Capitolo 1

REACH e *QSAR*

1.1 Il Regolamento, le *QSAR*s (*Quantitative Structure Activity Relationships*) e *QSPR*s (*Quantitative Structure Property Relationships*)

Nel giugno del 2007 è entrato in vigore il regolamento europeo CE 1907/2006, del 18 dicembre 2006, riguardante la *Registration, Evaluation, Authorisation and Restriction of Chemicals* (*REACH*). Esso è stato scritto con l'intenzione di sostituire e semplificare una quarantina di norme precedenti, e rappresenta il più grande intervento normativo mai attuato sulla chimica europea. Si tratta, pertanto, di un regolamento assai complesso, per di più in continua evoluzione (è prevista una prima revisione già nel 2014).

Gli obbiettivi del *REACH* sono dichiarati all'art.1 dello stesso regolamento:

- incrementare la protezione della salute umana e dell'ambiente dai rischi derivanti dall'uso di *chemicals*;
- incrementare la competitività dell'industria chimica europea, favorendo la ricerca di soluzioni ecocompatibili;
- promuovere l'uso di metodi alternativi di valutazione del rischio delle sostanze chimiche;
- assicurare la libera circolazione delle sostanze all'interno della comunità europea.

Al *REACH* sono sottoposti produttori, importatori ed utilizzatori a valle (a uso industriale) di sostanze o preparati (*chemicals*) e di articoli, così come definiti all'art.3 del regolamento.[1]

La serie degli obblighi previsti a carico di fabbricanti, importatori ed utilizzatori è assai articolata, e una discussione approfondita degli stessi esula dagli scopi di questa ricerca, tuttavia, vale la pena mettere in evidenza alcuni aspetti del regolamento.

[1]Sostanza: "...un elemento chimico e i suoi composti, allo stato naturale o ottenuti per mezzo di un procedimento di fabbricazione, compresi gli additivi necessari a mantenerne la stabilità e le impurità derivanti dal procedimento utilizzato, ma esclusi i solventi che possono essere separati senza compromettere la stabilità della sostanza o modificarne la composizione"; preparato: "...una miscela o una soluzione composta di due o più sostanze"; articolo: "...un oggetto a cui sono dati durante la produzione una forma, una superficie o un disegno particolari che ne determinano la funzione in misura maggiore della sua composizione chimica".

Il *REACH* prevede l'inversione dell'onere della prova rispetto alla legislazione precedente. Esso ha posto in capo alle aziende l'obbligo di fornire alla *European Chemicals Agency* (*ECHA*) le informazioni necessarie alla registrazione dei *chemicals* da esse utilizzati. Secondo il principio di precauzione (*no data no market*), in assenza delle informazioni richieste il prodotto non potrà essere immesso sul mercato. In precedenza, la norma prevedeva che fosse l'autorità pubblica a dover dimostrare la pericolosità di una sostanza, prima di poterla far ritirare dal mercato. Gli ingenti costi che queste ricerche normalmente comportano, hanno fatto sì che solo per una minima parte delle sostanze di uso industriale, in Europa oggi oltre 100000, si abbiano informazioni dettagliate circa la loro pericolosità per l'uomo e per l'ambiente.

Per effettuare la registrazione, a norma dell'art.10 del *REACH*, è necessario fornire a *ECHA* una serie di informazioni comprendenti sia dati chimico fisici, sia dati tossicologici e ecotossicologici, valutati secondo i criteri esposti nell'allegato I del regolamento. Priorità nella registrazione è stata data alle sostanze che sono: persistenti, bioaccumulabili o tossiche (PBT), molto persistenti e molto bioaccumulabili (*vPvB*), sensibilizzanti e/o cancerogene, mutagene o tossiche per la riproduzione (*CMR*), oppure già classificate come pericolose a norma della direttiva 67/548/CE. Per le sostanze comprese nell'allegato XIV del *REACH*, oltre alla registrazione, è prevista l'autorizzazione all'uso da parte dell'agenzia. Questa verrà data sulla base della effettiva necessità, impossibilità di sostituzione a breve termine nel ciclo produttivo, e sarà comunque un'autorizzazione temporanea, soggetta a scadenza. Tale clausola è stata molto criticata dai costruttori in considerazione della possibilità che alcuni impianti, oggi in funzione, debbano essere chiusi nel 2018, allo scadere dell'autorizzazione, prima ancora di aver recuperato gli investimenti fatti. Le sostanze incluse nell'allegato XVII, saranno soggette a restrizione e tolte dal mercato. La lista delle restrizioni comprende al momento pochissime sostanze. L'inserimento nell'allegato XVII è infatti un iter complesso, che tiene conto di aspetti, sia scientifici sia economici, e richiede

l'accordo di tutti gli stati membri allo scopo di evitare che vengano penalizzati specifici settori produttivi. Le restrizioni dell'allegato XVII non sono generiche. Vengono applicate a singole sostanze o preparati, oppure a gruppi di essi, identificati, per quanto possibile, in modo univoco tramite i numeri *Chemical Abstracts Service* (CAS), *European Inventory of Existing Commercial Chemical Substances* (*EINECS*), *European List of Notified Chemical Substances* (*ELINCS*). Sono inoltre specificati gli usi per cui le restrizioni si devono applicare. A due anni dall'entrata in vigore del *REACH*, i costi della sua implementazione sono controversi [Gilbert 2009; Hartung *et al.* 2009]. Le preregistrazioni eseguite finora riguardano circa 140000 sostanze (più di quante se ne ritenessero circolanti in Europa). Sino a oggi sono stati costituiti, tra le ditte interessate dal *REACH,* solo 2270 consorzi per lo scambio di informazioni (*SIEF*), riferiti a altrettanti *chemicals.*

La valutazione del rischio richiesta dal *REACH* (*ex* artt. 13 e 14), viene fatta considerando sia la pericolosità intrinseca di una sostanza, sia la possibilità che essa venga in contatto con l'uomo o con l'ambiente. Il *REACH* pertanto prevede vengano fornite informazioni non solo sulle proprietà dei *chemicals,* ma anche sugli scenari di esposizione[2] degli stessi (*ex* art. 3 del regolamento). Per abbattere i costi, il *REACH* ammette la possibilità di raccogliere informazioni relative alla pericolosità di una sostanza utilizzando, oltre ai *test in vivo,* anche test *in vitro* o *in silico,* alternativi ai test su animali, secondo i criteri esposti nell'allegato XI del regolamento.

Fra i metodi alternativi ai test sugli animali citati al paragrafo 1 dell'allegato XI vi sono: Relazioni Qualitative o Quantitative tra Struttura e Attività *QSAR*,[3] metodi *in vitro*, raggruppamento di sostanze e metodo del *Read Across*.

---

[2]Scenario d'esposizione:"*... l'insieme delle condizioni, comprese le condizioni operative e le misure di gestione dei rischi, che descrivono il modo in cui la sostanza è fabbricata o utilizzata durante il suo ciclo di vita e il modo in cui il fabbricante o l'importatore controlla o raccomanda agli utilizzatori a valle di controllare l'esposizione delle persone e dell'ambiente. Questi scenari d'esposizione possono coprire un processo o un uso specifico o più processi o usi specifici, se del caso*".

[3]Con il termine *SAR* viene di solito indicata una relazione qualitativa, con *QSAR* una quantitativa,(*Q*)*SAR* rappresenta una crasi delle due. Questo verrà specificato meglio più avanti citando la linea guida *ECHA* 2008, nel testo si è mantenuto il termine solo il QSAR , poiché oggetto del lavoro di tesi sono solo i modelli quantitativi.

Maggiori informazioni sui metodi *QSAR* e *Read Across*, con esplicito riferimento al *REACH*, si trovano nella "*Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals*", emessa da *ECHA* nel 2008 [ECHA2008] e disponibile in rete. La guida, che non è una norma imperativa, è stata redatta con lo scopo di aiutare gli *stakeholder*[4] a ottemperare agli obblighi previsti dal *REACH* evitando per quanto possibile i test sugli animali. Secondo quanto riportato in essa, per generare dati da "*non testing methods*" si possono seguire tre approcci principali:

- tecniche di raggruppamento (che includono *read across* e identificazione di categorie chimiche);
- *QSAR*, il termine "quantitativo" è opzionale;
- sistemi esperti (una combinazione di modelli *SAR*, *QSAR* e impiego di *data base*).

Lo sviluppo di *non testing methods* è basato sul *principio di similitudine,* ovvero sulla ipotesi che composti simili si comportino in modo simile dal punto di vista biologico. In realtà questo principio poggia su almeno due ipotesi di lavoro. La prima è la stazionarietà del fenomeno, ovvero che molecole uguali si comportino sempre alla stessa maniera e che altre molecole "simili" possano seguire lo stesso meccanismo. La seconda è che si possa definire in qualche modo la "similitudine" delle molecole, ovvero che esiste un criterio per identificare le molecole che seguono lo stesso meccanismo. Fra i metodi sopra citati, l'oggetto della nostra ricerca sono le *QSAR*.

Nella linea guida *ECHA* 2008 [*ECHA* 2008] i metodi SAR e *QSAR*, in sintesi (Q)SAR, sono così definiti:

1. *SAR* è una relazione qualitativa che correla una (sub)struttura alla presenza o all'assenza di una proprietà o di una attività di interesse.

[4]*Stakeholder* è il termine con cui si indicano tutti i soggetti latori di interessi in una iniziativa economica, il termine fu inventato da R.E. Freeman ma reso di uso comune da T. Blair [Stieb 2009]

2. *QSAR* è un modello matematico (spesso una correlazione statistica) che mette in relazione uno o più parametri quantitativi derivanti dalla struttura chimica o dalla misura quantitativa di una proprietà o attività biologica (e.g. un *endpoint* [5] eco-tossicologico). I modelli *QSAR* sono modelli quantitativi che forniscono risultati sia categorici (*label*), sia quantitativi. I parametri che si usano nel modello per descrivere le molecole sono chiamati *molecular descriptors* (*MD*).

Oltre alle *QSAR* esistono anche altri modelli: *quantitative structure property relationships* (*QSPRs*), tramite i quali si possono stimare proprietà chimico fisiche delle molecole sulla base del calcolo (o misura) di *MD* e di quantità note per molecole simili. Queste però sono state trascurate dalla guida *ECHA* in quanto, ai fini *REACH*, la misura diretta in molti casi risulta più agevole e economica. Da un esame della letteratura [ECHA 2008] si ricava che il termine "quantitativo" nelle *QSAR* può essere usato in modo ambiguo. Esso infatti può essere riferito al valore degli *endpoint* considerati quale variabile dipendente nel modello, oppure al valore dei descrittori usati quale variabile indipendente. La linea guida *ECHA* adotta questa seconda definizione sostenendo che sia corretto considerare *MD* quantitativi da cui ricavare *endpoint* qualitativi o quantitativi. Dal nostro punto di vista sarebbe tuttavia preferibile mettere in relazione due variabili (dipendente e indipendente) entrambe quantitative. Esistono in letteratura metodi statistici, quale l'analisi delle corrispondenze [Cuadras *et al*. 2007; Nenadic *et al*. 2006], che permettono di trattare variabili dipendenti di tipo non numerico (*label*) e quindi di ottenere *QSAR* qualitative, che tuttavia ci sembrano meno significative. Una discussione dell'influenza della variabile dipendente (*label* o numerica) sulla capacità predittiva di modelli *QSAR* applicati a sostanze cancerogene è stata fatta da Benigni [Benigni *et al*. 2008].

[5]La *Guidance document on the validation of (quantitative) structure activity relationships [QSAR] models* [OECD2007] riporta la seguente definizione *:" ...endpoint refers to any physico-chemical property, biological effect (human health or ecological) environmental fate parameter that can be measured and therefore modelled*".

Le tecniche più comuni per sviluppare una *QSAR* sono tecniche di analisi statistica dei dati: *regression analysis,* reti neurali e i metodi di classificazione. Fra gli esempi di regressione può essere citato il metodo dei minimi quadrati (nelle versioni, ordinaria, *multiple least squares* e *partial least squares)*. Fra i metodi di classificazione si possono citare la *discriminant analysis,* i *classification trees* e la *similarity analysis* con metodi basati sulla distanza sul piano euclideo.

Nel *REACH* e nei documenti ufficiali collegati, linea guida *ECHA* [ECHA2008] e linea guida della *Organisation for Economic Cooperation and Development* (*OECD*)[OECD 2007], non viene fornita alcuna indicazione su come sviluppare una *QSAR*. Vengono tuttavia poste delle condizioni al contorno, elencati principi cui attenersi, affinché i dati provenienti da un modello *QSAR* possano essere considerati un'accettabile alternativa ai dati sperimentali. Per poter essere usato un modello *QSAR* esso dovrebbe essere *relevant, reliable* e *adequate,*[6] inoltre i dati stimati dovrebbero essere accettati sulla base dello *weight of evidence approach*. Un elemento di incertezza è rappresentato dal fatto che i criteri elencati in ciascun documento [ECHA 2008, OECD 2007] per la validazione dei modelli, sono simili ma non identici.

Sul piano scientifico le *QSAR* forniscono uno strumento per esplorare e comprendere dati e per stimare l'attività di nuovi *chemicals* mai sperimentati prima (per esempio prima di introdurre una nuova tecnologia). Modelli *QSAR* sono usati da anni nell'ambito dell'industria farmaceutica per il *molecular modeling* [7] e il *drug design* [Hansch *et al*. 1962], rappresenta una novità invece il loro uso in ambito di igiene industriale e ambientale. Il postulato di quella che oggi è conosciuta quale la Hansch Analysis [Hansch 1995] è che le interazioni tra *ligand* e *biotarget* possano essere quantificate in termini di fattori sterici, idrofobici e elettronici (energie degli orbitali HOMO e

[6]Gli aggettivi usati: *relevant, reliable* e *adequate* sono tratti dall'edizione inglese della linea guida per l'applicazione del *REACH*, *ECHA* 2008 (esclusivamente presente in lingua veicolare).

[7]Informazioni sulle tecniche di *molecular modeling* e *drug design* possono essere reperite al sito della sezione di molecular modeling del Dipartimento di Scienze Farmaceutiche dell'Università di Padova *http://mms.dsfarm.unipd.it/index.html*. Una lista di pubblicazioni scientifiche che riguardano la materia delle *QSAR* è disponibile alll'indirizzo *http://www.qsarworld.com/literature-qsar-journals.php?lm=1*.

LUMO)[8] del *ligand*. L'esistenza di una relazione tra la reattività di una molecola e la presenza di uno o più gruppi funzionali è un'idea ben consolidata in chimica. È accettato il fatto che si possa stimare (almeno a livello qualitativo) il comportamento chimico di una data sostanza ,sulla base della presenza di tali gruppi funzionali. Per fenomeni complessi, quali la tossicità o l'attività biologica di una molecola, vi sono tuttavia molte più variabili in gioco e non è detto che tutto sia interpretabile sulla base della sola presenza di gruppi funzionali o della struttura molecolare. Questo fa sì che, nonostante siano passati quasi cinquant'anni da quando Hansch e colleghi proposero l'ipotesi dell'esistenza di *QSAR*, la *computational toxicology* a tutt'oggi può essere considerata una scienza giovane e imperfetta. Schultz [Schultz *et al*. 2006 b] ha proposto la seguente spiegazione: "*...the limitation of the applicability of organic chemistry to toxicology is that organic chemical reactions are often explaine QSAR d o QSAR n the basis of experimental evidence acquired in environments (e.g. temperature, solvents) very different from those found in live biological systems. Nevertheless, the general rules of chemical reactivity are a good starting point for defining reactivity toxicity*." Come evidenziato da Schultz, l'ambiente di reazione in questo caso ha un'importanza determinante nei confronti del risultato, il quale va sempre interpretato dal punto di vista del meccanismo senza limitarsi a ottenere "la migliore delle rette possibili" per interpolare i dati del *training set*.

Visti i limiti evidenziati dall'approccio statistico alla *QSAR*, si è scelto pertanto di adottare un approccio diverso; un approccio quantomeccanico alla *QSAR*, pochi descrittori locali calcolati con metodi a partire da principi primi. Secondo questo approccio:

- la scelta degli *MD* viene fatta sulla base del meccanismo ipotizzato;
- il loro calcolo viene effettuato tramite la *Density Functional Theory* (*DFT*);
- l'analisi statistica è usata per identificare la relazione (*QSAR*) e validare il modello.

[8] *HOMO* e *LUMO* corrispondono a *Highest Occupied Molecular Orbital* e *Lowest Unoccupied Molecular Orbital*, rispettivamente.

Incidentalmente, è opportuno sottolineare la necessità che la molecola di cui si stima la tossicità faccia parte del dominio del modello e non sia stato in qualche modo violato il principio di similitudine, questo può essere fatto solo con una ipotesi sul meccanismo. In ogni caso, la tossicologia computazionale è in continua evoluzione e non è detto che un progetto assolutamente ambizioso, quale quello della valutazione *in silico* della pericolosità di una molecola, non possa essere in futuro alla nostra portata. Per il momento appare più prudente non scordare quanto affermato da M. Planck: "*Experiments are the only means of knowledge at our disposal. The rest is poetry, imagination*". Inoltre merita di essere enfatizzato che i dati ottenuti da esperimenti *in silico* non sostituiscono del tutto i dati ottenuti da esperimenti in laboratorio, ma si appoggiano sempre su di essi. Le tecniche non sono alternative ma complementari, i metodi *in silico* aiutano a sfruttare al meglio le informazioni ottenute da esperimenti con metodi tradizionali. Concludendo, la tossicità di una molecola è un fenomeno complesso che non può essere compreso, e talvolta nemmeno definito, a prescindere dalla conoscenza del meccanismo di azione (*MOA*).

I metodi *in silico* quali sono le *QSAR* possono essere usati in due modi.

1. Quando non si conosce il *MOA* della tossicità del *chemical* dal confronto tra i dati sperimentali *in vivo* o *in vitro* con quelli *in silico* possono essere vagliate ipotesi di meccanismo isolando, ove possibile, il contributo di ciascun *MD* per comprendere qualcosa in più sul *MOA*.
2. Quando invece il *MOA* e la *QSAR* sono noti è possibile stimare l'*endpoint* considerato nel modello per una molecola incognita sulla base del calcolo, o della misura, degli *MD*. Che è propriamente quanto richiesto dal *REACH* alla *QSAR*.

1.2 Validazione di un modello *QSAR*

Uno dei punti cruciali nell'uso di una *QSAR* per la stima di un *endpoint* è la valutazione dell'incertezza della stima fornita dal modello. I requisiti richiesti dal *REACH* (annex XI) per poter usare un modello *QSAR* al fine di ottemperare agli obblighi previsti dal regolamento nell'ambito della registrazione, sono alquanto generici. Vi si chiede infatti che:

- i risultati derivino da modelli *QSAR* la cui validità scientifica sia stata stabilita;
- la sostanza in esame cada nel dominio del modello *QSAR* adottato;
- i risultati siano adeguati allo scopo, sia esso la classificazione e l'etichettatura della sostanza, oppure la valutazione del rischio;
- sia fornita adeguata e affidabile (*adequate and reliable)* documentazione riguardo al metodo adottato.

L'*OECD* nella propria guida alla validazione dei modelli *QSAR* [OECD 2007], elenca in modo analogo cinque principi; in base ai quali, per poter essere accettato, un modello dovrebbe poter fornire le seguenti informazioni:

- un *endpoint* definito;
- un algoritmo privo di ambiguità, lo scopo è quello di assicurare la trasparenza sull'algoritmo utilizzato dal modello;
- un dominio di applicabilità ben definito; ovvero stabilire per quali, molecole, strutture chimiche o meccanismo il modello possa essere ritenuto affidabile;
- una valutazione delle prestazioni statistiche del modello, tramite misure appropriate di *goodness-of-fit, robustness* e *predictivity;*
- un'interpretazione del meccanismo, "se possibile" .

I principi stabiliti da *OECD* sono stati studiati per essere applicati, oltre che alle *QSAR*, anche a altre tecniche, quali le reti neurali, i *decision trees,* i "sistemi esperti". Nelle due linee guida sopra

citate vi è un'ampia discussione sul significato da dare a ognuno dei cinque principi, con un elenco delle tecniche più comunemente usate per lo sviluppo di *QSAR* (tutte o quasi con approccio statistico), e un giudizio sulla loro conformità o meno ai cinque principi.

Sul piano scientifico la validazione di un modello *QSAR*, o QSPR, è una condizione essenziale affinché esso possa essere usato [Tropsha *et al*. 2003]. Per i modelli multivariati si possono eseguire due tipi di validazione: esterna e interna. Nella validazione esterna i dati tossicologici a disposizione vengono divisi in due gruppi, *training set* e *test set*, il primo viene usato per costruire il modello e il secondo per validarlo. Nel caso in cui le molecole siano troppo poche per poterle dividere, si effettua invece la validazione interna. La tecnica *leave one out,* o *leave many out,* prevede si tolgano una o più molecole dal *training set* e si costruisca il modello sulla base delle rimanenti. Il modello viene poi testato su quelle scartate, ripetendo il ciclo in modo iterativo fino a trovare la migliore delle correlazioni possibili. Per determinare le prestazioni statistiche dei modelli, sono stati sviluppati indici e test statistici [Martens *et al*. 1998; Faber 1999; Golbraikh *et al*. 2002; Faber *et al*. 2003; Öberg 2004; Aptula *et al*. 2005; Papa *et al*. 2005; Gramatica 2007; Manchester *et al*. 2009].

1.3 Come non sviluppare una *QSAR*

In cinquant'anni sono state pubblicate migliaia di relazioni QSAR *e* QSPR *(a settembre 2009 cercando con la parola chiave* "QSAR" *nella banca dati* ISIWEB of Science *si ottenevano più di 8500 articoli, aggiungendo la parola chiave* "DFT" *ne comparivano 124, di cui 16 nel 2009) e, nonostante la pubblicazione dei cinque principi adottati da* OECD*, il dibattito sulla validità dei modelli* QSAR *e sul modo migliore per ottenerli è ancora in corso. Inoltre, in letteratura si trovano varie pubblicazioni in cui sono stati affrontati i temi dell'affidabilità dei modelli* QSAR *e della loro liceità d'uso, sia a fini scientifici, sia normativi [Schultz et al. 2003, Oeberg 2004; Benigni et al. 2008; Zvinavashe et al. 2008; Bajorath et al. 2009].*

Il dibattito in corso testimonia sia della complessità dell'argomento, la stima dell'attività biologica di un chemical sulla base di informazioni strutturali può essere alquanto insidiosa, sia dell'onestà scientifica degli autori che vi si sono dedicati, i quali, di fronte al rinnovato interesse dato dal REACH *per questo tipo di tecniche, hanno saputo metterne in luce i limiti oltre che le qualità. Va sottolineato che alcuni degli autori citati sopra hanno contribuito alla stesura dei cinque principi adottati da* OECD *per l'accettabilità di un modello* QSAR *a fini normativi. Pertanto, detti principi nascono dall'esperienza accumulata dagli stessi autori in questo settore di ricerca. Nel giugno 2009 J. C. Dearden, M.T.D. Cronin e K.L.E. Kaiser,* DCK *nel seguito, hanno pubblicato un articolo [Dearden et al. 2009] dal titolo "How not to develop a quantitative-structure activity or structure-property relationship (*QSAR/QSPR*)" in cui si effettuava una rassegna critica dei più comuni (ventuno) tipi di errori che possono essere riscontrati in letteratura in articoli che presentano lo sviluppo di* QSAR*. Ognuno di questi errori è stato trattato con dovizia di esempi di letteratura, (anche opera degli stessi autori) e messo in relazione con una violazione di uno o più dei cinque principi definiti da* OECD*. L'articolo di* DCK *ha rappresentato un utile riferimento perché alcuni dei problemi sollevati dagli autori li si è dovuti affrontare noi stessi.*

Fra i problemi discussi da DCK *se ne segnalano due. Il primo è l'*overfitting *per cui un modello* QSAR *risulterebbe particolarmente adatto a stimare i dati del* training set *o* test set *usati per svilupparlo, ma sarebbe poco accurato al di fuori di essi. Il sospetto è che alcuni dei modelli che vantano ottime prestazioni statistiche, in realtà soffrano di questo problema. Il secondo è la mancanza di ipotesi sul meccanismo. Pur se* OECD *non lo pone come un criterio vincolante per la validità di una* QSAR*, una buona correlazione statistica tra* MD *e* endpoint *non garantisce affatto che gli* MD *possano spiegare il meccanismo di azione. La tossicità è un fenomeno complesso e multifattoriale, quello che noi possiamo misurare è l'effetto finale, ma sarebbe singolare che un solo* MD *fosse in grado di spiegarlo in modo significativo (e infatti i modelli monovariati non funzionano molto bene). Quando vi sono più meccanismi coinvolti nella tossicità, effetti sinergici o antagonisti possono agire quali fattori di confondimento.*

La ricerca presentata in questa tesi ha lo scopo di rispondere a due domande:

1) Vi è una relazione tra proprietà elettroniche molecolari e tossicità?
2) *É possibile usare la* DFT *per calcolare gli* MD *da usare nei modelli* QSAR*?*

Per rispondere è stato adottato un approccio meccanicistico alla QSAR*. Sono stati selezionati alcuni training set, comprendenti piccole molecole organiche di uso industriale sulla base del presunto ruolo svolto nella loro tossicità da un particolare recettore (*Aryl Hydrocarbon Receptor*) e sono state calcolate tramite la* DFT *alcune proprietà elettroniche molecolari utilizzate poi quali* MD *nello sviluppo dei modelli* QSAR*.*

Capitolo 2

Dati Biologici e Dominio di Applicabilità

2. 1 Come si costruisce una *QSAR*

Quattro sono gli elementi che compongono una *QSAR*.[9]

1) L'insieme delle proprietà tossicologiche che si vogliono indagare e che costituiscono la variabile dipendente del modello. Ad esse si fa normalmente riferimento con il termine *endpoint*. Il termine in tossicologia era inizialmente riferito all'accumulo di una sostanza tossica entro un organismo o organo bersaglio, in realtà l'azione tossica è molto più complessa per cui, in questo contesto, il termine va inteso come riferito al risultato di un test tossicologico qualsiasi.
2) L'insieme dei descrittori molecolari (*MD*), le variabili indipendenti, proprietà molecolari che si vogliono collegate alla tossicità, che devono esistere per ogni molecola e essere esprimibili in modo quantitativo.
3) L'insieme delle molecole per le quali si ritiene sia valido il modello *QSAR*, che normalmente viene diviso in due gruppi denominati *training set*, quello su cui si costruisce il modello, e *test set* quello sui cui si valida il modello. Per ogni molecola appartenente ad essi occorre siano note sia la variabile dipendente sia quella indipendente.
4) La relazione matematica che lega le due variabili dipendente e indipendente, cui viene dato il nome di *QSAR*.

Ovviamente, il problema è di per sé multidimensionale; conseguentemente, la rappresentazione grafica della soluzione implica la riduzione del numero delle variabili. Se si fissano *endpoint, MD* e molecole del *training set,* la *QSAR* può essere rappresentata nel modo descritto in figura 2.1.

[9]Come anticipato nel capitolo nel capitolo precedente, sono ammesse dal *REACH* sia relazioni qualitative tra attività e struttura *SAR*, sia relazioni quantitative tra attività e struttura *QSAR*, con la notazione (*Q*)*SAR* si fa riferimento sia a modelli qualitativi sia quantitativi. Poiché oggetto della ricerca sono solo i modelli quantitativi, a questo punto si ritiene di poter utilizzare solo la notazione *QSAR* intendendo con ciò fare riferimento solo a questi ultimi senza ambiguità.

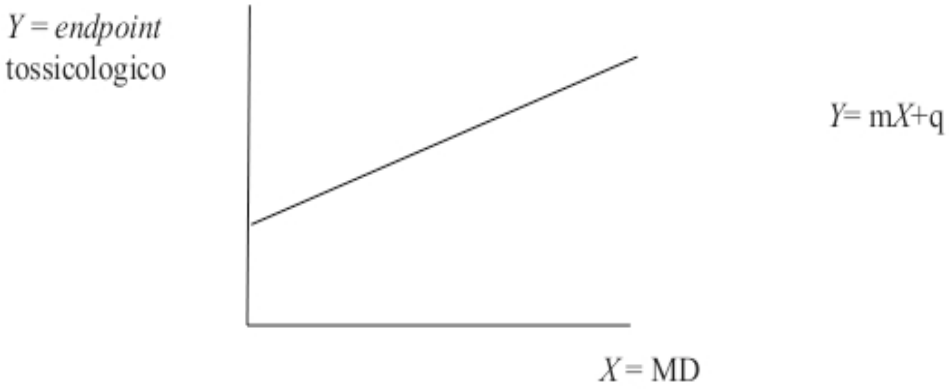

Figura 2.1. Schema di costruzione di una *QSAR* (lineare).

Se esiste una relazione, qui supposta lineare, le molecole del *training set* si distribuiranno lungo la retta e sarà possibile stimare dati di tossicità per molecole incognite (*test set*) sulla base del solo calcolo degli *MD* delle stesse. Questo permette di evitare la misura sperimentale diretta della tossicità per molecole incognite.

Come accennato nel primo capitolo vi sono vari approcci per costruire una *QSAR*. Qualunque sia il modello adottato esso non sarà valido per ogni molecola, per ogni *endpoint* e per qualsiasi *MD*. Questo infatti equivarrebbe a dire che per tutte le molecole basta misurare una proprietà qualsiasi e la si troverà collegata con la tossicità. Così non è, ovviamente. È ragionevole supporre, invece, che la presenza di una relazione quantitativa (soprattutto se lineare) tra proprietà molecolari e tossicità delle molecole sia in realtà un caso molto fortunato e che la validità dei modelli sia limitata a piccoli gruppi di molecole molto affini tra loro, nonché a proprietà e *endpoint* specifici. Stabilire quando un modello *QSAR* può essere usato equivale a definirne il "dominio di applicabilità" (*AD*). Occorre, in altre parole, definire per quali molecole, *endpoint* e *MD* si possa ritenere che la stima dell'*endpoint* fornita dal modello per una molecola incognita sia

sufficientemente accurata. I primi tre elementi, quindi, concorrono tutti insieme a determinare l'esistenza del quarto. In questo capitolo verranno affrontati i problemi della scelta delle molecole e dei dati tossicologici (*endpoint*) da inserire nel modello, postponendo la discussione sugli *MD* al capitolo quarto.

2.2 Approccio meccanicistico alla *QSAR* e dominio di applicabilità

Storicamente, al concetto di *AD* di un modello *QSAR* sono state date accezioni diverse, la definizione, infatti, non è univoca. Essa dipende dal modo in cui il modello *QSAR* è stato costruito. In letteratura [Netzeva *et al.* 2005] ne viene riportata la seguente definizione: *"The applicability domain of a QSAR model is the response and chemical structure space in which the model makes predictions with a given reliability"*. Con il termine *response* si intende ogni effetto chimico, fisico o biologico che si voglia stimare (in altre parole l'*endpoint* in esame); con *chemical structure* si intendono proprietà chimico fisiche, strutturali o topologiche delle molecole (gli *MD*). La definizione di Netzeva risente dei limiti dell'approccio classico alla *QSAR*: identificazione della variabile indipendente con la presenza di gruppi funzionali o particolari strutture nelle molecole, conseguentemente la definizione dell'*AD* del modello risulta bidimensionale, con identità tra *MD* e strutture molecolari (lo "spazio delle strutture chimiche"). L'*AD* del modello è invece tridimensionale e deve tener conto contemporaneamente delle prime tre componenti della *QSAR*: *endpoint, MD* e molecole. Un modello *QSAR* può infatti essere applicato con sufficiente affidabilità solo per la stima di uno specifico *endpoint* e per un limitato numero di strutture, proprietà chimico fisiche, meccanismi di azione[10] (*MOA*). Una volta scelti l'*endpoint*, gli *MD* e il *training set*, ci si possono aspettare stime attendibili solo per *chemicals* simili a quelli costituenti il *training set* (principio di similitudine). Si può ritenere che stime fatte per *chemicals* che giacciono al di fuori dell'*AD* del modello non siano attendibili; non è invece vero il contrario; la sola appartenenza di un *chemical* all'*AD* del modello, non è una garanzia sufficiente per poter affermare che le stime che lo riguardano siano attendibili. Il confine del dominio corrisponde alla violazione del principio di similitudine. Esso può a volte non essere così netto da poter includere o escludere una molecola in modo deterministico, per cui ne risulta un *AD* indefinito. Allo stesso modo, per un modello *QSAR*

[10]In questo caso, con il termine meccanismo di azione, si fa riferimento sia al modo (tossicocinetica) sia al meccanismo di azione (tossicodinamica).

non vi è, in generale, un confine assoluto tra stime affidabili e non; occorre ragionare in termini probabilistici. Così facendo, sulla definizione dell'*AD* di un modello *QSAR* influisce anche la scelta del metodo statistico usato per validarlo. La definizione del *AD* di un modello *QSAR* rappresenta pertanto uno dei problemi cruciali del suo sviluppo, alla cui soluzione concorrono tutte le informazioni in nostro possesso. La presenza di un *AD* definito è uno dei requisiti richiesti da OECD per il corretto utilizzo di un modello *QSAR*.

Alla base del principio di similitudine vi è il fatto di porre un'ipotesi sul meccanismo. La tossicità di una molecola può essere frutto sia di un singolo *MOA*, sia della combinazione di *MOA* diversi. Per quegli *endpoint* che possono essere rappresentativi degli effetti di *MOA* diversi si possono adottare due strategie:

- si costruiscono varie *QSAR* per un unico *endpoint,* una per ciascun meccanismo, ognuna con un *AD* molto ristretto;
- si costruisce, sulla base degli effetti osservati piuttosto che del *MOA*, una *QSAR* unica con un dominio vasto, comprendente molecole anche molto diverse tra loro.

La seconda scelta introduce una nuova approssimazione, si ammette infatti che gli stessi *MD* possano rendere conto di meccanismi diversi. Le due strategie sopra descritte rappresentano due approcci diversi alla *QSAR*.

- Il primo privilegia il ruolo del meccanismo, intendendo definire l'*AD* sulla base della similitudine di meccanismo.
- Il secondo privilegia il ruolo dell'effetto (*endpoint* considerato). Assumendo che ogni *endpoint* sia comunque frutto del contributo simultaneo di meccanismi diversi, non ben identificabili, l'*AD* del modello è definito in base alla similitudine di effetto.

Le due strategie e approcci sopra descritti, rappresentano le due scuole di pensiero, statistica e meccanicistica, già indicate nel primo capitolo. Occorre in ogni caso trovare un equilibrio tra

l'ampiezza dell'*AD* e la affidabilità delle stime che si vogliono ottenere da una *QSAR*. Modelli "specializzati" otterranno stime presumibilmente più affidabili.

Lo studio di metodi per la definizione dell'*AD* di un modello *QSAR* è un campo di ricerca tutt'ora in fase di esplorazione [Netzeva *et al*. 2005]. Nella linea guida OECD [OECD 2007] sono stati esaminati alcuni dei metodi presenti in letteratura per la definizione dell'*AD* di un modello *QSAR*. Nel seguito verranno presentati due di quelli citati dalla guida, il primo basato sulla presenza di strutture molecolari simili, il secondo sul campo di esistenza degli *MD* usati nel modello.

Un modo di definire l'*AD* è quello di raggruppare le molecole in base alla loro struttura, sviluppando il modello per molecole omologhe (e.g. una serie di alcoli alifatici, molecole aromatiche, congeneri del benzene sostituito, ecc.). La presenza di particolari strutture nella molecola può essere collegata tramite un "giudizio esperto" con portatori (*toxicophores*) o modulatori di tossicità. Collegare la tossicità con la presenza di strutture, frammenti, gruppi funzionali, è in ogni caso un processo delicato. Molecole omologhe possono presentare delle forti differenze nella tossicità dovute ad un cambiamento nel *MOA* [Deener *et al*. 1988]. Lo stesso frammento (e.g. metile) può in alcuni casi agire aumentando la tossicità [Russom *et al*. 1997] e in altri inibendola [Schultz *et al*., 2005]. A volte molecole strutturalmente diverse possono comportarsi in modo simile, per cui possono essere considerate ai fini di uno specifico *endpoint* un unico dominio chimico. È il caso delle ammine aromatiche e dei fenoli che si comportano similmente dal punto di vista del loro effetto narcotico su organismi acquatici, ma sono molto diversi per quanto riguarda la mutagenicità [Verhaar *et al*.1995]. In questi casi è conveniente definire il dominio in base al *MOA*, vincolando la scelta del dominio all'*endpoint* e al meccanismo considerati, piuttosto che alla presenza di strutture simili. Raggruppare le molecole in base al *MOA* richiede un "giudizio esperto" e una certa dose di conoscenze sulla tossicologia delle molecole in esame. La presenza simultanea di più gruppi funzionali può portare infatti ad effetti sinergici, antagonisti o a

cambiamenti nel *MOA* [Schultz *et al*. 2002]. Le molecole possono subire anche biotrasformazioni per cui responsabile della tossicità non è la molecola iniziale bensì il metabolita [O'Brien 1991]. Vista la difficoltà di tenere conto di tutti i fattori in gioco e il peso che il giudizio esperto può avere nella definizione di un "dominio meccanicistico", quale ausilio sono stati sviluppati anche dei software commerciali utili per la identificazione di potenziali tossicofori, modulatori e metaboliti (*Derek*, *Hazard Expert*, *Multicase*, *Topkat*, *Catabol*, *MetabolExpert*, *METEOR*, *META* ecc.).

Un modo alternativo modo di definire l'*AD* è quello di basarsi sul campo di esistenza delle grandezze chimico fisiche utilizzate come descrittori. Affinché il modello *QSAR* possa essere applicato ad un *chemical* per esso devono esistere (e si devono poter misurare o calcolare) gli *MD*. Più in generale si può definire l'*AD* in base allo spazio dei valori coperti dai descrittori del modello *QSAR*, detto spazio di interpolazione del modello.[11] Nel caso di un modello monovariato esso rappresenta l'intervallo compreso tra il valore minimo e massimo assunti dal descrittore. Nel caso di modelli con più *MD* (multivariati) lo spazio di interpolazione viene calcolato tramite tecniche di analisi statistica. La definizione dell'*AD* tramite lo spazio di interpolazione fa parte dell'approccio classico alla *QSAR*. Nel seguito verranno presentati alcuni metodi usati per definire lo spazio di interpolazione di un modello *QSAR* multivariato.

1. Confronto diretto. Il metodo più semplice è quello di farlo coincidere con il range dei valori assunti dagli *MD*. Questa soluzione si basa sulla ipotesi, non sempre vera, che i valori assunti dagli *MD* si distribuiscano secondo una normale e che la relazione *QSAR* esista per ogni valore assunto dagli *MD* (non ci siano spazi vuoti). Inoltre il metodo non tiene conto della eventuale correlazione tra *MD*.

[11]Nella linea guida OECD 2007 i termini "*applicability domain*", "*interpolation space*", "*descriptor space*" sono usati quale sinonimo del *AD*, inteso come lo spazio dei valori assunti dai descrittori, questo può generare delle ambiguità in quanto il concetto di "*AD*" di un metodo *QSAR* appare più correttamente riferito ai *chemical* per i quali un modello ha la capacità di fare delle stime di un *endpoint* definito [Netzeva *et al*. 2005].

2. Approccio geometrico [Jaworska *et al*. 2005]. Si può definire l'*AD* di un modello *QSAR* in base al calcolo della distanza tra un *chemical* indagato e un punto definito nello spazio dei descrittori del modello (ricavato con una delle tecniche statistiche di *cluster analysis*) [Sheridan 2000; Jaworksa *et al*. 2005; Stanforth *et al*. 2007; Wendt *et al*. 2008*;* Todeschini *et al*. 2009]. Dipendentemente dal metodo adottato, la distanza può essere definita in vari modi: "*euclidean*",[12] "*Mahalanobis*",[13] "*Manhattan*"[14] *distance*. Il principale vantaggio di questo approccio è che esso permette di definire degli intervalli di confidenza associati all'*AD* in base al luogo geometrico dei punti equidistanti da un punto predefinito nello spazio di interpolazione del metodo. Presenta invece lo svantaggio di richiedere che i dati di partenza giacciano secondo una distribuzione normale, essi sono infatti considerati solo in base alla loro distanza dal punto e non alla numerosità della popolazione di ciascuna regione. É opportuno sottolineare che la massima distanza accettabile è definita arbitrariamente dall'operatore.

Per effettuare l'analisi della distanza nell'approccio geometrico vengono normalmente utilizzate tecniche statistiche che sono utili non solo alla definizione dell'*AD*, secondo l'approccio classico, ma anche in fase di validazione del modello. Uno di questi metodi è rapresentato dal test di Hotelling e dall'analisi della *leverage*[15] associata [Gramatica *et al*. 2004 Pavan *et al*. 2005]. Questa esprime la distanza di un *chemical* dal centroide della proiezione delle molecole del *training set* sul piano *xy* della *QSAR*. È importante sottolineare che il significato del *leverage* varia se si sta

[12]La *Euclidean distance è* la linea dritta che unisce due punti in un piano. Se $P_1(x_1,y_1)$ e $P_2(x_2,y_2)$ sono due punti del piano, la loro distanza è $d^E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

[13]La *Mahalanobis distance* è la distanza tra due vettori $d^M(\vec{x},\vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$ dove *S* è la matrice covarianza [Mahalanobis 1936].

