

網羅的タンパク質間相互作用予測システム における判別精度の改良

○大上 雅史 松崎 裕介 松崎 由理 秋山 泰

東京工業大学

大学院情報理工学研究科 計算工学専攻

2009/09/17(木) 18th IPSJ SIG BIO



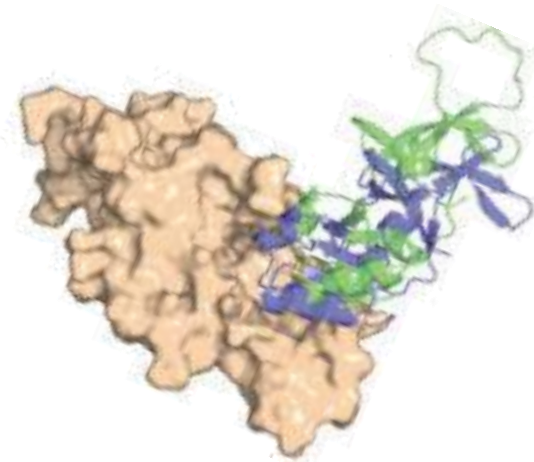
本研究について

● タンパク質間相互作用(PPI)ネットワーク

- 生命現象における中心的役割を担う
- 創薬ターゲットとして注目

● 計算機によるPPI予測の発展

- 配列モチーフからの予測
- 文献情報からの予測
- ドッキングシミュレーション
 - ー 分子動力学法
 - ー 形状相補性に基づいたタンパク質ドッキング(PPD)



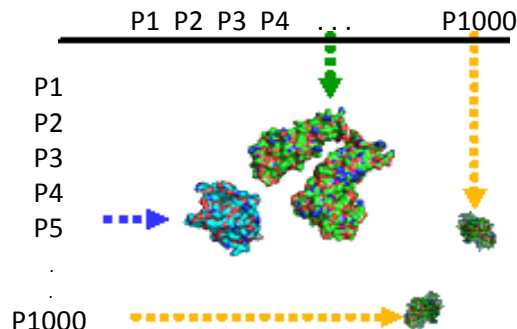
● タンパク質ドッキングによるPPIネットワーク予測

- 増えつつある立体構造データの活用
- PPI予測システムの確立へ



本研究について

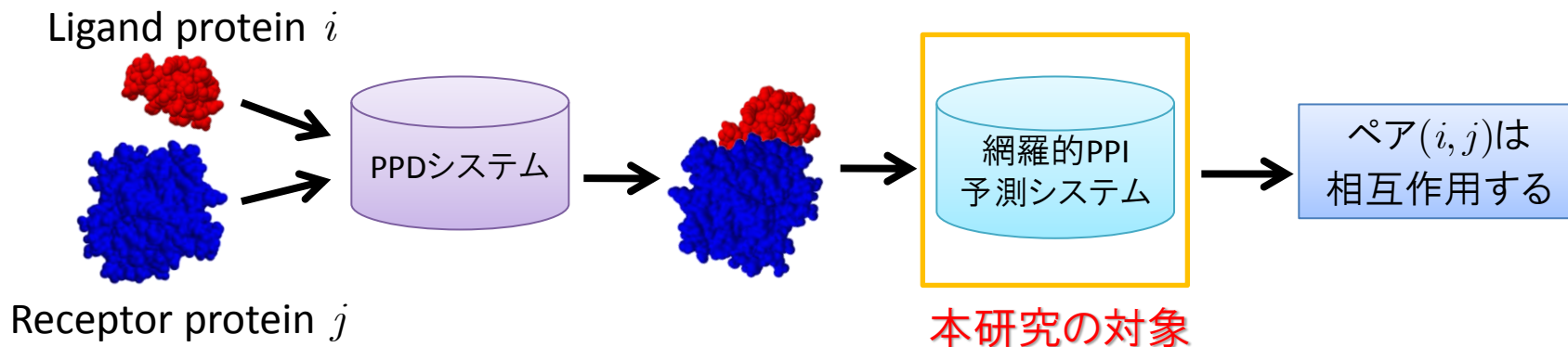
● PPDによって網羅的にPPI予測を行う



● PPDシステムは尤もらしい複合体を出力する

● タンパクペアが複合体を形成するかどうかは問題ではない

● ドッキングされた複合体から相互作用の判断が必要





● all-to-all PPD専用システム

- 計算の高速性に特化した網羅的計算専用PPDシステム
- 現行Ver.は2.1 (M.Ohue, IPSJ SIG-BIO 17(11), 2009)
- 形状相補性 G と静電的相互作用 E を考慮

