An integrative computational model for large-scale identification of metalloproteins in microbial genomes: A focus on iron–sulfur cluster proteins

ARTICLE *in* METALLOMICS · AUGUST 2014

Impact Factor: 3.59 · DOI: 10.1039/c4mt00156g · Source: PubMed

5 AUTHORS, INCLUDING:

Sandrine Ollagnier-de Choudens
Atomic Energy and Alternative Energies Co…
56 PUBLICATIONS **2,171** CITATIONS
SEE PROFILE

Myriam Smadja
Collège de France
2 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Marc Fontecave
Collège de France
222 PUBLICATIONS **8,362** CITATIONS
SEE PROFILE

Yves Vandenbrouck
Atomic Energy and Alternative Energies Co…
35 PUBLICATIONS **575** CITATIONS
SEE PROFILE

Metallomics

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c4mt00156g

An integrative computational model for large-scale identification of metalloproteins in microbial genomes: a focus on iron–sulfur cluster proteins†

Johan Estellon,‡[abc] Sandrine Ollagnier de Choudens,‡[def] Myriam Smadja,[ghi] Marc Fontecave[ghi] and Yves Vandenbrouck*[abc]

Metalloproteins represent a ubiquitous group of molecules which are crucial to the survival of all living organisms. While several metal-binding motifs have been defined, it remains challenging to confidently identify metalloproteins from primary protein sequences using computational approaches alone. Here, we describe a comprehensive strategy based on a machine learning approach to design and assess a penalized generalized linear model. We used this strategy to detect members of the iron–sulfur cluster protein family. A new category of descriptors, whose profile is based on profile hidden Markov models, encoding structural information was combined with public descriptors into a linear model. The model was trained and tested on distinct datasets composed of well-characterized iron–sulfur protein sequences, and the resulting model provided higher sensitivity compared to a motif-based approach, while maintaining a good level of specificity. Analysis of this linear model allows us to detect and quantify the contribution of each descriptor, providing us with a better understanding of this complex protein family along with valuable indications for further experimental characterization. Two newly-identified proteins, YhcC and YdiJ, were functionally validated as genuine iron–sulfur proteins, confirming the prediction. The computational model was then applied to over 550 prokaryotic genomes to screen for iron–sulfur proteomes; the results are publicly available at: http://biodev.extra.cea.fr/isph. This study represents a proof-of-concept for the application of a penalized linear model to identify metalloprotein superfamilies on a large-scale. The application employed here, screening for iron–sulfur proteomes, provides new candidates for further biochemical and structural analysis as well as new resources for an extensive exploration of iron-sulfuromes in the microbial world.

Received 6th June 2014,
Accepted 7th August 2014

DOI: 10.1039/c4mt00156g

www.rsc.org/metallomics

Introduction

Metalloproteins are found ubiquitously in all domains of life including bacteria, plants and animals,[1,2] and the total number of metal-binding structures annotated in a proteome scales linearly to both the genome size and protein-coding gene number.[3] Recently, it has been estimated that one-quarter to one-third of all proteins contain metals which are required for catalytic, regulatory and/or structural functions.[4,5] Despite being a large class of proteins, and representing a group of molecules of major importance for living organisms, the bioinformatics community has largely ignored metalloproteins compared to other families such as, for example, kinases or G protein-coupled receptors (GPCRs).[6] Nevertheless, large-scale predictive identification of metalloproteins in the ever-growing number of sequenced microbial genomes is of special interest as the main experimental techniques used to identify proteins – mass spectrometry and synchrotron-based approaches – are costly to implement and not routinely available for metalloproteins.[7,8] In the framework of systematic whole genome sequencing projects, protein function is usually inferred using bioinformatics tools based on sequence homologies or identification of conserved motifs. Although these approaches provide a primary level of functional annotation, around 30 to 40% of proteins are not assigned a function. Moreover, false annotations due to

[a] Univ. Grenoble Alpes, iRTSV-BGE, F-38000 Grenoble, France. E-mail: johan.estellon@gmail.com
[b] CEA, iRTSV-BGE, F-38000 Grenoble, France. E-mail: yves.vandenbrouck@cea.fr; Fax: +33 4 38 78 50 32; Tel: +33 4 38 78 26 74
[c] INSERM, BGE, F-38000 Grenoble, France
[d] Univ. Grenoble Alpes, iRTSV-LCBM, F-38000 Grenoble, France. E-mail: sandrine.ollagnier@cea.fr
[e] CNRS, IRTSV-LCBM, F-38000 Grenoble, France
[f] CEA, iRTSV-LCBM, F-38000 Grenoble, France
[g] CNRS, UMR8229, 74231 Paris Cedex 05, France. E-mail: marc.fontecave@cea.fr
[h] Collège de France, 11 place Marcelin Berthelot, 75231 Paris Cedex 05, France. E-mail: myriam.smadja@college-de-france.fr
[i] Université Pierre et Marie Curie, Paris, France

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4mt00156g

‡ These authors contributed equally to this study.

automatic bioinformatics-based function assignment procedures have been extensively reported and remain an important issue.[9,10] Despite the major importance of metalloproteins there is a dearth of truly specific, validated bioinformatics tools for their identification within sequence databanks.[11,12] Because the sequence of related proteins can diverge beyond the point where their relationship can be revealed by pairwise sequence comparisons, similarity detection can be improved by multiple alignment methods[13] or by position-specific iterative searches such as PSI-Blast.[14] However, the well-established degeneracy between primary sequences and structures limits these methods, which only rely on primary information.[15] Specifically for metalloproteins, predictive approaches based on molecular modelling and the design of 3D structural templates have been developed.[16–19] However, the relatively low number of resolved structures available in the non-redundant Protein Data Bank[20,21] restricts the availability of well-established structural features unambiguously discriminating between candidates in systematic large-scale screening approaches. Consequently, identifying potential metalloproteins solely on the basis of their primary genome sequences is extremely challenging; however, in terms of costs it remains an attractive alternative to systematic experimental identification. As far as we know, early bioinformatics attempts were based on straightforward individual motif searches[22] or regular expressions derived from multiple alignments of known metalloprotein sequences not associated with subsequent experimental validation.[23] Interestingly, iterative profile searching applied to the Radical-*S*-adenosylmethionine (SAM) superfamily was very efficient. This is probably due to the remarkably well-conserved metal-binding signature of this family, the CX_3CX_2C pattern (where C represents the metal-binding amino acid cysteine, X an amino acid residue other than cysteine, and the numbers correspond to the number of amino-acid residues between two neighbouring ligands).[24] However, it seems that what was considered the Radical SAM pattern paradigm does not hold true, as a growing number of newly-identified Radical SAM proteins harbour an atypical signature. These include ThiC, which has a CX_2CX_4C pattern in the C-terminal domain;[25] HmdB, with its CX_3CX_2C motif;[26] and PhnJ, with a $CX_2CX_{21}C$ motif.[27] To cope with this versatility, and the large diversity of metalloproteins, more targeted approaches have been developed, such as dedicated metal-binding pattern building[28–30] or prediction of metal binding using machine learning techniques.[17,31,32] However, despite these efforts to improve metalloprotein identification, there remains a need for significant methodological improvements to be able to fully exploit the large set of sequenced genomes in the field of bioinorganic chemistry.[8,11,33] Indeed, in a recent study using inductively coupled mass spectrometry (ICP-MS) to identify metals, combined with tandem mass spectrometry (MS/MS) for protein identification, Cvetkovic *et al.*[12] showed that around 45% of metalloproteins from a bacterial proteome were not covered when searching the Integrated Resource of Protein Domains and Functional Sites (InterPro) database.[34] Assuming that structure-based and modelling methods are not suitable for a large-scale approach (because they are too computationally costly), and considering that most current sequence-based approaches are not discriminatory enough,[12] we designed a computational model to predict metalloproteins based on a machine learning framework. To make the model directly interpretable, we restricted ourselves to penalized methods with variable selection over different categories of Fe–S protein descriptors; three of these categories were retrieved from publicly available resources (patterns and hidden Markov models derived from conserved protein domains), while the fourth consisted of tailored Fe–S profile hidden Markov models (HMMs), which allow for sequence diversity between Fe–S proteins. Then, to make integration of a potentially large number of descriptors possible and to focus only on features with the highest predictive power, we designed a generalized linear model. This model was trained on a PDB dataset with a penalty procedure to produce a sparse linear model automatically discarding irrelevant descriptors, if any were identified.[35] We observed that the four categories of descriptors used provided relevant and complementary information for Fe–S protein prediction. These descriptors were integrated into a single linear model which was applied to an independent dataset containing Fe–S protein sequences, the *Escherichia coli* proteome.

The results presented in this article show that our linear model clearly outperforms the sequence analysis tool InterPro, which is commonly used to identify protein family signatures.[34] We chose to focus on the iron–sulfur (Fe–S) protein family (*i.e.*, cofactors involving iron and inorganic sulphide, with mainly cysteinyl-iron coordination)[36] as these proteins are possibly one of the most ancient protein families, and are ubiquitous along with other structurally and functionally versatile biological prosthetic groups.[37,38] In addition, we have extensive expertise working with Fe–S proteins,[39–44] which made it possible to functionally validate candidate proteins identified by the computational model. Validation results confirmed the reliability of the computational model for two newly-identified proteins. Finally, the predictive model was applied to more than 550 bacterial proteomes, and the resulting set of predicted Fe–S proteins, classified by organism, was made publicly available through a dedicated website (http://biodev.extra.cea.fr/isph) for use by the community.

Materials and methods

Data sets

The training set is composed of all the protein sequences contained in the PDB70 (a filtered version of the PDB (release 04/2012)[20] with a maximum sequence identity of 70%). To avoid any bias between the training and testing phases, all the protein sequences from *E. coli* were discarded from the training set, resulting in 20 289 sequences from a total of 2323 bacterial species. After manual verification of all potential heterogeneous groups available in the PDB, 93 Fe–S proteins were identified through the annotation reported in HETNAM fields (*e.g.* "FS3", "SF4") (a list can be consulted at http://deposit.rcsb.org/het_dictionary.txt). The test set consisted of the *E. coli* K12 proteome (4408 protein sequences), retrieved from the HAMAP database[45] (based on UniProt[46] release 2012_10). The UniRef 50 database (release 2012_10) used for profile–profile HMM comparison (see below) is a filtered version

of the Uniprot database with a maximum sequence identity of 50%. After removing non-bacterial and *E. coli* protein sequences, the resulting databank contained 2 151 438 proteins covering 1043 bacterial species. Proteins with no homologues in this databank were removed from both test and training sets (see below).

Iron–sulfur descriptors

Fe–S-specific signatures were retrieved by keyword extraction (*e.g.* "iron–sulfur", "Rieske", "Fe–S") from the following databases: Pfam (v1.3, database v24.0), Superfamily (v22.2.02, database v1.73) and Prosite (v1.67, database v20.64). After manual curation, we produced a dedicated Fe–S protein signature databank (see Table S2, ESI†) against which all protein sequences were screened using an appropriate search tool: *pfscan* from the Pftools package for Prosite patterns;[47] *hmmscan* from the HMMER2 and HMMER3 package developed by S. Eddy (http://hmmer.janelia.org/) for Superfamily (SSF) and Pfam profile HMMs, respectively. To build tailored Fe–S distant homology profiles, the HHpred[48] toolbox based on profile–profile comparisons was used. For each Fe–S protein sequence in the training set, multiple sequence alignments (MSA) were built using 5 iterations of PSI-Blast[14] with a cut-off *E*-value of $< 10^{-5}$ against the UniRef 50 databank (see above). Secondary structure information (from structures assigned by DSSP[49] or predicted by PSIPRED (v2.45)[50]) was added to increase the sensitivity for remote, but structurally analogous, homologous proteins. These alignments were converted into profile HMMs using the HH-make algorithm. This produced 93 profile HMMs. To remove the redundancy between those Fe–S-specific profile HMMs, the Jaccard similarity coefficient was computed for each pair of profile HMMs, based on their MSA constituents (see ESI,† Section S1 for details). The result is a non-redundant database of Fe–S-specific profile HMMs, against which all profile HMMs built from query proteins can be tested (self-hits were removed) using the profile–profile comparison program HHsearch (v1.6).

