Protein Science (1995), *4*:103–112. Cambridge University Press. Printed in the USA.

A procedure for detecting structural domains in proteins

MARK B. SWINDELLS

Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

(RECEIVED May 18, 1994; ACCEPTED November 9, 1994)

Abstract

A procedure is described for detecting domains in proteins of known structure. The method is based on the intuitively simple idea that each domain should contain an identifiable hydrophobic core. By applying the algorithm described in the companion paper (Swindells MB, 1995, *Protein Sci 4*:93–102) to identify distinct cores in multidomain proteins, one can use this information to determine both the number and the location of the constituent domains. Tests have shown the procedure to be effective on a number of examples, even when the domains are discontinuous along the sequence. However, deficiencies also occur when hydrophobic cores from different domains continue through the interface region and join one another.

Keywords: domains; hydrophobic cores; protein structure; structural analysis

When we take time to appreciate the new structures that are now being published weekly, it is immediately apparent that proteins are constructed on a modular basis and that many of these modules or domains frequently have familiar topologies. However, although domains can sometimes be identified quite easily by eye, their detection by automated procedures is much more difficult. This can be attributed, at least in part, to the fact that there is no standard definition of what a domain really is, and, as a result, previously developed methods have varied enormously, with each researcher using a unique set of criteria.

One of the first algorithms developed for domain detection (Crippen, 1978) used a C^{α}–C^{α} distance map (Nishikawa et al., 1972) together with a clustering routine. By identifying regions of the map in which the C^{α} distances remained small, it was anticipated that the constituents of each domain might be estimated. However, due to practical difficulties in determining when a cluster was of sufficient size to constitute a domain, the results obtained were somewhat disappointing. In a separate experiment, Rose adopted the opposite approach and searched for domain boundaries rather than the domains themselves (Rose, 1985). In practice, this was achieved by identifying regions where the C^{α} packing density was at a minimum, so that the protein could be "sliced" into two continuous chain segments. Appropriate sites were determined using a projection of the C^{α} coordinates, and thus by repeating this process on the segments generated, a hierarchical tree of cutting points was generated. However, in addition to the difficulty of knowing which points on the tree corresponded to the intuitive domain boundaries, many of the necessary cutting points remained undetected.

A more sophisticated method for detecting boundaries in continuous domains was based on the calculated interface area between two chain segments cleaved at a residue i (Wodak & Janin, 1981). Interface areas were calculated by comparing surface areas of the cleaved segments with that of the native structure. By repeating the comparison for all values of i, a potential domain boundary could then be identified as a site where the interface area was at a minimum. In order to prevent an excessive number of minima being recorded, minima were only considered when they were at least 2σ lower than the largest interface area observed in other regions of the same protein. The value of σ describes the average variation in interface area (i.e., the noise) when progressing sequentially from one residue to another (i to $i+1$). In order to extend the principle to discontinuous domains, Wodak and Janin compared the interface areas of variable length segments, rather than single-residue positions, which were then plotted in two dimensions with the extra dimension describing all possible segment lengths. From the resulting contour plot, it was found that regions with a low interface area again corresponded with the generally accepted domain boundaries. However, this method also had limitations, because other minima were also identified in regions where no boundary was anticipated. In addition, the computational requirements for identifying discontinuous domains were rather high. Consequently, its suitability for use as an automated classification procedure has been limited.

Since this time, the assignment of domains from coordinate data has received little attention (Zehfus, 1994), although there

Reprint requests to Mark B. Swindells at his present address: Molecular Design Department, Yamanouchi Pharmaceutical Co. Ltd., 21 Miyukigaoka, Tsukuba, Ibaraki 305, Japan; e-mail: mark@yamanouchi.co.jp.

have been papers that try to predict domains from sequence data alone (Busetta & Barrans, 1984; Kikuchi et al., 1988). However, with the rapid increase in the number of structures now available, and the important observation that familiar topologies frequently recur in proteins with no detectable sequence similarity, the requirement for such automated procedures has never been greater. In particular, it would assist the classification of protein folds because only similarities between monomeric proteins are identified satisfactorily by automated structural alignment techniques (Orengo et al., 1993).

Although most of us have an intuitive idea of what a domain is, standard definitions are hard to find. Perhaps our expectations are best described by Richardson, who described domains as regions that could be expected to be stable as independent structures (Richardson, 1981). Using this description as our starting point, we can extend this theme and infer that, with the exception of extremely small proteins that are held together by numerous disulfide bridges, each domain will have a hydrophobic core. Therefore, the problem of identifying structural domains might in fact be better approached by searching for distinct hydrophobic cores in a multidomain protein. In the companion paper (Swindells, 1995), I have described a conceptually simple and computationally efficient algorithm for identifying hydrophobic cores in monomeric proteins. I now describe its application to the detection of domains using a variety of multidomain structures as examples.

Results

Details of core and domain assignments, from which the following ribbon diagrams are constructed, are given in Appendix 1.

All β protein: Elastase

Elastase is a member of the chymotrypsin superfamily of serine proteinases. These structures consist of two antiparallel β-barrel topologies and have their active site residues located in a cleft formed by the domain interface. As seen in Figure 1, the two domains of elastase are clearly distinguished by the algorithm.

