2006

Local Flexibility in Molecular Function Paradigm

Jag Bhalla
Georgetown University School of Medicine

Geoffrey B. Storchan
Georgetown University School of Medicine

Caitlin M. MacCarthy
Georgetown University School of Medicine

Vladimir N. Uversky
Indiana University School of Medicine, vuversky@usf.edu

Olga Tcherkasskaya
Georgetown University School of Medicine

Local Flexibility in Molecular Function Paradigm*

Jag Bhalla‡, Geoffrey B. Storchan‡, Caitlin M. MacCarthy‡, Vladimir N. Uversky§, and Olga Tcherkasskaya‡¶

It is generally accepted that the functional activity of biological macromolecules requires tightly packed three-dimensional structures. Recent theoretical and experimental evidence indicates, however, the importance of molecular flexibility for the proper functioning of some proteins. We examined high resolution structures of proteins in various functional categories with respect to the secondary structure assessment. The latter was considered as a characteristic of the inherent flexibility of a polypeptide chain. We found that the proteins in functionally competent conformational states might be comprised of 20–70% flexible residues. For instance, proteins involved in gene regulation, *e.g.* transcription factors, are on average largely disordered molecules with over 60% of amino acids residing in "coiled" configurations. In contrast, oxygen transporters constitute a class of relatively rigid molecules with only 30% of residues being locally flexible. Phylogenic comparison of a large number of protein families with respect to the propagation of secondary structure illuminates the growing role of the local flexibility in organisms of greater complexity. Furthermore the local flexibility in protein molecules appears to be dependent on the molecular confinement and is essentially larger in extracellular proteins. ***Molecular & Cellular Proteomics 5:1212–1223, 2006.***

Over the last 2 decades, the extent of structural research has led to a large number of three-dimensional (3D)[1] structures of biologically active macromolecules and their complexes. Enormous structural information (over 34,000 entries) is currently available from the Protein Data Bank (PDB) that includes details of protein organization, of their interactions with nucleic acids and ligands, and of their conformational behavior. Structural data provide the essential framework for characterizing molecular mechanisms of biological activity, for analyzing evolutionary relationships, and for illuminating our understanding of biological function (Ref. 1 and references therein).

Although proteins often are pictured as rigid entities corresponding to some average structure (immersed in a featureless solvent continuum), it has long been known that they have a rather fluid, dynamic structure with rapid conformational fluctuations (2). Subnanosecond dynamics of proteins studied by NMR (3), nitroxide spin labeling (4), dielectric relaxation (5), and fluorescence experiments (1) have advanced such descriptive terms as "breathing" (6), "relaxation," "segmental motion," and "mobile defect" (7) to portray the conformational mobility of proteins in functionally competent states. The presence of substantial >1-Å breathing motions has been recognized in early NMR studies on the flipping of the buried aromatic residues in the pancreatic trypsin inhibitor (8). Hemoglobin and myoglobin offer another striking example of the dynamic nature of biological activity where small structural fluctuations of the protein matrix allow O_2 molecule to move to and from the heme pocket (9–11). Buried water molecules, which are observed in most proteins, were shown to exchange with surface water molecules at the microsecond timescale, and that process necessitates large and correlated fluctuations in the host protein (12, 13). Furthermore the reduced activity of the protein mutants in some cases might be a consequence of reduced fluctuations and flexibility in the molecule away from that which has evolved for optimal functioning (1). Indeed the conformational lability in proteins, coordinated with the chemical requirements at each stage of their reactions, is a major component in enzyme catalysis, allosteric regulation, antigen-antibody interactions, and protein-DNA binding (1). The concept of inherent and correlated protein motions has become a landmark in biophysics and structural biology that underlies our understanding of molecular recognition (14, 15).

Although an importance of protein flexibility has been widely evoked in the literature, it has been more difficult to characterize experimentally. Proteins are composed of discrete atoms, which are constantly undergoing thermal fluctuations from rapid (picosecond) vibrations, through slower (multinanosecond) global reorientations and side chain isomerization, to long time scale (microsecond to second)

From ‡Biochemistry and Molecular & Cellular Biology, Georgetown University School of Medicine, Washington, D. C. 20007 and §Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202

Received, September 26, 2005, and in revised form, March 27, 2006

Published, MCP Papers in Press, March 29, 2006, DOI 10.1074/mcp.M500315-MCP200

[1] The abbreviations used are: used: 3D, three-dimensional; DSSP, a database of secondary structure assignments for all protein entries in the Protein Data Bank; GO, gene ontology; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; PDOC, PROSITE documentation; PROSITE, database of protein families and domains; MSDSD, Macromolecular Structure Database Search Database.

conformational changes (16). The reality of these fluctuations is evident in the PDB, which reports not only a set of fixed coordinates but also the temperature *B*-factors (Debye-Waller factors). The latter denotes the thermal fluctuations of the protein and provides information about the mobility of each atom in the structure (17–19). The crystallographic parameters have been successfully used to derive overall and intrinsic motions (20), to identify higher atomic mobility at the active site, and even to allocate a component in the amplitudes of atomic vibration that are derived from the overall global motion of the protein (21). Furthermore the analysis of the 3D structures of wild type proteins and their synthetic analogs (22) as well as the proteins that crystallize in a different space group (23) promotes the idea that the *B*-factors reveal the effect of different packing constraints on the protein flexibility. Further statistical-mechanical study of a large group of protein structures clearly demonstrated that the *B*-profile is, in fact, essentially determined by spatial variations in local packing density (24). Note that NMR data can also be used to characterize the flexibility of a protein (25), but in practice the number of atoms within a molecule is so large that drawing conclusions from the data is difficult. A simple method to predict protein flexibility using secondary chemical shifts has been developed recently that allows quantitative, site-specific mapping of protein backbone mobility without the need of a 3D structure or NMR relaxation experiments (26).

Numerous studies have attempted to identify flexible regions in proteins as well as to understand their role in overall protein dynamics and functionality. However, different groups use different definitions and various experimental approaches to identify this fold characteristic. One class of "intrinsically disordered" flexible regions was defined as the regions that are invisible in electron density maps of x-ray diffraction (27–31). Other researchers focus on extended ($>$70 consecutive residues) regions of a very low regular secondary structure that are particularly abundant in eukaryotic proteomes, conserved during evolution, and over-represented in regulatory and promiscuously interacting proteins (32–34). In fact, many proteins contain recognizable small "modules" that recur in other proteins in various combinations and in some cases can fold independently. They can be covalently linked to generate multimodular proteins and serve as self-directed structural units (35). Such domains can function independently, can be expressed in genomes, and are often rearranged through alternative splicing. These structural units are inferred to be a good evolutionary unit and are often used instead of whole proteins for annotations of the protein space (36). Yet if misplaced they can trigger dramatic biological consequences: oncoproteins comprising DNA-binding domains are capable of initiating transcription albeit being a small part of a largely unfolded chimeric polypeptide chain (37). In this regard, the crystallographic *B*-factors when considered over the length of a protein chain show that some segments undergo movements on a much larger scale than the rest of the protein, suggesting that the analysis of the *B*-distributions can be used to identify and predict flexible regions (38–41). Moreover the scheme was proposed to discriminate the amino acid residues according to their flexibility based on the *B*-factors of their C^{α} atoms (38, 39, 41). Altogether four categories with distinct flexibility were recognized that include low *B*-factor ordered regions, high *B*-factor ordered regions, and short and long disordered regions with the last two categories being the regions of missing electron density (31). The amino acid compositions of these categories differ significantly, whereas the biophysical properties of high *B*-factor ordered regions are relatively close to those of the short disordered regions. They provide a higher flexibility, hydrophilicity, and absolute net and total charge. The low *B*-factor ordered regions are enriched in hydrophobic residues and depleted in the total number of charged residues compared with the other categories (31). The "significantly-greater-than-chance" predictability of these categories from sequence suggests that they are, most likely, encoded at the primary structure level (31). The impact of local flexibility in proteins on biological activity remains unclear. It is clear, however, that even such a coarse grained aspect of protein structure as the secondary structure assigned from x-ray crystals captures flexibility relevant for protein function (42).

