Published in final edited form as:
Front Biosci. ; 14: 1143–1151.

Structural conservation of a short, functional, peptide-sequence motif

Susan Fox-Erlich, **Martin R. Schiller**, and **Michael R. Gryk**
Department of Molecular, Microbial & Structural Biology, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030-3305

1. ABSTRACT

Full length, eukaryotic proteins generally consist of several autonomously folding and functioning domains. Many of these domains are known to function by binding and/or modifying other partner proteins based on the recognition of a short, linear amino sequence contained within the target protein. This article reviews the many bioinformatic tools and resources which discover, define and catalogue the various, known protein domains as well as assist users by identifying domain signatures within proteins of interest. We also review the smaller subset of bioinformatic tools which catalogue and help identify the short linear motifs used for domain targeting. It has been suggested that these short, functional, peptide-sequence motifs are normally found in unstructured regions of the target. The role of protein structure in the activity of one representative of these short, functional motifs is explored through an examination of known structures deposited in the Protein Data Bank.

Keywords

Protein Domains; Protein Domains; Protein Structure; Bioinformatics; review

2. INTRODUCTION TO PROTEIN ARCHITECTURE

Recent advances in genomic sequencing have dramatically increased the volume of nucleic acid and protein sequence information available for hundreds of organisms spanning the entire taxonomic tree from bacteria to man. This sequence information has been immediately useful in demonstrating the overall similarity of all life on earth, with the degree of variation between species allowing for phylogenetic-based trees of life that provide an independent validation of anatomical phylogeny (1–3). Genome comparison of select individuals within an individual species allows for the mapping of population diversity, which when coupled with disease attributes, has been successful in the identification of genetic causes and susceptibilities of disease (4–8).

Send correspondence to: Michael R. Gryk, Department of Molecular, Microbial & Structural Biology, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030-3305, Tel: 860-679-4785, Fax: 860-679-3408, E-mail: gryk@uchc.edu.

At a molecular level, sequence comparisons of genes and gene products have been heavily relied upon to identify putative functions in newly-discovered genes of unknown function (9). The logic underlying such comparisons is that as protein/nucleic acid function is conferred through the specific activity of certain key residues, those residues must be conserved among gene products sharing the conserved function. In addition to the conservation of these functionally important, active-site moieties, a larger group of residues are also found to be conserved which are required to form the structural scaffold upon which the functional groups sit. Fortunately for the field of bioinformatics, the coupled need for conservation of structural and chemical groups to preserve function provides a large sequence 'fingerprint' which is exploited to identify putative function in otherwise unknown protein/nucleic acid sequences.

The activities ascribed to gene products within a cell can be roughly broken down into three general classes: enzymatic activity for catalyzing specific chemical reactions useful to the cell, structural scaffolding for maintaining the structural integrity and infrastructure of the cell, and regulatory control for ensuring the spatial and temporal utilization of the other two types of activities. (The latter classification is often blurry, as many enzymatic reactions such as ATP hydrolysis and phosphorylation are coupled to regulatory control). The regulatory elements greatly outnumber the enzymatic and structural elements and in higher organisms a single gene product is often composed of dozens of regulatory elements for controlling post-translational modifications, trafficking and signal transduction cascades. Each of these subclasses of functions require the specific recognition and binding of gene products, either nucleic acid – nucleic acid, protein – nucleic acid, or protein – protein interactions. Furthermore, the genes and gene products are modular in nature, composed of a series of smaller functional units each conveying a portion of function to the whole. In the case of regulatory elements, each separate unit provides a different regulatory mechanism to be imparted to the regulation of the gene/gene product.

This modular construction inherent to genes is very beneficial for bioinformatics analysis. Without it, we would need to observe sequence similarity over the entire gene/gene product in order to infer function. Because of this modular architecture, we can infer specific sub-functions from sequence conservation over much smaller portions of the gene sequence, and piece together the purpose of the whole from the sum of its parts. This carving up of the problem into smaller pieces is obviously critical for the analysis of DNA sequences, allowing for the identification of promoters, enhancers and other transcriptional control sites pulled from the enormity of an entire chromosomal sequence. This same reductionism is critical in the bioinformatic analysis of protein sequences, which will be the focus of the remainder of this review. While in rare instances the molecular biologist may discover a novel protein which is homologous in its entirety to another well-studied protein such that the function can be completely inferred (DNA polymerase activity, for example). However, it is more likely the case that portions of the novel protein will show homology to portions of a whole host of known proteins, making the role of the novel protein more enigmatic. It is for these latter cases which the bioinformatics tools described in this review are particularly useful, and the role of structure in imparting these modular sub-functions to proteins is the theme of this review.

