Impiego di tecniche Machine Learning per lo sviluppo di modelli QSAR predittivi applicati ad inibitori di BCL-2

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Farmacia e Medicina
Corso di Laurea in Chimica e Tecnologia Farmaceutiche
Tesi Sperimentale in Chimica Farmaceutica
A.A. 2016/2017

Laureanda: Ylenia Carta
Matricola: 1384915

Relatore: Prof. Rino Ragno

Per cominciare…

Overview

Ruolo di **BCL-2** nel processo apoptotico

Metodologie **QSAR** e procedura sperimentale

Risultati

L'apoptosi

Evento «*programmato*» di morte cellulare

Stimoli induttori
- Fisiologici
- Patologici

Squilibri
- \+ Malattie neurodegenerative
- \- Neoplasie

Vie di esecuzione
- Via estrinseca
- Via intrinseca

Il target: BCL-2

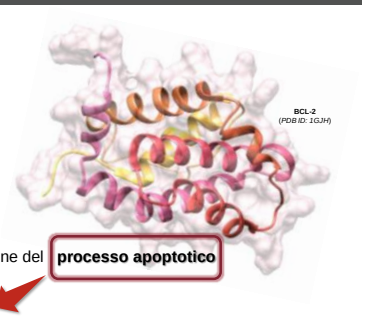
BCL-2
(*B-Cell Lymphoma 2*)

Proteine codificate

Gene

Ruolo biologico:
➪ Regolazione del **processo apoptotico**

Via intrinseca

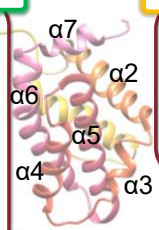


La famiglia BCL-2

Fattori pro-apoptotici

BH3 – only:
- BIM
- BAD
- BID
- PUMA
- NOXA

Multi BH domain:
- BAX
- BAK



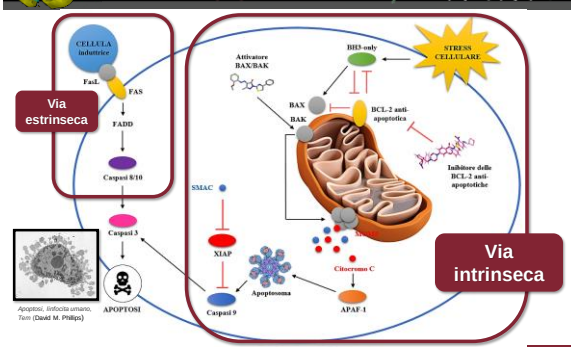
Fattori anti-apoptotici
- BCL-2
- BCL-X_L
- BCL-W
- MCL-1
- A1
- BCL-B

4 domini
9 α-eliche

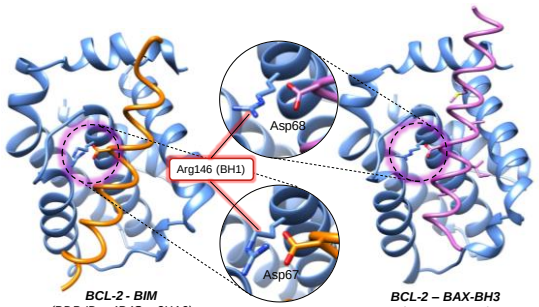
Fasi di esecuzione del processo apoptotico

Via estrinseca

Via intrinseca



BCL-2: cristalli PDB, interazioni chiave

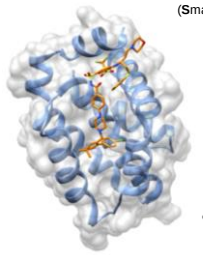


BCL-2 - BIM
(PDB IDs: 4B4S + 2XA0)

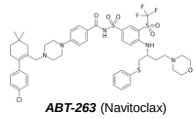
BCL-2 – BAX-BH3
(PDB ID: 2XA0)

Strategia farmacologica

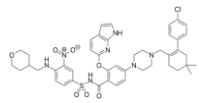
BH3 mimetici
(**S**mall **M**olecule **I**nhibitors, *SMIs*)



