Cell. Mol. Life Sci. (2009) 66:3625–3639
DOI 10.1007/s00018-009-0117-0

Cellular and Molecular Life Sciences

REVIEW

The protein meta-structure: a novel concept for chemical and molecular biology

Robert Konrat

Received: 9 April 2009 / Revised: 20 July 2009 / Accepted: 21 July 2009 / Published online: 19 August 2009

Abstract The ultimate goal of bioinformatics or computational chemical biology is the sequence-based prediction of protein functionality. However, due to the degeneracy of the primary sequence code there is no unambiguous relationship. The degeneracy can be partly lifted by going to higher levels of abstraction and, for example, incorporating 3D structural information. However, sometimes even at this conceptual level functional ambiguities often remain. Here a novel conceptual framework is described (the protein meta-structure). At this level of abstraction, the protein structure is viewed as an intricate network of interacting residues. This novel conception offers unique possibilities for chemical (molecular) biology, structural genomics and drug discovery. In this review some prototypical applications will be presented that serve to illustrate the potential of the methodology.

Keywords Meta-structure · Structural biology · Protein structure · Intrinsically unfolded protein · Disordered proteins · Drug discovery · Structural genomics

Introduction

Proteins are nature's robots and play pivotal roles in all living processes [1]. They are uniquely versatile in their capability to perform chemical transformations under mild conditions endurable for living organisms. The knowledge about underlying mechanistic principles of protein chemistry is increasingly being harnessed in modern molecular biology research and exploited to serve the demands of modern society (e.g., health care and biotechnology). Despite the fact that proteins are macromolecules, giant assemblies of small building blocks, they are like other molecules uniquely defined by their atomic composition and the way the individual components are linked together. To account for this enormous complexity, Linderstrom-Lang invented his classical description schemes in which he proposed a hierarchical system with four levels: primary, secondary, tertiary and quaternary structure. At the simplest level the primary structure denotes just the linear amino acid sequence. The secondary structure is the next 'level up' from the primary structure and is the local folding of regions within the polypeptide chain and primarily determined by hydrogen bonding patterns between the carbonyl oxygens and amide hydrogens of the peptide bond. The tertiary structure is the native (3D) conformation and results from a delicate balance between enthalpic and entropic contributions to the thermodynamic stabilities of the unfolded and folded state. The final level is the quaternary structure, the combination of individual polypeptide chains to an oligomeric arrangement (e.g., starting with the groundbreaking work of Perutz and coworkers on hemoglobin) [2]. Although Anfinsen demonstrated in his seminal work that the primary sequence contains all the information required for protein structure and function [3], the folding code (the relationship between primary sequence and 3D structure) is still elusive. It is common belief that increasing the level of complexity (from primary to quaternary) provides a more comprehensive understanding of protein functionality. As a result, in the past major investments were devoted to the elucidation of the molecular mechanisms of protein functions, which culminated in September 2000, when the National Institute

R. Konrat (✉)
Department of Structural and Computational Biology,
Max F. Perutz Laboratories, University of Vienna,
Vienna Biocenter Campus 5, 1030 Vienna, Austria
e-mail: robert.konrat@univie.ac.at

of General Medical Sciences launched the structural genomics initiative aiming at the determination of the 3D structures of approximately 10,000 proteins representative for the protein structure manifold. However, despite tremendous progress in the past, sometimes even at this conceptual level (3D structure) functional ambiguities often remain. Here a novel conceptual framework is presented (protein meta-structure). At this level of abstraction, the protein structure is viewed as an intricate network of interacting residues. This novel conception offers unique possibilities for chemical (molecular) biology, structural genomics and drug discovery. In this review some prototypical applications will be presented that serve to illustrate the potential of the methodology. For example, the performance of high-throughput structural genomics relies on efficient target selection and rapid refinement of protein constructs. Unstructured regions of proteins lacking well-defined structural elements are the most important causes for protein crystallization failures. Here applications of this novel concept for recognizing intrinsically unstructured regions in proteins are presented. The results obtained on protein targets of several organisms suggest that this approach represents a general in silico method for high-throughput target analysis with significant potential for structural genomics. Additionally, this novel approach reveals structural (and functional) similarities hidden and undetectable on the primary sequence level. Finally, to illustrate applications to rational drug design, examples will be given for the potential of the methodology to identify novel ligand scaffolds for protein targets solely based on the primary sequence of the target.

The meta-structure concept

The method for prediction of protein compactness and local structural features has been introduced recently [4]. In brief, the 3D structural information (encoded by the coordinate triples for the individual atoms) was transformed into topological space by calculating the network of residue interactions. In this network structure a node refers to a particular residue and edges indicate the existence of neighbourhood relationships. Two residues are considered as neighbors if the Cα–Cα distances are below a distance cutoff (typically 8 Å). The mutual topological relationship θ of two residues is quantified by the shortest path length connecting the two residues in the network (Fig. 1). The mutual topological relationship θ (θ is the shortest path length) between two residues (A, B) characteristically depends on the amino acid types (A, B) and their primary sequence distance, l_{AB}. A pictorial description of θ is shown in Fig. 1d. This characteristic topological relationship is statistically evaluated following a well-established statistical procedure [5] using the PDB subset of Bax and co-workers [6] and stored as pairwise statistical distribution functions $\rho(\theta, A, B, l_{AB})$. The PDB dataset comprises only single polypeptide chains and proteins of negligible structural similarities and sequence homologies (more details can be found in [6]). Most importantly, it was found that the subset of protein structures of Bax and co-workers was sufficient to reach statistical convergence and that increasing the number of protein structures in the dataset does not significantly change the pairwise distribution functions. It should be noted that for the derivation of $\rho(\theta, A, B, l_{AB})$, only well-structured parts of the different protein structures were considered. Structural disorder is thus entirely absent in the dataset. It also has to be emphasized that the meta-structure approach does not involve a protein training set, since the essential information (the shortest path length θ) is extracted from (a set of) available protein 3D structures following a precisely defined mathematical operation (a transformation from Cartesian coordinate space into topological space). Typical $\rho(\theta, A, B, l_{AB})$ are shown in Fig. 2. The statistical distribution functions $\rho(\theta, A, B, l_{AB})$ can subsequently be used to predict topological information solely based on the primary sequence. The input primary sequence is used to predict for each possible amino acid pair (of residue type A and B and separated by l_{ab} in the sequence) in the protein and based on $\rho(\theta, A, B, l_{AB})$ an average topology parameter, $d_{ij} = \Sigma\theta \times \rho(\theta, A, B, l_{AB})$, considering all possible θ values. The topological parameter d_{ij} is directly related to the most probable shortest path length between residues i and j. The subsequent summation (after normalization) leads to the compactness value $C_i = N_1 \times \Sigma(1/d_{ij}) - N_0, \ \forall(i \neq j)$, which is related to the inverse local residue exposure. N_0 and N_1 are empirical scaling factors that ensure that highly exposed residues display C_i values close to 0. Large C_i values are typically found for residues located in stable parts of the protein and on average deeply buried in the interior of the structure. In contrast, small values are found for flexible loop regions and/or intrinsically unfolded segments of the polypeptide chain. The prediction of local secondary structure elements is performed by applying only next neighbour distribution functions (restricted to primary sequence differences between residue pairs $\leq$4). The local secondary structure parameter S_i is defined as $S_i = {}^{\alpha}P_i - {}^{\beta}P_i$, where ${}^{\alpha}P_i$ and ${}^{\beta}P_i$ are defined as: ${}^{\alpha}P_i = N_A \times \gamma_2 \times \gamma_3 \times \gamma_4$, ${}^{\beta}P_i = N_B \times \delta_3 \times \delta_4$. $\gamma_i = \rho(1,i) \times (1.0 - \rho(2,i)/[\rho(2,i) \times (1.0 - \rho(1,i)]$; $\delta_i = \gamma_i^{-1}$. $\rho(1,i)$ and $\rho(2,i)$ are the probabilities of finding a shortest path length value of 1 and 2 between residues separated by i positions in the primary sequence. N_A and N_B are global scaling factors that ensure comparable magnitudes of compactness C_i and secondary structure S_i values. Typically, residues located in α-helices display positive S_i values, whereas for

