PROTEINS: Structure, Function, and Genetics 48:75–84 (2002)

Expanding the Nitrogen Regulatory Protein Superfamily: Homology Detection at Below Random Sequence Identity

Lisa N. Kinch and Nick V. Grishin*

Howard Hughes Medical Institute, and Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas

ABSTRACT **Nitrogen regulatory (PII) proteins are signal transduction molecules involved in controlling nitrogen metabolism in prokaryots. PII proteins integrate the signals of intracellular nitrogen and carbon status into the control of enzymes involved in nitrogen assimilation. Using elaborate sequence similarity detection schemes, we show that five clusters of orthologs (COGs) and several small divergent protein groups belong to the PII superfamily and predict their structure to be a $(\beta\alpha\beta)_2$ ferredoxin-like fold. Proteins from the newly emerged PII superfamily are present in all major phylogenetic lineages. The PII homologs are quite diverse, with below random (as low as 1%) pairwise sequence identities between some members of distant groups. Despite this sequence diversity, evidence suggests that the different subfamilies retain the PII trimeric structure important for ligand-binding site formation and maintain a conservation of conservations at residue positions important for PII function. Because most of the orthologous groups within the PII superfamily are composed entirely of hypothetical proteins, our remote homology-based structure prediction provides the only information about them. Analogous to structural genomics efforts, such prediction gives clues to the biological roles of these proteins and allows us to hypothesize about locations of functional sites on model structures or rationalize about available experimental information. For instance, conserved residues in one of the families map in close proximity to each other on PII structure, allowing for a possible metal-binding site in the proteins coded by the locus known to affect sensitivity to divalent metal ions. Presented analysis pushes the limits of sequence similarity searches and exemplifies one of the extreme cases of reliable sequence-based structure prediction. In conjunction with structural genomics efforts to shed light on protein function, our strategies make it possible to detect homology between highly diverse sequences and are aimed at understanding the most remote evolutionary connections in the protein world. Proteins 2002;48:75–84.**

Key words: PII nitrogen regulatory protein; BLAST database searches; homology detection; structure prediction; threading; protein classification; structural genomics

INTRODUCTION

Genomic sequencing efforts are providing an ever-increasing number of predicted gene products from a diverse set of organisms. The next step toward using this wealth of information to understand fundamental biological processes involves classifying these gene products into more meaningful functional and structural groups. The task of assigning potential functions and structures to new protein sequences relies heavily on detecting similarities to known protein sequences. Recently developed programs such as PSI-BLAST[1] and HMMer[2,3] have increased the sensitivity of such similarity searches and allowed for genomewide, automated assignments of potential protein functions. In addition, threading programs such as GenTHREADER[4] or the fold recognition method of Fischer[5] can automate the assignment of potential protein structure. Sequence similarity detection methods have been used successfully in the generation of functional clusters on a genomewide scale in the COG database,[6,7] which currently contains 2791 COGs, including proteins from the completed genomes of 6 archaea, 22 bacteria, and 3 eukaryotes. The COGs are classified into 17 broad functional categories, including one of "uncharacterized" function[7] representing 10–20% of assigned COGs. Often, the detection of remote homologs required to link uncharacterized clusters to those with known structure or function requires a unique combination of sequence and structure analysis methods that can only be performed manually at the present time. This article describes such a combination of approaches to expand a family with a known structure, the nitrogen regulatory (PII) proteins, to include several distant groups of sequences with no known structure or function.

PII proteins (COG0347) are signal transduction molecules involved in controlling nitrogen metabolism in prokaryotic cells and eukaryotic chloroplasts (reviewed in Refs. 8 and 9). PII proteins integrate the signals of

Abbreviations: PII, nitrogen regulatory proteins; COG, clusters of orthologous groups; BLAST, basic local alignment search tool; PDB, Protein Data Bank; NRII, histidine kinase; GS, glutamine synthase; ATP, adenosine triphosphate; kD, kilodalton; aa, amino acid.

*Correspondence to: Nick B. Grishin, Howard Hughes Medical Institute, and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX. 75390-9050. E-mail: grishin@chop.swmed.edu

Received 18 December 2001; Accepted 18 December 2001

intracellular nitrogen and carbon status into the control of enzymes involved in nitrogen assimilation. The nitrogen signal, glutamine, dictates post-translational modification of PII through binding its modifying enzymes. Low glutamine levels trigger the uridylylation (or phosphorylation for some species) of PII, decreasing its affinity for a histidine kinase (NRII) and ultimately upregulating the expression of nitrogen assimilation genes such as glutamine synthase (GS). The carbon signal, 2-ketoglutarate, binds to the PII protein synergistically with a second effector molecule, adenosine triphosphate (ATP). Fully saturated PII ceases to interact with NRII and an adenylyltransferase responsible for regulating the activity of GS. Thus, the function of PII regulator encompasses its protein effector interactions, its post-translational modification, and its small molecule-binding capacity.

Several bacterial genomes possess a second *PII* gene known as *GlnK*. In *E. coli* the two gene products retain 67% identity and function similarly (gi|1633299 and gi|5822483, respectively, in Fig. 2). The crystal structures of both PII and GlnK have been solved.[10–12] The two proteins adopt identical folds with differences confined to loop conformations and the extreme C-terminus. Both structures reveal closely associated trimers, with three ATP-binding sites found at the junctions between monomers [Fig. 4(A)]. Each PII monomer contains a core $(\beta\alpha\beta)_2$ secondary structural pattern [corresponding to aAbcBd, Fig. 4(A)] described in SCOP as a Ferredoxin-like fold[13] and in CATH as an α,β plait,[14] with two additional C-terminal β-strands [corresponding to ef, Fig. 4(A)] that help stabilize the quaternary structure. Highly conserved residues among the PII proteins (black highlights, Fig. 2) are mainly involved in the formation of the ATP-binding pocket, with residues from each of two individual monomers contributing to the ligand-binding site. Thus, homologs of the PII protein family can be assigned with an overall core $(\beta\alpha\beta)_2$ fold and will likely retain a similar ligand-binding pocket defined by conserved residues.

