

Massimo Mabilia

*M. Bacilieri • A. Bassan • L. Broccardo
E. Fioravanzo • S. Moro • L. Sartori • M. Stocchero*

CHEMOINFORMATICA

Soluzioni e strumenti per scienze
e tecnologie biomediche



Springer

SPRINGER BIOMED

CHEMOINFORMATICA

Soluzioni e strumenti per scienze e tecnologie biomediche

Massimo Mabilia

M. Bacilieri • A. Bassan • L. Broccardo •

E. Fioravanzo • S. Moro • L. Sartori • M. Stocchero

CHEMOINFORMATICA

Soluzioni e strumenti per scienze
e tecnologie biomediche

MASSIMO MABILIA
S-IN Soluzioni Informatiche
Vicenza

Serie Springer Biomed a cura di

MARIA RITA MICHELI
Dipartimento di Biologia Cellulare
e Ambientale
Università di Perugia
Perugia

RODOLFO BOVA
Dipartimento di Medicina Sperimentale
e Scienze Biochimiche
Università di Perugia
Perugia

ISBN 978-88-470-2408-3

e-ISBN 978-88-470-2409-0

DOI 10.1007/978-88-470-2409-0

© Springer-Verlag Italia 2012

Quest'opera è protetta dalla legge sul diritto d'autore, e la sua riproduzione è ammessa solo ed esclusivamente nei limiti stabiliti dalla stessa. Le fotocopie per uso personale possono essere effettuate nei limiti del 15% di ciascun volume dietro pagamento alla SIAE del compenso previsto dall'art. 68, commi 4 e 5, della legge 22 aprile 1941 n. 633. Le riproduzioni per uso non personale e/o oltre il limite del 15% potranno avvenire solo a seguito di specifica autorizzazione rilasciata da AIDRO, Corso di Porta Romana n. 108, Milano 20122, e-mail segreteria@aidro.org e sito web www.aidro.org.

Tutti i diritti, in particolare quelli relativi alla traduzione, alla ristampa, all'utilizzo di illustrazioni e tabelle, alla citazione orale, alla trasmissione radiofonica o televisiva, alla registrazione su microfilm o in database, o alla riproduzione in qualsiasi altra forma (stampata o elettronica) rimangono riservati anche nel caso di utilizzo parziale. La violazione delle norme comporta le sanzioni previste dalla legge.

L'utilizzo in questa pubblicazione di denominazioni generiche, nomi commerciali, marchi registrati, ecc. anche se non specificatamente identificati, non implica che tali denominazioni o marchi non siano protetti dalle relative leggi e regolamenti.

Responsabilità legale per i prodotti: l'editore non può garantire l'esattezza delle indicazioni sui dosaggi e l'impiego dei prodotti menzionati nella presente opera. Il lettore dovrà di volta in volta verificarne l'esattezza consultando la bibliografia di pertinenza.

9 8 7 6 5 4 3 2 1

2012 2013 2014 2015

Layout copertina: Simona Colombo, Milano

Impaginazione: Graphostudio, Milano
Stampa: Grafiche Porpora S.r.l., Segrate (MI)
Stampato in Italia

Springer-Verlag Italia S.r.l., Via Decembrio 28, I-20137 Milano
Springer fa parte di Springer Science+Business Media (www.springer.com)

Presentazione

Fino agli sgoccioli del XX secolo un ricercatore impegnato in laboratorio nella sintesi di nuovi composti o nella analisi di campioni chimici si poteva vantare di essere a proprio agio più con l'attrezzatura che con il computer. Questo strumento era, in quel tempo non lontano, spesso considerato solo un'evoluta macchina per scrivere, utile per riportare i risultati degli esperimenti. In realtà la chimica computazionale aveva già una storia alle spalle, e la chemiometria aveva già assunto un ruolo di primo piano nella chimica analitica, ma pratica sperimentale e uso degli strumenti informatici erano visti come ambiti separati, che potevano anche integrarsi, ma a patto di essere affrontati da persone diverse, formate con specifici percorsi culturali. Le cose sono cambiate velocemente, e oggi il tempo speso davanti al computer tende a superare quello dedicato alle attività sperimentali, anche per chi si senta più portato verso queste ultime.