$$\mathbf{R}_{l,m,n} = G_{l,m,n}^R + iE_{l,m,n}^R$$

$$\mathbf{L}_{l,m,n} = G_{l,m,n}^L + iwE_{l,m,n}^L$$

$$\mathbf{S}_{\alpha,\beta,\gamma} = \Re \left[\sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N \mathbf{R}_{l,m,n} \mathbf{L}_{l+\alpha,m+\beta,n+\gamma} \right]$$

- 形状相補性 G は独自のrPSC Scoreを利用
- 静電的相互作用 E はCHARMMを利用
- 2000個の複合体予測構造と、スコア $S_1 \sim S_{2000}$ を出力する



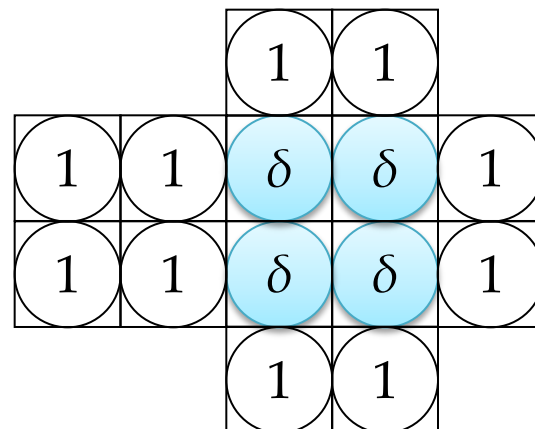
rPSC (real Pairwise Shape Complementarity)

(M.Ohue, IPSJ SIG-BIO 17(11), 2009)

1	2	3	3	3	2	1
2	3ρ	3ρ	3ρ	3ρ	3ρ	2
3	3ρ	9ρ	3ρ	3ρ	3ρ	2
3	3ρ	9ρ	3ρ	5	2	1
3	3ρ	9ρ	3ρ	5	2	1
3	3ρ	9ρ	3ρ	3ρ	3ρ	2
2	3ρ	3ρ	3ρ	3ρ	3ρ	2
1	2	3	3	3	2	1

Receptor rPSC

※ ρ は負, δ は正のパラメータ



Ligand rPSC



網羅的PPI予測～従来手法①

● η -method (M.Ohue, IPSJ SIG-BIO 17(11), 2009)

● $M \times M$ 通りのペア (i, j) に関する1位のスコア $S_{ij}^{(top)}$ を利用

$$\eta_{ij} = \frac{S_{ij}^{(top) \ 2/3}}{\min\{A_i, A_j\}}$$

● A_i, A_j はそれぞれタンパク質 i, j の表面積

● ここから M^2 個の η を母集団とする z-score を計算

$$z_{ij} = \frac{\eta_{ij} - \mu}{\sigma}$$

● ある閾値 z_τ を超えるペアは相互作用すると判定

$$z_{ij} > z_\tau \rightarrow (i, j) \text{ は相互作用する}$$



網羅的PPI予測～従来手法②

● AEP (Affinity Evaluation and Prediction)

(K.Tsukamoto, 2008)

● 以下の4ステップから構成

- 形状相補性によるPPD
- Groupingとよばれるクラスタリングに基づく独自の処理
- z-scoreの計算
- 親和性スコアと呼ばれるクラスタの重み付け和を用いた指標の計算