In this study we expand the PII superfamily to include several divergent paralogous groups that were detected by using various BLAST strategies, including a transitive PSI-BLAST approach and an alignment-seeded PSI-BLAST approach. Such strategies have increased the sensitivity of database searching over classic pairwise gapped-BLAST procedures and can detect sequences with as low as 7–10 % identity.[15] The diverse protein superfamily defined in this article includes sequences from several distinct groups that are clustered according to evolutionary distances. Remarkably, pairwise comparisons of sequences from different groups show below random identities (as low as 1%). We provide evidence for including each group of detected sequences in the PII superfamily in the form of BLAST statistics, fold recognition scores, secondary structure predictions, and hydrophobicity plots. Finally, we provide fold predictions for sequence groups, which otherwise have no known structure or function.

MATERIALS AND METHODS

Sequence Similarity Searches

To detect homologs of the PII protein family, we searched the non-redundant database (nr, May 16, 2001; 687,743 sequences-Aug 1,2001; 727,771 sequences, filtered for low-complexity regions) and the ERGO database (http://wit.integratedgenomics.com/ERGO/), which includes sequences from a total of 80 genomes. PSI-BLAST searches on the nr database with defined parameters (BLOSUM62 matrix, E-value threshold 0.01) were iterated to convergence starting with a single query sequence. Found homologs were grouped by using linkage clustering (score of 1 bit per site threshold, about 50% identity),[16] and representative sequences from each group were used as new queries for subsequent rounds of PSI-BLAST. The iterations were repeated until no new sequences were detected. To retrieve additional sequences not found in GenBank, representative sequences were also used to search the ERGO database (E-value threshold 0.001). We used the COG database (http://www.ncbi.nlm.nih.gov/COG/) to define orthologous groups of the detected sequences, and the SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/i) and CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/) databases to classify folds.

Multiple Sequence Alignments, Alignment-Seeded PSI-BLAST Searches, and Threading

We constructed multiple-sequence alignments for each detected group using the program T-COFFEE.[17] Secondary structure predictions (JPRED server,[18]) and patterns of hydrophobicity guided manual adjustments to the alignment, which served as input to generate a position-specific scoring matrix or profile (-B option in blastpgp) for a new round of BLAST searches. Each member sequence from the alignment was used as a query sequence to search the nr database in a single iteration of BLAST using the alignment generated profile. Hits to these individual query sequences were reported with the BLAST statistics (E-value) produced by this procedure. Weak hits were further justified by using the hybrid fold recognition method of Fischer[5] found on the BIOINBGU server (http://www.cs.bgu.ac.il/~bioinbgu), which incorporates evolutionary information into a traditional threading procedure. The multiple alignments of each group were merged into a global alignment by using secondary structure predictions, hydrophobicity plots, paired BLAST hit alignments, and fold recognition structure-sequence alignments as guides. The complete full-length alignment is available through anonymous FTP from ftp://iole.swmed.edu/pub/lkinch/PII.

Euclidian Space Mapping and Distance Diagram

We used the global multiple sequence alignment shown in Figure 2 to calculate identity fractions q_{ij} between sequence pairs i and j. We used the formula $d_{ij} = 1/((q_{ij} - q_{ij}^{ran})/(1 - q_{ij}^{ran}) - 1)$ in conversions of identity fractions to evolutionary distances, with q_{ij}^{ran} representing the expected identity between two random sequences with the same amino acid composition as the sequences i and j. Each sequence was represented as a point in a multidimen-

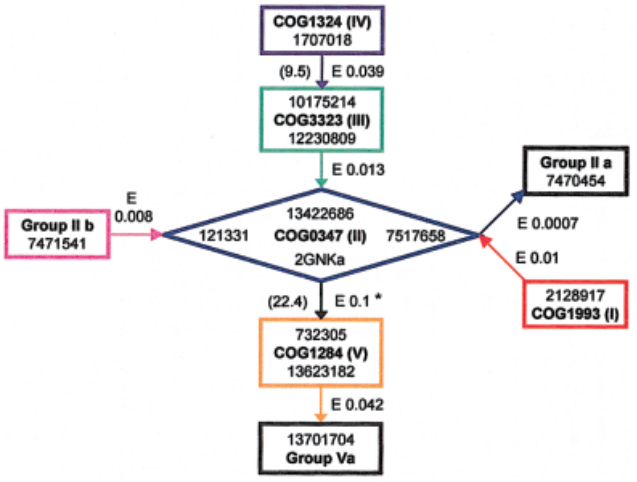

Fig. 1. Overview of sequence searching strategy. Boxes represent groups of closely related sequences and are labeled according to group number (see Fig. 3) and COG (if applicable). Query and hit sequences used to link groups are displayed in their representative boxes and labeled according to NCBI accession number (gi) or to PDB accession ID. Arrows point away from the query sequence toward the hit sequence produced with alignment-seeded BLAST searches (see text for methods and parameters). Boxes and arrows are colored according to alignments used as input for seeding, except where indicated by the asterisk (where the alignment included groups II, IIa, IIb, and I). E-values produces by seeded BLAST and consensus scores from fold recognition are indicated to the side of the arrows (consensus scores shown in parentheses, where applied).

sional Euclidian space in such a way that Euclidian distances $\bar{d}_{ij}$ between the points optimally approximated the estimated distances d_{ij} between the sequences: $\sum_{ij}(\bar{d}_{ij}^2 - d_{ij}^2)^2/d_{ij}^4 =$ min. These points were grouped by using the following procedure. Each point representing a sequence generated a Gaussian density in the Euclidian space having the following properties. The mean of each density was the point's coordinates and the variance σ^2 of each density was identical for all points. Starting from each point, the local maximum of the sum of such Gaussians was found. The points giving rise to the same local maximum were grouped together.

RESULTS AND DISCUSSION

Detection of Distant Homologs

To detect distant homologs of the PII family, we performed extensive transitive PSI-BLAST searches on the nr database. Multiple-sequence alignments were constructed and used to cluster sequences according to evolutionary distances in Euclidian space and to generate BLAST statistics to support links between clusters (see Materials and Methods for details). The overall process of detecting distant PII family sequences and the statistical relationships of these protein groups are summarized in Figure 1. Briefly, alignments of the central PII family (group II, COG0347) detect the closely related group IIa sequences and the distantly related group V (COG1284) sequences. Individual alignments of the group I (COG1993), group III (COG3323), and group IIb sequences link back to the central group II sequences, and alignments of the peripheral group IV (COG 1324) and group Va detect the group III and group V sequences, respectively. Below, we outline this process in more detail for each group of sequences.

