Modelli di relazione quantitativa struttura-attività (QSAR-Quantitative Structure Activity Realtionship):

Modelli di relazione quantitativa struttura-attività (QSAR-Quantitative Structure Activity Realtionship):

descrittori molecolari,

aspetti teorici ed applicativi di classificazione,

regressione e predizione.

Esempi pratici in free software dedicati.

introduzione_alla_chemiometria.pdf (unimib.it)

https://michem.unimib.it/wp-content/uploads/sites/43/2019/04/introduzione_alla_chemiometria.pdf

Introduzione alla chemiometria – R. Todeschini

LE STRATEGIE QSAR

Lo studio delle relazioni tra struttura molecolare e attività di un composto si fonda su un approccio razionale basato sull'**assunzione che esistano certe relazioni tra la struttura molecolare (S) e l'attività biologica (A) dei composti.**

In altre parole, si cerca di determinare la **relazione funzionale f(S, A)** che mette in relazione l'attività A con la struttura molecolare S di un composto.

L'obiettivo finale generale degli studi di relazioni struttura-attività (SAR, Structure-Activity Relationships) è quello di **comprendere i meccanismi dell'azione farmacologica o tossicologica,** suggerendo vie nuove per la sintesi di composti con attività biologica definita.

Le **interazioni farmaco-organismo rappresentano un sistema complesso che coinvolge un grande numero di processi molti dei quali sono sconosciuti**.

L'**effetto globale** dei singoli processi può essere visto come il manifestarsi di un'attività del composto: questo aspetto è la causa della natura statistica delle strategie *SAR*.

Le assunzioni fondamentali degli studi *SAR* sono:

a) a composti simili corrispondono simili proprietà biologiche
b) a modifiche simili nella struttura molecolare corrispondono cambiamenti simili delle proprietà.

Sugli stessi principi generali si basano anche le strategie per lo studio delle relazioni tra struttura molecolare e proprietà chimico-fisiche, note come strategie ***SPR*** (***Structure-Property Relationships***).

Schema generale entro cui si possono inquadrare gli studi delle relazioni attività/proprietà e struttura molecolare. Le funzioni α e μ rappresentano, rispettivamente, le *procedure sperimentali* mediante le quali determiniamo le proprietà biologiche, farmacologiche o tossicologiche e le proprietà chimico-fisiche, di un insieme di composti C:

$$\alpha: C \rightarrow A \qquad \mu: C \rightarrow M$$

$$\alpha(C) = A \qquad \mu(C) = M$$

L'insieme M è costituito da proprietà chimico-fisiche quali, ad esempio, il punto di ebollizione, il punto di fusione, il volume molare, il momento dipolare, la solubilità in acqua, i coefficienti di ripartizione ottanolo/acqua, aria/acqua, sedimento/acqua, i parametri di reattività chimica, ecc. L'insieme A è a sua volta costituito da quantità che sono legate all'attività biologica, farmacologica, tossicologica dei composti studiati.

Le attività biologiche esplicano le loro attività con effetti e a livelli molto diversi tra loro. Si possono, ad esempio, distinguere:

Livello macromolecolare:

forze di legame col recettore, costanti di inibizione, costanti di Michaelis

Livello cellulare:

mutagenicità, trasformazioni cellulari

Livello di organismi (effetti acuti):

risposta bologica di alghe, invertebrati, pesci, uccelli, mammiferi

Livello di organismi (effetti cronici):

neurotossicità ritardata, bioconcentrazione, biodegradazione, carcinogenicità, tossicità sulla funzionalità riproduttiva.

La funzione $\gamma_{3\,r}$ rappresenta l'insieme dei modelli con cui siamo in grado di calcolare le attività biologiche da proprietà chimico-fisiche dei composti:

$$\gamma_3(\mathrm{M}) = \mathrm{A}$$

Per molto tempo la strategia QSAR più frequente (prima dell'attuale riconosciuta importanza dei descrittori molecolari teorici) è stata improntata allaricerca di una correlazione diretta tra proprietà chimico-fische sperimentali e misure sperimentali di attività biologiche

$$\gamma_3 \mu(C) = \alpha(C)$$

In questo caso, evidentemente, non si fa ricorso ai descrittori di struttura molecolare e quindi è sempre assente ogni relazione funzionale sulla struttura molecolare.

