Protein Science (1999), *8*:1752–1764. Cambridge University Press. Printed in the USA.

Predicting conformational switches in proteins

MALIN YOUNG,[1,3] KENT KIRSHENBAUM,[1,4] KEN A. DILL,[1] AND STEFAN HIGHSMITH[2]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-0446
[2]Department of Biochemistry, University of the Pacific School of Dentistry, San Francisco, California 94115-2399

(RECEIVED January 8, 1999; ACCEPTED May 14, 1999)

Abstract

We describe a new computational technique to predict conformationally switching elements in proteins from their amino acid sequences. The method, called ASP (Ambivalent Structure Predictor), analyzes results from a secondary structure prediction algorithm to identify regions of conformational ambivalence. ASP identifies ambivalent regions in 16 test protein sequences for which function involves substantial backbone rearrangements. In the test set, all sites previously described as conformational switches are correctly predicted to be structurally ambivalent regions. No such regions are predicted in three negative control protein sequences. ASP may be useful as a guide for experimental studies on protein function and motion in the absence of detailed three-dimensional structural data.

Keywords: ASP; conformational switch; PHD; secondary structure preferences; structurally ambivalent sequence element

The biological activity of a protein often depends on its ability to undergo conformational rearrangements. Because protein sequences are far more readily available than structures, it would be useful to predict sites of conformational switches from amino acid sequence information alone. Here we describe such an algorithm.

Much evidence supports the idea that a given amino acid sequence can adopt different conformations, depending on its solvent environment (Zhong & Johnson, 1992; Waterhous & Johnson, 1994) or its environment inside a protein (Kabsch & Sander, 1984; Cohen et al., 1993; Han & Baker, 1996; Minor & Kim, 1996; Sudarsanam, 1998). Our method is based on ideas that originated with Chou and Fasman (1974). First, that secondary structures are partly predictable from local sequence information about intrinsic propensities to form a helix or strand. Second, that predictions of ambivalent propensities can help define the regions in peptides that are potentially capable of undergoing conformational change. Evaluation of peptide sequences allowed prediction of sites of environment-induced conformational changes in glucagon and preproparathyroid hormone (Chou & Fasman, 1975; Rosenblatt et al., 1980; Fasman, 1987). We extend these ideas from the study of peptides to the study of complete protein structures. Our approach follows the premise that if a protein subsequence lacks a strong intrinsic secondary structure preference, a change in its environment can alter its conformation.

Typically, secondary structure prediction methods rely upon associating homologies between new sequences and those of proteins of known structure. Whereas the Chou–Fasman rules were originally based on only 19 protein structures, now nearly 8,000 structures are known. The success rates of the best methods are now in the range of 70–75% (Rost & Sander, 1993). To evaluate secondary structure propensities, we use the PHD program of Rost and Sander (1993). For a given sequence, PHD finds and aligns homologs to that sequence. These are put into a neural network, which has been trained on a databank of structures. The output of PHD includes a scaled probability that each input residue will be found in a helix, strand, or other secondary structure (e.g., loop).

We introduce a program called ASP (Ambivalent Structure Predictor) that uses the scaled probabilities from PHD to identify structurally ambivalent sequence elements (SASEs). These sites are predicted to have a greater potential to exchange between secondary structural classes. In addition, we may also predict sites of more transient rearrangements involving disruption of a secondary structure element. We analyze the sequences of 16 proteins that undergo conformational switching as part of their biological activity. As a test of the method, we correlate the sequence elements predicted to be structurally ambivalent by ASP with the portions of protein structure previously shown by experiment to undergo conformational switching.

Results

Selection of test protein sequences

Our aim is to predict sites prone to conformational rearrangement in a diverse set of proteins that exhibit switches in secondary

Reprint requests to: Stefan Highsmith, Department of Biochemistry, University of the Pacific, 2155 Webster Street, San Francisco, California 94115-2399; e-mail: shighsmith@uop.edu.

[3]Current address: Sandia National Laboratories, Livermore, California 94550.

[4]Current address: Department of Chemistry and Chemical Engineering, Mail Code 210-41, California Institute of Technology, Pasadena, California 91125.

Abbreviations: ASP, the computer algorithm Ambivalent Structure Predictor; $\Delta Pr(k)$, residue ambivalence score for position k; $\mu(k)$, mean ambivalence score for residue k as part of a windowed subsequence; z, z score cutoff; SASE, structurally ambivalent sequence element.

structure as part of their activity. To minimize bias in the selection process, we sought a pre-existing compilation of suitable proteins. We wanted to evaluate comprehensively a wide selection of proteins that undergo backbone rearrangements. The Database of Macromolecular Motions, which categorizes over 120 motions, is such a compilation (Gerstein & Krebs, 1998). While most protein motions may not include substantial secondary structure rearrangements, the hierarchical classification in this database allowed us to identify the appropriate group for analysis. Protein motions are categorized on the basis of size and on the basis of packing (Gerstein & Krebs, 1998). We chose the domain category for size, excluding sequences less than 150 residues in length to minimize end effects. We considered all such proteins as candidates for our test set, except those with motions that are predominantly simple swivel, hinge, and/or shear (see Discussion).

All such sequences were submitted to the EMBL PHD mail server (http://www.embl-heidelberg.de/predictprotein/predictprotein.html) for secondary structure prediction (Rost & Sander, 1993; Rost, 1996). The reliability of secondary structure prediction by PHD is partially dependent on the number of sequence homologues in the SWISS-PROT data bank available as input to the neural net. Therefore, proteins with fewer than 10 homologues were not analyzed further. This selection process produced a test set of 15 proteins: four members of the serpin family, transducin-α, adenylyl cyclase activator α subunit, elongation factor Tu, NF-κB p50, TATA-box binding protein, $\gamma\delta$ resolvase, prior protein, topoisomerase I, aldose reductase, hemagglutinin, and myosin. We have analyzed more thoroughly results for those proteins with conformational transitions that are well characterized (the serpins, transducin-α, adenylyl cyclase activator α subunit, elongation factor Tu). Conformational rearrangements are suspected for the remaining proteins in the test set, but have not been as thoroughly described. We include analyses of these proteins, which may serve as true predictive test of method as future experimental findings become available.

Serpins

The serpins are a family of monomeric proteins, most of which function to inhibit serine protease activity (Potempa et al., 1994). The inhibitory serpins undergo large-scale conformational rearrangements, primarily following cleavage of the protein in the reactive center loop (RCL), which then becomes inserted as a strand into a large β-sheet (Stein & Chothia, 1991; Wisstock et al., 1998). Noninhibitory serpins, such as ovalbumin, do not demonstrate this rearrangement; ovalbumin is included in our negative control set.

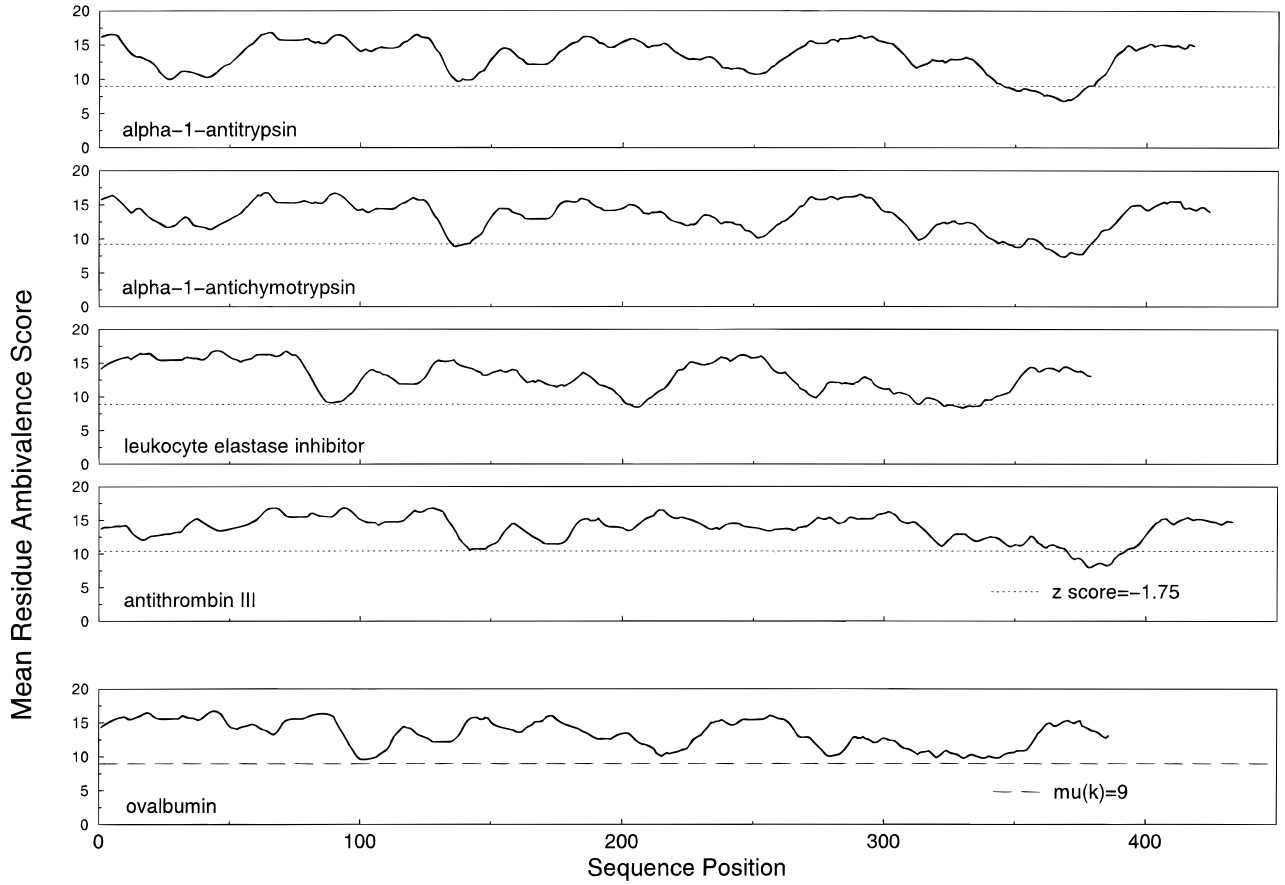

Fig. 1. Mean residue ambivalence scores, $\mu(k)$, for five serpin sequences. Smaller values are predicted to have an enhanced proclivity for conformational switching. Residues with scores below a *z* score cutoff of -1.75 standard deviations (dotted line) are defined to be structurally ambivalent. The sequences of four inhibitory serpins (α-1-antitrypsin, α-1-antichymotrypsin, LEI, antithrombin III) contain structurally ambivalent sequence elements (SASEs) corresponding to the reactive center loops. The noninhibitory serpin ovalbumin (bottom) does not contain any ambivalent sequence elements (all $\mu(k)$ scores > 9, dashed line).

