BIOINFORMATICS

Vol. 18 no. 7 2002
Pages 980–984

CX, an algorithm that identifies protruding atoms in proteins

Alessandro Pintar[1,*], *Oliviero Carugo*[1, 2] *and Sándor Pongor*[1,*]

[1]*Protein Structure and Bioinformatics Group, International Center for Genetic Engineering and Biotechnology (ICGEB), AREA Science Park, Padriciano 99, 34012 Trieste, Italy and* [2]*Department of General Chemistry, University of Pavia, Viale Taramelli 12, 27100 Pavia, Italy*

Received on December 12, 2001; revised on February 6, 2002; accepted on February 14, 2002

ABSTRACT

Motivation: A simple and fast algorithm is described that calculates a measure of protrusion (cx) for atoms in protein structures, directly useable with the common molecular graphics programs.

Results: A sphere of predetermined radius is centered around each non-hydrogen atom, and the volume occupied by the protein and the free volume within the sphere (internal and external volumes, respectively) are calculated. Atoms in protruding regions have a high ratio (cx) between the external and the internal volume. The program reads a PDB file, and writes the output in the same format, with cx values in the B factor field. Output structure files can be directly displayed with standard molecular graphics programs like RASMOL, MOLMOL, Swiss-PDB Viewer and colored according to cx values. We show the potential use of this program in the analysis of two protein–protein complexes and in the prediction of limited proteolysis sites in native proteins.

Availability: The algorithm is implemented in a standalone program written in C and its source is freely available at ftp.icgeb.trieste.it/pub/CX or on request from the authors.

Contact: pintar@icgeb.trieste.it; carugo@icgeb.trieste.it; pongor@icgeb.trieste.it

INTRODUCTION

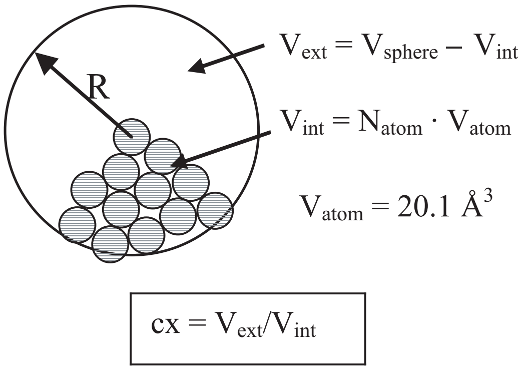
The analysis of protein–protein interfaces is a difficult task that has been tackled with a variety of computational approaches. In most cases, the analysis of protein surfaces has been aimed at finding cavities and clefts. This is important in the identification of binding sites for small molecules like cofactors, drugs, and peptides, but it represents only one face of the problem when protein–protein interactions are considered. The identification of protruding, or highly convex regions in proteins is important, on the other hand, not only in the study of protein–protein complexes, but also in the prediction of limited proteolysis cleavage sites and antigenic determinants.

Different approaches have been used to identify protruding regions in proteins. In Taylor's method (Taylor *et al.*, 1983) the overall shape of a protein is represented as an ellipsoid, and a residue protrusion index is calculated from a series of different ellipsoids each encompassing a different percentage of the Cα carbons. Nishikawa and Ooi (1986) used the number of Cα atoms within a certain distance from each Cα of the protein to characterize the exposure of a residue to the solvent. Connolly (1986) developed a method to measure the convexity or concavity of protein surface regions. In this method, a sphere is centered at any point of the protein surface and a numerical index (the solid angle Ω) that depends on the fraction of the sphere lying inside the protein is assigned to that point.

Nevertheless, none of the above methods has become a standard tool of the molecular modeling repertoire, probably because on one side, the residue-based indices are rather coarse descriptors of the real geometry of the protein surface and, on the other, methods based on molecular or solvent accessible surfaces are rather compute intensive and significantly depend on the parameters used in the computations (e.g. radius probe and atomic radii).

