Original Paper

The cleverSuite Approach for Protein Characterization: Predictions of Structural Properties, Solubility, Chaperone Requirements and RNA-Binding Abilities

Petr Klus[1,2], Benedetta Bolognesi[1,2], Federico Agostini[1,2], Domenica Marchese[1,2], Andreas Zanzoni[1,2] and Gian Gaetano Tartaglia[1,2,*]

1. Gene Function and Evolution, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.

2. Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

Server input

CM Input. The Clever Machine accepts two datasets (FASTA format). The first input set is considered positive (P) and - due to the fact that algorithm is comparative and based on strength difference - the second is called negative (N). However, despite the distinction between the datasets, there is no bias applied to their processing.

CC Input. CC accepts sequences in FASTA format. Apart from the dataset, there is one more required field – a reference to the model previously generate by CM (provided in the CM output).

Server output

Individual scale view (CC). The individual scale view shows the coverage, area under the ROC curve, Z-score and p-value for each physico-chemical propensity (Fig. S3). Depending on property enrichment or depletion, color-coded bars point to either positive or negative set. The colors reflect property's assignation to a group – for instance purple bars represent hydrophobicity. The view is interactive – the user can click on graph's bars and see which element of the table corresponds to it. Also, the table can be sorted according to the coverage, ROC, Z-score a p-value by clicking on the table header. The output is designed to be simple to interpret – e.g. the size of bars is proportional to their coverage.

Grouped property view (CC). The second section provides a view of individual scales grouped by class assignment. The server visualizes individual property consistencies and their relative strengths in separate plots (Fig. S3). Elements of the plots are interactive and linked to information about individual property coverage and Z-score. The scales that did not pass the signal strength filter are devoid of their group-specific color.

Propensity scale combinations (CC). We provide a plot showing relation between the number of combined scales and the individual dataset coverages for both positive and negative dataset (Fig. 4). We also report the statistics for each scale combination and its individual members. The user can click through the combination titles to reveal which scales are contained and their statistics.

Dataset assignment (CM). First part of the output is the dataset assignation (the output is similar to that presented in Fig. S1 and S3). The statistics is reported. For each protein of the query set, p-values are reported (Fig. S3). The relevant physico-chemical profiles are shown in the second part of the output (Fig. 4).

Independent validations

RePROF (Rost, 1996) For alpha helical assignment, we use minimum alpha-helical content of 50% (40% for beta-sheet proteins.

The *FoldIndex* (Prilusky *et al.*, 2005) scores were used to evaluate disordered proteins. Negative scores are associated with disordered proteins (positive scores for structured proteins).

NetSurfP (Petersen *et al.*, 2009) For alpha helical assignment, we use minimum alpha-helical content of 50% (40% for beta-sheet proteins). To detect disordered proteins, we use minimum coil content of 50%. For each residue, we consider structural assignment probability larger than 0.5.

Limbo (Van Durme *et al.*, 2009) We consider sequences that have at least one DnaK binding motif as positive. If the threshold is changed to 5 binding motifs, the true positive rate on GroEL substrates is 67%, the true positive rate on DnaK substrates is 55% and the true negative rate on independently folding proteins is 99%.

For *RNApred* (Kumar *et al.*, 2011) predictions, we used the default prediction threshold to determine if a protein is RNA-binding (SVM cutoff = -0.2).

For *PROSO II* (Smialowski *et al.*, 2012), we use the default score threshold value of 0.6, which matches the associated soluble/insoluble labels.

Table S1 - Links to CM and CC submissions and further related information.

Description	URL
Further information	
cleverSuite main portal	http://service.tartaglialab.com/clever_suite
Featured submissions	http://service.tartaglialab.com/clever_community
CM documentation	http://s.tartaglialab.com/static_files/algorithms/clever_machine/documentation.html
CM tutorial	http://s.tartaglialab.com/static_files/algorithms/clever_machine/tutorial.html
CC documentation	http://s.tartaglialab.com/static_files/algorithms/clever_classifier/documentation.html
CC tutorial	http://s.tartaglialab.com/static_files/algorithms/clever_classifier/tutorial.html

Table S2. Signal strength with respect to random set (same AA composition as reference sets).

