Research

Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences

Lakshminarayan M Iyer*, L Aravind*, Peer Bork†, Kay Hofmann‡, Arcady R Mushegian§, Igor B Zhulin¶ and Eugene V Koonin*

Addresses: *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. †EMBL, Biocomputing, Meyerhofstrasse 1, 69117 Heidelberg, Germany. ‡MEMOREC Stoffel GmbH, Köln D-50829, Germany. §Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64410, USA. ¶School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.

*Correspondence: Eugene V Koonin. E-mail: koonin@ncbi.nlm.nih.gov

Published: 13 November 2001

Genome ***Biology*** 2001, **2(12)**:research0051.1–0051.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/12/research/0051

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 3 July 2001
Revised: 7 September 2001
Accepted: 4 October 2001

Abstract

Background: Computational predictions are critical for directing the experimental study of protein functions. Therefore it is paradoxical when an apparently erroneous computational prediction seems to be supported by experiment.

Results: We analyzed six cases where application of novel or conventional computational methods for protein sequence and structure analysis led to non-trivial predictions that were subsequently supported by direct experiments. We show that, on all six occasions, the original prediction was unjustified, and in at least three cases, an alternative, well-supported computational prediction, incompatible with the original one, could be derived. The most unusual cases involved the identification of an archaeal cysteinyl-tRNA synthetase, a dihydropteroate synthase and a thymidylate synthase, for which experimental verifications of apparently erroneous computational predictions were reported. Using sequence-profile analysis, multiple alignment and secondary-structure prediction, we have identified the unique archaeal 'cysteinyl-tRNA synthetase' as a homolog of extracellular polygalactosaminidases, and the 'dihydropteroate synthase' as a member of the β-lactamase-like superfamily of metal-dependent hydrolases.

Conclusions: In each of the analyzed cases, the original computational predictions could be refuted and, in some instances, alternative strongly supported predictions were obtained. The nature of the experimental evidence that appears to support these predictions remains an open question. Some of these experiments might signify discovery of extremely unusual forms of the respective enzymes, whereas the results of others could be due to artifacts.

Background

The availability of a large number of protein sequences, including complete protein sets encoded in diverse genomes, and the rapidly growing database of protein structures have already greatly impacted on our understanding of the evolution of protein structure and function [1,2]. This process has been aided by the development of powerful algorithms and sensitive computational tools for detecting sequence and

structural similarities between proteins. In particular, methods that extract information from multiple alignments to construct various types of sequence profiles and use the resulting sequence profiles for iterative database searching, such as PSI-BLAST and Hidden-Markov-Model (HMM)-based approaches, have substantially improved the detection of subtle similarities between proteins that previously were amenable only to direct structural comparison [3,4]. The sensitivity and accuracy of these methods have been extensively tested and statistical approaches for validating the observed similarities are available [5-11].

Despite these achievements, detection and interpretation of relationships between homologous proteins that have limited sequence similarity remains a major challenge. Such studies typically require a case-by-case approach that is guided by a detailed understanding of protein sequence-structure patterns and is rooted in the biology of the proteins analyzed. Prediction of structures and function(s) of uncharacterized proteins is one of the principal outcomes of these analyses, and experimental verification of such predictions tends to increase confidence in the validity of sequence-structure comparative approaches. The negative feedback from experiments that failed to confirm a computational prediction is potentially even more important, because it could result in revision and refinement of the computational methods.

When examining cases of reported prediction followed by experimental validation, however, we encountered several paradoxical situations. In each of these, a prediction that has been reportedly confirmed by experiment was incompatible with results obtained with several standard computational procedures. More importantly, alternative predictions, supported by statistically significant sequence and/or structural similarity, were made in some of these cases. Here we present several such mysteries, describe the refutation of the original predictions and the new predictions, wherever feasible, and discuss the discrepancy between the computational and experimental results. The choice of the cases was not systematic; rather, those chosen were notable because they relied on novel computational techniques, exploited particularly subtle sequence or structural motifs, and dealt with crucial biological problems.

Results

MJ1477: a predicted archaeal cysteinyl-tRNA synthetase

Aminoacyl-tRNA synthetases (aaRSs) specific for 17 of the 20 amino acids are universally present in cellular life forms. The three exceptions are GlnRS, AsnRS and CysRS. GlnRS and AsnRS are missing in many bacteria and archaea because glutamine and asparagine are incorporated into proteins through transamidation of glutamate and aspartate, respectively. CysRS is missing in two archaeal methanogens whose genomes have been sequenced - *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii* [12]. No alternative mechanism for cysteine incorporation into proteins is known; hence the absence of CysRS in these organisms was an enigma.

Two solutions to this puzzle, both unusual, have recently been proposed and experimentally validated. One involves non-orthologous gene displacement, a situation in which the same essential function is carried out by distantly related or even unrelated proteins in different organisms [13,14]. It has been shown that *M. jannaschii* ProRS, a class II synthetase that is unrelated to the class I CysRS, substituted for the missing CysRS activity [15-17]. The other solution involved a new candidate for the role of CysRS, the MJ1477 protein from *M. jannaschii*. This protein and its orthologs (direct evolutionary counterparts related by vertical descent from a common ancestor) from the bacteria *Thermotoga maritima* and *Deinococcus radiodurans* were identified as 'distant orthologs' of the *Bacillus subtilis* CysRS by using a computational method specifically designed to detect distantly related orthologs [18]. The method is based on application of discriminant analysis to alignment scores, in order to separate the scores for pairs of functionally identical proteins from different genomes from the scores for proteins with different functions. This prediction was then validated experimentally by showing that MJ1477 had CysRS activity *in vitro* and that an ortholog of MJ1477 from *D. radiodurans*, DR0705, complemented a CysRS deficient, temperature-sensitive, lethal *E. coli* mutant strain [18]. An important corollary of these surprising findings is a rapid divergence of the MJ1477 family from CysRS, such that all the catalytic and otherwise functionally important residues characteristic of this enzyme, and also present in other class I aaRSs, have changed. Furthermore, MJ1477 and its orthologs do not have the accessory domains found in all known CysRS, namely the DALR domain (named after a distinct amino-acid signature), which is shared by aaRSs of several specificities, and another domain specific to CysRS [19].