[14]La *Manhattan distance* è la distanza tra due punti in un piano misurata lungo assi ortogonali. Se $P_1(x_1,y_1)$ e $P_2(x_2,y_2)$ sono due punti del piano $d^{Man} = |x_1 - x_2| + |y_1 - y_2|$.

[15]*Leverage* di un punto in una regressione multivariata è la sua capacità di influenzare la regressione pur giacendo fuori di essa, una misura della *leverage* è data dalla distanza di Mahalanobis. Secondo una regola empirica viene in genere stabilita una soglia di attenzione per la *leverage* pari a 3*p*/*n* dove *p* è il numero dei descrittori più uno e *n* il numero dei *chemical* del *training set*. Se la *leverage* supera tale valore è considerata grande.

costruendo il modello *QSAR* o se lo si sta applicando. Con specifico riferimento al primo caso, se un *chemical* ha un valore di *leverage* alto, significa che esso influenzerà la regressione costringendola a passargli vicino pur senza apparire un *outlier* dal punto di vista statistico. In questo caso può essere preferibile escluderlo dal *training set*. Se, alternativamente, il modello è utilizzato per fare delle stime di un endpoint, un *chemical* che abbia un alto valore di *leverage* sarà probabilmente situato al di fuori dello "spazio dei descrittori" del modello e conseguentemente il valore di *endpoint* per esso stimato non sarà affidabile .

2. 3 Ipotesi di meccanismo (*Aryl Hydrocarbon Receptor*)

Come più sopra descritto per l'*AD* di un modello *QSAR* si possono usare due approcci: statistico e meccanicistico. In questo lavoro di tesi si è optato per l'adozione dell'approccio meccanicistico che consente di raggruppare le molecole nel dominio in base al *MOA*. Tutti i modelli *QSAR* sviluppati sono pertanto fondati su di una particolare ipotesi di meccanismo. È stato ipotizzato che nella tossicità delle molecole giochi un ruolo determinante uno specifico recettore, l'*Aryl Hydrocarbon Receptor* (*AhR*) [McKinney 1985; McKinney *et al*. 1985; Mekenyan *et al*. 1996; Okey 2007]. *AhR* è un *transcription factor*, la cui esatta struttura non è stata ancora identificata,[16] che fungerebbe da *chaperon* per piccole molecole organiche, dotate di anelli aromatici, trasportandole all'interno della cellula attraverso la membrana nucleare fino al *DNA*. *AhR* è solubile, *ligand dependent* e media molti degli effetti biologici o tossicologici di *chemicals* idrofobi. L'esposizione ad agonisti di *AhR* genera una quantità di effetti, il più studiato dei quali è l'espressione di geni. In figura 2.2 viene riportato il meccanismo proposto da Denison [Denison *et al*. 2003].

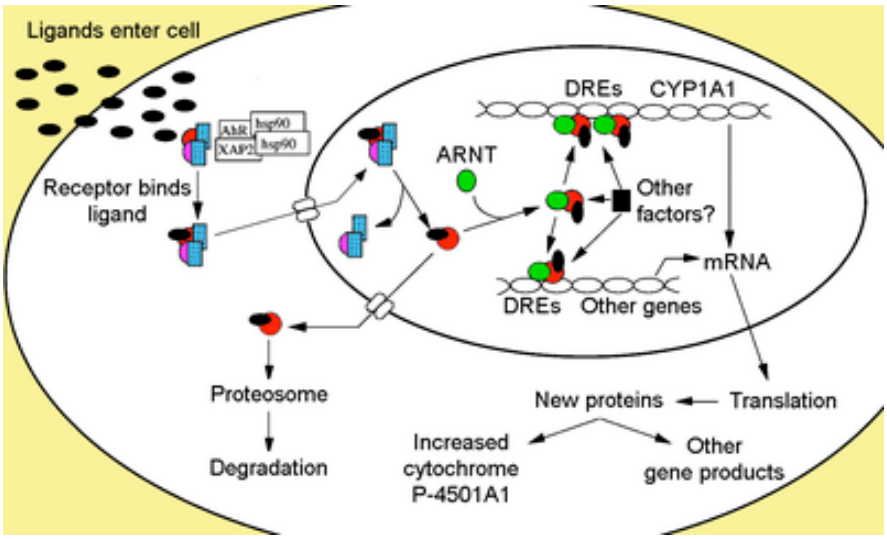

Figura 2.2. Meccanismo proposto per *AhR transcription factor*

[16]La struttura non risulta fra quelle inserite nel *Protein Data Bank* (*PDB*) *http://www.rcsb.org/pdb/home/home.do*

Inoltre, utilizzando l'approssimazione del *minimal cut set,*[17] si è assunto che fosse sufficiente valutare l'interazione di una molecola con *AhR* per spiegarne la tossicità.

Un'approssimazione simile al *minimal cut set* è la *elementary analysis of toxicity*. Quest'ultima presuppone che il modo di azione[18] di una molecola tossica possa essere scomposto in un certo numero di eventi tossicocinetici e cellulari che possono essere identificati in modo quantitativo all'interno del modello proposto per la tossicità di un *chemical* [Walum *et al*. 1992; De Jongh *et al*. 1999].

La tossicità è un fenomeno complesso, la prima domanda cui occorre rispondere è se esista un *rate determining step* e se questo possa essere identificato con l'interazione con *AhR*. Gli idrocarburi policiclici aromatici (PAHs) sono cancerogeni riconosciuti fin dai primi anni 60 [Dao *et al*. 1959; Huggins *et al*. 1961]. Questo ha suscitato interesse per il *MOA* della loro tossicità. Huggins scoprì che PAHs quali il 7,12-dimethylbenz-[a]-anthracene (DMBA) e il 3-methylcholanthrene (3-MC) causavano il cancro alla mammella. Riguardo al loro meccanismo d'azione esistono due scuole di pensiero. Secondo la prima i PAHs erano cancerogeni perché imitavano l'azione degli estrogeni (altri cancerogeni della mammella), mentre la seconda attribuiva la cancerogenicità degli estrogeni a loro metaboliti che agivano come i PAHs. In realtà, entrambe le ipotesi sono risultate corrette [Belous *et al*. 2007].

Studiando l'induzione di attività sul gene *CYP1A1* (noto anche come *Aryl Hydrocarbon Hydroxylase*), Nebert e Bausserman [Nebert *et al*. 1970] proposero: " *the process of hydroxylase induction involves a rate determining step, which may be the saturation of inducer-binding sites in the cell*". Poland [Poland *et al*. 1976], studiando la *2,3,7,8- tetrachlorodibenzo-p-dioxin* (*TCDD*), propose l'esistenza di un *induction receptor* per la *TCDD*, recettore che si sarebbe rivelato poi

[17] *Minimal cut set* è la minima catena di eventi per cui si possa verificare il guasto di una macchina, senza uno degli elementi della catena il guasto non avviene o avviene con modalità diverse, l'ipotesi permette di stimare la probabilità di guasto concentrando l'attenzione su uno o pochi eventi necessari al suo accadere.

[18] Il modo di azione è in questo caso traduzione di *mode of action* (*MOA*), riferito sia al modo di azione sia al meccanismo di azione.

essere *AhR* [Okey *et al*. 1979]. Esiste quindi un recettore, *AhR*, fondamentale per spiegare la tossicità di *TCDD*. Negli anni 90 si scoprì che *AhR*, pur comportandosi come un recettore di steroidi, mancava dello *zinc finger domain* tipico di tali recettori [Okey 2007]. Basandosi sul modello dei recettori di steroidi, anche il complesso *AhR*-legante avrebbe dovuto attraversare la membrana nucleare (come descritto in figura 2.2.). Nel tentativo di determinarne la struttura, Denison [Denison *et al*. 1986] ha identificato *AhR* quale un complesso macromolecolare di 250-280 kDa, che in determinate condizioni può scindersi in parti più piccole con massa di circa 120 kDa. Gli esperimenti di Denison portarono ad identificare macromolecole *AhR* simili ma leggermente diverse nei ratti e nelle cavie. Inoltre, le analisi di *AhR* presente nel *cytosol* e all'interno della membrana nucleare evidenziarono che dentro il nucleo esso è associato ad una proteina ulteriore denominata *Aryl Hydrocarbon Receptor Nuclear Translocator* (*ARNT*) di cui esiste una struttura in *Protein Data Bank* (*PDB*). Per contro, fuori dal nucleo *AhR* è associato ad una serie di proteine *chaperon* (*Hsp90, ARA9, p23*).[19] Secondo il meccanismo proposto per la interazione di *AhR* con la *TCDD* [Okey 2007], *AhR* risiederebbe nel *cytosol* legato ad Hsp90 e altre proteine *chaperon*, il legame con *TCDD* causerebbe la dissociazione delle proteine *chaperon*, il trasloco nel nucleo e la dimerizzazione di *AhR* con *ARNT*. Il complesso *TCDD-AhR-ARNT* sarebbe poi in grado di legarsi ad enzimi specifici (*AHRE*) presenti in numerosi geni. Gli enzimi, una volta occupati dal complesso *AhR-TCDD*, indurrebbero la sintesi di *mRNA* specifico. All'interno del nucleo il complesso *TCDD-AhR-ARNT* deve essere riconosciuto da siti specifici in modo da poter regolare l'espressione del gene. Denison identificò la sequenza nucleotidica specifica, denominata *AH Response Element* (*AHRE*), cui si lega il complesso sul gene *CYP1A1* [Denison *et al*. 1988 a) e b)]. *AHRE* è presente

[19]Il complesso citoplasmico di *AhR* consisterebbe di due molecole di Hsp90, almeno una di ARA9 (conosciuta anche come *Hepatitis B Virus-X associated protein 2*) ed il *cochaperone* p23 [Flaveny *et al*. 2009]. *Heat Shock Proteins* sono una famiglia di proteine, presenti in molte specie viventi, che nelle cellule in condizioni normali agiscono da *chaperones* nell'assemblaggio e nel *folding* nuove proteine. Sono così chiamate perché in condizioni di stress, quale uno shock termico, la loro produzione aumenta, il numero è invece associato alla massa espressa in kDa [Donati *et al*. 1990; Pockley 2003; Otaka *et al*. 2006].

in almeno 36 geni nei genomi di topo, ratto e uomo. Harper [Harper *et al*. 2006] ha dimostrato che *AhR in vivo* degrada secondo una moltitudine di fattori difficili da controllare e non solo per la presenza dello xenobiotico eventuale.

La domanda cui rispondere è da cosa venga regolato il "regolatore" (*AhR*). Riassumendo, dalla letteratura fin qui citata, su *AhR* si ottengono le seguenti informazioni:

- *AhR* è presente in molte specie, uomo compreso.
- In *Protein Data Bank* non ne è stata registrata la struttura tridimensionale.
- Il *binding site* di *AhR* è rigido e questo portò inizialmente ad ipotizzare che richiedesse molecole altamente planari, successivamente si scoprì che *AhR* può legare molecole di diverse forme e proprietà chimiche per cui viene oggi definito un recettore "promiscuo".
- *AhR* può legare esogeni strutturalmente diversi, si veda una lista in [Denison *et al*. 2003], è un recettore promiscuo ma non universale per tutti i contaminanti ambientali.
- Alta affinità con *AhR* non comporta alta tossicità, scarsa affinità con *AhR* indica una bassa probabilità di penetrare la membrana nucleare.
- Pur se la struttura 3D di *AhR* non è nota, se ne conosce l'organizzazione generale dei domini che è riportata nella figura 2.3 [Okey 2007]. Ipotesi sulla struttura del dominio PAS-B, interessato dall'interazione con *TCDD*, sono state fatte per analogia con la proteina *ARNT* [Pandini *et al*. 2007].

Posto che *AhR* si è rivelato essenziale per lo studio della tossicità della *TCDD*, due sono le domande che ci si è posti.

1. L'interazione con *AhR* può rappresentare quel passaggio senza il quale la tossicità dei composti *dioxin like* non si può esprimere (ipotesi del *minimal cut set*)?
2. Quali caratteristiche deve avere una molecola per poter interagire con *AhR*, è possibile costruire un modello *QSAR* sulla base di queste?

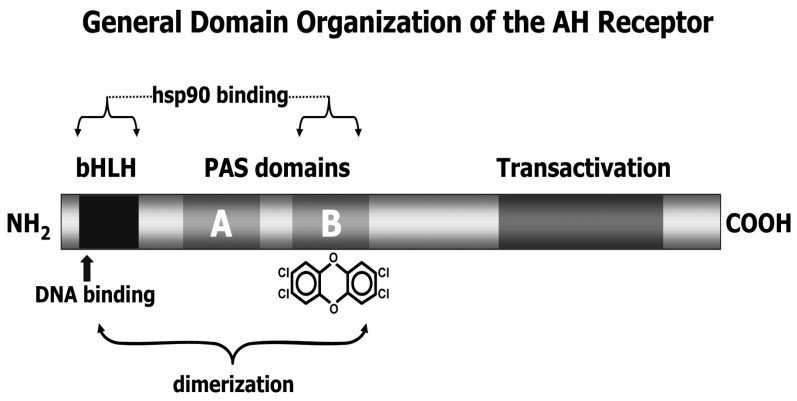

Figura 2.3. Organizzazione generale dei domini di *AhR*

In base agli studi citati da Okey [Okey 2007] la riposta alla prima domanda è affermativa. L'interazione con *AhR* è un passaggio fondamentale per spiegare la tossicità delle diossine, anche se essa da sola non è una garanzia di tossicità. La risposta alla seconda domanda è invece più complessa. Okey afferma che vi è un "*...astonishing range of AhR ligands*", per cui è difficile individuare una "struttura chimica", unica, caratterizzante tutte le molecole in grado di interagire con *AhR*. Si è deciso pertanto di focalizzare l'attenzione sulle proprietà elettroniche, che possono essere ben descritte su base quantomeccanica e con costi computazionali relativamente modesti mediante l'uso della *DFT*. Sempre secondo Okey, la tossicità delle diossine può essere considerata quale conseguenza di una aberrazione nell'espressione dei geni. Le diossine possono comunque interagire con centinaia di geni e non si conosce quali di questi vengano regolati da *AhR*, vi è quindi una difficoltà oggettiva nel determinare quale sia l'*endpoint* più adatto a gettare luce sul *MOA* della molecola potenzialmente tossica.

2. 3 Ricerca di dati tossicologici

Nonostante le incertezze sul ruolo svolto da *AhR* quale mediatore della tossicità di molecole *dioxin like*, la possibilità di una interazione con questo recettore è stata utilizzata quale criterio per definire il dominio chimico[20] dei modelli *QSAR* sviluppati. Si è cercato pertanto di reperire in letteratura dati tossicologici relativi a piccole molecole aromatiche di uso industriale da utilizzare quali *training set* per i modelli. La disponibilità di dati tossicologici per un certo numero di molecole da usare come *training set* è una condizione indispensabile per lo sviluppo di un modello *QSAR*. Una delle prime difficoltà con cui ci si è dovuti confrontare è stata la carenza di informazioni tossicologiche per un numero adeguato di molecole, utili cioè a definire un *training set* nell'ambito dello stesso meccanismo. Infatti, la maggior parte dei dati di letteratura riguardano molecole di interesse farmaceutico mentre, per le molecole di uso industriale, i dati sono molto più rarefatti. É opportuno sottolineare che uno degli obbiettivi del REACH è proprio quello di colmare questa lacuna.

La definizione del *training set,* e a maggior ragione dell'*AD*, ha una influenza enorme sui risultati ottenuti dal modello *QSAR*, per cui sarebbe assolutamente necessario poter disporre di dati di tossicità riproducibili e affidabili. In pratica questo è molto difficile! La definizione stessa del fenomeno tossicità è sfuggente e dipende molto dal modo in cui la si misura. In igiene industriale la definizione della tossicità oggi accettata si basa sul *no observable adverse effect level* (*NOAEL*). La dose per cui non si osservano effetti dannosi è, per sé stessa, una definizione "al di sotto del limite di rivelabilità del metodo" che, ovviamente, cambia con la sensibilità del metodo di analisi. In letteratura sono presenti molti metodi, *test,* saggi, indici per "misurare" la tossicità verso *endpoint*

[20]Con il termine dominio chimico s'intende l'insieme di tutte le molecole per le quali si ritiene possa essere applicato il modello *QSAR*. Analogamente possono essere definiti il dominio biologico (o tossicologico) come l'insieme degli *endpoint* per i quali le previsioni del modello sono ritenute affidabili, e il dominio dei descrittori l'insieme degli *MD* per i quali la *QSAR* esiste. Essi rappresentano le tre componenti dell'*AD* della *QSAR*, le prime due oggetto di discussione di questo capitolo la terza del quarto.

diversi, non direttamente confrontabili tra loro. Incidentalmente, la linea guida [OECD 2007] riporta un elenco degli *endpoint* più comunemente utilizzati per lo sviluppo di *QSAR*.

In questo lavoro si è fatto uso di banche dati quali: *Toxnet* della *US National Library of Medicine;*[21] *CDC,* Registry of Toxic Effects of Chemical Substances (RTECS);[22] *Danish QSAR Database;*[23] *DSSTox* della *US Environmental Protection Agency* (*EPA*);[24] *ACutetox*.[25] Inoltre, una lista di pubblicazioni contenenti dati biologici alla base di modelli *QSAR* e divisi per recettore, si può trovare sul sito della *Biograf 3R*.[26] Le banche dati presentano difficoltà nel controllo della omogeneità dei dati per cui, per i modelli studiati, si è fatto uso di dati ricavati direttamente da articoli di letteratura [Villeneuve *et al*. 2000; Falandysz *et al*. 2001; Siraki *et al*. 2004; Meerts *et al*. 2001; Yan *et al*. 2006; Hall *et al* 1989].

L'adeguatezza dei metodi oggi usati in tossicologia per la valutazione della pericolosità verso l'uomo e l'ambiente dei *chemicals* è un tema assai controverso [Hartung 2009]. Posto che si voglia valutare la tossicità di un *chemical* nel suo complesso, non è chiaro quale sia l'*endpoint* tossicologico da considerare e quale sia il test da utilizzare per la misura. Esistono in letteratura molti test diversi [Eisenbrand 2002], non è immediato decidere quale fornisca il risultato più significativo da utilizzare quale *endpoint*, ma è ragionevole ritenere che la scelta debba essere legata al *MOA* ipotizzato. Una prima distinzione si ha tra test *in vivo* e *in vitro*. Ovviamente, un elenco esaustivo delle tecniche esula dallo scopo di questa tesi. Tuttavia, con riferimento alla letteratura si può osservare che i test *in vivo* utilizzati possono essere divisi in due categorie.

1 Test su eucarioti.

[21] *http://toxnet.nlm.nih.gov/*
[22] *http://www.cdc.gov/niosh/rtecs/default.html*
[23] *http://ecbqsar.jrc.it/*
[24] *http://www.epa.gov/ncct/dsstox/About.html*
[25] *http://www.acutetox.org/publications/index.php*
[26] *http://www.biograf.ch/index.php?id=home*

1.1 Mortalità (*LC50*)[27] di *Pimephales promelas* (*fathead minnow*), un pesce appartenente alla famiglia dei ciprinidi, molto diffuso in USA e Canada, utilizzato per la banca dati EPA che impiega l'*LC50* verso questa specie.

1.2 Inibizione della crescita (*IGC50*)[28] di *Tetrahymena Pyriformis*, un protozoo ciliato del genere *tetrahymena* molto comune in acqua dolce.

1.3 Inibizione della crescita (*IGC50*) di *Chlorella Vulgaris*, un genere di alghe monocellulari.

2 Test su procarioti.

2.1 Inibizione della crescita (*IGC50*) di *Escherichia Coli*, un batterio presente nell'intestino di animali a sangue caldo.

2.2 Inibizione della crescita (*IGC50*) di *Vibrio Fischeri*, un batterio gram-negativo dalle proprietà luminescenti che vive in simbiosi con varie specie marine.

Molto più complesso è il panorama dei metodi di valutazione della tossicità dei *chemicals* che fanno uso di tecniche *in vitro*. Esse sono così definite perché i test sono condotti su colture cellulari fuori dall'organismo vivente. Per una review critica dei metodi il lettore interessato può fare riferimento al già citato articolo di Eisenbrand [Eisenbrand *et al*. 2002]. I test *in vitro* sono i più usati e i motivi del loro successo sono il basso costo, rispetto ai test *in vivo*, e il fatto che sono *mechanism oriented*, forniscono cioè informazioni specifiche sul *MOA*. Essi possono essere divisi in tre categorie.

1) Test per la valutazione della citotossicità.
2) Test per la valutazione dei *cellular responses* (*genomics, transcriptomics, proteomics e functional responses*).
3) Test per la costruzione di modelli tossicocinetici, del metabolismo e identificazione di *biomarkers*.

[27] *LC50* concentrazione letale per il 50% dei soggetti, anche espresso come *LD50* dose letale per il 50% dei soggetti.
[28] *IGC50* concentrazione che inibisce la crescita del 50% dei soggetti.

Analogamente ai *test in vivo* e *in vitro* anche i *target* considerati dai test possono essere divisi in più categorie.

- Sistemi subcellulari: macromolecole, organelli cellulari (e.g. ribosomi, mitocondri), frazioni citoplasmatiche.
- Sistemi cellulari: cellule primarie, cellule geneticamente modificate, cellule in diversi gradi di trasformazione e differenziazione.
- Organi o tessuti interi.

Con il termine citotossicità si intende la capacità di un *chemical* di causare la morte di una cellula per necrosi[29] o apoptosi.[30] La maggior parte dei test *in vitro* sono riferiti alla necrosi cellulare, ma anche l'apoptosi gioca un ruolo molto importante in alcuni processi, quali la cancerogenesi, per cui sono stati sviluppati test specifici. Gli *endpoint* più usati per la valutazione della citotossicità riguardano la permeabilità attraverso la membrana cellulare, cambiamenti nella morfologia o replicazione cellulare, riduzione della funzione dei mitocondri, frammentazione del *DNA*.

L'effetto a livello cellulare di un *chemical* tossico si traduce molto spesso in un impatto nell'espressione dei geni. La misura dei fattori di trascrizione può rivelare l'effetto tossico prima del livello patologico. Con il termine genomica vengono comprese molte tecniche differenti tutte legate in qualche modo ad informazioni contenute nella cellula (*DNA*, *RNA*). Secondo il modello corrente le informazioni contenute nei geni vengono copiate nell'*RNA* messaggero (*mRNA*) e poi tradotte in proteine funzionali. I due approcci principali di analisi della espressione genetica sono:

- la generazione di mappe di espressione del *mRNA* (transcrittomica);
- analisi del profilo di espressione delle proteine (proteomica).

[29]Necrosi è la morte cellulare non programmata dovuta ad una qualche forma di stress o trauma cellulare.
[30]Apoptosi è la morte cellulare programmata, che richiede consumo di energia, ed è una forma di manutenzione delle cellule volta ad eliminare quelle danneggiate in modo irreparabile oppure infettate da un virus.

La trascrizione di *mRNA* può essere analizzata in tempo reale tramite la *polymerase chain reaction* (PCR), dopo la trascrizione inversa. Metodi più recenti sono l'analisi seriale della espressione genetica (SAGE) e lo sviluppo di *cDNA/oligonucleotide microarrays*. Quest'ultima tecnica permette l'analisi dell'espressione di oltre 10000 geni per singolo esperimento [Watson *et al*. 1998; Duggan *et al*. 1999; Graves *et al*. 1999].

Mentre il genoma di un organismo è in qualche modo fisso e caratteristico, il proteoma cambia secondo le condizioni ambientali e di salute, per cui può essere usato per monitorare gli effetti di un cambiamento dovuto alla presenza di un *chemical* esogeno. Tramite tecniche di *matrix assisted laser desorption/ionization* (*MALDI*) e *matrix assisted laser desorption/ionization time of flight* (*MALDI/TOF*), *electrospray* e spettroscopia di massa, si possono analizzare le proteine contenute in tessuti o cellule per mettere in evidenza gli eventuali cambiamenti.

Lo scopo principale dell'analisi effettuata con queste tecniche (soprattutto se in modo combinato tra loro) è quello di stabilire, sotto certe condizioni sperimentali, la mappa dell'espressione genetica indotta da un certo *chemical* per compararla con quella indotta da sostanze tossiche note che agiscono secondo lo stesso meccanismo. Si assume pertanto che molecole che agiscono con lo stesso meccanismo comportino i medesimi effetti sulla espressione genetica.

La presenza di un *chemical* genera all'interno della cellula un cambiamento nella risposta quale l'incremento di *heat shock protein* (*Hsp*), perdita di glutatione (*GSH*), incremento di specie ossidanti (*ROS*), induzione di *stress activated protein kinases* (*SAPK*s) e *glucose regulated proteins* (Grps). Questi possono essere usati quali *endpoint* in *QSAR*.

I modelli tossicocinetici descrivono l'assorbimento, distribuzione, metabolismo, eliminazione di xenobiotici in funzione della dose e del tempo nell'organismo. Possono essere divisi in due categorie.

1. *Databased compartment models* sono tipi di modelli in cui il corpo è usualmente rappresentato da un sistema che descrive accumulo, distribuzione, scambio e metabolismo.
2. *Physiologicallly-based toxicokinetic* (PB-TK) *models,* il corpo viene diviso in distretti sulla base delle conoscenze di fisiologia e viene costruito un modello basandosi su ipotesi di meccanismo.

I *biomarkers* sono invece tutte quelle alterazioni nell'espressione genetica, alterazioni metaboliche, o altro, che possono segnalare l'effetto indotto dalla presenza del *chemical* esogeno. Il nutrito elenco delle tecniche di valutazione del rischio tossicologico fin qui citato (non certamente esaustivo) contrasta con la scarsezza dei dati tossicologici presenti in letteratura utilizzabili per lo sviluppo di modelli *QSAR*, carenza che è una delle ragioni d'essere del REACH. Come ampiamente discusso da Dearden [Dearden *et al*. 2009], i dati presenti in letteratura presentano numerose criticità soprattutto in relazione alla loro qualità. Per poter essere utilizzati per una *QSAR* i dati tossicologici dovrebbero:

1. essere omogenei tra loro;
2. descrivere un *endpoint* appropriato;
3. essere per quanto possibile completi (ovvero rappresentare tutti i dati presenti in letteratura per le stesse molecole senza omissioni o contraddizioni);
4. essere adeguati (e.g. le concentrazioni vanno espresse in moli o equivalenti e non in mg/litro per molecole di peso diverso);
5. essere esenti da ambiguità o ripetizioni inutili (e dannose) di molecole omonime;
6. rappresentare un *range* di valori per l'*endpoint* non troppo stretto, così da poter mettere in evidenza eventuali tendenze della *QSAR* verso gli *MD*.

In generale, si potrebbe affermare che da un punto di vista dello studio del meccanismo siano più utili dati tossicologici ricavati *in vitro* con test opportuni, piuttosto che le *LD50* o *LC50* ottenute

in vivo. Non potendo disporre di dati provenienti da misure sperimentali effettuate in proprio, si è dovuto fare uso dei dati di attività biologica presenti in letteratura. La necessità di avere dati provenienti da un'unica fonte, così da avere la ragionevole certezza che sia stato usato lo stesso protocollo sperimentale, ha limitato enormemente la numerosità dei *training set* disponibili.

Nei training set dei modelli *QSAR* sono state inserite molecole per la cui tossicità in letteratura è chiamato in causa il recettore *AhR*. Sono stati presi in esame: 13 policloronaftaleni [Villeneuve *et al*. 2000; Falandysz *et al*. 2001]; 14 benzochinoni [Siraki *et al*. 2004]; 8 polibromodifenileteri [Meerts *et al*. 2001]; 28 benzeni sostituiti [Yan *et al*. 2006], 89 benzeni sostituiti [Hall *et al*. 1989].

Capitolo 3

Teoria del Funzionale Densità

3.1 L'equazione di Schrödinger

Nella meccanica quantistica la funzione d'onda, Ψ, determina in modo completo lo stato di un sistema fisico. Ciò significa che questa funzione, data in un certo istante, descrive non soltanto tutte le proprietà in quell'istante, ma ne definisce anche il comportamento in tutti gli istanti successivi [Landau *et al*. 1969]. Quanto appena detto si esprime con l'equazione di Schrödinger [Schrödinger 1926a-g] dipendente dal tempo

$$i\hbar\frac{\partial\Psi(t)}{\partial t} = \hat{H}\,\Psi(t) \qquad (3.1.1)$$

dove $\hat{H}$ è l'operatore Hamiltoniano che descrive il sistema e $\Psi(t)$ è la funzione d'onda totale dipendente dal tempo.

Per un sistema isolato, atomico o molecolare, in cui sono presenti *N* elettroni, l'Hamiltoniano non contiene esplicitamente il tempo, in quanto, relativamente a tale sistema fisico, tutti gli istanti sono equivalenti. Perciò, quando non si è in presenza di interazioni tra atomi o molecole che dipendano dal tempo, si usa l'equazione di Schrödinger indipendente dal tempo [Schrödinger 1926a-e], che nella forma non relativistica e nell'ambito dell'approssimazione di Born-Oppenheimer [Born *et al*. 1927] è data da

$$\hat{H}_0\,\Phi = E\Phi \qquad (3.1.2)$$

dove E rappresenta l'energia elettronica totale, $\Phi = \Phi(\mathbf{x}_1,\mathbf{x}_2,.....,\mathbf{x}_N)$[31] la funzione d'onda e $\hat{H}_0$ l'Hamiltoniano indipendente dal tempo, definito dall'espressione[32]

[31] $\mathbf{x}_i$ include sia le coordinate spaziali **r** che le coordinate di *spin* σ_i dell'i^{mo} elettrone.

[32] Questa equazione è espressa in unità atomiche: l'unità di lunghezza è il raggio di Bohr - $a_0 = 5.2918 \times 10^{-11}$ m; l'unità di carica è la carica dell'elettrone - $e = 1.6022 \times 10^{-19}$ C; e l'unità di massa è la massa dell'elettrone - $m = 9.1095 \times 10^{-31}$ kg.

$$\hat{H}_0 = \sum_{i=1}^{N}\left(-\frac{1}{2}\nabla_i^2\right) + \sum_{i=1}^{N}\hat{v}(\mathbf{r}_i) + \sum_{i<j}^{N}\frac{1}{r_{ij}} \qquad (3.1.3)$$

dove il primo termine del secondo membro definisce l'energia cinetica degli *N* elettroni; il secondo termine, esplicitabile come

$$\hat{v}(\mathbf{r}_i) = \sum_{\alpha}^{M}\frac{Z_\alpha}{\vec{r}_{i\alpha}} \qquad (3.1.4)$$

è il potenziale esterno agente sull'elettrone i^{mo} dovuto agli *M* nuclei, ciascuno dei quali caratterizzato da una carica α; il terzo ed ultimo termine è relativo alla repulsione interelettronica.

L'equazione (3.1.3) può essere riscritta in forma più compatta come

$$\hat{H}_0 = \hat{T} + \hat{V}_{ne} + \hat{V}_{ee} \qquad (3.1.5)$$

dove

$$\hat{T} = \sum_{i=1}^{N}\left(-\frac{1}{2}\nabla_i^2\right) \qquad (3.1.6)$$

è l'operatore energia cinetica,

$$\hat{V}_{ne} = \sum_{i=1}^{N} v(\mathbf{r}_i) \qquad (3.1.7)$$

è l'operatore relativo all'interazione Coulombiana nucleo – elettrone, e

$$\hat{V}_{ee} = \sum_{i<j}^{N}\frac{1}{r_{ij}} \qquad (3.1.8)$$

è l'operatore relativo all'interazione Coulombiana elettrone - elettrone. È opportuno sottolineare che deve sempre essere inclusa l'energia di repulsione nucleo-nucleo. Nell'ambito dell'approssimazione di Born-Oppenheimer, questa quantità è una costante, ed è data da

$$\hat{V}_{nn} = \sum_{\alpha<\beta}^{M} \frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} \qquad (3.1.9)$$

L'equazione di Schrödinger (3.1.2) deve essere risolta con opportune condizioni al contorno:

i) per un atomo/molecola, Φ deve andare a zero all'infinito;

ii) per un solido cristallino ideale, Φ deve obbedire a specifiche condizioni di periodicità;

iii) $|\Phi|^2$ è una distribuzione di probabilità, nel senso che $\left|\Phi\left(\mathbf{r}^N, s^N\right)\right|^2 d\mathbf{r}^N$ corrisponde alla probabilità di trovare il sistema con coordinate spaziali comprese tra $\mathbf{r}^N$ e $\mathbf{r}^N + d\mathbf{r}^N$ e coordinate di *spin* uguali a s^N; $d\mathbf{r}^N = d\mathbf{r}_1 d\mathbf{r}_2 \ldots d\mathbf{r}_N$; $\mathbf{r}^N$ e s^N indicano l'insieme $\mathbf{r}_1$, $\mathbf{r}_2$,…, $\mathbf{r}_N$ e s_1, s_2,…, s_N, rispettivamente;

iv) le coordinate spaziali sono continue mentre quelle di *spin* sono discrete;

v) la funzione d'onda F deve essere antisimmetrica rispetto allo scambio delle coordinate di una qualsiasi coppia di elettroni (principio di antisimmetria).

Data Φ, il valore di aspettazione di qualsiasi quantità fisica è ottenibile da espressioni del tipo [Landau *et al*. 1969]

$$\left\langle \hat{A} \right\rangle = \frac{\int \Phi^* \hat{A} \Phi dx}{\int \Phi^* \Phi dx} = \frac{\langle \Phi | \hat{A} | \Phi \rangle}{\langle \Phi | \Phi \rangle} \qquad (3.1.10)$$

dove $\hat{A}$ è l'operatore Hermitiano[33] lineare associato ad una quantità fisica *A*. Se Φ è normalizzata i valori di aspettazione dell'energia cinetica e potenziale sono dati da

$$T[\Phi] = \langle \hat{T} \rangle = \int \Phi^* \hat{T} \Phi dx \qquad (3.1.11)$$

e

$$V[\Phi] = \langle \hat{V} \rangle = \int \Phi^* \hat{V} \Phi dx \qquad (3.1.12)$$

dove le parentesi quadre indicano che *T* e *V* sono funzionali di Φ.

È noto che gli unici sistemi atomici o "molecolari" per i quali siano disponibili soluzioni analitiche dell'equazione di Schrödinger (3.1.2) sono quelli caratterizzati dalla presenza di un singolo elettrone, mentre quelli non idrogenoidi necessitano l'impiego di metodi approssimati. I problemi legati alla soluzione di tale equazione sono dovuti alla presenza dell'operatore $\hat{V}_{ee}$ (3.1.8). A ciò si aggiunga che il moto degli elettroni è intimamente correlato, nel senso che, a causa della repulsione coulombiana, un elettrone tenderà ad evitare le regioni dello spazio in cui sia presente un altro elettrone. Spesso, si afferma che un elettrone crea intorno a sé un *buco di Coulomb,* caratterizzato da una minor probabilità di trovare un secondo elettrone. Incidentalmente, merita di essere enfatizzato il fatto che il principio di antisimmetria impone già una parziale correlazione nel moto degli elettroni aventi lo stesso *spin*. Per questi non è possibile occupare la stessa posizione spaziale, e ciascun elettrone si muove all'interno di un *buco di Fermi,* una regione in cui la probabilità di trovare un altro elettrone con lo stesso *spin* è nulla.

[33]Un operatore $\hat{A}$ è Hermitiano o autoaggiunto se $\hat{A}^+ = \hat{A}$. Per un operatore Hermitiano i diversi autovalori sono reali, e i diversi autovettori sono ortogonali.

3.2 Teoria del funzionale densità (*DFT*)

Durante gli ultimi quindici anni la Teoria del Funzionale Densità (*DFT*) non ha solo influenzato, ma si può dire anche rivoluzionato, l'applicazione dei principi della meccanica quantistica a sistemi complessi. Basata sui famosi teoremi di Hohemberg e Kohn [Hohenberg & Kohn (1964)], la *DFT* fornisce un punto di partenza per lo sviluppo di strategie computazionali che abbiano come fine ultimo quello di ottenere informazioni circa energie, struttura e proprietà di atomi e molecole con costi notevolmente inferiori rispetto alle tradizionali tecniche *ab-initio* [Geerlings *et al*. (2003)].

In questo paragrafo saranno descritti gli aspetti fondamentali di tale teoria a partire dal modello di Thomas-Fermi [Thomas (1927); Fermi (1928)], fino a giungere alla formulazione data da Kohn e Sham [Kohn & Sham (1965)].

3.2.1 Il modello di Thomas-Fermi (*TF*)

La teoria del funzionale densità affonda le proprie radici nel modello di Thomas-Fermi (*TF*) relativo agli atomi isolati [Thomas 1927; Fermi 1928; Wigner 1934; Weizsäcker 1935; Gombas 1949; March 1975; Lieb 1981]. *TF* partirono dal considerare la densità ρ_0 di un gas uniforme di elettroni liberi in funzione del momento di Fermi:[34]

$$\rho_0 = \frac{p_f^3}{3\pi^2\hbar^3} \qquad (3.2.1)$$

[34]Il momento di Fermi per il caso di livelli monoelettronici occupati ad alta energia è definito come $p_f = \hbar k_f$, dove $k_f = \left(3\pi^2 N/V\right)^{1/3}$ è il vettore d'onda di Fermi; l'energia corrispondente $\varepsilon_f = \hbar^2 k_f^2/2m$ è conosciuta come *energia di Fermi* [Ashcroft & Mermin (1976)].

ed hanno applicato tale relazione ad una situazione inomogenea come quella di solidi, atomi e molecole. L'importante risultato a cui giunsero è che l'energia elettronica di un sistema di *N* elettroni, $E[\Phi]$, è espressa come un funzionale, E_{TF}, della densità di carica $\rho(\mathbf{r})$:

$$E_{TF}[\rho(\mathbf{r})] = C_{TF}\int \rho^{\frac{5}{3}}(\mathbf{r})dr - \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + \frac{1}{2}\iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}d\mathbf{r}_1 d\mathbf{r}_2 \qquad (3.2.2)$$

L'aspetto rivoluzionario di tale approccio risiede nel fatto che per la prima volta l'energia dello stato fondamentale di un sistema risulta espressa in "funzione" di un osservabile come la densità elettronica, e non in funzione delle funzioni d'onda, che osservabili non sono.

È opportuno sottolineare che l'equazione (3.2.2), il cui primo membro di destra è il funzionale energia cinetica di *TF*, è una forma approssimata dell'espressione rigorosa per l'energia elettronica; infatti, il termine energia cinetica è qui espresso come funzionale della densità monoelettronica, e le sole interazioni prese in considerazione sono quelle elettrostatiche classiche di attrazione nucleo-elettrone e di repulsione elettrone-elettrone.

Da un punto di vista fisico la (3.2.2) implica che le proprietà elettroniche di un dato sistema possano essere determinate come funzionali della densità di carica applicando localmente relazioni appropriate ad un sistema omogeneo di elettroni liberi. Tale approssimazione è conosciuta come *local density approximation* (*LDA*).

La teoria *TF* fornisce una ragionevole descrizione della densità di carica in atomi pesanti. È infatti possibile dimostrare [Lieb *et al*. 1973] che tale teoria è esatta per il limite del numero atomico $Z \to \infty$. Tuttavia, essa fallisce se applicata a sistemi molecolari poiché incapace di predire l'esistenza di qualsiasi legame chimico: l'energia minima di un aggregato di atomi nell'ambito della teoria *TF* è sempre data dai nuclei posti a distanza infinita [Teller 1962].

3.2.2 I teoremi di Hohenberg e Kohn (*HK*)

Hohemberg e Kohn (*HK*) [Hohenberg *et al*. 1964] a soli due anni di distanza dal contributo di Teller, rivoluzionarono il mondo della chimica teorica mettendo in evidenza che il modello *TF* doveva essere considerato come la forma approssimata di una teoria esatta, nota oggi come *DFT*. Ciò avvenne mediante la formulazione di due teoremi, detti appunto teoremi di *HK*.

Il primo teorema di *HK* legittima l'uso della variabile $\rho(\mathbf{r})$ come variabile di base. Il teorema stabilisce quanto segue : "*Il potenziale esterno* $v(\mathbf{r})$ *è determinato, a meno di una costante additiva, dalla densità elettronica* $\rho(\mathbf{r})$." Poiché $\rho(\mathbf{r})$ determina anche il numero totale degli elettroni, attraverso l'espressione

$$\int \rho(\mathbf{r}) d\mathbf{r} = N \qquad (3.2.3)$$

ne consegue che essa determina anche la funzione d'onda Φ e così tutte le altre proprietà elettroniche del sistema.[35]

L'equazione (3.2.2) può essere riscritta utilizzando E_v al posto di E, esplicitandone la dipendenza da $v(\mathbf{r})$

$$E_v[\rho] = T[\rho] + V_{ne}[\rho] + V_{ee}[\rho] = \int \rho(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} + F_{\mathrm{HK}}[\rho] \qquad (3.2.4)$$

con

$$F_{HK}[\rho] = T[\rho] + V_{ee}[\rho]^{36} \qquad (3.2.5)$$

e

$$V_{ee}[\rho] = J[\rho] + \text{termini non classici} \qquad (3.2.6)$$

[35]Per la dimostrazione del primo e del secondo teorema di *HK* si faccia riferimento a [Parr & Yang (1989)].
[36]$F_{HK}[\rho]$ è indipendente dal potenziale esterno risultando così un funzionale universale di ρ.

dove $J[\rho]$ è il termine classico di repulsione coulombiana, mentre il contributo maggiore ai termini non classici è dato dall'energia di scambio e correlazione.

