Check My Blob

Come funziona? Pipeline di lavoro.

Laboratorio di Bioinformatica, Modulo 2 - Presentazione

Chiara Solito

Corso di Laurea in Bioinformatica
Università degli studi di Verona
A.A. 2021/22

1 Com'è stato costruito l'algoritmo Check My Blob?

Partiamo dalla collezione dei dati, per poi osservare come sono stati raggruppati i ligandi a partire dai dati raccolti e poi com'è stato addestrato il modello.

1.1 Data Collection

Il training set partiva da tutte le entry PDB scaricate il 19 Gennaio 2020. Dopo una necessaria "pulizia" dei dati, il training set finale consisteva in ligandi derivati da esperimenti di diffrazione a raggi X con almeno 4,0 Å di risoluzione.

1.2 Clustering dei Ligandi

Il passo successivo è stato clusterizzare i ligandi.
Diversi ligandi sono indistinguibili dalla sola densità elettronica. Ad esempio, per la maggior parte degli intervalli di risoluzione, è estremamente difficile distinguere tra 6, 7 o 8 elettroni (carbonio, azoto, ossigeno) o distinguere tra legami singoli, doppi o tripli in alcune configurazioni. Per migliorare la robustezza dell'algoritmo di classificazione, per Check My Blob è stato deciso di raggruppare i ligandi in base alla forma prevista dalla loro densità elettronica derivata dalle loro definizioni chimiche. Vale a dire, che si è considerata la connettività dell'atomo, la chiralità e le differenze significative nel numero di elettroni ma si è ignorato il tipo di atomo e l'ordine di legame in qualche equivalente configurazione. La procedura di raggruppamento è stata la seguente:

1. Raggruppare le molecole facendo corrispondere il numero di atomi, anelli, e anelli aromatici;
2. Raggruppare per connettività (abbinando le sottostrutture, utilizzando atomi generici e legami generici);
3. Controllare se la chiralità di tutte le combinazioni corrispondenti convalida;
4. Verificare se i modelli SMART corrispondono a posizioni equivalenti;
5. Controllare se gli atomi equivalenti sono nello stesso gruppo di numeri atomici.

1.3 Modello

Il modello di classificazione primario (GBM) è stato addestrato a riconoscere i 219 gruppi di ligandi più popolari (cluster). Questo numero è stato raggiunto limitando la formazione alle classi con almeno 100 esempi nel set di formazione. Tutti i ligandi che non erano in quei 219 gruppi sono stati etichettati come una classe separata chiamato "rara". Quando il modello di classificazione primario prevede "raro", l'esempio viene ulteriormente elaborato da un secondario modello (algoritmo 1-NN) addestrato solo su ligandi nel gruppo "raro".

1.4 Validazione del modello

Successivamente vi è stata una validazione del modello.

	10-fold CV	Hold out Set
Ligand instances	696 887	17 150
Mean resolution (Å)	2.2	2.5
Accuracy (%)	71.2(9)	58.9
Top-5 accuracy (%)	90.7(5)	87.2
Top-10 accuracy (%)	94.9(2)	92.5
Micro-averaged recall (%)	71.2(9)	58.9
Micro-averaged precision (%)	69.3(11)	62.7
Micro-averaged F1 (%)	69.3(11)	55.7
Cohen's kappa (%)	64.6(12)	46.2

L'hold out set era più impegnativo perché conteneva una risoluzione inferiore strutture di soluzione, ma i risultati sono comunque superiori a quelli segnalati per ligandi non clusterizzati con almeno due atomi di idrogeno.

2 Come funziona la pipeline di lavoro?

Stabilito il modo in cui è stato creato l'algoritmo Check My Blob vediamo cosa succede quando effettivamente si utilizza il web server.

