Vol.2016-MPS-108 No.19
Vol.2016-BIO-46 No.19
2016/7/5

ランク学習を用いたNTDs薬剤標的タンパク質候補選択の改良

曽我 光瑛[1,2,a)]　石田 貴士[1,2]

概要：寄生原虫感染症向け薬剤開発では，標的タンパク質候補が既に数百個以上知られており，標的タンパク質探索が大きな問題とならない一方，どのタンパク質を標的タンパク質として選択すべきかという問題がある．現在，寄生原虫のゲノム配列上に存在するタンパク質について構造情報，阻害化合物情報などを集約したデータベースが開発されており，標的タンパク質選択の大きな助けとなっている．しかし，非常に多様な情報が登録されているため，現在提供されている単純な検索機能や絞込機能では望ましい特徴を備えた標的タンパク質を得るのが困難であるという問題が生じている．そこで本研究では，ユーザに一部のタンパク質についてランク付けを行ってもらい，その情報から機械学習の一手法であるランク学習を用いて，ユーザの要求に特化された予測モデルを構築し，大量のタンパク質から望ましい標的候補を推薦する手法を提案する．その結果，ランク付け方針によっては高精度のランク予測が可能であり，訓練セットの数を減らしても高い精度でランクの予測が可能であることがわかった．

キーワード：寄生原虫症，薬剤標的タンパク質選択，ランク学習，シャーガス病，*Tripanosoma cruzi*

Improvement of NTDs drug target protein selection by learning to rank

Mitsuaki Soga[1,2,a)]　Takashi Ishida[1,2]

Abstract: In drug discovery for human parasitic protozoan diseases, high throughput experiments have revealed many drug target protein candidates. Thus, the selecting the most suitable target from them becomes a problem while many drug target protein candidates are available. Currently, several databases provide various information about those protozoa and we can use them for selecting suitable targets. However, the search functions in those systems are insufficient and it is difficult to obtain proteins with desirable features from many candidates. In this study, we proposed a new method to select proteins whose properties are suitable for a user based on learning to rank.

Keywords: Parasite protozoa , Drug target protein selection , Learning to rank , Chagas' disease

1. 導入

マラリアやシャーガス病などの寄生原虫感染症は，原因となる原虫がヒトに寄生することで引き起こされる感染症であり主に発展途上国で蔓延している [1]．主な感染地の経済的理由から薬剤開発研究自体が少なく，新薬もほとんど開発されてこなかった [2] ため一部の寄生原虫感染症は顧みられない熱帯病（Negrected Tropical Diseases, NTDs）[3] と呼ばれている．近年これらの疾患への対策が注目を集めており，創薬研究なども開始されつつある．しかし，これらの創薬研究では経済上の問題から一般的な薬剤開発よりもさらに効率的な薬剤開発を行うことが求められており，構造ベース創薬などの予め薬剤の標的となるタンパク質を同定した上で効率的に薬剤開発を行う手法 [4] の活用が期待されている．このような手法の利用には，まず薬剤の標的タンパク質の同定が必要となるが，寄生原虫感染症の治療薬の場合には寄生原虫の生存に関わるタンパク質は

[1] 東京工業大学 情報理工学院 情報工学系 知能情報コース
[2] 東京工業大学 情報生命博士教育院
[a)] soga@cb.cs.titech.ac.jp

Vol.2016-MPS-108 No.19
Vol.2016-BIO-46 No.19
2016/7/5

すべて薬剤の標的となる可能性があり，それらのタンパク質は大規模な生物学的実験によりかなり多数存在することが知られている [5]．例えば，アフリカ睡眠病の原因となる *Trypanosoma brucei*[6] では，全 7,435 タンパク質中生活環のすべてのステージで機能が失われると寄生原虫の生存に関わるタンパク質が 750 個存在することが知られている [5]．そのため，一般的な創薬で重要となる標的タンパク質の同定が大きな問題とならない一方で，その多数の候補からどのタンパク質を標的とするかが問題となっている．これは，薬剤開発の際にタンパク質によって創薬研究のしやすさが異なりうるため，より望ましい特徴、より多くの関連情報が存在するタンパク質を標的としたほうが効率的な創薬が可能となるためである．