Data representation and the encoding scheme

To develop an interpretable linear model for Fe–S protein prediction, we retained a set of descriptors able to detect Fe–S proteins. Thus, each protein entry is represented as a binary vector. The dimension of this vector corresponds to the total number of Fe–S descriptors. The binary vector encodes the result of a Fe–S hit for each descriptor. More precisely, the correlation score for each of the Fe–S descriptors used for a given protein sequence was set to 0 or 1 according to the significance of the hit with respect to a defined threshold. A large number of threshold settings were tested (see ESI,† Section S2) and we observed that our results converged with the threshold values generally recommended in the literature: *E*-value $< 10^{-5}$ for Pfam and Superfamily hits, probability ≥ 90 for HHsearch hits. The size of the resulting binary vector corresponds to the total number of Fe–S descriptors considered (*e.g.* 30 signatures and 52 profile HMMs for the mixed model).

Regression shrinkage and descriptor selection using the elastic net approach

The elastic net implementation available in the glmnet package (v1.6)[51] for R statistical software[52] was used to build the generalized linear model known here as the mixed model. Model fitting and parameter tuning were carried out on the training set by a ten-fold cross-validation procedure. The whole training set was randomly partitioned into ten groups of approximately equal size. To ensure that the training process is completely independent, the model is trained on nine groups and tested on the remaining group. Each group is chosen for assessment in turn. 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. The elastic net mixing parameter, α, was chosen as the minimum value giving the best *F*-measure on the training set (for details, see ESI,† Sections S3 and S4).

Extended predictive model

The extended predictive model was built in the same way as the previous mixed predictive model, using UniRef databanks and both training and testing protein sequence datasets no longer depleted of *E. coli* sequences. The non-redundant set of Fe–S HMM-profiles was also rebuilt based on both PDB70 and *E. coli* Fe–S protein sequences. The same values of parameters λ and α as those used with the mixed model were used with the extended model (see previous section).

Proteome comparison

Fe–S protein predictions returned by the extended model were compared with UniProt annotations (reviewed or not) for the following six bacterial proteomes: *Acinetobacter baumannii* (ACIBY),[53] *Bradyrhizobium* sp. ORS278 (BRASO),[54] *Clostridium difficile* 630 (CLOD6),[55] *Helicobacter pylori* B38 (HELPB),[56] *Neisseria meningitidis* NEM8013 (NEIM8),[57] *Vibrio splendidus* (VIBSL).[58] All proteomes were downloaded from HAMAP[45] in their updated version (05-2013). Fe–S annotations were retrieved from keywords and the "cofactor" field in "Sequence annotations" of the flat files.

Metrics

True positives (TP), false positives (FP) and false negatives (FN) were determined by comparing the predictions with Uniprot annotations and the literature. Recall, also known as sensitivity, is the fraction of correct predictions among all Fe–S proteins:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

while precision is the fraction of correct predictions among those that the computational model believes to belong to the Fe–S protein family:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

The *F*-measure corresponds to the weighted harmonic mean of precision and recall:

$$F_\beta = \frac{1 + \beta^2 \times (\text{Precision} \times \text{Recall})}{(\beta^2 \times \text{Precision} + \text{Recall})}$$

Here, we used the F_2-measure, which gives recall twice the weight of precision ($\beta = 2$), this metric reflects the efficiency of the model.

Expression and purification of YhcC and YdiJ from *E. coli*

The gene encoding YhcC was amplified by PCR from *E. coli* genomic DNA using the primers P1 (5′ GGGCCCCATATGCAGT TACAGAAATTAGTC 3′) and P2 (5′ AAATTTGGATCCCTACTCC GTTGGAGGTAG 3′). The PCR product was cloned into pET-15b (Invitrogen). The gene encoding YdiJ was amplified by PCR from *E. coli* genomic DNA using the primers P1 (5′ AAATTTG AATTCATGATTCCACAGATTTCCC 3′) and P2 (5′ AAATTTAA GCTTTTTAATAATCTCCAGTAAAGCCT 3′) for cloning into pET-22b (Invitrogen). The *E. coli* BL21(DE3) expression strain (Novagen) was transformed with pET15b-yhcC and pET22b-ydiJ plasmids. YhcC expression was induced in LB medium at OD_{600nm} = 0.5 by adding isopropyl-beta-D-galactopyranoside (IPTG) (0.5 mM final concentration). Induction was maintained for 3 h at 37 °C. YdiJ was induced for 2.5 h with 0.25 mM IPTG. In both cases, cells were harvested by centrifugation at 5000 rpm for 15 min. Cell pellets were stored at −80 °C. Protein purification was performed in a glove box (Jacomex B553 (NMT)) under nitrogen. Bacterial pellets containing YhcC were resuspended in degassed buffer A (100 mM Tris-HCl, 50 mM NaCl, pH 7.5) containing 0.6 mg mL^{-1} lysozyme and 1 mM PMSF; pellets containing YdiJ were resuspended in buffer B (0.1 M potassium phosphate pH 7.5, 50 mM NaCl, Pefabloc 239.69 g mol^{-1}). Resuspended cells were then transferred into ultracentrifuge tubes and rested for 45 min. Tubes were frozen rapidly in liquid nitrogen outside the glove box and thawed inside the glove box. This procedure was repeated 3 times to fully lyse the cells. Suspensions were ultracentrifuged (4 °C, 40 000 rpm, 75 min). Inside the glove box, supernatants were loaded onto a Ni-NTA column (5 mL, Qiagen) which had been degassed and equilibrated with the appropriate buffer the day before. After extensive washing with 100 mL of equilibration buffer containing 10 mM imidazole, proteins were eluted with a linear gradient of buffer (A or B) containing 1 M imidazole. Eluates were immediately desalted on a NAP-25 column (GE Healthcare) to remove imidazole. Pure proteins (as determined by SDS-PAGE) were concentrated, frozen in liquid nitrogen and stored at −80 °C until use. A small aliquot of all preparations were kept for protein quantitation and iron and sulphur analyses. Proteins were quantified using the Bradford method (Biorad) with Bovine Serum Albumin as a standard.

Fe–S reconstitution of YhcC

Fe–S cluster reconstitution was carried out as described previously[59] under strictly anaerobic conditions in a Jacomex NT glove box containing less than 2 ppm O_2.

SAM reduction

SAM reduction was performed under anaerobic conditions in a glove box. Apo-YhcC or Fe–S containing YhcC (10 μM) in 0.1 M Tris-HCl, 50 mM NaCl, pH 8, was incubated with 200 μM SAM and 2 mM dithionite. After 1 h at 37 °C, the reaction was quenched by acidification with sodium formate (3.5 M). The solution was centrifuged at room temperature for 10 min at 14 000 rpm. The supernatant was loaded onto a HPLC Zorbax SB-C18 column equilibrated with 0.1% trifluoroacetic acid. A linear gradient from 0 to 28% acetonitrile in 0.1% trifluoroacetic acid was run at 1 ml min^{-1} for 25 min. 5′-Deoxyadenosine (AdoH) was detected at 260 nm and identified by comparing its elution time (12.5 min) with that of a commercial sample.

Iron and sulfur quantitation

Iron was determined using a bathophenanthroline disulfonate assay after acid denaturation of the protein;[60] labile sulfur was determined by the Beinert method.[61]

Spectroscopic methods

UV-visible spectra were recorded using a Cary 1 Bio (Varian) spectrophotometer. EPR spectra were recorded on a Bruker EMX (9.5 GHz) EPR spectrometer equipped with an ESR 900 helium flow cryostat (Oxford Instruments).

Results and discussion

Experimental set-up and workflow

The general workflow for our approach is represented in Fig. 1. Briefly, the study can be broken down into three initial steps: (1) selection of a set of Fe–S descriptors, (2) training and tuning of generalized linear models on a dataset composed of Fe–S and non-Fe–S protein sequences from the PDB databank and (3) assessment of the performance of the resulting models on an independent dataset. To avoid any bias, the dataset and the databanks used for the training phase were different from those used during the testing phase. For the test set we retained the whole *Escherichia coli* proteome because of its high level of annotation as a widely studied model organism. The best-performing computational model was selected and its prediction capacity assessed by (4) experimentally validating selected newly-identified Fe–S proteins among the candidates drawn from the whole *E. coli* proteome, and (5) assessing how generic the predictive model is by applying it to six recently re-annotated bacterial proteomes.

Descriptors for iron–sulfur cluster proteins

To develop an interpretable linear model for prediction of Fe–S proteins, we need to define a set of descriptors representing Fe–S proteins in terms of sequence information and secondary structure. First, we considered three types of protein ''signature'' databases which could be relevant for Fe–S protein sequence identification. These were derived from the Integrated Resource of Proteins Domains and Functional Sites (InterPro) database[34] by taking advantage of an in-house set of 85 experimentally characterized Fe–S proteins from *E. coli*[41] (Table S1, ESI†). The protein sequences for this set of proteins were submitted locally to the iprscan program. This programme scans protein sequences against InterPro's signatures.[62] Results of this analysis showed that three categories of signature databases

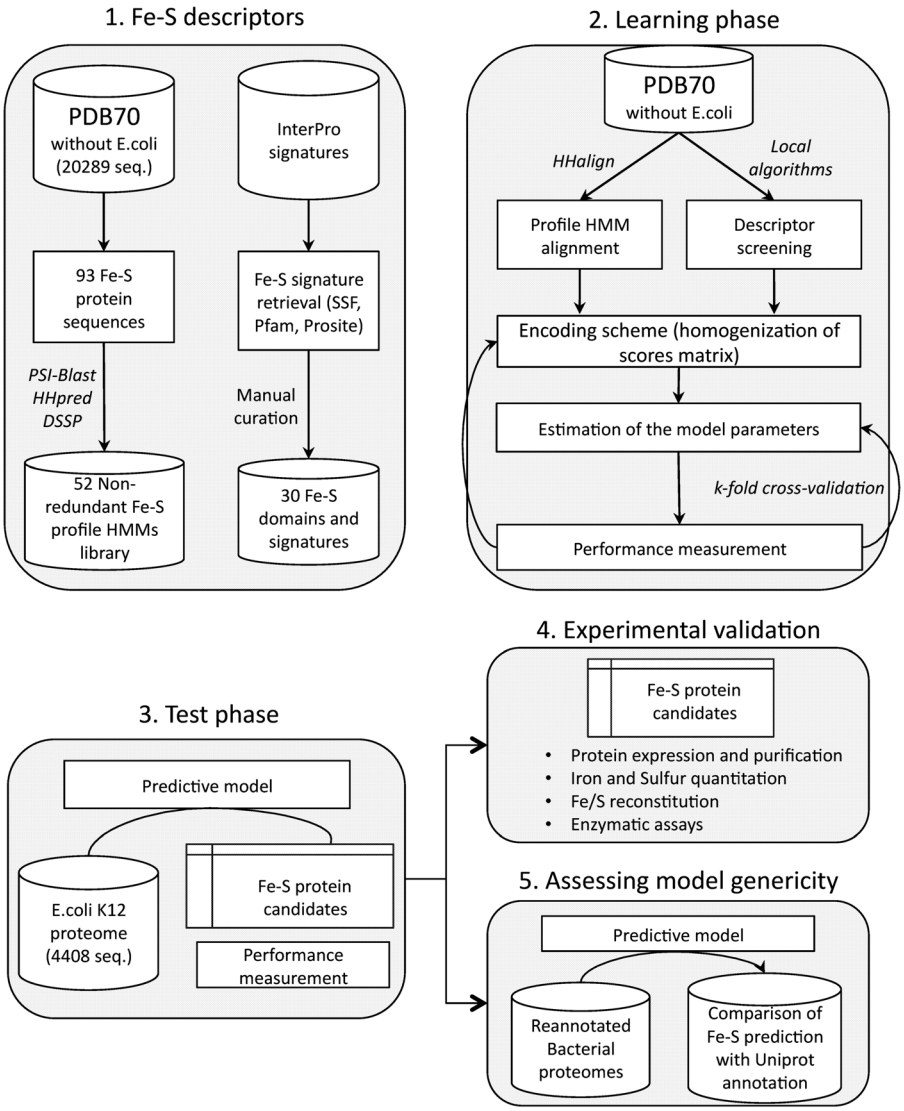

Fig. 1 Diagrammatic representation of the workflow for the design and assessment of the computational model for Fe–S protein prediction. (1) Fe–S protein-specific descriptors are built then considered into two distinct databases: (i) a database of remote homology profiles (ii) a database of motifs, conserved domains and structural protein domains. (2) The predictive model is trained with all protein sequences from the PDB70 after removing *E. coli* sequences. (3) The final performance of the trained model is evaluated on an independent test set, the complete *E. coli* K12 proteome. (4) Experimental characterization of new *E. coli* Fe–S protein candidates. (5) In parallel, the predictive model is applied to six recently re-annotated microbial proteomes to assess how generic it is by comparing output predictions and UniProt annotations.

cover around 95% of InterPro hits, indicating either complementarity or overlap between these resources (data not shown). The three categories of signature databases are as follows: conserved domains from Pfam,[63] structural protein domains from Superfamily (SSF),[64] and motifs from Prosite.[65] The remaining hits were found to be fully redundant with these three categories of sequence signatures, therefore the other databanks in the InterPro database were not further considered as sources of descriptors. Of the set of experimentally validated Fe–S proteins, the overall coverage by the InterPro signatures does not exceed 60% (see list of InterPro Metal hits in Table S1, ESI†). This confirms the trend observed by Cvetkovic *et al.*,[12] who reported that only 55% percent of metalloproteins from *Pyrococcus furiosus* identified by MS/MS combined with ICP-MS are present in InterPro. After manual inspection of annotation by these three sources of sequence descriptors (Prosite, Pfam and SSF), 30 signatures or domains were annotated as specific to Fe–S clusters, with 9, 13 and 8 signatures selected from Prosite, Pfam and SSF, respectively (Table S2, ESI†). Pfam contains multiple alignments and HMM-based profiles (profile HMMs) of complete protein domains,[63] SSF is based on a collection of HMM representing structural protein domains sharing an evolutionary relationship and grouped together at a superfamily level,[66] and Prosite patterns are short conserved sequences contained within regions known to be important, or harbouring biologically significant residue(s).[65] Due to the relatively weak sensitivity of these signatures, we also designed descriptors accounting for the versatility of Fe–S proteins.

