28

Linked Open Dataと配列データを組み合わせた配列分類手法

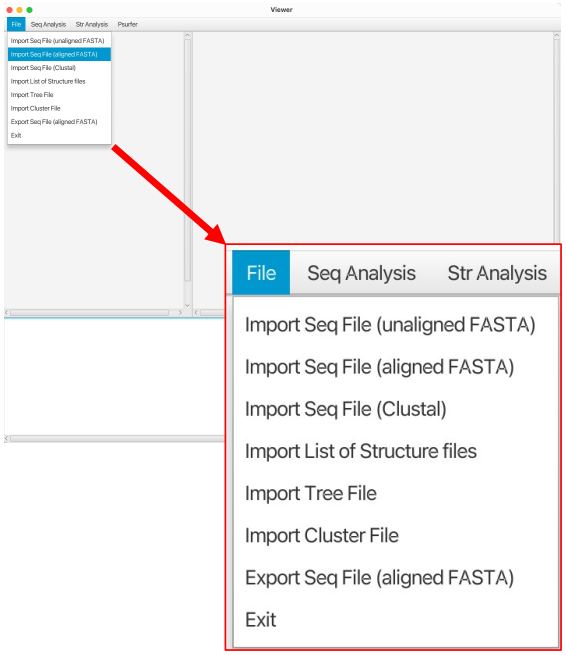
トーゴーの日シンポジウム2021

◯藤　博幸[1]、 山口　敦子[2]

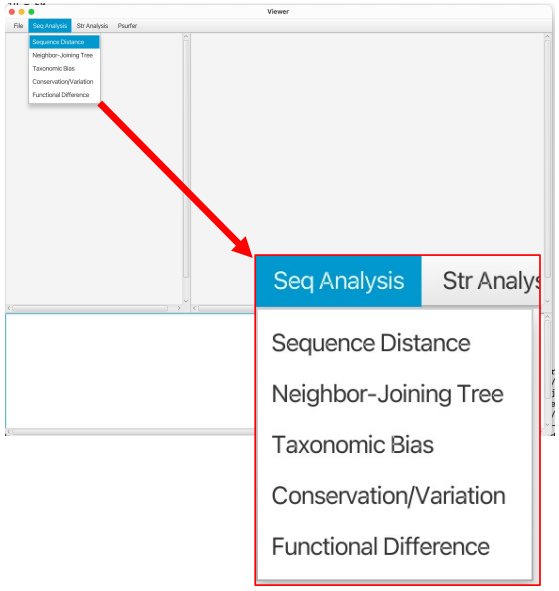
1. 関西学院大学生命環境学部、2. 東京都市大学大学院総合理工学研究科

近年、タンパク質のアミノ酸配列や立体構造データが大量に蓄積され、それらからのデータマイニングの大きな基盤が形成されてきている。一方、そのようなデータの大量性は、従来のアラインメントや系統樹から、それらのデータの持つ意味の把握やそれに基づく知識抽出を困難にしている。この問題を解決するために、新たなアラインメントビューアの開発を行なっている。
このビューアの大きな特徴の一つは、Linked Open Dataを探索し、そこから得られた情報を系統樹や配列の上に可視化できる点にある。今回、ビューアの新機能として、これらのLOD情報とアラインされている配列情報を組み合わせて、配列をクラスタ分類する手法を開発した。LOD情報の内、分子機能、生物学的プロセス、細胞の構成要素から、Jaccard係数あるいはSimpson係数に基づく距離行列を、また生物の階層情報からはハミング距離に基づく距離行列を作成した。
これらを、配列間距離の距離行列と組み合わせて３階のテンソルを構築し、テンソル分解を行なった。分解した結果をx-means法でクラスタ数まで含め、自動分類できるようにした。本手法について議論する.