Il secondo teorema di *HK* fornisce il principio variazionale per l'energia stabilendo che, data una densità elettronica di prova $\tilde{\rho}(\mathbf{r})$ con $\tilde{\rho}(\mathbf{r}) \geq 0$ e ricordando la (3.2.3),

$$E_0 \leq E_v[\tilde{\rho}] \tag{3.2.7}$$

dove $E_v[\tilde{\rho}]$ è il funzionale energia dato dalla (3.2.4) ed E_0 l'energia esatta dello stato fondamentale.

Assunta la differenziabilità di $E_v[\tilde{\rho}]$, l'applicazione del principio variazionale dato dalla (3.2.7) implica che la densità elettronica dello stato fondamentale soddisfi il principio di stazionarietà dato da

$$\delta\{\mathrm{E}_v[\rho] - \mu[\int \rho(\mathbf{r})d\mathbf{r} - N]\} = 0 \tag{3.2.8}$$

che fornisce l'equazione di Eulero - Lagrange

$$\mu = \frac{\delta E_v[\rho]}{\delta\rho(\mathbf{r})} = v(\mathbf{r}) + \frac{\delta F_{HK}[\rho]}{\delta\rho(\mathbf{r})} \tag{3.2.9}$$

Nonostante l'importanza dei due teoremi di *HK*, è doveroso menzionare che il risultato da essi fornito è soltanto parziale. Il secondo, in particolare, è "semplicemente" un teorema di esistenza e complessivamente non fornisce alcuna indicazione su come costruire il funzionale energia per lo stato fondamentale. Tuttavia, l'esistenza di una teoria esatta giustifica la ricerca di sempre nuovi funzionali che, sebbene approssimati, siano sempre più accurati. Per una dettagliata descrizione dello sviluppo storico della *DFT* è possibile fare riferimento alle seguenti *review* , testi o tesi

[Williams *et* al. 1983, Callaway *et al*. 1984, Dahl *et al*. 1984, Jones 1987, Becke 1988a-b, Jones *et al*. 1989, Parr *et al*. 1989, Ziegler 1991].

3.2.3 Vantaggi della *DFT* e le equazioni di Kohn-Sham (*KS*)

Il modello di *TF*, descritto in precedenza, costituisce un approccio diretto; ossia si cerca di ottenere in modo esplicito forme approssimate del funzionale energia cinetica $T[\rho]$ e del funzionale che rappresenta l'interazione elettrone-elettrone $V_{ee}[\rho]$.[37] Questa procedura ha il vantaggio di avere equazioni che dipendono esclusivamente da ρ, ma non consente di andare oltre il livello di approssimazione insito nel modello di partenza.

Kohn e Sham (*KS*) hanno ideato nel 1965 [Kohn *et al*. 1965] un ingegnoso metodo indiretto per trattare il funzionale energia cinetica $T[\rho]$, rendendo in questo modo la *DFT* uno strumento utilissimo per calcoli rigorosi. *KS* proposero "semplicemente" di introdurre gli orbitali in modo tale che fosse possibile calcolare l'energia cinetica con buona approssimazione.

Dalla teoria della matrice densità [Parr *et al*. 1989] si ricava che l'espressione corretta per l'energia cinetica dello stato fondamentale è:

$$T[\rho] = \sum_{i=1}^{N} n_i \langle \psi_i | -\frac{1}{2}\nabla^2 | \psi_i \rangle \qquad (3.2.10)$$

dove ψ_i e n_i sono, rispettivamente, gli *spin* orbitali naturali ed i loro numeri di occupazione. Il principio di esclusione di Pauli [Pauli 1925] implica che $0 \le n_i \le 1$; inoltre, il primo teorema di *HK* ci assicura che l'energia cinetica sia un funzionale della densità di carica totale data da:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} n_i \sum_{s} |\psi_i(\mathbf{r},s)|^2 \qquad (3.2.11)$$

[37] Vedi equazioni (3.2.5) e (3.2.6)

Per un dato sistema ci possono essere un numero infinito di termini del tipo (3.2.10) e (3.2.11). *KS* partirono dal considerare espressioni più semplici quali:

$$T_s = \sum_{i=1}^{N} \langle \psi_i | -\frac{1}{2}\nabla^2 | \psi_i \rangle \qquad (3.2.12)$$

e

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} \sum_{s} |\psi_i(\mathbf{r},s)| \qquad (3.2.13)$$

dove $n_i = 1$ per N orbitali e $n_i = 0$ per i rimanenti. Le equazioni (3.2.12) e (3.2.13) sono casi speciali delle equazioni (3.2.10) e (3.2.11). La rappresentazione semplificata della densità di carica e dell'energia cinetica fornita da *KS* descrive esattamente un sistema di N elettroni non interagenti ai quali sia associata una singola funzione d'onda determinantale.

È possibile dimostrare [Parr *et al*. 1989] che, data una $\rho(\mathbf{r})$, continua, non negativa e normalizzata, questa può essere sempre decomposta secondo la (3.2.13). Il problema consiste però nell'ottenere un'unica decomposizione orbitalica tale da fornire un unico valore di $T_s[\rho]$.

In analogia con la definizione di *HK* del funzionale universale $F_{HK}[\rho]$, *KS* definirono un sistema di riferimento costituito da particelle non interagenti, con densità $\rho(\mathbf{r})$ per lo stato fondamentale e Hamiltoniano

$$\hat{H}_s = \sum_{i}^{N} \left(-\frac{1}{2}\nabla_i^2 \right) + \sum_{i}^{N} v_s(\mathbf{r}) \qquad (3.2.14)$$

che non include alcun termine repulsivo. Lo stato fondamentale di un simile sistema di riferimento è esattamente descritto da una funzione d'onda determinatale

$$\psi_s = \frac{1}{\sqrt{N!}}\det[\psi_1\psi_2\cdots\psi_N] \tag{3.2.15}$$

dove ψ_i sono i primi N autovettori dell'Hamiltoniano monoelettronico

$$\hat{h}_s\psi_i = \left[-\tfrac{1}{2}\nabla^2 + v_s(\mathbf{r})\right]\psi_i = \varepsilon_i\psi_i \tag{3.2.16}$$

L'energia cinetica è semplicemente $T_s[\rho]$, data dalla (3.2.12)

$$T_s[\rho] = \langle\psi_s|\sum_i^N\left(-\tfrac{1}{2}\nabla_i^2\right)|\psi_s\rangle = \sum_i^N\langle\psi_i|-\tfrac{1}{2}\nabla^2|\psi_i\rangle \tag{3.2.17}$$

e la densità elettronica è decomposta in accordo con la (3.2.13).

Prima di procedere, è opportuno sottolineare che $T_s[\rho]$ non corrisponde all'energia cinetica esatta $T[\rho]$. Il successo del metodo *KS* è che $T_s[\rho]$, il cui calcolo mediante la (3.2.17) è ora relativamente semplice, è esattamente l'energia cinetica presa in considerazione per il calcolo dell'energia dello stato fondamentale. Questo risultato può essere ottenuto scrivendo:

$$F[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho] \tag{3.2.18}$$

$$J[\rho] = \tfrac{1}{2}\iint\frac{1}{r_{12}}\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2 \tag{3.2.19}$$

$$E_{xc}[\rho] \equiv T[\rho] - T_s[\rho] + V_{ee}[\rho] - J[\rho] \tag{3.2.20}$$

dove la quantità $E_{xc}[\rho]$ prende il nome di energia di scambio e correlazione ed include la differenza fra $T_s[\rho]$ e $T[\rho]$, di solito piccola, ed i contributi non classici di $V_{ee}[\rho]$.

L'equazione di Eulero - Lagrange (3.2.9) diventa quindi*:*

$$\mu = v_{eff}(\mathbf{r}) + \frac{\delta T_s[\rho]}{\delta \rho(\mathbf{r})} \tag{3.2.21}$$

dove il potenziale effettivo di *KS* ($v_{eff}(\mathbf{r})$) è definito come segue

$$v_{eff}(\mathbf{r}) = v(\mathbf{r}) + \frac{\delta J[\rho]}{\delta \rho(\mathbf{r})} + \frac{\delta E_{xc}[\rho]}{\delta \rho(\mathbf{r})} = v(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathbf{d}r' + v_{xc}(\mathbf{r}) \tag{3.2.22}$$

con il potenziale di scambio e correlazione $v_{xc}(\mathbf{r})$ dato da

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho]}{\delta \rho(\mathbf{r})} \tag{3.2.23}$$

L'equazione (3.2.21), con il vincolo $\int \rho(\mathbf{r}) d\mathbf{r} = N$, è proprio l'equazione che si otterrebbe applicando la *DFT* ad un sistema di *N* elettroni non interagenti in un campo di potenziale esterno $v_s(\mathbf{r}) = v_{eff}(\mathbf{r})$. Questo significa che per un dato $v_{eff}(\mathbf{r})$, è possibile ottenere una $\rho(\mathbf{r})$ che soddisfi l'equazione (3.2.21) semplicemente risolvendo le *N* equazioni monoelettroniche

$$\left[-\tfrac{1}{2}\nabla^2 + v_{eff}(\mathbf{r}) \right] \psi_i = \varepsilon \psi_i \tag{3.2.24}$$

con il vincolo

$$\rho(\mathbf{r}) = \sum_i^N \sum_s |\psi_i(\mathbf{r}, s)|^2 \tag{3.2.25}$$

È importante sottolineare che $v_{eff}(\mathbf{r})$ dipende da $\rho(\mathbf{r})$ attraverso la (3.2.23) e di conseguenza le equazioni (3.2.22), (3.2.24) e (3.2.25) devono essere risolte in modo autoconsistente. Le equazioni (3.2.22) - (3.2.25) sono chiamate equazioni di *KS*.

Il funzionale energia (3.2.4) a questo punto può essere riscritto come

$$E[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho] + \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \qquad (3.2.26)$$

che, inserendo l'espressione di KS per l'energia cinetica, diventa

$$E[\rho] = \sum_{i}^{N} \sum_{s} \int \psi_i^*(\mathbf{r})\left(-\frac{1}{2}\nabla^2\right)\psi_i(\mathbf{r})d\mathbf{r} + J[\rho] + E_{xc}[\rho] + \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \qquad (3.2.27)$$

dove la densità elettronica è data dalla (3.2.25), così che il funzionale energia risulti espresso in termini di *N* orbitali.

3.2.4 Correzioni non locali

Per quanto riguarda il metodo di *KS* rimane il problema di definire il potenziale di scambio e correlazione funzionale della densità di carica. L'accuratezza dei calcoli *DFT* dipende in gran parte dalla precisione delle approssimazioni fatte per $v_{xc}(\mathbf{r})$ (3.2.23) e per il relativo $E_{xc}[\rho]$ (3.2.20).

La prima approssimazione del potenziale di scambio e correlazione è stata quella formulata nell'ambito della teoria di *TF* del gas omogeneo di elettroni, la *local density approximation* (*LDA*).

Successivamente sono stati elaborati schemi più complessi volti al superamento di tale approssimazione, permettendo di ovviare, almeno in parte, agli inconvenienti che essa comporta. Poiché l'errore compiuto nel calcolo di E_{xc} tramite l'approssimazione *LD* è principalmente ascrivibile alla componente di scambio, la maggior parte dei contributi si sono concentrati sulla ricerca di opportune correzioni per questa quantità [Langreth *et al*. 1983; Becke 1983; Perdew 1985; Perdew *et al*. 1986; Becke 1986; De Pristo *et al*. 1987; Becke 1988a-c; Tschinke *et al*. 1989; Perdew *et al*. 1989].

In tempi relativamente recenti, Becke [Becke 1983], ha ricavato una espressione non locale per l'energia di scambio in termini di gradiente della densità elettronica di *spin* (*generalized gradient approximation - GGA*). In questa espressione l'energia di scambio è calcolata sommando

al termine ottenuto dall'approssimazione *LD* le cosiddette correzioni non locali (E_x^{NL}), che dipendono dal gradiente della densità elettronica e sono quindi correlabili alla sua non omogeneità in un sistema reale.

$$E_x = E_x^{LDA} + E_x^{NL} \tag{3.2.28}$$

Le *gradient corrections* E_x^{NL} hanno la seguente forma generale:

$$E_x^{NL} = \sum_{\gamma} \int g\left(\chi^{\gamma}\right)\left[\rho_1^{\gamma}(\mathbf{r}_1)\right]^{\frac{4}{3}} d\mathbf{r}_1 \tag{3.2.29}$$

dove $g\left(\chi^{\gamma}\right)$ è una funzione del parametro adimensionale

$$\chi^{\gamma} = \frac{\left|\nabla\rho^{\gamma}(\mathbf{r}_1)\right|}{\left[\rho^{\gamma}(\mathbf{r}_1)\right]^{\frac{4}{3}}} \tag{3.2.30}$$

A conferma della notevole attenzione dedicata alla ricerca di *gradient corrections* in grado di fornire valori per l'energia di scambio (3.2.28) sempre più prossimi a quello esatto, va precisato che sono reperibili in letteratura diverse forme per tali correzioni non locali [Langreth *et al*. 1983; Perdew *et al*. 1986); Becke 1986; DePristo *et al*. 1987; Becke 1988a-c; Perdew *et al*. 1989].

Prima di procedere si ritiene opportuno mettere in risalto l'esistenza di alcuni casi per i quali l'introduzione delle correzioni non locali fino ad ora descritte non ha portato giovamento alcuno. A titolo di esempio è utile menzionare le proprietà di risposta che dipendono sensibilmente dal comportamento del potenziale di scambio e correlazione.[38] Una delle cause dell'inadeguatezza delle correzioni non locali nel caso appena citato è da ricercare nello scorretto comportamento asintotico

[38]Vedi paragrafo **3.3.2**, ed in particolare le equazioni (3.3.31), (3.3.32), (3.3.33).

del potenziale *LDA*, caratterizzato da un decadimento esponenziale, in contrasto con il decadimento reale che va come $-\frac{1}{r}$. A prescindere dal comportamento asintotico ci sono altri aspetti del potenziale che richiedono un'approssimazione più accurata come ad esempio il comportamento attorno alle *shell* di confine nella regione prossima al nucleo e nel guscio di valenza.

Per ovviare a tali problemi sono stati elaborati nuovi modelli nell'approssimazione del potenziale *xc*, tra cui il LB94 [Leeuwen *et al*. 1994], il SAOP [Gritsenko *et al*. 1999; Schipper *et al*. 1999] ed il GRAC [Grüning *et al*. 2000].

3.3 Sistemi dipendenti dal tempo e teoria dei responsi lineari

Lo studio di stati eccitati svolge un ruolo fondamentale, in particolar modo nella descrizione delle proprietà di transizione. Per questo motivo sono stati introdotti metodi per il calcolo di tali proprietà che si basano sulla dinamica del sistema, in altre parole metodi dipendenti dal tempo. Tali approcci sono legati alla descrizione delle proprietà di risposta (statiche e dinamiche) del sistema in seguito ad una perturbazione esterna. Si è interessati, quindi, a comprendere il modo in cui un sistema risponde ad una sollecitazione esterna, come il campo oscillante della radiazione elettromagnetica, o meglio, a conoscere come l'interazione radiazione-materia modifica un particolare stato.

In questo paragrafo, dopo un approfondimento della Teoria dei Responsi Lineari, verranno presentati i principi generali della teoria *DF* per sistemi dipendenti dal tempo.

3.3.1 Teoria dei Responsi Lineari

Da un punto di vista formale, si tratta di risolvere l'equazione di Schrödinger dipendente dal tempo (3.1.1) [Schrödinger 1926a-g], ricorrendo alla teoria delle perturbazioni [Atkins 1983].

L'operatore Hamiltoniano di un sistema quantomeccanico in presenza di una perturbazione è dato dall'espressione [McWeeny 1985; McWeeny 1989]:

$$\hat{H}(t) = \hat{H}_0 + \hat{H}'(t) \qquad (3.3.1)$$

dove $\hat{H}_0$ è l'Hamiltoniano del sistema indipendente dal tempo, prima di applicare la perturbazione, e $\hat{H}'(t)$ è l'Hamiltoniano della perturbazione, che rappresenta l'interazione tra il sistema ed un campo esterno variabile nel tempo.

Si può esprimere $\hat{H}'(t)$ come

$$\hat{\mathrm{H}}'(t) = \hat{\mathrm{A}}F(t) \qquad (3.3.2)$$

dove $F(t)$ è il campo esterno applicato, e $\hat{\mathrm{A}}$ è l'operatore Hermitiano indipendente dal tempo, che descrive le variabili del sistema accoppiate al campo $F(t)$. Nel caso in cui il campo perturbante sia il campo elettrico oscillante della radiazione elettromagnetica $\mathbf{E}(t)$, l'equazione (3.3.2) diventa $\hat{\mathrm{H}}'(t) = -\boldsymbol{\mu}\cdot\mathbf{E}(t)$ dove $\boldsymbol{\mu}$ è il momento di dipolo elettrico del sistema definito come $\boldsymbol{\mu} = \sum_i q_i \mathbf{r}_i$ dove q_i è la carica della i^{ma} particella nella posizione $\mathbf{r}_i$.

La funzione d'onda dipendente dal tempo $\Psi(t)$ può essere espressa in termini del set completo di autofunzioni $\{\psi_n\}$ dell'Hamiltoniano non perturbato $\hat{\mathrm{H}}_0$ attraverso un'opportuna combinazione lineare:

$$\Psi(t) = \sum_n c_n(t)\psi_n(t) \qquad (3.3.3)$$

dove $\psi_n(t)$ è

$$\psi_n(t) = \psi_n e^{-iE_n t} \qquad (3.3.4)$$

In linea di principio, quindi, la conoscenza delle funzioni d'onda degli stati elettronici stazionari per un sistema permette di ottenere informazioni sulle proprietà sia statiche che dinamiche di un sistema.

É conveniente considerare la funzione d'onda in termini di frequenze di eccitazione ω. A tale scopo viene introdotta la funzione $\tilde{\Psi}$:

$$\tilde{\Psi}(t) = e^{iE_0(t-t_0)}\Psi(t) = \sum_n c_n(t)e^{-i\omega_{n0}(t-t_0)}\psi_n \qquad (3.3.5)$$

dove $\omega_{n0} = E_n - E_0$ è una delle frequenze di eccitazione del sistema non perturbato. L'equazione di Schrödinger dipendente dal tempo (3.1.1) in termini di $\tilde{\Psi}$ diventa:

$$i\frac{\partial\tilde{\Psi}(t)}{\partial t}=(\hat{H}_0-E_0)\tilde{\Psi}(t)+\hat{H}'\tilde{\Psi}(t) \qquad (3.3.6)$$

che rappresenta l'evoluzione temporale di $\tilde{\Psi}(t)$. Per esplicitare l'equazione (3.3.6), è necessario valutare come variano i coefficienti dell'espansione (3.3.3) $c_n(t)$ rispetto al tempo. La variazione temporale di un generico c_k con $0 \le k \le n$ è:

$$\begin{aligned}\frac{dc_k(t)}{dt} &= \frac{1}{i}\langle\psi_k|\hat{H}'|\tilde{\Psi}\rangle e^{i\omega_{k0}(t-t_0)} = \\ &= \frac{1}{i}\sum_n c_n(t)e^{i\omega_{kn}(t-t_0)}\langle\psi_k|\hat{H}'|\psi_n\rangle\end{aligned} \qquad (3.3.7)$$

e quindi dipende da tutti i $c_n(t)$ dell'espansione di (3.3.5). Per determinare i diversi coefficienti, si ricorre ad approssimazioni successive attraverso l'introduzione di un parametro λ

$$c_k(t)=c_k^{(0)}(t)+\lambda c_k^{(1)}(t)+\lambda^2 c_k^{(2)}(t)+... \qquad (3.3.8)$$

Ciò corrisponde ad esprimere la funzione d'onda con un termine di ordine zero e con correzioni successive di primo ordine, secondo ordine e così via. Sostituendo la (3.3.8) nella (3.3.7) e separando secondo i diversi ordini di correzione, si ottengono una serie di equazioni

$$\begin{aligned}&(0) \qquad \frac{dc_k^{(0)}(t)}{dt} = 0 \\ &(1) \qquad \frac{dc_k^{(1)}(t)}{dt} = \frac{1}{i}\sum_n \langle\psi_k|\ \hat{H}'|\psi_n\rangle e^{i\omega_{kn}(t-t_0)}c_n^{(0)}(t) \\ &\qquad\qquad \vdots\end{aligned} \qquad (3.3.9)$$

Come si può notare i coefficienti di ordine zero, prima equazione, sono costanti e quindi non dipendono dal tempo; i loro valori esprimono le condizioni iniziali, specificando lo stato del sistema prima di applicare la perturbazione. Si suppone che inizialmente il sistema si trovi nello stato fondamentale ψ_0, perciò i coefficienti di ordine zero sono nulli, tranne $c_0^{(0)}(t)$, e quindi $c_n^{(0)}(t) = \delta_{n0}$. Integrando la seconda equazione della (3.3.9) si ottengono i coefficienti del primo ordine $c_k^{(1)}(t)$, che risultano

$$\begin{aligned} c_k{}^{(1)}(t) &= \frac{1}{i}\int_{t_0}^{t} dt' \langle \psi_k | \hat{\mathrm{H}}' | \psi_0 \rangle e^{i\omega_{k0}(t-t_0)} \\ c_0{}^{(1)}(t) &= \frac{1}{i}\int_{t_0}^{t} dt' \langle \psi_0 | \hat{\mathrm{H}}' | \psi_0 \rangle \end{aligned} \qquad (3.3.10)$$

In modo analogo, si può rappresentare la funzione d'onda $\tilde{\Psi}(t)$ per approssimazioni successive

$$\tilde{\Psi}(t) = \tilde{\Psi}^{(0)}(t) + \tilde{\Psi}^{(1)}(t) + \tilde{\Psi}^{(2)}(t) + \ldots \qquad (3.3.11)$$

e, separando i diversi ordini di correzione, si ottiene

$$\begin{aligned} &(0) \quad \tilde{\Psi}^{(0)}(t) = \psi_0 \\ &(1) \quad \tilde{\Psi}^{(1)}(t) = c_0{}^{(1)}(t)\psi_0 + \sum_{n\neq 0} c_n{}^{(1)} e^{-i\omega_{n0}(t-t_0)}(t)\psi_n \\ &\qquad\qquad \vdots \end{aligned} \qquad (3.3.12)$$

Lo scopo della Teoria dei Responsi Lineari è di determinare la risposta al primo ordine del valore di attesa di un qualche osservabile, rappresentato dall'operatore $\hat{B}$, in seguito alla perturbazione descritta dall'operatore $\hat{\mathrm{A}}$ (equazione 3.3.2).

Si può esprimere il valore di attesa di $\hat{B}$ come

$$\langle B\rangle = \langle \Psi(t)|\hat{B}|\Psi(t)\rangle = \langle \tilde{\Psi}(t)|\hat{B}|\tilde{\Psi}(t)\rangle \tag{3.3.13}$$

e, sostituendo l'equazione (3.3.11), si ottiene

$$\begin{aligned}\langle B\rangle &= \left\langle \tilde{\Psi}^{(0)}+\tilde{\Psi}^{(1)}+\tilde{\Psi}^{(2)}+..\middle|B\middle|\tilde{\Psi}^{(0)}+\tilde{\Psi}^{(1)}+\tilde{\Psi}^{(2)}+..\right\rangle = \\ &= \left\langle \tilde{\Psi}^{(0)}(t)\middle|B\middle|\tilde{\Psi}^{(0)}(t)\right\rangle + \left\langle \tilde{\Psi}^{(0)}(t)\middle|B\middle|\tilde{\Psi}^{(1)}(t)\right\rangle + \left\langle \tilde{\Psi}^{(1)}(t)\middle|B\middle|\tilde{\Psi}^{(0)}(t)\right\rangle + ...\end{aligned} \tag{3.3.14}$$

Si definisca la fluttuazione lineare del responso al tempo *t* come

$$\delta\langle B\rangle = \langle B\rangle - \langle B\rangle_0 \tag{3.3.15}$$

e, considerando le equazioni (3.3.14) e (3.3.12),

$$\begin{aligned}\delta\langle B\rangle &= \left\langle \tilde{\Psi}^{(0)}(t)\middle|B\middle|\tilde{\Psi}^{(1)}(t)\right\rangle + \left\langle \tilde{\Psi}^{(1)}(t)\middle|B\middle|\tilde{\Psi}^{(0)}(t)\right\rangle = \\ &= c_0^{(1)}(t)\langle\psi_0|B|\psi_0\rangle + \sum_{n\neq 0} c_n^{(1)}(t)\langle\psi_0|B|\psi_n\rangle e^{-i\omega_{n0}(t-t_0)} + \\ &+ c_0^{(1)*}(t)\langle\psi_0|B|\psi_0\rangle + \sum_{n\neq 0} c_n^{(1)*}(t)\langle\psi_n|B|\psi_0\rangle e^{i\omega_{n0}(t-t_0)}\end{aligned} \tag{3.3.16}$$

Ricordando che $c_0^{(1)*} = -c_0^{(1)}$, il primo ed il terzo termine della (3.3.16) si annullano. Inoltre, sostituendo al posto dei coefficienti $c_n^{(1)}$ e $c_n^{(1)*}$ le espressioni esplicite ottenute in precedenza (3.3.10), ne consegue che, in base alla (3.3.2)

$$\delta\langle B\rangle = \int_{-\infty}^{\infty} dt'\theta(t-t')K\left(BA|t-t'\right)F(t') \tag{3.3.17}$$

dove $\theta(t-t')K\left(BA|t-t'\right)$ è un propagatore, o meglio la forma temporale di un propagatore, con

$$K\left(BA|t-t'\right) = \frac{1}{i}\sum_{n\neq 0}\left[\langle\psi_0|B|\psi_n\rangle\langle\psi_n|A|\psi_0\rangle e^{-i\omega_{n0}(t-t_0)} - \langle\psi_n|B|\psi_0\rangle\langle\psi_0|A|\psi_n\rangle e^{i\omega_{n0}(t-t_0)}\right]$$

$$\theta(t-t') = \begin{cases} 1 & t > t' \\ 0 & t < t' \end{cases} \qquad (3.3.18)$$

dove $K\left(BA|t-t'\right)$ è la funzione di risposta lineare, che mette in relazione la fluttuazione di B ad un tempo *t* con l'intensità della perturbazione dovuta all'accoppiamento tramite A del campo F ad un tempo t' precedente; $\theta(t-t')$ è la funzione *STEP* di Heaviside, introdotta allo scopo di estendere l'integrale della (3.3.17). La funzione di risposta è definita solo per $t > t'$, in accordo con il principio di casualità, ed è funzione solo della differenza fra i due tempi, $\tau = t - t'$. Il responso dell'osservabile *B* del sistema, $\delta\langle B\rangle$, può essere visto come il risultato di un treno di impulsi $F(t')$ applicati a tutti i tempi da -∞ a *t*, dove il fattore di proporzionalità, $\theta(t-t')K\left(BA|t-t'\right)$, descrive come questi impulsi si propagano nel tempo producendo il loro effetto sul valore di attesa dell'osservabile $\hat{B}$,$\langle B\rangle$.

In alcuni casi è conveniente considerare, invece della (3.3.2), una perturbazione oscillante di frequenza w. Esprimendo, quindi, *F*(*t'*) in termini della sua trasformata di Fourier, l'Hamiltoniano della perturbazione, applicata in modo graduale attraverso un fattore di convergenza $e^{\eta t}$ con il limite per $\eta \to 0$, diventa

$$\hat{H}'(t) = \lim_{\eta\to 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega\, f(\omega) \frac{\hat{A}_\omega e^{-i(\omega+i\eta)t} + \hat{A}_{-\omega} e^{i(\omega-i\eta)t}}{2} \qquad (3.3.19)$$

In questo caso, il responso lineare $\delta\langle B\rangle$ diventa [McWeeny 1985; McWeeny 1989]

$$\delta\langle B\rangle_\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega f(\omega) \frac{1}{2}\left[\Pi\left(BA_\omega|\omega\right)e^{-i\omega t} + \Pi\left(BA_{-\omega}|-\omega\right)e^{i\omega t}\right] \qquad (3.3.20)$$

dove

$$\Pi\left(BA_\omega|\omega\right) = \lim_{\eta\to 0} e^{\eta t} \sum_{n\neq 0}\left(\frac{\langle 0|\hat{B}|n\rangle\langle n|\hat{A}_\omega|0\rangle}{\omega-\omega_{n0}+i\eta} - \frac{\langle 0|\hat{A}_\omega|n\rangle\langle n|\hat{B}|0\rangle}{\omega+\omega_{n0}+i\eta}\right) \qquad (3.3.21)$$

è la polarizzabilità[39] dipendente dalla frequenza (o suscettività) di B rispetto ad A. Esiste una relazione tra l'equazione (3.3.17) in termini di funzione di risposta lineare $K(BA|t - t')$ e la (3.3.20) in termini di polarizzabilità dipendente dalla frequenza $\Pi(BA_\omega|\omega)$; infatti il propagatore $\theta(t - t')$ $K(BA|t - t')$ e la polarizzabilità $\Pi(BA_\omega|\omega)$ sono trasformate di Fourier, vale a dire

$$\begin{aligned}\Pi\left(BA_\omega|\omega\right) &= \int_{-\infty}^{\infty} d\tau\theta\left(\tau\right)K\left(BA_\omega|\tau\right)e^{i\omega\tau} \\ \theta\left(\tau\right)K\left(BA_\omega|\tau\right) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\Pi\left(BA_\omega|\omega\right)e^{-i\omega\tau}\end{aligned} \qquad (3.3.22)$$

Per questo motivo, $\Pi\left(BA_\omega|\omega\right)$ viene denotata anche come espressione del propagatore in termini di frequenza. Infatti essa descrive come una perturbazione del tipo $A\cos\omega t$ si propaga e come viene rilevata dall'operatore B.

3.3.2 Teoria del funzionale densità dipendente dal tempo (*TD-DFT*)

Nel 1984 Runge e Gross [Runge *et al*. 1984] hanno posto le basi teoretiche per la Teoria del Funzionale Densità dipendente dal tempo (*TD-DFT*), dimostrando gli analoghi teoremi di *HK* [Hohenberg *et al*. 1964] per sistemi dipendenti dal tempo. In particolare, mostrarono che "*le densità* $\rho(\mathbf{r},t)$ *e* $\rho'(\mathbf{r},t)$ *di due sistemi che si sviluppano dallo stesso stato iniziale* $\Psi(t_0)$ *sotto l'influenza, rispettivamente, dei potenziali* $\upsilon(\mathbf{r},t)$ *e* $\upsilon'(\mathbf{r},t)$*, entrambi espansi secondo la serie di Taylor attorno al tempo finito* t_0 *e che differiscono per più di una funzione* $c(t)$ *dipendente solamente dal tempo, rimarranno sempre differenti*.". In altre parole, il potenziale esterno dipendente dal tempo, riferito a

[39]La polarizzabilità misura l'attitudine di una molecola a rispondere ad un campo elettrico $\mathbf{E}(t)$ e ad acquistare un momento di dipolo elettroco $\boldsymbol{\mu}$ [Atkins 1983].

una certa densità $\rho(\mathbf{r},t)$, è unico a meno di una funzione $c(t)$ che dipende solamente dal tempo, e determina la funzione d'onda totale, che a sua volta è unica a meno di un fattore di fase $\alpha(t)$:

$$\Psi(t)=e^{-i\alpha(t)}\tilde{\Psi}[\rho](t) \qquad (3.3.23)$$

Nel precedente paragrafo, si è visto che per calcolare le proprietà di risposta di un sistema (come la polarizzabilità dipendente dalla frequenza e le energie di eccitazione) è necessario ricorrere a metodi dipendenti dal tempo. Per questo motivo il formalismo TD-DFT*, che ha avuto origine dall'estensione dalla* DFT *a sistemi dipendenti dal tempo, è di notevole importanza. In quanto segue si forniranno alcuni dettagli sullo sviluppo di tale teoria. Più specificatamente, partendo dalle equazioni KS dipendenti dal tempo, si definirà il responso lineare della densità e si otterrà la soluzione perturbativa di queste equazioni.*

Analogamente al caso di un sistema indipendente dal tempo, si può scegliere come riferimento un sistema di elettroni non interagenti, ed introdurre un set di equazioni KS dipendenti dal tempo:

$$i\frac{\partial}{\partial t}\psi_j(\mathbf{r},t)=\left(-\frac{\nabla^2}{2}+\upsilon_s[\rho](\mathbf{r},t)\right)\psi_j(\mathbf{r},t) \qquad (3.3.24)$$

dove la densità è ottenuta dagli orbitali non interagenti

$$\rho(\mathbf{r},t)=\sum_{j=1}^{N}\left|\psi_j(\mathbf{r},t)\right|^2 \qquad (3.3.25)$$

Il potenziale $\upsilon_s[\rho](\mathbf{r},t)$ della (3.3.24) è generalmente chiamato potenziale *KS time-dependent* e viene definito come

$$\upsilon_s[\rho](\mathbf{r},t)=\upsilon(\mathbf{r},t)+\int\frac{\rho(\mathbf{r}',t)}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'+\upsilon_{xc}(\mathbf{r},t) \qquad (3.3.26)$$

dove $v(\mathbf{r},t)$ è il campo esterno e $v_{xc}(\mathbf{r},t)$ è il potenziale di scambio e correlazione dipendente dal tempo. Ora si consideri un potenziale esterno v_{ext} di forma

$$v_{ext}(\mathbf{r},t) = \begin{cases} v_0(\mathbf{r}) & per \quad t \le t_0 \\ v_0(\mathbf{r}) + v_1(\mathbf{r},t) & per \quad t > t_0 \end{cases} \tag{3.3.27}$$

In modo similare alla (3.3.11), è possibile espandere la densità $\rho(\mathbf{r},t)$ in serie di Taylor

$$\rho(\mathbf{r},t) = \rho^{(0)}(\mathbf{r}) + \rho^{(1)}(\mathbf{r},t) + \rho^{(2)}(\mathbf{r},t) + ... \tag{3.3.28}$$

dove $\rho^{(0)}(\mathbf{r})$ è la densità imperturbata per $t \le t_0$, che può essere ottenuta dalle equazioni *KS* per lo stato fondamentale con il potenziale $v_0(\mathbf{r})$, e la densità dipendente dal tempo al primo ordine $\rho^{(1)}(\mathbf{r},t)$, calcolata dall'esatta funzione χ di risposta lineare

$$\rho^{(1)}(\mathbf{r},t) = \int d\mathbf{r}' \int dt' \chi(\mathbf{r},t;\mathbf{r}',t') v_1(\mathbf{r}',t') \tag{3.3.29}$$

Incidentalmente, la funzione di risposta lineare è data dal funzionale

$$\chi(\mathbf{r},t;\mathbf{r}',t') = \left. \frac{\partial \rho[v_{ext}](\mathbf{r},t)}{\partial v_{ext}(\mathbf{r},t)} \right|_{v_0} \tag{3.3.30}$$

Nel caso del sistema *KS* di elettroni non interagenti, la densità al primo ordine è [Gross *et al*. 1996]

$$\rho^{(1)}(\mathbf{r},t) = \int d\mathbf{r}' \int dt' \chi_s(\mathbf{r},t;\mathbf{r}',t') v_{s,1}(\mathbf{r}',t') \tag{3.3.31}$$

dove χ_S è la funzione di risposta lineare non interagente, e $v_{s,l}$ è il potenziale *KS* al primo ordine del campo esterno, dato da

$$\upsilon_{s,1}(\mathbf{r},t) = \upsilon_1(\mathbf{r},t) + \int d\mathbf{r} \frac{\rho_1(\mathbf{r},t)}{|\mathbf{r}-\mathbf{r}'|} + \int d\mathbf{r} \int dt' f_{xc}[\rho_0](\mathbf{r},t;\mathbf{r}',t')\rho_1(\mathbf{r}',t') \quad (3.3.32)$$

con f_{xc} corrispondente alla derivata del potenziale $\upsilon_{xc}(\mathbf{r}, t)$ di scambio e correlazione dipendente dal tempo rispetto alla densità $\rho(\mathbf{r}, t)$

$$f_{xc}(\mathbf{r},t;\mathbf{r}',t') = \frac{\partial \upsilon_{xc}(\mathbf{r},t)}{\partial \rho(\mathbf{r}',t')} \quad (3.3.33)$$

Passando dal dominio del tempo a quello delle frequenze ω, si ottiene la funzione di risposta *KS* in termini di orbitali *KS* $\psi_j(\mathbf{r})$ imperturbati, i cui numeri di occupazione sono f_j[40] e le energie orbitaliche ε_j

$$\chi_s(\mathbf{r},\mathbf{r}',\omega) = \sum_{j,k} (f_k - f_j) \frac{\psi_j(\mathbf{r})\psi_k^*(\mathbf{r})\psi_j^*(\mathbf{r}')\psi_k(\mathbf{r}')}{\omega - (\varepsilon_j - \varepsilon_k) + i\eta} \quad (3.3.34)$$

ed η è un infinitesimo positivo. Se si prende in considerazione la risposta reale della densità di una molecola in un campo elettrico applicato, è possibile scegliere gli orbitali reali *KS* $\psi_j(\mathbf{r})$ e l'infinitesimo η può essere fissato a zero. Inoltre, si può osservare che, se *j* e *k* sono entrambi o occupati o virtuali, non ci sono contributi da *jk*. La funzione di risposta può essere quindi riscritta come

$$\chi_s(\mathbf{r},\mathbf{r}',\omega) = \sum_i^{occ} \sum_a^{virt} \psi_a(\mathbf{r})\psi_i^*(\mathbf{r})\psi_a^*(\mathbf{r}')\psi_i(\mathbf{r}') \times \left\{ \frac{2(\varepsilon_i - \varepsilon_a)}{(\varepsilon_i - \varepsilon_a)^2 + \omega^2} \right\} \quad (3.3.35)$$

Definita la funzione di risposta lineare in termini di orbitali reali di *KS* (3.3.35), si può ottenere la soluzione perturbativa delle equazioni di *KS* [Gisbergen 1998; Gisbergen *et al*. 1999].[41]

[40] I numeri di occupazione assumono i valori 0/1 per gli orbitali non occupati/occupati, rispettivamente.

[41] Verranno usati orbitali reali scartando i complessi coniugati e, sebbene all'inizio sarà mantenuto l'indice di *spin* σ, la discussione poi si limiterà al caso *spin-restricted* dove $\psi_{i_\uparrow}(\mathbf{r}) = \psi_{i_\downarrow}(\mathbf{r})$.

La densità al primo ordine della (3.3.31) può essere scritta in termini di prodotti tra *spin* orbitali *KS* occupati $\psi_{j\sigma}(\mathbf{r})$ e *spin* orbitali *KS* virtuali $\psi_{a\sigma}(\mathbf{r})$

$$\rho_{\sigma}^{(1)}(\mathbf{r},\omega) = \sum_{i,a}\left[P_{ia}^{\sigma}(\omega)\psi_{a\sigma}(\mathbf{r})\psi_{i\sigma}(\mathbf{r}) + P_{ai}^{\sigma}(\omega)\psi_{a\sigma}(\mathbf{r})\psi_{i\sigma}(\mathbf{r})\right] \qquad (3.3.36)$$

dove *P* è la matrice densità al primo ordine sulla base di autofunzioni. Come la densità di ordine zero dell'equazione (3.2.13)[42] contiene solo i prodotti degli orbitali occupati, così la densità al primo ordine può essere scritta esclusivamente in termini di prodotti tra orbitali occupati ed orbitali virtuali.[43] Per questa ragione, solo le componenti P_{ai} e P_{ia} sono diverse da zero e sono state incluse nella sommatoria.

