BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 17 2011, pages 2368–2375
doi:10.1093/bioinformatics/btr396

Structural bioinformatics

Advance Access publication June 30, 2011

Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins

Alexander Goncearenco[1,2] and Igor N. Berezovsky[1,*]

[1]Computational Biology Unit, Uni Research and [2] Department of Informatics, University of Bergen, N-5008 Bergen, Norway

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Enzymes are complex catalytic machines, which perform sequences of elementary chemical transformations resulting in biochemical function. The building blocks of enzymes, elementary functional loops (EFLs), possess distinct functional signatures and provide catalytic and binding amino acids to the enzyme's active sites. The goal of this work is to obtain primordial prototypes of EFLs that existed before the formation of enzymatic domains and served as their building blocks.

Results: We developed a computational strategy for reconstructing ancient prototypes of EFLs based on the comparison of sequence segments on the proteomic scale, which goes beyond detection of conserved functional motifs in homologous proteins. We illustrate the procedure by a CxxC-containing prototype with a very basic and ancient elementary function of metal/metal-containing cofactor binding and redox activity. Acquiring the prototypes of EFLs is necessary for revealing how the original set of protein folds with enzymatic functions emerged in predomain evolution.

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Contact: igor.berezovsky@uni.no

Received on April 1, 2011; revised on June 7, 2011; accepted on June 26, 2011

1 INTRODUCTION

Current evolution of proteins takes place via mutation and recombination of protein domains, leading to both divergence and convergence. It is rather obvious, however, that evolution of protein structure and function did not start from full protein folds (Iwasaki, 2010; Koonin, 2003), and the latter has to have been formed in a predomain stage of evolution. Therefore, one of the most important tasks in studies of the evolution of protein function is to find how the first set of folds with the most basic biochemical functions was formed and to determine the set of ancient functional peptides that gave rise to the above repertoire of folds (Berezovsky *et al*., 2003a, b; Trifonov *et al*., 2001).

Previous studies have demonstrated that closed loops of 25–30 amino acid residues are universal building blocks of protein folds (Berezovsky, 2003; Berezovsky and Trifonov, 2001; Berezovsky *et al*., 2000). According to the same studies, the closed loops emerged from ring-like peptides, primitive proteins or protein-like molecules in prebiotic evolution, and served as building blocks of the first protein folds. Similarly to folds that are built from closed loops as structural units, the enzymatic functions can also be decomposed into combinations of elementary units of protein function. The latter are closed loops that possess one or a few functional residues and bring them to the active site (de Gennes, 1990), which we call Elementary Functional Loops (EFLs; Goncearenco and Berezovsky, 2010). The functional signatures of some EFLs are shared between (super)families of proteins with different biochemical functions and even between different folds. Their prototypes presumably underwent recombination at the predomain stage of protein evolution. Recently, we have developed a computational procedure for finding sequence profiles of widely spread EFLs with characteristic functional signatures (Goncearenco and Berezovsky, 2010). In this procedure, we obtained sequence profiles from complete proteomes. Many of the profiles correspond to their originating protein families [such as those described in PFAM (Bateman *et al*., 2004) or CDD (Marchler-Bauer *et al*., 2005)] and represent family-specific functional signatures [as in Prosite, (Sigrist *et al*., 2010)], while others reveal connections between different folds and functions. Here, we would like to proceed further and find a way to obtain ancient prototypes of EFLs. Therefore, we developed a procedure for reconstructing prototypes, which served as basic units of the first folds/domains with enzymatic functions.

The procedure for obtaining prototypes is very different to that of ancestor reconstruction. The latter typically requires a phylogenetic tree built from the alignment of related (super)family members and an evolutionary model with particular mutation and amino acid substitution rates (Cai *et al*., 2004; Harms and Thornton, 2010; Mirkin *et al*., 2003). Here, on the contrary, we work with short sequence fragments belonging to phylogenetically unrelated proteins from remote (super)families, which presumably were involved in the design of the first enzymes in predomain evolution (Trifonov *et al*., 2001).

We illustrate the computational strategy that we have developed by reconstructing an ancient prototype with redox-active/metal-binding elementary function. The most generic signature of this prototype reads Cys-Xaa-Xaa-Cys (-CxxC- for brevity). We start with a minimal set of non-redundant profiles from which we reconstruct the prototype and its derivatives. Using these profiles in a profile-sequence search over complete proteomes and the protein databank, we detect a diversity of EFLs with metal/metal-containing cofactor binding and redox activity in different folds and show that they take part in different biochemical functions.

*To whom correspondence should be addressed.

2 METHODS

Procedure for obtaining and charactering profiles of EFLs: we introduce here a computational procedure for reconstructing the most ancient and generic prototypes of EFLs. Previous studies were mostly focused on detecting conserved functional motifs among homologous proteins with known functions. The distinctive feature of our method is that no preliminary assumptions about the homology or function are made, and prototypes are reconstructed from the comparison of all sequence segments on a proteomic scale. First, we cut several proteomes into initial 50-residue long segments (size of closed loop 25–30 residues plus 10-residue flanks) (Berezovsky *et al*., 2000), iteratively compare them to non-redundant proteomic sequences, and thus obtain sequence profiles represented as position-specific scoring matrices (PSSM) [for details see (Goncearenco and Berezovsky, 2010)]. Some of the obtained sequence profiles have clear functional signatures and correspond to EFLs from various non-homologous proteins. We reconstruct prototypes from profiles that have similar functional signatures using clustering procedure described below.

In order to characterize the profiles and their elementary functions we identified the corresponding EFLs in enzymatic domains with known biochemical function and structure by performing a profile-sequence search in the CDD database (Marchler-Bauer *et al*., 2005). In some cases it was possible to annotate the elementary function of individual residues in a profile's signature by analyzing the CDD annotation on the domain level, in other cases we used databases such as IBIS (Shoemaker *et al*., 2010), MACiE (Holliday *et al*., 2009), Prosite (Sigrist *et al*., 2010) and PDBeMotif (Golovin and Henrick, 2008).