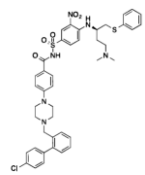
BCL-2 – ABT-263
(PDB ID: 4LVT)



ABT-263 (Navitoclax)



ABT-199 (Venetoclax)



ABT-737

Relazioni quantitative struttura-attività

(Q)SAR
=
(Quantitative) Structure – Activity Relationship

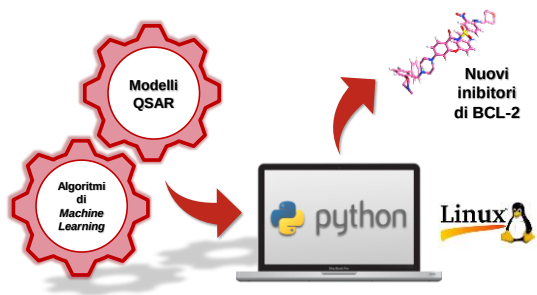
Struttura, proprietà chimico-fisiche

$$[y] = f([x])$$

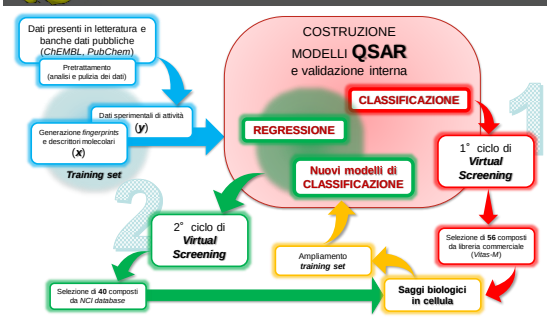
Attività, effetto tossico

IN SILICO

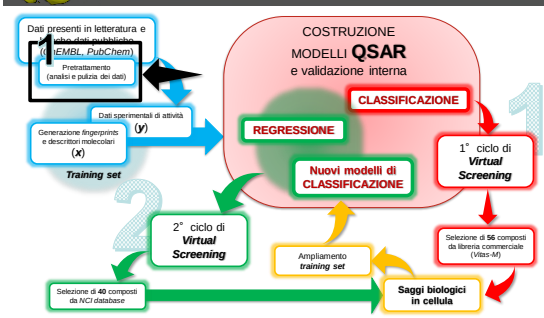
Scopo della Tesi



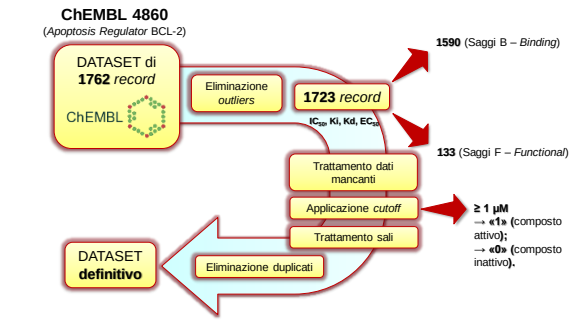
Procedura sperimentale



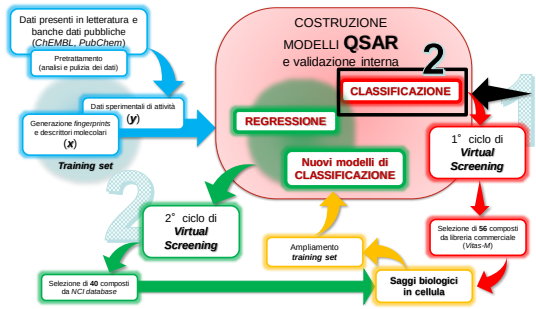
Procedura sperimentale



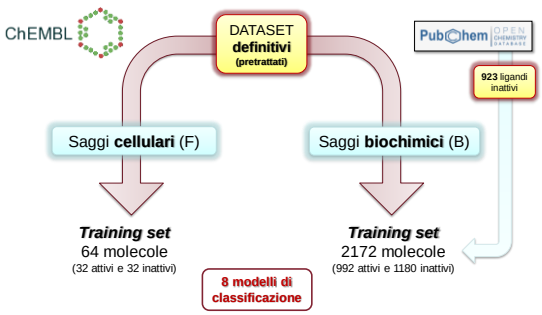
Pretrattamento dati



Procedura sperimentale



Primo round di classificazione: i *training set*



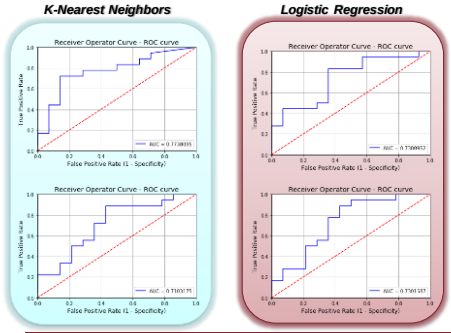
Modelli di classificazione per saggi cellulari (F)

Validazione interna | Ottimizzazione con *Grid Search* | Cross-Validazione con approccio *Leave-Half-Out* (50% del *training* set, 32 molecole) | *StratifiedShuffleSplit* (n_split = 100)

Tipo di X	Algoritmo	Miglior set di parametri	*Accuracy*	MCC	ROC AUC	F_1 *score*