Structural complexity of biological macromolecules allows for a large variety of mechanisms to regulate the molecular recognition, including key-lock binding, template-assisted folding, folding by association, conformational selection, etc. Whether the recognition involves inducible or constitutive binding, the interaction *per se* depends on and affects the secondary structure of the individual protein (43–47). It seems reasonable therefore to analyze the deviation of the secondary structure (assigned by crystallography and NMR data) in various protein families with specific emphasis on the residues embedded in configurations with high flexibility. Such residues do not necessarily represent the class of natively disordered residues with high inherent flexibility but constitute a more flexible class than rigid helical or stranded regions. Mining conventional biophysical data might facilitate the discovery of underlying trends in structure-function relationships that were missed previously (48, 49). Here we examine the impact of local flexibility in proteins on molecular recognition and structure-function relationships utilizing gene-regulating molecules as a starting point.

Transcription factors (TFs) are modular molecules with separate domains mediating DNA binding, transcriptional activation, and repression. They are regulators of cell cycle progression, differentiation, and survival and are often altered in human diseases. Eukaryotic transcription factors contain a minimum of two domains that are responsible for the sequence-specific DNA recognition and transcriptional activation (50, 51). Most DNA-binding domains adopt folded structures in the absence of DNA, and conformational transitions induced by DNA binding are rather local (43–45). Yet tran-

scriptional activation domains might be both non-globular and unstructured and become partly ordered upon binding. Thus, local disorder appears to be an important facet of proteins involved in gene regulation, allowing for a variety of mechanisms in molecular networking (43, 45).

Despite the successes in protein domain assignment and comparison (50), there have been only a few breakthroughs in the area of quantifying the structure-function relationship. In fact, most of the databases classify the protein space by occurrence of a particular type of secondary structure (*e.g.* all α, all β, etc.) or global fold (*e.g.* Rossman, ferrodoxin, α/β-cylinder, $\alpha+\beta$-plaits, etc.). Analyzing the secondary structure in proteins with various biological activities will aid greatly in understanding the role of structure-function relationships in evolutionary biology. In this report, we introduce the parameter of the effective flexibility in proteins and analyze its variation through the functional space. We elucidate the impact of molecular confinement on inherited conformational lability of proteins and demonstrate the growing role of intrinsic flexibility in organisms of higher complexity that require intricate molecular networking (52).

EXPERIMENTAL PROCEDURES

Data Procurement—Structural data were collected using the annotations reported in the PBD and, later, utilizing an automated search with cross-referencing of protein structure with protein function, taxonomy, and gene ontology. Initially complete entries for high resolution structures (containing *B*-values, secondary structure indexes, experimental conditions, resolution characteristics, etc.) were taken from an unbiased selection of the PDB entries. In the case of TFs, this procedure resulted in 353 structures that included individual TF chains and multisubunit complexes with DNA. Given the significant redundancy of the PDB, the recovered data set was revised manually. Structures determined with resolution >3 Å and *R*-factor ≥0.25 were excluded from analysis as was NMR data averaged over less than 20 models, synthetic constructs, mutated proteins, and short polypeptides (<50 amino acids). The multiple entries were substituted by the average values of appropriate parameters. Of the remaining 112 structures of TF chains and single chain-DNA complexes, 98 structures (98 chains) were complete in all respects and were found to be acceptable (Table I). A similar approach was utilized to generate the structural set for oxygen transporters. In this case, a PDB query provided the original set of 497 entries (1,275 chains). Exclusion of structures with poor resolution, synthetic constructs, mutated proteins, duplicates, and multimeric complexes (>2 chains) left only 41 structures (57 chains) that were found acceptable for the analysis (Table II).

An automated data search was performed utilizing Macromolecular Structure Database Search Database (MSDSD) provided by the Macromolecular Structural Database Group at the European Bioinformatics Institute (www.ebi.ac.uk/msd). This database uses a generic search interface with a standard query language server that allows data exchange with the Oracle platform (Oracle RDBMS Version 9.2.0.1). Oracle9i software is widely used to manage large amounts of data, to import and export new data in the generated databases, to create tables, to select specific sets of data, and to combine data from different sources. Standard query language offers a powerful tool to connect relational databases. Data-housing capabilities of Oracle9i together with standard query language allow users to efficiently access and organize information reported in PDB: surfing 34,000 structures to reveal target correlations only takes a few minutes. Applying this approach, we examined the secondary structure composition in various protein families and in evolutionarily distinct organisms. Note that the MSDSD is derived from PDB entries and contains an extensive set of links to other biological databases including Swiss-Prot, SCOP (Structural Classification of Proteins), PROSITE, GO, and many others. These built-in, tightly integrated databases were used as 1) a source of input data, 2) an independent instrument for calculating various structural characteristics (*e.g.* B_α-factors, secondary structure composition, frequency functions, etc.), and 3) a tool for data analysis. Importantly MSDSD allows integration of PDB-derived data with the Gene Ontology database at the European Bioinformatics Institute, UK (www.geneontology.org) and the

TABLE I

PDB codes of 98 high resolution structures of TF proteins selected for analysis

36 structures resolved by crystallography and used for *B*-factor analysis are shown in bold. The subset of the best 21 structures resolved with resolution of ≤2.5 Å and *R*-values of <0.25 are underlined.

1ARD	**1BF5**	**1BGF**	1BHI	**1BM8**	1BOR	1BQV
1BVO	**1CDW**	**1CIT**	1CMO	1COK	1D5V	1D8J/K[a]
1DL6	**1DP7**	**1DXS**	**1E2X**	**1E3O**	**1EAN**	1ETC/D[a]
1F62	1FLI	1FTT	1FU9	1FV5	**1FYK**	1GAT/U[a]
1GCC	1GNF	1H95	1HDP	**1HH2**	1HKS/T[a]	**1HZ4**
1I11	1I1S	**1I27**	**1I4W**	**1I6A**	1J4W	1J5K
1JN7	1K1V	1K99	1KHM	1KQ8	1L3G	**1LFB**
1LV2	**1M1H**	1NPR	**1MB1**	**1MN4**	**1MNN**	1NFA
1NHA	**1OCT**	1OCP	**1ODH**	**1OPC**	1P97	1PFJ
1POU	1PXE	1PYC	**1PZW**	**1Q1H**	**1QQH**	1RLY/O4[a]
1RXR	**1SKN**	1SP1	1SP2	1TF3	1TFB	**1TGH**
1TTU	**1UBD**	1UL4	1UL5	**1UUS/R**[a]	1V63	1V64
1VD4	1YUI/J	2ADR	2/3GAT[a]	2/3GCC[a]	2HDC	2HFH
2HTS	2JHB	2LEF	2LFB	3HSF	**3HTS**	4/5GAT[a]

[a] Denotes double entries, *e.g.* 1D8J and 1D8K and similar entries across the table. This usually includes NMR structures averaged over ≥20 samples and that reported for Min Average. For double entries, the average value of molecular parameters was used in consequent analysis.