Sequence profile clustering procedure: pairwise profile comparison and clustering have $O(N^2)$ complexity and memory requirements (Murtagh, 1984). Thus, processing of $10^4 - 10^5$ sequence profiles becomes unfeasible without a specifically developed method. We implemented a parallel program for pairwise comparison and UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering (Sokal, 1958) of the initial set of profiles. Based on the properties of the distance matrix we introduced a heuristic, which allowed speeding the computation up significantly. In particular, identification of disconnected components in the distance matrix (above a certain profile–profile distance threshold) splits a large unclustered dataset into several smaller sets that can be clustered independently. The program source code in C using MPI is available on request.

Overlap-based profile–profile comparison: profiles a and b with sets of matches A and B overlap if they match the same protein (same Gene Id) and positions of the matches are not further than 20 residues apart from each other. The distance between profiles is calculated using the Jaccard distance (Lipkus, 1999) $J(a,b) = 1 - |A \cap B|/|A \cup B|$, assuming that the intersection is the number of overlapping matches, and the union is the total number of unique matches in both profiles. If $J = 0$, then all the matches of profiles overlap; if $J = 1$, then profiles do not have any overlapping matches.

PSSM-based comparison of profiles: the distance between two profiles is calculated by comparing their PSSMs. In order to account for possible profile–profile alignments, we calculate superpositions of PSSMs with maximal offset ± 20 and compare the corresponding 30-residue windows. The distance between the windows is calculated as the Euclidean distance between aligned PSSM positions weighted by the total information gain (Kullback–Leibler divergence, D_{KL}) of amino acid frequencies in the aligned positions relative to the average proteomic frequencies of amino acids (Kullback and Leibler, 1951). From all possible superpositions of two profiles, the one giving the minimal distance between 3-residue windows a and b is used. The pairwise distance reads:

$$d = \sum_i \left(D_{KL}(a_i) + D_{KL}(b_i)\right) \|a_i - b_i\|.$$

Benchmark preparation: sequences of 40 archaeal proteomes (Supplementary Table ST1) were compared using BLASTP (Altschul *et al*., 1990) (E-value ≤ 0.0001) against a non-redundant (up to 40% pairwise sequence identity) SCOP/ASTRAL database of structural domains, release 1.75 (Murzin *et al*., 1995) in order to identify structural domains in the proteomes. Homologous domains were identified in 40% of the sequences (48 269 domains in 36 718 sequences). Redundancy between the identified domains has been removed with cd-hit program (Li and Godzik, 2006) down to 70% sequence identity. The benchmark required a set of superfamilies for which the coverage could be robustly measured. Therefore, we selected 150 SCOP superfamilies (TOP150) that were represented by more than 50 non-redundant domains in our dataset. The fold census of the TOP150 set is shown in Supplementary Figure S1.

3 RESULTS

The goal of this work is to obtain prototypes of elementary functions and their derivatives. We define a prototype as the most generic and ancient functional signature, and its derivatives as descendants (EFLs) with a diversity of contemporary functions. We develop a special benchmark for sequence search with profiles of EFLs, introduce and evaluate a new PSSM-based measure for profile-profile comparison, and reconstruct prototypes by clustering the profiles. As a case study, we explore a prototype with -CxxC- containing signature and its derivatives.

3.1 Sequence profile search benchmark

The functional diversity of a prototype can be estimated by the number of SCOP superfamilies found by its associated profiles of EFLs (assuming that the superfamilies themselves are unrelated). Since it is important to select out random hits from real matches to superfamilies we need a way to estimate the sensitivity (detection power) of a profile in the sequence search. A true match to a superfamily can be asserted by a sufficient coverage of the superfamily by sequence search. As in any sequence–sequence or profile–sequence search, we resort to the E-value parameter to control the number of expected false positives. We need to determine the minimal E-value, which gives the maximal number of true positive hits. Already existing benchmarks can only be used for evaluating the sequence–sequence and sequence–profile search on the level of whole proteins or domains, and not on the level of short sequence fragments. Therefore, we were prompted here to create the TOP150 benchmark, which we use to show that E-value calculated in our procedure is a correct estimate of the expected number of false positives. Figure 1A shows that our search procedure provides a clear separation between true and false hits by a plateau phase as function of E-value. The plateau widens with increasing coverage from 0% (black line) to 70% (green line), resulting in only true hits in superfamilies with high coverage (above 50%, yellow line). As a result, by controlling the coverage we obtain the number of 'true' superfamilies corresponding to the prototype. This level is indicated by an arrow in Figure 1A, showing where the red, blue and yellow plateaus are aligned. In the routine search, where coverage is not controlled (black line), the minimal E-value can be determined as the level at which the number of covered superfamilies reaches an already determined 'true superfamily plateau'. Correspondingly, the end of the plateau on the black line designates the maximal E-value. The interval between the plateau on the black line and the one on red/blue lines (few hits) determines the false positive rate expected in the search with no control of coverage. At lower E-values ($E < 0.0005$) several hits will be missed as 'false negatives'. At higher E-values ($E > 1$) false positive rate will increase exponentially (see the black line in the figure). The TOP150 benchmark helps to determine the optimal

