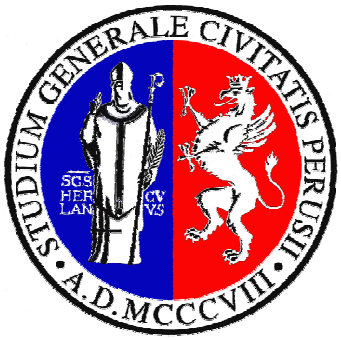


UNIVERSITÀ DEGLI STUDI DI PERUGIA

Facoltà di Scienze Matematiche Fisiche e Naturali

Corso di Laurea Specialistica in Scienze Chimiche

Metodi Computazionali e Applicazioni di Ricerca Virtuale di Molecole Biologicamente Attive

Laureando:
Giuseppe Marco Randazzo

Relatore:
Prof. Gabriele Cruciani

Anno Accademico 2009/2010

Dedicata alla mia famiglia… Mamma, Papà, Salvatore

Grazie!

"La materia non ragiona!

Lo Spirito immortale trionfa sulla materia.

I materialisti proclamano la progettazione completa, meravigliosa, fantastica, stupenda, più esatta, precisa, perfetta e poi dicono la parola più idiota che uomo possa pronunciare: che è il caso!!! Nulla avviene per caso! Il caso non esiste! […]

C'è un'unica causa fuori dello spazio, perché spazio collegante non c'è, fuori dal tempo, perché tempo utilizzabile non c'è, fuori di quegli esseri perché è una causa che ha dato all'essere di esistere così come è. Quindi padrona dell'intrinsecità dell'essere e della sua essenza una causa fuori dello spazio, fuori del tempo, fuori della materia padrona dell'essere, io la chiamo e la proclamo Mio Dio: io ti adoro. Questa è la mia fede."

dai discorsi di *Enrico Medi*

Ringrazio Dio per …

Mamma Concetta e Papà Santo Francesco, due sguardi che si sono incontrati per donarmi questa preziosa e strepitosa vita…

mio Fratello Salvatore, senza di lui non sarei stato da nessuna parte. È per te: “A volte ti sarà sembrato che sono veramente lontano da te, o forse avrai sentito ingratitudine. Ti chiedo perdono :)”

tutta la mia famiglia, Ziu Nnano, Zia Maria, Zio Nello, Zia Graziella, Cummare Maria, Cumpare Nnanuzzu , Nonnu Pippinu, Nonna Niluzza, Nonna Giovanna, Nonnu Ture, Giuseppe, Gianluca, Giuseppe, Salvo, Antonella.

i ragazzi di Monteripido per avermi sopportato in questi due anni… ma anche per l’affetto e l’amore che mi hanno donato… ma un particolare grazie ad Ezio che come dice lui “mi ha vestito” :D

la GiFra di Monteripido perché mi hanno fatto conoscere la vera verità, l’amore verso i fratelli. Ogni momento trascorso con voi mi arricchisce, mi sconvolge perché è inverosimile :)

I Frati: Fra Alessandro Mantini perché lui mi ha messo dentro il seme della verità, Fra Paolo perché lui è stato ed è tutt’ora colui che cura la crescita di questo fiore, Fra Bernardino che mi ha dissetato quand’ero arido, Fra Giulio per la fiducia che ripone su di me, Fra Luigi per il sole che mi dona, Frate Rino per l’affetto, Frate Francesco Treccia detto “Muchacho” perché lui sa come tirarti su!!

Il prof. Cruciani per l’opportunità di crescita che mi ha donato, per il suo supporto nei momenti difficili di questa tesi e per la fiducia che ha riposto in me

Emanuele perché ha riposto su di me tanta fiducia, ma soprattutto perché mi ha donato il suo tempo per la mia crescita a 360°

Daniela per la simpatia e l’affetto

Massimo per l’eccezionali e affascinanti lezioni che ci ha tenuto durante il suo corso, nei momenti di sconforto tra un malloc(), un free() e un valgrind, e per le piacevoli pause caffè trascorse :)

Paolo per la sua temperanza :), per la sua simpatia

Monica per la sua simpatia e genuinità

Laura per la sua intraprendenza ed audacia

Il Prof. Clementi per l'accoglienza che mi ha donato, ho sentito l'odore di casa mia…

Fabio per i suoi consigli e la stima che ha di me, e anche per l'eccezionale persona che è!! :)

Jean perché come me è del SUD:P, ma anche per la sua calma e temperanza.

Francesco detto da me "ciccio" ("perché 'checco' te l'ho spiegato com'è… :)") perché mi ha sopportato, per la sua straordinaria compagnia, non potevo desiderare meglio che un compagno di banco come Ciccio! E per la sua sincerità…

Lydia perché anche lei è del SUD :P , per la sua simpatia e la delicata presenza

e per finire… tutte le persone che ho incontrato nella mia vita perché ognuno mi ha donato un pezzettino di se per compormi così come sono, la mia vita come un puzzle dove ogni pezzo è legato all'altro per dare una forma autentica ed unica. Continuerò a comporlo, non so se è finito o infinito, ma poco conta perché rimarrà sempre unico e che non stanca!

Indice

1. Il Virtual Screening

1.1. Una Panoramica Generale

I prodotti farmaceutici svolgono un ruolo molto importante nella società tanto che numerose malattie sono curate in maniera non invasiva attraverso di essi. La scoperta di un nuovo farmaco è il risultato di una complessa attività di ricerca e sviluppo, talvolta molto onerosa sotto il profilo economico. Attualmente secondo i dati riportati[1], il prezzo di tale processo si avvicina intorno agli 800-1000 milioni di dollari per lo sviluppo di un farmaco. Pertanto, non sorprende che le case farmaceutiche siano in costante ricerca di nuovi metodi che superino i confini del laboratorio chimico tradizionale attraverso la realizzazione di strumenti nuovi e ad alto contenuto tecnologico che permettano di ottimizzare i costi di ricerca e produzione, non sottovalutando nello stesso tempo sicurezza ed efficacia in termini di attività biologica. Tra gli altri, un ruolo di primo piano è certamente svolto dalla chimica computazionale che ha il potenziale per permettere la contrazione dei tempi e la riduzione dei costi nonché uno sviluppo più razionale di nuovi e più selettivi composti bioattivi. In questo scenario, la moderna ricerca farmaceutica si pone due obiettivi chiave:

- identificazione di nuovi composti (*leads*) che mostrino attività contro specifici targets biologici.
- progressiva ottimizzazione dei composti individuati al fine di migliorarne affinità e proprietà ADMET (Assorbimento, Distribuzione, Metabolismo, Eliminazione, Tossicità).

A tale scopo le istituzioni, entrano in sinergia con le industrie, allo scopo di contribuire alla ricerca di nuovi farmaci, attraverso lo sviluppo di nuove soluzioni.

Le attuali metodiche o tecniche usate nella ricerca di nuovi composti a potenziale attività farmaceutica sono due:

- High-Throughput Screening (HTS): una tecnica di screening di un grande numero di composti contro un target biologico per l'identificazione di composti guida (lead compound) necessari per la ricerca di nuovi strutture;
- Virtual Screening: una metodica computazionale che permette di valutare in maniera rapida un grosso numero di database contenenti strutture molecolari allo scopo di identificare quelle strutture che si legano al target macromolecolari di interesse, un recettore o un enzima.

Generalmente la tecnica HTS permette di testare circa 100.000^{2} composti al giorno, un numero non indifferente che però comporta un costo non indifferente. Questa tecnica tuttavia, oltre alle controindicazioni di carattere economico, presenta delle problematiche in relazione ai tests effettuati, quali scarsa solubilità dei composti, composti che rimangono attaccati alle pareti dei pozzetti dove vengono condotte le reazioni, precipitazioni etc… tutti problemi che producono molto 'rumore' nei dati prodotti, generando numerosi falso positivi e falsi negativi, facendo definire questa tecnica come una tecnica veloce ma sporca.

Un' approccio alternativo è quello conosciuto come Virtual Screening[3-4], metodologia che tramite dei test "*virtuali*" su molecole presenti in un database contro un target biologico virtuale, permette di effettuare delle previsioni in termini di attività. Da una rapida analisi effettuata in letteratura, circa il 30% dei composti predetti *in-silico* risulta essere attivo verso il target biologico (Vedi Tabella 1):

Articolo	Numero Molecole Testate	Attive	% Attive
J. Med. Chem. **2010**	30	9	30 %
J. Med. Chem **2007**	32	3	9.37 %
J. Chem. Inf. Model. **2010**	24	10	41,6 %
J. Med. Chem. **2008**	44	18	40.90 %
J. Med. Chem. **2009**	50	5	10 %
Bioorg- Med. Chem. **2009**	6	3	50 %
J. Med. Chem. **2010**	37	8	21.62 %
J. Med. Chem. **2010**	106	33	31 %
ACS Chem. Biol. **2009**	525	40	7.61 %
J. Med. Chem. **2010**	33	17	51.5 %

Tabella 1 Ricerca bibliografica sui dati di virtual screening

Tale metodologia ha avuto e sta avendo un grande successo proprio per la semplicità di trovare o prevedere nuovi ligandi da acquistare e testare, senza ricorrere a numerosi tests sperimentali e in maniera più economica.

Tuttavia anche questa tecnica presenta dei limiti, quali la presenza di molecole predette come false positive e/o false negative, la definizione del dominio sperimentale spesso ristretto, la disponibilità di composti chimici non brevettati, la velocità del calcolo in relazione al problema che si ha a disposizione.

Infatti, uno dei punti più deboli del virtual screening è la velocità del calcolo in relazione al problema che si ha a disposizione e al tempo necessario per effettuare misure sperimentali in competizione con quelle *in-silico*. La tempistica di un calcolo generalmente è funzione della velocità di un computer e delle risorse hardware a disposizione ma è anche dipendente dalla tipologia del problema in studio. Anche se l'industria informatica è in continuo sviluppo, si sente la necessità di focalizzarci sul problema per ottimizzare tempi di calcolo e risultati. Con gli attuali metodi e risorse computazionali è stato evidenziato che un progetto di virtual screening applicato ad un database di 800.000 molecole

necessita di più di un anno di calcolo. Tale stima è stata fatta tenendo conto che per un calcolo completo su ogni molecola vengono impiegati circa 3 secondi.
Nonostante ciò il virtual screening offre un rapido percorso per la scoperta di composti nuovi (leads compound) per la ricerca farmaceutica, riscuotendo da sempre grande successo per il risparmio di risorse economiche.

1.2. Approcci di Virtual Screening

Nella metodologia del virtual screening, le librerie molecolari giocano un ruolo importante. Generalmente sono distinti tre tipi di librerie molecolari[5]:

- *focused library*, costituita da composti indirizzati ad agire su una classe di targets biologici (ad esempio serin-proteasi)

- *targeted library*, costituita da composti indirizzati ad agire su uno specifico target biologico (ad esempio la trombina)

- *general library*, costituita da varie classi di composti.

L'approccio di *virtual screening*, implica che siano disponibili informazioni riguardanti il sito di legame del recettore, o la biomolecola che svolge/induce l'attività biologica nota anche come ligando, o nella migliore ipotesi entrambi. Il *virtual screening* comprende un vasto numero di tecniche computazionali, dalle più semplici alle più sofisticate, e quindi può sfruttare differenti tipi di informazioni relative al recettore o al ligando.
Da un punto di vista pratico i fattori da valutare in un' analisi di *virtual screening* sono:

- i tempi di calcolo
- l'accuratezza del calcolo delle proprietà dei composti
- i software e gli hardware richiesti per il calcolo e l'archiviazione dei dati.

L'obiettivo di questo approccio è la selezione, a partire da un vasto set iniziale di composti, di un sottoinsieme rappresentativo che contenga possibili "hits", composti iniziale promettenti anche se con una bassa attività biologica. È importante che il campione di composti sottoposto all'analisi sia realmente rappresentativo in termini di proprietà chimico-fisiche e di diversità molecolare; questo è, infatti, un presupposto necessario, sebbene non sufficiente, per il campionamento di un ampio dominio di possibili risposte biologiche e quindi di

composti potenzialmente bioattivi. Il diagramma di flusso in Figura 1 mostra le varie fasi di analisi di *virtual screening* fino alla sintesi chimica in laboratorio e, quindi, ai saggi biologici. Lo schema circolare del diagramma di flusso dimostra come la fase di sintesi chimica e quella di ottimizzazione mediante tecniche computazionali siano fortemente integrate di modo che non si possa in alcun modo stabilire priorità dell'una o dell'altra. Una prima selezione dei composti può essere effettuata sulla base di caratteristiche bidimensionali; le molecole prescelte verranno quindi convertite in strutture tridimensionali energeticamente stabili per essere sottoposte ad un'ulteriore fase di valutazione e quindi di scrematura.

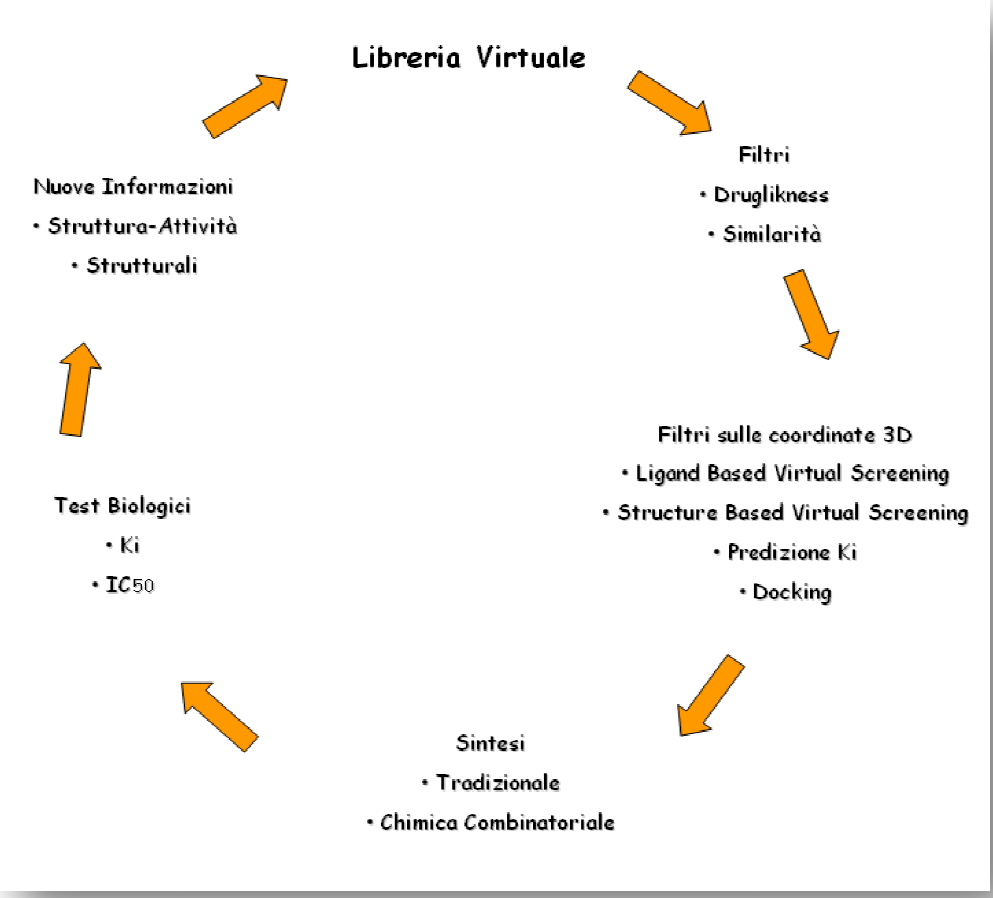


Figura 1 Diagramma di flusso di un'analisi di virtual screening

Essenzialmente si possono distinguere due grandi categorie di Virtual Screening:

- Ligand-Based
- Structure-Based

I due approcci si differenziano sulla base del punto di partenza da cui opera il virtual screening.

Nel caso dell'approccio Ligand-Based si necessita di informazioni strutturali riguardo i ligandi del target in oggetto di studio, mentre nel caso dell'approccio Structure-Based si necessita di informazioni riguardo il sito attivo del target.

Tali informazioni sono racchiuse nei "*descrittori molecolari*" che possono essere estratti attraverso procedure computazionali fornite in maniera appropriata da alcuni software.

Presso il laboratorio di chemiometria e chemioinformatica dell'Università di Perugia sono stati sviluppati dei software che permettono di calcolare vari descrittori molecolari quali GRID, VolSurf e FLAP che verranno discussi nel capitolo riguardante gli strumenti computazionali.

1.2.1. Il Virtual Screening Ligand Based

La metodologia Ligand-Based permette di predire nuovi composti come attivi facendo uso di un set di molecole strutturalmente differenti che si legano al recettore in studio. Tali molecole sono note come composti di riferimento e le nuove strutture predette aventi attività identica o molto simile possono essere progettate *in-silico* oppure ottenute da una collezione di molecole fisicamente disponibili che esibiscono una somiglianza con tali composti di riferimento. Successivamente questi composti predetti possono essere testati in vitro. La scelta di un'appropriata misura della similarità dipende dal target in studio e dal suo contesto. Per tale approccio sono disponibili due fonti complementari: un database di strutture fisicamente disponibili e una libreria di strutture "virtuali" che a sua volta include un grosso numero di librerie combinatoriali. Oggi esistono molte banche dati commerciali tra cui si ricordano:

- World Drug Index (WDI di Derwent Information London, UK) che contiene oltre 60000 composti farmaceutici provenienti da tutte le fasi di sviluppo
- MDDR (Drug Data Report di MDL Information System USA) che rappresenta una collezione di oltre 100000 strutture con rispettive attività nelle prime fasi dello sviluppo di un farmaco
- Comprehensive Medicinal Chemistry database (CMC di MDL Information System, USA) con oltre 75000 strutture farmaceutiche.
- Zinc Database (Zinc, a free database for virtual screening di Shoichet Laboratory, USA) con più di 13 milioni di strutture commercialmente disponibili.
- ChemDiv (ChemDiv, the chemistry of cures, USA) con più di 800.000 molecole commercialmente disponibili.
- Specs (Specs, chemistry solutions for drug discovery, USA) con oltre 200.000 molecole commercialmente disponibili.
- Sigma Aldrich(Life Science Technologies and Specialty Chemicals, USA) con più di 300.000 molecole.

Inoltre, diverse compagnie farmaceutiche offrono grosse librerie di composti, e generalmente tali collezioni si aggirano numericamente intorno ad un range di 700000 fino a qualche milione di molecole. Ciò è stato possibile grazie alle migliorie apportate alla sintesi combinatoriale che oggi attraverso sistemi automatici rende possibile produrre composti ad alta purezza.

Per attuare la metodologia Ligand-Based occorre conoscere informazioni riguardo l'attività di questi composti, informazioni chimico-fisiche (coefficienti di ripartizione Acqua/Ottanolo cLogP, peso molecolare, etc…), informazioni riguardanti la struttura molecolare in 3D. Nella maggior parte dei casi molte informazioni chimico-fisiche non possono essere estratte dai database, per cui si ricorre a dei modelli QSAR, modelli lineari di relazione quantitativi tra le proprietà molecolari e la loro struttura che permettono di effettuare delle predizioni di queste. Le informazioni 3D invece vengono estratte attraverso procedure computazionali, come ad esempio la procedura GRID**.** Successivamente queste informazioni vengono analizzate in maniera da effettuare una classificazione delle molecole e razionalizzare l'attività di questi composti.

1.2.2 Il Virtual Screening Structure-Based

Nel precedente capitolo è stato accennato l'approccio ligand-based che ha come obbiettivo l'identificazione di molecole con proprietà chimico-fisiche simili ai ligandi conosciuti che interagiscono con il target in studio. E' stato evidenziato quindi che la strategia ligand-based limita la diversità dei nuovi candidati farmaci alle proprietà dei ligandi conosciuti.

L'approccio Structure-Based[3-6-7-8] invece sfrutta il riconoscimento molecolare tra un ligando ed una proteina target selezionando le entità chimiche che si legano fortemente ed identificando i siti attivi di rilevanza biologica del target proteico stesso. La condizione necessaria per questo tipo di approccio è che deve esser nota la struttura 3D del target macromolecolare in studio. Se tale struttura non è nota si può ricorrere ad un "homology model" ovvero un modello per omologia noto anche "modello comparativo di una proteina". La condizione necessaria per poter fare un homology model è che deve essere nota la sequenza amminoacidica della proteina ed almeno una struttura 3D di una proteina omologa. L'algoritmo agisce in maniera da individuare una o più strutture di una proteina per allineamento della sua sequenza amminoacidica con la sequenza amminoacidica della struttura proteica omologa di cui è nota la conformazione 3D. E' stato visto che se il 20% della sequenza non riesce ad allinearsi con quella della struttura omologa, allora la struttura proteica avrà un'alta probabilità di differire dalla struttura proteica omologa[9]. Quindi la qualità del modello per omologia (Homology model) è dipendente dalla qualità della sequenza allineata e dalla struttura della proteina omologa. Dalla struttura 3D del recettore attraverso un'analisi con GRID è possibile estrarre i punti di interazione target-ligando più importanti e risalire ad un così detto "farmacoforo", ovvero un arrangiamento di proprietà molecolari o frammenti molecolari che formano una condizione necessaria, ma non sempre sufficiente, affinché ci sia un'attività biologica. Il farmacoforo ha una dimensione, forma e proprietà ben precise che sono responsabili della sua attività farmacologica. Attraverso questo farmacoforo

è possibile effettuare il virtual screening, ovvero proiettare molecole all'interno di esso per analizzarne le interazioni che si instaurano. Tutto questo viene operato attraverso diverse tipologie di algoritmi. Qui ne riportiamo alcuni:

- Docking : l'algoritmo di docking[10-11] si occupa della previsione della conformazione di un ligando assunta all'interno del sito attivo di un recettore. Tale algoritmo ha maturato in maniera significativa nel corso degli anni affrontando il limite di flessibilità del ligando e della proteina. Tuttavia rimangono irrisolte questioni come il cambio di conformazione della proteina indotto dalle condizioni esterne, cambiamenti di conformazione dovuti all'interazione con un ligando o alla partecipazione delle molecole d'acqua nelle interazioni ligando-proteina. Ogni ligando che viene analizzato da questo algoritmo avrà uno score[8-12] calcolato per via empirica o semiempirica, e tale score cercherà di stimare le complesse interazioni ligando-proteina. Molte di queste funzioni di score implementate in vari programmi di docking effettuano diverse ipotesi per poter determinare caratteristiche complesse, semplificando alcune caratteristiche fisiche come gli effetti idrofobici, entropici e di solvatazione al fine di determinare il riconoscimento molecolare. Proprio questa stima delle interazioni nell'algoritmo di docking è lo step critico ma allo stesso tempo di successo.

- MOE[13]: le caratteristiche farmacoforiche sono generalmente rappresentate da punti 3D nello spazio, ciascuno associato ad un raggio di tolleranza. Essenzialmente un farmacoforo può essere costituito da atomi, come l'azoto, gruppi di atomi come il carbonile, o molto frequentemente da gruppi funzionali[14]. I gruppi funzionali comunemente usati sono definiti in termini di Polarità-Carica-Hidrofobicità (PCH) includendo interazioni quali legame a idrogeno accettore (Acc), legame a idrogeno donatore (Don), specie cationica (Cat), specie anionica (Ani),

centro aromatico (Aro), e area idrofobica (Hyd). Inoltre possono essere specificate in maniera approssimata le direzioni dei legami a idrogeno, o la preferenza planare di un anello aromatico. Permette di costruire a livello interattivo il farmacoforo permettendo all'utente di personalizzare i raggi di tolleranza, la posizione delle sfere di tolleranza. Infine le molecole di un database vengono sovrapposte all'interno del farmacoforo andando a valutare se ci sono o meno interazioni favorevoli con le sfere di tolleranza, e poi applicare uno score in base al tipo di interazioni globali che si instaurano con la molecola.

- Catalyst[15-1617]: tale algoritmo è stato introdotto nel 1991, ed è stato il primo a tentare il riconoscimento di un farmacoforo. I due algoritmi popolari implementati in catalist sono HypoGen e HipHop. HypoGen prima utilizza un training set costituito da composti attivi per generare un insieme di possibili siti attivi farmacoforici, e dopo attraverso un training set di inattive rimuove tutti quei siti non favorevoli. HipHop invece identifica le proprietà comuni tra i vari composti senza considerarne l'attività. Attraverso una collezione di modelli conformazionali delle molecole e un set di gruppi funzionali, HipHop produce una serie di allineamenti delle molecole ed identifica una o più configurazioni in cui tutte le proprietà identificate si trovano in comune. Successivamente le varie ipotesi sono ordinate sulla base del numero di molecole che fittano il farmacoforo e sulla frequenza con cui questo si presenta.

1.3. Strumenti Computazionali

1.3.1. GRID

GRID è una procedura computazionale che permette di determinare i siti favorevoli di legame in molecole la cui struttura tridimensionale è nota[18]. Può essere usato nello studio individuale di molecole come un farmaco, un set di piccole molecole, o macromolecole come proteine , acidi nucleici, glicoproteine e polisaccaridi. Concettualmente GRID opera attraverso una sonda chimica anisotropica, chiamata "probe", che può essere uno specifico atomo, un gruppo di atomi in una molecola o un'intera molecola che interagisce in maniera "puntuale" con la molecola in studio in modo da evidenziare i punti di massima interazione energetica. Essendo anisotropici i probe differiscono l'uno dall'altro ed evidenziano interazioni energetiche differenti. Ad esempio un probe carbonio sp2 aromatico differisce dal probe carbonio sp2 di un gruppo carbonilico per il semplice motivo che il carbonio sp2 è influenzato da un'intorno chimico differente. Nel carbonio carbonilico si ha l'influenza dell'ossigeno che cambia le proprietà chimico-fisiche del sistema atomico come la polarizzabilità, la carica elettrostatica, etc… Il carbonio sp2 aromatico invece è influenzato da altri atomi di carbonio che modificano a loro volta sempre proprietà chimico-fisiche come polarizzabilità, carica elettrostatica, etc... Ciò lo si può evincere nella figura 2 che rappresenta i campi di interazione dei due probe differenti di carbonio applicati interagenti con la stessa molecola di aspirina:

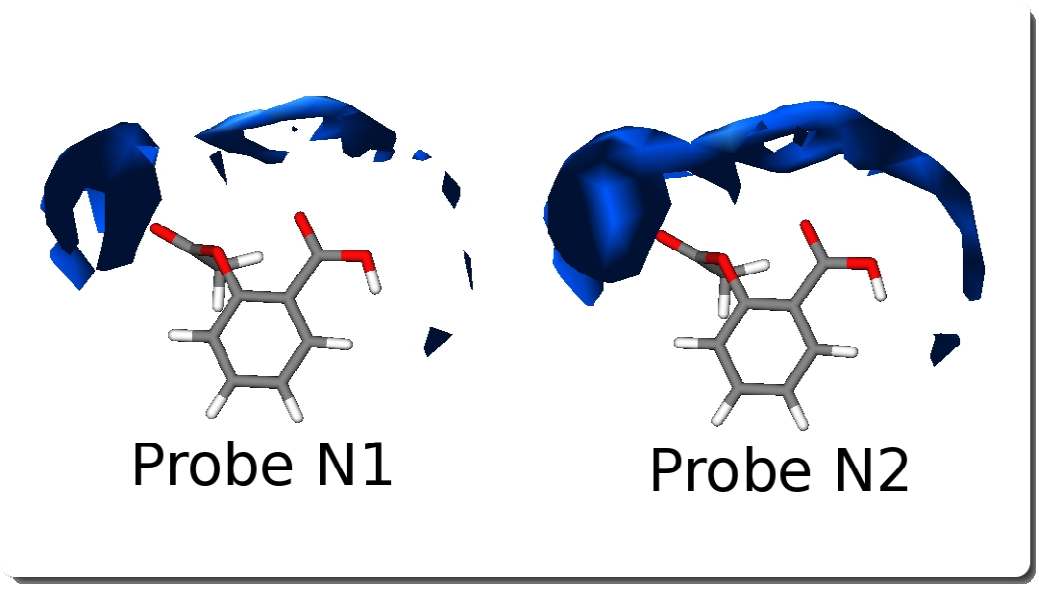


Figura 2 MIF dell'aspirina ottenuti con due probe differenti dell' azoto ma visualizzati alla stessa energia.

Inizialmente GRID costruisce attorno alla molecola in studio una gabbia parallelepipida, suddividendola in piccole griglie ed essendo stato ideato principalmente per lo studio sui sistemi biologici, viene assunto a priori che l'ambiente che circonda tale molecola sia l'acqua. Successivamente attraverso il proprio force-field empirico e con il probe calcola delle interazioni puntuali molecola-probe e da queste interazioni vengono estratti i "Campi di Interazione Molecolare" (Molecular Interaction Field – MIF). Nello specifico l'energia di interazione nel punto P(x,y,z) viene calcolata attraverso la seguente formula:

$$E = E_{LJ} + E_Q + E_{HB} + S$$

dove:

E_{LJ} è il termine energetico di Lennard-Jones calcolato come $(Ad^i – Bd^j)F$ in cui i = -12 , j = -6 e F=1; A e B sono delle costanti positive che vengono scelte e "d" è la distanza tra il probe in un punto della griglia e l'atomo del target in studio.