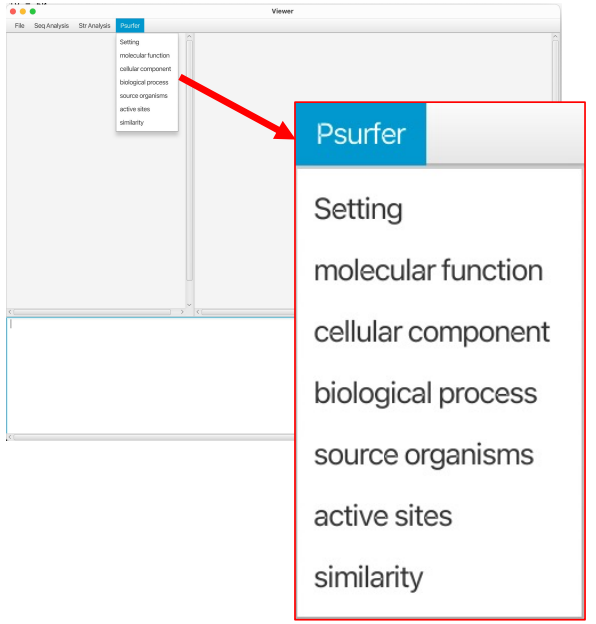
Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）



配列データを読み込む.
アラインメントされていない
場合は内部でmafftを実行し
アラインメントを作成する.

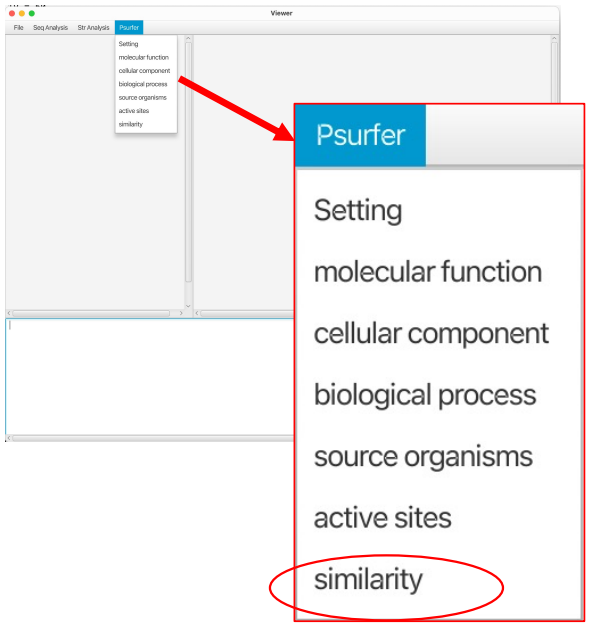


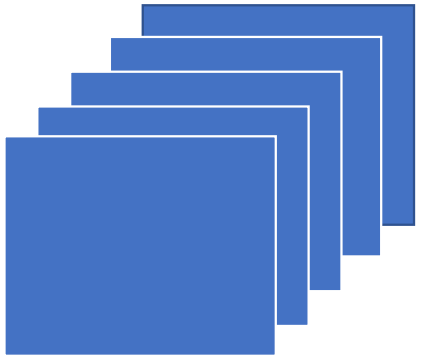
アラインメントから距離行列
を計算する.



アラインメントの各配列のUniProt ID
を用いて、PsurferでGO termおよび
Sourceの情報を収集

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）





(1) 配列データから得られた距離行列を類似度行列に変換

(2) タンパク質間のGO termの共有度をJaccard係数あるいはSimpson係数でもとめ、それを要素とした類似度行列を求める

Molecular Function, Biological Process, Cellular Componentのそれぞれについて求め、３種の類似度行列を得る。

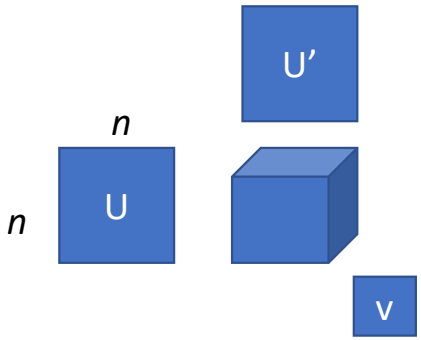
Jaccard 係数 = |A∩B|/|A∪B|
Simpson係数= |A∩B|/min(|A|, |B|)

(3) Psurfer[1]で得られた種名から界門綱目科属種の情報をViewer内部で求め、タンパク質間での階層分類のハミング距離を求め、それを類似度に変換した行列を求める