For this, we chose the HHpred method[67] which is based on profile–profile HMM comparison. HHpred includes known or predicted secondary structure information in the HMM and makes it possible to identify distant homologous relationships. HHpred has proven to be faster and more sensitive than other sequence similarity search methods, as it can detect homologous relationships far beyond the twilight zone (*i.e.*, below 20% sequence identity).[48,68] To take advantage of this, a specialized set of profile HMMs was built from 93 Fe–S protein sequences retrieved from the PDB, excluding *E. coli* protein sequences. This was done using the HHpred toolbox, as described in Materials and methods. To avoid redundancy and consecutive bias due to overrepresented Fe–S protein subtypes, all nearly identical profile HMM descriptors were filtered out. This resulted in a final collection of 52 Fe–S descriptors (see ESI,† Section S1) in addition to the previously selected signatures.

Overlap and complementarity between descriptor categories

Each category of descriptor (Prosite, Pfam, SSF, and tailored Fe–S Profiles HMM) was tested separately for its predictive power against a dataset composed of protein sequences from PDB70 (January 2013 release). This database covers all bacterial species, except *E. coli*, and includes 93 Fe–S protein sequences (after removal of self-hits for HHpred) (see Materials and methods and Table S3, ESI†). With this method, a Fe–S protein is predicted as such when at least one hit is considered significant according to a given score with one of the four categories of descriptors (see ESI,† Section S2 for score threshold settings). The Fe–S protein sequences dataset from the PDB covered by each category of descriptor is shown in Fig. 2. Screening for motifs (Prosite, Pfam, SSF) with appropriate search tools on the Fe–S protein sequence dataset provides a recall (also known as sensitivity) of 0.57, with a false positive rate of 34%. Results show that a large proportion of Fe–S proteins are retrieved equally well by the three methods, suggesting that these three categories of descriptors (Prosite, Pfam and SSF), although considered independently, actually shared common determinants and captured very little specific information. For instance, we noticed that predictions based on Prosite patterns are fully covered by Pfam and SSF descriptors. In contrast, remote homologies and fold recognition recognised by profile HMM building with the HHpred toolbox identified 16 additional Fe–S proteins which were not identified by the three other descriptor categories. Of the 93 Fe–S proteins contained in the training dataset, this represents an added value in recall of 18%. This added value must, however, be weighed against the poor precision of the method, as the HHsearch program retrieved 1172 false positives. Therefore, taking all the descriptors from the four categories into account in a single scoring scheme should enhance the recall/precision ratio. In total, 81 proteins (out of the 93 expected) are correctly predicted to be Fe–S proteins, while 1290 false positives were returned (Fig. 2).

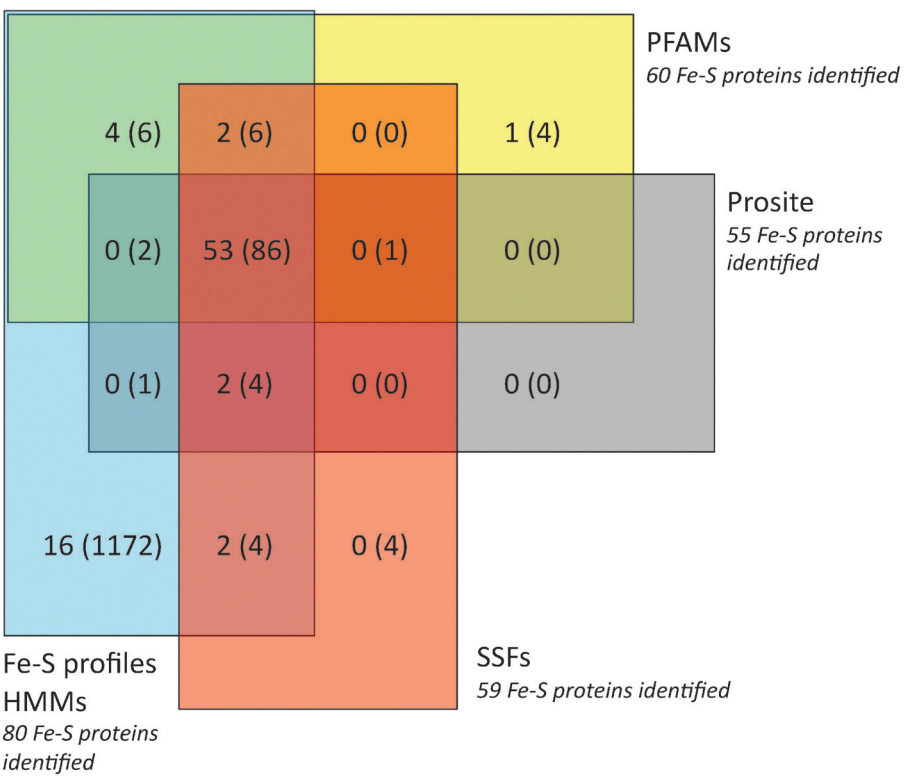

Fig. 2 Predictive power of each category of Fe–S descriptors. The Venn diagram displays the number of Fe–S proteins identified by one or more categories of descriptors. The total number of true positives is reported in each area with respect to each category of descriptors, false positive numbers are given in brackets. The dataset is composed of 20 289 protein sequences from the PDB70, including 93 Fe–S proteins.

Building a predictive model

To increase the signal/noise ratio (*i.e.*, decrease the number of false positives while avoiding an excessive false negative rate), each of the four categories of descriptors was considered either independently or grouped together into a penalized linear model. In this model, we wish the descriptors with a better predictive power to be highly weighted, and to remove the ones with excessive noise levels during the learning step. To do this, we used the elastic net procedure, which has a remarkable ability to deal with high numbers of redundant and correlated descriptors.[35] Elastic net is a linear model which can be trained with an L1 norm (lasso) and an L2 norm (ridge regression) prior to regularisation. This combination produces a sparse model where few of the weights are non-zero, like lasso, while still maintaining the regularisation properties of ridge regression. The elastic net model is particularly suitable when multiple correlated descriptors are used, as with Fe–S protein sequence features which contain partially overlapping information. Lasso is likely to pick one of these descriptors at random, while elastic net is likely to pick two. In other words, the elastic net retains more variables than lasso, but its behaviour is more selective than ridge regression.[51] The elastic net procedure estimates a linear model aiming to explain the Fe–S character of a given protein (from the test set) based on a set of proteins with known Fe–S ligands (the training set). The training set was composed of 20 289 protein sequences from the PDB70 (filtered version of the Protein Data Bank with a maximum sequence identity of 70%). As the test dataset to independently evaluate the performances of the resulting predictive model was *E. coli*, all the *E. coli* protein sequences (4408 sequences) were removed from the training set to avoid introducing bias. To assess performance, we used common metrics (precision, recall and F_2-measure) which are defined in the Material & methods section. Two parameters are of primary importance during the training phase: (i) the λ penalty and (ii) the mixing parameter, α. λ refers to the penalties attributed to each features (*i.e.*, Fe–S descriptors), while α represents the ratio between the lasso and ridge penalties. Thus, α controls the trade-off between these two models.[35,51] Tuning parameters and model fitting were optimized on the training set by a ten-fold cross-validation procedure. The value of the λ penalty parameter 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. The elastic net mixing parameter, α, was chosen as the minimum value giving the best F_2-measure on the training set (for details, see ESI,† Sections S3 and S4).

Performance of the predictive model

Five models with different tuning were built and evaluated: four corresponding to each category of descriptors considered separately (Prosite, Pfam, SSF, and Fe–S profile HMMs), and what we have called the mixed model, which integrates the four categories of Fe–S descriptors (Table 1, left part). In this table, each protein entry is represented as a binary vector, the dimension of the vector corresponding to the number of descriptors considered; for instance, for the model based on Pfam descriptors, the dimension of the vector will be 13, whereas for the mixed model, the dimension is 82 (see Materials and methods). At first, we observed an opposite trend between signature descriptors (Prosite, Pfam, SSF) and profile HMMs (HHpred). The first three had middle-range precisions (from 58.3% to 62.2%) and recalls (from 59.6% to 64.9%), while the other provided a higher recall (73.4%) but significantly lower precision (17.1%) (Table 1, left part). Interestingly, the mixed model combining the whole set of descriptors outperforms each individual category, with a recall of 83% and an F_2-measure of 0.74. The five models were subsequently assessed using an independent test dataset, the complete *E. coli* proteome, containing 4408 protein sequences (Table 1, right part). This organism was chosen because it is currently the best-annotated proteome to our knowledge. The full list of Fe–S proteins predicted or missed by the mixed model is presented in Table S4 (ESI†). To compute the recall and precision, Fe–S proteins predicted by the model that have been experimentally validated and/or are annotated as Fe–S proteins based on Uniprot keywords (*e.g.*, Fe_2S_2, Fe_3S_4, Fe_4S_4, iron–sulfur, *etc.*) are considered to be true positives (TP). Fe–S proteins predicted by the model with no supporting information in Uniprot are considered as false positives (FP), while Fe–S proteins which are either experimentally validated or annotated as Fe–S in Uniprot but were missed by the predictive model are considered as false negatives (FN). Once again, we observed an inverse trend between motifs and domain descriptors

Table 1 Performance of each predictive model. The six models used were: four models representing each individual category of descriptors, the mixed model integrating all Fe–S descriptors, and the extended model which comprises the overall set of Fe–S descriptors including those built from known *E. coli* Fe–S proteins. All models were trained on the PDB70 dataset without *E. coli* sequences and tested on the whole *E. coli* proteome. True positives (TP), false positives (FP) and false negatives (FN) were determined by comparison of the prediction with Uniprot annotations and literature citations. Precision (Pre.), recall (Rec.) and F_2-measure are reported. F_2-measure reflects the efficiency of the model

	Training set (PDB70 without *E. coli*)			Test set (*E. coli* proteome)					
Predictive models	Precision (%)	Recall (%)	F_2-measure	TP	FP	FN	Precision (%)	Recall (%)	F_2-measure
Prosite	62.20	59.60	0.6	53	2	83	96.4	38.9	0.44
Pfam	60.40	64.90	0.64	76	3	60	96.2	55.8	0.61
SSF	58.30	63.80	0.63	72	4	64	94.7	52.9	0.58
Profile HMM	17.10	73.40	0.44	90	63	46	58.8	66.2	0.64
Mixed model	51.00	83.00	0.74	90	14	46	86.5	66.2	0.69
Extended model	—	—	—	109	15	27	87.9	80.1	0.81

(Prosite, Pfam, SSF) and tailored Fe–S profiles (HHpred) with regard to the precision-recall balance; the Prosite motifs model provided the lowest recall (Table 1, right part). Surprisingly, the precision of the three models based on Prosite, Pfam or SSF descriptors is better for the test dataset than for the training set. This suggests that Fe–S protein sequences from *E. coli* were considered in the set of seed alignments during the procedure to build patterns or profiles for these descriptors.[66,69] Nevertheless, the most significant result was that the mixed model attained the maximal observed F_2-measure (0.69) and achieved a good level of precision and recall (86.5% and 66.2% respectively) compared to the other four. This indicates that these different categories of descriptors contain relevant complementary information for Fe–S protein prediction, and that the features selection procedure performed by the elastic net can produce a good trade-off between the sensitivity provided by tailored Fe–S profile HMMs and the precision of the descriptors contained in Pfam, SSF and Prosite databanks.