E_Q è il termine elettrostatico calcolato come "q_1q_2/dD" in cui "q_1" e "q_2" sono rispettivamente le cariche del probe e la carica dell'atomo della molecola in studio, "D" è la costante dielettrica e "d" è la distanza fra le due cariche.

E_{HB} è il termine energetico di legame a idrogeno in cui si usa soltanto quando uno degli atomi della molecola può accettare o donare legame a idrogeno e viene calcolato proprio come si calcola il termine energetico di Lennard-Jones ma con la differenza nelle costanti A e B e nella funzione F che dipende dall'ibridizazione e da un termini che descrivono la direzionalità dell'interazione.

Il termine S è il contributo entropico, che tiene conto della riorganizzazione delle molecole in cui è immerso il target in studio.

1.3.2 I Descrittori Molecolari

Per "Descrittori Molecolari" si intendono dei vettori risultato di una procedura logica e matematica, che trasformano informazioni chimiche codificate all'interno di una rappresentazione simbolica di una molecola. Il fine di questi descrittori quindi è dare un'interpretazione delle proprietà molecolari.
I Descrittori Molecolari si dividono in diverse classi:

- Descrittori molecolari parziali: si usano quando c'è un'alta similarità tra le molecole. Descrivono le molecole in maniera parziale perché quando tutto l'insieme delle molecole presenta lo stesso *scaffold* di base, questa informazione viene omessa nel descrittore molecolare risultante e vengono considerati soltanto i sostituenti presenti nello scaffold.

- Descrittori Chimico-Statistici: questa tipologia di descrittori vengono anche chiamati proprietà principali e queste proprietà vengono ottenute tramite applicazione dei metodi statistici a descrittori molecolari classici.

- Descrittori molecolari globali di derivazione teorica o sperimentale: questi tipi di descrittori vengono usati per molecole differenti facendo uso di proprietà molecolari come massa molecolare, logP, solubilità, etc... in questo caso i parametri usati possono derivare da misure sperimentale o da calcoli teorici. I calcoli teorici ci daranno queste proprietà tramite lo studio degli effetti elettronici. Tale approccio permette di fare predizioni anche per molecole che non sono disponibili a livello sintetico e permettono di valutare l'eventuale necessità di sintetizzarle (GRID).

- Descrittori topologici: questi descrittori topologici sono riferiti agli atomi che costituiscono la molecola e quindi forniscono una descrizione più dettagliata della molecola (Es. quanti atomi donatori di legame ad

idrogeno vi sono nella molecola, quante lone-pairs vi sono, etc...). Questi descrittori topologici ci danno informazioni su come e in che modo siano legati fra di loro gli atomi quindi ci permette di relazionare l'attività di una molecola in base alle loro proprietà atomiche.

1.3.3 VolSurf+

Volsurf+ è una procedura computazionale che comprime le informazioni presenti nei MIFs ottenuti tramite GRID, per produrre descrittori numerici, ottimizzati per modelli AMDE e per virtual screening[19-20-21]. Il software può essere utilizzato per la messa a punto di modelli statistici volti alla descrizione delle caratteristiche farmacocinetiche ovvero dello studio quantitativo dell'assorbimento, distribuzione, metabolismo e eliminazione dei farmaci, e delle caratteristiche farmacodinamiche cioè lo studio degli effetti biochimici e fisiologici dei farmaci sull'organismo, ed il loro meccanismo d'azione. Infatti i descrittori prodotti da tale procedura rappresentano tutte quelle proprietà che rilevano il profilo ADME di un farmaco.

Questo strumento computazionale ha quindi come obbiettivo la predizione delle proprietà ADME al fine di selezionare o escludere possibili farmaci candidati nella fase preliminare dello studio.

VolSurf+ offre inoltre la possibilità di utilizzare strumenti chemio metrici come PCA e PLS per analizzare i descrittori molecolari e le relazioni tra proprietà/attività e la struttura delle molecole.

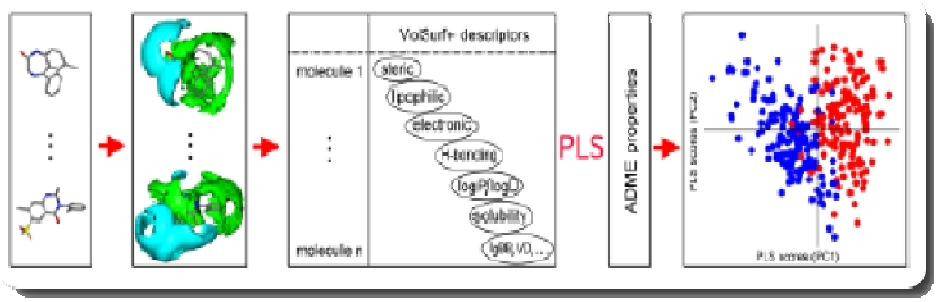


Figura 3 Le caratteristiche di VolSurf+

1.3.3.1 I descrittori di VolSurf

Il concetto di forma molecolare è di fondamentale importanza per poter valutare i vari descrittori molecolari. Dal momento che, nel senso stretto di quantomeccanica, le molecole non hanno una superficie ben definita, le superfici convenzionali delle molecole sono comunque diventate degli importanti descrittori nell'interpretazione delle proprietà e dei processi molecolari.
VolSurf+ produce diversi tipi di descrittori:

- Descrittori di dimensione e forma:
    - Volume molecolare: rappresenta il volume in $Å^3$ incluso nella superficie accessibile all'acqua calcolato da GRID al valore energetico di +0.20 kcal/mol
    - Superficie molecolare: rappresenta la superficie accessibile in $Å^2$ tracciata dal probe acqua al valore energetico di +0.20 kcal/mol
    - Rugosità (Rapporto volume/superficie): è una misura della rugosità molecolare, calcolata come rapporto tra volume e superficie molecolare. Più piccolo è il rapporto, più grande è la rugosità
    - Sfericità molecolare: è definita come S/S_{equiv} dove S_{equiv} è l'area di superficie di una sfera di volume V. La sfericità è uguale a 1.0 per le molecole perfettamente sferiche. Assume valori maggiori di 1.0 per le molecole realmente sferoidali. La sfericità è anche correlata con la flessibilità molecolare.

- Descrittori delle regioni idrofiliche
    - Descrittori idrofilici: le regioni idrofiliche sono definite come l'involucro molecolare che è accessibile e che attrae le molecole d'acqua. Il volume di questo involucro varia con il livello delle energie di interazione. In generale i descrittori idrofilici calcolati dai campi molecolari tra -1.0 e -0.2 kcal/mol spiegano la

polarizzabilità delle forze di dispersione, mentre i descrittori calcolati dai campi molecolari tra -2.0 e -6.0 kcal/mol spiegano la polarità e le regioni di legami a idrogeno donatore-accettore.

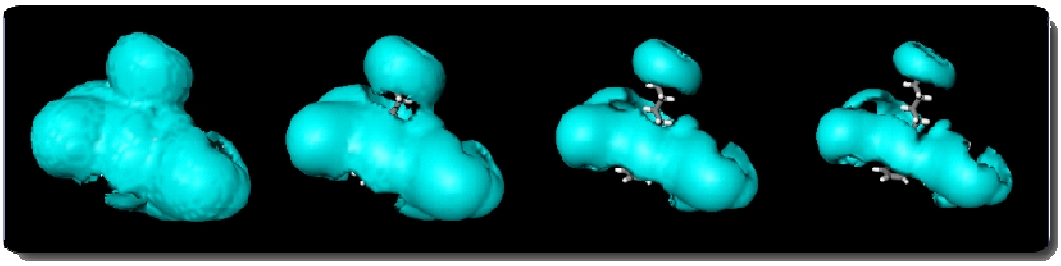


Figura 4 Campi idrofobici di una molecola calcolati a 4 differenti livelli di energia

- Fattori di capacità: rappresentano il rapporto della superficie idrofilia sulla superficie molecolare totale. In altre parole, rappresentano la superficie idrofilia per unità di superficie. I fattori di capacità sono calcolati ad otto diversi livelli energetici, gli stessi livelli usati per calcolare i descrittori idrofilici.

- Descrittori delle regioni idrofobiche: in GRID è usato un probe speciale chiamato DRY per generare un campo lipofilico 3D. In analogia con le regioni idrofiliche, le regioni idrofobiche possono essere definite come l'involucro molecolare che genera interazioni idrofobiche attrattive. VolSurf+ calcola i descrittori idrofobici ad otto diversi livelli energetici adattati all'abituale campo di energia di interazioni idrofobiche (per esempio da 0.0 a -2.0 kcal/mol)

- Descrittori dei momenti di energia di interazione (Integy moments): come i momenti dipolari, gli integy moments esprimono lo sbilanciamento tra il centro della massa di una molecola e il baricentro delle sue regioni idrofiliche. Gli integy moments quando sono riferiti alle regioni idrofiliche, sono dei vettori che puntano dal centro della massa al centro delle regioni idrofiliche. Quando l'integy moment è alto, c'è una chiara

concentrazione di regioni polari solamente in una parte della superficie molecolare. Se l'integy moment è basso, le parti polari sono entrambe vicine al centro della massa o si bilanciano alle estremità opposte della molecola e il loro baricentro risultante è vicino al centro della molecola. Quando si riferiscono alle regioni idrofobiche, gli integy moments misurano lo sbilanciamento tra il centro della massa di una molecola e il baricentro delle regioni idrofobiche.

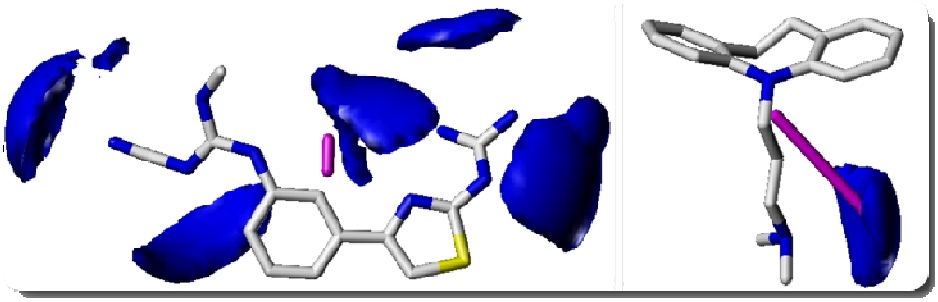


Figura 5 Esempio di differenti Integy moments

- Descrittori misti
    - Equilibrio idrofilico-lipofilico: è il rapporto tra le regioni idfofiliche misurate a -4 kcal/mol e le regioni idrofobiche misurate a -0.8 kcal/mol. L'equilibrio descrive quale effetto domina nella molecola, o se approssimativamente sono ugualmente bilanciati.
    - Momento anfifilico: è definito come un vettore che punta dal centro del dominio idrofobico al centro del dominio idrofilico. La lunghezza del vettore è proporzionale alla forza del momento anfifilico e può determinare l'abilità di un composto a penetrare una membrana
    - Critical packing parameter: definisce un rapporto tra la parte idrofilica e lipofilica di una molecola. Contrariamente all'equilibrio idrofilico-lipofilico, critical packing si riferisce solo alla forma della molecola. I calcoli lipofilici e idrofilici sono eseguiti rispettivamente a -0.6 e -3.0 kcal/mol. Critical packing è un buon

parametro per predire processi come il molecular packing responsabili della formazione di micelle, e può anche essere rilevante negli studi di solubilità dove il punto di fusione gioca un ruolo importante.

- Legami ad idrogeno: questi parametri descrivono la capacità di formare legami idrogeno di una molecola target, con un probe polare. Il probe acqua presenta un'abilità ottimale nel donare o accettare legami ad idrogeno con il target. Quando viene usato un diverso probe polare, l'interazione può essere più o meno favorevole e in ogni caso differisce a seconda della natura del probe usato. Il probe N1 per esempio è solo donatore di legame ad idrogeno, mente il probe O è solo un' accettore di legame ad idrogeno.
- Polarizzabilità: La polarizzabilità è determinata dalla media delle polarizzabilità molecolari, calcolate secondo un metodo basato sulla struttura dei composti (e non in ogni campo molecolare) ed è quindi indipendente dal numero e dal tipo di probes usati. La correlazione tra la polarizzabilità molecolare sperimentale e quella calcolata con VoilSurf+ è molto buona.

- Descrittori chimico-fisici
  - logP, logD, PSA: questi descrittori molecolari sono di semplice e comune comprensione come il logP, il logD e PSA. Il logP è il logaritmo in base 10 del coefficiente di partizione tra le fasi n-ottanolo e acqua a pH neutro; il logD è il logaritmo in base 10 del coefficiente di distribuzione tra le fasi n-ottanolo e acqua ad un dato pH, mentre la PSA è l'area della superficie polare, calcolata sulla base dei campi idrofilici usata molto spesso da programmi di virtual screening e quindi introdotta di recente. In aggiunta, i valori del logD sono riportati a 7 diversi valori di pH, da 5 a 10.

- Soly, sol-pH: sono descrittori molecolari di solubilità e sono ottenuti attraverso la predizione da un modello PLS sviluppato *in-house* applicando la procedura VolSurf+ a circa 1300 dati sperimentali derivati accuratamente dalla letteratura scientifica. Oltre alla solubilità intrinseca, misurata per composti neutri, sono stati derivati dei descrittori di solubilità apparente che permettono di ricostruire il profilo di solubilità di un composto (quando presenta centri ionizzabili) in funzione del pH.

- Descrittori modelli ADME
  - VD, CACO2, MET-STAB, BBB, SKIN: anche questi descrittori molecolari vengono ottenuti attraverso la predizione da modelli PLS *in-house* applicando la procedura VolSurf+ a diversi dati sperimentali derivati dalla letteratura scientifica. Questi modelli riguardano il volume di distribuzione (VD), il passaggio della membrana Caco-2 (CACO2), la stabilità metabolica rispetto all'enzima maggiormente presente nel fegato CYP3A4 (MET-STAB), il passaggio della barriera emato-encefalica (BBB) e la permeabilità della pelle (SKIN).

1.3.4 FLAP: Un'algoritmo di Virtual Screening Ligand-Based e Structure-Based

FLAP acronimo di "*Fingerprints for Ligands And Proteins*" è un software concepito come algoritmo per virtual screening, con lo scopo di descrivere ligandi e proteine in termini di 4 punti farmacoforici. Il concetto di "*farmacoforo*" è un concetto chiave, ed è definito come un arrangiamento di proprietà molecolari o frammenti molecolari che formano una condizione necessaria, ma non sempre sufficiente, affinché ci sia un'attività biologica. Attraverso i campi di interazione molecolare (MIFs) calcolati attraverso GRID è possibile associare tipi atomici specifici alle proprietà chimiche di un ligando: questi atomi selezionati vengono usati all'interno di FLAP per costruire tutte le possibili combinazioni di 4 punti farmacoforici della molecola in studio. Un'approccio similare può essere applicato nello studio delle proteine: FLAP può costruire i 4 punti farmacoforici presenti nel sito attivo della proteina utilizzando dei punti rappresentativi di esso. Tali punti rappresentativi sono calcolati dai MIFs e indicano le interazioni favorevoli tra un probe e la regione della proteina investigata. Tutti i potenziali 4 punti farmacoforici 3D provenienti dai ligandi o dai recettori sono calcolati tenendo conto della flessibilità conformazionale e della forma molecolare o del recettore e attraverso 4 punti farmacoforici può essere valutata la chiralità. Più nello specifico in una piccola molecola, un farmacoforo può essere definito da atomi in cui vi sono delle interazioni critiche con un recettore, mentre in una macromolecola un farmacoforo viene definito come una combinazione di tutti i punti rappresentativi localizzati nel sito attivo della macromolecola stessa. Essendo un algoritmo di recente sviluppo, esiste una sola pubblicazione relativa al funzionamento di FLAP[22], mentre nuove funzionalità vengono sviluppate continuamente. FLAP presenta diverse applicazioni: può essere usato come un tool di docking, per il virtual screening ligand-based (LBVS), per il virtual screening structure-based (SBVS), per calcolare descrittori da usare in un'analisi chemiometrica e per investigare sulla similarità delle proteine. Dal punto di vista

operativo FLAP necessita di una o più molecole da utilizzare come riferimento per effettuare il processo di virtual screening, ed un database di molecole da poter analizzare. Inizialmente FLAP costruisce un suo database. La molecola viene importata e minimizzata, e da essa vengono generati dei conformeri sui quali poi vengono calcolati i campi di GRID. Queste informazioni poi sono salvate all'interno di un file binario con estensione .fdb. Nello step successivo FLAP effettua il virtual screening ligand-based o structure-based utilizzando come informazioni soltanto quelle registrate all'interno del database prima creato, dunque informazioni indipendenti dal file di input originale. Scelta una molecola di riferimento o una serie di molecole su cui effettuare il virtual screening, FLAP sovrappone i 4 punti farmacoforici del database con quelli della/e molecola/e di rifermento. Questo processo è computazionalmente molto complesso e passa attraverso diversi stadi. Nel primo stadio viene importata la molecola di riferimento e viene sottoposta da un trattamento di minimizzazione della struttura e se si vuole anche generazione di varie conformazioni. Come risultato dell'analisi della molecola di riferimento viene creato un modello matematico che rappresenta il sistema di riferimento comune che avevamo menzionato precedentemente. Questo sistema di riferimento è registrato in una stringa di bit virtuale, per rendere più facili e veloci i confronti futuri. Per ogni ligando analizzato la stringa di bit generata da FLAP sarà $n \cdot 10^p$ dove "p" è il numero di probe utilizzati ed "n" è il numero di molecole di riferimento. Inizialmente tutti i bit sono uguali a 0, e ogni sovrapposizione tra la molecola del dataset e la molecola di riferimento "accende" un bit ponendolo uguale a 1. Partendo dalle coordinate degli atomi della molecola di riferimento o partendo dai punti estratti dai suoi MIFs, vengono definiti i punti farmacoforici della molecola. Nel modello matematico della molecola di riferimento sono definite tutte le possibili combinazioni di 4 punti farmacoforici. A questo punto ogni ligando del database precedentemente creato viene sovrapposto alla molecola di riferimento, in particolare vengono sovrapposti in maniera ottimale i 4 punti farmacoforici in maniera iterativa e scegliendo soltanto le migliori sovrapposizioni. Maggiore è il

numero di punti sovrapposti, maggiore è la similarità tra la molecola del database e la molecola di riferimento e tale similarità viene quantificata attraverso un punteggio (score). Va precisato che non si potrà mai avere una perfetta sovrapposizione ma, usando un accettabile grado di approssimazione nello spazio cartesiano, due punti si considerano sovrapposti se sono meno distanti di 1 Å l'uno dall'altro. In presenza di una lista di molecole di riferimento, FLAP può effettuare un processo di "Training" attraverso l'uso di un'attività segnata come 0 per molecole inattive e 1 per molecole attive. Dunque data una lista di 10 possibili molecole di riferimento, nella fase di training ogni molecola di queste viene utilizzata per sovrapporla alle molecole di cui si conosce l'attività fornendo un ranking differente. La molecola che meglio classifica le molecole per le quali è specificata l'attività, sarà scelta dal metodo come molecola di riferimento. Contemporaneamente alla scelta della molecola di riferimento, FLAP valuta il valore di distanza tra la molecola stessa di riferimento da usare come soglia, per distinguere le molecole attive dalle inattive. Sia per la scelta della molecola di riferimento che per la definizione del valore soglia di distanza, il programma utilizza un processo di ottimizzazione nel quale viene minimizzato l'errore percentuale, cioè la somma dei falsi positivi e dei falsi negativi, divisi per il numero totale di molecole nella lista. Per definire una molecola come vero positivo, vero negativo, falso positivo o falso negativo, si confronta l'attività conosciuta con l'attività predetta. Quest'ultima infine è ottenuta dal confronto degli scores con il valore di soglia citato. L'equazione per il calcolo del punteggio (score) è la seguente:

$$Score = Peso_{probe1} \cdot Distanza_{probe1} + Peso_{probe2} \cdot Distanza_{probe2} + \ldots +$$

FLAP varia sistematicamente tutti i pesi, rielaborando di volta in volta il valore di errore percentuale. Il processo utilizza l'algoritmo del simplesso ed infine è raggiunto il minor valore di errore percentuale possibile. In assenza di

ottimizzazione (ciò nel caso in cui è l'utente a fornire le molecole di riferimento), i pesi valgono tutti 1.0.

1.3. Lo scopo di questa tesi

Nel capitolo 1.1 è stato appurato che il virtual screening ha riscosso e continua a riscuotere grande successo per il risparmio di tempo e di risorse nel campo della scoperta dei nuovi composti a potenziale attività farmaceutica. Tuttavia non è una metodologia statica e consolidata, ma in continuo sviluppo ed ancora in fase di miglioramento.

Perciò, lo scopo di questa tesi è stato quello di sviluppare nuove metodologie di virtual screening focalizzate allo studio di algoritmi di filtraggio che tengano in conto il target (recettore) da raggiungere e che permettano di selezionare un set ridotto di molecole da analizzare in maniera da diminuire sensibilmente i tempi di calcolo e di aumentare l'efficienza del processo. Il nostro punto di partenza è stato l'uso di un sistema chiamato Global Positional System GPS che in analogia al sistema satellitare GPS localizza in maniera precisa una molecola in uno spazio chimico. L'idea è stata quella di effettuare analisi in questo spazio chimico attraverso metodi basati su metodi di clustering[23] e sulla Grid-based partitioning. Per fare questo sono stati sviluppati algoritmi e software *in-house* con linguaggi di programmazione C e Python. Nei capitolo successivo si parlerà dei vari metodi di virtual screening, e degli algoritmi che sono stati sviluppati, riportando esempi di applicazione.

2. Nuovi Metodi di Virtual Screening

Nel precedente capitolo è stata esposta la tecnica di virtual-screening analizzando i requisiti necessari (libreria di molecole, dataset di composti attivi, criteri di selezione) e gli utilizzi generali. In questo capitolo ci occuperemo di approfondire i criteri e metodi di virtual screening presentando due nuove metodologie.

2.1. I Criteri di Selezione molecolare

I Criteri di selezione molecolare, generalmente denominati filtri, sono dei criteri che permettono di discriminare a monte il numero di molecole da analizzare attraverso la tecnica di virtual screening. In pratica questi criteri agiscono scartando tutte quelle molecole che non presentano le caratteristiche ricercate. Esistono vari tipi di criteri di selezione e i più usati possono essere così suddivisi:

- Filtri sulle proprietà di "*druglikeness*": con il termine "*druglikeness*" si indica un ampio range di caratteristiche strutturali di un composto, come stabilità, solubilità e lipofilia, che influenzano le proprietà ADMET del composto stesso. Un semplice filtro basato sulle proprietà di "*druglikness*" ad esempio è dato dalla *Regola di Lipinski*[24], detta anche *Regola del 5*. La Regola di Lipinksi è una semplice regola dedotta empiricamente analizzando l'immensa mole di farmaci in commercio accumulatesi dopo decenni di ricerca, e questa risulta molto importante per la progettazione e lo sviluppo di un farmaco. Consente di delineare profili di biodisponibilità sulla base dei valori di peso molecolare, di lipofilia e di numero di donatori e accettori di legami ad idrogeno. Tale regola si basa su tre semplici fondamenti:

- le molecole non devono avere una massa molecolare maggiore di 500 (PM < 500).;

- il numero di siti donatori di legame a idrogeno non deve essere superiore a 5.

- il numero di accettori di legame a idrogeno (di solito atomi di ossigeno, azoto) e/o di donatori non deve essere superiore a 5;

- la molecola deve avere un log P minore di 5.

Se durante la fase di progettazione due di questi quattro punti non sono soddisfatti, molto probabilmente il composto presenterà problemi di assorbimento orale. Infatti se ad esempio le molecole fossero troppo voluminose e pesanti sarebbero difficilmente assimilabili ed incontrerebbero troppa difficoltà nel processo di diffusione. Troppi donatori di legame a idrogeno rendono la molecola eccessivamente polare, impedendone quindi la diffusione nelle parti lipofile.

- Filtri in presenza di un farmacoforo specifico per il target: la conoscenza della struttura tridimensionale del farmacoforo, quale unità strutturale di ligandi noti responsabile dell'attività biologica, può fornire un ulteriore criterio di selezione dei composti nell'analisi di *virtual screening.* È, dunque, necessario prendere in considerazione le proprietà steriche ed elettroniche dei gruppi funzionali facenti parte del farmacoforo, oltre che le caratteristiche strutturali espresse in termini di distanze fra gruppi accettori o donatori di legami ad idrogeno, centri lipofilici e gruppi aromatici. Tale filtro può essere applicato attraverso procedure structure-based come il docking, o la procedura FLAP.

- Filtri sulla similarità di uno o più ligandi al target: tali metodi si basano sul principio che composti simili hanno proprietà (attività) simile. Questi operano attraverso due fasi: il calcolo di descrittori molecolari, ossia proprietà chimico-fisiche misurabili direttamente o indirettamente che siano traducibili in attributi numerici, e la successiva stima del valore di similarità. I confronti di similarità generalmente eseguibili mediante due strategie:

    - *Clustering*[23]: un'insieme di tecniche di analisi multivariata dei dati volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati sfruttando il concetto di similarità. Tutte le tecniche di *clustering* si basano sul concetto di distanza tra due elementi. Infatti la bontà delle analisi ottenute dagli algoritmi di *clustering* dipende molto dalla scelta della metrica, e quindi da come è calcolata la distanza. Le tecniche di *clustering* si suddividono in due metodologie:

        - metodo gerarchico: viene costruita una gerarchia di partizioni caratterizzate da un numero (de)crescente di gruppi, visualizzabile mediante una rappresentazione ad albero (dendrogramma), in cui sono rappresentati i passi di accorpamento/divisione dei gruppi. Questo metodo si suddivide in due classi in funzione del tipo di algoritmo che usa:
            - *Agglomerativo*: questi algoritmi assumono che inizialmente ogni cluster (foglia) contenga un singolo punto; ad ogni passo, poi, vengono fusi i cluster più "vicini" fino ad ottenere un singolo grande cluster. Questi algoritmi necessitano di misure per valutare la similarità tra clusters, per

scegliere la coppia di cluster da fondere ad ogni passo.

- *Divisivo*: questi algoritmi partono considerando lo spazio organizzato in un singolo grande cluster contenente tutti i punti, e via via lo dividono in due. Ad ogni passo viene selezionato un cluster in base ad una misura, ed esso viene suddiviso in due cluster più piccoli. Normalmente viene fissato un numero minimo di punti sotto il quale il cluster non viene ulteriormente suddiviso (nel caso estremo questo valore è 1). Questi tipi di algoritmi necessitano di definire una funzione per scegliere il cluster da suddividere.

▪ metodo partitivo: *detto anche k-clustering*, il metodo partitivo definisce l'appartenenza ad un gruppo utilizzando una distanza da un punto rappresentativo del cluster (centroide, medioide ecc...), avendo prefissato il numero di gruppi della partizione risultato. Gli algoritmi di questa famiglia creano una partizione delle osservazioni minimizzando una certa funzione di costo.

- *Grid-based partitioning*[23]: l'approccio *grid-based partitioning* che consiste nel dividere lo spazio molecolare in celle distinte, mediante una griglia. Ad ogni molecola della libreria di partenza è assegnata una posizione nella suddetta griglia in accordo con le proprietà chimico-fisiche calcolate. Molecole simili fra loro si troveranno nella stessa cella, o in celle vicine. Può essere

selezionata una molecola per ogni cella o per ogni gruppo di celle (Figura 6).

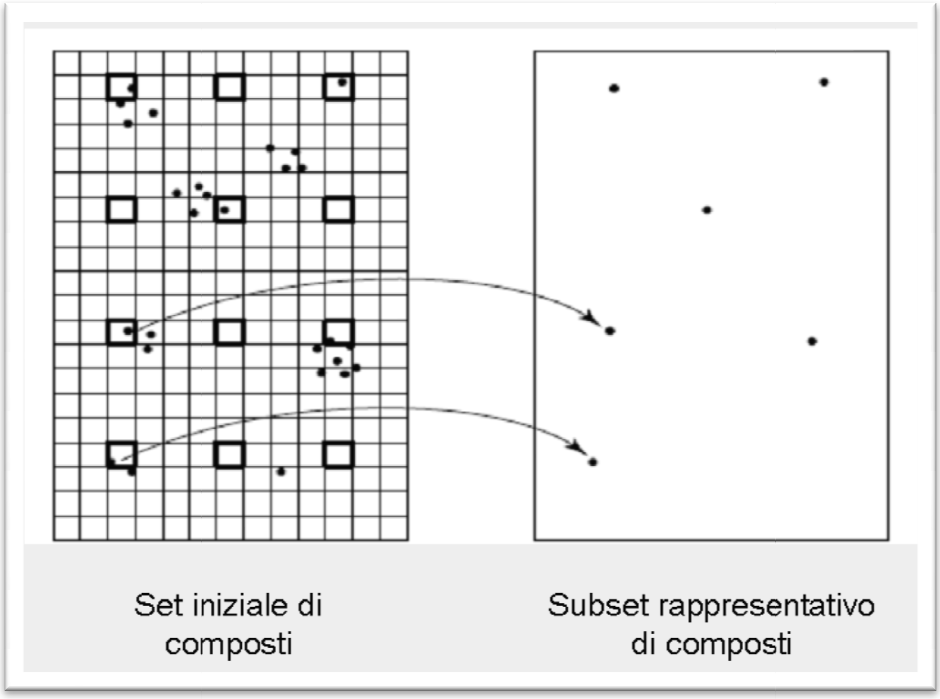


Figura 6 Schema generale di un generico metodo di selezione grid-based

2.2. Strumenti Utilizzati

2.2.1. I modelli globali: il GPS come sistema di navigazione sullo spazio chimico

T. Oprea[25] nel 2001 ha introdotto il concetto di navigare attraverso lo spazio chimico con l'obbiettivo di selezionare composti locati in particolari regioni dello spazio chimico, chiamato chem-GPS. In analogia con il sistema di posizionamento satellitare Navstar GPS[26], chem-GPS consiste in un'insieme di molecole (satelliti) che sono il sistema di riferimento standard (longitudine e latitudine) di un mondo di molecole costituito da una moltitudine di composti di interesse chimico-farmaceutico. A livello pratico questo sistema di riferimento proposto dal prof. Oprea utilizza gli scores delle proprietà principali (PC) come coordinate per la navigazione nello spazio chimico essendo il modello PCA il sistema di riferimento (cioè la mappa da utilizzare) per il posizionamento delle molecole, ossia la mappa che definisce lo "spazio chimico".