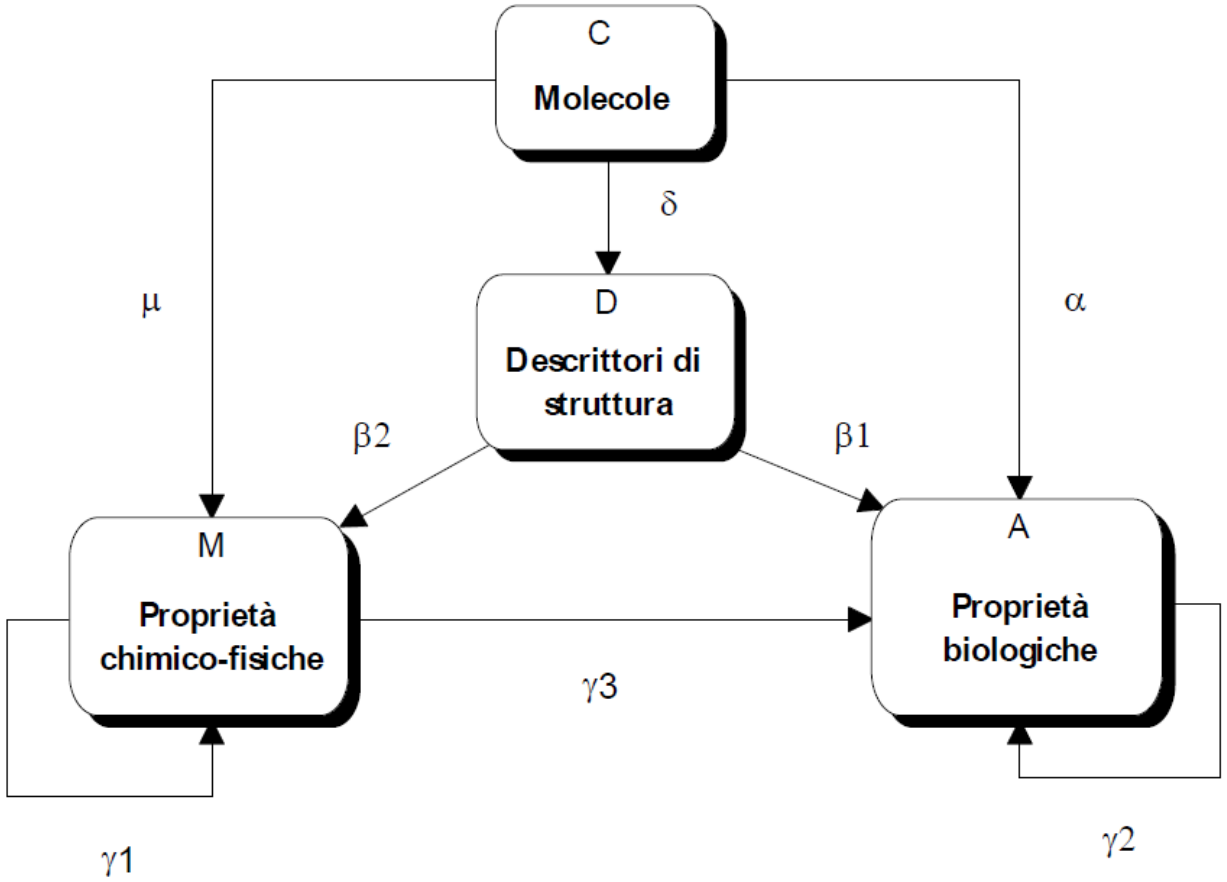
L'utilizzo della funzione γ_1 riflette le strategie secondo le quali si vuole predire una proprietà chimico-fisica da altre proprietà chimico-fisiche:

$$\gamma_1(M) = M'$$

dove M' rappresenta una proprietà in M, diversa dalle proprietà in M utilizzate come descrittori in γ_1 .