(4) 得られた5個の類似度行列を3階のテンソル $(n \times n \times 5)$ にまとめ、HOSVDでテンソル分解する。nは配列の本数
HOSVD部分はRをViewer内部で呼び出して計算

1. Yamaguchi, Toh Implementing LOD Surfer as a Search System for the Annotation of Multiple Protein Sequence Alignment, The 8th Joint International Semantic Technology Conference, Lecture Notes in Computer Science vol. 11341, 418-426, Awaji, Japan, 26-28 Nov. 2018

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）



対称行列で作成されたテンソルのため U = U’
Uの要素を使ってxmeansでクラスタ分析
x-meansのRのコードはhttp://www.rd.dnc.ac.jp/~tunenori/src/xmeans.prog より入手

乱数が使用されるので、100回クラスタ分析を実施し、タンパク質ペアが同じクラスタに分類される数をカウント

カウント数が閾値以上のペアは同じクラスタに属すものとした。
閾値はカウント数の四分位数を用いた。

以下、lysozyme familyに適用した結果を示す

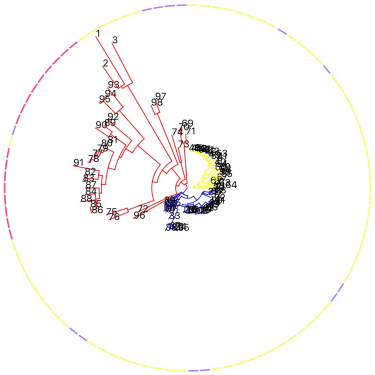
※ 類似度行列をテンソルにまとめ、テンソル分解に基づいてクラスタ分析を行う手法は蔵川、馬場 (2016)の著者名寄せの方法[2]に従った

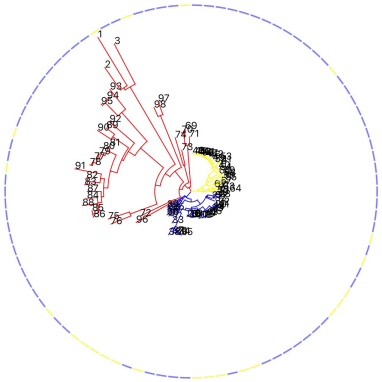
2. 蔵川圭、馬場康維　(2016) 情報処理科学会第78回全国大会 7D-01 1-495

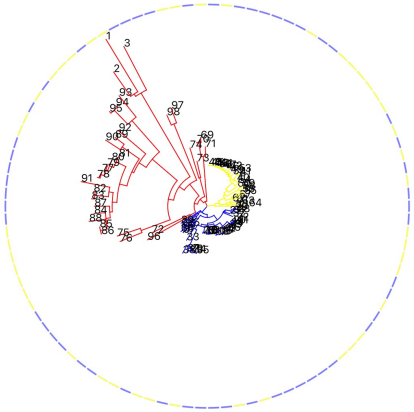
Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）