L'analisi delle componenti principali combina opportunamente la matrice derivante dai descrittori molecolari utilizzati: viene assegnato loro un peso relativo ai rispettivi contributi, e di conseguenza viene ridotta la dimensionalità del problema in poche variabili latenti dette componenti principali, ortogonali ed indipendenti tra loro, con l'obiettivo di estrarre gran parte dell'informazione.

In geografia il sistema di posizionamento Navstar GPS si basa sull'uso di 24 satelliti geostazionari che orbitano intorno alla terra. Tramite sistemi di triangolazione, si può risalire in pochi istanti alla posizione di un oggetto in qualsiasi punto del pianeta.

Analogamente, il sistema di riferimento chemGPS messo appunto da Oprea comprende un set di "strutture core" e un set di "strutture satellite". Il primo set contiene 202 strutture molecolari (core) rappresentative di farmaci in commercio che tengono il modello focalizzato e bilanciato nello spazio chimico

delle molecole drug-like ed anche nei confronti delle strutture satellite. Queste molecole rappresentative sono state scelte sulla base della loro permeabilità passiva intestinale (>10 %).

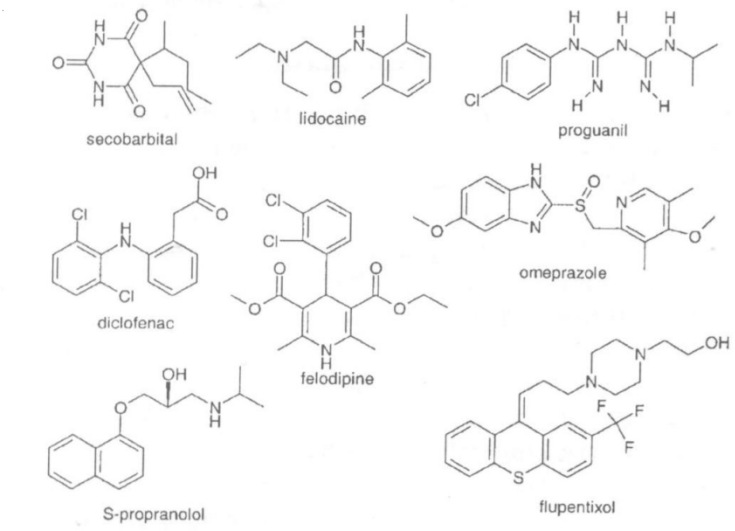


Figura 7 Alcune strutture core utilizzate nel modello GPS di Oprea

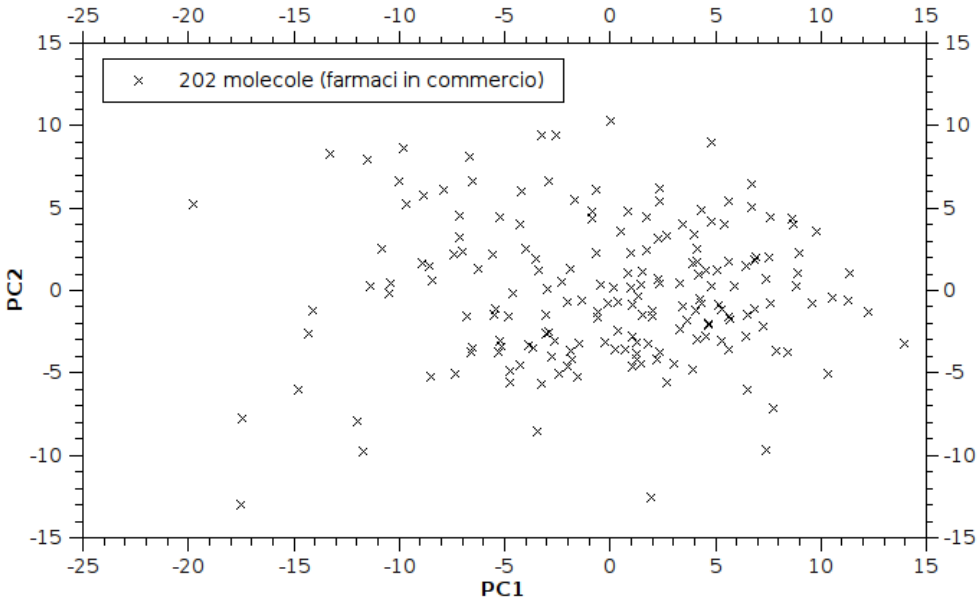


Figura 8 plot delle prime due componenti principali di VS+ delle strutture core utilizzate nel modello GPS di Oprea.

221 strutture satellite invece sono intenzionalmente piazzati al di fuori dello spazio, includendo molecole che possiedono valori estremi in almeno una delle PC (proprietà principali).

Graficamente tali strutture possono essere immaginate come ai confini dello spazio chimico; in modo da focalizzare l'attenzione al centro, dove sono presenti le molecole core drug-like,

rendendo così il modello più generale e rappresentativo. Le tipiche molecole satellite contengono sottostrutture tipiche delle molecole drug-like.

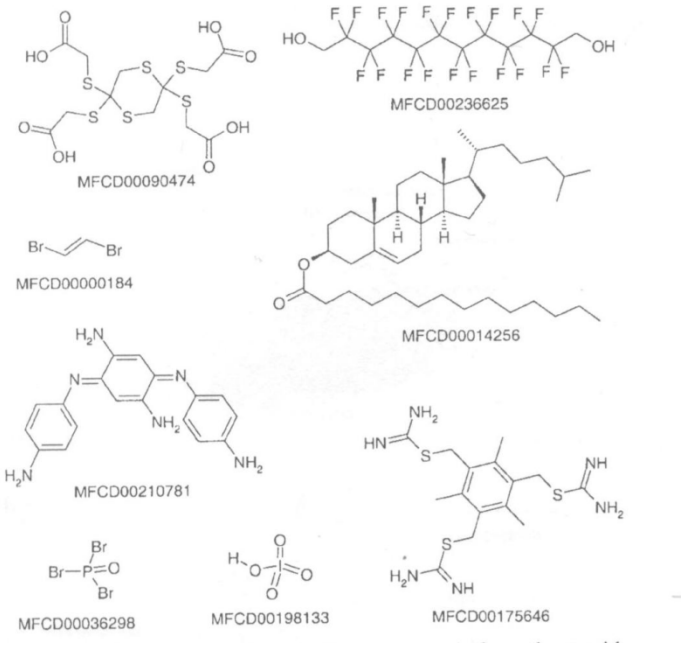


Figura 9 Alcune strutture satellite utilizzate nel modello GPS di Oprea

Il modello PCA costruito può essere usato per proiettare nuove molecole. Gli scores risultanti dall'analisi statistica delle molecole del chemGPS forniscono un sistema di riferimento standard se ad esempio sulla mappa vengono proiettate nuove molecole.

Nei lavori di Oprea, i descrittori molecolari usati nell'analisi statistica multivariata PCA sono di due diverse tipologie: descrittori ricavati dal software SaSA e quelli derivati dal software Volsurf+. Vengono infatti sviluppati in due diversi lavori due modelli: chemGPS e GPSVS basati su due differenti tipologie di descrittori

molecolari rispettivamente 2D e 3D e in seguito confrontati. In entrambi, i modelli PCA viene costruito utilizzando in totale 423 molecole: 202 strutture core e 221 strutture satellite.

Nel nostro caso invece vengono utilizzati due modelli globali le cui proprietà principali sono state calcolate attraverso i descrittori farmacocinetici di VolSurf+ e quelli farmacoforici di FLAP. L'uso di VS+ assicura una descrizione 3D delle molecole, in maniera da delucidarne le proprietà farmacocinetiche, mentre l'uso di FLAP permette di evidenziarne le proprietà di complementarietà con i targets recettoriali.

Lo strumento chemiometrico da noi impiegato per l'interpretazione di queste informazioni chimiche ancora una volta è l' Analisi delle Componenti Principali (PCA), in particolare i t-scores derivati dal metodo di NIPALS[27] (acronimo di non-linear iterative partial least squares).

2.2.2. PCA: uno strumento chemiometrico per l'interpretazione dei descrittori molecolari

L'*analisi in componenti principali* o *PCA* è una tecnica per la semplificazione dei dati utilizzata in ambito della statistica multivariata. Lo scopo primario di questa tecnica è la riduzione di un numero più o meno elevato di variabili (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti. Tale analisi può essere calcolata attraverso diversi algoritmi, tuttavia il più efficiente risulta essere il metodo di NIPALS[27].

Applicato nello specifico caso chemiometrico, questo algoritmo opera effettuando un'analisi sulla matrice dei descrittori calcolati per ogni molecola, nota anche come matrice X multiproprietà. Tale matrice X quindi è costituita da "n" oggetti e "p" variabili. Il metodo scompone la matrice dei dati X in due parti: una parte di struttura ed una di rumore:

$$X = TP^T + E = \text{Structure} + \text{Noise}$$

Il modello PC è il prodotto della matrice TP^T ed è quindi la parte di struttura. L'assunzione che viene fatta è che la matrice X sia scomponibile come somma della matrice struttura TP^T e della matrice del rumore E, come evidenziato nella figura 10.

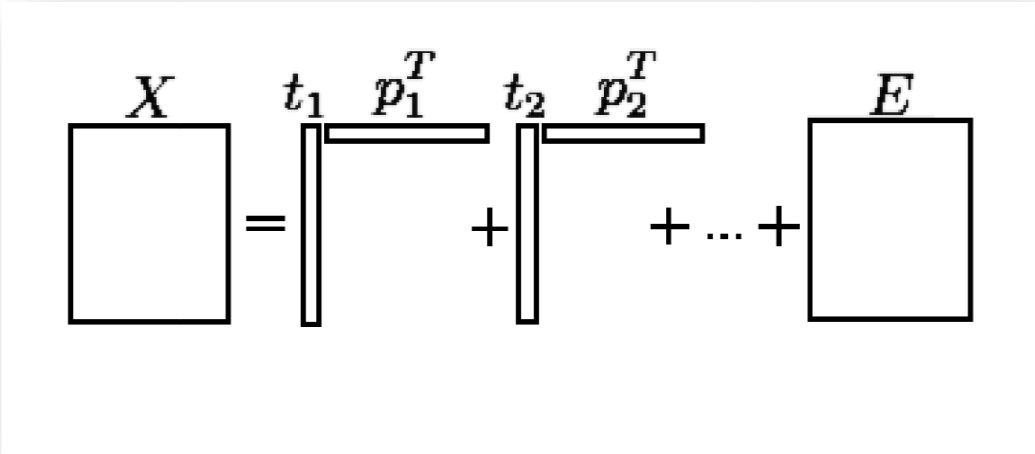


Figura 10 Rappresentazione degli scores e dei loadings

Per quanto riguarda il numero di componenti principali (PC), c' è un valore massimo che è costituito dal valore minimo di "n" o "p" della matrice dati X. Successivamente la matrice TP^T viene scomposta in una sottomatrice t-scores che rappresenta gli scores e una matrice p-loadings che rappresenta i loadings. Gli scores a livello pratico rappresentano quanto differiscono le righe della matrice ovvero quanto differiscono le molecole rispetto alle altre. I loadings invece rappresentano l'influenza delle variabili p nella matrice X, quindi rappresentano l'influenza dei descrittori. Dal plot dei loadings è possibile evidenziare quali variabili sono responsabili dei modelli trovati nella matrice t-score.

2.3. Le Metodologie sviluppate

In questo lavoro vengono proposte due metodologie di virtual screening che differiscono dal criterio di selezione, ma che operano con le tecniche precedentemente citate ovvero la Ligand-Based e la Structure-Based. Entrambi i criteri sfruttano il concetto di similarità e impiegano lo strumento di analisi PCA e lo spazio chimico GPS.

Tutti gli algoritmi che operano questi criteri di selezione sono stati sviluppati in questo laboratorio in linguaggi di programmazione C e python.

La prima metodologia di cui parleremo fa uso di una strategia *Grid-based partitioning*, mentre la seconda fa uso di una tecnica di clustering partitiva. Entrambi i due criteri sono stati successivamente validati attraverso test su dati di letteratura e infine uno di questi è stato scelto ed applicato ad un progetto di virtual-screening: uno studio sulla scoperta di nuovi composti pro-apoptotici non peptidici che inibiscano la proteina XIAP a livello mitocondriale.

2.3.1. "Cmapt": Il Metodo della Griglia

"Cmapt" acronimo di *Cluster MAP Test* è un metodo di tipo Grid-based partitioning nato con lo scopo di sviluppare "mappe" dello spazio chimico per evidenziare le regioni popolate dai composti attivi/inattivi.

Cmapt si basa sulla teoria del campo minato: opera suddividendo lo spazio chimico a 2 componenti principali (2 PCs) in quadrati o rettangoli 2D (dette anche celle), costruendo così una griglia.

Ogni cella popolata da molecole attive e/o inattive assumerà uno score compreso fra 0 ed 1 funzione del numero di molecole che la popolano e dal numero di celle vicinali con almeno una molecola. Vengono distinte quattro tipologie di celle:

- Cella Centrale: 8 possibili celle vicinali
- Cella Laterale Sinistra e Destra: 5 possibili celle vicinali
- Cella Spigolo Sinistra e Destra: 3 possibili celle vicinali
- Cella Superiore e Inferiore: 5 possibili celle vicinali

Score = (Popolazione Cella) X (Numero Celle Vicine Con Almeno Una Molecola)

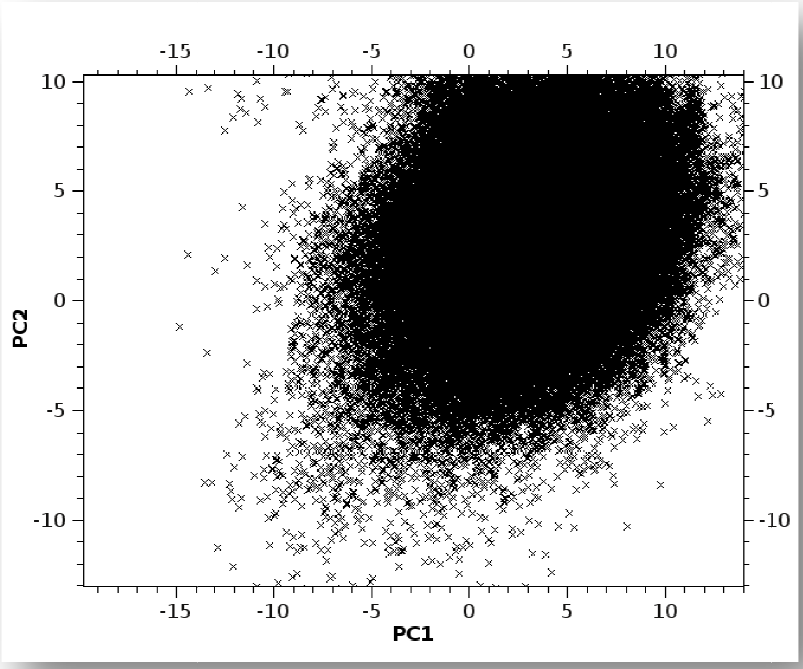


Figura 11 Plot delle prime due componenti principali di VolSurf+ del GPS

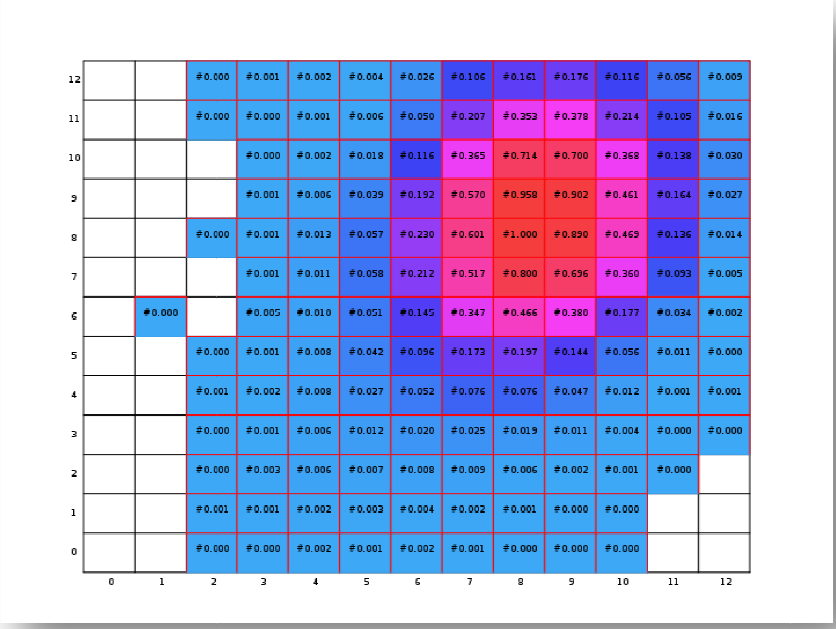


Figura 12: Mappa ottenuta tramite l'algoritmo CMAPT relativa al precedente plot GPS

L'algoritmo viene applicato su 2 delle 10 componenti principali GPS delle molecole (Cmapt permette di scegliere fra le 10 le 2 componenti principali da utilizzare) ottenendo così una mappa colorata con degli score per ogni cella, in funzione della disposizione delle molecole stesse nello spazio chimico GPS. La mappa quindi classifica le molecole simili sotto forma di clusters, evidenziando in rosso la zona "calda" dello spazio GPS dove esse si posizionano.

Ottenuta questa speciale mappa successivamente diventa possibile proiettare molecole non contenute nella mappa per predire le posizioni che loro assumono. In base alle posizioni che vengono assunte si può fornire una categoria di appartenenza alle molecole.

Da un punto di vista pratico in seguito alla proiezione delle molecole sulla mappa dello spazio chimico GPS verranno evidenziate tutte le molecole che popolano le celle colorate in rosso o con una gradazione in rosso (le zone calde della mappa, quelle interessanti per l'oggetto di studio, che presentano uno score prossimo ad 1) e queste molecole verranno successivamente analizzate con i metodi di virtual screening Ligand Based o Structure Based.

Questo criterio di selezione viene suggerito quando si ha a disposizione un numero sufficientemente grande di composti attivi/inattivi su cui sviluppare una *hit-map* appropriata, in maniera da evidenziare facilmente la/le zona/e dove si vanno a collocare le molecole proiettate.

Cmapt è disponibile in versione command-line e possiede diverse opzioni evidenziate nella seguente figura:

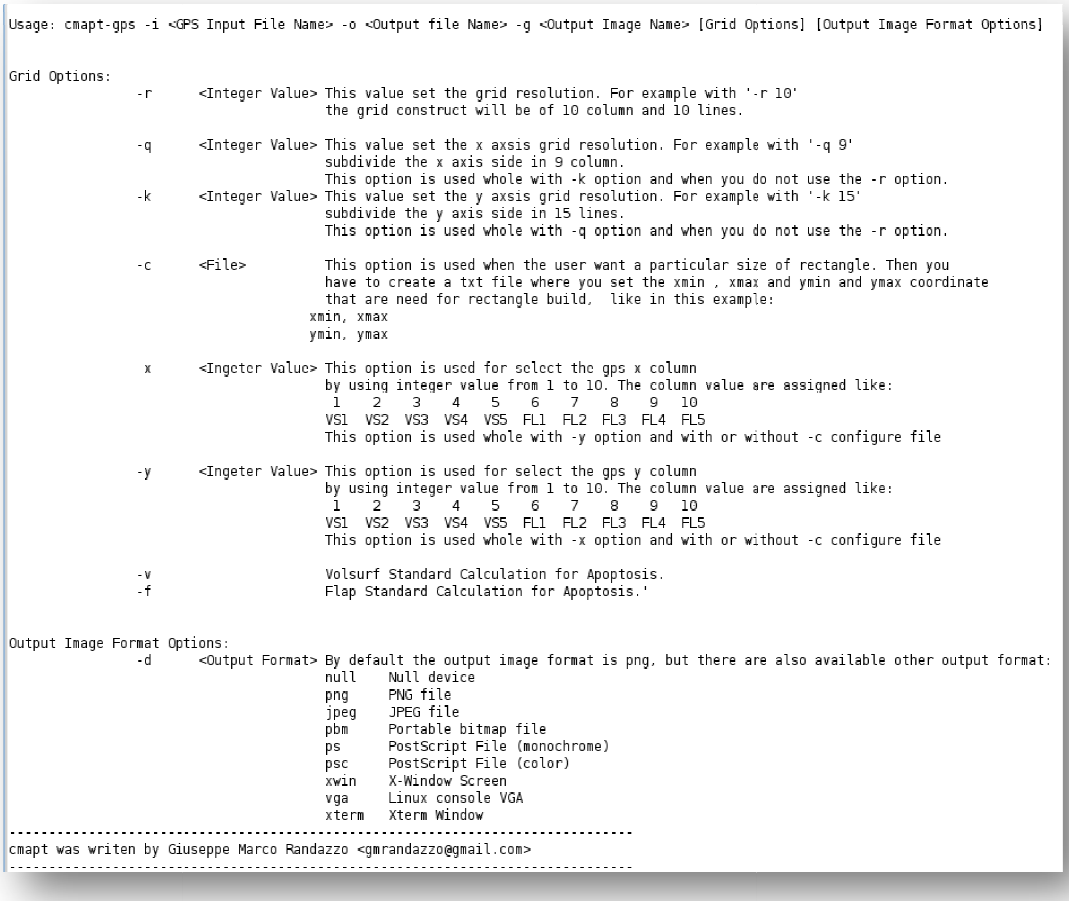


Figura 13 Opzioni linea di comand-line del software cmapt-gps

2.3.1.1. Validazione del metodo cmapt.

In questo capitolo ci occuperemo di validare il metodo Cmapt attraverso un test set raccolto dalla letteratura. Il target scelto è stato il recettore dell'istamina H4[28-29-30-31].

L'istamina è una molecola naturale che esplica la sua funzione fisiologica attraverso quattro proteine G conosciute: H1, H2,H3 e H4. Tali recettori sono stati studiati ed è stato visto che ognuno di essi svolge una funzione differente:

- Il recettore H1 regola le risposte infiammatorie allergiche. Tra i farmaci più usati nel trattamento di queste risposte infiammatorie si ricorda "clarityn"(loratadina)
- Il recettore H2, regola le secrezioni acide a livello gastrico. La "cimetidine" è un farmaco che viene utilizzato nel trattamento delle ulcere gastriche.
- Il recettore H3 si trova abbondantemente nel sistema nervoso centrale e regola il rilascio e la sintesi dell'istamina per modulare i neurotrasmettitori. Gli antagonisti di questo recettore modulano le capacità di attenzione e conoscenza, e quindi vista la sua importanza come target farmacologico oggi molti composti antagonisti si trovano nella fase dei test clinici.
- Il recettore H4, è stato scoperto di recente (2000) ed è stata evidenziata che gli antagonisti di tale recettore svolgono proprietà anti-infiammatorie.

Da un' analisi dalla letteratura sono stati scelti degli articoli da cui sono state classificate molecole antagoniste del recettore H4 in base alla loro attività sperimentale. Complessivamente, il set a disposizione consiste di 25 molecole la cui attività in termini di pK_i è compresa tra 6.0 e 8.5. La seguente tabella riporta le strutture in 2D e i relativi pK_i: Il metodo Cmapt è stato valutato attraverso un

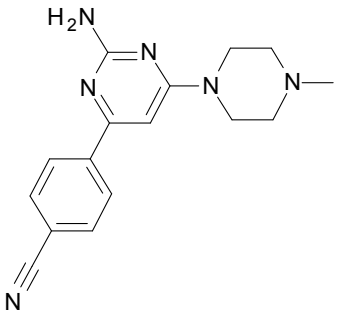
test con dati di letteratura. Il target selezionato è stato quello per il recettore dell'istamina H4 [32-33-34-35].

L'istamina è una molecola naturale che esplica la sua funzione fisiologica interagendo almeno con quattro recettori accoppiati alla proteina G chiamati H1, H2, H3 e H4. Ognuno di questi recettori è implicato in funzioni biochimiche diverse:

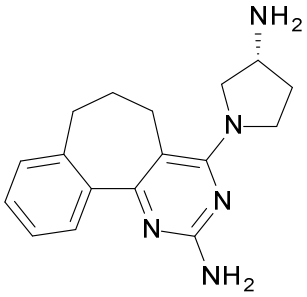
- Il recettore H1 regola le risposte infiammatorie allergiche. Tra i farmaci più usati nel trattamento di queste risposte infiammatorie si ricorda il farmaco chiamato "clarityn"(loratadina)
- Il recettore H2, regola le secrezioni acide a livello gastrico. La "cimetidina" è un farmaco che viene utilizzato nel trattamento delle ulcere gastriche.
- Il recettore H3 si trova abbondantemente nel sistema nervoso centrale e regola il rilascio e la sintesi dell'istamina per modulare i neurotrasmettitori. Gli antagonisti di questo recettore modulano le capacità di attenzione e coscienza, e quindi vista la sua importanza come target farmacologico oggi molti composti antagonisti si trovano nella fase dei test clinici.
- Il recettore H4, è stato scoperto di recente (2000) ed è stato evidenziato che gli antagonisti di tale recettore svolgono proprietà anti-infiammatorie.

La validazione dell'algoritmo è stata effettuata in diverse fasi: prima di tutto abbiamo selezionato delle molecole di letterature attive contro il recettore H4.