Il processo che ha gradualmente condotto alla necessità di integrare competenze chimiche e informatiche è probabilmente iniziato con la consultazione dei documenti bibliografici e con l'esigenza di disegnare formule e reazioni da inserire in presentazioni o pubblicazioni. Poi, grazie ai progressi degli algoritmi e alla semplificazione dell'interfaccia di programmi dedicati a funzioni utili per il chimico, la separazione dei due ambiti si è progressivamente ridotta, e oggi è impensabile non prevedere, nel bagaglio culturale di una persona interessata alla chimica, la conoscenza di strumenti per catalogare, ricercare e analizzare informazioni sulla struttura dei composti, o per stimare in modo facile e veloce proprietà chimiche di vario interesse. Al contrario, la prospettiva che si intravede in questo momento è quella di un'integrazione sempre più completa ed efficiente di diversi tipi di informazione chimica, con lo sviluppo di modelli predittivi per proprietà sempre più complesse. Integrazione e sviluppo che consolideranno ulteriormente il ruolo funzionale della chemoinformatica nelle situazioni in cui il chimico si trovi a prendere decisioni importanti per i propri progetti. È pertanto più che mai necessario prevedere, nei processi di formazione per ricercatori, tecnici e personale delle autorità regolatorie, momenti specificamente dedicati all'apprendimento delle basi della chemoinformatica e alle sue principali applicazioni.

Semplificando una situazione estremamente sfaccettata, gli strumenti che la chemoinformatica mette a disposizione si possono suddividere in tre gruppi, caratterizzati rispettivamente dall'impiego dell'analisi statistica, della simulazione e della gestione dei dati. Questi strumenti possono essere utili nella progettazione di composti finalizzata a ottimizzarne certe proprietà, nell'interpretazione

di dati analitici, nella previsione del destino metabolico o dell'impatto ambientale di composti chimici, oppure nella catalogazione e classificazione di grandi elenchi di sostanze. Ciò che contraddistingue i metodi chemoinformatici è la gestione razionale di diversi tipi di informazione, legati alla natura chimica di composti o miscele, in senso estremamente lato. L'informazione chimica, la cui complessità rende necessario l'impiego di mezzi informatici per gestirla, può essere costituita da semplici stringhe che riportano in forma concisa la struttura dei composti (atomi e loro connettività), così come da grandi matrici di dati che ne descrivono le proprietà spaziali; oppure da valori numerici relativi a proprietà dei composti, sia di origine sperimentale che derivanti dall'applicazione di algoritmi all'informazione strutturale; oppure da codici che permettono di stabilire relazioni tra i campi di banche dati contenenti documenti di vario genere. In questo ambito i processi di costruzione, valutazione e impiego di modelli costituiscono il momento centrale dell'impiego di tutta questa informazione, almeno per quanto riguarda le relazioni tra la struttura dei composti chimici e le loro proprietà. Generalizzando, si potrebbe affermare che gran parte del lavoro scientifico consiste nella creazione e formalizzazione di modelli.

Un modello costituisce una rappresentazione della realtà funzionale all'analisi, all'estrapolazione e alla progettazione. In genere, qualunque processo decisionale si serve dell'applicazione di modelli alle informazioni disponibili, per desumere o convalidare delle ipotesi relative ai sistemi che si stanno studiando e inferire le possibili conseguenze di un intervento sul sistema stesso. Per esempio, ci serviamo di modelli quando cerchiamo di prevedere la tossicità di una particolare ammina aromatica, o quando analizziamo lo spettro di massa di un composto sottoposto all'azione metabolica di un preparato microsomiale. A volte il modello è il risultato dell'analisi statistica di dati conosciuti a priori, e il suo impiego consiste nell'applicare a nuovi composti una funzione calibrata su un insieme opportunamente scelto di informazioni. In questo caso è vitale non solo disporre di strumenti statistici adatti ai dati di natura chimica, tipicamente complessi, multidimensionali e ridondanti, ma anche valutare con grande attenzione l'utilità, la rilevanza e l'ambito di applicazione degli elementi del modello (funzioni e parametri) che sono stati scelti nella fase di calibrazione. Non si insisterà mai a sufficienza sull'importanza culturale, ma anche pratica, della fase in cui si selezionano le informazioni necessarie per costruire il modello, valutandone l'impatto sulla attendibilità delle inferenze che ne risultano; questa fase è comunemente nota, nel mondo chemoinformatico, con l'espressione "disegno sperimentale".

In certi casi, invece, la proprietà che si desidera stimare può essere desunta dall'informazione strutturale di una certa entità chimica (può trattarsi di una singola molecola, ma spesso di un insieme di molecole, per esempio una proteina, un farmaco e del solvente) applicando leggi fisiche più o meno approssimate allo studio dei possibili stati del sistema. Si parla allora di simulazione, che è l'azione più comunemente applicata nel campo della cosiddetta modellistica molecolare. In realtà esiste un nesso indissolubile tra modelli e simulazione, nel senso che quest'ultima consiste sempre nell'applicazione di modelli, per cui, a dispetto delle evidenti differenze tecniche tra QSAR, stima di proprietà chimico-fisiche e

modelli molecolari, il loro impiego prevede analoghe problematiche, che è utile e opportuno trovare trattate in un unico volume, con linguaggio facilmente comprensibile e grande attenzione agli aspetti pratici.