提案手法

● ブースティングによる判別モデル生成

● 機械学習メタアルゴリズム

● 多数の弱学習器の重みつき多数決による判別

● 過学習が起こりにくいと考えられている

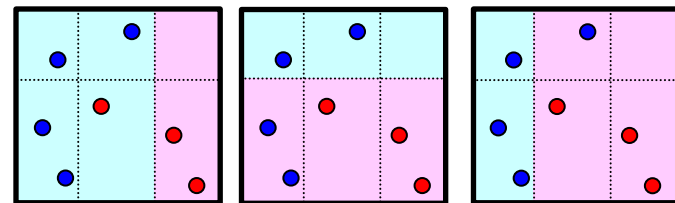
● 実装

● スキーム: **LogitBoost**

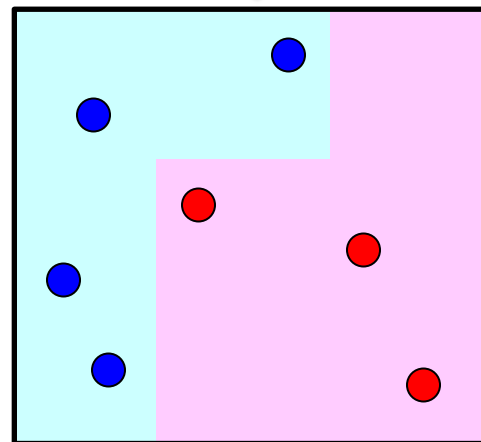
● サンプルのノイズや外れ値に強い

● 弱学習器: **Decision Stump**

● 高さ2の決定木学習



多数決





問題設定

- PPD出力の複合体からPPI予測を行う
- 以下について記述
 - 扱うデータセット
 - 特徴量の選出
 - 正例と負例のバランスについて
 - 評価指標



問題設定～データセット

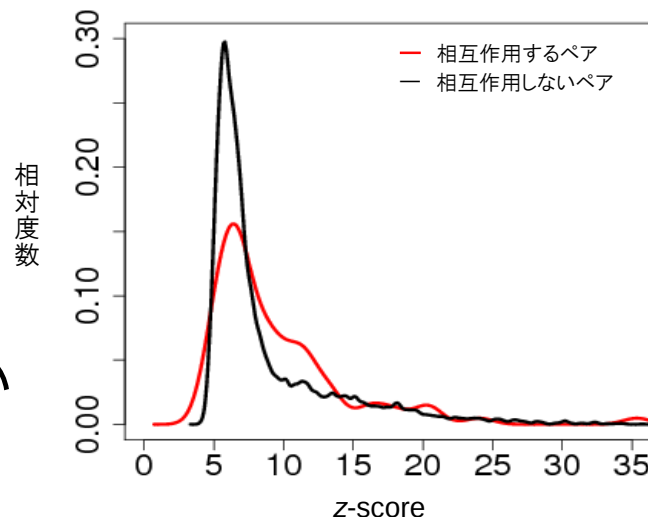
- PPD Benchmark 2.0 (J. Mintseris, 2005)
 - 網羅的PPI予測の研究で広く用いられている
 - 84組のペアにおけるレセプターとリガンドの立体構造データをMEGADOCK 2.1によって網羅的にドッキングさせたもの
 - 23個の複合体によるサブセット(サンプル数:529)
 - 84個の全複合体によるデータセット(サンプル数:7056)
 - 42個と42個に分割した交差検定用サブセット(サンプル数:1764)



問題設定～特徴量の選出

● ドッキングスコアS

- 複合体構造の良さを示すためにPPDシステムが出力
 - タンパク質の大きさにある程度依存
 - 1位～2000位まで
- 2000次元だとサンプル空間がスパースになり過学習の恐れあり
- 特徴選択(フィルタ法)
 - Reliefアルゴリズムによってリランキング
 - ⇒ 結果ほぼ変わらず
 - 相関係数によって類似度を評価
 - ⇒ 1位とそれ以外で明らかな差
 - ⇒ 今回は S_1 のみ利用
- ドッキングスコアだけでは判別が難しい
 - 他の特徴量の検討