3. BIOINFORMATIC TOOLS FOR DOMAIN/MOTIF DETECTION

In 1957, Christian Anfinsen demonstrated that the function of ribonuclease depends not only on its chemical sequence, but also on the structure those residues adopt when properly folded (10). This settled the ongoing argument as to whether proteins were amorphous blobs, active solely based on their chemical constituents, or whether a particular spatial orientation of those components was required for activity. The precision of the spatial orientation of folded proteins was revealed with the x-ray determined crystal structure of myoglobin in 1958 (11).

It is perhaps not surprising then to discover that in complex proteins containing multiple modules for performing specific functions, those independent modules adopt unique and reproducible structures required for the conveyed function. Identifying these smaller, functional modules within the context of a large peptide chain is therefore an important topic in bioinformatics, and many tools exist to assist in this venture. Some of the more popular search tools for this purpose are Prosite, Pfam, SMART, PRINTS, BLOCKS, DOMO, Prodom, SCOP, Globplot, Scansite, ELM and MnM. The key features and uses of these search tools are shown in Table 1.

3.1. Multiple Sequence Alignments

Prosite is a member of the ExPASy (Expert Protein Analysis System: http://www.expasy.ch) suite of tools provided by the Swiss Institute of Bioinformatics which also contains the protein sequence databases, Swiss-Prot and TrEMBL (12). Prosite is self-described as a “motif descriptor database” which contains sequence signatures for protein families and domains, protein families being defined as groups of proteins which are evolutionarily related through the conservation of a protein domain. The Prosite motif descriptors take two forms: patterns and profiles (13).

A Prosite pattern is a regular expression defining all peptide sequences which are to be considered a match. As illustrated below, the regular expression is a positional based array composed of either (a) amino acids which are to be considered matches (ex. [ILV]), (b) amino acids which are not to be considered matches (ex. {P}), (c) variable length gaps (ex. x(7)), and (d) identifiers for specifying the pattern must be found at N- or C-terminal of the protein (ex. <,>). Note that the regular expression thus defined provides a binary outcome. A given sequence string either is a match or is not a match of the motif descriptor pattern. No effort is made to rank non-matches based on how ‘close’ they are to fitting the pattern. Therefore, patterns are useful for identifying stretches within proteins which are strictly conserved in linear sequence. Conversely, to usefully define a domain or functional site by a Prosite pattern requires that all representatives share strictly conserved residues over the entire family. Prosite patterns are typically ten to twenty amino acids in length, providing an inherent statistical significance in their observation.

<M[RK]x(2,3)Px(2)P{FYW}

In the above Prosite pattern, the pattern will only match N-terminal sequences that begin with methionine, followed by a basic residue, two to three variable residues, proline, exactly two variable residues, proline, and finally a non-aromatic residue. The string MKGDPQEPQ would be considered a match at the N-terminus while the strings MWGDPQEPQ (W at the second position), MKGDAQEPQ (missing first proline) and MKGDPQEPY (aromatic residue in last position) would not be considered matches.

A Prosite profile, on the other hand, is a probability matrix of finding each individual amino acid at any given position in the sequence, along with probabilities for gaps and insertions at specific positions. When attempting to match a given profile within an arbitrary sequence, all probabilities are multiplied which yields a probability score of the likelihood of the match being valid. Unlike the yes/no answer for a pattern, such a probability ranking is similar to the methodology behind sequence homology searches such as the Basic Local Alignment and Search Tool (BLAST) (14).

The developers of Prosite view profiles and patterns as complementary approaches. Profiles are ideal in identifying structural domains, in which many residues are required to form the

structure, but the relative importance of each individual residue is small. In other words, structural profiles are vast enough to accommodate a large number of conservative changes along with a small amount of more significant amino acid substitutions. Patterns, on the other hand, are ideal for mapping the catalytic or functional residues within the structural domain, in order to identify whether the putative function normally ascribed to a structural domain is possible for the identified representative. An example of their combined use would be the paralogue mammalian receptors, ephrin B4 and ephrin B6 (12). Both match the Prosite profile indicating a structural kinase domain. However, only ephrin B4 contains the required Asp residue at position 740 for kinase activity. The kinase domain of ephrin B6 lacks this residue and is known to be inactive (15).

Interestingly, although the obvious rationale behind the effectiveness of Prosite descriptors is the conservation of protein structure through evolution, the profiles themselves are primarily determined through multiple sequence alignments without reference to the observed structures within the Protein Data Bank (PDB) (16). Several groups studied this apparent deficiency and discovered that including the known structure in the matching algorithm provided a higher success rate for identifying motifs within proteins, as judged by the reduction of false positives and increase in true positives in their studies (17,18). The newly defined combined pattern (ComPat) of structure and sequence elements are highly promising for refining the detection of structural domains in proteins (18).

