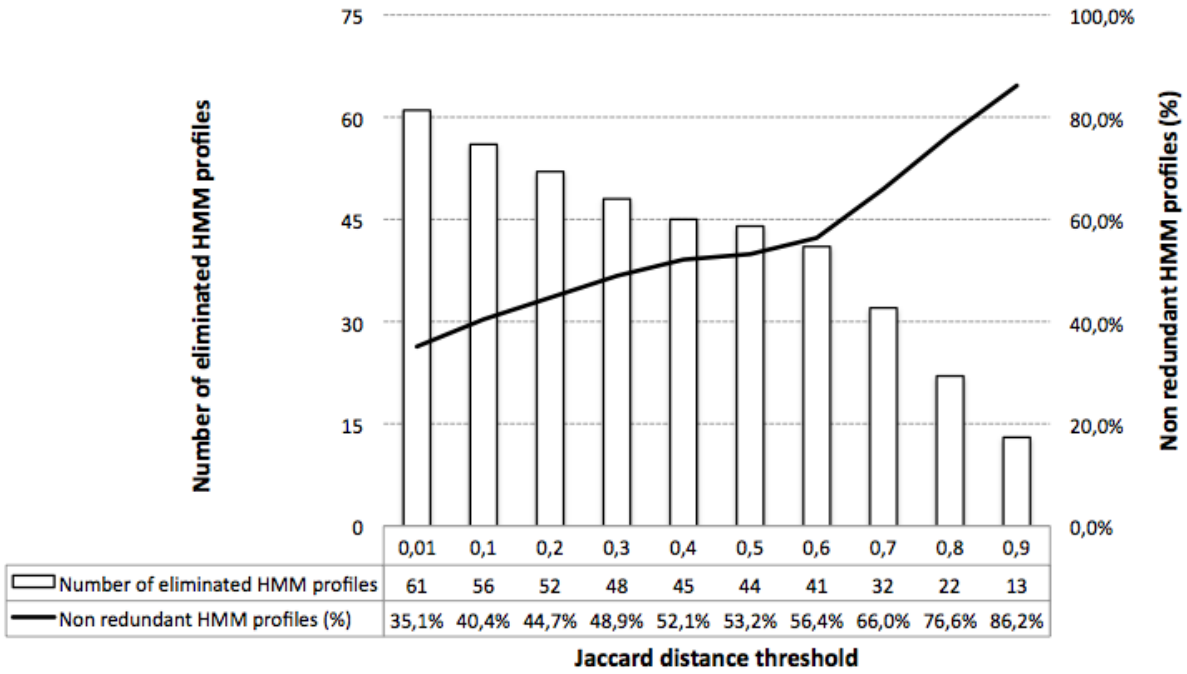
Supplementary materials (Estellon et al.,)

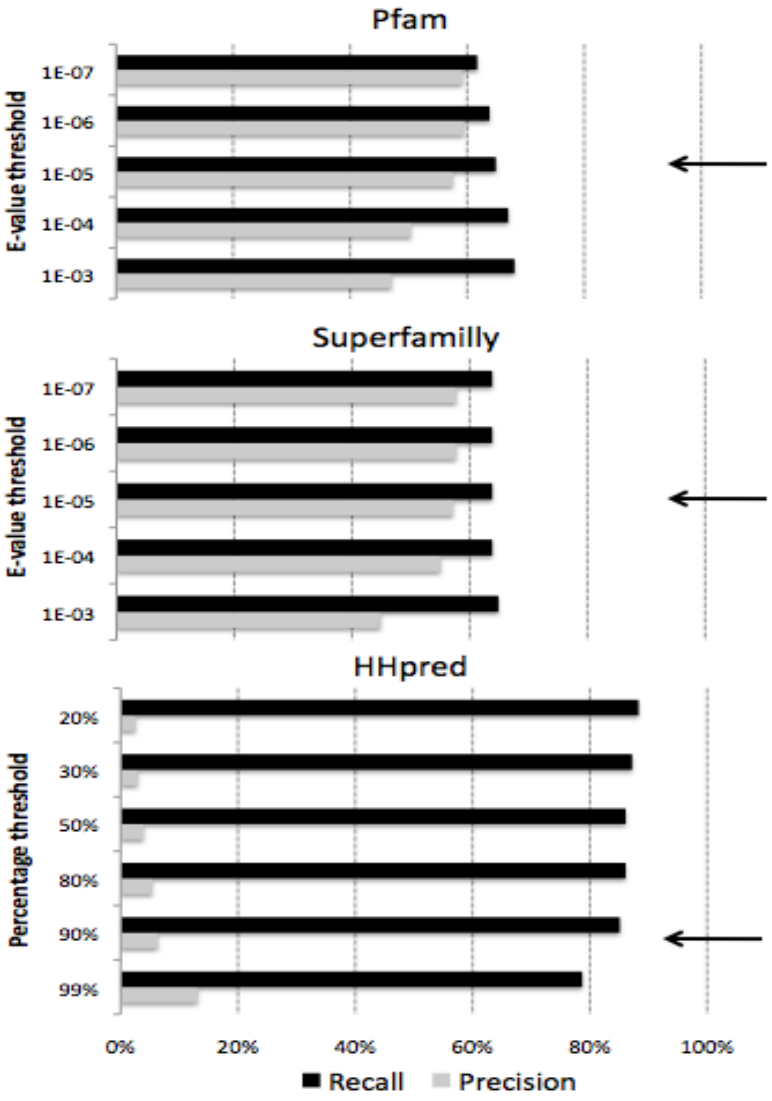
1. Fe-S profile HMM redundancy filtering procedure