Table 1. *Predicted structurally ambivalent sequence elements (SASEs)*

Test protein	WS[a]	Accession number[b]	SASEs
α-1-antitrypsin	21	A1AT_HUMAN	346–377
α-1-antichymotrypsin	21	AACT_HUMAN	136–141, 348–353, 361–378
Antithrombin III	21	ANT3_BOVIN	370–392
LEI	21	ILEU_HORSE	204–207, 322–337
Gsα	5	GBAS_HUMAN	126–128, 161–170, 201–203, 233–234, 236–242, 251–258, 280–282
Gtα	5	GBT1_BOVIN	68, 138–142, 173–176, 205–213, 224–230, 251–255
EF-Tu	5	EFTU_THEAQ	46–51, 64–66, 98–100, 175–179, 224–226, 293–300, 314–315
NF-κB p50	5	KBF1_MOUSE	132–138, 144–147, 209–213, 304–308
$\gamma\delta$-resolvase	5	TNP1_ECOLI	137, 168–173, 177–179
TATA-box binding protein	5	TF2D_YEAST	64–67, 105, 151, 166–172
DNA topoisomerase I	5	TOP1_ECOLI	33–36, 68–70, 219–221, 236–238, 434–439, 509–519, 672–673, 698–700, 718–722, 826–839
Prion protein	5	PRIO_MOUSE	120–122, 128–133, 214, 246–250
Aldose reductase	5	ALDR_HUMAN	14–16, 40–43, 102–105, 251–255, 274–277, 295–297
Haemagglutinin	21	HEMA_IAAIC	325–341, 360–372
Myosin heavy chain	21	MYSS_CHICK	38–51, 235–247, 251–256, 279–296, 310–316, 357–372, 404–416, 459–481, 484–489, 501–508, 580–587, 691–708, 754–769
Control protein	**WS[a]**	**Accession number[b]**	**SASEs**
Ovalbumin	21	OVAL_CHICK	None
Collagen	5–21	CA13_BOVIN	None
Myosin rod region	5–21	MYSS_CHICK	None

[a]WS is the window size used in the analysis, as discussed in Methods.
[b]The accession numbers identify the sequences obtained from the Swiss-Prot Data Bank.

We analyzed the sequences of five serpins (Fig. 1), four of which are serine proteinase inhibitors (α1-antitrypsin, antithrombin-III, α1-antichymotrypsin, and leukocyte elastase inhibitor (LEI)), and one of which is noninhibitory (ovalbumin). For the four inhibitory serpins, ASP identifies structurally ambivalent sequence elements (Table 1). These SASEs map to the reactive center loops in the crystal structures (α1-antitrypsin: 346–377; antithrombin-III: 370–392; α1-antichymotrypsin: 348–353, 361–378, LEI: 332–337) (Table 2;

Table 2. *Comparison of known and predicted switch locations*

Protein	Switch	Published residue numbers	Predicted residue numbers
α-1-antitrypsin	RCL[a]	366–382	346–377
α-1 antichymotrypsin	RCL	366–382	361–378
Antithrombin III	RCL	378–394	370–392
LEI	RCL	328–344	322–337
Gsα[b]	Switch I	199–207	201–203
	Switch II	224–240	233–234, 236–242
	Switch III	256–264	251–258
Gtα[c]	Switch I	173–183	173–176
	Switch II	195–215	205–213
	Switch III	227–238	224–230
EF-Tu[d]	Switch I	40–62	46–51, 64–66
	Switch II	80–100	98–100
Myosin[e]	Switch I	235–245	235–247
	Switch II	466–506	459–481, 484–489, 501–508

[a]For the serpins, the switch is defined as the residues in the reactive center loop (RCL) C-terminal to strand 5A and N-terminal to the PI′ residue.
[b]Sunahara et al. (1997).
[c]Lambright et al. (1996).
[d]Abel et al. (1996).
[e]Putative switches as defined by Smith and Rayment (1996a).

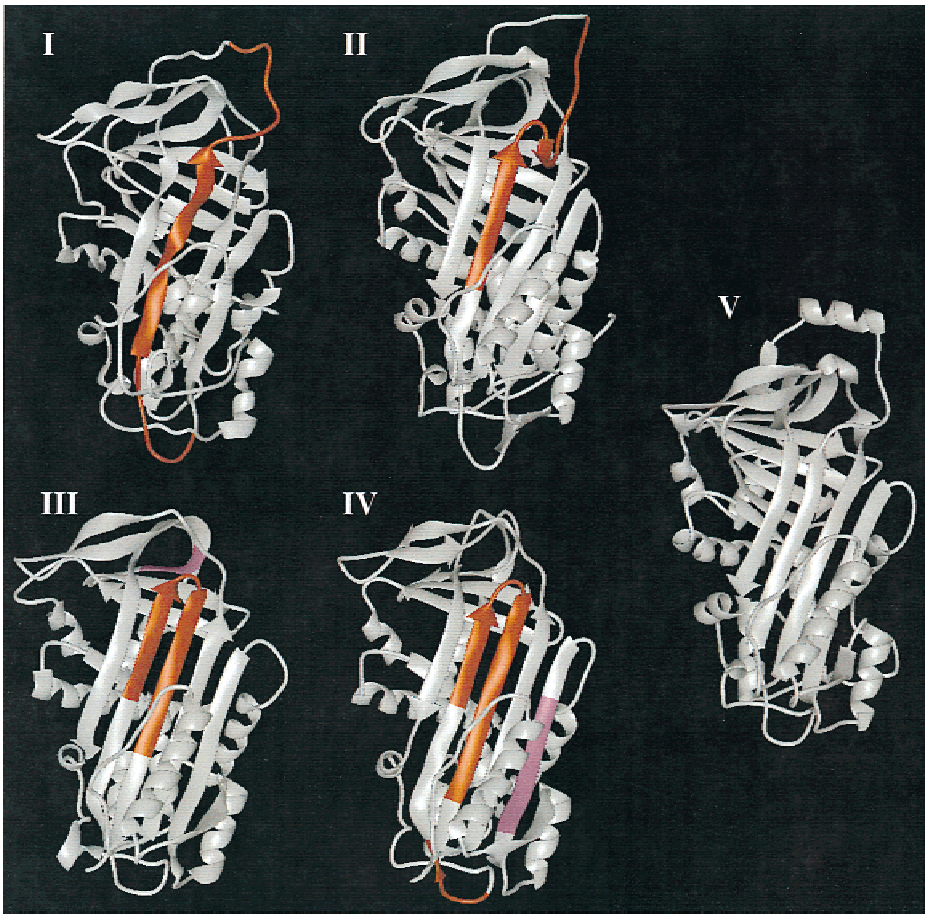

Fig. 2. Sequence elements predicted to be ambivalent are mapped onto the crystal structures of four inhibitory serpins. In each case, the reactive center loop changes to β-sheet upon cleavage (see text). The loop conformation can be seen in α-1-antitrypsin (I) and antithrombin III (II). The sheet form is seen in LEI (III) and α-1-antichymotrypsin (IV). Predicted SASEs in the reactive center loop are shown in red. Additional SASEs are shown in magenta. No SASEs were predicted for the control sequence of the noninhibitory serpin ovalbumin (V).

Fig. 2). In contrast, no SASEs are found in the ovalbumin sequence, consistent with the lack of conformationally labile regions expected for this control sequence.

In the cases of LEI and α1-antichymotrypsin, one additional subsequence is predicted to be ambivalent. The LEI SASE (204–207) stretches across one edge of the β-sheet involved in the serpin conformational transformation (the A-sheet), is located at the base of the RCL (Fig. 2, in magenta), and may be involved in the insertion of the cleaved RCL into the β-sheet. The α1-antichymotrypsin SASE maps to strand s2A (Fig. 2, in magenta). Comparison of serpin structures shows that s2A moves upon cleavage of the RCL (Stein & Chothia, 1991; Elliott et al., 1998).