TABLE II

PDB codes selected for the analysis of oxygen transporters

Search input for PDB[a]: oxygen transporters; single chains and dimers; search output: 41 entries/57chains						
1A4F	1A6N	1ASH	1B0B	1C40	1CG8	1ECO
1EMY	1FHJ	1FSL	1G0A	1H97	1HBR	1HDA
1HLB	1ITH	1KR7	1L2K	1LHT	1MBS	1MOH
1MYT	1NGK	1PBX	1QPW	1S5Y	1SCT	1SPG
1T1N	1V4X	1WMU	1Y0D	1YMC	2HBG	2HMZ
2LH7	2LHB	2MHR	4HHB	4SDH	5MBA	

Search input for MSDSD[a]: GO, oxygen transporters; taxonomy: human; search output: 49 entries/182 chains						
1A3N	1A9W	1B86	1BIJ	1BUW	1BZ0	1BZ1
1CLS	1COH	1DKE	1FDH	1FN3	1G9V	1GBV
1GZX	1HAB	1HAC	1I3D	1I3E	1IRD	1J3Y
1J3Z	1J40	1J41	1JEB	1K0Y	1KD2	1LFL
1LFQ	1LFT	1LFV	1LJW	1M9P	1MKO	1NEJ
1O1N	1QSH	1QSI	1QXD	1QXE	1R1X	1R1Y
1RPS	1RQ3	1RQ4	1SDK	1SDL	1UIW	6HBW

[a] Mutants and synthetic constructs were both removed from the data set.

taxonomy database at the National Center of Biotechnology Information (www.ncbi.nlm.nih.gov). We took advantage of MSDSD in our study, focusing on biophysical parameters of proteins with respect to their functional ability.

Altogether comparing the manual and automated searches revealed the independence of the recovered characteristics on the algorithm in use: launching the MSDSD for oxygen transporters containing single chain or protein dimers yielded an experimental set similar to that from manual exploration of the PDB. Likewise searching for human proteins (taxonomy index 9606) involved in oxygen transport (GO:0005344) provided 49 entries (182 chains) associated mainly with tetrameric complexes. Table II gathers both sets of the recovered samples. Furthermore using the MSDSD/Oracle9i allowed us to perform automatic assignment of proteins and to avoid the manual curation of the generated databases.

Secondary Structure Assignment—Each protein chain in PDB is indexed with respect to the secondary structure composition based on appropriate crystallography or NMR data as assigned by the DSSP (53). Specifically eight secondary structure classifications are recognized, *i.e.* α-helix, 3_{10}-helix, β-strand, β-turn, bend, bulge, π-helix, and coil, and grouped into three categories: α-helix, β-strand, and coil. Three types, α-helix, 3_{10}-helix, and π-helix, are classified as α-helix. The second category encompasses residues in extended β-strand configurations, whereas remaining residues are regarded as "coil." In the PDB annotation, all helices shorter than five residues and all strands shorter than three residues are reassigned to coil (54). The secondary structure assignment provided by the MSDSD ignores the requirement for the minimum size of the secondary structure element, thereby yielding an ~7% smaller fraction of coiled residues.

Rational for Data Analysis—Manual and automated searches were performed for a large number of protein families with respect to the secondary structure composition. Specifically each protein chain within the distinct family was assigned with the parameter of effective flexibility (see below), and the frequency distribution of this parameter as computed for a bin of 10 units (otherwise stated) was generated for the entire ensemble. In addition, for each recovered distribution normalized by a constituency number, we calculated the ensemble average and applied it as a flexibility index to a protein family. The distributions of the frequency function and the average flexibility parameters were used to analyze the effect of molecular interactions on the protein flexibility and the effective flexibility in proteins with various biological activities and/or in various organisms, *e.g.* prokaryotes and eukaryotes.

RESULTS AND DISCUSSION

Parameter of Local Flexibility—Intrinsic flexibility in molecular systems allows for a number of definitions that all imply a freedom of internal rotations. The mechanism of the molecular flexibility is usually rather complicated, involving both small scale isomeric rotations of atomic groups within a particular sequence fragment and large scale rotations of individual structured domain. Given the structural hierarchy in biological molecules, one might consider the lack of secondary structure as an indicator of the local flexibility. The latter can be measured as the fraction of residues approaching "coiled" configurations. To justify the idea that some secondary structures (recognized by DSSP) can be considered as flexible joints, we examined 36 high resolution structures of TF molecules (Table I) determined by x-ray crystallography with respect to the deviation of *B*-factors. This parameter represents smearing of atomic electron densities around their equilibrium positions, and a larger *B*-value corresponds to a larger uncertainty in the location of the appropriate residue or atom (55, 56). Altogether the entire set of data contained 36 structures, 6,481 residues associated with seven secondary structure types (π-helix was missing), and appropriate *B*-values. For each *i*-th chain comprising N_i residues we computed the entire single chain average B_α-factor, $\langle B_{\mathrm{ch},i}\rangle$, with α-carbons being used as representative sites and the B_α^k values identifying every *k* residue.

$$\langle B_{\mathrm{ch},i}\rangle = \frac{\sum_{k=1}^{N_i} B_\alpha^{k,i}}{N_i} \qquad \text{(Eq. 1)}$$

The standard deviation of B_α-distribution was also calculated for every chain.

$$\sigma^2(B_{\mathrm{ch},i}) = \frac{\sum (B_\alpha^{k,i} - \langle B_{\mathrm{ch},i}\rangle)^2}{N_i} \qquad \text{(Eq. 2)}$$

In addition, all residues of *i*-th chain were sorted by secondary structure type, and the average B_α-values, $\langle B_{i,j}\rangle$, were computed for each secondary structure subcategory.

$$\langle B_{i,j}\rangle = \frac{\sum_{j=1}^{N_{i,j}} B_\alpha^{i,j}}{N_{i,j}} \qquad \text{(Eq. 3)}$$

In Equation 3, index $N_{i,j}$ denotes the number of residues involved in the *j*th type of secondary structure. The deviation of $\langle B_{i,j}\rangle$ around the chain average $\langle B_{\mathrm{ch},i}\rangle$ was expressed in units of standard deviation and used in further analysis (41, 55, 56).

$$\Delta B_{i,j} = \frac{\langle B_{i,j}\rangle - \langle B_{\mathrm{ch},i}\rangle}{\sigma(B_{\mathrm{ch},i})} \qquad \text{(Eq. 4)}$$

Positive $\Delta B_{i,j}$ values imply a greater uncertainty in the positioning of the residues associated with a particular secondary structure type in comparison with the chain average value and vice versa. Finally for the ensemble of 36 structures we computed the average of $\Delta B_{j,j}$ values using the frequency of the *j* type, f_j.

$$\Delta B_j = \frac{\sum_{i=1}^{36} \Delta B_{i,j}}{f_j} \qquad \text{(Eq. 5)}$$

Note that comparing *B*-factors among different proteins necessitates protein-by-protein normalizations to avoid the systematic differences in the experimental data (38–41, 57–59). Indeed the *B*-distributions are highly irregular when viewed protein by protein, mainly because of differences in the refinement methods used (60) and the degree of care taken to determine accurate *B*-factor values (61). These normalizations are usually made assuming the Gaussian distri-

TABLE III
Average values of B_α factors associated with the specific secondary structure type, ΔB_j, as calculated using 36 high resolution structures (6481 residues) of transcription factors

Parameter	Coil	β-Turn	Bend	3_{10}-Helix	Bulge	β-Strand	α-Helix
ΔB_j	0.34	0.17	0.15	0.06	−0.08	−0.11	−0.11
Residues (%)	22	12.7	8.5	2.3	1.2	19	34.4

bution of *B*-factors (38, 39), although other statistical models such as Gumbel distributions (extreme value distributions) also can be used (41). Given that Gaussian distributions of *B*-factors were previously used for the development of successful flexibility predictors (38, 39), we believe that the overall differences between the Gaussian and Gumbel distributions are relatively small, and the results of our studies are not quantitatively affected by using the Gaussian rather than Gumbel distributions.