閾値＝第一四分位数

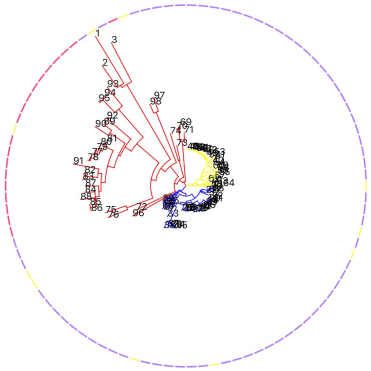
Jaccard係数

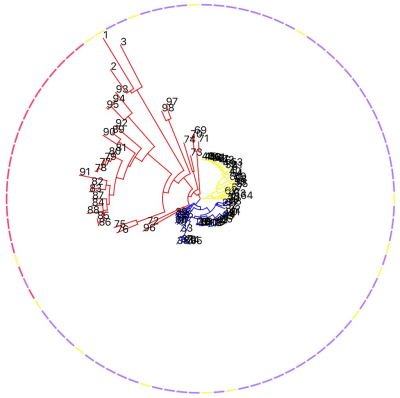


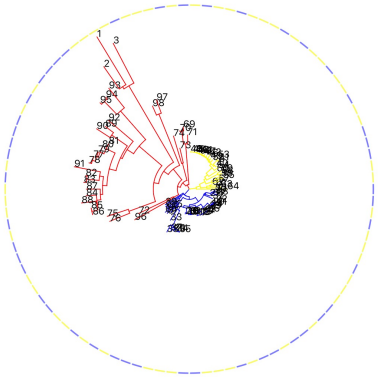




Simpson係数







全情報

GO term+配列

分類情報＋配列

全情報を使うとαラクトアルブミンのクラスタを同定できる
Jaccard係数よりも、Simpson係数の方が見落としが少ない

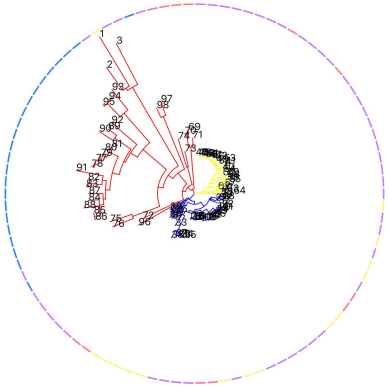
GO情報と配列情報だけだと、近縁なライソザイムをαラクトアルブミンが一つのクラスタになってしまう。