Nitrogen regulatory (PII) proteins

We set out to define all possible PII sequences using the COG database and transitive PSI-BLAST searches. The COG database defines 35 sequences from 21 different genomes as belonging to the PII orthologous cluster, including one distinct sequence from *Aquifex aeolicus* (gi|7517658) that appears to have lost many of the conserved residues required for PII function. Transitive BLAST searches of the nr database identified 72 additional group II family members in the first round of iterations (E-value cutoff 0.01) and the distinct *Aquifex aeolicus* sequence (gi|7517658) in the second round. To expand the diversity this *Aquifex aeolicus* sequence brings to the PII family, we searched the ERGO database for similar sequences not contained in GenBank (group IIa sequences 13 and 14, Fig. 2). Using these new sequences to initiate BLAST searches, we detected members of the PII family and a new sequence corresponding to a hypothetical protein (gi|7451867) that belongs to a different COG (COG1993).

Group I (COG1993) sequences

The COG database classifies this group of seven proteins from three archaea and three bacteria, with one bacterial species possessing two paralogs, as an "uncharacterized ancient conserved region." Hits from transitive BLAST searches starting with a group I query sequence (gi|7451866) combined with hits from the ERGO database extended COG1993 to include six additional homologs of unknown function. In addition, one of the representative proteins contains three similar domains linked together to form a single protein (gi|7477475). We included each of these domains separately in a multiple alignment of the entire family (representative sequences shown in Fig. 2, group I). This alignment was used to seed BLAST searches, which produced a significant hit (E-value 0.01) to the PII sequence from *Aquifex aeolicus* (gi|7517658) with the COG1993 query sequence (gi|2128917).

Peripheral group II sequences

The inclusion of COG1993 in the PII family provides position-specific sequence diversity to the global alignment and potentates the discovery of additional homologs. Therefore, we repeated BLAST searches using the new global alignment to generate a profile. This analysis yielded only one additional sequence from *Synechocystis* sp. (gi|7470454) with significant statistics (E-value of 7e-4 to gi|7517658). On the basis of distance mapping, we grouped this sequence and a similar sequence from *Anabaena* sp. found in the ERGO database (RAN00858) with the other divergent PII sequences (group IIa, Fig. 2). During the course of PSI-BLAST runs we also noticed a couple of potentially meaningful weak hits to the PII proteins that retained many of the conserved small hydrophobic residues across a significant portion of their BLAST

