Article No. jmbi.1999.2653 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 287, 1023–1040

JMB

Gleaning Non-trivial Structural, Functional and Evolutionary Information About Proteins by Iterative Database Searches

L. Aravind[1,2] and Eugene V. Koonin[1]*

[1]*National Center for Biotechnology Information National Library of Medicine National Institutes of Health Bethesda, MD 20894, USA*

[2]*Department of Biology, Texas A&M University, College Station, TX 77843, USA*

**Corresponding author*

Using a number of diverse protein families as test cases, we investigate the ability of the recently developed iterative sequence database search method, PSI-BLAST, to identify subtle relationships between proteins that originally have been deemed detectable only at the level of structure-structure comparison. We show that PSI-BLAST can detect many, though not all, of such relationships, but the success critically depends on the optimal choice of the query sequence used to initiate the search. Generally, there is a correlation between the diversity of the sequences detected in the first pass of database screening and the ability of a given query to detect subtle relationships in subsequent iterations. Accordingly, a thorough analysis of protein superfamilies at the sequence level is necessary in order to maximize the chances of gleaning non-trivial structural and functional inferences, as opposed to a single search, initiated, for example, with the sequence of a protein whose structure is available. This strategy is illustrated by several findings, each of which involves an unexpected structural prediction: (i) a number of previously undetected proteins with the HSP70-actin fold are identified, including a highly conserved and nearly ubiquitous family of metal-dependent proteases (typified by bacterial O-sialoglycoprotease) that represent an adaptation of this fold to a new type of enzymatic activity; (ii) we show that, contrary to the previous conclusions, ATP-dependent and NAD-dependent DNA ligases are confidently predicted to possess the same fold; (iii) the C-terminal domain of 3-phosphoglycerate dehydrogenase, which binds serine and is involved in allosteric regulation of the enzyme activity, is shown to typify a new superfamily of ligand-binding, regulatory domains found primarily in enzymes and regulators of amino acid and purine metabolism; (iv) the immunoglobulin-like DNA-binding domain previously identified in the structures of transcription factors NFκB and NFAT is shown to be a member of a distinct superfamily of intracellular and extracellular domains with the immunoglobulin fold; and (v) the Rag-2 subunit of the V-D-J recombinase is shown to contain a kelch-type β-propeller domain which rules out its evolutionary relationship with bacterial transposases.

Keywords: iterative database search; PSI-BLAST; structure prediction; DNA ligase; sialoglycoprotease

Introduction

Protein structure determination inevitably lags far behind the explosive quantitative and qualitative (thanks to the determination of genome sequences of taxonomicaly diverse organisms) growth of sequence databases. It has been observed, however, that newly determined structures increasingly tend to fall into already known

Abbreviations used: OSGP, O-sialoglycoprotease; MCE, mRNA capping enzyme; ACT, aspartokinase, chorismate mutase and TyrA; TIG, transcription factor IG; NR, non-redundant.

E-mail address of the corresponding author: koonin@ncbi.nlm.nih.gov

0022-2836/99/151023–18 $30.00/0

structural folds (Murzin, 1996, 1998). This indicates that the number of folds (the basic types of globular domains) is finite and is unlikely to exceed a few thousand (Chothia, 1992; Orengo *et al.*, 1994). Moreover, while it is difficult to estimate the total number of folds with a greater precision, it seems clear that for most of the widespread folds, representative structures are already available. Thus, it is highly probable that for any new protein sequence that does not have a significant compositional bias and, accordingly, is likely to form a globular domain(s) (Wootton, 1994), a structure with the same fold is present in the protein data bank (PDB; Bernstein *et al.*, 1977). In order to obtain structural information about a given protein domain, all one needs is to establish a reliable alignment with the sequence of one of the domains with a known structure. More frequently than not, however, this task is not trivial. Major transitions in the evolution of life appear to have been accompanied (or in part driven) by the origin of new protein families from preexisting ones when sequences rapidly diverge, while the structure remains basically conserved (Doolittle, 1995). This erosion of sequence information in the course of evolution is the major obstacle in making structural predictions using homology inferred from sequence similarity. Accordingly, a number of unexpected connections between protein families originally thought to be unrelated have been recently established by comparison of experimentally determined three-dimensional structures (Holm & Sander, 1996, 1997; Murzin, 1996, 1998; Murzin & Bateman, 1997).

In order to maximize the rate of structural prediction from protein sequences, increasing sensitivity of sequence comparison methods is critical. The subtle relationships discovered by structure-structure comparison may be considered the golden standard for sequence analysis methods. Those methods that are sufficiently powerful to detect at least some of the connections originally perceived as ''structural only'' should be expected to routinely produce non-trivial structural predictions. Most of the advanced sequence database search methods utilize information contained in multiple alignments. The recently developed PSI (Position-Specific Iterating)-BLAST method constructs a multiple alignment from the BLAST hits, converts it into a position-specific weight matrix and iterates the search using this matrix as the query (Altschul *et al.*, 1997; Altschul & Koonin, 1998). Several in-depth studies of protein families as well as benchmarking experiments suggest that given the new level of protein sequence diversity coming from whole genome sequencing, this method may significantly increase our ability to detect subtle sequence similarities and, in particular, to make non-trivial structure predictions (Aravind & Koonin, 1998; Aravind *et al.*, 1998; Huyney *et al.*, 1998; Mushegian *et al.*, 1997; Rychlewski *et al.*, 1998; Wolf *et al.*, 1999).

Here, using several previously described cases of relationships between protein families that have been deemed to be detectable only by structure-structure comparison, we show that with appropriate starting points, PSI-BLAST is capable of detecting, at the sequence level, many of these subtle similarities. We demonstrate that typically, the best starting points for the iterative search are those that produce the greatest diversity of hits in the first BLAST pass. We then investigate several new examples of unexpected structural inferences for highly conserved protein domains that have important functional and evolutionary implications.

Results and Discussion

The strategy for protein superfamily analysis using PSI-BLAST

For assessing the ability of PSI-BLAST to detect subtle similarity between proteins, we chose several cases where a relationship originally has been discovered by structure-structure comparison and has been deemed undetectable at the sequence level (Table 1). The examples include the classical case of structural similarity between actins, the HSP70 class of molecular chaperones and sugar kinases (Bork *et al.*, 1992), as well as more recently described relationships, such as those between antibiotic nucleotidyltransferases and DNA polymerase β (Holm & Sander, 1995; Aravind & Koonin, 1999), ClpP protease and enoyl dehydratase (Murzin, 1998), and the nicking-rejoining domains of type I and type II DNA topoisomerases (Berger *et al.*, 1998). A detailed examination of these examples showed that PSI-BLAST detects many, though not all, relationships originally thought to be tractable only at the structural level (Table 1).

This analysis also highlighted a major problem that must be taken into account in order to optimize the detection of protein superfamilies with iterative database search methods, such as PSI-BLAST. The position-specific weight matrix in PSI-BLAST is built using a single starting query sequence, and detailed analysis of the examples included in Table 1, as well as a number of other cases (L.A. and E.V.K., unpublished observations), indicates that the results dramatically depend on the choice of the query. In each investigated example, some queries were much more effective than others in the delineation of the respective superfamily by PSI-BLAST searches (Table 1). Furthermore, for certain superfamilies, particularly large ones such as the HSP70-actin-like enzymes and the nucleotidyltransferases, no single query could retrieve all members. Accordingly, the results produced with different queries had to be combined in order to fully characterize the respective superfamily. So far, there is no rigorous criterion to determine the effectiveness or quality of a query sequence. We noticed, however, an intui-

Table 1. Detection of protein superfamilies using iterative database search (PSI-BLAST) and different queries

Protein or domain superfamily defined by structure comparison (references)	Representatives, including proteins with known structure and ''structural only'' similarity to each other[a]	Number of hits with e<0.01 in first pass (e-value of last hit)	Number of iterations for detection of maximal representative diversity	Number of clusters formed by hits[d]	Diversity in terms of families within the superfamily detected by the given query
ClpP/enoyl dehydratase (Murzin, 1998)	ClpP protease[b] (116523)	50 (10^{-7})	3	2	Four families: ClpP, protease IV, enoyl dehydratase, and Acetyl coA carboxylase. All these families are detected by each of the queries
	Enoyl-CoA dehydratase (1dub)	124 (0.001)	5	3	
	Protease IV (2826302)	38 (10^{-5})	2	4	
Toprim (catalytic domain of topoisomerases and primases) (Aravind *et al.*, 1998)	Topoisomerase IA (1ecl)	43 (0.007)	C3[c]	1	Five families: topoisomerase IA, topoisomerase II, primases, small primase-like proteins and Old family nuclease. Representatives of all families are detected only by BB0626
	Topoisomerase II (1bgw)	198 (0.004)	C2	1	
	Small, primase-like protein BB0626 from *Borrelia burgdorferi* (2688557)	**6 (0.004)**	4	**2**	
ATP Grasp (Galperin & Koonin, 1997)	Synapsin (1aux)	29 (10^{-41})	C1	1	*ca* Ten families. The DD ligase starting point detects all families, for which sequence conservation could be shown by detailed comparative analysis (Galperin & Koonin, 1997)
	D-alanine D-alanine ligase (1iow)	**104 (0.004)**	**4**	**4**	
	Biotin carboxylase (3328523) from *Chlamydia trachomatis*	124 (0.008)	6 (does not recover all families, eg synapsin)	2	
Polβ-type nucleotidyl transferases (Holm & Sander, 1995; Aravind & Koonin, 1999)	Kanamycin nucleotidyl transferase (1kan)	8 (10^{-19})	C1	1	Nine families. None of the queries tried so far detects the entire superfamily; this requires transitive searches with several queries. YOL115w is the best query which detects 4 families.
	Polymerase β (1bpe)	37 (0.009)	C2	1	
	Putative yeast nucleotidyl transferase YOL115w (1077298) (L.A. and E.V.K., unpublished observations)	**15 (0.005)**	**C6**	**2**	
HSP70/actin domain (duplicated RnaseH module) (Bork *et al.*, 1992; Reizer *et al.*, 1993; Koonin, 1994; see also Figures 1 and 2)	Hexokinase (1hkb)	35 (0.008)	C3	1	At least 12 families all of which could not be detected by any single query (see text). The best query is actin which detected four families as opposed to two detected by DnaK and one by hexokinase.
	Actin (1atn)	**515 (10^{-25})**	**6**	**1, 3 (at iteration 3)[e]**	
	DnaK (1dkg)	659 (10^{-30})	C5	1, 2 (at iteration 3)[e]	
Double-stranded β-helix (Gane *et al.*, 1998)	Vicilin (1cax)	83 (0.006)	C4	3	Several distinct families many of which have lost the characteristic histidine residues (L.A., M. Y. Galperin and E.V.K., unpublished observations). Representatives of all these families are detected by CurC whereas AraC and Vicilin at convergence detect only members of their own families.
	AraC (2aac)	8 (10^{-43})	C1	1	
	CurC(729227)	**16 (0.008)**	**4**	**6**	