α/β Proteins: Rhodanese, arabinose binding protein, alcohol dehydrogenase

Rhodanese, arabinose binding protein, and alcohol dehydrogenase are all mixed α/β proteins. Rhodanese contains a pair of α/β-type domains that are highly conserved, despite the absence of any sequence homology. In contrast, the domains observed in arabinose binding protein are structurally unique, despite belonging to the α/β class. Alcohol dehydrogenase also contains two domains, one of which is a nicotinamide adenine dinucleotide (NAD) binding Rossmann fold. In all three cases, the domains are extrapolated from the hydrophobic cores (Fig. 1).

All α domains: Papain and thermolysin

Both papain and thermolysin structures contain an α-helical domain. In papain, the first α-helical domain is followed by an antiparallel β-barrel, whereas in thermolysin, an open-faced β-sheet precedes the small α Greek key. In Figure 2, differentiation of the α-helical structures from the adjoining domain can be clearly identified.

Aspartate aminotransferase

Aspartate aminotransferase (AAT) catalyzes the reversible transfer of an amino group between aspartate and α-ketoglutarate to give oxaloacetate and glutamate. Transamination also requires the coenzyme, pyridoxal 5′-phosphate (PLP), which subsequently forms a Schiff base, with a lysine ε-amino group located at the active site. Domain closure is crucial for the operation of this enzyme and recently a comparison of both open and closed forms has enabled these movements to be quantified (McPhalen et al., 1992b). Although AAT is usually described as containing two domains, the automated procedure consistently identified three distinct regions, irrespective of whether coordinates for the open or closed form were used. Though small, this third domain is distinct from the other two domains and seems to confer stability on the movements associated with substrate binding and domain closure. Thus, when a single AAT subunit is observed with these domain assignments the structure is somewhat reminiscent of a hinge (Fig. 3).

Aconitase

Aconitase is a large multidomain enzyme that catalyzes the conversion of citrate to isocitrate. The structure, which contains more than 750 residues, has been described by Robbins and Stout (1989) as a four-domain protein. This is closely mirrored by the automated assignments, except that the N-terminal 200 residues, which the authors consider a single domain, is split into two discontinuous domains by the algorithm (see Fig. 3 and Appendix 1 for details). This is because only one core residue is assigned to each of the β-strands consisting of residues 68–72 and 95–98, and this is insufficient to tie the two domains together. Despite this variation, the automated assignments still yield reasonably sized domains at the N-terminus, with even the smaller domain consisting of more than 60 residues and retaining a compact globular fold.

Assignment difficulties with pepsin and lactate dehydrogenase

From the preceding examples, it is clear that the program can generally identify domains in a wide range of proteins. However, the program inevitably has limitations, and the following examples are presented in order to emphasize problems that should be anticipated.

In aspartic proteinases such as pepsin, only a single domain is assigned, despite the well-known two-fold symmetry that corresponds to the dimeric interface of the distantly homologous human immunodeficiency virus (HIV) proteinase (Lapatto et al., 1989). The reason for this failure is that a central β-sheet and hydrophobic core passes straight through the anticipated boundary. Figure 4 shows details of the interface region, from which it is clear that Leu-6 enables the two potentially separate cores to merge. Although the detection of a single hydrophobic core leads to this domain assignment, a single core passing through the interface is in line with earlier analyses of the HIV proteinase, which also noted this characteristic (Lapatto et al., 1989).

A second case where automated assignments yield unanticipated results is in lactate dehydrogenase, where core residues also span the domain interface. As a result, the Rossmann fold domain is merged with the second domain (see Appendix 1). As

Fig. 1. Ribbon diagrams showing the domain assignments for elastase (top left), rhodanese (top right), arabinose binding protein (bottom left), and alcohol dehydrogenase (bottom right). Domains are colored yellow and blue. N-termini are highlighted with a black ball and C-termini with a white ball.

in the previous case, domain assignments fail, because the algorithm is unable to delineate the protein's constituent cores.

Discussion

With the rapid increase in newly determined structures and the recurrence of familiar topologies, new methods for analysis and prediction are required. In this and the companion paper (Swindells, 1995), I have described an algorithm that is capable of producing intuitive assignments for both hydrophobic cores and domains in proteins of known structure. It is hoped that this program will be useful for a wide range of applications, particularly in the areas of fold classification and recognition (Orengo et al., 1993). Detecting structural similarities at the domain level is important because unanticipated functional similarities generally occur at this hierarchical level. However, although rigorous automated structural alignment programs are available for comparing monomeric structures (Taylor & Orengo, 1989; Holm & Sander, 1993), these methods frequently encounter problems with multidomain proteins (Orengo et al., 1993). Thus, by first separating each protein into its constituent domains, it should then be possible to complete a more comprehensive classification

Fig. 2. Ribbon diagrams showing the domain assignments for papain (left) and thermolysin (right).

of each domain using the structural alignment programs referred to above. The suitability of this algorithm to a large-scale classification of multidomain proteins is currently under investigation.