A specialized database of profiles HMM was built from the Fe-S protein sequences retrieved from the PDB70 depleted of *E.coli* sequences. As a profile HMM is built from a multiple sequence alignment (MSA), two different proteins can generate a similar profile HMM by aggregating similar sequences. This redundancy can be removed using the Jaccard similarity coefficient (defined as the size of the intersection divided by the size of the union of the sample sets). This coefficient was computed for each pair of profile HMMs, based on the MSA sequence composition. On the graph shown below, the dip at 0.6 in the metric distribution was used as a threshold above which two HMM-profiles were considered similar. When two highly similar profile HMMs were detected, only the one with the highest number of aligned protein sequences in the MSA was retained. Out of 93 initial HMM profiles, 41 were considered to be redundant (44%) and were thus discarded. This resulted in a final collection of 52 Fe-S-specific and non-redundant profile HMMs (see figure below).



2. Threshold settings for Pfam, SSF and profile HMM descriptors

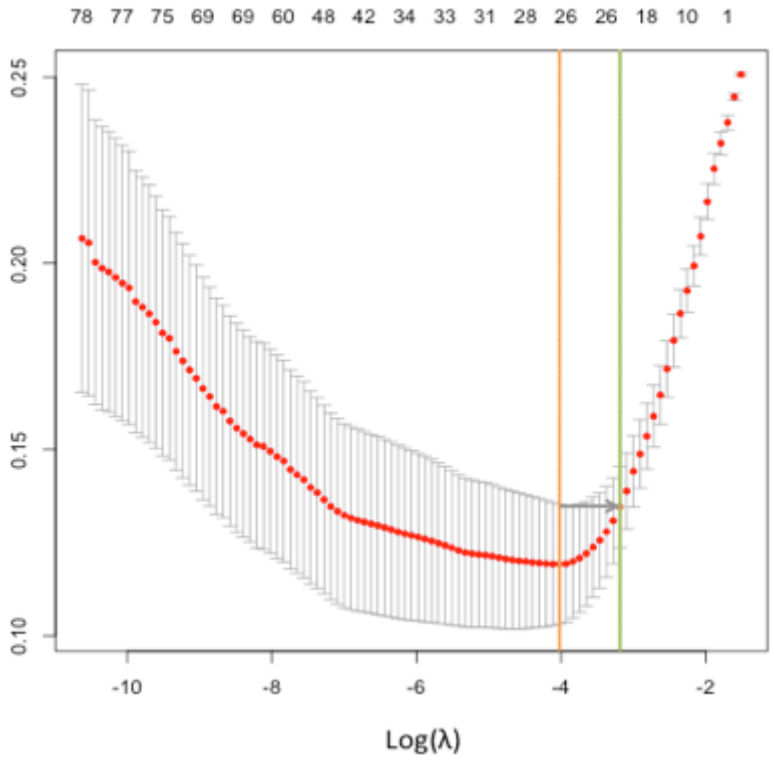
Searches using Pfam and Superfamily return hits with an associated E-value. A hit is usually considered valuable when the e-value is less than 10^{-5}. In our case, we want to increase the precision/recall ratio of these two to improve Fe-S protein identification, we therefore tested different values for the E-value threshold. The best ratio was attained with a threshold of 10^{-5} for Pfam, and 10^{-6} for Superfamily after running the hmmscan program. Given the very slight difference between performances at 10^{-6} and 10^{-5}, we finally kept the recommended 10^{-5} threshold for both descriptors. To determine a cut-off threshold value for the HHpred tool, we considered the value of the probability that each match was a true positive rather than the E-value, because it has been reported that E-values returned by most tools can be very unreliable [1]. As sensitivity (also known as recall) is essential, we retained a probability score of 90% as threshold (i.e. the highest one before a drop in recall). These results are presented on the figure below, where recall (black bar) and precision (grey bar) are plotted according to a range of E-values and probability score thresholds. The arrow on each bar graph indicates the cut-off value retained for each tool (Pfam, SSF and HHpred).