Here we present a very simple and fast algorithm that calculates a numeric measure of protrusion or convexity for each protein atom, the cx index, that can be directly visualized with the commonly used molecular graphics programs like RASMOL (Sayle and Milner-White, 1995), Swiss-PDB Viewer (Guex and Peitsch, 1997), and MOLMOL (Koradi *et al.*, 1996). We show that the cx index is a sensitive visual indicator of protruding atoms within protein/protein (or protein/DNA) interfaces, and that its use can be extended to the prediction of proteolysis sites in proteins.

*To whom correspondence should be addressed.



R
$V_{ext} = V_{sphere} - V_{int}$
$V_{int} = N_{atom} \cdot V_{atom}$
$V_{atom} = 20.1$ Å^3
$cx = V_{ext}/V_{int}$

Fig. 1. Schematic representation of the CX algorithm. N_{atom} is the number of non-hydrogen atoms found within a distance R around a non-hydrogen protein atom. The default radius of the spherical probe is 10 Å. V_{atom} is the average volume of a heavy atom in a protein, its value is 20.1 Å^3 (Richards, 1974). Given this approximation, the volume occupied by the protein from the sphere, V_{int} can be calculated, and compared with V_{ext}, the portion of the sphere left free by the protein. For protein atoms, the ratio cx $= V_{ext}/V_{int}$ is a number between 0 and ∼15, protruding atoms having higher cx values.

ALGORITHM

The principle of the algorithm is illustrated in Figure 1. For each heavy (non-hydrogen) atom in a protein structure, the program calculates the number of heavy atoms within a fixed distance R (the default value is 10 Å). The number of atoms within the sphere is multiplied by the mean atomic volume found in proteins (20.1 ± 0.9 Å^3; Richards, 1974), which gives the volume occupied by the protein within the sphere, V_{int}. The remaining volume of the sphere, V_{ext}, is calculated as the difference between the volume of the sphere and V_{int}. The cx value is then defined by V_{ext}/V_{int}.

IMPLEMENTATION

A simple, standalone C-program was written that reads standard PDB coordinate files as the input. The program reads only ATOM lines. Thus, HETATM lines describing non-standard residues, cofactors, metal ions, and water molecules are not taken into account. By default, the program treats each chain in the PDB file as an independent molecule (i.e. the atoms of chain B are not taken into account when calculating the protrusion index for the atoms of chain A) but the results are written into a single file. The output is a coordinate file in PDB format in which the atomic displacement parameter (B-factor) is replaced by the cx value. The output files can be thus displayed using most molecular graphics programs, and atoms colored by their cx value in a straightforward manner. On an SGI R10000 (195 MHz) processor, the program requires ∼1.5 s cpu time for a 1000 atom protein. The program is deposited in ftp.icgeb.trieste.it/pub/CX and is also available from the authors upon request (pintar@icgeb.trieste.it, carugo@icgeb.trieste.it, pongor@icgeb.trieste.it).

RESULTS

As shown in Figure 1, the value of cx will be large for those atoms that have few neighbors in their vicinity. As this occurs in protruding parts of proteins (or, in other words, convex parts of the protein surface), cx can be considered as an approximate measure of protrusion or convexity. To determine the empirical maximum of the cx index in proteins, we selected with PDBSELECT (Hobohm and Sander, 1994) a set of 475 non-homologous (sequence identity lower than 25%) protein crystal structures determined at better than 2.0 Å resolution, and containing more than 400 atoms. We found a maximum of 13.89 for the NZ atom of a lysine side chain in 1cru. The minimum of the cx value is expected to be zero. However, slightly negative values can also be observed occasionally for buried atoms. The minimal value found in the same protein data set is −0.20. For practical purposes, negative cx values are reset to zero, so in the case of protein structures, cx can be roughly considered as a numerical index varying between 0 and 15.

As a numerical evaluation of CX, we calculated the cx value of surface Cα carbons (atomic solvent accessible surface >2.0 Å^2) for the same set of non-homologous proteins, and compared them with the: (i) Ooi number; (ii) the Cα-factor; (iii) the residue solvent accessibility (Å^2); and (iv) the relative residue solvent accessibility (%). We found correlation coefficients of: (i) −0.67; (ii) 0.32; (iii) 0.49; and (iv) 0.60, respectively.