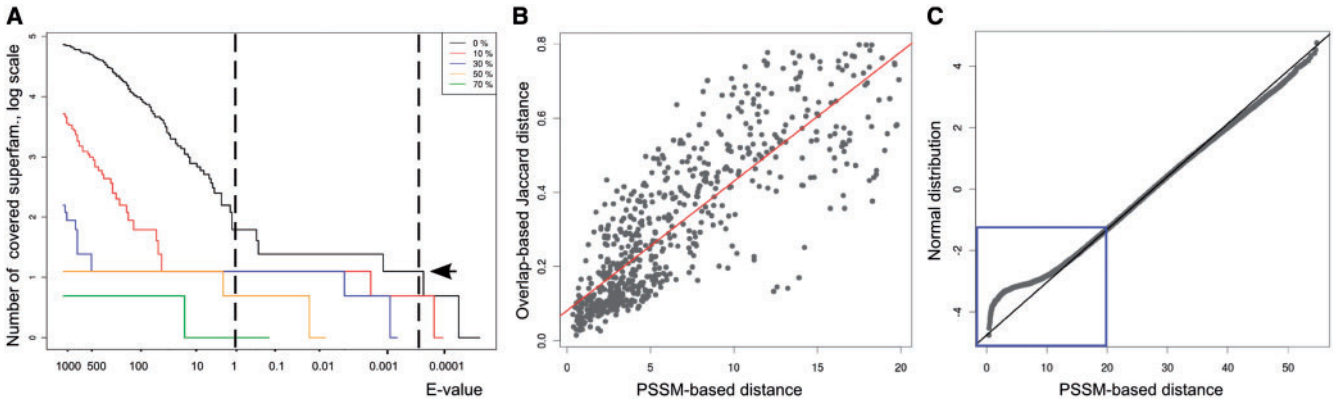

Fig. 1. Benchmark for profile-sequence search and profile–profile comparison. (**A**) Dependency between the *E*-value of profile-sequence search and the number of TOP150 superfamilies found with a certain level of coverage (color lines). The plateau phase observed at high coverage (shown with an arrow) indicates the number of 'true positive' superfamilies. The corresponding range of *E*-values ($0.0005 < E < 1$, shown between the vertical dashed lines) gives a realistic estimate of number of observed false positives. The interval between the plateau on the black line and the one on red/blue lines (a few hits) determines the false positive rate expected in the search with no control of coverage. (**B**) Correlation between the PSSM-based distances (d) and overlap-based distances (J) for the range of $d < 20$ (shown as blue box in C); (**C**) Quantile–quantile plot comparing the distribution of PSSM-based distances (d) with normal distribution (corresponding to distances between random profiles, see Supplementary Figure S2) shows that the former deviates significantly from the latter in the range of $d < 10$ (confidence level $\alpha_{d=10} = 0.016$).

range of *E*-values for profile–sequence search. We show that *E*-value provides a correct estimate of the number of false positives, and the benchmark can be used to determine the optimal range of *E*-values for profile–sequence search.

We also show here how to select a group of related derivatives from the complete set of profiles and how to reconstruct the prototype. In order to do that it is necessary to compare profiles with each other, supposing that the most similar profiles will be derivatives of the same prototype. The natural way to compare profiles would be to compare their matches. However, this comparison depends on the number of matches and may not correctly reflect the similarity between profiles in the case of low sequence coverage. An alternative way to compare profiles is to compare the PSSMs representing them. The advantage of this comparison is that it is independent of the coverage. However, it is necessary to find out how well matrix comparison reflects the match-based similarity between profiles, and to determine the range of profile–profile distances where the matrix comparison is applicable. Here, in order to evaluate the performance of the PSSM-based distance measure, we compare it with the overlap-based measure (see Methods). Figure 1B shows that PSSM-based distance (d) correlates well ($R^2 = 0.7$) with the overlap-based distance (J) in the range of distances $d < 20$ and $J < 0.8$. We use the distribution of pairwise distances between reshuffled profiles to estimate statistical significance of profile–profile distances (Supplementary Figure S2). Figure 1C shows that pairs of profiles with statistically significant similarities are observed in a narrow distance range below PSSM-based distance $d = 10$, where confidence level is 0.016.

3.2 Reconstruction of the -CxxC- prototype

We illustrate the computational strategy of prototype reconstruction with an example of a prototype with a very simple -CxxC- signature (two cysteines separated by two other residues). We reconstructed this prototype from a set of 11 derivatives, which we found to be related by profile clustering. These profiles formed a group with low profile–profile pairwise distances (in Supplementary Figure S2 the green line shows the density plot for pairwise distances in this group). Unrelated derivatives, as well as reshuffled profiles, have higher pairwise distances between them. We analyzed the representatives of the -CxxC- prototype, found by its derivatives in annotated protein domain families (Marchler-Bauer *et al.*, 2005; Shoemaker *et al.*, 2010), and concluded that the prototype has elementary functions of metal/metal-containing cofactor binding and redox activity.

Figure 2 shows the -CxxC- prototype (logo in the center), its 11 derivatives (red diamonds), and the CDD superfamilies (blue and green ovals) where representatives of the prototype and the derivatives (corresponding to EFLs) were found. CDD superfamilies with unknown function (corresponding to PFAM DUFs) were excluded from the Figure 2 for clarity. However, matches found in these superfamilies can be considered as predictions of elementary functions (listed in Supplementary Table ST2). Superfamilies, where more than one derivative has matches, are placed in the central circle. Most of the derivatives also have specific sets of superfamilies, which are shown in the outer circle. The specificity of these superfamilies is determined by the peculiarities of their signatures and the corresponding elementary functions. For instance, Derivative 1 matches medium chain reductase/zinc-dependent alcohol hydrogenase (MDR) where the corresponding EFL has a unique signature -CxxCxxCxxGx(4)C- and is responsible for binding of zinc ions (Figure 2f). Another example is one of the matches of Derivative 4 in Cytochrome C, where the specific functional signature –CxxCH- binds a heme (Figure 2a). The signature of copper binding -GM[TH]CXXC- represented by Derivative 8 is found in heavy metal-associated domains (HMA, Figure 2j). It is important to note that the specificity of one derivative prevents finding the matching EFLs by other derivatives. Also, it is impossible to find all related EFLs by the very prototype with -CxxC- signature. Even in a search with high *E*-values, the prototype is unable to detect the whole diversity of EFLs.