Interpretability of the model

As the mixed model built with the elastic net procedure was linear, it is possible to understand how the predicted accuracy is computed, and to assess the importance of the different descriptors by examining the weights of the model associated with each feature. Fig. 3 shows a graphic representation of the weights attributed to each descriptor when the elastic net is trained on the PDB70 sequences depleted of *E. coli* proteins (also detailed in Table S5, ESI†). We provide a precise quantitative assessment of the contribution of each feature: the weight of a feature illustrates its individual contribution to the final predicted Fe–S protein status. Among the 82 descriptors included in the mixed model, 47 were selected by the elastic net procedure and assigned a non-null weight. We observed that patterns from Prosite are mainly (5/9) eliminated by the model, probably because they are too specific, suggesting a poor ability to account for the great diversity of Fe–S protein sequences described by J. Meyer.[21] In fact, the information carried by these patterns encoded as regular expression could be either too strict or could be redundant with information provided by the profile HMMs from the Pfam and SSF databanks. Only four Fe–S protein descriptors have a weight greater than 1 (with a maximum recorded weight of 2.5). These descriptors are particularly relevant in the case of conserved Fe–S protein sub-families bearing specific functional motifs. Indeed, the weights awarded to the conserved domain PF04055 (1.288) and the structural protein domain SSF102114 (1.368) clearly reflect their contribution to the model predicting the Radical-SAM subfamily. Generally, these proteins are identified by the well-conserved consensus CX_3CX_2C, but sometimes there is a greater distance between the first two cysteine residues.[24,25] This probably explains the lack of efficacy of the corresponding Prosite pattern, PS01278 (methylthiotransferase Radical-SAM domain signature),

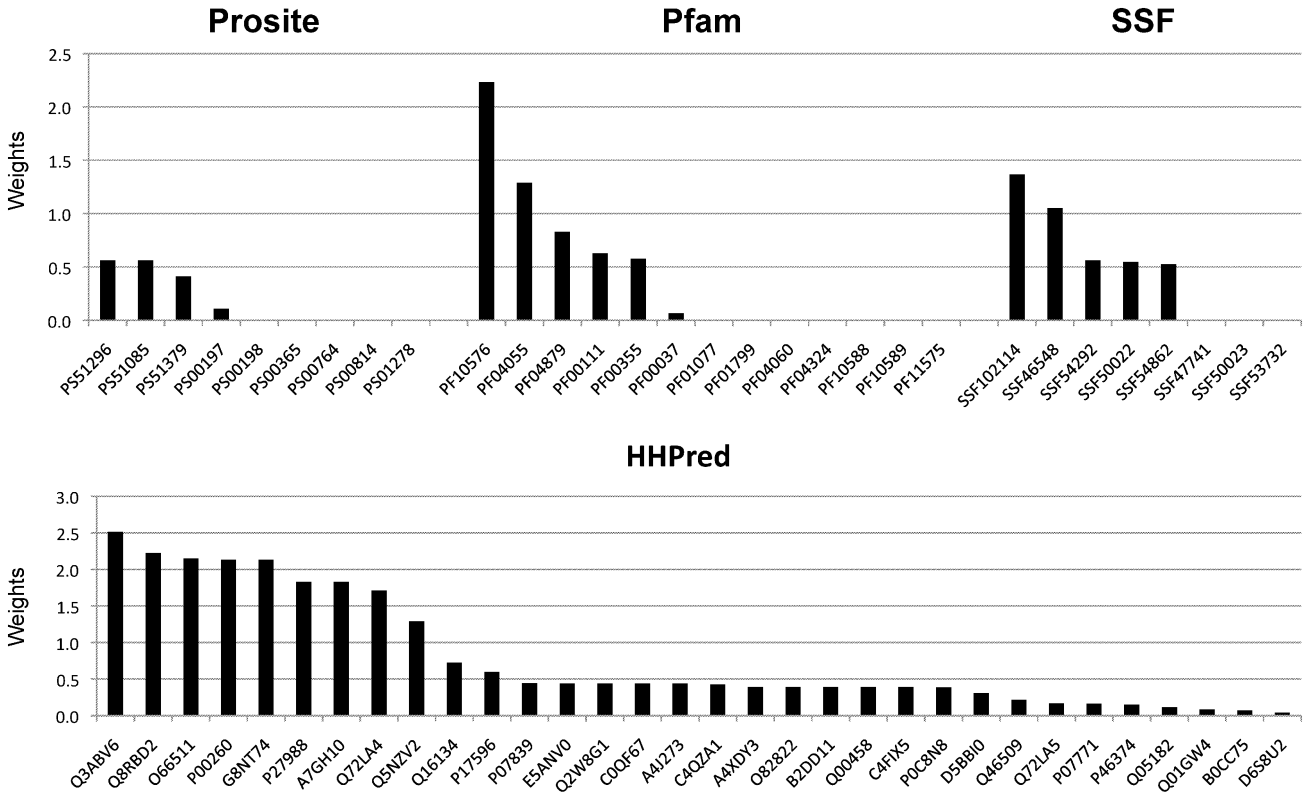

Fig. 3 Contribution of Fe–S descriptors according to the mixed predictive model. The elastic net procedure attributes weights to the descriptors during the training phase depending on their correlation with the expected prediction, and thus reflecting their influence on the overall Fe–S protein prediction. Weight values are plotted, grouped by category of descriptors (Prosite, Pfam, SSF and HHPred).

which carries a null weight. Among the highest weighted descriptors, we also noticed PF10576, which corresponds to the iron–sulfur binding domain of endonuclease III. This Fe–S protein binds a single Fe_4S_4 cluster that does not seem to be important for catalytic activity, but is probably involved in the appropriate positioning of the enzyme on the DNA strand, and may thus be subjected to evolutionary pressure.[70] The SSF46548 descriptor, with a weight of 1.05, corresponds to the alpha-helical ferredoxin domain containing two Fe_4S_4 clusters, typical of bacterial ferredoxin. This structural domain is present in several proteins involved in redox reactions, especially proteins implicated in the prevention of reactive oxygen formation (aerobic respiration) or catalysing the final step in anaerobic respiration.[71] All of these are vital processes. On the other hand, according to the weights of the model, tailored Fe–S profile HMMs contribute more to Fe–S prediction. Most of these descriptors carry little weight and can be used in combination with other features highlighting their complementarity, thanks to the grouping effect of the elastic net procedure.[72] However, 9 HHpred profile HMMs have weights greater than 1. Their individual specificities towards, structural (Fe_2S_2 and Fe_4S_4 are mostly represented) and functional (*e.g.* oxidoreductase, glycosylase, hydrogenase, thioredoxin-like, high potential iron–sulfur protein) Fe–S subclasses are very different, each reflecting some of the diversity of iron–sulfur proteins.[38] This reinforces the idea that as much information as possible should be gathered to extend the coverage of such a versatile protein family. Table S5 (ESI†) provides the list of iron–sulfur protein descriptors along with their weights.

Towards an extended model

The Fe–S proteins missed by the mixed model (35 experimentally validated proteins labelled as false negatives in Table S4, ESI†) were checked to determine how they were classified into the five functional classes proposed by Fontecave and collaborators.[41] Analysis of the model and the functional class to which false negative proteins belong reveals that the ''metallosite assembly'' or the ''non-redox catalysis'' functional classes are completely overlooked by the mixed model, but that the ''electron transfer'' and ''redox catalysis'' functional classes are well identified, with a coverage of 83% (Table 2). This is perhaps not totally surprising since members of these functional classes are either under-represented in public databanks and/or are not annotated as Fe–S proteins. As this additional knowledge could better account for Fe–S functional versatility and sequence/structure diversity, we assessed their contribution towards prediction by designing an extended predictive model including these additional *E. coli* sequences in the training set and in the descriptor bank. To do this, we re-constructed the Fe–S profile HMM descriptors combining the set of Fe–S protein sequences used for the mixed model and the set of experimentally validated *E. coli* Fe–S protein sequences. This gave rise to a total of 97 features (30 motif descriptors and 67 non-redundant tailored Fe–S HMM-profiles after filtering for HMM redundancy), instead of the initial 82 encoded into the mixed model. To ensure consistent comparison between the extended and the mixed model, none of the parameters tuned for the mixed model were changed (see ESI†). The resulting model thus contained 67 descriptors with non-null weights. The recall and precision computations were performed for the new model using the same definitions for TP, FP, and FN as described above (see ''Performance of the predictive model'' section). Overall performance on the complete *E. coli* proteome showed improved performance compared to the mixed model (Table 1, right part), reaching 80.1% in recall and 87.9% in precision, along with an F_2-measure of 0.81. As expected, Fe–S proteins from the ''metallosite assembly'' and ''non-redox catalysis'' functional classes not covered by the mixed model were better predicted by the extended model, with a gain of 66.6% and 60%, respectively (Table 2). In general, the misclassification rate for the five Fe–S functional classes was decreased with this enriched model.

Table 2 Functional classes of missed Fe–S proteins in *E. coli*. The number of Fe–S proteins from *E. coli* not predicted by the mixed or the extended model is reported for each functional protein class (percentage of misclassified Fe–S proteins in brackets). Only experimentally validated Fe–S proteins of known functional class (as defined in ref. 41 see Table S1, ESI) are listed

		Misclassifications (% coverage)	
Functional classification	Number of proteins	Mixed model	Extended model
Metallosite assembly	6	6 (100%)	2 (33.3%)
Nucleic acid binding protein	9	7 (77.7%)	6 (66.6%)
Non-redox catalysis	10	10 (100%)	4 (40%)
Redox catalysis	12	2 (16.6%)	1 (8.3%)
Electron transfer	49	10 (16.6%)	6 (12.2%)

What do false negative Fe–S proteins suggest?

Surprisingly, the extended model missed 27 Fe–S proteins, annotated as false negative in Table 1. Table S6 (ESI†) gives a detailed list of them based on functional class only (19 proteins). These proteins have been experimentally validated as Fe–S proteins, or are annotated as such by UniProt keywords/cofactor. We can discuss the reasons why some sets of proteins were missed by the model. One set of false negative proteins corresponds to SufB and IscU (NifU), which are involved in ''metallosite assembly''. *In vitro*, these two proteins assemble both Fe_2S_2 and Fe_4S_4 clusters and have the ability to transfer them to diverse apo-target proteins.[59,73,74] SufB and IscU are probably endowed with some plasticity allowing transient assembly of Fe–S clusters. Protein ligands involved in Fe–S coordination and whether they are identical for both Fe_2S_2 and Fe_4S_4 clusters are not yet known. Possible reasons why these two proteins were not covered by the extended model relate to the lack of structural information available and their under-representation in *E. coli*. Another set of proteins missed by the model contains four proteins, NadA, IlvD, FumB and TdcG, belonging to the ''non-redox catalysis'' functional class (Table 2). All these proteins catalyse dehydration reactions in which one

of the four iron sites of the cluster, which is not coordinated by a protein amino acid but rather by a water molecule, plays a crucial role in catalysis. Indeed, this accessible iron site is able to bind substrate(s) and thanks to its Lewis acid properties, contributes to dehydration reactions.[42,75–77] This group of proteins displays similar functional properties to aconitases Acon1 and Acon2 (which the model predicted correctly, see Table S6, ESI†). Aconitases use a similar cluster (a Fe_4S_4 cluster coordinated by only 3 cysteines), but comparison reveals that the unidentified group and aconitases neither share the same rare structural fold nor the same iron binding motif. This may explain the false negative classification. Even though IspH belongs to the ''redox catalysis'' class, it shares the same structural fold as NadA[78] and contains a cluster with similar properties to that of NadA, FumB, IlvD and TdcG. This may explain why it also escaped the model. Finally, the model missed a set of six proteins which are members of the ''nucleic acid binding proteins class'', DinG, Fnr, IscR, NsrR, RumA and SoxR. Four of these, Fnr, IscR, NsrR and SoxR, are sensor proteins, expression of which is regulated by environmental conditions. The model was probably unable to detect these proteins because no strong InterPro Fe–S descriptor is used by the linear model associated with nucleotide binding proteins. The only potentially relevant descriptor has a weight of 2.2 (PF10576) and is associated with endonuclease III DNA repair enzyme, but is probably too specific for this protein family. Although potentially interesting, these observations reflect how difficult it is to predict metalloproteins based on primary sequence information alone.