Si può dimostrare che, in seguito all'espansione delle equazioni di *KS* (3.3.24) nel campo applicato, la densità al primo ordine può essere ottenuta dalla soluzione del seguente set di equazioni lineari per gli elementi della matrice densità P_{jb}^{σ} e P_{bj}^{σ} [Casida 1995; Bauernschmitt *et al.* 1996]

$$\sum_{jb\tau}\left[\delta_{\sigma\tau}\delta_{ij}\delta_{ab}(\varepsilon_{a\sigma} - \varepsilon_{i\sigma} + \omega) + K_{ia\sigma,jb\tau}\right]P_{jb}^{\tau} + \sum_{jb\tau}K_{ia\sigma,bj\tau}P_{bj}^{\tau} = -\left[\delta\upsilon_{ext}\right]_{ia\sigma} \qquad (3.3.37)$$

$$\sum_{jb\tau}\left[\delta_{\sigma\tau}\delta_{ij}\delta_{ab}(\varepsilon_{a\sigma} - \varepsilon_{i\sigma} - \omega) + K_{ai\sigma,bj\tau}\right]P_{bj}^{\tau} + \sum_{jb\tau}K_{ai\sigma,jb\tau}P_{jb}^{\tau} = -\left[\delta\upsilon_{ext}\right]_{ai\sigma} \qquad (3.3.38)$$

dove δ_{ij} è il delta di Kronecker, ω è la frequenza del campo applicato, $\varepsilon_{a\sigma}$ e $\varepsilon_{i\sigma}$ sono le energie degli *spin* orbitali. Gli elementi di matrice dei campi elettrici esterni sono dati da

$$\left[\delta\upsilon_{ext}\right]_{ia\sigma} = \left[\delta\upsilon_{ext}\right]_{ai\sigma} = \int d\mathbf{r}\psi_{i\sigma}(\mathbf{r})\delta\upsilon_{ext}(\mathbf{r})\psi_{a\sigma}(\mathbf{r}) \qquad (3.3.39)$$

[42]Altro non è che la densità *SCF* convergente di un ordinario calcolo *DFT* per lo stato fondamentale.
[43]Secondo la convenzione *a* denota un orbitale virtuale, mentre *i* un orbitale occupato.

e la matrice *K* delle equazioni (3.3.37) e (3.3.38), chiamata matrice di accoppiamento, è costituita da una parte coulombiana ed una parte di scambio e correlazione

$$K_{ij\sigma,kl\tau} = K^{Coul}_{ij\sigma,kl\tau} + K^{xc}_{ij\sigma,kl\tau} \tag{3.3.40}$$

dove

$$K^{Coul}_{ij\sigma,kl\tau}(\omega) = \int d\mathbf{r} \int d\mathbf{r}' \psi_{i\sigma}(\mathbf{r})\psi_{j\sigma}(\mathbf{r}) \frac{1}{|\mathbf{r}-\mathbf{r}'|} \psi_{k\tau}(\mathbf{r}')\psi_{l\tau}(\mathbf{r}') \tag{3.3.41}$$

e

$$K^{xc}_{ij\sigma,kl\tau}(\omega) = \int d\mathbf{r} \int d\mathbf{r}' \psi_{i\sigma}(\mathbf{r})\psi_{j\sigma}(\mathbf{r}) f^{\sigma\tau}_{xc}(\mathbf{r},\mathbf{r}',\omega)\psi_{k\tau}(\mathbf{r}')\psi_{l\tau}(\mathbf{r}') \tag{3.3.42}$$

Qui, è stata introdotta la trasformata di Fourier del cosiddetto *xc kernel* f_{xc}, che è la derivata del potenziale di scambio e correlazione dipendente dal tempo per elettroni di *spin* s $\upsilon^{\sigma}_{xc}(\mathbf{r},t)$ rispetto alla densità *time-dependent* per gli elettroni τ $\rho_{\tau}(\mathbf{r}',t')$

$$f^{\sigma\tau}_{xc}(\mathbf{r},\mathbf{r}',t-t') = \frac{\delta\upsilon^{\sigma}_{xc}(\mathbf{r},t)}{\delta\rho_{\tau}(\mathbf{r}',t')} \tag{3.3.43}$$

Questo *kernel* determina il cambiamento al primo ordine nel potenziale *xc time-dependent* dovuto alla perturbazione elettrica applicata.

Nell'ambito dell'approssimazione *LD* adiabatica, *ALDA* [Zangwill 1980; Zangwill *et al.* 1981], il funzionale della (3.3.43) è ridotto ad una funzione reale spazialmente locale, indipendente dalla frequenza, e valutata alla densità locale $\rho_0(\mathbf{r})$.

$$f^{ALDA,\sigma\tau}_{xc}(\mathbf{r},\mathbf{r}',\omega) = \delta(\mathbf{r}-\mathbf{r}') \frac{d\upsilon^{LDA,\sigma}_{xc}}{d\rho_{\tau}}\bigg|_{\rho_{\tau}=\rho_{0,\tau}(\vec{r})} \tag{3.3.44}$$

Si assume così che la derivata dell'equazione (3.3.43) sia diversa da zero solo se $t = t'$, vero nel caso di processi che dipendono lentamente dal tempo.

Se si considera la risposta reale della densità (sufficiente per energie di eccitazione e forze dell'oscillatore) è possibile semplificare le equazioni (3.3.37) e (3.3.38) in modo considerevole, utilizzando le proprietà di simmetria della matrice K di accoppiamento. Nel caso si usi *ALDA* si ha che, a causa della scelta per gli orbitali reali di *KS*, $K_{ia\sigma,jb\tau} = K_{ia\sigma,bj\tau}$ .Quindi, sostituendo nelle espressioni (3.3.37) e (3.3.38), e trasformandole in un set di equazioni per $P^{\tau}_{jb} + P^{\tau}_{bj}$ e $P^{\tau}_{jb} - P^{\tau}_{bj}$, si ottiene per la parte reale degli elementi della matrice densità

$$\sum_{bj\tau}\left[\delta_{\sigma\tau}\delta_{ab}\delta_{ij}(\varepsilon_i - \varepsilon_a) - 2K_{ia\sigma,jb\tau} - \omega^2\frac{\delta_{\sigma\tau}\delta_{ab}\delta_{ij}}{(\varepsilon_i - \varepsilon_a)}\mathrm{Re}\,\delta P^{\tau}_{jb}(\omega)\right] = \left[\delta\upsilon_{ext}(\omega)\right]_{ia\sigma} \qquad (3.3.45)$$

che può essere scritta in notazione vettoriale come

$$[\Delta - 2K](\mathrm{Re}\,\delta P) = \delta\upsilon_{ext} \qquad (3.3.46)$$

dove la matrice Δ è una matrice diagonale

$$\Delta_{ia\sigma,ij\tau} = \delta_{\sigma\tau}\delta_{ab}\delta_{ij}\left[(\varepsilon_i - \varepsilon_a) - \frac{\omega^2}{(\varepsilon_i - \varepsilon_a)}\right] \qquad (3.3.47)$$

La parte reale della matrice densità *P* del primo ordine, ottenuta dalla soluzione dell'equazione lineare (3.3.46) permette di ottenere la polarizzabilità dipendente dalla frequenza [Gross *et al*. 1990; Karna *et al*. 1991; Gisbergen *et al*. 1995], mentre per calcolare le energie di eccitazione, si fa uso di un'equazione derivante dalla (2.4.46) [Casida 1995; Jamorski *et al*. 1996]

$$\Omega\mathbf{F}_i = \omega_i^2\mathbf{F}_i \qquad (3.3.48)$$

dove le componenti della matrice Ω sono date da

$$\Omega_{ia\sigma,ij\tau} = \delta_{\sigma\tau}\delta_{ab}\delta_{ij}(\varepsilon_i - \varepsilon_a)^2 + 2\sqrt{(\varepsilon_a - \varepsilon_i)}K_{ia\sigma,ij\tau}\sqrt{(\varepsilon_b - \varepsilon_j)} \qquad (3.3.49)$$

Le energie di eccitazione desiderate sono uguali a ω_i, e le intensità sono date dalle forze dell'oscillatore,[44] ottenute dagli autovettori $\mathbf{F}_i$. La soluzione diretta di queste equazioni risulta impraticabile, poiché richiede di calcolare e memorizzare tutti gli elementi della matrice K di accoppiamento. Per questo motivo è preferibile risolvere le equazioni in modo iterativo e l'algoritmo generalmente impiegato è quello proposto da Davidson [Davidson 1975], risultato particolarmente efficiente [Olsen *et al*. 1988]. Il metodo di Davidson permette di risolvere in modo iterativo un problema agli autovalori con ridotti costi computazionali per matrici estese. È importante sottolineare che questo processo iterativo utilizza prodotti matrice-vettore per il calcolo degli autovettori. Inoltre per risolvere iterativamente l'equazione agli autovalori (3.3.48), è necessario stabilire un'ipotesi iniziale, che potrebbe essere ad esempio $K = 0$ come ragionevole punto di partenza. Nel primo ciclo le energie di eccitazione sono uguali alla differenza tra gli autovalori degli orbitali KS occupati e quelli non occupati. Per maggiori informazioni riguardo a questo algoritmo è possibile fare riferimento agli articoli di [Davidson 1989; Davidson 1993; Stathopoulos *et al*. 1994].

[44]La forza dell'oscillatore è una grandezza adimensionale e costituisce una misura dell'intensità di assorbimento. Per una transizione $n \leftarrow 0$ è data da $f_{n0} = \left(\frac{4\pi m_e \nu_{n0}}{3e^2\hbar}\right)|\mu_{n0}|^2$, dove μ_{n0} è il momento di dipolo della transizione e ν_{n0} è la frequenza di transizione [Atkins 1983].

Capitolo 4

Descrittori Molecolari

4. Descrittori molecolari (*MD*)

Lo scopo di questo capitolo è quello di fornire informazioni circa la variabile indipendente di un modello *QSAR*, analogamente a quanto fatto nel secondo capitolo per la variabile dipendente. Gli *MD* rappresentano quelle proprietà molecolari che si ritiene siano correlate in qualche modo con la tossicità dei *chemicals* appartenenti al dominio del modello. Gli *MD* per modelli *QSAR* fino ad oggi proposti in letteratura sono migliaia [Todeschini *et al.* 2009]. In linea di principio si può affermare che qualsiasi proprietà molecolare esprimibile con un numero può essere proposta quale *MD*. Una discussione adeguata di ognuno di essi esula dallo scopo del presente lavoro. Ciononostante, scopo finale di questo capitolo è quello di:

3. fornire alcune indicazioni circa i descrittori citati nelle linee guida ufficiali [*OECD* 2007];
4. operare una divisione in classi degli *MD* reperibili in letteratura;
5. mettere in evidenza alcune criticità nella scelta degli *MD*;
6. fornire informazioni circa gli *MD* utilizzati nella ricerca.

4.1 Descrittori molecolari e linee guida *OECD*

Storicamente, la prima *QSAR* proposta da Hansch [Hansch *et al.* 1962] fu costruita mettendo in relazione l'attività biologica di una serie di acidi fenossiacetici con la costante di Hammet,[45] σ, e il coefficiente di partizione ottanolo/acqua, logK_{OW}, detto anche log*P*. Successivamente, Hansch e Fujita [Hansch *et al*. 1964] proposero l'uso della loro costante di sostituzione idrofobica[46] per descrivere l'effetto di un sostituente sulla idrofobicità di una molecola. Il coefficiente di partizione log*P*, e la costante di Hansch e Fujita sono *MD* ancora oggi molto usati in quanto alcuni aspetti di farmacocinetica sono legati alla idrofobicità/lipofilicità della molecola. Un approccio alternativo è il *Taft's solvatochromic approach* [Taft 1956; Kamlet *et al.* 1985]. Taft e collaboratori postularono di poter spiegare l'attività biologica sulla base di fattori quali il volume molecolare, il momento dipolare, la capacità di accettare o donare H (acidità o basicità della molecola).

La linea guida per lo sviluppo di *QSAR* emessa da *OECD* nel 2007 riporta la seguente definizione: "*...a molecular descriptor is a structural or physicochemical property of a molecule, or part of a molecule, which characterises a specific aspect of a molecule and is used as independent variable in a QSAR".* La definizione data da *OECD* comprende sia quantità misurate sperimentalmente sia calcolate. Gli *MD* possono infatti essere anche indici non direttamente collegabili ad una proprietà chimico fisica, ma legati, ad esempio, alla presenza di determinate strutture nella molecola. In tabella 6.1 della guida *OECD* è riportato un elenco di *MD* normalmente usati nelle *QSAR.* Di questa viene riportata di seguito una breve sintesi.

1. Il coefficiente di partizione ottanolo/acqua, log*P*.
2. La costante di sostituzione idrofobica di Hansch e Fujita.

[45] La costante di Hammett σ esprime l'effetto dell'introduzione di un sostituente sull'acidità dell'acido benzoico in modo tale che il rapporto tra le costanti di dissociazione con e senza sostituente sia uguale alla costante σ moltiplicata per un fattore ϱ secondo l'equazione $\log(K'/K) = \varrho\sigma$

[46] La costante di sostituzione idrofobica esprime l'effetto su log*P* della sostituzione di un idrogeno con un gruppo idrofobico, è la differenza tra i log*P* del composto sostituito e idrogenato ($\log P_{R\text{-}X} - \log P_{R\text{-}H}$).

3. La costante di Hammett.
4. Il parametro sterico di Taft [Taft 1956], descrive il contributo alla velocità di reazione degli effetti sterici intermolecolari secondo l'equazione:

$$\log k = \log k_0 + \rho\sigma + \delta E_s \qquad (4.1)$$

dove ρ e σ sono due costanti .

5. La solubilità in acqua S_{aq}, corrispondente alla massima concentrazione ottenibile di un composto completamente sciolto in acqua ad una certa temperatura e pressione, all'equilibrio.
6. *Molecular refractivity* (*MR*) [Vogel 1948; Pauling 1960; Agin *et al.* 1965; Hansch *et al.* 2003]. Questo *MD* è stato sviluppato nel tentativo di combinare proprietà strutturali quali la massa molecolare *MW* e la densità in fase liquida, con proprietà elettroniche quali la polarizzabilità che è legata all'indice di rifrazione di un liquido puro da una equazione tipo Lorentz-Lorentz.

$$MR = \left[\frac{\left(n^2 - 1\right)}{\left(n^2 + 1\right)}\right]\frac{MW}{\rho} \qquad (4.2)$$

dove *n* è l'indice di rifrazione del liquido, *MW* la massa molecolare del liquido, ρ la densità.

7. La costante di dissociazione pK_a della molecola tossica in esame.
8. Il momento di dipolo permanente μ che descrive la separazione di carica nella molecola, è connesso alla sua idrofobicità.
9. Il *Molecular Electrostatic Potential* (*MEP*), viene calcolata la carica nello spazio attorno la molecola per identificare siti di possibile interazione elettrostatica.

10. Le energie degli orbitali *HOMO* e *LUMO*, utilizzate quali indicatori della nucleofilicità e elettrofilicità della molecola.
11. *L'Hydrogen Bonding*, è stato proposto quale *MD* la capacità di una molecola di stabilire ponti ad idrogeno con altre molecole.
12. I *Molecular Weight* e *Molecular Volume*, descrittori fisici della molecola.
13. La *Molecular Surface Area* (*MSA*), viene calcolato utilizzando le superfici di Van der Waals della molecola.
14. I descrittori topologici. Applicando la teoria dei grafi alle molecole sono state ricavate delle tavole di connettività tra gli atomi che permettono di definirne la forma, le dimensioni, la flessibilità. Sono stati proposti vari indici (che portano i nomi dei rispettivi inventori).
15. I descrittori elettrotopologici. Sono ottenuti aggiungendo agli *MD* topologici, sopra descritti, informazioni circa le interazioni elettroniche tra gli atomi.
16. Gli *MD* calcolati tramite la Teoria del Funzionale Densità (*DFT*), sono i descrittori oggetto di questa tesi.

L'elenco degli *MD* citati nella linea guida *OECD* non è ovviamente esaustivo di tutti quelli esistenti in letteratura, per i quali tuttavia è possibile una divisione in tre classi.

1. *MD* topologici o substrutturali, a loro volta divisi in topostrutturali (TS) e topochimici (TC) dipendentemente dal loro contenuto di informazioni, solo sulla struttura o anche sulle proprietà elettroniche. Il loro sviluppo è legato ad una applicazione della teoria dei grafi alle molecole per cui esse sono assimilate ad insiemi (e gli atomi ad elementi di questi insiemi). Applicando i teoremi della topologia matematica (di qui il nome degli *MD*) si ottengono degli indici di connettività tra insiemi che rappresentano i legami nelle molecole [Randić *et al.* 2001; Gallegos *et al.* 2005; Todeschini *et al.* 2009].
2. I descrittori tridimensionali (MD-3D) quantificano la forma, la grandezza e altre caratteristiche strutturali delle molecole che derivano dalla disposizione degli atomi nello

spazio. Descrittori MD-3D sono stati usati per esempio per introdurre quale variabile indipendente del modello la chiralità di una molecola [Natarajan *et al.* 2007].

3. Descrittori quantomeccanici (MD-QM). Sono tutti quelli che vengono calcolati tramite tecniche di calcolo quantomeccanico con metodi semiempirici o *ab initio* (*DFT*, Hartree Fock ecc.). Sono legati alle proprietà elettroniche della molecola e sono quelli utilizzati nell'ambito di questo lavoro.

Un elenco degli *MD* più usati con informazioni circa la loro classificazione si trova in [Basak *et al.* 2001; Todeschini *et al.* 2009]. Gli *MD* possono essere ulteriormente classificati in base alla loro complessità e alla loro richiesta di risorse di calcolo. Ordinandoli secondo complessità crescente si trova la seguente gerarchia:

$$MD\text{-}TS < MD\text{-}TC < MD\text{-}3D < MD\text{-}QM$$

Così, per poter calcolare un descrittore topostrutturale saranno necessarie minori risorse di calcolo che per uno quantomeccanico. Questo aspetto è particolarmente importante per quelle *QSAR* che richiedono il calcolo simultaneo di centinaia di *MD* diversi e che potrebbero quindi diventare esose nelle risorse di calcolo richieste.

La scelta di una specifica classe di *MD* per la costruzione di una *QSAR* è cosa tutt'altro che ovvia. Idealmente, sarebbe necessario tenere conto contemporaneamente di tutte le informazioni strutturali ed elettroniche, e quindi poter usare il maggior numero di *MD* possibili, e di tutti i tipi. In realtà, il numero degli *MD* è limitato dalla statistica ad un quinto del numero delle molecole nel campione,[47] e la scelta del tipo di *MD* da usare dipende dall'approccio scelto per la *QSAR*. Secondo l'approccio geometrico-strutturale si assume che la reattività delle molecole sia essenzialmente

[47] Secondo la regola di Topliss e Costello [Topliss *et al*. 1972] il numero dei regressori nella regressione lineare multivariata dovrebbe rimanere inferiore ad un quinto del numero delle molecole nel *training set*.

legata alla loro forma, ovvero che la reattività dipenda soprattutto dagli effetti sterici. Questo approccio può essere adatto per descrivere molecole molto grandi (quali le proteine) ove la presenza di sacche, tasche, canali, avvallamenti, può essere necessaria per permettere ad una molecola più piccola di raggiungere il sito di reazione. Questo approccio è adottato ad esempio nelle tecniche di *docking* molecolare. Ove invece siano più importanti le interazioni a corto raggio, che implichino un trasferimento di carica, può essere più adeguato l'approccio di tipo quantomeccanico.

Qualunque sia l'approccio adottato, la scelta degli *MD* da utilizzare è sempre vincolata all'*endpoint* indagato. La tossicità è legata alla capacità di una molecola esogena di farsi riconoscere da un sistema biologico e di perturbarne in qualche modo i processi biochimici. Sulla base della loro capacità di stabilire legami con siti molecolari di interesse biologico gli esogeni possono essere divisi in due classi.

- Esogeni ad interazione generica.
- Esogeni ad interazione specifica.

Gli esogeni con azione generica interagiscono con strutture cellulari protoplasmatiche, in modo spesso reversibile, con effetti anestetici di narcosi tipici ad esempio di alcuni solventi organici. La caratteristica di queste sostanze è la non selettività e i loro effetti possono essere spiegati sulla base di proprietà globali e aspecifiche, quali la massa, il volume molecolare, log*P*. Per contro, gli esogeni con azione specifica possiedono particolari caratteristiche strutturali o elettroniche e interagiscono su specifici *target* biologici. Più l'interazione è specifica più saranno favoriti i descrittori locali, quali i descrittori *QM*.

4. 2 Criticità nella scelta degli *MD*

Come anticipato nel capitolo primo, l'approccio adottato in questa ricerca è stato quello di utilizzare esclusivamente *MD* quantomeccanici calcolati tramite *DFT*. Si è supposto infatti che data la scarsa selettività strutturale di *AhR*, in grado di legarsi a molecole molto diverse tra loro, fossero le caratteristiche elettroniche delle molecole esogene a giocare un ruolo fondamentale. Va in ogni caso ricordato che la costruzione di un modello *QSAR* efficiente richiede tre condizioni simultanee.

- Dati tossicologici di partenza (*endpoint*) di buona qualità.
- Un set di *MD* in grado di cogliere l'essenza della proprietà indagata.
- Appropriate tecniche statistiche per la validazione del modello.

Merita di essere sottolineato che, molto recentemente, Dearden [Dearden *et al.* 2009] ha svolto un'analisi delle principali criticità emerse dalla consultazione dei modelli *QSAR* presenti in letteratura in relazione a violazioni dei cinque principi stabiliti da *OECD*. A proposito della scelta degli *MD* egli ha segnalato la presenza dei seguenti errori.

3. Uso di *MD* collineari. La collinearità si presenta quando si effettua una regressione multivariata utilizzando quali variabili indipendenti *MD* che non sono perfettamente indipendenti tra loro. *MD* collineari non aggiungono informazioni al modello ma ne peggiorano le prestazioni statistiche. Il loro uso simultaneo va quindi evitato. L'analisi statistica dei dati permette attraverso l'indice di correlazione di Pearson[48] di valutare la presenza di collinearità. Nel nostro caso avendo utilizzato *MD* che sono tutti calcolati con la DFT, e quindi hanno origine comune nella densità elettronica, la valutazione della collinearità si è resa quanto mai opportuna.

[48] Il coefficiente di correlazione (lineare) di Pearson (detto anche di Bravais-Pearson) tra due variabili aleatorie o due variabili statistiche x e y è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili. Assume valori compresi tra -1 e +1, è zero per variabili perfettamente indipendenti tra loro, condizione necessaria ma non sufficiente per la non collinearità.

4. Uso di *MD* incomprensibili. A causa del gran numero di *MD* proposti in letteratura, alcuni dei quali assolutamente non riconducibili ad un significato chimico fisico immediato, anche per gli operatori più esperti è difficile comprenderne la natura e l'eventuale importanza. Si tenga presente che alcune tecniche attribuiscono ad ogni *MD* un coefficiente di regressione su pure basi statistiche, senza alcuna considerazione chimica a monte. Nella nostra ricerca si sono utilizzati *MD* dal significato chimico fisico consolidato.
5. Errori nel calcolo degli *MD*. Per quanto nessun calcolo sia avulso da un certo margine di errore, l'uso di tecniche di calcolo quantomeccanico quali la *DFT* permette di effettuare calcoli accurati di ciascuna grandezza. Inoltre l'analisi dei residui in fase di validazione del modello può mettere in luce la presenza di errori sistematici.
6. Uso di un eccessivo numero di *MD* nella regressione. Secondo la regola di Topliss e Costello [Topliss *et al.* 1972], un numero eccessivo di regressori rispetto ai campioni nel *training set* riduce i gradi di libertà del sistema e peggiora le prestazioni statistiche.
7. Mancanza di un adeguato *autoscaling*[49] degli *MD* usati quali regressori che spesso assumono valori molto diversi tra loro e vanno perciò scalati per poterne valutare l'importanza nella regressione. Nella costruzione dei modelli *QSAR* multivariati è stato effettuato l'*autoscaling mean centering* dei regressori tramite il software R,[50] ed in particolare il pacchetto software *PLS* [Mevik *et al*. 2007], usato per i calcoli statistici.

[49]Per scalare il valore assunto da un *MD* (usato quale regressore nella regressione multivariata) vi sono varie tecniche, una delle più usate è il *mean centering* per cui si toglie al valore la media e si divide per la *standard deviation*. $x = (x_i - x_m)/SD$

[50]*The R Project for Statistical Computing http://cran.r-project.org/*.

4. 3 Gli MD utilizzati nella ricerca

Tutti gli *MD* utilizzati sono stati calcolati tramite la *DFT* e sono legati quindi alle proprietà elettroniche della molecola (più in particolare alla densità elettronica). Questo tipo di *MD* è quello di più recente introduzione nelle *QSAR*, ma l'importanza delle proprietà elettroniche è stata evidenziata da Hansch [Hansch *et al*. 2003] *"...no matter how one approaches QSAR , electronic interactions must be considered if we are to begin to develop a science of chemical-biological interactions"*.

In letteratura, sono stati proposti quali *MD* da correlare con la tossicità di molecole per le quali viene proposto il meccanismo con *AhR* le seguenti proprietà molecolari: la polarizzabilità e l'anisotropia della polarizzabilità [Hansch *et al.* 2003; Hirokawa *et al.* 2005; Hinchliffe *et al.* 2006; ChengGangGu *et al.* 2007], gli indici di elettrofilicità globale e locale e la *chemical hardness* [Roy *et al.* 2006; Sarkar *et al.* 2005, Chattaraj *et al.* 2006; Yan *et al.* 2006; Schultz *et al.* 2006 a)].

4.3.1 Polarizzabilità

La polarizzabilità è la capacità di un sistema di rispondere ad un campo elettrico esterno, oscillatorio o statico, generato da un apparato esterno o da un'altra molecola o ione e di acquisire un momento di dipolo. È un tensore di secondo grado che rappresenta la derivata seconda dell'energia rispetto al campo elettrico esterno [Karplus *et al.* 1971; Atkins 1985]. L'applicazione di un campo **E** orientato lungo l'asse *z* può indurre un dipolo che ha componenti lungo gli assi *x*, *y*, *z* (normalmente l'asse *z* parallelo al campo rappresenta la componente più importante). La determinazione sperimentale della polarizzabilità molecolare può essere assai difficile soprattutto per molecole con bassa simmetria. Vi sono tuttavia vari modi per calcolarla. Un elenco di formule contenenti la polarizzabilità (utili per il suo calcolo a partire da grandezze misurabili) è riportato in *CRC Handbook of Chemistry and Physics*, *84th Edition.*

Ogni mezzo ottico è caratterizzato da una quantità denominata indice di rifrazione n[51] che rappresenta il rapporto tra le velocità che la luce assume nel vuoto e nel mezzo. La propagazione di un campo elettromagnetico ad alta frequenza nel mezzo induce una polarizzazione. In fase gas il valore medio della polarizzabilità molecolare, di atomi o molecole non polari, ad una data frequenza ν può essere determinato dalla misura dell'indice di rifrazione della luce alla frequenza ν tramite l'equazione di Lorentz Lorentz [Lorentz 1880; Lorentz 1906; Sylvester-Hvid *et al.* 1999; Hinchliffe *et al.* 2006]:

$$\alpha(\nu) = \frac{3V}{4\pi N}\left[\frac{n^2(\nu)-1}{n^2(\nu)+2}\right] \qquad (4.3)$$

Dove N è il numero delle molecole in un volume V e n l'indice di rifrazione alla frequenza ν.

In fase condensata le interazioni tra le molecole non possono essere trascurate per cui, oltre al contributo dovuto al campo esterno, vi è anche quello dovuto al campo generato dalle molecole circostanti. Questo comportamento è descritto dall'equazione di Clausius-Mossotti [Millikan 1897; Oughstun *et al*. 2003]:

$$\alpha(\nu) = \frac{3}{4\pi N}\left[\frac{\varepsilon_r^2(\nu)-1}{\varepsilon_r^2(\nu)+2}\right] \qquad (4.4)$$

Dove N è il numero delle molecole in un volume V e ε_r la permittività del mezzo.

Il momento dipolare indotto da un campo elettrico $\vec{E}$ è definito da:

[51] Si è usata la notazione n per l'indice di rifrazione in luogo del più comune η per evitare l'uguaglianza con il simbolo che verrà usato nel testo per la *chemical hardness*.

$$\langle \mu \rangle = \mu_0 + \alpha \cdot \mathbf{E} + \frac{1}{2}\beta : \mathbf{E}^2 + \ldots \qquad (4.5)$$

Dove $\boldsymbol{\mu}_0$ appresenta il momento di dipolo permanente della molecola, $\boldsymbol{\alpha}$ è la polarizzabilità, $\boldsymbol{\beta}$ l'iperpolarizzabilità. Se $\mathcal{E}$ rappresenta l'energia e $\mathbf{E}$ il campo elettrico, possiamo definire:

$$\mu_0 = -\left(\frac{\delta \mathcal{E}(\mathbf{E})}{\delta \mathbf{E}}\right)_0 \qquad (4.6)$$

$$\alpha = -\left(\frac{\delta^2 \mathrm{E}(\mathbf{E})}{\delta \mathbf{E}^2}\right)_0 \qquad (4.7)$$

$$\beta = -\left(\frac{\delta^3 \mathrm{E}(\mathbf{E})}{\delta \mathbf{E}^3}\right)_0 \qquad (4.8)$$

Normalmente, l'iperpolarizzabilità è trascurabile. L'unità di misura della polarizzabilità è $J^{-1}\ C^2\ m^2$, tuttavia viene spesso espressa come α' la polarizzabilità di volume:

$$\alpha' = \frac{\alpha}{4\pi\varepsilon_0} \qquad (4.9)$$

α' le dimensioni di un volume, tipicamente i suoi valori sono dell'ordine di 10^{-24} cm^3. Una unità di misura molto utilizzata per esprimere la polarizzabilità sono le *atomic units a.u.*,[52] (le piccole molecole organiche studiate in questo lavoro presentano valori dicompresi tra 100 e 300 *a.u.*). Nonostante la natura tensoriale di α per molti scopi è sufficiente utilizzare la media degli elementi diagonali del tensore.

[52] La polarizzabilità espressa in cm^3 può essere convertita in $a.u.\ a_0^3$, ove a_0 è il raggio atomico di Bohr, tramite l'equazione $\alpha\ [cm^3] = 0{,}148184 \times 10^{-24}\ \alpha\ [a_0^3]$

$$\bar{\alpha} = \frac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}) \qquad (4.10)$$

Nel presente lavoro quale *MD* per la polarizzabilità è stato utilizzato il valore medio della polarizzabilità statica, espressa in *a.u.* e calcolata secondo l'equazione 4.10. Nel seguito, se non diversamente specificato, con il termine polarizzabilità α si farà riferimento a tale valore.

Il tensore polarizzabilità può essere calcolato per via teorica in vari modi. Una rassegna dei metodi per il calcolo della polarizzabilità molecolare per mezzo della DFT, per molecole derivate da benzene, furano e tiofene, e dell'importanza di questa proprietà per spiegarne la reattività è riportata da Hinchliffe [Hinchliffe *et al.* 2006]. Un metodo semiempirico è stato proposto da Miller [Miller 1990]. Un'applicazione del modello di Thole per un calcolo *ab initio* è stata proposta da Devries [Devries *et al.* 1997] e da Van Duijnen [Vanduijnen *et al*. 1998]. Calcoli DFT su molecole di benzene e naftalene sono stati proposti da Millefiori [Millefiori *et al*. 1998].

La suite di programmi di calcolo quantomeccanico utilizzata in questo lavoro di tesi, *Amsterdam Density Functional* (*ADF*), permette il calcolo della polarizzabilità molecolare e delle altre proprietà sopra nominate. Dettagli circa i metodi utilizzati nel calcolo, sono reperibili sul sito della casa produttrice di *ADF* all'indirizzo *http://www.scm.com*. Fra le pubblicazioni di riferimento possono essere citati i seguenti lavori [Van Gisbergen *et al*. 1995; Van Gisbergen *et al*. 1996; Osinga *et al.* 1997; Champagne *et al.* 1998; Van Gisbergen *et al*. 1998; Van Gisbergen *et al*. 1999; Schipper e*t al.* 2000; Gruening *et al.* 2002].

Un primo tentativo di correlare la polarizzabilità molecolare con processi biologici è stato fatto da Pauling [Pauling *et al.* 1945]. Successivamente, fu dimostrato da Kutter [Kutter *et al.* 1969] che l'ipotesi di Pauling era eccessivamente restrittiva e gli *MD* legati a fattori sterici non potevano essere trascurati al fine di descrivere il fenomeno studiato da Pauling. Ci furono, tuttavia, casi in cui l'uso della polarizzabilità molecolare, come *MD*, portò a risultati di notevole successo. Agin [Agin

et al. 1965] mise in relazione la capacità neurotossica di 39 molecole nei confronti delle rane con due *MD*: polarizzabilità molecolare e potenziale di ionizzazione (*IP*). Il *set* di molecole studiato comprendeva molecole strutturalmente assai differenti e non facilmente riconducibili ad una classe di composti chimici. Ciononostante, il modello ottenuto da Agin si sovrapponeva molto bene con i dati sperimentali e la polarizzabilità risultava il migliore fra i due *MD*, potendosi quasi trascurare il contributo di *IP*. Più di recente, la polarizzabilità molecolare è stata proposta quale descrittore per comprendere le affinità di legame con *AhR* di un set di policlorodibenzo-*p*-diossine [Fraschini *et al.* 1996]. Similmente, la polarizzabilità molecolare è stata usata quale *MD* per descrivere la tossicità di policlorodibenzofurani da Hirokawa e Chengang Gu [Hirokawa *et al.* 2005; Chengang Gu *et al.* 2007]. Complessivamente, i lavori di Pauling, Hansch e Agin sembrano suggerire che la polarizzabilità possa, in alcuni casi, essere un *MD* adeguato per spiegare l'attività biologica di specifici sistemi molecolari. Per contro, in altri casi, sembrano prevalere fattori quali idrofobicità, fattori sterici ecc. La scelta di utilizzare quale *MD* non è quindi valida *a priori,* essa deve essere necessariamente legata al *MOA* delle sostanze.

4.3.2 Anisotropia della polarizzabilità

L'anisotropia della polarizzabilità rappresenta la differenza tra le componenti del tensore polarizzabilità parallele e perpendicolari al campo esterno **E**.

$$\gamma^2 = \frac{1}{2}\left[(\alpha_{xx} - a_{yy})^2 + (\alpha_{xx} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{yy})^2\right] \qquad (4.11)$$

Nel seguito del testo l'anisotropia della polarizzabilità verrà indicata con $\Delta\alpha$ L'anisotropia della polarizzabilità può essere ottenuta per un liquido puro o per un gas da misure di *light scattering* in

presenza di un forte campo elettrico esterno per mezzo della formula di Rayleigh [Boettcher 1973; Boettcher *et al.* 1973; Alms *et al*. 1975; Chrissanthopoulos *et al.* 2000; Shuvaeva 2007].

La polarizzabilità e la sua anisotropia sono stati i primi *MD* utilizzati per i modelli *QSAR* sviluppati in questa tesi, non solo per l'importanza che veniva attribuita a questa proprietà molecolare in letteratura [Hansch *et al.* 2003; Hirokawa *et al.* 2005; Hinchliffe *et al.* 2006; Chengang Gu *et al.* 2007], ma anche perché la disponibilità di dati sperimentali [Millefiori *et al.* 1998; Miller 2004] ha permesso una verifica della affidabilità dello schema di calcolo *DFT* adottato.

4.3.3 *Chemical Hardness*

L'elettronegatività di una specie chimica è la tendenza di un atomo in una molecola ad attrarre elettroni [Pauling 1960]. Dal punto di vista quantitativo esistono in letteratura varie scale di elettronegatività [Mulliken 1934; Pauling 1945; Gordy 1946; Gordy 1951; Walsh 1951; Sanderson 1955; Allred *et al*. 1958; Pauling 1960]. Posto che $\mathcal{E}(N)$ sia la funzione che lega l'energia elettronica di una specie nel proprio stato fondamentale al numero di elettroni *N*, la derivata di $\mathcal{E}(N)$ rispetto al numero di elettroni *N*, a potenziale esterno v[53] costante, rappresenta il potenziale chimico μ (oppure il valore della elettronegatività assoluta χ cambiata di segno) [Pauling 1945; Pauling 1960; Iczowski *et al.* 1961; Parr *et al.* 1978; Parr *et al*. 1983; Sen *et al.* 1987]. Questa definizione corrisponde esattamente alla definizione del potenziale chimico data dalla *DFT* [Parr *et al.* 1978].

$$\mu = -\chi = \left(\frac{\delta \mathrm{E}}{\delta N}\right)_{v} \qquad (4.12)$$

[53]Si è usato il simbolo v per indicare il potenziale esterno e quello *v* per la frequenza nelle equazioni 4.3 e 4.4.

Da un punto di vista pratico si può calcolare l'elettronegatività (secondo Pauling e Mulliken) come la media del potenziale di prima ionizzazione *IP* e dell'affinità elettronica (*EA*) [Mulliken 1934]. Questa definizione resta valida anche nel caso non esista in ogni punto una funzione continua e derivabile $\mathcal{E}(N)$ [Perdew *et al.* 1982].

$$\chi = \frac{1}{2}(IP + EA) \tag{4.13}$$

La derivata seconda di $\mathcal{E}(N)$ rispetto al numero di elettroni, a potenziale esterno costante, rappresenta la resistenza opposta dal potenziale chimico al cambiamento del numero di elettroni ovvero l'*hardness* del sistema [Parr *et al.* 1983; Pearson 2005].

$$\eta = \left(\frac{\delta^2 \mathrm{E}}{\delta N^2}\right)_V \tag{4.14}$$

I concetti di *hardness* e *softness* sono stati introdotti nel passato per spiegare il comportamento degli acidi e delle basi di Lewis [Pearson 1963].

$$\mathrm{A} + :\mathrm{B} \longrightarrow \mathrm{A}:\mathrm{B}$$

In riferimento alla stabilità del prodotto A:B, formato per reazione con alcune basi di riferimento, acidi e basi di Lewis sono stati divisi in due categorie. Dal punto di vista delle specie che donano elettroni, possono essere definite basi di Lewis *soft,* quelle basi che hanno bassa elettronegatività e alta polarizzabilità, per cui donano facilmente gli elettroni. Sono invece basi *hard* quelle che pur

avendo bassa elettronegatività sono scarsamente polarizzabili e quindi si oppongono al cambiamento.

Dal punto di vista delle *QSAR,* sia elettronegatività (o potenziale chimico), sia la polarizzabilità possono essere buoni descrittori della capacità di una specie di cedere o acquistare elettroni. L'elettronegatività renderà conto degli aspetti termodinamici, la polarizzabilità di quelli cinetici. Poiché *hardness* e polarizzabilità di una molecola sono proprietà legate tra loro, nel caso vengano contemporaneamente inserite quali regressori di una regressione multivariata, occorrerà utilizzare opportuni test statistici per verificare che sia rispettato il requisito di indipendenza dei regressori.

Si può definire *softness* il reciproco dell'*hardness.*

$$S = \frac{1}{\eta} \tag{4.15}$$

Tramite l'approssimazione delle differenze finite si può avere anche una definizione operativa dell'*hardness* facendo uso del potenziale di ionizzazione *IP* e dell'affinità elettronica *EA* [Parr *et al.*1983].

$$\eta = \frac{1}{2}\left(IP - EA\right) \tag{4.16}$$

Posto che il potenziale di ionizzazione è sempre maggiore o uguale all'affinità elettronica [Nalewajski *et al.* 1982], ne segue che il valore minimo dell'*hardness* è zero e corrisponde ad un cambiamento nullo del potenziale chimico per una variazione del numero di elettroni. Per il calcolo

dell'*hardness* sono stati effettuati i calcoli di *IP* e *EA* secondo l'approssimazione del *transition state* di Slater [Slater 1972; Parr & Yang 1989], è poi stata utilizzata l'equazione 4.16.

4.3.4 Indici di elettrofilicità globale e locale

Pur se il *MOA* di una sostanza tossica non è conosciuto esattamente è ragionevole ritenere che, qualunque sia l'azione esplicata, essa debba comportare la formazione e la rottura di legami chimici. La formazione o rottura di un legame chimico comporta lo spostamento di cariche (elettroni) da una specie ad un'altra o da un frammento molecolare ad un altro. Alla specie (o al frammento) ricco di elettroni viene dato il nome di nucleofilo, viceversa, alla specie o al frammento povero di elettroni viene dato il nome di elettrofilo. Un nucleofilo è in grado di farsi attrarre da un centro di carica positiva (in una specie o frammento elettrondeficiente) e di stabilire con esso un legame; condividendo o cedendo elettroni, un elettrofilo è in grado di reagire in modo analogo acquistando elettroni da un centro di carica negativa. Anche nel caso di rottura omolitica, con formazione di radicali liberi, alle specie formatesi può essere attribuita una elettrofilicità/nucleofilicità dipendentemente dalla loro tendenza ad attaccare siti ad alta o bassa densità di carica. Pur non trascurando l'importanza di altri fattori (fattori sterici, effetti termodinamici, cinetici o di trasporto), si può affermare che la maggior parte delle reazioni può essere analizzata attraverso l'elettrofilicità e nucleofilicità delle specie coinvolte [Lowry *et al.* 1987; Smith *et al.* 2006; Carey *et al.* 2007]. In generale, vi è corrispondenza tra una specie elettrofila (nucleofila) e un acido (base) di Lewis. Tuttavia, con il termine elettrofilo, si fa tradizionalmente riferimento agli aspetti cinetici, mentre con il termine acido di Lewis a quelli termodinamici. Il concetto di elettrofilicità è conosciuto fin dagli anni 30 del secolo scorso. Al riguardo, merita di essere sottolineato che si ritiene [Ingold 1933; Ingold 1934] sia stato Ingold il primo a proporre una scala per la elettrofilicità globale delle specie.