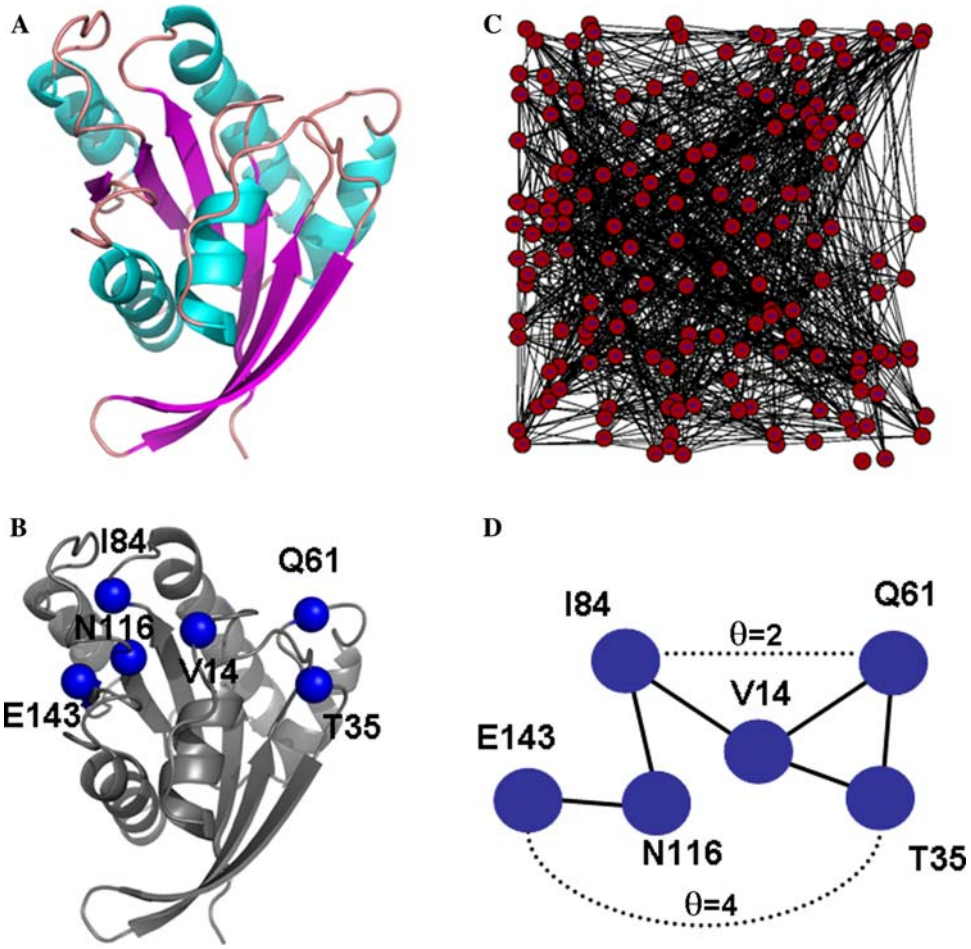

Fig. 1 The protein meta-structure concept. The 3D structural information is transformed into topological space by calculating the network of residue interactions. In this network structure a node refers to an amino acid and edges indicate the existence of neighborhood relationships. Two residues are considered as neighbors if the Cα–Cα distances are below a distance threshold. (**a**) 3D structure of H-ras p21 [52]. (**b**) Location of the network comprising Q61–T35–V14–I84–N116–E143 of H-ras p21. (**c**) Topological network of H-ras p21 calculated from the 3D structure (5p21; Cα–Cα distance cutoff: 8 Å). (**d**) Subset of the topological network of H-ras p21 (by zooming into the network graph of Fig. 1**c**) demonstrating the topological relationship (shortest path length θ) between residues. Solid lines indicate the spatial relationship between residues (Cα–Cα distance below 8 Å). The topological relationships θ between different nodes/residues in the network graph are defined as the shortest path lengths θ (*dashed lines*) through the network (see text). For example, Q61 and I84 are connected via a single node (V14) and thus yielding the shortest path length of $\theta = 2$, whereas T35 and E116 are linked via three nodes ($\theta = 4$). The figure was created using the programs pymol (http://www.pymol.org) and Visone 1.1.1 (http://www.visone.de)

residues located in extended regions (β-strands) significantly smaller (negative) values were observed. It has to be emphasized that the sequence analysis provides quantitative information about the local secondary structure and residue compactness for each residue position. Most importantly and in contrast to existing sequence analysis tools, the new approach reveals how a particular local structural element is embedded in the context of the 3D fold, either deeply buried in the interior of the structure or located on the surface and thus exposed to the solvent. The combination of quantitative secondary structure values and residue compactness is called the meta-structure. The term meta-structure also reflects the fact that in this (relational) approach, individual residues are not treated as independent units but rather building blocks that form an intricate network displaying significant interdependence. As a prototypical example, Fig. 3 shows the meta-structure analysis of the N-terminal SH2 domain of the PI3-kinase p85. As can be seen from the figure, there is a convincing agreement between predicted secondary structural elements and the location in the 3D structure. In addition the spatial distribution of secondary structure elements is correctly identified; β-strands deeply buried in the interior of the 3D fold display larger compactness values, whereas in contrast surface exposed α-helices are indicated by small compactness values. In this particular SH2 domain fold, adjacent secondary structure elements are characterized by an alternating inside-out behavior, from which a mixed α/β fold with a core β-sheet

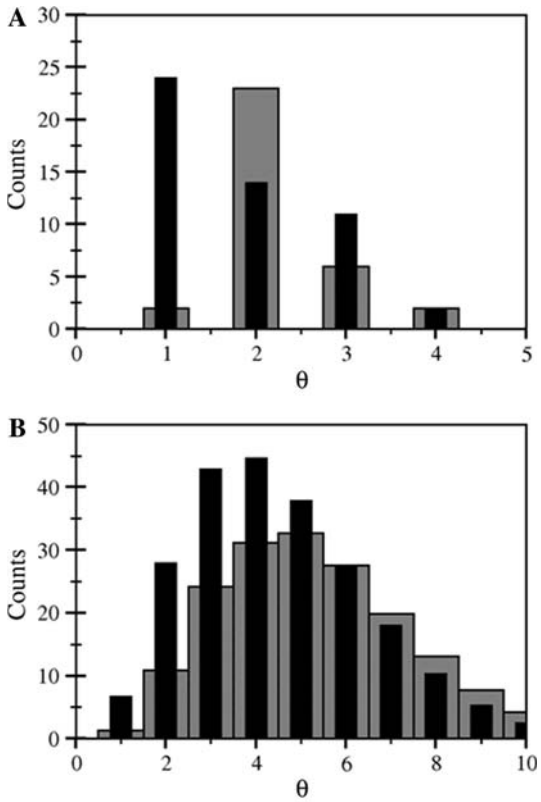

Fig. 2 Typical pairwise distribution functions $\rho(\theta, \text{A}, \text{B}, l_{\text{AB}})$ extracted from the PDB database. (**a**) The distribution of shortest path lengths observed for residue pairs separated by four residues in the primary sequence are shown, (*grey*: $\rho(\theta$, Asp, Glu, 4); *black*: $\rho(\theta$, Ile, Leu, 4). (**b**) Long-range (primary sequence difference $l_{\text{AB}} \geq 5$) shortest path length distributions (*grey*: $\rho(\theta$, Asp, Glu, n); *black*: $\rho(\theta$, Ile, Leu, n). (In **b** the counts are divided by 1,000). It can be seen that Ile–Leu pairs are typically clustered in protein structures (smaller θ values), whereas Asp–Glu tends to be more distant (larger θ values)

and flanking α-helices can be deduced. Of course the meta-structure does not provide (highly refined) information about the 3D protein structure with atomic resolution. The information content of the meta-structure is rather of topological nature as it gives a (quantitative) measure of how the individual secondary structure elements are embedded in the 3D fold. The example with the SH2 domain illustrates how the meta-structural features can be transformed into protein fold information.

Results

Identification of protein disorder

It is now well accepted that an increasing number of proteins are lacking stably folded tertiary structures [7–9] and that this inherent protein conformational flexibility has significant impact on biological functionality (e.g., molecular recognition events). Estimations were made suggesting that up to nearly 50% of proteins from eukaryotic organisms comprise conformationally flexible (disordered) amino acid stretches longer than about 30 residues [10]. This conformational flexibility requires us to reassess the central structure-function paradigm in molecular structural biology [11, 12]. A great challenge in structural proteomics is thus the rapid and facile identification of conformational flexibility and the subsequent exploitation of this information for protein construct optimizations [13]. For the sequence-based prediction of protein disorder, several approaches have been developed (primarily based on neural network approaches) [14–19]. The meta-structure approach differs from these bioinformatics approaches in the sense that it provides quantitative information about compactness of the polypeptide chain and local secondary structure (and not folded/unfolded) defined on a per residue basis, which can be used to analyze the fine details of protein folds with residue resolution.