Affinché l'impiego di un modello sia produttivo, è necessario che esso permetta di cogliere le caratteristiche essenziali del sistema reale cui si riferisce, in modo efficace ed efficiente. In caso contrario l'utente troverà il modello fuorviante, inutile, pleonastico, o eccessivamente oneroso in termini di risorse spese per costruirlo ed applicarlo. Egli potrebbe così cadere nella tentazione di generalizzare, confondendo l'uso errato di uno strumento con la scarsa utilità di investire in risorse informatiche. D'altra parte, per poter costruire, analizzare o applicare i modelli implementati nel software disponibile a problemi di natura chimica, è necessaria la conoscenza delle radici teoriche su cui essi sono basati. Con lo sviluppo delle interfacce grafiche, è oggi relativamente facile per chiunque usare un programma per l'analisi multivariata, la stima di proprietà o la modellistica molecolare. Può essere più difficile reperire facilmente una descrizione chiara, completa ed accessibile della logica e delle basi che sottendono al loro impiego, e che è necessario considerare per valutare il campo di applicazione dei programmi in uso. Data la numerosità e la eterogeneità degli strumenti chemoinformatici, nella letteratura didattica e scientifica è facile trovare trattati completi, ma difficili da leggere dall'inizio alla fine, o monografie ben strutturate, ma dedicate ad applicazioni specifiche. Non è altrettanto facile trovare opere di consultazione che forniscano una panoramica sulle possibili applicazioni della chemoinformatica, senza cadere nella tentazione di concentrarsi sulle tecniche più recenti trascurando le nozioni di base, necessarie per chi si avvicina da neofita a questo campo, ma utili anche a chi si è abituato ad un uso piuttosto acritico degli strumenti disponibili.

Il presente testo è un riuscito tentativo di colmare questa lacuna, particolarmente apprezzabile dai lettori di lingua italiana. I primi tre capitoli introducono in maniera graduale il lettore all'analisi statistica di dati chimici, alla costruzione e all'impiego di modelli predittivi; il quarto capitolo, dedicato alla stima di proprietà chimico-fisiche, appare come complemento e conseguenza di quanto esposto nei precedenti. Il quinto capitolo, sulla modellistica molecolare, fornisce una panoramica sulle tecniche più comunemente utilizzate, senza entrare in dettagli che andrebbero cercati in un testo di chimica computazionale, ma focalizzandosi sulla meccanica molecolare e sul suo impiego nella progettazione di farmaci. Il lettore viene indirizzato a eventuali approfondimenti tramite un'aggiornata bibliografia. L'ultimo capitolo chiude il cerchio e, come spesso accade nelle opere ben congegnate, invoglia a rileggere gli altri con una rinnovata ottica, dedicando una particolare attenzione alla relazione tra l'informazione chimica, così come è raccolta nelle banche dati virtuali, e i modelli tramite i quali si cerca di analizzare, interpretare o prevedere il comportamento dei sistemi chimici.

Parma, novembre 2011

Marco Mor
Professore Ordinario di Chimica Farmaceutica
Università degli Studi di Parma

Prefazione

“Chemoinformatica. Soluzioni e strumenti per scienze e tecnologie biomediche” è stato pensato, coordinato e scritto da “addetti ai lavori” nel settore chemoinformatico, con la fondamentale supervisione dei curatori della collana Springer Biomed. Benché il prefisso “Chem” nel titolo possa far pensare altrimenti, questo volume non è rivolto innanzitutto a chimici, i quali hanno già a disposizione riviste, pubblicazioni, corsi e congressi relativi alle varie aree di applicazione della chemoinformatica, bensì a studenti, docenti e professionisti che desiderino arricchire le proprie conoscenze in questo settore.

Scopo dell’opera, nel suo insieme, e di ogni suo capitolo è offrire al lettore una chiave di accesso e una facilitazione per affrontare specifici argomenti o applicazioni. Al termine di ogni capitolo il lettore sarà in grado di decidere se e come approfondire ulteriormente un tema di interesse, grazie ai numerosi riferimenti bibliografici. Benché ogni capitolo possa essere considerato a sé stante e possa quindi essere letto indipendentemente dagli altri, la sequenza segue un percorso logico.

Ogni ambito settoriale e specializzato di un particolare sapere scientifico e tecnologico sviluppa necessariamente una sua terminologia; nel caso specifico, il “gergo” che ne deriva è contaminato da termini in lingua inglese e da numerosi acronimi e sigle: i vari autori hanno cercato di limitare al minimo i termini stranieri – e comunque di tradurli ove opportuno – e di spiegare al lettore gli acronimi più criptici; al tempo stesso ci è parso opportuno mantenere nella lingua originale, cioè l’inglese americano, quei termini che sono quasi intraducibili nella nostra lingua o che comunque sono entrati nell’uso comune anche nella comunità dei “chemoinformatici” italiani: pertanto, è bene che il lettore li apprenda così.