Transducin-α ($G_{t\alpha}$)

$G_{t\alpha}$ is the α-subunit of a heterotrimeric G protein involved in modulating the vertebrate light response via rhodopsin-induced GTP-binding. Upon exchange of GDP for GTP, $G_{t\alpha}$-GTP undergoes a conformational change. It disassociates from the $\beta\gamma$ subunits and subsequently activates its effector, cGMP-specific phosphodiesterase. Hydrolysis of the GTP to GDP eventually triggers reassociation of the G protein heterotrimer with rhodopsin, thus completing the cycle of activation. The conformational changes observed in $G_{t\alpha}$ upon binding GTP have been localized to three switch regions, named switch I, switch II, and switch III (Noel et al., 1993; Sondek et al., 1994; Lambright et al., 1996).

The ASP profile for bovine $G_{t\alpha}$ indicates there are five SASEs with length greater than one: 138–142, 173–176, 205–213, 224–230, and 251–255 (Fig. 3; Table 1). Three of the five SASEs correspond to switches (Table 2) previously identified (Lambright et al., 1996). SASE 173–176 maps to switch I which is involved in phosphate bindng. SASE 205–213 maps to switch II and includes a neighboring effector-binding region 209–212. SASE 224–230 overlaps the N-terminal portion of switch III.

When the two other SASEs (138–142, 251–255) are mapped onto the crystal structure of $G_{t\alpha}$ (Fig. 4), it can be seen that SASE 138–142 is in van der Waals contact with SASE 224–230. Similarly, SASE 138–142 is proximal to residues 145–149, which are known to be involved in binding the guanine of GTP (Noel et al., 1993). SASE 251–255 is the $G_{t\alpha}$ $\alpha 3/\beta 5$ surface loop, which has been implicated in phosphodiesterase binding and activation (Noel et al., 1993). It, too, is in van der Waals contact with one of the three switches, specifically switch II. Residue F254 is, in addition, one of a series of interhelical contacts linking surface loops $\alpha 2/\beta 4$, $\alpha 3/\beta 5$, and $\alpha 4/\beta 6$ (Noel et al., 1993). The interhelical contacts in

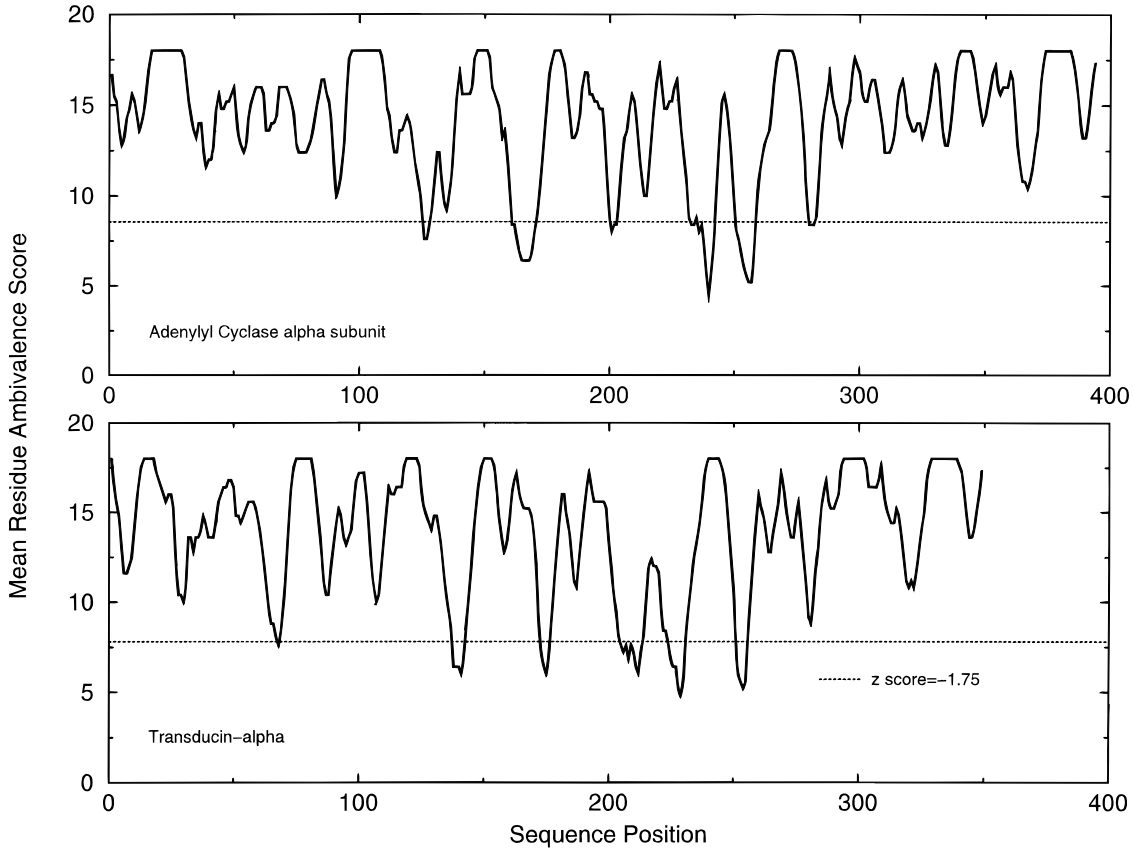

Fig. 3. Mean residue ambivalence scores, $\mu(k)$, for two G protein α subunits. Both proteins (adenylyl cyclase α subunit, transducin-α) have SASEs (z scores < -1.75 standard deviations, below dotted line) that correspond to known conformational switches in the tertiary structures.

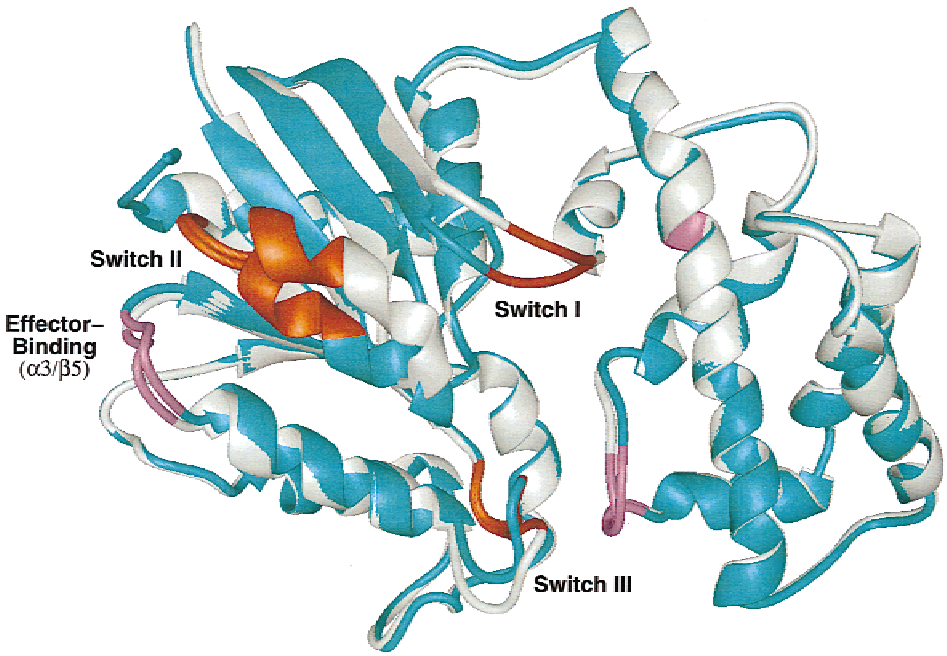

Fig. 4. Structurally ambivalent sequence elements for transducin-α mapped onto the aligned crystal structures of the GDP-bound (1tag, white) and the GTP-bound (1tnd, cyan) conformations. Sites of conformational rearrangement correspond to the locations of predicted SASEs (in red and magenta). The SASEs mapping to the previously identified Switches I, II, and III are shown in red (Noel et al., 1993; Lambright et al., 1996).

these three loops are thought to transmit information from the phosphate-binding $\alpha 2$ helix to the effector binding loops $\alpha 3/\beta 5$ and $\alpha 4/\beta 6$. SASE 251–255 may, therefore, function in the transmission of information from the GTP-binding site to switch II.

Adenylyl cyclase activator α subunit ($G_{s\alpha}$)

$G_{s\alpha}$, like $G_{t\alpha}$, is an α-subunit of a heterotrimeric G-protein complex, which activates adenylyl cyclase upon binding GTP. The tertiary structures of $G_{s\alpha}$ and $G_{t\alpha}$ can be superimposed with a root-mean-square deviation (RMSD) of 1.8 Å over 303 Cα positions (Holm & Sander, 1998). The sequence identity, on the other hand, is only 36.5%. $G_{s\alpha}$ contains three switch regions that undergo similar conformational rearrangements as those described for $G_{t\alpha}$. A valuable test of our method would be the identification of similar SASEs in $G_{s\alpha}$ and $G_{t\alpha}$ in the absence of substantial sequence identity.

ASP identified six SASEs in human $G_{s\alpha}$: 126–128, 161–170, 201–203, 233–234/236–242, 251–258, and 280–282 (Table 1; Fig. 3). Three of the six SASEs correspond to the three known conformational switches, I, II, and III, in $G_{s\alpha}$ (Sunahara et al., 1997; Tesmer et al., 1997; Wall et al., 1998) (Table 2; Fig. 4). SASE 201–203 corresponds to switch I. SASE 233–242 maps to switch II and SASE 251–258 maps to switch III. SASE 280–282 is part of the $\alpha 3/\beta 5$ surface loop, which binds and activates adenylyl cyclase.

ASP identified all three switch regions, as in the case of $G_{t\alpha}$, even though the sequence identity shared by the two proteins is relatively low. When the six SASEs are mapped onto the crystal structure (Fig. 5), five form a coherent region on the protein surface that markedly resembles the interaction footprint previously described by Sunahara et al. (compare to Fig. 3 of Sunahara et al., 1997).