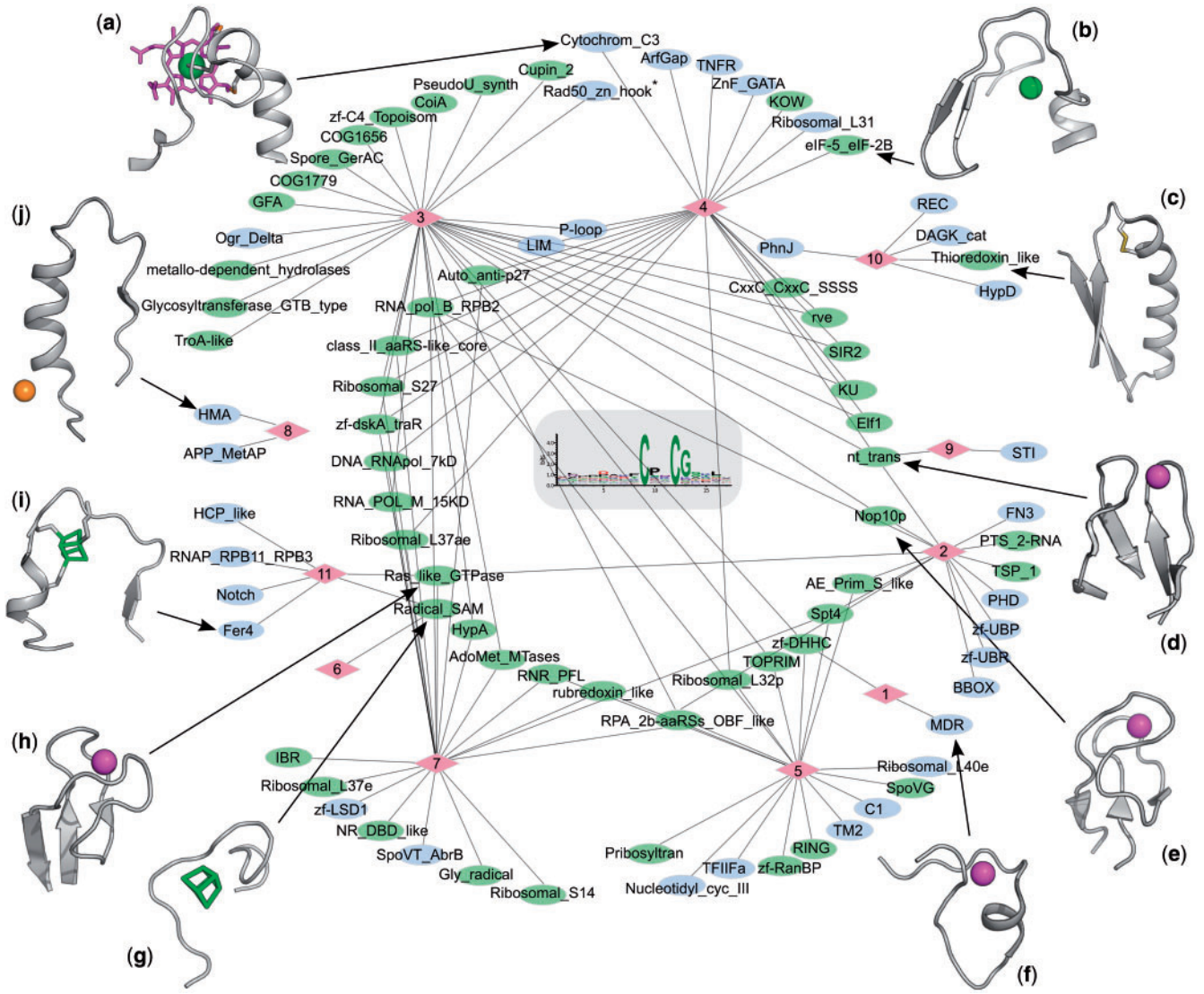

Fig. 2. The -CxxC- prototype (logo in the center), its derivatives (red diamonds), and their hits in CDD superfamilies (ovals and structures). The color of the ovals represents whether the superfamily is found by the prototype (green) or if it can only be found by the prototype's derivative (blue). *E*-value threshold of profile-sequence search is 1 for the derivatives and 145 for the prototype. The inner circle represents superfamilies containing matches from several derivatives. The structural fragments are shown with arrows pointing to the corresponding CDD superfamily, and are taken from the following proteins: (**a**) Cytochrome C553 with bound heme (PDB 1c75); (**b**) Archaeal translation factor Aif2beta which is a C4 zinc finger (PDB 1nee); (**c**) Thioredoxin from Protein disulfide oxidoreductases Thioredoxin superfamily (PDB 2trx); (**d**) Isoleucyl-tRna Synthetase from Nucleotidyl transferase superfamily (PDB 1jzs); (**e**) Archaeal Box HACA SRNP NOP10-Cbf5 complex from Nucleolar RNA-binding protein superfamily (PDB 2aus); (**f**) Alcohol Dehydrogenase from Medium chain reductase/dehydrogenase superfamily (PDB 1e3e); (**g**) MoaA molybdenum cofactor biosynthesis protein from the S-adenosylmethionine (SAM)-dependent radical enzyme superfamily (PDB 1tv8); (**h**) Large gamma subunit of Initiation Factor Eif2 from Ras-like GTPase superfamily which includes GTP translation factors (PDB 1kk3); (**i**) domain from a Ferredoxin-Cytochrome complex belonging to Fer4 [4Fe-4S] binding domain (PDB 1dwl_A); (**j**) PDB 1k0v, copper-transport CopZ protein domain from heavy-metal-associated domain superfamily (HMA). The bound cofactors are shown in colored spheres or sticks: magenta spheres, zinc ions; orange, copper ion; green sphere, iron ions; green sticks, [4Fe-4S] clusters; magenta sticks, heme C. Disulfide bond is shown in orange sticks in structure (c). (Asterisk) Rad50_zn_hook superfamily is not listed in the CDD release 2.25, and it was amended to the graph because it brings a new structural example (see Fig. 3). The graph was visualized with Cytoscape 2.8.1 (Shannon *et al*., 2003) and protein structures – with PyMol 1.3 (DeLano Scientific).

Blue ovals in Figure 2 show superfamilies where the prototype has no matches with *E*-value threshold 145. Green ovals show superfamilies where EFLs are detected by the prototype at this threshold. Notably, the inner circle is almost completely green, whereas the outer one contains mostly the specific superfamilies missed by the prototype. In general, even though all derivatives possess the -CxxC- signature, they differ in the other positions contributing to the information content of their profiles (logos in

Supplementary Table ST3). Derivatives with a higher number of informative positions have representatives in fewer superfamilies (see the third column in Table ST3). More generic signatures (closer to that of the prototype) always have fewer informative positions. For example, Derivative 3, whose signature has less than 8 informative positions (-CP[KRE]CG[GSA]x[LMV]-), is the derivative with the maximal number of superfamily matches. Thus, the prototype describes a very generic elementary function, but it looses the detection power for finding its representatives. The set of derivatives, in turn, complements the prototype allowing one to find representative EFLs in contemporary proteins.

