aspetti legati al tipo di scelta ed implementazione di tali prodotti.

In primo luogo si è voluto indagare sulla diffusione di programmi *open source*, ovvero distribuiti, spesso gratuitamente, sottoforma di codice sorgente, che ben si prestano ad essere modificati per meglio rispondere alle esigenze dell'utilizzatore finale. L'interesse nei confronti di questa particolare tipologia di software ha consentito di valutare la diffusione di competenze nell'ambito della programmazione e, posto che i progetti *open source* sono sempre in continuo sviluppo, la disponibilità all'aggiornamento del software all'interno dell'azienda. Diversamente la diffusione dei programmi *freeware*, rilasciati gratuitamente, ma senza il corredo del codice sorgente, possono fornire un'idea sulla sensibilità dell'azienda alle

CHIMICA & FARMACEUTICA

voci di spesa destinate alle metodologie computazionali. Infine, allo scopo di indagare in maggiore dettaglio la presenza di competenze nell'ambito della programmazione, sono stati richiesti i dati relativi all'eventuale sviluppo interno (*in house*) di metodologie o strumenti di calcolo. Le 2 domande poste in questa sezione prevedono la compilazione in percentuale di tre campi:

3.1 *Indicare la percentuale di utilizzo del software chimico-computazionale*

- a) *open source*
- b) *freeware*
- c) *proprietario e/o commerciale*

3.2 *Se si sviluppa un proprio software indicare la percentuale della modalità*

- a) *sviluppo all'interno dell'azienda*
- b) *sviluppo in collaborazione con software house*
- c) *partecipazione a progetti open source.*

La quarta sezione è dedicata ad informazioni sulle risorse hardware. Anche in questo caso si è preferito investigare gli aspetti legati alle tipologie di calcolo intensivo, di architettura e di workstation grafiche presenti nei laboratori di chimica computazionale delle aziende coinvolte anziché entrare nel merito dei prodotti hardware commerciali. Attraverso tre domande, pertanto, si è chiesto di indicare le percentuali di diffusione degli approcci predefiniti:

4.1 *Indicare le percentuali delle risorse di calcolo intensivo utilizzate a scopo di ricerca*

- a) *cluster Linux*
- b) *GRID computing*
- c) *super computer propri dell'azienda*
- d) *tempo di calcolo a noleggio*
- e) *personal computer*

4.2 *Indicare le percentuali delle architetture di calcolo intensivo utilizzate a scopo di ricerca*

a) *32 bit*

b) *64 bit*

4.3 *Indicare se gli applicativi per la manipolazione grafica risiedono prevalentemente su*

- a) *Workstation Unix/Linux*
- b) *Workstation Windows®*
- c) *Workstation Mac OS®.*

Risultati del questionario

Le 10 aziende farmaceutiche italiane consultate sono geograficamente situate nell'Italia centro-settentrionale, area che tradizionalmente vede una più alta densità industriale.

Dall'analisi della sezione 1 del nostro questionario si osserva come tutte fanno uso di tecniche computazionali nel settore di ricerca e sviluppo chimico-farmaceutico. Nel grafico riportato in Fig. 1 si evince che la maggior parte di esse ha esperienza pluriennale con le suddette metodologie.

Il numero degli strutturati impiegati mediamente nel settore chimico-computazionale è, nella maggior parte dei casi, compreso tra 2 e 4 unità. È, però, da segnalare la presenza nel nostro campione di due realtà agli antipodi, ovvero una sola azienda può contare su oltre 15 ed un'altra su una sola unità di addetti. Osservando le risposte relative al numero di non strutturati emerge la tendenza ad evitare l'impiego di personale estraneo all'azienda, probabilmente per questioni di riservatezza (Fig. 2).

Relativamente al numero medio degli occupati in attività *in silico* rispetto al numero medio di occupati in laboratorio sperimentale e in altre attività legate alla ricerca e sviluppo, la percentuale ammonta a circa il 13% (Fig. 3), valore sicuramente in crescita rispetto al passato, anche recente, quando la figura professionale del ricercatore computazionale era assente, o quasi, nelle aziende farmaceutiche italiane.