La chemoinformatica, in generale, ha sicuramente come oggetto di indagine preferenziale la chimica e, in particolare, utilizza modelli, proprietà e descrittori relativi a strutture molecolari, siano esse molecole organiche, peptidi, proteine o acidi nucleici o altri tipi di molecole. È altrettanto vero che storicamente i primi utilizzatori di queste “tecniche assistite da calcolatore” sono stati dei chimici. Ma come emerge già dalla lettura di questa prefazione, la chemoinformatica è un settore multidisciplinare e soprattutto interdisciplinare, in cui cioè non solo sono presenti, ma concorrono e si intrecciano varie scienze e conoscenze: matematica (algebra lineare, calcolo differenziale, topologia, etc.), chimica, chimica-fisica, fisica (meccanica classica e quantistica,

termodinamica, etc.), statistica, informatica e altre scienze a seconda delle aree di applicazione: chimica farmaceutica, farmacologia, biologia molecolare, tossicologia, fisiologia molecolare, etc., e le relative figure professionali.

Una definizione vera e propria, chiara e distinta del termine “chemoinformatica” non è possibile, poiché tale termine si riferisce a un settore “aperto” che (come molti altri) è in continua, rapida evoluzione ed espansione: definire significa sì chiarire, ma anche porre un confine, dei limiti, e questo contraddirebbe quanto appena affermato. È però possibile proporre delle descrizioni e, seppur succintamente, alcuni elementi storici che possano aiutare il lettore a inquadrare meglio ciò che si intende con il suddetto termine e che cosa esso implichi da un punto di vista conoscitivo e operativo.

“Chemoinformatica” (talvolta scritto “Cheminformatica” e, più raramente, “Chemioinformatica”) è un termine relativamente recente, composto dai nomi di due discipline, la chimica (che ha lo stesso etimo di alchimia, dall’arabo *al chema* cioè “il segreto”) e l’informatica (contrazione di informazione automatica). Una delle prime occorrenze pubbliche del termine “chemoinformatica”, forse la prima, risale a Brown, nel 1998 e recita: “L’uso della gestione e tecnologia della informazione (Information Technology and Management) è divenuto una componente essenziale del processo di scoperta di un farmaco (*drug discovery*). La chemoinformatica è l’insieme di quelle risorse atte a trasformare dati in informazioni e informazioni in conoscenza, con lo scopo di prendere decisioni migliori più velocemente nell’identificazione e ottimizzazione di nuovi potenziali farmaci” (Brown, 1998).

Quindi, il termine “chemoinformatica”, che potremmo anche designare “informatica chimica”, è recente, come abbiamo visto, ma la storia della collaborazione sinergica fra informatica e chimica vanta una lunga storia e tradizione, addirittura anteriore alla nascita e all’utilizzo del termine informatica!

Gli studi di QSAR (relazione quantitativa fra struttura chimica e attività biologica) iniziano nel XIX secolo; i primi calcoli di meccanica quantistica risalgono agli anni ’20 del secolo scorso; negli anni ’30 viene sviluppato il concetto di relazioni lineari di energia libera; i primi calcoli di meccanica molecolare vengono compiuti negli anni ’40.

Negli stessi anni, durante la Seconda Guerra Mondiale, nasce Eniac, capostipite della prima generazione di calcolatori elettronici: 30 tonnellate, 17mila valvole: non si andava al calcolatore, ma si “entrava” nel calcolatore! Si sperimentano le prime applicazioni dei transistor che vanno a sostituire le vecchie valvole. Con la terza generazione di calcolatori, a partire dagli anni ’60 e dagli USA, e in particolare grazie all’avvento dei terminali in sostituzione delle schede perforate, molti più utenti, fra cui studenti, ricercatori e docenti di diverse discipline (fra queste chimica teorica e meccanica quantistica) possono accedere direttamente alle risorse di calcolo, senza dover perforare schede e consegnarle all’operatore, gestendo così direttamente i vari programmi software e l’immissione di dati (input) e potendo comodamente controllare il risultato dei calcoli (output).

La Legge empirica di Moore osserva che da oltre cinquanta anni – e preve-

de che per almeno altri dieci – il numero di transistor che possono essere collocati su un circuito integrato raddoppia circa ogni due anni. La velocità di una CPU (central processing unit), le prestazioni della grafica computerizzata e molte altre capacità di strumenti elettronici sono fortemente correlati alla Legge di Moore. Questa crescita esponenziale ha aumentato in modo eclatante l'impatto e la diffusione dell'elettronica in ogni settore dell'economia, ricerca e sviluppo mondiale e, va da sé, in modo particolare in quelle applicazioni che sono nate e dipendono dalle prestazioni dell'hardware.