We examined the protein sequences of MJ1477 and its homologs using more traditional computational techniques. Almost all these proteins contain amino-terminal signal peptides readily identifiable by using the SignalP program [20], but do not contain any predicted transmembrane segments, and, accordingly, are predicted to be secreted from the cells (Figure 1). Furthermore, iterative database searches using the PSI-BLAST program [9] showed statistically significant sequence similarity between these proteins and an experimentally characterized endo α-1,4-polygalactosaminidase from *Pseudomonas* species [21]. For example, in a search initiated with the sequence of MJ1477 and a profile inclusion cut-off of 0.01, the polygalactosaminidase sequence was retrieved from the database in the second iteration, followed by other bacterial proteins predicted to possess the same activity. This protein family has several conserved motifs, including a characteristic Dxhp signature (h, hydrophobic

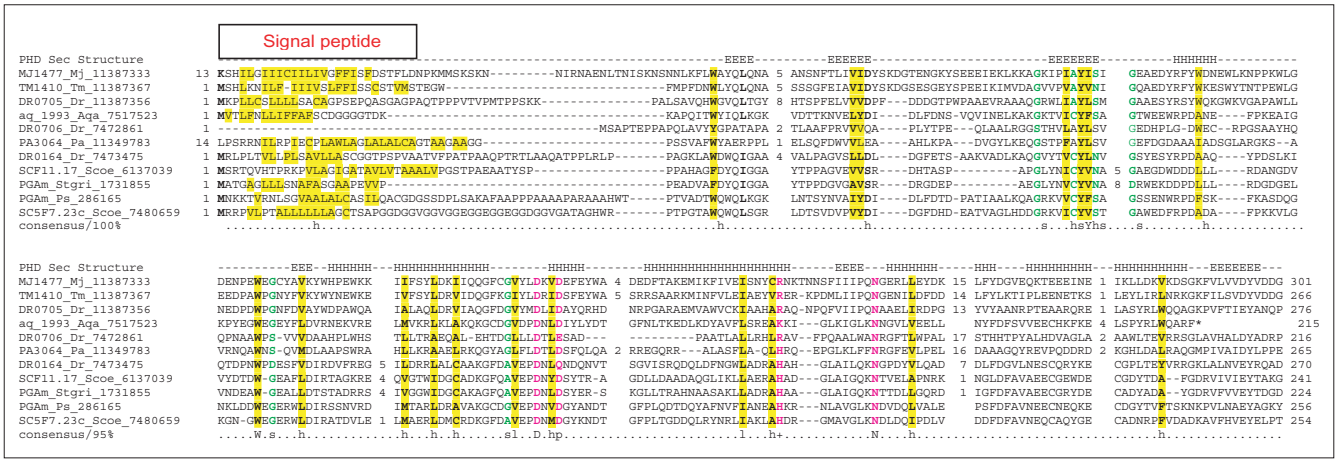

Figure 1
Multiple alignment of the polygalactosaminidase family that includes MJ1477, the alleged archaeal CysRS. Proteins are denoted by their gene name, followed by their species abbreviations and GenBank identifier (GI) numbers. The coloring reflects the 100% consensus. The consensus abbreviations and coloring scheme used in this and subsequent figures are as follows. Hydrophobic residues (h; LIYFMWACV) and aliphatic (l; LIAV) residues are shaded yellow. Colored magenta are alcohol (o; ST), charged (c; KERDH), basic (+; KRH), acidic (- ; DE), and polar (p; STEDRKHNQ) residues. Small (s; SAGDNPVT) residues are colored green and big (b; LIFMWYERKQ) residues are shaded gray. The hydrophobic residues of the signal peptide are highlighted in yellow. In the Secondary Structure line, H indicates a helix and E indicates extended conformation (b strand). Aqa, *Aquifex aeolicus*; Dr, *Deinococcus radiodurans*; Mj, *Methanococcus jannaschii*; Pa, *Pseudomonas aeruginosa*; Ps, *Pseudomonas* species; Scoe, *Streptomyces coelicolor*; Strgi, *Streptomyces griseus*; Tm, *Thermotoga maritima.*

residue; p, polar residue), in which the conserved aspartate is likely to directly participate in catalysis (Figure 1). The hybrid-fold-recognition method, which combines sequence-profile analysis with alignment-based secondary-structure prediction [22] and the 3D-PSSM method [23] both suggested a likely α-amylase-like triosephosphate isomerase (TIM) barrel structure for this protein family. Thus, although the identification of MJ1477 as a secreted polygalactosaminidase or a related polysaccharide hydrolase with a different specificity awaits experimental verification, it shows all the signs of a correct computational prediction: statistically significant similarity between the analyzed protein and an experimentally characterized enzyme; conservation of distinct motifs implicated in catalysis; potential presence of a structural fold compatible with the experimentally demonstrated enzymatic activity; and confident prediction of the extracellular localization that is, again, compatible with a polysaccharide hydrolase activity involved in environmental carbohydrate utilization or capsular metabolism. None of this evidence is offered by the analysis that led to the CysRS prediction for MJ1477.

Therefore we are forced to conclude that MJ1477 and its homologs are not related to CysRS and there is nothing in the computational analysis of these proteins that would point to an aaRS activity. In contrast, we predict these proteins to be extracellular polygalactosaminidases or similar polysaccharide hydrolases. The polysaccharide hydrolase and aaRS functions seem to be essentially incompatible. First, a secreted enzyme is unlikely to function as an aaRS whose site of action is, by definition, intracellular. Second, even if an entirely new class of aaRSs is postulated, the reaction catalyzed by this new aaRS does not resemble polysaccharide hydrolysis or its reversal. Aminoacyl-tRNA synthetases catalyze a succession of reactions, which involve: hydrolysis of the α-β phosphate bond in ATP; condensation of AMP with the cognate amino acid, resulting in the formation of an aminoacyl-adenylate; displacement of the AMP moiety of the aminoacyl-adenylate with the cognate tRNA, producing aminoacyl-tRNA. Even if the two condensation reactions, in very general terms, could be considered a reversal of the polysaccharide hydrolysis reaction, there is no indication that polysaccharide hydrolases could bind and hydrolyze ATP, and the multiple alignment of the MJ1477 family did not include any conserved signatures typical of potential phosphate-binding loops (Figure 1). Neither does this family contain any recognizable RNA-binding domains. Finally, *M. thermoautotrophicum* does not encode any homologs of MJ1477, ruling out the possibility that this family encompasses CysRS of both archaeal methanogens. Taken together, these observations appear to effectively refute the prediction of a CysRS activity, thus pitting computational results against experimental data.

MJ0301: a predicted dihydropteroate synthase

Dihydropteroate synthase (DHPS) catalyzes the condensation of *p*-aminobenzoic acid with 7,8-dihydro-6-hydroxymethylpterin pyrophosphate to give 7,8-dihydropteroate, an intermediate in folate metabolism. The protein from *Staphylococcus*, a Gram-positive bacterium, has been crystallized

and shown to adopt a TIM-barrel structure [24]. Although it has been indicated that no DHPS could be detected in archaeal genomes [25], orthologs of bacterial DHPS are readily identifiable in all archaea; this enzyme is missing only in animals and in several intracellular bacterial pathogens, such as *Rickettsia prowazekii*, spirochetes and mycoplasmas (COG0294 in the database of Clusters of Orthologous Groups of proteins (COGs)) [26]. Most archaea have a distinct version of DHPS that shows relatively low sequence similarity to the bacterial orthologs and contains an additional uncharacterized carboxy-terminal domain. This previously undetected domain is also present in some other enzymes of pterin biosynthesis, such as tetrahydromethanopterin-*S*-methyltransferase from *Streptomyces* (L.M.I., L.A. and E.V.K., unpublished observation). Some archaeal species, including *Thermoplasma* and *Halobacterium*, have the bacterial-type DHPS, which was probably acquired by horizontal gene transfer and displaced the original archaeal version. Despite the relatively low sequence similarity to bacterial DHPS, all archaeal orthologs have the conserved catalytic residues identified in DHPS (Figure 2) and are confidently predicted, by the hybrid-fold-recognition method, to assume the same fold as DHPS from *Pneumocystis carinii* and *Staphylococcus aureus* whose crystal structures have been determined.