Dai risultati ottenuti nella sezione conoscitiva, si può concludere che il campione è abbastanza rappresentativo per consentire l'analisi delle

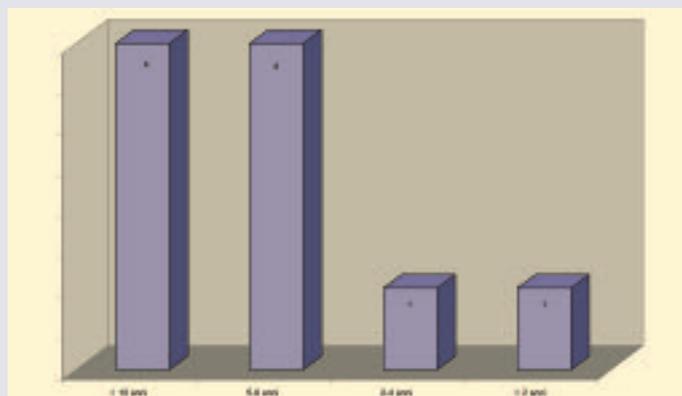


Fig. 1 - Esperienza nel settore chimico computazionale delle aziende campionate

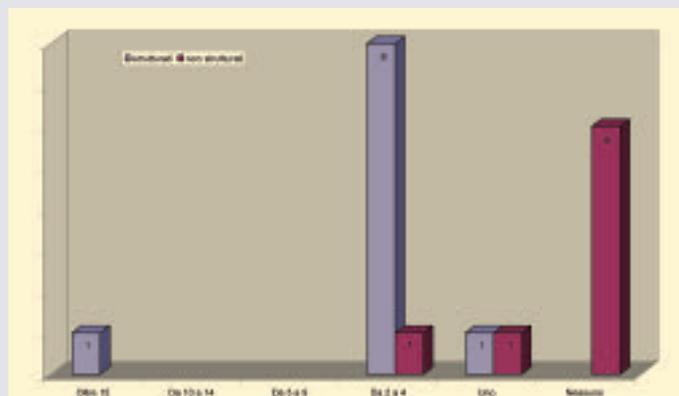


Fig. 2 - Distribuzione del numero di strutturati e non nel settore chimico computazionale delle aziende campionate

15 tecniche computazionali sulla base dei tre elementi richiesti nel questionario: l'utilizzazione, l'interesse e l'impatto (Fig. 4).

La seconda sezione ha fornito indicazioni sulle 15 tecniche con tre modalità di risposta differenti. Quelle relative all'interesse e all'impatto rappresentano le potenzialità culturali e applicative che ci si attende dalle metodologie investigate. Tra esse vi è una modesta correlazione (r^2 uguale a 0,68) dovuta probabilmente all'evoluzione di impiego delle tecniche più classiche (*ab initio*, dinamica molecolare, QSAR classica) o di altre più recenti che hanno dimostrato alcuni limiti nel processo del *drug discovery* (de novo design). Su queste si registra un generico calo dell'impatto potenziale rispetto all'interesse che esse suscitano.

I dati sull'utilizzazione sono invece specchio del reale impiego medio delle 15 tecniche nelle aziende campionate. L'analisi dell'utilizzo percentuale delle metodiche è stato condotto offrendo la possibilità di aggiungere altre tecniche a quelle indicate. Due aziende su 10 hanno indicato due voci legate rispettivamente all'analisi di sequenze e alla creazione delle banche dati molecolari che, per permettere un diretto confronto con i primi dati raccolti, sono state rispettivamente considerate nei metodi di multiallineamento proteico e nelle librerie chimiche. La correlazione tra utilizzo ed interesse e tra utilizzo ed impatto è sensibilmente inferiore al già modesto dato ottenuto in precedenza. Non sorprende che ciò possa avvenire poiché possono verificarsi altri aspetti che condizionano il reale utilizzo delle tecniche computazionali a livello industriale. Allo scopo di focalizzare con maggiore attenzione tali elementi è stato analizzato lo scarto tra l'utilizzo e gli altri due descrittori: interesse ed impatto (Fig. 5).

Confrontando i dati riportati nelle Fig. 4 e 5 si evince come le tecniche dotate di più ampio impiego, quali la meccanica molecolare, il *docking*, la 3D-QSAR e le librerie chimiche, risultino quasi sovrautilizzate rispetto all'interesse ed all'impatto che le stesse possono vantare.