Nel 1962 all'Università dell'Indiana a Bloomington nasce il Quantum Chemistry Program Exchange e, negli stessi anni, si moltiplicano le pubblicazioni e gli incontri scientifici che nel titolo contengono "Computer programs for chemistry" o "Computer applications in chemistry" o ancora "Computational chemistry". Negli anni '70 nasce la chemiometria: il termine è stato coniato da Swante Wold ed è stato da lui definito come "l'arte di estrarre informazioni chimiche pertinenti da dati prodotti da esperimenti chimici, in analogia con biometria, econometria, etc. utilizzando modelli matematici e statistici" (S. Wold 1995); vengono inoltre scritti i principali algoritmi, tuttora usati in molti programmi per l'analisi multivariata dei dati, che appartengono alla chemoinformatica. Agli stessi anni risalgono la maggior parte degli indici topologici e descrittori molecolari ancora oggi usati.

A cavallo fra anni '70 e '80 vengono sviluppati i primi sistemi di software (insiemi di programmi, procedure, algoritmi e grafica molecolare) presso alcune università negli USA. Nascono, spesso ad opera degli stessi professori universitari talvolta con finanziatori esterni, le prime società per commercializzare prodotti software "per la chimica". I sistemi di grafica molecolare si sviluppano velocemente ed evolvono in un nuovo strumento: la stazione grafica (graphics workstation) intesa come integrazione a livello hardware di mini- o micro-computer e sistemi grafici. Una delle prime società nel settore, oggi definito "Chemoinformatica", già agli inizi degli anni '80 distribuiva un sistema di *computer-aided molecular modeling / design* (CAMM/CAMD) e dei database per strutture e reazioni chimiche. Uno dopo l'altro, tutti i principali laboratori di ricerca farmaceutica, a partire dagli USA, poi in Gran Bretagna e nel resto dell'Europa e quindi anche in Italia, creano un gruppo di *computer-aided drug design & discovery* (CADD).

Perché quella che oggi chiamiamo "chemoinformatica" si è sviluppata inizialmente e preferenzialmente nel settore farmaceutico? Basti pensare che una multinazionale del farmaco per arrivare a registrare un nuovo farmaco e avere una molecola di backup, porta in media 5 composti nei test clinici (clinical trials) ma per identificare quelle 5 molecole ne ha valutate fra 50mila e 100mila. I fattori "tempo" e "costi" sono determinanti: di conseguenza, ogni metodo o procedura che permetta di ridurre i tempi e/o di abbattere i costi per arrivare prima in fase clinica viene impiegato.

La gamma di algoritmi, metodi e strumenti di calcolo continua ad espandersi fino a raggiungere la "maturità" verso la fine del XX secolo e gli inizi del nuovo. Non sarebbe allora un caso che proprio in quegli anni venga coniato e

utilizzato il termine “chemoinformatica”: ma questa rimane una ipotesi. A differenza di un gruppo di *computer-aided drug design* degli anni '80 che poteva contare su risorse hardware e software limitate rispetto ad oggi, un gruppo degli anni '90 o un gruppo di “chemoinformatica” di oggi ha generalmente a disposizione una vasta gamma di soluzioni software: programmi per la progettazione di esperimenti (*design of experiments*); programmi per generare indici e descrittori molecolari topologici e 3D; analisi statistiche multivariate (MVA – multivariate data analysis) ed eventuali altre soluzioni per studiare relazioni quantitative struttura-attività (QSxR, Quantitative Structure – Activity / Property – Relationship); vari programmi e procedure basate su calcoli di meccanica quantistica ma soprattutto meccanica molecolare per studi di analisi conformazionale, dinamica molecolare, ricerca di farmacoforo (*ligand-based design*), per simulare l'interazione farmaco-ricettore (*structure-based design / docking studies*); procedure per lo screening virtuale anche di milioni di molecole (virtual screening); software per la predizione di proprietà chimico-fisiche (quali acidità, lipofilia e solubilità) e proprietà ADMET (assorbimento, distribuzione, metabolismo, escrezione, tossicità); soluzioni per la predizione e simulazione di spettri (UV-Vis, IR, NMR, MS, etc.); strumenti informatici per l'archiviazione, gestione e ricerca di dati alfanumerici, chimico-strutturali (per struttura, sottostruttura, similarità, etc.), chimico-analitici, spettrali (per spettro, regione spettrale, picco, etc.); soluzioni specifiche per il data mining, nonché una gamma inesauribile di strumenti per la generazione, gestione e ricerca (electronic database management) di documenti, relazioni, quaderni di laboratorio, etc.