Elongation factor Tu (EF-Tu)

EF-Tu is a G protein that plays a role in polypeptide elongation. EF-Tu undergoes a series of conformational changes during protein synthesis, which are due to its interactions with GDP, GTP, aminoacyl-tRNA, and elongation factor Ts. X-ray crystallographic studies of variously liganded EF-Tu complexes have identified two conformational switch regions (Abel et al., 1996; Polekhina et al., 1996). A seven amino acid portion of switch I (53–59) has been shown to undergo an α-helix to β-strand transition upon GTP hydrolysis.

Seven SASEs predicted for the primary sequence of EF-Tu. They are: 46–51, 64–66, 98–100, 175–179, 224–226, 293–300, 314–315 (Table 1; Fig. 6). Three of the seven SASEs map to the two known conformational switches in EF-Tu (Table 2). SASEs 46–51 and 64–66 correspond to residues at the N-terminus and near the C-terminus of switch I. SASE 46–51 includes residues Y47 and D51, which are involved in GTP-binding. Surprisingly, the seven residue portion of switch I that undergoes a transition from α-helical to β-strand structure upon GTP hydrolysis is not identified as a structurally ambivalent sequence element. However, SASEs 46–51 and 64–66, which ASP does identify, flank these seven residues. These flanking regions contain residues (Asp51 and Asn64) that have been described to undergo "flips" that initiate the $\alpha \rightarrow \beta$ structural transitions (Abel et al., 1996).

SASE 98–100 is located at the C-terminus of switch II. SASE 175–179 is part of the GTP binding site. The side chain of S174 and the backbone amides of A175 and L176 are involved in hydrogen-bonding GTP (Sobolev et al., 1996). The remaining two ambivalent regions, SASEs 224–226 and 293–300, are both located in domain 2, and they are in van der Waals contact with one another.

NF-κB p50

NF-κB p50 is a prototypical member of a family of mammalian transcription factors containing a ~280 amino acid "Rel homology region" (RHR) found within the amino-terminal fragment of the protein. NF-κB p50 forms a functional dimer that binds to specific sequences in the DNA major and minor grooves. Residues in a loop near the core of the dimer interface make sequence-specific

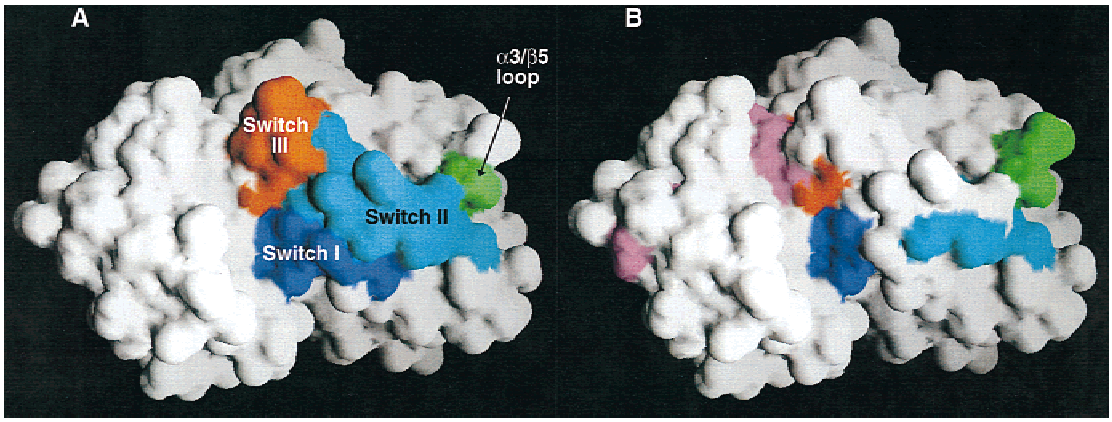

Fig. 5. Comparison of the known $G_{s\alpha}$ conformational switch regions with the predicted structurally ambivalent sequence elements. **A:** Surfaces representing conformational switch I, II, and III are mapped in color onto the solvent-accessible surface along with the $\alpha 3/\beta 5$ effector-binding loop. These surfaces were identified as part of the $G_{s\alpha}$ "interaction footprint" (Sunahara et al., 1997). **B:** The three SASEs predicted for $G_{s\alpha}$ corresponding to switches I, II, and III are residues 201–203 (blue), 233–234/236–242 (cyan), and 251–258 (red), respectively. The SASE mapping to the $\alpha 3/\beta 5$ loop consists of residues 280–282 (green). Two additional SASEs are shown in magenta (see text).

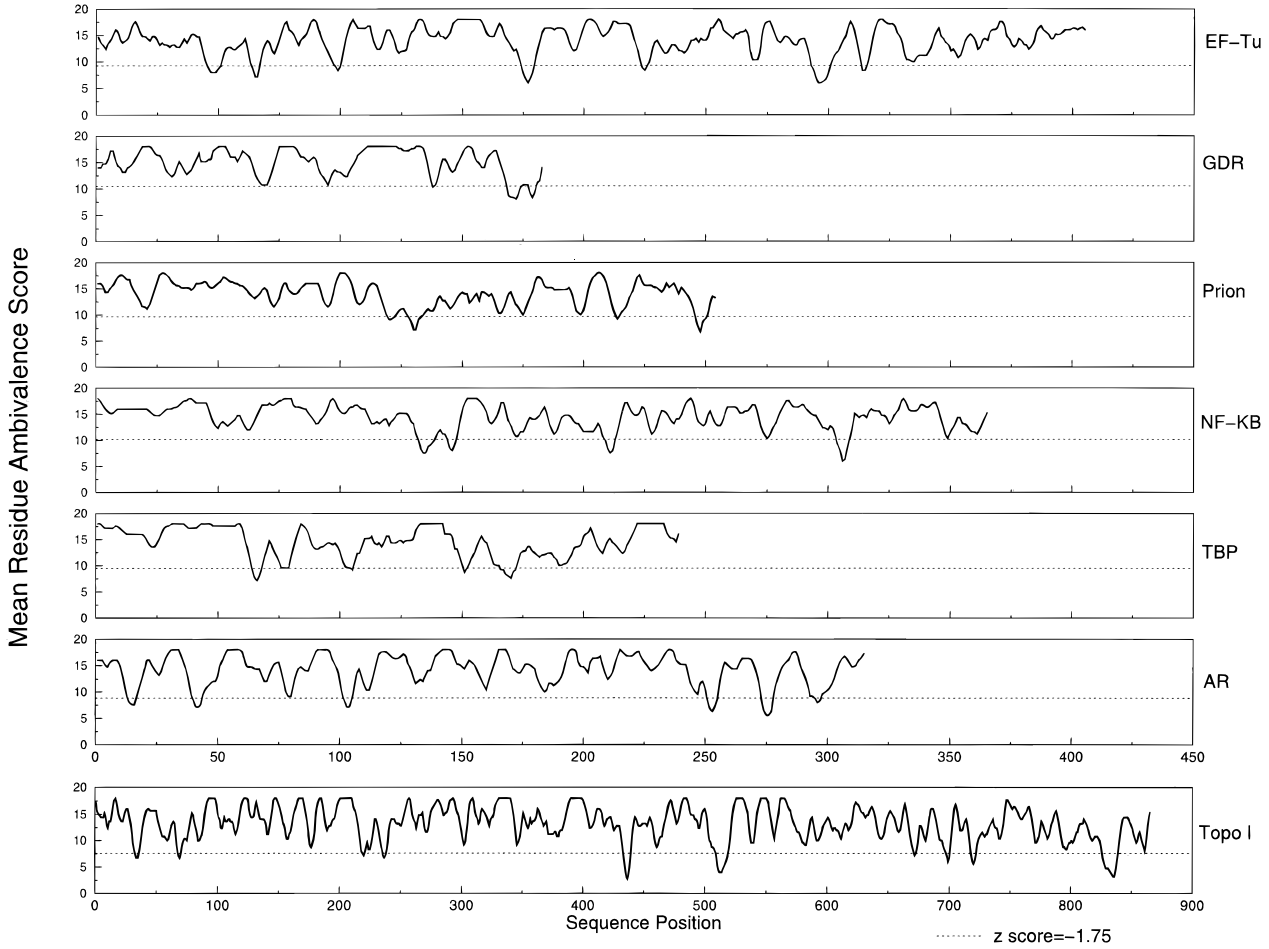

Fig. 6. Mean residue ambivalence scores, $\mu(k)$, for seven test proteins that undergo localized conformational rearrangements. From top to bottom they are: elongation factor Tu, $\gamma\delta$-resolvase, prion protein, NF-κB p50, TATA-box binding protein, aldose reductase, and topoisomerase I. Predicted SASEs consist of residues with z scores less than -1.75 standard deviations (below dotted line).

contacts to DNA (Ghosh et al., 1995; Muller et al., 1995) in the major groove. Muller et al. have postulated the presence of an allosteric switch in which conformational changes at the dimer interface would be communicated to adjacent residues that contact the DNA (Muller et al., 1995). Residues in the minor groove binding site form a loop that contacts the DNA backbone phosphate groups. These loops "may be more flexible" in order that they may "adjust to different sequences, different binding partners, and possibly to neighboring transcription factors" (Ghosh et al., 1995).