[a] The PDB code (whenever available) or the Gene Identification number in the NR database is indicated in parentheses; the best query is shown by bold type.
[b] The structure of ClpP was used for this comparison (Murzin, 1998), although it is not yet in PDB.
[c] Cn indicates convergence after n iterations
[d] The sequences with e-values <0.01 at the first BLAST pass were retrieved from the NR database and single-linkage clustering by sequence similarity was performed using the GROUPER script of the SEALS package (the cut-off for clustering was adjusted individually for each superfamily and was in the range of 0.5-0.75 bit/position)
[e] In these searches, the diversity of the hits appears only in the third iteration.

tively plausible, positive correlation between the diversity among the sequences retrieved in the first pass (or less frequently in a subsequent iteration) and the success in retrieval of the superfamily members by a given query in the iterative search. In each of the examples, the sequences whose hits formed the greatest number of distinct clusters fared best in terms of the recovery of the entire superfamily using PSI-BLAST (Table 1). Thus it appears that the optimal strategy for protein superfamily analysis using PSI-BLAST should include either an exhaustive iterative search using all known members or, particularly in the case of large superfamilies, the selection of a set of queries that are likely to perform well on the basis of the diversity criterion.

It should be emphasized that the problem of choosing the optimal query for iterative database searching is completely different from the well-known problem of weighting sequences prior to position-dependent matrix construction. When the query used for a search is a member of a large family within a protein superfamily, weighting is critical in order to avoid skewing the matrix towards this particular family (Vingron & Sibbald, 1993). A simple but apparently effective weighting procedure is incorporated in the PSI-BLAST program (Altschul *et al.*, 1997; Henikoff & Henikoff, 1994). However, the only thing any weighting scheme can do is reducing the bias in the data; it cannot increase the ultimate diversity of the data that seems to determine, at least in part, the searching power of the constructed matrix (Table 1).

The issue of the choice of the optimal query(s) is particularly relevant when structure prediction using iterative database search with PSI-BLAST is considered. Indeed, the observations discussed above make it clear that starting the search with the sequence of a protein whose structure is available is not necessarily the best way to detect subtle structural relationships. Nor is it sufficient to start the search with the sequence of a structurally uncharacterized protein and look for similarity to a sequence from the PDB. This is exemplified by the case of synapsin whose relationship to the ATP-grasp enzyme superfamily was originally detected by structure comparison (Esser *et al.*, 1998). A search initiated with the synapsin sequence does not detect any non-trivial relationships. In contrast, several sequences of ATP-grasp proteins readily retrieve the synapsin sequence from the NR database at a statistically highly significant level (Table 1).

Thus the use of optimal starting points, at least in some case, has the potential to significantly increase the chance of revealing relationships between structurally uncharacterized protein families and known structures represented in the PDB. Below we describe the application of this strategy to several biologically interesting protein superfamilies; these examples further illustrate the potential of the iterative sequence search in detecting non-trivial relationships between proteins and the importance of the optimal selection of starting points.

Non-trivial structural inferences from iterative sequence database searches

New families within the HSP70-actin fold

This fold includes a number of ATP-dependent enzymes, some of which possess molecular chaperone and other additional activities, such as the nearly ubiquitous chaperone Hsp70, the eukaryotic cytoskeletal protein actin, sugar kinases, phosphatases (e.g. *Escherichia coli* Ppx) and proteins whose exact activities are not known, e.g. bacterial cell cycle proteins MreB and FtsA (Bork *et al.*, 1992; Reizer *et al.*, 1993; Koonin, 1994). Our analysis using selected starting points identified a number of new proteins that are predicted to possess the HSP70-actin fold, some of which are highly conserved in taxonomically diverse species (Table 1 and Figure 1). The most notable of these are the O-sialoglycoproteases (OSGPs). This protein family is represented in all organisms whose genomes have been sequenced so far, and the *E. coli* and *Bacillus subtilis* OSGPs are essential for bacterial growth (Arigoni *et al.*, 1998). OSGPs possess metal-dependent protease activity (Abdullah *et al.*, 1992; Mellors & Lo, 1995). Consistent with this, they contain a highly conserved histidine residue dyad which is typical of metal coordination sites of other metal-dependent proteases. A PSI-BLAST search started with most of the OSGP sequences converged after retrieving the members of this family. In contrast, a search started with the protein AF1959 (gi:2648583, Figure 1), the predicted *R*-hydroxyglutaryl-CoA dehydratase activator from the archaeon *Archaeoglobus fulgidus*, retrieved the first representative of the OSGP family in the second iteration with an *e*-value of 2×10^{-6} and most of the known and several new (see below) protein families that possess the HSP70/DnaK fold in subsequent iterations; this search did not produce any obvious false positives. An additional test using the ZEGA method predicted that OSGP and DnaK have the same fold with a *p*-value of $\sim 10^{-4}$. A multiple alignment analysis using the Gibbs sampling procedure indicated that OSGPs and HSP70 share several conserved motifs with associated probability of chance occurrence in the range of $\sim 10^{-6}$ to 10^{-20} (Figure 1). Mapping of these motifs onto the three-dimensional structure of HSP70 shows that they correspond to highly conserved structural elements of the HSP70/DnaK fold. The basic scaffold of the domain consists of two structurally similar halves which, in the SCOP database, are classified under the ''RNaseH-like fold'' (Hubbard *et al.*, 1999). The ATP molecule is sandwiched between these symmetrically placed halves; this interaction involves two conserved loops flanked on either side by long β-strands. Both these loops and the strands of the ATP

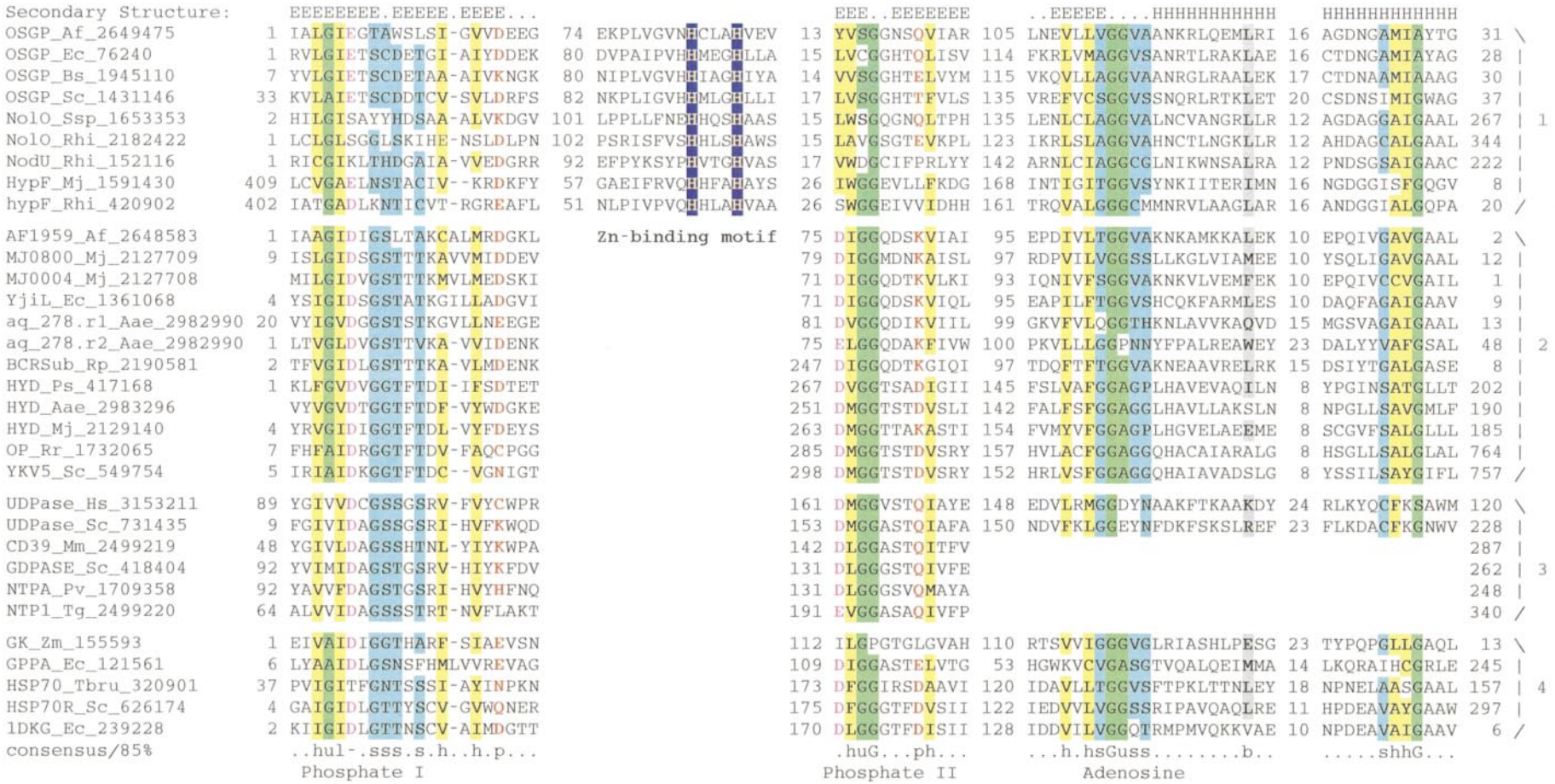

Figure 1. Multiple alignment of the new protein families of the HSP70-actin fold. The alignment was constructed using the Gibbs sampling option of the MACAW program and modified on the basis of the PSI-BLAST search results. The numbers indicate the distances from the protein termini to the proximal and distal aligned blocks, and the distances between the blocks. The sequences are grouped by similarity: 1, OSGP and related proteins; 2, newly identified proteins with the HSP70-actin fold including a benzoyl-CoA reductase (BCR), hydantoinase (HYD) and 5-oxoprolinase (OP) subunits; 3, UDPases and extracellular ATPases (apyrases); 4, classic HSP70 and sugar kinases. The alignment includes only selected, diverse sequences from each of these groups. The shading and coloring of conserved residues is according to the consensus that is shown below the alignment and includes residues conserved in at least 85 % of the aligned sequences; h indicates hydrophobic residues (A, C, F, I, L, M, V, W, Y; yellow background), s indicates small residues (A, C, S, T, D, N, V, G, P; blue background), u indicates ''tiny'' residues (G, A, S, cyan background), b indicates big residues (F, I, L, M, V, W, Y, K, R, E, Q; gray background), p indicates polar residues (D, E, H, K, N, Q, R, S, T; dark red), and (-) indicates negatively charged residues (D, E; magenta). The predicted active-site histidine residues of OSGP are shown in yellow, with dark blue shading. The designation of each protein includes its name (for uncharacterized proteins, the name in the NR database is indicated) followed by the species abbreviation and the Gene Identification number; 1DKG is the PDB code for the X-ray structure of the *E. coli* DnaK protein. The secondary structure elements are shown above the alignment according to the 1DKG structure; E indicates extended conformation (β-strand) and H indicates α-helix. The species abbreviations are: Aae, *A. aeolicus*; Af, *A. fulgidus*; Bs, *B. subtilis*; Ec, *E. coli*; Hs, *Homo sapiens*; Mj, *Methanococcus jannaschii*; Mm, *Mus musculus*; Ps, *Pseudomonas sp.*; Pv, *Phaseolus vulgaris*; Rhi, *Rhizobium sp.*; Rr, *Rattus rattus*; Sc, *S. cerevisiae*; Tg, *Toxoplasma gondii*; Tbru, *Trypanosoma brucei*; Zm, *Zymomonas mobilis*.