Prosite is not the only database tool designed to identify domain signatures in peptide sequences. The Protein Family database (PFam) is generically similar to that of the Prosite profiles (19,20). Probabilities of amino acid occurrence, insertions and deletions are modeled as a Markov chain. The Hidden Markov probabilities are then determined from multiple sequence alignments. In the context of these multiple sequence alignment (MSA)-based motif searches, the term “family” is used to refer to a collection of protein sequences which are anticipated to have the same function (and likely similar structures) due to the high degree of similarity in their protein sequences. The term “domain” is used to refer to a single instance of the domain family represented in an individual protein sequence. However, as illustrated above with the kinase domain of ephrins, individual domains belonging to one structural family may not have the same function. This use of the term “domain” is in contrast to that used in structural biology in which a “domain” is defined as an autonomously folding protein unit.

SMART (A Simple Modular Architecture Research Tool) is another database of domains defined as probability profiles determined from multiple sequence alignment (21). It was originally introduced in 1998 to complement the existing domain identification tools by concentrating on domains of proteins involved in signal transduction – a noted weak spot of the other existing tools. It has subsequently been expanded to cover over 600 domain families (22).

The PRINTS database also relies on MSA, but it differs from the other aforementioned tools by assigning multiple, smaller signatures to a given domain family (23). Analogous to a fingerprint comparison in which several small, characteristic patterns must be present to confirm a match, PRINTS ranks domain matches based on how many of the sub-signatures are present. The PRINTS protein fingerprint database therefore represents yet another mechanism for defining protein domain families through the use of multiple sequence alignment of known domain instances.

While each of the preceding databases was generated in a semi-automated approach, requiring human annotation, three additional domain detection web-tools, ProDom, DOMO and BLOCKS use completely automated MSA approaches to define domain signatures (24–26).

BLOCKS is similar to PRINTS in that uses multiple sequence alignments to identify many short conserved fragments which may comprise one structural/functional domain (26). Identifying such small segments gives BLOCKS the largest coverage of any of the domain identification tools; however, as it uses a completely automated technique for discovering the conserved sequence blocks, it relies on the previous domain databases for functional annotations. As such, it represents a unification of the other domain search tools (27).

3.2. Structural classification

Globplot is a bioinformatics tool which heuristically classifies regions of any given protein sequence based on their likelihood of being natively structured (globular) or unstructured (disordered) (28). Unlike many of the tools reviewed here, Globplot does not make its classifications based on reference to an annotated database – SMART/Pfam domains are reported for convenience, but such domain identifications are not utilized by the Globplot heuristic. Rather, the web-accessible tool uses a simple table of propensities for each of the amino acids to exist in a globular state and performs a rather complex windowing of these propensities over the entire sequence. The result is a simple graph plotting the propensity to be globular (or conversely, disordered) versus the residue number of the given protein sequence.

The Structural Classification of Proteins (SCOP) Database provides a hierarchical classification of all protein domain families of known structure (29). In contrast to the MSA-based tools, members of a SCOP domain family may have very little sequence identity among each other (as little as 15%) as long as they have similar folds as determined by X-ray crystallography or NMR. The SCOP database further classifies groups of similar domain families into superfamilies, similar superfamilies into folds, and finally each fold is contained within a class defined by the secondary structure elements observed in each. The logical rationale for such a classification is to trace the evolutionary development of each protein family or fold assuming that the conservation of structure is a predominant force in the conservation of sequence. Despite its underpinnings to structural and evolutionary biology, SCOP is also commonly used for bioinformatics searches due to its ability to identify putative domain signatures within a given protein sequence (based on sequence similarity to annotated SCOP domains).

Fragment Finder and the Sequence, Motif & Structure (SMS) database are both web-accessible tools for identifying similar structural elements within non-homologous proteins (30–31). Both tools use a repository of three-dimensional structures of non-homologous proteins (32). Fragment Finder is a web-accessible search tool which retrieves structural fragments from the non-homologous dataset which match their three-dimensional structure to the reference structure, as judged by a comparison of backbone torsion angles, phi and psi. The SMS database is a repository containing both the sequence and three-dimensional coordinates of all 5–10 residue peptide fragments that are found in at least three instances within the set of non-homologous proteins. Users can query the database by inputting a 5–10 residue sequence, and the database reports back any known instances of the reference sequence in the structure set. If the sequence is represented, all structures are available, as well as a simple classification: (1) all sequences exhibit similar structures, (2) all sequence exhibit different structures, or (3) the sequences exhibit a mixture of similar and different structures. Similarity of structure is determined using the program STAMP (33).