Interessante quando la proprietà che si vuole predire è meno conveniente delle proprietà utilizzate perpredire la prima (es. la proprietà misurata è affetta da elevato rumore sperimentale, richiede una procedura sperimentale complessa, costosa, lunga; richiede costi alti per la sua determinazione; non è accessibile per alcune condizioni sperimentali). Analogamente

$$\gamma_2(A) = A'$$



La classe più importante delle strategie *SAR* e *SPR* si basa sulla creazione di funzioni δ in grado di trasformare l'informazione chimica contenuta nelle formule di struttura, nella topologia molecolare, nelle rappresentazioni 3D delle molecole, cioè nei termini delle loro coordinate spaziali x,y,z, in descrittori di struttura (D): $\delta(\mathrm{C}) = \mathrm{D}$

Approccio di Free-Willson

Si tratta di un particolare approccio alla costruzione di modelli di regressione in cui i descrittori sono variabili indicatrici (*dummy variables*) del tipo di sostituente e della sua posizione.
Ad esempio, si consideri il toluene e i due siti di sostituzione adiacenti al metile (posizioni orto o 2 e meta o 3) e supponiamo che i sostituenti siano il fluoro, il bromo e lo iodio. In Tab.12-1 sono riportati cinque possibili composti. La variabile sito1 - F indica la presenza (1) o l'assenza (0) del fluoro in posizione orto rispetto al metile del toluene; la variabile sito1 - Br indica la presenza (1) o l'assenza (0) del bromo in posizione orto rispetto al metile del toluene; e così via.

	sito 1: orto / pos.2			*sito 2: meta / pos.3*		
composto	*F*	*Br*	*I*	*F*	*Br*	*I*
toluene	0	0	0	0	0	0
2-iodo-toluene	0	0	1	0	0	0
3-iodo-toluene	0	0	0	0	0	1
2,3-difluoro-toluene	1	0	0	1	0	0
2-bromo, 3-fluoro-toluene	0	1	0	1	0	0

TAB. 12-1

Il numero di variabili indipendenti è dato dal totale dei gruppi sostituenti considerati per i siti di sostituzione (nell'esempio: 3 sostituenti x 2 siti = 6). Ogni composto è definito da un vettore di zeri e uno e la matrice dei descrittori è quindi interamente costituita da valori zero e uno; il modello di regressione cerca di mettere in relazione la risposta sperimentale Y con le posizioni e il tipo di sostituenti.
Questo tipo di approccio presenta il vantaggio di essere facilmente applicabile e di non dipendere dalla conoscenza di alcuna proprietà chimico-fisica; tuttavia la tipologia delle variabili è tale da presentare normalmente notevoli problemi di predittività.

Sebbene le **prime relazioni tra le proprietà chimico-fisiche delle molecole e la loro attività biologica** si debbano far risalire ai lavori di **Meyer** (1899) e di **Overton** (1901), il grande sviluppo di questo campo deriva dalle ricerche sviluppate da **Hansch e collaboratori** a partire dal **1963**. Il postulato sul quale si basa l'approccio di Hansch afferma che quando una sostanza biologicamente attiva entra in contatto con il sistema molecolare di un organismo vivente, la probabilità che essa raggiunga un sito recettore e induca quindi una determinata risposta sull'organismo - l'effetto - è funzione delle proprietà della sostanza stessa.

sistema biologico + sostanza attiva = risposta

ovvero

$$risposta = f_1(L) + f_2(E) + f_3(S) + f_4(M)$$

dove le quattro funzioni sono rispettivamente funzioni di **proprietà lipofile (L**), di **proprietà elettroniche (E),** di **proprietà steriche (S)** e di eventuali **altre proprietà molecolari (M)** necessarie per una completa descrizione dell'effetto biologico considerato. Tutte queste proprietà sono le proprietà della molecola considerata e la risposta dipende additivamente da esse. La validità di questo postulato è indissolubilmente legata ad un altro postulato, noto come **principio di congenericità**. Secondo questo principio, le strategie *QSAR* sono applicabili soltanto a classi di composti "simili", ove per composti simili si intendono:

a) composti che abbiano uno scheletro-base comune

b) i sostituenti dello scheletro-base differiscano tra loro in modo da non influenzare in modo decisivo le proprietà globali della molecola.

In base ai principi su cui si basa l'approccio di Hansch, lo sviluppo di questa strategia si è decisamente indirizzato verso la parametrizzazione delle proprietà dei gruppi sostituenti (gruppi funzionali, frammenti molecolari), piuttosto che sulle misure delle proprietà di tutta la molecola. Questa impostazione, una volta note le proprietà dei gruppi funzionali, consente un'ampia e facile applicabilità a moltissimi problemi *QSAR*. Tuttavia, appaiono anche evidenti e indiscutibili i limiti di questo approccio che presume la possibilità di modellare le risposte biologiche mediante uno schema lineare puramente additivo utilizzando solo l'informazione locale insita nei gruppi sostituenti, indipendente quindi dalle proprietà globali di ciascuna molecola.