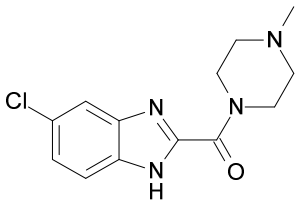
Sono state scelte molecole con attività antagonista del recettore H4 in base alla loro attività sperimentale. Complessivamente, il set a disposizione consiste di 25 molecole la cui attività in termini di pK_i è compresa tra 6.0 (poco attiva) e 8.5 (molto attiva). La seguente tabella riporta le strutture in 2D e i relativi pK_i:



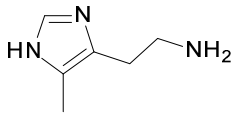
8.53



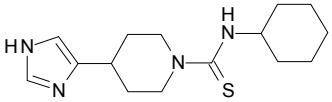
8.33



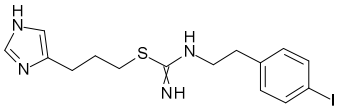
7.1



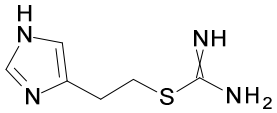
7.3



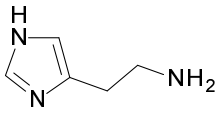
6.49



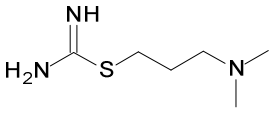
8.1



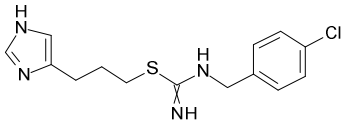
8.44



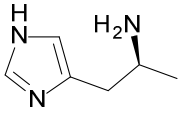
7.87



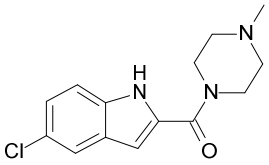
7.04



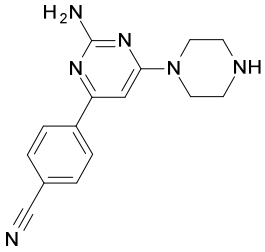
8.30



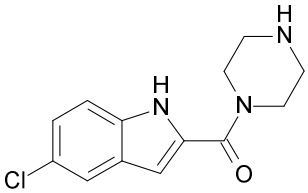
6.73



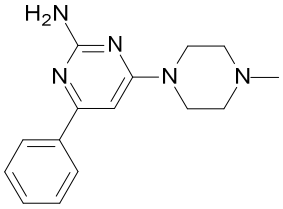
7.92



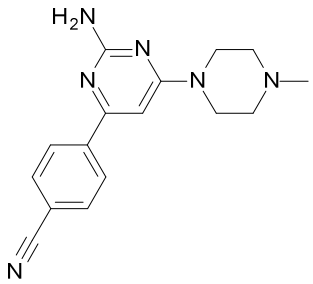
7.38



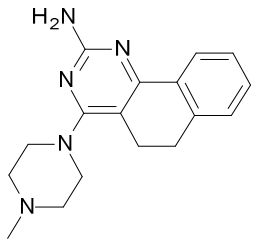
7.59



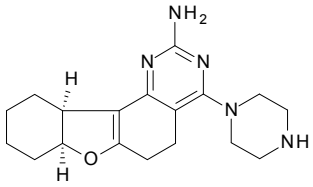
8.39



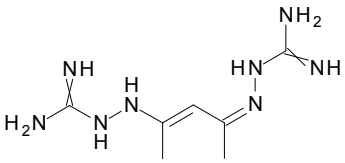
7.46



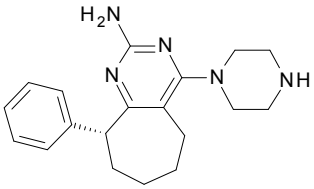
7.81



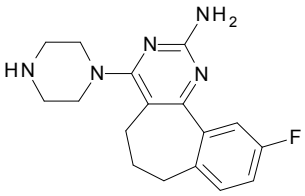
8.24



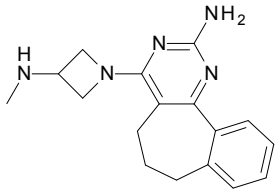
6.04



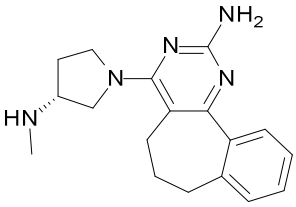
7.40



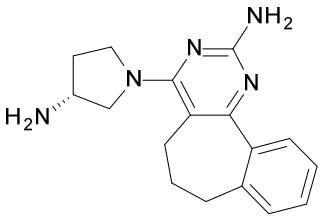
8.29



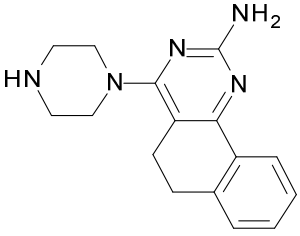
7.96



8.46



8.33



7.28

Nello step successivo tramite i descrittori farmacocinetici di VolSurf+ , quelli farmacoforici di FLAP e uno script in python sviluppato nel precedente lavoro di tesi[36], sono state calcolate le coordinate GPS costituite da 10 componenti principali di queste molecole, per poi plottarele ed analizzarle con il nostro algoritmo Cmapt. Nella figura 14 mostriamo il plot delle prime due componenti principali GPS delle molecole:

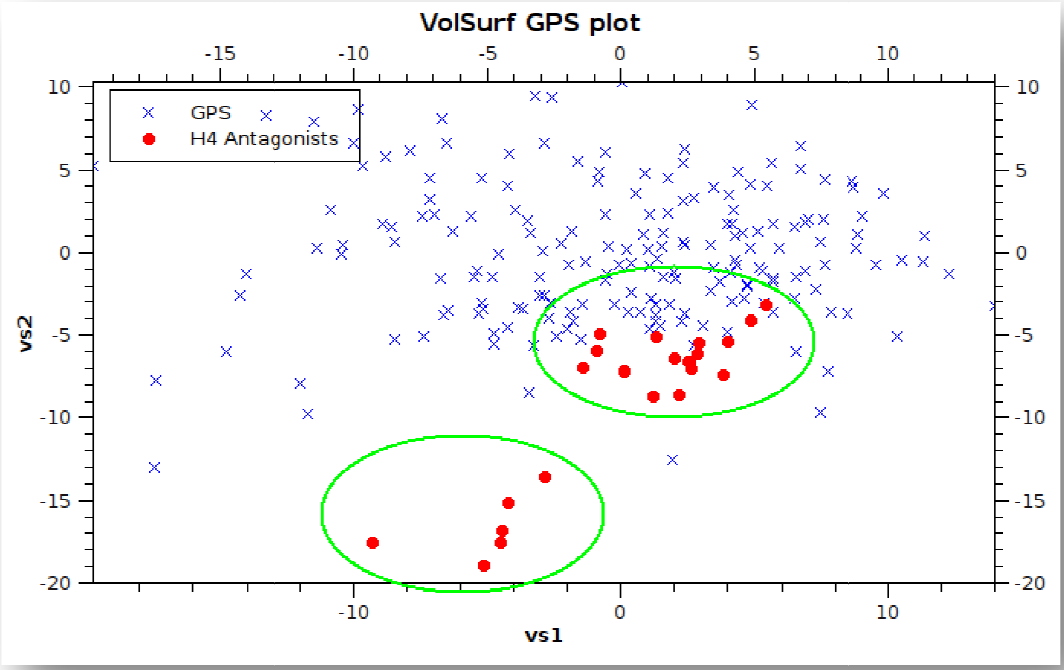


Figura 14 Plot GPS delle molecole antagoniste del recettore H4

La figura mostra che gli antagonisti-H4 si localizzano in due clusters ben definiti e compatti. Si possono ipotizzare due differenti tipologie di binding per il recettore H4, evidenziati dalla disposizione in due clusters che queste molecole assumono nello spazio GPS. Purtroppo la mancanza di dati sperimentali specifici non permette speculazioni aggiuntive su queste ipotesi.

Applicando l'algoritmo CMAPT utilizzando una griglia costituita da 15 righe e 15 colonne e attenendoci ai limiti dello spazio gps delle prime due componenti ottenute dai descrittori di VolSurf+ si ottiene la mappa mostrata nella successiva figura:

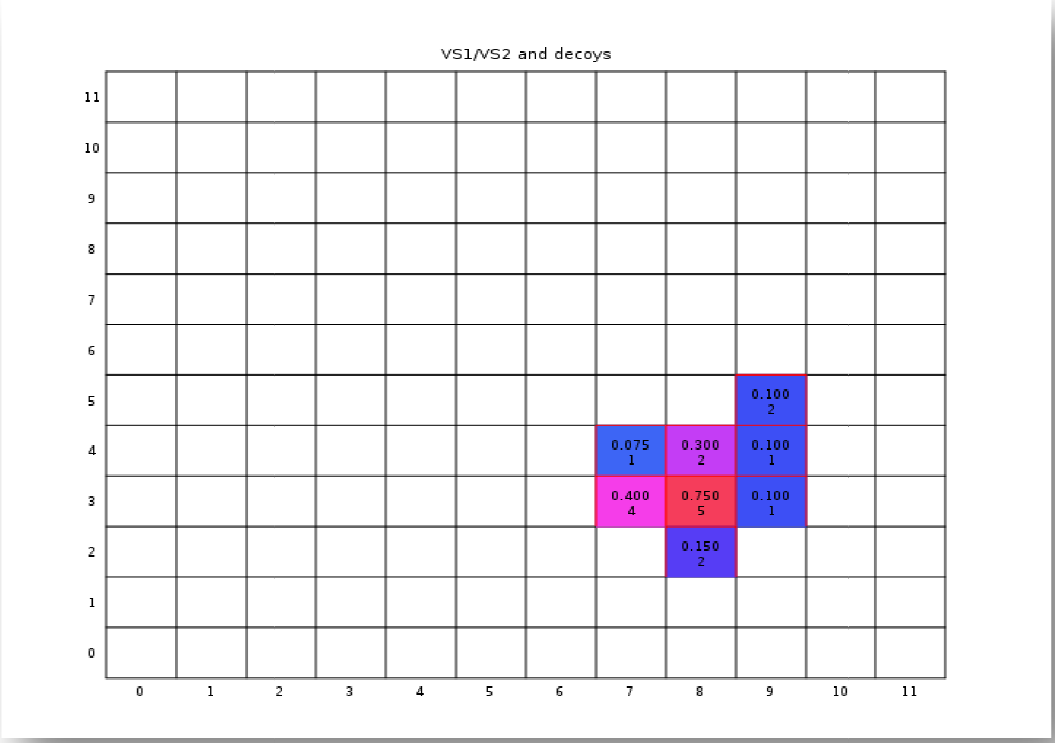


Figura 15 Mappa dello spazio chimico occupato dagli antagonisti del recettore H4

Da notare l'assenza del gruppo di molecole che si trovano al di fuori del limite dello spazio gps, perciò l'algoritmo è stato in grado di riconoscere queste molecole e classificarle come “outliers”.

Infine come detto precedentemente, su queste mappe vengono proiettate le componenti principali GPS delle molecole di un database e attraverso il tool “ctestmm” che fa parte del pacchetto cmapt-gps vengono selezionate soltanto quelle molecole che si trovano all'interno delle celle evidenziate in rosso o gradazione di rosso (celle con un alto valore di score) per poi analizzarle con l'approccio ligand-based o structure-based.

In questo esempio sono state proiettate le componenti principali delle molecole del database ChemDiv e sono state selezionate le molecole che occupano la cella (8;2) come mostrato in Figura 15 (cella con score pari a 0.75). Attraverso questo algoritmo sono state quindi selezionate 4181 molecole da analizzare che corrisponde allo 0.52% dell’intero database, quindi una riduzione notevole.

2.2.3 "CLAN": Il Metodo di Clustering

"CLAN" acronimo di "Cluster ANalysis" è un metodo di clustering partitivo in cui vengono classificati gruppi di molecole attraverso le coordinate GPS delle molecole. CLAN è stato pensato per lavorare con un training set minore di 20 molecole in modo da permettere facilmente la scelta di una molecola rappresentativa del training set da cui poi effettuare confronti di similarità.
CLAN può operare tramite due metodi:

- Primo metodo:
  - Scelta di una molecola rappresentativa del data set delle attive
  - Calcolo della distanza euclidea *n*-dimensionale tra ogni molecola del database e la molecola scelta.
  - Plot delle funzioni statistiche *sensitivity* e *abundance*

- Secondo metodo:
  - Calcolo della distanza euclidea *n*-dimensionali tra ogni molecola del database e tutte le molecole attive del data set.
  - Selezione della minima distanza fra tutte le distanze (funzione del numero del data set delle molecole attive) che ogni molecola del database ha.
  - minimizzazione della funzione statistica *abundance*.

Nel primo metodo le distanze N-Dimensionali sono calcolate utilizzando le coordinate GPS dei due modelli globali, il modello di VolSurf+ e il modello di FLAP, in maniera tale da farli entrare in sinergia per estrarre il massimo delle informazioni. Per ogni modello vi sono 5 componenti principali quindi in totale la distanza euclidea sarà una distanza a 10 dimensioni e verrà calcolata attraverso la funzione:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}.$$

dove $p_1,p_2,...,p_n$ sono le coordinate del punto d'origine P $(p_1,p_2,...,p_n)$ e $q_1,q_2,...,q_n$ coordinate del punto $Q(q_1,q_2,...,q_n)$. Ricordiamo che ogni punto rappresenta una molecola descritta attraverso dei valori numerici che sono appunto i "descrittori molecolari". Da un punto di vista pratico l'origine da cui calcolare le distanze è data dalle coordinate GPS della molecola scelta.

Nel punto successivo si calcolano a partire da questi valori di distanza delle funzioni statistiche, la *sensitivity* e *l'abundance*.

La sensitivity è una misura della percentuale della molecole attive che sono correttamente identificate come tali ad un certo valore di distanza. Da un punto di vista matematico la sensitivity è definita come:

$$\frac{Numero\ Molecole\ Attive\ al\ valor\ e\ X\ di\ distanza}{Numero\ Molcole\ attive\ + Numero\ Molecole\ False\ Inattive\ al\ valore\ X\ di\ distanza}$$

L'abundance invece rappresenta una misura della proporzione di molecole presenti rispetto al totale ad un certo valore di distanza. Si derivata dalla funzione sensitivity e si traduce in formula matematica secondo questa equazione:

$$\frac{Numero\ di\ Molecole\ al\ valore\ X\ di\ distanza}{Numero\ Totale\ di\ Molecole}$$

I risultati di queste due funzioni statistiche *abundance* e *sensitivity* sono rappresentati attraverso plot 2D dove nelle ascisse si riporta la distanza da una molecola del dataset delle attive, mentre nell'ordinata si riporta il valore che la funzione assume. Da questi, in ultima analisi viene scelta una distanza limite in funzione di un valore percentuale della quantità delle molecole del database da analizzare attraverso il metodo FLAP (generalmente questo valore è compreso

tra l' 1 o il 10 % del database), e dal valore che la *sensitivity* assume in quel valore di distanza è possibile valutare la probabilità di trovare molecole attive.

Nel secondo modo invece, scelte le "*m*" molecole attive del data set, da queste verranno calcolate le distanze *n*-Dimensionali per ogni molecola del database, ottenendo così "m" distanze euclidee per ogni molecola del database.

Successivamente attraverso un semplice algoritmo vengono comparate le "m" distanze euclidee che ogni molecola ha e viene scelta la distanza più piccola (Vedi Figura 16). Questa operazione viene fatta perché ogni molecola attiva del training set rappresenta una struttura a se stante ma con una attività comune alle altre, e quindi permette di analizzare da un database un set vario di molecole che probabilmente non sono simili fra loro, ma sono simili alla molecola del training set a cui esse sono vicine.

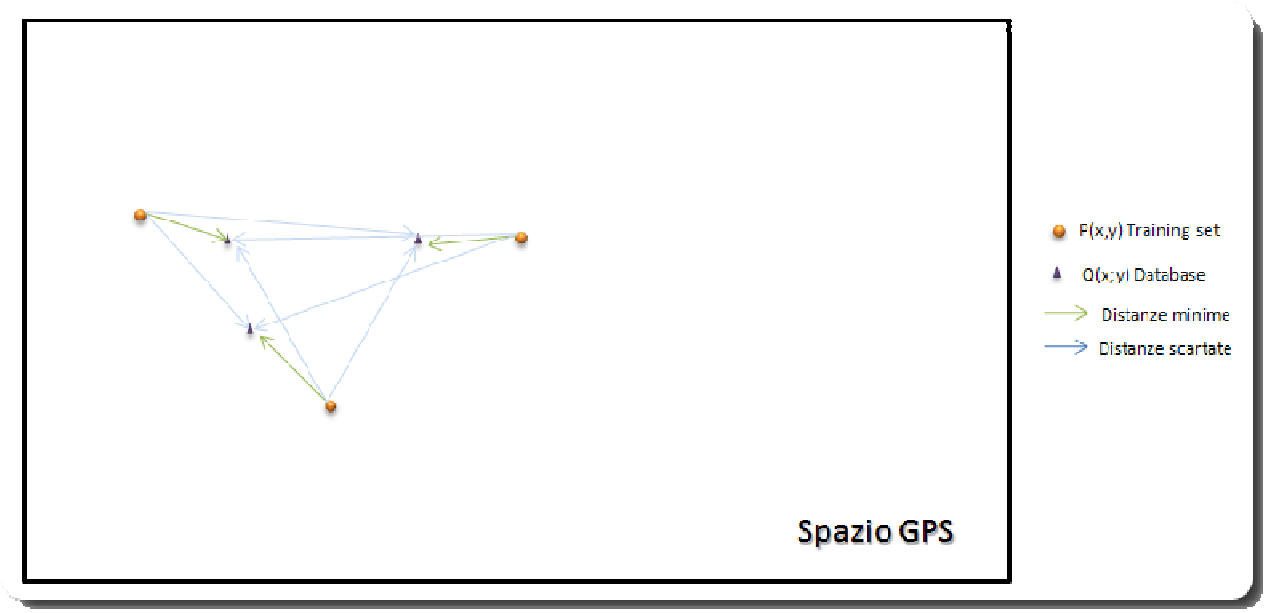


Figura 16 Esempio del calcolo delle distanze minime

Infine viene scelto un set ridotto di molecole attraverso la rappresentazione grafica 2D della funzione statistica *abundance.*

2.3.2.1 Validazione del metodo "CLAN"

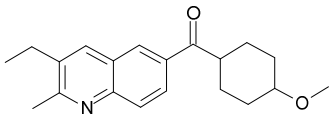
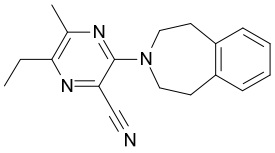
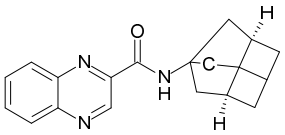
Per validare questo metodo sono stati analizzati due casi estratti dalla letteratura scientifica. Nel primo caso ci siamo occupati di "ricerca di antagonisti non competitivi dei recettori metabotropici del glutammato"[37]. I recettori accoppiati alla proteina G (GPCRs) rappresentano una grande famiglia di proteine che gioca un ruolo importante in molti processi fisiologici e patofisiologici. Fanno parte di questa superfamiglia i recettori metabotropici del glutammato mGluRs (in particolare appartenenti alla famiglia C). Tali recettori mediano le trasmissioni eccitatorie nella superficie cellulare legando inizialmente il glutammato per poi influenzare la forza delle trasmissioni sinaptiche. Tali recettori sono ampiamente distribuiti in molte regioni del cervello, compreso l'ippocampo, il cerebellum, il nucleo talamico, e la corda spinale. La stimolazione dei recettori **mGluR1** e **mGluR5** porta all'idrolisi dell'enzima phosphoinositide (PI) e all'innalzamento dei livelli di calcio intracellulare. I recettori mGluR1 sembrano agire nei danni celebrali e nel dolore, mentre i recettori mGluR5 sono stati visti sovraespressi in diverse patologie neurologiche come l'ansia, la depressione, il morbo di Parkinson etc…

Dei due target è stato scelto di analizzare il recettore mGluR1 che presenta una distribuzione migliore tra molecole attive ed inattive:

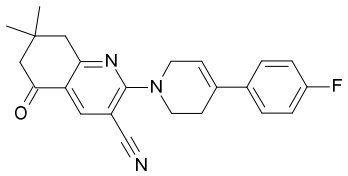
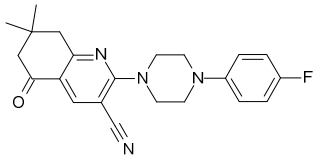
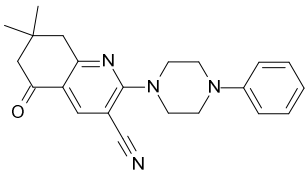
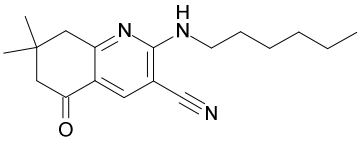
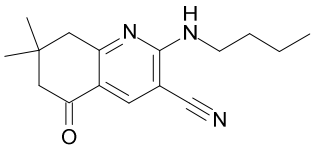
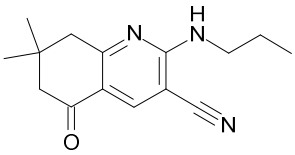
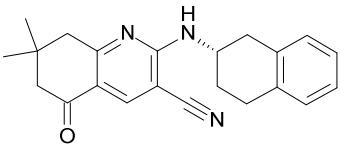
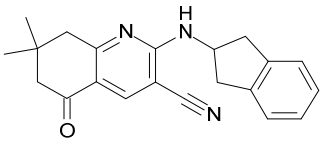
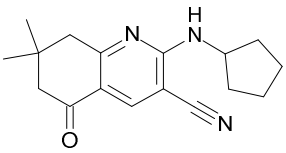
Tabella 2 Data set a disposizione

Training Set	Attive	Inattive
3	9	2

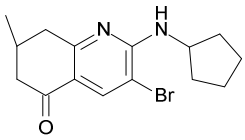
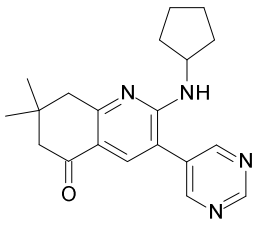
Training set:

Attive:

Inattive:

I plot delle proiezioni sul GPS (vedi figure 17 e 18) mostrano che le molecole attive sono vicine alle molecole del training set, nonostante lo scheletro molecolare principale sia significativamente diverso.

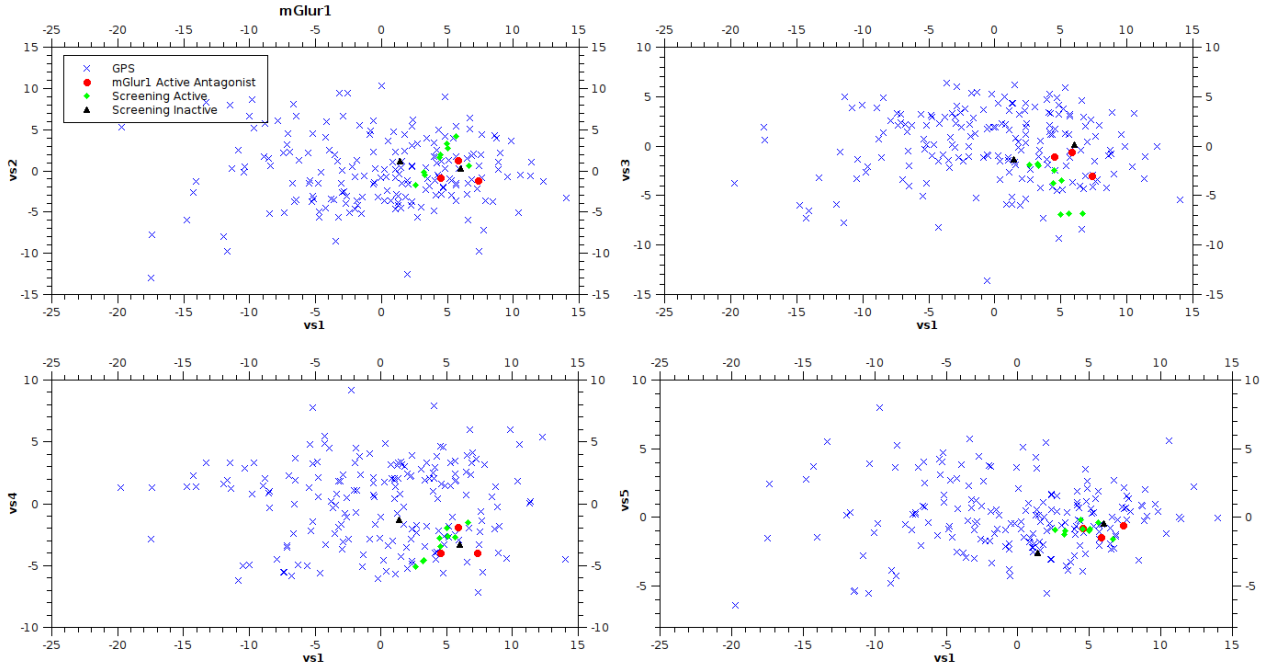


Figura 17 Plot delle componenti principali del GPS ottenuto con VolSurf+

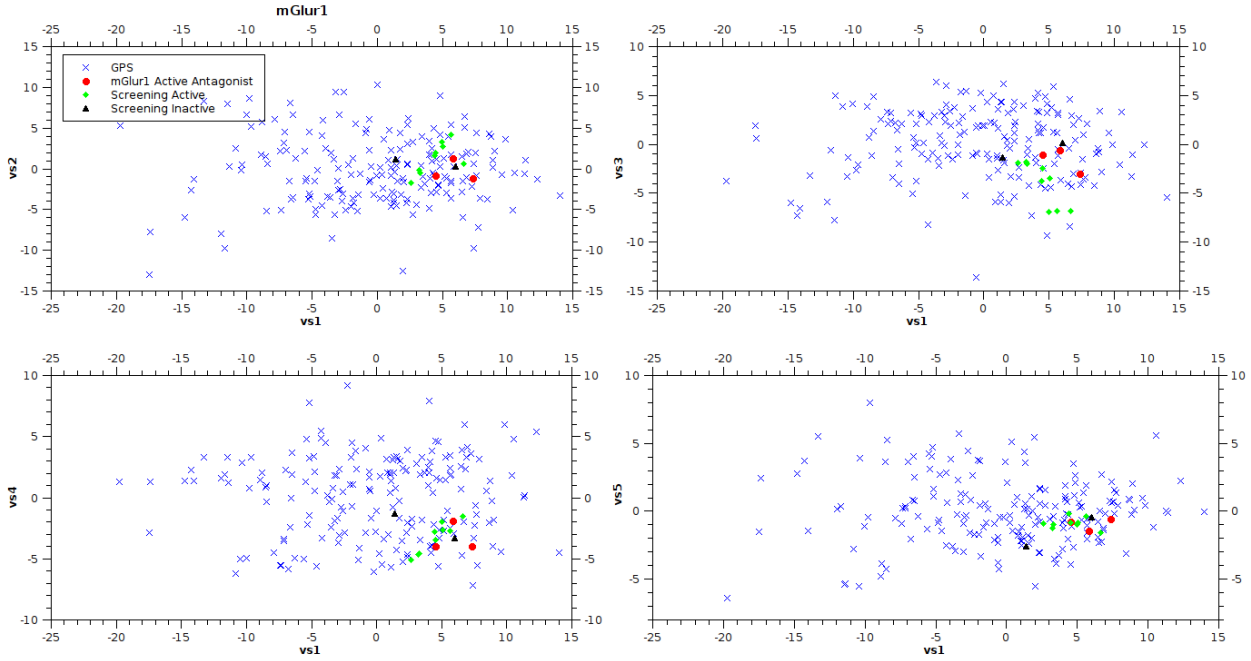


Figura 18 Plot delle componenti principali del GPS ottenuto con FLAP

Applicando il secondo metodo dell'algoritmo CLAN sul dataset delle molecole attive/inattive e sul database ChemDiv, che consiste nel calcolo delle distanze minime delle molecole del database da ogni molecola del dataset, si procede

graficando la funzione abundance e scegliendo tre valori di distanza limite relativi all'1, al 5 e al 10 % del valore della funzione abundance. Infine verranno selezionate ed estratte tutte le molecole del database che stanno al di sotto di questi valori di soglia per poi analizzarle con i metodi di virtual screening Ligand Based o Structure Based. Nel caso scelto otteniamo i seguenti risultati:

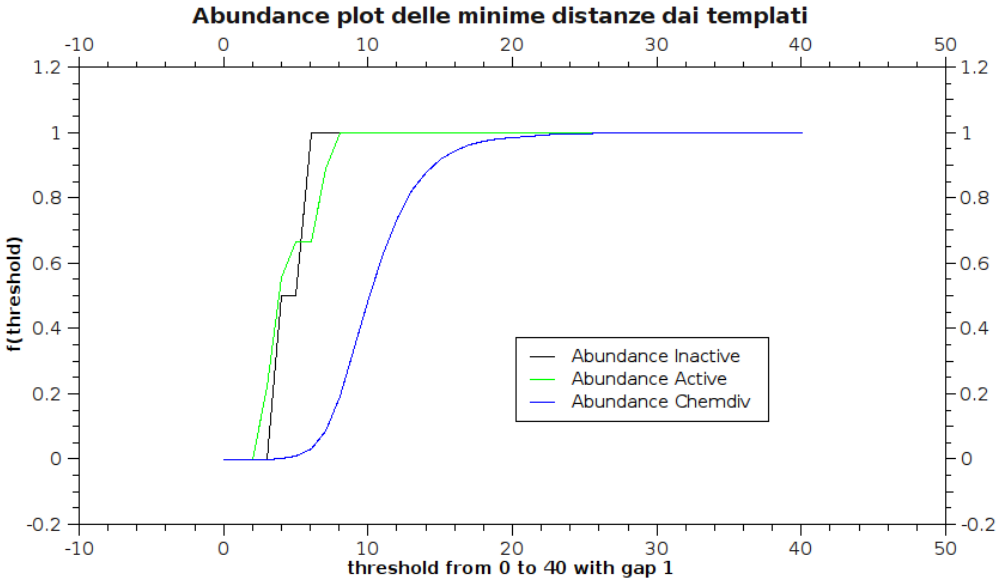


Figura 19 Grafico delle abbondanze

Tabella 3 Tabella riassuntiva del numero di molecole attive presenti a varie % di *abbundance* del database, delle attive ed inattive.

Chemdiv	Active	Inactive	Distance	No. Active	No. Chemdiv
1%	65%	49%	4.88	6/9	7275/699903
5%	75%	100%	6.39	7/9	34215/699903
10%	90%	100%	7.18	8/9	71455/699903

Da questo metodo è emerso che scelta la distanza limite corrispondente al 10% dell'abundance del database ChemDiv sarebbero rientrate all'interno di essa 8/9 molecole dello screening risultate sperimentalmente attive, un risultato notevole che permette di risparmiare tempo con la minima perdita di potenziali candidati attivi. Ciò verifica il principi di similarità basato sulla distanza, in cui molecole

simili hanno proprietà simili, e se sono simili queste saranno vicine l'una con l'altra e quindi confinate in una regione ben precisa dello spazio chimico GPS.

Il secondo caso che è stato affrontato per la validazione del metodo riguarda la "ricerca di inibitori dell'enzima Carbonato deidrasi IX"[38].