Are false positives potential Fe–S proteins?

Of the 124 Fe–S proteins predicted by this extended model (*i.e.*, InFeS Score > 0.5, see Table S6, ESI†), 15 are annotated as false positives in Table 1 because they were not previously known as Fe–S (either based on bibliographic references to experimental validation or UniProt keywords/cofactor annotation). These predicted Fe–S proteins are presented in Table 3, and their corresponding annotations are listed in Table S7 (ESI†). According to their annotation in UniProt (with no reference to any iron–sulfur properties keywords), these protein sequences were assigned to the false positives category. However, they could represent potential Fe–S proteins which have been not yet annotated in UniProt. Indeed, some of them are annotated as uncharacterized in Uniprot (see Table S7, ESI†). Based on the weights of the model, we chose to focus on the following unknown proteins of particular interest (Table 3): *E. coli* PreT (also referred to as YeiT) is a dehydrogenase whose activity has been suggested to be related to iron–sulfur metabolism,[79] *E. coli* YdiJ is an oxidoreductase enzyme coded by a gene just downstream of the *suf* operon – which is involved in Fe–S proteins biosynthesis – it could therefore have a functional link with it. Other hypothetical proteins are also good candidates, such as *E. coli* YhcC which has a very high confidence score (see next section). The remaining Fe–S protein candidates are already annotated; 5 are oxidoreductase enzymes, a functional class in which Fe–S clusters are highly represented. Among these, *E. coli* BisC was awarded a very high confidence score; this protein is thought to be involved in reducing a spontaneous oxidation product of biotin (BSO or BDS) to regenerate biotin.[80] The molybdenum cofactor identified could be used in conjunction with an Fe–S cluster to achieve the electron carrier activity, or electrons could be sequestered and delivered by an Fe–S cluster, as recently proposed in ref. 81. Moreover, BisC belongs to the prokaryotic molybdopterin-containing oxidoreductase family, in which most members bind an Fe–S cluster (590 proteins out of the 611 from this family are annotated as Fe–S binding in UniProt). The two last Fe–S protein candidates (AhpF and NirD) potentially bind NAD cofactors, and are therefore likely to be genuine false positives since it has been hypothesized that NAD/P cofactors can substitute for Fe–S clusters in proteins.[82] Thus, common conserved domains could result in possible homology. However, it appears that further investigation would be needed to clarify this hypothesis, since NirD has been shown to contain iron–sulfur clusters in *E. coli* by electron spin resonance spectroscopy studies.[83]

Comparison with InterPro

The results of the method were compared with those obtained with InterPro which is commonly used for automatic proteome annotation and is also available for batch analysis.[62] To make the comparison fair, both methods must be tested using the same data. We checked the InterPro resource by applying the irpscan program to the list of Fe–S proteins in *E. coli* and only retained those that have been directly experimentally confirmed. This resulted in a list of 93 Fe–S proteins (Table S8, ESI†). Because the Fe–S protein status assigned to proteins in Uniprot may be based on InterPro prediction, performance could be biased, which would make the comparison meaningless. To overcome potential bias we manually checked each protein entry related to an Fe–S protein and discarded entries from the *E. coli* proteome for which the field ''protein existence'', ''cofactor'' and ''metal-binding sites'' were annotated as ''inferred by similarity'', ''predicted'', ''inferred from electronic annotation'', ''probable'' or ''potential'' in UniProt. Only proteins which are explicitly referred to by a literature citation were retained (see Table S8, ESI†). Consequently, false positives were eliminated, and the recall was used as the only measure of performance for the comparison. We obtained a 63% recall for InterPro *vs.* 79.3% for the extended model. Thus, the extended model provides significantly improved prediction quality in terms of sensitivity. The increased sensitivity for the extended model emphasizes the added value of tailored Fe–S profile HMMs as new descriptors. It also underlines the benefit of combining profile HMMs with known descriptors to produce a generalized linear model, enriching the model beyond a single-pattern hit-based approach. Finally, these profiles may provide a new basis from which dedicated InterPro descriptors could be derived in order to improve current iron–sulfur protein coverage in this useful public resource.

Experimental validation of two Fe–S protein candidates

Among the fifteen Fe–S candidates from the *Escherichia coli* proteome that have not been annotated in Uniprot or predicted

Table 3 *E. coli* Fe–S protein candidates and their associated Fe–S descriptors. Fe–S protein candidates (columns) returned by the extended model are represented along with each descriptor that allowed for their prediction (lines), the weight associated with each descriptor is italicized. Detailed annotation of each Fe–S protein candidate is provided in Table S7

Descriptor identifier	Description	AHPF_ ECOLI	GLPA_ ECOLI	MUG_ ECOLI	NIRD_ ECOLI	TTDB_ ECOLI	YHCC_ ECOLI	BISC_ ECOLI	TORA_ ECOLI	TORZ_ ECOLI	YIDL_ ECOLI	YJIM_ ECOLI	YHAM_ ECOLI	PRET_ ECOLI	YDIJ_ ECOLI	YCBX_ ECOLI
PF00111	2Fe–2S iron–sulfur cluster binding domain															*0.29*
PF04055	Radical SAM superfamily						*0.96*									
PF04324	BFD-like [2Fe–2S] binding domain		*1.40*													
SSF102114	Radical SAM enzymes superfamily						*1.06*									
SSF46548	Alpha-helical ferredoxin superfamily													*1.35*	*1.35*	
SSF50022	[Rieske] ISP domain superfamily				*0.62*											
SSF54292	2Fe–2S ferredoxin-like superfamily															*0.70*
A5F890	Fumarate and nitrate reduction regulatory protein										*1.53*					
A7ZJJ0	Molybdenum cofactor biosynthesis protein A						*0.46*									
A8GH11	NADH-quinone oxidoreductase							*0.74*	*0.74*	*0.74*						
A8I8B7	Glutaredoxin	*1.35*														
B0BRR9	Lipoyl synthase						*0.26*									
B1Y6A6	Periplasmic nitrate reductase							*0.89*	*0.89*	*0.89*						
C4FIX5	Arsenite oxidase, small subunit				*0.47*											
Q5SJ65	Uracil-DNA glycosylase superfamily			*2.45*												
P08201	Nitrite reductase [NAD(P)H] large subunit	*1.18*	*1.18*											*1.18*		
P08500	Ubiquinol-cytochrome *c* reductase iron–sulfur subunit				*0.37*											
P09152	Respiratory nitrate reductase 1 alpha chain							*0.89*	*0.89*	*0.89*						
P0ABR8	Putative dioxygenase subunit alpha yeaW				*0.46*											
P14407	Fumarate hydratase class I, anaerobic					*2.71*										
P18775	Anaerobic dimethyl sulfoxide reductase chain A							*0.80*	*0.80*	*0.80*						
P25550	Anaerobic sulfatase-maturating enzyme homolog AslB						*0.63*									
P30140	2-Iminoacetate synthase						*0.10*									
P32131	Oxygen-independent coproporphyrinogen-III oxidase						*0.09*									
P42630	L-Serine dehydratase tdcG												*2.99*			
P44428	2Fe–2S ferredoxin															*0.30*
P44893	Fumarate reductase iron–sulfur subunit														*0.40*	*0.40*

Table 3 (*continued*)

Descriptor identifier	Description	AHPF_ ECOLI	GLPA_ ECOLI	MUG_ ECOLI	NIRD_ ECOLI	TTDB_ ECOLI	YHCC_ ECOLI	BISC_ ECOLI	TORA_ ECOLI	TORZ_ ECOLI	YIDL_ ECOLI	YJJM_ ECOLI	YHAM_ ECOLI	PRET_ ECOLI	YDIJ_ ECOLI	YCBX_ ECOLI
O54050	Xanthine dehydrogenase iron–sulfur-binding subunit															*0.58*
B8HK01	Oxidoreductase FAD/NAD(P)-binding domain protein						*0.51*									*0.30*
Q2LSB7	Iron only hydrogenase large subunit															*0.49*
P07839	Ferredoxin															*0.54*
Q12PT7	Ribosomal RNA large subunit methyltransferase N						*0.10*									
Q188I6	Subunit of oxygen-sensitive 2-hydroxyisocaproyl-CoA dehydratase C											*2.77*				
Q31XV0	3-Phenylpropionate/cinnamic acid dioxygenase ferredoxin subunit				*0.46*											
Q47GC4	Rieske (2Fe–2S) region				*0.48*											
Q48FA7	Ribosomal protein S12 methylthiotransferase RimO						*0.51*									
Q5NZV2	Ethylbenzene dehydrogenase, alpha subunit							*0.51*	*0.51*	*0.51*						

to be Fe–S proteins, we chose to experimentally validate two hypothetical proteins for which confidence scores were high, guided by descriptor coverage for each protein, as reported in Table 3. The rationale for this choice was that YhcC protein is covered by four descriptors (PF04055, SSF102114, P25550 (AslB) and Q48FA7 (RimO)) (Table 3), strongly suggesting that it could be a new Radical-SAM enzyme with an atypical Fe–S cysteinyl ligation motif. Indeed, it contains a $CX_{11}CX_2C$ motif, rather than the CX_3CX_2C motif characteristic of the Radical-SAM family.[24] For YdiJ, we were excited by the fact that the *ydiJ* gene is located just downstream the *suf* operon in the *E. coli* genome and therefore may be linked to Fe–S assembly machinery. In addition, this protein is identified by the descriptor SSF46548 which is derived from members of the ferredoxin superfamily (Table 3).

YhcC and YdiJ as two new *E. coli* Fe–S proteins

yhcC and *ydiJ* were both inserted into a pET vector allowing their expression with a His-tag. Both proteins were expressed aerobically in BL21(DE3) cells (Fig. S1, ESI†). As potential Fe–S proteins, they were purified under anaerobic conditions using a Ni-NTA column. Under these conditions, pure YhcC (inset Fig. 4A) was colourless and devoid of any iron or sulfur atoms. It exhibited a tendency to precipitate which could be due to the absence of its stabilizing metallic cofactor. To confirm this, YhcC was chemically reconstituted under anaerobic conditions with an excess of ferrous iron and inorganic sulphide. This technique has been extensively used in the past with Fe–S proteins in our laboratory.[59] After incubation and desalting, the brownish stable protein displayed a UV-visible spectrum characteristic of an Fe_4S_4 protein with a main absorption band at 420 nm (Fig. 4). Quantitation revealed the protein to contain 3 iron and 3.6 sulphide molecules per polypeptide chain, thus strongly suggesting the presence of one Fe_4S_4 cluster per monomer. Radical-SAM enzymes catalyse the reductive cleavage of *S*-adenosyl-methionine (SAM) into methionine and 5′-deoxyadenosine (AdoH).[84] When Fe–S-containing YhcC was incubated with SAM and dithionite for 1 hour, AdoH was produced (Fig. 4B). In contrast, no AdoH was produced in the presence of the apo-YhcC or in the absence of YhcC. These results indicate that (i) YhcC can cleave SAM into AdoH and methionine like Radical-SAM proteins, and (ii) Fe–S-containing YhcC is the active form of the protein. Taken together, these properties identify YhcC as a Radical-SAM. It is interesting that YhcC cleaved SAM in the absence of any additional substrate (still unknown), this is also the case for a few other Radical-SAM enzymes, including Fo synthase, HydG, PqqE and BioB.[85–88]

YdiJ also expresses well in *E. coli* cells (Fig. S1, ESI†). Anaerobically purified YdiJ was yellow-red in colour and contained 4 iron and 5 sulfur atoms per polypeptide chain. The UV-visible spectrum of the as-isolated YdiJ exhibited absorption bands at 280 nm, 360 nm, 400 nm, 440 nm and 459 nm, and a shoulder was visible on the spectrum at 500 nm (Fig. 5). This spectrum is not typical for classical Fe–S cluster proteins which usually have absorption bands at 420 nm (for Fe_4S_4), or 320 nm,