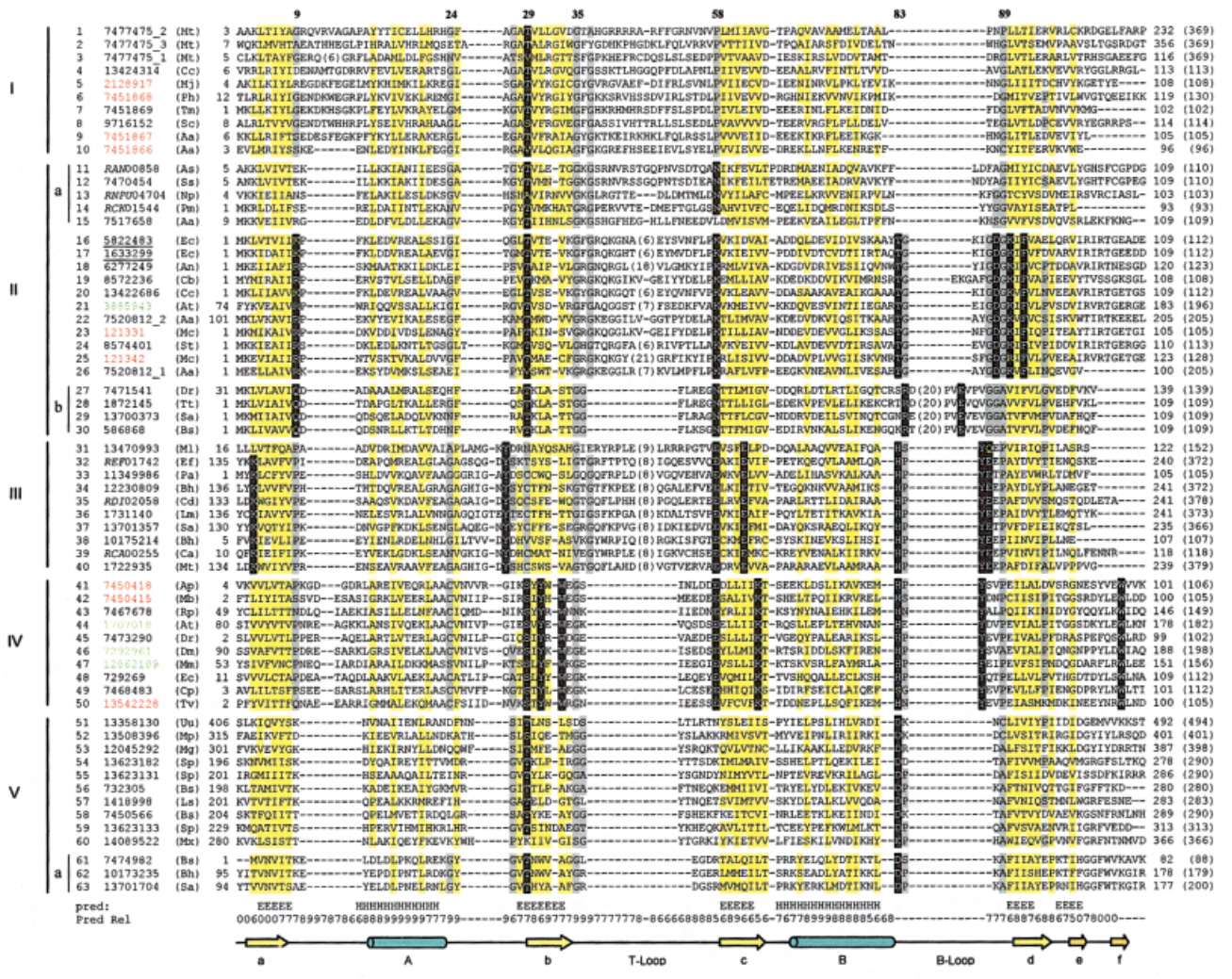

Fig. 2. Multiple-sequence alignment of the nitrogen regulatory protein superfamily. The five classes of sequences (I–V) shown are grouped and numbered according to Euclidian distance mapping, with subgroups (IIa, IIb, and Va) corresponding to BLAST-derived groups. Group I sequences contain members of COG1993; group II sequences contain members of the nitrogen regulatory family (COG0347) and are divided into two additional subclasses. Group III, groupIV, and group V sequences contain members of COG3323, COG1324, and COG1260, respectively. Each sequence is identified by the NCBI gene identification number (gi) or the ERGO gene identifier (italics) and colored according to superkingdom in black (bacterial), red (archaeal), or green (eukaryotic). The sequence identifiers corresponding to known structures are underlined, and the residue numbers marked above the alignment correspond to these sequences. The secondary structural element diagram shown below the alignment, with β-strands and α-helices shown in yellow arrows and blue cylinders, respectively, is based on known structures (2GNK and 2PII). The secondary structural elements (E for β-sheets and H for α-helices) predicted by a component of JPRED (Pred) and the reliability of this prediction (rel) are based on the entire multiple-sequence alignment and are also shown below the alignment. The first and last residue numbers of the shown sequences are indicated before and after each sequence, with the total length of the sequences following in parentheses. Some nonconserved residues in loops are omitted, and the number of omitted residues is shown in parentheses. Residues conserved among groups are highlighted black, uncharged residues at mainly hydrophobic positions are highlighted in yellow, and conserved small residues are highlighted in gray. The species name abbreviations are as follows: Aa, *Aquifex aeolicus*; An, *nitrogen-fixing bacterium ANFK33*; Ap, *Aeropyrum pernix*; As, *Anabaena* sp.; At, *Arabidopsis thaliana*; Bh, *Bacillus halodurans*; Bs, *Bacillus subtilis*; Ca, *Clostridium acetobutylicum*; Cb, *Clostridium beijerinckii*; Cc, *Caulobacter crescentus*; Cd, *Corynebacterium diphtheriae*; Cp, *Chlamydophila pneumoniae*; Dm, *Drosophila melanogaster*; Dr, *Deinococcus radiodurans*; Ec, *Escherichia coli*; Ef, *Enterococcus faecalis*; Lm, *Listeria monocytogenes*; Ls, *Lactobacillus sakei*; Mb, *Methanothermobacter thermautotrophicus*; Mc, *Methanococcus thermolithotrophicus*; Mg, *Mycoplasma genitalium*; Mj, *Methanococcus jannaschii*; Ml, *Mesorhizobium loti*; Mm, *Mus musculus*; Mp, *Mycoplasma pneumoniae*; Mt, *Mycobacterium tuberculosis*; Mx, *Mycoplasma pulmonis*; Np, *Nostoc punctiforme*; Pa, *Pseudomonas aeruginosa*; Ph, *Pyrococcus horikoshii*; Pm, *Prochlorococcus marinus*; Rp, *Rickettsia prowazekii*; Sa, *Staphylococcus aureus*; Sc, *Streptomyces coelicolor*; Sp, *Streptococcus pyogenes*; Ss, *Synechocystis* sp.; St, *Streptococcus thermophilus*; Tm, *Thermotoga maritima*; Tt, *Thermus thermophilus*; Tv, *Thermoplasma volcanium*; Uu, *Ureaplasma urealyticum*.

alignments. For example, the second iteration of PSI-BLAST with the query PII sequence (gi|121384) detected a hypothetical 12-kD protein from *Thermus thermophilus* (gi|1872145, E-value of 0.025) and an unknown conserved protein from *Bacillus halodurans* (gi|12230809, E-value of 1.5). Persuing these weak hit as potential homologs, we searched the COG database and initiated transitive BLAST searches. The first weak hit (gi|1872145) does not belong to a COG, and its first round of PSI-BLAST iterations defined a group of six bacterial proteins, with subsequent rounds

yielding no additional homologs (group IIb, Fig. 2). However, a slightly higher cutoff (E-value of 0.03) in the iterative analysis retrieved the same group of proteins in the first round, and the PII proteins in subsequent rounds, without detecting any erroneous hits. Seeding BLAST searches with an alignment of these sequences produces a significant hit (E-value of 0.008) to a nitrogen regulatory protein (gi|121331) with the bacterial protein query (gi|7471541). These new sequences are included in the multiple-sequence alignment as a subset of group II (group IIb, Fig. 2).

Group III (COG3323) sequences

The amino acid sequence of the second weak hit to the nitrogen regulatory proteins is significantly larger (372 aa) than a typical PII protein domain (around 112 aa), and the paired BLAST alignment extends across the middle portion of the sequence (aa 149–aa 222). Indeed, COG analysis of this hypothetical protein sequence reveals it to contain more than one domain, with the aligned center portion belonging to COG3323 and the remaining divided sequence belonging to COG0327. The aligned COG3323 is classified as another uncharacterized bacterial conserved region. Transitive PSI-BLAST searches picked up all family members in the first round of iterations and nitrogen regulatory proteins in the second round (E-value cutoff of 0.01), and seeding BLAST searches with a group alignment produced significant hits to several PII family members. For example, the COG3323 query sequence (gi|13422686) detects a PII protein (gi|12230809) with an E-value of 0.013.

Group IV (COG1324) sequences

In addition to detecting PII sequences, individual PSI-BLAST searches with members of the group III proteins produced weak hits to several divalent metal ion tolerance proteins (CutA). These weak hits encompassed the C-terminal half of both the query and the subject sequences and included the conserved sequence motif ($_{83}$HPYEXP$_{89}$) downstream from a conserved charged residue (E_{58}), without gaps (Fig. 2). Therefore, we produced an alignment of all members of the CutA family (COG1324) using the same procedures and criteria used for the previous groups. Seeding BLAST with this alignment linked a CutA query sequence (gi|1707018) to the group III family (gi|10175214) with a modest E-value (0.036). Given this moderate BLAST statistic, we sought to provide additional support for including group IV sequences in the PII superfamily by using the fold recognition method of Fischer[5] found on the BIOINBGU server (http://www.cs.bgu.ac.il/~bioinbgu/). This method identified the ferredoxin fold of the splicesomal U1A protein (PDB entry 1URN chain A) as the top hit (consensus score 9.5) when given the group IV query sequence (gi|729269). The 1URN_A adopts the same fold as the PII structure (ferredoxin secondary structure pattern $(\beta\alpha\beta)_2$), with a small C-terminal helix replacing strand e and strand f of PII. Although the consensus score for this prediction (9.5) is below the confidence limit of the method (<12), no other ranked folds approached this score (next highest score 5.9). Although the structural fold prediction for the group IV sequences does not imply an evolutionary link to splicesomal U1A, it further supports linking this group to the rest of the PII superfamily.