このような理由から、寄生原虫感染症の薬剤標的タンパク質を選択を補助するシステムが開発されている．TriTrypDB[7][8] は 2009 年に Eukaryotic Pathogen Bioinformatics Resource Center(EuPathDB.org)[9][10] と GeneDB[11][12] の研究者たちが共同で開発したデータベースで，寄生原虫感染症の原因となる原虫のゲノム情報や各タンパク質のアノテーション情報を集約している．TDRTargets[13][14] は 2008 年に世界保健機関 (WHO) による国際熱帯病研究特別計画 (Special Programme for Research and Training in Tropical Diseases) によって薬剤標的タンパク質の推薦や優先順位付けを行うことを目的として開発されたもので，ゲノム情報や化合物情報などの多様な情報を扱っている．また、iNTRODB[15] は 2012 年から我々の研究グループで開発されている寄生原虫向け創薬標的タンパク質検索を行うシステムで，特に構造ベース創薬に合った標的選択を目的として，寄生原虫のゲノム情報，タンパク質立体構造情報，化合物情報などを集約している．以上のシステムの問題点としては，登録された情報と検索機能の多さにより目的のタンパク質の検索が困難となっている点があげられる．TriTrypDB では様々な検索、絞込の機能が提供され，TDRTargets では重み付けからターゲットの推薦を行う機能などが備えられている．それらのシステムではタンパク質の検索は属する集団や各情報についてのパラメータの調整によって行われるが，そのような検索では与えられた情報全体について知っていないとパラメータの調整をしづらく，登録情報とそれらについてのパラメータ調整の種類の多さから検索が煩雑である．

本研究では，寄生原虫の一種である *Trypanosoma cruzi* の薬剤標的タンパク質選択について，これまでのシステムで問題となっていたタンパク質選択の困難さの原因であったパラメータ調整などを行わなくても利用者の興味に応じた標的タンパク質候補を選択するシステムの開発を目的とする．

2. 提案手法

これまでのシステムでは多くの情報とそれについての検索機能の煩雑さから標的タンパク質の選択が困難であった．本研究では，これまで大量の情報を扱う情報検索（Information Retrieval, IR）の分野で情報の推薦として用いられてきたランク学習 [16] の手法を薬剤標的タンパク質選択に適用し，この検索を容易にする手法を提案する．これはまずユーザに標的タンパク質候補の中から幾つかのタンパク質についてランク付けをしてもらい，そのタンパク質の順位の情報を学習して，全体の中からユーザにとってより好ましいタンパク質を提示するものである（図 1）．これによって、ユーザの要求に特化された予測モデルを構築することが可能となり，このモデルを用いることで大量のタンパク質から望ましい標的候補を得ることが可能となる．