Fase di apprendimento: I blob non interpretati vengono "tagliati" dalle mappe di densità elettronica generate tramite la sezione "polymer-only-portion" di PDB.
I blob vengono trovati automaticamente analizzando tutti i picchi di densità elettronica positivi all'interno della mappa Fo-Fc. Per mitigare il problema dei ligandi divisi in più blob, il sistema rileva i massimi locali e scheletizza la densità elettronica all'interno dell'isosuperficie di ciascun blob e combina i blob adiacenti se la distanza tra i massimi locali o i nodi dello scheletro è inferiore a 2,15 Å. Infine, tutti i frammenti di densità elettronica nell'isosuperficie del blob che si sovrappongono all'isosuperficie degli atomi del biopolimero modellato vengono ritagliati dal blob. In pratica, CheckMyBlob è in grado di rilevare ligandi costituiti da decine di candidati blob.

Descrizione del blob: Ogni blob è descritto tramite un set di feature numeriche, che sono date in pasto all'algoritmo di machine learning (un classificatore).

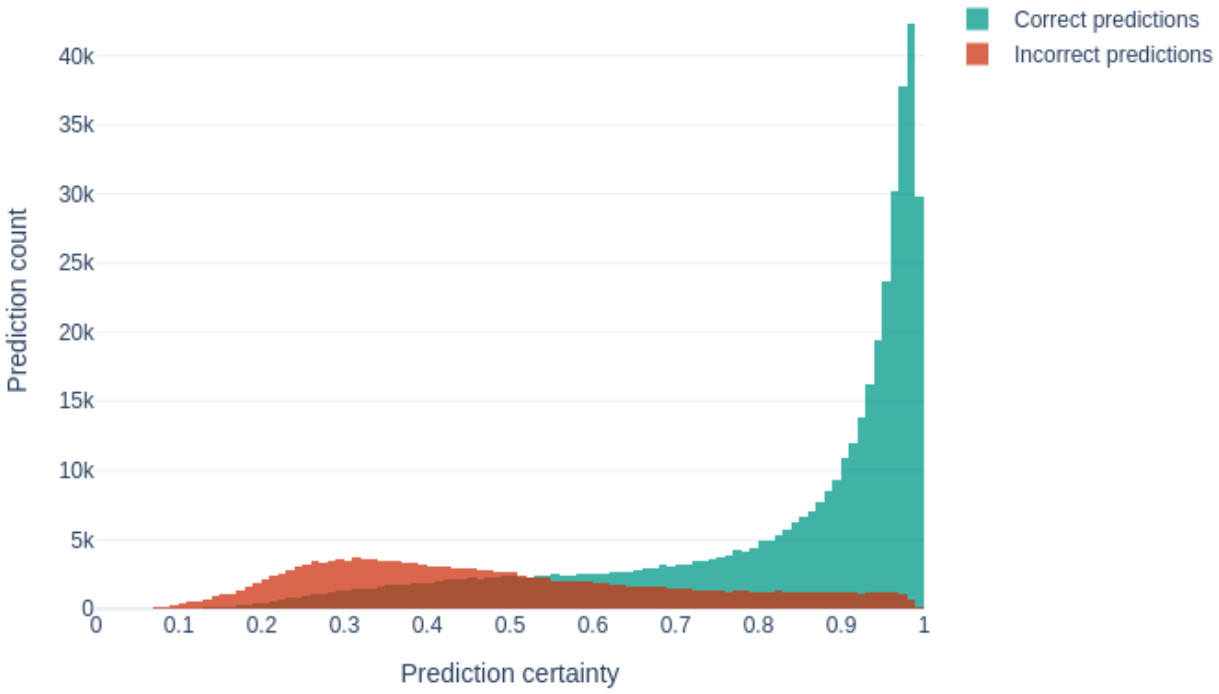
Classificazione: Il classificatore crea una funzione (un modello di classificazione) che predice i migliori ligandi basandosi sulle feature numeriche.