Group V (COG1284) and Va sequences

With the growing diversity of the PII superfamily, we decided to seed new rounds of BLAST searches with different combinations of PII superfamily subsets. A subset containing all group I and group II sequences (group 1, IIa, II, and IIb sequences found in nr, ERGO, and PDB databases) produced weak hits to the C-terminal half ($\approx$80 aa) of several predicted integral membrane proteins (lowest E-value of 0.10 to gi|732305). Each of these weak hits belongs to COG1284, which is another uncharacterized bacterial conserved region. Members of this COG contain six predicted transmembrane spans in the N-terminus, followed by the C-terminal BLAST-detected domain. A seeded PSI-BLAST search with the COG1284 C-terminal domain multiple alignment detects an additional smaller group of sequences (gi|10173235) with the query (gi|1418998) with modest statistics ($E = 0.055$) but does not detect any PII proteins. It is surprising that threading with the fold recognition method of Fischer[5] using a member of the group V sequences (gi|732305) produces a nitrogen regulatory protein structure (PDB entry 1PIL) as the top hit with a consensus score (22.4) above the confidence threshold of the method. This strong fold prediction provides compelling support for including the weakly linked group V sequences in the PII superfamily.

Sequence grouping: euclidian space mapping and distance diagram

Our extensive searching procedures have linked several groups of sequences to the PII family. These groups are quite diverse, with pairwise sequence identities between members of distant groups as low as 1%. When combined with the relatively short length of the member sequences (about 100 aa), the extended diversity of the PII superfamily makes complete phylogenetic analysis unreliable. Therefore, to understand and visualize the relationships between superfamily members, we chose to represent approximations of evolutionary distances between sequences as points in Euclidian space.[19] We use these points to group similar sequences (see Materials and Methods for details). Each of the sequence clusters (I–V) produced by distance mapping corresponds to a single COG and broadly follows the BLAST-derived sequence groups highlighted in Figure 1. The distance diagram illustrated in Figure 3 shows the intermediate sequences (group IIa, 11–15; blue triangles) responsible for linking the group I sequences (red circles) to the PII sequences (blue circles). Appearing further back in the plane of the diagram, the group III sequences (green circles) cluster between the PII sequences and the group IV sequences (yellow circles), whereas the group V sequences (orange circles) appear further forward in the plane of the diagram, nearest to the PII sequences. Although the group Va sequences are linked to the other members of group V with

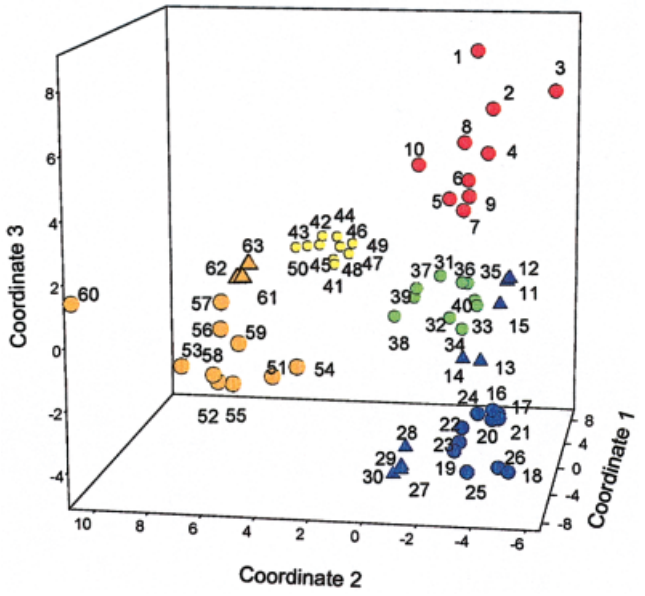

Fig. 3. Euclidian distance mapping. Coordinates 1, 2, and 3 are three dimensions of a maximal scatter of points in multidimensional space. The data points represent the sequences shown in Figure 3 and are numbered accordingly. The symbols correspond to grouped sequences: filled red circles (group I, 1–10); filled blue circles (group II PII, 16–25); filled blue triangles (group IIa, 11–15 and group IIb, 26–29); filled green circles (group III, 30–39); filled yellow circles (group IV, 40–49); filled orange circles (group V, 50–59); and filled orange triangles (group Va, 60–63).

moderate BLAST statistics (E-value = 0.042), the distance mapping classifies these sequences (61–63, orange triangles) within group V. In fact, the sequence (gi|14089533, 60), which is detected by automatic BLAST searching procedures, appears to be more distantly related to other group members than these sequences.

Structural and Functional Implications of Expanding the PII Family

Although a great deal of functional and structural information is available for the PII signal transduction proteins, virtually nothing is known about the other groups of proteins found in this study. The groups are all described as "uncharacterized" in the COG database and have no available structural information. By linking these diverse groups to the PII proteins, we can make fold predictions based on known PII structures. Such predictions may provide functional information about otherwise unknown proteins. For the PII proteins, trimer formation establishes the ATP-binding sites. Therefore, we consider the quaternary structural propensity of new PII superfamily members by using secondary structure predictions and hydrophobicity conservations. Because structurally and functionally important residues do not tend to change, we can also gather information about the ligand-binding site of new PII superfamily members using residue conservations in the global multiple sequence alignment (Fig. 2). Finally, we consider individual group fold predictions and gene linkage information in terms of protein function.

Trimer formation

As observed in the PII structures, individual residues involved in trimer interface interactions are not strictly conserved among the group II sequences. The primary force driving trimer formation appears to be dictated by the hydrogen bonding of backbone strand-strand interactions between neighboring monomers. These interactions result in the formation of a β-sheet surrounding a central cavity in the PII trimer [Fig. 4(A)]. In the trimer interface, strand e interacts with a neighboring strand d at one end of the cavity; and the second half of strand b interacts with the first half of a neighboring strand b at the other end of the cavity, with a bend in the center of strand b mediating the interface.[11] Therefore, the presence of these two elements, a bend (or possibly an insertion) in strand b and the existence of strand e, provide the best evidence for the prediction of trimer formation within the different PII groups.