Calcoli, simulazioni e ricerche in database compiute su calcolatori relativamente costosi e che richiedevano tempi proibitivi una decina di anni fa, oggi possono essere condotti in tempi ragionevoli su un portatile. Di conseguenza, calcoli, simulazioni e ricerche via via più complessi possono essere condotti utilizzando *graphics workstation* multiprocessore, *cluster* (letteralmente “grappolo”) di computer connessi tra loro (*cluster computing*), oppure usando infrastrutture di calcolo distribuito (*grid computing*) o un insieme di tecnologie che permettono l'utilizzo di risorse distribuite in rete (*cloud computing*).

Hardware sempre più veloci, l'accumulo di esperienze e lo sviluppo continuo di nuovi algoritmi gradualmente hanno migliorato e continuano a migliorare il livello di precisione di molte proprietà calcolate rispetto ai valori sperimentali, al punto che alcune proprietà chimico-fisiche non vengono più misurate sperimentalmente, ma predette (in silico) a livello di calcolo a una frazione del costo sperimentale (e talvolta gratuitamente on-line) e in pochi secondi per molecola.

Interfacce grafiche (GUI – Graphical User Interface) sempre più accattivanti e relativamente semplici rendono accessibili strumenti di calcolo sofisticati e complessi a un numero crescente di utenti, sia esperti che occasionali. Un pericolo è in agguato: per pigrizia, per cultura o per necessità le soluzioni software possono diventare una “scatola nera” (black box): si fa “click” su alcuni bottoni, si inseriscono alcuni dati e si ottiene sempre un risultato, senza

conoscere né apprezzare cosa sia successo “dentro”. Lascio al lettore ogni ulteriore riflessione, semplicemente ricordando come monito la sigla GIGO: *garbage in, garbage out!*

Ma questa “chemoinformatica” dà certezze di successo? In primo luogo, il successo dovuto all'utilizzo di tecniche computer-assisted in campo farmaceutico (quello storico e più “collaudato”) deve essere valutato sul numero di hit (composti con attività o affinità al di sopra di un valore di soglia fissato) e sul numero di lead (potenziali farmaci) identificati e non sul numero di farmaci effettivamente registrati. In secondo luogo, è bene tenere presente che un gruppo di chemoinformatica è costituito da: hardware, software e “cervelli umani”. Questi ultimi sono da tutti i punti di vista la componente più importante, e comunque condizione necessaria, ma non sufficiente per il successo. Altre condizioni necessarie perché la chemoinformatica sia uno strumento efficace sono dati dalla cultura e dall'organizzazione della ricerca in cui strumenti e gruppo chemoinformatici sono inseriti, e dalla scelta strategica degli obiettivi (target) e dal grado di conoscenza dei target in termini di struttura molecolare e meccanismo biochimico. Di fatto, va osservato che il numero di nuovi farmaci sviluppati a livello mondiale per anno è in graduale diminuzione: questa situazione meriterebbe una riflessione particolare, ma queste e altre considerazioni collegate esulano dagli scopi di questo volume. Fatte queste precisazioni, ci sono comunque dei successi (nell'accezione sopra menzionata) raccontati ai congressi e talvolta riportati in letteratura; non sempre però sono a disposizione tutte le informazioni e gli elementi per discernere in modo certo i successi reali, dovuti al contributo specifico della chemoinformatica, dalle “operazioni di immagine”. Quando certi prerequisiti sono soddisfatti (qualità degli strumenti di calcolo e capacità degli operatori) e le condizioni di contorno lo permettono (efficaci ed efficienti organizzazioni di ricerca) l'utilizzo sapiente di queste tecniche permette di generare ed eliminare ipotesi in tempi rapidi e di aumentare in modo significativo la probabilità di “successo”.

Qual è il significato del termine “chemoinformatica” oggi? Seguendo l'indicazione di Ludwig Wittgenstein, filosofo del linguaggio del secolo scorso, per il quale “il significato di una parola è il suo uso nel linguaggio”, scopriamo che ai nostri giorni il termine “chemoinformatica” è utilizzato in un numero crescente di situazioni e da un numero crescente di figure professionali. Infatti, le aree di applicazione della chemoinformatica si sono gradualmente e successivamente estese non solo a tutti gli aspetti del processo di ricerca e sviluppo di nuovi farmaci, ma interessano anche molti altri settori, dall'agro-alimentare alla cosmetica, dalla biologia molecolare alla scienza dei materiali, dalla tossicologia ambientale alle scienze biomediche; in queste ultime le sinergie fra bioinformatica e chemoinformatica, in particolare fra genomica e proteomica, sono notevoli e in continuo sviluppo. Pertanto, non è casuale che del termine chemoinformatica siano state fornite descrizioni più ampie e generali, ad esempio (Gasteiger, 2003): “La chemoinformatica è l'uso di metodi informatici per risolvere problemi chimici”. Descrizione decisamente gene-

rare, ma che ha il merito di essere semplice, di rispondere al vero e di indicare il potenziale utilizzo degli strumenti, dei metodi e delle procedure chemoinformatiche in quei settori e sicuramente nelle scienze e tecnologie biomediche, i quali richiedano conoscenze anche di natura chimica e utilizzino come oggetti di ricerca e analisi strutture molecolari, dati e informazioni ad esse collegati e da esse derivati.