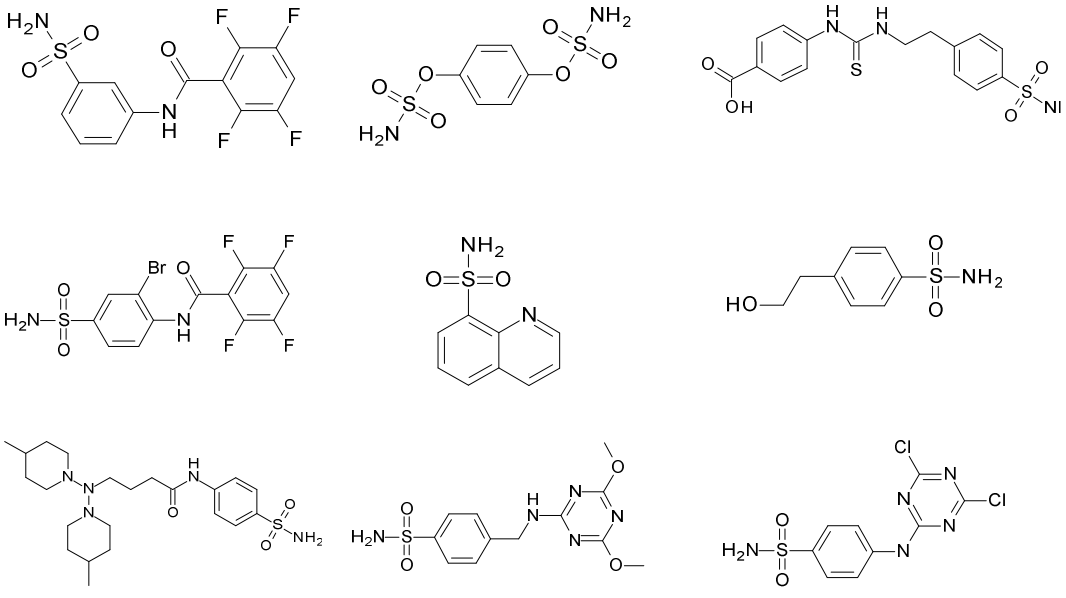
L'ipossia nei tumori è uno dei microeventi che caratterizzano la crescita e lo sviluppo di un tumore in cui la cellula tumorale viene privata dall'ossigeno. Nonostante ciò il tumore continua a crescere sviluppandosi nei tessuti e sottraendo ossigeno ad essi, portando così ad una crescita incontrollata del tumore stesso. È stato visto che l'enzima carbonato deidrasi IX risulta essere sovra espresso nella maggior parte dei tumori umani. Per cui tale enzima risulta essere un target terapeutico importante al fine di bloccare la crescita incontrollata dei tumori stessi per poi debellarli con gli attuali mezzi.

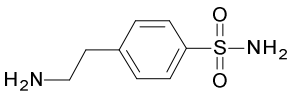
Il set di dati che presenta questo caso è costituito da:

Tabella 4 Data set a disposizione

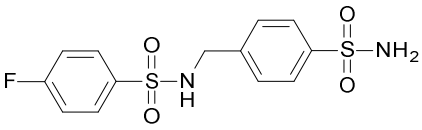
Training set	Attive	Inattive
10	3	3

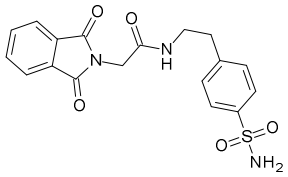
Training set:

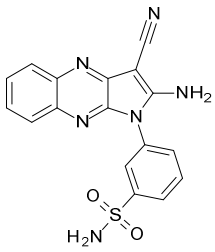




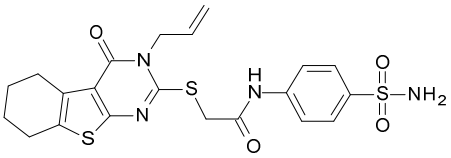
Attive:

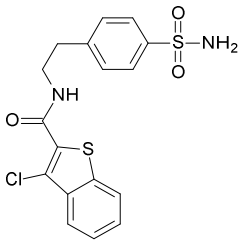


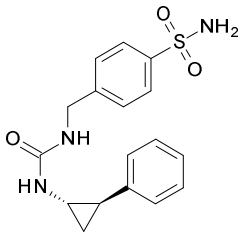




Inattive:







In questo caso il training set è costituito da molecole differenti che, nonostante abbiano tutte il gruppo solfonamidico in comune, si posizionano in parti differenti dello spazio. Le molecole risultate attive dai test sperimentali sono vicine solo ad alcune di esse.

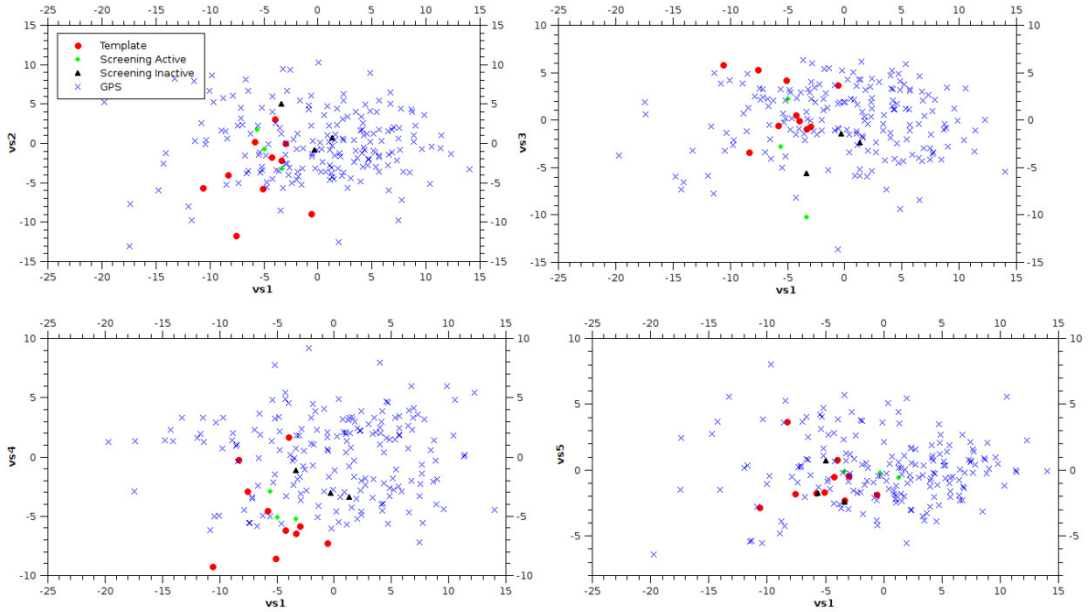


Figura 20 Plot delle componenti principali del GPS ottenuto con VolSurf+

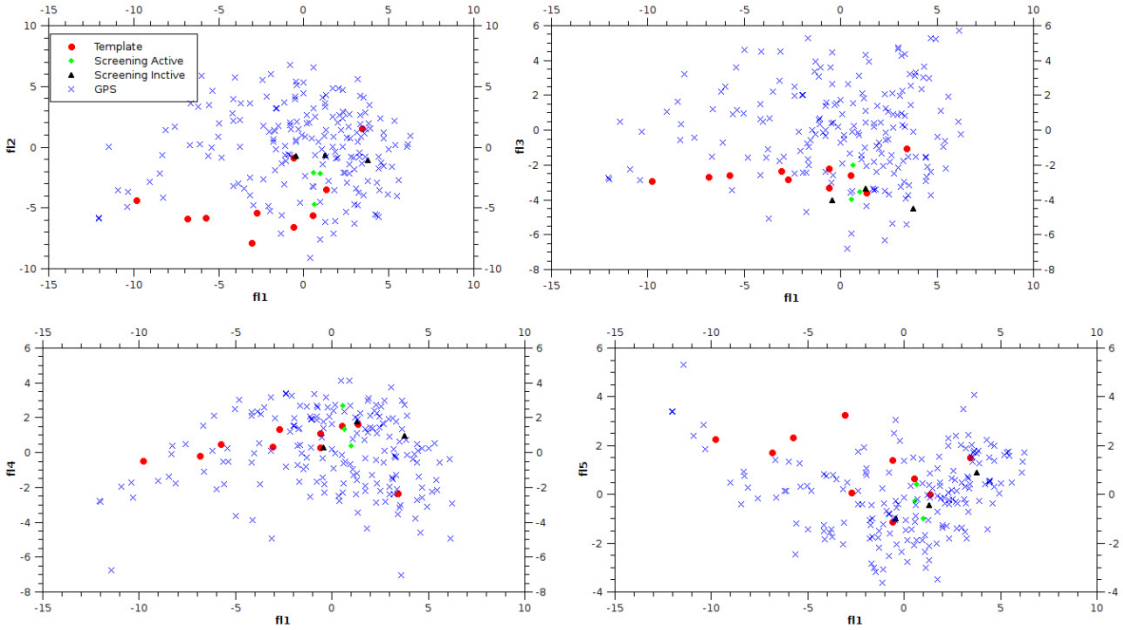


Figura 21 Plot delle componenti principali del GPS ottenuto con FLAP

Applicando anche in questo caso in studio il secondo metodo dell'algoritmo CLAN sul dataset delle molecole attive/inattive e sul database ChemDiv, si procede graficando la funzione abundance e si scelgono tre valori di distanza limite relativi all'1, al 5 e al 10 % del valore della funzione abundance. Infine verranno selezionate ed estratte tutte le molecole del database che stanno al di sotto di questi valori di distanza limite per poi analizzarle con i metodi di virtual

screening ligand-based o structure-based. In questo caso scelto per la validazione del metodo CLAN otteniamo i seguenti risultati:

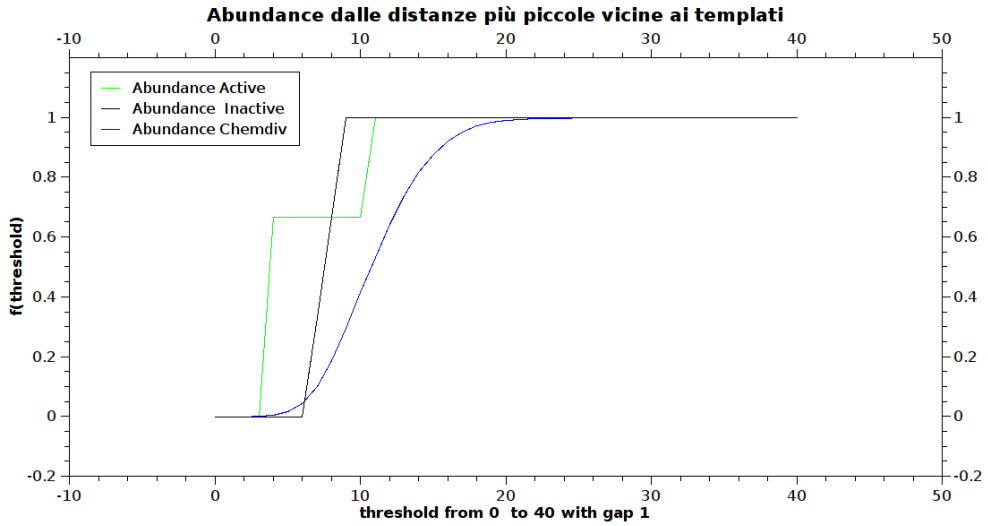


Figura 22 Grafico delle abbondanze

Tabella 5 Tabella riassuntiva del numero di molecole attive presenti a varie % di *abbundance* del database, delle attive ed inattive.

Chemdiv	Active	Inactive	Distance	No. Active	No. Chemdiv
1%	66%	0%	4.73	2/3	8935/699903
5%	66%	5%	6.21	2/3	37720/699903
10%	66%	34%	7.02	2/3	80625/699903

Ancora una volta scegliendo la distanza limite corrispondente al 10% dell'abundance del database ChemDiv permette di far rientrate all'interno di essa 2/3 molecole dello screening risultate sperimentalmente attive, verificando quindi solo una lieve perdita di molecole possibili candidate attive con il beneficio del notevole risparmio in termini di tempistica di calcolo.

3. Applicazione delle metodologie sviluppate

3.1. Ricerca di nuove molecole pro apoptosi che inibiscano il dominio Bir3 della proteina XIAP

3.1.1. L'Apoptosi

Apoptosi deriva dal greco *"apo"*e *"ptosis"* che significa "*la caduta delle foglie e dei petali dei fiori*", e questo processo indica una forma di morte cellulare programmata attraverso un meccanismo ben preciso. La regolazione di questo processo è assai critica e di estrema importanza per la crescita di un individuo. Le disfunzioni di questo meccanismo possono portare ad una varietà di malattie tra cui malattie neurodegenerative ed in particolare a neoplasie maligne (cancro), in quanto queste disfunzioni conferiscono resistenza alle cellule cancerogene durante il trattamento clinico con agenti terapeutici, con la conseguenza del fallimento terapeutico stesso. Tali disfunzioni in genere sono dovute alla famiglia di proteine IAP (Inibitor of apoptosis protein) che inibiscono all'interno della cellula il rilascio di proteine chiamate *caspasi* (contrazione di proteasi cistein-dipendente aspartasi-specifica) che, a loro volta, sono dei messaggeri che inducono la morte della cellula stessa. Recentemente è stata identificata la chiave dell'inibizione dell'apoptosi, la proteina XIAP[39-40-41-42-43-44] (X-linked IAP) umana che inibisce in maniera diretta una famiglia di caspasi (caspasi-3, caspasi-7, caspasi-9). La proteina XIAP contiene tre domini baculoviral IAP repeats (BIR1, BIR2 e BIR3), e ognuno di questi è costituito da circa 70 amminoacidi. Mediante cristallografia ai raggi-X e studi all'NMR è stato ipotizzato il meccanismo di

inibizione del terzo dominio (BIR3) della XIAP sulla caspasi-9. Il dominio BIR3 si lega selettivamente all'amminoacido N terminale della caspasi-9, esponendola così a proteolisi e convertendola in procaspasi-9. Per quanto riguarda invece il secondo dominio (BIR2) della XIAP, si sa soltanto che inibisce selettivamente sia la caspasi-3 che la caspasi-7, ma ancora oggi non se ne conosce il meccanismo. Poiché le caspasi giocano un ruolo fondamentale nel meccanismo apoptotico mitocondriale, il processo di inibizione di queste molecole ad opera della proteina XIAP previene la formazione di specie catalitiche che attivano il processo apoptotico stesso. Quindi la proteina XIAP è ritenuta un'efficiente inibitore dell'apoptosi. Recentemente è stata scoperta una proteina nei mammiferi dal nome Smac (Second Mitochondrial acrivator of caspases) che svolge la funzione di antagonista alla proteina antiapoptotica XIAP, che presenta un basso punto isoelettrico e viene rilasciata dai mitocondri nel citosol in seguito ad una risposta a messaggi apoptotici. La proteina Smac non fa altro che promuovere l'attivazione della caspasi-9 competendo con la caspasi stessa per il sito attivo della IAP nel dominio BIR3. La Smac in particolare si lega al dominio BIR3 attraverso i primi 4 residui amminoacidici AVPI (Alanina-Valina-Prolina-Isoleucina). Prendendo come punto di partenza questi 4 residui amminoacidici sono state sviluppate diverse molecole che mimano l'effetto della proteina Smac e che possono essere utili nella cura del cancro. Avendo come target la proteina XIAP, molecole che la inibiscono permettono il rilascio delle caspasi stesse, al fine di rallentare e controllare la crescita delle cellule tumorali. Il meccanismo apoptotico è quindi oggi un target promettente per la scoperta e lo sviluppo di nuovi farmaci, dato che molte malattie umane sono legate al malfunzionamento di questo.

In precedenti lavori pubblicati in importanti riviste internazionali (Nature) è stato presentato lo sviluppo di molecole di derivazione peptidica AVPI; tuttavia, il campo è interessante e c'è spazio per nuove idee. Avendo in corso un progetto di ricerca con i Prof R. Mannhold (Università di Dusseldorf) e S.

Fulda (Università di Ulm) proprio in questo campo, ci si è posti l'obiettivo di ricercare molecole AVPI mimetiche di tipo non peptidico attraverso l'ausilio di una delle nostre metodologie sviluppate in questo periodo di tesi.

3.1.1.1. L'interazione Smac-BIR3

Nel novembre del 2000 è stata ottenuta ai raggi X la struttura della proteina XIAP (Codice PDB: 1G73[45]) permettendo così di identificare il ligando che inibisce questa proteina ed evidenziare le interazioni che vengono messe in gioco con esso.

La figura 23 mette in luce tale interazione, il tetrapeptide AVPI appartenente alla proteina Smac con il dominio BIR3 della proteina XIAP.

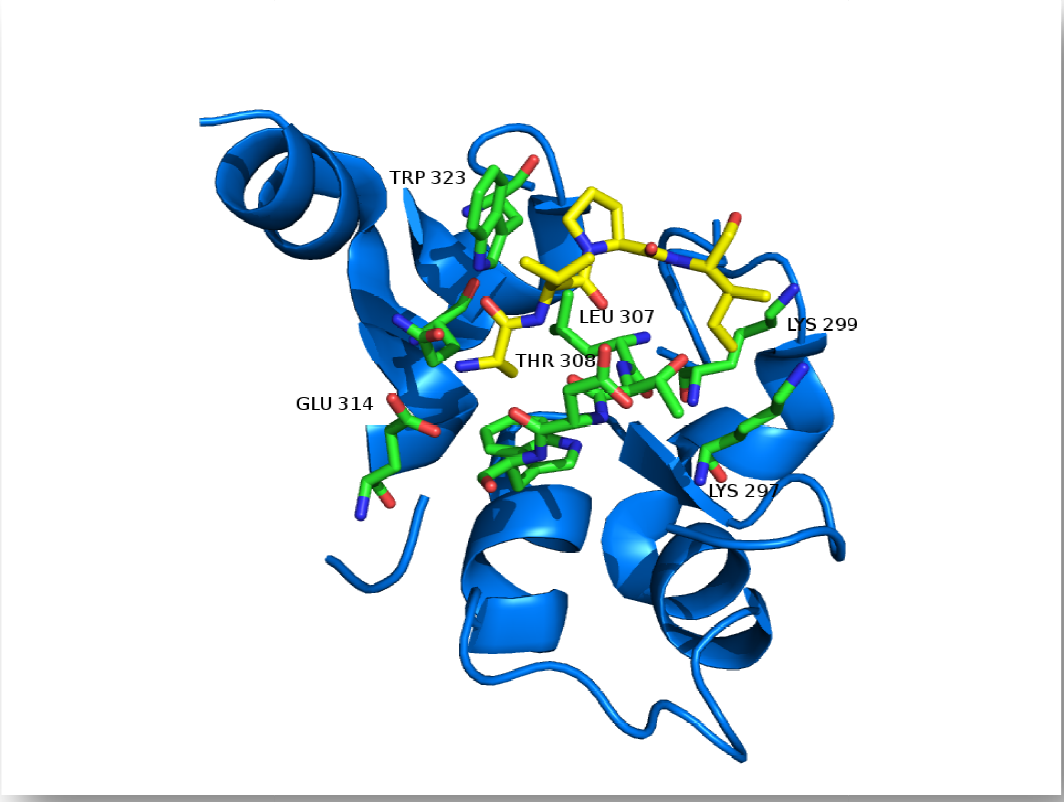


Figura 23 Interazioni fra i residui aminoacidici del dominio BIR3 e il tetrapeptide AVPI

Per mezzo di studi SAR e studi sulle strutture cristallografiche ed NMR, sono state analizzate tutte le interazioni dei singoli amminoacidi che avvengono all'interno del dominio BIR3[46-47]:

- Alanina (ALA 1): è stato visto che la carica positiva sul residuo N-terminale è critica affinché tutto il peptide AVPI si possa legare al dominio stesso tramite un'interazione carica-carica con la carica negativa situata nel

gruppo carbossilico della catena laterale del residuo aminoacidico GLU 314. Inoltre il metile della catena laterale va ad inserirsi all'interno di una piccola tasca idrofobica.

- Valina (VAL 2): il legame peptidico alanina-valina è di importanza critica in quanto questo forma due legami a idrogeno con il residuo aminoacidico THR 308: uno con il doppietto elettronico sul carbonile come accettore, ed uno con l'idrogeno del gruppo amminico come donatore. Inoltre la catena laterale della valina non ha specifiche interazioni con la proteina, e questo è stato visto anche nella struttura cristallografica XIAP-Smac.

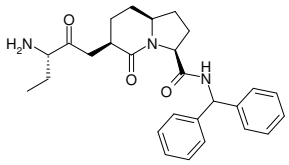
- Prolina (PRO 3): instaura un'interazione idrofobica con il dominio BIR3 attraverso i residui LEU 307 e TRP 323, più in particolare con i 5 atomi dell'anello della prolina.

- Isoleucina (ILE 4): instaura una interazione idrofobica con i residui LYS 297 e LYS 299; inoltre, il gruppo carbonilico non ha specifiche interazioni con la proteina ma viene a contatto con il solvente. Tuttavia, si pensa che questo gruppo possa contribuire nell'affinità di legame fra Smac-XIAP andando a controllare l'orientazione della catena idrofobica laterale della isoleucina, riducendo così la perdita di entropia in seguito alla formazione del complesso Smac-XIAP.

3.1.2. Il Data Set

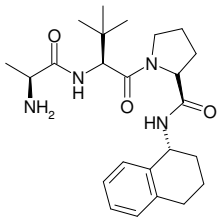
Il data set che viene usato per studiare nuovi composti AVPI mimetici è stato ottenuto da dati di letteratura[48-49-50-51-52-53-54] ed è costituito da 25 molecole classificate come attive e 55 molecole come inattive. L'attività di questi composti è stata valutata attraverso i valori di K_i riportati dalla letteratura. Composti con un valore di K_i > 1 µM sono stati considerati inattivi mentre quelli con valore di K_i < 1 µM sono stati considerati attivi.

Nella tabella che segue riportiamo le strutture 2D dei 25 composti attivi assieme all'articolo di provenienza e all'attività biologica K_i:



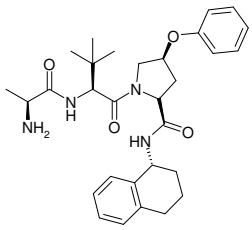
K_i = 0.35 µM ± 0.01

2004 J. Med. Chem. SUN[46]



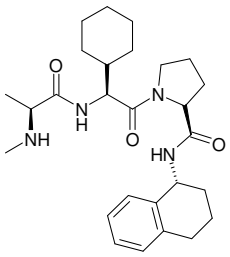
K_i = 0.012 µM

2004 J. Med Chem. Oost[45]



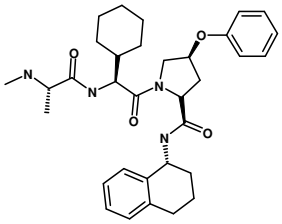
K_i = 0.005 µM

2004 J. Med Chem. Oost[45]



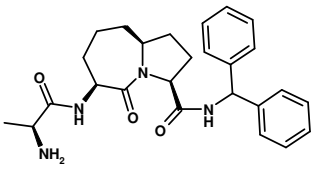
K_i = 0.016 µM

2004 J. Med Chem. Oost[45]



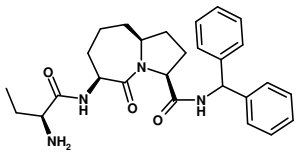
K_i = 0.006 µM

2004 J. Med Chem. Oost[45]



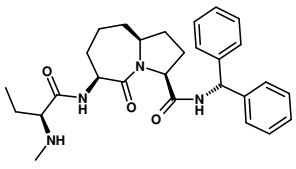
K_i = 0.06 µM ± 0.02

2008 J. Med. Chem. SUN[42]



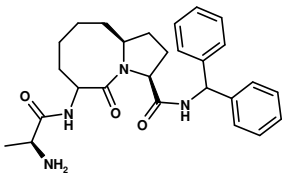
K_i = 0.025 µM ± 0.004

2008 J. Med. Chem. SUN[42]



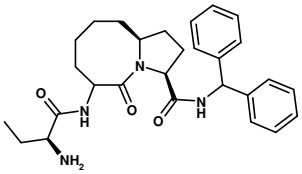
K_i = 0.061 µM ± 0.006

2008 J. Med. Chem. SUN[42]



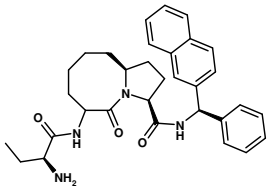
K_i = 0.014 µM ± 0.003

2008 J. Med. Chem. SUN[42]



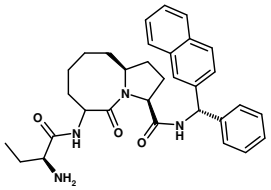
K_i = 0.039 µM ± 0.004

2008 J. Med. Chem. SUN[42]



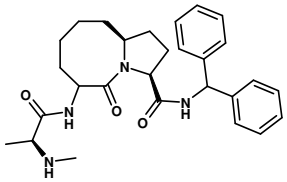
K_i = 0.759 µM ± 0.16

2008 J. Med. Chem. SUN[42]



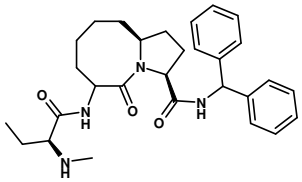
K_i = 0.047 µM ± 0.009

2008 J. Med. Chem. SUN[42]



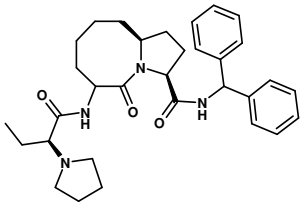
K_i = 0.026 µM ± 0.005

2008 J. Med. Chem. SUN[42]



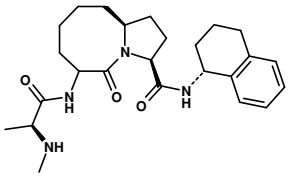
K_i = 0.067 µM ± 0.018

2008 J. Med. Chem. SUN[42]



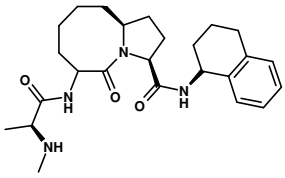
K_i = 0.856 µM ± 0.075

2008 J. Med. Chem. SUN[42]



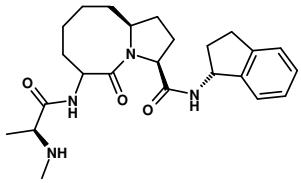
K_i = 0.014 µM ± 0.005

2008 J. Med. Chem. SUN[42]



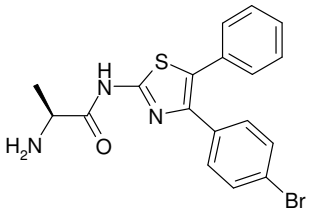
K_i = 0.274 µM ± 0.037

2008 J. Med. Chem. SUN[42]



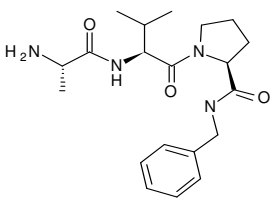
K_i = 0.015 µM ± 0.008

2008 J. Med. Chem. SUN[42]



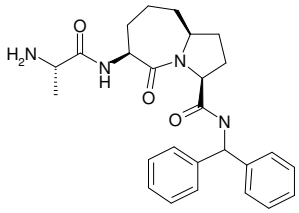
K_i = 0.74 µM

2005 Bio. Med. Chem. Lett. Park[43]



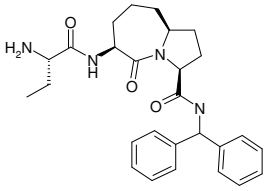
K_i = 0.29 µM ± 0.06

2004 J. Med. Chem. SUN[46]



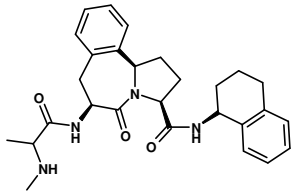
Ki = 0.060 µM ± 0.02

2004 J. Med. Chem. SUN[46]



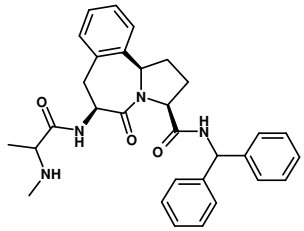
K_i = 0.025 µM ± 0.004

2004 J. Med. Chem. SUN[46]



K_i = 18 nM ± 10

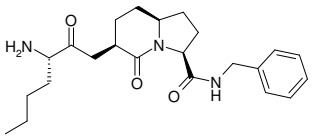
2008 J. Med. Chem. Zhang[48]



K_i = 30 nM ± 4.4

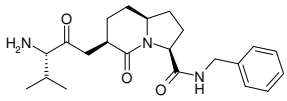
2008 J. Med. Chem. Zhang[48]

Inoltre riportiamo le strutture 2D del data set delle 55 molecole inattive sempre con l'articolo di provenienza e l'attività biologica:



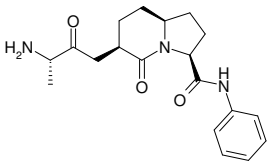
K_i = > 100 µM

2004 J. Med. Chem. SUN[46]



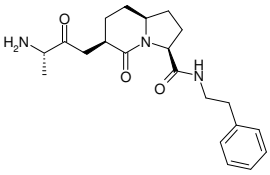
K_i =43.11 µM ± 1.51

2004 J. Med. Chem. SUN[46]



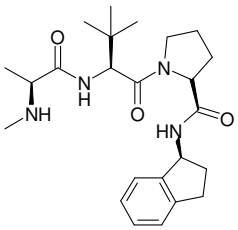
K_i > 100 µM

2004 J. Med. Chem. SUN[46]



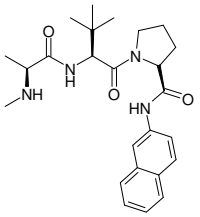
K_i = 22.4 µM ± 1.87

2004 J. Med. Chem. SUN[46]



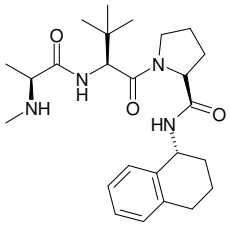
K_i > 1 µM

2004 J. Med Chem. Oost[45]



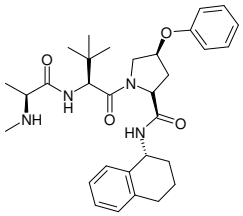
K_i > 1 µM

2004 J. Med Chem. Oost[45]



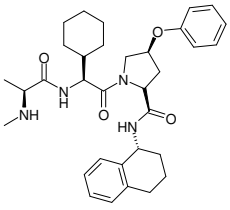
K_i > 1 µM

2004 J. Med Chem. Oost[45]



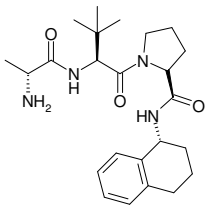
K_i > 1 µM

2004 J. Med Chem. Oost[45]



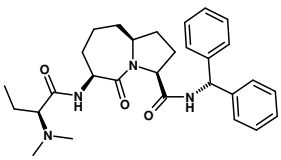
K_i > 1 µM

2004 J. Med Chem. Oost[45]



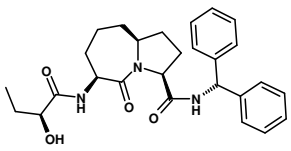
K_i > 1 µM

2004 J. Med Chem. Oost[45]



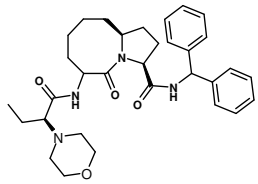
K_i = 14.4 µM ± 0.6

2008 J. Med. Chem. SUN[42]



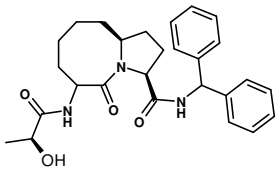
K_i = 29.0 µM ± 1.4

2008 J. Med. Chem. SUN[42]



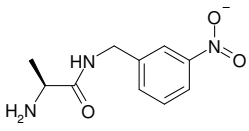
K_i = 32.87 µM ± 4.18

2008 J. Med. Chem. SUN[42]



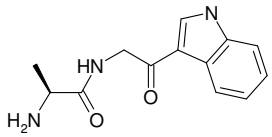
K_i = 12.97 µM ± 2.02

2008 J. Med. Chem. SUN[42]



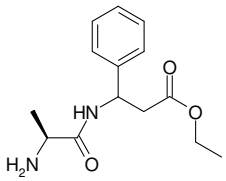
K_i N/T

2005 Bio. Med. Chem. Lett. Park[43]



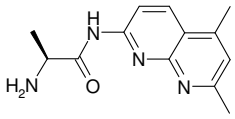
K_i N/T

2005 Bio. Med. Chem. Lett. Park[43]



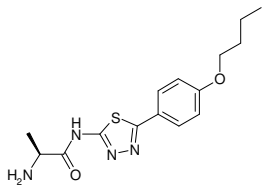
K_i N/T

2005 Bio. Med. Chem. Lett. Park[43]



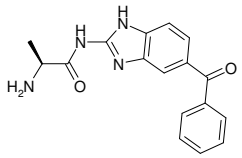
K_i = 20 µM

2005 Bio. Med. Chem. Lett. Park[43]



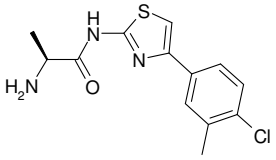
K_i = 25 µM

2005 Bio. Med. Chem. Lett. Park[43]



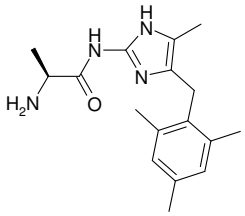
K_i = 30 µM

2005 Bio. Med. Chem. Lett. Park[43]



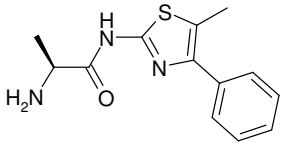
K_i = 30 µM

2005 Bio. Med. Chem. Lett. Park[43]



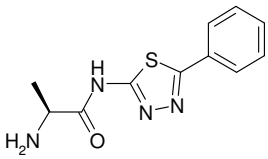
K_i = 2 µM

2005 Bio. Med. Chem. Lett. Park[43]



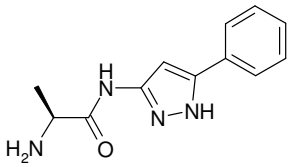
K_i = 15 µM

2005 Bio. Med. Chem. Lett. Park[43]



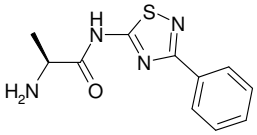
K_i = 161 µM

2005 Bio. Med. Chem. Lett. Park[43]



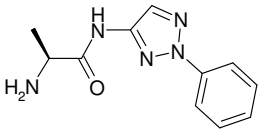
K_i = 170 µM

2005 Bio. Med. Chem. Lett. Park[43]



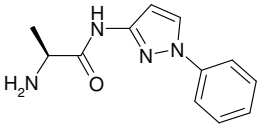
K_i = 84 µM

2005 Bio. Med. Chem. Lett. Park[43]



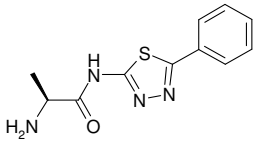
K_i = 40 µM

2005 Bio. Med. Chem. Lett. Park[43]



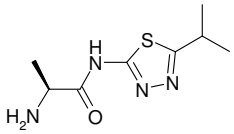
K_i = 84 µM

2005 Bio. Med. Chem. Lett. Park[43]



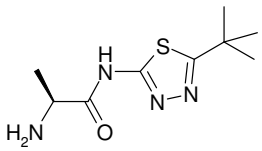
K_i > 2000 µM

2005 Bio. Med. Chem. Lett. Park[43]



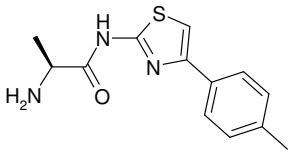
K_i = 63 µM

2005 Bio. Med. Chem. Lett. Park[43]



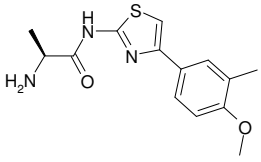
K_i = 21 µM

2005 Bio. Med. Chem. Lett. Park[43]



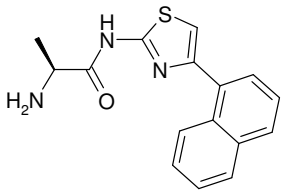
K_i = N/T

2005 Bio. Med. Chem. Lett. Park[43]



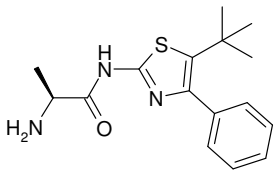
K_i = N/T

2005 Bio. Med. Chem. Lett. Park[43]



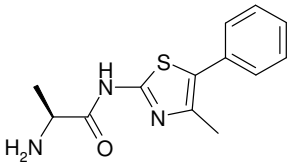
K_i = N/T

2005 Bio. Med. Chem. Lett. Park[43]



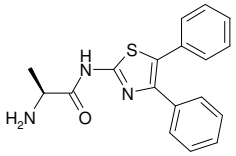
K_i = 4.8 µM

2005 Bio. Med. Chem. Lett. Park[43]



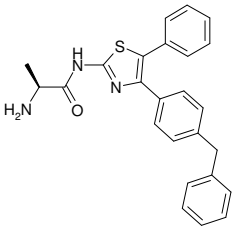
K_i = 15 µM

2005 Bio. Med. Chem. Lett. Park[43]



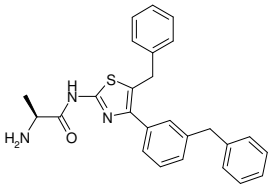
K_i = 1.1 µM

2005 Bio. Med. Chem. Lett. Park[43]



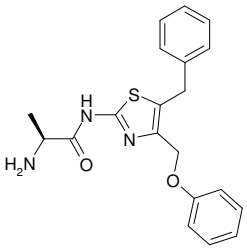
K_i = 4.42 µM

2005 Bio. Med. Chem. Lett. Park[43]



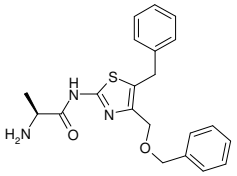
K_i > 100 µM

2005 Bio. Med. Chem. Lett. Park[43]



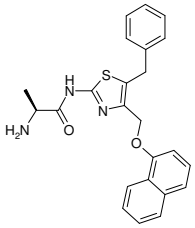
K_i = 4.1 µM

2005 Bio. Med. Chem. Lett. Park[43]



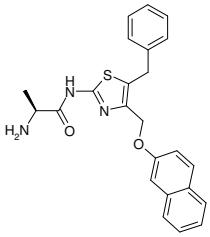
K_i = 4.1 µM

2005 Bio. Med. Chem. Lett. Park[43]



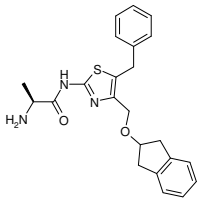
K_i = 21 µM

2005 Bio. Med. Chem. Lett. Park[43]



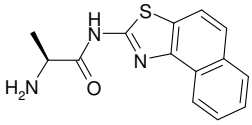
K_i = 14 µM

2005 Bio. Med. Chem. Lett. Park[43]



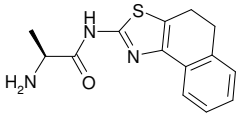
K_i = 22 µM

2005 Bio. Med. Chem. Lett. Park[43]



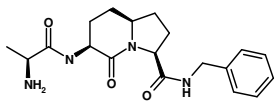
K_i = 7.1 µM

2005 Bio. Med. Chem. Lett. Park[43]



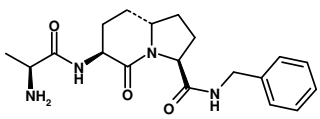
K_i = 5.6 µM

2005 Bio. Med. Chem. Lett. Park[43]



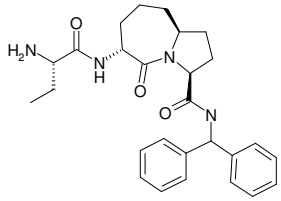
K_i = 4.47 µM ± 0.65

2004 J. Med. Chem. SUN[46]



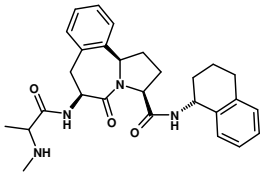
K_i >100 µM

2004 J. Med. Chem. SUN[46]



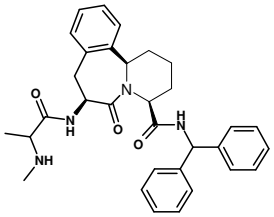
K_i = 16.9 µM ± 0.60

2004 J. Med. Chem. SUN[46]



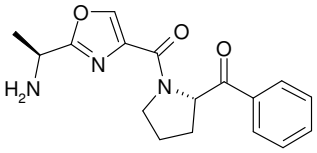
K_i = 1200 nM ± 500

2008 J. Med. Chem. Zhang[48]



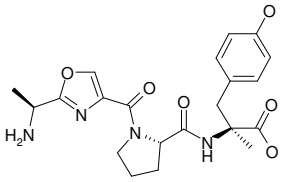
K_i = 690 nM ± 200

2008 J. Med. Chem. Zhang[48]



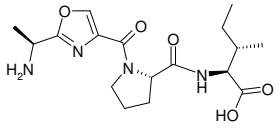
K_i = 11 µM ± 3

2007 Bio. Med. Chem. Wist[50]



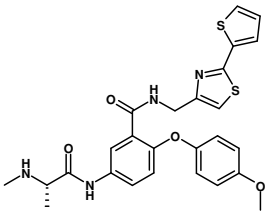
K_i = 120 µM ± 60

2007 Bio. Med. Chem. Wist[50]



K_i = 300 µM ± 60

2007 Bio. Med. Chem. Wist[50]



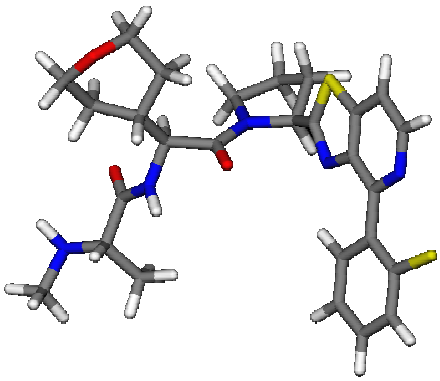
K_i = 2.5 µM

2008 Med. Chem. Huang[48]

Osservando attentamente le strutture del training set delle attive che delle inattive si nota sempre la presenza di un gruppo basico terminale (amminico), prova che la carica positiva in quel gruppo è fondamentale affinché, come detto nel capitolo 3.1.1.1, si instauri l'interazione col dominio BIR3 della proteina XIAP.

3.1.3. Il Virtual Screening

Per l'ottenimento di nuove entità chimiche che siano AVPI mimetiche è stata progettata e realizzata una metodologia di Virtual Screening Ligand Based incentrata interamente sulla ricerca di molecole simili alla molecola PS1[55] (Figura 24) .Tale molecola risulta attiva in più una proteina, la proteina c-IAP1, la c-IAP2 e il nostro target XIAP-Bir3 (Vedi Tabella 6).



a **b**

Figura 24 (**a**) Struttura 2D della molecola PS1; (**b**) Struttura 3D della molecola PS1

Tabella 6 Valori di affinità di binding della PS1 nei 3 recettori in cui è attiva

c-IAP1 BIR3	c-IAP2 BIR3	XIAP BIR3
0.036 µM	0.096 µM	0.033 µM

La procedura è stata effettuata sul database CHEMDIV, costituito da 800.000 molecole, attraverso 3 step:

- Calcolo delle coordinate GPS dell'intero database CHEMDIV e del nostro data set di attive/inattive, come spiegato in precedenza nel capitolo 2.2.1
- Applicazione del criterio di selezione CLAN usando come origine del cluster la molecola PS1 e selezione del set ridotto

- Applicazione del metodo Flap Ligand-Based al set ridotto di molecole e scelta degli hits da testare per via sperimentale

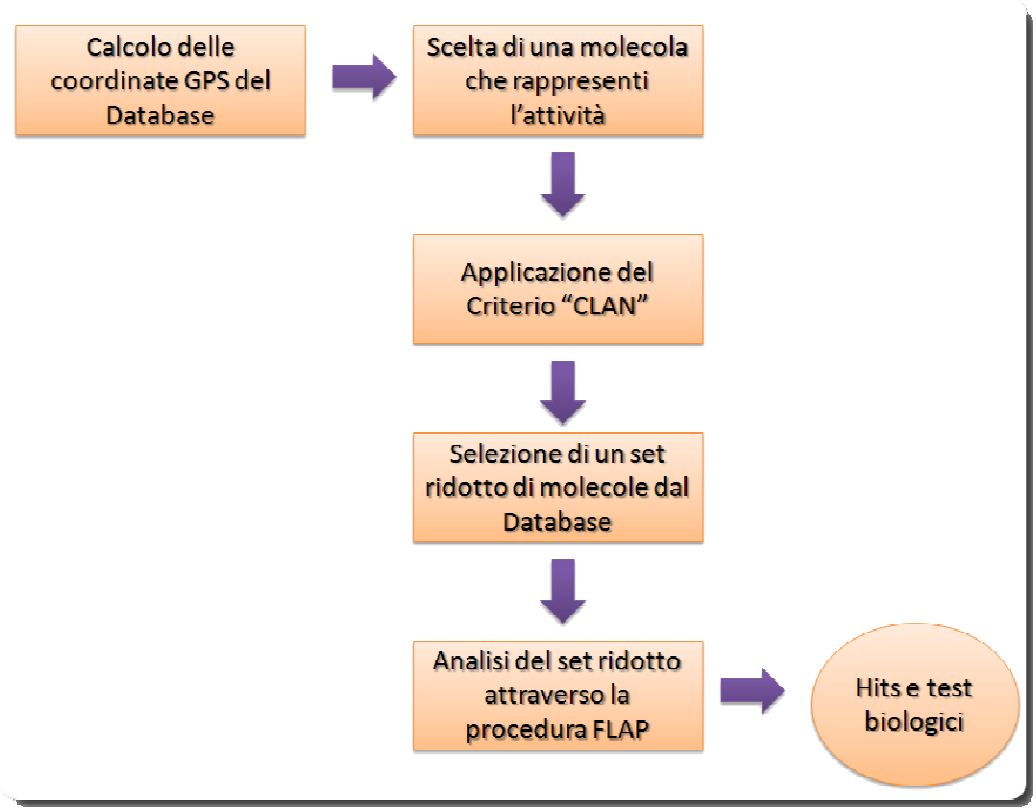


Figura 25 Schema della procedura di Virtual-Screening applicata

Segue adesso un'analisi dettagliata step by step.

Nel primo step le molecole del database CHEMDIV sono state convertite attraverso "mizer" dal formato "SDF", al formato "MOL2", un formato molecolare 3D, in modo da valutare i descrittori molecolari con i metodi GRID e VolSurf+ . Le coordinate GPS sono state calcolate tramite uno script python command-line sviluppato in un precedente lavoro di tesi citato nel capitolo dei metodi. Le proiezioni delle componenti principali GPS del nostro data set, della molecola PS1 e del tetrapeptide AVPI vengono riportate nelle figure 26 e 27:

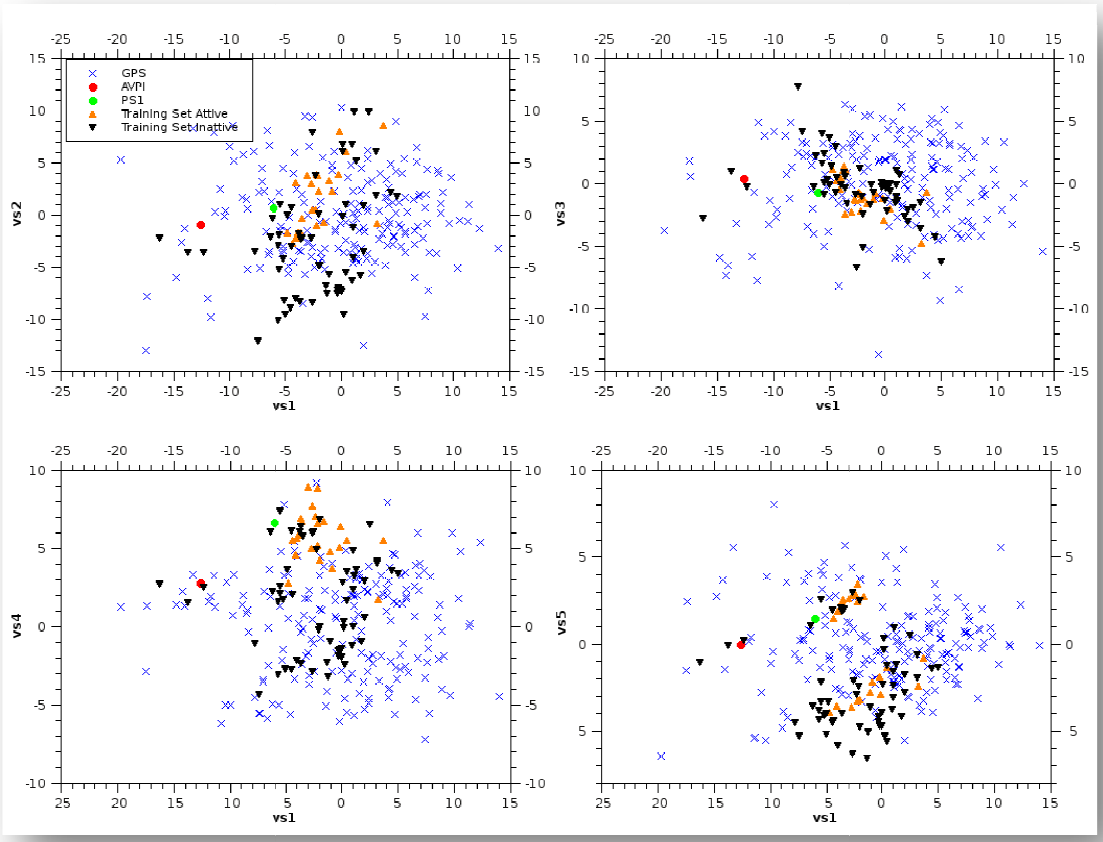


Figura 26 Proiezione delle componenti principali del training set, dell'AVPI e PS1 sullo spazio GPS ottenuto con VolSurf+

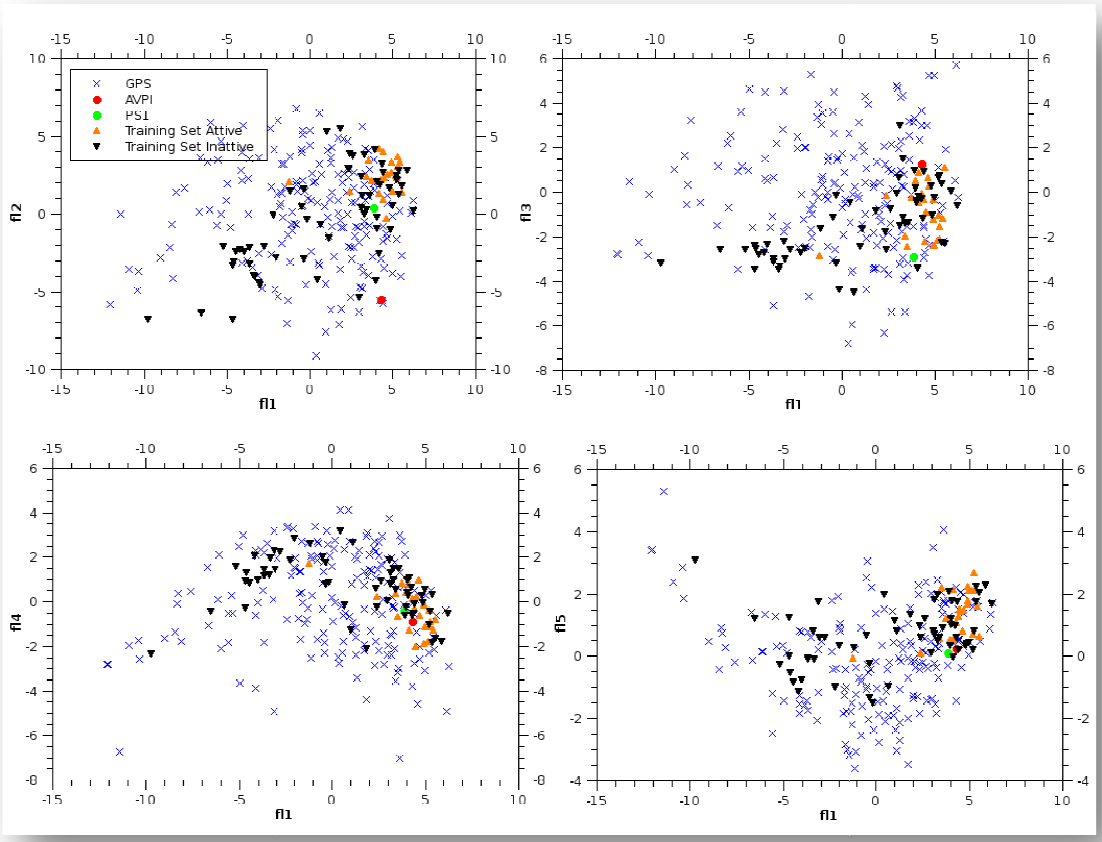


Figura 27 Proiezione delle componenti principali del training set, dell'AVPI e PS1 sullo spazio GPS ottenuto con FLAP

Questi plot evidenziano chiaramente che le molecole del data set attive/inattive sono simili fra loro e simili alla molecola PS1, per questo si raggruppano in una regione definita dello spazio GPS di VolSurf+ e di FLAP.

Da notare che l' AVPI è isolata rispetto a tutte le altre molecole. Questo è dovuto al fatto che, essendo un tetrapeptide, esso ha una struttura abbastanza diversa dalle molecole che compongono il GPS, che sono principalmente molecole drug-like non peptidiche.

Infine i plots non evidenziano le differenze nelle attività dei composti verificando così i limiti dello strumento di analisi PCA e dello spazio GPS.

Nel secondo step è stata applicata la procedura CLAN attraverso gli script python in questo modo:

- Preparazione dei file di input con il tool “cfc”: su ogni file del database da computare vengono introdotte le coordinate GPS della molecola PS1.
- Calcolo delle distanze a 10 dimensioni con “cndist”: per ogni file vengono calcolate distanze a 10 dimensioni usando come origine le coordinate GPS della molecola PS1.
- Analisi delle distanze attraverso le funzioni statistiche abundance e sensitivity con “StaRaCalc”: si analizza la distribuzione delle distanze e la distribuzione delle molecole del training set e del database.
- Grafico delle funzioni statistiche abundance, sensitivity e scelta di un valore di distanza che permetta di estrarre il data set ridotto con “LimFilter” e “MolExtract” (Figura 28)

Inoltre è stato effettuato un’ulteriore studio sulla distribuzione delle distanze dalla molecola PS1 del training set delle attive e inattive (Vedi Figura27).

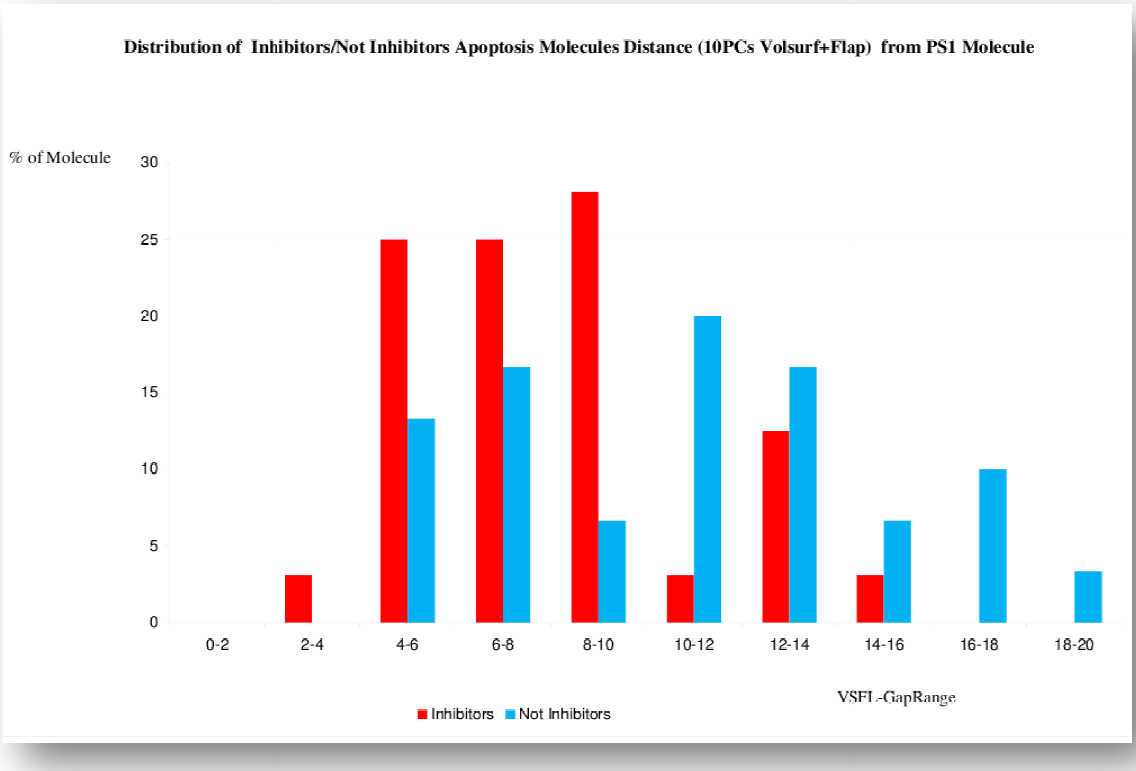


Figura 28 Distribuzione delle molecole in funzione della distanza a partire dalla molecola PS1

È interessante notare che in prossimità della molecola PS1 la distribuzione delle inattive è molto bassa, non ché nulla in un range di distanza circa 5.