To investigate the potential of the meta-structure approach as a general in silico method for high-throughput target analysis, a large-scale survey of sequences from the PDB database was performed. As a global measure for protein foldedness, the average compactness value of the protein (*A*verage *R*esidue *C*ompactness, ARC $= 1/N \Sigma C_i$, N being the number of residues) was used. On average the ARC value for a well-folded protein is about 300. Significantly smaller values (ARC < 200) are found for structurally flexible proteins (intrinsically disordered/unstructured), whereas considerably larger values (ARC > 400) are typically observed for membrane proteins or oligomeric proteins. The average residue compactness approach was applied to the protein sequences of several organisms from different kingdoms (archea, prokaryote and eukaryotes). As observed earlier by neural network based disorder predictors, a significant number of proteins exist as intrinsically unstructured/disordered proteins (IUPs) [10, 14–19]. Consistent with these findings we found a significant fraction of proteins displaying ARC values smaller than 200 (or comprising longer stretches of locally disordered residues), and thus they most likely do not form stable tertiary structures in solution. A summary of the results is given in Table 1. Overall the results agree very well with previous analysis based on position specific score matrices [20]. Lower organisms have only a few unstructured proteins, IUPs/NUPs, (archea: 1.7% and prokaryotes: 1.7–3.5%), whereas for eukaryotes a significant fraction of the proteome falls into this category (13.9–21.5%). The results obtained on protein of several organisms suggest that this approach represents a general in silico method for high-throughput sequence analysis of IUPs/NUPs in different genomes. It is important to note, however, that the meta-structure approach provides per residue information about foldedness and/or residue compaction and thus goes beyond a binary (folded/

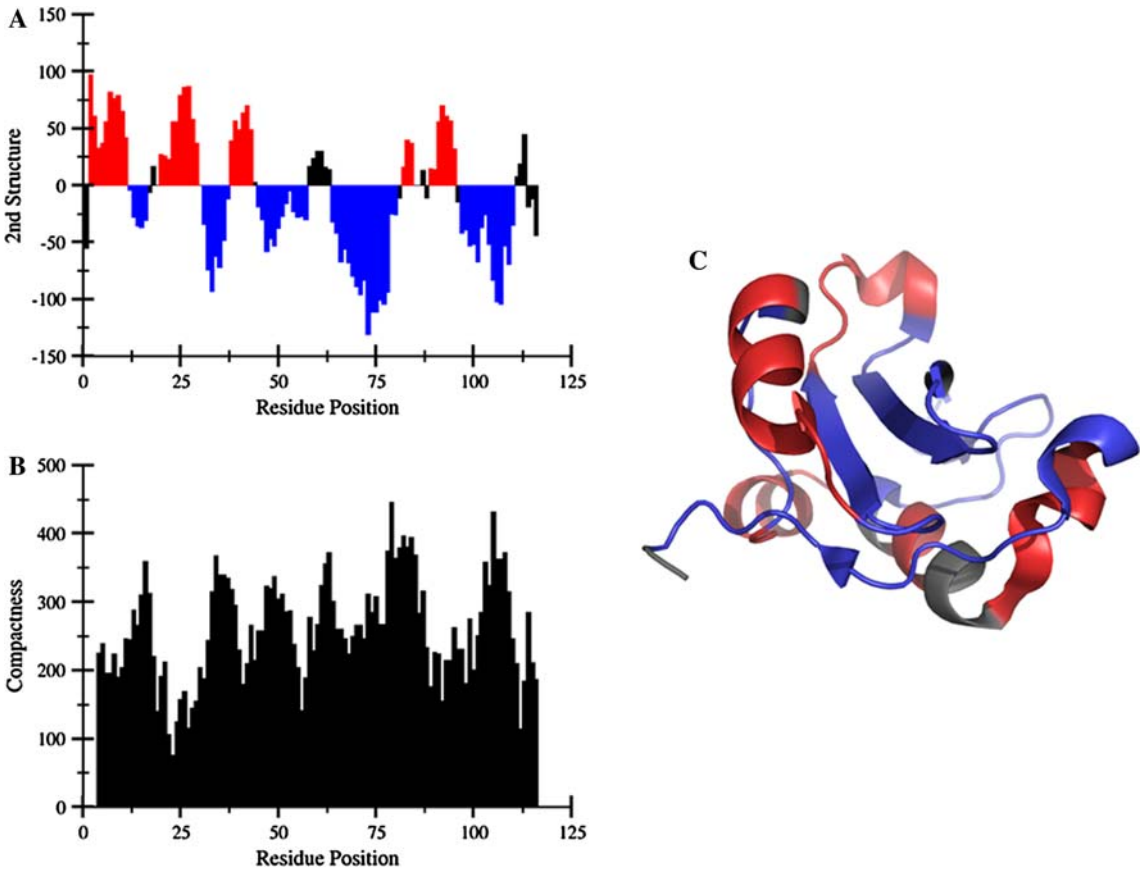

Fig. 3 The protein meta-structure. Residue secondary structure and compactness plot of the PI3-kinase p85 N-terminal SH2 domain. (PDB: 2IUG). Comparison of predicted local secondary structural features (**a**) and compactness (**b**) as a function of residue position, and 3D protein structure (**c**). Positive secondary structure values are indicative of α-helical segments (shown in *red*). In contrast, continuous negative values are typical for extended or β-strand regions (shown in *blue*). Residues of loosely defined secondary structure are shown in black. Large compactness values indicate residue positions typically buried in the interior of the 3D structure, whereas small values are found for residues exposed to the solvent. In the 3D structure of the SH2 domain (**c**) residues are color coded according to the meta-structure (local secondary structure) results (α-helix: *red*; β-sheet: *blue*). The figure was prepared using the program pymol (http://www.pymol.org)

unfolded) descriptor scheme. The availability of residue-specific compactness and secondary structure information allows for a detailed (sequence specific) analysis of the overall and local structural propensities of natively unstructured proteins. Fig. 4 shows examples for intrinsically unstructured proteins from different kingdoms as identified by the meta-structure approach. Interestingly, the orthology relationship between the two Prefoldin subunits from *Methanobacterium thermoautotrophicus* and *C. Glabrata* is clearly visible in the meta-structure analysis.

Sequence-based protein construct optimization for structure determination purposes

Structural genomics is undoubtedly a key of modern proteomics. The performance of high-throughput structural genomics, however, relies on efficient target selection and rapid refinement of protein construct definition. Unstructured regions of proteins lacking well-defined structural elements are the most important causes for protein crystallization failures. NMR spectroscopy, as an alternative, also suffers from protein disorder as flexible parts of a protein tend to give rise to severe spectral overlap, which impedes signal and NOE assignments, respectively. As a prototypical example for residue specific protein construct refinement applications, we present data obtained on ICln, a cytosolic protein that modulates volume-regulated anion current associated with hypotonic cell swelling [21]. The average ARC value of ICln was 221, thus indicative of a reasonably folded polypeptide chain, although the ARC values are somehow smaller than average values typical for proteins in the PDB. The residue plot (C_i vs. residue position, Fig. 5), however, allowed for a detailed analysis of the conformational preferences of the polypeptide and

Table 1 Genome-wide protein meta-structure analysis. Overview of intrinsically unstructured (natively unfolded) proteins (IUP/NUP) in different organisms. The definition of IUP/NUP is based on the average residue compactness (ARC). A protein is annotated as IUP/NUP if the global ARC is smaller than 200 or alternatively 30 consecutive residues display a local average compactness value smaller than 150

Organism	Kingdom	Size	% IUPs/NUPs
Methanobacterium thermoautotrophicum	Archea	1,866	1.7
Thermotoga maritima	Prokaryote	1,852	2.8
E. coli	Prokaryote	4,336	1.6
Listeria monocytogeneso	Prokaryote	2,842	3.4
Streptococcus pneumoniae	Prokaryote	2,075	3.2
Yarrowia lipolytica	Eukaryote	6,485	18.8
Yeast	Eukaryote	6,430	19.1
Candida glabrata	Eukaryote	5,146	20.2
Debaryomyces hansenii	Eukaryote	6,282	18.8
Kluyveromyces lactis	Eukaryote	5,286	18.2
Cryptococcus neoformans	Eukaryote	6,405	21.5
Human	Eukaryote	12,091	13.9