The entire set of ΔB_j values are shown in Table III together with the pertinent secondary structure type and the fraction of residues involved. As can be seen, the coiled configurations provide the largest positive ΔB_j deviation of 0.34, whereas α-helical and β-stranded conformations yield the largest negative ΔB_j deviation of −0.11. The other secondary structure type with a negative ΔB_j value was found to be bulges representing only 1.2% of the entire ensemble. Unexpectedly the 3_{10}-helical configuration also provides a positive ΔB_j value of 0.06, thereby indicating notable uncertainty in the positioning of the residues involved. Thus, the residues located in α-helical or β-stranded segments consistently provide smaller B_α-factors than ensemble average, thereby demonstrating the superior structural rigidity in comparison with others. Hence α-helical and β-stranded configurations could be considered relatively rigid, whereas all other configurations could be considered flexible.

To examine overall positional ambiguity associated with intrinsic flexibility, we analyzed the scattering of ΔB_j values around the ensemble average as computed for flexible, ΔB_{fl}, and rigid, ΔB_r, residues. In this case, all secondary structure types that provided positive values of scaled *B*-factor, ΔB_j, were assigned as flexible, and those yielding negative values were assigned as rigid (see Table III). In addition, for every *i*th chain the average B_α-values were computed utilizing the set of flexible or rigid residues.

$$\langle B_i(\tilde{d})\rangle = \frac{\sum_{k=1}^{N_i(\tilde{d})} B_\alpha^k(\tilde{d})}{N_i(\tilde{d})} \qquad \text{(Eq. 6)}$$

In Equation 6, index $\tilde{d}$ indicates rigid or flexible residues and other parameters as denoted above. The deviation of the mean B_α-factors associated with flexible and/or rigid configurations around the chain average $\langle B_{ch,i}\rangle$ was expressed in units of standard deviation (41, 55, 56).

$$\Delta B_i(\tilde{d}) = \frac{\langle B_i(\tilde{d})\rangle - \langle B_{ch,i}\rangle}{\sigma(B_{ch,i})} \qquad \text{(Eq. 7)}$$

Fig. 1 displays the $\Delta B(\tilde{d})$ profiles for rigid (*curve 1*) and flexible (*curve 2*) configurations as calculated for 36 structures that are identified by the PDB code. The flexible configurations provided the average value of ΔB_{fl} of 0.22, whereas for rigid fragments one obtains −0.16 (*solid lines 2* and *1*, respectively). Using the subset of the first 21 structures with better resolution (<2.5 Å) yielded greater distinction of rigid and flexible ensembles given that $\Delta B_{fl} = 0.27$ and $\Delta B_r = -0.18$ (*thin lines*). This data corroborates the idea that residues in protein molecules present two independent groups that consistently exhibit quite different uncertainties in location and, therefore, can be regarded as relatively rigid and flexible populations.

Local Flexibility in Protein Family—Each selected protein was indexed by a parameter of flexibility, $\tilde{d}$, that implies the fraction of residues comprised in coiled configurations. The entire group was divided into subcategories in accord with the value of $\tilde{d}$ parameter (<10, <20, etc.), and the frequency of occurrence of proteins in each 10-unit bin, $F(\tilde{d})$, was counted. In addition, the frequency function was normalized to the total number of protein chains selected for the analysis. As a characteristic of a frequency function distribution, $F(\tilde{d})$, we utilized the weighted average, $\langle\tilde{d}\rangle$,

$$\langle\tilde{d}\rangle = \frac{\sum_i \tilde{d}_i F_i(\tilde{d}_i)}{\sum_i F_i} \qquad \text{(Eq. 8)}$$

where F_i and $\tilde{d}_i$ are the frequency index and the appropriate flexibility parameter, respectively.

To begin with, we explored the local flexibility in the TF family using readily available data reported in the PDB. An exhaustive cleansing (see "Data Procurement") left 73 single TF chains and 25 complexes with DNA fragments. Among them 36 structures were resolved by x-ray crystallography, and 62 structures were determined by NMR spectroscopy. The selected set of 98 structures was sufficient to permit reasonable statistical tendencies to be determined. For each protein chain we retrieved the secondary structure indexes and calculated the flexibility index as a fraction of residues involved in flexible configurations.

Fig. 2*A* shows the distributions of the flexibility parameter computed for the entire structural ensemble (*open circles*), individual TF chains (*gray circles*), and TF in complex with

DNA (*black circles*). One sees that the flexible residues constitute over half of the TF chains, providing the ensemble average $\langle\tilde{d}\rangle$ value of 61%. Formation of intermolecular complexes with DNA induces only a slight ordering of the TF proteins given that the fraction of flexible residues decreases from 63 to 58%. The recovered distributions are nearly symmetrical with a half-width of the frequency function of ~20%.

When searching for structural determinants that are indicative for protein families, it is imperative to elucidate inherent limitations of generated databases and the experimental techniques utilized for the data acquisition. The proteins analyzed here are those for which structures have been determined at atomic resolution. This selection inevitably introduces a bias, principally toward smaller proteins (given that the mean chain length in the PDB is about 370 residues) and those with rigid structures (35). Clearly crystallography reports mostly on proteins that are inherently ordered and therefore can be crystallized, whereas a much wider structural diversity can be studied by NMR method. In agreement with this supposition, we found (see Fig. 2*B*) that the average flexibility (as characterized by the content of coiled residues) in structures determined by NMR (*black circles*) is larger by 15% than that in crystallized samples (*gray circles*). Both the NMR and the x-ray sets, however, brought in relatively large $\langle\tilde{d}\rangle$ values of 66 and 51%, respectively, thereby pointing to a large portion of the TF residues that exhibit substantial uncertainty in their locations. Whether this finding indicates the significant amplitude of thermal motions (dynamic disorder) allowed in packed structures or reflects quenched disorder, *i.e.* trapped coil-like configurations, requires further investigation.

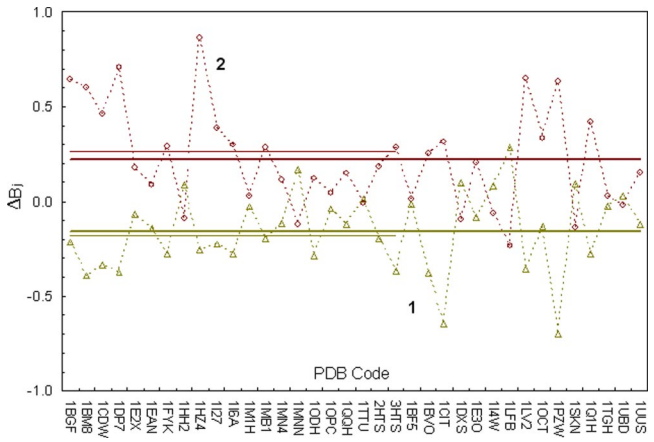

FIG. 1. **Chain-averaged B_α-factors calculated for the rigid (*1*) and flexible (*2*) configurations in 36 high resolution structures of transcription factors.** ΔB is expressed in the units of standard deviation. Appropriate ensemble averaged B_α-factors are identified by the *horizontal lines*. *Thin lines* show the ensemble-averaged B_α-factors calculated for 21 structures with better resolution. The PDB codes of the selected proteins are listed in Table I.