3.3. Short linear motifs

With the exception of the last two, the aforementioned bioinformatic tools are primarily useful for identifying large motifs within proteins which are often the signature of autonomously folding domains. “Mini-motifs”, on the other hand, are short, linear, amino acid stretches within

proteins that have been identified as mediating numerous biologically important events (34–36). If one were to consider protein domains to be the 'locks', mini-motifs could be considered the 'keys'. There are dozens of web-accessible databases containing such short motif signatures which are known as targets for proteolysis, phosphorylation and other post-translational modifications. However, there are only three bioinformatic tools (Scansite, Eukaryotic Linear Motif Resource (ELM) and MiniMotif Miner (MnM)) available for identifying a broad range of different types of short motifs within a given protein sequence (34–36). Scansite is unique as a resource in that it defines motifs as a position-specific scoring matrix, in which the occurrence of each amino acid at a specific position is given a numerical weight representative of the likelihood for the individual residue occurring at that position within the motif. The other two motif search engines define these short, linear motifs using the regular expression syntax used for Prosite patterns.

Due to the short nature of these motifs, all three research groups struggled with the large number of motif hits within any arbitrary protein sequence. Scansite, ELM and MnM allow the user to filter the reported motifs based on taxonomy, while ELM and MnM also allow filtering based on cell compartment localization. MnM allows for two additional filters, one based on a prediction scheme as to whether the motif is predicted to be found on the surface of the folded protein, or buried on the inside –the logic being that functional motifs would need to be surface accessible in order to engage in protein-protein interactions. The other MnM filter performs a multiple sequence alignment of the given protein sequence between orthologues from different species. Specific interpretation of these results is left to the user, but the so-called 'evolutionary conservation score' is an attempt to determine if the identified short motif is conserved through evolution in a similar manner to structural domains.

ELM provides an additional, structural-based filter in which short motifs located within identified, globular domains can be filtered out. This group notes that such short motifs are rarely found within globular domains, and if they are, they tend to be in exposed, presumably flexible, loops. Fuxreiter *et al.* have posited that these mini-motifs are unstructured in their unbound state, speculating that the mini-motif only conforms to a defined structure when bound to its binding partner (37).

4. ANALYSIS OF STRUCTURAL CONSERVATION OF YXN MOTIF

To investigate this hypothesis that mini-motifs are unstructured, we have examined the three-dimensional structures the YXN mini-motif adopts in a subset of the YXN containing protein structures deposited in the PDB. YXN, in which the tyrosine is phosphorylated (p), is one of the canonical binding motifs for several SH2-containing proteins (38). This motif, with an unphosphorylated tyrosine, is heavily represented within the PDB, often appearing multiple times within individual molecules (Table 2). Because the SH2-pYXN binding interaction is involved in initiating a biological cascade that appears to have a high level of specificity, we surmised that the majority of the YXN motifs in the PDB were not involved in this particular phenomenon (38). However, it is not readily apparent by what mechanism(s) within the cell different YXN motifs are biologically discriminated. Numerous possible mechanisms exist, including the necessity of the motif being located on the surface of the molecule, additional amino acid residues influencing secondary and tertiary structure, the hydrophobicity of the motif residues, and the susceptibility of the tyrosine to phosphorylation. In this paper, we review the possibility that the pYXN motifs involved in SH2 binding adopt a unique three-dimensional structure in the unbound state that provides precise specificity for the appropriate SH2 binding partner.

Several x-ray crystallographic determinations of SH2 domains bound to tyrosine phosphorylated YXN peptides have been determined (39–42). Using MolMol to superimpose

these structures resulted in very close geometric overlaps that were within the experimental error of the measurements, despite the structure determinations having been made independently by several laboratories (Figure 1) using different techniques (43). A key determinant in the recognition of the pYXN motif by the SH2 domain appears to be the spatial orientation between the phosphorylated tyrosine and the asparagine, which forms a specific beta-turn 'structural motif'. We sought to identify other molecules within the PDB that had similar geometric orientations of their YXN motifs even when the tyrosine was unphosphorylated.

Examination of the overlapped structures depicted in Figure 1, revealed that there were numerous geometric characteristics that could be used to search a database of YXN geometries. We chose to select structurally similar YXN motifs based on two measurements present in the tyrosine phosphorylated YXN peptides bound to SH2 sites: the distance between the alpha carbon atoms of the Y and the X+3 residue had to be less than 7 Angstroms (a characteristic of beta-turns) and a pseudo-torsion angle between the C-alpha and C-beta atoms of the Y and the N had to be constrained between 10 and 70 degrees, allowing for similar orientations of these key residues (44). Using these criteria, only ~10% of the 30,186 YXN motifs within the PDB were selected, including all of those that involved SH2 binding sites bound to phosphorylated YXN peptides (Table 3).

This observation that many biologically active, SH2-binding YXN motifs are found at beta-turns, led us to identify motifs that overlapped with a beta-turn as determined by the Define Secondary Structure of Proteins (DSSP) database, a more rigorous criteria than the 7 Angstrom distance between the C-alpha's of the Y and the X+3 residues. This selection further reduced the number of possible motifs within the result set, but it also removed some molecules that were known to be present in the SH2-binding, biologically active subset of molecules containing YXN motifs. In general, the identification of a beta-turn by the DSSP program involves a number of determinants outside the motif itself that may or may not be relevant to the YXN motif's biological activity.