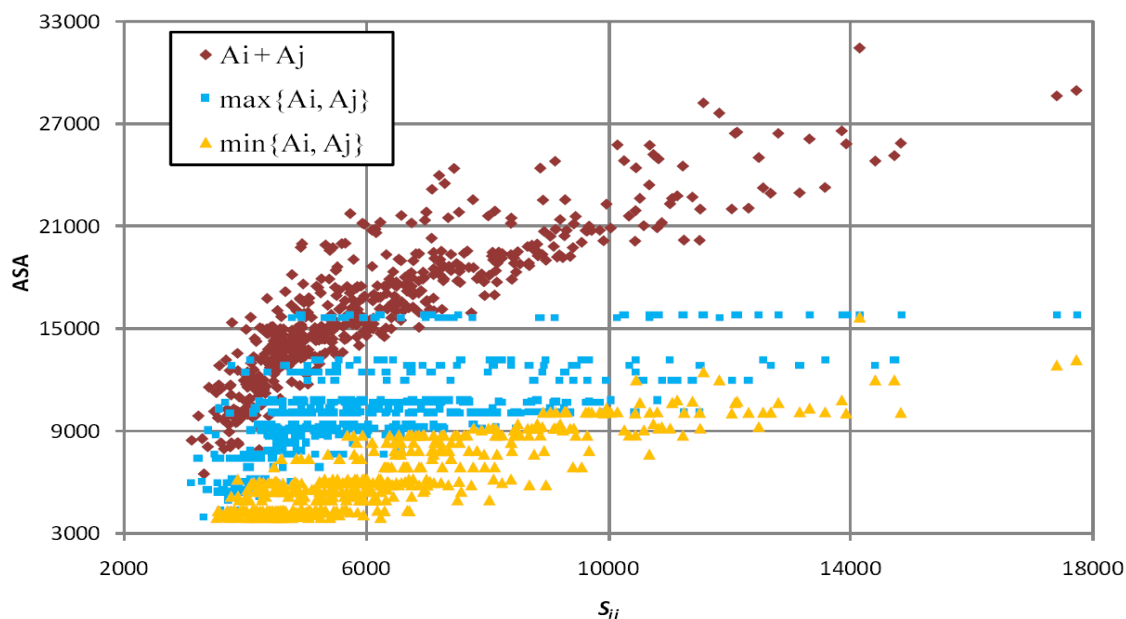




問題設定～特徴量の選出

タンパク質の表面積

● Dockingスコアは表面積に依存



● それぞれの表面積 A_R , A_L を利用

● 表面積による判別を選択する弱学習器が少ない

⇒ ReceptorとLigandそれぞれの重複によるものと考えられる

● η -methodの η を特徴として追加

- PPI予測の正例と負例⇒PDBに登録されているか否か
 - 行列の対角要素が正例, それ以外は負例
 - $M \times M$ のPPDでは正例:負例=1:M-1
 - ⇒多くの弱学習器が「すべて負例である」という予測を導く
 - 数学的には(予測精度の面で)正しいが実用上は無意味
 - 正例と負例を同等に扱う必要がある
 - オーバーサンプリング vs. アンダーサンプリング
 - 正例が極めて少ないためアンダーサンプリングではサンプル数が激減

● オーバーサンプリング

● Random Over Sampling

- ランダムに少数派サンプルと同一サンプルを追加
- 本研究では重みを一律にするため利用せず

● SMOTE

- k -最近傍サンプルを基に新たなサンプルを生成
- 生成された新サンプルに対する説明が不可能
- 予備実験で試してみたが精度改善されず

● 正例をそれぞれコピーして、数を $M-1$ 個ずつにする

- 重みを一律にすることが可能
- ノイズや外れ値も増幅される



問題設定～評価指標

● 性能評価に用いる値

- ⊕ TP 相互作用すると予測されたペアの複合体がPDBにある
- ⊕ FP 相互作用すると予測されたペアの複合体がPDBにない
- ⊕ FN 相互作用しないと予測されたペアの複合体がPDBにある
- ⊕ TN 相互作用しないと予測されたペアの複合体がPDBにない
- ⊕ precision 「相互作用する」と予測した中で当たっている割合
- ⊕ recall 正解すべき複合体のうち「相互作用する」と判定された割合
- ⊕ F-score precision と recall の調和平均

$$\text{precision} = \frac{TP}{TP+FP} \quad \text{recall} = \frac{TP}{TP+FN}$$

$$F\text{-score} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$$



問題設定～評価指標

● F -score について

● 既存研究では F -scoreによる評価

● 新たな相互作用発見のためにはFPの存在も重要

— TPとFPを小さくしてprecisionを高めても実用上意味がない

— recallの値が重要

● 正例負例のバランスの悪さから, precisionは非常に低い値になりがち

⇒recall重視の指標の必要性

● F_{β} -score (C.J.Van Rijsbergen, 1979)

● recallに一定の重みを付加

$$F_{\beta}\text{-score} = \frac{1 + \beta^2 \text{ recall} \cdot \text{precision}}{\text{recall} + \beta^2 \text{ precision}}$$