The presence of strand e within each sequence group is best justified by secondary structural predictions. Although secondary structure predictions of the core $(\beta\alpha\beta)_2$ correspond nicely to both the predicted and the experimentally determined PII secondary structures, predictions of the extreme C-terminus (strands e and f) diverge. The final β-strand (strand f) is absent from all group predictions, including those based on the PII sequences. The β-strand (strand e) providing essential interactions for trimer formation is also predicted with less reliability. The predictions for group I and group V include this strand, whereas predictions for group III and group IV do not include this strand. Although these individual secondary structural predictions vary, one component of JPRED (Pred) predicts strand e when given the global multiple-sequence alignment as input (Fig. 2). In addition, several residues become buried on interactions of strand d with strand e in PII trimer formation (I_{91}, V_{93}, and V_{99}). Correspondingly, the positions of these three residues remain quite hydrophobic throughout the PII superfamily multiple-sequence alignment (yellow highlights, Fig. 2).

The second structural element important for trimer formation is the bend in the second β-strand (strand b). The β-bulge shown in the multiple-sequence alignment (indicated by insertions, strand b, Fig. 2) is consistent with a bend in this strand. This bulge is required to optimize the alignment of sequences in groups I, II, III, and V. One residue positioned in strand b (T_{29}) becomes buried on trimer formation, lining up with a residue (G_{35}) from the neighboring monomer strand b. These residues also line the back of the ATP-binding pocket, providing essential interactions with the adenine base and sugar. The first residue (T_{29}) is almost invariant in several PII groups (groups I, II, and V), with conservative residue replacements in the remaining groups (C_{29} in group III and S_{29} in group IV). The second residue (G_{35}) is the most conserved position in the PII superfamily and provides the 1% sequence identity observed between the most distant superfamily members. Thus, the overlap between structural (trimer formation) and functional (ligand-binding site establishment) requirements of these two residues is

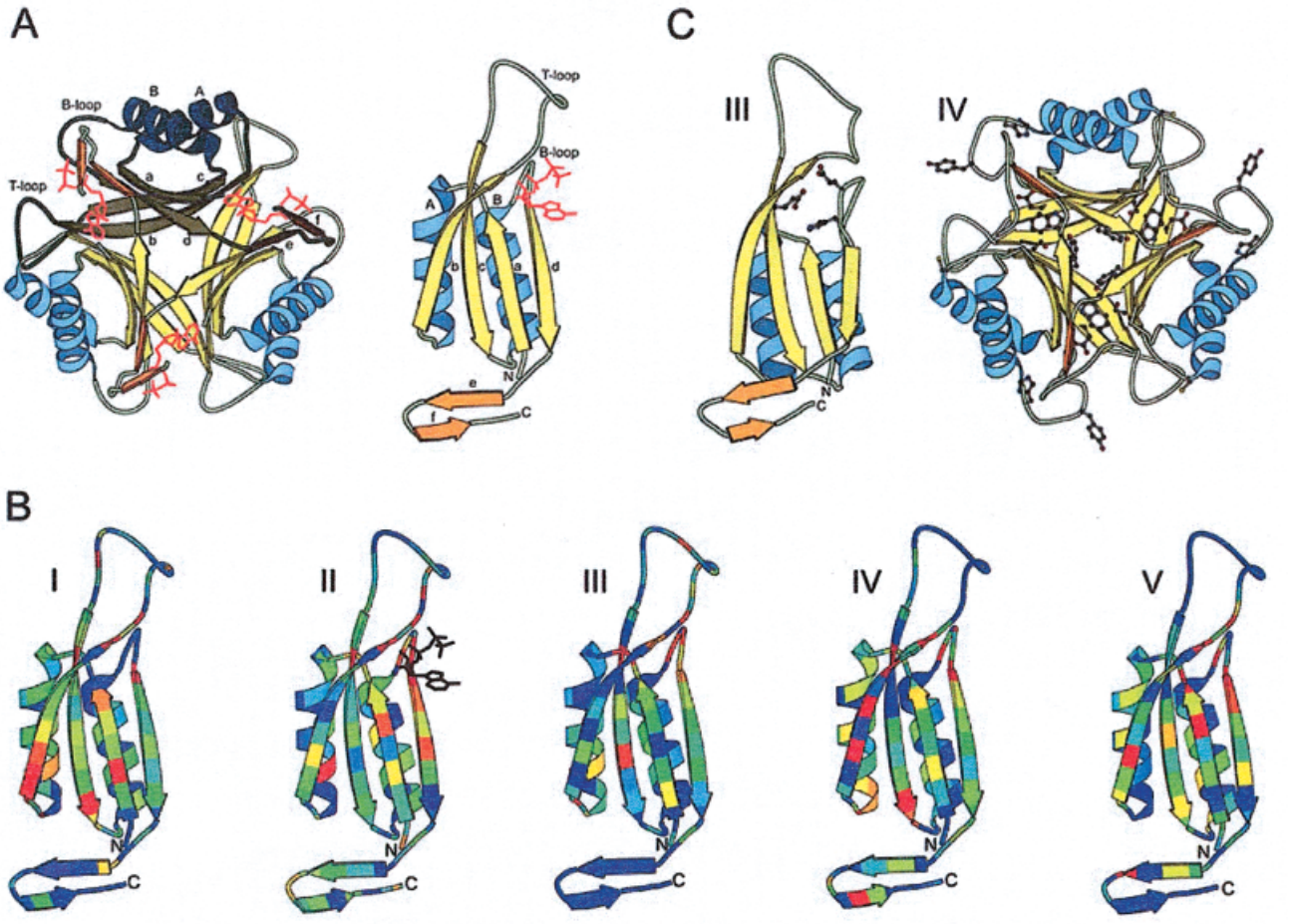

Fig. 4. Structural diagrams of the nitrogen regulatory protein superfamily. The structural diagrams were produced with the program BOBSCRIPT[32] based on the *E. coli* PII protein structure (PDB id 2PII) and the *E. coli* GlnK ATP ligand (PDB id 2GNK). **A:** Ribbon diagrams of the PII trimer and monomer are labeled according to secondary structural elements with lowercase (β-sheet) and uppercase (α-helix) letters. One monomeric subunit of the trimer is shaded to help illustrate interface interactions described in the text. The β-sheets and α-helices of the core ferridoxin $(\beta\alpha\beta)^2$ fold are colored yellow and blue, respectively; with the two C-terminal β-strands colored orange. ATP is displayed in red lines. **B:** Conservations of the PII superfamily groups (I–V) are projected onto the PII ribbon diagram. Conservations were calculated from individual group alignments (excluding subgroups) by using the program AL2CO[33] and projected onto the corresponding residues of the PII structure. Conserved residues are depicted in a rainbow scale from red (most conserved) to blue (least conserved). The ATP ligand is in black lines (group II). C. Fold predictions of the group III monomer and the group IV trimer are shown. Corresponding residues of the PII structure were replaced with group III or group IV residues using the Biopolymer module of the InsightII graphics package. Conserved side-chains discussed in the text are represented as ball and stick.

reflected in their conservation throughout the PII superfamily.