	Signal strength		
	CM	**CC (P)**	**CC (N)**
Alpha-beta	0.5	0.4	0.4
Disorder	0.4	0.4	0.2
Solubility	0.5	0.5	0.1
Chaperones	0.3	0.2	0.2
mRNA	0.5	0.4	0.1

The signal strength ranges from 0 (no discrimination between prediction and random sets) to 0.5 (complete separation)

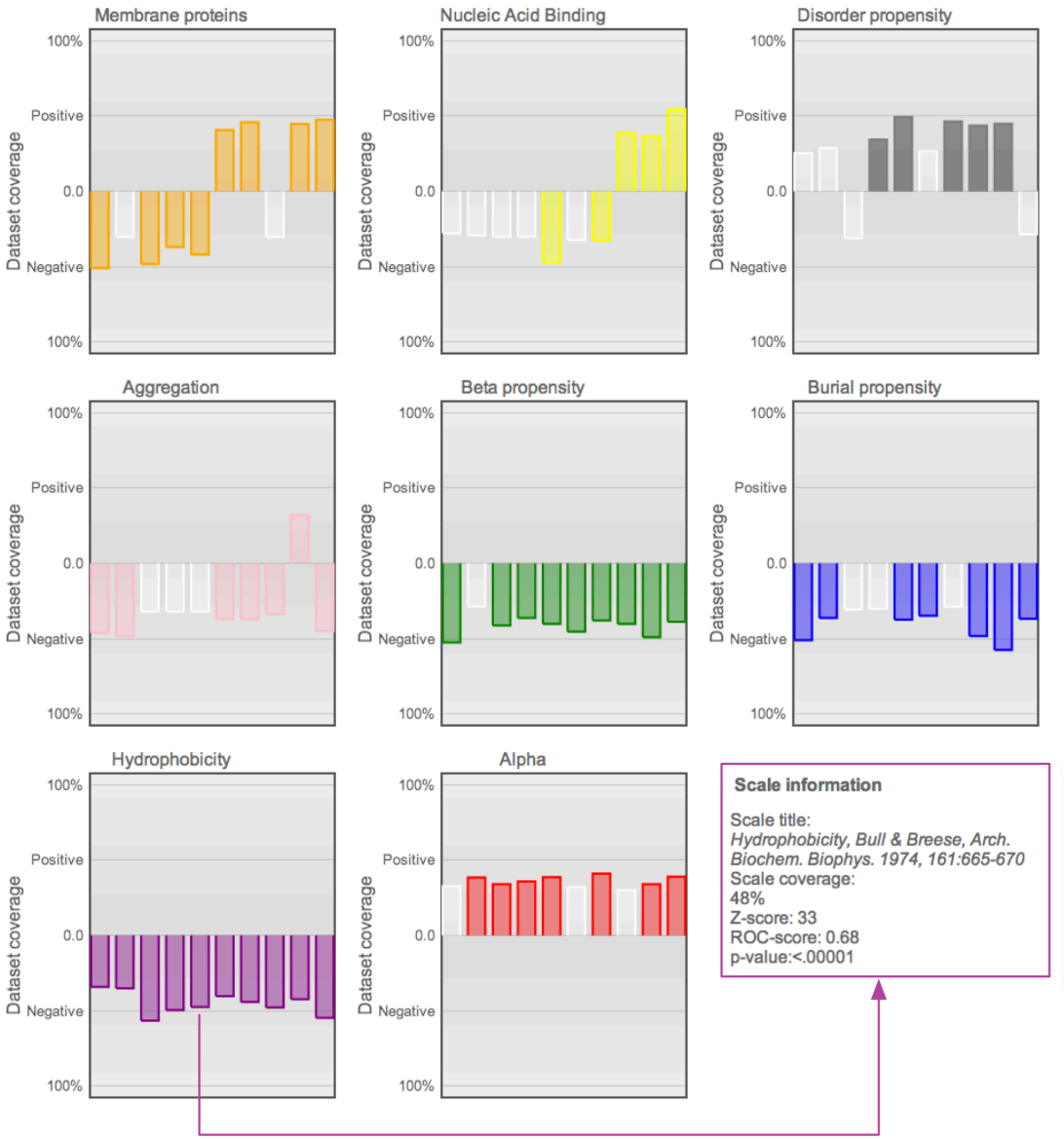

Fig. S1. *Grouped property view*. 8 Properties are grouped by class assignment and color. Low-significance properties (Z-score < Z_{th}; p>0.01;) are devoid of color. In the webserver, this view is interactive and shows information about each scale after clicking (the *E. coli* solubility analysis is used as example).