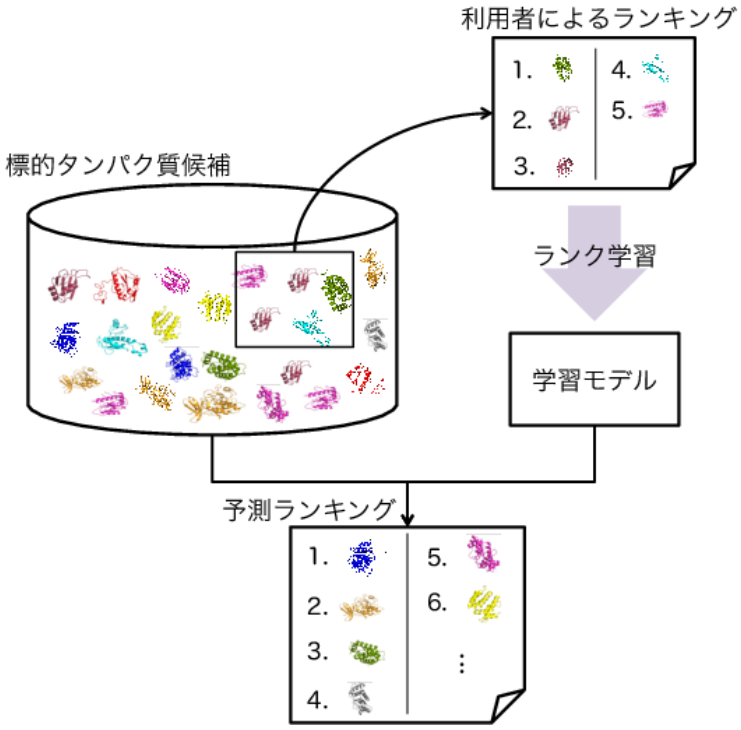

図 1 提案手法の概略図

2.1 ランク学習

ランク学習は主に情報検索の分野で用いられてきた手法 [16][17] で，あらかじめクエリの特徴量に対する関連度の学習モデルを作成し，入力に対してその入力がどのくらい上位に来るのかを関連度として予測する．本研究では，各標的タンパク質候補に付与されたアノテーションの情報を特徴量として用いる．

2.2 特徴量

寄生原虫向け創薬標的検索を行うデータベースであ

る，iNTRODB で得られる情報を用いる．iNTRODB には ChEMBL[18][19] や PDB[20] に登録されているタンパク質との配列相同性情報，ハイスループットでの Phenotyping 実験，RNA-interference target sequencing（RIT-seq）の情報などが収められている．

また，現在公開されているバージョンの iNTRODB には含まれていないが，予測立体構造モデルのアラインメント情報と予測立体構造モデルの信頼性の情報も対象とした．これは，登録されている *Trypanosoma cruzi* の全タンパク質に対して MODELLER[21] による比較モデリングを適用して作成したもので，予測立体構造モデルを構築できた全 10,334 個中 3,728 個のタンパク質のみに与えられる情報である．

2.2.1 対象となるアノテーション情報

iNTRODB では様々なデータベースの情報を集約しおり，以下の情報を寄生原虫のタンパク質を検索するときに得ることができる．本研究では，薬剤候補化合物データベース ChEMBL からは，薬剤標的タンパク質として登録されている Target Protein の中で一番相同性の高かったタンパク質とのアラインメントスコアの E-value（ChEMBL e-value）と，Target Protein の中で最も相同性が高かったタンパク質に関連した実験情報のある化合物の数（ChEMBL ncompound），タンパク質立体構造データベース Protein Data Bank （PDB）からは，PDB に登録されている全タンパク質の中で最も相同性が高かったタンパク質との配列アラインメントスコアの e-value （PDB e-value）とアラインメント長がタンパク質配列長に占める割合である coverage （PDB coverage），生物種を絞って相同性を検索し最も相同性が高かったタンパク質とのアラインメントスコアの e-value として寄生原虫では Leishmania major と *Tripanosoma brucei* ，さらにヒト（Homo sapiens）と哺乳類（Mammal）のタンパク質との相同性情報，さらに生物学的実験の情報としてタンパク質を検索した時にそのタンパク質と最も相同性が高かった *Trypanosoma brucei* のタンパク質の RIT-seq の結果を特徴量として用いた．

2.2.2 予測立体構造情報

MODELLER による比較モデリングでは，テンプレートタンパク質とのアラインメントの情報と作成された予測立体構造の安定性を示すエネルギーの情報が得られる．比較モデリングにおいて予測立体構造の質はアラインメントの質によるため，予測立体構造の信頼性としてアラインメントの情報とエネルギーの情報，予測立体構造を得る際にテンプレートとの間に得られたアラインメントのアラインメントスコアの e-value （e-value），完全に一致している配列数が短い方のタンパク質の配列長に占める割合（Identity），アラインメント配列長がタンパク質配列長に占める割合（coverage）をアラインメントの質として用い，MODELLER の立体構造のエネルギー計算に利用されるエネルギー値（DOPE[22]）と，その値をタンパク質配列長でスケーリングした値（LDOPE）を特徴量として用いた．DOPE は本来 MODELLER が予測立体構造を生成する際に，最終的な立体構造の候補が複数得られた際にその候補間で比較をするためのエネルギースコア [23] であるため，複数のタンパク質間で比較を行うためにはタンパク質の配列長に対する依存を排除しなければならないため LDOPE も立体構造モデルの質を評価するための情報とした．