A literature review of the original papers cited in the PDB for the structural determinations of several YXN motifs sorted into groups according to which amino acid was present between the Y and the N indicated that there were numerous other biological functions that the YXN motif was associated with, including binding calcium ions, binding cofactors and sugars, to name a few. Most motifs, however, had no identified function which may or may not be due to the fact the motif is in fact biologically inactive.

As might be expected, there were highly conserved motifs within biologically similar molecules such as the phosphatases. However, although there was considerable overlap with low RMSDs of these 'structural motifs' there were slight areas of difference that might preclude the ability of the tyrosine in the motif to be phosphorylated. In fact, several of them are known to bind SH2 domains in the absence of an external phosphorylated binding partner. Comparison of the structures using a visualization tool to examine the Van der Waal's radii showed similar local configurations in terms of how the individual amino acids were exposed on the surface of the molecule. Moreover, in conserved sequences there were often times conservative substitutions e.g. of phenylalanine for tyrosine and aspartic acid for asparagines with a retention of the 'structural motif' although these substitutions would preclude a similar biological activity as a phosphorylated pYXN sequence might have in the SH2 cascades previously described.

Many structures with significant overlap of their backbone atoms had differences in the orientation of their tyrosine side chains. Because it is well known that a tyrosine ring can flip,

orientations where the angle of the C-alpha-C-beta bond was similar but the ring is oriented differently could represent different biological sets of similarly activated molecules.

Interestingly, in one instance a nearly identical YXN structural motif was identified that was not present on the surface of the molecule although both the Y and the N have been identified as being essential for biological activity: thymidylate synthase, a key enzyme in DNA synthesis that catalyzes the formation of dTMP from dUMP. In this protein, TYR 261 in the sequence YVN is highly conserved and is involved in hydrogen binding to dUMP. Recent studies have documented that mutation of TYR 261 greatly affects enzyme activity without affecting dUMP binding (45). These studies indicate that TYR 261 is necessary for maintaining the structure of the molecule in key places.

The YVN sequence of thymidylate synthase adopts a structure that is remarkably similar to GRB-2 phosphopeptides and tyrosine phosphatase internal SH2 binding sites although the side chains are not accessible for recognition from the surface (Figure 2). Moreover, mutation of the tyrosine to an alanine retains this structure (45). The phi and psi angles, and all other measured geometric values are similar to these other molecules.

5. CONCLUSION

The preceding example illustrates one specific case (Grb2-like SH2 domains) in which a short, functional, peptide-sequence motif (pYXN) is recognized not solely by its linear amino acid sequence, but also by the particular three-dimensional structure the sequence adopts upon binding. It is too soon to speculate as to whether such structure-assisted recognition will prove to be the exception for short, functional motifs or the rule. However, considering how frequently nature exploits the three-dimensional structure of biomolecules (particularly proteins) in molecular recognition, such a possibility is not easily discounted.

Acknowledgements

The authors would like to thank Dr. Mark W. Maciejewski, Mr. Krishna Kaduvera and Mr. Jay Vyas for useful discussions. This research was funded in part by US National Institutes of Health grants EB001496 and GM079689.

References

1. Kosiol C, Bofkin L, Whelan S. Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. J Biomed Inform 2006;39:51–61. [PubMed: 16226061]
2. Pollock DD. Genomic biodiversity, phylogenetics and coevolution in proteins. Appl Bioinformatics 2002;1:81–92. [PubMed: 15130847]
3. Whelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 2001;17:262–272. [PubMed: 11335036]
4. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet 2007;39:S30–36. [PubMed: 17597779]
5. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet 2007;8:857–868. [PubMed: 17943193]
6. Kiryluk K, Martino J, Gharavi AG. Genetic susceptibility, HIV infection, and the kidney. Clin J Am Soc Nephrol 2007;2 Suppl 1:S25–35. [PubMed: 17699508]
7. Kullo IJ, Ding K. Mechanisms of disease: The genetic basis of coronary heart disease. Nat Clin Pract Cardiovasc Med 2007;4:558–569. [PubMed: 17893684]
8. Agrawal S, Khan F. Human genetic variation and personalized medicine. Indian J Physiol Pharmacol 2007;51:7–28. [PubMed: 17877289]
9. Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol 2006;16:368–373. [PubMed: 16679011]