An analysis using ORF, a program developed to recognize folds by comparing predicted secondary structures of proteins ([27]; we are unaware of a published detailed description of this method), identified MJ0301 as a homolog of DHPS, although, given the low sequence similarity, a convergent origin of the relationship between MJ0301 and DHPS was deemed likely (there seems to be a terminological confusion involved here, but we are quoting the results of the original computational analysis of this protein as they have been presented). It was acknowledged that MJ0107 (a member of COG0294) could be identified as a possible homolog of DHPS by sequence-based methods, and this protein was assayed for dihydropteroate synthase activity, but none was detected [25]. In contrast, DHPS activity (albeit relatively low) was shown *in vitro* for the partially purified MJ0301 protein [25]. However, MJ0301 has been shown to belong to the metallo-β-lactamase superfamily of enzymes and, in the evolutionary classification of metallo-β-lactamases, belongs to an archaea-specific family (Figure 2; COG1237) [28]. Metallo-β-lactamases encompass a wide range of metal-dependent hydrolytic and oxidoreductase activities with a variety of substrates and are particularly abundant in archaea where some of them are involved in RNA processing [28]. None of these enzymes catalyzes a reaction resembling the condensation reaction catalyzed by DHPS. The characteristic motifs of metallo-β-lactamases, which mostly include metal-binding histidines, are highly conserved in MJ0301 and its orthologs (Figure 3). In contrast, most of the MJ0301 residues described as equivalent to the functionally important residues of *Escherichia coli* dihydropteroate synthase are not conserved, even among the archaeal orthologs of this protein. Finally, the β-lactamase fold consists of two subdomains of the β4-α-β-α topology whose β sheets are sandwiched against each other; in structural terms, these domains are completely different from the TIM-barrel, with which the ORF program matched the MJ0301 structural prediction. Taken together, these observations are sufficient to reject the proposed relationship between MJ0301 and dihydropteroate synthases.

MJ0757: a predicted thymidylate synthase

Thymidylate synthase is a central enzyme of pyrimidine metabolism that catalyzes the formation of deoxythymidine monophosphate (dTMP) from deoxyuridine monophosphate

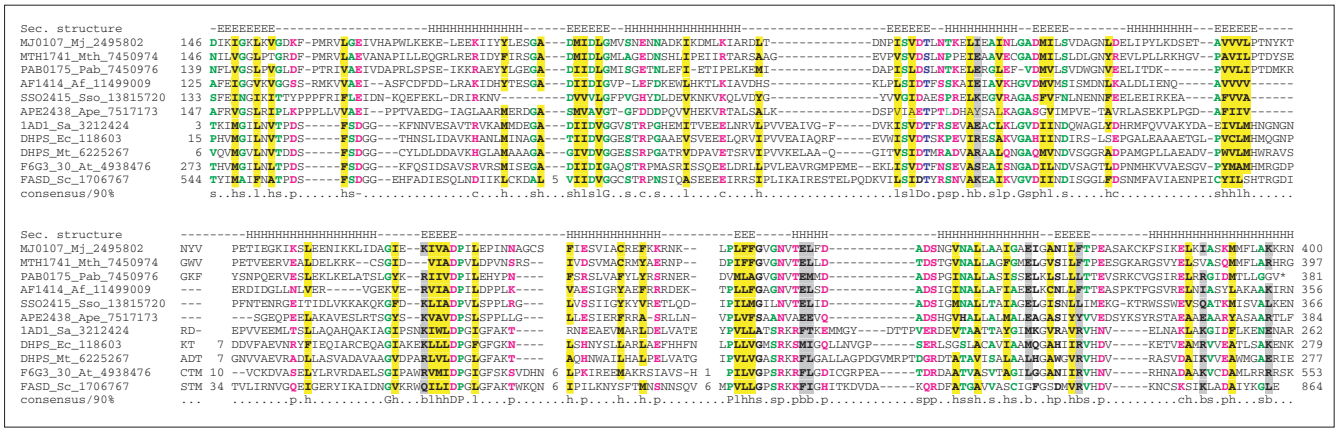

Figure 2
Multiple alignment of predicted archaeal dihydropteroate synthases. The scheme for displaying multiple alignments is as described in the legend to Figure 1. The consensus secondary structure was derived from the crystal structures of the *Staphylococcus aureus*, *Mycobacterium tuberculosis* and *Escherichia coli* DHPS (Protein Data Bank ID: 1AD1, EYE, 1AJ0). Residues are colored at 90% consensus. Af, *Archaeoglobus fulgidus*; Ape, *Aeropyrum pernix*; At, *Arabidopsis thaliana*; Ec, *Escherichia coli*; Mj, *Methanococcus jannaschii*; Mt, *Mycobacterium tuberculosis*; Mth, *Methanobacterium thermoautotrophicum*; Sa, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Pab, *Pyrococcus abyssi*.