While this paper was under review, Holm and Sander (1994) published another domain recognition algorithm. Their method, which uses a simple harmonic model to approximate oscillations in a multidomain system, is based on the concept that domains will have the highest concentration of atomic interactions, whereas domain boundaries will have the lowest. In order to prevent the algorithm from producing too many subdivisions, a number of additional filters are included. These include lower limits for the number of residues constituting a domain, con-

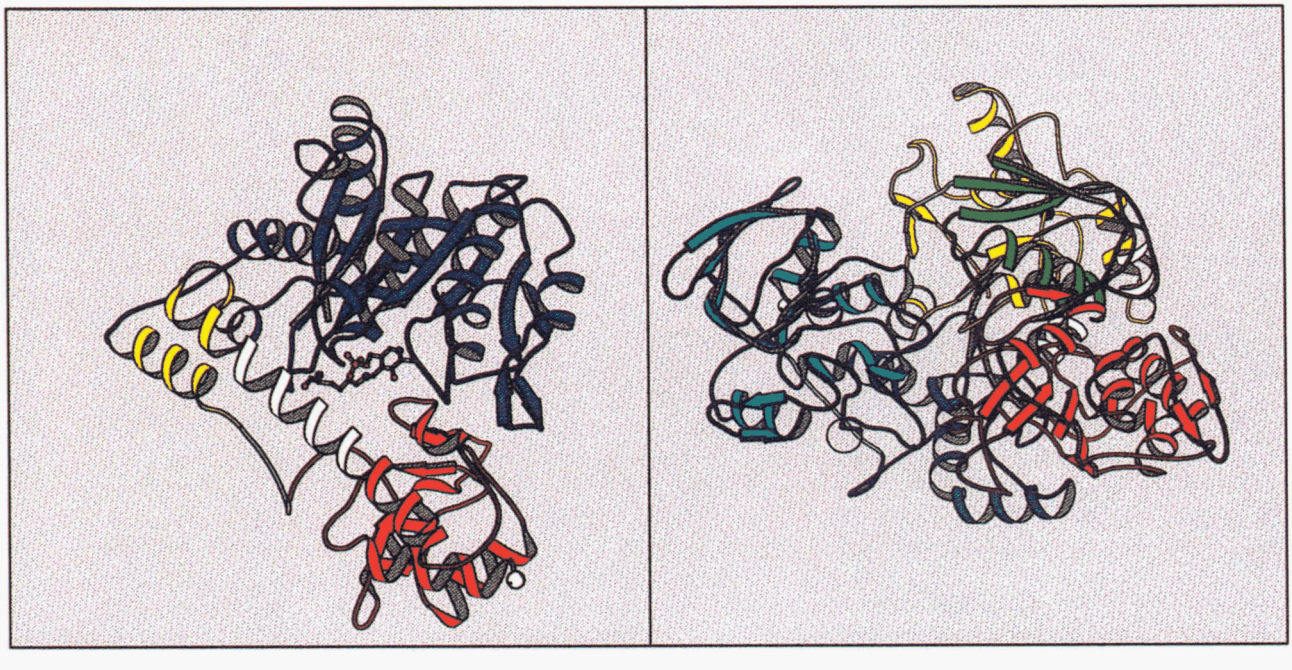

Fig. 3. Ribbon diagrams showing the domain assignments for aspartate aminotransferase (left) and aconitase (right). The coloring system is described with reference to the domain numbers written in Appendix 1. Aspartate aminotransferase is colored yellow, blue, and red for domains 1, 2, and 3, respectively. Aconitase is colored yellow, blue, green, red, and cyan for domains 1–5.

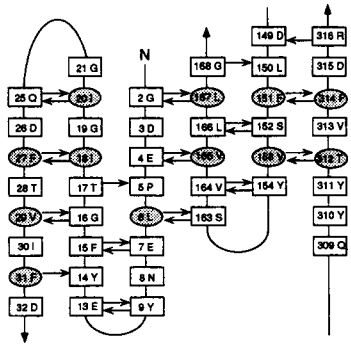

Fig. 4. Problems encountered when assigning pepsin domains. This schematic diagram shows residues of the β-sheet at the interface region, with Leu-6 creating the undesired connection between two otherwise separate domains. Ellipsoids denote core residues.

ditions for globularity, as well as rules for flexible loops and central β-sheet locations. Tests of their algorithm on a wide number of proteins suggested that it will be of significant use to those concerned with structural analysis. Nevertheless, ambiguous domain assignments were also reported, which confirm the difficulties associated with designing a robust automated procedure.

Methods

Domains are assigned using the following eight-step procedure, which is most easily understood by using the simplified example in Figure 5.

1. Hydrophobic cores are calculated using the algorithm described in the companion paper (Swindells, 1995), with an accessibility cutoff of 7%.

2. Isolated core sites are identified and removed from consideration. For instance, if site i is assigned to a core, but its sequentially adjacent core sites (one toward the N-terminus and another toward the C-terminus) are different, site i is classified as isolated and removed.

3. Cores now consisting of less than five sites are removed from consideration.

4. Of the assignments that remain, each site must now have at least one sequentially adjacent site belonging to the same core. These are joined together and effectively initiate the domain assignment process.

The next steps are concerned with extending these initial domain assignments over the entire molecule.