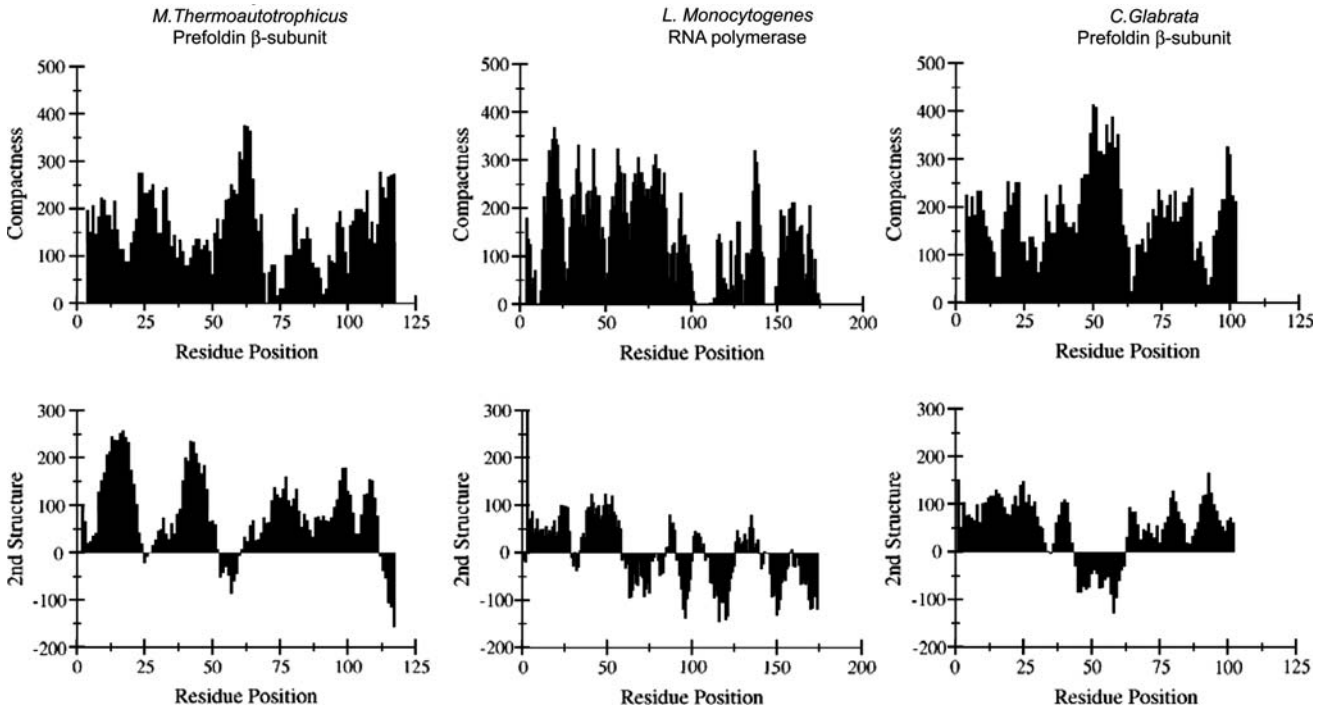

Fig. 4 Intrinsically unstructured proteins (IUP) from different kingdoms. A selection of prototypical IUPs/NUPs identified in archea (*left*), prokaryotes (*middle*) and eukaryotes (*right*) are shown. The following examples are shown: (*left*) ARC: 152.5, *Methanobacterium thermoautotrophicus*, SwissProt: O26774, Prefoldin-β-subunit; (*middle*) ARC: 135.3, *Listeria monocytogenes*, SwissProt: Q8Y494, Probable DNA-directed RNA polymerase; (*right*) ARC: 177.7, *Candida Glabrata*, SwissProt: Q6FY96, Prefoldin-β-subunit. The IUPs/NUPs have been identified based on the average residue compactness (ARC) approach (see text). A protein is annotated as IUP/NUP if the global ARC is smaller than 200 or alternatively 30 consecutive residues display a local average compactness value smaller than 150. Predicted compactness (*upper part*) and local secondary structure (*lower part*) are shown. Positive secondary structure values are indicative of α-helical segments, whereas continuous negative values are typical for extended conformations (β-strand or polyproline II)

revealed distinctly different areas in ICln. Most importantly, residues 85–105 showed considerably smaller compactness values compared to other parts of the molecule. Additionally, residues located in the C-terminal region (adjacent to the C-terminal α-helix) also display significantly smaller compactness values. We have recently determined the solution structure of ICln by solution NMR spectroscopy [21]. ICln exhibits a pleckstrin homology

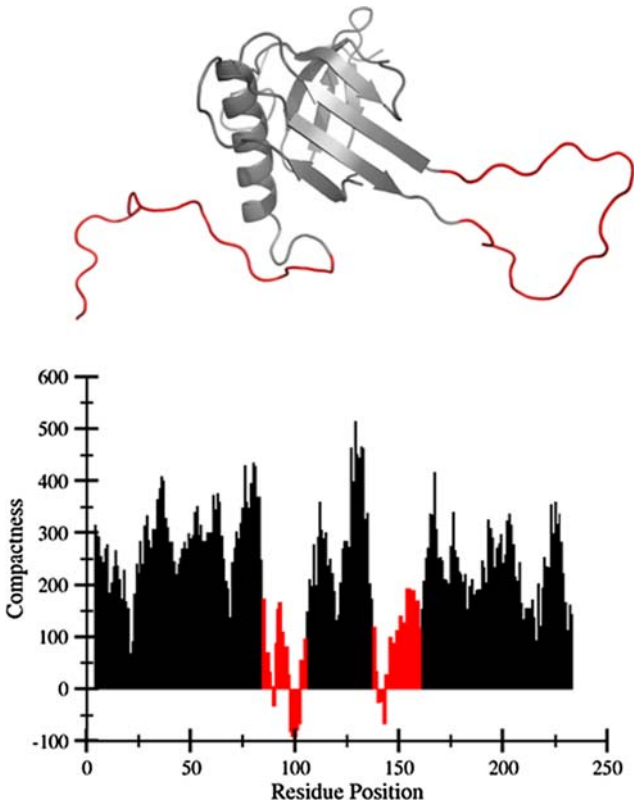

Fig. 5 *Top*: Solution structure of ICln [21]. Conformationally flexible parts of the protein are indicated in *red*. *Bottom*: Residue plot showing compactness C_i vs. residue position. The conformationally flexible parts (*red*) are correctly identified by significantly reduced compactness values. The conformational flexibility of these regions has been independently verified by NMR spin relaxation analysis [21]. The protein construct used for structure determination only comprised residue 1–165; the part adjacent to the C-terminal α-helix was missing in the construct for structure determination

(PH) domain topology. It consists of a pair of nearly orthogonal β-sheets of four and three anti-parallel strands forming a collapsed β-barrel capped on one side by a C-terminal α-helix. PH domains are a very large family of structurally homologous protein motifs of moderate to low sequence homology and are found in many proteins involved in signal transduction. The loop regions of PH domains vary considerably from protein to protein and presumably allows for the differential modes and specificities of ligand binding observed (e.g., predominantly binding of phosphatidylinositols). In the solution structure of ICln two regions with considerable internal dynamics were observed (loop region E98-D112 connecting β6 and β7, as well as the residue stretch following the C-terminal α-helix). The lack of stable conformations for these parts in ICln was experimentally validated by ^{15}N NMR relaxation measurements. Figure 6 shows an overlay of ^{15}N–^{1}H^{N} HSQC spectra of both wild-type ICln (black) and a truncated form lacking residues 98–112 (yellow), which were predicted to be highly flexible and largely disordered.

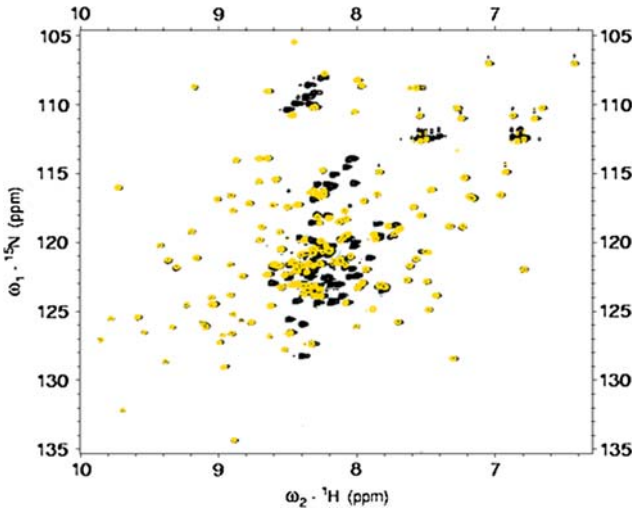

Fig. 6 Overlay of ^{15}N–^{1}H HSQC spectra for wild-type (*black*) [21] and truncated (elimination of residues 85–105 located in the flexible linker region, *yellow*) ICln, respectively. Overall cross peak positions are unchanged, thus indicating unchanged solution structures of both ICln constructs. Cross peaks in the *black* dataset, which are absent in the *yellow* dataset, correspond to residues in the flexible linker region (eliminated in the truncated version)

Overall the nearly unchanged correlation spectrum indicates that the elimination of the highly dynamic loop region does not alter the overall fold of the protein. However, the reduction of conformationally flexible parts in the protein may increase the likelihood of crystallization and thus the amenability for structure determination by X-ray protein crystallography.

Local structure assessment in intrinsically unstructured proteins

There is growing evidence that intrinsically or natively unstructured/unfolded proteins (IUP/NUP) display local structure formation and—although globally "unfolded"—exist as flexible polypeptides with distinct secondary structural features. This structural preformation is thought to be of functional relevance as these local structural motifs serve as recognition sites for biologically relevant protein–protein interaction events. The structural preformation of interaction motifs is particularly advantageous as it relieves the entropic penalty associated with the formation of protein complexes involving intrinsically flexible protein partners. In recent years NMR spectroscopy has found widespread application for the analysis of the structural dynamics of conformationally flexible proteins in solution [22, 23]. Several NMR parameters exist that sensitively probe the local secondary structure of proteins in solution, the most important being the chemical shift of backbone carbons ^{13}Cα and ^{13}Cβ, because of its straightforward experimental determination and interpretation. Recently, it