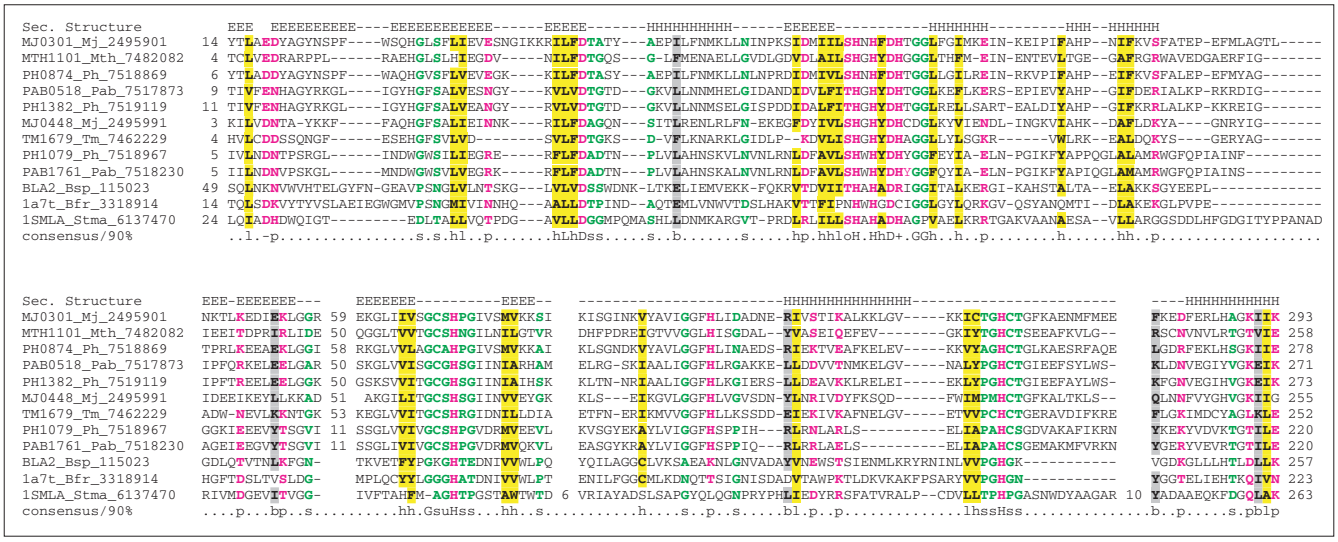

Figure 3
Multiple alignment of the archaea-specific family of predicted metallo-β-lactamase superfamily hydrolases that includes the alleged archaeal dihydropteroate synthase, MJ0301. The scheme for displaying multiple alignments is as described in the legend to Figure 1. A consensus secondary structure was derived from the crystal structure metallo-β-lactamases from *Stenotrophomonas maltophilia* (1SML) and *Bacteroides fragilis* (1A7T). Residues are colored at 90% consensus. Bfr, *Bacteroides fragilis*; Bsp, *Bacillus* species 170; Mj, *M. jannaschii*; Mth, *M. thermoautotrophicum*; Pab, *P. abyssi*; Ph, *P. horikoshii*; Stma, *S. maltophilia*; Tm, *Thermotoga maritima.*

(dUMP) by transfer of a methyl group to its pyrimidine ring. This reaction is catalyzed by at least two unrelated enzymes. The canonical thymidylate synthase (TS), such as the *E. coli* ThyA, is a protein with a distinct α/β-fold that transfers a methyl group to dUMP from 5,10-methylenetetrahydrofolate [29]. This classic TS is readily identifiable in many (but not all) bacteria, eukaryotes and three archaeal species, *Archaeoglobus fulgidus*, *M. jannaschii*, and *M. thermoautotrophicum* (COG0207). The archaeal members of the TS family share with their bacterial orthologs all the conserved residues involved in catalysis (Figure 4).

An alternative TS or its subunit is predicted to be encoded by a gene from *Dictyostelium* that rescues a slime mold mutant auxotrophic for thymidylate [30]. This protein is not homologous to the canonical TS, but its orthologs in bacteria and archaea show an almost perfect complementary phyletic distribution (COG1351).

In a screen for the TS in *M. jannaschii*, the ORF method picked the MJ0757 protein as the most likely homolog of the canonical TS family [27]. In the validation experiment, MJ0757 overexpressed in *E. coli* was shown to possess TS activity [25]. Sequence searches show that MJ0757 belongs to a small family of euryarchaea-specific proteins of uncharacterized function (COG1810). Of the 17 residues reported to be conserved between MJ0757 and the TS family, only seven were conserved thoughout the MJ0757 family (Figure 5). Moreover, a comparison of the secondary structure elements derived from the reported three-dimensional model of MJ0757 [27] and those derived from a prediction generated using a multiple alignment query with the structure-prediction program PHD (such predictions typically exceed 70% accuracy), showed an overlap of just two of the 16 or so secondary structural elements (Figure 5). Conversely, several sequence motifs that are characteristic of the MJ0757 family did not overlap with the conserved regions in the MJ0757-TS alignment (Figure 5). Furthermore, some, but not all, members of the MJ0757 family contain an amino-terminal insertion of a small, metal-chelating module (Figure 5), which was used to improve the alignment with the *E. coli* TS [25], although this region was variable even within the MJ0757 family itself. On the basis of these observations, a relationship between MJ0757 and the canonical TS has to be rejected. The actual fold and function of MJ0757 and its homologs cannot be predicted at present. However, these proteins have several features that suggest that they might be metal-dependent enzymes potentially involved in redox reactions. These suggestive features include the fusion with a ferredoxin domain seen in the *M. thermoautotrophicum* member MTH601, the insertion of the metal-binding module in certain members, including MJ0757 (see above), and the presence of three cysteines that are conserved throughout this family.

Cmpp16: a plant 'paralog' of plant viral movement proteins

Viral movement proteins (MPs) are encoded by diverse, unrelated families of plant viruses, such as positive-strand

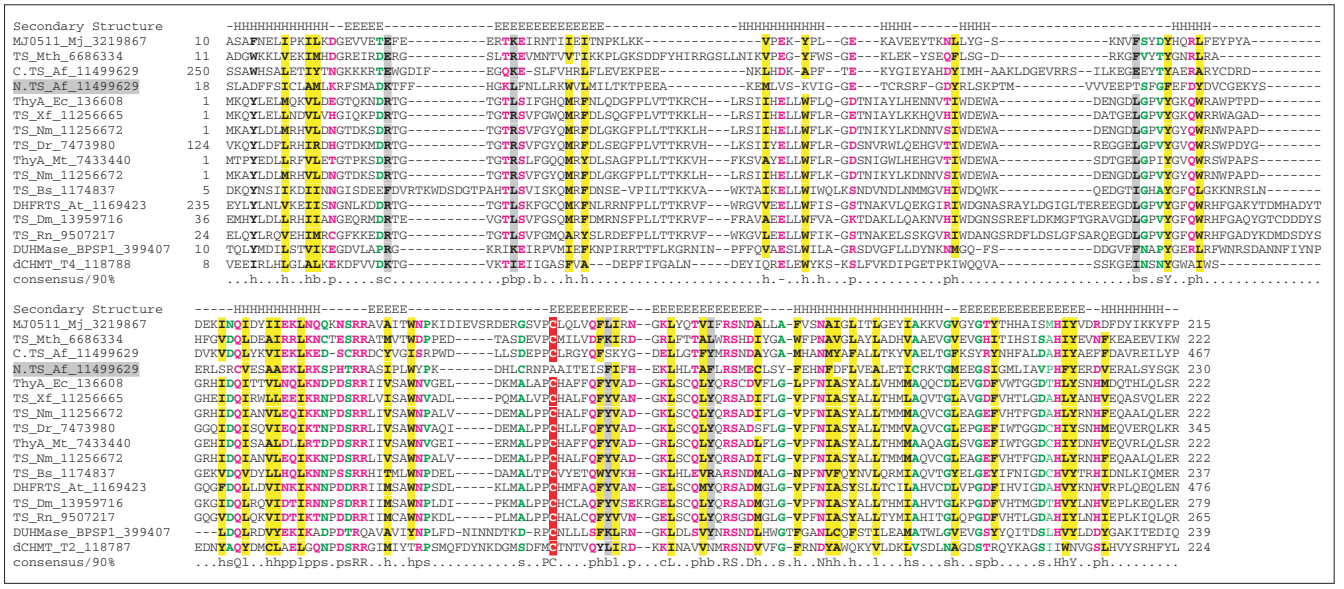

Figure 4
Multiple alignment of predicted archaeal thymidylate synthases (TS). The scheme for displaying multiple alignments is as described in the legend to Figure 1. Residues are colored at 90% consensus. A consensus secondary structure was derived using known TS structures from *R. norvegicus*, *E. coli* and bacteriophage T4 deoxycytidylate hydroxymethyltransferase (1B5D). The *Archaeoglobus fulgidus* TS has a duplication of the TS domain and the amino-terminal domain (N.TS_Af; shaded gray) is predicted to be inactive. Af, *Archaeoglobus fulgidus*; At, *Arabidopsis thaliana*; BPSP1; bacteriophage SP1; Bs, *B. subtilis*; Dm, *Drosophila melanogaster*; Dr, *D. radiodurans*; Ec, *E. coli*; Mj, *M. jannaschii*; Mt, *M. tuberculosis*; Mth, *M. thermoautotrophicum*; Nm, *Neisseria meningitidis*; Rn, *R. norvegicus*; T2, bacteriophage T2; Xf, *Xylella fastidiosa*.