ASP analysis predicts four SASEs in NF-κB p50: 132–138, 144–147, 209–213, 304–308 (Table 1; Fig. 6). Mapping the SASEs onto the crystal structure of NF-κB reveals that we identify two DNA-binding loops (Fig. 7A). Residues 304–308 correspond to the major-groove binding site. Residues 144–147 includes all residues that contact DNA in the minor groove. The remaining SASEs (132–138 and 209–213) are proximal to the DNA-binding loop in the minor groove (around 144–147) and flank the helical segments in the NF-κB structure. These helices are thought to lie near the binding site of co-activating transcription factors (Ghosh et al., 1995; Muller et al., 1995).

$\gamma\delta$ Resolvase

$\gamma\delta$ Resolvase is a homodimeric recombinase that catalyzes the site-specific catenation of double-stranded DNA. The crystal structure of resolvase has been determined as a dimer in complex with a 34-base pair segment of DNA containing a recombination site (Yang & Steitz, 1995). The dimer displayed substantial asymmetry and the catalytic residue was distant from the cleavage site. Therefore, this study concluded that conformational lability may be important in resolvase function. Structural adaptability would also allow resolvase to recognize different DNA-binding sites in distinct configurations while retaining high affinity.

Consistent with these results, ASP analysis identified two SASEs with length greater than 1: 168–173 and 177–179 (Table 1; Fig. 6). Residues 168–173 and 177–179 are part of a three-helix bundle in the C-terminal domain. These subsequences correspond to residues observed to make sequence specific contacts in the DNA major groove (Yang & Steitz, 1995; Fig. 7B). Conformational changes in this region have also been demonstrated by solution NMR studies of resolvase (Pan et al., 1997). In addition, a single-amino acid SASE, residue G137, was predicted by ASP. It is near the minor-

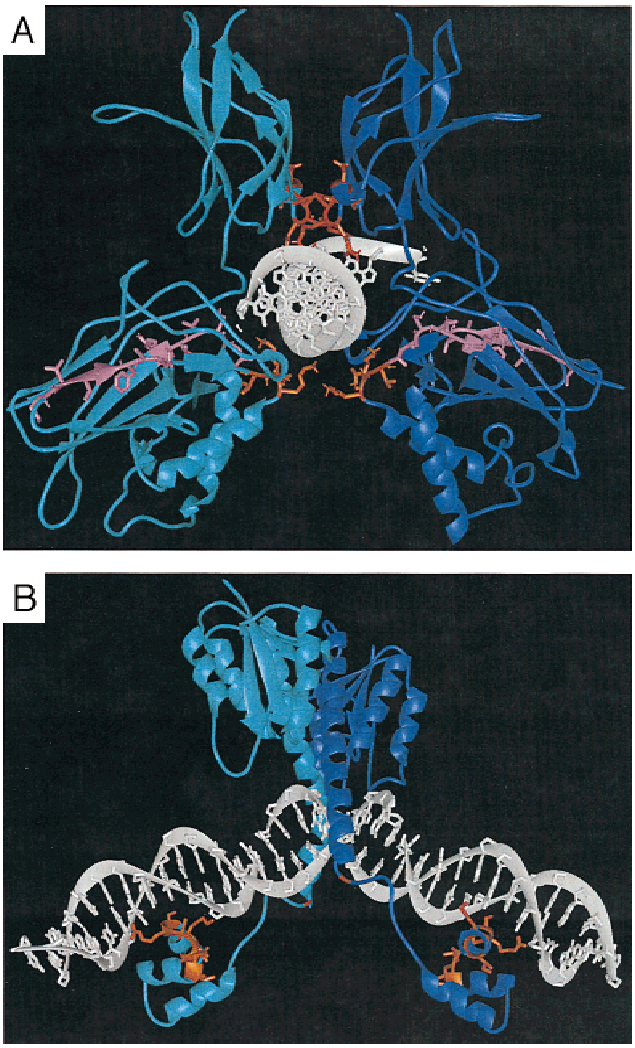

Fig. 7. Predicted SASEs mapped onto the crystal structures of two homodimeric DNA-binding proteins (Ghosh et al., 1995; Yang & Steitz, 1995). **A:** SASEs predicted for NF-κB p50. SASEs corresponding to DNA-binding sites are shown in red. Additional SASEs are shown in magenta. **B:** SASEs predicted for $\gamma\delta$-resolvase are shown in red. These SASEs correspond to regions thought to undergo conformational rearrangement to allow sequence-specific recognition of DNA (see text).

groove binding site of $\gamma\delta$ resolvase and is located at the C-terminus of a central helix (Yang & Steitz, 1995). This portion of the helix is thought to be present only in the presence of DNA and may constitute an important conformational switch upon DNA-binding.

TATA-box binding protein (TBP)

TBP is a major component of the transcription factor TFIID, which is involved in transcriptional regulation. Inspection of an asymmetric structure in the absence of DNA indicates that the twisting of a central sheet changes the relative position of the two TBP domains by $\sim 10^\circ$ (Chasman et al., 1993).

Two SASEs with lengths greater than two residues were predicted by ASP for the sequence of yeast TBP: 65–68, 167–173 (Table 1; Fig. 6). Residues 65–68 map to strand s1 in the crystallographic structure (1tpb). This strand occupies a central position in the flexible sheet. Site-directed mutagenesis studies showed that one amino acid, L67, in this region is involved in DNA binding and transcription (Kim JL et al., 1993). The other SASE spans residues 167–173. L172 and the nearby L175 are two additional residues shown to be critical in DNA binding and transcription (Kim Y et al., 1993).

DNA topoisomerase I

DNA topoisomerase I is a member of a family of enzymes that catalyze alterations of DNA topology. These reactions entail the passage of a DNA segment through a transient break in another DNA strand. An X-ray crystallographic structure has been obtained for the N-terminal 67 kDa fragment (residues 2–590) of the 97 kDa *Escherichia coli* protein (Lima et al., 1994). The study concluded that topoisomerase I would need to undergo substantial conformational changes to expose an active-site tyrosine and allow DNA strand passage. It was postulated that this involves a "gate opening" motion in the toroidal architecture of the protein, allowing for rigid body rotation of two of the four topoisomerase I domains. NMR solution structures on the C-terminal DNA-binding domain of *E. coli* topoisomerase I (residues 745–865) have also been obtained (Yu et al., 1995, 1996).

ASP analysis of the full length *E. coli* sequence identified 10 SASEs: 33–36, 68–70, 219–221, 236–238, 434–439, 509–519, 672–673, 698–700, 718–722, 826–839 (Table 1; Fig. 6). The first six SASEs are located in the N-terminal fragment studied by X-ray crystallography. The bulk of these residues (219–221, 236–238, and 509–519) are located in strands at opposite sides of sequence elements connecting topoisomerase domains II and IV. Conformational transitions in these strands may trigger the rigid-body rotation necessary for "gate-opening."

One SASE (826–839) is in the C-terminal fragment studied by NMR (Yu et al., 1995, 1996). This region corresponds to residues on the DNA-binding face that were found to have altered mobility in the complexed form. Changes in the topoisomerase I secondary structure upon DNA binding may be focused at these residues. The remaining three SASEs are located in the intermediate portion of topoisomerase I for which a structure is not yet available.

Prion proteins

Prions are a family of proteinaceous infectious agents for a variety of neurodegenerative illnesses. A mechanism has been postulated for prion pathogenesis in which a conformational transition converts a largely α-helical cellular isoform, PrPc, to an infectious isoform, PrPsc, containing substantial β-sheet structure. Although no high resolution structure is available for the PrPsc form, structures of intact soluble PrPc form PrP(23–231), and its fragments have been obtained by NMR. Both the mouse and hamster prion proteins have residues in the region 121–131, which form an "unusual and dynamic structural feature" (James et al., 1997; Riek et al., 1998). It was postulated that a large number of glycine residues (five between positions 123 and 131) permits many alternative conformations (James et al., 1997).

ASP analysis of the mouse prion protein identifies three SASEs greater than one residue (Table 1; Fig. 6). Two, 120–122 and 128–133, are in the N-terminal region studied by NMR. These sites are also found in analysis of the hamster sequence (data not shown). The SASEs identified by ASP are located in the "domain with marginally stable polymorphic structure" (James et al., 1997).

Aldose reductase

Aldose reductase (AR) is a NADPH-dependent enzyme. Although its complete physiological role is still undetermined, AR is capable of catalyzing the reduction of glucose to sorbitol and has been implicated in various diabetic complications. Three-dimensional structures of AR have been obtained (Bohrani et al., 1992; Wilson et al., 1992; Urzhumtsev et al., 1997). These studies demonstrated that binding of NADPH or inhibitors requires conformational rearrangements to expose a cleft in the protein, allowing the ligands access to otherwise closed pockets.

ASP analysis of human AR found six SASEs: 14–16, 40–43, 102–105, 151–155, 274–277, and 295–297 (Table 1; Fig. 6). These residues may undergo structural rearrangements involved in enzyme activity and regulation.

Haemagglutinin (HA)

Influenza virus HA is a membrane glycoprotein responsible for interacting with target cell surfaces. HA is comprised of two subunits, HA_1 and HA_2, which are derived from proteolytic processing of the intact HA_0 precursor. The conformation of the cleaved HA is metastable and undergoes extensive structural reorganization at the low pH of endosomes, enabling membrane fusion. Among these changes is the ~100 Å movement of the N-terminal "fusion peptide" region of HA_2, which relocates to the end of a triple-stranded coiled-coil (Bullough et al., 1994). Recent studies have compared the three-dimensional structures of HA and HA_0 to determine why the uncleaved HA_0 form is incapable of a pH-induced conformational change (Chen et al., 1998). One possibility is that local rearrangements upon cleavage result in burial of ionizable residues in an adjacent surface cavity, thus setting "a low-pH trigger." Another possibility is that motions in the boundary between HA_1 and HA_2 may be restricted, preventing extensive conformational rearrangements until liberated by cleavage (Chen et al., 1998).