2.3 学習アルゴリズム

本研究では，ランク学習のアルゴリズムとして Ranking SVM[24] を用いた．Ranking SVM は Support Vector Machine（SVM）を用いて順序予測を行う手法であり，入力ベクトルを順序に応じた関連度に写像するモデルパラメータを学習するために SVM を用いる．主に情報検索（Information Retrieval : IR）分野で用いられてきた手法である．本研究では以上で挙げたタンパク質の情報を入力ベクトルとして，関連度を求めるモデルを ranking SVM の実装の一つである SVM rank [25] によって学習しランク予測を行う．

3. 評価実験

3.1 データセット

寄生原虫の一種である *Trypanosoma cruzi*（*T. cruzi*）のタンパク質のうち，最も相同性が高かった *T. brucei* のタンパク質が RIT-seq で生活環の全ステージで著しい減少（0-0-0-0）または生活環のうち血流内で著しい減少（0-0-1-0）という 2 つの結果が得られたタンパク質を対象にする．その *T. cruzi* に対して，iNTRODB と予測立体構造情報を特徴量として付加する．

今回の実験で用いる特徴量は，それぞれ取りうる値が異なるため，Z-score 化によるスケーリングを行った．Z-score は，

$$Z - score_i = \frac{x_i - \mu}{\sigma}$$

によって計算される．ここで μ は母集団平均，σ は母集団の標準偏差である．iNTRODB では，アノテーション情報が一部得られていないタンパク質が存在するため，母集団平均と母集団標準偏差は情報が得られたものを用いて計算し，Z-score を計算する際には情報を補う必要がある．ChEMBL compounds , PDB coverage , Identity , Coverage , DOPE , LDOPE , Poten Length で情報が得られていないタンパク質には 0 を補い，アラインメントスコアの e-value（ChEMBL evalue , PDB e-value , Leishmania major , Trypanosoma brucei , Homo Sapiens , Mammal , e-value）で情報が得られていないものは，1e+2 を補った．e-value の値は非常に小さい値であり情報が得られていても値が切り捨てられて 0.0 と登録されているものが存

在するが，iNTRODB に登録されている e-value と予測立体構造を作製するときに得られたアラインメントの e-value では小さい値の切り捨て方が違うため，iNTRODB に登録されている e-value（ChEMBL e-value，PDB e-value，Leishmania major，Trypanosoma brucei，Homo sapiens，Mammal）には 10^{-200} を，比較モデリングを行う際のアラインメントスコアの e-value（e-value）には 10^{-30} をそれぞれ 0.0 の代わりに補った．

以上の値を特徴量とした *T. cruzi* のタンパク質から，訓練セット 50 個，テストセット 20 個をそれぞれランク付けしたものを用いる．

3.1.1 データセットの順位付け

訓練データは 5 段階で順位付けを行い，テストデータはすべてのタンパク質に異なる順位を付与した．実際の順位付けでは，好ましい順の昇順になるが，学習や予測の評価の際には降順の関係である関連度のほうが望ましいため，i 番目に好ましいと思った順位が上から i 番目の大きさの値になるように順位スコアを定義した．具体的には，訓練セットでは最も望ましい物が 5 で最も望ましくないものが 1 になるような降順になっていて，テストセットでは最も望ましい物が 20 の降順になるようにした．

今回行った順位付けの方針は以下のとおり

- A. 予測立体構造を重視したランク付け．
  構造モデルの質の高さと，それを得るための配列相同性の高さだけを考慮し，残りは無視して順位を付けており，学習は容易であると考えられる．
- B. 製薬会社らとの創薬共同研究で使用した基準に基づくランク付け．
  質の高い構造モデルが得られていることを重視し，副作用を考慮してヒトのタンパク質との配列相同性ができるだけ低いものを優先し，化合物実験関連の情報はあまり重視しない．

3.2 実験方法

3.2.1 パラメータ探索

訓練データセットの三分割交差検定によってカーネル関数（線形カーネル，ガウスカーネル）とハイパーパラメータを定めた．対象となるハイパーパラメータは誤認識率とマージンの比重である C と，rbf カーネルの式

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$$

の中に出てくる変数 γ である．これら 2 つのハイパーパラメータに対し，それぞれ

$$C = \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$$
$$\gamma = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$$