Figure 2a and b show the complex formed by the amino-terminal domain of the HIV-1 capsid protein p24 and the antibody fragment Fab25.3 (PDB: 1afv Momany *et al.*, 1996) and a detail of the interface, respectively, colored by the cx protrusion index. The interface is made by two loops of Fab (residues 29–35 and 101–105) ‘grabbing’ the C-terminal part of one of the p24 α-helixes (residues 65-85). The largest cx values are in Y32 and S103 of Fab, and R82 on p24, so the key residues of the interaction are correctly highlighted (Figure 2a, b). Fab Y32 lies in a cavity flanked by S102 and R100, whose base is formed by A77 and G101, while the α-helix from p24 is using the side chain of R82 to ‘land’ in a hydrophobic pocket of Fab.

Figure 2c and d show the complex between a serine protease inhibitor (serpin) and trypsin (PDB code:1i99; Ye *et al.*, 2001) and a detail of the interface, respectively. In this complex, a 20 residue long loop is protruding from the serpin molecule, and interacts with a rather large concave region in trypsin. CX identifies this loop as the most protruding region in serpin, and the side chains of I350 and K353 as the ones having the highest cx

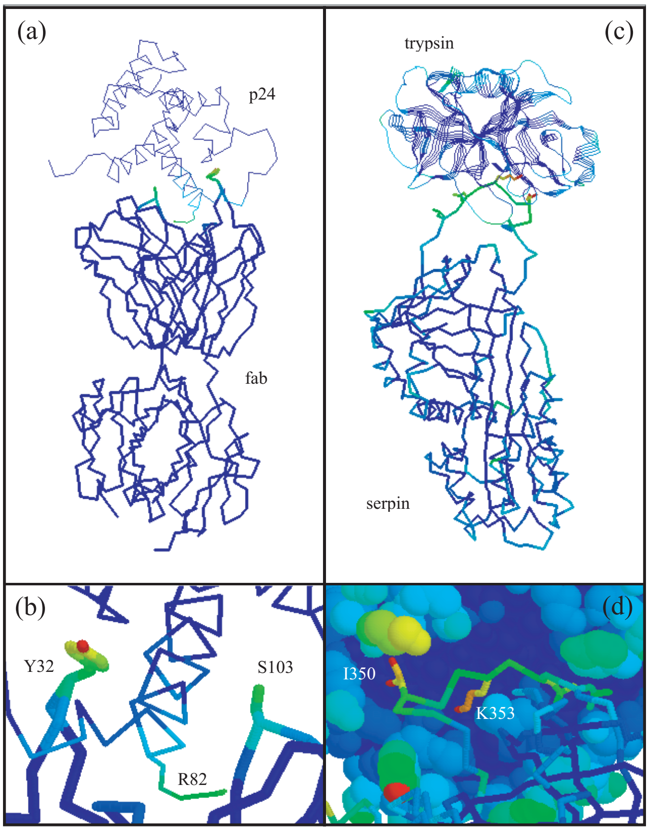


Fig. 2. Analysis of protein–protein complexes. Left: (a) the complex between the amino-terminal domain of the HIV-1 capsid protein p24 (thin line) and the antibody fragment Fab25.3 (thick line) (PDB: 1afv); to highlight the protein–protein interface only, bonds are colored according to the difference of the cx values measured in the complex and in the isolated proteins; (b) a detail of the interface. Right: (c) the complex between serpin (shown as ribbon) and trypsin (PDB: 1i99); (d) a detail of the interface (trypsin is shown as a CPK model). The color scale is from blue (low cx) to red (high cx). Molecules were displayed using RASMOL (Sayle and Milner-White, 1995).

values (Figure 2c, d). Serpin I350 is flanked by a number of hydrophobic side chains of trypsin residues: W215, K175, L99, and K97. Serpin K353 is deeply buried in a hydrophobic pocket, and contacts the catalytic residue of trypsin, D189. The side chains of trypsin L355, L357 are also involved in filling the cavity.

As a further example of potential applications, we analyzed the trypsin cleavage sites in a set of 8 native proteins of known structure, for which experimental limited proteolysis data are available (Hubbard *et al.*, 1998). For this set of proteins, predictions of the potential cleavage sites have also been carried out using NICKPRED. This predictive algorithm takes into account several weighted conformational parameters: solvent accessibility, Taylor's protrusion index, residue-averaged temperature factors, Ooi numbers, secondary structure elements, and main chain hydrogen bonding.