RNA, negative-strand RNA, single-stranded DNA and double-stranded DNA viruses, and are essential for cell-to-cell movement of all these viruses [31,32]. To isolate potential host homologs of the red clover necrotic mosaic virus (RCNMV) MP, antibodies to this protein were used to screen phloem extracts of *Cucurbita maxima,* resulting in the detection of a protein designated Cmpp16. This protein was identified as a 'paralog' (generally, this term refers to homologous genes related by duplication within the same genome) of the viral MPs on the basis of sequence similarity detected using the Megalign program [33]. Subsequently, Cmpp16 was shown to bind RNA, which is a common property of viral MPs, and to induce an increase of the size-exclusion limit of plasmodesmata, also a mechanism associated with the MPs [33].

However, computational analysis of the Cmpp16 sequence reveals a picture that is incompatible with a homologous relationship with MPs. Cmpp16 consists mostly of a C2 domain that is readily detected by PSI-BLAST or by profile-searching engines such as the CD-search [34]. The Cmpp16 sequence contains all critical residues of the C2 domain (Figure 6). C2 domains bind a variety of substrates, such as Ca^{2+}, phospholipids, inositol polyphosphates and other proteins, but apparently not RNA [35]. There is no detectable similarity between C2 domains and the MPs, and conserved motifs in the published alignment of Cmpp16 and the RCNMV MP do not correspond to those in C2 domains; moreover, many of the residues described as conserved in Cmpp16 and MP are not conserved within the viral movement protein family itself. Thus, we conclude that viral MPs and Cmpp16, a C2-domain protein, are not homologs. Subsequently, a similar methodology has been employed to detect a relationship between Cmpp36 (a cytochrome B5 reductase), Cmpp16 and the RCNMV movement protein [36]. As in the above case of Cmpp16, this relationship of a cytochrome B5 reductase with the viral movement proteins appears to be spurious (data not shown).

Human activating transcription factor-2 (ATF-2): a predicted histone acetyltransferase

Histone acetyltransferases (HAT) are key regulators of eukaryotic transcription. GCN5-like HATs, which modulate chromatin-associated transcription, belong to a vast superfamily of amino-group acetyl- and myristoyl-transferases with extremely diverse functions [37]. ATF-2 is a basic leucine zipper (b-ZIP) family transcription factor that binds to cyclic AMP-response elements (CRE) and activates transcription [38]. Vertebrate ATF-2 also has an amino-terminal zinc finger, which is involved in transcription activation [39]. Non-vertebrate orthologs of ATF-2, in *Drosophila, Caenorhabditis elegans* and yeasts, lack the zinc finger. In experiments designed to isolate ATF-2-associated HAT, ATF-2 alone was shown to be sufficient for the acetyltransferase activity.

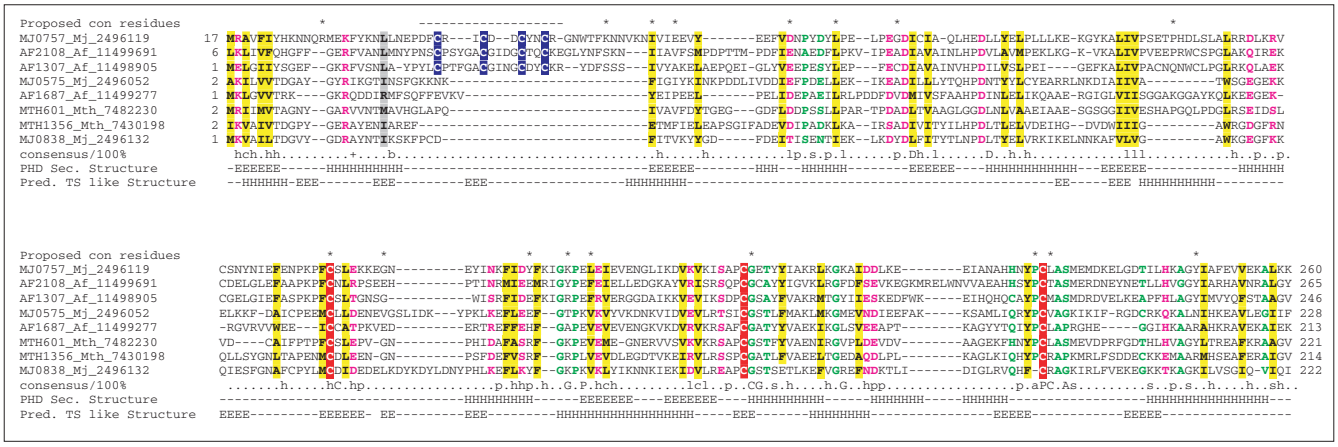

Figure 5
Multiple alignment of the uncharacterized archaeal protein family that includes the alleged archaeal thymidylate synthase, MJ0757. The scheme for displaying multiple alignments is as described in the legend to Figure 1. Residues are colored at 100% consensus. In addition, metal-chelating residues in an inserted module shared by orthologs of MJ0757 are shaded blue. The asterisks denote residues in MJ0757 that were predicted to be conserved between MJ0757 and TS. Also shown are predicted secondary structures for the MJ0757 family that were obtained by using the PHD program, and the TS-like secondary structure predicted for MJ0757 in [25]. Af, *A. fulgidus*; Mj, *M. jannaschii*; Mth, *M. thermoautotrophicum*.

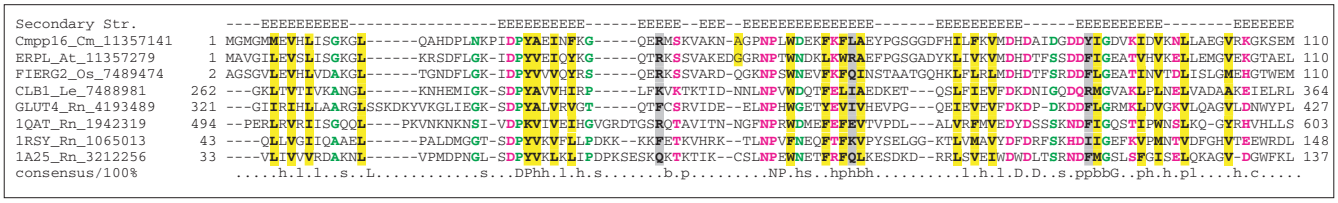

Figure 6
Multiple alignment of a selection of C2 domains including the alleged 'paralog' of plant virus movement proteins, Cmpp16. The scheme for displaying multiple alignments is as described in the legend to Figure 1. Residues are colored at 100% consensus. A consensus secondary structure was derived from known structures of the C2 domains in phospholipase C-δ1 (1QAT), synaptotagmin (1RSY), and protein kinase C (1A25). At: *A. thaliana*, Cm: *Cucurbita maxima*, Le: *Lycopersicon esculentum*, Os: *Oryza sativa*, Rn: *R. norvegicus*.