Fingerprints	*KNN*	*K = 6*, *Metric = Hamming*, *Weights = Distance*	0.78	0.57	0.77	0.79
Fingerprints	*Log. Reg.*	*C = 1*, *Penalty = l2*	0.76	0.54	0.74	0.78
Descrittori	*KNN*	*K = 13*, *Metric = Cosine*, *Weights = Distance*	0.70	0.43	0.71	0.75
Descrittori	*Log. Reg.*	*C = 0.5*, *Penalty = l2*	0.70	0.44	0.73	0.74

Receiver Operator Curves (F)

***K*-Nearest Neighbors** | ***Logistic* Regression**

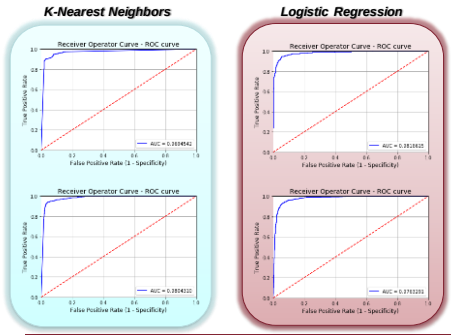


Modelli di classificazione per saggi biochimici (B)

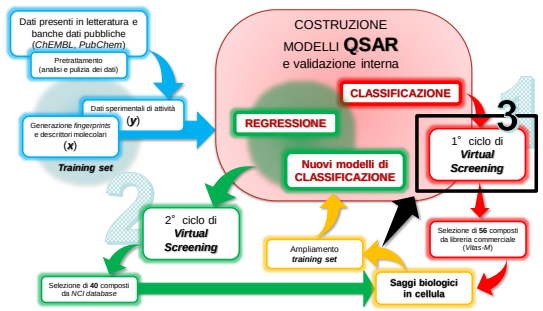
Validazione interna | Ottimizzazione con *Grid Search* | Cross-Validazione con approccio *Leave-Many-Out* (90% del *training* set, 1955 molecole) | *StratifiedShuffleSplit* (n_split = 100)

Tipo di X	Algoritmo	Miglior set di parametri	*Accuracy*	MCC	ROC AUC	F_1 *score*
Fingerprints	*KNN*	*K = 4*, *Metric = Hamming*, *Weights = Distance*	0.93	0.86	0.97	0.92
Fingerprints	*Log. Reg.*	*C = 0.5*, *Penalty = l2*	0.93	0.87	0.98	0.93
Descrittori	*KNN*	*K = 5*, *Metric = Cosine*, *Weights = Distance*	0.93	0.85	0.98	0.91
Descrittori	*Log. Reg.*	*C = 1*, *Penalty = l2*	0.93	0.85	0.98	0.90

Receiver Operator Curves (B)