Nel corso degli anni sono state proposte varie scale delle elettrofilicità delle molecole, ottenute per via empirica. Una scala è stata proposta da Mayr [Mayr *et al*. 1994; Roth *et al.* 1995; Mayr *et al*. 2001; Mayr *et al*. 2002; Bug *et al.* 2003; Lemek *et al.* 2003; Tokuyasu *et al.* 2004; Minegishi *et al*. 2005] secondo il modello Mayr -Patz:

$$\log K = s\left(\tilde{N} + \tilde{E}\right) \qquad (4.17)$$

dove K rappresenta la costante di equilibrio nella reazione tra la specie elettrofila e nucleofila, s e $\tilde{N}$[54] rappresentano parametri legati alla nucleofilicità, $\tilde{E}$ alla elettrofilicità.

Un'altra scala empirica è stata proposta da Legon e Millen [Legon et al. 1987; Legon 1999; Jaramillo et al. 2006; Cedillo et al. 2007] basata sulla costante di forza k, ricavata tramite spettroscopia IR, del legame tra una specie nucleofila B e una serie di composti alogenati. Nel modello proposto da Legon si ha:

$$K = c\ddot{E}\ddot{N} \qquad (4.18)$$

dove c è una costante, $\ddot{E}$[55] e $\ddot{N}$ sono rispettivamente il valore di elettrofilicità di RX (con X alogeno) e il valore di nucleofilicità di B. Non ci risulta, tuttavia, che alcuna *QSAR* sia stata basata sui modelli di Mayr e di Legon.

Solo di recente, Parr ha proposto un indice per esprimere il concetto di elettrofilicità in modo quantitativo [Maynard *et al.* 1998; Parr *et al.* 1999]. L'indice di elettrofilicità (globale) ω è stato definito da Parr come il quadrato della elettronegatività (potenziale chimico) diviso per la *chemical*

[54]Sono stati usati i simboli $\tilde{N}$ e $\tilde{E}$ per i parametri di elettrofilicità e nucloefilicità di Mayr, per evitare ambiguità di notazione con la 4.12.

[55]Analogamente a quanto fatto nella 4.17, sono stati usati i simboli $\ddot{E}$ e $\ddot{E}$ per i parametri di elettrofilicità e nucleofilicità di Legon per evitare omonimie con simboli già usati nel testo.

hardness, e misura il cambiamento di energia al second'ordine di un elettrofilo quando viene saturato di elettroni. L'indice di elettrofilicità locale ω_k è definito invece come il prodotto dell'indice di elettrofilicità globale per la funzione di Fukui [Fukui 1982] dell'atomo. Gli indici di elettrofilicità di Parr (globale e locale) si sono rivelati molto utili per descrivere, sia la reattività delle molecole [Chattaraj *et al.* 2006], sia la tossicità di idrocarburi policiclici aromatici (*PAHs*) [Roy *et al.* 2006; Sarkar *et al.* 2005; Sarkar *et al.* 2006; Schultz *et al*. 2006; Chattarj *et al.* 2007]. Essi sono stati perciò utilizzati quali *MD* nei modelli *QSAR* sviluppati nel presente lavoro.

Nell'ambito della DFT sono state definite sia l'*hardness* η sia l'elettronegatività χ (si vedano le equazioni 4.12 e 4.14). L'indice di elettrofilicità globale ω può quindi essere definito come:

$$\omega = \frac{\mu^2}{2\eta} \qquad (4.19)$$

Operativamente, per il calcolo di ω e ω_k si può usare l'approssimazione delle differenze finite (si vedano le equazioni 4.13 e 4.16). In questo caso poi, dovendosi elevare μ al quadrato, la 4.13 può essere usata direttamente. In uno studio recente, Pérez [Pérez *et al.* 2002; Pérez 2003] ha effettuato un confronto tra le elettrofilicità teoriche, ricavate secondo la definizione di Parr, e quelle sperimentali, ricavate con il modello empirico di Mayr, riscontrando una buona correlazione tra valori teorici e sperimentali.

Come per il calcolo dell'*hardness* η anche per il potenziale chimico μ nei calcoli di *IP* e *EA* è stata usata l'approssimazione del *transition state* di Slater. Esaminando i modelli *QSAR* precedentemente citati [Roy *et al.* 2006; Sarkar *et al.* 2006; Chattaraj *et al.* 2007], si osserva che l'approssimazione più usata per il calcolo di *IP* e *EA* è rappresentata dal teorema di Koopmans, per cui il potenziale di ionizzazione *IP* viene considerato pari all'energia del più alto orbitale occupato (*HOMO*) e l'affinità elettronica *EA* pari all'energia del più basso orbitale non occupato (*LUMO*),

cambiati di segno. Secondo il teorema di Koopmans quindi si possono calcolare elettronegatività e *hardness* secondo le equazioni:

$$\chi = -\frac{1}{2}\left(\varepsilon_{HOMO} + \varepsilon_{LUMO}\right) \tag{4.20}$$

$$\eta = \frac{1}{2}\left(\varepsilon_{LUMO} - \varepsilon_{HOMO}\right) \tag{4.21}$$

Nel presente lavoro si è optato per l'uso dell'approssimazione del *transition state* di Slater in quanto più accurata.

L'indice di elettrofilicità globale ω rappresenta una proprietà globale della molecola. Talvolta, per spiegare la reattività di una molecola è necessario fare riferimento alle sue proprietà locali. La funzione di Fukui [Fukui 1982; Parr *et al.* 1984], rappresenta un descrittore della reattività locale definito come [Yang *et al.* 1986; Cioslowski *et al.*1993, Roy *et al.* 2006]:

$$f(\mathbf{r}) = \left(\frac{\delta\rho(\mathbf{r})}{\delta N}\right)_{v(\mathbf{r})} = \left(\frac{\delta\mu}{\delta V_{(\mathbf{r})}}\right)_{N} \tag{4.22}$$

Ovvero, la funzione di Fukui rappresenta la variazione del potenziale chimico al variare del campo esterno senza variazione del numero totale di elettroni *N*. Per definire la funzione di Fukui occorre distinguere tre casi, corrispondenti a tre distinti tipi di reazioni: attacco nucleofilo, elettrofilo e radicalico. Vengono di seguito riportate le rispettive equazioni.

$$f^{+}(\mathbf{r}) = \rho_{N+1}(\mathbf{r}) - \rho_{N}(\mathbf{r}) \tag{4.23}$$

$$f^{-}(\mathbf{r}) = \rho_{N}(\mathbf{r}) - \rho_{N-1}(\mathbf{r}) \tag{4.24}$$

$$f^0(\mathbf{r}) = \frac{[\rho_{N+1}(\mathbf{r}) - \rho_{N-1}(\mathbf{r})]}{2} \qquad (4.25)$$

L'indice di elettrofilicità locale ω_k è definito come il prodotto dell'indice di elettrofilicità globale ω per la funzione di Fukui.

$$\omega = \omega f_i \qquad (4.26)$$

Dove, alla funzione di Fukui f_i corrisponde una delle 4.23, 4.24 o 4.25.

Distinguere *a priori* tra molecole tossiche in grado di effettuare uno dei tre tipi di attacco non è sempre possibile. La maggior parte delle molecole studiate in questa ricerca (dibenzofurani, nitrobenzeni ecc.) sono conosciute come tossici di tipo elettrofilo. Tuttavia, poiché alcuni fra i *training set* includevano sistemi molecolari con sostituenti elettrondonatori (ammine aromatiche, tolueni ecc.) si è preferito adottare un approccio più pratico. Le funzioni di Fukui f_i sono state calcolate considerando il massimo contributo al *LUMO* dello stato fondamentale [Gorelski *et al.* 2004; Makedonas *et al.* 2006] e tramite la 4.26 è stato poi calcolato l'indice ω_k utilizzato come MD.

Capitolo 5

Modelli *QSAR* sviluppati

5.1 Modelli QSAR sviluppati

Nei capitoli precedenti sono stati illustrati i tre elementi necessari allo sviluppo di un modello *QSAR*:

1. la disponibilità di un certo numero di molecole per le quali sia ipotizzabile una similitudine di *MOA* nel determinare la loro tossicità;
2. la disponibilità di dati tossicologici adeguati da utilizzare quali *endpoint* del modello, nell'accezione di Dearden [Dearden *et al*. 2009];
1. la disponibilità di *MD* in grado di descrivere il fenomeno della tossicità di ogni singola molecola, ovvero *MD* per i quali esista una relazione quantitativa con la tossicità.

Come a questo punto sarà chiaro, la coesistenza di queste tre condizioni non è scontata. Lo scopo di questo capitolo è quello di consentire al lettore di ripercorrere lo stesso percorso logico che ha portato allo sviluppo di alcuni modelli *QSAR* per mezzo della *DFT*, cercando di mettere in luce vantaggi e svantaggi delle soluzioni adottate.

5.2 Strumentazione

Per poter sviluppare *QSAR* di tipo quantomeccanico occorre idealmente avere a disposizione due tipi di strumentazioni diverse:

1. un laboratorio in grado di effettuare i test tossicologici. La raccolta dei dati di attività biologica è da considerarsi la parte della ricerca economicamente più onerosa [Hartung 2009]. Non avendo disponibilità di una simile struttura ci si è affidati a dati di letteratura;
2. risorse di calcolo a due livelli diversi.

1. Per il calcolo degli *MD* tramite *DFT* sono necessari sistemi di calcolo parallelo complessi, meglio se in *grid computing.* Per effettuare i calcoli *DFT* si sono utilizzate le risorse del *LICC* (Laboratorio Interdipartimentale di Chimica Computazionale) dell'Università di Padova ed il software commerciale *ADF*.[56]

2. Per lo sviluppo del modello *QSAR* e la sua analisi statistica sono stati usati un normale *personal computer* ed il *software opensource* R.[57] È opportuno osservare che il costo delle risorse per il calcolo di *MD* con approccio statistico e quantomeccanico non è uguale. La minor richiesta di risorse è forse uno dei motivi principali per cui le *QSAR* di tipo statistico si sono sviluppate prima e più velocemente di quelle quantomeccaniche. Le *QSAR* sono infatti nate quale *decision tool* in ambito industriale farmaceutico. La possibilità di far girare software *QSAR oriented* su mezzi *hardware* di potenza ridotta ne ha favorito senz'altro la diffusione.

Un'ulteriore osservazione va fatta a proposito dell'accuratezza necessaria nel calcolo degli *MD*. Dearden [Dearden *et al.* 2009] ha identificato fra i possibili errori nello sviluppo di *QSAR* la mancanza di accuratezza nei calcoli degli *MD*. La mancanza di accuratezza nel calcolo delle variabili indipendenti, o nella misura della variabile dipendente, può infatti nascondere la presenza

[56] *Amsterdam Density Functional*, *www.scm.com*.
[57] *The R Project for Statistical Computing http://cran.r-project.org*

della *QSAR*. Nel caso dell'analisi multivariata di centinaia di descrittori il problema dell'accuratezza nel calcolo degli *MD* è in genere meno evidente di quanto non lo siano gli errori nella misura dell'*endpoint.* Tuttavia, non va trascurato il fatto che il peso attribuito a ciascun *MD* dall'analisi multivariata non può essere accettabile qualora sia basato sui presupposti di un calcolo sbagliato. Pertanto, va impiegata la massima cura nella messa a punto del metodo di calcolo degli *MD* e la valutazione della sua accuratezza.

5.3 Dettagli computazionali

Tutti i calcoli delle proprietà molecolari (*MD*) sono stati effettuati utilizzando il software *ADF* (*Amsterdam Density Functional*), un *software* commerciale la cui documentazione di riferimento è disponibile sul sito www.scm.com. Inoltre gli esperimenti numerici sono stati condotti utilizzando il funzionale *PBE*, [Perdew et al. 1996; Perdew et al. 1998; Zhang et al. 1998], e *basis sets TZP* (triplo ξ con una funzione di polarizzazione). Questo approccio è diverso da quelli già adottati in letteratura. Al riguardo, merita di essere sottolineato che Hirokawa e Chengang Gu [Hirokawa *et al.* 2005; Chengang Gu *et al.* 2007] hanno sviluppato modelli *QSAR* per i dibenzofurani (*DBF*) facendo uso di *MD* legati a proprietà elettroniche (polarizzabilità, momenti di dipolo ecc.), calcolati tramite *DFT* utilizzando il funzionale ibrido *B3LYP*, [Becke 1993; Lee *et al.* 1988; Vosko *et al.*1980; Stephens *et al.* 1994] con un set di base *6-31G*** (funzioni Gaussiane con due funzioni di polarizzazione, notazione di Pople). I funzionali ibridi hanno la caratteristica di trattare nel potenziale di scambio o correlazione, la componente dello scambio a livello Hartree Fock (da cui l'aggettivo ibrido). Un confronto tra i risultati ottenuti con approcci computazionali differenti sulla molecola del naftalene è riportato da Millefiori [Millefiori *et al.* 1998]. L'uso congiunto di *PBE/TZP* è stato scelto perché ha dato risultati comparabili con i dati sperimentali.

5.4 Esperimenti Numerici Preliminari

Sulla base dei lavori di Hirokawa e Chengang Gu [Hirokawa *et al.* 2005; Chengang Gu *et al.* 2007] è stato deciso di utilizzare quali *MD* la polarizzabilità α e la sua anisotropia $\Delta\alpha$ calcolate tramite *DFT* con funzionale *PBE* e *basis sets TZP*. Inizialmente, sono stati effettuati una serie di esperimenti numerici preliminari per verificare l'accuratezza della scelta fatta. In Tabella 5.1 sono presentati i risultati del calcolo delle polarizzabilità (espresse in *atomic units*) confrontati con i valori sperimentali riportati in letteratura [Miller 2004].

Tabella 5.1. Confronto tra polarizzabilità (*a.u.*) sperimentali e calcolate con *PBE/TZP*.

Molecola	α_{exp}	α_{theo}	$\alpha_{theo}/\alpha_{exp}$
benzene	67,5	66,4	98,4
benzene	69,6	66,4	95,3
benzene	72,5	66,4	91,6
toluene	79,7	80,8	101,5
anilina	81,7	78,5	96,1
naftalene	111,3	117,3	105,4
naftalene	118,1	117,3	99,3
naftalene	117,4	117,3	99,9
nitrobenzene	99,2	85,4	86,1
nitrobenzene	87,2	85,4	98,0
p-dinitrobenzene	124,2	105,3	84,8

L'ispezione della tabella 5.1 mette in evidenza che, per quelle molecole per cui vi è più di un dato sperimentale a disposizione, le differenze tra valori teorici e sperimentali sono paragonabili, e a

volte inferiori, a quelle fra i dati sperimentali stessi. Pur se non vi è alcuna garanzia che lo schema di calcolo possa essere altrettanto accurato per altri *MD* di cui non si abbiano dati sperimentali, questo risultato tenderebbe a confermare la validità della scelta fatta.

5.5 Modelli di regressione monovariata

La disponibilità di dati tossicologici per un certo numero di molecole da usare come *training set* per un modello *QSAR,* è una *conditio sine qua non* per lo sviluppo di simili modelli. Per ottenere una *QSAR,* il rapporto ottimale tra numero di molecole del *training set* e numero di descrittori molecolari considerati dovrebbe essere di 5:1;[58] questa condizione limita enormemente il numero di *MD* utilizzabili con piccoli *training set* di molecole. Pertanto, si è optato per la costruzione di modelli *QSAR* monovariati con l'intenzione di isolare un'eventuale relazione della polarizzabilità α e della sua anisotropia $\Delta\alpha$ con la tossicità, al fine di aggiungere nuove informazioni sul meccanismo.

Sono stati identificati in letteratura i seguenti *training set* per i quali non esistevano calcoli *DFT*: 13 policloronaftaleni (*PCN*) [Villeneuve *et al.* 2000; Falandysz *et al.* 2001]; 14 benzochinoni (*BQ*) [Siraki *et al.* 2004]; 8 polibromodifenileteri (*PBDE*) [Meerts *et al.* 2001]. Per ognuno dei *training set* nominati erano riportati in letteratura più indici di tossicità utilizzabili quali *endpoint*.

Nel seguito vengono riportati i risultati delle regressioni lineari monovariate dei dati tossicologici per *PCN*, *BQ* e *PBDE* rispetto alla polarizzabilità e alla sua anisotropia che, pur essendo correlata alla prima, è meno sensibile di questa al tipo di approccio computazionale adottato nel calcolo. I dati relativi alla tossicità sono stati raggruppati in modo da provenire da un'unica fonte per ridurre al minimo l'incertezza dovuta alle differenze di protocollo sperimentale seguito dai ricercatori.

Per i *PCN,* i dati di tossicità sono tratti da Villeneuve [Villeneuve *et al.* 2000], e da Falandysz [Falandysz *et al.* 2001]. Gli indici di tossicità citati in entrambi gli studi sono stati ricavati tramite saggi *in vitro*. Questo tipo di tecniche richiede l'uso di particolari linee cellulari, dette immortali, che sono prive per natura del meccanismo dell'apoptosi, come le cellule tumorali o embrionali, o

[58] Regola di Topliss e Costello [Topliss *et al*. 1972].

che ne vengono private artificialmente. L'esposizione di queste cellule alla sostanza indagata per un tempo prefissato può, in alcuni casi, ucciderne una parte. Vengono così costruite delle curve dose/risposta[59] da cui si ricavano in genere gli indici *LC50*, [60] *LD50* in perfetta analogia con i *test in vivo*. Nel caso dei *test in vitro* su linee cellulari è possibile anche la titolazione per via citofluorometrica della quantità di sostanza che si lega alle cellule, là dove esistono nella cellula particolari recettori, presenti naturalmente o inseriti artificialmente, che possono essere attivati tramite enzimi. Da questo *test* si ottengono gli *EC50.* Villeneuve ha utilizzato nei *test* due tipi di cellule diverse, *PLHC-1*, cellule cancerose di fegato di topo, e *H4IIE* modificate in *H4IIE-EROD*, attivata tramite enzima Etossiresorufinadietilasi *EROD,*[61] e *H4IIE-luc*, cellule di epatoma di ratto, modificate per avere in sé la luciferasi. Come bianco (negativo) è stata utilizzata sieroalbumina bovina mentre come standard (positivo) è stata utilizzata la 2,3,7,8 tetraclorodibenzo-p-diossina (*TCDD*). I dati sono stati espressi come potenze relative riferite alla *TCDD*. Delle due linee cellulari quella risultata più utile è la *H4IIE*, essendo la *PLCH1* relativamente insensibile al *test*. Falandysz nel proprio lavoro ha riportato dati tossicologici provenienti dalla letteratura [Blankenship *et al.* 1999; Blankenship *et al.* 2000; Hanberg *et al.* 1990, Hansch *et al.* 1995; Villalobos *et al.* 2000; Villeneuve *et al.* 2000], citando gli stessi *test* e, in parte, le stesse molecole presenti nel lavoro di Villeneuve. Per i *PCN* possono essere identificati 75 congeneri. Dai lavori di Villeneuve e Falandysz possono essere ricavate informazioni tossicologiche per 24 molecole, per le quali sono stati calcolati gli *MD*, α e $\Delta\alpha$. In seguito si è deciso di utilizzare per le regressioni lineari solo i dati citati da Villeneuve, essendo possibile per questi risalire alla fonte di provenienza.

[59] Il termine inglese *response* è in questo caso, ai fini della QSAR, sinonimo di *endpoint*.

[60] *LC50* è la concentrazione letale per il 50% delle cellule, *LD50* la dose letale per il 50% delle cellule, *IC50%*, la concentrazione, o la dose, che inibisce la crescita per il 50% delle cellule, *IG50* la concentrazione, o la dose, che inibisce la crescita per il 50% soggetti pluricellulari, *EC50* è la concentrazione che causa una risposta del saggio pari al 50% tra la massima e la minima ottenute. Per ulteriori dettagli si veda il capitolo due.

[61] Nel test d'induzione di EtossiResorufina (O) Dietilasi (*EROD*), il substrato etossiresorufina viene idrolizzato a resorufina, un composto stabile e fluorescente.

In Tabella 5.2 vengono riportati i risultati del calcolo degli *MD*, i dati di tossicità riportati da Villeneuve, espressi come log*(EC50TCDD/EC50PCN)* per i due test, *EROD* e *Luciferase*, utilizzati nelle regressioni. Le regressioni sono mostrate nelle Figure 5.1-5.4. I dati contrassegnati con l'asterisco * sono stati giudicati da Villeneuve come affetti da alta incertezza nei risultati. Si noti che in alcuni casi i dati sperimentali sono diversi per una stessa molecola.

Tabella 5.2 Risultati dei calcoli degli *MD* e dati di tossicità provenienti da Villeneuve. In prima colonna i dati provenienti da Villeneuve in seconda quelli di letteratura. Il risultato dei *test* corrisponde al log(*EC50*TCDD/*EC50*PCN *nM in well*). I numeri in colonna *PCN* si riferiscono alla posizione del sostituente *Cl*.

PCN	α	$\Delta\alpha$	*testEROD-V.*	*testEROD-al.*	*test Luc.-V.*	*test Luc-al.*
1,4	144,21	112,06	-8,51*		-7,46	-6,70*
1,2,3,6,7	197,64	194,98	-4,12		-3,19	-3,77
1,2,3,7,8	193,62	178,47	-4,66		-4,34	
1,2,4,5,6	192,36	168,57	-5,80		-5,46	
1,2,3,4,5,6	207,71	183,99		-2,70		
1,2,3,4,5,7	208,83	180,89	-4,70	-4,70		
1,2,3,4,6,7	211,22	196,41	-3,20		-2,59	-2,41
1,2,3,5,6,7	210,69	196,17	-3,54	-2,70		-3,00
1,2,3,5,6,8	209,16	182,70	-2,70	-2,70		-3,82
1,2,3,5,7,8	208,89	182,92		-2,70		
1,2,3,6,7,8	211,41	198,97	-2,68		-2,00	-3,23
1,2,4,5,6,8	208,28	178,68		-5,15*		
1,2,3,4,5,6,7	225,50	202,89	-3,34	-2,52	-3,16	-3,00

Nelle figure seguenti, Figure 5.1-5.4, vengono presentati i risultati delle regressioni. L'analisi della Figura 5.1 permette di osservare che la *QSAR* è fortemente influenzata dal valore a più bassa polarizzabilità, relativo a 1,4-dicloronaftalene, che è pure caratterizzato da un'alta incertezza sperimentale.

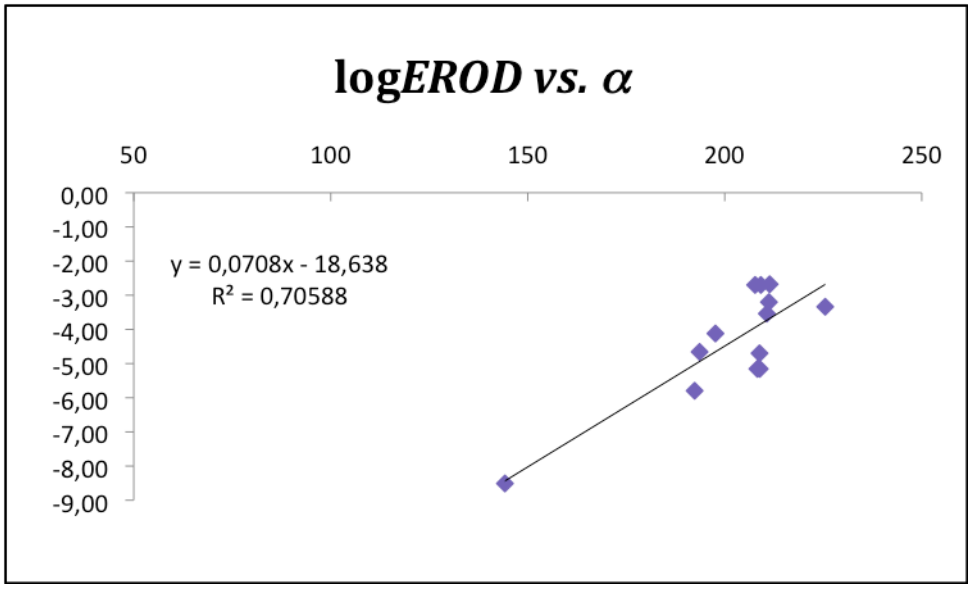

Figura 5.1 log(*EC50*TCDD/*EC50* molecola *nM in well*) *vs*. polarizzabilità α, test *EROD*, dati Villenueve; in ascissa α in (*a.u.*).

In Figura 5.2 è riportata la relazione con l'anisotropia della polarizzabilità $\Delta\alpha$.

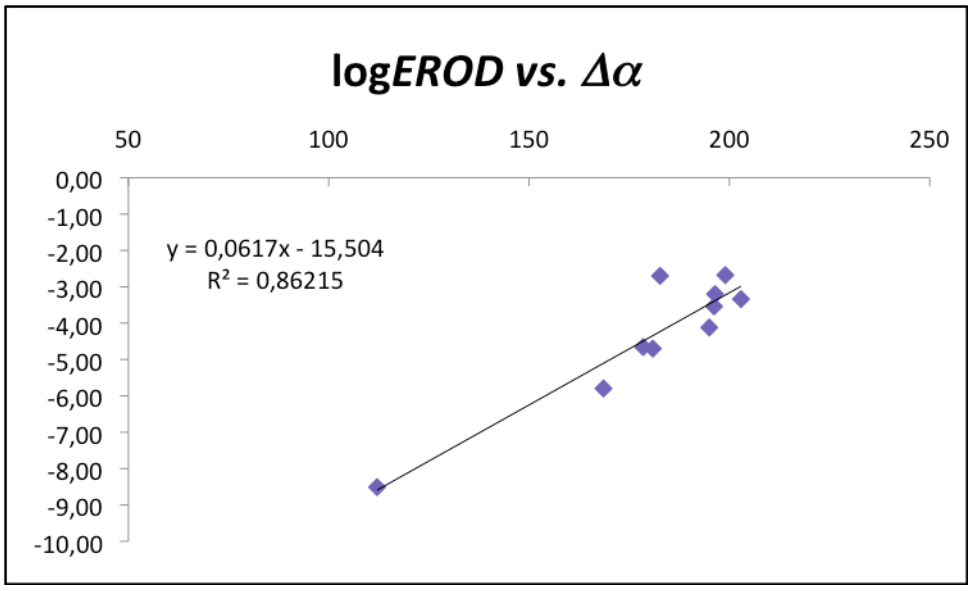

Figura 5.2 log(EC50TCDD/EC50 molecola nM in well) vs. anisotropia della polarizzabilità Δα, test EROD, dati Villenueve; in ascissa Δα in (a.u.).

Dall'analisi della Figura 5.2 si osserva quanto già sottolineato a proposito della regressione in Figura 5.1.

Nelle Figure 5.3 e 5.4 vengono esposti i risultati di regressione delle stesse molecole e *MD* verso un diverso *endpoint*, il *test* Luciferasi.

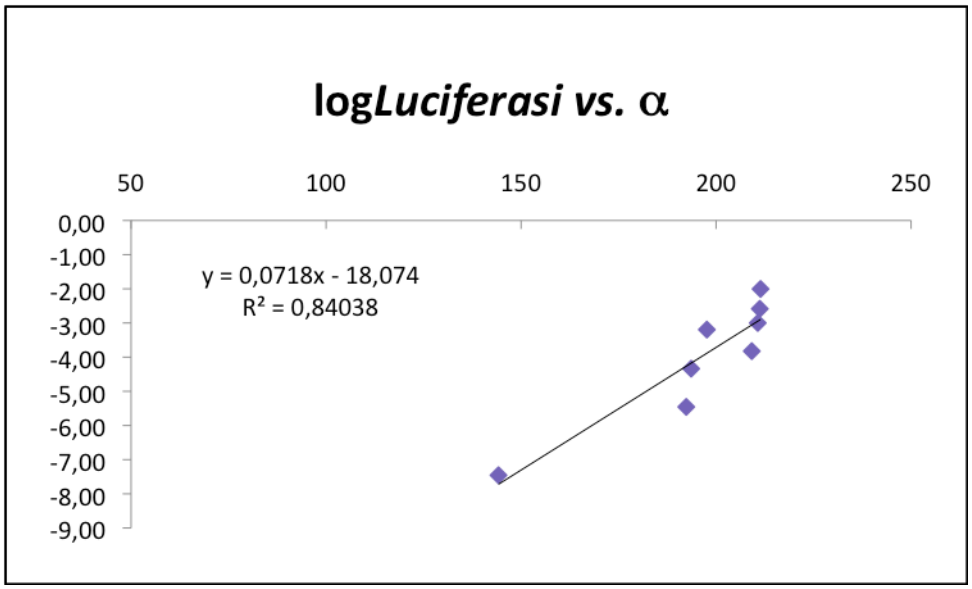

Figura 5.3 log(*EC50*TCDD/*EC50* molecola *nM in well*) *vs*. polarizzabilità α, *test* Luciferasi, dati Villenueve; in ascissa la polarizzabilità in (*a.u.*).

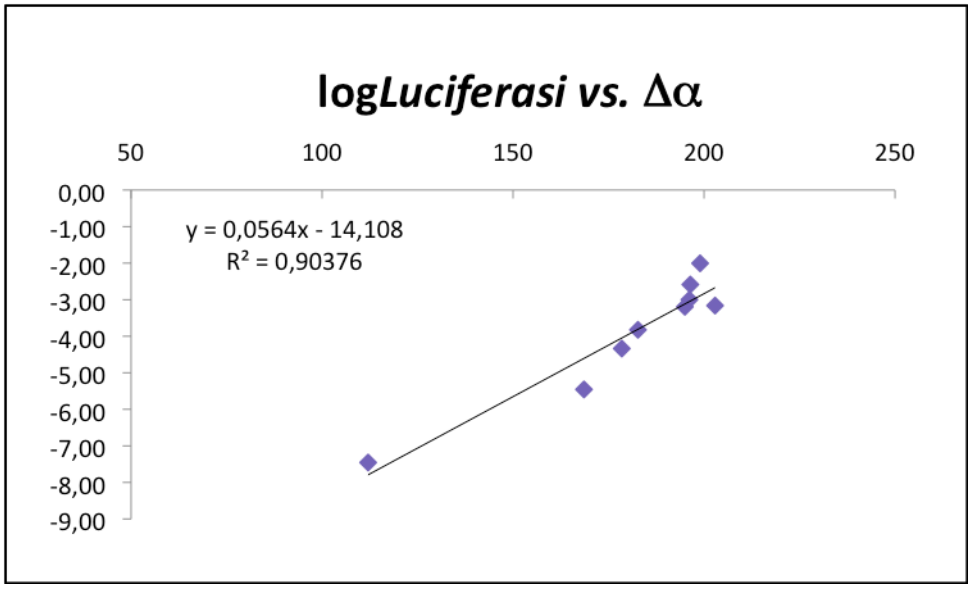

Figura 5.4 log(*EC50*TCDD/*EC50* molecola *nM in well*) *vs*. anisotropia della polarizzabilità $\Delta\alpha$, *test* Luciferasi, dati Villenueve; in ascissa $\Delta\alpha$ in (*a.u.*).

Siraki [Siraki *et al.* 2004] ha preso in esame una serie di 14 p-benzochinoni (non propriamente dei congeneri) ed ha proposto una *QSAR* utilizzando quattro descrittori molecolari: volume molecolare, affinità elettronica, costante di Hammet ed il potenziale redox monoelettronico

(un parametro strutturale e tre parametri legati alla elettrofilicità). Nel lavoro di Siraki sono riportati i dati di citossicità (espressi come concentrazione in μM di xenobiotico) riferiti alla *LD50* (a 120 minuti, saggio *MTT*),[62] alla formazione di specie ossidanti *ROS* (*EC*200[63] a 30 minuti, test *DCFH DA*), [64] e alla inibizione di *GSH* (*EC50* a 120 minuti) determinata per fluorescenza con o-ftalaldeide, [Hissin *et al.* 1976]. I *test* sono stati condotti su due linee cellulari diverse, epatociti di fegato di ratto e *PC12*.[65] Fra le molecole in esame il durochinone è stato identificato come *outlier* all'interno del modello. Le due linee cellulari scelte da Siraki non sono equivalenti. Gli epatociti sono cellule normali non immortalizzate, più delicate e simili alle condizioni *in vivo* di quanto non lo siano le *PC12*, tumorali e prive di apoptosi. Pur se la matrice dei dati riportata da Siraki non è completa per tutte le molecole, si hanno complessivamente sei possibili *endpoint.* In Tabella 5.3 si riportano i dati relativi a 4 *endpoint*: tossicità e consumo di *GSH* per entrambe le linee cellulari, più i risultati del calcolo degli *MD*, α e $\Delta\alpha$.

Dall'analisi dei dati in Tabella 5.3, si osserva come le ultime molecole delle serie abbiano un comportamento assolutamente diverso dalle prime, per cui non appare opportuno inserirle nello stesso dominio chimico. Questo andamento può essere meglio evidenziato attraverso dalle relazioni tra *LD50* e polarizzabilità, riportate nella Figura 5.5 per gli epatociti e nella Figura 5.6 per le *PC12*.

[62] Yellow MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, a tetrazole)
[63] *EC200* la concentrazione che causa il raddoppio del rapporto di risposta massimo. [Davidov *et al*. 2000]
[64] *DCFH-DA* (2',7'-dichlorofluoroscindiacetate) viene fatto reagire con le cellule di epatociti e viene determinato per fluorescenza.
[65] *PC12* è una linea cellulare cancerosa derivata da *pheochromocytoma* di midollo di topo [Greene *et al*. 1976].

Tabella 5.3 *LD50*(120') e EC50(120') per epatociti e *PC12* dati estratti da Siraki , α e Δα (a.u.) *PBE/TZP*. *ID* riporta la numerazione assegnata da Siraki alle molecole.

ID	α	Δα	LD50HEP	GSHHEP	LD50PC12	GSHPC12
1	131,69	106,12	18	16	10,1	12
2	105,13	98,09	23,3	17,1	14,2	31,3
3	157,22	149,8	32,5	24,4	20	18,1
4	127,11	70,33	32,5	31,8	12,5	100
5	135,31	102,67	43,2	18,3	16,3	26,2
6	88,81	69,60	45,7	37,2	25,2	22,1
7	75,02	67,61	56,6	25,4	20,8	63
8	119,13	92,24	81,5	71,1	13,7	100
9	102,73	71,42	80,5	63	18,4	140
10	103,07	76,27	90,5	80,2	19,5	150
11	203,94	132,48	513	236	62,5	353
12	273,9	164,81	506	75,3	55	135
13	129,16	80,33	800	> 800	388	>800
14	179,55	69,85	877	> 900	218	>800

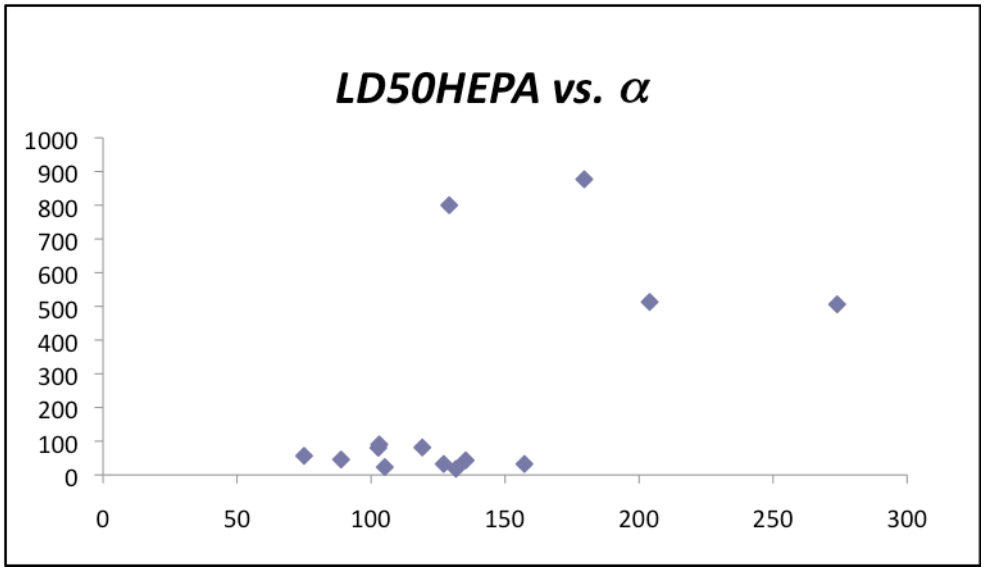

Figura 5.5 *LD50*(120') (μM) epatociti *vs.* α (a.u.).

Dall'analisi della figura si osserva chiaramente l'andamento anomalo delle ultime quattro molecole rispetto alle prime otto.

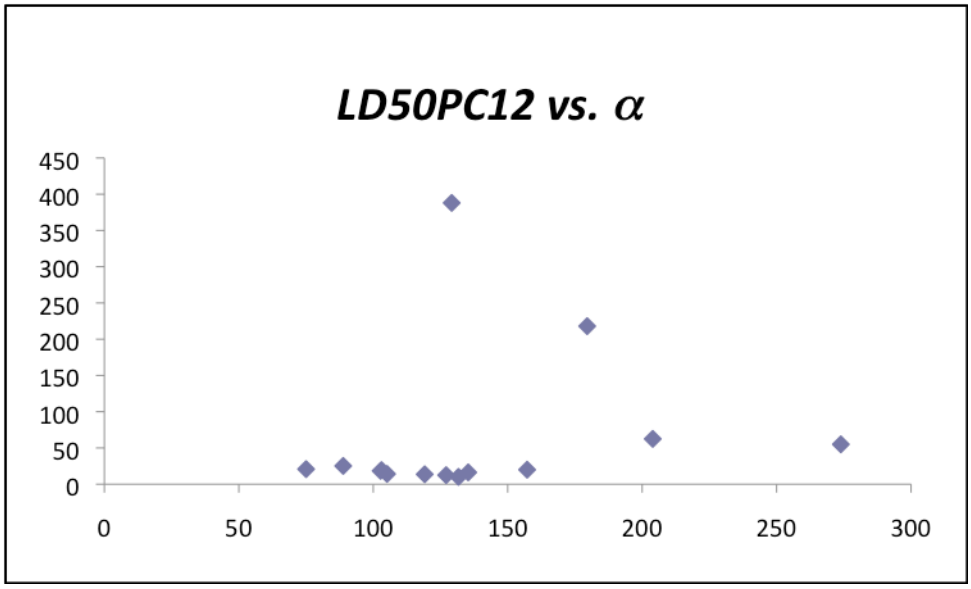

Figura 5.6 *LD50*(120') (μM) *PC12 vs.* α (a.u.).

Complessivamente, dall'ispezione delle Figure 5.5 e 5.6 non emerge alcuna relazione lineare tra l'*endpoint* e l'*MD*. Viene piuttosto confermato il comportamento da *outlier* del durochinone già evidenziato da Siraki. La necessità di considerare la presenza di *outliers* è molto frequente nello sviluppo di *QSAR*.

Il terzo gruppo di molecole che è stato preso in considerazione riguarda un gruppo di *PBDE* tratti da [Meerts *et al.* 2001]. Meerts ha utilizzato tre linee cellulari diverse *T47D-Luc*, *293-ER-*α*-Luc* e *293-ER-*β*-Luc* per studiare l'affinità di legame di una serie di *PBDE* e bisfenoli nei confronti del recettore degli estrogeni. Vista la somiglianza che viene attribuita nel *MOA* di questo recettore con *AhR* , utilizzare i dati di Meerts per sviluppare una *QSAR* con *MD* α e $\Delta\alpha$. Si è scelto di

utilizzare quali *endpoint* la potenza relativa rispetto all'estradiolo, *REC50*, [66] e *RLOEC*. [67] In Tabella 5.4 sono esposti i valori di log *REC50* e log*RLOEC* e polarizzabilità utilizzati.

Tabella 5.4 *LogRLOEC*, l*ogREC* per i *PBDE* tratti Meerts. α e $\Delta\alpha$ (*a.u.*) *PBE/TZP. ID* riporta la numerazione assegnata da Meerts alle molecole.

PBDE	α	Δα	logEC50	logREC50
28	210,94	135,94		
30	209,89	128,56	5,31E-01	-5,54E+00
32	208,29	119,27	7,08E-01	-5,72E+00
51	230,45	99,95	4,91E-01	-5,49E+00
71	230,27	133,45	8,63E-01	-5,85E+00
75	233,7	143,93	4,62E-01	-5,46E+00
85	258,34	165,63		
119	254,72	136,21	5,91E-01	-5,59E+00

In Figura 5.7 viene mostrata la regressione dei dati tossicità dei log*REC50* (*PBDE)* verso la polarizzabilità.