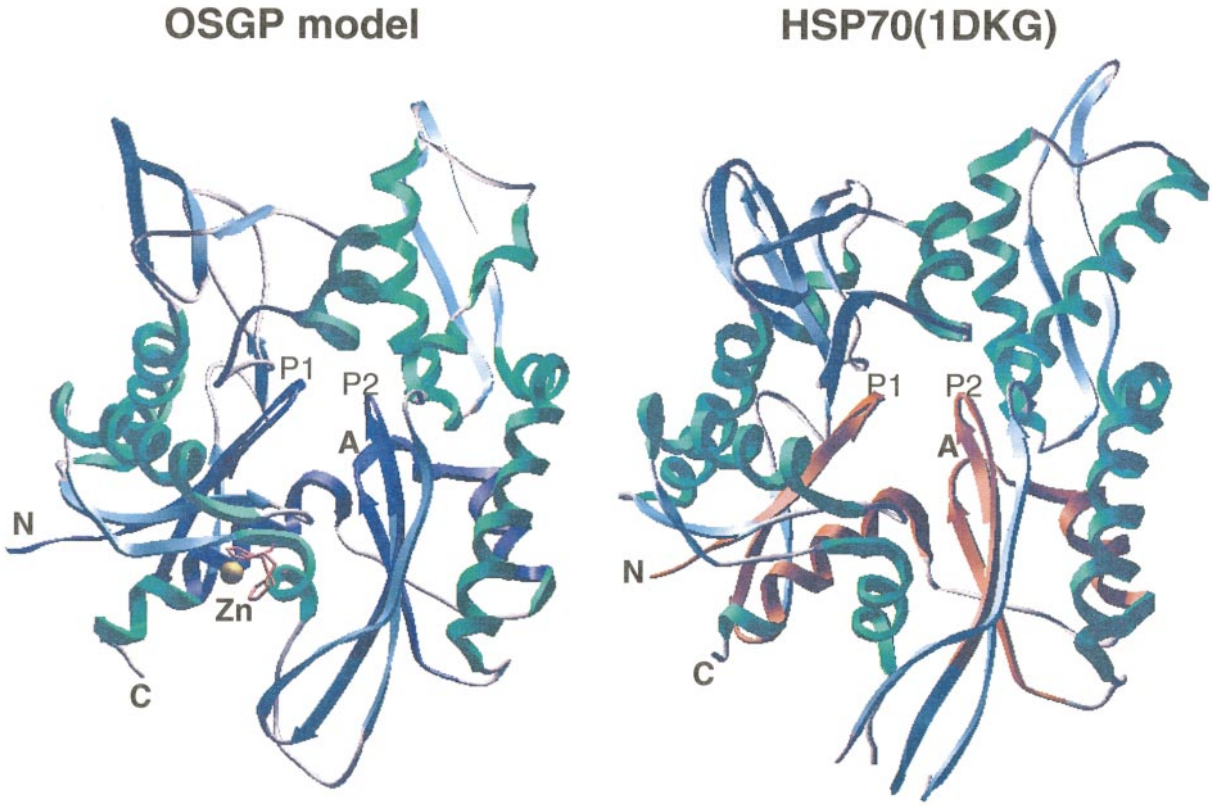

Figure 2. Comparison of the HSP70 structure and a structural model of the O-sialoglycoprotease. The conserved motifs from the alignment in Figure 1 are shown by dark red coloring in the 1DKG structure and by dark magenta coloring in the OSGP model. P1, P2 and A indicate the two phosphate binding site and the adenosine binding site, respectively. In the OSGP model, a Zn atom and two chelating histidine residues are shown.

binding site are conserved in the OSGPs (Figures 1 and 2).

Even a crude model of the OSGP structure generated using the HSP70 structures (1dkg and 1ngc) as templates provides some insight into the possible interaction between the active site of the protease and the ATPase domain. The predicted metal-chelating histidine residue dyad of the protease is located in a helical region which belongs to the linker between the two RNaseH-like halves of the molecule (Figure 2). This part of the molecule has a deep cleft that is predicted to accommodate both the metal atom and the peptide substrate; the protease active site is distinct from the ATP-binding cleft and points away from it (Figure 2). This represents a remarkable adaptation of the HSP70-actin fold to the protease function by grafting a metal-binding motif onto its structural framework. It appears that this motif has evolved by mutation rather than by insertion and, accordingly, the similarity between the active sites of OSGPs and other metal-dependent proteases is purely convergent. These findings suggest that OSGP is an as yet uncharacterized ATP-dependent protease. The conservation of the two phosphate and adenosine binding sites (Figures 1 and 2) suggests that OSGP binds ATP similarly to the HSP70/DnaK fold proteins, which may result in a conformational change that could affect the protease active site. The protein-protein interaction domain of the HSP70/DnaK-class molecular chaperones is located in the C-terminal region of these proteins (Martin & Hartl, 1997) that is not conserved in OSGPs; therefore the prediction of the HSP70 fold may not directly indicate a chaperone function for OSGP. Nevertheless, the obvious analogy to other ATP-dependent proteases, such as the Clp system, Lon and FtsH in bacteria (Gottesman, 1996; Gottesman & Maurizi, 1992), suggests that like these proteases, OSGP may possess a chaperone-type activity. The nearly universal conservation of OSGP and its essential role in bacteria indicate that it has a critical cellular function, perhaps in the ATP-dependent degradation of some classes of misfolded proteins, that remains to be identified experimentally.

The Hsp70-actin superfamily domain of the OSGP-type, with the conserved histidine residue dyad, also occurs in multidomain bacterial and archaeal proteins, such as HypF and NolO (Figure 1). In HypF, the OSGP domain is combined with the small acyl phosphatase domain, Zn fingers and an uncharacterized, conserved ''SUA5'' domain, suggesting that this protein has multiple

activities. As HypF participates in the biogenesis of the hydrogenase complex, which involves proteolytic steps (Colbeau *et al.*, 1998), it seems likely that the OSGP domain contributes to this process both directly as a protease and as a chaperone. The NolO and NodU proteins are carbamoyl transferases (Jabbouri *et al.*, 1998), and their predicted ATP-binding domain may participate in the formation of carbamoyl phosphate, though the role of the predicted protease activity remains unclear.

In addition to OSGP, iterative database searches started with the HSP70-actin fold protein sequences detected significant relationships with the subunits of hydantoinases (Watabe *et al.*, 1992), oxo-prolinases (Ye *et al.*, 1996) and benzoyl-CoA reductases (Gibson *et al.*, 1997). In each of these cases, all the diagnostic motifs and residues typical of the HSP70 domain were conserved, and the respective proteins are predicted to be active ATPases (Figure 1). ATP dependence has been observed for the oxo-prolinases and certain hydantoinases (Ye *et al.*, 1996), which suggests that the HSP70-like domain is indeed the domain involved in ATP utilization by these enzymes. In the case of the benzoyl-CoA reductases, at least one biochemically characterized, purified enzyme shows the requirement of ATP for conformational changes that are involved in its conversion into an active form; furthermore, in the absence of the aromatic substrate, the enzyme behaves as an ATPase (Boll & Fuchs, 1995). This suggests a specific chaperone-like role for the HSP70 fold subunit of this enzyme. Generally, the identification of HSP70-actin fold in subunits of enzymes with diverse activities (Figure 1) seems to indicate that these proteins may be adapted for the previously under-appreciated role of activators/chaperones for specific enzyme systems.

An unusual relationship was observed between the HSP70-actin superfamily and eukaryotic UDPases (Wang & Guidotti, 1998) and secreted NTPases (apyrases), such as CD39 (Komoszynski & Wojtczak, 1996). These sequences were retrieved in the PSI-BLAST searches and showed striking conservation of the N-terminal phosphate-binding motif of the HSP70 superfamily (Figure 1). The distal parts of these proteins, however, are highly divergent and the counterparts of the motifs that are conserved in the second RNaseH-like lobe of HSP70, in particular the adenosine-binding motif, could not be identified in the apyrases (Figure 1). This may reflect the fact that these enzymes are general nucleotide phosphatases that do not specifically recognize adenosine, and their C-terminal domain has diverged accordingly.

Thus iterative database searches using PSI-BLAST with optimal starting points significantly expanded the HSP70-actin fold superfamily, and showed that this ancient ATPase domain has diversified and adapted in the course of evolution to perform a greater variety of functions than previously suspected.

Unification of ATP-dependent and NAD-dependent DNA ligases

DNA ligases are among the central enzymes of DNA replication and repair. Previous sequence comparisons, as well as structural and functional studies, have led to the conclusion that the ATP-dependent DNA ligases that are seen predominantly in eukaryotes, archaea, some viruses, and only sporadically in bacteria, were unrelated to the NAD-dependent ligases that are ubiquitous in bacteria (Shuman & Schwer, 1995). In contrast, it has been demonstrated that ATP-dependent ligases share several conserved sequence motifs with the mRNA capping enzymes (MCEs) and RNA ligases (Shuman & Schwer, 1995). The latter enzymes are confidently predicted to possess the same fold as the ATP-dependent ligases whose structure has been recently determined (Shuman, 1996; Subramanya *et al.*, 1996). Using PSI-BLAST searches, we demonstrated a statistically significant relationship between the NAD-dependent and ATP-dependent ligases. For example, a search using the *A. fulgidus* ATP-dependent ligase I sequence as the query recovered the *E. coli* NAD-dependent ligase in the 5th iteration with an *e*-value of 10^{-4}. Examination of the multiple alignment of the ligases constructed using the GIBBS sampling procedure showed that all the conserved elements which had unified the ATP-dependent ligases and the capping enzymes are detectable in the NAD-dependent ligases (Figure 3), with the probability of these motifs being detected by chance in the range of 10^{-5} to 10^{-20}.