Fanno eccezione le tecniche *ab initio* ed ADME il cui utilizzo rispecchia in maniera fedele le aspettative dell'azienda e dei ricercatori. È interessante notare come il *docking* sopravanzi, nell'utilizzo, il *Virtual Screening*, sia bi che tridimensionale, che viceversa presenta elevato sia l'interesse che l'impatto relativi. Tale osservazione sembra quasi suggerire che le aziende aspettino un'evoluzione dei codici *Virtual Screening*, ad oggi carenti sotto il profilo dell'accuratezza rispetto al *docking*. Per contro, tecniche quali le semiempiriche, la dinamica molecolare, il QSPR, l'*homology modelling* ed il *de novo design*, pur diffuse, trovano un impiego inferiore alle loro potenzialità di interesse ed impatto.

I dati relativi alle diverse tipologie di software presenti nelle 10 aziende coinvolte nella nostra indagine, rivelano (Fig. 6) un utilizzo prevalente dei codici proprietari e/o commerciali.

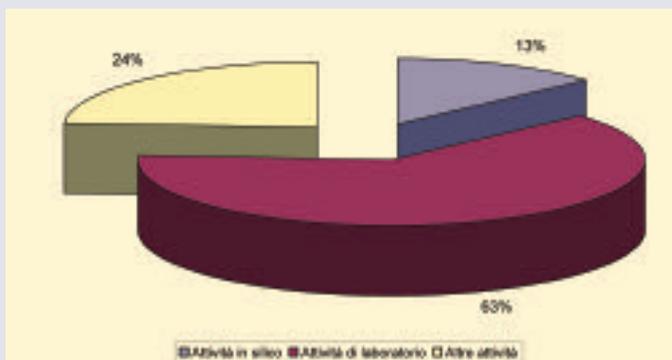


Fig. 3 - Distribuzione % delle mansioni nel settore ricerca e sviluppo delle aziende campionate

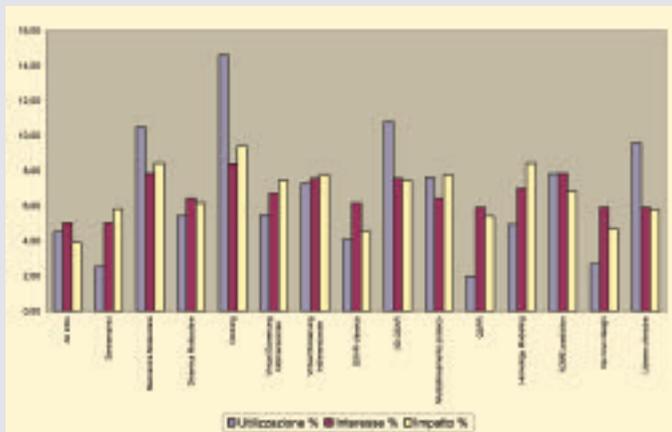


Fig. 4 - Media % della valutazione relativa all'utilizzazione, all'interesse ed all'impatto delle 15 tecniche computazionali oggetto del questionario

Il dato, associato al minore impiego di software sviluppato in proprio, descrive il deciso interesse nell'ottenimento del risultato (sviluppo di nuovi farmaci) utilizzando metodologie ampiamente valutate ed accettate dalla comunità scientifica. Tali prodotti sembrano rappresentare una forma di garanzia per l'azienda pur essendo una non trascurabile voce di spesa. D'altro canto tutte le aziende interpellate fanno uso di codici *open source*, in un caso l'utilizzo di questi software è addirittura preponderante. Questa osservazione lascia intendere un interesse verso nuove metodologie anche personalizzate. Con ogni probabilità la bassa diffusione di codici propri può essere attribuita alla scarsità di risorse umane proposte allo sviluppo metodologico. È chiaro, infatti, che apportare delle modifiche a codici già esistenti (*open source*) rappresenti un onere minore rispetto alla realizzazione integrale del prodotto. Si registra, inoltre, la singolare diffusione dei prodotti *freeware* che, pur essendo quasi sempre presenti, non sembrano rivestire un ruolo di primario interesse per nessuna delle aziende. Infine per quanto riguarda i programmi propri, il cui utilizzo si attesta intorno al 9%, la maggioranza delle aziende predilige lo sviluppo interno o, in minima parte, in compartecipazione a progetti di *open source*, mentre nessuno risulta coinvolto in collaborazione con *software house*.