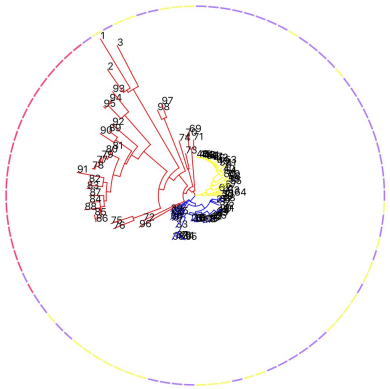
分類情報と配列情報だけだと、αラクトアルブミンは一つのクラスタにまとまらない。

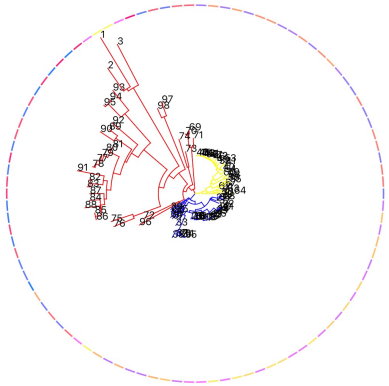
ライソザイムのクラスタは系統樹上まとまらない。

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）

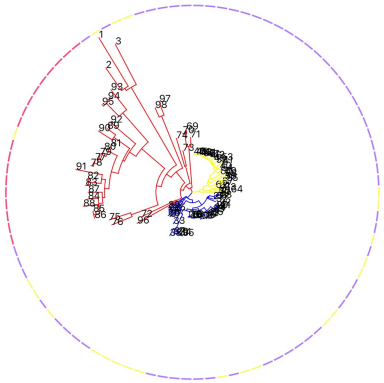
閾値＝第二四分位数
Jaccard係数

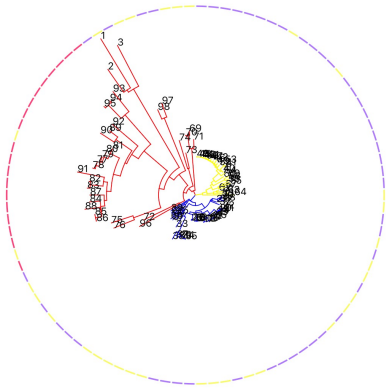


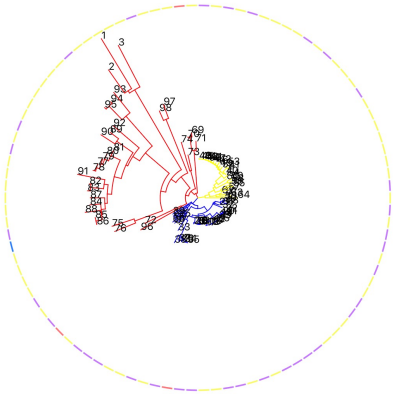




Simpson係数







全情報

GO term+配列

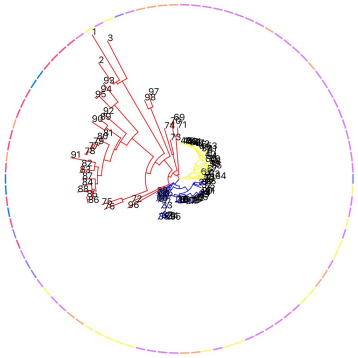
分類情報＋配列

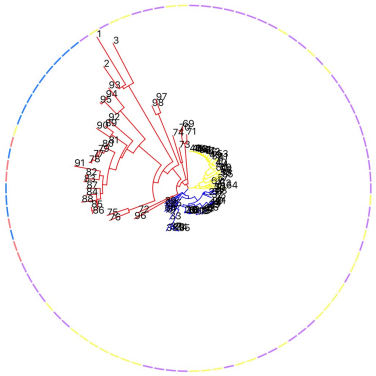
全情報を用いた場合、Jaccard係数はαラクトアルブミンのクラスタにライソザイムが含まれるが, Simpson係数の場合は、αラクトアルブミンのみでクラスタが形成そのクラスタに含まれないαラクトアルブミンがある。
GO termと配列情報を用いた場合、αラクトアルブミンのクラスタにライソザイムが含まれる。
分類情報と配列情報を用いた場合αラクトアルブミンは一つのクラスタにまとまらない。

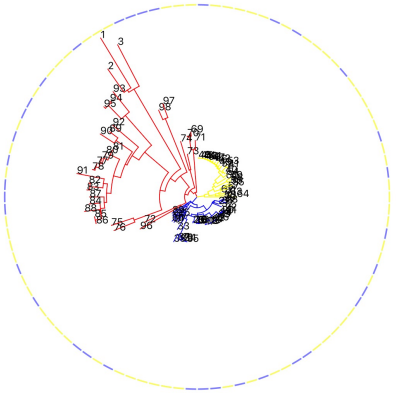
ライソザイムのクラスタは系統樹上まとまらない。

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）

閾値＝第三四分位数
Jaccard係数

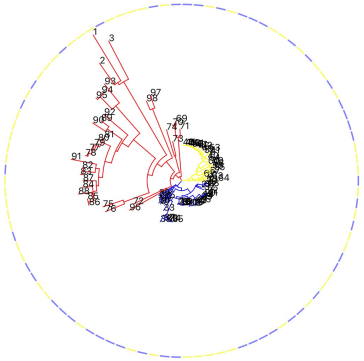


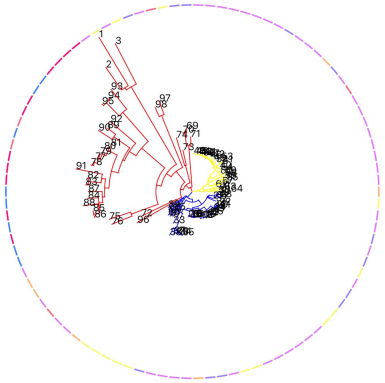


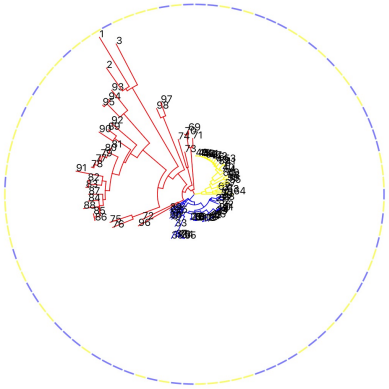


クラスタが多くなるが、
系統的に近縁な部分で
形成されるものがない

Simpson係数







全情報　　GO term+配列　　分類情報＋配列

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）

結論

全情報を用い、Simpson係数で類似度行列を作ると、αラクとアルブミンのクラスタを捉えやすくなるが
自動分類としてはまだ不十分である。

GO termが十分に割り当てられていないタンパク質や、そのタンパク質にのみ割り当てられた情報があることが原因と考えられる。
配列類似度で重みづけして、GO termを割り当ててやることで類似度をならしてやることを検討

Licensed under a Creative Commons 表示4.0国際ライセンス
©2021藤　博幸　（関西学院大学生命環境学部）