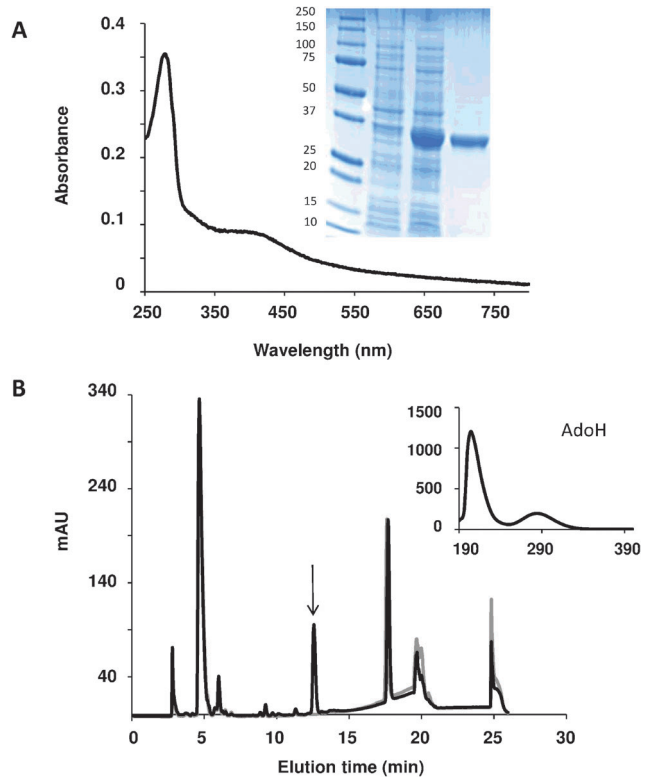

Fig. 4 Characterization of YhcC. (A) UV-vis spectrum of anaerobically reconstituted Fe_4S_4 YhcC, inset: SDS-PAGE showing purification of YhcC on Ni-NTA column (1: standard molecular weights; 2: BL21(DE3) cells before inducing protein expression; 3: BL21(DE3) cells 3 h after inducing protein expression; 4: as-purified YhcC after imidazole elution from Ni-NTA column); (B) HPLC and UV-vis detection of AdoH produced by YhcC-dependent SAM cleavage. The chromatogram corresponds to analysis of assay mixture containing YhcC (10 μM) in 50 mM Tris-HCl pH 8, 50 mM NaCl, 2 mM dithionite and 200 μM SAM (black trace). A control without YhcC was also performed (grey trace). HPLC elution conditions are described in "Materials and methods". The arrow corresponds to AdoH produced. Inset: UV-vis detection of AdoH. mAU: milli-arbitrary units.

420 nm and 460 nm (for Fe_2S_2). Thus, YdiJ may contain an additional cofactor (Fig. 5). The yellowish colour of the solution suggests the presence of a flavin cofactor. After aerobic heat treatment and centrifugation of the protein solution, the UV-visible spectrum recorded for the supernatant resembled that of a free flavin cofactor, with characteristic absorption bands at 370 nm and 450 nm. MS analysis performed on the as-isolated YdiJ protein unambiguously established the exclusive presence of flavin adenine dinucleotide FAD (Fig. 5C). Using the extinction coefficient, ε, at 450 nm for free FAD (11 300 mM^{-1} cm^{-1}), we calculated a ratio of about 1.1 mol of FAD per mol of protein from the UV-visible spectrum of as-isolated YdiJ. The UV spectrum of the YdiJ Fe–S cluster was obtained by subtracting the FAD spectrum from that of the as-isolated protein (Fig. 5A). Given that YdiJ contains 4 iron and 5 sulfur molecules per polypeptide chain, it could either contain two Fe_2S_2 clusters or one Fe_4S_4 cluster. To discriminate between Fe_2S_2 and Fe_4S_4 clusters, we performed EPR spectroscopy (Fig. 5). Upon reduction with dithionite, the protein solution became instantaneously colourless in agreement with the loss of all the absorption bands in the UV-visible spectrum, suggesting that both the flavin and the Fe–S cluster were reduced (Fig. 5B). No band at 550 nm, characteristic of reduced Fe_2S_2 clusters, was observed. An axial EPR signal with g values at 1.96 and 2.01 was obtained. The temperature dependence of this signal strongly suggested the presence of an Fe_4S_4 cluster (Fig. 5). Altogether, our results on YdiJ indicate that it is an Fe_4S_4 FAD-containing protein.

Genericity of the predictive model

To check whether the main descriptors used are not merely a bias due to the training dataset used, we also applied the model to other recently re-annotated complete proteomes to determine its robustness (*i.e.*, how generic the model is over different proteomes). The following six recently re-annotated complete proteomes were selected: *Acinetobacter baumannii* (ACIBY),[53] *Bradyrhizobium* sp. (BRASO),[54] *Clostridium difficile* (CLOD6),[55] *Helicobacter pylori* (HELPB),[56] *Neisseria meningitidis* (NEIM8),[57] *Vibrio splendidus* (VIBSL).[58] In these tests, we shall not refer to false positives and negatives, but rather to differences between database annotations and computational predictions. Thus, false positives are candidate new Fe–S proteins and can be used to quantify the added value of the predictive model. False negatives are not necessarily missed Fe–S proteins, since UniProt annotations cannot be fully relied upon owing to some annotations being inferred automatically by other predictive approaches. Fig. 6 shows the contingency tables for the Fe–S proteins predicted by the extended model compared to their UniProt Fe–S annotations (including proteins with an "unreviewed" status to enhance coverage). For all the proteomes analysed, we noticed very few annotation conflicts (from 0 to 7 undetected Fe–S proteins per complete proteome). In addition, the extended model was able to propose new Fe–S candidates (from 30 for *H. pylori* to 116 for *C. difficile*), suggesting that the extended model could be helpful in completing/extending current Fe–S protein annotations by proposing new candidates even though the actual number of Fe–S proteins is likely to be underestimated, as stated elsewhere.[30] The results of all of these predictions can be downloaded from the IronSulfurProteHome website (see below).

Proteome-wide screening for Fe–S proteins in prokaryotes

Andreini *et al.*[30] have shown that non-heme iron binding proteins account for, on average, 3.9 ± 1.6% of all proteins in bacteria, and do not exceed 10%. In another study, Major *et al.*[22] observed that a high optimal growth temperature range or an anaerobic mode of life (oxygen intolerance) correlated well with a high proportion of Fe_4S_4 proteins. To explore the potential benefit of our model in detecting these trends, we applied it to a set of 556 prokaryotic proteomes (covering both Bacteria and Archaea kingdoms) by randomly selecting one organism per species. The resulting prediction was used to determine the ratio between the number of predicted Fe–S proteins relative to the proteome size for each organism. The Fe–S protein content averaged at 2.37 ± 1.31%, with large

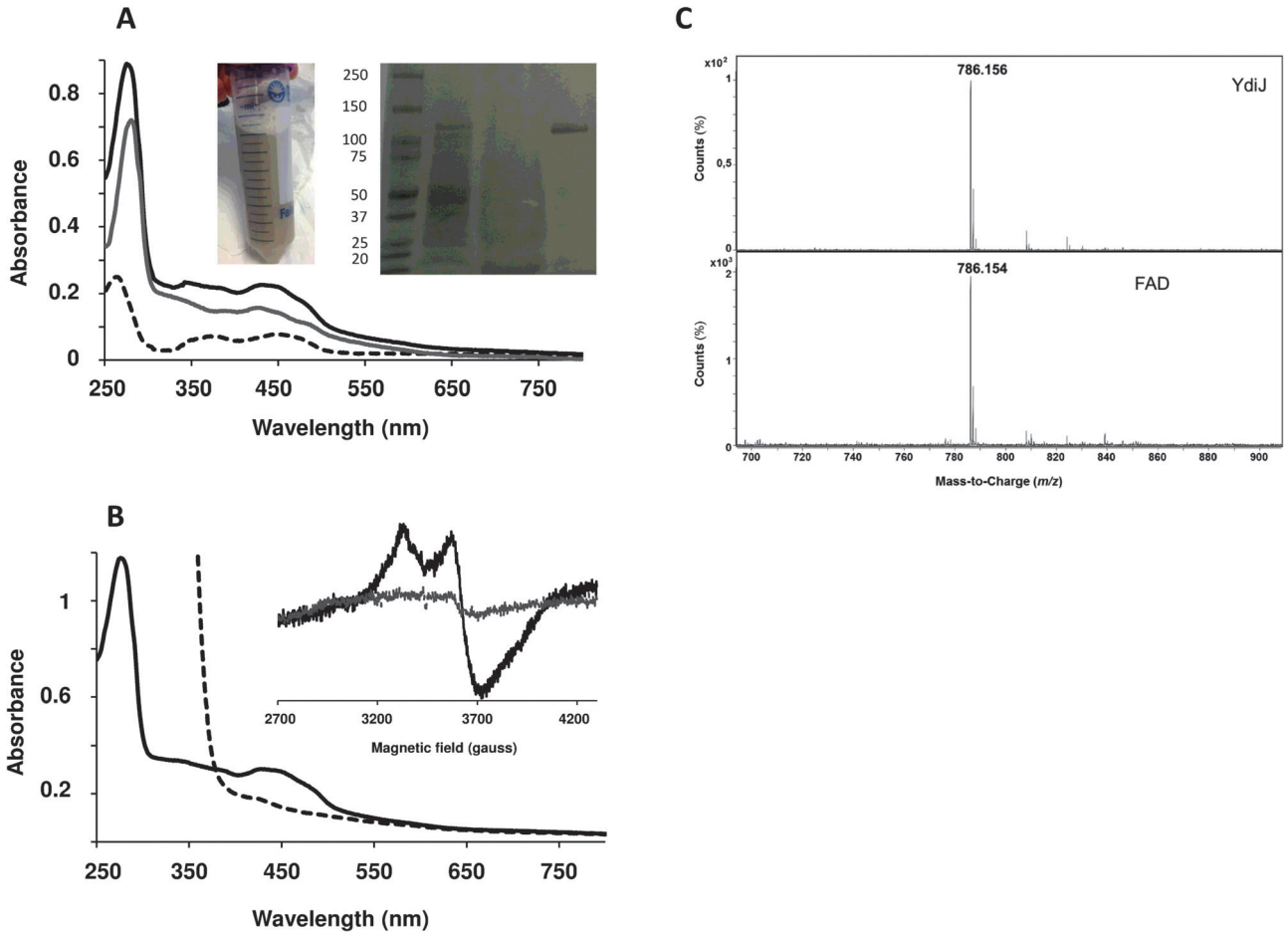

Fig. 5 Characterization of YdiJ. (A) UV-vis spectrum of anaerobically purified YdiJ (black trace); UV-vis spectrum of heated YdiJ supernatant (dashed black trace) showing FAD cofactor; UV-vis spectrum of Fe–S YdiJ (grey trace) produced by subtracting the UV-vis spectrum of FAD from the UV-vis spectrum for anaerobically purified YdiJ. Insets: cells after 2.5 h YdiJ over-expression (left) and SDS-PAGE showing YdiJ purification on Ni-NTA column (right) (1: standard molecular weights; 2: soluble extracts after over-expression; 3: Ni-NTA washing; 4: YdiJ after elution with imidazole). (B) UV-vis and EPR spectra of dithionite-reduced YdiJ. Anaerobically purified YdiJ (90 μM, black trace) was reduced with 1 mM dithionite for 20 min (dashed trace) before analysis by EPR spectroscopy. Conditions: Mod: 10 gauss; temperature: 4 K (black trace) and 20 K (grey trace), gain: 2×10^5. (C) ESI mass spectrum of anaerobically purified YdiJ (top) and commercial FAD (bottom).

variations from one species to another. The highest normalized Fe–S protein content was 6.94% in *Dehalococcoides* sp. *CBDB1*, anaerobic bacteria, while the lowest was 0% in *Mycoplasma* species, which are obligate intracellular organisms (host-associated) and facultative aerobes. We noticed that the difference in Fe–S protein content is more pronounced when considering Archaea and the Bacteria separately, with average values of 3.84 ± 1.21% and 2.07 ± 1.12%, respectively. We also ranked the results from the highest normalized Fe–S proteins content to the lowest and observed that, for the first 20 organisms, this value apparently positively correlates with an anaerobic mode of life (Table S9, ESI†). This observation is supported by previous studies.[22,89] This analysis also highlighted that Fe–S protein content could correlate either with ecological niche, such as an aquatic environment, or with living conditions, such as a high-temperature. These preliminary results also indicate that the model trained on PDB and *E. coli* sequences selected features that were relevant to other non-model species, even with different modes of life, evolutionary history and thus different sequence compositions. Further studies on the overall set of prokaryotic species are ongoing to explore these trends in detail with regard to their mode of life and evolution.