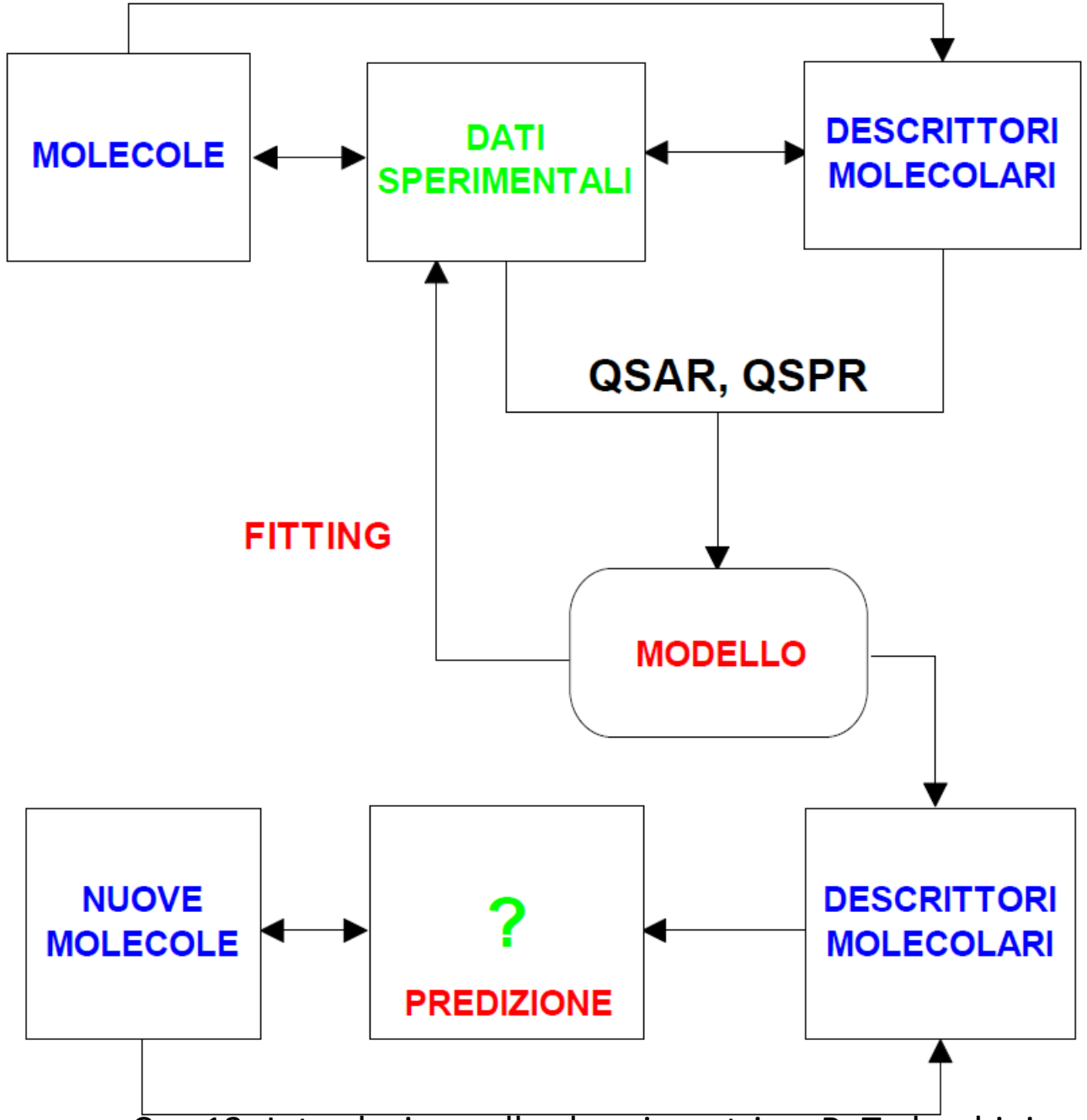
Schema generale che rappresenta la strategia di lavoro tipica nella ricerca delle relazioni attività-struttura molecolare.