Local Flexibility and Biological Activity—Toward understanding the role of fold specifics in structure-function relationships, the correlation between the protein flexibility and the protein function needs to be addressed using samples with distinct functional activity. The proteins involved in the transport of oxygen in biological systems present the best candidates for this study given a large number of structures determined at high resolutions. Our initial PDB search provided 454 entries (1,345 chains), which included individual proteins and multimeric complexes. The manual refinement left only 41 structures (57 chains) suitable for the analysis, which included individual chains and protein dimers (Table II). The *inset* in Fig. 3 displays the $\tilde{d}$-distributions for the representative 57 chains (*open triangles*) and for the entire structural set (*filled triangles*) as computed for a bin of five units. We found that this functional category can tolerate only a small portion of the flexible residues given that $\langle\tilde{d}\rangle$ of about 34%

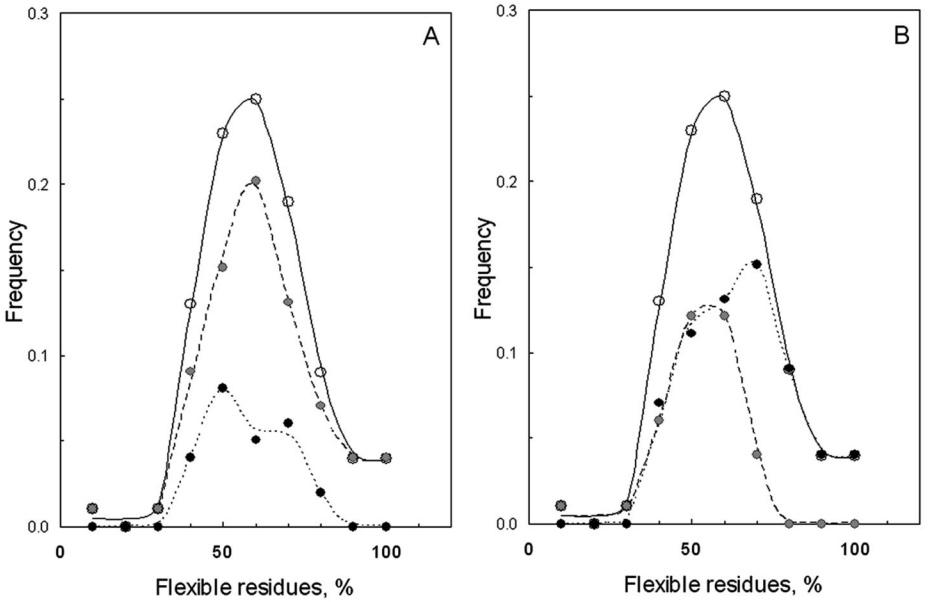

FIG. 2. **Effective flexibility in the family of the transcription factors (*open circles*).** *A* shows the contributions from individual chains (*gray circles*) and complexes with DNA (*black circles*). *B* displays the contributions from the structures resolved by x-ray crystallography (*gray*) and those by NMR spectroscopy (*black*). The PDB codes of the selected samples are listed in Table I.

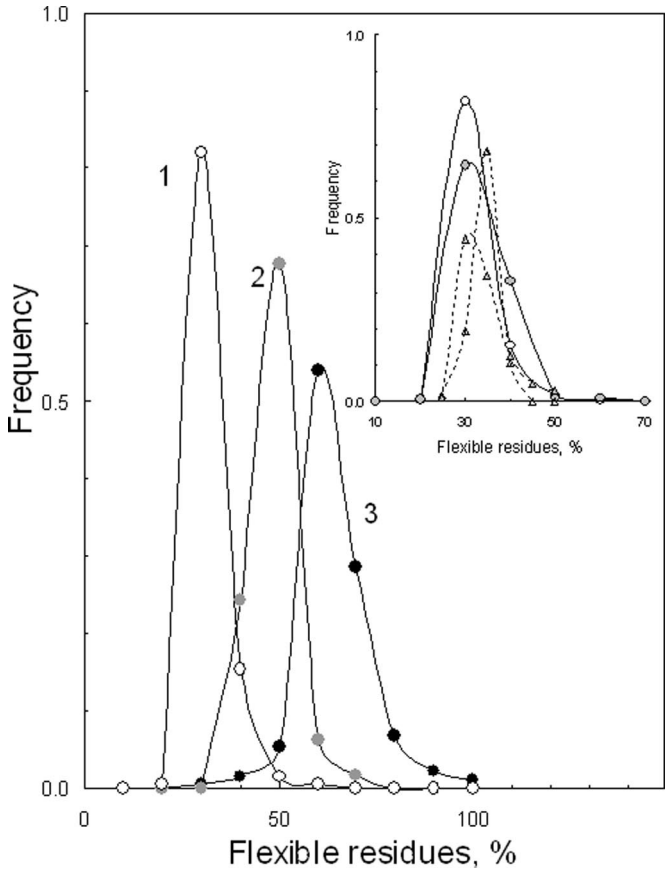

FIG. 3. **Distribution of flexibility parameter calculated for oxygen transporters (*1*), proteins with tyrosine phosphatase activity (*2*), and proteins with serine type endopeptidase activity (*3*).** The *inset* shows the distributions generated for oxygen transporters utilizing manual PDB search (*triangles*) and automated search through MSDSD (*circles*). Effects of cleansing are also shown (*filled symbols*).

was obtained for both data sets. Furthermore the $\tilde{d}$-distributions generated for oxygen transporters exhibit narrowly defined functions with a half-width of about 5% for the revised set and 10% for the entire set retrieved from the PDB. Once again, we have established that the incorporation of protein molecules into intermolecular complexes has very little effect on such fold specifics as intrinsic flexibility.

Using the MSDSD allows for automated analysis of the large data domain with assignment of gene ontology and taxonomy indexes to the protein samples. We utilized this approach to generate the database comprising human proteins (taxonomy index 9606) that are involved in the oxygen transport as identified by GO function (GO:0005344). The *inset* in Fig. 3 displays the $\tilde{d}$-distributions generated for the originally identified 428 chains (*open circles*) and for the 182 chains that were left after cleansing (*filled circles*). The average flexibility parameter was found to be 32 and 34%, respectively. It appears that secondary structure specificity might be robust enough to survive the coarse grained screening of protein families with respect to the functional activity. To provide further support to this notion, we computed the $\tilde{d}$-distributions for a number of human protein families with distinct functions as identified by the GO database. Fig. 3 displays the distributions of flexibility parameter computed for oxygen transporters (*curve 1*, GO:0005344), proteins with tyrosine phosphatase activity (*curve 2*, GO:0004725), and proteins with serine type endopeptidase activity (*curve 3*, GO: 0004252). A data search provided 116, 86, and 307 PDB entries that contained 428, 111, and 454 protein chains, correspondingly. As can be appreciated, each protein family provides rather specific $\tilde{d}$-distributions in terms of position of the maximum and the bandwidth: increasing flexibility correlates with the increasing width of the distribution. Overall these findings point to a great potential of coarse grained analysis of fold specifics to reveal the structure-function correlations and their diversity in evolutionary distinct organisms.

Phylogenic Comparison of Functional Categories—To elucidate the effect of evolution on protein structure we analyzed the proteins with similar functional ability in organisms of different complexity. Our initial set of the TFs included 20 structures from prokaryotic proteins and 68 from eukaryotes selected manually from the PDB (Table I). The decomposition of the $\tilde{d}$-distribution into the contributions from eukaryotic and prokaryotic proteins revealed that average flexibility in eukaryotes is larger by 11% (data are not shown), pointing thereby to the increasing role of the conformational lability in the functioning of the organism with larger complexity. At this point, a serious question arises: to what extent does the PDB data provide a comprehensive set suitable for phylogenic comparison of functional categories? To address this issue we analyzed the TF family using the PROSITE database to explore the molecular compositions of the similar functional categories. A search provided 58 entries with various PROSITE documentation indexes, PDOC. For each PDOC entry we generated a taxonomic tree view of all Swiss-Prot/TrEMBL entries matching the given PDOC code and counted the samples associated with specific taxonomy. The representative set of TFs in eukaryotic and prokaryotic proteins are shown in Fig. 4. One can see that the composition of the TF family associated within evolutionarily distinct organisms differs significantly as follows from the uneven distribution of appropriate sequences in the functional space. For instance, zinc fingers, which are largely unfolded macromolecules, comprise almost 17% of the human proteins, and they are lost in simple organisms (Fig. 4). (All zinc fingers are *italic* in Fig. 4.) Whether this reflects the effect of evolution on a protein family intended to perform a specific function or reflects the bias of experimental data remains unclear. It is clear, however, that structural comparisons lacking the taxonomy of samples might be misleading toward understanding structure-function relationships.