の範囲でパラメータ探索を行い評価をする．

3.3 評価方法

予測の評価として、情報検索などの性能評価などに用いられる $nDCG$（Normalized Discounted Cumulated Gain）[26] と、順位の相関の指標としてスピアマンの順位相関係数を用いた．

3.3.1 $nDCG_k$(normalized Discounted Cumulative Gain)[26]

DCG_k は，予測された順序の i 番目の関連度を rel_i として、

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \tag{1}$$

と計算される．この指標は関連度が高い値を示すものが上位に来ると大きな値となる．
DCG_k の計算には

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{2}$$

という計算式も用いられる．式 2 は上位に関連度の高いものが正しい順序で予測されたことを式 1 より強く評価する．その場合、関連度の高いものが上位に一つあるだけで、その他の上位に予測された関連度の評価が低くともそれが評価の高さに大きく影響する．一方、式 1 の場合は一つ一つの上位に予測されたものが大きくは影響しないが、関連度が高いものが全体として上位にあることがより評価の高さに影響する．今回は、一番よいタンパク質を一つ選択すればよいのではなく望ましいタンパク質はなるべく上位に予測したいので式 1 を用いて DCG_k を計算した．

取りうる最大の DCG_k を $IDCG_k$ としたときに

$$nDCG_k = \frac{DCG_k}{IDCG_k} \tag{3}$$

と表され、予測されたランキングが完全に一致した時に 1 となる．今回，テストセットでは 20 個のタンパク質について順位付けを行ったが，上位によりよいものが来るのが望ましいので順位付けの評価をする際は上位 10 個だけを見る $nDCG_{10}$ で評価を行った．

3.3.2 スピアマンの順位相関係数

スピアマンの順位相関係数は，順位データから求められる相関の指標であり，比較する 2 つの順位データの分布に何も仮定せずに用いることができる．スピアマンの順位相関係数 ρ は、順位データ x と y の i 番目の順位をそれぞれ x_i、y_i、値のペアの数を n とすると

$$\rho = 1 - \frac{6\sum (x_i - y_i)^2}{n^3 - n} \tag{4}$$

によって計算される．

4. 結果

4.1 訓練セットを全て用いた場合

cross validation の分割の仕方を変えて 5 回実行して

$nDCG_{10}$ の値が中央値となった予測順位を示している．

表 1 case:A

予測関連度	順位スコア
2.19	20
1.73	19
1.61	14
1.56	16
1.50	18
1.21	13
0.993	11
0.985	12
0.952	15
0.782	17

表 2 case:B

予測関連度	順位スコア
1.09	13
1.07	15
0.686	16
0.659	12
0.628	14
0.548	17
0.513	7
0.365	10
0.163	18
0.132	20

A は $nDCG_{10}$ の値が 0.982 の場合のランク付けの結果で（平均： 0.976 標準偏差： 0.0125），この時カーネル関数は rbf カーネルが選択され，ハイパーパラメータは $\{C, \gamma\} = \{0.01, 0.1\}$ であった．また，スピアマンの順位相関係数は $\rho = 0.515$ であり，精度の良い順位付けであるといえる．ランク付けの方針が単純であれば，順位付けも精度よく行うことができるといえる．

B は $nDCG_{10}$ の値が 0.844 の場合のランク付けの結果で（平均：0.842 標準偏差：0.00482），この時カーネル関数は線形カーネルが選択され，ハイパーパラメータは $C = 0.01$ であった．また，スピアマンの順位相関係数は $\rho = -0.285$ であり，弱い負の相関が見られた．ランク付け方針の複雑さをうまく学習できていない結果であると考えられる．

4.2 訓練セットの数を減らした場合

case A の場合，つまり単純な基準でのランク付けの場合には良い精度の予測結果が得られたが，ランク付けのコストを減らした場合にどうなるかを調べた．具体的には，訓練セットの数が 30，10 の場合にテストセットの順位付けにどう影響が出るかを調べた．

表 3 訓練 30

予測関連度	順位スコア
0.17363	14
0.17360	20
0.14900	16
0.13965	18
0.12400	19
0.11591	15
0.11516	17
0.08804	13
0.06008	12
0.03352	10