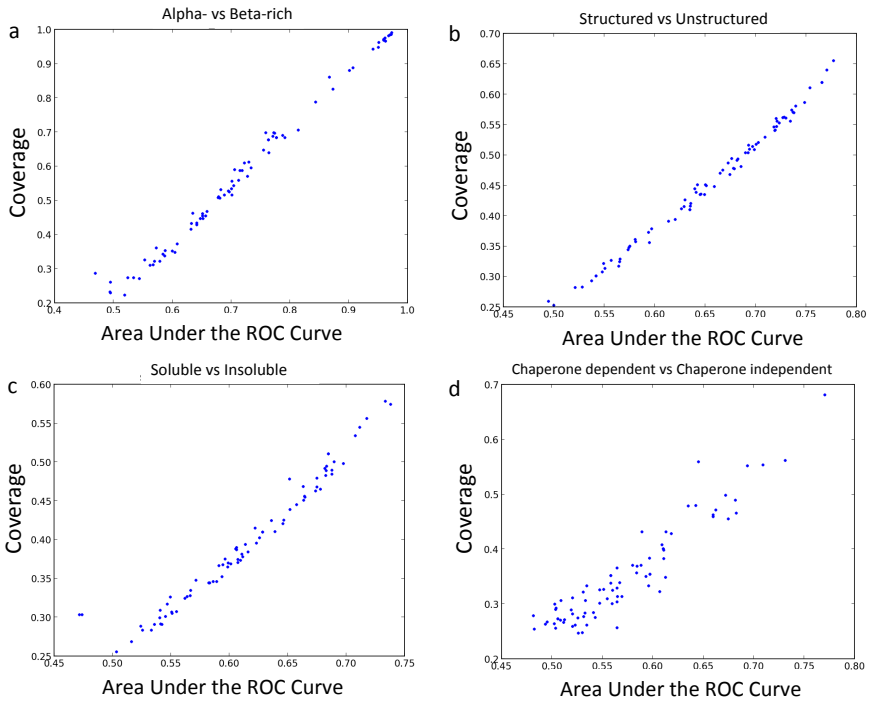

Fig. S2. *Coverage vs Area Under the ROC Curve (AUC)*. For each of the examples presented in the main text, coverage and AUC are highly correlated. a) Alpha-helix vs beta sheet proteins (Pearson's correlation r=0.97); b) Structured vs unstructured proteins (r=0.99); c) Soluble vs insoluble proteins (r=0.95); d) Chaperone-dependent vs independently folding proteins (r=0.85).

Cov	ROC	Scale title	Z	p-value
57.92%	0.73	Hydrophobicity, Sweet et al., J. Mol. Biol. 1983, 171:479-488	44.6	<.00001
47.67%	0.68	Hydrophobicity, Bull & Breese, Arch. Biochem. Biophys. 1974, 161:665-670	33.3	<.00001
49.92%	0.68	Aggregation - Conchillo-Sole, BMC Bioinformatics 2007, 8:65	33.3	<.00001
55.32%	0.71	Hydrophobicity, Fauchere et al., Eur. J. Med. Chem. 1983, 18:369-375	33.1	<.00001
49.08%	0.68	Hydrophobicity, Black, Anal. Biochem. 1991, 193:72-82	32.4	<.00001
46.23%	0.67	DisProt, Dunker AK, Protein Pept Lett. 2008; 15(9): 956-963	32.2	<.00001
57.87%	0.73	Propensity to be buried inside, Wertz-Scheraga, Macromolecules 1978, 11:9-15	31.8	<.00001
50.50%	0.70	Hydrophobicity, Roseman, J. Mol. Biol. 1988, 200:513-522	30.6	<.00001
49.15%	0.68	Mean area buried on transfer, Rose et al., Science 1985, 229:834-838	30.5	<.00001
53.17%	0.71	Nucleic Acid Binding - unknownRBD, Castello et al., Cell 2011, 149:1393-1406	29.5	<.00001

Fig. S3. *Individual properties output.* A) Depending on enrichment or depletion of physico-chemical properties, associated color-coded bars point to either positive or negative set; B) Coverage and Area under the ROC curve (Cov and ROC) and statistical significance (Z-score and P-value) are reported per each property (the *E. coli* solubility analysis is used here as example). References from literature are reported. The web-view is interactive – the user can click on graph's bars and see which element of the table corresponds to it.