Gli studi QSAR (Quantitative Structure-Activity Relationships) sono basati su approcci matematici e statistici con lo scopo di trovare ***modelli quantitativi*** nelle relazioni tra struttura molecolare e attività.

Le strategie principali si fondano comunemente sulle seguenti fasi:

a) una rappresentazione delle strutture molecolari mediante opportuni **descrittori**.
b) la ricerca di **relazioni** quantitative specifiche tra descrittori ed attività (biologica, farmacologica, tossicologica) utilizzando principalmente l'analisi statistica multivariata e i metodi chemiometrici.
c) la **predizione dell'attività di nuovi composti con una struttura predefinita** utilizzando i modelli matematici trovati.

Anche QSPR,QSRR, QSTR…



Chemiometria e modelli QSAR

Lo sviluppo della chemiometria ha messo in luce delle nuove potenzialità nello sviluppo dei modelli QSAR.

In primo luogo, la logica della validazione consente di sviluppare modelli anche complessi (presenza di molte variabili, modelli non-lineari) per i quali è possibile valutare realisticamente la qualità predittiva. Accanto alle procedure di validazione, il cui scopo è quello di trovare la complessità ottimale del modello in grado di fornire il massimo potere predittivo, in ***molti metodi chemiometrici si è posto l'accento sulla ricerca della rilevanza di ciascuna variabile nel modello*** (loadings in PCA, coefficienti standardizzati in regressione, potere modellante di una variabile in PLS, metodi di selezione di un sottoinsieme di variabili, ecc.).

Da tutto ciò emerge la ***possibilità di utilizzare nella fase iniziale della ricerca di un modello non più alcune variabili preselezionate ad-hoc***, ***ma un numero più elevato (anche molte centinaia) di variabili candidate***: le variabili che si manifesteranno come poco o per nulla rilevanti nel modello saranno successivamente eliminate a favore di quelle variabili specifiche che sono correlate con la risposta studiata.

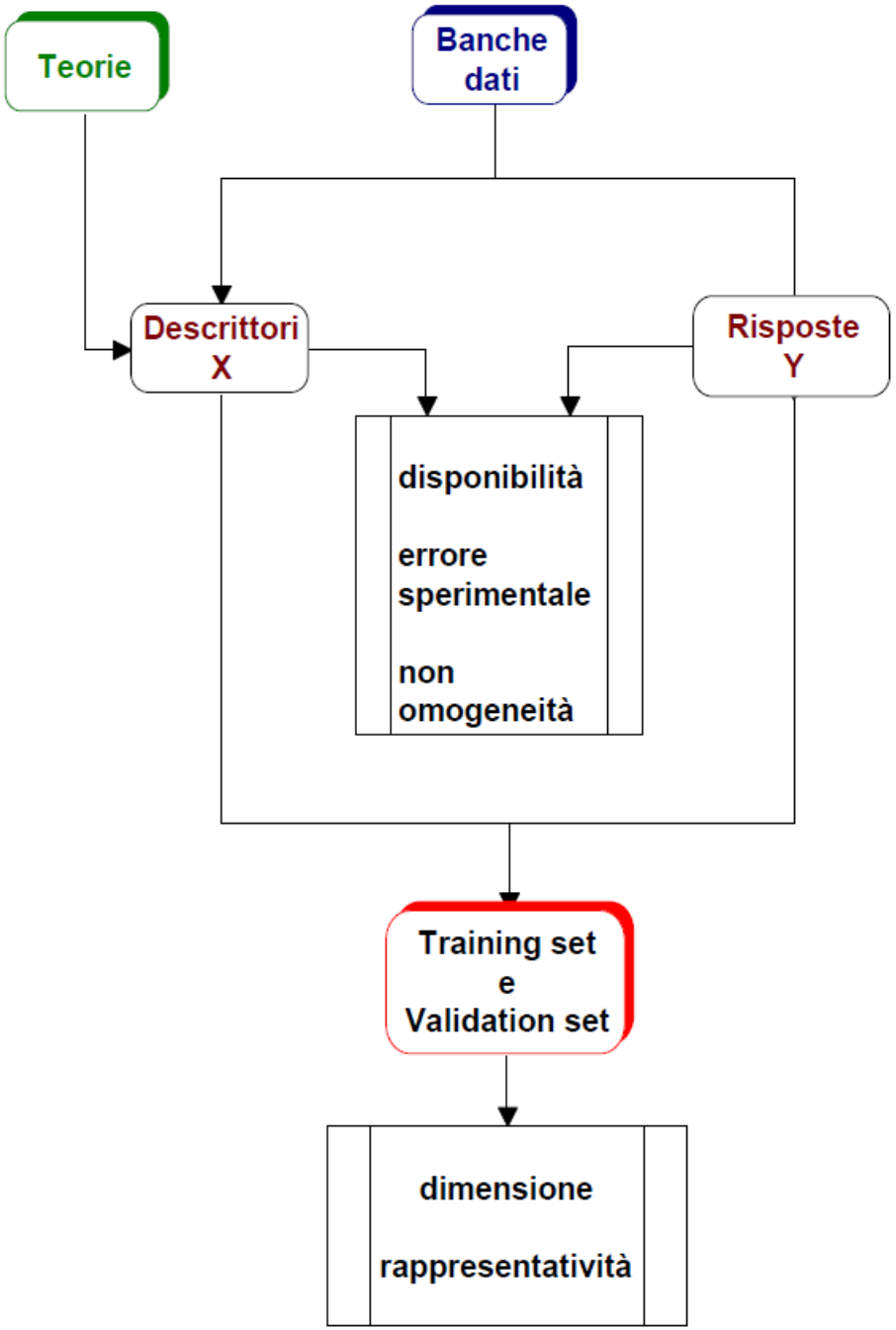
In particolare, nei problemi *QSAR*, lo sviluppo di descrittori molecolari teorici (ad esempio, **topologici e tridimensionali**) gioca un ruolo fondamentale nella ricerca rivolta a **predire risposte sperimentali complesse** da grandezze calcolabili teoricamente.