Mapping of the conserved motifs onto the ATP-dependent ligase and MCE structures shows that both subdomains of the pincer-like structure of the ligases (classified as a version of the ''ATP-grasp'' fold in the SCOP database) are present in the NAD-dependent ligase (data not shown). The ATP molecule is bound at the interface of the two subdomains, and the conservation of both subdomains suggests that NAD is bound in essentially the same fashion. The strongest sequence conservation was seen in the elements corresponding to the active site and the β-strands that form the scaffold of the ligase domain. The lysine residue that covalently binds AMP at the intermediate step of the ligase reaction is bounded by two hydrophobic β-strands, and this arrangement is clearly conserved in both types of ligases (Figure 3). Two highly conserved acidic residues in motifs 2 and 4 that are probably necessary for the catalytic mechanism of the ATP-dependent ligases are also seen in the NAD-dependent ligases (Figure 3). The similarity of the motifs around the catalytic lysine residue in the ATP-dependent and NAD-dependent ligases has been noticed previously, but given the purported absence of other conserved features it has been attributed to convergence (Shuman & Schwer, 1995). The detection of five conserved motifs and the overall similarity that suggests a common fold indicates that two types of ligases

Figure 3. Multiple alignment of ATP-dependent and NAD-dependent DNA ligases, capping enzymes and RNA ligases. The sequence groups are as follows: 1, typical ATP-dependent DNA ligases; 2, a newly identified group of predicted ATP-dependent DNA ligases present in archaea and *A. aeolicus*; 3, newly predicted nucleotidyltransferases with unknown specificity; 4, mRNA capping enzymes; 5, NAD-dependent DNA ligases; 6, RNA ligases. The secondary structure assignments were derived from the X-ray structure of a viral capping enzyme (1CKN). The catalytic lysine residue that covalently binds AMP is shown in yellow, with dark blue shading. Additional species abbreviations: BPT7, bacteriophage T7; VAC, vaccinia virus; Mtu, *Mycobacterium tuberculosis*; Kl, *K. lactis*; Lm, *Leishmania major*; NPV, nuclear polyhedrosis virus; Sp, *Schizosaccharomyces pombe*; ASFV, African swine fever virus; PBCV, *Paramecium bursaria* Chlorella virus; Mge, *Mycoplasma genitalium*. The other designations are as described in the legend to Figure 1.

have evolved from a common ancestor and have a similar catalytic mechanism. The divergence may have largely arisen from the need to accommodate different nucleotide cofactors.

The iterative searches initiated with ATP-dependent ligase sequences also identified a novel family of predicted ATP-dependent ligases in archaea and *Aquifex aeolicus* (Figure 3; Altschul & Koonin, 1998). Further searches started with diffferent members of the ligase superfamily, such as the MCEs, also detected several uncharacterized proteins that are predicted to have the same fold and possess nucleotidyl transferase activity. These include proteins encoded by bacteriophage T4, nuclear polyhedrosis virus and the killer plasmid of *Kluyveromyces lactis* (Figure 3). Finally, the detection of one of such predicted nucleotidyl transferases in *Leishmania* is particularly interesting as this organism shows RNA editing which requires an RNA ligase (Benne, 1993). The newly detected, diverged enzyme of the ligase superfamily may be a candidate for this role.

ACT: a novel ligand-binding domain

Iterative searches seeded with the small subunit of acetolactate synthase (IlvN), an enzyme that catalyzes the synthesis of acetolactate from pyruvate, revealed significant similarity between the core globular domain of this protein and a variety of other proteins and domains, most of which, directly or indirectly, are involved in amino acid and purine metabolism. For example, in a search started with the *E coli* IlvN sequence, an aspartokinase sequence was detected in the third iteration with an *e*-value of $\sim 10^{-3}$, a homoserine dehydrogenase sequence was retrieved in the fourth iteration with an *e*-value of $\sim 10^{-5}$ and chorismate mutase and RelA were detected in the sixth iteration with *e*-values <0.01. The proteins in which this previously unknown domain was identified are: (i) aspartokinases, (ii) chorismate mutases; (iii) prephenate dehydrogenases (TyrA); (iv) prephenate dehydratases; (v) homoserine dehydrogenases; (vi) malate dehydrogenases; (vii) phosphoglycerate dehydrogenases; (viii) phenylalanine and tryptophan-4-monooxygenases; (ix) phosphoribosylformylglycinamidine synthase (PurQ); (x) uridylyl transferase and removing enzyme (GlnD); (xi) GTP pyrophosphokinase/phosphohydrolase (SpoT/RelA); (xii) tyrosine and phenol metabolism operon regulators (TyrR), (xiii) several uncharacterized proteins from archaea, bacteria and plants that contain from one to four copies of this domain (Figure 4). We named this conserved and widespread domain the ACT domain after aspartokinase, chorismate mutase and TyrA.

The structure of 3-phosphoglycerate dehydrogenase (3PGDH) has been solved (Schuller *et al.*, 1995) and provides insight into the structure and functions of the ACT domain. In 3PGDH, ACT is a C-terminal regulatory domain that is well separated from the classic oxidoreductase domain and forms a β-sheet with appressed helices (a version of the ferredoxin fold according to SCOP). This domain binds L-serine which is the final product of the respective pathway and an allosteric regulator of 3PGDH (Grant *et al.*, 1996). The most conserved portion of the ACT domain is the region at the interface between the first strand and the first helix (Figure 4). Mapping of this conserved motif onto the structural model shows that it is likely to be critical for ligand binding (Figure 5). The characteristic glycine residue followed by a hydrophobic residue in the helix are necessary for maintaining the conformation of the strand-helix interface. In the third position N-terminal of this conserved doublet is a small, polar residue which typically, an aspartate or an asparagine residue (Figure 4). The position at the junction between helix 1 and strand 2 of this domain is again occupied by a small polar amino acid, most frequently an asparagine residue. Both of these residues form hydrogen bonds with the ligand (serine residue) in 3PGDH, and site-directed mutagenesis of these positions alleviates the allosteric inhibition by serine (Grant *et al.*, 1996).

These observations suggest a common ligand-binding mode for all ACT domains whereby the ligand is held in the vicinity of the strand-1-helix-1 interface by means of hydrogen bonds with the two conserved polar residues (Figure 5). The distribution of the ACT domain in enzymes is remarkable in that several of them, e.g. 3PGDH, aspartokinase (Patte *et al.*, 1976) and acetolactate synthase (Vyazmensky *et al.*, 1996), are classic examples of allosteric regulation by the end products of the respective pathways. The presence of the ACT domain in several enzymes involved in the metabolism of different amino acids and in the purine metabolism enzyme PurQ is compatible with the hypothesis of a common origin of allosteric regulation in these functionally diverse enzymes. According to such a scenario, a conserved, evolutionarily mobile module was independently fused to a variety of enzymes, which made them susceptible to the regulation by the respective ligands. This fusion model is consistent with the C-terminal location of the ACT domain in most of these enzymes. The presence of the ACT domain in transcriptional regulators of amino acid metabolism, such as TyrR (Pittard, 1996; Wilson *et al.*, 1995), again indicates that this domain has been recruited for the recognition of the respective amino acids by these regulators. The detection of the ACT domain in GlnD and SpoT/RelA is particularly notable because of the role these proteins play in sensing environmental conditions in the regulation of glutamine synthesis and in stringent response, respectively (Rhee *et al.*, 1985; Cashel *et al.*, 1996). It is likely that the catalytic domains of these enzymes are regulated in response to yet unidentified ligands bound by their ACT domains. The uncharacterized proteins that contain single or multiple copies of the ACT domain may be novel