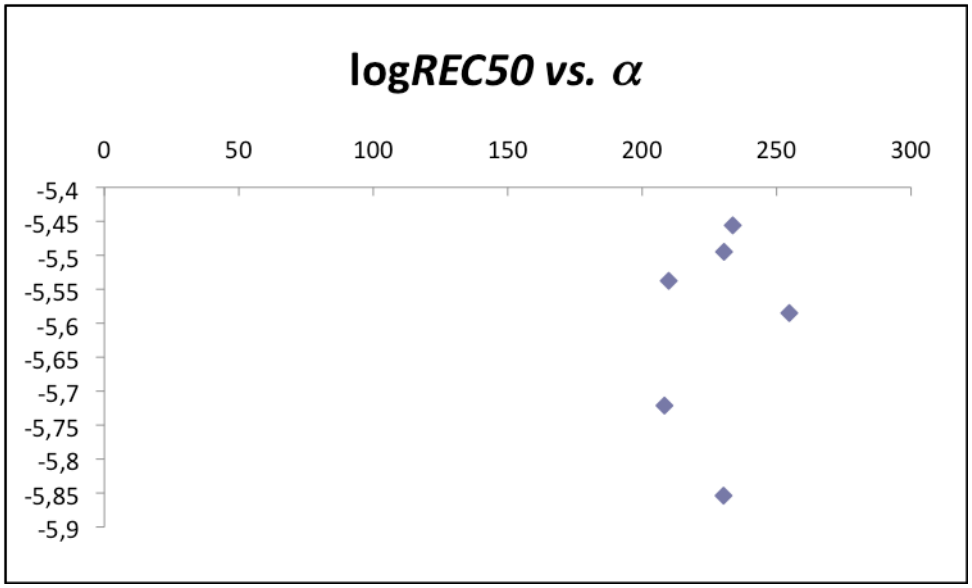

Figura 5.7 logR*EC50*, saggio *ER CALUX* con cellule *TD47DLuc*,[68] potenze relative riferite all'estradiolo, dati biologici tratti da Meerts vs. α (*a.u*).

[66]*REC50* rappresenta il rapporto tra *EC50* dell'estradiolo e quello del *PBDE,* in analogia a quanto visto per le diossine.
[67]*RLOEC* rappresenta il rapporto tra la concentrazione minima alla quale si osserva attività per l'estradiolo e quella del composto in esame.
[68]*Estrogen Receptor* (*ER*) *Chemical Activated Luciferase gene expression* saggio *(ER-CALUX). TD47DLUC* cellule di adenocarcinoma della mammella.

Non vi è alcuna relazione apparente tra le due variabili.

In Figura 5.8 è esposta la relazione tra log*RLOEC* e α.

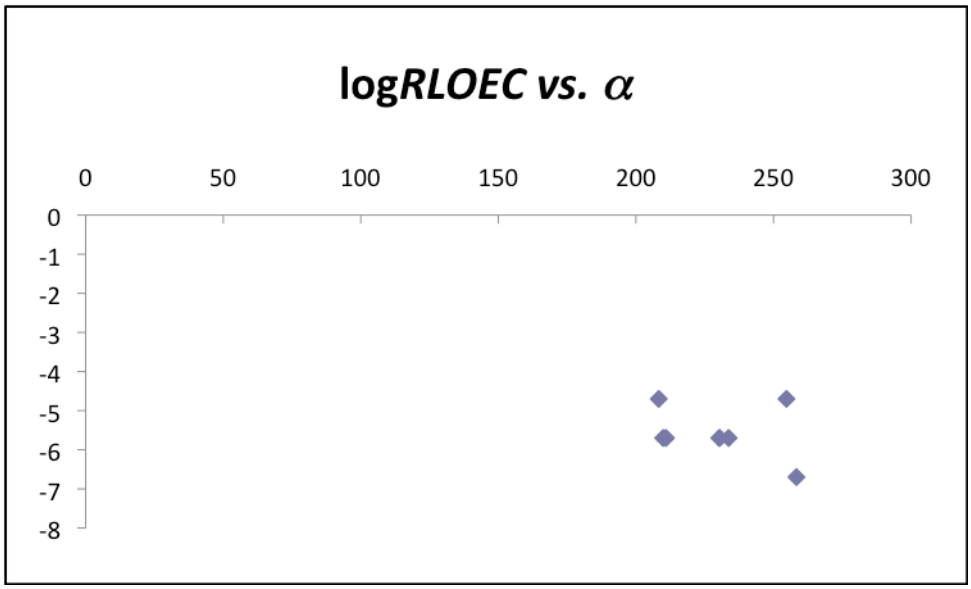

Figura 5.8 log*REC50*, saggio *ER CALUX* con cellule *TD47DLuc*,[69] potenze relative riferite all'estradiolo, dati biologici tratti da Meerts vs. polarizzabilità (*a.u*).

Dall'analisi della Figura 5.8, si osserva che i dati di tossicità sono pochi per poter sostenere l'esistenza di una relazione di qualche tipo con la polarizzabilità. Tuttavia, va notato che le tossicità riscontrate da Meerts sono basse (*EC50* dell'ordine di 10E-5 μM), molto simili tra loro, quindi non adatte a mettere in luce tendenze particolari dovute all'effetto della struttura molecolare su di esse. La presenza di dati biologici non adeguatamente dispersi è stata discussa da Dearden [Dearden *et al.* 2009] fra le cause di errore per i modelli *QSAR*.

I risultati delle regressioni sinora esposte portano a escludere, in due casi su tre almeno, l'ipotesi di una relazione lineare tra la tossicità e l'*MD* considerato. Inoltre, essi mettono in luce una serie di criticità.

Primo: i dati costituenti i *training set* sono pochi. La causa è la necessità di usare *set* di dati provenienti da una stessa fonte.

[69] *Estrogen Receptor* (*ER*) *Chemical Activated Luciferase gene expression* saggio *(ER-CALUX). TD47DLUC* cellule di adenocarcinoma della mammella.

1. Secondo: nel caso dei *PCN* i dati biologici sono male distribuiti in relazione al campo di esistenza della variabile indipendente, il che rende il modello eccessivamente sensibile ai valori estremi.

2. Terzo: nel caso dei *PBDE* i dati biologici sono appiattiti sugli stessi valori e non mostrano effetti diversi da molecola a molecola, quindi non sembrano adatti per essere sfruttati per questo tipo di indagine.

3. Nello sviluppo dei modelli occorre tenere conto della possibile presenza di "*outliers*", il cui inserimento fra le molecole del *training set* può influenzare i risultati. Un tipico esempio di *outlier* per le diossine è rappresentato dalla 2,3,7,8-TCDD notevolmente più tossica dei propri congeneri.

L'insuccesso dei primi modelli monovariati poteva essere ascritto a tre cause (non esclusive tra loro):

1. dati biologici dei *training set* potevano essere poco accurati. Essendo dati di letteratura, e non essendo riportate, nelle fonti, informazioni precise riguardo alle tecniche utilizzate per valutarne l'accuratezza, non era possibile effettuare nessun controllo;
2. il calcolo effettuato tramite la *DFT* dei valori degli *MD* poteva essere poco accurato;
3. gli *MD* scelti, polarizzabilità e sua anisotropia, potevano non avere alcuna relazione con la tossicità delle molecole inserite nei training set.

Al fine di escludere un errore dovuto alla scarsa accuratezza nel calcolo degli *MD,* si è deciso di tornare sul *training set* dei *DBF*, utilizzato sia da Hirokawa sia da ChengangGu; di costruire un modello utilizzando quale *MD* la polarizzabilità, calcolata con un potenziale di scambio e correlazione *PBE* e utilizzando un *set* di base *TZP* [Perdew *et al.* 1996; Perdew *et al.* 1998, Zhang *et al.* 1998]; e di confrontare, infine, i risultati ottenuti con quelli calcolati via *B3LYP/6-31G***

[Vosko *et al.* 1980; Lee *et al.* 1988; Becke *et al.* 1993; Stephens *et al.* 1994]. Entrambi gli autori citati riportavano, infatti, risultati secondo i quali α era un *MD* adeguato per la tossicità dei *DBF*.

Dal confronto dei risultati ottenuti con i due metodi si è ricavato che con *PBE/TZP* la polarizzabilità veniva stimata un po' più alta (18% circa), rispetto a *B3LYP/6-31G***. L'andamento era costante e riproducibile per gran parte delle molecole. In accordo a ciò l'anisotropia della polarizzabilità, invece, era identica per entrambi gli approcci. Pur se ai funzionali ibridi vengono normalmente riconosciute maggiori capacità di stima delle proprietà a lungo raggio rispetto a quelli puri (come il *PBE*), nel caso della polarizzabilità, il *PBE/TZP* si è rivelato in grado di riprodurre adeguatamente i dati sperimentali (Tabella 5.1). Inoltre, poiché ha il vantaggio di richiedere un tempo di calcolo inferiore agli ibridi, si è deciso di mantenere questo approccio.

A ulteriore conferma dell'accuratezza del metodo adottato (*PBE/TZP*), si sono effettuate due regressioni monovariate dei dati di tossicità dei *DBF* con i dati di polarizzabilità calcolati secondo i due diversi approcci di calcolo. I dati di tossicità dei *DBF* sono quelli riportati nel lavoro di ChengGangGu e sono riferiti a tre test diversi *BA*[70]*, EROD*[71] e *AHH.*[72] Si riportano in Tabella 5.5 i valori relativi ai dati tossicologici e agli *MD* calcolati con *PBE/TZP*. In Figura 5.9 e Figura 5.10 sono presentate le regressioni con i dati di *binding affinity,* quelli che costituiscono il *training set* più numeroso.

[70]*Binding Affinity BA* è riferito al saggio su *rat* hepatic cytosol.
[71]Saggio con EtossiResorufina-o-Dietilasi (*EROD*)su cellule cancerose di fegato di ratto *H-4-II E*.
[72]Saggio con *Aryl Hydrocarbon Hydroxylase AHH* su cellule cancerose di fegato di ratto *H-4-II E*.

Tabella 5.5 p*EC50 BA,* p*EC50AHH,* p*EC50EROD* per i *DBF*, tratti da Chengang Gu, α e $\Delta\alpha$ (*a.u.*) calcolate con *PBE/TZP*.

DBF	p*EC50BA*	p*EC50AHH*	p*EC50 EROD*	α	$\Delta\alpha$
123478DF	6,64	9,50	9,42	242,62	241,45
12348DF	6,92	6,68	6,79	222,93	210,90
123678DF	6,57	8,83	8,91	242,10	241,25
1236DF	6,46	4,00	4,00	206,86	193,04
12378DF	7,13	8,60	8,51	227,76	240,03
1237DF	5,00	4,92	7,07	211,98	224,80
124678DF	5,08	7,37	4,03	238,61	216,54
12467DF	7,17	6,49	7,53	222,80	204,28
12468DF	5,51	5,00	6,46	220,96	191,07
12478DF	5,89	6,98	4,92	224,76	211,92
12479DF	4,70	7,42	6,83	222,88	201,87
1248DF	6,96	4,57	7,42	205,78	183,62
13478DF	6,70	8,80	4,20	227,12	227,21
13678DF	6,70	8,10	8,85	226,55	222,88
136DF	5,36	5,60	8,24	192,07	175,02
138DF	4,07	4,71	5,47	193,48	193,81
234678DF	7,33	9,16	4,52	243,24	239,68
2346DF	6,46	5,88	9,24	207,72	190,91
23478DF	7,82	9,59	5,95	229,50	238,67
2348DF	6,70	7,38	7,83	210,14	206,97
234DF	4,72	6,82	7,43	193,10	183,75
2368DF	6,66	5,98	6,61	211,60	207,65
2378DF	7,39	8,41	6,11	215,28	238,98
23DF	5,33	5,60	5,81	179,34	184,10
26DF	3,61	4,21	5,32	176,61	166,15
28DF	3,59	4,40	5,50	177,76	173,78
2DF	3,55		4,20	161,88	157,08
3DF	4,38		4,40	164,32	170,93
4DF	3,00	5,00		160,57	142,23
DF	3,00			147,49	136,63

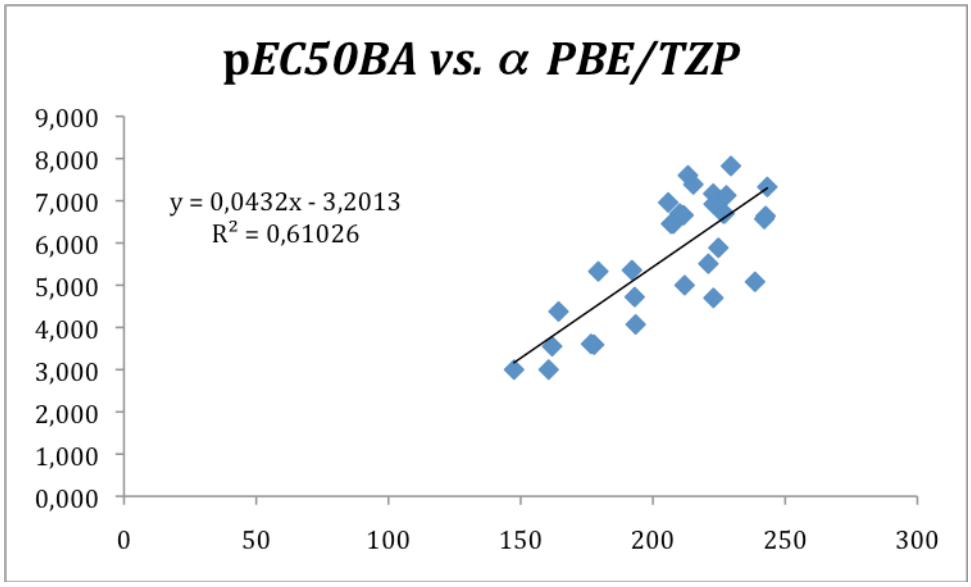

Figura 5.9 Regressione p*EC50*, test *BA*, dati biologici tratti da Chengganggu, con la α (*a.u*) calcolata con l'approccio di calcolo *PBE/TZP* (da noi adottato).

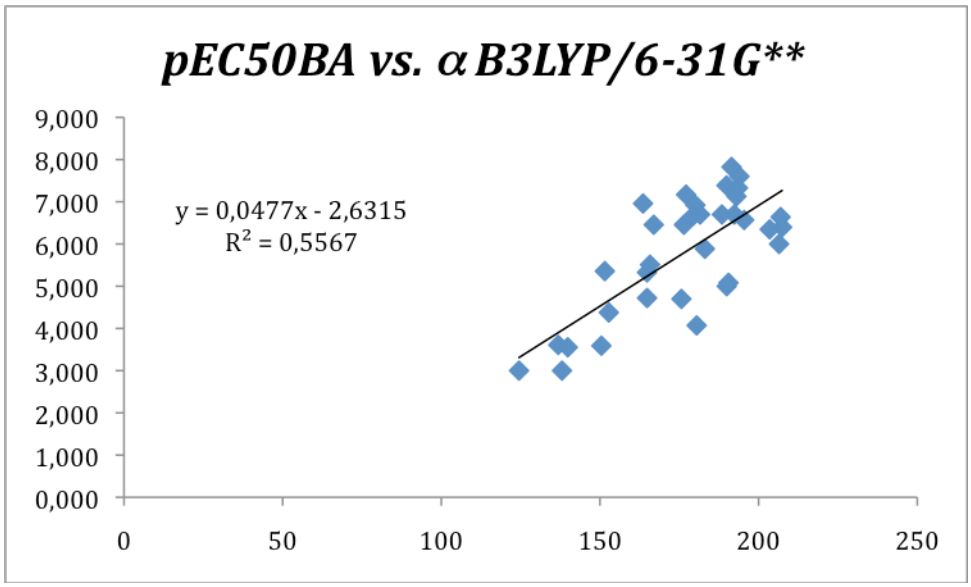

Figura 5.10 Regressione p*EC50*, test *BA*, dati biologici tratti Chengganggu, con la polarizzabilità α (*a.u.*) calcolata secondo l'approccio *B3LYP/ 6-31G*** utilizzato da Chenggang Gu.

Dall'ispezione delle Figure 5.9 e 5.10 non si osservano variazioni evidenti tra i due approcci di calcolo *PBE/TZP* e *B3LYP/6-31G***. Simili risultati si sono ottenuti considerando tutti gli *endpoint,* anche per la anisotropia della polarizzabilità $\Delta\alpha$. Si riportano in Tabella 5.6 i risultati ottenuti per il coefficiente di determinazione R^2 delle regressioni, al fine di effettuare un confronto tra le linearità dei modelli.

Tabella 5.6 Riassunto coefficiente di determinazione R^2

	PBE/TZP			*B3LYP/6-31G***		
	TEST			*TEST*		
MD	*BA*	*AHH*	*EROD*	*BA*	*AHH*	*EROD*
α	*0,61*	*0,56*	*0,55*	*0,60*	*0,56*	*0,55*
$\Delta\alpha$	*0,65*	*0,68*	*0,65*	*0,75*	*0,66*	*0,66*

Sulla base del solo R^2 non vi è alcun vantaggio apparente ad usare un metodo piuttosto che un altro. In ogni caso R^2 è compreso tra 0.55 e 0.75; la relazione tra la tossicità e α o $\Delta\alpha$ non è da ritenersi lineare.

Lo stesso *set* di *DBF* utilizzato da Hirokawa e Chengang Gu è stato studiato anche da altri autori [Sarkar *et al.* 2005; Roy *et al.* 2006]. Questi ultimi hanno identificato relazioni tra la tossicità dei *DBF* e altri *MD*. Si è deciso pertanto di calcolare per i *DBF* anche alcuni di questi *MD.* Sono state quindi calcolate le componenti del momento di quadrupolo (Q_{xx}, Q_{yy}, Q_{zz}), espresse in *a.u.* secondo la convenzione di Buckingam [Van Gisbergen 1999; Swart 2001], nonché gli indici di elettrofilicità globale ω [Parr *et al.* 1999] e locale ω_k. Il calcolo degli indici di elettrofilicità è stato effettuato nel modo descritto al **Capitolo 4**, utilizzando il metodo del *transition state* [Slater 1972].

Va osservato che nei lavori di letteratura sui *DBF,* pur se gli autori affermano di essersi basati sugli stessi dati biologici [Bandiera *et al.* 1984; Mason *et al.* 1985], i *training set* sono leggermente diversi, come se alcune molecole fossero state escluse dagli autori in quanto *outliers*. Merita di essere ricordato che l'incompletezza delle informazioni sui dati biologici è un'altra delle fonti di errore nelle *QSAR* discusse da Dearden. Da un confronto tra i dati di tossicità (p*IC50*) dei *DBF* citati in [Sarkar *et al*. 2006] e [Chenggang Gu *et al.* 2007] e ottenuti sulle stesse molecoleß con due saggi diversi per ottenere lo stesso indice, si ricava che tra i dati vi è un *bias* costante del 15%.

Per quanto riguarda la variabile indipendente, da un confronto tra gli indici ω e ω_k calcolati secondo lo schema di calcolo da noi adottato (*PBE/TZP*) con quello adottato da Sarkar [Sarkar *et al*. 2006] (*B3LYP/6-31G***), è risultato che tra i due schemi vi è un *bias*, quello da noi adottato fornisce valori del 20% inferiori. Al momento tuttavia non abbiamo elementi per stabilire quale dei due approcci sia il più accurato.

Vengono riportate nel seguito, Figure 5.11 e 5.12, due regressioni monovariate dei dati di tossicità (p*EC50BA*), dello stesso *training* set utilizzato da Sarkar, con gli *MD* ω e ω_k , calcolati secondo lo schema *PBE/TZP*.

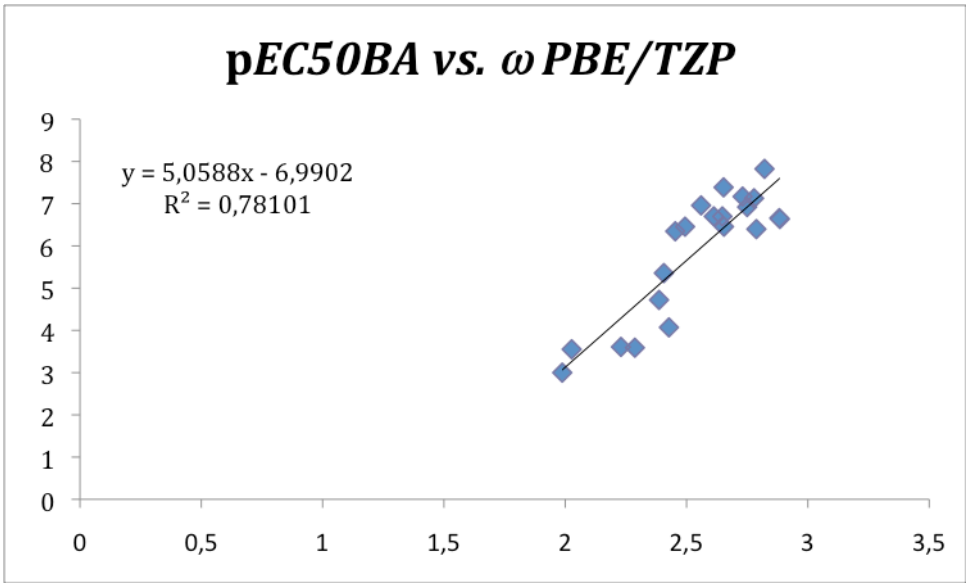

Figura 5.11 Regressione p*EC50 DBF*, test *BA*, dati biologici tratti da [Chengganggu et al. 2007], con l'indice di elettrofilicità globale calcolato con schema *PBE/TZP*.

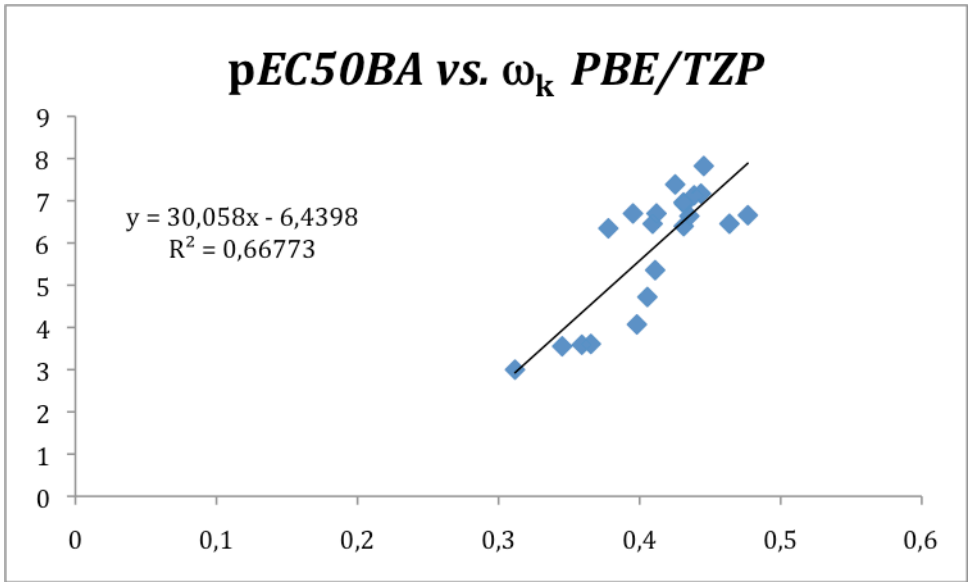

Figura 5.12 Regressione p*EC50 DBF*, test BA, dati biologici tratti da [Siraki *et al.* 2006] con l'indice di elettrofilicità locale ω_k calcolato con schema PBE/TZP.

Le regressioni monovariate con ω e ω_k hanno mostrato un comportamento analogo a quello di α e $\Delta\alpha$. Nessuno degli *MD* fin qui considerati ha mostrato una capacità particolarmente elevata di spiegare la tossicità delle molecole; non si può escludere l'esistenza di una relazione tra *MD* e dati biologici, ma non vi sono elementi sufficienti per sostenere, ammesso che esista, la linearità della relazione.

5.6 Modelli di regressione multivariata.

Gli esperimenti numerici condotti sui sistemi presi in considerazione hanno dimostrato che la teoria *DFT*, accoppiata al funzionale *PBE* con un *set* di base TZP, è sufficientemente accurata per il calcolo di descrittori molecolari quali α, $\Delta\alpha$, ω, ω_k, η. I risultati conseguiti permettono di correlare questi *MD*, pur se in modo approssimato (tramite regressione lineare monovariata), con la tossicità dei *DBF* e *PCN*; per contro, non vi è una correlazione evidente con la tossicità di *BQ* e *PBDE*.

I risultati dei calcoli, tuttavia, sembrano suggerire che un modello monovariato non sia sufficiente a spiegare un fenomeno complesso quale quello della tossicità. Ipotizzando quindi che quest'ultima sia funzione di più variabili contemporaneamente, è necessario utilizzare tecniche di analisi dei dati multivariata. Si è deciso quindi di costruire per i *PCN* un modello *QSAR* tramite una regressione lineare multivariata che tenesse conto di tutti i descrittori a nostra disposizione.

Il limite di questo tipo di analisi è dato sempre dal numero di molecole richiesto per il *training set.* Visto il numero esiguo di molecole (13) che componevano il *training set* per i *PCN*, questo ha limitato il numero di descrittori che potevano essere utilizzati.

Esistono molti pacchetti software in grado di effettuare simili calcoli. In questo lavoro di tesi si è utilizzato il *software open source R,* che offre la possibilità di effettuare regressioni lineari multivariate, *Principal Component Analysis*, *PCA*, *Partial Least Square Regression*, *PLSR*. La documentazione è resa disponibile dai programmatori sul sito di *R*.

Dai calcoli fin qui eseguiti avevamo a disposizione per i 13 *PCN* 11 descrittori molecolari: α, $\Delta\alpha$, α_{xx}, α_{yy}, α_{zz}, Q_{xx}, Q_{yy}, Q_{zz}, ω, ω_k, η.

É stata quindi fatta una regressione lineare multivariata dei valori di log(*EC50TCDD/EC50PCN*) dei *PCN* (dati ricavati da Villenueve con il test *EROD*) in funzione degli 11 descrittori molecolari. Di seguito viene riportata l'equazione di regressione ottenuta

utilizzando tutte le componenti in *Cross Validation (CV)* (vedi equazione 5.1) e anche i risultati dell'analisi della varianza (*ANOVA*) (vedi Tabella 5.3).

$$\log EC50 = 0,79\alpha - 0,10\Delta\alpha + 32\omega + 34\omega_k + 41\eta - 34Q_{xx} - 1,05Q_{yy} - 0,018\alpha_{xx} - 0,008\alpha_{yy} - 0,01\alpha_{zz} - 188 \qquad (5.1)$$

Tabella 5.7 analisi ANOVA modello 13 *PCN* 11 regressori

Errore residuo standard: 0.8701	gradi di libertà (*DF*): 1
statistica F: 4.186	*p-value*: 0.3654
R^2 multiplo: 0.9787	R^2_{adj} multiplo : 0.7449

L'analisi statistica mette in evidenza che, nonostante il valore del coefficiente di determinazione R^2 [Everitt 2000] inizialmente trovato sia apparentemente ottimo (0.98), il coefficiente R^2_{Adj}[73] corretto per i gradi di libertà presenta un valore nettamente inferiore (0.75). Inoltre il *p-value* [Sterne 2001] indica che l'ipotesi nulla (esistenza di una relazione lineare tra *MD* e tossicità) è in realtà da rigettare. Quello che emerge dall'analisi è che, usando tutti gli *MD*, i gradi di libertà sono troppo pochi.

Sulla base dei risultati conseguiti, si è deciso di incrementare il numero dei gradi di libertà eliminando alcuni dei regressori. Ovviamente sono stati eliminati quelli ritenuti meno significativi sia dal punto di vista chimico che da quello statistico in conseguenza della mutua correlazione. Sono state mantenute solo cinque variabili, α, $\Delta\alpha$, η, ω, ω_k. Di seguito si riportano la regressione multivariata (equazione 5.2), e l'*ANOVA*, Tabella 5.8.

$$\log EC50 = 0,34\alpha - 0,06\Delta\alpha - 10,7\omega + 20,4\omega_k + 16,5\eta - 34Q_{xx} \qquad (5.2)$$

Tabella 5.8 analisi ANOVA modello 13 *PCN* 5 regressori

[73] *Adjusted R squared* è una variante di R^2 che tiene conto del numero di regressori o variabili indipendenti impiegate nella regressione lineare.

Errore residuo standard: 0.701	gradi di libertà (*DF*): 7
statistica F: 13.09	*p-value*: 0.001934
R^2multiplo: 0.9034	R^2_{adj} multiplo : 0.8344

L'aumento dei gradi di libertà rende accettabile il *p-value*, autorizzando l'ipotesi dell'esistenza di una qualche relazione tra la tossicità e gli *MD* scelti.

Dai risultati esposti nelle equazioni 5.1 e 5.2 e tabelle 5.7 e 5.8 emergono due importanti risultati:

1. l'accettabilità (statistica) dell'esistenza di una relazione tra tossicità dei *PCN* e *MD* dipende dal numero dei regressori e dai gradi di libertà del sistema, in modo determinante, come richiesto dalla regola di Topliss e Costello;
2. il significato "chimico" degli *MD* non può essere ricavato dalle equazioni di regressione se prima non si è effettuato un opportuno *autoscaling* rispetto alla varianza. L'importanza di questa operazione è stata sottolineata anche Dearden nella propria analisi degli errori comuni nelle *QSAR*.

Cinque regressori con sole 13 molecole sono ancora troppi per poter ottemperare alla regola di Topliss e Costello. Conseguentemente, occorre utilizzare una tecnica di riduzione delle variabili. La *Principal Component Analysis* (*PCA*) [Shlens 2005] è una della tecniche statistiche che permettono di ridurre il numero delle variabili indipendenti in una regressione multivariata alle loro componenti principali, permettendo poi di effettuare la *Principal Component Regression* (*PCR*). Una tecnica analoga, più recente, è la *Partial Least Squares Regression* (*PLSR*). Quest'ultima è la tecnica utilizzata per l'analisi statistica in questo lavoro di tesi. Considerando il modello con 13 *PCN* e 11 regressori, sono stati presi in esame i dati di tossicità log(*EC50*TCDD/*EC50*PCN) dei 13 *PCN* e degli 11 descrittori sopra citati ed è stata eseguita una *PLSR* con 11 componenti principali (*PC*), allo scopo di esaminare la percentuale di varianza spiegata da ciascuna di esse.

Tabella 5.9 Percentuale varianza spiegata da ciascuna componente principale (13 *PCN*, 11 regressori).

PC	1	2	3	4	5	6	7	8	9	10	11
%	31,6	77,8	95,4	99,8	100,3	100,0	100,0	100,0	100,0	100,0	100,0
log*EC50*	51,4	63,4	69,5	82,6	84,3	87,8	91,0	91,5	95,6	97,2	97,9

I risultati riportati nella Tabella 5.9 dimostrano che non occorre prendere in esame tutte le componenti principali possibili. Già le prime tre sono sufficienti a spiegare il 95 % della varianza.

Nella figura 5.13 è riportato il risultato del test di validazione con *PLSR* a tre componenti:

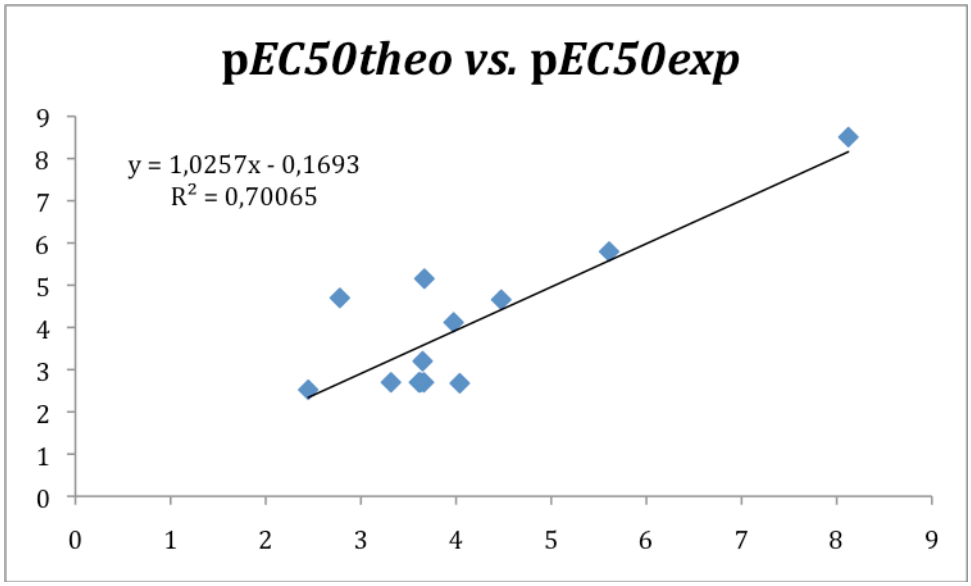

Figura 5.13 Test di validazione modello, (*response plot*). Analisi *PLSR* 13 *PCN,* 5 regressori, 3 *PC*, p(*EC50*TCDD/*EC50PCN*), test *EROD*, dati tratti da Villenueve.

Il grafico fornisce informazioni circa il "*valore predittivo"* del modello; più specificamente, una relazione lineare tra valori previsti e misurati sarebbe indicativa della validità del modello adottato. In questo caso il confronto tra valori previsti e misurati mostra alcuni valori che giacciono fuori dalla retta. Complessivamente, il risultato è di qualità nettamente inferiore a quanto ci si sarebbe aspettati dalla PLSR in confronto alla regressione multivariata. Infatti, esso va confrontato con quelli riportati in Tabella 5.8, e non era stato messo in evidenza dalle precendenti analisi con

regressione lineare mono e multivariata. Dal confronto con i risultati della regressione monovariata Figura 5.1 la PLSR non risulta avere particolari vantaggi anzi il *response plot* sembra confermare l'andamento della regressione monovariata (tenendo conto dei segni opposti del log).

Il *training set* di *DBF* considerato da Hirokawa e Chengang Gu era composto (nella versione più numerosa) da 34 molecole, quello di *PCN* preso in considerazione nel presente lavoro da 13. Per poter escludere il fatto che l'insuccesso apparente del modello sui *PCN* fosse dovuto al numero esiguo di molecole, piuttosto che a differenze nelle proprietà intrinseche tra *DBF* e *PCN*, era necessario poter considerare un *set* più vasto. X.F. Yan e altri nel 2006 effettuarono uno studio su 28 molecole di composti nitroaromatici [Yan *et al.* 2006]. Per ognuna delle molecole i dati di tossicità verso la specie *Fathead Minnow*, espressi come la concentrazione letale per il 50% dei soggetti (*LC50*) a 96h di distanza, furono tratti da un lavoro di Hall [Hall *et al.* 1989]. Yan *et al.* utilizzarono i seguenti *MD*: la differenza di energia E_{LUMO} - E_{HOMO}, la carica Q_c (*net atomic charge*) del *C* che fosse risultata massima fra i *C* sostituiti con un nitro gruppo, la carica del nitrogruppo Q_{NO2}, il momento di dipolo μ. Per il calcolo degli *MD* gli autori adottarono quattro approcci computazionali diversi, *AM1*, *PM3, HF/6-31G** e *DFT* (*B3LYP 6-311G***), così da metterne in luce eventuali differenze nelle prestazioni, quindi costruirono le rispettive *QSAR*. I modelli riportati nello studio di Yan sono tuttavia di difficile interpretazione; sono state, infatti, riportate solo le quattro migliori *QSAR* (quelle che, per ciascun approccio di calcolo, hanno dato i risultati di correlazione migliori). I modelli risultano, quindi, costruiti con *MD* diversi, per cui è difficile comprendere quali differenze di risultato siano da ascrivere agli *MD* e quali ai diversi approcci di calcolo. Si è deciso pertanto di utilizzare gli stessi dati di tossicità riportati nello studio di Yan per costruire una *QSAR*, con cinque *MD* (α, $\Delta\alpha$, η, ω e ω_k) calcolati con la *DFT* (*PBE/TZP*). L'equazione 5.3 è il risultato della regressione lineare multivariata di p*LC50* (*mMol/l*) con 5 componenti, con *cross* validazione interna; in Tabella 5.10 viene mostrata l'*ANOVA*.

$$pLC50 = 0{,}048\alpha - 0{,}022\Delta\alpha + 0{,}704\omega - 0{,}841\omega_k 1{,}092\eta - 5{,}714 \qquad (5.3)$$

Gli *MD* sono stati scalati rispetto alla deviazione standard (*mean centered scaling*), in modo che i loro coefficienti rispecchino il peso effettivo di ciascuno di essi nella regressione.

Tabella 5.10 ANOVA della regressione p*LC50* verso 5 regressori

Errore residuo standard: 0.52	gradi di libertà (DF): 22
statistica F: 8.565	p-value: 0.0001268
R^2 multiplo: 0.66	R^2_{adj} multiplo : 0.58

Il primo parametro da prendere in considerazione è il *p*-value. Dall'esame del *p value* (globale) si osserva come si abbia solo una probabilità di 1/10000 che nessuno degli *MD* considerati abbia una relazione con la tossicità, per cui si può ritenere probabile esista una *QSAR*. Incidentalmente, quale sia il *p value* minimo accettabile è una decisione arbitraria che dipende dal contesto; di solito non si accettano *p value* superiori al 5%. Il *p value* viene utilizzato nelle *QSAR* con approccio statistico anche per la scelta degli *MD*; vengono infatti scartati in modo iterativo quelli con *p value* maggiore del 5% (*stepwise linear regression*). Avendo seguito l'approccio quantomeccanico, gli *MD* sono stati in questo caso scelti sulla base di informazioni chimiche piuttosto che del *p value*. Il coefficiente di correlazione multiplo R^2_{adj} (che per le regressioni multivariate ha solo valore qualitativo) tuttavia è basso. Se quindi è probabile che la *QSAR* esista, è anche probabile che la relazione non sia lineare. In figura 5.14 si riporta il confronto tra i dati sperimentali espressi come p*LC50* (*mMol/l*) tratti da X.F.Yan e quelli stimati con il nostro modello *QSAR*.

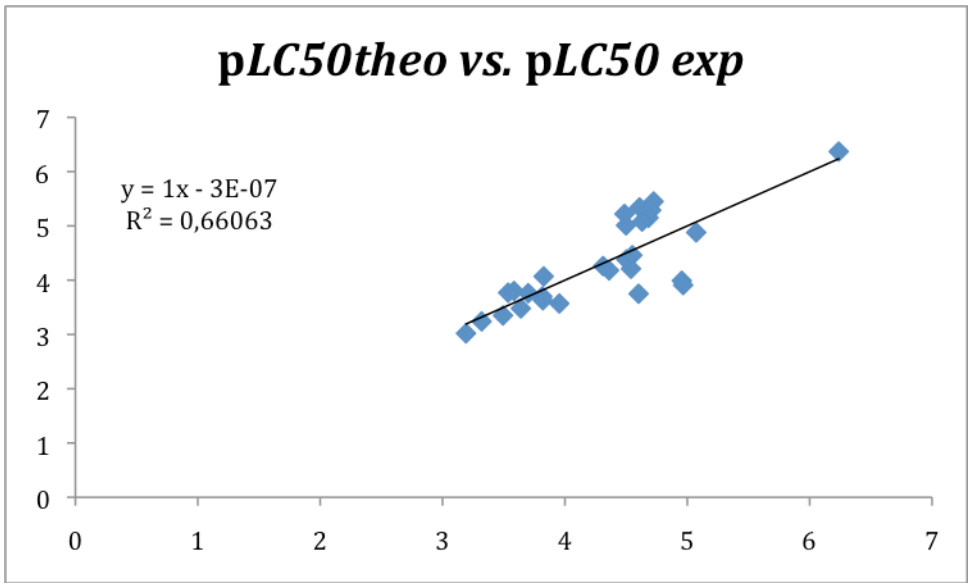

Figura 5.14 Valori di p*LC50* (mMol/l) , 28 *NB*, 5 *MD*, 5*PC* teorico verso sperimentale (*response plot*), dati sperimentali da X.F.Yan.

I dati riportati in Figura 5.14 sono congruenti con il fatto che per alcune molecole la relazione tra dati stimati e sperimentali non sia lineare. La tossicità delle molecole (2,4 dinitrotoluene; 2,6 dinitrotoluene; 3,5 dinitrotoluene) che si trovano sotto la retta viene sovrastimata dal modello; viceversa le molecole (1,4 dinitrobenzene; 1,3,5 trinitrobenzene; 2,3 dinitrotoluene; 2,5 dinitrotoluene; 3,4 dinitrotoluene; 2 metil 3,6 dinitroanilina) che si trovano sopra la retta vengono sottostimate dal modello. La conclusione è che le molecole che non giacciono sulla retta non vengono accuratamente descritte dal modello. Fra i motivi più frequenti di fallimento dei modelli *QSAR* vi è la presenza di *outliers*[74] e di *activity cliffs*[75] [Maggiora 2006]. Metodi per la identificazione ed il trattamento di *outliers* ed *activity cliffs* sono presenti in letteratura [Lipnick 1991, Jouan-Rimbaud *et al.* 1999, Verma *et al.* 2005; Rajarshi *et al.* 2008; Bajorath *et al*. 2009]. Fra questi vi sono l'analisi dei residui e del *leverage*[76] [Gramatica *et al.* 2005] che sono presentati in

[74]*Outlier* è una molecola che nel modello presenta un *residual standard error* particolarmente elevato.
[75]*Activity cliffs* sono comportamenti non lineari delle attività biologiche per cui certe molecole mostrano attività molto diverse dalle loro simili.
[76]Il *leverage* (detto anche *hat value h*) rappresenta la capacità di un dato di influenzare il modello con la sua presenza: più alto è *h* e maggiore l'influenza del dato sul modello. Il massimo valore di *leverage* accettabile è arbitrario e di

Figura 5.15 e Figura 5.16. In Figura 5.15 viene presentato il grafico dell'analisi dei residui verso i valori stimati di p*LC50* (*mM/l*). La numerazione attribuita alle molecole è la stessa dell'articolo di Xian [Xian *et al.* 2006] così da permettere un confronto diretto.