La **costruzione di adeguato *training set* per costruire modelli sufficientemente rappresentativi del problema esaminato** costituisce una fase di importanza fondamentale per le strategie *QSAR*.

I dati vengono reperiti non solo dalle banche dati esistenti e dalla letteratura, ma possono anche essere calcolati per via teorica.

I ***fattori limitanti* più importanti nella costruzione del *training set* riguardano la disponibilità di banche date aggiornate, l'omogeneità e l'accuratezza dei dati sperimentali,** la disponibilità e la rappresentatività di descrittori teorici.

Il *training set* deve essere costituito da un numero di dati sufficientemente numerosi da garantire di rappresentare il problema in modo adeguato e devono essere sufficientemente accurati al fine di evitare che il rumore in essi presente possa sovrastare l'informazione che si vuole estrare da essi.


Teorie
Banche dati
Descrittori X
Risposte Y
disponibilità
errore sperimentale
non omogeneità
Training set e Validation set
dimensione
rappresentatività

3D-QSAR in generale

L'approccio tridimensionale allo studio delle relazioni attività-struttura è fortemente legato all'approccio tradizionale *QSAR*, ove viene fatto un largo uso dei metodi chemiometrici. Esso tuttavia si differenzia dalle strategie tradizionali per il fatto che **i descrittori molecolari di cui si fa uso considerano in qualche modo gli aspetti 3D della molecola**.

Viene quindi dato un largo spazio alle informazioni che tengono conto degli aspetti tridimensionali delle molecole, della geometria molecolare e degli aspetti conformazionali. Questo comporta inevitabilmente un pesante aggravio di lavoro in quanto non è più possibile definire la molecola semplicemente mediante il suo grafo molecolare (2D), ma è **necessario effettuare calcoli che portino ad una "vera" struttura tridimensionale di minima energia.**

Si deve tener conto che si possono avere più conformazioni di minima energia, geometricamente anche molto diverse tra loro, e che **l'azione biologica potrebbe essere espletata non dalla conformazione di minima energia, ma da un suo stato di transizione.**

La progettazione di un nuovo composto con proprietà o attività farmacologiche predefinite, la sua sintesi, la ricerca sperimentale farmacologica ed i test clinici necessari sono un impegno estremamente rilevante e i cui esiti sono comunque incerti nella maggior parte dei casi.