In order to identify and characterize the elementary function of the prototype -CxxC- we explore the diversity of roles of the conserved cysteines in different biochemical contexts via the derivatives of the prototype and via the corresponding EFLs. The short motif with two cysteine residues separated by two other residues can be found in a variety of proteins and its function is well-studied (Chivers *et al*., 1996, 1997; Fomenko and Gladyshev, 2003; Quan *et al*., 2007). Early studies were mostly centered around the redox function performed by the CxxC signature. The focus on the redox role was achieved in two different ways: via sequence analysis of complete genomes and by selection of pairs of proteins with alternate redox states from the protein databank. In the first case, Gladyshev and colleagues derived a collection of sequences containing CxxS, SxxC, CxxT and TxxC motifs, avoiding motifs involved in metal binding. Furthermore, they analyzed occurrence, conservation and function-related modifications of the CxxC motif and its variants in proteins of thioredoxin fold (thioredoxins, glutaredoxins, protein disulfide isomerases and nucleoredoxins) (Fomenko and Gladyshev, 2003). Fan *et al.* (2009), on the contrary, scanned the protein databank, searching for redox-active Cys–Cys pairs. They described four major classes of structural changes between alternate redox states: disulfide oxidation following expulsion of metal; reorganization of polypeptide backbone; order/disorder transition; and changes in quaternary structure. They proposed that the above changes are associated with physiologically relevant redox activity. The results obtained by Gladyshev and colleagues and Fan *et al.* have important functional and evolutional implications, but are limited to redox function and obtained from, and applicable to, redox-related functional (super)families. Our goal is to reconstruct prototypes of EFLs, which are not restricted by conserved functional motifs in homologous proteins. Therefore, we derive the prototypes on a proteomic scale without any preliminary assumptions about sequence/structure homology or similarity of the function. As a result, we obtained the most generic prototype, which describes all redox functions along with metal and metal-containing cofactor binding ones. By analyzing EFLs representing the prototype in different biochemical functions and connections between remote functions, we unraveled long evolutionary history hidden in the signatures of the CxxC prototype and its derivatives. We show below that metal/metal-containing cofactor binding and redox elementary functions are intimately connected, and the -CxxC–containing prototype is their common evolutionary root.

3.2.1 Metal and metal-containing cofactor binding The cysteines in the -CxxC- signature often interact with metal ions (typically zinc and iron, but also other metals) and metal-containing cofactors, such as heme and [Fe-S] clusters. Metals and metal-containing cofactors, in turn, can assist biochemical reactions. The cysteine residues in -CxxC- and additional residues such as methionine, as in the case of Derivative 8 and HMA, can also bind other metals, such as copper, cadmium and cobalt. One of the important biological functions implied by metal and metal-containing cofactor binding EFLs is detoxification of heavy metals by binding and transferring them in metal chaperones. A special structural role of cysteine-coordinated metals is revealed in zinc fingers (Figure 2b, d and e) where the metal (not necessarily a zinc ion) stabilizes the minimalistic fold. Many zinc finger-containing proteins are also involved in catalysis as DNA- or RNA-affecting enzymes (polymerases, isomerases, primases, nucleotidyl transferases). For example, the topoisomerase-primase (TOPRIM) superfamily (Derivatives 3 and 5 in Figure 2) contains an ATP-dependent reverse gyrase. Some of the derivatives also bind metal-containing cofactors, such as iron–sulfur cluster, which is a key electron carrier in central metabolic and energetic pathways (Johnson *et al*., 2005). The cluster has different potentials in different configurations, such as [Fe-S], [2Fe-2S], [4Fe-4S] and [3Fe-4S] (Meyer, 2008). Ferredoxin (Fer4) is a very compact and short [4Fe-4S] binding domain. It has less than 60 amino acid residues, and is structurally composed of three closed loops. Binding of iron-sulfur clusters by ferredoxins (as exemplified by the EFL structures in Figure 3g and i) is an ancient and essential function (Dupont *et al*., 2010). Another cofactor, which is typically bound by EFLs with the -CxxC- functional signature, is heme. For instance, Heme C is covalently connected to cysteines in Cytochrome C553 (Derivative 4, Figure 3a), which are port of a -CxxCH- signature. The functionally important histidine, which is part of the signature, coordinates the iron in the heme.

3.2.2 Redox chemistry of the -CxxC- signature The -CxxC-containing EFLs are able to perform redox chemistry without any cofactors, as exemplified by thioredoxins (Figure 2c). Depending on the biochemical context, the role of the disulfide/dithiol in -CxxC- can be different. For example, members of the Thioredoxin-like superfamily contains redox active dithiols/disulfide bond within the -CxxC- motif. Redox-active dithiols in thioredoxins (derivative 10, structure c in Figure 2) play a protective role as antioxidants. For example, oxidation of cysteine residues that bind zinc in transcription factors can have deleterious implications for gene expression (Wilcox *et al*., 2001). Thioredoxin and Glutaredoxin with -CxxC- signature have potential functions as facilitators and regulators of protein folding and chaperone activity. They bind unfolded proteins and act as chaperones and isomerases of disulfides to generate a native fold (Berndt *et al*., 2008). The differences in redox properties of Thioredoxin proteins are attributed to variation of the -CxxC- motif (Atkinson and Babbitt, 2009; Fomenko and Gladyshev, 2003). Besides, the -CxxC- prototype (as shown in logo in the center of Figure 3) contains more than just two cysteines. The proline and glycine (in the CPxCG signature) can play an important structural role, providing the conformation necessary for redox chemistry.

3.2.3 Connections between remote superfamilies revealed by the -CxxC- prototype and its derivatives Figure 3 shows the relationships between the prototype, its derivatives and the corresponding EFLs. Derivatives 1, 4 and 10 (Figure 3), described below in detail, illustrate connections between remote superfamilies. An example of EFL structure and function is shown for each

