秋山研究室新入配属生向けバイオインフォマティクス実習課題

2019年度版　作成：大上雅史

以下の諸注意を読み，１～４の課題を実施せよ．

※処理を行う言語については特に指定をしないが、Perl，Python，Ruby等のスクリプト言語を用いることが望ましい。特に好みがなければPythonを勧める。anacondaで必要なものは大概揃う。

※プログラミング、実行等の作業はできるだけ秋山研のサーバ (ghostgw, 192.168.1.180) を利用すること。

※「バイオインフォマティクスのためのPerl入門（Tisdall著，水島監訳）」の第5章，8章，9章，11章を参考にした。必要に応じて参照せよ。

１．配列情報処理

1. ヒト21番染色体の先頭コンティグのDNA配列を秋山研サーバ/mnt/fs/ohue/newcomer/NT\_113952.1.fasta から取得し，A/T/G/Cの各塩基数を出力せよ．
2. NT\_113952.1.fastaの逆相補鎖を生成するプログラムを作成せよ．
3. ウィンドウ幅*w*，ステップ幅*s*で，ウィンドウ内のGC含量を出力するプログラム（図も参照のこと）を作成し，NT\_113952.1.fastaに*w* = 1000, *s* = 300で適用した結果をgnuplot, R, python (matplotlib) 等，Microsoft Excel以外のツールでプロットせよ．



図: *w* = 5, *s* = 3の例

1. 引数で与えられた部分配列を検索し，何文字目に現れたかを表示するプログラムを作成せよ（逆相補鎖上も検索すること）．部分配列をGAATTCおよびATGとし，NT\_113952.1.fastaに適用せよ．
2. NT\_113952.1.fastaを6つの読み枠でアミノ酸配列に変換せよ．Stopコドンはアンダースコア‘\_’で表示すること．

２．タンパク質構造情報処理

1. PyMOLソフトウェアを利用し，ヒトのヘモグロビン (ID: 1BUW) の構造をチェインごとに異なる色で表示し，4量体であることを確認せよ．構造データはProtein Data Bank (PDB) より取得すること．また，Aチェインだけを表示し，タンパク質鎖をcartoon表示して2次構造に従って色付けし，結合するヘムをstick表示して原子ごとに色分けせよ．これらの2つの表示結果はpng形式で画像出力すること．

※PyMOLソフトウェアは無償で利用可能である．WindowsではEducational版バイナリを取得するか、Pythonによってソースビルドすること．Mac/Linuxはhomebrew, apt, yum等のパッケージ管理ツールかソースビルドでインストールすること．

1. PDBファイル名とチェイン名を引数にして，その回転半径 (Radius of Gyration) を計算するプログラムを作成せよ．  
   　※回転半径は高分子の大きさを示す為に用いられる値で重心からの平均半径のこと．ここでは全て  
   　　の原子ではなく，Cα原子(CA)のみから近似的に計算して良い．
2. 上記のプログラムの結果を利用し，PyMOLで重心から回転半径の範囲内にある原子を赤で，範囲外にある原子を青で色づけせよ．

※PyMOL上で色がついて表示できていれば，その実現の仕方はどのような方法でも構わない．  
外部ファイルを作成したり，PyMOLコマンドを発行するスクリプトを書いたりなど，各自工夫せよ．

３．機械学習

間欠泉のデータセットionosphereを/mnt/fs/ohue/newcomer/ionosphere.scaleから取得せよ．

1. Pythonのscikit-learnのSVMでionosphereデータの10-fold Cross Validationを実施せよ．予測性能としてprecision, recall, MCC, ROC曲線のAUC値 (AUROC)，F-scoreを求めよ．カーネルはRBFカーネルとし，パラメータは適宜定めよ．
2. ionosphereデータにおいて，より良いRBFカーネルパラメータ*γ*とコストパラメータ*C*の値を探索せよ．評価方法は10-fold Cross Validationとし，評価基準はAUROCとF-scoreの2通りを試すこと．

４．創薬情報処理・機械学習

1. （準備1）PythonにRDKitをインストールし、化合物ファイル (SDFファイル) を読み込んで構造式が出力できることを確認せよ。（anacondaを利用するのが楽かも）
2. （準備2）論文Leung SSF, *et al*. *J Chem Inf Model* 56: 924-929, 2016. (<https://pubs.acs.org/doi/10.1021/acs.jcim.6b00005>) のSupporting Informationより、2D SDFファイル (TXT) とPDFファイルをダウンロードせよ。Table S1の実験値「RRCK Log *P*app」の値をパース（転記）し、CSVファイル等で準備せよ。（このあと、Data Set 1, 3, 4, 8を利用する。）
3. （準備3）所望の化合物に対し、RDKitのECFP4 fingerprint（GetMorganFingerprint()でradius=2とする）を計算できるようにせよ。（この後、機械学習の特徴ベクトルとして用いる。）
4. RRCK Log *P*appを目的変数、ECFP4 fingerprintを特徴ベクトルとして、Data Set 3 (医薬品104化合物) に関して回帰（10-fold Cross Validation）を行う機械学習プログラムを作成せよ。学習器はSupport Vector Regressionとし、カーネルはRBFカーネルとすること。
5. 10-fold Cross Validationによって、RRCK Log *P*appの予測値との平均2乗誤差 (RMSE) が最も小さくなるパラメータを探索せよ。またそのときのRMSEと*R*2値を求めよ。
6. (15)で決めたパラメータの学習器で、Data Set 1 (環状ペプチド7化合物)、Data Set 4 (環状ペプチド16化合物)、Data Set 8 (医薬品様分子22化合物) のRRCK Log *P*appの予測値を求めよ。それぞれのData SetごとにRMSEと*R*2値を求め、予測値と実験値の散布図を描け。