表 4 訓練 10

予測関連度	順位スコア
1.76	20
1.25	14
1.15	16
1.07	18
1.03	19
0.813	15
0.744	17
0.673	12
0.634	13
0.353	11

訓練 30 は，$nDCG_{10}$ の値が 0.958 の場合のランク付けの結果で（平均：0.945 標準偏差：0.0521），この時カーネル関数は rbf カーネルが選択され，ハイパーパラメータは $\{C, \gamma\} = \{0.001, 0.01\}$ であった．また，スピアマンの順位相関係数は $\rho = 0.636$ であった．

訓練 10 は $nDCG_{10}$ の値が 0.909 の場合のランク付けの結果で（平均： 0.932 標準偏差： 0.0333），この時カーネル関数は線形カーネルが選択され，ハイパーパラメータは $C = 0.001$ であった．また，スピアマンの順位相関係数は $\rho = 0.697$ であった．

訓練データを減らした場合でも，順位相関係数の値は悪化しなかったが $nDCG_{10}$ の値は僅かに減少したが，訓練データセットの数が case B の五分の一であってもより精度の高い予測順位が得られた．

5. 考察

異なる方針による 2 つの順位付けによる比較では，単純な順位付けのほうが順位の予測精度が良いという結果が得られた．どちらの場合にも上位 10 個のランクと予測されたものには上位 10 個の順位スコアを持つタンパク質が多く含まれる結果となったが，順位の相関という点では複雑な順位付けは悪い結果となった．今回順位付けの精度の評価として用いた $nDCG_{10}$ の計算には式 1 を用いたが，式 2 のほうが上位に高い順位スコアを持つものが来た時により強く評価する．そこで，評価を式 2 で行い学習モデルを作製することで正の相関を持つランク付けを期待したが，スピアマンの順位相関係数で $\rho = -0.0667$ と無相関という結果が得られた．

6. まとめ

6.1 まとめ

本研究では，NTDs の薬剤開発における標的タンパク質の選択という問題に対して，ランク学習を用いる効率的な標的タンパク質の提案を試みた．単純な方針でのランク付けでは $nDCG_{10}$ の値と正の相関を持つ順位の予測を行うことができたが，実用で使うような複雑な方針でのランク付けにおいては現状では精度に課題があることがわかった．また，ランク付けの方針によっては訓練データの個数が少ない場合でも正しい順位の予測が可能であることが示された．

6.2 今後の課題

本研究の課題としてタンパク質全体についてのランク付けと，訓練データのランク付けに対する工夫という二点が挙げられる．

今回の実験では予め用意したテストケースについてのランク付けであったが，実際にデータベースなどで利用する際には，作成した学習モデルでの生物種ごとのタンパク質全体に対するランク付けが想定される．学習モデルの作成

方法やどのように順位を評価するかなどは今後の課題としたい．

ランク付けの方針によっては訓練データの個数を少なくしても利用者によっては正しい順位の予測が可能であることが示されたが，ランク付けのコストを考えるとより少ない数で訓練できたほうが利用者にとって好ましいと言える．利用者の方針によってランク付けの個数を変化させるか，方針によって特徴量の重みを変化させるなどの方法が考えられる．また，今回は 5 段階のランクによって評価を行ったが訓練データのランク付け方法も今後の課題としたい．