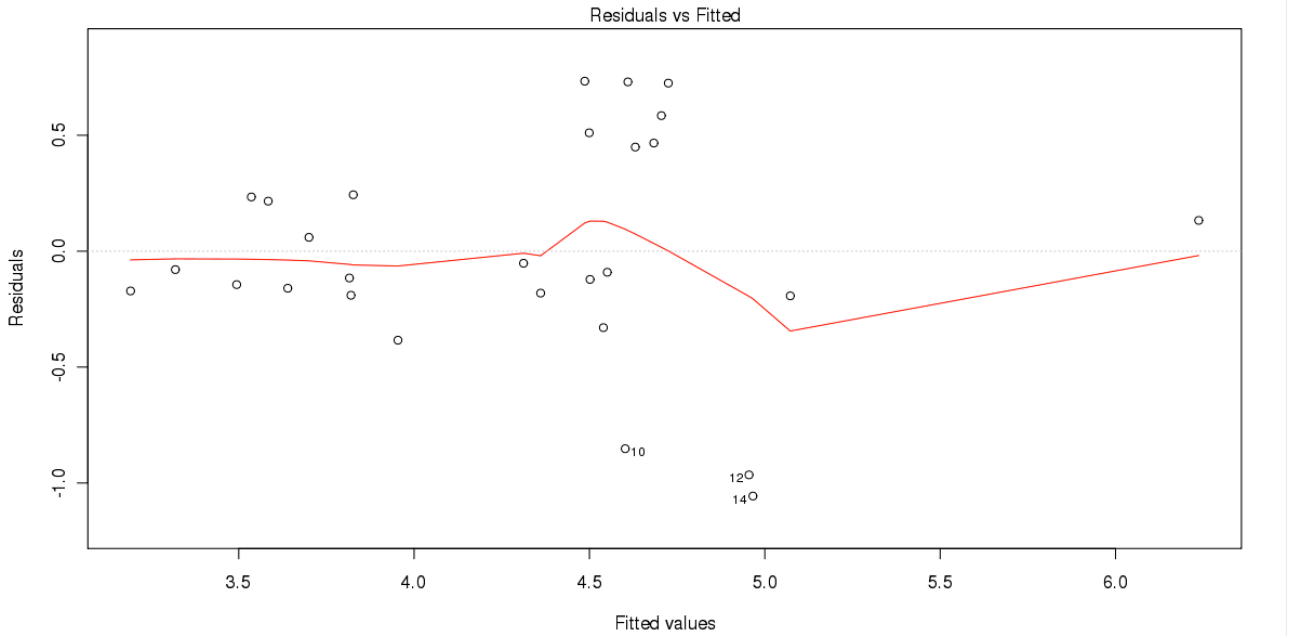

Figura 5.15 Grafico dei residui verso i valori di p*LC50* (mM/L) stimati (teorici).

I residui rappresentano lo scarto tra valore osservato e valore stimato. Dal grafico si osserva come per i composti 10 (2,4 dinitrotoluene), 12 (2,6 dinitrotoluene), 14 (3,5 dinitrotoluene), il modello sovrastimi del 20-25% la tossicità. La linea di tendenza (colorata in rosso) mette in evidenza l'eventuale eteroscedasticità (la dipendenza dei residui dal valore misurato). Il crescere del residuo al crescere del valore assunto dall'*endpoint* può indicare la presenza di errori sistematici; in questo caso non sembra vi sia una tendenza chiara.

In Figura 5.16 viene proposto un tipo di analisi alternativo dei residui, il cosiddetto diagramma di Williamson: residui standardizzati verso *leverage*. Lo scopo è quello di mettere in rilievo se quelle molecole per cui il residuo è alto (ovvero quelle che giacciono fuori dal modello

solito viene preso pari a *3p/n* (dove *p* è il numero degli *MD* considerati più uno ed *n* il numero delle molecole del *training set*).

lineare), siano o meno in grado di influenzare il modello avendo anche un alto *leverage*. I residui sono stati scalati dividendoli per la deviazione standard.

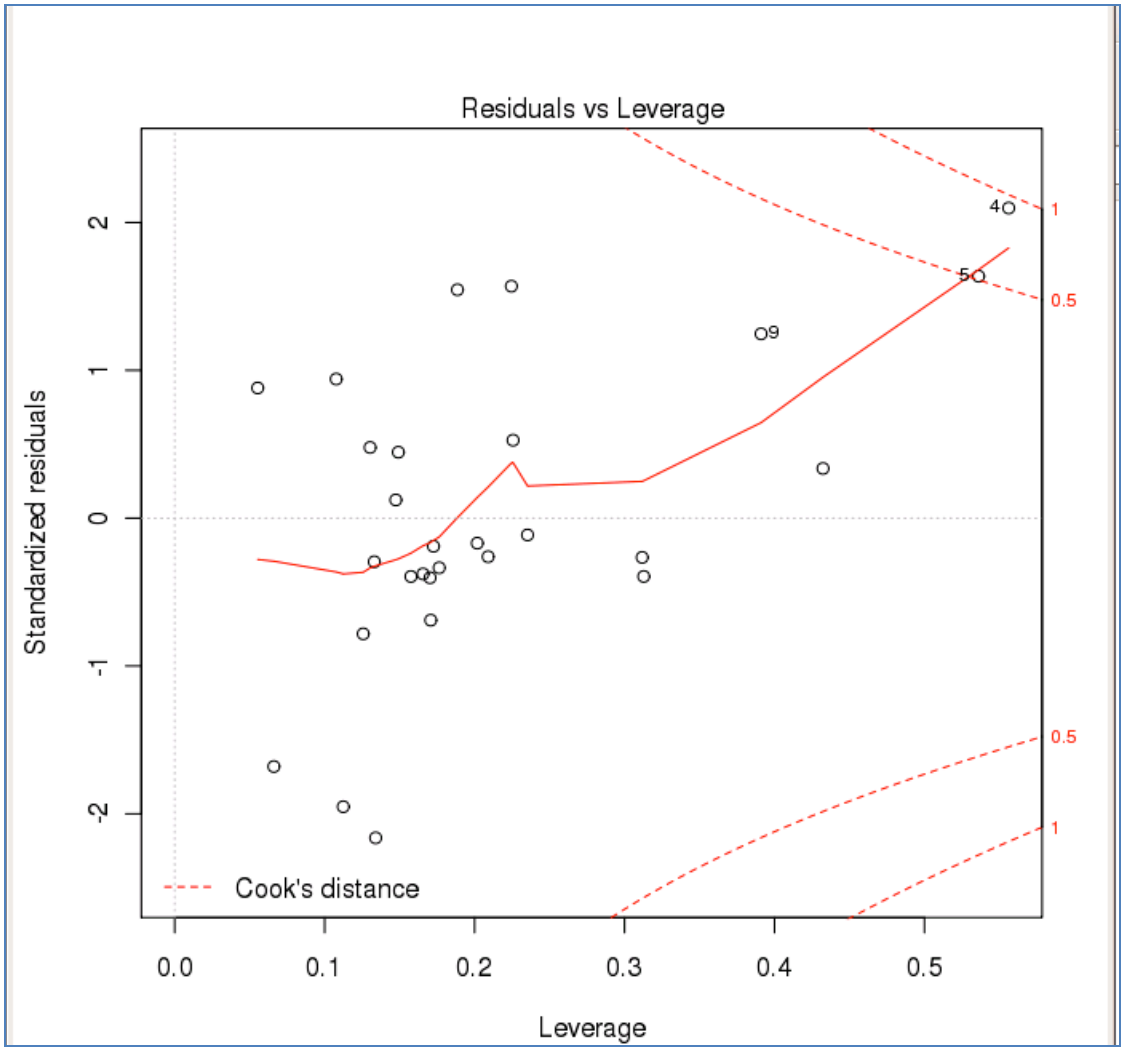

Figura 5.15 Residui standardizzati (rispetto alla DS) verso *leverage*.

Per il nostro modello il massimo valore *h* accettabile è circa 0,5. Dall'ispezione del grafico emerge che le molecole 4 (1,4 dinitrobenzene) e 5 (1,3,5 trinitrobenzene) con più alto valore di *h* vengono sottostimate dal modello per ciò che riguarda la tossicità. Se il residuo è alto il modello non riesce a spiegare il dato, se il *leverage* è alto il dato influenza molto il modello (la combinazione residui e *leverage* alti è deleteria per il modello *QSAR*).

Un'altra utile analisi è quella dell'indice di correlazione di Pearson, proposta in Figura 5.17. Qualora vi sia un sospetto che le variabili indipendenti non lo siano veramente tra loro, l'indice di

Pearson permette di mettere in luce eventuali casi di collinearità. Nel nostro caso, tutti i descrittori molecolari sono calcolati tramite la *DFT* ed in qualche misura derivano dalla densità elettronica, quindi è opportuno eseguire il test di Pearson per la collinearità.

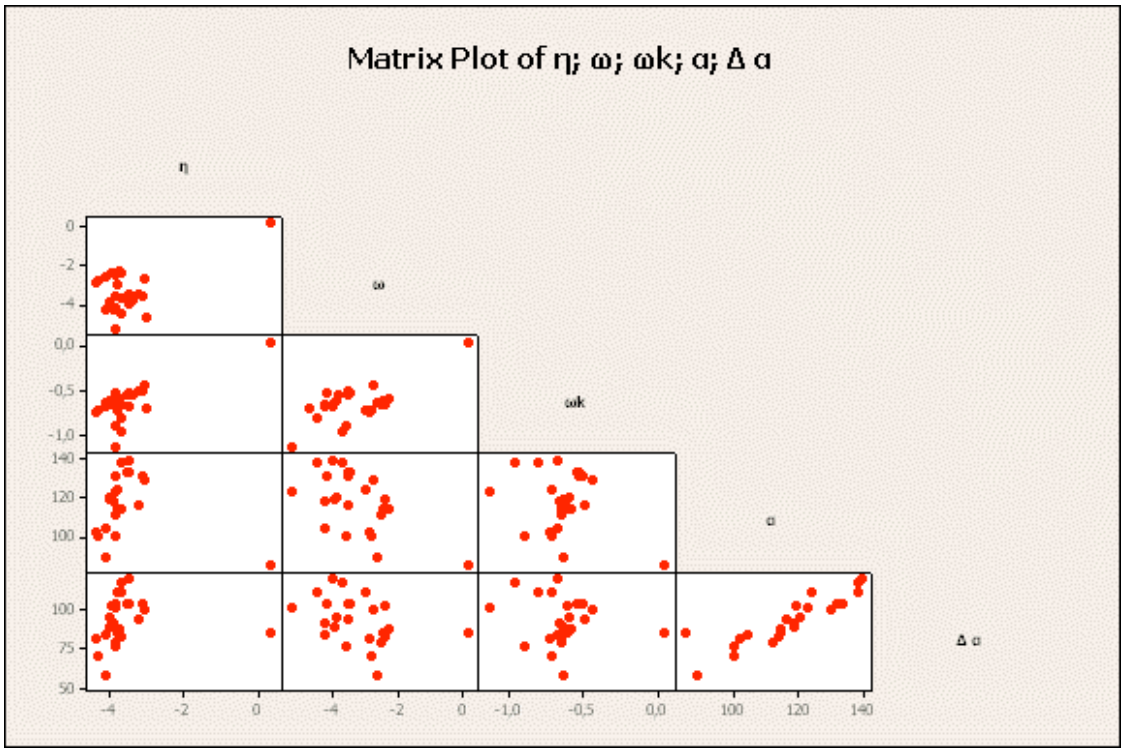

Figura 5.17 *Matrix plot* dei regressori, indice di correlazione di Pearson, 28 *NB*.

La lettura del grafico non è immediata ed è necessario procedere come segue: si incrociano le righe con le colonne, se le due variabili indipendenti sono correlate saranno rappresentate da una distribuzione di punti lineare. In questo caso si osserva come *hardness* e polarizzabilità, polarizzabilità e la sua anisotropia, siano correlate tra loro in qualche modo. Apparentemente gli *MD*, nonostante vengano calcolati tutti a partire dalla densità elettronica, non sono per questo correlati tra loro necessariamente. L'uso di *MD* altamente collineari è sconsigliato in quanto non aggiunge informazioni al modello ma solo rumore statistico [Dearden *et al.* 2009].

Merita di essere osservato che nel modello studiato da X.F. Yan non sono state inserite dall'autore tutte le molecole citate nello studio di Hall [Hall *et al.* 1989]. Lo studio di Hall nomina

105 molecole e non è chiaro come siano state selezionate le 28 comprese nel modello di X.F. Yan. Si è deciso quindi di costruire una *QSAR* utilizzando tutte le molecole disponibili nello studio di Hall. Tuttavia, per tre molecole la nomenclatura riportata da Hall non ha permesso di risalire alla struttura in modo univoco. In altri casi, il calcolo degli *MD* ha dato problemi di convergenza, per cui nel modello sono state inserite solo 89 molecole, per le quali erano stati considerati degli *MD* sufficientemente accurati.

Sull'intero *set* di molecole sono state eseguite la regressione lineare multivariata e la *PLSR* con *cross* validazione interna. Avendo poi a disposizione un numero di molecole sufficiente, si è provveduto a dividerle in due parti, in modo casuale, con lo scopo di effettuare una validazione esterna del modello. Le prime 70 sono state usate come *training set*, le rimanenti 19 sono state usate come *test set*. In equazione 5.4 e Tabella 5.11 sono riportate la regressione multivariata verso 5 regressori e l'*ANOVA* (*training set* di 89 elementi).

$$pLC50 = 0{,}0172\alpha + 0{,}0086\Delta\alpha - 0{,}180\omega - 0{,}176\omega_k - 0{,}521\eta - 1{,}0 \qquad (5.4)$$

I coefficienti degli *MD* sono stati scalati con la tecnica *mean centering,* per cui si può immediatamente osservare che il peso maggiore nella regressione è attribuito all'*hardness*, potendo quasi trascurare la polarizzabilità e la sua anisotropia Merita di essere sottolineato che η, α, $\Delta\alpha$, rappresentano, in genere, grandezze correlate tra loro.

Tabella 5.11 ANOVA della regressione p*LC50* verso 5 regressori.

Errore residuo standard: 43	gradi di libertà (DF): 83
statistica F: 8.565	p-value: 0.000
R^2 multiplo: 0.32	R^2_{adi} multiplo : 0.28
PRESS: 41,62	R^2_{pred}: 0.22

In questo caso sono stati inseriti nell'ANOVA anche il valore del *Predicted Residual Error Sum of Squares* (*PRESS*),[77] e del coefficiente R_{pred}[78] che estendono l'analisi al comportamento del modello sul *test set*. I risultati della regressione multivariata non si sono rivelati incoraggianti e, anche se il *p-value* indica che non si può rigettare l'ipotesi dell'esistenza di una relazione tra gli *MD* considerati e la tossicità, la relazione non sembra essere di tipo lineare. A questo punto si è ritenuto opportuno estendere l'analisi dei *p-value* ai singoli regressori. In Tabella 5.12 sono riportati i *p-value* ed i coefficienti per ciascun *MD*.

Tabella 5.12 ANOVA coefficienti di regressione e *p-value* dei singoli MD.

MD	α	$\Delta\alpha$	ω	ω_k	η
coefficiente	0,02	0,01	-0,18	-0,18	-0,52
p-value	0,14	0,48	0,18	0,74	0,06

Dai risultati esposti in tabella 5.12 si osserva che la possibilità che vi sia una relazione tra la tossicità e i singoli *MD* può essere presa in considerazione come una ipotesi da non rigettare limitatamente all'*hardness*. Questo risultato potrebbe essere dovuto ad un'alta correlazione tra le variabili indipendenti. Una delle condizioni perche sì possa fare una regressione lineare multivariata è che le variabili indipendenti siano veramente tali. Le variabili indipendenti altamente correlate tra loro non vanno usate perché portano poca informazione ed aumentano il rumore statistico. Uno dei modi per riconoscerle è l'analisi del coefficiente di correlazione di Pearson, i cui risultati sono riportati in Figura 5.18. In presenza di variabili altamente correlate tra loro è conveniente l'uso

[77] *PRESS*: è una misura della distanza tra il valore stimato per ciascun punto sulla base di un modello che comprenda tutti gli altri meno lo stesso punto (la somma totale dei quadrati degli errori). Molto usato in questo tipi di analisi è anche il coefficiente di correlazione *cross* validato ve *SD* è la deviazione standard. $q^2 = 1 - \frac{PRESS}{SD}$

[78] *Pred R-quadro*: una misura della quantità di varianza del *test set* spiegata dal modello.

della *PLSR* in luogo della regressione semplice. Le tecniche di riduzione delle variabili come la PCR o la PLSR attenuano il rumore contrastando anche l'effetto della correlazione.

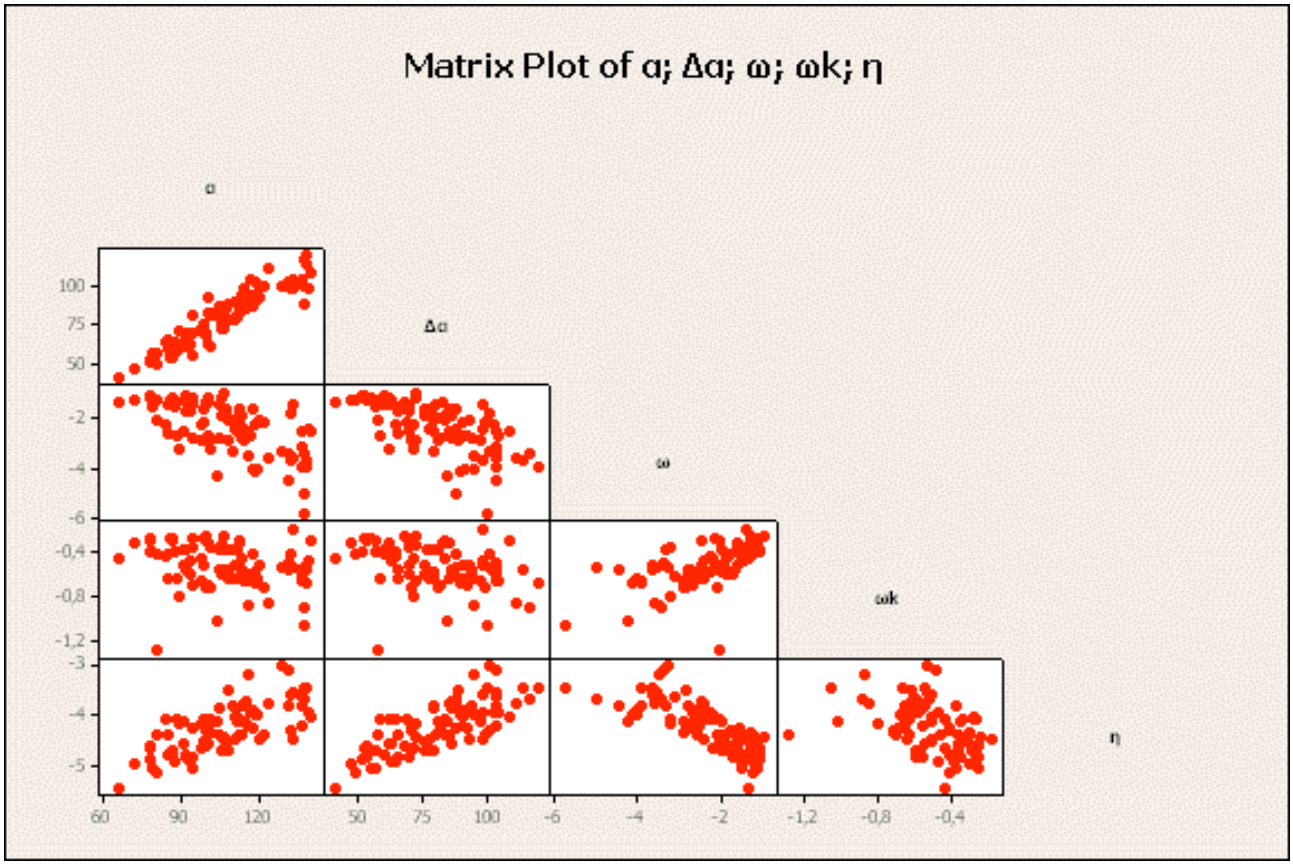

Figura 5.18 *Matrix plot* dei regressori, indice di correlazione di Pearson, 89 *NB*.

Dal confronto tra la Figura 5.17 e la Figura 5.18 si osserva come in quest'ultima la collinearità delle variabili indipendenti è più evidente a causa del *training set* più numeroso rispetto al caso dei 28 aromatici tratti da X.F. Yan.

La tecnica della regressione multivariata non sembra, perciò, adatta a risolvere il problema con una così alta correlazione tra gli *MD*, per cui si è effettuata la *PLSR*.

In Figura 5.19 vengono riportati i risultati della *PLSR,* più in particolare il *response plot* con due componenti, per il set completo in cross validazione interna con il metodo *Leave One Out* (LOO).

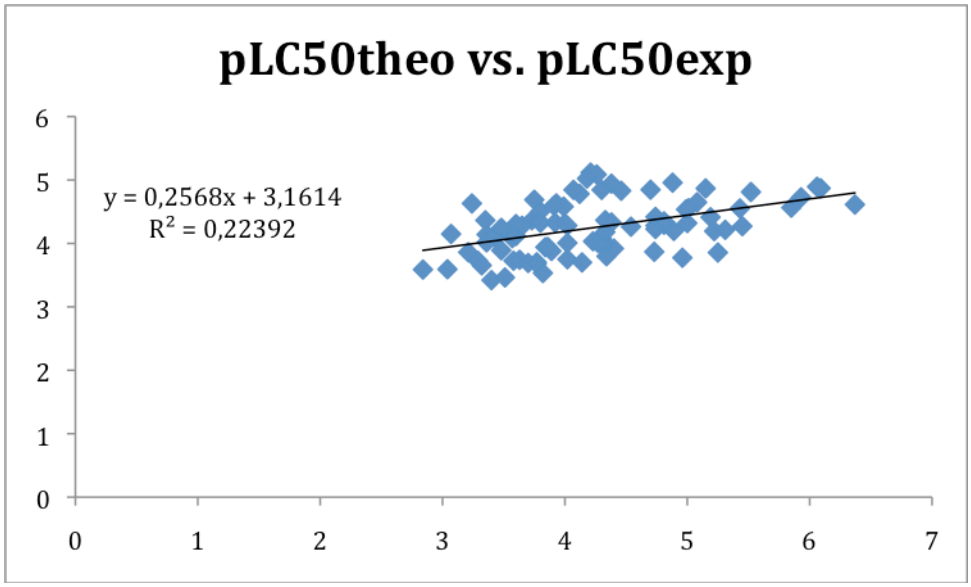

Figura 5.19 *Response plot PLSR* a due componenti, *cross* validazione interna LOO, 89NB.

L'ispezione della Figura 5.19, mette in evidenza che il valore di *endpoint* stimato dal modello è coincidente con quello sperimentale solo in pochi casi, risultando nella maggior parte degli altri alternativamente sopra o sottostimato.

Successivamente, è stata eseguita la *PLSR* con validazione esterna, utilizzando le prime 70 molecole quale *training set* e le ultime 19 quale *test set*. In Figura 5.20 si riporta il grafico della validazione dei risultati (*response plot,* valori stimati per l'*endpoint* verso i dati sperimentali) relativi al *test set* di 19 elementi.

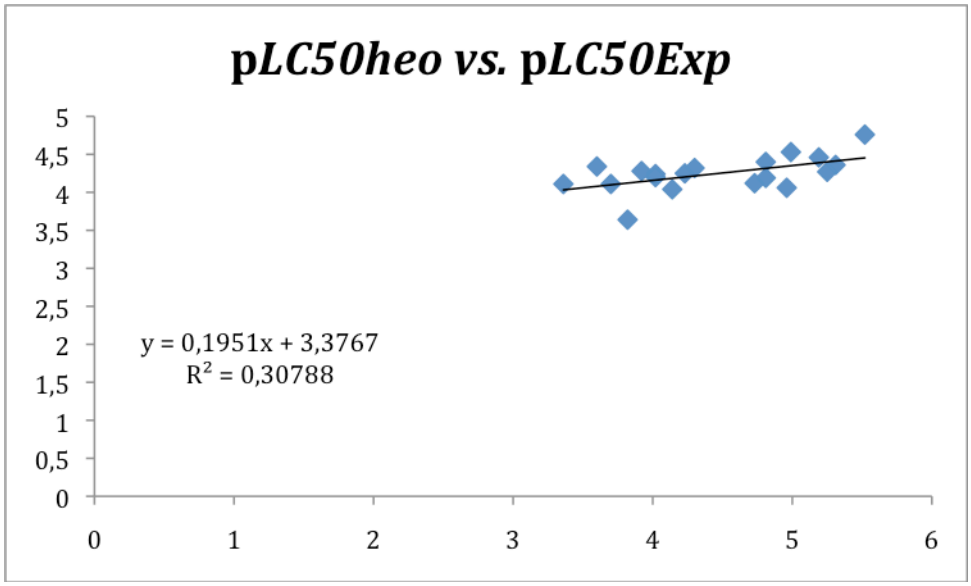

Figura 5.20 *Response plot test set*, PLSR a cinque componenti, validazione esterna.

Ancora una volta la correlazione tra dati teorici e sperimentali è modesta e il modello non riesce a stimare in modo soddisfacente gli *endpoint* per il *test set.*

Dall'esame dei risultati riportati in Figura 5.20, confronto tra valori teorici e sperimentali, si osserva che l'intero *set* di dati può essere diviso in due parti: una comprendente i dati per cui il modello sovrastima il risultato ed una per cui lo sottostima. Si sono divisi i due insiemi di dati ripetendo per entrambi la regressione mutivariata. L'equazione di regressione per l'insieme dei dati sovrastimati dal modello globale, quelli che hanno dato il risultato peggiore in termini di correlazione lineare, sono è la 5.5, nelle Tabelle 5.13 e 5.14 sono riportati i risultati dell'*ANOVA*.

Regressione multivariata p*LC50* verso 5 regressori, insieme delle molecole sovrastimate dal modello globale.

$$pLC50 = 0{,}027\alpha - 0{,}012\Delta\alpha - 0{,}150\omega - 0{,}429\omega_k - 0{,}294\eta - 0{,}081 \qquad (5.5)$$

Tabella 5.13 ANOVA della regressione p*LC50* verso 5 regressori.

Errore residuo standard: 40	Gradi di libertà (DF): 45
statistica F: 18,43	*p-value*: 0.000
R^2 multiplo: 0,67	R^2_{adj} multiplo : 0,66
PRESS: 3,15	R_{pred} : 0,61

Come si può osservare dalla equazione 5.5 e dalla Tabella 5.8, restringendo il set ai soli dati sovrastimati i risultati della regressione migliorano notevolmente. L'ipotesi di una relazione lineare non è da rigettare ed anche il coefficiente R^2 migliora notevolmente.

In Tabella 5.14 si riporta il confronto tra i *p-value* e i coefficienti di regressione.

Tabella 5.14 *ANOVA* coefficienti di regressione e *p-value* dei singoli *MD*, insieme sovrastimato.

MD	α	$\Delta\alpha$	ω	ω_k	η
coefficiente	0,03	-0,12	-0,15	-0,43	-0,29
p-value	0	0,06	0,02	0,08	0,07

L'analisi dei dati in Tabella 5.14 mostra che, in base al loro *p-value,* a tutti i regressori viene attribuita una correlazione molto più alta con la tossicità di quanto non avvenisse per l'intero *set* di dati, con la polarizzabilità che presenta un'alta probabilità di essere un buon *MD* per questo *set*.

I risultati per il *set* di dati sottostimato dal modello globale sono analoghi. Lo scopo di questa analisi è quello di mostrare la forte dipendenza di questo tipo di modelli dal numero e dalla natura delle molecole che vengono comprese nel *training set*. Questo fatto, non necessariamente negativo, potrebbe significare che il *MOA* nei due *set* di molecole non è lo stesso e quindi esse non possono essere inserite nello stesso modello. È stata fatta perciò un'analisi, per confronto, delle molecole contenute nei due *set* (sovrastimato e sottostimato) per verificare, ad esempio, che ai due estremi non vi fossero le molecole con sostituenti elettrondonatori, da una parte, e elettronattrattori, dall'altra. La presenza contemporanea di sostituenti di entrambi i tipi sulle molecole rende l'analisi possibile solo a livello qualitativo. Merita comunque di essere osservato che la molecola per la quale il modello globale ha sovrastimato l'*endpoint* in modo più evidente è il toluene, e, per contro, quella per cui la sottostima dell'*endpoint* è risultata maggiore è il 2,3,6-trinitrotoluene. Pur con notevoli eccezioni, si potrebbe osservare che fra le molecole derivate dal benzene, comprese nel *set*

identificato da Hall [Hall *et al.* 1989], il modello *QSAR* sviluppato tende a sovrastimare la p*LC50* di quelle con sostituenti elettrondonatori (come $–NH_2$), mentre tende a sottostimare l'*endpoint* per quelle con sostituenti alogeni o elettronattrattori (come $–NO_2$).

A conclusione di quest'analisi si ribadisce ancora una volta l'importanza della conoscenza del *MOA* nella valutazione dell'utilizzo di un modello *QSAR*.

Capitolo 6

Conclusioni

Lo scopo di questa tesi è stato dichiarato al capitolo primo. Prendendo lo spunto dall'entrata in vigore del *REACH*, si è voluto verificare la possibilità dell'utilizzo della Teoria del Funzionale Densità per il calcolo di *MD* da inserire in modelli *QSAR* per la valutazione della pericolosità intrinseca di sostanze d'interesse industriale.

L'idea di utilizzare la *DFT* per la costruzione di modelli *QSAR* con *MD* quantomeccanici non è nuova, è presente in letteratura in vari lavori [Roy *et al.* 2006; Sarkar *et al.* 2005] già citati nel corso della tesi. La novità di questo lavoro di tesi è rappresentata, invece, da due idee:

1. l'applicazione delle tecniche *QSAR*, sviluppate nell'industria farmaceutica a partire dagli anni 60 [Hansch et al.1962] a problemi di igiene industriale;
2. il passaggio dalla tossicologia computazionale alla tossicologia teorica.

L'entrata in vigore del *REACH* nel 2007, con il carico di lavoro di ricerca richiesto alle aziende per la implementazione del regolamento, ha posto il problema della necessità di tecniche utili per minimizzare e razionalizzare gli esperimenti necessari. A questo scopo erano nate le *QSAR*, fin dai tempi della *Hansch analysis*, è stato quindi naturale che si pensasse a una loro applicazione anche in campo *REACH*. Tuttavia, merita di essere osservato che solo una minima parte delle *QSAR* sviluppate negli ultimi anni è stata orientata al *REACH*. Nella tabella 6.1 sono riportati i risultati di una ricerca nella banca dati *ISIweb of Science* riguardo alle pubblicazioni fatte sull'argomento *QSAR* dal 1990. Le parole chiave utilizzate erano *QSAR, QSAR and REACH, QSAR and DFT*.

Tabella 6.1 Numero di pubblicazioni riguardanti *QSAR* e *REACH* a partire dal 1990, fonte *ISI*.

	1990-2009	2007-2009	2009
QSAR	8399	2430	722
QSAR and REACH	62	43	5
QSAR and DFT	123	75	18

Dall'ispezione della Tabella 6.1 si possono trarre almeno tre conclusioni:

1. circa un terzo delle pubblicazioni con oggetto le *QSAR*, pubblicate negli ultimi 20 anni sono state pubblicate durante questa tesi di dottorato;
2. l'entrata in vigore del *REACH* non ha influenzato significativamente lo studio delle *QSAR*;
3. l'utilizzo della *DFT* per lo sviluppo di *QSAR*, rappresenta un campo di ricerca nuovo e più di metà delle pubblicazioni fatte in proposito sono state fatte nel corso di questa tesi di dottorato.

Lo scarso impatto che il *REACH* ha avuto sull'interesse scientifico per le tecniche di valutazione *in silico* come le *QSAR*, può anche essere spiegato dal fatto che il *REACH*, pur rappresentando una rivoluzione normativa per il settore chimico, non si applica all'industria farmaceutica, che è regolata da normativa specifica.

Il ruolo che la *DFT* può avere nell'igiene industriale, e nella tossicologia in generale, merita un approfondimento ulteriore.

La *DFT* deve il suo enorme successo nell'ambito della chimica computazionale alla capacità di fornire soluzione a problemi di enorme complessità a un costo computazionale accessibile, essa permette di stimare e interpretare su base teorica risultati sperimentali relativi a sistemi complessi non affrontabili con altre metodologie teoriche. Nella tossicologia non è chiaro quali siano i risultati sperimentali che debbano essere confrontati con la *DFT*. L'apparente insuccesso dei modelli presentati nel capitolo 5 non è certamente dovuto alla inadeguatezza della *DFT* nel calcolo di proprietà molecolari quali la polarizzabilità, l'*hardness* ecc., ma alla inadeguatezza degli *endpoint* che con queste proprietà molecolari vengono messi in relazione. Nello sviluppo di *QSAR* con approccio quantomeccanico si dovrebbe evitare di utilizzare per gli assi delle due variabili del modello, indipendente e dipendente, due scale diverse. Da una parte la *DFT* permette una discussione approfondita delle proprietà molecolari su scala nanoscopica, dall'altra non vi è chiarezza su come possa essere definita, o esclusa, la tossicità di una molecola e si vorrebbe trovare una relazione (lineare) di queste proprietà molecolari con una quantità di *endpoint* diversi. Qualsiasi

modello che venga sviluppato su queste basi è destinato a sollevare forti perplessità sulla propria validità scientifica. Se si misurano, o si calcolano, proprietà molecolari sull'asse delle *x* occorre avere *endpoint* molecolari altrettanto adeguati sull'asse delle *y*. Chi scrive ha avuto modo di sviluppare una simile opinione nel corso della tesi, ma l'idea di dover valutare necessariamente la tossicità chimica su basi molecolari è ben radicata in letteratura [McKinney 1985].

Nella tossicologia industriale, fino agli anni 60 del secolo scorso, ogni valutazione veniva fatta prevalentemente sul piano clinico. Con lo sviluppo di nuove tecniche di Chimica Analitica vi è stato spazio per un forte ruolo dei chimici, volto prevalentemente a determinare presenza e quantità delle sostanze oggetto d'indagine. Oggi le tecniche che vengono comunemente raggruppate con i nomi di proteomica e metabolomica, permettono di vedere quali sono gli effetti di un *chemical* sugli esseri viventi a livello cellulare. Queste devono essere utilizzate per definire gli *endpoint* da usare nei modelli *QSAR* con approccio quantomeccanico. È sorprendente come una gran quantità di modelli *QSAR* in letteratura si basino in realtà su pochi dati tossicologici, anche datati. Molti di coloro che hanno studiato i *DBF* si sono basati di fatto sui medesimi lavori di Bandiera, Mason, McKinney. L'acquisizione dei dati biologici di partenza sembra essere uno dei punti cruciali dello sviluppo di modelli *QSAR*. Questo problema è strettamente legato alla difficoltà di definire su scala quantitativa la tossicità. Per esempio, se si considera la dose di 0,12 mg/L di naftalene, questa risulta essere il *NOEL* per il salmone (a 40 giorni di distanza), mentre ne basta una di 0.11 mg/L per uccidere il 50% delle trote d'acqua dolce che vi sono esposte (*LD50*). Altre volte, invece, per uno stesso indice si trova un intervallo di numeri in letteratura, ad esempio *EC50* a 48h del naftalene verso gli invertebrati è compreso tra 2,1 e 24 mg/L. In questo caso quindi il dato numerico dipende dalla convenzione adottata sulla tossicità. Senza un riferimento opportuno si pone un problema di scala che potrebbe nascondere la *QSAR* con un qualsivoglia *MD*. Una soluzione potrebbe essere quella di eseguire modelli multivariati anche sulla variabile dipendente oltre che su quella indipendente per considerare più *endpoint* contemporaneamente. Questa possibilità è stata oggetto

di indagine nell'ultimo anno di tesi ma il problema rimane, al momento, senza soluzione (per lo meno non è nota a chi scrive).

In questa tesi sono state utilizzate tecniche di analisi statistica dei dati ma si è volutamente omessa una discussione approfondita delle stesse, non essendo oggetto dell'argomento di tesi. Il lettore interessato potrà trovare informazioni sulle tecniche statistiche usate nelle *QSAR* e sui problemi relativi al trattamento di *outliers* e *activity cliffs* negli articoli citati nel testo e nei seguenti: [Wold 1985; Lipnick *et al.* 1991; Wold *et al.* 1998; Rimbaud *et al.* 1999; Wold 2001; Wold *et al.* 2001 a); Wold *et al.* 2001 b); Golbraikh *et al.* 2002; Forina *et al.* 2003; Aptula *et al.* 2005; Verma *et al.* 2005; Guha *et al.* 2008; Tetko *et al.* 2009; Bajorath *et al.* 2009]. Per quanto riguarda i problemi del dominio di applicabilità e la validazione statistica dei modelli *QSAR* si vedano [Gramatica 2004; Tropscha *et al.* 2003; *ECHA* 2008; Horvath *et al.* 2009].

In questo lavoro di tesi è stata posta particolare enfasi sul ruolo del *MOA*. Un'ipotesi sul meccanismo è necessaria per determinare se esiste o no il *minimal cut set*, postulato della nostra ricerca sulle *QSAR*, che porta fenomeni su scala nanometrica (come le proprietà molecolari) a ottenere un effetto macroscopico come la tossicità. Il maggior risultato di questo lavoro di tesi è la raggiunta consapevolezza che il quinto principio stabilito da *OECD* per l'accettazione dei modelli *QSAR* va enunciato in senso più restrittivo: solo se si è in grado di porre un'ipotesi di meccanismo alla base del modello *QSAR*, si può definirne correttamente l'*AD* e ottenere stime affidabili per l'*endpoint.* Anche alla luce di quanto rilevato in letteratura [Cronin 2003, Dearden 2009] sui "trabocchetti" che lo sviluppo di modelli QSAR può presentare, si può affermare che, benché non sia sicuro che un modello QSAR possa gettare luce sul meccanismo della tossicità, tuttavia è certamente rischioso costruire un modello senza porre alla base un'ipotesi di meccanismo. In altre parole la conoscenza del *MOA* permette di aumentare enormemente la capacità di stima e l'affidabilità del modello, viceversa, la non conoscenza del *MOA* pone seri vincoli alla sua affidabilità.

Concludendo, negli ultimi anni si è sviluppato un dibattito molto acceso sull'accettabilità delle tecniche *QSAR*. Il lettore interessato potrà trovare su questo argomento molti articoli in letteratura [Mc Kinney *et al.* 2000; Schultz *et al*. 2000; Cronin *et al.* 2003 a) Cronin *et al*. 2003 b) Cronin *et al*. 2003 c); Eriksson *et al.* 2003; Jaworska *et al*. 2003; Maggiora 2006; Vedani *et al.* 2006; Gramatica 2007; Zvinavashe *et al*. 2008; Benigni *et al.* 2008]. Fra le questioni emerse due meritano senz'altro di essere sottolineate:

1. il dominio di applicabilità di una QSAR è limitato, prima di intraprendere una ricerca occorre stabilire se i propri obbiettivi vi rientrino, per non correre il rischio di cercare una soluzione a problemi impossibili;
2. le *QSAR* sono per sé stesse materia multidisciplinare, per avere successo occorre ottenere la collaborazione di persone con competenze diverse.

A proposito della multidisciplinarietà delle QSAR sembra opportuno concludere questa tesi con una citazione tratta da Hansch [Hansch *et al.* 2003].

> *"The most unfortunate aspect of the theoretical approach to understanding how chemicals effect living organisms (or their parts from DNA to protein, enzymes, etc. and organisms from bacteria to man), is that one needs all of the mechanistic chemistry and biology that can be stuffed into one's head, plus the expertise to do meaningful modeling with the 'proper' system. This is something that cannot be attained in a few years. This problem has led to compartmentalization in QSAR studies. One needs synthetic chemists to make the chemicals, biologists to test them and computational experts to understand the chemical−biological interactions. Often there is not close cooperation between these three groups. There are three major types of interactions that the modeler must deal with: hydrophobic, electronic and steric. Of course, the chemist making compounds must incorporate into his set of congeners a good range in these properties, otherwise their*

importance or lack of it cannot be established. The buzzword these days in drug research is ADME (absorption, distribution, metabolism and elimination). Here the experience of the biologists becomes essential and needs to be organized. QSAR can do it."

Corwin Hansch 2003

Bibliografia

Abraham M.H., Sànchez Moreno R., Cometto Muñiz J.E; (2007) *Chem. Senses* **32**:711.

Agin D., Hersh L., Holtzman D.; (1965) *PNAS* **53**:952.

Aldenberg T.; (2004) "*Review of methods for assessing the applicability domains of SARs and QSARs*" pubblicazione per conto di "*The European Commission Joint Research Centre*" *ECVAM*.

Allred A.L., Rochow G.; (1958) *J. Inorg. Nucl. Chem.* **5**:264.