CHIMICA & FARMACEUTICA

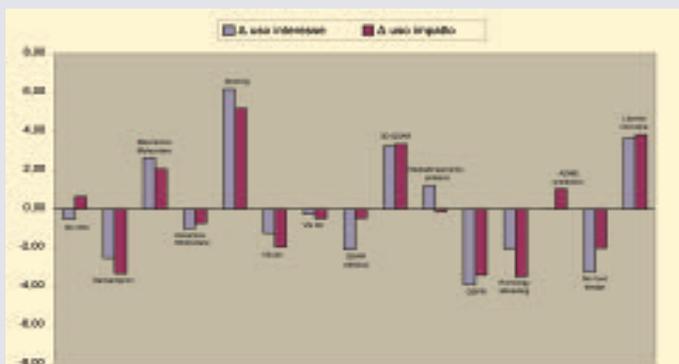


Fig. 5 - Scarto tra le medie percentuali tra utilizzo e interesse e tra utilizzo e impatto delle 15 tecniche computazionali oggetto del questionario

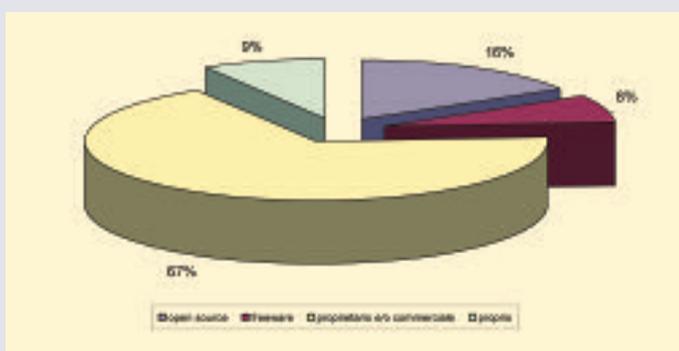


Fig. 6 - Utilizzo % delle diverse tipologie di software oggetto del questionario

I dati relativi alle risorse di calcolo intensivo (Fig. 7) indicano come l'azienda, a differenza di altri centri di ricerca, non si rivolga mai all'esterno per svolgere le proprie simulazioni.

Il tempo di calcolo a noleggio non rappresenta una risorsa utilizzabile forse per motivi di riservatezza. Un'unica azienda che contribuisce nella percentuale globale per circa il 3% svolge attività di calcolo intensivo mediante il *GRID computing*, che prevede lo svolgimento dell'elaborazione su risorse note e fidate che potrebbero trovarsi sempre all'interno dell'azienda ma in edifici diversi. Non sorprende, visti i progressi in ambito di potenza computazionale, che tutte le aziende facciano uso anche dei personal computer per svolgere le simulazioni necessarie. La stessa motivazione, associata al costo relativamente contenuto, può essere adottata per spiegare l'ampia diffusione, 7 aziende su 10, dei cluster dedicati al calcolo scientifico che, spesso, sono un aggregato di personal computer. Motivazioni opposte possono, invece, spiegare la scarsa diffusione dei super computer interni. Al giorno d'oggi tali architetture sono richieste solo per particolari applicazioni (es. *ab initio*) che comunque possono essere svolte, magari in tempi più lunghi ma con costi inferiori anche di 100 volte, sui cluster.

Nell'ambito dell'infrastruttura hardware il panorama descritto dalle aziende riguardo alla diffusione di architetture a 32bit e 64bit rivela un sostanziale equilibrio, specchio dell'odierno mercato (Fig. 8).

Nel passato anche recentissimo (2/3 anni addietro) l'architettura a 64bit era riservata al calcolo mentre quella a 32bit all'analisi dei risultati. La prima, infatti, rappresentava una scelta obbligata sia per simulazioni che necessitavano di allocare più di 2,5 GB di memoria per processo, sia perché alcuni software proprietari erano disponibili solo per siffatto hardware. Oggi, grazie alla diffusione di sistemi operativi *open source* UNIX-like ed al *porting* su tali piattaforme dei software più diffusi, è possibile utilizzare anche un'architettura a 32bit per scopi che prima le erano preclusi.