10. Sela M, White FH Jr, Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. Science 1957;125:691–692. [PubMed: 13421663]
11. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 1958;181:662–666. [PubMed: 13517261]
12. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Res 2006;34:D227–230. [PubMed: 16381852]
13. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3:265–274. [PubMed: 12230035]
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410. [PubMed: 2231712]
15. Matsuoka H, Iwata N, Ito M, Shimoyama M, Nagata A, Chihara K, Takai S, Matsui T. Expression of a kinase-defective Eph-like receptor in the normal human brain. Biochem Biophys Res Commun 1997;235:487–492. [PubMed: 9207182]
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]
17. Kasuya A, Thornton JM. Three-dimensional structure analysis of PROSITE patterns. J Mol Biol 1999;286:1673–1691. [PubMed: 10064723]
18. Jonassen I, Eidhammer I, Grindhaug SH, Taylor WR. Searching the protein structure databank with weak sequence patterns and structural constraints. J Mol Biol 2000;304:599–619. [PubMed: 11099383]
19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–141. [PubMed: 14681378]
20. 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. [PubMed: 9847196]
21. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 1998;95:5857–5864. [PubMed: 9600884]
22. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 2002;30:242–244. [PubMed: 11752305]
23. Attwood TK, Beck ME. PRINTS--a protein motif fingerprint database. Protein Eng 1994;7:841–848. [PubMed: 7971946]
24. Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. Nucleic Acids Res 1998;26:323–326. [PubMed: 9399865]
25. Gracy J, Argos P. DOMO: a new database of aligned protein domains. Trends Biochem Sci 1998;23:495–497. [PubMed: 9868374]
26. Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. Nucleic Acids Res 1991;19:6565–6572. [PubMed: 1754394]
27. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 2000;28:228–230. [PubMed: 10592233]
28. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 2003;31:3701–3708. [PubMed: 12824398]
29. 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. [PubMed: 7723011]
30. Ananthalakshmi P, Kumar Ch K, Jeyasimhan M, Sumathi K, Sekar K. Fragment Finder: a web-based software to identify similar three-dimensional structural motif. Nucleic Acids Res 2005;33:W85–88. [PubMed: 15980587]
31. Balamurugan B, Roshan MN, Michael D, Ambaree M, Divya S, Keerthana H, Seemanthini M, Sekar K. SMS: sequence, motif and structure--a database on the structural rigidity of peptide fragments in non-redundant proteins. In Silico Biol 2006;6:229–235. [PubMed: 16922686]

32. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524. [PubMed: 8019422]
33. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323. [PubMed: 1409577]
34. Balla S, Thapar V, Verma S, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shinn JH, Mohler WA, Maciejewski MW, Gryk MR, Piccirillo B, Schiller SR, Schiller MR. Minimotif Miner: a tool for investigating protein function. Nat Methods 2006;3:175–177. [PubMed: 16489333]
35. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630. [PubMed: 12824381]
36. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31:3635–3641. [PubMed: 12824383]
37. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics 2007;23:950–956. [PubMed: 17387114]
38. Roque AC, Lowe CR. Lessons from nature: On the molecular recognition elements of the phosphoprotein binding-domains. Biotechnol Bioeng 2005;91:546–555. [PubMed: 15959902]
39. Ettmayer P, France D, Gounarides J, Jarosinski M, Martin MS, Rondeau JM, Sabio M, Topiol S, Weidmann B, Zurini M, Bair KW. Structural and conformational requirements for high-affinity binding to the SH2 domain of Grb2(1). J Med Chem 1999;42:971–980. [PubMed: 10090780]
40. Cho S, Velikovsky CA, Swaminathan CP, Houtman JC, Samelson LE, Mariuzza RA. Structural basis for differential recognition of tyrosine-phosphorylated sites in the linker for activation of T cells (LAT) by the adaptor Gads. Embo J 2004;23:1441–1451. [PubMed: 15029250]
41. Kimber MS, Nachman J, Cunningham AM, Gish GD, Pawson T, Pai EF. Structural basis for specificity switching of the Src SH2 domain. Mol Cell 2000;5:1043–1049. [PubMed: 10911998]
42. Ogura K, Tsuchiya S, Terasawa H, Yuzawa S, Hatanaka H, Mandiyan V, Schlessinger J, Inagaki F. Solution structure of the SH2 domain of Grb2 complexed with the Shc-derived phosphotyrosine-containing peptide. J Mol Biol 1999;289:439–445. [PubMed: 10356320]
43. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–55. 29–32. [PubMed: 8744573]
44. Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 1968;6:1425–1436. [PubMed: 5685102]
45. Newby Z, Lee TT, Morse RJ, Liu Y, Liu L, Venkatraman P, Santi DV, Finer-Moore JS, Stroud RM. The role of protein dynamics in thymidylate synthase catalysis: variants of conserved 2′-deoxyuridine 5′-monophosphate (dUMP)-binding Tyr-261. Biochemistry 2006;45:7415–7428. [PubMed: 16768437]