To approach this problem we used the MSDSD provided by the European Bioinformatics Institute (62). Briefly MSDSD is a relational database that offers a single access point for protein and nucleic acid structures and related information. Impor-

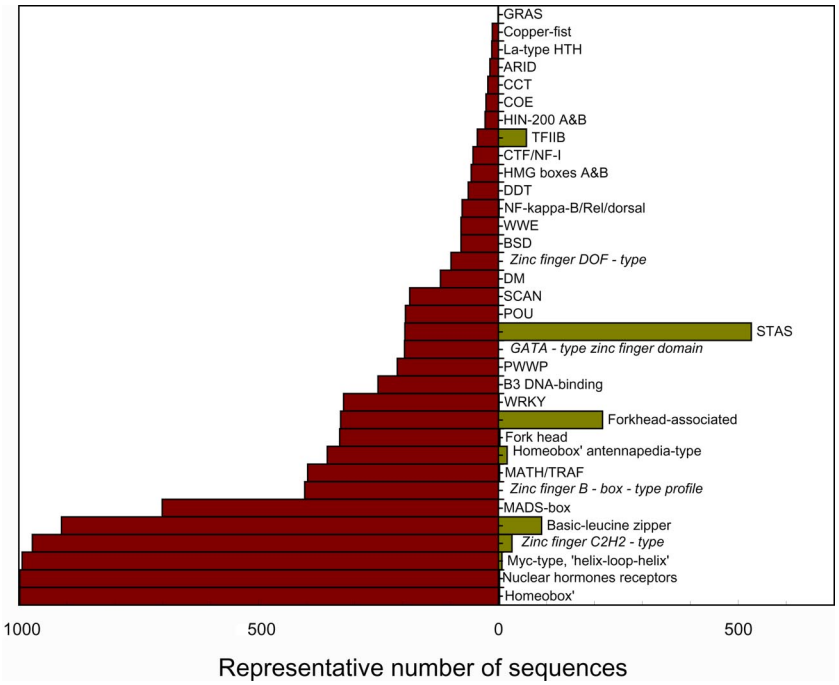

FIG. 4. **Subclassification of transcription factors in prokaryote (*right*) and eukaryote (*left*) as generated with PROSITE database (us. expasy.org).**

tantly the MSDSD allows combining the PDB-derived structural data with ontology and taxonomy databases. Using the GO database, we identified sets of proteins with resolved 3D structure using human and *Escherichia coli* proteins as targets. To minimize sampling errors, only GO functions that were supported by >50 PDB entries were used. The search with a larger cutoff provided similar results. In addition, for a given organism we calculated the flexibility parameter for proteins with identical GO functions. Altogether we examined the set of 70 functional categories associated with human proteins and 21 categories relevant for *E. coli*, comprising a total of 12,107 structures (22,109 chains). For each functional category, we calculated the average flexibility parameter and organized data in ascending manner that shows increasing flexibility with respect to the identified function (Fig. 5).

We found that the fraction of residues with higher flexibility in human proteins ranges from 25% (*e.g.* oxygen transporters) to 65% (*e.g.* trypsin and plasminogen activators), whereas for *E. coli* proteins it scatters around 35–45%. It appears that in organisms of greater complexity (human *versus E. coli*), the molecular flexibility becomes more significant. As expected the "function" as a physical process (*e.g.* binding) shows no distinctiveness with respect to the molecular flexibility. In fact, it implies a response to the presence of another counterpart: DNA-binding proteins exhibit the average flexibility parameter of <40%, whereas heparin and zinc binding requires 50 and 60% of flexible residues, respectively. It seems likely that protein fold specifics are receptive for the relevant biological process that is a series of events involving proteins (series of functions) and/or for the molecular localization indicating where the interactions occur. To this end, Fig. 5*B* displays the effective flexibility in proteins that are identified by their localization. Two examples, human proteins (*curve 1*) and *E. coli* proteins (*curve 2*), are considered. As can be appreciated, there is a clear correlation between the confinement of human proteins and the content of coiled residues: local flexibility increases as proteins move out from the interior (Golgi apparatus, nucleus, mitochondria, cytoplasm, etc.) to the exterior of the cell. On the contrary, the *E. coli* proteins show no significant changes of these fold specifics upon the cellular locations. Once again, we observe a striking correlation between the cell complexity and the flexibility of relevant molecular forms.

CONCLUSIONS

To date, most of our knowledge about proteins is derived from sequence-related databases. Sequence comparisons, however, fail to identify many relationships that emerge from known protein structures. Here we address correlations between protein fold characteristics and protein functional activities. Several families were investigated with an approach relying on readily available structural data. The primary observation was that natively folded proteins might have a significant fraction of residues that exhibit uncertainty in residue positions and, therefore, can be regarded as flexible. The diversity of this structural parameter within the protein family can be significant. Some examples are provided in Fig. 6. Given that many PDB entries are truncated chains and crystallized proteins that are biased toward a higher content of rigid regions, our assessment of the local flexibility in protein families is most likely underestimated. One should also consider the less ordered conformational states (*e.g.* molten glob-