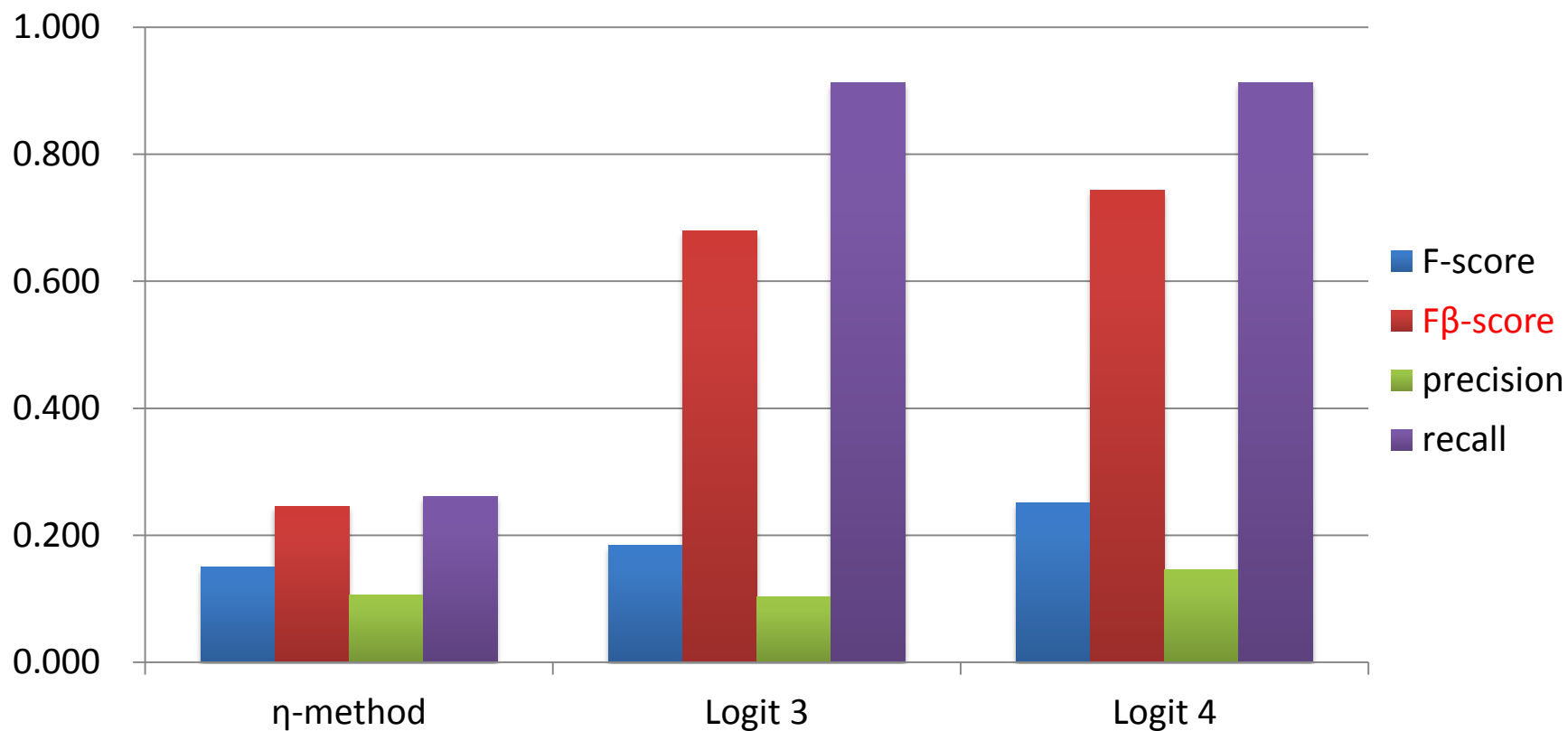
● β の値は系の大きさに合わせて $\beta = \sqrt{M-1}$ とする



実験① 23 × 23 PPI Prediction

● 比較対象

- η -method (M.Ohue, 2009)
- LogitBoost(S_1, A_R, A_L) **Logit3**
- LogitBoost(S_1, A_R, A_L, η) **Logit4**