謝辞 本研究は JSPS 科研費 15K16082 の助成を受けたものである．

参考文献

[1] World Health Organization,
入手先 ⟨http://www.who.int/en/⟩（参照 2016-02-04）
[2] Yoshino, R., Yasuo, N., Inaoka, D.K., Hagiwara, Y., Ohno, K., Orita, M., Inoue, M., Shiba, T., Harada, S., Honma, T. and Balogun, E.O..:Pharmacophore Modeling for Anti-Chagas Drug Design Using the Fragment Molecular Orbital Method,*PLos ONE*,10(5),e0125829,2015.
[3] Eisai ATM Navigator,
入手先 ⟨http://atm.eisai.co.jp/ntd/⟩（参照 2016-02-04）
[4] 長野哲雄, 岩槻壮市, 高木淳一, 古谷利夫編: 融合発展する構造生物学とケミカルバイオロジーの最前線, 共立出版株式会社 (2009).
[5] ALSFORD, Sam, et al.:High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome,*Genome research*, 21.6: 915–924,2011.
[6] Tollitt,M E.:Trypanosoma brucei Plimmer & Bradford,*Bulletin of Zoological Nomenclature*,43:348–349,1986.
[7] Aslett, Martin, Cristina Aurrecoechea, Matthew Berriman, John Brestelli, Brian P. Brunk, Mark Carrington, Daniel P. Depledge *et al.*:TriTrypDB: a functional genomic resource for the Trypanosomatidae,*Nucleic acids research*,38: D457–D462,2010.
[8] TriTrypDB Kinetopiastid Genomicd Resource,
入手先 ⟨http://tritrypdb.org/tritrypdb/⟩（参照 2016-02-04）
[9] Aurrecoechea, Cristina, *et al.*:ApiDB:integrated resources for the apicomplexan bioinformatics resource center,*Nucleic acids research*,35,D427–D430,2007.
[10] EuPathDB Eukaryotic Pathogen Database Resources,
入手先 ⟨http://eupathdb.org/eupathdb/⟩（参照 2016-02-04）
[11] Logan-Klumpler,Flora J.,*et al.*:GeneDB-an annotation database for pathogens,*Nucleic acids research*,40.D1,D98–D108,2012.
[12] GeneDB,
入手先 ⟨http://www.genedb.org/Homepage⟩（参照 2016-02-04）
[13] Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, Crowther GJ, Doyle MA, Hertz-Fowler C, Hopkins AL, McAllister G, Nwaka S, Overington JP, Pain A, Paolini GV, Pieper U, Ralph SA, Riechers A, Roos DS, Sali A, Shanmugam D, Suzuki T, Van Voorhis WC, Verlinde CL.:Genomic-scale prioritization of drug targets: the TDR Targets database.,*Nature Reviews Drug Discovery* 7,900–7,2008.
[14] TDRTargets,
入手先 ⟨http://tdrtargets.org/⟩（参照 2016-02-03）
[15] iNTRODB Integrated Neglected TROpical disease DataBase,
入手先 ⟨http://www.bi.cs.titech.ac.jp/introdb/⟩（参照 2016-02-04）
[16] Hang, L. I. .:A short introduction to learning to rank,*IEICE TRANSACTIONS on Information and Systems*,94,10,1854-1862,2011.
[17] T.Y. Liu.:Learning to rank for information retrieval,*Foundations and Trends in Information Retrieval*,vol.3,no.3,225-331,2009.
[18] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J.P. Overington.:The ChEMBL bioactivity database: an update,*Nucleic Acids Res*,42,1083–1090,2014.
[19] ChEMBL,
入手先 ⟨https://www.ebi.ac.uk/chembl/⟩（参照 2016–06–01）
[20] RCSB PROTEIN DATA BANK,
入手先 ⟨http://www.rcsb.org/pdb/home/home.do⟩（参照 2016-06-01）
[21] A. Sali,T.L. Blundell.:Comparative protein modelling by satisfaction of spatial restraints.,*J. Mol. Biol.*, 234, 779-815, 1993.
[22] M.Y. Shen, A. Sali.:Statistical potential for assessment and prediction of protein structures.*Protein Sci 15*, 2507–2524, 2006.
[23] Andrés Colubri,Abhishek K. Jha,Min-yi Shen, Andrej Sali, R. Stephen Berry,Tobin R. Sosnick,Karl F. Freed,.:Minimalist Representations and the Importance of Nearest Neighbor Effects in Protein Folding Simulations,*Journal of Molecular Biology*,363,835–857,2006.
[24] Herbrich, Ralf, Thore Graepel, and Klaus Obermayer,.:Support vector learning for ordinal regression.*Artificial Neural Networks ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*,Vol.1, 1999.
[25] T. Joachims,.:Optimizing Search Engines Using Clickthrough Data,*Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*,2002.
[26] Kalervo Jarvelin, Jaana Kekalainen,.:Cumulated gain-based evaluation of IR techniques.,*ACM Transactions on Information Systems 20*,4, 422–446,2002.