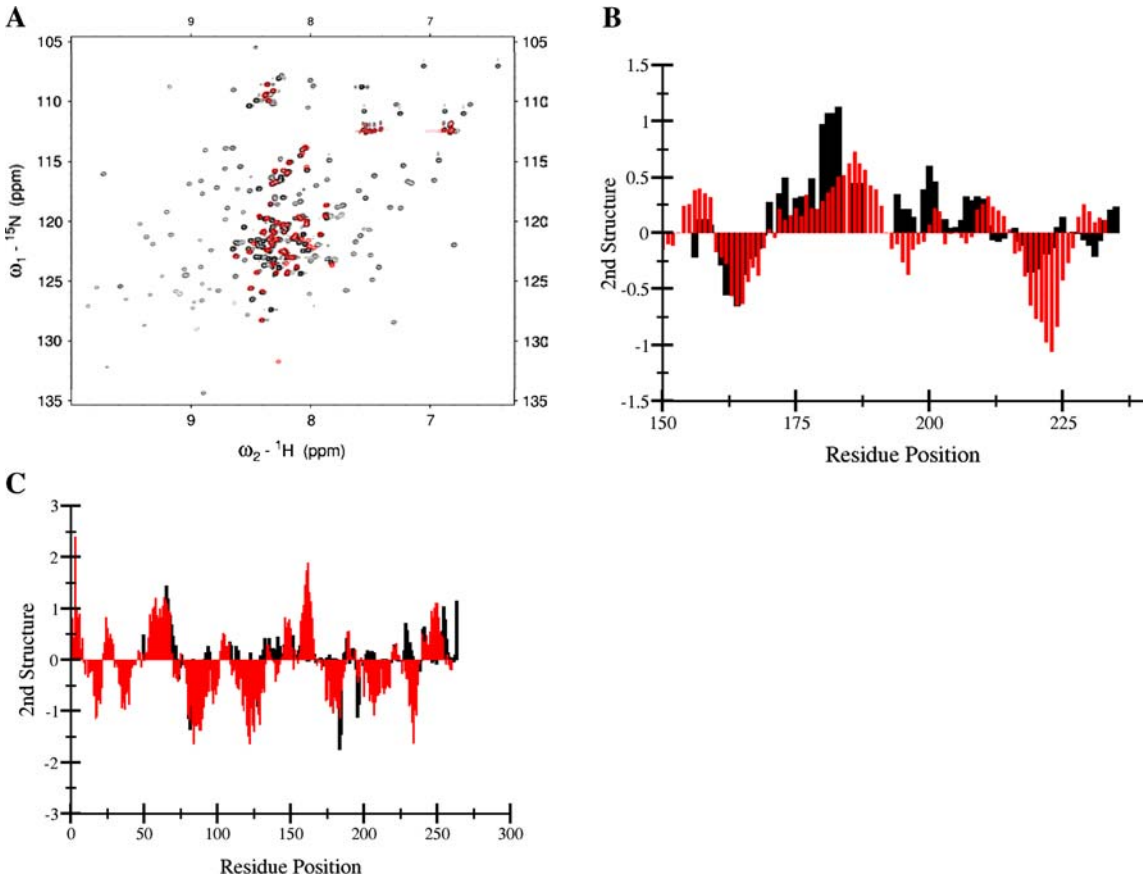

Fig. 7 Meta-structure-based assessment of local secondary structure in natively unfolded proteins. (**a**) Overlay of ^{15}N–^{1}H HSQC spectra for full-length ICln (1–237, *black*) and the C-terminal domain of ICln (158–237, *red*). The nearly identical peak positions in the C-terminal domain of ICln, CTD-ICln, (*red*) indicates that CTD-ICln exists as a largely unfolded polypeptide chain in solution. Comparison between (*red*) meta-structure and (*black*) NMR derived local secondary structure elements for (**b**) CTD-ICln and (**c**) Osteopontin [27, 28]. Positive and negative values indicate the existence of local α-helical segments or β-strands. NMR results were obtained from $\Delta^{13}C\alpha$–$\Delta^{13}C\beta$ secondary shifts (see text).

was shown that also residual dipolar couplings (RDCs) and paramagnetic relaxation enhancement data (PRE) are versatile spectral probes for the characterization of the structural dynamics of intrinsically/natively unfolded proteins [24, 25]. The $^{13}C\alpha/^{13}C\beta$ chemical shift analysis is based on the deviation of experimental chemical shifts $^{13}C\alpha$[exp] from tabulated random coil values $^{13}C\alpha$[rc], ($\Delta^{13}C\alpha = {}^{13}C\alpha[\text{exp}] - {}^{13}C\alpha[\text{rc}]$), so-called secondary chemical shifts ($\Delta^{13}C\alpha$, $\Delta^{13}C\beta$). Typically, downfield/upfield shifts are found for $^{13}C\alpha/^{13}C\beta$ in α-helices, whereas the opposite (upfield/downfield) trend is observed in β-sheets. Because of this reciprocal behavior differential secondary shifts ($\Delta^{13}C\alpha - \Delta^{13}C\beta$) are used as an indicator for secondary structure formation. Positive difference secondary shifts are indicative for α-helices, whereas negative values are found for β-sheets [26]. A particular advantage of NMR spectroscopy for the investigation of IUPs/NUPs lies in the fact that NMR probes the time-averaged structural ensemble of proteins and thus reports on conformational equilibria.

The potential of meta-structure analysis for investigation of local structure preformation in natively unstructured proteins is illustrated with a comparison to NMR-derived local structure assignments. The following proteins were chosen: (1) the C-terminus of ICln [21] and (2) Osteopontin, a cytokine and cell attachment protein implicated in tumorigenesis [27, 28]. While the N-terminal domain ICln (1–145) forms a stably folded pleckstrin homology (PH) domain topology in solution, the C-terminal domain of ICln (158–237) appears to be natively unfolded. From the ^{15}N–^{1}H HSQC overlay of full-length ICln (1–237) and ICln N-terminal domain (1–145) (Fig. 7a), it can be concluded that the C-terminal extension of ICln does not interact with the autonomously folded N-terminal domain of ICln. A more detailed NMR analysis of the C-terminal domain (A. Schedlbauer, University of Vienna) demonstrated that the C-terminal extension of ICln appears to be unfolded under native conditions, but nonetheless displays distinct secondary structural features (α-helical segment between 170–190). The dynamic properties of the C-terminal

extension were independently corroborated by NMR spin relaxation. The meta-structure analysis correctly identified the C-terminal domain of ICln as a flexible polypeptide, as the average residue compactness value for this segment was calculated to 197.5, which is well below the PDB average of 300. Additionally, as can be seen from the comparison in Fig. 7b, the location of preformed secondary structure elements was correctly predicted by the meta-structure approach.

A second example is given with osteopontin (Fig. 7c). Avian fibroblasts simultaneously transformed by *v-myc* and *v-mil*(*raf*) display significantly elevated levels of the transcription factor complex AP-1, which subsequently leads to upregulation of specific target genes, such as the osteopontin (OPN) encoding gene (*OPN*), *126MRP*, and *rac2* [27]. OPN is an arginine–glycine–aspartate (RGD)-containing glycoprotein that was first identified as a major sialoprotein in bones [29, 30]. It comprises two receptor-recognition sites, a central integrin-binding domain and a C-terminal CD44-binding site, which are separated by a protease-hypersensitive cleavage site. Although quail OPN exists as an intrinsically unstructured protein in solution [28], significant ^{13}C secondary shifts were observed for several parts of the polypeptide chain (α-helix: Q59-S74, β-strands: E77-F87 and V117-R132). ^{15}N relaxation data also suggest that OPN is rather flexible and lacking a stable 3D structure in solution. The obtained average residue compactness ARC value of quail OPN was found to about 172, and thus supports the NMR-derived inherent conformational flexibility of OPN. Again, as can be seen from the comparison in Fig. 7c, the location of preformed secondary structure elements (α-helix and β-strands) was correctly predicted by the meta-structure approach. Summing up, the data shown in Fig. 7 convincingly demonstrate the potential of the meta-structure concept to correctly identify natively unfolded proteins and also provide residue specific local secondary structure elements with high accuracy. Interestingly, even on a quantitative level the meta-structure predicted secondary structure values reasonably agree with NMR secondary chemical shifts, which additionally reflect the convincing accuracy and precision of the approach.

Protein meta-structure alignment

The availability of a quantitative (and sequence-based) scheme for representing the (meta) structural features of proteins provides a new and rich avenue for large-scale sequence comparison. Sequence alignment tools have offered tremendous opportunities and transformed molecular biology research. Identification of protein homologues is nowadays a routine task and an indispensable guiding principle in biological research. Despite its tremendous success in the past, sequence alignment fails in cases of low amino acid similarities. Nonetheless, even in these cases of insignificant sequence identity, structural and functional homologies may exist that remain unnoticed/undetectable by conventional methodology. The meta-structure approach has the potential to provide a solution to this important problem, as it reveals (meta-) structural similarities that are better correlated with functionality than primary sequence information. Meta-structure-based sequence alignment can be implemented in a straightforward manner. In brief, our strategy follows the dynamic programming algorithm of Needleman and Wunsch [31] for aligning sequences based on pairwise amino acid similarities. Instead of using well-established measures for amino acid similarities such as BLOSUM62 [32], calculated meta-structure parameters derived from the primary sequence (see above) are applied to define pairwise similarity matrices. For the calculation of a suitable scoring function, both secondary structure and compactness values are used. A residue–residue match was considered acceptable if the pairwise C_i and S_i differences were below an empirically optimized threshold value (ΔC_i: 250; ΔS_i: 150). For the quantification of alignments, stretches of 30 amino acids were considered. No penalty or moderate penalties were introduced if more than 27 (weight: 100) or between 15 and 27 residues (weight: 30) were considered as acceptable; otherwise, penalties were introduced (fewer than 15 matches: -25; no match: -50). The gap penalty value was optimized to 1.