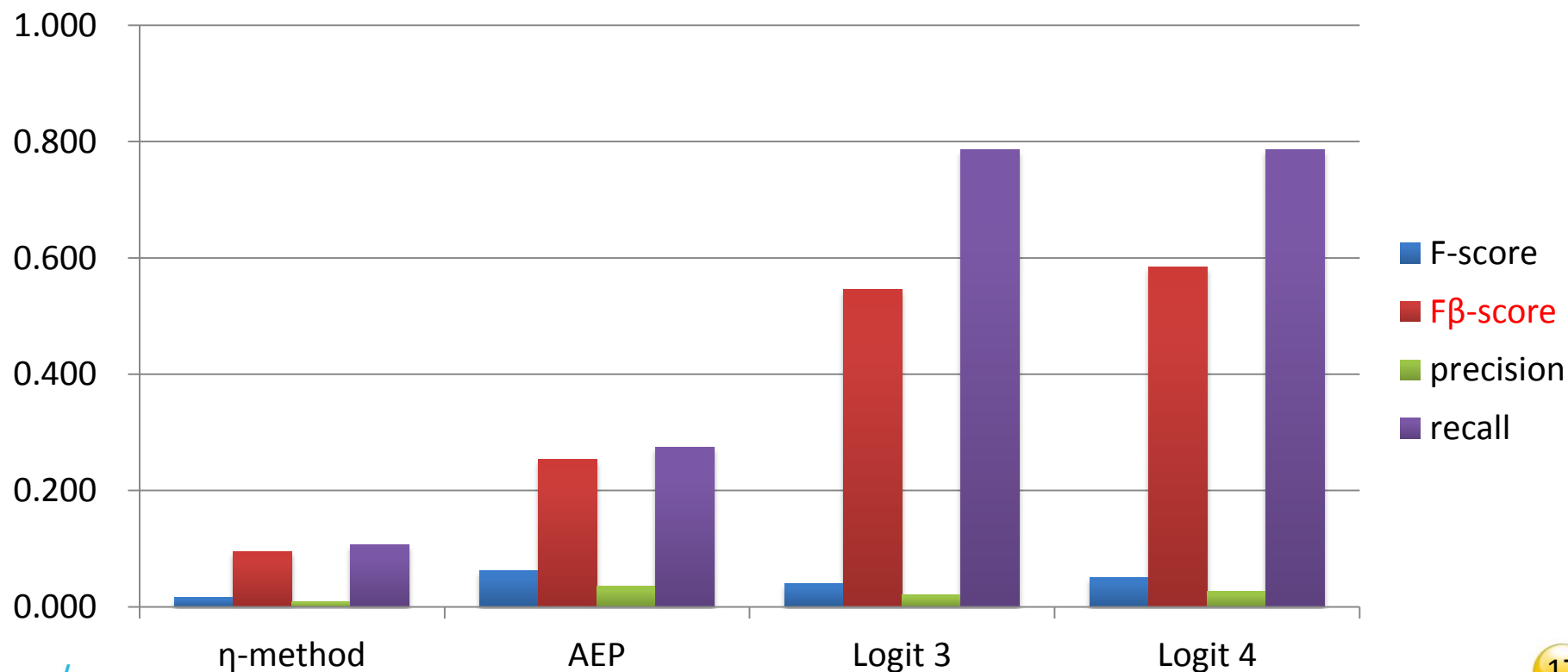




実験② 84 × 84 PPI Prediction

● 比較対象

- η -method (M.Ohue, 2009)
- AEP (T.Yoshikawa, 2009)
- LogitBoost(S_1, A_R, A_L) **Logit3**
- LogitBoost(S_1, A_R, A_L, η) **Logit4**