3. Setting the λ parameter for the elastic net procedure

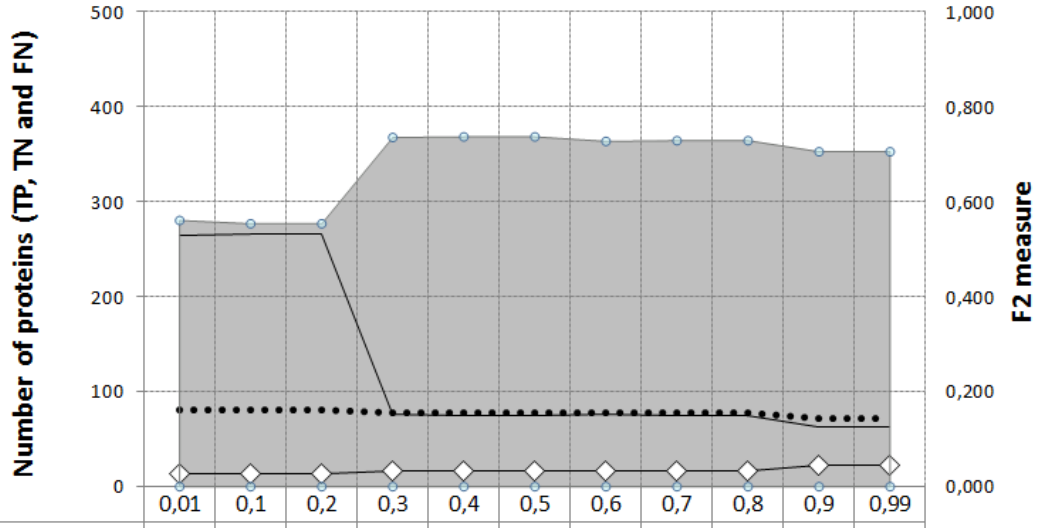
λ represents the shrinkage parameter; as λ increases the coefficients are shrunk ever more strongly. The risk of producing a false prediction is estimated by the mean error square as a function of the penalty (actually *Log(λ)*). The value of the penalty parameter, λ, at which predictions are required was chosen as the maximum value for which the 10-fold cross-validation estimation of the error does not exceed one standard deviation of the minimal mean square error (for practical details see [2]). The minimal mean square error of 0.119 is reached for Log(λ)=-4.013, thus λ=0.018 (see below). In the graph, the x-axis corresponds to the Log(λ) value, while the y-axis corresponds to the mean square error. Across the top of the plot, the number of descriptors considered as relevant according to the elastic net procedure as a function of Log(λ) is indicated. At one standard deviation (between the orange and green vertical bars), we obtained a threshold value for λ of 0.041 (mean square error 0.135 and

Log(λ)=-3.194). Therefore, the risk is minimized, while also producing the most parsimonious model.



4. Tuning the α parameter for the elastic net procedure

Parameterisation of the mixed model was done during the learning phase. True positives and negatives, false positives and negatives were determined by comparing the predicted proteins from the PDB70 (depleted of *E. coli* sequences) with Uniprot annotations and the literature. The α parameter of the elastic net regression is defined as the elastic net mixing parameter. It sets the degree of mixing between the ridge regression penalty (α = 0) and the lasso penalty (α = 1) [3]. This penalty is particularly useful in our case, where there are many correlated variables (i.e., Fe-S descriptors). To be generic, the model should keep as many descriptors as possible. Therefore, the α parameter chosen was the minimum value giving the maximum F_2-measure. This latest metric is defined as the weighted harmonic mean of precision and recall; it reflects the efficiency of the mixed model. Here, the maximum F_2-measure was reached with α = 0.4 (as shown in the figure below); this value is therefore retained for both mixed and extended models.



	0,01	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,99
F2 measure	0,561	0,554	0,554	0,736	0,737	0,737	0,728	0,729	0,729	0,706	0,706
True Positives (TP)	81	80	80	78	78	78	77	77	77	72	72
False Positives (FP)	265	266	266	76	75	75	76	75	75	62	62
False Negatives (FN)	13	14	14	16	16	16	17	17	17	22	22

Value of the α parameter

References:

1. Söding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21**:951–60.

2. **The glmnet package reference manual** [http://cran.r-project.org/web/packages/glmnet/glmnet.pdf]

3. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33**:1–22.