The intact hemagglutinin precursor sequence (HEMA_IAAIC) was analyzed with ASP. The results identified two SASEs: 325–341 and 360–372 (Fig. 8; Table 1), corresponding to residues 309–325 of HA_1 and 15–27 of HA_2. These sequences flank the cleavage site and include part of the highly mobile fusion peptide in the N-terminus of HA_2 (Bullough et al., 1994). The sequence element including the cleavage site was not predicted to conformationally labile. Proteolytic cleavage at the HA_1-HA_2 boundary may thus enable the localized backbone rearrangements flanking this site and potentiate further refolding upon transfer to a low pH regime.

Myosin

The myosins are a family of motor proteins that interact with actin to produce force. The motor domain is thought to undergo substantial conformational rearrangements that includes rotation of a segment about a fulcrum site during this process (Dominguez et al., 1998). The nature of this lever arm rotation is complex. Rotation is coordinated by the ATP and actin binding at site distant from the fulcrum site. Several crystal structures of the myosin motor domain have been solved (Rayment et al., 1993; Smith & Rayment, 1996a; Dominguez et al., 1998). Two switches near the ATP site have been postulated (Smith & Rayment, 1996b; Kull et al., 1996). However, neither the details of the myosin motor switch rearrangements nor the pathways for communicating between the ATP site, actin site, and the fulcrum site are understood.

We have performed ASP analyses on chicken skeletal muscle myosin. There are 13 SASEs predicted, all of which are in the 1–843 N-terminal residues of the heavy chain, which is the motor domain (Table 1). The two putative switches (Smith & Rayment, 1996b; Kull et al., 1996) are among the subsequences predicted to undergo conformational changes, as is the fulcrum site (Table 2). In addition, the SASEs are near one another in the tertiary structure and form plausible allosteric connections between the ATP site, the actin binding site, and the fulcrum site. Detailed results and analysis of the allosteric connections are presented in an accompanying paper (Kirshenbaum et al., 1999).

Evaluation of proteins lacking conformational switches

ASP analyses of ovalbumin, bovine collagen, and the rod region of chicken myosin II heavy chain were performed as negative controls. Ovalbumin is a globular protein with both α-helical and β-sheet secondary structure. As discussed above, this noninhibitory serpin is thought not to undergo conformational rearrangement. ASP does not identify any SASEs in this sequence. Ovalbumin

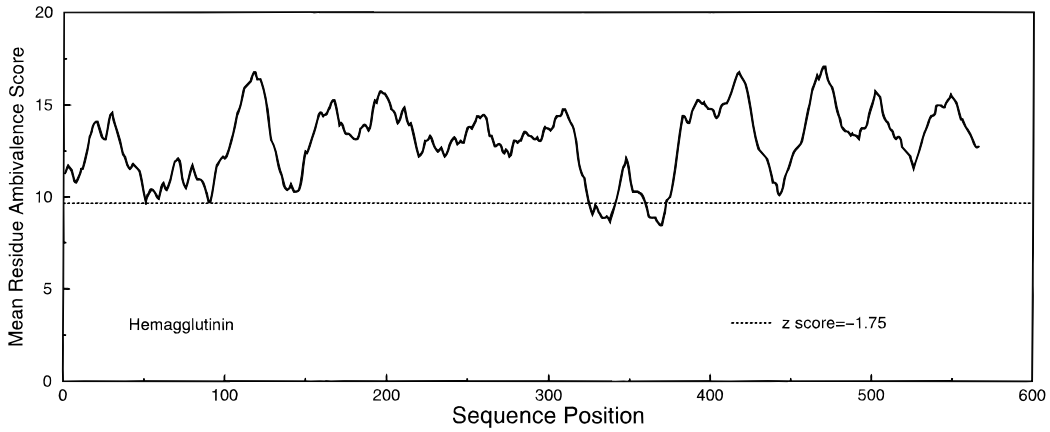

Fig. 8. Mean residue ambivalence scores, $\mu(k)$, for haemagglutinin, a protein that undergoes extensive conformational rearrangements (window size = 21). Predicted SASEs consist of residues with z scores less than −1.75 standard deviations (below dotted line).

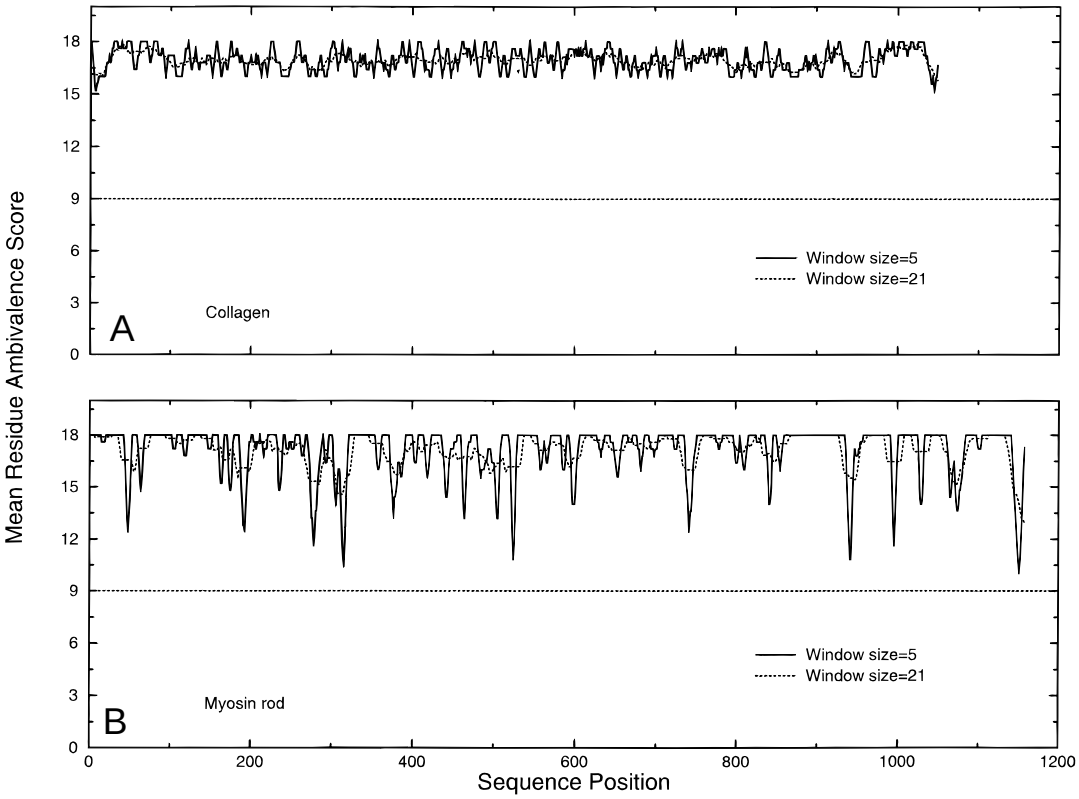

Fig. 9. Mean residue ambivalence scores, $\mu(k)$, for two negative control proteins (see also ovalbumin in Fig. 1): (**A**) collagen and (**B**) myosin II heavy chain rod region. Profiles are shown with window sizes of 5 and 21. No SASEs were identified for either protein (all $\mu(k) > 9$).

is a useful control because it can be compared directly to the inhibitory serpins (see above).

Collagen is a triple helix and the myosin rod is a coiled coil of α-helices. These proteins lack enzymatic function or conformational switch regions. The $\mu(k)$ profile for collagen clearly indicates that there are no ambivalent subregions, as all of the $\mu(k)$ scores for this sequence were $\geq$14, using window sizes of 5 and 21 amino acids (Fig. 9). Similarly, the $\mu(k)$ profile for the rod region (amino acids 780–1938) of the chicken skeletal muscle myosin heavy chain shows uniformly high $\mu(k)$ scores for window sizes of 5 and 21 amino acids in width (Fig. 9). All $\mu(k)$ scores for the myosin rod region are $\geq$10.

Discussion

We investigate whether secondary structure prediction methods may be used to predict potential sites of protein conformational rearrangement from primary sequence data. Using the output of the algorithm PHD (Rost & Sander, 1993), we formulated the program ASP to identify "structurally ambivalent sequence elements" that lack a strong intrinsic preference for any particular secondary structure type.

To evaluate the predictive ability of ASP, we have analyzed a catalog of test protein sequences for which function is known to require substantial backbone structural rearrangements. To be as objective as possible, we used a subcategory of proteins thought to contain conformational switches from a comprehensive database of macromolecular motions (Gerstein & Krebs, 1998). Our catalog of test proteins includes four inhibitory serpins, $G_{t\alpha}$, $G_{s\alpha}$, EF-Tu, NF-κB p50, TBP, GDR, prion protein, topoisomerase I, aldose reductase, hemagglutinin, and myosin. Three sequences known to lack conformational switches (ovalbumin, collagen, and the rod portion of myosin) were evaluated as negative controls.