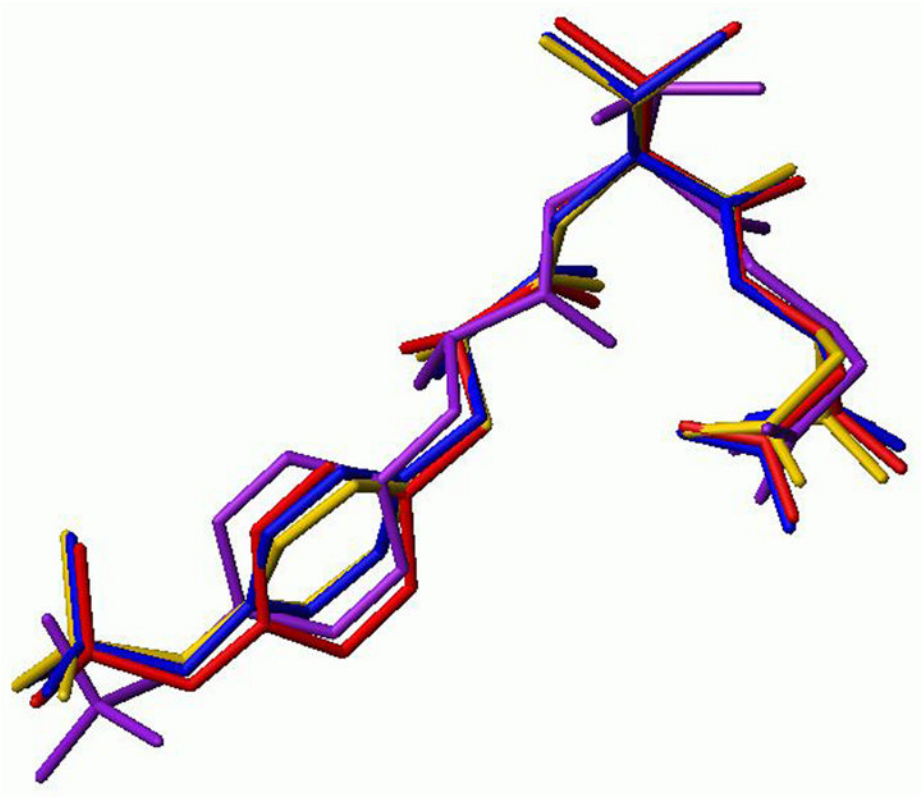

Figure 1.
Structural overlay of the pYXN motif for several different SH2-YXN complexes. In red, Src (1F1W (41)); in yellow, GADS (1R1P (40)); in blue, Grb2 structure determined by X-ray diffraction (1BM2 (39)); in magenta, Grb2 structure determined by NMR (1QG1 (42)).