5. Analyze the remaining unassigned sites to see whether they make atomic contacts with any sites that *have* been assigned. In practice, this is done by analyzing the array $h_{(i,j)}$ described in the preceding paper (Swindells, 1995). If site i has not yet been assigned to a domain, all values of j that *have* been assigned are searched for contacts. If contacts are found and they consistently relate to the same domain, the site under consideration is tentatively assigned to the same domain.

6. After all tentative assignments have been made, isolated sites are again removed using the principles described in step 2. In this case, however, we also remove pairs of sites when their residues are sequentially adjacent, yet isolated from all other assignments. The purpose of this criterion is to eliminate the effects of two adjacent residues in a helix making similar contacts.

7. Once again sites are joined together in the manner described in step 4.

8. Most residues should now be assigned to a domain. In order to tidy up the remaining unassigned residues, we perform two tasks. Where possible, sites are extended to the ends of their appropriate secondary structures. Finally, assignments are extended to their N- and C-termini.

Full atom coordinates were obtained from the Protein Data Bank (Bernstein et al., 1977). Coordinate sets used in this work are: elastase, 3est (Meyer & Cole, 1988); rhodanese, 1rhd (Ploegman et al., 1978); alcohol dehydrogenase, 8adh (Colanna-Cesari et al., 1986); arabinose binding protein, 8abp (Quiocho & Vyas, 1984); thermolysin, 3tln (Holmes & Matthews, 1982); papain, 9pap (Kamphuis et al., 1984); aspartate aminotransferase, 7aat (McPhalen et al., 1992a); aconitase, 5acn (Robbins & Stout, 1989); pepsin, 5pep (Cooper et al., 1990); lactate dehydrogenase, 6ldh (Abad-Zapatero et al., 1987).

As stated in the companion paper (Swindells, 1995), this algorithm requires a set of full atom coordinates that have been refined at high resolution. Because calculations involving side chains are required, these should be well defined in the structure. The use of incomplete data sets (which frequently occur in low-resolution crystal structures) will inevitably have a dele-

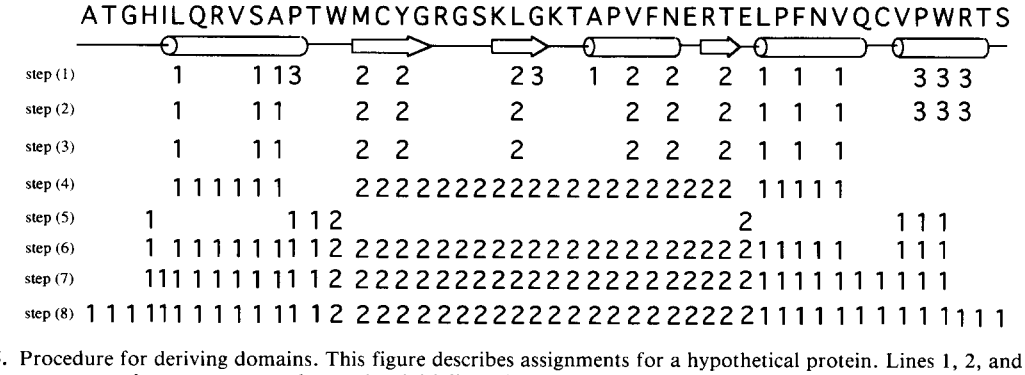

Fig. 5. Procedure for deriving domains. This figure describes assignments for a hypothetical protein. Lines 1, 2, and 3 show the sequence, secondary structure, and core sites initially assigned to the protein. The eight steps required for constructing a domain relate to the descriptions given in the Methods section.

terious effect on the algorithm's performance and are therefore not recommended.

All diagrams were made using Molscript (Kraulis, 1991).

Acknowledgments and program availability

I thank Drs. K. Nishikawa and Y. Kuroda for many helpful discussions during my work at PERI, Dr. David Jones and Prof. Janet Thornton for ideas on how to improve the algorithm, and Dr. Judith Healy for critically reading the manuscript. It is intended to make a computer program based on this work available. For more information, please contact the author at mark@yamanouchi.co.jp

References

Abad-Zapatero C, Griffith JP, Sussman JL, Rossmann MG. 1987. Refined crystal structure of M4 apo-lactate dehydrogenase. *J Mol Biol 198*:445-467.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 122*:535-542.

Busetta B, Barrans Y. 1984. The prediction of protein domains. *Biochim Biophys Acta 790*:117-124.

Colanna-Cesari F, Perahia D, Karplus M, Eklund H, Branden CI, Tapia O. 1986. Interdomain motion in liver alcohol dehydrogenase. Structural and energetic analysis of the hinge bending mode. *J Biol Chem 261*:15273-15280.

Cooper JB, Khan G, Taylor G, Tickle IJ, Blundell TL. 1990. X-ray analyses of aspartic proteases. II. *J Mol Biol 214*:199-222.

Crippen GM. 1978. The tree structural organisation of proteins. *J Mol Biol 126*:315-332.

Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123-138.

Holm L, Sander C. 1994. Parser for protein folding units. *Proteins Struct Funct Genet 19*:256-268.

Holmes MA, Matthews BW. 1982. Structure of thermolysin refined at 1.6 Å resolution. *J Mol Biol 160*:623-639.