Table 1. Trypsin cleavage sites in native proteins of known structure

PDB	cleavage (exp.)	Rank (CX)	Rank (NICK.)
1sno	K5 K48 K49	(a) 2/27 4/27	(a) 1/25 2/25
3est	R125	3/15	1/14
1tgn	K145	1/16	1/15
1thv	R119 K163	1/23 2/23	9/21 1/21
2cst	K19 R25	18/47[b] 2/47[c]	9,23/94 2,7/94
1mup	R12	1/17	1/17
1hcy	K174 K175	3/63 1/63	1/59 2/59
2sh1	R13	1/7	(d)

The experimentally determined proteolysis sites were compared with the ranking calculated by CX and NICKPRED. Of all the possible trypsin cleavage sites, the first in the ranking is the most probable one. The number of potential sites can be different in CX and NICKPRED because NICKPRED takes into account sequence requirements that are not considered by CX; for example, sites containing R/K followed by P are not cleaved by trypsin. Proteins are: 1sno, staphylococcal nuclease from *S. aureus*; 3est, porcine elastase; 1tgn, bovine trypsinogen; 1thv, thaumatin from *T. daniellii*; 2cst, aspartate aminotransferase from pig heart (homodimer); 1mup, major urinary protein from mouse; 1hcy, hemocyanin subunit A from spiny lobster; 2sh1, SH-I neurotoxin from anemone (NMR structure).
[a] not in PDB ATOM list
[b] 29,30/94 as homodimer
[c] 1,2/94 as homodimer
[d] not applicable.

The first step in the cleavage of the peptide bond by trypsin is the nucleophilic attack of the oxygen of the catalytic serine to the carbon of the amide bond following a lysine or arginine residue. We thus selected and ranked the cx values for the carbonyl C atom of lysine/arginine residues in the chosen proteins (Table 1). We find that most of the experimental cleavage sites correspond to residues having a high cx value at the C atom, and are in the top ranking positions calculated by CX. Despite the simplicity of the CX method, the results of CX compare well with the predictions made by the more sophisticated NICKPRED (Hubbard *et al.*, 1998, Table 1).

DISCUSSION

The approach used by CX is conceptually similar to the solid angle method described by Connolly (1986), but takes advantage of the simplicity of the approach used by

Nishikawa and Ooi (1986). As a result, it is more accurate than the Ooi number because it provides information on all heavy atoms of a protein, including side chains; at the same time, the calculation of cx values is computationally much less demanding than the calculation of surface curvature. Furthermore, cx values are atomic, and not surface properties, so they can be handled and analyzed in a much simpler way. As the output is a standard PDB file, structures can be displayed, colored, and analyzed in a straightforward manner using the most popular molecular graphics programs, like RASMOL (Sayle and Milner-White, 1995), Swiss-PDB Viewer (Guex and Peitsch, 1997) and MOLMOL (Koradi *et al.*, 1996). The analysis of protein structures in terms of cx values does not require solid rendering of a surface on the graphics terminal, and is thus accessible even to low-end personal computers. As a demonstration, Figure 2 was prepared using the 8-bit version of RASMOL (Sayle and Milner-White, 1995).

A limitation of this algorithm is that, while it can identify convex regions, it cannot distinguish an atom that is in a concave region from one that is just buried. For the identification of cavities and clefts in proteins, additional approaches are necessary. For example, atoms in surface cavities may be identified by the fact that they have low cx values but at the same time are accessible to the solvent.