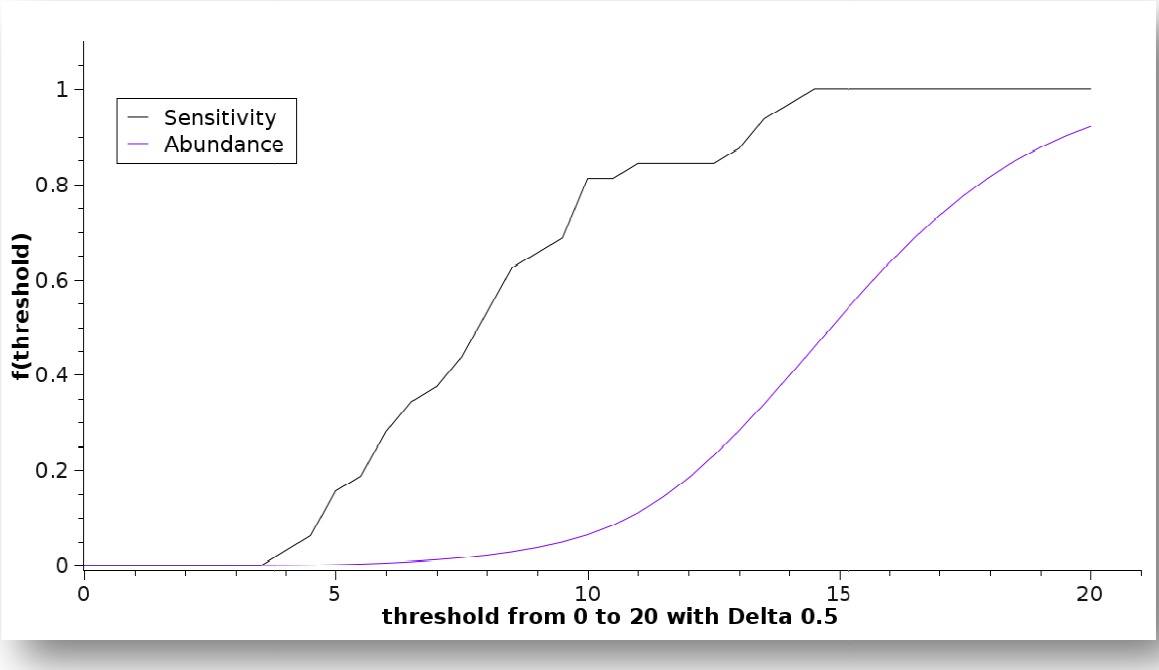


Figura 29 Plot dell'abundance del database ChemDIV e della sensitivity del training set delle attive

Analizzando i grafici di abundance e sensitivity è stato selezionato il 5% del database pari a 38000 molecole su un totale di 800.000. Quindi è stata annotato il valore di distanza corrispondente al 5% del database ed è stata estratta la lista di nomi delle molecole comprese dentro questo range di distanza attraverso i tool "LimFilter" e "MolExtract". Infine queste 38000 molecole sono state analizzate attraverso la procedura FLAP Ligand-Based, attraverso dei confronti di similarità per sovrapposizione alla molecola PS1.

3.1.4. Risultati e Discussioni

Visto il nostro target, ovvero il dominio BIR3 della proteina XIAP, la sovrapposizione delle 38000 molecole con la PS1 è stata fatta passando per uno step importante, ovvero la scelta della conformazione della PS1 stessa.

Per mezzo della struttura ai raggi X della proteina XIAP con la biomolecola antagonista SMAC reperibile presso il database PDB con il codice 1G73, è stato possibile estrarre la conformazione del tetrapeptide AVPI e con la metodologia FLAP la molecola PS1 è stata sovrapposta su di essa, in maniera da poter analizzare l'orientazione di quest'ultima all'interno del dominio BIR3 ed effettuare comparazioni sulle interazioni dei campi molecolari con il dominio BIR3 (Vedi Figura 30).

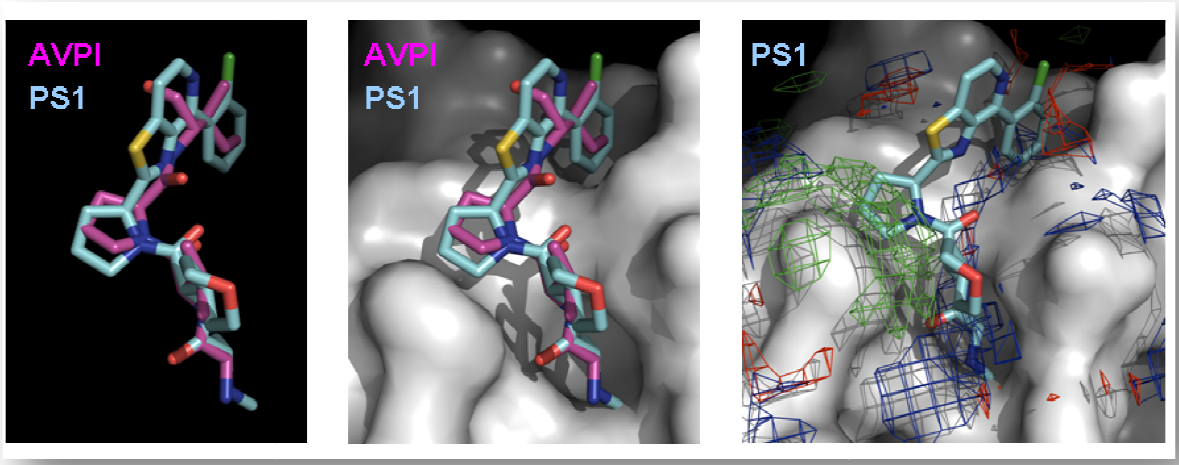
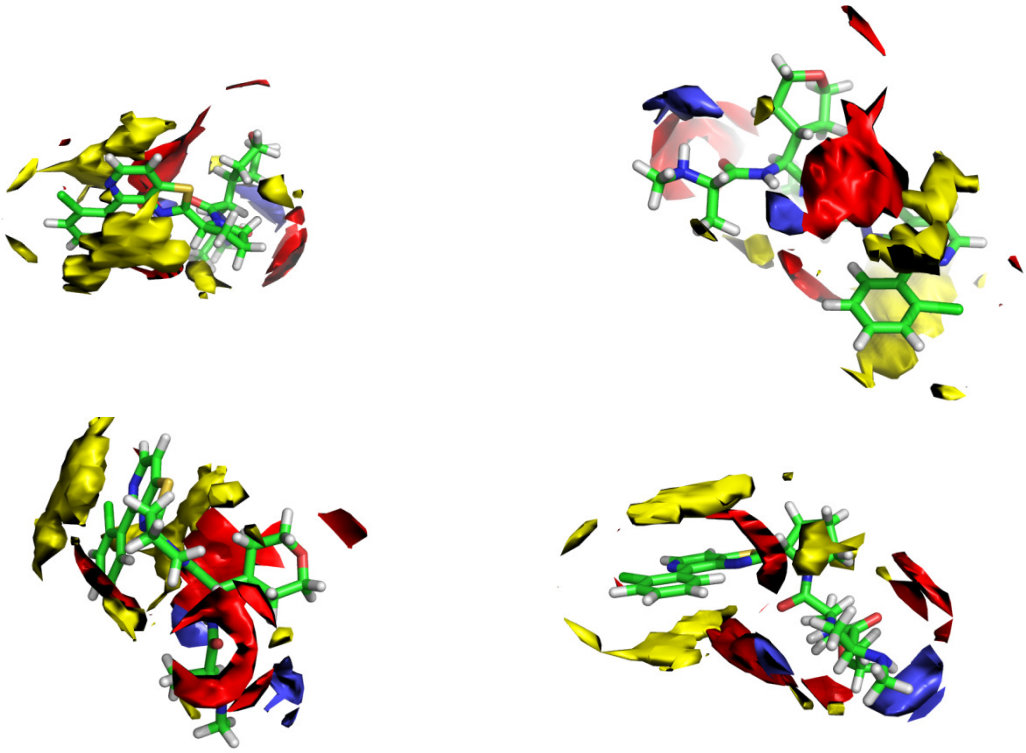


Figura 30 Sovrapposizione della molecola PS1 sul tetrapeptide AVPI e analisi dei campi di interazione molecolare

Dopo aver selezionato quindi la conformazione più adatta che la molecola PS1 dovesse assumere è stata applicata alle 38000 molecole la procedura FLAP Ligand-Based da cui sono state scelte ed acquistate 31 molecole per poi effettuarvi dei test biologici sulla linea cellulare SH-EP neuroblastoma cells. Nell'appendice si riporta una tabella con le strutture in 2D e le sovrapposizioni delle molecole scelte con il ligando PS1. È interessante notare che tutte le molecole sono state disposte da flap sulla molecola PS1 prediligendo specifici

punti farmacoforici affinché si possa avere un'interazione di tipo legante con il sito attivo XIAP-BIR3. Questi sono riportati nella figura seguente:

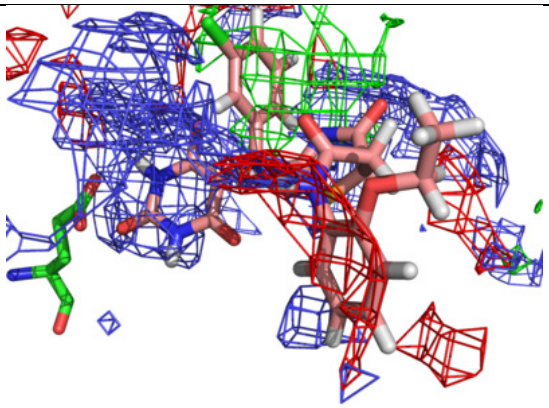


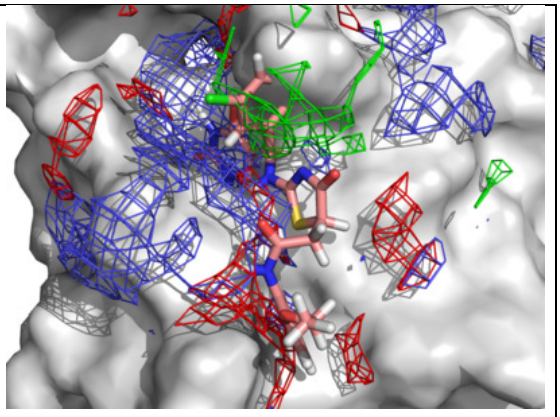
- In giallo i campi DRY: queste due regioni si evidenziano le interazioni.
- In blu i campi HBD: queste due regioni si evidenziano interazioni di legame a idrogeno come donatore. È fondamentale importanza l'N-H

marcato con il numero 1 perché responsabile dell'interazione con il carbossile dell'amminoacido GLU 314 del dominio BIR3.

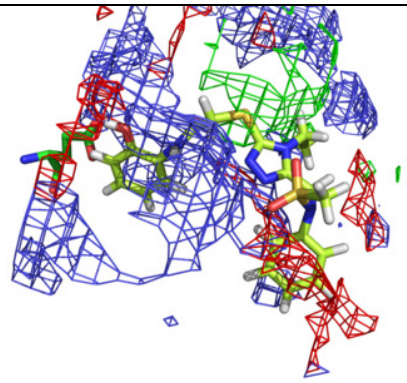
- In rosso i campi HBA: in queste tre regioni si evidenziano le interazioni di legame a idrogeno come accettore.

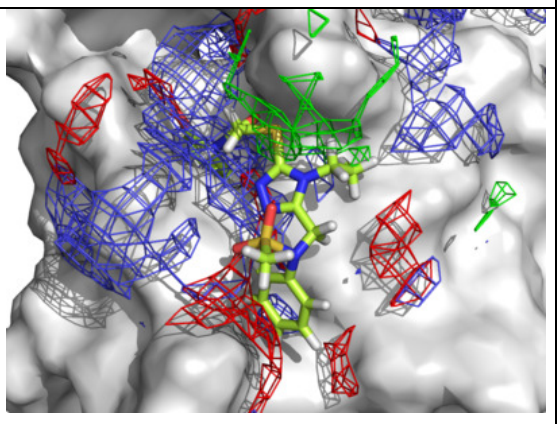
Infine è stata effettuata una verifica sulla predizione dei 31 composti attraverso l'analisi delle conformazioni predette all'interno della cavità BIR3 ed esaminare che le regioni DRY, HBD e HBA si trovino correttamente sovrapposte ai campi di interazione della proteina. Di seguito si riportano 4 risultati di queste analisi:



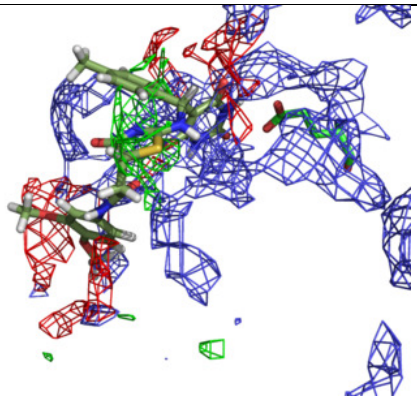


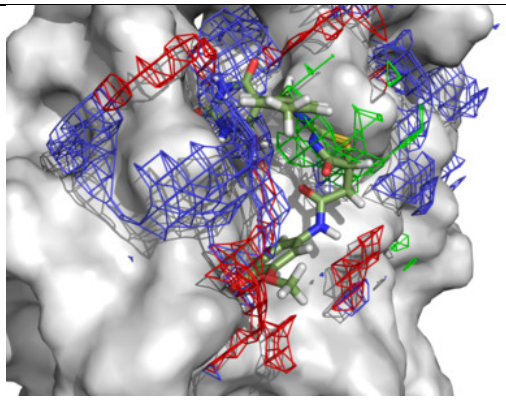
C202-2081



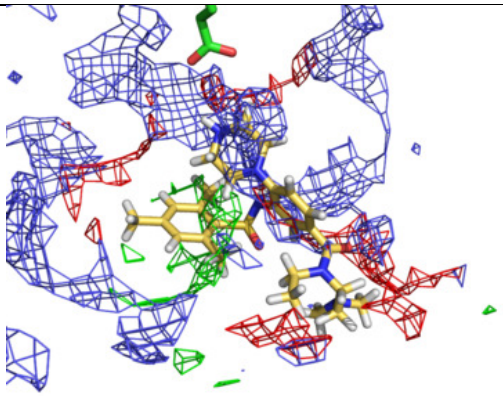


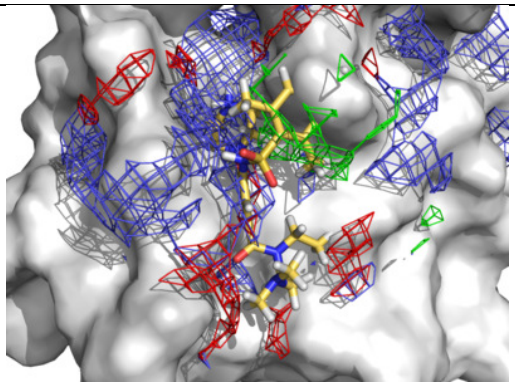
D585-0150





E755-0025





F919-0613

In rosso si evidenziano i campi accettori di legame a idrogeno, in blu i campi donatori di legame a idrogeno e in verde i campi idrofobici. Dalle immagini si evince che quasi tutti i campi del sito attivo BIR3 sono sovrapposti alle corrispondenti regioni accettori/donatori di legame a idrogeno e regioni idrofobiche delle molecole oggetto dello studio.

4. Conclusioni

In questo lavoro di tesi sono stati sviluppati dei nuovi approcci di Virtual-Screening che permettono velocizzare l'analisi dei dati andando ad esaminare un set ridotto di molecole che presentano proprietà simili alle molecole targets. Dopo aver testato la procedura in sets di letteratura, il metodo è stato applicato per la ricerca di molecole pro-apoptotiche. In seguito all'attenta valutazione di un cospicuo numero di candidati, sono state scelte 31 molecole per test sperimentali, condotti dalla Prof. Simone Fulda del Children's Hospital Ulm University. Purtroppo, vista la complessità delle procedure sperimentali, i risultati dei test in vitro sulle linee cellulari tumorali non sono ancora disponibili.

Tuttavia, si può comunque affermare che è stato raggiunto con successo l'obbiettivo prestabilito, ovvero l'ottimizzazione dei tempi di calcolo attraverso la messa a punto di metodologie che permettono di orientarsi in maniera veloce ed intelligente nello spazio chimico GPS, cioè permettono di selezionare ed analizzare un set ridotto di molecole.

Da una stima sui tempi di calcolo utilizzando i parametri di default di FLAP (Tabella 7 e Figura 30) è emerso che per l'intero dabatase ChemDiv (800.000 molecole) con la procedura FLAP su di un computer con processore Dual Core frequenza di Clock 3.00 GHz sarebbero stati necessari 1 anno e tre mesi, mentre con la procedura sviluppata sono bastati circa 2 mesi, guadagnando circa il 90% del tempo. Si ricorda ovviamente che queste stime sono fortemente dipendenti dal tipo di computer in uso, dal tipo di molecole che si stanno analizzando e dai parametri che sono stati impostati per effettuare il virtual screening.

Tuttavia, molte cose restano da ottimizzare per quanto riguarda l'uso dello spazio chimico "*pre-calcolato*" per ricerca virtuale, e ci si propone in futuro di ottimizzare le metodologie sviluppate e di effettuare altre applicazioni.

Tabella 7: Dati dei tempi di calcolo necessari a FLAP per computare delle molecole scelte in maniera random dal database ChemDiv, attraverso la modalità Ligand-Based con templato la molecola PS1, su di un computer Intel Pentium 4 Dual Core 3.00 GHz

Numero Molecole	Tempo (secondi)
2	63
4	125
6	206
8	271
11	362
20	595
50	2.596
250	13.592
400	7.920
500	28.153

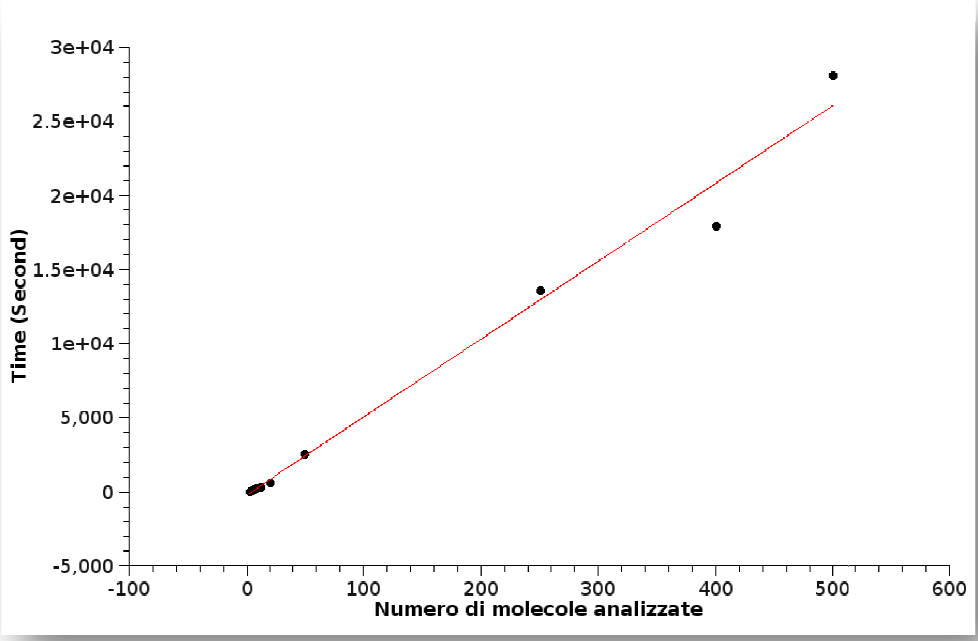


Figura 31 Dalla distribuzione dei dati viene evidenziato un andamento lineare della crescita del tempo computazionale, quindi è stato applicato un fit lineare dei dati ottenendo i seguenti valori di coefficiente angolare ed intercetta all'asse del tempo: **intercetta = $-1.97 \cdot 10^{2} \pm 4.98 \cdot 10^{1}$; Coefficiente Angolare = $5.26 \cdot 10^{1} \pm 2.28$**

5. Algoritmi e software

5.1. Cmapt

All'interno del CD troverete il pacchetto .tar.bz2 che contiene il sorgente del software compilabile su qualsiasi sistema UNIX Like che soddisfi le dipendenze: cmake, glib2 e plplot. Inoltre sono disponibili i pacchetti binari relativi alle distribuzioni Linux Ubuntu, Debian e Fedora.

Cmapt è distribuito sotto licenza BSD.

5.1.1. Utilizzo

Cmapt utilizza come file di input un file ASCII contenente le coordinate GPS a 10 dimensioni. Di seguito portiamo un esempio della struttura di questo file:

Molecule	GPS-VS1	VS2	VS3	VS4	VS5	GPS-FL1	FL2	FL3	FL4	Fl5
JNJ-7777120	3.79	-7.37	-1.12	0.52	-3.42	-7.46	-2.67	-1.22	-0.20	-0.75
dimaprit	-4.23	-15.14	8.50	0.73	-1.91	-9.67	-5.84	-2.43	-1.40	1.93
histamine	-5.16	-18.90	5.57	-4.01	-5.04	-11.07	-6.38	-3.14	-2.54	3.19

Figura 32 Esempio dell'input ASCII File di cmapt-gps, ctestmm e CLAN

Ogni riga del file è costituita da 11 colonne. La prima colonna rappresenta il nome della molecola, mentre le successive colonne rappresentano le 10 componenti principali ottenute dal GPS. Generalmente la prima riga rappresenta l' intestazione che definisce il nome delle colonne e questa viene ignorata dal programma.

Quindi a partire da un file con estensione .txt (file.txt) che presenta questa struttura, è possibile applicare il programma come segue in questo esempio:

 # cmapt-gps -i file.txt –o output.txt –v

dove l' opzione "-v" applica le opzioni di default:

- Algoritmo applicato sulle prime due componenti di VolSurf+ VS1 e VS2
- Limiti dello spazio di analisi definiti dai valori di massimo e di minimo delle componenti principali ottenute dalle strutture core del GPS (capitolo 2.2).
- Costruzione di una griglia costituita da 13 righe e 13 colonne
- Generazione di un grafico come immagine png con il nome di vs1-vs1.png

Inoltre possibile personalizzare tutte questi valori di default come meglio si vuole. Per esempio per costruire una griglia costituita da 10 righe e 10 colonne basta aggiungere l'opzione "-r 10", oppure per costruire una griglia di 10 righe e 14 colonne basta aggiungere le opzioni "-q 14 –k 10", o ancora definire i limiti dello spazio GPS attraverso un file ASCII "conf.txt" che presenta la seguente struttura:

xmin,	xmax
ymin,	ymax

e quindi passarlo al programma con l'opzione "-c conf.txt".

Per qualsiasi altra informazione digitare da linea di comando il nome del programma "cmapt-gps" ed esso restituirà un output a video con tutte le opzioni disponibili.

Il file di output di cmapt-gps è un file ASCII dove sono memorizzate le coordinate dell'origine degli assi cartesiani delle componenti principali, il numero di righe e colonne , e i valori di score per ogni cella della griglia.

Per proiettare delle molecole esterne dentro le mappe ottenute con cmapt-gps occorre utilizzare il file di output della mappa (output.txt era il nome del precedente esempio) e calcolare le coordinate GPS di queste molecole

ottenendo un file di input come riportato nella figura 32 ed utilizzare il programma ctestmm come nell' esempio che segue:

 # ctestmm -i file.txt -v output.txt -o output.txt -g -d pngcairo -n immagine.png

dove:

- "-v" rappresenta l'opzione per passare la mappa relativa alle prime due componenti di VolSurf+
- "-o" per scrivere un file di output ASCII contenente il nome delle molecole , lo score che queste prendono e il numero di riga e colonna dove queste molecole si posizionano.
- "-g" per generare un grafico 2D che rappresenta la mappa con il numero di molecole presenti per ogni cella e quindi "-d pngcairo" per definire il tipo di file dell'immagine del grafico e "-n immagine.png" per dare un nome al grafico.

Per qualsiasi altra informazione riguardo l'utilizzo di questo programma basta digitare da linea di comando "ctestmm" che darà come output a video le opzioni disponibili.

5.2. CLAN

CLAN è disponibile in versione command-line ed è costituito da una suite di script python riportati in un archivio tar.bz2 all'interno di questo cd:

- cfc: preparazione dei file di input per i vari tool.
- cndist: calcolo della distanza n-dimensionale a partire dalle coordinate GPS delle molecole
- bdist: filtro delle migliori distanze
- StaRaCalc: analisi statistica dei dati quale abundance, sensitivity etc…
- LimFilter e MolExtract: tool di analisi dei risultati statistici ottenuti tramite StaRaCalc

Clan è supportato su qualsiasi sistema UNIX Like provvisto dell' interprete python>=2.5 ed è distribuito sotto licenza BSD. Per l'installazione degli script basta copiarli manualmente all'interno della directory /usr/local/bin oppure eseguirli all'interno della directory stessa anteponendo il comando "./".

5.2.1. Utilizzo

CLAN utilizza lo stesso tipo di file input di Cmapt come riportato in figura 32. A differenza di Cmapt, CLAN è costituito da una suite di python script che si interagiscono fra di loro per effettuare un'analisi di clustering. Passiamo adesso ad un'esempio pratico dell'utilizzo di questa suite di script.

Partendo da un database del quale sono state calcolate le coordinate gps (molecule-gps-file.txt), se per esempio avrò 10 molecole otterrò 10 file di output la cui prima riga di ogni file corrisponde ad una molecola appartenente al file template-gps-file.txt che quindi verrà utilizzata come origine. Successivamente questi file potranno essere utilizzati i tool "cndist" per calcolare le distanze euclidee a 10 dimensioni, "bdist" per scegliere le migliori distanze ed operare con il primo metodo dell'algoritmo CLAN, "StaRaCalc" e "StaRaCalc-gps" per effettuare l'analisi statistica delle distanze. Di seguito si riporta un'esempio dell'utilizzo di questi programmi per le due metodologie CLAN:

- Prima metodologia

N.B.: in questa metodologia viene scelta solo una molecola come origine del cluster.

 # cfc template-gps-file.txt molecule-gps-file.txt gps-molecule-from_template.txt

 # cndist txt gps-molecule-from_template.txt distance-from_template.txt*

 # StaRaCalc-gps -d distance-from_template.txt -o abundance.txt –t 0-40-1 –b

- Seconda metodologia

 # cfc template-gps-file.txt molecule-gps-file.txt gps-molecule-from_template.txt

 # for i in gps-molecule_from-template.txt; do sufx=`echo ${i%%.*} | cut –d "_" –f "2"`;cndist ${i} dist-molecule-from_${sufx}.txt; done*

 # bdist dist-molecule-from_*.txt best-distance-molecule.txt 3

 # StaRaCalc -d best-distance-molecule.txt -o abundance.txt -t 0-40-1 -b -c 1

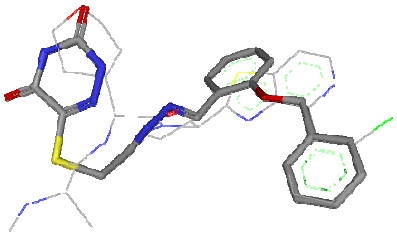
In entrambe le metodologie si otterrà un file ASCII di output (abundance.txt) costituito da due colonne dove la prima colonna rappresenta la distanza che và da 0 a 40, mentre l'ordinata rappresenta il valore della funzione abundance e che questo file potrà essere graficato attraverso qualsiasi programma che permetta di importare e graficare file ASCII.