Kamphuis IG, Kalk KH, Swarte MBA, Drenth J. 1984. Structure of papain refined at 1.65 Å resolution. *J Mol Biol 179*:233.

Kikuchi T, Némethy G, Scheraga HA. 1988. Prediction of the location of structural domains in globular proteins. *J Protein Chem 7*:427-471.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946-950.

Lapatto R, Blundell T, Hemmings A, Overington J, Wilderspin A, Wood S, Merson JR, Whittle PJ, Danley DE, Geohegan KF, Hawrylik SJ, Lee SE, Scheld KG, Hobart PM. 1989. X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among the retroviral enzymes. *Nature 342*:299-302.

McPhalen CA, Vincent MG, Jansonius JN. 1992a. X-ray structure, refinement and comparison of three forms of mitochondrial aspartate aminotransferase. *J Mol Biol 225*:495-517.

McPhalen CA, Vincent MG, Picot D, Jansonius JN, Lesk AM, Chothia C. 1992b. Domain closure in mitochondrial aspartate aminotransferase. *J Mol Biol 227*:197-213.

Meyer EF, Cole G. 1988. Structure of native porcine pancreatic elastase at 1.65 Å resolution. *Acta Crystallogr Sect B 44*:26-38.

Nishikawa K, Ooi T, Isogai Y, Saito N. 1972. Tertiary structure of proteins. I. *J Phys Soc Jpn 32*:1331-1337.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485-500.

Ploegman JH, Drent G, Kalk KH, Hol WGJ. 1978. Structure of bovine liver rhodanese. I. Structure determination at 2.5 Å resolution and a comparison of the conformation and sequence of its two domains. *J Mol Biol 123*:557-594.

Quiocho FA, Vyas NK. 1984. Novel specificity of the L-arabinose binding protein. *Nature 310*:381-386.

Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem 34*:167-339.

Robbins AH, Stout CD. 1989. The structure of aconitase. *Proteins Struct Funct Genet 5*:289-312.

Rose GD. 1985. Automatic recognition of domains in globular proteins. *Methods Enzymol 115*:430-440.

Swindells MB. 1995. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci 4*:93-102.

Taylor WR, Orengo C. 1989. Protein structure alignment. *J Mol Biol 208*:1-22.

Wodak SJ, Janin J. 1981. Location of structural domains in proteins. *Biochemistry 20*:6544-6552.

Zehfus MH. 1994. Binary discontinuous compact domains in proteins. *Protein Eng 7*:335-340.

Appendix 1

Details of the domain assignments for structures discussed in the text. Data are presented in the following manner:

Line 1: Sequence.
Line 2: Kabsch and Sander secondary structure assignments.
Line 3: Domain assignments. Residues belonging to the same domains are assigned the same numbers. Unassigned residues are left blank.

Elastase

```
VVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTLIRQNWVMTAAHCVDRELTFRVVVGEHNLNQNNGTEQYVGVQKIVVH
 BT EE  TTS TTEEEEEEEETTEEEEEEEEEEETTEEEE GGGG S   EEEEES  BTTS  S  EEEEEEEEEE
222222222222221111111111111111111111111111111111111111111111111111111111111111

PYWNTDDVAAGYDIALLRLAQSVTLNSYVQLGVLPRAGTILANNSPCYITGWGLTRTNGQLAQTLQQAYLPTVDYAICSS
TT  TT GGG    EEEEESS    BTTB       TT    TT  EEEEESS SSTT    SB EEEE  EE HHHHTS
1111111111111111111111111111111  222222222222222222222222222222222222222222222222

SSYWGSTVKNSMVCAGGDGVRSGCQGDSGGPLHCLVNGQYAVHGVTSFVSRLGCNVTRKPTVFTRVSAYISWINNVIASN
TTTTGGG  TTEEEE  BSS B  TT TT EEEEEETTEEEEEEEEEE  TT SS BT  EEEEEGGGTHHHHHHHHHT
2222222222222222222222222222222222222222222222222222222222222222222222222111111111111
```

Rhodanese

```
VHQVLYRALVSTKWLAESVRAGKVGPGLRVLDASWYSPGTREARKEYLERHVPGASFFDIEECRDKASPYEVMLPSEAGF
          EE HHHHHHHHHTT BTTTEEEEE      TTT  HHHHHTTSB TT EE  TTSSS TTSSSS     HHHH
11111111111111111111111111111111111111111111111111111111111111111111111111111111

ADYVGSLGISNDTHVVVYNGDDLGSFYAPRVWWMFRVFGHRTVSVLNGGFRNWLKEGHPVTSEPSRPEPAIFKATLNRSL
HHHHGGGS  TTSEEEE   STTS SSHHHHHHHHHHT    EEEETTHHHHHHHHT   B                  TTS
11111111111111111111111111111111111111111111111111111111111111111111111111111111

LKTYEQVLENLESKRFQLVDSRAQGRYLGTQPEPDAVGLDSGHIRGSVNMPFMDFLTENGFEKSPEELRAMFEAKKVDLT
EE TTHHHHHHHH  SEEEE S HHHHTT S  SS SS    B TT EE  TTTTB TTS B  HHHHHHHHHHTT  TT
22222222222222222222222222222222222222222222222222222222222222222222222222222222

KPLIATCRKGVTACHIALAAYLCGKPDVAIYDGSWFEWFHRAPPETWVSQGKG
S EEEE SSSSTTHHHHHHHHHTT TT EEETTTHHHHHHHS GGGSB SS
22222222222222222222222222222222222222222222222222222
```