Examining the region of ATF-2 that showed HAT activity, the authors found some sequence similarity and at least one motif resembling the acetyltransferase superfamily and concluded that ATF-2 contained a GCN5-like acetyltransferase domain [40]. Subsequent site-directed mutagenesis supported the importance of the reported acetyltransferase motifs for the HAT activity of ATF-2.

However, profile-based sequence searches and attempts at fold recognition failed to detect any relationship between ATF-2 and the acetyltransferase superfamily. The region designated as having HAT activity and containing the acetyltransferase domain shows poor conservation between orthologs and closely related paralogs of the ATF-2 family, especially in the sequence identified as the most prominent A motif of the acetyltransferase family (Figure 7). Furthermore, complexity analysis using the SEG program, with the parameters adjusted for decomposition of a protein into globular and non-globular regions [41], predicted that the entire region of the ATF-2 protein between the amino-terminal zinc finger and the carboxy-terminal helical b-ZIP was unstructured. This is consistent with the structural prediction derived using the PHD program that indicated no regular secondary structure in this region. Thus, the relationship between ATF-2 and the GCN5-like acetyltransferase superfamily seems to be invalid, leaving the structural basis for the reported acetyltransferase activity of ATF-2 an open issue.

Predicted PAS domain in the phytochrome-interacting transcription factor PIF3

PAS domains are sensory modules in various signal transduction proteins from all major lineages of cellular life [42]. PAS domains are typically implicated in sensing oxygen, redox potential, light and small ligands [43]. In addition,

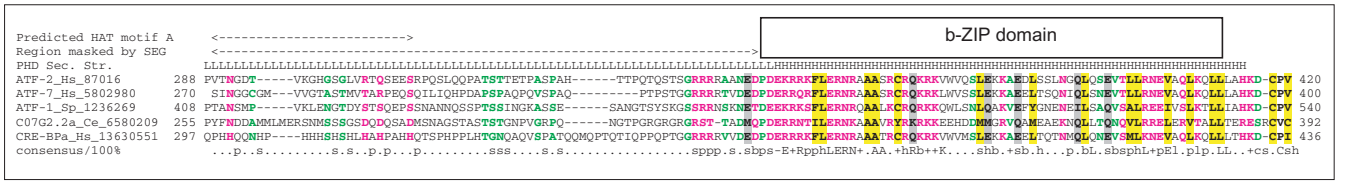

Figure 7
Multiple alignment of the region of the ATF-2 transcription factor and its homologs identified as a GCN5-like acetyltransferase domain. The scheme for displaying multiple alignments is as described in the legend to Figure 1. Residues are colored at 100% consensus. Ce: *Caenorhabditis elegans,* Hs: *Homo sapiens*, Sp: *Schizosaccharomyces pombe*.

PAS domains are sites for protein-protein interactions and are responsible for the formation of homo- and heterodimers in several signal transduction pathways that involve transcriptional activation. A PAS domain has been reported in the transcription factor PIF3 from *Arabidopsis*, which interacts with a phytochrome photoreceptor and transduces light signals to photoresponsive plant genes [44]. It has been hypothesized that the purported PAS domain of PIF3 directly interacts with the PAS domains of the phytochrome [44]. This hypothesis was later tested experimentally and evidence was presented that the PAS domain of PIF3 indeed was a major contributor to the interaction between the two proteins [45].

PIF3 belongs to a plant-specific family of basic helix-loop-helix (bHLH)-domain- containing proteins that, in addition to the bHLH domain, have an uncharacterized conserved domain at the amino terminus present in single or duplicate copies (L.M.I., I.Z., L.A. and E.V.K., unpublished observations). The PIF3 family currently consists of about eight paralogous proteins in *Arabidopsis* and an ortholog from rice. The region predicted to be a PAS domain is poorly conserved in the rice ortholog of PIF3 and the paralogs from *Arabidopsis*. An alignment with the rice ortholog indicated that the proposed PAS domain was a rapidly diverging, compositionally biased sequence (Figure 8). Complexity analysis using the SEG program showed that the reported PAS domain mapped to a region that was predicted to be entirely nonglobular. All attempts to objectively detect a PAS domain in PIF3 using sensitive profile methods based on PSI-BLAST-derived scoring matrices or Hidden Markov Models (HMM) failed. Additionally, secondary-structure prediction for the proposed PAS region using PHD indicated that this region is largely unstructured. These observations appear to be sufficient to reject the presence of a PAS domain in PIF3 although the region thought to be a PAS domain could indeed be involved in the interaction with phytochrome.

Discussion and conclusions

In the six cases described above, we provide evidence for rejecting the homologous relationships and functional predictions inferred for the proteins in question by using computational methods. The number of examples in this category could be increased, and some have already been considered in the literature, for example the spurious discovery of a 'functional PDZ domain' in the molecular chaperone ClpA ([46], see refutation in [47]) or the finding of an ATPase domain and death effector domains in the apoptosis-associated protein FLASH ([48], see refutation in [49]). The common and most striking aspect of all these cases is that the predictions based on apparently erroneous computational analysis were supported by experiments. What are the solutions to this clash between computational and experimental evidence?

We envisage three main possibilities. The first, experiment-centered view would hold that experimental evidence always has the upper hand and that, even if the alternative computational solutions that we describe here seem more plausible than the original predictions, the latter are correct insofar as they are supported by experiment. Epistemologically, this argument is not sound because hypotheses (computational predictions in this case) cannot be proved by the success of the experiments they prompt. They can only be falsified by experiments producing results incompatible with the predictions [50]. Simply put, the experiments could have worked for a wrong reason. For example, this seems particularly likely in the case of the site-directed mutagenesis of the transcription factor ATF-2 discussed above. The mutagenized residues probably are indeed important for the function of this protein, but not because they are part of a GCN5-like acetyltransferase domain, which this protein does not contain. Similar logic applies to the case of the predicted, but apparently nonexistent, PAS domain in the transcription factor PIF3. More important, however, computational predictions are falsifiable within the realm of computational analysis itself. Falsification is offered by alternative, unequivocally supported predictions that are incompatible with the original ones. In four of the six cases described (CysRS, DHPS, TS and MP), such evidence was obtained by computational methods.