Molte strategie teoriche basate sui principi su cui si fondano gli studi delle relazioni struttura-attività vengono oggi utilizzate col proposito di **evitare, almeno in parte, pesante e costosa attività sperimentale.**

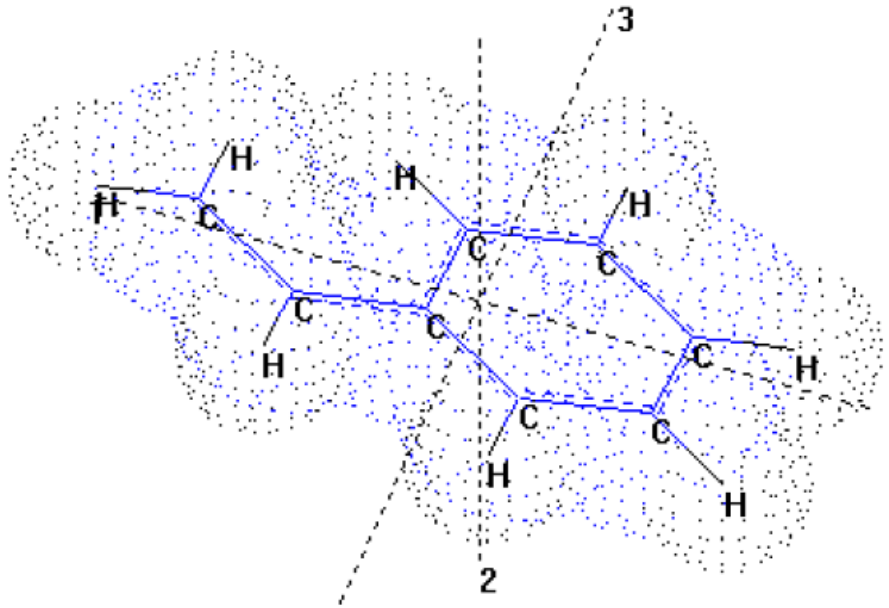
Studiare problemi a livello molecolare e di **interazione molecola-recettore** anche mediante visualizzazioni grafiche (Computer-Aided Molecular Design, CAMD).

I due metodi più noti sono GRID e CoMFA:

i campi di interazione di GRID rappresentano le energie totali (somma delle interazioni date dai potenziali elettrostatici, di Lennard-Jones, di legame idrogeno),

nel metodo CoMFA (COmparative Molecular Field Analyis) i potenziali sono di tipo elettrostatico o sterico. In entrambi i casi, i campi otttenuti possono essere utilizzati come descrittori puntuali della struttura molecolare e delle sue interazioni, particolarmente adatti allo studio di ***interazioni di binding***

Sono disponibili programmi che utilizzano per il calcolo delle proprietà molecolari, diversi metodi caratteristici della chimica teorica (metodi quantistici semi-empirici (Es http://openmopac.net/), metodi di meccanica molecolare, eccetera). Utilizzando i risultati di questi calcoli è possibile "vedere" la molecola rappresentata da superfici di risposta che modellano le proprietà molecolari e permettono di studiare visivamente le singole molecole



Descrittori molecolari

Un **descrittore molecolare** in chimica è un sistema per caratterizzare una molecola che permette di comparare molecole diverse e cercare molecole affini in una banca dati. Esso consiste nella rappresentazione matematica formale di una molecola in grado di trasformare l'informazione chimica, codificata all'interno di una rappresentazione simbolica della stessa molecola, in un valore numerico utile.[

Il linguaggio con cui esprimiamo i concetti e descriviamo la realtà non coincide con la realtà stessa, ma svolge tuttavia una parte attiva nel modellare la realtà di cui intendiamo parlare. L'interpretazione di ogni "fatto" è quindi indissolubilmente legata ai modi con cui il fatto viene descritto attraverso la mediazione del linguaggio ed il significato di ogni termine utilizzato dipende dal contesto teorico in cui si trova.

In questo contesto, i descrittori costituiscono gli elementi del linguaggio con cui rappresentiamo l'oggetto studiato, sia esso un sistema chimico, fisico o biologico.