The meta-structure alignment was applied to two sets of proteins, and the obtained results are shown in Fig. 8. The first example shows a comparison between the DNA mismatch repair protein PMS2 (PDB code: 1EA6) and TM1457, a protein of unknown function from *T. maritima* (PDB code: 1S12). These proteins were recently discussed to illustrate the difficulties of finding structural homologues [33]. Based on the 3D structure, TM1457 was described as a new fold [34]. The 3D structure alignment, however, (Fig. 8, top, right) clearly shows the significant structural similarity between these seemingly unrelated proteins [35, 36]. Despite that Tm1457 was classified as a new fold, it is obvious that the 3D structural motifs of TM1457 are present in the structure of PMS2. This significant similarity is correctly identified using the meta-structure approach.

The second example (Fig. 8, bottom) provides an application of the meta-structure approach to proteins of similar structure but distinct sequences. It shows the alignment of the A chain of a putative isomerase from *Rhodopseudomonas palustris* (Midwest Center of Structural Genomics, to be published, PDB: 3DM8) and the B chain of limonene-1,2-epoxide hydrolase from *Rhodococcus erythropolis* (PDB: 1NWW) [37]. Despite the low sequence similarity between the two proteins, the meta-structure alignment correctly identifies the structural similarity.

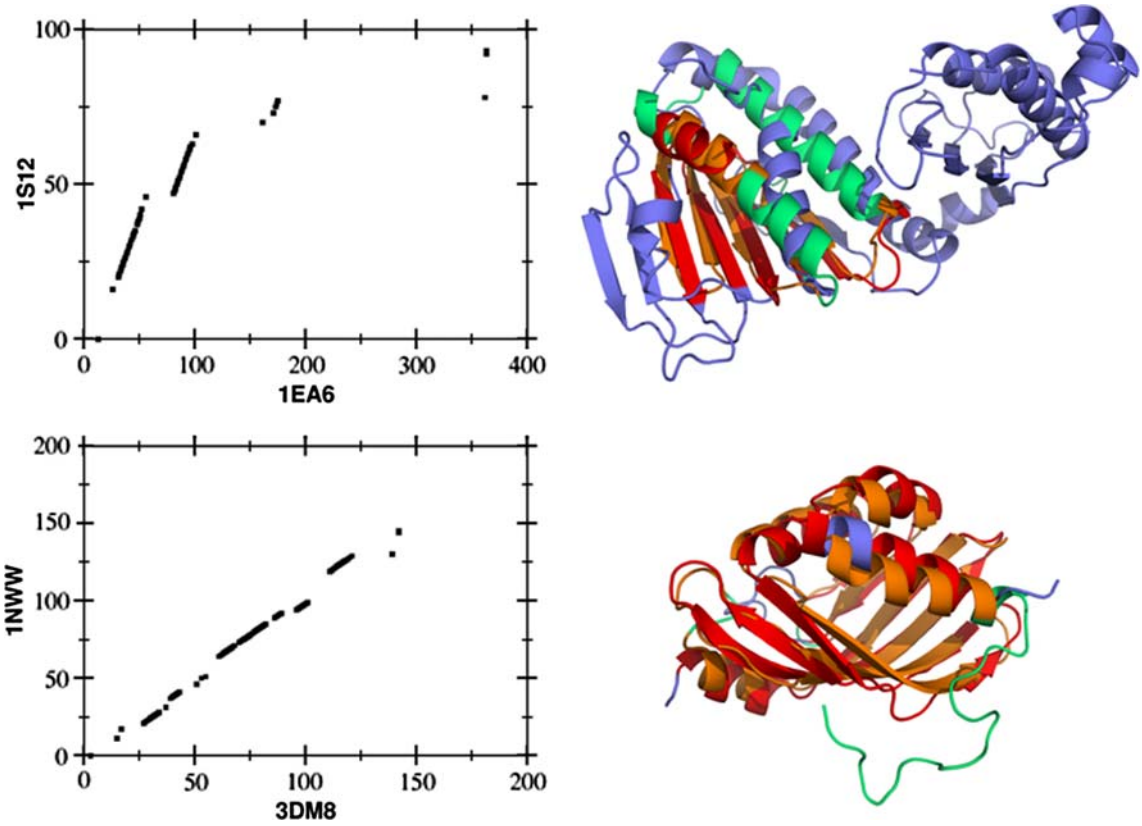

Fig. 8 Protein meta-structure alignment. The pairwise protein sequence alignment is based on calculated meta-structure parameters. The scoring function for obtaining the optimal sequence match involves compactness and secondary structure values. Comparison between meta-structure alignments (*left*) and structural superpositions (*right*). *Top* Meta-structure (*left*) and 3D structure alignment (*right*) of TM1457 (PDB:1S12) and the DNA mismatch repair protein PMS2 (PDB:1EA6). *Bottom*: Meta-structure (*left*) and 3D structure alignment (*right*) of the A chain of a putative isomerase from *Rhodopseudomonas palustris* (Midwest Center of Structural Genomics, to be published, PDB:3DM8) and the B chain of limonene-1,2-epoxide hydrolase from *Rhodococcus erythropolis* (PDB:1NWW) [38]. The structure alignment was performed with the program TopMatch [35, 36]. The protein structures are shown in *blue* (*top*: 1EA6, *bottom*: 3DM8) and *green* (*top*: 1S12, bottom: 1NWW), and the regions of similar structure are colored *red* and *orange*. The figure was generated using Pymol (http://www.pymol.org)

Identification of distant orthologs (cross-species homologues)

In view of the rapidly increasing number of sequenced organisms, the identification of cross species homologues is obviously an extremely important task in present (and future) molecular biology research. Powerful bioinformatics tools exist that allow for rapid and efficient genome-wide identification of orthologs, genes or proteins of different species that fulfill similar functions and are closely related in primary sequence [38, 39]. Despite tremendous successes in the past, these approaches rely on primary sequence homologies identifiable by BLAST methodology and thus potentially fail in cases of low sequence similarities. Since the meta-structure encodes 3D (meta-) structural information and can thus find structural/functional homologues even in cases where there is no significant pairwise sequence similarity, it might seem feasible to use meta-structure-based alignment strategies to search for distant homologs across species. As a prototypical example, the ortholog screening results for *S. cerevisiae* Zip1p are presented. Finding the *A. thaliana* ortholog of Zip1p from *S. cerevisiae* was a tremendously difficult task and required the development of a novel search strategy combining principles from genomics, proteomics and morphometric analysis of subcellular structures [40]. The duplicated *Arabidopsis* genes ZYP1a/ZYP1b encode homologous proteins with similarities to the transverse filament proteins of the synaptonemal complex (SC) [41]. As a starting point all sequences (32740) of the *A. thaliana* proteome were subjected to a meta-structure calculation. In a next step the *S. cerevisiae* Zip1p query sequence was aligned with all sequences from *A. thaliana*. The pairwise meta-structure similarities were quantified based on compactness and secondary structure values. The results are stored as an ordered list. The seven best scoring hits from *A. thaliana* are given in Table 2. As expected, the meta-structure approach did reveal the homology between Zip1p and ZYP1 (sixth

Table 2 Genome-wide meta-structure search for orthologs

Gene index	Similarity score	Protein/locus
AT1G22260.1	**589**	**'ZYP1a', chr1:7860149-7865131**
AT1G22275.1	**592**	**'ZYP1b', chr1:7867234-7872055**
AT5G46070.1	594	'GTPase', chr5:18700695-18705624
AT1G63300.1	603	PTHR23160,chr1:23485858-23489732
AT5G52280.1	607	chr5:21244185-21247335
AT1G05320.1	608	'Prefoldin', chr1:1554855-1557657
AT3G54670.1	680	'SMC1', 'ATSMC1', 'TTN8', chr3:20246796-20254679

The query sequence *S. cerevisiae* Zip1p was aligned with all protein sequences from *A. thaliana*. The pairwise meta-structure similarities were quantified based on compactness and secondary structure values. The results are stored as an ordered list, and only the best scoring hits (large similarity scores) from *A. thaliana* are given. The score is related to the total number of residue matches (similar compactness and secondary structure values). Details of the alignment procedure are given in the text. It should be noted that the absolute similarity score depends on the number of residues. The experimentally verified ortholog ZYP1a,b scores at the sixth and seventh position (*bold*), thus indicating the reliability of the meta-structure approach

and seventh position). It has to be noted that this search example is particularly challenging as both *S. cerevisiae* Zip12p and *A. thaliana* ZYP1 are large coiled-coil proteins, and thus lacking specific and distinct protein domain features. It is thus anticipated that the meta-structure approach can be successfully applied to large-scale genome-wide screening projects and will be a valuable extension and/or complement of existing sequence analysis methodology for ortholog searches.