D'altro canto, e qui sta l'equilibrio di cui sopra, i maggiori produttori hardware, AMD®, per prima, con i processori Opteron, e, in seguito, Intel® con Itanium e Xeon, stanno spostando il loro *target* commerciale su architetture a 64bit dal costo contenuto. Sulla base di quanto affermato, oggi è possibile disporre di una tecnologia a 64bit altamente performante anche su comuni personal computer. Allo stato attuale è facile prevedere una sempre maggiore diffusione del 64bit che, grazie anche alla retrocompatibilità, nei prossimi anni surclasserà definitivamente il 32bit (il cui sviluppo già oggi è da considerarsi concluso, quanto meno nell'ambito del calcolo scientifico).

Per quanto concerne l'utilizzo delle risorse hardware deputate all'analisi grafica dei risultati, possiamo riportare (Fig. 9) un netto predominio dei sistemi UNIX/Linux che può essere attribuito in parte alla diffusione delle workstation proprietarie che fino a pochi anni addietro venivano equipaggiate con sistemi UNIX e che oggi sono fornite con Linux, ed in parte all'installazione post-vendita del sistema operativo Linux sui personal computer.

Ad ogni modo, non può essere trascurata la presenza delle workstation Microsoft Windows® la cui diffusione in parte è dovuta a motivi commerciali ed in parte al globale miglioramento che tale sistema operativo ha incontrato nel corso degli anni. È, infine, sorprendente osservare lo scarso utilizzo, ristretto ad 1 sola azienda, delle workstation Apple® Mac OS/OSX.