Arabinose Binding Protein

```
NLKLGFLVKQPEEPWFQTEWKFADKAGKDLGFEVIKIAVPDGEKTLNAIDSLAASGAKGFVICTPDPKLGSAIVAKARGY
 EEEEEEES TTSHHHHHHHHHHHHHHHHHTEEEEEEE  SHHHHHHHHHHHHHTT  EEEEE S GGGHHHHHHHHHT
11111111111111111111111111111111111111111111111111111111111111111111111111111111

DMKVIAVDDQFVNAKGKPMDTVPLVMLAATKIGERQGQELYKEMQKRGWDVKESAVMAITANELDTARRRTTGSMDALKA
T EEEEESS  B TTS B TTS EEEE HHHHHHHHHHHHHHHHHHHT  GGGEEEEEEE TTSHHHHHHHHHHHHHHHH
11111111111111111111111222222222222222222222222222222222222222222222222222222222

AGFPEKQIYQVPTKSNDIPGAFDAANSMLVQHPEVKHWLIVGMNDSTVLGGVRATEGQGFKAADIIGIGINGVDAVSELS
HT  GGGEEEEE SSSSHHHHHHHHHHHHTT TT SEEEEE SSHHHHHHHHHHHHHTT  GGGEEEEEESSGGGHHHHT
22222222222222222222222222222222222222222222222222222222222222222222222222222222

KAQATGFYGSLLPSPDVHGYKSSEMLYNWVAKDVEPPKFTEVTDVVLITRDNFKEELEKKGLGGK
SSS  SEEEEEE  HHHHHHHHHHHHHHHHHH      SEEEE   EEEETTTHHHHHHHTT
222222222222  111111111111111111       22222222222222222222222222
```

Alcohol Dehydrogenase

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTPLPVIAGHEAAGIVESIGEGV
  TTS EEEEEEEB STTS  EEEEEEE    TTEEEEEEEEEE  HHHHHHHHTSS   SSB     EEEEEEEE TT
11111111111111111111111111111111111111111111111111111111111111111111111111111111

TTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPRGTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKI
 S  TT EEEE SS   S SHHHHSSS    TT SSSS   S TTS  SEEETTEE B STTT  SBSEEEEEGGGEEE
11111111111111111111111111111111111111111111111111111111111111111111111111111111

DAASPLEKVCLIGCGFSTGYGSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC
 TT  TTTGGGGGTHHHHHHHHHHTTT   TT EEEEE  SHHHHHHHHHHHHTT SEEEEE S GGGHHHHHHT SEE
11111111111111222222222222222222222222222222222222222222222222222222222222222222

VNPQDYKKPIQEVLTEMSNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQNLSMNPMLLLSGRTWKGAIFG
E GGG SS HHHHHHHHTTS BSEEEE S  HHHHHHHHHTB TTT EEEE    TT       THHHHTT EEEE SGG
22222222222222222222222222222222222222222222222222222222222222222222222222221111

GFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFDLLRSGESIRTILTF
G  HHHHHHHHHHHHHTTSS  GGGEEEEEEGGGHHHHHHHHSS  SEEEEE
11111111111111111111111111111111111111111111111111111
```

Papain

```
IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRSYGCNGGYPWSALQLVAQYGI
  S EETTTTT      B  SSB HHHHHHHHHHHHHHHHHHS    B HHHHHHH TTS TTB   HHHHHHHHHHT B
11111111       22222222222222222222222222222222222222222222222222222222222222222

HYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHA
 BTTTS   SS     SGGG S SB  SEEEE  SS HHHHHHHHHHS EEEEE   SHHHHT  SSEE     S   EE
2222222222222222222222222222222211111111111111111111111111111111111111111111111

VAAVGYGPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
EEEEEEETTEEEEE SB TTSTBTTEEEEE  SS SS GGGTTS  EEEE
1111111111111111111111111111111111111111111112222
```

Thermolysin

```
ITGTSTVGVGRGVLGDQKNINTTYSTYYYLQDNTRGDGIFTYDAKYRTTLPGSLWADADNQFFASYDAPAVDAHYYAGVT
   EEEEEEEE TTS EEEEEEEESSSEESEE SSTT EEEEE TTSSSSS EE EESSSEE SGGGHHHHHHHHHHHHH
1111111111111111111111111111111111111111111111111111111111111111111111111111111

YDYYKNVHNRLSYDGNNAAIRSSVHYSQGYNNAFWNGSEMVYGDGDGQTFIPLSGGIDVVAHELTHAVTDYTAGLIYQNE
HHHHHHHH   TTTSS   EEEEES SSS   EEE SSSEEE    SSSBS GGG HHHHHHHHHHHHHHHTT   SSHH
1111111111111111111111111111111111111111111111111111111111111111111111222222222