Protein meta-structure similarity clustering (PMSSC)

In the following the potential of meta-structure based sequence alignments is additionally demonstrated with an application to the analysis of protein ligand complexes. Here sequence similarities identified by the meta-structure approach are used as starting points for the identification of ligand scaffolds relevant for drug development programs. It is well established that sequence homology provides valuable information about protein ligand complexes and is thus widely used in drug development programs [42–44]. However, the major drawback of this concept is its reliance on evolutionary conserved molecular recognition sequence motifs and the resulting limitation to close sequence homologs [45]. Since spatial structures are more conserved than sequences [46], the group of Waldmann has recently introduced a novel and powerful concept for structure-based drug development called protein-structure similarity clustering (PSSC) [47–49]. The rationale behind their approach is that conservation of structural motifs in the ligand sensing region of proteins can be used as a classifying principle to group proteins into clusters with similar ligand-binding properties. Structures of ligands binding to one member of the cluster are thus valid starting points for ligand development for other cluster members. Several examples on a diverse set of protein targets have been provided and convincingly demonstrate the applicability of this approach [47–49]. Although a very successful tool for drug development programs, the limited structural information available for therapeutically relevant protein targets reduces its general applicability. Thus, alternative methods would be highly desirable to fill this gap. Here it is demonstrated that the meta-structure approach can circumvent the requirement for highly resolved protein structures for structure-based drug design. The strategy is based on the consideration that meta-structure similarities in ligand interaction sites provide valuable starting information for the identification of chemical scaffolds and guiding structures in ligand development programs without the requirement of high-resolution protein structures. An outline of the protein meta-structure similarity clustering (PMSSC) approach is shown in Fig. 9. Suitable protein target sequences for meta-structure alignment can be taken from, for example, the DRUGBANK database, a public repository of biologically relevant protein targets with experimentally verified inhibitory ligands [50].

The PMSSC approach is demonstrated with an application to a protein target Tm0936 from *Thermotoga maritima* for which conventional bioinformatics analysis did not reveal any information and therefore the 3D structure of the protein had to be determined [51]. The subsequent structure-based docking analysis revealed a list of adenine analogues, which appeared to undergo C6-deamination. Out of the identified hit list, 5-methylthioadenosine (MTA) and S-adenosylhomocysteine (SAH) were tested as substrates and found to be active. It was thus concluded that Tm0936 acts as an MTA/SAH deaminase in a previously uncharacterized SAH degradation pathway. As a benchmark test for the new approach the sequence of Tm0936 was screened against the targets of the DRUGBANK database. The nine best hits are shown in Table 3. Despite the lack of sequence similarity to Tm0936, the meta-structure approach correctly identified homologs of the enzymes revealed by the protein 3D structure. Most importantly, S-adenosylmethionine

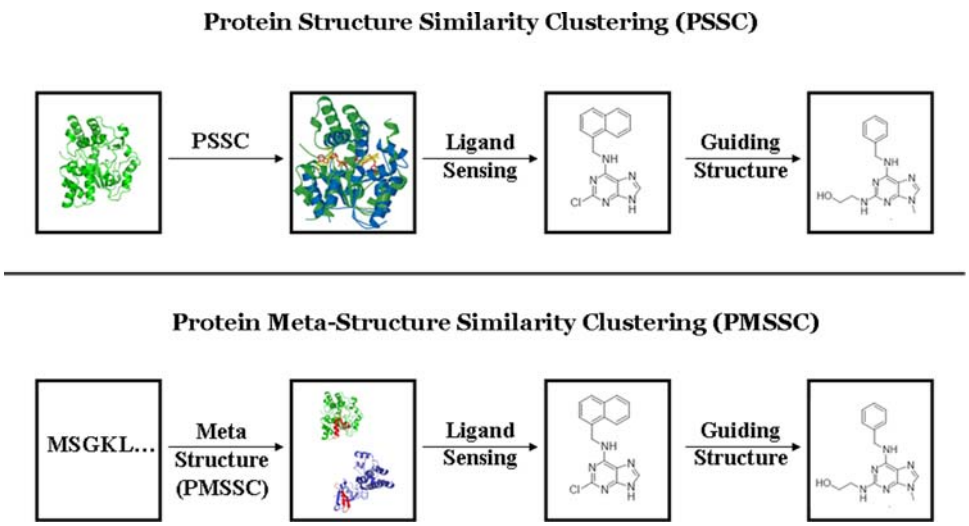

Fig. 9 Overview of the protein meta-structure similarity clustering (PMSSC) approach for ligand development. *Top*: 3D Structure-based similarity clustering approach developed by H. Waldmann and co-workers [47–49]. Conservation of structural motifs in the ligand sensing region of proteins is used as a classifying principle to group proteins into clusters with similar ligand-binding properties. Structures of ligands binding to one member of the cluster are valid starting points for ligand development for other cluster members. *Bottom*: Meta-structure similarities provide valuable starting information for the identification of chemical scaffolds and guiding structures in ligand development programs without the requirement of high-resolution protein structures

(SAM) synthetase and *S. cerevisiae* cytosine deaminase were found to be prominent meta-structure homologues of Tm0936, which is in very good agreement with the experimental X-ray/computer-docking/biochemistry approach. The following additional meta-structure homologs were found: glucarate dehydratase, serine hydroxyl-methyltransferase and β-glucosidase. The chemical structures of the ligands of some of the identified meta-structure homologs of Tm0936 are shown in Fig. 10. It is particularly noteworthy to also consider the ligand scaffolds of the other Tm0936 meta-structure homologs. For example, the carboxylic groups of the glucarate dehydratase ligands resemble the phosphate group of ADP. Even the serine hydroxyl-methyltransferase ligand comprises the characteristic chemical moieties of SAM, a N-heterocycle linked (via a phenyl-ring) to a highly negatively charged group. These examples nicely illustrate the potential of the meta-structure approach to identify possible chemical fragments that can be subsequently used for fragment-based drug design. Of course, the major benefit of the methodology is that it is exclusively based on primary sequence information and that no 3D protein structure is required; broad usage in the biotechnology and pharmaceutical sectors can thus be anticipated.

Table 3 Meta-structure search for DRUGBANK homologs

DRUGBANK-entry	Similarity score	Protein
EXPT03025	235	Cytochrome P450
APRD00932	237	Tubulin β-1 chain
EXPT02642	238	S-adenosylmethionine synthetase
EXPT01594	240	Glucarate dehydratase
NUTR00040	250	Alpha-N-acetylgalactosaminidase
EXPT01777	252	Cytosine deaminase
EXPT03045	268	Ser-hydroxymethyltransferase
EXPT02374	270	Beta-glucosidase

The query sequence *T. maritima* Tm0936 was aligned with all protein sequences from the DRUGBANK database, a public repository of protein sequences with their ligands. The score is related to the total number of residue matches (similar compactness and secondary structure values). Details of the alignment procedure are given in the text. The results are stored as an ordered list, and only the nine best scoring hits from are given. The DRUGBANK entry, the similarity score and protein name are given

Discussion and outlook

A novel concept for protein sequence analysis has been described. In this conceptual framework a protein is viewed as an intricate network of interacting residues organized as a multi-spherical entity. This novel conception offers unique possibilities for chemical (molecular) biology, structural genomics and drug discovery. In this review some prototypical applications were presented that serve to illustrate the potential of the methodology for molecular and chemical biology research. It has been shown that meta-structure analysis is an effective tool for large-scale (genome-wide) protein sequence analysis, target selection