IronSulfurProteHome, a website dedicated to bacterial iron–sulfur proteomes

All the results of this study have been made publicly available through an interactive web platform (http://biodev.extra.cea.fr/isph). This web resource can be used to retrieve predicted Fe–S proteins either by the name of the organism (556 prokaryotic species), or by a UniProt accession number (for a total of 36 799 entries). Together with these results, the IronSulfurProteHome website also provides additional resources such as a downloadable list of predicted Fe–S proteins for each microbial organism and ESI,† including the present datasets used to select Fe–S proteins (http://biodev.extra.cea.fr/isph/ref.html).

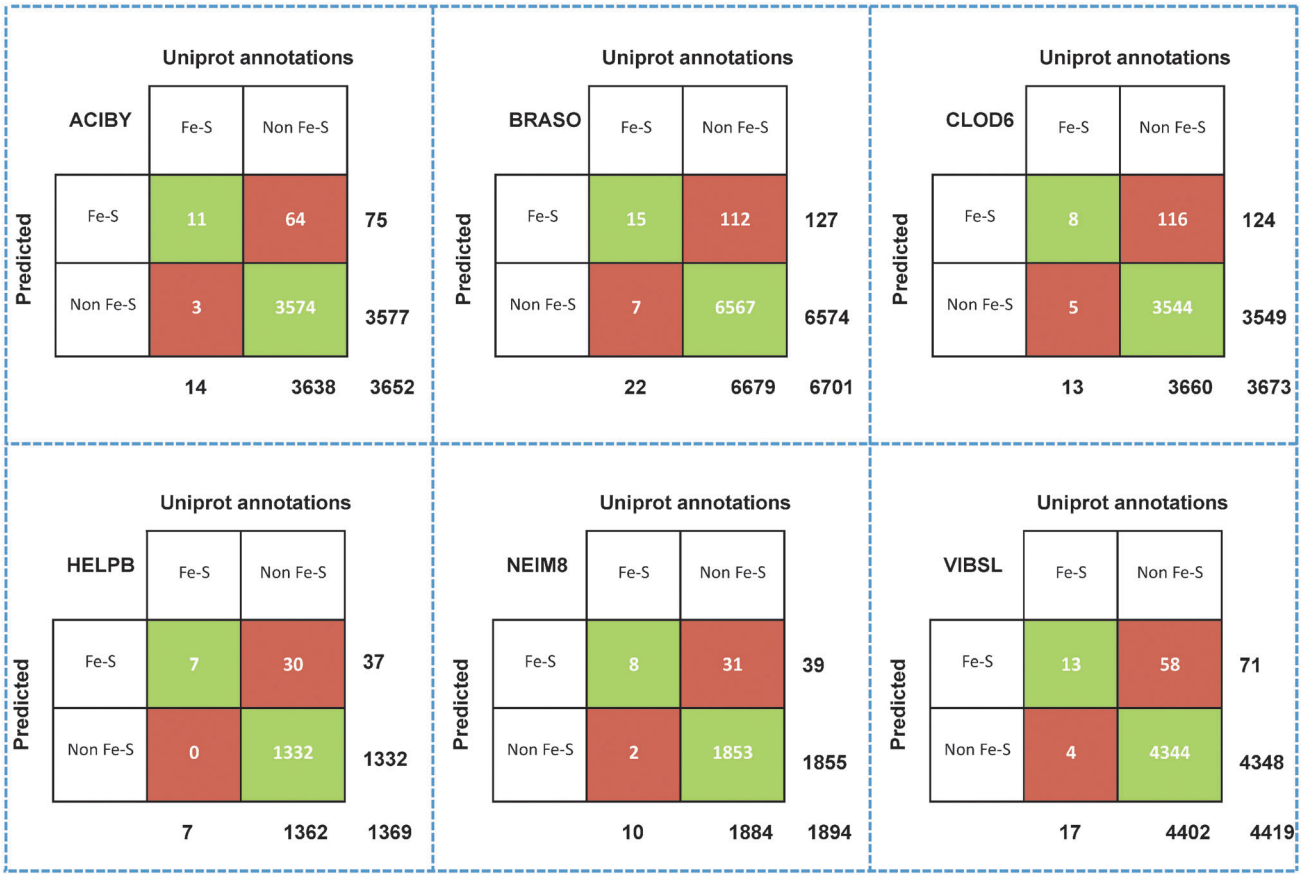

Fig. 6 Comparison of Fe–S prediction with database annotations on six re-annotated bacterial proteomes. Fe–S proteins predicted by the extended model were compared with UniProt annotations (reviewed or not) for six recently re-annotated bacterial species. ACIBY: *Acinetobacter baumannii*, BRASO: *Bradyrhizobium* sp., CLOD6: *Clostridium difficile*, HELPB: *Helicobacter pylori*, NEIM8: *Neisseria meningitidis*, VIBSL: *Vibrio splendidus*. Values in the contingency tables give the concordance (green cells) and conflicts (red cells) between annotations and predictions, respectively; the lower-left counts give the added value of the extended model and the upper-left counts give the number of Fe–S proteins missed by the predictive model. For each organism, the total number of proteins is reported for each line (predicted Fe–S and non-Fe–S proteins, right part) and each column (annotated Fe–S and non-Fe–S proteins, lower part of the table).

Conclusion and perspectives

Along with mass spectrometry techniques and high-throughput X-ray absorption spectroscopy, computational bioinformatics analysis has been considered as the third pillar of the metalloproteomics field.[5,90] Here, we assessed a probabilistic approach by designing a sensitive and interpretable linear model to predict Fe–S proteins in prokaryotic proteomes. This study represents a proof-of-concept of the benefit of penalized generalized linear models that provide a good compromise between precision and recall compared to motif-based approaches. While domain- and motif-based approaches provide a good level of specificity for metal-binding pattern recognition, profile HMMs designed using the HHpred toolbox supply increased sensitivity (recall). This added value is mainly due to the inclusion of structural information in these descriptors. In addition, the use of pairwise alignment of profile HMMs make this approach particularly effective for the detection of distant remote homologies, even those falling below the twilight zone of sequence similarity (20% sequence similarity).[67,68,91] Considering these properties, the use of such profiles could renew the way metalloprotein descriptors are designed. Using the penalized linear model we have shown how the weights provide a quantitative estimate of the most relevant descriptors for Fe–S protein identification. We also noticed that some commonly used strict signatures (such as CX_nCX_pC motifs) should be used with caution when seeking to identify Fe–S proteins. Interestingly, the model showed that no less than 67 descriptors are needed to nearly cover the known part of the iron–sulfur proteome of *E. coli* while also suggesting new candidates. This observation confirms the extraordinary diversity of this metalloprotein family due to the structural and electronic plasticity of Fe–S clusters. It also highlights the complexity of the process required for metalloprotein identification by computational approaches.[7,12,92] This limitation is probably due to an incomplete understanding of the complex determinants of metal-binding specificity in proteins. Along this line, new Fe–S proteins delivered with direct structural and functional assays are expected to help in understanding their constants and specificity towards Fe–S cluster ligation, which in return could be ideally used to feed a model based on a machine

learning approach. From a methodological point of view, the dataset and computational model can be viewed as two sides of the same issue; both might be considered continuously in parallel during the overall experimental workflow. First, well-annotated training and testing datasets must be established. Even though this is time-consuming, improvements are necessary both in terms of size and reliability to better account for the diversity of metalloproteins. Secondly, the application of penalized linear models to versatile objects such as metalloproteins clearly reflects their ability to capture more complex and subtle patterns than single conserved domain hits combined with a transparent scoring scheme. Thus, efforts towards producing publicly available Fe–S protein datasets with confident annotations should be encouraged so as to improve current computational approaches. Importantly, the close collaboration we have developed throughout this work has been fruitful not only because we shared knowledge and data during the design of the model, but also by filling the gap between prediction and experimental validation, which often limits bioinformatics outcomes. In fact, the penalized linear model guided us in our choice of Fe–S protein candidates for experimental characterization based on weights and annotation of the related contributing descriptor(s). From the biological point of view, the experimental validation of YhcC and YdiJ – which are currently annotated 'uncharacterized' in public databases – as two genuine Fe–S proteins, provides new information on this protein family while also contributing to the long road towards completing the *E. coli* iron–sulfur proteome.[40] In the same spirit, the large-scale screening of more than 550 prokaryotic proteomes establishes some new foundations for more comprehensive exploration of iron–sulfuromes in non-model organisms. Investigations questioning how these fascinating objects have pervaded the world through the evolution of living organisms are underway. Finally, we think that this bioinformatics approach may apply to other family of metalloproteins (*e.g.* copper proteins, zinc proteins) provided that (i) datasets are carefully established, (ii) descriptors are properly selected and designed and (iii) the parameters of the linear model are rightly optimized during the training phase.

Metalloproteomics promises to deliver novel insights into fundamental biological processes in the future, but it is clear that further methodological advances are necessary to exploit the full potential of this emerging research area.[8,92] By making these results available, we hope that this work will stimulate the community working in the bioinorganic chemistry field, and will open new avenues towards methodological development for high-throughput identification of metalloproteins.

Conflict of interests

All the authors of this work declare that they have no conflict of interest.

Authors' contributions

JE, SOC and YV designed the experiments. JE implemented the predictive model. JE and YV performed the computational analyses and developed the website. SOC and MF produced the curated datasets. SOC and MS performed the experimental validation. JE, SOC, MF and YV analysed the results. SOC and YV wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

JE is funded by the International PhD Program at the Life Sciences department (DSV) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA). We thank Dr C. Médigue who advised us on selection of the re-annotated bacterial genomes and Dr A. Viari for his technical advice during model construction. We thank the DSV/IBITEC-S/GIPSI team, particularly Arnaud Martel, for technical assistance and hosting the server installation. We are grateful to Dr L. Signor (IBS Grenoble) for MS analysis on YdiJ and to Maighread Gallagher-Gambarelli for suggestions on language usage.

References

1 P. J. Kiley and H. Beinert, *Curr. Opin. Microbiol.*, 2003, **6**, 181–185.
2 K. J. Waldron, J. C. Rutherford, D. Ford and N. J. Robinson, *Nature*, 2009, **460**, 823–830.
3 C. L. Dupont, S. Yang, B. Palenik and P. E. Bourne, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17822–17827.
4 K. J. Waldron and N. J. Robinson, *Nat. Rev. Microbiol.*, 2009, **7**, 25–35.
5 C. A. Blindauer, J. P. Barnett and D. J. Scanlan, *Anal. Bioanal. Chem.*, 2012, **402**, 3311–3322.
6 K. Degtyarenko, *Bioinformatics*, 2000, **16**, 851–864.
7 I. Bertini and G. Cavallaro, *Metallomics*, 2010, **2**, 39–51.
8 W. Shi and M. R. Chance, *Curr. Opin. Chem. Biol.*, 2011, **15**, 144–148.
9 W. R. Gilks, B. Audit, D. De Angelis, S. Tsoka and C. A. Ouzounis, *Bioinformatics*, 2002, **18**, 1641–1649.
10 A. Valencia, *Curr. Opin. Struct. Biol.*, 2005, **15**, 267–274.
11 C. Andreini, I. Bertini and A. Rosato, *Acc. Chem. Res.*, 2009, **42**, 1471–1479.
12 A. Cvetkovic, A. L. Menon, M. P. Thorgersen, J. W. Scott, F. L. Poole, F. E. Jenney, W. A. Lancaster, J. L. Praissman, S. Shanmukh, B. J. Vaccaro, S. A. Trauger, E. Kalisiak, J. V Apon, G. Siuzdak, S. M. Yannone, J. A. Tainer and M. W. W. Adams, *Nature*, 2010, **466**, 779–782.
13 J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard and C. Chothia, *J. Mol. Biol.*, 1998, **284**, 1201–1210.
14 S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
15 C. Chothia and M. Gerstein, *Nature*, 1997, **385**, 579.
16 M. Babor, S. Gerzon, B. Raveh, V. Sobolev and M. Edelman, *Proteins*, 2008, **70**, 208–217.
17 M. Lippi, A. Passerini, M. Punta, B. Rost and P. Frasconi, *Bioinformatics*, 2008, **24**, 2094–2095.