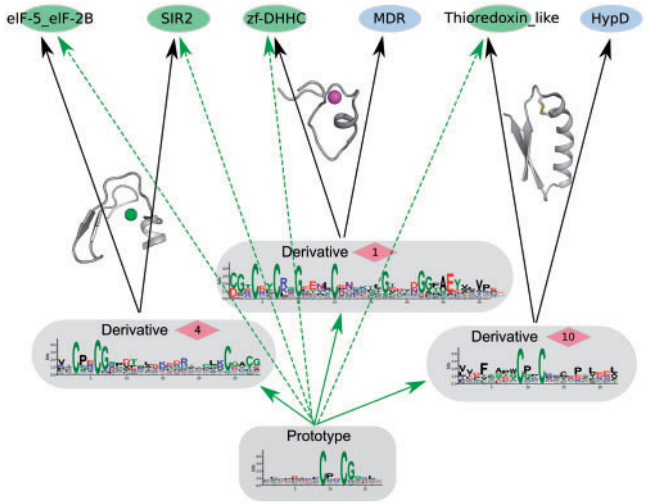

Fig. 3. Connections between remote biochemical functions established via relations between the prototype, its derivatives and their representative EFLs in CDD superfamilies. Black arrows represent the hits between derivatives (shown with sequence logo) and superfamilies (ovals) found by the profile-sequence search. Dashed green lines show hits of the very prototype (green ovals). Superfamilies detected only by the derivatives, but not by the prototype are shown as blue ovals. The elementary functions of derivatives are: 1, binding of zinc ion by two cysteine residues; 4, binding of zinc ion by four cysteines; 10, contains two redox active cysteine residues. Examples of CDD superfamilies include: eIF-5_eIF-2B, translation initiation factor IF2B/IF5 (PDB 1nee); SIR2, silent information regulator 2 (PDB 1ici); zf-DHHC, DHHC zinc finger domain; MDR, medium chain reductase/dehydrogenase and zinc-dependent alcohol dehydrogenase-like superfamily (PDB 2eih); Thioredoxin_like, a superfamily characterized by Thioredoxin fold and including disulfide oxidoreductases and protein disulfide isomerases (PDB 2trx); HypD, superfamily of enzymes required for maturation of hydrogenases (PDB 2z1d). Structural examples of the enzymes and the corresponding EFLs are shown in Supplementary Figures S3–S9.

derivative. Derivative 4 is responsible for zinc ion coordination provided by its pair of -CxxC- signatures, and the EFLs of this derivative work in different biochemical functions. For example, this zinc finger is important for correct recognition of the AUG codon in translation initiation factor eIF-5_eIF-2B, as well as in transcription factors, where it is responsible for recognition of sequence-specific double-stranded nucleic acids (Gutierrez *et al*., 2004). An example of another biochemical function is SIR2 superfamily, where zinc ion coordination provided by the EFL of derivative 4 contributes to creating and maintaining the substrate-binding site. The biochemical function of the SIR2 family is NAD-dependent protein deacetylation, which is involved in transcriptional silencing, X-chromosome silencing and suppression of ribosomal DNA recombination (Min *et al*., 2001). Derivative 1 corresponds to EFLs coordinating structural zinc ion. In case of the MDR superfamily, which includes alcohol dehydrogenases, this loop supports quaternary structure of metalloenzymes. A similar role of zinc ion coordination is performed by this loop in zf-DHHD domain. Although, the biochemical functions of proteins containing the zf-DHHD domain are not completely understood, the zinc-coordinating loop presumably assists protein–protein and protein–DNA interactions (Putilina *et al*., 1999). Derivative 10 describes EFLs with two redox-active cysteines, which can form a disulfide bond or be reduced to dithiol. The Thioredoxin superfamily represents a large group of proteins with diverse redox biochemical functions. All the enzymes in this superfamily are characterized by a small Thioredoxin fold with a redox active -CxxC- signature. The elementary function of the -CxxC- containing loop in connected HypD superfamily is similar to the one in thioredoxin reductases, but the biochemical function of HypD is much more complex (Watanabe *et al*., 2007). Enzymes in the HypD superfamily are involved in maturation of [NiFe] hydrogenases—a multi-step process, which includes insertion and cyanilation of the iron center. The cyanilation is provided by the EFL of Derivative 10, in which a -CxxC- disulfide bond should be preliminarily reduced.

3.2.4 Role of EFLs of -CxxC- prototype in different biochemical functions Figure 4 exemplifies how EFLs of the CxxC prototype (blue) contribute to different biochemical functions together with other EFLs. Cytochrome C553 (Figure 4A) is an example of one of the simplest membrane-bound electron-transfer proteins, which is presumably involved in respiratory metabolism, and consists of two EFLs only (Benini *et al*., 2000). Cysteines in the loop containing the -CxxCH- signature (blue) are covalently bound to heme, whereas histidine in this signature is one of the two axial ligands to the heme Fe atom. The second loop provides another axial ligand (methionine), forming the entire fold together with the EFL of Derivative 4.

Methyltransferase rRNA modification enzyme RlmN (Figure 4B) belongs to a diverse superfamily of Radical SAM enzymes, which includes more than 3000 members and is characterized by the presence of an iron–sulfur cluster and s-adenosylmethionine (SAM) (Frey *et al*., 2008). SAM is bound to one of the Fe atoms in the cluster. The [4Fe-4S] cluster, in turn, is ligated by three cysteine residues forming a characteristic -CxxxCxxC- signature (represented by Derivatives 6 and 11). Methyltransferase function also involves methylation of a cysteine residue (Boal *et al*., 2011), which is provided by another typically unstructured loop (the modeled loop is shown in red color in Figure 4B).

Figure 4C contains a remarkable example of the complex biochemical function of DNA recombination and repair. First, the -CxxC- containing loop (blue) performs a structural function of zinc ion coordination in the coiled-coil zinc hook (Hopfner *et al*., 2002), where it assists dimerization. Secondly, Rad50-ATPase domains together with Mre11 domains located at the coiled-coil termini perform the very DNA processing (recombination and repair) driven by ATP hydrolysis (Hopfner *et al*., 2001). Rad50-ATPase is an example of an ATP-binding cassette (ABC), consisting of EFLs with characteristic Walker-A (red) and Walker-B (green) signatures. It is important to note, however, that contrary to the typical ATP binding cassettes of membrane-bound ABC transporters, the Rad50-ATPase is a complex domain formed from two subunits located on different alpha-helices of the coiled coil linker. This observation emphasizes the role of EFLs as independent elementary units of protein function, showing how EFLs can be used in simple domains [formed by the continuous polypeptide chain, as in membrane-bound ABC transporters (Davidson *et al*., 2008)], as well as in the complex domains exemplified here (Rad50-ATPase).

Figure 4D contains an example of the function performed at the interface in the two-domain protein, involving -CxxC- containing (blue), FAD-binding (red) and NADPH-binding (green) EFLs.