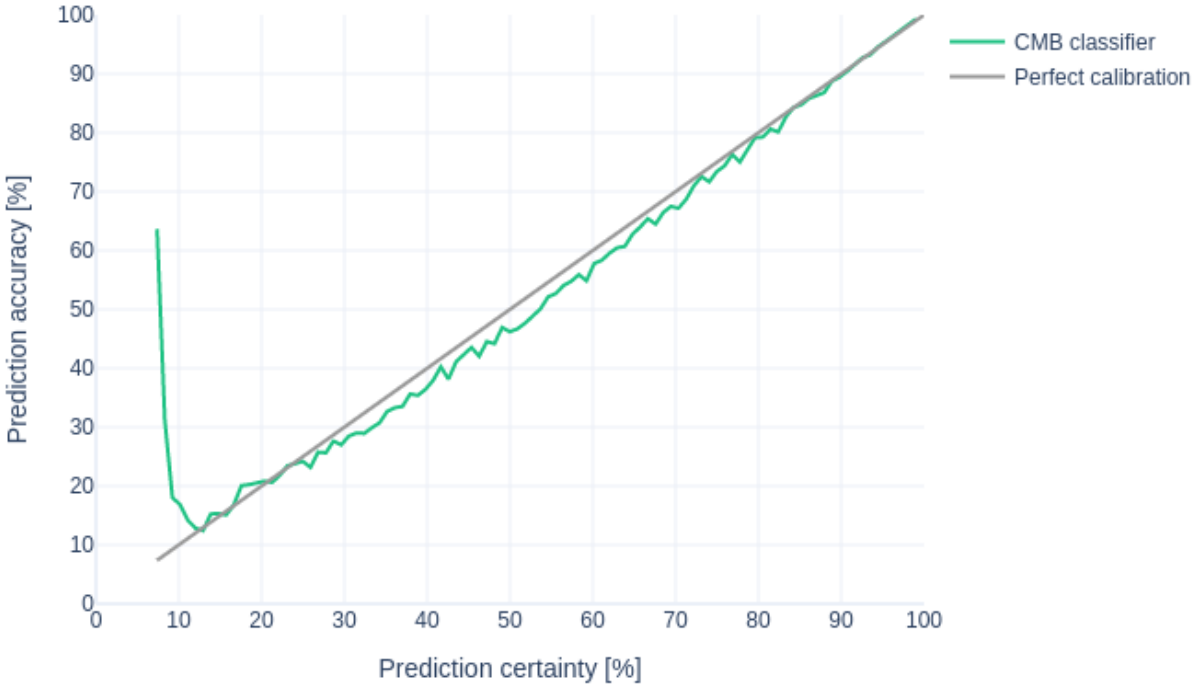
3 Performance e attendibilità dei risultati

Anche se le istanze del ligando utilizzate per l'addestramento sono selezionate in base a diversi criteri di qualità, ogni previsione di CheckMyBlob dovrebbe essere trattata come un suggerimento che necessita di ulteriori indagini.
La predizione si basa sulla mappa di densità degli elettroni, mentre la conoscenza dei ligandi che potrebbero essere presenti nel cristallo è basata su:

- condizioni di cristallizzazione
- componenti del protein buffer
- componenti aggiuntive
- componenti che potrebbero essere stati eliminati durante la purificazione proteica
- tutte le reazioni chimiche tra questi

L'affidabilità di una previsione concreta può essere misurata dalla certezza della previsione (probabilità). Ogni previsione del server è accompagnata da una probabilità percentuale che il server abbia ragione. L'istogramma e il line plot di seguito mostrano la frequenza con cui il server è stato corretto per un determinato livello di certezza, in termini di valori assoluti e relativi. Entrambi i grafici mostrano che le previsioni con maggiore certezza sono in realtà molto probabili, mentre le previsioni con valori di certezza inferiori hanno una maggiore probabilità di essere errate.


Correct predictions
Incorrect predictions
Prediction count
40k
35k
30k
25k
20k
15k
10k
5k
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prediction certainty


CMB classifier
Perfect calibration
Prediction accuracy [%]
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Prediction certainty [%]