Among the test proteins are cases for which structures have been solved in two distinct conformations, and also cases where only one conformation is known but structural rearrangements are suspected. The results of our analyses indicate that ASP is able to predict sites of conformational rearrangements, as categorized below:

(i) *Structurally ambivalent sequence elements were predicted for all test proteins that undergo conformational rearrangements.*

All test protein sequences were found by ASP to contain subsequences that were ambivalent with regard to secondary structure (Table 1; Figs. 1, 3, 6, 8). Conversely, control proteins known to have stable secondary structures were found to be devoid of such ambivalent sequence elements ($\mu_k > 9$ threshold) (Figs. 1, 9). In the serpin protein family, ambivalent sequence elements were found only for those inhibitory members that undergo a conformational rearrangement. In the homologous but noninhibitory ovalbumin sequence, ambivalent sequence elements were not detected. No SASEs were predicted for the collagen triple helix. For the myosin sequence, which contains a well-defined subunit subject to rearrangements (the motor domain), and a relatively rigid subunit

(the rod), ambivalent sequence elements were found only in the motor domain (Table 1).

(ii) *Regions of protein structure shown by experiment to be conformational switches are accurately predicted by ASP to be structurally ambivalent.*

For those proteins in which localized "conformational switches" have been characterized (inhibitory serpins, $G_{t\alpha}$, $G_{s\alpha}$, EF-Tu), sequence elements identified by ASP ($z < -1.75$) correlated well to the known sites of conformational change (Table 2). Ambivalent subsequences mapped into or overlapped all three of the conformational switches described in $G_{t\alpha}$ (Fig. 4), all three switches in $G_{s\alpha}$ (Fig. 5), both of the switches in EF-Tu, and the two putative switches in myosin. Similarly, for all four of the inhibitory serpins studied, ambivalent sequence elements mapped into the reactive center loop (Fig. 2). This loop region undergoes a transition to a strand of β-sheet.

(iii) *For those proteins with conformational transitions that are not as well characterized, ambivalent sequence elements map to regions suggested by experiment to undergo structural rearrangement.*

In the case of TBP, prion protein, topoisomerase I, and haemagglutinin, ambivalent sequence elements were identified that map to elements of the protein structure previously described as putative sites of structural heterogeneity. The two putative switches in myosin are identified (Table 2). For two test proteins that bind to specific DNA sequences (NF-κB, $\gamma\delta$-resolvase), ambivalent sequence elements mapped to regions of the protein thought to undergo structural transitions that facilitate binding and that mediate intermolecular interactions (Fig. 7).

(iv) *For allosteric proteins, ASP predicts ambivalent sequences that are spatially proximal, indicative of regions with coordinated motions.*

In the case of $G_{s\alpha}$ and $G_{t\alpha}$, ambivalent sequence elements were identified at sites in addition to the conformational switches. With one exception, the sequence elements were found to be in van der Waals contact with a known switch. These contacts have been previously suggested to function in the transmission of information from ligand binding sites to effector binding sites. When all of the ambivalent sequence elements identified in $G_{s\alpha}$ are mapped onto the protein surface, these residues constitute a region that markedly resembles the "interaction footprint" previously described by experiment (Sunahara et al., 1997). Similarly, residues are identified in EF-Tu, which would allow for interaction between the effector region (switch I) and the nucleotide binding region (switch II). In the case of myosin, when structurally ambivalent sequence elements are mapped to the crystal structure, they are found to provide for a pathway between the actin binding site, the nucleotide binding site, and the fulcrum site (Kirshenbaum et al., 1999). An ambivalent sequence element was also identified in NF-κB that contacts the corresponding residues at the interface of the homodimer structure in complex with DNA (Muller et al., 1995; Fig. 7A). This region has been suggested to be the site of allosteric communication between the NF-κB monomers.

The above results demonstrate that ASP can predict localized elements of protein sequence that have an enhanced proclivity for undergoing substantial backbone conformational rearrangements.

ASP relies upon identifying sequence elements compatible with more than one type of secondary structure. Thus, it was not expected a priori to identify regions of structural disorder in which a transition from one type of secondary structure class to another does not occur. For example, ASP would not be expected to identify mobile regions of flexible loops, in which other secondary structure types are not sampled. Similarly, ASP may not identify shear-type or hinge-bending motions, especially when these occur in regions smaller than the sequence window size used by the algorithm (5 and 21 residues). For this reason, sequences of proteins known to undergo shear or hinge-bending motions were removed from consideration in the present study.

The ASP algorithm has only three adjustable parameters: the $\mu(k)$ threshold value, the window size, and the z score cutoff. In the present study, the values for $\mu(k)$ and z were fixed for all protein sequences. Although this provides for internal consistency, setting the z score cutoff to be -1.75 standard deviations constrains the predicted ambivalent subsequences to $\sim$4% of the total sequence. This strategy is effective for identifying protein sequence elements that are most likely to undergo transitions, such as conformational switches. However, using this z score cutoff, ASP cannot predict all sites in proteins that undergo global conformational changes. In the case of haemagglutinin, for example, a limited number of residues are predicted to be ambivalent, whereas the observed pH-induced conformational changes constitute refolding most of the tertiary structure (Bullough et al., 1994). The sequences that ASP predicts for hemagglutinin correspond to localized structural elements that potentiate the subsequent more extensive conformational changes (Chen et al., 1998). In the case of other proteins that undergo only highly localized conformational changes, a z score cutoff of -1.75 may overpredict the portion of the sequence that is structurally ambivalent. If the extent of conformational change is known for a protein, adjustment of the ASP z score cutoff may be appropriate.

The findings presented here indicate that sites of conformational rearrangement in proteins are amenable to prediction from primary sequence data alone. ASP identifies subsequences that are inclined to move from one type of secondary structure toward another. The extent of these motions can vary. In some instances a complete transition has been observed. For example, subsequences of the inhibitory serpins are present in crystal structures either as loop or β-strand (Stein & Chothia, 1991; Wisstock et al., 1998). In other cases, there is no crystallographic evidence for complete transitions between secondary structure types. In these cases the movement may consist of backbone excursions away from canonical secondary structure (Garcia, 1992). The proximity of multiple predicted structurally ambivalent subsequences (Figs. 4, 5) suggests that these small concerted movements could serve as connections between allosterically communicating sites (Hilser et al., 1998).

The ASP algorithm may assist in studies of proteins known to undergo conformational switching in the absence of detailed structural information. Here we have used ASP to predict known conformational switches and also sites of conformational rearrangements in proteins for which limited experimental data are available. An accompanying paper demonstrates that ASP can identify putative pathways of allosteric communication between the nucleotide, actin binding, and fulcrum sites of myosin (Kirshenbaum et al.,

1999). The potential exists to use ASP in elucidating protein signaling pathways, enzyme function and regulation, and molecular mechanisms of protein-folding diseases. In addition, future studies could develop ASP as a tool to assist in rational drug design by identification of subregions in protein binding sites predicted to undergo conformational changes.

Methods

The input for our computational method is the result of a secondary structure prediction for a sequence of interest. We use the program PHD (Rost & Sander, 1993; Rost, 1996), which assigns to each residue one of three possible secondary structure types, helical (H), strand (E), or loop (L), or makes no prediction. For each residue, PHD also outputs a prediction probability for each secondary structure type; i.e., prH, prE, and prL for helix, strand, and loop, respectively. The prediction probabilities range from 0 (low probability) to 9 (high probability), with the constraint that the sum of the three probabilities must equal 9.

We use these prediction probabilities to identify structurally ambivalent sequence elements (SASEs). SASEs are subsequences lacking a strong preference for a particular secondary structure type. We evaluate the residue ambivalence score, $\Delta Pr(k)$, of a particular amino acid at position k in the sequence using Equation 1:

$$\Delta Pr(k) = |prH - prE| + |prH - prL| + |prE - prL|. \quad (1)$$

The values of $\Delta Pr(k)$ range from 0 (maximum ambivalence) to 18 (minimum ambivalence). A $\Delta Pr(k)$ of 0 is obtained when a residue has been assigned prediction probabilities of 3 by PHD for each secondary structure type. A score of 18 indicates that one secondary structure type is predicted with a probability of 9, and the other two secondary structure types are assigned probabilities of 0.

Once residue ambivalence scores are generated, we proceed to identify ambivalent subsequences. Isolated residues are not good candidates to function as switches by conformational transitions between different secondary structures. We hypothesize that subsequences comprised of multiple residues with low $\Delta Pr(k)$ scores are more likely to have switch like character. We therefore evaluate the residues as members of a subset defined by a particular window size. The window is centered about each amino acid in the sequence, and the mean ambivalence score $\mu(k)$ is calculated using Equation 2:

$$\mu(k) = \Sigma \Delta Pr(i)/S \quad (2)$$

where S is the window size, and where $k - [(S - 1)/2] \leq i \leq k + [(S - 1)/2]$. A tract of residue with sufficiently low mean ambivalence scores is identified as an ambivalent subsequence. We investigated window sizes of 5, 11, 15, 21, and 25 to ascertain how best to identify regions with persistent ambivalent character (data not shown). Smaller window sizes encapsulate more limited subsequence features because fewer neighboring residues are used in the $\mu(k)$ calculation. For our analyses, we find that a window size of 5 was suitable for proteins that are known to exhibit localized conformational changes, while a window size of 21 is appropriate for proteins that undergo more extensive structural rearrangements. However, the most ambivalent subsequences are identified using either window size.

We use a two-step process to identify SASEs. We first use a coarse analysis to discriminate which proteins might be expected to display backbone conformational rearrangements. For this analysis, we define proteins with $\mu(k)$ scores of 9 or less to contain conformationally-switching regions, whereas proteins with globally high $\mu(k)$ scores are defined to not have these regions. Only proteins that contain subsequences with mean residue scores of 9 or less are analyzed further.

