PROJET LONG

Sujet: Reconnaissance de repliements d'Unité Protéique par apprentissage automatique des propriétés des acides aminées de familles alignées

Proposé par Jean-Christophe Gelly

L'idée est de réaliser une méthode de reconnaissance de repliement basée sur un algorithme d'apprentissage automatique.
La reconnaissance de repliement se fera sur des Unités Protéiques, un échelon de taille inférieure aux domaines et pertinent pour la prédiction de structures protéiques.
La base d'apprentissage sera constituée à partir de l'alignement de séquence des Unités Protéiques. La classe des unités protéiques choisi sera la classe des unités protéiques de taille 25 à 35.

Deux jeux de données seront donc utiles :

- Un jeu de données d'alignement positif, c'est à dire contenant des alignements pertinents
- Un jeu de données d'alignement négatif, c'est à dire contenant des alignements non pertinents.

Chaque séquence d'Unité Protéique, issue d'un alignement multiple, sera transformé en vecteur de propriétés issues de la banque AAINDEX (Kawashima et al 2008).

Construction des jeux de données

Alignements positifs des Unités Protéiques :

Ils sont disponibles à cette adresse :
http://www.dsimb.inserm.fr/~gelly/data/orion_files.tar.gz

Pour chaque sous famille, il y a un répertoire contenant :
- un fichier contenant l'alignement de séquences de la sous famille basé sur un alignement : *.fas. Il s'agit des données d'apprentissage de classe positive (ex : upload_orion_files/2PTRA_10_33_58/2PTRA_10_33_58_mafft2.fas).
Il faudra filtrer cet alignement pour que chaque séquence contienne moins de 30% de gap.
Ne seront pris en compte dans cette étude que les familles d'Unités protéiques de taille 25 à 35.
Les unités à prendre en ligne en compte sont ici :
http://www.dsimb.inserm.fr/~gelly/data/PU_25_35.mcl_out_rscop

Chaque ligne est un groupe d'Unités protéiques. Les familles d'Unités Protéiques présentant des similarités structurales sont sur la même ligne qui représente un groupe.
Si notre méthode fonctionne, on devrait donc être capable de détecter avec l'algorithme d'apprentissage automatique cette similarité.

Pour les besoins de cette étude, seuls les groupes compris entre les lignes 10 et 20 seront pris en compte.

Alignement négatif des Unités Protéiques :

Il est nécessaire de générer des alignements négatifs
Génération des alignements négatifs :
Pour chaque famille d'Unités Protéiques, on va aligner cette famille à toutes les autres familles n'appartenant pas au même groupe, c'est à dire aux familles n'étant pas sur la même ligne dans le fichier PU_25_35.mcl_out_rscop

Le programme mafft sera utilisé (à télécharger ici :
http://mafft.cbrc.jp/alignment/software/).

Il sera lancé avec la commande suivante :
ginsi --thread 4 --addfull PU_negative.fas --keeplength PU_positive.fas >tmp

Ex :
ginsi --thread 4 --addfull upload_orion_files/1PN2A/1VPBA_10_1_31_mafft2.fas --keeplength
upload_orion_files/1PN2A/1PN2A_10_235_261/1PN2A_10_235_261_mafft2.fas
>output_alignement_neg_1VPBA_10_1_31_1PN2A_10_235_261.fas

Cet alignement sera filtré pour éliminer les séquences correspondant à la famille d'Unités Protéiques positives (il s'agit des n première séquences, avec n les nombre de lignes du fichier PU_positive.fas)

Choix des AAINDEX représentatifs :

D'après l'étude de Van Westen et al. il existe 58 propriétés AAINDEX plus pertinentes que les autres.
Elles sont présentées en annexe 1.

Ce seront les propriétés utilisées dans le reste de cette étude. Il sera nécessaire de les extraire à partir du fichier « aaindex_selected »
(http://www.dsimb.inserm.fr/~gelly/data aaindex1_reformated)

Traduction des alignements de séquences en profil AAINDEX

Le programme fasta2 vector_wgap.pl sera utilisé.
Ex : ./fasta2vector_wgap.pl
upload_orion_files/3IPIA_10_209_224/3IPIA_10_209_224_mafft2.fas selected_ aaindex1_reformated >output
Ce programme est disponible à l'adresse : http://www.dsimb.inserm.fr/~gelly/data/ fasta2vector_wgap.gz

Pondération des séquences

Les séquences familles d'Unités Protéiques peuvent être partiellement redondantes. Pour éviter que le poids de certaine sous-famille de séquences ne soient trop important, elles seront pondérés en utilisant le programme compute_weight_sequence_position.pl (disponible ici : http://www.dsimb.inserm.fr/~gelly/data/weight.tar.gz).
Ex : ./compute_weight_sequence_position.pl 3IPIA_10_209_224_mafft2.fas
La pondération de la séquence est donnée par la valeur « Normalized weight ».
Ce poids sera nommé Wseq.