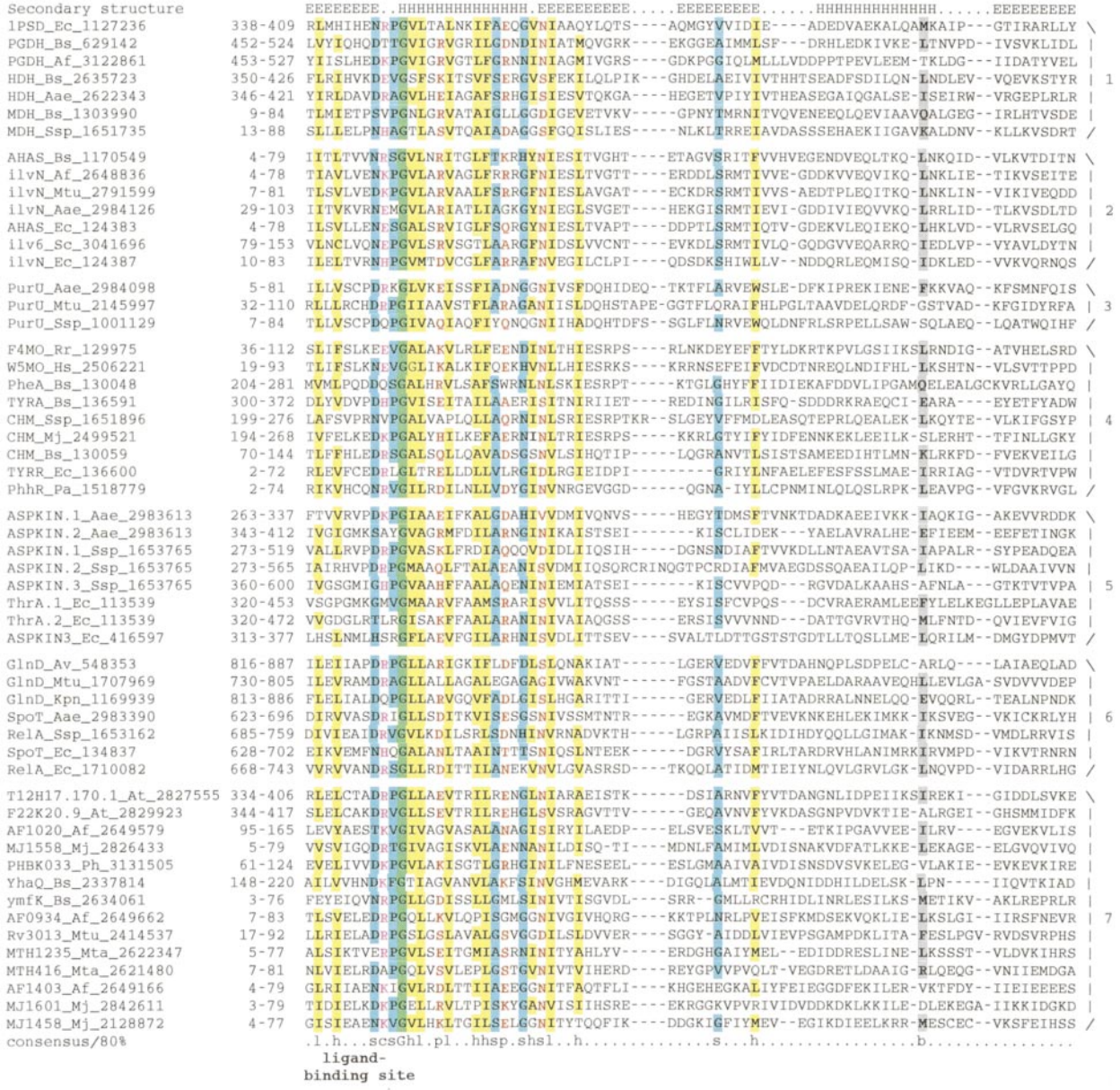

Figure 4. Multiple alignment of the ACT domain superfamily. The sequence groups are as follows: 1, 3-phosphoglycerate dehydrogenase and related dehydrogenases (HDH, homoserine dehydrogenase; and MDH, malate dehydrogenase); 2, small subunits of acetolactate synthases; 3, formyltetrahydrofolate deformylases (PurU enzymes); 4, prephenate dehydratases (PheA, TYRA), chorismate mutases (CHM) and regulators of aromatic amino acid biosynthesis operon expression (TYRR, PhhR); 5, aspartokinases; 6, uridylyl transferases (GlnD), guanosine polyphosphate 3′-pyrophosphohydrolases (SpoT) and GTP pyrophosphokinases (RelA); 7, uncharacterized proteins containing predicted ACT domains. The secondary structure assignments were derived from the 3-phosphoglycerate dehydrogenase X-ray structure (1PSD). The positions of the aligned regions in each of the sequences are indicated by numbers in front of the alignment. The other designations are as described in the legend to Figure 1. Additional species abbreviations: Ssp, *Synechocystis* sp.; Pa, *Pseudomonas aeruginosa*; Av, *Azotobacter vinelandii*; Ph, *Pyrococcus horikoshii*; Mta, *Methanobacterium thermoautotrophicum*.

sensors or regulators that bind specific ligands, primarily amino acids.

The immunoglobulin-like domain in transcription factors (the TIG domain)

The transcription factors of the rel/dorsal/NFκB family have been shown to possess a bipartite DNA binding structure which has two distinct β-strand-rich domains (Ghosh *et al.*, 1995; Muller *et al.*, 1995). The N-terminal domain is a β-barrel similar to that seen in other DNA-binding domains proteins, such as p53, T-box and the STATs (cytochrome F fold in SCOP). The C-terminal domain is an unusual immunoglobulin (Ig) fold domain (Ig superfamily, type E according to SCOP) which we designated the TIG domain, after transcription factor IG. Sequence and structural comparisons

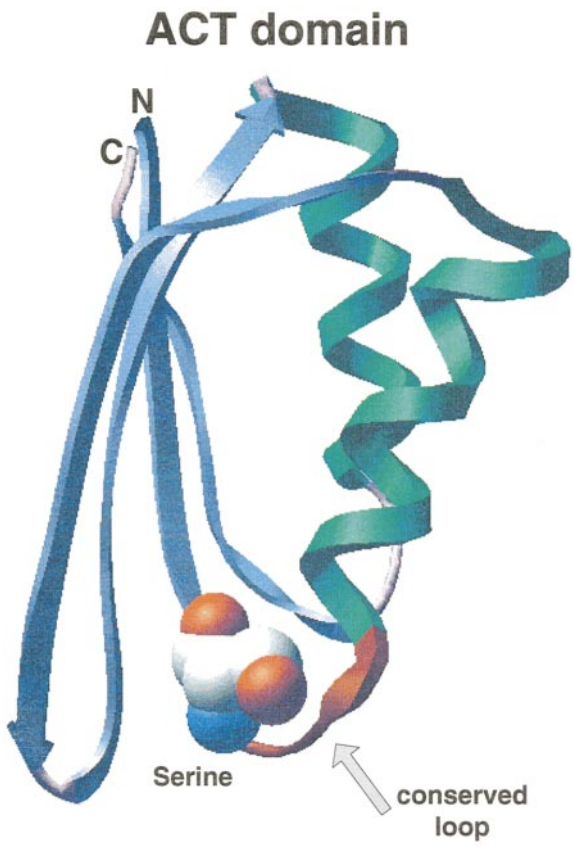

Figure 5. A structural model of the ACT domain with a bound ligand. The domain structure was extracted from the PDB entry for 3-phosphoglycerate dehydrogenase (1PSD). The conserved loop between strand-1 and helix-1 which is the primary determinant of ligand binding is shown in red.

clearly indicate that the NFAT transcription factors possess a two-domain DNA-binding structure similar to that in NFκB (Chen *et al.*, 1998). However, a direct sequence-based identification of relationships of this C-terminal Ig-like domain beyond the obvious connection between NFκB and NFAT has not been reported so far.

In order to investigate the relationships and possible origin of these domains, we conducted a systematic analysis with different starting queries. These PSI-BLAST searches recovered a vast superfamily of proteins with a conserved domain similar to the C-terminal Ig-like domain of NFAT and NFκB. The searches initiated with different starting points consistently retrieved essentially the same set of sequences at *e*-values < 0.01, which reinforced the significance of the observed relationships. This was also supported by secondary structure predictions using the PHD program for the newly identified members. The diverse TIG domain superfamily (Figure 6) includes transcription factors of different families such as the Olf-1/Unc-3 family (Prasad *et al.*, 1998), including the *Saccharomyces cerevisiae* SPT23 and MGA2 (Yiv1; Zhang *et al.*, 1997), and the CBF1(Suppressor of *hairless*) family (Fortini & Artavanis-Tsakonas, 1994; Schweisguth, 1995). Multiple copies of the TIG domain were also detectable in the extracellular regions of several proteins, such as the tyrosine kinases of the MET/RON family (Vande Woude *et al.*, 1997) and probable adhesion molecules such as plexin (Satoda *et al.*, 1995) and SEX (Maestrini *et al.*, 1996), as well as an extracellular protein which is the major virulence determinant of the fish pathogenic actinomycete *Renibacterium* (Barton *et al.*, 1997). In addition, the Sec5 subunit of the animal secretory exocyst complex (Kee *et al.*, 1997) also contains a TIG domain at its extreme N terminus. Some of the searches started with the TIG domains also detected the Ig-like domain of the bacterial cyclodextrin glucan transferases that has been identified previously on the basis of structural comparisons (Hofmann *et al.*, 1989).

Inspection of the multiple alignment of TIG domains showed that the sequence conservation mapped to hydrophobic residues that form the structural basis of the β-strands (Figure 6) as well as a characteristic N-terminal loop between strands 2 and 3. The identification of the Ig-like TIG domains in some of these proteins has important functional and evolutionary implications. Proteins containing the TIG domain have been shown to contact DNA either as dimers (NFκB; Ghosh *et al.*, 1995; Muller *et al.*, 1995) or as monomers (NFAT; Chen *et al.*, 1998). The TIG domain plays a major role in the dimerization of NFκB (Ghosh *et al.*, 1995; Muller *et al.*, 1995). The differences in the structural context of this domain in NFAT and NFκB suggest that the TIG domain is versatile in its DNA and protein contacting activities, with loops between the strands of the Ig domain being crucial for this process (Chen *et al.*, 1998). Olf-1 binds DNA as a dimer and has an additional Zn-binding motif which is located upstream of the TIG domain and is required for specific DNA binding (Hagman *et al.*, 1995). It seems likely that in Olf-1, similarly to NFκB, the TIG domain performs a dual role, being involved both in dimerization and in DNA binding. The TIG domain is a likely candidate for a role in non-specific DNA binding by Olf-1 rather than the purported helix-loop-helix (HLH) domain (Wang & Reed, 1993), for which no statistical support could be obtained (L.A., unpublished observations).

The CBF family of transcription factors are DNA-binding proteins that act as repressors of transcription in the Notch pathway (Fortini & Artavanis-Tsakonas, 1994; Schweisguth, 1995). These proteins bind to NFκB-like target sequences (Shirakata *et al.*, 1996) and interact with NFκB and C/EBP in the IL-6 gene regulatory region (Kannabiran *et al.*, 1997). Taken together with the presence of the TIG domain, this suggests a mode of DNA binding and protein-protein interactions similar to that of NFκB. Finally, the Arabidopsis protein F1N21.9 appears to be the ortholog of the