Massimo Mabilia

Elena Fioravanzo

Ringraziamenti

Il primo ringraziamento, sia doveroso che spontaneo, va ai due curatori della Collana Springer BioMed, la Dott.ssa Maria Rita Micheli e il Dott. Rodolfo Bova dell'Università degli Studi di Perugia, per aver intuito e caldeggiato l'opportunità di un volume dedicato alla Chemoinformatica.

Un ringraziamento va a ciascuno degli autori che hanno contribuito alla stesura dei vari capitoli, per aver messo a disposizione del lettore la propria esperienza professionale, cercando di coniugare al meglio il rigore scientifico con la relativa semplicità richiesta da un'opera divulgativa. Un ulteriore ringraziamento è rivolto ai collaboratori del Prof. Moro dell'Università di Padova, Dott. Marco Fanton, Dott. Matteo Floris, Dott. Giorgio Cozza e Dott. Andrea Cristiani per l'attento ed efficace lavoro di revisione del capitolo 5, ai colleghi (in particolare al Dott. Remo Calabrese) che a vario titolo hanno letto, commentato e offerto suggerimenti relativi a uno o più capitoli.

Senza il coordinamento della Dott.ssa Lorenza Broccardo questo volume non sarebbe arrivato alle stampe! A nome degli autori porgo un sentito ringraziamento.

Massimo Mabilia

Elenco degli Autori

Magdalena Bacilieri

Dipartimento di Scienze Farmaceutiche
Sezione di Modellistica Molecolare
Università degli Studi di Padova

Arianna Bassan

S-IN Soluzioni Informatiche
Vicenza

Lorenza Broccardo

S-IN Soluzioni Informatiche
Vicenza

Elena Fioravanzo

S-IN Soluzioni Informatiche
Vicenza

Stefano Moro

Dipartimento di Scienze Farmaceutiche
Sezione di Modellistica Molecolare
Università degli Studi di Padova

Luca Sartori

IEO – Istituto Europeo di Oncologia
Dipartimento di Oncologia Sperimentale
Unità di Drug Discovery
Milano

Matteo Stocchero

S-IN Soluzioni Informatiche
Vicenza

Indice

Capitolo 1 – Il disegno sperimentale	1
Introduzione alla metodologia	1
Definizione di alcuni termini di uso frequente	2
Metodo classico di sperimentazione e metodo multivariato	4
Il flusso di lavoro	9
Il modello matematico quale rappresentazione della realtà	12
Gli obiettivi di una sperimentazione	15
Obiettivo “Screening”	16
Obiettivo “Ottimizzazione”	16
Obiettivo “Test di robustezza”	16
Tipologie di piani sperimentali	17
Piani fattoriali completi	17
Piani fattoriali frazionari	18
Piani fattoriali composti	20
Relazioni tra piani fattoriali completi, frazionari e composti	21
Formulazione del problema	23
Definizione degli obiettivi	24
Identificazione delle risposte	24
Scelta dei fattori	24
Identificazione della strategia	25
Analisi dei dati e stima del modello	26
Ottimizzazione delle condizioni di crescita del lievito	
<i>Pachysolen tannophilus</i>	29
Conclusioni	32
Lecture consigliate	34
 Capitolo 2 – Analisi statistica multivariata di dati	 37
Introduzione	37
Tabelle di dati	38
I metodi proiettivi	38
Principali classi di problemi che possono essere affrontati	
con i metodi proiettivi	40
Correlazione e causalità	42
Proiezione e tabelle di dati	43

Modello e decomposizione della tabella di dati	46
Vantaggi nell'uso dei metodi proiettivi	47
Tecniche unsupervised e supervised	48
Scaling e centratura delle variabili	48
Analisi delle componenti principali (PCA)	50
Quando utilizzare la PCA	50
Note tecniche	51
Interpretazione dei modelli mediante grafici	52
Metodo della classificazione SIMCA (Soft Independent Modeling of Class Analogy)	56
Quando utilizzare la tecnica SIMCA	56
Note tecniche	56
Interpretazione dei modelli mediante grafici	57
Metodo di regressione PLS (Projections to Latent Structures by Partial Least Squares)	59
Quando utilizzare la PLS	59
Note tecniche	59
Interpretazione dei modelli mediante grafici	62
Metodo di classificazione PLS-DA (<i>PLS-Discriminant Analysis</i>)	64
Quando utilizzare la PLS-DA	64
Note tecniche	65
Interpretazione dei modelli mediante grafici	66
Lecture consigliate	69