Alms G. R., Burham A.K., Flygare W.H.; (1975) *J. Chem. Phys.* **63**:3321.

Aptula A.O., Jeliazkova N.G., Schultz T.W., Cronin M.T.D.; (2005) *QSAR Comb. Sci.* **24**:385.

Atkins P.W. (1983) *Molecular Quantum Mechanics* 2nd Edition chapter 13 Oxford U. Press (UK).

Bajorath J.; Peltason L.; Wawer M.; Guha R.; Lajines M.S.; Van Drie J.H.; (2009) *Drug Discovery Today* **14**:698.

Bandiera S., Sawyer T., Romkes M., Zmuzdka B., Safe L., Mason G., Keys B., Safe S.; (1984) *Toxicology* **32**:131.

Bauernschmitt R., Ahlrichs R.; (1996) *Chem. Phys. Lett.* **256**:454.

Basak S.C., Mills D.; (2001) *SAR QSAR Environ. Res.* **12**:481.

Becke A.D.; (1993) *J.Chem.Phys.* **98**:5648.

Belous A. R., Hachey D.L., Dawling S., Roodi N., Parl F.F.; (2007) *Cancer Res.* **67**:812.

Benigni R., Bossa C.; (2008) *J. Chem. Inf. Model.* **48**:971.

Blaauboer B.J., Barrat M.D., Houston J.B.; (1999) *ATLA* **27**:229.

Bonati L., Fraschini E., Lasagni M., Palma Modoni E., Pitea D.; (1995) *J. Mol Struct. THEOCHEM* **340**:83.

Bonati L., Fraschini E., Lasagni M., Pitea D.; (1994) J. Mol. Struct. *THEOCHEM* **303**:43.

Bonchev D.; (2001) *J. Mol. Graphics Modell.* **20**: 65.

Born M., Oppenheimer R.; (1927) *Ann. Phys.* **84**:457.

Borth D.M.; (1996) *Chemom. Intell. Lab. Sys.* **32**:25.

Borth D.M., Wilhelm M.S.; (2002) *Chemom. Intell. Lab. Sys.* **63**:117.

Boettcher C.J.F.; (1973) *Theory of electric polarization* 2nd Ed. Vol 1. Elsevier Amsterdam.

Boettcher C.J.F., Bordewijk P.; (1973) *Theory of electric polarization* 2nd Ed. Vol 2. Elsevier Amsterdam.

Bug T., Hartnagel M., Schlierf C., Mayr H.; (2003) *Chem. Eur. J.***9**:4068.

Callaway J., March N.H.; (1984) *Solid State Phys.* **38**:135.

Carey F.A., Sundberg R.J.; (2007) "*Advanced Organic Chemistry:part B Reaction and Synthesis*". 5th Edition Springer (NY).

Casida M.E.; (1995) "*Recent Advances in Density Functional Methods*" Ed. D.P. Chong World Scientific Singapore.

CDC, "*Registry of Toxic Effects of Chemical Substances*" (*RTECS*) *http://www.cdc.gov/niosh*.

Champagne B., Perpète E.A., Van Gisbergen S.J.A., Baerends S.J., Snijders J.G., Soubra Ghaoui C., Robins K.A., Kirtman B.; (1998) *J. Chem. Phys.* **109**:10489.

ChemIDplus Advanced banca dati a cura della *U.S. National Library of Medicine http://chem.sis.nlm.nih.gov/chemidplus/*

Cedillo A., Contreras R., Galván M., Aizman A., Andrés J., Safont V.S.; (2007) *J. Phys. Chem. A* **112**:2442.

Ceperly D.M., Alder B.J.; (1980) *Can. J. Phys.* **58**:1200.

Chattaraj P.K., Sarkar U., Roy D.R.; (2006) *Chem. Rev.* **106**:2065.

Chattaraj P.K.; (2009) "*Chemical Reactivity Theory: a Density Functional View*". CRC Press.

Chermette H. ; (1999) *J. Comput. Chem.* **20**:129.

Chenggang G., Jang X., Xuehai J., Guifen Y., Yongrong B.; (2007) *Chemosphere* **67**:1325.

Chrissanthopoulos A., Hohm U., Wachsmuth U.; (2000) *J. Mol. Struct.* **526**:323.

Ciosloswki J., Martinov M., Mixon S.T.; (1993) *J. Phys. Chem.* **97**:10948.

Cronin M.T.D., Schultz T.W.; (2003) *J. Mol. Struct. THEOCHEM* **622**:39.

Cronin M.T.D., Jaworska J.S., Walker J.D., Comber M.H.I., Watts C.D., Worth A.P.; (2003) *Env. Health Persp.* **111**: 1376. 1391.

Cronin M.T.D., Jaworska J.S., Walker J.D., Comber M.H.I., Watts C.D., Worth A.P.; (2003) *Env. Health Persp.* **111**:1391.

Cuadras C.M., Cuadras D., Greenacre M.J.; (2006) *CmmStB* **35**:447.

Dahl J.P., Avery J.; (1984) *"Local Density Approximationin Quantum Chemistry and Solid State Physics"* Plenum *NY*.

Dao T. L., Sunderland H.; (1959) *J. Natl. Cancer Inst.* **23**:567.

Davidov Y., Rozen R., Smulski D.R., Van Dyk T.K., Vollmer A.C., Elsemore D.A., La Rossa R.A., Belkin S.; (2000) *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **466**:97.

Davidson E.R.; (1975) *J. Comput. Phys.* **17**:87.

Dearden J.C.; (2006) *Expert Opin. Invest. Drugs* **1**: 31.

Dearden J.C., Cronin M.T.D., Kaiser K.L.E.; (2009) *SAR QSAR Environ. Res.* **3-4**: 241.

Deneer, J.W., W. Seinen, J.L.M. Hermens; (1988) *Aquat. Toxicol.* **12**:185.

Denison M.S., Vella L.M., Okey A.B.; (1986) *J. Biol. Chem.* **261**:3987.

Denison M.S., Fisher J.M., Whitlock J. P. *Jr.*; (1988) *PNAS* **85**:2528.

Denison M.S., Fisher J.M., Whitlock J. P. *Jr.*; (1988) *J. Biol. Chem.* **263**:17221.

Denison M.S., Nagy S.R.; (2003) *Rev. Pharmacol. Toxicol.* **43**:309.

De Jong S., Ter Braak C.J.F.; (1994) *J. Chemometrics* **8**:169.

De Jongh J., Forsby A., Houston J.B., Beckman M., Combes R., Blaauboer B.J.; (1999) *Toxicol. in Vitro* **13**:549.

Devillers J., Doré J.C.; (2002) *SAR QSAR Environ. Res.* **13**:409.

Devillers J., Bintein S., Domine D., Karcher W.; (1995) *SAR e QSAR Environ. Res.* **4**:29.

Devries A.H., Van Duijnen P.T., Juffer A.H., Rullmann J.A.C., Dijkman J.P., Merenga H., Thole B.T.; (1997) *J. Comput. Chem.* **86**:49.

Dimitrov S., Dimitrova G., Pavlov T., Dimitrova N., Patlewicz G., Niemala J., Mekenyan O.; (2005) *J. Chem. Inform. Comput. Sci.* **45**:839.

Donati Y.R.A., Slosman D.O., Polla P.S.; (1990) *Biochem. Pharmacol.* **40**:2571.

Draper, N.R., Smith H.; (1991) *Applied Regression Analysis,* John Wiley and Sons, *NY*, (*USA*).

ECB (2005). *Scoping study on the development of a technical guidance document on information requirements on intrinsic properties of substances*. Rapporto preparato dal *CEFIC*, *DK-EPA*, *Environmental Agency of Wales and England*, *ECETOC*, *INERIS*, *KemI* e *TNO*. *European Chemicals Bureau, Joint Research Centre*, *European Commission*, Ispra, Italia. *http://ecb.jrc.it*.

Dreizler R.M., Gross E.K.U.; (1990) *Density Functional Theory: An Approach to the Quantum Many-Body Problem*. Springer Verlag Berlin.

ECHA (2008) *Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals . www.echa.eu*.

Enslein K.; (1988) *Toxicol. Ind. Health* **4**:479.

Eisenbrand G., Pool-Zobel B., Baker V., Balls M., Blaauboer B.J., Boobis A., Carere A., Kevekordes S., Lhuguenot J.C., Pieters R., Kleiner J.; (2002) *Food Chem. Toxicol.* **40**:193.

Eriksson L., Johansson E., Kettaneh-Wold N.; (2001) "*Multi and Megavariate data Analysis-Principle and A*pplications".Umea, Sweden Umetrics *AB*.

Eriksson L., Johansson E., Müller M., Wold S.; (2000) *J. Chemom.* **14**:599.

Eriksson L., Jaworska J.S., Worth A.P., Cronin M.T.D., McDowell R.M., Gramatica P.; (2003) *Env. Health Perspect.* **111**:1361.

Everitt, B.S.; (2002) *Cambridge Dictionary of Statistics* (2nd Edition).

Faber N.K.M.; (1999) *Chemom. Intell. Lab. Syst.* **49**:79.

Faber N.M., Song X. H., Hopke P.K.; (2003) *Trends Anal. Chem.* **22**:330.

Falandysz J., Puzyn T., Szymanowska B., Kawano M., Markuszewski M., Kaliszan P., Skurski P., Blazejowski J., Wakimoto T; (2001) *Polish Journal of Environmental Studies* **4**:217.

Feher M., Ewing T.; (2009) *QSAR Comb. Sci.* **28**:850.

Ferguson J.; (1939) *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **127**:387.

Fermi E.; (1928) *Z. Physik* **48**:73

Flaveny C., Perdew G.H., Miller C.A.; (2009) *Toxicol Lett.* **189**:57.

Fock V.; (1930) *Z. Physik* **61**:126.

Fock V.; (1930) *Z. Physik* **62**:795.

Forina M., Casolino C., Almansa E.M.; *Chemom. Intell. Lab. Sys.* **68**:29.

Fraschini E., Bonati L., Pitea D.; (1996) *J. Phys. Chem.* **100**:10564.

Fukui K.; (1982) *Science* **218**:747.

Furusjö E., Svenson A., Rahmberg M., Andersson M.; (2006) *Chemosphere* **63**:99.

Gallegos Saliner A., Gironés X.; (2005) *J. Mol. Struct. THEOCHEM* **727**: 97.

Gedeck P., Rohde B., Bartels C.; (2006) *J. Chem. Inf. Model.* **46**:1924.

Geerlings P., De Proft F., Langenaeker W.; (2003) *Chem. Rev.* **103**:1793.

Gilbert N.; (2009) *Nature* **460**:1065.

Golbraikh A., Tropsha A.; (2002) *J. Mol. Graphics Modell.* **20**:269.

Golbraikh A., Tropsha A.; (2002) *J. Comput. Aided Mol. Des.* **16**:357.

Golbraikh A., Shen M., Xiao Z., Xiao Y.D., Lee K.H., Tropsha A. (2003) *J. Comput. Aided Mol. Des.* **17**: 241.

Gordy W.; (1946) *Phys. Rev.* **69**:604.

Gordy W.; (1951) *J. Chem. Phys.* **19**:702.

Gorelsky S. I., Lever A.B.P.; (2001) *J. Organomet. Chem.* **635**:187.

Gorelsky, S. I.; in "*Comprehensive Coordination Chemistry-II*" McCleverty, J. A., Meyer T.J.; (2003) Elsevier Science 590.

Gorelsky S.I; *AOMix: Program for Molecular Orbital Analysis. http://www.sg-chem.net/*, University of Ottawa, 2009

Gramatica P.; (2004) "*Evaluation of different statistical approaches for the validation of quantitative structure-activity relationships*". Rapporto finale per il *JRC Contract ECVA- European Chemicals Bureau*, *Joint Research Centre*, *European Commission*, Ispra, Italia *http://ecb.jrc.it*.

Gramatica P., Pilutti P., Papa E; (2004) *J.Chem Inf. Comput. Sci.* **44**:1794.

Gramatica P., Papa E; (2005) *QSAR & Combinatorial Science* **24**:953.

Gramatica P.; (2007) *QSAR Comb. Sci.* **26**:694.

Greene L.A., Tischler A.S.; (1976) *PNAS* **73**:2424.

Gross E.K.U., Kohn W.; (1990) *Adv. Quantum Chem.* **21**:255.

Gross E.K.U., Ullrich C.A., Grossmann U.J.; (1995) "*Density Functional Theory of Time Dependent System*" volume **337** di *NATO ASI Ser. B* Plenum NY.

Gruening M., Gritsenko O.V., Baerends E.J.; (2002) *J. Chem. Phys.* **116**:6435.

Guha R., Van Drie J.H; (2008) *J. Chem. Inf. Model* **48**:646.

Hall L.H., Maynard E.L., Kier B.; (1989) *Environ. Toxicol. Chem.* **8**:431.

Halle W.; (2003) *ATLA* **31**:89.

Hansch C., Maloney P. P., Fujita T., Muir R. M.; (1962) *Nature* **194**:178.

Hansch C., Fujita T.; (1964) *J.Am. Chem. Soc.* **86**:1616.

Hansch C., Leo A., Hoeckman D.; (1995) "*Exploring QSAR:Hydrofobic, Electronic and Steric Constants*" Ed. American Chemical Society *USA*.

Hansch C., Steinmetz W.E., Leo A.J., Mekapati S.B., Kurup A., Hoekman D.; (2003) *J. Chem. Inf. Comput. Sci.* **43**:120.

Harper P.A., Riddick D.S., Okey A.B.; (2006) *Biochem. Pharmacol.* **72**:267.

Hartree D.R.; (1928) *Proc. Cambridge Phyl. Soc.* **24**:89.

Hartree D.R.; (1928) *Proc. Cambridge Phyl. Soc.* **24**:111.

Hartree D.R.; (1928) *Proc. Cambridge Phyl. Soc.* **24**:426.

Hartung T., Bremer S., Casati S., Coecke S., Corvi R., Fortaner S., Gribaldo L., Halder M., Hoffmann S., Janusch Roi A., Prieto P., Sabbioni E., Scott L., Worth A., Zuang V.; (2004) *ATLA* **32**: 467.

Hartung T.; (2009) Nature **460**:208.

Hartung T., Rovida C.; (2009) *Nature* **460**:1080.

Hemmateennejad B., Javadnia K., Elyasi M.; (2007) *Anal. Chim. Acta* **592**:72.

Hinchliffe A.; Mkadmh A., Nikolaidi B., Soscùn H.J., Abu-Awwad F.M.; (2006) CEJC **4**: 743.

Hirokawa S., Imasaka T., Imasaka T.; (2005) *Chem. Res. Toxicol.* **18**:232.

Hissin P., Hilf, R.; (1976) *Anal. Biochem.* **74**:214.

Hohenberg P., Kohn W.; (1964) *Phys. Rev.* **136**:864.

Horvath D. , Marcou G., Varnek A.; (2009) *J.Chem. Inf. Model.* **49**: 1762.

Huggins C., Grand L. C., Brillantes F.P.; (1961) *Nature* **189**:204.

Iacus S., Masarotto G.; (2003) *Laboratorio di statistica con R*. McGraw-Hill, Milano.

Iczowski R.P., Margrave J.L.; (1961) *J.Am.Chem. Soc.* **83**:3547.

Ildiko E.F.; (1995) *Chemom. Intell. Lab. System.* **27**:1.

Ingold C.K.; (1933) *J. Chem Soc*. 1120.

Ingold C.K.; (1934) *Chem. Rev.* **15**:225.

IUBMB Enzyme Nomenclature. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)

IUPAC: IUPAC International Chemical Identifier (InChI™) *http://www.iupac.org/inchi/*.

Jaramillo P., Pérez P., Contreras R., Tiznado W., Fuentealba P.; (2006) *J. Phys. Chem. A* **110**:8181.

Jaworska J.S., Comber M., Auer C., Van Leeuwen C.J.; (2003) *Env. Health Perspect.* **111**: 1358.

Jaworska J.S., Nikolova-Jeliazkova N & Aldenberg T (2005). *ATLA* **33**:445.

Jensen L., Schmidt O.H., Mikkelsen K.V., Astrand P.O.; (2000) *J. Phys. Chem.* **104**:10462.

Jensen L., Astrand P.O., Osted A., Kongsted J., Mikkelsen K.V.; (2002) *J. Chem. Phys.* **116**:4001.

Jouan-Rimbaud D., Bouveresse E., Massart D.L., De Noord O.E.; (1999) *Anal. Chim. Acta* **388**:283.

Kaiser K.L.E.; (2003) *J. Mol. Struct. THEOCHEM* **622**: 85.

Kamlet M.J., Abboud J.L.M., Abraham M.H., Taft R.W.; (1983) *J. Org. Chem.* **48**:2877.

Karna S.P., Dupuis M.; (1991) *J. Comput. Chem.* **12**:487.

Karplus I.M., Porter R.H.; (1971) *Atoms and molecules* Ed. W.A. Benjamin Menlo Park (CA).

Katritzky A.R. , Petrukhin R., Tatham D., Basak S., Benfenati E., Karelson M., Maran U.; (2001) *J. Chem . Inf. Comput . Sci.* **41**:679.

Kohn W., Sham L.; (1965) *J. Phys. Rev.* **140**:1133.

Kohn W., Becke A.D., Parr R.G.; (1996) *J. Phys. Chem.* **100**:12974.

Kubinyi H.; (2002) *J. Brazil Chem. Soc.* **13**:717.

Kutter E., Hansch C.; (1969) *Arch. Biochem. Biophys.* **135**:126.

Landau (1969)

Leach A. R.; (2001) *Molecular Modelling Principles and Applications* 2nd Edition Pearson Education Limited Edinburgh Gate Harlow England.

Lebedew P.; (1891) *Ann. Phys.* **280**:288.

Lee C., Yang W., Parr R.G.; (1988) *Phys. Rev. B* **37** :785.

Legler, J. Van den Brink, C.E. Brouwer, A., Murk A.J., Van der Saag P.T., Vethaak A.D., and Van der Burg B.; (1999) *Toxicol. Sciences*, **48**:55.

Legon A.C., Millen D.J.; (1987) *J. Am. Chem. Soc.* **109**:356.

Legon A.C. (1999) *Angew. Che. Int. Ed.* **38**:2686.

Lemek T., Mayr H.; (2003) *J. Org. Chem.* **68**:6880.

Leonard J.T., Roy K.; (2006) *QSAR Comb. Sci.* **25**:235.

Van Leeuwen D.A., Van Ruitenbeek J.M., De jongh L.J.; (1994) Phys. Rev. Lett. **73**:1432.

Levine I.N.; *Quantum Chemistry* 5th edition Prentice Hall cap. 15 -16.

Lieb E. H., Simon B.; (1973) *Phys. Rev. Lett.* **31**:681.

Lin I.C., Von Lilienfeld O.A., Coutinho-Neto M.D., Tavernelli I., Rothlisberger U.; (2007) *J. Phys. Chem.* **111**:14346.

Lipnick R. L.; (1991) *Sci. Tot. Environ.* **109/110**: 131.

Livingstone D.; (1995) *Data analysis for Chemists* Oxford University Press.

Lorentz H.A.; (1880) *Ann. Phys.* **9**:641.

Lorenz L.; (1880) *Ann. Phys.* **11**:70-103.

Lorentz H.A. (1906) *Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern*. Teubner Verlag Leipzig (D).

Lowry T.H., Richardson K.S.; (1987) "Mechanism and Theory in Organic Chemistry". Harper& Row NY (USA) terza edizione.

Lucius R., Loos R., Mayr H., (2002) *Angew. Chem. Int. Ed.* **41**:91.

Manchester J.; Czermiński R.; J.Chemi. Inf. Model. (2009) **49**:1449.

Maggiora G.M.; (2006) J. Chem. Inf. Model. **46**:1535.

Mahalanobis P.C.; (1936). *Proceedings of the National Institute of Sciences of India* **2**:49.

Makedonas, C.; Mitsopoulou, C. A.; (2006) *Eur. J. Inorg. Chem.* **3**:590.

Martens H. A., Dardenne P.; (1998) *Chemom. Intell. Lab. Syst.* **44**:99.

Mason G., Sawyer T., Keys B., Bandiera S., Romkes M., Piskorska-Pliszczynska J, Zmuzdka B., Safe S. (1985) *Toxicology* **37**:1.

Maynard, A. T., Huang M.; Rice W. G., Covel, D. G.; (1998) *PNAS* **95**:11578.

Mayr H., Patz M.; (1994) *Angew. Chem. Int. Ed.* **33**:938.

Mayr H., Bug T., Gotta M.F.; Hering N., Irrgang B., Janker B., Kempf B., Loos R., Ofial A.R., Remenikov G., Schimmel H.J.; (2001) *J.Am. Chem. Soc.* **123**:9500.

Mayr H., Lang G., Ofial A.R.; (2002) *J. Am. Chem. Soc.* **124**:4076.

Mc Kinney J.D.; (1985) *Environ. Health Perspect.* **61**:5.

Mc Kinney J.D. Fawkes J., Jordan S., Chae K., Oatley S., Coleman R.E., Briner W.; (1985) *Environ. Health Perspect.* **61**:41.

Mc Weeny R.; (1989)*"Methoids of Molecular Quantum Mechanics"*. Academic Press NY.

Meerts I.A.T.M., Letcher R.J., Hoving S; (2001) *Environ. Health Perspect.* **109**:399.

Mekenyan O.G., Veith G.D., Call D.j., Ankley G.T.; (1996) *Env. Health Perspect.* **104**:1302.

Mevik B.H., Wherens R.; (2007) *J.Stat. Soft.* **18**:1.

Meyer, T. J.; (2004) *AOMix Software Manual* Elsevier Amsterdam, Vol. **2**, pag. 651.

Millefiori S. , Alparone A; (1998) *J. Mol. Struct. THEOCHEM* **422**:179.

Miller K.J.; (1990) *J. Am. Chem. Soc.* **112**:8543.

Miller T.M.; (2004) in *CRC Handbook of Chemistry and Physics*, 84th Edition.

Millikan R.; (1897) *Ann. Phys. Chem.* **296**:376.

Mills E.J.; (1884) *Phil. Mag. Ser.* **5**:173.

Minegishi S., Loos R., Kobayashi S., Mayr H.; (2005) *J. Am. Chem. Soc.* **127**:2641.

Møller C., Plesset M.S.; (1934) *Phys. Rev.* **46**:618.

Moro G., Bonati L., Bruschi M., Cosentino U., De Gioia L., Fantucci P. C., Pandini A., Papaleo E., Pitea D., Saracino G.A.A., Zampella G.; (2007) *Theor. Chem. Acc.* **117**:723.

Mulliken R.S. (1934) *J. Chem. Phys.* **2**:782.

Nebert D.W., Bausserman L.L.; (1970) *Mol. Pharmacol.* **6**:304.

Nalewajski R.F., Capitani J.F.; (1982) *J. Chem. Phys.* **77**:2514.

Natarajan R., Basak S.C., Harriss D.K., Magnuson V.R.; (2007) *J. Chem. Inf. Model.***47**:771.

Nenadic O., Greenacre M; (2007) *J. Stat. Soft.* **20**:1.

Netzeva T.I., Worth A.P., Aldenberg T., Benigni R., Cronin M.T.D., Gramatica P., Jaworska J.S., Kahn S., Klopman G., Marchant C.A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G.Y., Perkins R., Roberts D.W., Schultz T.W., Stanton D.T., Van de Sandt J.J.M., Tong W., Veith G., Yang C.; (2005) ATLA **33**:155.

Nikolova N, Jaworska J (2003) *QSAR & Combinatorial Science* **22**:1006.

Oeberg T.; (2004) Chem Res. Toxicol. **17**:1630.

O'Brien, P.J.; (1991) *Chem.-Biol. Interact.* **80**:1.

OECD (2004). *The Report from the Expert Group on (Quantitative) Structure Activity Relationship ([Q]SARs) on the Principles for the Validation of (Q)SARs*. OECD Series on Testing and Assessment No. 49. ENV/JM/MONO(2004)24. Organisation for Economic Cooperation and Development, Paris, France. 206 *http://www.oecd.org*

OECD (2005) *Guidance Document on the Validation and International Acceptance of New or Updated Test*. *http://www.oecd.org*

OECD (2005) *Methods for Hazard Assessment*. *http://www.oecd.org*

OECD (2006). *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative)Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals*. Organisation for Economic Cooperation and Development. Paris, France. http://www.oecd.org.

OECD (2007). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR]Models*. *OECD Series on Testing and Assessment* No. 69. ENV/JM/MONO(2007)2. Organisation for Economic Cooperation and Development, http://www.oecd.org.

Okey A.B., Bondy G.P., Mason M.E., Kahl G.F., Eisen H.J., Guenthner T.M., Nebert D.W.; (1979) *J. Biol. Chem.* **254**:11636.

Okey A. B.; (2007) *Toxicol. Sci.* **98**:5.

Olsen J. , Jensen H.J.A., Jørgensen P.;(1988) *J. Comput. Phys.* **74**:265.

Osinga V.P., Van Gisbergen S.J.A., Snijders J.G., Baerends E.J.; (1997) *J. Chem. Phys.* **106**:5091.

Otaka M., Odashima M., Watanabe S.; (2006) *Biochem. Biophys. Res. Commun.* **348**:1.

Oughstun K.E., Cartwright N.A.; (2003) *Optics Express* **11**:1542.

Pandini A., Denison M. S., Soshilov A.A., Song Y., Bonati L.; (2007) *Biochemistry* **46**:696.

Papa E., Villa F., Gramatica P.; J. Chem. Inf. Model. (2005) **45**:1256.

Parr R.G., Donnelly R.A., Levy M., Palke W. E.; (1978) *J. Chem. Phys.* **68**:3801.

Parr R.G., Pearson R.; (1983) *J. Am. Chem. Soc.* **105**:7512.

Parr R.G., Yang W.; (1984) *J. Am. Chem Soc.* **106**:4049.

Parr R.G., Yang W.; (1989) "*Density Funtional Theory of Atoms and Molecules*" Oxford University Press NY.

Parr R. G., Szentpály L.V., Liu S.; (1999) *J. Am. Chem. Soc.* **121**:1922.

Pauli W.J.; (1926) *Z. Physik* **31**:765.

Pauling L.; (1932) *J. Am. Chem. Soc.* **54**:3570.

Pauling L., Pressman D.; (1945) *J. Am. Chem. Soc.* **67**:1003.

Pauling L.; (1960) *The nature of the chemical bond* Cornell University Press. 3rd Edition.

Pavan, M., Netzeva T.I., Worth A.P.; (2006) *SAR QSAR Environ. Res.* **17**:147.

Pavan, M., Worth A., Netzeva T.I.; (2005) *Preliminary Analysis of an Aquatic Toxicity Dataset and Assessment of QSAR Models for Narcosis,* JRC Report EUR 21479 EN, European Commission, Joint Research Centre, Ispra, Italy.

Pearson R.G.; (1963) *J. Am. Chem. Soc.* **85**:3533.

Pearson R.G.; (1997) "*Chemical hardness applications from molecules to solids*". Wiley VCH Weinheim.

Pearson R.G.; (2005) *J. Chem. Sci.* **117**:369.

Perdew J.P., Parr R.G., Levy M., Balduz J.L. Jr. (1982) *Phys. Rev. Lett.* **49**:1691.

Perdew J.P., Burke K., Ernzerhof M.; (1996) *Phys. Rev. Lett.* **77**:3865.

Perdew J.P., Burke K., Ernzerhof M.; **(1998)** *Phys. Rev. Lett.* **80**:891.

Pérez P., Aizman A., Contreras R., (2002) *J. Phys Chem. A* **106**:3964.

Pérez P.; *J. Org. Chem.* (2003) **68**:5886.

Petersilka M., Grossmann U.J., Gross E.K.U.; (1996) *Phys. Rev. Lett.* **76**:1212.

Poland A., Glover E., Kende A. S.; (1976) *J. Biol. Chem.* **251**:4936.

QSAR WORLD sito *web http://www.qsarworld.com/ Strand Life Sciences http://www.strandls.com*

Raevsky O.A., Grigor'ev V.Y., Weber E.E., Dearden J.C.; (2008) *QSAR Comb. Sci.* **27**:127.

Rajarshi G., Van Drie J. H.; (2008) *J. Chem. Inf. Model.* **48**:646.

Randić O.A., Basak S.C.; (2001) *J. Comp. Inf. Sci.* **41**:614.

Regolamento CE 1907/2006 pubblicato su Gazzetta Ufficiale della Comunità Europea il 30.12.2006 e poi corretto il 29.05.2007.

Richnon A.B., Young S.S.; *An introduction to QSAR methodology* Network Science http://www.netsci.org/Science/Compchem/feature19.html.

Roth M., Mayr H.; (1995) *Angew. Chem. Int. Ed.* **34**:2250.

Roy D.R; Sarkar U.; Chattaraj P.K.; (2006) *Molecular Diversity* **10**:119.

Roy P.P., Leonard J.T., Roy K.; (2008) *Chemom. Intell. Lab. Sys.* **90**:31.

Runge E., Gross E.K.U.; (1984) *Phys. Rev. Lett.* **52**:997.

Rydbergh H., Dion M., Jacobson N., Schroeder E., Hyldgaard P., Simak S.I., Langreth D.C., Lundqvist B.I.; (2003) *Phys. Lett.* **91**:126402.

Russom C.L., Bradbury S.P., Broderius S.J., Hammermeister D.E., Drummond R.A.; (1997) *Environ. Toxicol. Chem.* **16**:948.

Sanderson R.T.; (1955) *J. Chem. Phys.* **23**:2467.

Sarkar U., Roy D.R., Chattaraj P.K., Parthasarathi R., Padmanabahn J., Subramanian V.; (2005) *J. Chem. Sci.* **117**: 599.

Schervish M.J. (1996) *The American Statistician* **50**:203.

Schipper P.R.T., Gritsenko O.V., Van Gisbergen S.J.A., Baerends S.J.; (2000) *J. Chem. Phys.* **112**:1344.

Schneider G., Wrede P.; (1998) *Progress in Biophysics & Molecular Biology* **70**:175

Schroedinger E.;(1926) *Ann. Phys.* **79**:361.

Schroedinger E.;(1926) *Ann. Phys.* **79**:489.

Schroedinger E.;(1926) *Ann. Phys.* **79**:734.

Schroedinger E.;(1926) *Ann. Phys.* **80**:437.

Schroedinger E.;(1926) *Ann. Phys.* **79**:109.

Schultz, T.W., Yarbrough J. W., Pilkington B.T.; (2002) *Chem. Res. Toxicol.* **15**:1602.

Schultz T.W., Carlson R. E.; Cronin M. T. D; Hermens J. L. M., Johnson R.; O'Brien P. J.; Roberts D. W.; Siraki A.; Wallace K. B.; Veith G. D. (2006) *SAR QSAR Environ. Res.* **17**:413.

Schultz T.W., Yarbrough J.W., Koss S.K.; (2006) *Cell Biol. Toxicol.* **22**: 339.

Scripps Institute (USA) indirizzo web: *http://www.scripps.edu/e_index.html*

Selassie C. D., Rajni G., Kapur S., Kurup A., Verma R. P., Mekapati S.B., Hansch C.; (2002) *Chem. Rev.* **102**:2585.

Sen K.D., Jorgenson C.K.; (1987) "*Structure and bonding. Vol. 66:electronegativity*" Springer Verlag Berlin.

Sen K.D., Mingos D.M.P.; (1993) "*Structure and bonding. Vol. 80:hardness*" Springer Verlag Berlin

Sheridan R.P.; (2000) *J. Chem. Inf. Comput. Sci.* **40:**1456.

Shlens J.; A *tutorial on Principal Component Analysis* (2005) disponibile in rete shlens@salk.edu

Shuvaeva O.V.; (2007) *Russ. J. Phys. Chem. A* **81**:798.

Siraki A.G., Chan T.S., O'Brien P.J. (2004) *Toxicol. Sci.* **81**:148.

Sito web del gruppo di ricerca: Milano Chemometrics and QSAR Research Group *http://www.disat.unimib.it/chm/default.htm ; http://www.moleculardescriptors.eu*

Sito web del laboratorio di modeling molecolare della Facoltà di Farmacia dell'Università di Padova *http://mms.dsfarm.unipd.it*

Sito web dello European Chemicals Bureau *http://ecb.jrc.it*

Slater, J.C. (1972) *Adv. Quantum Chemistry* **6**:1.

SMILES™ programma della Daylight Chemical Information Systems Inc. http://www.daylight.com.

Smith M.B., March J.; (2006) "*March's Advanced Organic Chemistry: Reactions, Mechanisms, and structure*". *Wiley Interscience* sesta edizione.

So S.S., Karplus M.; (1997) *J. Med. Chem.* **40**:4360.

Stanforth R.W., Kolossov E., Mirkin B.; (2007) *QSAR Comb. Sci.* **26**:837.

Stephens P.J., Devlin F.J., Chabalowski C.F., Frisch M.J.; (1994) *J.Phys.Chem.* **98:**11623.

Sterne J.A.C., Smith G.D. (2001) *BMJ* **322**:226.

Stieb J.A.; (2009) *Journal of Business Ethics* **87**:401.

Swart M.; Van Duijnen P.Th.; Snijders J.G.; (2001) *J. Comput. Chem.* **22**:79.

Sylvester-Hvid K.O., Astrand P.O., Ratner M.A., Mikkelsen K.V.; (1999) *J. Phys. Chem A.* **103**:1818.

Taft R.W. (1956) *Separation of Polar, Steric and Resonance Effects in Reactivity* in M.S. Newman *Steric Effects in Organic Chemistry*, Ed. Wiley NY pp. 556-675.

Teller E.;(1962) *Rev. Mod. Phys.* **34**:627.

Tetko I.V., Sushko I., Pandey A.K., Zhu H., Tropsha A., Papa E., Öberg T., Todeschini R., Fourches D., Varnek A.; (2008) *J. Chem. Inf. Model.* **48**:1733.

Todeschini R., Consonni V.; (2003) *Descriptors from molecular geometry* in: J. Gasteiger (ed.), Handbook of Chemoinformatics, vol.**3** . Wiley-VCH, Weinheim (GER) 1004-1033.

Todeschini R., Consonni V., Mauri A., Pavan M.; (2004) *Anal. Chim. Acta* **515**:199.

Todeschini R.; Consonni V.; (2009). *Molecular descriptors for chemoinformatics*. Wiley & Sons Ltd. NY.

Tokuyasu T., Mayr H.; (2004) *Eur. J. Org. Chem.* 2791.

Topliss J.G., Costello R.J.; (1972) *J. Med. Chem.* **15**:1066.

Thomas L.H., (1927) *Proc. Cambridge Phyl. Soc.* **23**:542.

Thorburn W.M.; (1915) *Mind* **24**:287.

Thorburn W.M.; (1918) Mind **27**:345.

Tropsha A., Gramatica P., Gombar V.K.; (2003) *QSAR Comb. Sci.* **22:**69

Tute M.S.; (1971) *Principles and practice of Hansch analysis: a guide to structure activity correlation for the medicinal chemist* in *Advances in drug research* di Harper N.J, Simmonds A.B.; Academic Press London.

Van Duijnen P.T., Swart M.; (1998) *J. Phys. Chem A* **102**:2399.

Van Gisbergen S.J.A., Snijders J.G., Baerends E.J.; (1995) *J. Chem. Phys.* **103**:9347.

Van Gisbergen S. J. A., Osinga V.P., Gritsenko O.V., Van Leeuwen R., Snijders J.G., Baerends E.J.; (1996) *J. Chem. Phys.* **105**:3142.

Van Gisbergen S. J. A., Kootstra F., Schipper J.G., Gritsenko O.V., Snijders J.G., Baerends E.J.; (1998) *Phys. Rev. A* **A57**:2556.

Van Gisbergen S.J.A., Snijders J.G., Baerends E.J.; (1998) *J. Chem. Phys.* **109**:10644.

Van Gisbergen S.J.A., Snijders J.G., Baerends E.J.; (1999) *Comp. Phys. Commun.* **118:**119. (S.J.A. van Gisbergen, Ph.D. thesis Amsterdam (1998))

Vedani A., Dobler M., Lill M.A.; (2006) *Basic Clin. Pharmacol. Toxicol.* **99**:195.

Verhaar, H.J.M., Mulder W., Hermens J.L.M.; (1995) "QSARs for Ecotoxicity", in J.L.M. Hermens(ed.), *Overview of Structure-Activity Relationships for Environmental Endpoints. Part 1: General Outline and Procedure*, Report Prepared within the Framework of the Project "QSAR for Prediction of Fate and Effects of Chemicals in the Environment", Contract with the European Commission EV5V-CT92-0211.

Verma R.P.; (2005) *Bioorg. Med. Chem.* **13**:237.

Verma R.P., Hansch C.; (2005) *Bioorg. Med. Chem.* **13**:4597.

Villenueve D. L., Kannan K., Khim J.S., Falandysz J., Nikiforov V.A., Blankenship A.L., Giesy J.P. ; (2000) *Arch. Environ. Contam. Toxicol.* **39**:273.

Vogel A.I.; (1948) *J. Chem. Soc.* **XXIII**:1833.

Vogel W.T., Cresswell W.T, Jeffery G.J. Leicester J.; (1950) *Chem. Ind.* 358.

Vosko S.H., Wilk L., Nusair M.; (1980) *Can. J. Phys.* **58**:1200.

Vracko M., Bandelj V., Barbieri P., Benfenati E., Chaudhry Q., Cronin M.T.D., Devillers J., Gallegos A., Gini G., Gramatica P., Helma C., Mazzatorta P., Neagu D., Netzeva T., Pavan M., Patlewicz G., Randic M., Tsakovska I., Worth A.; (2006). SAR QSAR Environ. Res. **17**:265.

Wahl A. C., Das. G.; (1977) *Methods of Electronic Structure Theory*. H.F. Schaefer, Plenum , New York.

Waller C.L, McKinney J.D. (1995) *Chem. Res. Toxicol.***8**:847.

Walum E., Balls M., Bianchi V., Blaauboer B.J., Bolefsoldi G., Guillouzo A., Moore G.A., Odland L., Reinhard C., Spielmann H.; (1992) *ATLA* **20**:406.

Walsh A.D.; (1951) *Proc. Roy. Soc.* (London) **A207**:13.

Wendt B., Cramer R.D.; (2008) *J. Comput. Aided Mol. Des.* **22**:541.

Williams R.D., Koonin S.E. 1983 *Phys Rev C* **27:1817.**

Wilson, A. P.; (2000) *Cytotoxicity and Viability Assays in Animal Cell Culture: A Practical Approach*, 3rd ed. (ed. Masters, J. R. W.) Oxford University Press, Vol. 1.

Wilson E.B. (1965) non pubblicato.

Wold S.; (1995) *Chemom. Intel. Lab. Sys.* **30**:109.

Wold S.; (2001) *Chemom. Intel. Lab. Sys.* **58**:83.

Wold S., SjöStröm M, Eriksson L.; (2001) *Chemom. Intel. Lab. Sys.* **58**:109.

Worth A., Bassan A., Gallegos A., Netzeva T.I., Patlewicz G., Pavan M., Tsakovska I., Vracko M.; (2005). *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*. European Commission report EUR 21866 EN. European Commission-Joint Research Centre, Ispra, Italy. Available from ECB website: *http//ecb.jrc.it/QSAR*.

Worth A. et al. (2007) P*ossible applications of QSAR methods. In: A Compendium of Case Studies that helped to shapethe REACH Guidance on Chemical Categories and Read Across*. Worth A & Patlewicz G (Eds). EUR report no 22481EN. Available from ECB website: *http://ecb.jrc.it/qsar/*

Worth A.P. & Patlewicz G., Eds (2007). *A Compendium of Case Studies that helped to shape the REACH Guidance on Chemical Categories and Read Across. European Commission* report EUR22481 EN. European Commission-Joint Research Centre, Ispra, Italy. Available from ECB *http://ecb.jrc.it*.

Yan X.F., Yao H.M., Gong X.D., Ju X.H.; (2006) *J. Mol. Struct. THEOCHEM* **764**:141.

Yang W., Mortier W.J.; (1986) *J. Am. Chem. Soc.* **108**: 5708.

Young D., Marin T., Venkatapathy R., Harten P.; (2008) *QSAR Comb. Sci.* **27**:1337.

Zangwill A., Soven P.; (1980) *Phys. Rev. Lett.* **45**:204.

Zangwill A., Soven P.; (1981) *Phys. Rev. B* **24**:4121.

Zhang Y., Yang W.; (1998) *Phys. Rev. Lett.* **80**:890.

Ziegler T. Rauk A.; 1977 *Theor. Chim. Acta* **46**:1.

Ziegler T.; (1991) *Chem. Rev.* **91**:651.

Ziegler T., Tschinke V., Becke A. (1989) *J.Am. Chem. Soc.* **109:**1351

Ziegler T., Tschinke V., Baerends E.J., Snijders J.G., Ravenek W.; (1989) *J. Phys. Chem.* **93**:3050.

Zupan J., Gasteiger J.; (1999) *Neural Networks in chemistry and drug design* seconda ed. Wiley.

Zvinavashe E., Murk A.J., Rietjens I.M.C.M.; (2008) *Chem. Res. Toxicol.* **21**:229.