SGAINEAISDIFGTLVEFYANKNPDWEIGEDVYTPGISGDSLRSMSDPAKYGDPDHYSKRYTGTQDNGGVHINSGIINKA
HHHHHHHHHHHHHHHHHHHHTSS  SEESTTTB TTSSS  SEESS GGGGT  SSGGG   SSHHHHTTTTHHHHHHH
2222222222222222222222222222222222222222222222222222222222222222222222222222222

AYLISQGGTHYGVSVVGIGRDKLGKIFYRALTQYLTPTSNFSQLRAAAVQSATDLYGSTSQEVASVKQAFDAVGVK
HHHHHH EEETTEEE    HHHHHHHHHHHHHHT  TT  HHHHHHHHHHHHHHHH TTSHHHHHHHHHHHTT
2222222222222222222222222222222222222222222222222222222222222222222222222222222
```

Aspartate Aminotransferase

```
SSWWSHVEMGPPDPILGVTEAFKRDTNSKKMNLGVGAYRDDNGKPYVLNCVRKAEAMIAAKKMDKEYLPIAGLADFTRAS
  SSTT       HHHHHHHHHHH   TT EE    S   TTS     HHHHHHHHHHHHTT       TT  HHHHHHH
333333333333333333333333333333333333333333     11111111111112222222222222222222

AELALGENSEAFKSGRYVTVQGISGTGSLRVGANFLQRFFKFSRDVYLPKPSWGNHTPIFRDAGLQLQAYRYYDPKTCSL
HHHHH TT HHHHTT EEEEEEEHHHHHHHHHHHHHHHH TT  EEEEEES  TTHHHHHHHTT EEEEEE EETTTTEE
2222222222222222222222222222222222222222222222222222222222222222222222222222222

DFTGAMEDISKIPEKSIILLHACAHNPTGVDPRQEQWKELASVVKKRNLLAYFDMAYQGFASGDINRDAWALRHFIEQGI
 HHHHHHHHTTS TT EEEEESSS TTT     HHHHHHHHHHHHHTT EEEEEES TTTTTS HHHHTHHHHHHHTT
2222222222222222222222222222222222222222222222222222222222222222222222222222222

DVVLSQSYAKNMGLYGERAGAFTVICRDAEEAKRVESQLKILIRPMYSNPPMNGARIASLILNTPELRKEWLVEVKGMAD
  EEEEE TTTS  GGG EEEEEEE SSHHHHHHHHHHHHHHHHHHHSS  HHHHHHHHHHHH HHHHHHHHHHHHHHH
2222222222222222222222222222222222222222222222222222  11111111111

RIISMRTQLVSNLKKEGSSHNWQHITDQIGMFCFTGLKPEQVERLTKEFSIYMTKDGRISVAGVASSNVGYLAHAIHQVT
HHHHHHHHHHHHHHHHTT  S  HHHHH  SSEEE    HHHHHHHHHHH EE  TT EEEGGG  TTTHHHHHHHHHHHH
    333333333333333333333333333333333333333333333333333333333333333333333333333
```

Aconitase

```
 RAKVAMSHFEPHEYIRYDLLEKNIDIVRKRLNRPLTLSEKIVYGHLDDPANQEIERGKTYLRLRPDRVAMQDATAQMAM
  S  BS TT TTSB HHHHHHHHHHHHHHH S    HHHHHHHTTBS TTT    TTTSEEEE  SEEEEEHHHHHHHH
 111111111111111111111111111111111111111111111111111111111111111111111112222222222

LQFISSGLPKVAVPSTIHCDHLIEAQLGGEKDLRRAKDINQEVYNFLATAGAKYGVGFWRPGSGIIHQIILENYAYPGVL
HHHHHH  SS SS EEEE  SS   BSS HHHHHHHHHHHHHHHHHHHHHHHHHT EEE TTSB HHHHHHHT   TT E
222222222222222222222222222222222222222222222222222222222222222          1111111

LIGTDSHTPNGGGLGGICIGVGGADAVDVMAGIPWELKCPKVIGVKLTGSLSGWTSPKDVILKVAGILTVKGGTGAIVEY
EEESSTTGGGGGGGT EEEE  HHHHHHHHHT  EEEE  EEEEEEEES   TT   HHHHHHHHHHH TTTTTTEEEEE
11111111111111111111111111111111111111   333333333333333333333333333333333333333

HGPGVDSISCTGMATICNMGAEIGATTSVFPYNHRMKKYLSKTGRADIANLADEFKDHLVPDPGCHYDQVIEINLSELKP
ESSGGGGS HHHHHHHHHHGGGGT SEEE    HHHHHHHHHTT HHHHHHHHHTGGGSS  TT    SEEEEEESTT
333333333333333311111 333 111333333       11111111111111111111111113333333333333

HINGPFTPDLAHPVAEVGSVAEKEGWPLDIRVGLIGSCTNSSYEDMGRSAAVAKQALAHGLKCKSQFTITPGSEQIRATI
EEE SSSS SEEEHHHHHHHHHHHHT  BEEEEEEETTTB SHHHHHHHHHHHHHHHTTT    SSEEEEE SBHHHHHHH
44444444444444444444444444444444444444444444444444444444444444444444444444444444