For those proteins, a statistical criterion is applied to identify consistently the residues of a particular protein that are most likely to undergo changes in secondary structure. We use a simple z score cutoff for each windowed score, $\mu(k)$, relative to the mean for the entire sequence, $\mu(\text{seq})$, to identify the most ambivalent residues. The z score is calculated according to Equation 3, so that the maximal values corresponded to regions of minimum ambivalence:

$$z = [\mu(k) - \mu(seq)]/\sigma(seq) \quad (3)$$

where $\sigma(seq)$ is the standard deviation of $\mu(seq)$. Thus, we can calculate the z score for each residue and identify those regions that are significantly ambivalent. By using a z score cutoff of -1.75 standard deviations, approximately 4% of the sequence is defined to be ambivalent.

Acknowledgments

We thank ID Kuntz, David Baker and his research group, David Miller, and Keith Dunker for helpful discussions. K.K. thanks Chiron Corp. and Ron Zuckerman for their generosity. M.Y. and K.K. are supported by training grants from the National Institutes of Health, GM08388 and GM07175, respectively. S.H. is supported by AR42895. M.Y. and K.K. contributed equally to the work.

References

Abel K, Yoder MD, Hilgenfeld R, Jurnak F. 1996. An alpha to beta conformational switch in EF-Tu. *Structure 4*:1153–1159.

Bohrani DW, Harter, TM, Petrash JMJ. 1992. The crystal structure of the aldose reductase·NADPH binary complex. *J Biol Chem 267*:24841–24847.

Bullough PA, Hughson FM, Skehel JJ, Wiley DC. 1994. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature 371*:37–43.

Chasman DI, Flaherty KM, Sharp PA, Kornberg RD. 1993. Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proc Natl Acad Sci USA 90*:8174–8178.

Chen J, Lee KL, Steinhauer DA, Stevens DJ, Shehal JJ, Wiley DC. 1998. Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and origin of the labile conformation. *Cell 95*:409–417.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Chou PY, Fasman GD. 1975. The conformation of glucagon: Predictions and consequences. *Biochemistry 14*:2536–2541.

Cohen BI, Presnell SR, Cohen FE. 1993. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci 2*:2134–2145.

Dominguez R, Freyzon Y, Trybus KM, Cohen C. 1998. Crystal structure of a vertebrate smooth muscle myosin motor domain and its complex with the essential light chain: Visualization of the pre-power stroke state. *Cell 94*:559–571.

Elliott PR, Abrahams J-P, Lomas DA. 1998. Wild-type alpha1-antitrypsin is in the canonical inhibitory conformation. *J Mol Biol 275*:419–425.

Fasman GD. 1987. The road from poly(alpha-amino acids) to the prediction of protein conformation. *Biopolymers 26*:S59–S79.

Garcia AE. 1992. Large-amplitude nonlinear motions in proteins. *Phys Rev Lett 68*:2696–2699.

Gerstein M, Krebs W. 1998. A database of macromolecular motions. *Nucleic Acids Res 26*:4280–4290.

Ghosh B, Van Duyne G, Ghosh S, Sigler PB. 1995. Structure of NF-κB p50 homodimer bound to a κB site. *Nature 373*:303–310.

Han KF, Baker D. 1996. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA 93*:5814–5818.

Hilser VJ, Dowdy D, Oas TG, Freire E. 1998. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc Natl Acad Sci USA 95*:9903–9908.

Holm L, Sander C. 1998. Touring protein fold space with Dali/FSSP. *Nucl Acids Res 26*:316–319.

James TL, Liu H, Ulyanov NB, Farr-Jones S, Zhang H, Donne DG, Kaneko K, Groth D, Mehlhorn I, Prusiner SB, Cohen FE. 1997. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc Natl Acad Sci USA 94*:10086–10091.

Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA 81*:1075–1078

Kim JL, Nikolov DB, Burley SK. 1993. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature 365*:520–512.

Kim Y, Geiger JH, Hahn S, Sigler PB. 1993. Crystal structure of a yeast TBP/TATA-box complex. *Nature 365*:512–520.

Kirshenbaum K, Young M, Highsmith S. 1999. Predicting allosteric switches in myosins. *Protein Sci 8*:1806–1815.

Kull FJ, Sablin EP, Lau R, Fletterick RJ, Vale RD. 1996. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature 380*:550–555.

Lambright DG, Sondek J, Bohm A, Skiba NP, Hamm HE, Sigler PB. 1996. The 2.0 Å crystal structure of a heterotrimeric G protein. *Nature 379*:311–319.

Lima CD, Wang JC, Mondragon A. 1994. Three-dimensional structure of the 67K N-terminal fragment of *E. coli* DNA topoisomerase I. *Nature 367*:138–146.

Minor DL Jr, Kim PS. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature 380*:730–734.

Muller CW, Rey FA, Sodeoka M, Verdine GL, Harrison SC. 1995. Structure of the NF-κB p50 homodimer bound to DNA. *Nature 373*:311–317.

Noel JP, Hamm HE, Sigler PB. 1993. The 2.2 Å crystal structure of transducin-alpha complexed with GTPgammaS. *Nature 366*:654–663.

Pan B, Deng Z, Liu D, Ghosh S, Mullen GP. 1997. Secondary and tertiary structural changes in $\gamma\delta$ resolvase: Comparison of the wild-type enzyme, the I110R mutant, and the C-terminal DNA binding domain in solution. *Protein Sci 6*:1237–1247.

Polekhina G, Thirup S, Kjeldgaard M, Nissen P, Lippmann C, Nyborg J. 1996. Helix unwinding in the effector region of elongation factor EF-Tu-GDP. *Structure 4*:1141–1151.

Potempa J, Korzus E, Travis J. 1994. The serpin superfamily of proteinase inhibitors: Structure, function, and regulation. *J Biol Chem 269*:15957–15960.

Rayment I, Rypniewski WR, Schmidt-Base K, Smith R, Tomchick DR, Benning MM, Winkelmann DA, Wesenberg G, Holden HM. 1993. Three-dimensional structure of myosin subfragment-1: A molecular motor. *Science 261*:50–58.

Riek R, Wider G, Billeter M, Hornemann S, Glockshuber R, Wuthrich K. 1998. Prion protein NMR structure and familial spongiform encephalopathies. *Proc Natl Acad Sci USA 95*:11667–11672.

Rosenblatt M, Beaudette NV, Fasman GD. 1980. Conformational studies of the synthetic precursor-specific region of preproparathyroid hormone. *Proc Natl Acad Sci USA 77*:3983–3987.

Rost B. 1996. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol 266*:535–539.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Smith CA, Rayment I. 1996a. X-ray structure of the magnesium(II)·ADP·vanadate complex of the *Dictyostelium discoideum* myosin motor domain to 1.9 Å resolution. *Biochemistry 35*:5404–5417.

Smith CA, Rayment I. 1996b. Active site comparisons highlight structural similarities between myosin and other P-loop proteins. *Biophys J 70*:1590–1602.

Sobolev V, Wade RC, Vriend G, Edelman M. 1996. Molecular docking using surface complementarity. *Proteins 25*:120–129.

Sondek J, Lambright DG, Noel JP, Hamm HE, Sigler PB. 1994. GTPase mechanism of G-proteins from the 1.7 Å crystal structure of transducin alpha-GDP-AlF4. *Nature 372*:276–279.

Stein P, Chothia C. 1991. Serpin tertiary structure transformation. *J Mol Biol 221*:615–621.

Sudarsanam S. 1998. Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations. *Proteins 30*:228–231.

Sunahara RK, Tesmer JJG, Gilman AG, Sprang SR. 1997. Crystal structure of the adenylyl cyclase activator Gsalpha. *Science 278*:1943–1947.

Tesmer JJG, Sunahara RK, Gilman AG, Sprang SR. 1997. Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gs$\alpha\cdot$GTPγS. *Science 278*:1907–1916.

Urzhumtsev A, Tete-Favier F, Mitschler A, Barbanton J, Barth P, Urzhumtseva L, Biellmann JF, Podjarny A, Moras D. 1997. A "specificity" pocket inferred from the crystal structures of the complexes of aldose reductase with the pharmaceutically important inhibitors tolrestat and sorbinil. *Structure 5*:601–612.

Wall MA, Posner BA, Sprang SR. 1998. Structural basis of activity and subunit recognition in G protein heterotrimers. *Structure 6*:1169–1183.

Waterhous DV, Johnson WC Jr. 1994. Importance of environment in determining secondary structure in proteins. *Biochemistry 33*:2121–2128.

Wilson DK, Bohren KM, Gabbay KH, Quiocho FA. 1992. An unlikely sugar substrate site in the 1.65 Å structure of the human aldose reductase holoenzyme implicated in diabetic complications. *Science 257*:81–84.

Wisstock J, Skinner R, Lesk AM. 1998. An atlas of serpin conformations. *Trends Biochem Sci 23*:63–67.

Yang W, Steitz TA. 1995. Crystal structure of the site-specific recombinase $\gamma\delta$ resolvase complexed with a 34 bp cleavage site. *Cell 82*:193–207.

Yu L, Zhu CX, Tse-Dinh YC, Fesik SW. 1995. Solution structure of the C-terminal single-stranded DNA-binding domain of *Escherichia coli* topoisomerase I. *Biochemistry 34*:7622–7628.

Yu L, Zhu CX, Tse-Dinh YC, Fesik SW. 1996. Backbone dynamics of the C-terminal domain of *Escherichia coli* topoisomerase I in the absence and presence of single-stranded DNA. *Biochemistry 35*:9661–9666.

Zhong L, Johnson WC Jr. 1992. Environment affects amino acid preference for secondary structure. *Proc Natl Acad Sci USA 89*:4462–4465.