The second possibility is that, although the computational predictions described here are correct, whereas the original ones are wrong, the experimental evidence is also solid. In

Figure 8
A comparison of the multiple alignments of PIF3, its rice ortholog, and PAS domain proteins. The scheme for displaying multiple alignments is as described in the legend to Figure 1. Residues are colored at 90% consensus. A consensus secondary structure was derived from those available for FixL (1EW0) and photoactive yellow protein (3PYP). Aa, *A. aeolicus*; Af, *A. fulgidus*; At, *A. thaliana*; Av, *Azotobacter vinelandii*; Bs, *B. subtilis*; Dm, *D. melanogaster*; Ec, *E. coli*; Eh, *Ectothiorhodospira halophila*; Nc: *Neurospora crassa*; Os, *O. sativa*, Rm: *Rhizobium meliloti*.

each of the described cases, this would elevate the biochemical activities identified through these experiments to the status of major, unexpected discoveries, because the chemistry underlying them would have to be extremely unusual. In particular, if the identification of the *M. jannaschii* cysteinyl-tRNA synthetase is indeed correct, this enzyme would have to be a derivative of a specific family of polysaccharide hydrolases containing a signal peptide but no recognizable ATP-binding or RNA-binding domains.

The third explanation is that the original computational predictions triggered over-interpretation of the experimental results that, in reality, might have been obtained as a result of nonspecific activities, contamination or other artifacts. In this regard, it is important to realize that not only computational predictions, but biological experiments also, are intrinsically error-prone and open to conflicting interpretations. The probabilistic nature of computational analyses is well realized (and at times, perhaps, overrated) by most researchers, probably because explicit calculation of probability or likelihood is at the core of most widely used computer methods for sequence and structure analyses. In this regard, it is prudent to note that the alternative computational predictions presented here should be considered to be 'more likely' than the original ones, rather than to contradict the latter in an absolute sense. As we attempted to show above, however, the difference in the likelihood of two mutually incompatible predictions can be overwhelming, with one supported by multiple lines of evidence as opposed to the other. In contrast to computational studies, experimental ones are often, consciously or unconsciously, treated as demonstration of 'final truth'. In reality, however, probabilistic inference is inherent in practically any interpretation of experimental results when questions are asked such as "How likely is it that the protein under study has a particular biochemical activity *in vivo*?" or "How central is this activity for the *in vivo* function of the protein under study, given the results of a surrogate *in vitro* assay?" Thus, certain experimental designs may not be appropriate to ascertain the actual *in vivo* biochemistry of a protein. Furthermore, even if the particular activities detected under these conditions are genuine, the likelihood of these being relevant *in vivo* needs to be additionally assessed. Accordingly, when strong computational predictions seem not to be borne out by experiment, the conditions and design of the experiments deserve special scrutiny: they might have given a negative result for a wrong reason. A case in point is the MJ0107 protein, the apparent archaeal ortholog of DHPS, which failed to show dihydropteroate synthase activity [25]. We strongly believe that this issue needs to be revisited. All this considered, the results of independent application of computational and experimental techniques tend to be complementary, and useful in adding or reducing confidence in the biological conclusions of a particular study.

Finally, it should be emphasized that these cautionary notes on application of computational methods in protein function prediction in no way suggest that new computational approaches that depart sharply from more established ones are doomed to failure. Indeed, the most popular advanced search methods based on sequence profiles - PSI-BLAST and Hidden Markov Model (HMM) search - are rather recent innovations [11,51,52]. Furthermore, methods based on a different principle, such as protein sequence-structure threading, have a recent history of success despite uncertainties in their statistical foundations [22,53-56]. It does seem, however, that when a structurally and functionally plausible prediction is produced, with a high confidence, by a well

tested, statistically sound computational method, an incompatible prediction yielded by a new method without a clear statistical foundation is most likely to be incorrect.

Materials and methods

The non-redundant protein-sequence database at the National Center for Biotechnology Information (NCBI) was searched using the gapped version of the BLAST program [9]. Sequence-profile searches were carried out using the PSI-BLAST program, with the cut-off for inclusion of sequences into the profile set at $E = 0.01$ [3,9], and the HMMer program package [57]. Multiple alignments of amino-acid sequences were generated using the T_Coffee program [58]. Protein secondary-structure predictions were generated using the PHD program [59,60], with multiple alignments of individual protein families used as queries. Sequence-structure threading was carried out using the combined-fold-prediction algorithm [22] or the 3D-PSSM algorithm based on the use of a three-dimensional position-specific scoring matrix [23]. Signal peptides in protein sequences were predicted using the SignalP program [61]. The COG database [62,63] was used as a source of information on orthologous relationships between proteins.