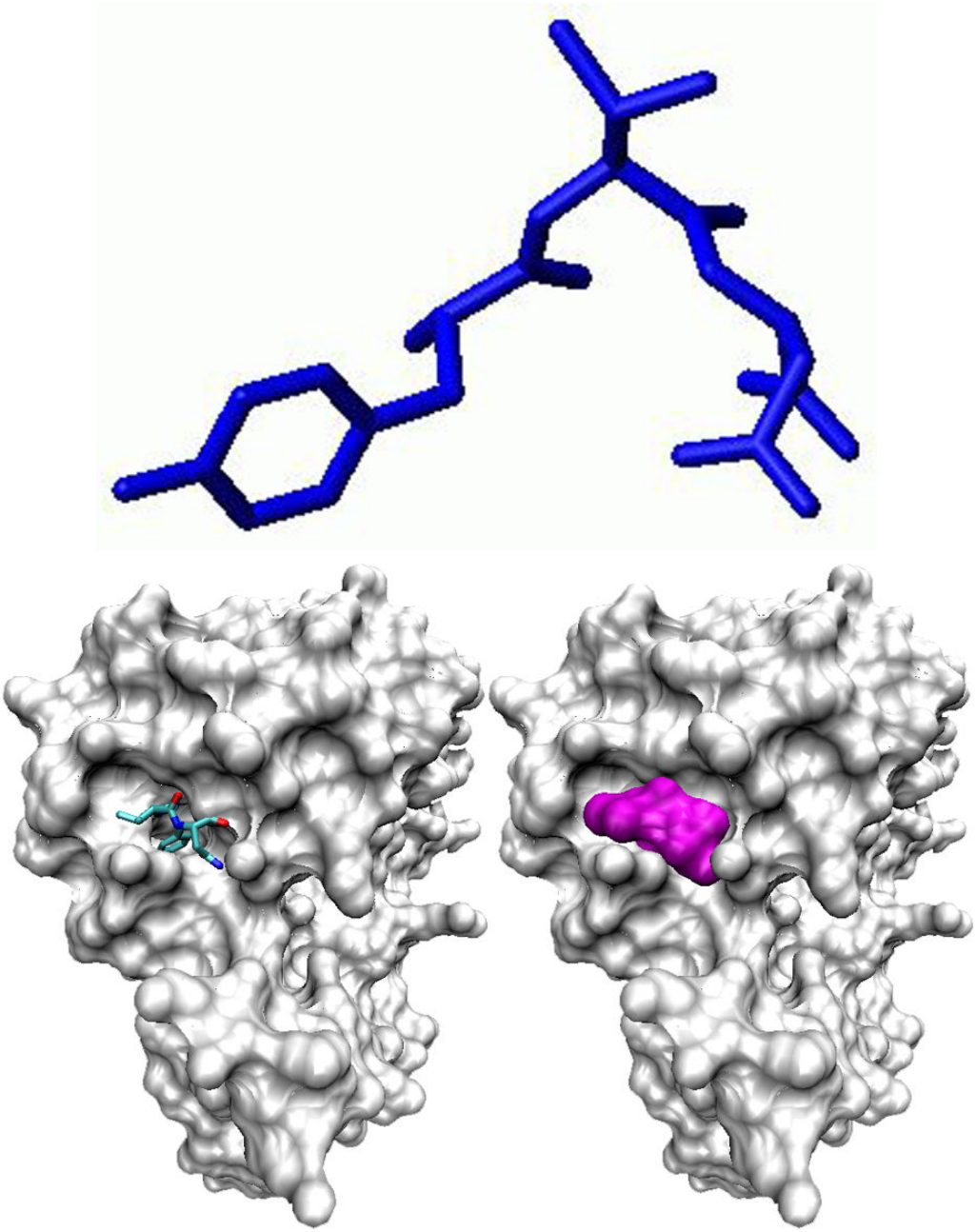

Figure 2.
A. Structure of YVN in thymidylate synthase. B. Solvent exposure of YVN of thymidylate synthase. Note the sidechain moieties of Y and N are both pointed into the interior of the protein and are therefore not recognizable from the surface.

Table 1

Bioinformatic tools for identifying various protein 'motifs'

Tool	URL	Ref	Motif Type	Underlying Method
Prosite	www.expasy.ch	(12)	Patterns: functional Profiles: structural	MSA with annotation
PFam	pfam.sanger.ac.uk	(20)	HMMs	MSA with annotation
SMART	smart.embl-heidelberg.de	(21)	Profiles/HMMs	MSA with annotation
PRINTS	www.bioinf.manchester.ac.uk/dbbrowser/sprint	(23)	Multiple Patterns	MSA with annotation
BLOCKS	blocks.fhcrc.org	(26)	Small Patterns	Fully-automated MSA
ProDom	prodom.prabi.fr/	(24)	Pre-aligned domains	Fully-automated MSA
DOMO	iubio.bio.indiana.edu/	(25)	Pre-aligned domains	Fully-automated MSA
GlobPlot	globplot.embl.de	(28)	Structural	Heuristic
SCOP	scop.mrc-lmb.cam.ac.uk/scop	(29)	Structural	Sequence similarity to annotated domain families identified demonstrating structural similarity
Scansite	scansite.mit.edu	(36)	Probability Matrix	Experimentally determined matrix
ELM	elm.eu.org	(35)	Pattern (< 10 aa)	Experimentally validated consensus
MnM	mnm.engr.uconn.edu	(34)	Pattern (< 10 aa)	Experimentally validated consensus

Table 2

Frequency of Occurrence of YXN motif within protein sequences in the PDB

# YXN's per protein	# proteins in PDB	Total # of YXN's
1	16115	16115
2	4333	8666
3	1129	3387
4	363	1452
5	177	885
6	21	126
7	29	203
8	2	16
9	0	0
10	0	0
11	1	11
>11	0	0
Total	22170	30186

Table 3

Statistics of YXN representations when selecting for particular geometric relationships

Selection Criteria	Total Observed	Average CA-CA distance (Angstroms)	Average Pseudo-torsion angle (°)
None	30341	7.31 ± 2.10	96 ± 48
CACB_Torsionangle >10	29087	7.24 ± 2.10	100 ± 45
CACB_Torsionangle <70	10080	8.59 ± 1.88	38 ± 20
CACB_Torsionangle >10 and <70	8826	8.52 ± 1.91	31 ± 17
CA-CA distance < 7	14694	5.49 ± 0.54	115 ± 40
CA-CA distance < 7 & CACB_Torsionangle >10	14509	5.48 ± 0.54	117 ± 38
CA-CA distance < 7 & CACB_Torsionangle <70	2280	5.79 ± 0.51	43 ± 20
CA-CA distance < 7, CACB_Torsionangle >10 & <70	2095	5.78 ± 0.50	46 ± 17
PTYR Bound Peptide	18	5.36± 0.28	45 ± 9