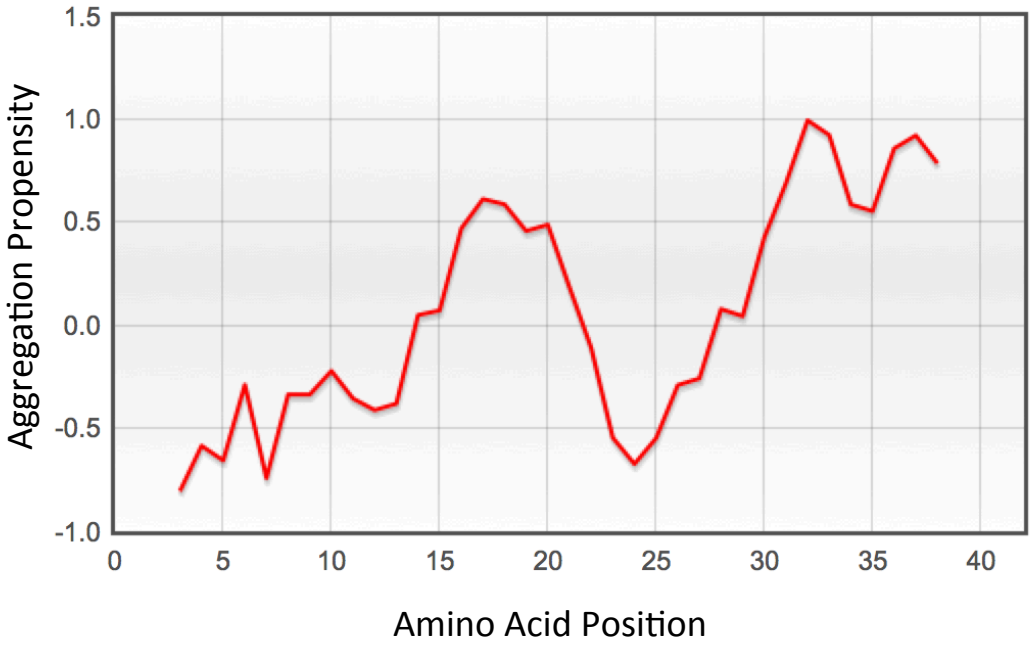

Fig. S4. *Physico-chemical profiles.* The CC algorithm generates physico-chemical profiles for each input protein. Here, we show the aggregation propensity of Alzheimer's Aβ42 calculated with the scale derived by Ventura and coworkers. The profile significantly correlates with the one predicted by Zyggregator and highlights regions that are important for the aggregation process.

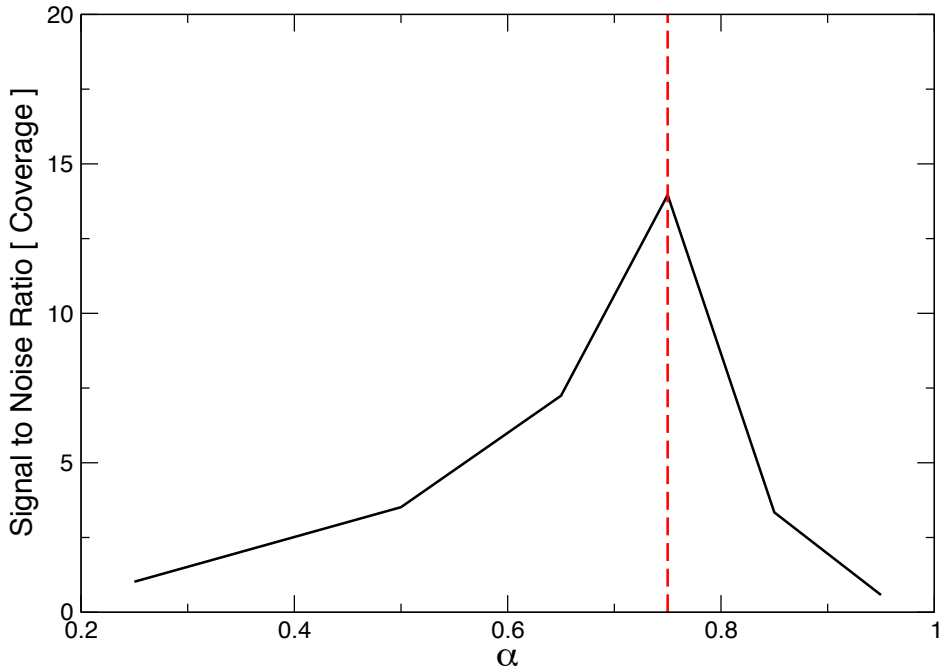

Fig. S5. *Optimization of the* internal parameter α. For each of the 5 cases presented in this work, we calculated the coverage enrichment with respect to the shuffled sets: $\delta = coverage(P, N) - \frac{1}{R}\sum_r coverage(P_r, N_r)$ (Methods). The signal to noise ratio (i.e., ratio of mean to standard deviation) was evaluated at different values of α. We found that the optimal discrimination between signal and noise corresponds to $\alpha = 0.75$.

Van Durme,J. *et al.* (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS Comput. Biol.*, **5**, e1000475.
Kumar,M. *et al.* (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit. JMR*, **24**, 303–313.
Petersen,B. *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.
Prilusky,J. *et al.* (2005) FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
Rost,B. (1996) [31] PHD: Predicting one-dimensional protein structure by profile-based neural networks.
Smialowski,P. *et al.* (2012) PROSO II – a new method for protein solubility prediction. *FEBS J.*, **279**, 2192–2200.