Only two independent parameters are used by CX: the average atomic volume, and the sphere radius. Both can be modified in the program. The default value for the average atomic volume used here (20.1 Å^3) is a good approximation for the buried atoms constituting the core of a protein (Richards, 1974). However, it should be kept in mind that different atom types have different average atomic volumes. As a consequence, the average atomic volume can vary from 15.9 Å^3 for histidine to 22.3 Å^3 for methionine, depending on the residue type (Pontius *et al.*, 1996). Not taking into account the relative amino acid abundance, the overall average value would be close to 18 Å^3. This is ~10% lower than the default value used by CX. On the other hand, standard atomic volumes have been derived for buried atoms only, and in high resolution protein structures. Simulations showed that atoms exposed at the surface, with the exception of charged atoms, are ~6% larger than buried atoms (Gerstein *et al.*, 1995). Moreover, experimental conditions and crystallographic resolution also affect atomic volumes. Indeed, atomic volumes can be used as a quality measure for protein crystal structures (Pontius *et al.*, 1996). We can conclude that, given the approximate nature of our method and its purposes, slight variations in the average atomic volume do not affect the results in a remarkable way.

The second parameter, the sphere radius, is chosen in a rather empirical way and can be tuned according to the needs: smaller values will make CX more sensitive to the local environment, whereas larger values will make it more sensitive to the global fold of the protein (Figure 4). In the solid angle method proposed by Connolly, a default radius of 6 Å was used (Connolly, 1986). Nishikawa and Ooi found that a sphere of 8 Å includes only the residues that are in contact with a given residue, while a sphere of 14 Å is better suited to describe the global structural features of a protein (Nishikawa and Ooi, 1986). We found that the default radius used by CX (10 Å) is a good compromise to highlight both backbone and side chain protruding atoms in most applications. In some instances, however, like in a complex between a large protein and a short peptide, it might be desirable to run CX with different *R* values, according to the different size of the molecules to be studied.

It has been shown that the PI (protrusion index), the solvent accessibility, the B-factors, and the Ooi number are all correlated to some extent (Thornton *et al.*, 1986). In a similar way, we find some correlation between the cx value on one hand, and the Ooi number, B-factors, and the residue solvent accessibility, respectively, on the

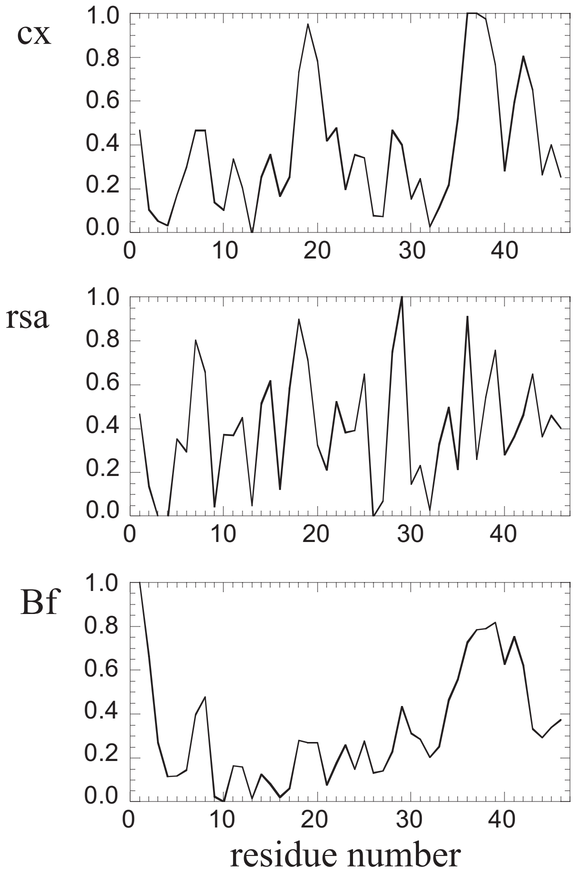


Fig. 3. cx values of Cα atoms, residue solvent accessibility (rsa), and B-factors (Bf) of Cα atoms calculated for crambin (PDB: 1crn). Values were normalized between 0 and 1 for better comparison.