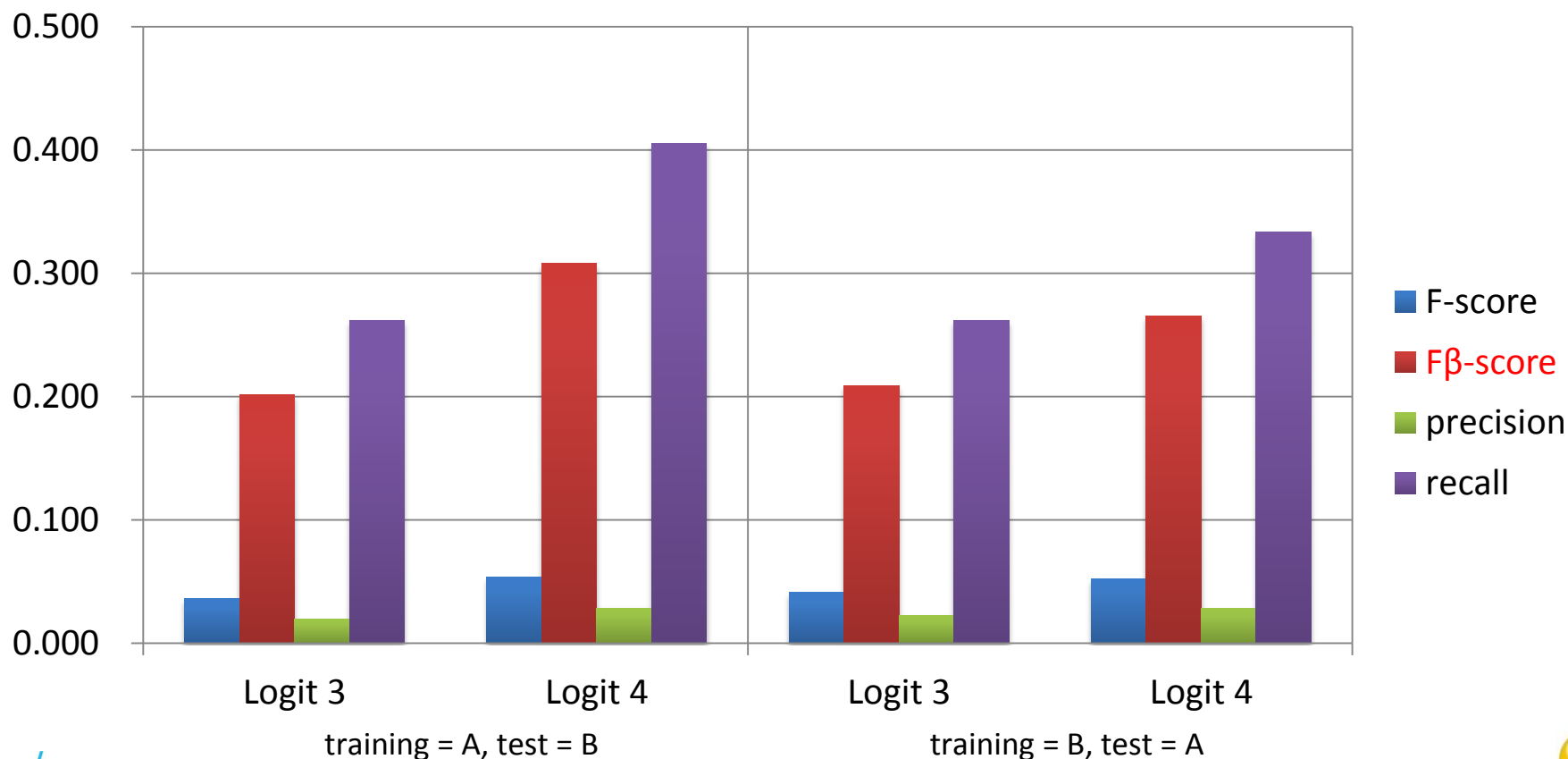
実験③ 42×42 two subset cross validation

● 比較対象

● $\text{LogitBoost}(S_1, A_R, A_L)$ **Logit3**

● $\text{LogitBoost}(S_1, A_R, A_L, \eta)$ **Logit4**

● subset A と subset B を用いて交差検定





● 23 × 23

- F -score, F_{β} -score両方において精度の改善を示す
- Logit 3よりLogit 4の方が高精度

● 84 × 84

- F -scoreではAEPに劣る(AEP $F=0.062$, Logit4 $F=0.051$)
- F_{β} -scoreでは倍以上の精度改善
- Logit 3よりLogit 4の方が高精度

● Cross Validation

- Logit 3よりLogit 4の方が高精度
- LogitBoostである程度の精度が出ることを確かめた
 - trainingサンプルの数(特に正例)を増やしての実験
 - 特徴量の追加を検討



おわりに

- 網羅的PPI予測の性能向上を行った
 - 相互作用判定を機械学習によって構成
 - LogitBoostによる学習, PPD Benchmark 2.0での評価実験
 - 正例負例の偏りを考慮した評価法の検討
 - 従来手法に比べて精度向上
 - 実問題を想定した交差検定では精度はまだ不十分

- さらに向上のために

- 特徴量の検討
- 構造アンサンブルによる正例サンプルの追加
- 学習データセットの検討
 - 実際の生物系, 局在等を考慮したデータセットの構築