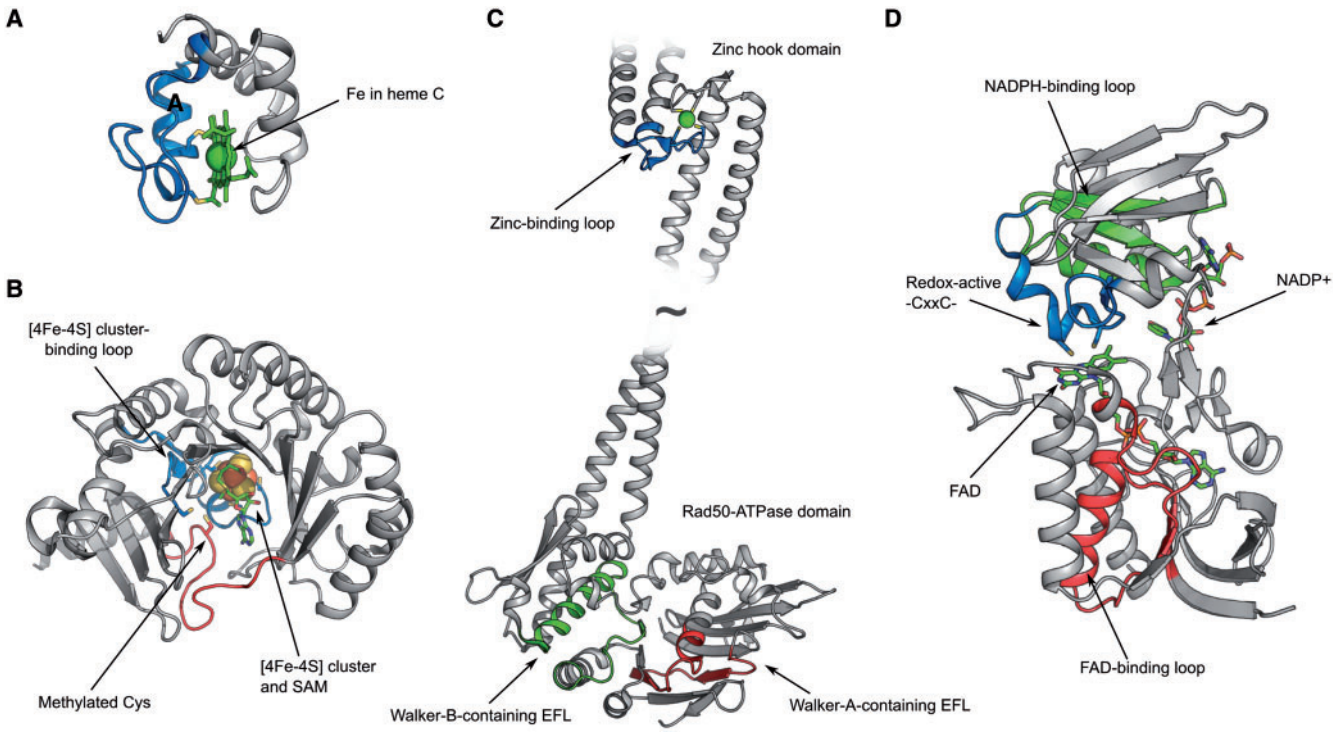

Fig. 4. Examples of enzymes with EFLs–descendants of the -CxxC- prototype (blue). (**A**) Cytochrome C553 (Derivative 4, Cytochrom_C3 in Figure 2, PDB 1c75); (**B**) SAM (S-adenosylmethionine)-dependent enzyme RlmN (Derivatives 6 and 11, Radical_SAM in Figure 2, PDB 3rfa); (**C**) Mre11 Binding Rad50-ATPase/Coiled-coil domain working DNA recombination and repair (Derivative 3, Rad50_zn_hook in Figure 2). Parts of the heterotetrameric DNA processing head and a double coiled-coil linker are shown (PDB 1l8d and 1ii8, see also Supplementary Figures S10, S11 for the whole structures). Each CxxC-containing EFL (blue) contributes two cysteines coordinating the zinc ion. In addition, Rad50-ATP contains EFLs characteristic for ABC domains with typical Walker A (red) and Walker B (green) signatures; (**D**) Thioredoxin reductase (Derivative 10, Thioredoxin_like superfamily in Figures 2 and 3, PDB 1tdf) consists of two domains and involves two redox active catalytic cysteines as part of the EFL with CxxC signature (blue), along with EFLs of NADP+ binding (green) and FAD binding (red).

This thioredoxin reductase disrupts disulfide bonds in its substrate thioredoxin (Waksman *et al*., 1994). In the conformation shown in Figure 4D, the disulfide of -CxxC- containing elementary functional loop stacks against the isoalloxazine ring system of FAD and is reduced at the expense of oxidized FAD. Next, rotation between domains takes place, and as a result, the nicotinamide ring of NADPH is brought in close contact to FAD, hydride is transferred between the two dinucleotides, and NADPH reduces FAD. At the same time, the redox active dithiol of the -CxxC- containing elementary functional loop is on the surface of the protein and is accessible to the substrate (thioredoxin).

4 CONCLUSIONS

We have developed a computational strategy for reconstructing the most ancient prototypes of elementary functions and their derivatives, allowing one to establish evolutionary relations between a prototype and its descendants in contemporary proteins. The key steps of this strategy are: (i) obtaining a non-redundant set of sequence profiles with signatures of elementary functions; (ii) finding a set of derivatives of a prototype; (iii) reconstructing a generalized prototype with the most generic elementary function; and (iv) analyzing and showing relations between the prototype, its derivatives and their descendants in modern folds and functions. This computational approach allows one to tackle the problem of hidden evolutionary relations between different folds and remote biochemical functions. The above strategy is exemplified here by the -CxxC- prototype of the EFLs with metal/metal-containing cofactor binding and redox activities, represented in more than 90 superfamilies. We go beyond the analysis of functional conservation in homologous proteins by dividing the enzymes into units of elementary functions. Figure 4 contains examples of proteins where EFLs of -CxxC- prototype occur. In these proteins elementary functions of the prototype work together with other elementary functions resulting in different biochemical transformations such as radical SAM-dependent methyl transfer, thioredoxin reductase, and ATP-dependent nuclease in a DNA double-strand break repair complex. Figure 4 illustrates how one can study predomain evolution of proteins applying the strategy developed in this work. Ultimately, we would like to obtain the set of ancient prototypes of elementary functions and to determine the set of first enzymes emerged as a result of their recombination in predomain evolution.