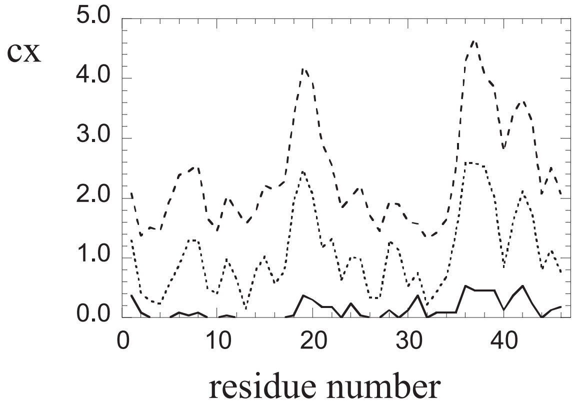


Fig. 4. cx values for Cα atoms of crambin (PDB: 1crn) calculated using a sphere radius of 5 Å (---), 10 Å (····) and 15 Å (—).

other. This is not surprising. However, the Ooi number provides very limited information, as it is restricted to Cα atoms. The atomic solvent accessibility, as calculated for example by NACCESS (Hubbard and Thornton, 1993), gives a strictly local information, which is limited by the small radius (normally 1.4 Å) of the rolling sphere used for the calculation, whereas CX takes into account long range interactions. The relative residue solvent accessibility can give results that are sometimes close to those obtained by CX, but in this case the atomic information is lost. B-factors are also expected to be higher in highly protruding regions, but this is not always the case. The different profiles obtained for the cx values of Cα atoms, the relative residue solvent accessible surface, as calculated by NACCESS, and the Cα B-factors for a small protein (PDB: 1crn, crambin) are shown in Figure 3. It should also be stressed that the physical principles underlying the calculation of cx, accessibility, or B-factors are totally different and the information given by these parameters can be thus considered as complementary.

Other potential uses of this program might be the prediction of antigenic epitopes (Barlow *et al.*, 1986) for the production of antibodies and vaccines, a rough estimation of protein packing in 3D structures and the generation of 2D protein profiles (Nishikawa and Ooi, 1986) that might be used in fold recognition.

CONCLUSION

We developed a simple program that calculates a protrusion index (cx) for heavy atoms based on the volume occupied by the protein and the free volume around each heavy atom in the protein. Cx values can be read by most molecular graphics programs in a straightforward manner, and structures colored accordingly. This can greatly facilitate the visual analysis of protein–protein complexes and of protruding regions in proteins. As protruding regions are good candidates for cleavage sites in limited proteolysis, CX can also be used as a predictive tool.

REFERENCES

Barlow,D.J., Edwards,M.S. and Thornton,J.M. (1986) Continuous and discontinuous protein antigenic determinants. *Nature*, **322**, 747–748.

Connolly,M.L. (1986) Measurement of protein surface shape by solid angles. *J. Mol. Graph.*, **4**, 3–6.

Gerstein,M., Tsai,J. and Levitt,M. (1995) The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.*, **249**, 955–966.

Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Hubbard,S.J., Beynon,R.J. and Thornton,J.M. (1998) Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.*, **11**, 349–359.

Hubbard,S.J. and Thornton,J.M. (1993) NACCESS, Department of Biochemistry and Molecular Biology, University College.

Koradi,R., Billeter,M. and Wüthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph*, **14**, 29–32, 51–55.

Momany,C., Kovari,L.C., Prongay,A.J., Keller,W., Gitti,R.K., Lee,B.M., Gorbalenya,A.E., Tong,L., McClure,J., Ehrlich,L.S. *et al.* (1996) Crystal structure of dimeric HIV-1 capsid protein. *Nat. Struct. Biol.*, **3**, 763–770.

Nishikawa,K. and Ooi,T. (1986) Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem. (Tokyo)*, **100**, 1043–1047.

Pontius,J., Richelle,J. and Wodak,S.J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, **264**, 121–136.

Richards,F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, **82**, 1–14.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Taylor,W.R., Thornton,J.M. and Turnell,W.G. (1983) An ellipsoidal approximation of protein shape. *J. Mol. Graph.*, **1**, 30–38.

Thornton,J.M., Edwards,M.S., Taylor,W.R. and Barlow,D.J. (1986) Location of ’continuous’ antigenic determinants in the protruding regions of proteins. *Embo J.*, **5**, 409–413.

Ye,S., Cech,A.L., Belmares,R., Bergstrom,R.C., Tong,Y., Corey,D.R., Kanost,M.R. and Goldsmith,E.J. (2001) The structure of a Michaelis serpin–protease complex. *Nat. Struct. Biol.*, **8**, 979–983.