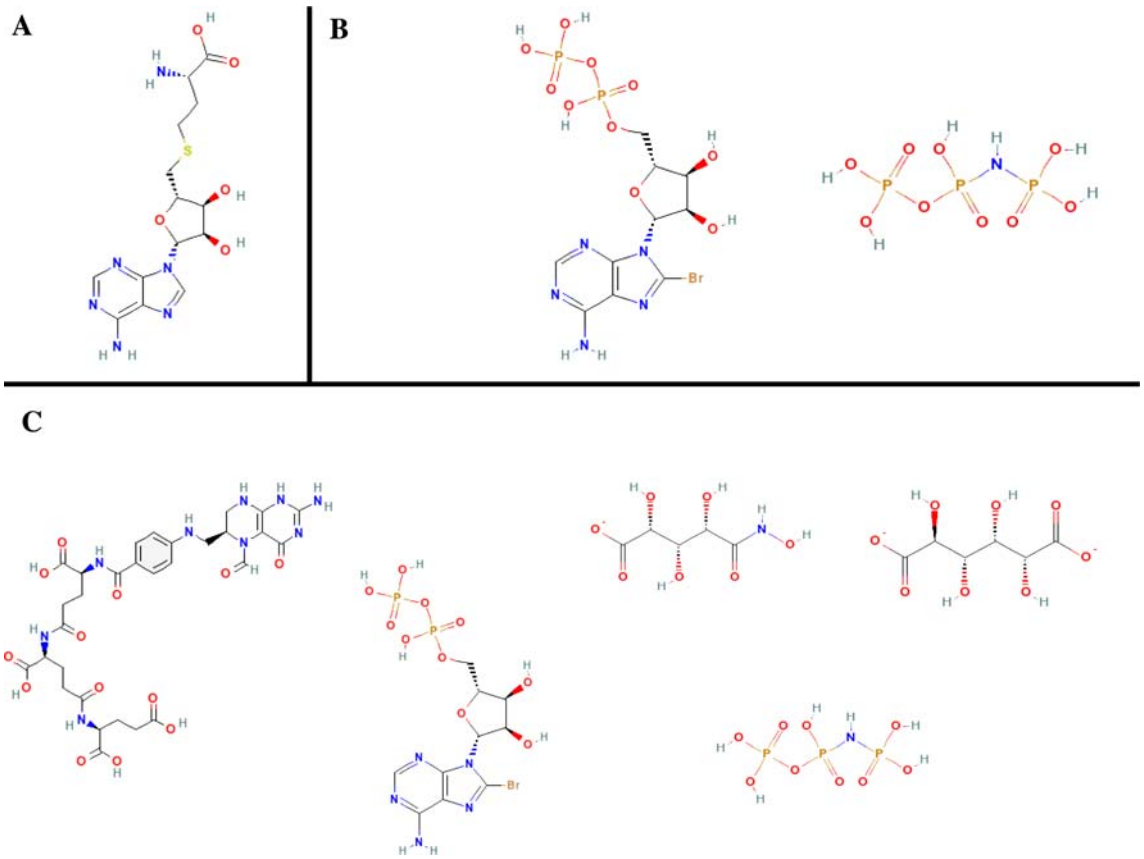

Fig. 10 Chemical structures of ligand scaffolds identified by the meta-structure approach based for Tm0936 from *Thermotoga maritima*. (**a**) Ligand scaffold identified by the 3D structure-based approach using the crystal structure of Tm0936 and computational docking and virtual screening [52], (**b**) known ligands for S-adenosylmethionine-synthetase and (**c**) ligand scaffolds identified using the meta-structure approach (exclusively based on primary sequence information). The meta-structure approach was based on a (pairwise) sequence-to-sequence screen against the targets of the DRUGBANK database [51]

for structural genomics and the identification of intrinsically unstructured (unfolded) proteins. It also constitutes the conceptual basis for a novel sequence alignment approach which can be used to identify distant homologues without identifiable primary sequence similarities. These novel sequence analysis tools provide rich avenues for large scale (genome-wide) explorations of protein sequences and together with the rapidly growing knowledge about biological activities allows for the design of powerful ultra-high throughput sequence functional inference strategies.

Acknowledgements This work was supported in part by the FWF (SFB-17), WWTF (LS162) and EU-BACRNA. The author thanks Andreas Schedlbauer (MFPL) for providing the NMR data of ICln, Peter Schlöglhofer (MFPL) for providing the *A.Thaliana* sequences and helpful discussion about orthology problems, Manfred Sippl (University of Salzburg) for valuable help with sequence alignment problems, and Karin Kloiber, Martin Tollinger and Nicolas Coudevylle (University of Vienna) for helpful discussions.

References

1. Tanford C, Reynolds J (2003) Nature's robots. A history of proteins. Oxford University Press, Oxford
2. Perutz MF, Muirhead H, Cox JM, Goaman LC (1968) Three-dimensional Fourier synthesis of horse oxyhemoglobin at 2.8 Å resolution: the atomic model. Nature 219:131–139
3. Epstein CJ, Goldberger RF, Anfinsen CB (1963) The genetic control of tertiary protein structure. Model systems. Cold Spring Harb Symp Quant Biol 28:439–449
4. Mayer O, Rajkowitsch L, Lorenz C, Konrat R, Schroeder R (2007) RNA chaperone activity and RNA-binding protein properties of the E.coli protein StpA. Nucl Acid Res 35:1257–1269
5. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213:859–883
6. Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Meth Enzym 394:42–78
7. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208

8. Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12:54–60
9. Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15:35–41
10. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signalling and cancer-associated proteins. J Mol Biol 323:573–584
11. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. J Mol Biol 293:321–331
12. Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579:3346–3354
13. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459
14. Jones DT, Ward JJ (2003) Prediction of disordered regions in proteins from position specific score matrices. Proteins 53:573–578
15. Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701–3708
16. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequenze complexity of disordered proteins. Proteins 42:38–48
17. Li X, Romero P, Rani M, Bunker AK, Obradovic Z (1999) Predicting disorder for N–, C– and internal regions. Genome Inform 10:30–40
18. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434
19. Zeev-Ben-Mordehai T, Rydberg EH, Solomon A, Toker L, Auld VJ, Silman I, Botti S, Sussmann JL (2003) The intracellular domain of the Drosophila cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. Proteins 53:758–767
20. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645
21. Fürst J, Schedlbauer A, Gandini R, Garavaglia ML, Saino S, Gschwentner M, Sarg B, Lindner H, Jakab M, Ritter M, Bazzini C, Botta G, Meyer G, Kontaxis G, Tilly BC, Konrat R, Paulmichl M (2005) ICln159 folds into a pleckstrin homology domain-like structure. Interaction with kinases and the splicing factor LSm4. J Biol Chem 280:31276–31282
22. Dyson HJ, Wright PE (2004) Unfolded proteins and protein folding studied by NMR. Chem Rev 104:3607–3622
23. Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. Curr Opin Struct Biol 17:3–14
24. Shortle D, Ackerman MS (2001) Persistence of native-like topology in a denatured protein in 8 M urea. Science 293:487–489
25. Gillespie JR, Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. J Mol Biol 268:170–184
26. Wishart D, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbour effects. J Biomol NMR 5(6):7–81
27. Hartl M, Karagiannidis AI, Bister K (2006) Cooperative cell transformation by Myc/Mil(Raf) involves induction of AP-1 and activation of genes implicated in cell motility and metastasis. Oncogene 25:4043–4055
28. Schedlbauer A, Ozdowy P, Kontaxis G, Hartl M, Bister K, Konrat R (2008) Backbone assignment for osteopontin, a cytokine and cell attachment protein implicated in tumorigenesis. Biomol NMR Assign 2:29–31
29. Denhardt DT, Giachelli CM, Rittling SR (2001) Role of osteopontin in cellular signaling and toxicant injury. Annu Rev Pharmacol Toxicol 41:723–749
30. Rangaswami H, Bulbule A, Kundu GC (2006) Nuclear factor inducing kinase: a key regulator in osteopontin-induced MAPK/IkappaB kinase dependent NF-kappaB-mediated promatrix metalloproteinase-9 activation. Trends Cell Biol 16:79–87
31. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
32. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–10919
33. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: implications for the nature of 'Äfold space', and structure and function prediction. Curr Opin Struct Biol 16:393–398
34. Shin DH, Lou Y, Jancarik J, Yokota H, Kim R, Kim SH (2005) Crystal structure of TM1457 from *Thermotoga maritima*. J Struct Biol 152:113–117
35. Sippl MJ (2008) On distance and similarity in fold space. Bioinformatics 24:872–873
36. Sippl MJ, Wiederstein M (2008) A note on difficult structure alignment problems. Bioinformatics 24:426–427
37. Arand M, Hallberg BM, Zou J, Bergfors T, Oesch F, van der Werf MJ, de Bont JAM, Jones TA, Mowbray SL (2003) Structure of rhodococcus erythropolis limonene-1, 2, expoxide hydrolase reveals a novel active site. EMBO J 22:2583–2592
38. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282:2022–2028
39. Rubin GM, Yandell MD, Wortman JR, Gabor-Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S (2000) Comparative genomics of the eukaryotes. Science 287:2204–2215
40. Bogdanov YF, Dadashev SY, Grishaeva TM (2002) Comparative genomics and proteomics of *Drosophila*, Brenner's nematode, and *Arabidopsis*: identification of functionally similar genes and proteins of meiotic chromosome synapsis. Russ J Genet 38:908–917
41. Higgins JD, Sanchez-Moran E, Armstrong SJ, Jones GH, Franklin FCH (2005) The *Arabidopsis* synaptonemal complex protein ZYP1 is required fro chromosome synapsis and normal fidelity of crossing over. Genes Dev 19:2488–2500
42. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci 43:391–405
43. Frye SV (1999) Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem Biol 6:R3–R7
44. Jacoby E, Schuffenhauer A, Floersheim P (2003) Chemogenomics knowledge-based strategies in drug discovery. Drug News Perspect 16:93–102
45. Gerlt JA, Babitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Ann Rev Biochem 70:209–246
46. Grishin NV (2001) Fold change in evolution of protein structrures. J Struct Biol 134:167–185

47. Breinbauer R, Vetter I, Waldmann H (2002) From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. Angew Chem Int Ed 41:2878–2890
48. Koch MA, Breinbauer R, Waldmann H (2003) Protein structure similarity as guiding principle for combinatorial library design. Biol Chem 384:1265–1272
49. Koch MA, Waldmann H (2005) Protein structure similarity clustering and natural product structures as guiding principles in drug discovery. Drug Discov Today 10:471–483
50. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34:D668–D672
51. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448:775–779
52. Pai EF, Krengel U, Petsko GA, Goody RS, Kabsch W, Wittinghofer A (1999) Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis. EMBO J 9:2351–2359