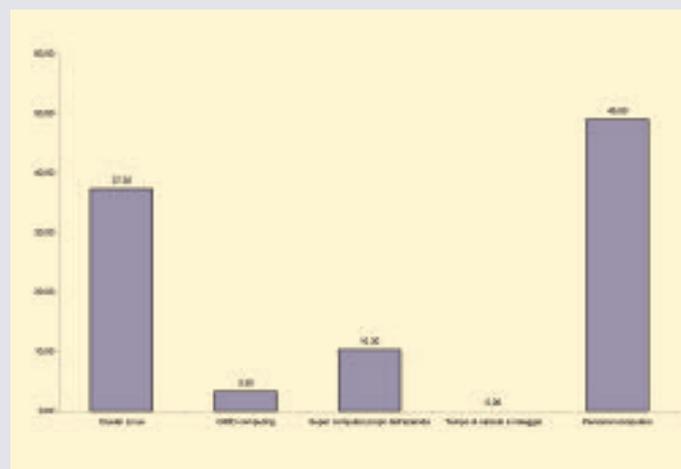


Fig. 7 - Utilizzo % delle risorse di calcolo intensivo oggetto del questionario

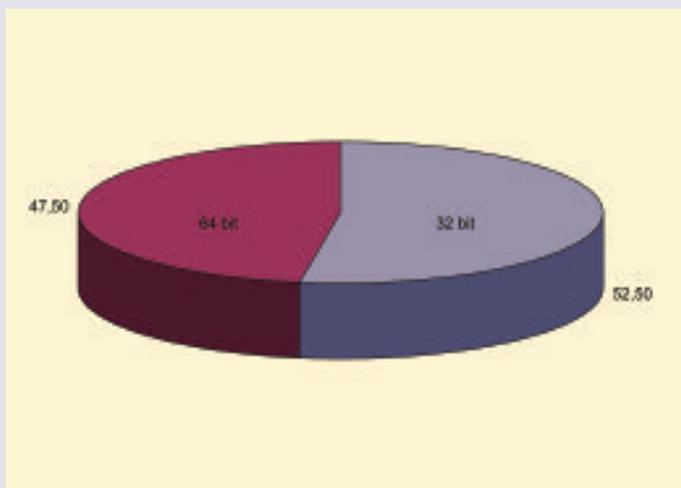


Fig. 8 - Diffusione % delle architetture hardware nel nostro campione di aziende

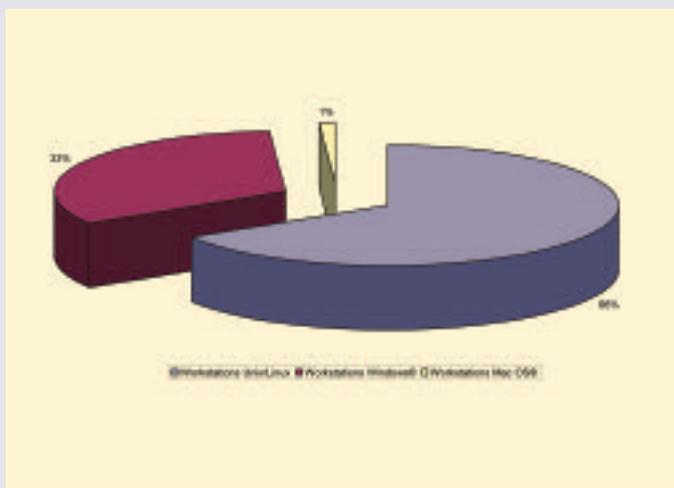


Fig. 9 - Distribuzione % delle risorse hardware riservate all'analisi grafica dei dati

Conclusioni

L'indagine statistica è stata ispirata dalla volontà di valutare la diffusione e la rilevanza delle tecniche di chimica computazionale in ambito industriale. A tale scopo è stato creato un apposito questionario che, nel pieno rispetto della privacy, è stato sottoposto ad un campione rappresentativo di 10 aziende farmaceutiche italiane localizzate in aree ad alta densità industriale. Preziosa e fondamentale si è rivelata la collaborazione con il Direttivo della Divisione di Chimica Farmaceutica che ha sostenuto l'iniziativa.

I risultati ottenuti dimostrano come, nelle aziende farmaceutiche italiane, i livelli di interesse e di impatto associati a ciascuna delle metodiche computazionali non siano uno specchio fedele dell'utilizzo reale che si fa nella routine. Emerge, infatti, un quadro che vede in particolare sulle tecniche del *docking*, della meccanica molecolare, della 3D-QSAR e delle librerie chimiche il massimo impiego a livello industriale. Probabilmente la tendenza chemioinformatica a catalogare in appositi database il parco molecole dell'azienda sposta fortemente l'uso delle tecniche verso gli strumenti più adatti a sfruttare tali informazioni.

I risultati salienti delle sezioni dedicate a software e hardware affermano l'importanza di competenze su più piattaforme e siste-

mi operativi. Di particolare rilievo risultano le conoscenze delle architetture HPC (*High Performance Computing*), di cluster multiprocessore e di sistemi operativi UNIX-Linux. Risulta inoltre interessante ed indicativo il ricorso a risorse software di tipo *open source* o, in minore misura, *freeware* per la risoluzione di problematiche o aspetti tecnici spesso non adeguatamente affrontati nei prodotti commerciali. Di minore diffusione si è, invece, rivelato l'utilizzo di codici di calcolo di produzione interna all'azienda.

Ne consegue quindi che il chimico farmaceutico computazionale, interessato all'attività di ricerca a livello industriale in Italia e probabilmente anche fuori frontiera, deve avere alcune ben definite competenze sia teoriche che tecniche per essere inserito nel team che si occupa di *drug discovery & development*. Queste competenze sono importati per la definizione di percorsi formativi adeguati nel curriculum degli studenti dei Corsi di Laurea in Chimica e Tecnologia Farmaceutiche ed in Farmacia, soprattutto per l'alta formazione che oggi risulta indispensabile introdurre in appositi moduli dei Corsi di Dottorato di Ricerca o nei Master attivi presso le Facoltà di Farmacia italiane, dove questa specifica informazione metodologica non è ancora patrimonio culturale a disposizione dei giovani.

Bibliografia

- [1] P. von Raguè Schleyer *et al.*, Encyclopedia of Computational Chemistry, John Wiley & Sons, 1998.
- [2] K.B. Lipkowitz *et al.*, Reviews in Computational Chemistry, Wiley-VCH, volume 1, 1990, volume 25, 2007.
- [3] T. Clark, A Handbook of Computational Chemistry, John Wiley & Sons, 2004.
- [4] V. Hornak, C. Simmerling, *Drug Discov Today*, 2007, **12**, 132.