Capitolo 3 – Relazioni quantitative struttura-attività/proprietà

Introduzione	71
Cosa si intende per modello struttura-attività/proprietà	71
I descrittori molecolari	75
Definizione di descrittore molecolare	76
Alcune famiglie di descrittori molecolari	78
Analisi della matrice dei descrittori	87
Tecniche multivariate per la costruzione dei modelli	89
Reti neurali artificiali (Artificial Neural Networks)	89
Classificatore naïve bayesiano	90
Partizione ricorsiva (<i>Recursive Partitioning</i>)	91
Applicazioni	91
Studio della permeabilità della barriera emato-encefalica ...	91
Costruzione di modelli indipendenti per la predizione della inibizione hERG	95
Lecture consigliate	101

Capitolo 4 – Predittori di pKa, lipofilia e solubilità	103
Introduzione	103
pKa di un composto chimico	103
Modelli per predire in silico la pKa	105
Lipofilia di un composto chimico	107
Lipofilia in funzione del pH	109
Modelli per predire in silico la lipofilia	110
Solubilità acquosa di un composto chimico	113
Modelli per predire in silico la solubilità acquosa	114
Come migliorare l'accuratezza delle predizioni	115
Alcuni predittori freeware disponibili via web	116
Letture consigliate	117
 Capitolo 5 – Modellistica molecolare	119
Virtualizzazione della struttura molecolare	119
Rappresentazione computazionale delle strutture molecolari	119
Banche dati di strutture molecolari	122
<i>Ligand-based drug design</i>	122
Modello farmacoforico	123
<i>Structure-based drug design</i>	125
Meccanica molecolare e campo di forza	125
<i>Docking</i> molecolare	131
Predizione della struttura terziaria di una proteina attraverso la tecnica dell' <i>homology modeling</i>	134
Identificazione dei templates e allineamento delle sequenze ..	136
Costruzione di un modello tramite la tecnica di <i>homology modeling</i>	137
Validazione dei modelli	138
Dinamica molecolare classica	139
Simulazioni di dinamica molecolare	139
Limitazioni e tecniche innovative	142
Letture consigliate	143
 Capitolo 6 – Archiviazione e analisi di dati di tipo chimico	145
Introduzione	145
Banca dati (database)	148
Esempi di banche dati	151
Database pubblici	152
Database commerciali	152
Database specifici per il campo chimico-farmaceutico	153
Quaderno di laboratorio elettronico	156

Analisi dei dati (<i>Data Mining</i>)	158
Conclusioni	161
Lecture consigliate	162
 Glossario	 163
 Indice analitico	 181

CAPITOLO 1

Il disegno sperimentale

Lorenza Broccardo

Introduzione alla metodologia

Il metodo scientifico di incremento della conoscenza di un sistema per il quale non è noto il modello teorico che definisce la relazione tra le sue variabili, prevede le seguenti fasi:

- formulazione delle ipotesi, in base alle informazioni disponibili;
- deduzioni sul comportamento del sistema;
- acquisizione di nuove informazioni mediante l'esecuzione di prove sperimentali;
- analisi dei dati e loro interpretazione;
- verifica delle ipotesi.

Nello studio di sistemi naturali, quali ad esempio i sistemi biologici, l'acquisizione di nuove informazioni e l'analisi dei dati possono costituire due fasi critiche dello sviluppo del processo cognitivo. Ciò è dovuto al fatto che tali sistemi dipendono, generalmente, da più di due variabili: i sistemi naturali sono, cioè, multivariati. La definizione delle relazioni tra numerose variabili implica la necessità di eseguire un numero elevato di prove sperimentali con conseguenti oneri in termini di materie prime, strumentazione e personale impiegato. Inoltre, maggiore è il numero di variabili in esame, maggiore è la probabilità che si verifichino fenomeni di interazione dovuti al ruolo combinato di due o più variabili, non rilevabili osservando l'effetto di una variabile alla volta. La stima delle interazioni è fondamentale per il controllo di un sistema e tuttavia è possibile solo mediante un'appropriata organizzazione delle prove sperimentali. La complessità di un sistema si riflette necessariamente nella complessità dei dati che lo descrivono: per un'adeguata analisi in grado di estrarre l'informazione utile, è necessario disporre di metodi in grado di analizzare tali dati nel loro insieme, di separare l'informazione dal rumore, di gestire la correlazione e di presentare i risultati mediante grafici riassuntivi di facile interpretazione. Gli aspetti critici dovuti alla multidimensionalità di un problema sono superati grazie all'impiego di metodi chemiometrici quali il disegno sperimentale e l'analisi statistica multivariata (la chemiometria può essere definita come