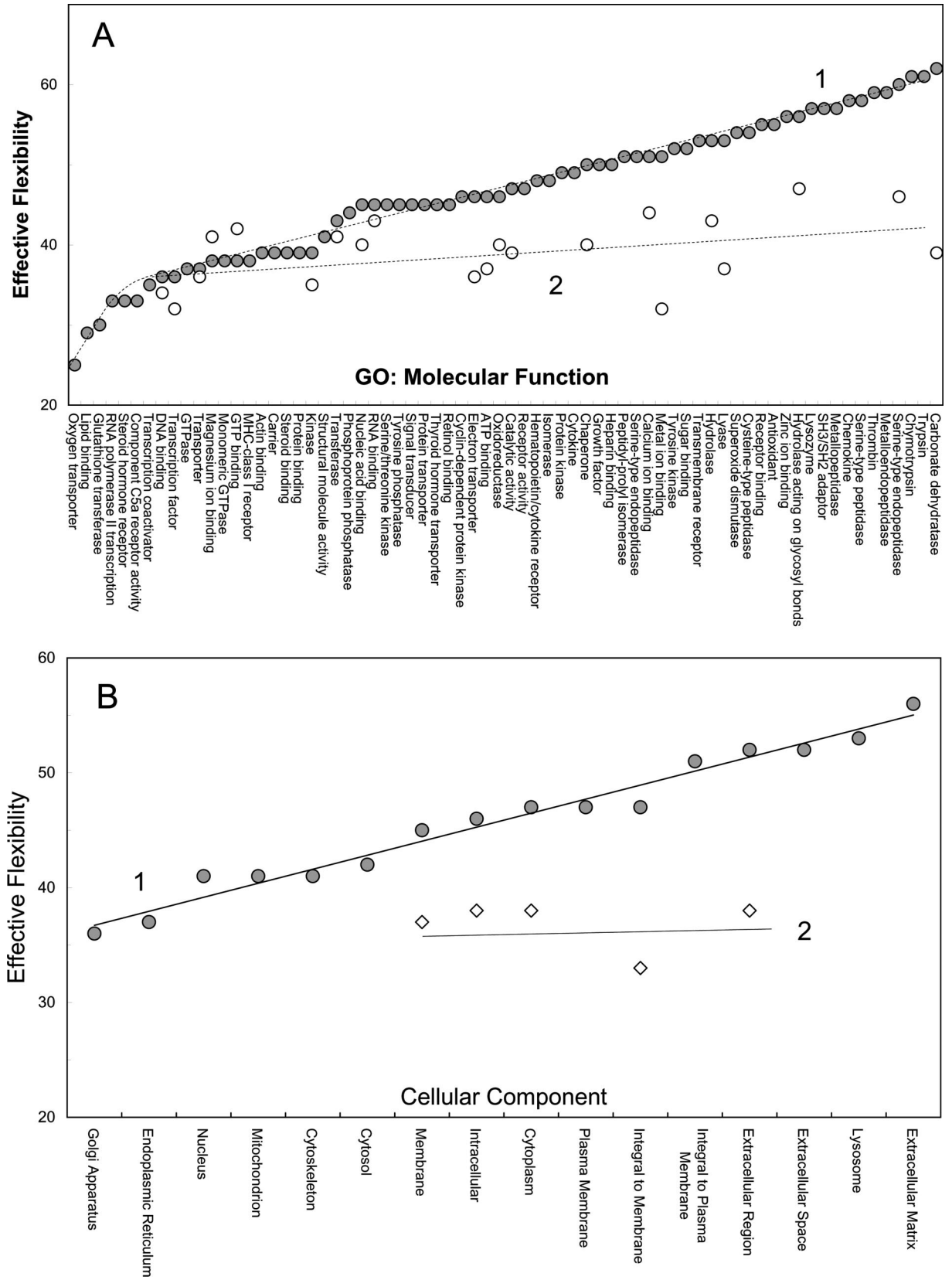

FIG. 5. **Effective flexibility in human (*1*) and *E. coli* (*2*) proteins with respect to the molecular function (*A*) and cell localization (*B*).** The proteins were identified by taxonomy index 9606 (human) and 562 (*E. coli*).

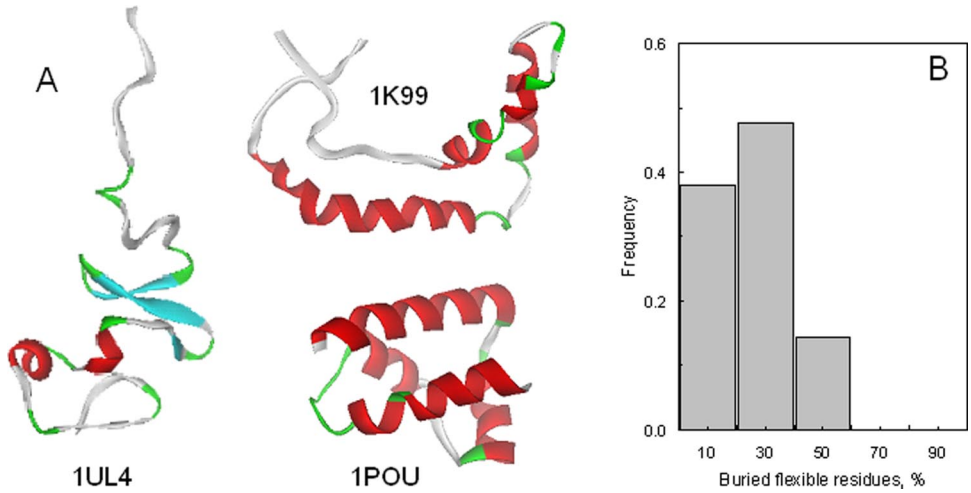

FIG. 6. *A*, representative structures of Oct-1/Pou-specific domain (*1POU*), the first Hmg box domain in human upstream binding factor (*1K99*), and DNA-binding domains of *Arabidopsis* Sbp family transcription factors (*1UL4*) that contain 35, 55, and 77% of the flexible residues, respectively. Proteins are identified by the PDB code. *B*, distribution of buried flexible residues in the family of the transcription factors (see Table I).

ules) that might be involved in protein functioning. At this juncture, a question arises as to the location of flexible regions within the 3D structures. Our analysis of transcription factors (see Table I) indicated that over 30% of flexible residues are, in fact, buried in the interior of the molecule and have no direct exposure to the solvent (Fig. 6*B*).

Presence of significant coiled regions in protein molecules challenges the conventional views of molecular recognition and protein function as well as our understanding of structure-function relationships. The higher flexibility in eukaryotic proteins points to an increasing role of conformational lability in biological systems of greater complexity with respect to molecular recognition and molecular assembly as well as protein modification, etc. Clearly certain biological activities, such as enzyme catalysis, immunological recognition, or molecular discrimination by receptors, demand exquisite control of 3D structure. Other functions such as signaling can be achieved by linear sequences. Locally disordered/flexible segments, which are induced to fold by interactions with other molecules, offer several important advantages for molecular signaling and cellular regulation. For instance, inherently labile segments can easily be shaped by their environment and might be able to recognize a large number of biological targets without sacrificing specificity. Different sequences might provide comparable binding sites (63), and similarly different sequences can fold into similar structures but have different functions (64–67). In fact, numerous efforts to annotate function based on structure and sequence homology alone often lead to misannotations (68, 69). Most importantly, complex organisms might retain their intricate physiologies without dramatically increasing their genome size (note that the number of genes in the human genome proved to be deceptively low). It may be the case that conformational lability of small sequences manifests the capacity to provide various binding sites, recognition templates, and signaling pathways and points to multidimensional structure-function relationships. Using available structural data, we also recognize that protein structures are closely tied to the location of the protein: intracellular proteins on average seem to be more rigid than those located closer to the extracellular milieu. Another direction yet to be explored is to consider function as a biological process that assumes locations, interactions, and counterparts. Indeed incorporation of fold-specific correlations into evolutionary models would provide a new tool for the organization of biological data, thereby expanding the conventional views of biological function.

Acknowledgments—We thank Drs. Eugene Shakhnovich and Angela M. Gronenborn for thorough discussions and helpful suggestions. We express our gratitude to Drs. P. Keller and M. John of the European Bioinformatics Institute for providing access to the Macromolecular Structure Database.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ To whom correspondence should be addressed. Tel.: 202-687-1303; Fax: 202-687-7186; E-mail: ovt@georgetown.edu.

REFERENCES

1. Dodson, G., and Verma, C. S. (2006) Protein flexibility: its role in structure and mechanism revealed by molecular simulations. *Cell. Mol. Life Sci.* **63**:207–219
2. Cooper, A. (1976) Thermodynamic fluctuations in protein molecules. *Proc. Natl. Acad. Sci. U. S. A.* **73,** 2740–2741
3. Allerhand, A., Doddrell, D., Glushko, V., Cochran, D., Wenkert, E., Lawson, P., and Gurd, F. (1971) Conformation and segmental motion of native and denatured ribonuclease A in solution. Application of natural-abundance carbon-13 partially relaxed Fourier transform nuclear magnetic resonance. *J. Am. Chem. Soc.* **93,** 544–546
4. Likhtenshtein, G. I., Grebenshchikov, Yu. B., and Avilova, T. V. (1972) An investigation of the microrelief and conformational mobility of proteins by