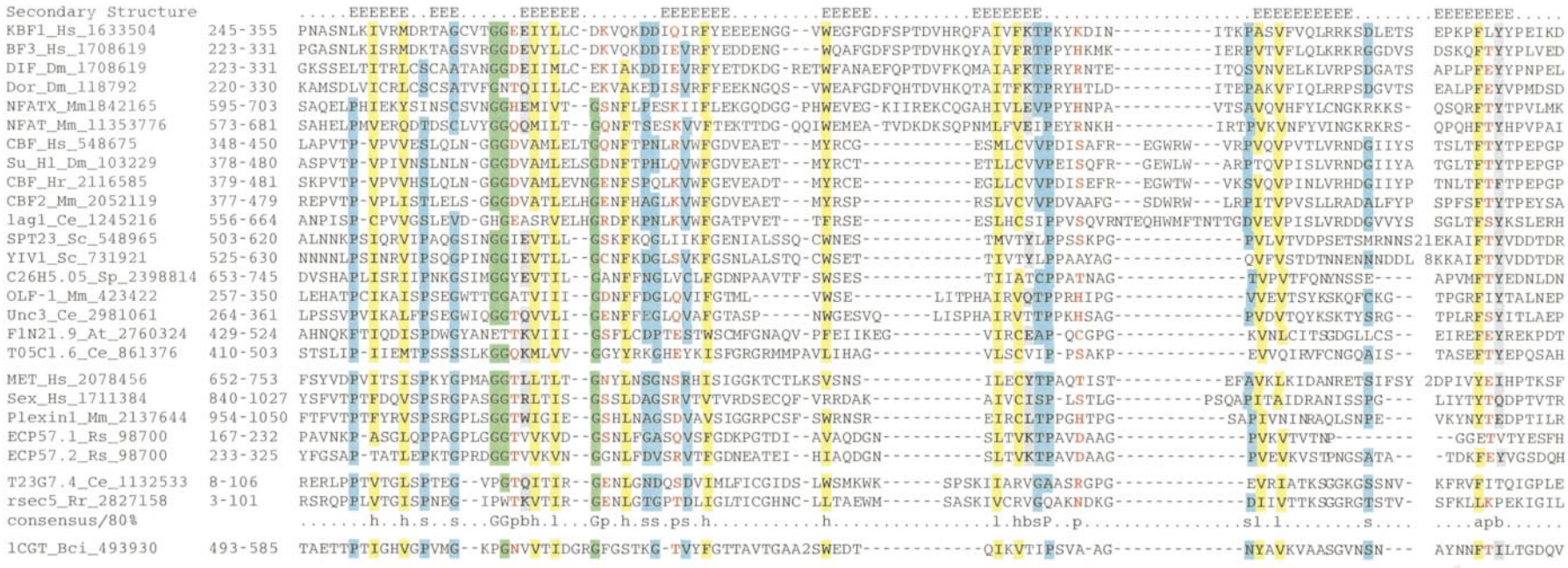

Figure 6. Multiple alignment of the TIG domain superfamily. The sequence groups are as follows: 1, transcription factors; 2, receptor tyrosine kinases and other membrane proteins; 3, exocyst complex subunits (Sec5). The sequence below the consensus line is from cyclodextrin glycosyltransferase. The secondary structure assignments are a consensus of the secondary structure elements extracted from the X-ray structures of NFκB (1SVC) and NFAT (1A02). The other designations are as described in the legend to Figure 1. Additional species abbreviations: Dm, *Drosophila melanogaster*; Hr, *Halocynthia roretzi*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*; Rs, *Renibacterium salmonarium*; Bci, *Bacillus circulans*.

CG-1 protein from parsley (whose cDNA has been cloned only partially), which is a light-induced DNA-binding protein with a specificity towards the CGCG motif (da Costa e Silva, 1994), and an uncharacterized *Caenorhabditis elegans* protein. This observation is of interest as it suggests that TIG domain-containing transcription factors are ubiquitous at least in the crown group of eukaryotes.

The presence of multiple copies of the TIG domain (hitherto unnoticed) in the extracellular regions of the Met family receptor tyrosine kinases, plexins and the related SEX receptor molecules (which, instead of the kinase domain, contain an intracellular Ras GAP domains; L.A., unpublished observations) is consistent with the traditional extracellular role of Ig-like domains. The tyrosine kinases of this family function as receptors for hepatocyte growth factors and also interact with the extracellular matrix; the TIG domains likely mediate some of these interactions (Vande Woude *et al.*, 1997). The TIG domain superfamily is notable in that a clear relationship detectable at the sequence level was established between extracellular Ig-like domains and the intracellular ones seen in the transcription factors.

A β-propeller domain in Rag-2

The diversity of antigen receptors (namely the immunoglobulins and T-cell receptors) in vertebrates depends on combinatorial shuffling of individual modules at the DNA level mediated by the so-called V-D-J recombinase. This recombinase, which also possesses a transposase activity, consists of two subunits, RAG-1 and RAG-2 (Oettinger, 1996; van Gent *et al.*, 1995, 1996). The crystal structure of RAG-1 revealed that it combines a RING finger with a C2H2 Zn finger into a novel DNA-binding structure without recognizable similarity to any other recombinases (Bellon *et al.*, 1997). The second subunit, RAG-2, does not show significant similarity to any other proteins in standard database searches. However, PSI-BLAST searches initiated with different β-propellers of the kelch-repeat type (Bork & Doolittle, 1994), retrieved the previously well-characterized proteins, such as kelch itself, HCF, fungal galactose oxidase, scruin and a family of poxvirus proteins, as well as Rag-2. Rag-2 emerged with the same level of statistical significance as galactose oxidase with a known β-propeller structure in a search initiated with the N-terminal propeller domain of β-scruin (*e*-value $\sim 10^{-4}$ in the third iteration).

On the basis of the galactose oxidase structure, the position of the characteristic glycine doublets typical of the kelch domain were identified in Rag-2 and the individual β-barrel repeats of the propeller were demarcated (Figure 7). Inspection of the multiple alignment of the repeats from Rag-2 with those of other kelch proteins shows that while the repeats in Rag-2 are divergent, they maintain the conserved hydrophobic residues corresponding to the individual strands of the β-barrel (Figure 7). Rag-2 contains five clearly detectable kelch repeats, but additional, permuted copies may be present at the ends of the proper repeats which would result in a six or seven-bladed β-propeller structure for Rag-2. The linker regions between the kelch repeats in Rag-2 are of similar length to those in the fungal galactose oxidase, which suggests that they curl around the next repeat and the individual repeats are placed at a deeper angle with respect to one another than in such proteins as Kelch or HCF. In addition to the β-propeller domain, the only other portion of Rag-2 that is predicted to possess globular structure is a putative cysteine-rich, metal-binding domain located near the C terminus. The dissection of the Rag-2 sequence into these domains has two important implications. Firstly, like other β-propellers off this class, Rag-2 is expected to be capable of versatile protein-protein interactions and is likely to play a central role in the formation of multisubunit complexes involved in recombination. Secondly, there was extensive speculation that Rag proteins may derive from the genes of some transposable element that has been inserted in the vertebrate germ line (Agrawal *et al.*, 1998). The identification of typically "cellular" eukaryotic domains in these proteins makes this hypothesis highly unlikely. Rag-2 may have evolved from a pre-existing cellular kelch-repeat protein. Given the catalytic activity of a number of kelch-repeat β-propellers, for example, galactose oxidase and sialidases (Bork & Doolittle, 1994), it is possible that Rag-2 plays a structural as well as a catalytic role in the recombinase reaction. We are aware of an independent analysis that arrived to very similar conclusions on the domain architecture of Rag-2 using a different computational technique (Callebaut & Mornon, 1998).

Concluding remarks

It must be emphasized that this work by no means presents a comprehensive bench-marking of PSI-BLAST. Some efforts in this direction have been recently published by this and other groups (Huynen *et al.*, 1998; Rychlewski *et al.*, 1998; Wolf *et al.*, 1999). Clearly, much additional work is required in order to fully evaluate the benefits and pitfalls of using iterative database search at large scale and in an automated regime and to establish the optimal strategy for such applications. Furthermore, the present analysis involved a relatively permissive cut-off for inclusion of sequences into profiles and the procedure to some extent varied for different protein superfamilies. Thus, we describe here an approach to protein superfamily analysis that should be applied in a human-controled fashion, which involves careful examination of diagnostic sequence and structural motifs, rather than a protocol for automated analysis.

These limitations and cautionary notes notwithstanding, the results presented here demonstrate

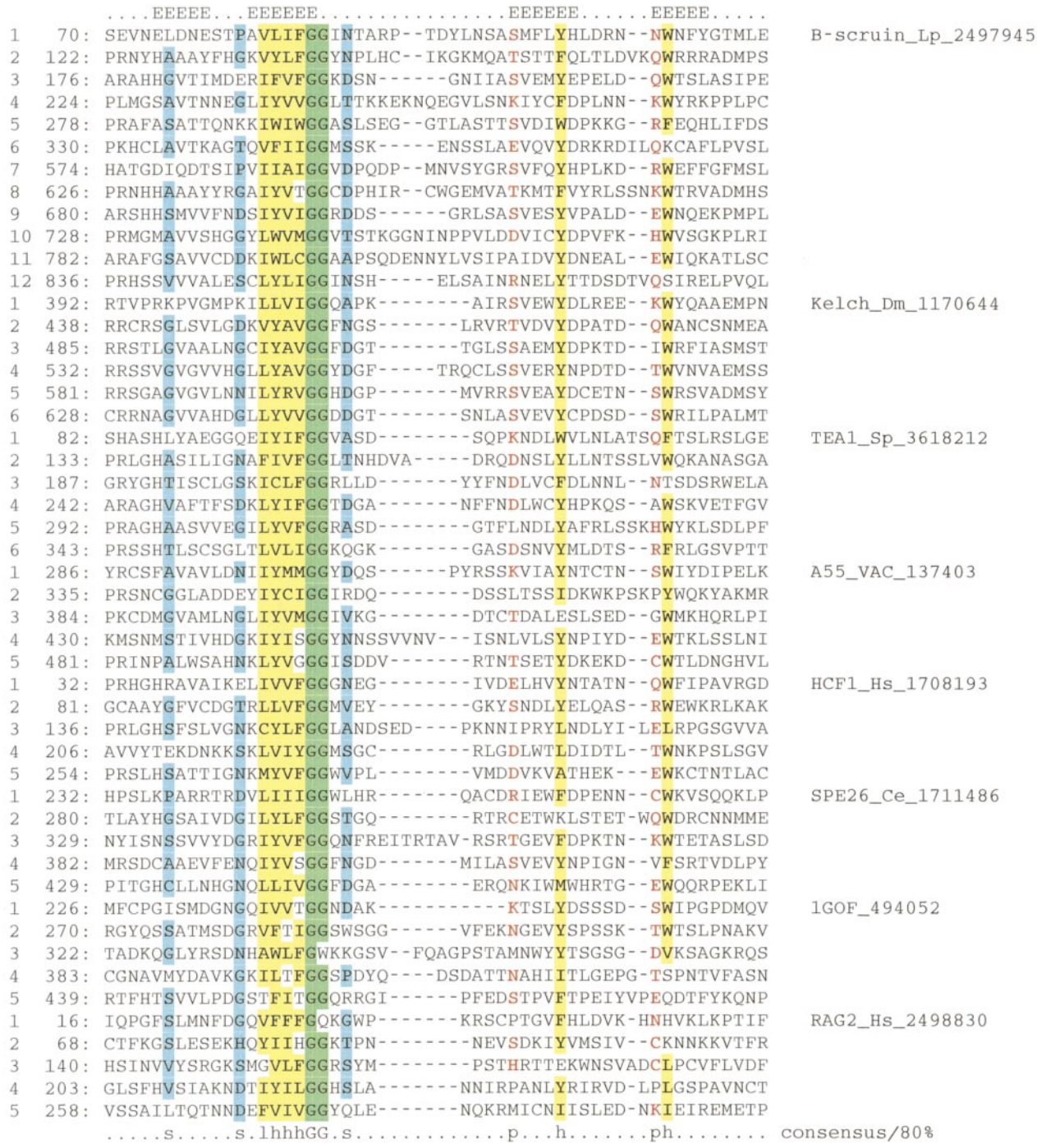

Figure 7. Multiple alignment of the kelch-repeats detected in Rag-2 with known kelch repeats. The alignment shows all repeats identified in Rag-2 and selected proteins known to contain the kelch domain; the repeats from each of the proteins are numbered consecutively from the N to the C terminus. The numbers in the second column indicate the number of the first residue of each repeat in the respective protein sequence. TEA1 is a cell polarity protein, HCF1 is host cell factor-1 (a protein required for the transcription of immediate early genes of herpes simplex virus), 1GOF is galactose oxidase from *Hypomyces rosellus*. The secondary structure assignments were derived from the 1GOF structure. The other designations are as described in the legend to Figure 1. Additional species abbreviation: Lp, *Limulus polyphemus*.

