Novozymes
酶穩定性預測

409410197 周騏軍 409410197 劉柏辰
409410296 王諠傑 409416541 蔡敔

目錄

研究動機

環保議題一直是人類共同要面對的一大問題，隨著近百年來工業發展的迅速，地球上的資源幾乎快被人類掏空了，若要朝著永續發展前進，我們勢必要利用更少的資源製造更大的價值。

因此我們這組朝著環保議題或永續發展這類型的方向找資料，找到了有一篇學術研究，內容在講NOVOZYMES在自然界中發現酶，酶在優化後就可以取代化學品並加速生產過程，以利人類能節省能源、減少對自然界的傷害，同時又可以創造更大的價值，聽起來超級符合未來的趨勢，所以我們這組就決定做這個主題。

酵素酶的功能

洗滌劑

酶可以用在洗衣和洗碗的洗滌劑中，它可以在去除污漬並實現低溫洗滌。

食物品質

酶可以用在人類的食物上，例如改善麵包與葡萄酒的品質，也可以用在提高動物飼料的營養價值。

生物燃料

酶在生產生物燃料時，可以將生物值中的澱粉或纖維素轉為醣類，然後在發酵成乙醇。

需克服的問題

大部分的酶其實只是勉強穩定而已，這限制了科學家們在嚴苛應用條件下的性能，
而不穩定性則會降低細胞可產生的蛋白質數量。
然而，開發預測蛋白質穩定性的有效計算方法，需要巨大的技術。

這次的專題，我們將會從以前的科學數據中，
試圖找尋有效的計算方法，並預測蛋白質的穩定性。

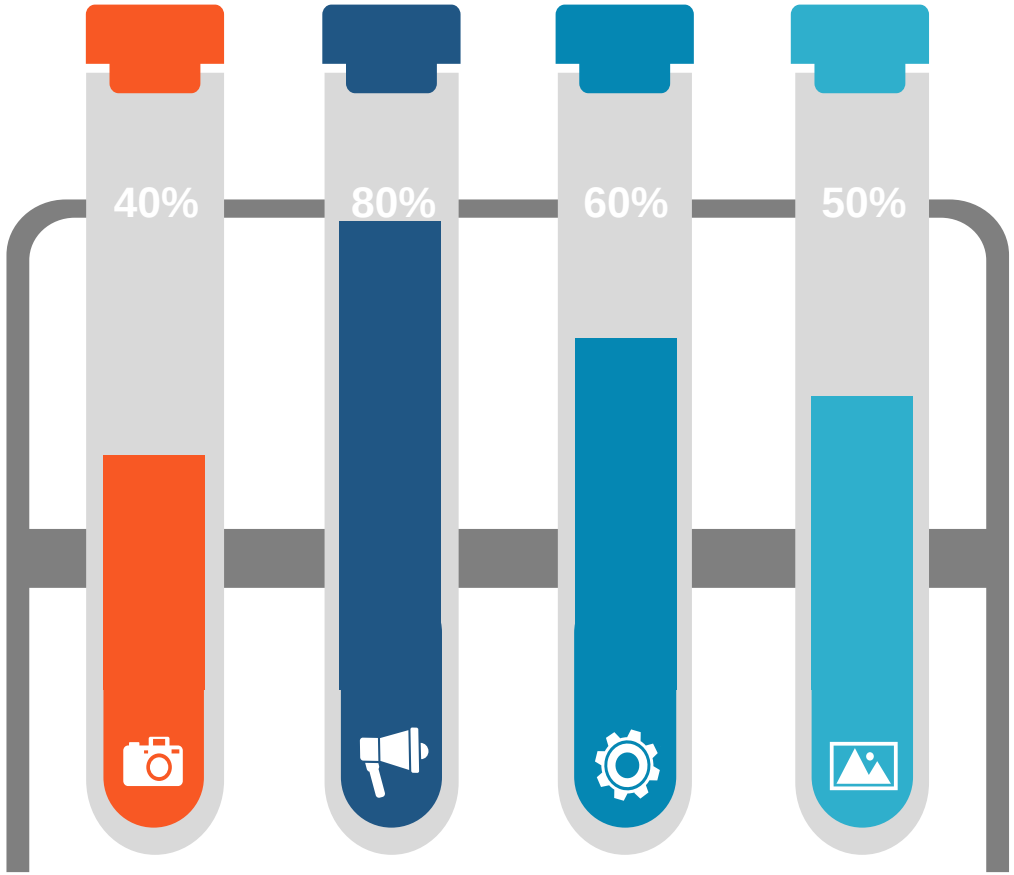
研究方法–資料前處理

01 原始資料中有錯因此先照著官方指示將原資料中的錯誤給修正。

02 發現資料中PH值有缺失並以平均值補上。

03 將資料欄位中的蛋白質序列做拆解並計算每個序列出現的次數，以此做為判斷不同酶的依據。

研究方法-資料預測



首先先使用gridsearch對超參數做search

先將目標欄位分割出來以及做資料切割。

我們使用XGBRegressor做資料預測，xgb跟隨機森林相似，然而隨機森林是決策樹每每獨立，而xgb為每棵樹相互影響，但是xgb的優勢在於更多的超參數和更快的迭代速度，提供了更多參數調整的可能性，因此我們最終決定用xgb進行資料預測。

```
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8,
             early_stopping_rounds=None, enable_categorical=False,
             eval_metric=None, feature_types=None, gamma=0, gpu_id=0,
             grow_policy='depthwise', importance_type=None,
             interaction_constraints='', learning_rate=0.1, max_bin=256,
             max_cat_threshold=64, max_cat_to_onehot=4, max_delta_step=0,
             max_depth=5, max_leaves=0, min_child_weight=1, missing=nan,
             monotone_constraints='()', n_estimators=800, n_jobs=0,
             num_parallel_tree=1, predictor='auto', random_state=0, ...)
```

成 果

如右圖所示，最終訓練的分數還算不錯，在如此短的時間內將大量的樣本進行訓練並取得如此成績，甚是感慨。

```
MSE train: 2.296, test: 60.291
R^2 train: 0.984, test: 0.594
spearmanr train: SpearmanrResult(correlation=0.9869671934316924, pvalue=0.0)
spearmanr test: SpearmanrResult(correlation=0.5773911818453205, pvalue=0.0)
```

我們可以看到這次我們取得了1691的成績，就結果而言還算不錯，希望能藉此機會幫助到自然環境。

1691

1691	**TsaiYu**		0.076	10	1d


THANK YOU
第15組