ERDGYAQVLRDVGGIVLANACGPCIGQWDRKDIKKGEKNTIVTSYNRNFTGRNDANPETHAFVTSPEIVTALAIAGTLKF
HHTSHHHHHHHTT EE  SS GGGTT B   SS TT   EEEEESSS   TTTTTS TTEEEEE  HHHHHHHHH  BTT
44444444444444444444444444444444444444444444444444444444444444444444444444444444

NPETDFLTGKDGKKFKLEAPDADELPRAEFDPGQDTYQHPPKDSSGQRVDVSPTSQRLQLLEPFDKWDGKDLEDLQILIK
 TTT  EE TTS EE        SS SS     SS EE   S  TT      TT SSB          SS EEEEEEEEE
44444444444444444444444444444444444442222222222222222222222       5555555555555555

VKGKCTTDHISAAGPWLKFRGHLDNISNNLLIGAINIENRKANSVRNAVTQEFGPVPDTARYYKQHGIRWVVIGDENYGE
B S  BHHHHS  GGGGGGTB HHHHGGGTTTTSBBTTTS BT EE TTT  EE HHHHHHHHHTT   EEEE  SSBTB
55555555555555555555555555555555555555555555555555555555555555555555555555555555

GSSREHSALEPRHLGGRAIITKSFARIHETNLKKQGLLPLTFADPADYNKIHPVDKLTIQGLKDFAPGKPLKCIIKHPNG
S   THHHHHHHTTEEEEEEES B HHHHHHHHHTT EEEEESSGGGGGG  TT EEEEE  TT  TT  EEEEEE TTS
55555555555555555555555555555555555555555555555555555555555555555555555555555555

TQETILLNHTFNETQIEWFRAGSALNRMKELQQK
 EEEEEEE     HHHHHHHHHTSHHHHHHHHTT
5555555555555555555555555555555555
```

Pepsin

```
IGDEPLENYLDTEYFGTIGIGTPAQDFTVIFDTGSSNLWVPSVYCSSLACSDHNQFNPDDSSTFEATSQELSITYGTGSM
 EEEEEEEETTTEEEEEEEEETTTTEEEEEEEEEETT    EEE BTT   SHHHHSS   B GGG TT EEEEEEEEEE SS EE
11111111111111111111111111111111111111111111111111111111111111111111111111111111

TGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGILGLAYPSISASGATPVFDNLWDQGLVSQDLFSVYLSSNDD
EEEEEEEEEETTEEEEEEEEEEEEE   SSHHHH S   EEEE S GGG GGG    HHHHHHHTT SSSSEEEEE   STT
11111111111111111111111111111111111111111111111111111111111111111111111111111111

SGSVVLLGGIDSSYYTGSLNWVPVSVEGYWQITLDSITMDGETIACSGGCQAIVDTGTSLLTGPTSAIANIQSDIGASEN
S EEEEET    TTSBSS   EEEE SSTTTTEEEE EEEETTEEEE TT EEEEE TT   SEEE HHHHHHHHHHHT EE
11111111111111111111111111111111111111111111111111111111111111111111111111111111

SDGEMVISCSSIASLPDIVFTINGVQYPLSPSAYILQDDDSCTSGFEGMDVPTSSGELWILGDVFIRQYYTVFDRANNKV
TTS EEE TTTTTT    EEEEETTEEEEE HHHHEEEETTEEEESEEE       TT S EEE HHHHTTEEEEEEETTTTEE
11111111111111111111111111111111111111111111111111111111111111111111111111111111

GLAPVA
EEEEB
111111
```

Lactate Dehydrogenase

```
ATLKDKLIGHLATSQEPRSYNKITVVGVGAVGMACAISILMKDLADEVALVDVMEDKLKGEMMDLQHGSLFLHTAKIVSG
  HHHHH    S       SSEEEEE  SHHHHHHHHHHHTTT  SEEEEE S HHHHHHHHHHHHHGGGG   S EEEE
1111111111111111111111111111111111111111111111111111111111111111111111111111111

KDYSVSAGSKLVVITAGARQQEGESRLNLVQRNVNIFKFIIPNIVKHSPDCIILVVSNPVDVLTYVAWKLSGLPMHRIIG
SSGGGGTT SEEEE       TT  HHHHHHHHHHHHHHHHHHHHHH TT EEEE  SSHHHHHHHHHHHHT  GGGEEE
1111111111111111111111111111111111111111111111111111111111111111111111111111111

SGCNLDSARFRYLMGERLGVHSCSCHGWVIGEHGDSVPSVWSGMNVASIKLHPLDGTNKDKQDWKKLHKDVVDSAYEVIK
 TTHHHHHHHHHHHHHHHTS TTTEE  B BSSSTT  B GGG EETTEEHHHHSS  SSSSSSTHHHHHHHHHHHHHHH
1111111111111111111111111111111111111111111111111111111111111111111111111111111

LKGYTSWAIGLSVADLAETIMKNLCRVHPVSTMVKDFYGIKDNVFLSLPCVLNDHGISNIVKMKLKPNEEQQLQKSATTL
HHSS  HHHHHHHHHHHHHHHTT  EEEEEEEE TTSSS  SS EEEEEEEEETTEEEEE      HHHHHHHHHHHHHH
1111111111111111111111111111111111111111111111111111111111111111111111111111111

WDIQKDLKF
HHH   S
111111111
```