the significant potential of iterative sequence database search in detecting subtle but functionally and evolutionarily important structural relationships between proteins. At the same time, they show that detailed examination of protein superfamilies that allows an optimal choice of queries to initiate the iterations is critical for the realization of this potential. Thus, in principle, a comprehensive exploration of structural relationships between proteins using sequence analysis should proceed by

systematically identifying all protein superfamilies in the NR database and then performing iterative searches with appropriately selected starting points, in order to detect diverged members of these superfamilies and join some of them into higher level classes. A corollary of this is that protein classification and structural characterization may be regarded as an autocatalytic processes: the better the resolution of superfamily analysis, the greater the opportunities for detecting ever more subtle relationships. Given the parallel progress in structure determination and sequence and structure comparison methods that is currently observed, a complete hierarchical classification of the protein universe, with a reliable structural prediction for each family, however challenging a task, may be in sight.

Material and Methods

Databases

Standard database searches were performed using the non-redundant (NR) protein database at the NCBI. The structural databases used here were PDB and SCOP (Structural Classification of Proteins; Murzin *et al.*, 1995; Hubbard *et al.*, 1999). SCOP employs a manual process to identify structural relationships between proteins and classifies them into a four-level hierarchy. This hierarchy from top to bottom reflects the protein structural class in terms of secondary structural elements (α-helices and β-strands), a general structural relationship in the arrangement of these elements (fold), an inferred evolutionary relationship (superfamily), and a statistically highly significant sequence similarity between proteins (family). In contrast, the FSSP database classifies proteins by clustering them according to *Z*-scores, a measure of alignment between the backbones of two structures. Throughout this analysis, we adopted the SCOP classification of folds. Coordinates for protein structures were obtained from PDB.

Database searches

The principal search tool used in this study was PSI-BLAST (Altschul *et al.*, 1997; Altschul & Koonin, 1998). Briefly, the program constructs a position-dependent weight matrix from multiple alignments generated from the BLAST hits above a certain expectation value (e-value) and carries out iterative database searches using the information derived from this matrix. PSI-BLAST-C and R options were used to save and retrieve the position-dependent weight matrices (profiles), respectively. Typically, the profiles were built either by searching with a query sequence against the NR database for a fixed number of iterations or to convergence, or alternatively, against a data set comprised of proteins known to belong to a given superfamily. Generally, an expectation value threshold of 0.01 was used for inclusion of sequences into the matrix for the next iteration. In some cases, however, profiles were built with variable thresholds for each iteration in order to ensure the exclusion of apparent false positives. In order to minimize the risk of including false positives into profiles, the searches were typically carried out using the sequences of the predicted globular domains only. The likely globular domains were delineated by masking compositionally biased regions with the following programs: SEG, with the parameters window size 45, trigger complexity 3.4 and extension complexity 3.75 (Wootton & Federhen, 1996), for detection of different types of regions with a low compositional complexity; COILS, for coiled coil regions (Lupas, 1996); and PHDhtm for hydrophobic transmembrane helices (Rost *et al.*, 1995). All these procedures as well as batch database searches and clustering of sequences by similarity were carried out using the scripts of the SEALS package (Walker & Koonin, 1997). The currently accepted default cut-off for inclusion of sequences into profiles by PSI-BLAST is 0.001 rather then 0.01 as employed in this analysis (Altschul & Koonin, 1998; http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi blast). The protocol described here is generally not suitable for large-scale, completely automated analyses using PSI-BLAST and must be applied in a controled manner. More specifically, this analysis included examination of the final or, if necessary, intermediate results of the PSI-BLAST searches for the conservation of sequence and structural motifs that are diagnostic of a particular protein superfamily (Bork & Koonin, 1998; Altschul & Koonin, 1998). Additional precautions, such as masking compositionally biased regions in the database and running a limited number of iterations may be required for implementing an automated procedure based on PSI-BLAST (e.g. Wolf *et al.*, 1999).

Multiple alignments

For constructing multiple alignments, the high-scoring segment pairs (HSPs) generated by PSI-BLAST were fed into the multiple alignment program CLUSTALX (Thompson *et al.*, 1994) and re-aligned using different gap opening and extension parameters and the BLOSUM series of matrices. This procedure was particularly effective for compact domains that do not contain large insertions or deletions. Alternatively, for domains with variable-size insert and gap regions, the GIBBS sampling procedure as implemented in the MACAW and MGIBBS programs was used to identify conserved motifs (Neuwald *et al.*, 1995; Schuler *et al.*, 1991). For all constructed alignments, the key motifs were mapped on the known three-dimensional structures and the alignments were extended and modified (if necessary) on the basis of the compatibility with the structures.

Structure manipulations

PDB files were visualized using the SWISSPDB-Viewer program. This program was also used for constructing structural alignments between a target sequence and a template structure and submitting them for crude structural modeling using the PROMODII program which applies the GROMOS energy minimization method (Pietsch, 1996). The structural alignments between the target and the template were manually improved in order to achieve a global reduction in the potential. This modeling protocol does not aim at predicting fine structural details of the target proteins but allows one to visualize both the general similarity to the template and major distinctions, such as large insertions and deletions. Secondary structure predictions were carried out using the PHD program with multiple alignment inputs (Rost & Sander, 1994). Additional assessments of the structural relationships were performed using the Zega procedure which computes the probability of two aligned

sequences adopting the same structure (Abagyan & Batalov, 1997).

Acknowledgments

We are grateful to Michael Rozanov for his participation in the early stage of the HSP70 superfamily analysis.

References

Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.

Abdullah, K. M., Udoh, E. A., Shewen, P. E. & Mellors, A. (1992). A neutral glycoprotease of *Pasteurella haemolytica* A1 specifically cleaves O-sialoglycoproteins. *Infect. Immun.* **60**, 56-62.

Agrawal, A., Eastman, Q. M. & Schatz, D. G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature,* **394**, 744-751.

Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444-447.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Aravind, L. & Koonin, E. V. (1998). Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucl. Acids Res.* **26**, 3746-2752.

Aravind, L. & Koonin, E. V. (1999). DNA polymerase β-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucl. Acids Res.* **27**, in the press.

Aravind, L., Leipe, D. D. & Koonin, E. V. (1998). Toprim-a conserved catalytic domain in tyupe IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucl. Acids Res.* **26**, 4205-4213.

Arigoni, F., Talabot, F., Peitsch, M., Edgerton, M. D., Meldrum, E., Allet, E., Fish, R., Jamotte, T., Churchod, M.-L. & Loferer, H. (1998). A genome-based approach for the identification of essential bacterial genes. *Nature Biotechnol.* **16**, 851-856.

Barton, T. A., Bannister, L. A., Griffiths, S. G. & Lynch, W. H. (1997). Further characterization of *Renibacterium salmoninarum* extracellular products. *Appl. Environ. Microbiol.* **63**, 3770-3775.

Bellon, S. F., Rodgers, K. K., Schatz, D. G., Coleman, J. E. & Steitz, T. A. (1997). Crystal structure of the RAG1 dimerization domain reveals multiple zinc-binding motifs including a ovel zinc binuclear cluster. *Nature Struct. Biol.* **4**, 586-591.

Benne, R. (1993). RNA editing in mitochondria of *Leishmania tarentolae* and *Crithidia fasciculata*. *Semin. Cell Biol.* **4**, 241-249.

Berger, J. M., Fass, D., Wang, J. C. & Harrison, S. C. (1998). Structural similarities between topoisomerases that cleave one or both DNA strands. *Proc. Natl Acad. Sci. USA,* **95**, 7876-7881.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Boll, M. & Fuchs, G. (1995). Benzoyl-coenzyme A reductase (dearomatizing), a key enzyme of anaerobic aromatic metabolism. ATP dependence of the reaction, purification and some properties of the enzyme from *Thauera aromatica* strain K172. *Eur. J. Biochem.* **234**, 921-933.

Bork, P. & Doolittle, R. F. (1994). *Drosophila kelch* motif is derived from a common enzyme fold. *J. Mol. Biol.* **236**, 1277-1282.

Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.* **18**, 313-318.

Bork, P., Sander, C. & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl Acad. Sci. USA,* **89**, 7290-7294.

Callebaut, I. & Mornon, J. P. (1998). The V(D)J recombination activating protein RAG2 consists of a six-bladed properller and a PHD fingerlike domain, as revealed by sequence analysis. *Cell Mol. Life Sci.* **54**, 880-891.

Cashel, M., Gentry, D. R., Hernandez, V. J. & Binella, D. (1996). The stringent response. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F. C., III, R, C., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M. & Umbarger, H. E., eds), pp. 1458-1496, ASM Press, Washington, DC.

Chen, L., Glover, J. N., Hogan, P. G., Rao, A. & Harrison, S. C. (1998). Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature,* **392**, 42-48.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature,* **357**, 543-544.

Colbeau, A., Elsen, S., Tomiyama, M., Zorin, N. A., Dimon, B. & Vignais, P. M. (1998). *Rhodobacter capsulatus* HypF is involved in regulation of hydrogenase synthesis through the HupUV proteins. *Eur. J. Biochem.* **251**, 65-71.

da Costa e Silva, O. (1994). CG-1, a parsley light-induced DNA-binding protein. *Plant Mol. Biol.* **25**, 921-924.

Doolittle, R. F. (1995). The origins and evolution of eukaryotic proteins. *Phil. Trans. Roy. Soc. ser. B,* **349**, 235-240.

Esser, L., Wang, C. R., Hosaka, M., Smagula, C. S., Sudhof, T. C. & Deisenhofer, J. (1998). Synapsin I is structurally similar to ATP-utilizing enzymes. *EMBO J.* **17**, 977-984.