In addition to the conservation of these two structural elements required for the PII quaternary structure, the domain organization found in multi-domain members of the superfamily is also consistent with trimer formation. The N- and C-termini map to one surface of the assembled trimer, whereas residues thought to be important for PII protein-effector interactions (contained in B- and T-loops) and the ligand-binding site map to the other side. Thus, the unusual group II sequence that contains a second PII core $(\beta\alpha\beta)_2$ domain fused to its N-terminus is compatible with the formation of a trimer that retains PII function. Trimer formation also allows for an appropriate spatial arrangement of the N-terminal membrane spanning segment of the group V sequences, and the N- and C-terminal extensions of the group III sequences. In such an arrangement, additional domains would not be predicted to interfere with either small molecule binding or protein effector interactions.

Ligand-binding pocket

PII protein sequences (excluding group IIa and group IIb sequences, Fig. 2) retain a remarkable degree of sequence similarity. Several small residues are conserved across the length of the sequence (Fig. 2, gray highlights). Many of these residues line the ATP-binding pocket, providing van der Waals contacts with the adenine sugar and base (G_{27} and $G_{35}XG_{37}$) and forming hydrogen bonds with the ATP phosphate and sugar oxygens ($G_{87}XG_{89}$, and T_{29}, respectively), whereas others maintain loop structures ($T_{83}G$ in the B-loop).[11] Conservations of a subset of these small, hydrophobic residues extend across the entire superfamily (T_{29}, G_{35}, and G_{89}). PII sequences also include conserved

charged and large hydrophobic residues (Fig. 2, black highlights). Several of these residues line the ATP-binding pocket (F_{92}, K_{58}, K_{90}), whereas others maintain loop conformations (D_{88}, K_{58}, K_9).[11] In contrast to most residues that provide ATP contacts, the strand e residues (R_{101} and R_{103}) that provide hydrogen bonds with the gamma phosphate of ATP are not well conserved among PII sequences [Fig. 4(B)]. As illustrated in the multiple-sequence alignment, the conservation of these mainly charged PII residues involved in binding ATP and maintaining the loop structures do not extend to other groups of the PII superfamily, including the closely related sequences (groups IIa and IIb) that cluster with the PII proteins in distance groupings. This lack of specific residue conservation suggests that proteins belonging to other superfamily groups do not bind ATP.

Despite the loss of ATP-binding capacity, the conservation patterns of the remaining groups of the PII superfamily suggest a preservation of the ligand-binding site. Figure 4(B) illustrates the sequence alignment conservations of each main group (I–V) projected onto the PII structure in a rainbow color scheme, with red being the most conserved residue positions and blue being the least conserved residue positions. It is of interest that many of the conservations contained within the different groups map to the same positions. For example, the projected conservations of the group V sequences almost replicate the PII protein conservations, especially surrounding the ligand-binding pocket. Some of these conserved positions are also retained in the group I sequences. These group I and group V sequence conservations are limited to mainly hydrophobic residues, making functional interpretations difficult. Alternatively, the group III and group IV conservations contain many invariant charged residues, allowing for more insight into possible functions.

Group III and group IV fold prediction: potential metal ion coordination

The conserved motif ($_{83}$HPYEXP$_{89}$) possessed by both group III and group IV maps to the B-loop of the PII structure. The close spatial proximity of another conserved residue (E_{58}) to the residues found in this motif provides potential candidates for metal ion coordination. Although the differences between the PII B-loop and the group III B-loop composition and length make precise structural predictions difficult, the fold prediction model for the group III sequences [Fig. 4(C)] shows a possible transformation of the ATP-binding site of the PII proteins (residues K_{58}, T_{83}, and K_{90}) to a metal-binding site (residues E_{58}, H_{83}, and E_{89}).

Although inspection of conserved residues in the predicted group III structure suggests a possible metal-binding site, the biological role of this protein remains unknown. Often, functional information can be discerned from the presence of domain fusions[20,21] or from the conserved spatial clustering of sequences in genomes termed "functional coupling."[22] The second domain found in several of the group III sequences corresponds to COG0327. Although this COG is also classified as "uncharacterized," it contains a yeast gene (NIF3) whose encoded protein interactions have been determined in genomewide two-hybrid screens.[23,24] The yeast NIF3p interacts with a nuclear inport/export protein (Srp1p) and a ras-like GTPase (Tem1p), which are both required for proper exit from mitosis in the cell cycle.[25,26] Another member of COG0347 contains an N-terminal SWIB domain involved in the regulation of chromatin structure. Because domain fusions often imply functional interactions of the individual proteins,[20,21] the group III sequences may also be involved in regulating chromatin structure in the bacterial cell cycle. Ortholog chromosome clustering available on the EGRO database (http://wit.integratedgenomics.com/ERGO/) defines conserved gene clusters that have been shown to convey functional coupling between genes present in these clusters.[22] Such analysis clusters these genes to those of several nucleic acid-binding proteins, including RNA polymerase sigma factor rpoD (4.96), endonuclease IV (3.82), and DNA primase dnaG (3.76) with coupling scores of high significance (scores > 1).[22]

Linked to the PII sequences only through group III, the group IV sequences appear to deviate from the rest of the superfamily. The most divergent secondary structure predictions occur for the group IV sequences. Although the core $(\beta\alpha\beta)_2$ domain is present, consensus predictions contain an additional helix at the C-terminus, possibly replacing the last β-strand (strand f) of PII. The predicted length of the first α-helix (helix A) is longer than its corresponding helix in the PII structure, and its hydrophobicity pattern differs. These altered properties make both the global alignment and the fold prediction of the region less reliable than that of other regions. We based the placement of the group IV helix in the final global alignment on the overall conservation of residues within the group. The results of fold recognition and the alignment of the conserved group IV Cys with the conserved Gly also support this placement. The altered hydrophobicity pattern of this helix suggests that it packs against the rest of the protein differently than the corresponding PII helix. Inspection of the PII structure does indicate that this altered packing can be tolerated, because helix A does not contribute directly to either trimer formation or the ligand-binding pocket.