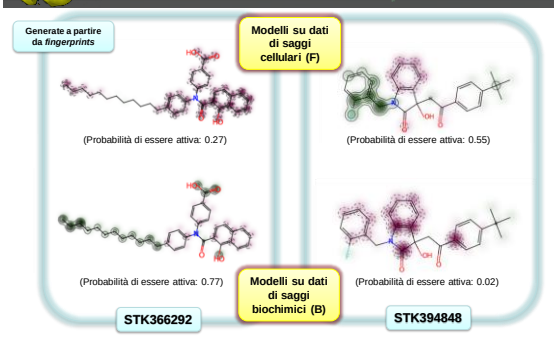
Procedura sperimentale



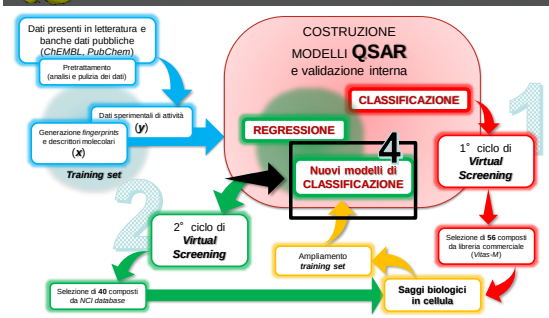
Primo ciclo di *Virtual Screening*

MOLID	Struttura	DOSE (μM)	POZZETTO 1	POZZETTO 2	Mortalità cellulare media (%)	COMMENTO
STK366292		5	3.2	4.2	3.7	
STK366292		20	95.1	93.1	94.1	
STK366292		50	98	96.5	97.25	
STK367596		5	4	3.8	3.9	
STK367596		20	82.8	91.6	87.2	
STK367596		50	98	98.1	98.05	

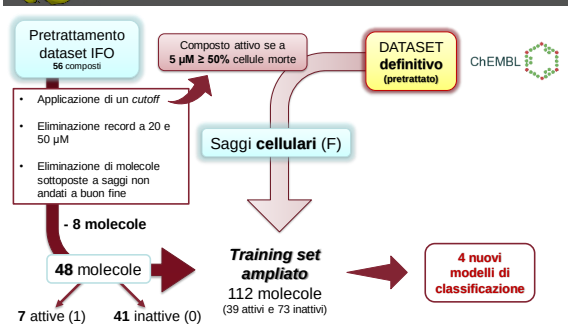
Mappe di similarità



Procedura sperimentale



Secondo round di classificazione: *training set*



Nuovi modelli di classificazione per saggi cellulari (F)

Validazione interna

Ottimizzazione con ***Grid Search***

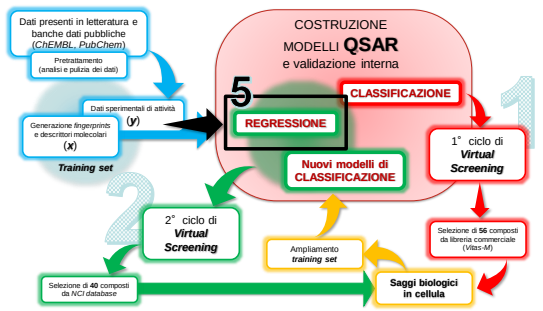
Cross-Validazione con approccio ***Leave-Half-Out*** (**50**% del *training set*, 56 molecole)

StratifiedShuffleSplit (n_split = 100)

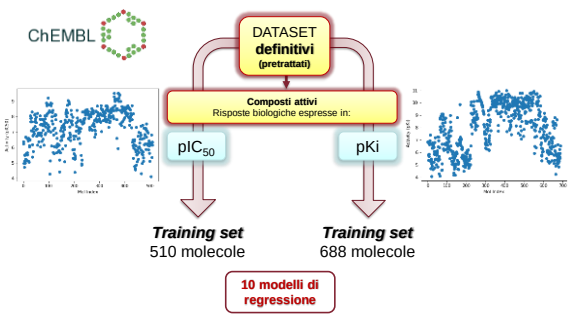
Tipo di X	Algoritmo	Miglior set di parametri	Accuracy	MCC	ROC AUC	F_1 score
Fingerprints	***KNN***	*K = 5* *Metric = Hamming* *Weights = Uniform*	**0.81**	**0.60**	**0.85**	0.72
Fingerprints	***Log. Reg.***	*C = 1* *Penalty = l2*	**0.77**	0.49	**0.76**	0.66
Descrittori	***KNN***	*K = 10* *Metric = Cosine* *Weights = Uniform*	**0.76**	**0.50**	**0.86**	0.68
Descrittori	***Log. Reg.***	*C = 0.5* *Penalty = l1*	**0.75**	**0.50**	**0.86**	0.68