La più comune classificazione dei descrittori, in accordo con l'approccio di Hansch, fa riferimento a tre gruppi di proprietà fondamentali:

a) idrofobicità (ad esempio, logP)

b) parametri elettronici (ad esempio, le energie HOMO e LUMO)

c) parametri sterici, di forma e dimensione (ad esempio, il peso molecolare)

Descrittori	*Fonte principale*
COMPOSIZIONALI	SPERIMENTALI
CHIMICO-FISICI	SPERIMENTALI
QUANTO-MECCANICI	TEORICI
DI GRUPPI SOSTITUENTI	CALCOLATI DA DATI SPERIM.
GLOBALI GEOMETRICI	CALCOLATI
LOCALI GEOMETRICI	CALCOLATI
BINARI	CALCOLATI
DI PUNTEGGIO	CALCOLATI
ENUMERATIVI	DEFINITI DALLA STRUTTURA
MATRICIALI DI CONNETTIVITÀ	DEFINITI DALLA STRUTTURA
TOPOLOGICI	TEORICI
DI CORRELAZIONE STRUTTURALE	TEORICI
WHIM E G-WHIM	TEORICI
DIFFERENZIALI	TEORICI
CROMATOGRAFICI	SPERIMENTALI
SPETTROSCOPICI	SPERIMENTALI
DI INTERAZIONE A CAMPI SCALARI	TEORICI
DI SIMILARITÀ MOLECOLARE	TEORICI
DI REATTIVITÀ CHIMICA E PROCESSO	SPERIMENTALI
DI ATTIVITÀ BIOLOGICA	SPERIMENTALI
CHEMO-AMBIENTALI	SPERIMENTALI
DI PROPRIETA' PRINCIPALI	CALCOLATI

I descrittori molecolari possono presentare le caratteristiche e le tipologie più diverse: in particolare, possono

(a) provenire da misure sperimentali, da calcoli teorici e da semplici operazioni di conto o di somma;

(b) rappresentare l'intera molecola o un particolare frammento molecolare o un sostituente in un sito definito;

(c) richiedere la conoscenza della struttura 3D della molecola, oppure il suo grafo molecolare o semplicemente la formula bruta;

(d) essere definiti da uno scalare, da un vettore, da un campo di scalari, etc.

I descrittori molecolari si possono dividere, ad esempio, in descrittori monodimensionali (1D), bidimensionali (2D), tridimensionali (3D), a livello microscopico e macroscopico, ecc.

m.kode-solutions.net/pf/dragon-7-0/

COMPANY PRODUCTS SERVICES PROJECTS PUBLICA

DRAGON HAS BEEN DISCONTINUED. IF YOU CURRENTLY OWN A DRAGON LICENSE AND NEED TECHNICAL SUPPORT, PLEASE CONTACT US AT CHM@KODE-SOLUTIONS.NET

Dragon is the world-wide most used application for the calculation of **molecular descriptors**. Its new version, Dragon 7.0, provides an improved user interface, new descriptors and additional features such as the calculation of fingerprints and the support for disconnected structures.

☑ **Molecular descriptors**: Dragon calculates 5,270 molecular descriptors, covering most of the various theoretical approaches. The list of descriptors includes the simplest atom types, functional groups and fragment counts, topological and geometrical descriptors, three-dimensional descriptors, but also several properties estimation (such as logP) and drug-like and lead-like alerts (such as the Lipinski's alert). The wide range of different approaches and theories for descriptors calculation, and the correctness and precision of their implementation are ensured by the scientific supervision of the Milano Chemometrics

Journal of Cheminformatics

Home About Articles Submission Guidelines About The Editors Calls For Papers

Software | Open Access | Published: 06 February 2018

Mordred: a molecular descriptor calculator

Hirotomo Moriwaki ✉, Yu-Shi Tian, Norihito Kawashita & Tatsuya Takagi

Journal of Cheminformatics **10**, Article number: 4 (2018) | Cite this article

52k Accesses | **330** Citations | **47** Altmetric | Metrics

Journal of Cheminformatics

Home About Articles Submission Guidelines About The Editors Calls For Papers

Software | Open Access | Published: 09 December 2015

ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation

Jie Dong, Dong-Sheng Cao ✉, Hong-Yu Miao, Shao Liu, Bai-Chuan Deng, Yong-Huan Yun, Ning-Ning Wang, Ai-Ping Lu, Wen-Bin Zeng ✉ & Alex F. Chen

Journal of Cheminformatics **7**, Article number: 60 (2015) | Cite this article

14k Accesses | **162** Citations | **5** Altmetric | Metrics