Pondération de la famille positive

Le poids de la famille positive sera multiplié par le nombre de familles négative :
Wpos=Nbre de famille négative
Le poids total des séquence positive sera donc Wpos x Wseq
Les poids des séquences négative ne sera pas modifié.

Bilan des jeux de données

Pour chaque famille d'Unités Protéiques il y aura :
(i) Un fichier contenant les exemples positif et négatif constitué des exemples issus de l'alignement brut UP_POSITIVE_mafft.fas transformé en valeur aaindex
(ii) Un fichier contenant les concaténations de tous les fichiers de toutes les familles négatives alignées sur l'alignement UP_POSITIVE_mafft.fas et transformé en valeur aaindex
(iii) un fichier contenant les deux fichiers précédents concaténés
(iv) Un fichier contenant les poids (un par ligne) de chacune des séquences dans l'ordre du fichier concaténé précédent
(v)Un fichier contenant la classe (+1 ou -1 par exemple) (un par ligne) de chacune des séquences dans l'ordre du fichier concaténé précédent

Apprentissage

L'apprentissage se fera avec les fichiers des jeux de données en utilisant la bibliothèque python scikit-learn et utilisant la méthode *random forrest* . La pondération sera utilisé (fit(X, y[, sample_weight])).

Contact : Jean-Christophe Gelly (jean-christophe.gelly@univ-paris-diderot.fr)

Références :

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Research. 2008;36(Database issue):D202-D205. doi:10.1093/nar/gkm998.

Van Westen GJ, Swier RF, Wegner JK, IJzerman AP, van Vlijmen HW, Bender A. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. Journal of Cheminformatics. 2013;5:41. doi:10.1186/1758-2946-5-41.

Annexe 1 : Les 58 propriétés AAINDEX les plus représentatives

1	ARGP820103	Membrane-buried preference parameters
2	BAEK050101	Linker index
3	BHAR880101	Average flexibility indices
4	CASG920101	Hydrophobicity scale from native protein structures
5	CHAM810101	Steric parameter
6	CHAM820101	Polarizability parameter
7	CHAM830101	The Chou-Fasman parameter of the coil conformation
8	CHAM830107	A parameter of charge transfer capability
9	CHAM830108	A parameter of charge transfer donor capability
10	CHOP780201	Normalized frequency of alpha-helix
11	CHOP780202	Normalized frequency of beta-sheet
12	CHOP780203	Normalized frequency of beta-turn
13	CIDH920105	Normalized average hydrophobicity scales
14	COSI940101	Electron-ion interaction potential values
15	FASG760101	Molecular weight
16	FAUJ880102	Smoothed upsilon steric parameter
17	FAUJ880103	Normalized van der Waals volume
18	FAUJ880104	STERIMOL length of the side chain
19	FAUJ880105	STERIMOL minimum width of the side chain
20	FAUJ880106	STERIMOL maximum width of the side chain
21	FAUJ880109	Number of hydrogen bond donors
22	FAUJ880110	Number of full nonbonding orbitals
23	FAUJ880111	Positive charge
24	FAUJ880112	Negative charge
25	FAUJ880113	pK-a(RCOOH)
26	GRAR740102	Polarity
27	JANJ780102	Percentage of buried residues
28	JANJ780103	Percentage of exposed residues
29	JOND920102	Relative mutability
30	JUNJ780101	Sequence frequency
31	KLEP840101	Net charge
32	KOEP990101	Alpha-helix propensity derived from designed sequences
33	KOEP990102	Beta-sheet propensity derived from designed sequences
34	KRIW790101	Side chain interaction parameter
35	KYTJ820101	Hydropathy index
36	LEVM760102	Distance between C-alpha and centroid of side chain
37	LEVM760103	Side chain angle theta(AAR)
38	LEVM760104	Side chain torsion angle phi(AAAR)
39	LEVM760105	Radius of gyration of side chain
40	LEVM760106	van der Waals parameter R0
41	LEVM760107	van der Waals parameter epsilon
42	MITS020101	Amphiphilicity index
43	MONM990201	Averaged turn propensities in a transmembrane helix
44	NISK800101	8 A contact number
45	NISK860101	14 A contact number
46	PONP800101	Surrounding hydrophobicity in folded form
47	PONP930101	Hydrophobicity scales
48	RACS770103	Side chain orientational preference
49	RADA880108	Mean polarity
50	ROSG850101	Mean area buried on transfer
51	ROSG850102	Mean fractional area loss
52	ROSM880102	Side chain hydropathy, corrected for solvation
53	TAKK010101	Side-chain contribution to protein stability
54	VINM940101	Normalized flexibility parameters (B-values), average
55	WARP780101	Average interactions per side chain atom
56	WOLR810101	Hydration potential
57	ZHOH040102	Relative stability scale extracted from mutation experiments
58	ZHOH040103	Buriability