ACKNOWLEDGEMENT

We thank Simon Mitternacht for critical reading of the manuscript.

Funding: FUGE II Programme in Functional Genomics, Research Council of Norway.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Atkinson,H.J. and Babbitt,P.C. (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput. Biol.*, **5**, e1000541.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Benini,S. *et al.* (2000) Crystal structure of oxidized Bacillus pasteurii cytochrome c553 at 0.97-A resolution. *Biochemistry*, **39**, 13115–13126.

Berezovsky,I.N. (2003) Discrete structure of van der Waals domains in globular proteins. *Protein Eng.*, **16**, 161–167.

Berezovsky,I.N. *et al.* (2000) Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.*, **466**, 283–286.

Berezovsky,I.N. *et al.* (2003a) Protein sequences yield a proteomic code. *J. Biomol. Struct. Dyn.*, **21**, 317–325.

Berezovsky,I.N. *et al.* (2003b) Spelling protein structure. *J. Biomol. Struct. Dyn.*, **21**, 327–339.

Berezovsky,I.N. and Trifonov,E.N. (2001) Van der Waals locks: loop-n-lock structure of globular proteins. *J. Mol. Biol.*, **307**, 1419–1426.

Berndt,C. *et al.* (2008) Thioredoxins and glutaredoxins as facilitators of protein folding. *Biochim. Biophys. Acta*, **1783**, 641–650.

Boal,A.K. *et al.* (2011) Structural basis for methyl transfer by a radical SAM enzyme. *Science*, **332**, 1089–1092.

Cai,W. *et al.* (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.

Chivers,P.T. *et al.* (1996) The CXXC motif: imperatives for the formation of native disulfide bonds in the cell. *EMBO J.*, **15**, 2659–2667.

Chivers,P.T. *et al.* (1997) The CXXC motif: a rheostat in the active site. *Biochemistry*, **36**, 4061–4066.

Davidson,A.L. *et al.* (2008) Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiol. Mol. Biol. Rev.*, **72**, 317–364.

de Gennes,P.G. (1990) *Introduction to Polymer Dynamics*. Cambridge University Press pp.17–26.

Dupont,C.L. *et al.* (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc. Natl Acad. Sci. USA*, **107**, 10567–10572.

Fan,S.W. *et al.* (2009) Conformational changes in redox pairs of protein structures. *Protein Sci.*, **18**, 1745–1765.

Fomenko,D.E. and Gladyshev,V.N. (2003) Identity and functions of CxxC-derived motifs. *Biochemistry*, **42**, 11214–11225.

Frey,P.A. *et al.* (2008) The radical SAM superfamily. *Crit. Rev. Biochem. Mol. Biol.*, **43**, 63–88.

Golovin,A. and Henrick,K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.

Goncearenco,A. and Berezovsky,I.N. (2010) Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics*, **26**, i497–i503.

Gutierrez,P. *et al.* (2004) Structure of the archaeal translation initiation factor aIF2 beta from Methanobacterium thermoautotrophicum: implications for translation initiation. *Protein Sci.*, **13**, 659–667.

Harms,M.J. and Thornton,J.W. (2010) Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.*, **20**, 360–366.

Holliday,G.L. *et al.* (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.*, **390**, 560–577.

Hopfner,K.P. *et al.* (2002) The Rad50 zinc-hook is a structure joining Mre11 complexes in DNA recombination and repair. *Nature*, **418**, 562–566.

Hopfner,K.P. *et al.* (2001) Structural biochemistry and interaction architecture of the DNA double-strand break repair Mre11 nuclease and Rad50-ATPase. *Cell*, **105**, 473–485.

Iwasaki,T. (2010) Iron-sulfur world in aerobic and hyperthermoacidophilic archaea Sulfolobus. *Archaea*, **2010**, 842639.

Johnson,D.C. *et al.* (2005) Structure, function, and formation of biological iron-sulfur clusters. *Annu. Rev. Biochem.*, **74**, 247–281.

Koonin,E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, **1**, 127–136.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 142–143.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lipkus,A.H. (1999) A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.*, **26**, 263–265.

Marchler-Bauer,A. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.

Meyer,J. (2008) Iron-sulfur protein folds, iron-sulfur chemistry, and evolution. *J. Biol. Inorg. Chem.*, **13**, 157–170.

Min,J. *et al.* (2001) Crystal structure of a SIR2 homolog-NAD complex. *Cell*, **105**, 269–279.

Mirkin,B.G. *et al.* (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.

Murtagh,F. (1984) Complexities of hierarchic clustering algorithms: the state of the art. *Comput., Stat. Q.*, **1**, 101–113.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Putilina,T. *et al.* (1999) The DHHC domain: a new highly conserved cysteine-rich motif. *Mol. Cell. Biochem.*, **195**, 219–226.

Quan,S. *et al.* (2007) The CXXC motif is more than a redox rheostat. *J. Biol. Chem.*, **282**, 28823–28833.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shoemaker,B.A. *et al.* (2010) Inferred biomolecular interaction server–a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.*, **38**, D518–D524.

Sigrist,C.J. *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.

Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**, 1409–1438.

Trifonov,E.N. *et al.* (2001) Distinct stages of protein evolution as suggested by protein sequence analysis. *J. Mol. Evol.*, **53**, 394–401.

Waksman,G. *et al.* (1994) Crystal structure of Escherichia coli thioredoxin reductase refined at 2 A resolution. Implications for a large conformational change during catalysis. *J. Mol. Biol.*, **236**, 800–816.

Watanabe,S. *et al.* (2007) Crystal structures of [NiFe] hydrogenase maturation proteins HypC, HypD, and HypE: insights into cyanation reaction by thiol redox signaling. *Mol. Cell*, **27**, 29–40.

Wilcox,D.E. *et al.* (2001) Oxidation of zinc-binding cysteine residues in transcription factor proteins. *Antioxid Redox Signal*, **3**, 549–564.