References

1. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back.** *J Mol Biol* 1998, **283:**707-725.
2. Koonin EV, Aravind L, Kondrashov AS: **The impact of comparative genomics on our understanding of evolution.** *Cell* 2000, **101:**573-576.
3. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287:**1023-1040.
4. Murzin AG: **Progress in protein structure prediction.** *Nat Struct Biol* 2001, **8:**110-112.
5. Karlin S, Bucher P, Brendel V, Altschul SF: **Statistical methods and insights for protein and DNA sequences.** *Annu Rev Biophys Biophys Chem* 1991, **20:**175-203.
6. Karlin S, Brendel V: **Chance and statistical significance in protein and DNA sequence analysis.** *Science* 1992, **257:**39-49.
7. Karlin S, Altschul SF: **Applications and statistics for multiple high-scoring segments in molecular sequences.** *Proc Natl Acad Sci USA* 1993, **90:**5873-5877.
8. Karlin S: **Statistical studies of biomolecular sequences: score-based methods.** *Phil Trans R Soc Lond B* 1994, **344:**391-402.
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
10. Altschul SF, Bundschuh R, Olsen R, Hwa T: **The estimation of statistical parameters for local alignment score distributions.** *Nucleic Acids Res* 2001, **29:**351-361.
11. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, MA: Cambridge University Press; 1998.
12. Ibba M, Soll D: **The renaissance of aminoacyl-tRNA synthesis.** *EMBO Rep* 2001, **2:**382-387.
13. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12:**334-336.
14. Galperin MY, Walker DR, Koonin EV: **Analogous enzymes: independent inventions in enzyme evolution.** *Genome Res* 1998, **8:**779-790.
15. Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, Ibba M, Soll D: **One polypeptide with two aminoacyl-tRNA synthetase activities.** *Science* 2000, **287:**479-482.
16. Lipman RS, Sowers KR, Hou YM: **Synthesis of cysteinyl-tRNA(Cys) by a genome that lacks the normal cysteine-tRNA synthetase.** *Biochemistry* 2000, **39:**7792-7798.
17. Stathopoulos C, Jacquin-Becker C, Becker HD, Li T, Ambrogelly A, Longman R, Soll D: ***Methanococcus jannaschii* prolyl-cysteinyl-tRNA synthetase possesses overlapping amino acid binding sites.** *Biochemistry* 2001, **40:**46-52.
18. Fabrega C, Farrow MA, Mukhopadhyay B, de Crecy-Lagard V, Ortiz AR, Schimmel P: **An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes.** *Nature* 2001, **411:**110-114.
19. Wolf YI, Aravind L, Grishin NV, Koonin EV: **Evolution of aminoacyl-tRNA synthetases - analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events.** *Genome Res* 1999, **9:**689-710.
20. Nielsen H, Brunak S, von Heijne G: **Machine learning approaches for the prediction of signal peptides and other protein sorting signals.** *Protein Eng* 1999, **12:**3-9.
21. Tamura J-I, Kaname H, Kadowaki K, Igarashi Y, Kodama T: **Molecular cloning and sequence analysis of the gene encoding an endo α-1,4 polygalactosaminidase of *Pseudomonas* sp. 881.** *J Ferment Bioeng* 1995, **80:**305-310.
22. Fischer D: **Hybrid fold recognition: combining sequence derived properties with evolutionary information.** *Pac Symp Biocomput* 2000:119-130.
23. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299:**499-520.
24. Hampele IC, D'Arcy A, Dale GE, Kostrewa D, Nielsen J, Oefner C, Page MG, Schonfeld HJ, Stuber D, Then RL: **Structure and function of the dihydropteroate synthase from *Staphylococcus aureus*.** *J Mol Biol* 1997, **268:**21-30.
25. Xu H, Aurora R, Rose GD, White RH: **Identifying two ancient enzymes in Archaea using predicted secondary structure alignment.** *Nat Struct Biol* 1999, **6:**750-754.
26. **COG: Phylogenetic classification of proteins encoded in complete genomes** [http://www.ncbi.nlm.nih.gov/COG/]
27. Aurora R, Rose GD: **Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons.** *Proc Natl Acad Sci USA* 1998, **95:**2818-2823.
28. Aravind L: **An evolutionary classification of the metallo-β-lactamase fold proteins.** *In Silico Biology* 1998, 1:8; available at [http://www.bioinfo.de/isb/1998/01/0008/].
29. Matthews DA, Appelt K, Oatley SJ: **Crystal structure of *Escherichia coli* thymidylate synthase with FdUMP and 10-propargyl-5,8-dideazafolate.** *Adv Enzyme Regul* 1989, **29:**47-60.
30. Dynes JL, Firtel RA: **Molecular complementation of a genetic marker in *Dictyostelium* using a genomic DNA library.** *Proc Natl Acad Sci USA* 1989, **86:**7966-7970.
31. Mushegian AR, Koonin EV: **Cell-to-cell movement of plant viruses. Insights from amino acid sequence comparisons of movement proteins and from analogies with cellular transport systems.** *Arch Virol* 1993, **133:**239-257.
32. Melcher U: **The '30K' superfamily of viral movement proteins.** *J Gen Virol* 2000, **81:**257-266.
33. Xoconostle-Cazares B, Xiang Y, Ruiz-Medrano R, Wang HL, Monzer J, Yoo BC, McFarland KC, Franceschi VR, Lucas WJ: **Plant paralog to viral movement protein that potentiates transport of mRNA into the phloem.** *Science* 1999, **283:**94-98.
34. **CD-search** [http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi]
35. Ponting CP, Parker PJ: **Extending the C2 domain family: C2s in PKCs delta, epsilon, eta, theta, phospholipases, GAPs, and perforin.** *Protein Sci* 1996, **5:**162-166.
36. Xoconostle-Cazares B, Ruiz-Medrano R, Lucas WJ: **Proteolytic processing of CmPP36, a protein from the cytochrome b(5) reductase family, is required for entry into the phloem translocation pathway.** *Plant J* 2000, **24:**735-747.
37. Neuwald AF, Landsman D: **GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein.** *Trends Biochem Sci* 1997, **22:**154-155.
38. Hai TW, Liu F, Coukos WJ, Green MR: **Transcription factor ATF cDNA clones: an extensive family of leucine zipper proteins able to selectively form DNA-binding heterodimers.** *Genes Dev* 1989, **3:**2083-2090.

39. Nagadoi A, Nakazawa K, Uda H, Okuno K, Maekawa T, Ishii S, Nishimura Y: **Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain.** *J Mol Biol* 1999, **287:**593-607.
40. Kawasaki H, Schiltz L, Chiu R, Itakura K, Taira K, Nakatani Y, Yokoyama KK: **ATF-2 has intrinsic histone acetyltransferase activity which is modulated by phosphorylation.** *Nature* 2000, **405:**195-200.
41. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18:**269-285.
42. Taylor BL, Zhulin IB: **PAS domains: internal sensors of oxygen, redox potential, and light.** *Microbiol Mol Biol Rev* 1999, **63:**479-506.
43. Anantharaman V, Koonin EV, Aravind L: **Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains**. *J Mol Biol* 2001, **307:**1271-1292.
44. Ni M, Tepperman JM, Quail PH: **PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein.** *Cell* 1998, **95:**657-667.
45. Zhu Y, Tepperman JM, Fairchild CD, Quail PH: **Phytochrome B binds with greater apparent affinity than phytochrome A to the basic helix-loop-helix factor PIF3 in a reaction requiring the PAS domain of PIF3.** *Proc Natl Acad Sci USA* 2000, **97:**13419-13424.
46. Levchenko I, Smith CK, Walsh NP, Sauer RT, Baker TA: **PDZ-like domains mediate binding specificity in the Clp/Hsp100 family of chaperones and protease regulatory subunits.** *Cell* 1997, **91:**939-947.
47. Neuwald AF, Aravind L, Spouge JL, Koonin EV: **AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes.** *Genome Res* 1999, **9:**27-43.
48. Imai Y, Kimura T, Murakami A, Yajima N, Sakamaki K, Yonehara S: **The CED-4-homologous protein FLASH is involved in Fas-mediated activation of caspase-8 during apoptosis.** *Nature* 1999, **398:**777-785.
49. Koonin EV, Aravind L, Hofmann K, Tschopp J, Dixit VM: **Apoptosis. Searching for FLASH domains.** *Nature* 1999, **401:**662-663.
50. Popper K: *The Logic of Scientific Discovery*. New York/London: Routledge; 1999.
51. Altschul SF, Koonin EV: **PSI-BLAST - a tool for making discoveries in sequence databases.** *Trends Biochem Sci* 1998, **23:**444-447.
52. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
53. Bryant SH, Altschul SF: **Statistics of sequence-structure threading.** *Curr Opin Struct Biol* 1995, **5:**236-244.
54. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.** *J Mol Biol* 1999, **287:**797-815.
55. Panchenko A, Marchler-Bauer A, Bryant SH: **Threading with explicit models for evolutionary conservation of structure and sequence.** *Proteins* 1999, **37:**133-140.
56. Panchenko AR, Marchler-Bauer A, Bryant SH: **Combination of threading potentials and sequence profiles improves fold recognition.** *J Mol Biol* 2000, **296:**1319-1331.
57. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6:**361-365.
58. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
59. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction-based threading.** *J Mol Biol* 1997, **270:**471-480.
60. Rost B, Sander C, Schneider R: **PHD - an automatic mail server for protein secondary structure prediction.** *Comput Appl Biosci* 1994, **10:**53-60.
61. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Int J Neural Syst* 1997, **8:**581-599.
62. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-617.
63. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.