Fortini, M. E. & Artavanis-Tsakonas, S. (1994). The suppressor of hairless protein participates in notch receptor signaling. *Cell,* **79**, 273-282.

Galperin, M. Y. & Koonin, E. V. (1997). A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity. *Protein Sci.* **6**, 2639-2643.

Gane, P. J., Dunwell, J. M. & Warwicker, J. (1998). Modeling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. *J. Mol. Evol.* **46**, 488-493.

Ghosh, G., van Duyne, G., Ghosh, S. & Sigler, P. B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. *Nature,* **373**, 303-310.

Gibson, J., Dispensa, M. & Harwood, C. S. (1997). 4-Hydroxybenzoyl coenzyme A reductase (dehydroxylating) is required for anaerobic degradation

of 4-hydroxybenzoate by *Rhodopseudomonas palustris* and shares features with molybdenum-containing hydroxylases. *J. Bacterial.* **179**, 634-642.
Gottesman, S. (1996). Proteases and their targets in *Escherichia coli*. *Annu. Rev. Genet.* **30**, 465-506.
Gottesman, S. & Maurizi, M. R. (1992). Regulation by proteolysis: energy-dependent proteases and their targets. *Microbiol. Rev.* **56**, 592-621.
Grant, G. A., Schuller, D. J. & Banaszak, L. J. (1996). A model for the regulation of D-3-phosphoglycerate dehydrogenase, a Vmax-type allosteric enzyme. *Protein Sci.* **5**, 34-41.
Hagman, J., Gutch, M. J., Lin, H. & Grosschedl, R. (1995). EBF contains a novel zinc coordination motif and multiple dimerization and transcriptional activation domains. *EMBO J.* **14**, 2907-2916.
Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.
Hofmann, B. E., Bender, H. & Schulz, G. E. (1989). Three-dimensional structure of cyclodextrin glycosyltransferase from *Bacillus circulans* at 3.4 Å resolution. *J. Mol. Biol.* **209**, 793-800.
Holm, L. & Sander, C. (1995). DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* **20**, 345-347.
Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595-603.
Holm, L. & Sander, C. (1997). New structure-novel fold? *Structure,* **5**, 165-171.
Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. (1999). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **27**, 254-256.
Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. & Bork, P. (1998). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323-326.
Jabbouri, S., Relic, B., Hanin, M., Kamalaprija, P., Burger, U., Prome, D., Prome, J. C. & Broughton, W. J. (1998). nolO and noeI (HsnIII) of *Rhisobium* sp. NFR234 are involved in 3-*O*-carbamoylation and 2-*O*-methylation of Nod factors. *J. Biol. Chem.* **273**, 12047-12055.
Kannabiran, C., Zeng, X. & Vales, L. D. (1997). The mammalian transcriptional repressor RBP (CGF1) regulates interleukin-6 gene expression. *Mol. Cell Biol.* **17**, 1-9.
Kee, Y., Yoo, J. S., Hazuka, C. D., Peterson, K. E., Hus, S. C. & Scheller, R. H. (1997). Subunit structure of the mammalian exocyst complex. *Proc. Natl Acad. Sci. USA,* **94**, 14438-14443.
Komoszynski, M. & Wojtczak, A. (1996). Apyrases (ATP disphosphohydrolases, EC 3.6.1.5): function and relationship to ATPases. *Biochim. Biophys. Acta,* **1310**, 233-241.
Koonin, E. V. (1994). Yeast protein controlling interorganelle communication is related to bacterial phosphatases containing the Hsp70-type ATP-binding domain. *Trends Biochem. Sic.* **19**, 156-157.
Lupas, A. (1996). Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**, 513-525.
Maestrini, E., Tamagnone, L., Longati, P., Cremona, O., Gulisano, M., Bione, S., Tamanini, F., Neel, B. G., Toniolo, D. & Comoglio, P. M. (1996). A family of transmembrane proteins with homology to the MET-hepatocyte growth factor receptor. *Proc Natl Acad. Sci. USA,* **93**, 674-678.
Martin, J. & Hartl, F. U. (1997). Chaperone-assisted protein folding. *Curr. Opin. Struct. Biol.* **7**, 41-52.
Mellors, A. & Lo, R. Y. (1995). O-sialoglycoprotease from *Pasteurella haemolytica*. *Methods Enzymol.* **248**, 728-740.
Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature,* **373**, 311-317.
Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386-394.
Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.
Murzin, A. G. & Bateman, A. (1997). Distant homology recognition using structural classification of proteins. *Proteins: Struct. Funct. Genet.* **1**, 105-112.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
Mushegian, A. R., Bassett, D. E., Jr, Boguski, M. S., Bork, P. & Koonin, E. V. (1997). Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl Acad. Sci. USA,* **94**, 5831-5836.
Neuwald, A. F., Liu, J. S. & Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618-1632.
Oettinger, M. A. (1996). Cutting apart V(D)J recombination. *Curr. Opin. Genet. Dev.* **6**, 141-145.
Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631-634.
Patte, J. C., Richaud, C., Boy, E., Reinisch, F., Richaud, F. & Cassan, M. (1976). Regulation of lysine biosynthesis in *Escherichia coli* K12. *Acta Microbiol. Acad. Sci. Hung.* **23**, 121-128.
Peitsch, M. C. (1996). ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* **24**, 274-279.
Pittard, J. (1996). The various strategies within the TyrR regulation of *Escherichia coli* to modulate gene expression. *Genes Cells,* **1**, 717-725.
Prasad, B. C., Ye, B., Zackhary, R., Schrader, K., Seydoux, G. & Reed, R. R. (1998). unc-3, a gene required for axonal guidance in *Caenorhabditis elegans*, encodes a member of the O/E family of transcription factors. *Development,* **125**, 1561-1568.
Reizer, J., Reizer, A., Saier, M. H., Jr, Bork, P. & Sander, C. (1993). Exopolyphosphate phosphatase and guanosine pentaphosphate phosphatase belong to the sugar kinse/actin/hsp 70 superfamily. *Trends Biochem. Sci.* **18**, 247-248.
Rhee, S. G., Park, S. C. & Koo, J. H. (1985). The role of adehylyltransferase and uridylyltransferase in the regulation of glutamine synthetase in *Escherichia coli*. *Curr. Top. Cell Reg.* **27**, 221-232.
Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55-72.
Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521-533.
Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Design,* **3**, 229-238.
Satoda, M., Takagi, S., Ohta, K., Hirata, T. & Fujisawa, H. (1995). Differential expression of two cell surface

proteins, neuropilin and plexin, in *Xenopus olfactory* axon subclasses. *J. Neurosci.* **15**, 942-955.
Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet.* **9**, 180-190.
Schuller, D. J., Grant, G. A. & Banaszak, L. J. (1995). The allosteric ligand site in the Vmax-type cooperative enzyme phosphoglycerate dehydrogenase. *Nature Struct. Biol.* **2**, 69-76.
Schweisguth, F. (1995). Suppressor of Hairless is required for signal reception during lateral inhibition in the *Drosophila* pupa notum. *Development*, **121**, 1875-1884.
Shirakata, Y., Shuman, J. D. & Coligan, J. E. (1996). Purification of a novel MHC class I element binding activity from thymus nuclear extracts reveals that thymic RBP-Jkappa/CBF1 binds to NF-kappaB-like elements. *J. Immunol.* **156**, 4672-4679.
Shuman, S. (1996). Closing the gap on DNA ligase. *Structure,* **4**, 653-656.
Shuman, S. & Schwer, B. (1995). RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.* **17**, 405-410.
Subramanya, H. S., Doherty, A. J., Ashford, S. R. & Wigley, D. B. (1996). Crystal structure of an ATP-dependent DNA ligase from bacteriophage T7. *Cell,* **85**, 607-615.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.
Vande Woude, G. F., Jeffers, M., Cortner, J., Alvord, G., Tsarfaty, I. & Resau, J. (1997). Met-HGF/SF: tumorigenesis, invasion and metastasis. *Ciba Found. Symp.* **212**, 119-130, 130-132, 148-154.
Van Gent, D. C., McBlane, J. F., Ramsden, D. A., Sadofsky, M. J., Hesse, J. E. & Gellert, M. (1995). Initiation of V(D)J recombination in a cell-free system. *Cell,* **81**, 925-934.
van Gent, D. C., Mizuuchi, K. & Gellert, M. (1996). Similarities between initiation of V(D)J recombination and retroviral integration. *Science,* **271**, 1592-1594.
Vingron, M. & Sibbald, P. R. (1993). Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc. Natl Acad. Sci. USA,* **90**, 8777-8781.
Vyazmensky, M., Sella, C., Barak, Z. & Chipman, D. M. (1996). Isolation and characterization of subunits of acetohydroxy acid synthase isozyme III and reconstitution of the holoenzyme. *Biochemistry,* **35**, 10449-10346.
Walker, D. R. & Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences. *Intell. Syst. Mol. Biol. (ISMB),* **5**, 333-339.
Wang, M. M. & Reed, R. R. (1993). Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature,* **364**, 121-126.
Wang, T. F. & Guidotti, G. (1998). Golgi localization and functional expression of human uridine diphosphatase. *J. Biol. Chem.* **273**, 11392-11399.
Watanabe, K., Ishikawa, T., Mukohara, Y., Nakamura, H., Boll, M., Albracht, S. S. & Fuchs, G. (1992). Cloning and sequencing of the genes involved in the conversion of 5-substituted hydantoins to the corresponding L-amino acids from the native plasmid of *Pseudomonas sp*. strain NS671. *J. Bacteriol.* **174**, 962-969.
Wilson, T. J., Argaet, V. P., Howeltt, G. J. & Davidson, B. E. (1995). Evidence for two aromatic amino acid-binding sites, one ATP-dependent and the other ATP-independent, in the *Escherichia coli* regulatory protein TyrR. *Mol. Microbiol.* **17**, 483-492.
Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17-26.
Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexing measures. *Comput. Chem.* **18**, 269-285.
Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.
Ye, G. J., Breslow, E. B., Meister, A. & Gjo-jie, G. E. (1996). The amino acid sequence of rat kidney 5-oxo-L-prolinase determined by cDNA cloning. *J. Biol. Chem.* **271**, 32293-32300.
Zhang, S., Burkett, T. J., Yamashita, I. & Garfinkel, D. J. (1997). Genetic redundancy between SPT23 and MGA2: regulators of Ty-induced mutations and Ty1 transcription in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **17**, 4718-4729.

Edited by J. M. Thornton

(Received 29 September 1998; received in revised form 11 February 1999; accepted 23 February 1999)