The strongest sequence similarity between the group IV sequences and the rest of the PII superfamily is the presence of the conserved motif ($_{83}$HPYEXP$_{89}$) downstream from the conserved residue (E_{58}) proposed to bind metal ions in the group III fold predictions. However, one of the potential coordinating residues (E_{89}) is not conserved in the group IV sequences, leading to the possibility that the preceding residue (Y_{88}) provides the third coordinating group. Alternatively, the group IV sequences may bind another ligand. As can be seen in the conservation projection [Fig. 4(B); group IV), conserved residues fall within the ligand-binding pocket. These residue sidechains are compared with those of the PII structure in the group IV fold prediction model [Fig. 4(C)]. Another striking feature of this model is the conserved residue (Y_{31}) within the central channel of the trimer cavity [Fig. 4(C)]. The group III sequences also include conserved residues

lining the channel (K_3 and E_{62}, not shown). When considered in conjunction with the presence of ordered water molecules lining the central cavity of the PII structure, these conservations leave open the possibility of the group III or the group IV trimers functioning as some sort of channels. Genetic evidence provides further support for group IV binding of metal ions. Elimination of the gene locus (*cutA*) of one member of the group IV sequences (gi|729269) from *E. coli* leads to sensitivity to divalent metal ions including copper, zinc, nickel, and cobalt.[27] The *CutA* gene is functionally coupled to thiol disulfide isomerase *dsbD* (4.41) and C4-dicarboxlyate transporter *dcuA* (0.59) in bacteria (http://wit.integratedgenomics.com/ERGO/) and is suggested to localize with acetylcholinesterase at the surface of mammalian cells.[28]

CONCLUSIONS

As is seen with PSI-BLAST detection of other diverse superfamilies,[29] no individual query sequence detects all members of the emerging PII superfamily. Therefore, sequence-searching strategies appear to be extremely important in the detection of distant homologs. The inclusion of increasing numbers of sequences to generate greater diversity, the production of accurate multiple alignments of these sequences, and grouping these sequences into similar clusters become essential elements of detection. The most diverse members of the PII superfamily retain only 1% sequence identity over the entire domain (about 100 aa). Such diversity suggests not only an equally varied set of functions for the different PII superfamily groups but a lack of precise sequence requirements for the overall fold. The main structural requirements for the PII protein fold include strands necessary for trimer formation and conservations of hydrophobicity found within secondary structural elements. This same diversity can be found in other protein folds with sequence-detected homology. The kelch repeat β-propeller proteins (kelch, galactose oxidase, nuclear protein HCF, and Rag-2) assume a variety of functions, interacting with many different proteins and small molecules. Like the PII superfamily sequences, the most striking conservations found among the folding units of these proteins are restricted to patterns of hydrophobicity.[29,30] Proteins with TIM barrel folds also perform a wide range of functions. Evolutionary links have been detected between many members of this large fold group.[31] Like the proposed PII protein superfamily, these proteins use particular backbone positions for functionally important residues and retain similar active site positions.

In summary, we were able to unify five COGs and several small divergent protein groups into a large and very diverse superfamily, clarifying their evolutionary history and making functional preditions. This analysis pushes the limits of sequence homology detection and is likely to exemplify one of the extreme cases of reliable sequence-based structure prediction. The comprehensive homology detection procedure described here is an in silico equivalent to the structural genomic efforts. Structure predicted with high confidence becomes a low-resolution reflection of the true structure, which can be successfully used in functional predictions. Our strategies make it possible to detect remote sequence homology and can be broadly used in protein classification schemes and for structural genomics target selection refinement.

REFERENCES

1. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1007;25:3389–3402.
2. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 1994;235:1501–1531.
3. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
4. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
5. Fischer D. Hybrid fold recognition, combining sequence derived properties with evolutionary imformation. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, editors. Pacific Symp on Biocomputing. Hawaii: World Scientific; 2000. p 119–130.
6. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28:33–36.
7. Tatusov RL, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–28.
8. Arcondeguy T, Jack R, Merrick M. P(II) signal transduction proteins, pivotal players in microbial nitrogen control. Microbiol Mol Biol Rev 2001;65:80–105.
9. Ninfa AJ, Atkinson MR. PII signal transduction proteins. Trends Microbiol 2000;8:172–179.
10. Vasudevan SG, et al. Escherichia coli PII protein: purification, crystallization and oligomeric structure. FEBS Lett 1994;337:255–258.
11. Xu Y, et al. GlnK, a PII-homologue: structure reveals ATP binding site and indicates how the T-loops may be involved in molecular recognition. J Mol Biol 1998;282:149–165.
12. Cheah E, Carr PD, Suffolk PM, Vasudevan SG, Dixon NE, Ollis DL. Structure of the Escherichia coli signal transducing protein PII. Structure 1994;2:981–990.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
14. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thorton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
15. Pei J, Grishin NV. GGDEF domain is homologous to adenylyl cyclase. Proteins 2001;42:210–216.
16. Walker DR, Koonin EV. SEALS: a system for easy analysis of lots of sequences. Proc Int Conf Intell Syst Mol Biol 1997;5:333–339
17. 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.
18. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14:892–893
19. Forster M, Heath A, Afzal M. Application of distance geometry to 3D visualization of sequence relationships. Bioinformatics 1999;15: 89–90.
20. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
21. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.
22. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999;96:2896–2901.
23. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98:4569–4574.
24. Uetz P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623–627.

25. Shirayama M, Matsui Y, Toh EA. The yeast TEM1 gene, which encodes a GTP-binding protein, is involved in termination of M phase. Mol Cell Biol 1884;14:7476–7482.
26. Loeb JD, Schlenstedt G, Pellman D, Kornitzer D, Silver PA, Fink GR. The yeast nuclear import receptor is required for mitosis. Proc Natl Acad Sci USA 1995;92:7647–7651.
27. Fong ST, Camakaris J, Lee BT. Molecular genetics of a chromosomal locus involved in copper tolerance in Escherichia coli K-12. Mol Microbiol 1995;15:1127–1137.
28. Perrier AL, et al. Two distinct proteins are associated with tetrameric acetylcholinesterase on the cell surface. J Biol Chem 2000;275:34260–34265.
29. 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.
30. Paoli M. Protein folds propelled by diversity. Prog Biophys Mol Biol 2001;76:103–130.
31. Copley RR, Bork P. Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. J Mol Biol 2000;303:627–641.
32. Esnouf RM. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. J Mol Graph Model 1997;15:132–134, 112–133.
33. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700–712.