18 B. K. Kuntal, P. Aparoy and P. Reddanna, *Protein Pept. Lett.*, 2010, **17**, 765–773.
19 K. Goyal and S. C. Mande, *Proteins*, 2008, **70**, 1206–1218.
20 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
21 J. Meyer, *JBIC, J. Biol. Inorg. Chem.*, 2008, **13**, 157–170.
22 T. A. Major, H. Burd and W. B. Whitman, *FEMS Microbiol. Lett.*, 2004, **239**, 117–123.
23 R. Thilakaraj, K. Raghunathan, S. Anishetty and G. Pennathur, *Bioinformatics*, 2007, **23**, 267–271.
24 H. J. Sofia, G. Chen, B. G. Hetzler, J. F. Reyes-Spindola and N. E. Miller, *Nucleic Acids Res.*, 2001, **29**, 1097–1106.
25 A. Chatterjee, Y. Li, Y. Zhang, T. L. Grove, M. Lee, C. Krebs, S. J. Booker, T. P. Begley and S. E. Ealick, *Nat. Chem. Biol.*, 2008, **4**, 758–765.
26 S. E. McGlynn, E. S. Boyd, E. M. Shepard, R. K. Lange, R. Gerlach, J. B. Broderick and J. W. Peters, *J. Bacteriol.*, 2010, **192**, 595–598.
27 S. S. Kamat, H. J. Williams, L. J. Dangott, M. Chakrabarti and F. M. Raushel, *Nature*, 2013, **497**, 132–136.
28 C. Andreini, I. Bertini and A. Rosato, *Bioinformatics*, 2004, **20**, 1373–1380.
29 C. Andreini, L. Banci, I. Bertini and A. Rosato, *J. Proteome Res.*, 2006, **5**, 196–201.
30 C. Andreini, L. Banci, I. Bertini, S. Elmi and A. Rosato, *Proteins*, 2007, **324**, 317–324.
31 J. S. Sodhi, K. Bryson, L. J. McGuffin, J. J. Ward, L. Wernisch and D. T. Jones, *J. Mol. Biol.*, 2004, **342**, 307–320.
32 N. Shu, T. Zhou and S. Hovmöller, *Bioinformatics*, 2008, **24**, 775–782.
33 I. N. Kasampalidis, I. Pitas and K. Lyroudia, *Proteins*, 2007, **68**, 123–130.
34 S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. a. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats, *Nucleic Acids Res.*, 2009, **37**, D211–D215.
35 H. Zou and T. Hastie, *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 2005, **67**, 301–320.
36 H. Beinert, *JBIC, J. Biol. Inorg. Chem.*, 2000, **5**, 2–15.
37 M. K. Johnson, *Curr. Opin. Chem. Biol.*, 1998, **2**, 173–181.
38 R. Lill, *Nature*, 2009, **460**, 831–838.
39 M. Fontecave and S. Ollagnier-de-Choudens, *Arch. Biochem. Biophys.*, 2008, **474**, 226–237.
40 M. Fontecave, *Nat. Chem. Biol.*, 2006, **2**, 171–174.
41 M. Fontecave, B. Py, S. Ollagnier de Choudens and F. Barras, in *Escherichia coli and Salmonella.*, ed. T. P. Begley, 2008, ch. 3.6.3.14, DOI: 10.1128/ecosalplus.3.6.3.14.
42 A. Chan, M. Clémancey, J.-M. Mouesca, P. Amara, O. Hamelin, J.-M. Latour and S. Ollagnier de Choudens, *Angew. Chem., Int. Ed.*, 2012, **51**, 7711–7714.
43 M. Atta, E. Mulliez, S. Arragain, F. Forouhar, J. F. Hunt and M. Fontecave, *Curr. Opin. Struct. Biol.*, 2010, **20**, 684–692.
44 F. Forouhar, S. Arragain, M. Atta, S. Gambarelli, J.-M. Mouesca, M. Hussain, R. Xiao, S. Kieffer-Jaquinod, J. Seetharaman, T. B. Acton, G. T. Montelione, E. Mulliez, J. F. Hunt and M. Fontecave, *Nat. Chem. Biol.*, 2013, **9**, 333–338.
45 T. Lima, A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. de Castro, C. Lachaize, D. Baratin, I. Phan, L. Bougueleret and A. Bairoch, *Nucleic Acids Res.*, 2009, **37**, D471–D478.
46 T. U. Consortium, *Nucleic Acids Res.*, 2010, **38**, D142–D148.
47 P. Bucher, K. Karplus, N. Moeri and K. Hofmann, *Comput. Chem.*, 1996, **20**, 3–23.
48 J. Söding, A. Biegert and A. N. Lupas, *Nucleic Acids Res.*, 2005, **33**, W244–W248.
49 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
50 L. J. McGuffin, K. Bryson and D. T. Jones, *Bioinformatics*, 2000, **16**, 404–405.
51 J. Friedman, T. Hastie and R. Tibshirani, *J. Stat. Softw.*, 2010, **33**, 1–22.
52 R Development Core Team, *Vienna Austria R Found. Stat. Comput.*, 2008, 1, ISBN 3–900051–07–0.
53 D. Vallenet, P. Nordmann, V. Barbe, L. Poirel, S. Mangenot, E. Bataille, C. Dossat, S. Gas, A. Kreimeyer, P. Lenoble, S. Oztas, J. Poulain, B. Segurens, C. Robert, C. Abergel, J.-M. Claverie, D. Raoult, C. Médigue, J. Weissenbach and S. Cruveiller, *PLoS One*, 2008, **3**, e1805.
54 E. Giraud, L. Moulin, D. Vallenet, V. Barbe, E. Cytryn, J.-C. Avarre, M. Jaubert, D. Simon, F. Cartieaux, Y. Prin, G. Bena, L. Hannibal, J. Fardoux, M. Kojadinovic, L. Vuillet, A. Lajus, S. Cruveiller, Z. Rouy, S. Mangenot, B. Segurens, C. Dossat, W. L. Franck, W.-S. Chang, E. Saunders, D. Bruce, P. Richardson, P. Normand, B. Dreyfus, D. Pignol, G. Stacey, D. Emerich, A. Verméglio, C. Médigue and M. Sadowsky, *Science*, 2007, **316**, 1307–1312.
55 M. Monot, C. Boursaux-Eude, M. Thibonnier, D. Vallenet, I. Moszer, C. Medigue, I. Martin-Verstraete and B. Dupuy, *J. Med. Microbiol.*, 2011, **60**, 1193–1199.
56 J.-M. Thiberge, C. Boursaux-Eude, P. Lehours, M.-A. Dillies, S. Creno, J.-Y. Coppée, Z. Rouy, A. Lajus, L. Ma, C. Burucoa, A. Ruskoné-Foumestraux, A. Courillon-Mallet, H. De Reuse, I. G. Boneca, D. Lamarque, F. Mégraud, J.-C. Delchier, C. Médigue, C. Bouchier, A. Labigne and J. Raymond, *BMC Genomics*, 2010, **11**, 368.
57 C. Rusniok, D. Vallenet, S. Floquet, H. Ewles, C. Mouzé-Soulama, D. Brown, A. Lajus, C. Buchrieser, C. Médigue, P. Glaser and V. Pelicic, *Genome Biol.*, 2009, **10**, R110.
58 F. Le Roux, M. Zouine, N. Chakroun, J. Binesse, D. Saulnier, C. Bouchier, N. Zidane, L. Ma, C. Rusniok, A. Lajus, C. Buchrieser, C. Médigue, M. F. Polz and D. Mazel, *Environ. Microbiol.*, 2009, **11**, 1959–1970.
59 S. Wollers, G. Layer, R. Garcia-Serres, L. Signor, M. Clemancey, J.-M. Latour, M. Fontecave and S. Ollagnier de Choudens, *J. Biol. Chem.*, 2010, **285**, 23331–23341.
60 W. W. Fish, *Methods Enzymol.*, 1988, **158**, 357–364.

61 H. Beinert, *Anal. Biochem.*, 1983, **131**, 373–378.
62 E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez, *Nucleic Acids Res.*, 2005, **33**, W116–W120.
63 M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, *Nucleic Acids Res.*, 2012, **40**, D290–D301.
64 D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia and J. Gough, *Nucleic Acids Res.*, 2009, **37**, D380–D386.
65 N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni and C. J. a. Sigrist, *Nucleic Acids Res.*, 2006, **34**, D227–D230.
66 J. Gough, K. Karplus, R. Hughey and C. Chothia, *J. Mol. Biol.*, 2001, **313**, 903–919.
67 J. Söding, *Bioinformatics*, 2005, **21**, 951–960.
68 G. Yona and M. Levitt, *J. Mol. Biol.*, 2002, **315**, 1257–1275.
69 E. L. Sonnhammer, S. R. Eddy and R. Durbin, *Proteins*, 1997, **28**, 405–420.
70 T. P. Hilbert, W. Chaung, R. J. Boorstein, R. P. Cunningham and G. W. Teebor, *J. Biol. Chem.*, 1997, **272**, 6733–6740.
71 T. M. Iverson, C. Luna-Chavez, L. R. Croal, G. Cecchini and D. C. Rees, *J. Biol. Chem.*, 2002, **277**, 16124–16130.
72 C. De Mol, E. De Vito and L. Rosasco, *J. Complex.*, 2009, **25**, 201–230.
73 J. N. Agar, C. Krebs, J. Frazzon, B. H. Huynh, D. R. Dean and M. K. Johnson, *Biochemistry*, 2000, **39**, 7856–7862.
74 A. Saini, D. T. Mapolelo, H. K. Chahal, M. K. Johnson and F. W. Outten, *Biochemistry*, 2010, **49**, 9402–9412.
75 D. H. Flint, M. H. Emptage, M. G. Finnegan, W. Fu and M. K. Johnson, *J. Biol. Chem.*, 1993, **268**, 14732–14742.
76 B. M. a. van Vugt-Lussenburg, L. van der Weel, W. R. Hagen and P.-L. Hagedoorn, *PLoS One*, 2013, **8**, e55549.
77 J. D. Burman, R. L. Harris, K. a. Hauton, D. M. Lawson and R. G. Sawers, *FEBS Lett.*, 2004, **576**, 442–444.
78 E. V Soriano, Y. Zhang, K. L. Colabroy, J. M. Sanders, E. C. Settembre, P. C. Dorrestein, T. P. Begley and S. E. Ealick, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1685–1696.
79 H. Mihara, R. Hidese, M. Yamane, T. Kurihara and N. Esaki, *Biochem. Biophys. Res. Commun.*, 2008, **372**, 407–411.
80 D. E. Pierson and a. Campbell, *J. Bacteriol.*, 1990, **172**, 2194–2198.
81 P. C. Dos Santos and D. R. Dean, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 11589–11590.
82 R. M. Daniel and M. J. Danson, *J. Mol. Evol.*, 1995, **40**, 559–563.
83 R. Cammack, R. H. Jackson, a. Cornish-Bowden and J. a. Cole, *Biochem. J.*, 1982, **207**, 333–339.
84 P. a. Frey, A. D. Hegeman and F. J. Ruzicka, *Crit. Rev. Biochem. Mol. Biol.*, 2008, **43**, 63–88.
85 L. Decamps, B. Philmus, A. Benjdia, R. White, T. P. Begley and O. Berteau, *J. Am. Chem. Soc.*, 2012, **134**, 18173–18176.
86 J. K. Rubach, X. Brazzolotto, J. Gaillard and M. Fontecave, *FEBS Lett.*, 2005, **579**, 5055–5060.
87 S. R. Wecksler, S. Stoll, H. Tran, O. T. Magnusson, S.-P. Wu, D. King, R. D. Britt and J. P. Klinman, *Biochemistry*, 2009, **48**, 10151–10161.
88 S. Ollagnier-de Choudens, Y. Sanakis, K. S. Hewitson, P. Roach, E. Münck and M. Fontecave, *J. Biol. Chem.*, 2002, **277**, 13449–13454.
89 V. Gennis and R. B. Stewart, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. F. C. Neidhardt, American Society for Micro-biology, Washington, DC, 2nd edn, 1996, pp. 217–261.
90 W. Shi and M. R. Chance, *Cell. Mol. Life Sci.*, 2008, **65**, 3040–3048.
91 A. Hildebrand, M. Remmert, A. Biegert and J. Söding, *Proteins*, 2009, **77**(suppl 9), 128–132.
92 S. M. Yannone, S. Hartung, A. L. Menon, M. W. W. Adams and J. A. Tainer, *Curr. Opin. Biotechnol.*, 2012, **23**, 89–95.