Licenza BSD

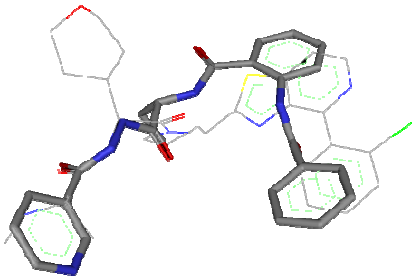
MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL <copyright holder> BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Appendice

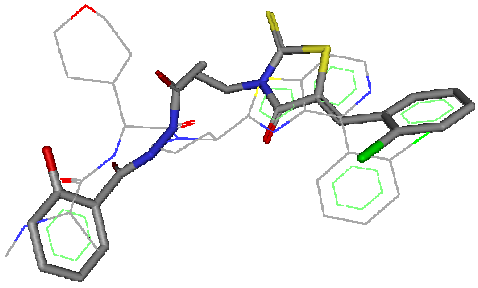
Le 31 molecole ottenute dalla procedura di Virtual Screening



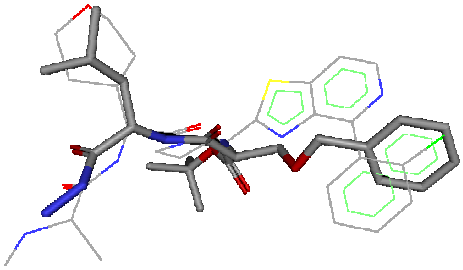
2279-4360



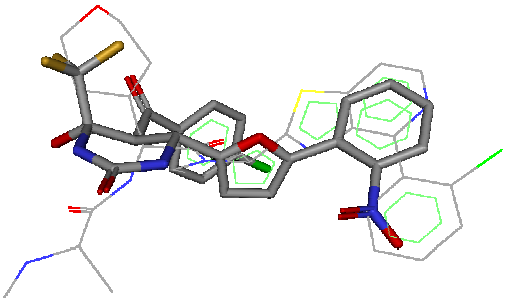
4072-3079



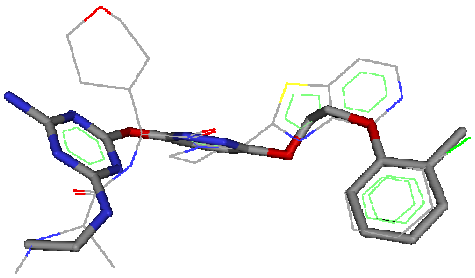
4358-5605



4464-0969

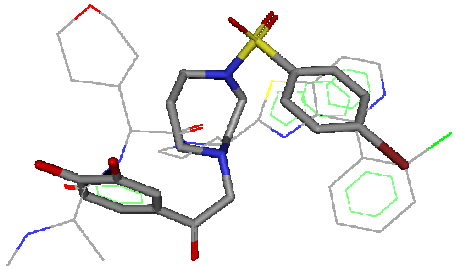


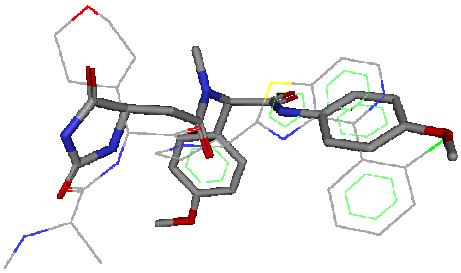
4618-0155



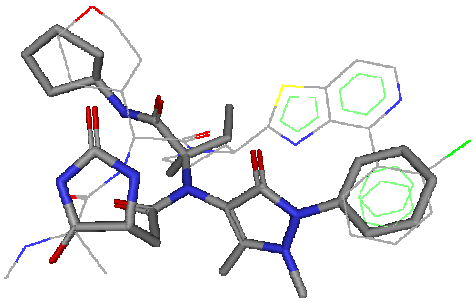
8017-4052

C036-0014



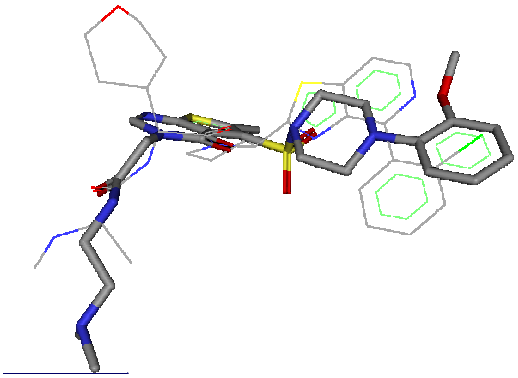


C094-0329

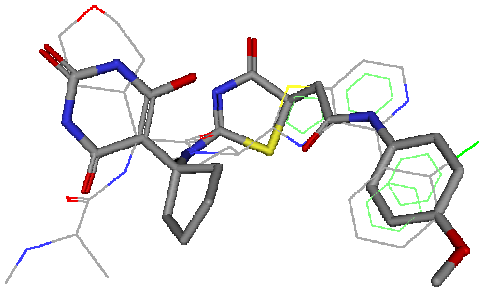


C094-1625

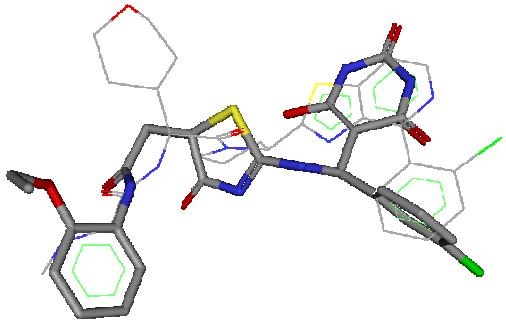
C096-0070



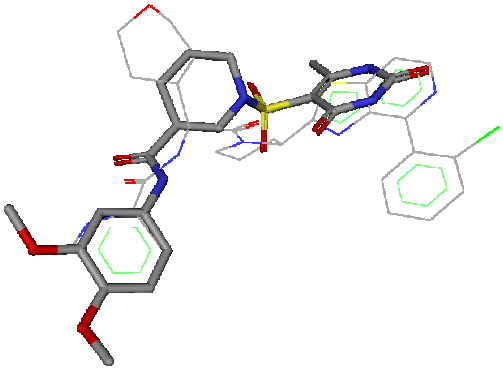
C202-1919



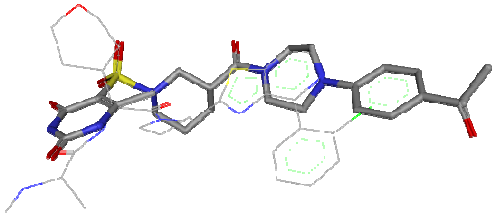
C202-2081



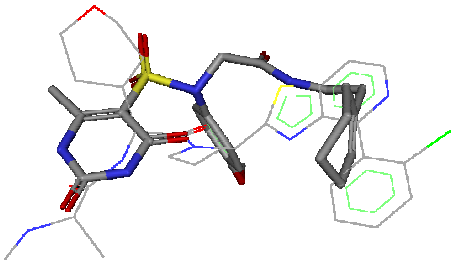
C274-7965



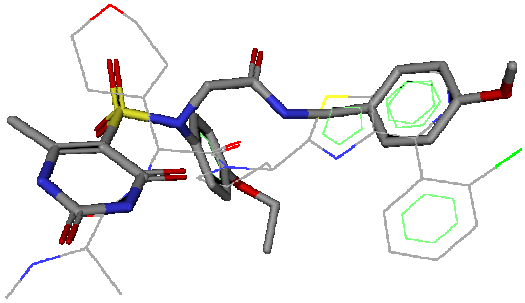
C274-8098



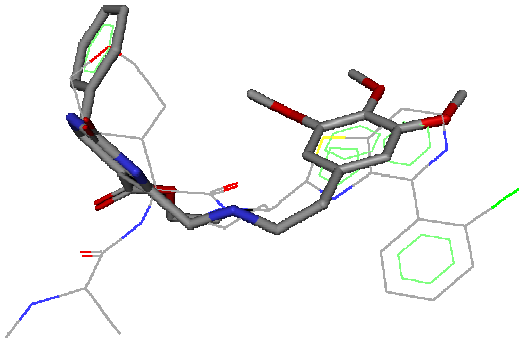
C548-0300



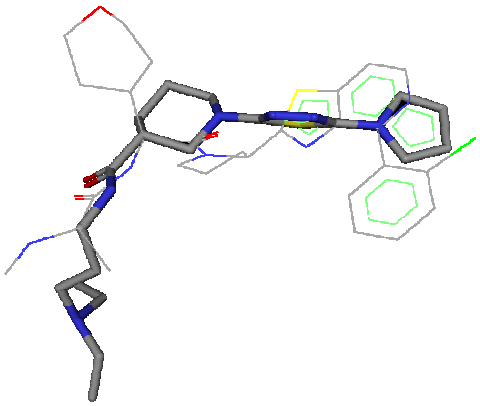
C848-0809



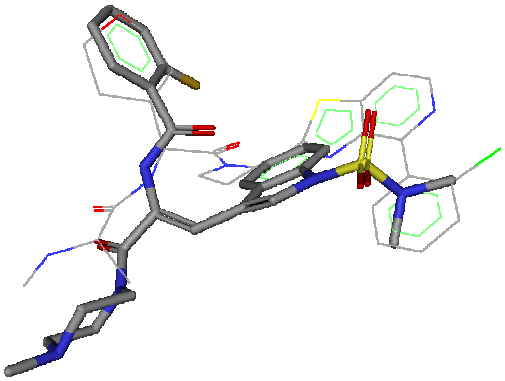
C561-2455



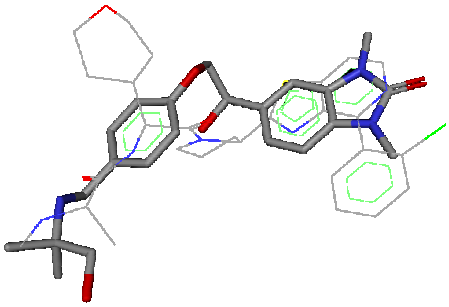
C587-0331



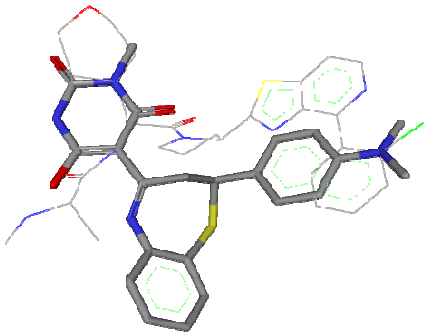
D009-0044



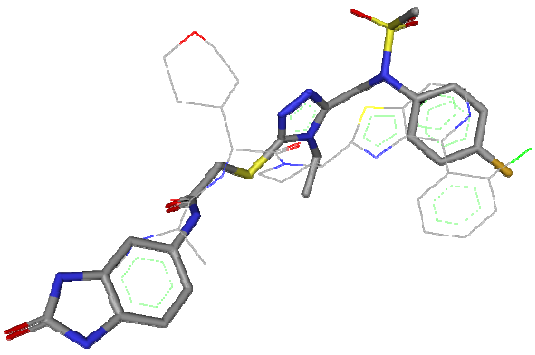
D300-0274



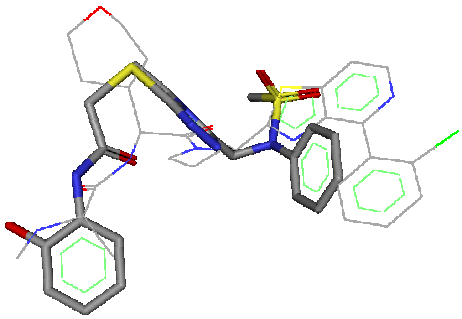
D364-2459



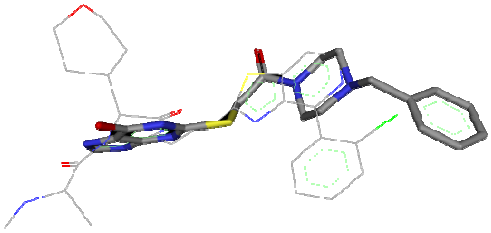
D532-0341



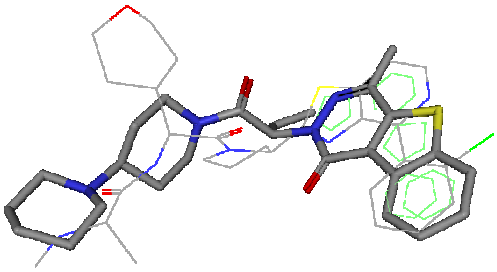
D583-0763



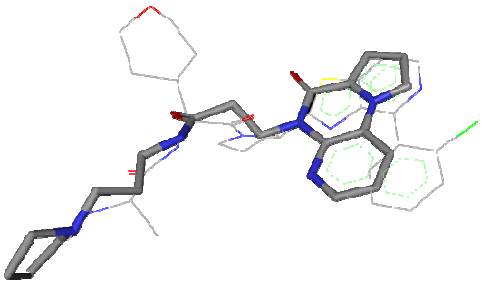
D585-0150



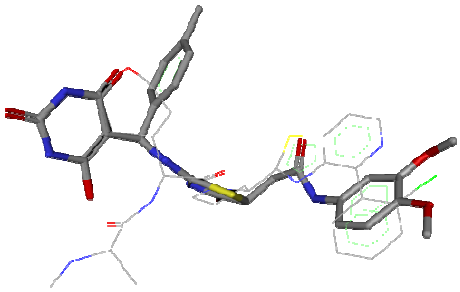
E019-0619



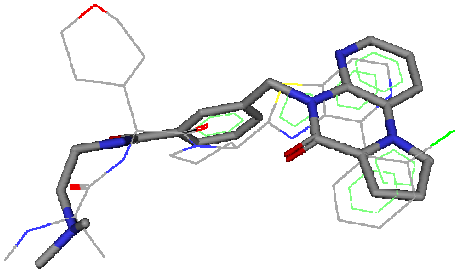
E524-1124



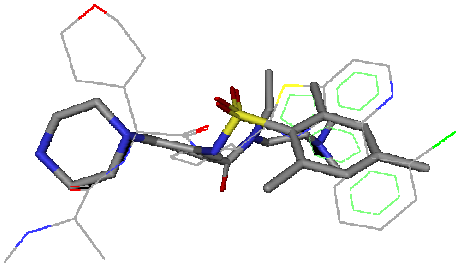
E755-0025



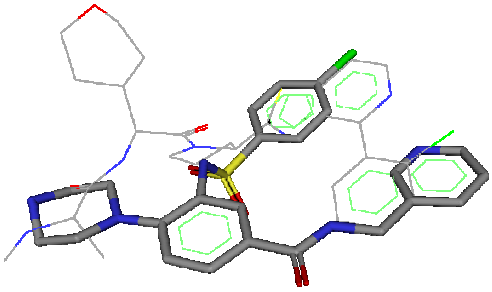
F752-0032



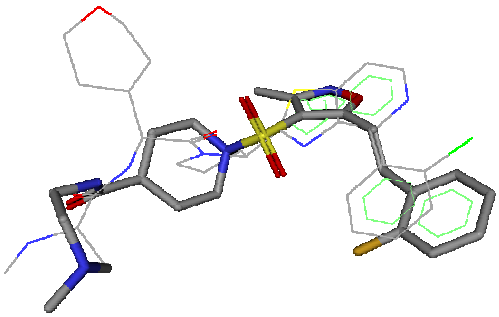
F919-0613



F919-0795



G637-0899



Bibliografia

[1] Werner J. G., Kevin E. G., Watson M. ; Elsevier (February **2006**) Vol. 11, No. ¾

[2] Robert P Hertzberg and Andrew J Pope *; High-throughput screening: new technology for the 21st century*; Current Opinion in Chemical Biology **2000**, Vol. 4, Pag. 445–451

[3] Brian K. Shoichet; *Virtual screening of chemical libraries*; Nature 16 DECEMBER **2004**, VOL 432, Pag 862-865

[4] Hongmao Sun; *Pharmacophore-Based Virtual Screening;* Current Medicinal Chemistry, **2008**, Vol. *15, Pag*. 1018-1024

[5] Walters, W. P., Stahl, M. T., Murcko M. A.; *Virtual screening – an overview*.; DDT Rev. **1998**, Vol. 3, No. 4, 160-178

[6] Gerhard Klebe; Recent *developments in structure-based drug design*; J. Mol. Med. **2000**, Vol. 78, Pag.269–281

[7] Sutapa Ghosh, Aihua Nie, Jing An and Ziwei Huang; Structure-based virtual screening of chemical libraries for drug discovery; *Current Opinion in Chemical Biology*, **2006**, Vol. 10 Pag. 194–202

[8] Douglas B. Kitchen,Hélène Decornez, John R. Furr and Jürgen Bajorath; *DOCKING AND SCORING IN VIRTUAL SCREENING FOR DRUG DISCOVERY: METHODS AND APPLICATIONS*; Nature Revievs: Drug Discovery, November **2004** , Vol. 3, Pag. 935-449

[9] Cyrus Chothial and Arthur M.Lesk2; *The relation between the divergence of sequence and structure in Proteins*; The EMBO Journal, **1986**, vol.5 no.4 pp.823-826

[10] Brooijmans N, Kuntz ID; *Molecular recognition and docking algorithms*.; Annu Rev Biophys Biomol Struct **2003**, Vol. 32, Pag. 335-373.

[11] Halperin I, Ma B, Wolfson H, Nussinov R; Principles of docking: an overview of search algorithms and a guide to scoring functions; Proteins, **2002**, Vol. 47, Pag. 409-443.

[12] Wang R, Lu Y, Wang S; *Comparative evaluation of 11 scoring functions for molecular docking*.; J Med Chem. , **2003**, Vol. 46, Pag. 2287-2303.

[13] MOE; Chemical Computing Group Inc: Montreal, Quebec, Canada **2005**

[14] Dror O., Shulman-Peleg A., Nussinov R., Wolfson H.; J. *Curr. Med. Chem.*, **2004**, Vol. *11*, Pag. 71.

[15] Biogen Inc. (P. Sprague, Z. Zheng, S.P. Adams, A. Castro, J. Singh, H. Van Vlijmen), Molecular model for VLA-4 inhibitors US6552216, **2003**

[16] Catherine Michaux a, Jean-Michel Dogne´, Ste´phanie Rolin, Bernard Masereel, Johan Wouters a, Franc¸ois Durant; A pharmacophore model for sulphonyl-urea (-cyanoguanidine) compounds with dual action, thromboxane receptor antagonists and thromboxane synthase inhibitors; European Journal of Medicinal Chemistry, **2003** Vol 38, Pag. 703_ 710

[17] Akinori Hirashima, Masako Morimoto, Eiichi Kuwano, Eiji Taniguchi, Morifusa Eto; *Three-dimensional common-feature hypotheses for octopamine agonist arylethanolamines*; Journal of Molecular Graphics and Modelling **2002**, Vol. 21, Pag. 81–87

[18] P. J. Goodford; A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules; J. Med. Chem., **1985** Vol. 28, Pag. 849-857

[19] Cruciani, G., Crivori P.; Molecular fields in quantitative structure permeation relationships: the Volsurf approach; J.Mol Struct **2000**,Vol. THEOCHEM 503, Pag. 17-30

[20] G. Cruiciani; Use of MIF-based VolSurf descriptors in Physicochemical and Pharmacokinetic Studies in Molecular Interaction Fields, **2005**, Wiley, Pag. 173-196

[21] Cruciani G., Pastor M., Guba W.; *Volsurf: a new tool for the pharmacokinetic optimization of lead compounds*; Eur. J. Pharm. Sci. **2000**, Vol. 1, Pag. S29-S39

[22] Baroni M., Cruciani G.; Common Reference Framework for analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application; J. Chem. Inf. Model. , **2007**, Vol. 47 (2), Pag. 279-294

[23] Todeschini, R.; *Introduzione alla chemiometria*. EdiSES, **2003**.

[24] Lipinski C.A., Lombardo F., Dominy B.W., Feeny P.J.;. Adv Drug Delivery Rev, **1997**, Vol. 23, Pag. 3

[25] Tudor I. Oprea and Johan Gottfries*; Chemography: The Art of Navigating in Chemical Space*; J. Comb. Chem*.* **2001**, Vol. 3, Pag. 157-166

[27] Geladi, Paul; Kowalski, Bruce; *Partial Least Squares Regression:A Tutorial*; Analytica Chimica Acta **1986**, Vol. 185 Pag. 1–17

[28] Róbert Kiss, Béla Kiss, Árpád Könczöl, Ferenc Szalai, Ivett Jelinek, Valéria László, Béla Noszál, András Falus, and György M. Keseru; *Discovery of Novel Human Histamine H4 Receptor Ligands by Large-Scale Structure-Based Virtual Screening*; J. Med. Chem. **2008**, Vol. 51, Pag. 3145–3153

[29] Herman D. Lim, Rogier A. Smits, Remko A. Bakker, Cindy M. E. van Dam, Iwan J. P. de Esch, and Rob Leurs; *Discovery of S-(2-Guanidylethyl)-isothiourea (VUF 8430) as a Potent Nonimidazole Histamine H4 Receptor Agonist*; J. Med. Chem. **2006**, Vol. 49, Pag. 6650-6651

[30] Marlon D. Cowart, Robert J. Altenbach, Huaqing Liu, Gin C. Hsieh, Irene Drizin, Ivan Milicic, Thomas R. Miller, David G. Witte, Neil Wishart, Shannon R. Fix-Stenzel, Michael J. McPherson, Ronald M. Adair, Jill M. Wetter, Brian M. Bettencourt, Kennan C. Marsh, James P. Sullivan, Prisca Honore, Timothy A. Esbenshade, and Jorge D. Brioni; *Rotationally Constrained 2,4-Diamino-5,6-disubstituted Pyrimidines: A New Class of Histamine H4 Receptor Antagonists with Improved Druglikeness and in Vivo Efficacy in Pain and Inflammation Models*; J. Med. Chem. **2008**, Vol. 51, Pag. 6547–6557

[31] Huaqing Liu, Robert J. Altenbach, Tracy L. Carr, Prasant Chandran, Gin C. Hsieh, La Geisha R. Lewis, Arlene M. Manelli, Ivan Milicic, Kennan C. Marsh, Thomas R. Miller, Marina I. Strakhova, Timothy A. Vortherms, Brian D. Wakefield, Jill M. Wetter, David G. Witte, Prisca Honore, Timothy A. Esbenshade, Jorge D. Brioni, and Marlon D. Cowart; *cis-4-(Piperazin-1-yl)-5,6,7a,8,9,10,11,11a-octahydrobenzofuro[2,3-h]quinazolin-2-amine (A-987306), A New Histamine H4R Antagonist that Blocks Pain Responses against Carrageenan-Induced Hyperalgesia*; J. Med. Chem. **2008**, Vol. 51, Pag. 7094–7098

[32] Róbert Kiss, Béla Kiss, Árpád Könczöl, Ferenc Szalai, Ivett Jelinek, Valéria László, Béla Noszál, András Falus, and György M. Keseru; *Discovery of Novel Human Histamine H4 Receptor Ligands by Large-Scale Structure-Based Virtual Screening*; J. Med. Chem. **2008**, Vol. 51, Pag. 3145–3153

[33] Herman D. Lim, Rogier A. Smits, Remko A. Bakker, Cindy M. E. van Dam, Iwan J. P. de Esch, and Rob Leurs; *Discovery of S-(2-Guanidylethyl)-isothiourea (VUF 8430) as a Potent Nonimidazole Histamine H4 Receptor Agonist*; J. Med. Chem. **2006**, Vol. 49, Pag. 6650-6651

[34] Marlon D. Cowart, Robert J. Altenbach, Huaqing Liu, Gin C. Hsieh, Irene Drizin, Ivan Milicic, Thomas R. Miller, David G. Witte, Neil Wishart, Shannon R. Fix-Stenzel, Michael J. McPherson, Ronald M. Adair, Jill M. Wetter, Brian M. Bettencourt, Kennan C. Marsh, James P. Sullivan, Prisca Honore, Timothy A. Esbenshade, and Jorge D. Brioni; *Rotationally Constrained 2,4-Diamino-5,6-disubstituted Pyrimidines: A New Class of Histamine H4 Receptor Antagonists with Improved Druglikeness and in Vivo Efficacy in Pain and Inflammation Models*; J. Med. Chem. **2008**, Vol. 51, Pag. 6547–6557

[35] Huaqing Liu, Robert J. Altenbach, Tracy L. Carr, Prasant Chandran, Gin C. Hsieh, La Geisha R. Lewis, Arlene M. Manelli, Ivan Milicic, Kennan C. Marsh, Thomas R. Miller, Marina I. Strakhova, Timothy A. Vortherms, Brian D. Wakefield, Jill M. Wetter, David G. Witte, Prisca Honore, Timothy A. Esbenshade, Jorge D. Brioni, and Marlon D. Cowart; *cis-4-(Piperazin-1-yl)-5,6,7a,8,9,10,11,11a-octahydrobenzofuro[2,3-h]quinazolin-2-amine (A-987306), A New Histamine*

H4R Antagonist that Blocks Pain Responses against Carrageenan-Induced Hyperalgesia; J. Med. Chem. **2008**, Vol. 51, Pag. 7094–7098

[36] Tesi Francesco Sirci: Esplorazione dello Spazio Chimico-Farmaceutico mediante metodologie In Silico **2008/09**

[37] Maksims Vanejevs,Claudia Jatzke,Steffen Renner,Sibylle Müller, Mirko Hechenberger,Tanja Bauer,Anna Klochkova, Ilya Pyatkin, Denis Kazyulkin, Elena Aksenova, Sergey Shulepin, Olga Timonina, Ariane Haasis, Aleksandrs Gutcaits, Christopher G. Parsons, Valerjans Kauss, and Tanja Weil; *Positive and Negative Modulation of Group I Metabotropic Glutamate Receptors*; *J. Med. Chem.* **2008**, Vol. *51, Pag.* 634–647

[38] Anne Thiry, Marie Ledecq, Alessandro Cecchi, Raphael Frederick, Jean-Michel Dogné, Caudiu T. Supuran, Johan Wouters, Bernard Masereel; *Ligand-based and structure-based virtual screening to identify carbonic anhydrase IX inhibitors*; Bioorganic & Medicinal Chemistry, **2009,** Vol. 17, Pag. 553–557

[39] Deveraux Q. L. and Reed J. C.; *IAP family proteins-suppressor of apoptosis.*; Genes. Dev. **1999**, Vol. 1, Pag. 239-252

[40] Salvesen G. S. and Duckett C. S.; *IAP proteins: blocking the road to death's door. Nat*.; ReV. Mol. Cell Biol. **2002,** Vol. 3, Pag. 401-410.

[41] Riedl S. J. and Shi Y. M.*; Molecular mechanisms of caspase regulation during apoptosis.*; Nat. ReV. Mol. Cell Biol., **2004**, Vol. 5, Pag. 897-907.

[42] LaCasse E. C., Baird S., Korneluk R. G. and MacKenzie A. E.; *The inhibitors of apoptosis (IAPs) and their emerging role in cancer.*; Oncogene 1998, Vol. 17, Pag. 3247-3259

[43] Holcik M., Gibson H., and Korneluk R. G.; *XIAP: Apoptotic brake and promising therapeutic target.;* Apoptosis **2001**, Vol. 6, Pag. 253-261

[44] Huang Y., Park Y., Rich R., Segal D., Myszka D., and H. W.; *Structural basis of caspase inhibition by XIAP: di_erential roles of the linker versus the BIR domain*; Cell, **2001**, Vol. 104, Pag. 781-790

[45] Geng Wu, Jijie Chai, Tomeka L. Suber , Jia-Wei Wu, Chunying Du, Xiaodong Wang & Yigong Shi; *Structural basis of IAP recognition by Smac/DIABLO*; **2000** Nature Vol. **408**, Pag. 1008-1012

[46] Sun H., Stuckey J. A., Nikolovska-Coleska Z., Qin D., Meagher J. L., Qiu O. S., Lu Yang C.Y., Saito N. G., and Wang S.; *Structure-Based Design, Synthesis, Evaluation, and Crystallographic Studies of Conforma-tionally Constrained Smac Mimetics as Inhibitors of the X-linked Inhibitor of Apoptosis Protein (XIAP).;* J. Med. Chem., **2008**, Vol. 51, Pag. 7169-7180

[47] Park C.M., Sun C., Olejniczakand E. T., Wilson A. E., Meadows R. P., Betz S. F., Elmorea S. W., and Fesika S. W.; *Non-peptidic small molecule inhibitors of XIAP*.; Bioorg. Med. Chem. Lett., **2005,** Vol. 15, Pag. 771-775

[48] Kerry Zobel, Lan Wang, Eugene Varfolomeev, Matthew C. Franklin, Linda O. Elliott, Heidi J. A. Wallweber, David C. Okawa, John A. Flygare, Domagoj Vucic, Wayne J. Fairbrother, and Kurt Deshayes; *Design, Synthesis, and Biological Activity of a Potent Smac Mimetic That Sensitizes*

Cancer Cells to Apoptosis by Antagonizing IAPs; ACS CHEMICAL BIOLOGY **2006**, VOL.1 NO.8 , Pag. 525-534

[49] Thorsten K. Oost, Chaohong Sun, Robert C. Armstrong, Ali-Samer Al-Assaad, Stephen F. Betz, Thomas L. Deckwerth, Hong Ding, Steven W. Elmore, Robert P. Meadows, Edward T. Olejniczak, Andrew Oleksijew, Tilman Oltersdorf, Saul H. Rosenberg, Alexander R. Shoemaker, Kevin J. Tomaselli, Hua Zou, and Stephen W. Fesik; *Discovery of Potent Antagonists of the Antiapoptotic Protein XIAP for the Treatment of Cancer; J. Med. Chem.* **2004, Vol**. *47,* Pag. 4417-4426

[50] Haiying Sun, Zaneta Nikolovska-Coleska, Chao-Yie Yang, Liang Xu, York Tomita, Krzysztof Krajewski, Peter P. Roller, and Shaomeng Wang; *Structure-Based Design, Synthesis, and Evaluation of Conformationally Constrained Mimetics of the Second Mitochondria-Derived Activator of Caspase That Target the X-Linked Inhibitor of Apoptosis Protein/Caspase-9 Interaction Site; J. Med. Chem.* **2004,** Vol .47, Pag.4147-4150

[51] Bin Zhang,§ Zaneta Nikolovska-Coleska,‡,# Yan Zhang, Longchuan Bai,‡,# Su Qiu,‡,# Chao-Yie Yang, Haiying Sun, Shaomeng Wang, and Yikang Wu; *Design, Synthesis, and Evaluation of Tricyclic, Conformationally Constrained Small-Molecule Mimetics of Second Mitochondria-Derived Activator of Caspases*; *J. Med. Chem.* **2008,** Vol. *51,* Pag. 7352–7355

[52] Jui-Wen Huang, Ziming Zhang, Bainan Wu, Jason F. Cellitti, Xiyun Zhang, Russell Dahl, Chung-Wai Shiau, Kate Welsh, Aras Emdadi, John L. Stebbins, John C. Reed, and Maurizio Pellecchia; *Fragment-Based Design of Small Molecule X-Linked Inhibitor of Apoptosis Protein Inhibitors; J. Med. Chem.*, **2008**, Vol. 51 (22), Pag. 7111-7118

[53] Cheol-Min Park, Chaohong Sun,b Edward T. Olejniczak, Alan E. Wilson, Robert P. Meadows, Stephen F. Betz, Steven W. Elmore and Stephen W. Fesik; *Non-peptidic small molecule inhibitors of XIAP*; Bioorganic & Medicinal Chemistry Letters, **2005**, Vol. 15, Pag.771–775

[54] Aislyn D. Wist, Lichuan Gu, Stefan J. Riedl,_ Yigong Shi and George L. McLendon; *Structure–activity based study of the Smac-binding pocket within the BIR3 domain of XIAP*; Bioorganic & Medicinal Chemistry **2007**, Vol.15, Pag. 2935–2943

[55] Chudi Ndubaku, Eugene Varfolomeev, Lan Wang, Kerry Zobel, Kevin Lau, Linda O. Elliott, Brigitte Maurer, Anna V. Fedorova, Jasmin N. Dynek, Michael Koehler, Sarah G. Hymowitz, Vickie Tsui, Kurt Deshayes, Wayne J. Fairbrother, John A. Flygare and Domagoj Vucic; *Antagonism of c-IAP and XIAP Proteins Is Required for Efficient Induction of Cell Death by Small-Molecule IAP Antagonists.* ACS Chem. Biol., **2009** ,Vol. 4 (7), Pag. 557–566