Performance

Procedura sperimentale



Regressione: i *training set*



Modelli di regressione

Validazione interna

Fingerprints

Ottimizzazione con ***Grid Search***

Cross-Validazione con approccio ***Leave-Half-Out*** (**50**% del *training set*)

255 molecole (pIC_{50})

344 molecole (pKi)

StratifiedShuffleSplit (n_split = 100)

	Miglior set di parametri	*Random Forest* r^2	q^2_{LHO}	SDEC	SDEP
pIC_{50}	*N_estim. = 100*	0.97	0.75	0.20	0.55
pKi	*N_estim. = 100*	0.98	0.86	0.23	0.65
		KNN			
		r^2	q^2_{LHO}	SDEC	SDEP
pIC_{50}	*K = 7* *Metric = Hamming* *Weights = Distance*	0.99	0.73	0.02	0.59
pKi	*K = 5* *Metric = Hamming* *Weights = Distance*	0.99	0.85	0.01	0.69
		Ridge Regression			
		r^2	q^2_{LHO}	SDEC	SDEP
pIC_{50}	*α = 10*	0.96	0.76	0.23	0.53
pKi	*α = 25*	0.95	0.87	0.41	0.63

Modelli di regressione

Validazione interna

Descrittori

Ottimizzazione con ***Grid Search***

Cross-Validazione con approccio ***Leave-Half-Out*** (**50**% del *training set*)

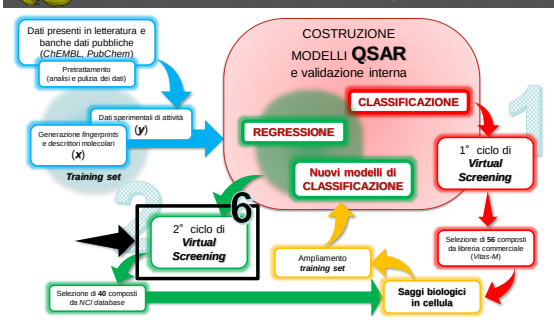
255 molecole (pIC_{50})

344 molecole (pKi)

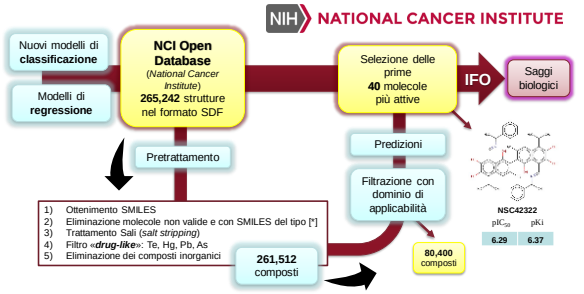
StratifiedShuffleSplit (n_split = 100)

	Miglior set di parametri	*Random Forest* r^2	q^2_{LHO}	SDEC	SDEP
pIC_{50}	*N_estim. = 100*	0.97	0.75	0.20	0.55
pKi	*N_estim. = 100*	0.98	0.86	0.23	0.65
		KNN			
		r^2	q^2_{LHO}	SDEC	SDEP
pIC_{50}	*K = 7* *Metric = Minkowski* *Weights = Distance*	0.99	0.73	0.02	0.59
pKi	*K = 10* *Metric = Minkowski* *Weights = Distance*	0.99	0.85	0.01	0.69

Procedura sperimentale

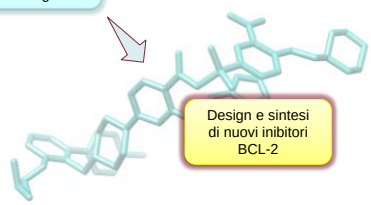


Secondo ciclo di *Virtual Screening*



Conclusioni e obiettivi futuri

- Modelli **QSAR** ad alta performance;
- Cicli di screening



Design e sintesi di nuovi inibitori BCL-2

Grazie per l'attenzione