the ESR method. *Mol. Biol.* **6,** 52–60
5. Pennock, B. E., and Schwan, H. P. (1969) Further observations on the electrical properties of hemoglobin-bound water. *J. Phys. Chem.* **73,** 2600–2610
6. Englander, S. W., Downer, N. W., and Teitelbaum, H. (1972) Hydrogen exchange. *Annu. Rev. Biochem.* **41,** 903–924
7. Eftink, M. R., and Ghiron, C. A. (1975) Dynamics of a protein matrix revealed by fluorescence quenching. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 3290–3294
8. Wüthrch, K., and Wagner, G. (1975) NMR investigations of the dynamics of the aromatic amino acid residues in the basic pancreatic trypsin inhibitor. *FEBS Lett.* **50,** 265–268
9. Perutz, M. F., and Mathews, F. S. (1966) An x-ray study of azide methaemoglobin. *J. Mol. Biol.* **21,** 199–202
10. Frauenfelder, H., Petsko, G. A., and Tsernoglou, D. (1979) Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **280,** 558–563
11. Case, D. A., and Karplus, M. (1979) Dynamics of ligand binding to heme proteins. *J. Mol. Biol.* **132,** 343–368
12. Denisov, V. P., Peters, J., Horlein, H. D., and Halle, B. (1996) Using buried water molecules to explore the energy landscape of proteins. *Nat. Struct. Biol.* **3,** 505–509
13. Verma, C. S., and Fischer, S. (2005) Protein stability and ligand binding: new paradigms from *in-silico* experiments. *Biophys. Chem.* **115,** 295–302
14. Schulz, G. E. (1979) Nucleotide binding proteins, in *Molecular Mechanisms of Biological Recognition* (Balaban, M., ed) pp. 79–94, Elsevier, Amsterdam
15. Dunker, A. K., Garner, E., Guillot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J. E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* **3,** 473–484
16. Karplus, M., and McCammon, J. A. (1981) The internal dynamics of globular proteins. *CRC Crit. Rev. Biochem.* **9,** 293–349
17. Rasmussen, B. F., Stock, A. M., Ringe, D., and Petsko, G. A. (1992) Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature* **357,** 423–424
18. Trueblood, K. N., Bürghi, H.-B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U., and Abrahams, S. C. (1996) Atomic displacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr. Sect. A* **52,** 770–781
19. Drenth, J. (1994) *Principles of Protein Crystallography*, Springer-Verlag, New York
20. Sternberg, M. J., Grace, D. E., and Phillips, D. C. (1979) Dynamic information from protein crystallography: an analysis of temperature factors from refinement of the hen egg-white lysozyme structure. *J. Mol. Biol.* **130,** 231–252
21. Artymiuk, P. J., Blake, C. C., Grace, D. E., Oatley, S. J., Phillips, D. C., and Sternberg, M. J. (1979) Crystallographic studies of the dynamic properties of lysozyme. *Nature* **280,** 563–568
22. Phillips, G. N., Jr. (1990) Comparison of the dynamics of myoglobin in different crystal forms. *Biophys. J.* **57,** 381–383
23. Davies, D. R., and Cohen, G. H. (1996) Interactions of protein antigens with antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 7–12
24. Halle, B. (2002) Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 1274–1279
25. Ishima, R., and Torchia, D. A. (2000) Protein dynamics from NMR. *Nat. Struct. Biol.* **7,** 740–743
26. Berjanskii, M. V., and Wishart, D. S. (2005) A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* **127,** 14970–14971
27. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity—linking function and disorder. *Nat. Biotechnol.* **19,** 805–806
28. Romero, P., Obradovic, Z., and Dunker, A. K. (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* **462,** 363–367
29. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins* **53,** Suppl. 6, 566–572
30. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., Garner, E., Guilliot, S., and Dunker, A. K. (1998) Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **3,** 437–448
31. Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., Lawson, J. D., and Dunker, A. K. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.* **13,** 71–80
32. Liu, J., Tan, H., and Rost, B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322,** 53–64
33. Liu, J., and Rost, B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.* **31,** 3833–3835
34. Schlessinger, A., and Rost, B. (2005) Protein flexibility and rigidity predicted from sequence. *Proteins* **61,** 115–126
35. Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. G. (1999) Proteins folds, functions and evolution. *J. Mol. Biol.* **293,** 333–342
36. Shakhnovich, B. E., Harvey, J. M., Comeau, S., Lorenz, D., DeLisi, C., and Shakhnovich, E. (2003) ELISA: structure-function inferences based on statistically significant and evolutionarily inspired observations. *BMC Bioinformatics* **4,** 34–41
37. Uren, A., Tcherkasskaya, O., and Toretsky, J. A. (2004) Recombinant EWS-FLI1 oncoprotein activates transcription. *Biochemistry* **43,** 13579–13589
38. Karplus, P. A., and Schulz, G. E. (1985) Prediction of chain flexibility in proteins. *Naturwissenschaften* **72,** 212–213
39. Vihinen, M., Torkkila, E., and Riikonen, P. (1994) Accuracy of protein flexibility predictions. *Proteins* **19,** 141–149
40. Kundu, S., Melton, J. S., Sorensen, D. C., and Phillips, G. N., Jr. (2002) Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.* **83,** 723–732
41. Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., and Zhu, G. (2003) Improved amino acid flexibility parameters. *Protein Sci.* **12,** 1060–1072
42. Andersen, C. A. F., Palmer, A. G., Brunak, S., and Rost, B. (2002) Continuum secondary structure captures protein flexibility. *Structure* (*Lond.*) **10,** 175–185
43. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331
44. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins. *Curr Opin. Struct. Biol.* **12,** 54–60
45. Dyson, H. J., and Wright, P. E. (2004) Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* **104,** 3607–3622
46. Dunker, A. K., Brown, C. J., Lawson, J. D., lakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function. *Biochemistry* **41,** 6573–6582
47. Demchenko, A. P. (2001) Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recognit.* **14,** 42–61
48. Tcherkasskaya, O., and Uversky, V. N. (2003) Polymeric aspects of protein folding: a brief overview. *Protein Pept. Lett.* **10,** 239–245
49. Tcherkasskaya, O., Davidson, E. A., and Uversky, V. N. (2003) Biophysical constraints for protein structure prediction. *J. Proteome Res.* **2,** 37–42
50. Kant, S., Bagaria, A., and Ramakumar, S. (2002) Putative homeodomain proteins identified in prokaryotes based on pattern and sequence similarity. *Biochem. Biophys. Res. Comm.* **299,** 229–232
51. Brivanlou, A. H., and Darnell, J. E. (2002) Transcription—signal transduction and the control of gene expression. *Science* **295,** 813–818
52. Fernandez, A., Scott, R., and Berry, R. S. (2004) The non-conserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 2823–2827
53. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637
54. Eisenhaber, F., Imperiale, F., Argos, P., and Frommel, C. (1996) Prediction of secondary structural content of proteins from their amino acid composition alone. 1. New analytic vector decomposition methods. *Proteins Struct. Funct. Genet.* **25,** 157–168
55. Parthasarathy, S., and Murthy, M. R. N. (1997) Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.* **6,** 2561–2567
56. Yuan, Z., Zhao, J., and Wang, Z. X. (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.* **16,** 109–114
57. Carugo, O., and Argos, P. (1997) Correlation between side chain mobility and conformation in protein structures. *Protein Eng.* **10,** 777–787

58. Carugo, O., and Argos, P. (1997) Protein-protein crystal-packing contacts. *Protein Sci.* **6,** 2261–2263
59. Carugo, O., and Argos, P. (1998) Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* **31,** 201–213
60. Tonrud, D. E. (1996) Knowledge-based B-factor restraints for the refinement of proteins. *J. Appl. Crystallogr.* **29,** 100–104
61. Stroud, R. M., and Fauman, E. B. (1995) Significance of structural changes in proteins: expected errors in refined protein structures. *Protein Sci.* **4,** 2392–2404
62. Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S., and Vranken, W. (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.* **31,** 458–462
63. Finkelshtein, A. V., and Ptitsyn, O. B. (2002) *Physics of Proteins*, pp. 9–11, Academic Press, London
64. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540
65. Wise, E., Yew, W. S., Babbitt, P. C., Gerlt, J. A., and Rayment, I. (2002) Homologous (β/α)8-barrel enzymes that catalyze unrelated reactions: orotidine 5′-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase. *Biochemistry* **41,** 3861–3869
66. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321,** 741–765
67. Jurgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000) Directed evolution of a ($\beta\alpha$)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9925–9930
68. Bork, P., and Koonin, E. V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.* **18,** 313–318
69. Hegyi, H., and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288,** 147–164