Rothfield L I , Zhao C R (1996) How do bacteria decide where to divide? Cell **84**, 183-186

Zhou X Z , Madhusudan Whiteley J M , Hoch J A , Varughese K I (1997) Purification and preliminary crystallographic studies on the sporulation response regulatory phosphotransferase protein, SpoOB, from *Bacillus subtilis* Proteins - Structure Function and Genetics **27**, 597-600

Extending Protein Families through Structural Studies

J. A. BRANNIGAN

Chemistry Dept , University of York, York YO1 5DD, UK

Key words: protein structures, evolution

Structural biology is in an exciting era Improved methods of protein production, crystallisation, X-ray data collection and analysis have greatly reduced the time and effort required to generate 3-D structures of proteins The number of known structures now forms a useful database to search for relatedness of form and function Fuelled by the explosion of primary sequences available from genome DNA sequencing projects, protein structures are revealing biochemical and evolutionary links

Two examples are described The structure of a DNA helicase suggests how the large number (currently over 300) of known proteins with a characteristic sequence motif "signature" may have evolved different activities and substrate specificity

In contrast, the emerging Ntn (N-Terminal Nucleophile) hydrolase family is used as an example of protein structures revealing an evolutionary relatedness between proteins which have diverged to such an extent that no sequence homology can be detected, even including their active sites

The apparent ease with which protein structures are determined is due to parallel advances in both the molecular and structural biology techniques involved Recombinant protein production and purification is now more reliable, leading to greater chances of success for protein crystallisation The dissection of large structures into their constituent domains has also proved to be a useful route to building up a native protein structure Improved methods of data collection, notably crystal freezing to minimise damage from increasingly brilliant X-ray beams and detection by CCD optimises the use of synchrotron time These advances mean that the biochemists can produce protein in which methionine residues are replaced by selenomethionine, and structural biologists can use Multiwavelength Anomalous Dispersion phasing within a reasonable time-frame to solve the protein structure, thus avoiding the requirement for heavy atom derivatives and concomitant problems with non-isomorphism

The two structures described here exemplify the wealth of information which can be gleaned from protein structure The first describes how the structure of one member of a protein super-family not only reveals biochemical and functional data, but also gives clues about the family evolution and acts as a guide to understanding how different activities and functionalities can be derived from elaborations of a common structural framework DNA helicase from *Bacillus stearothermophilus* was the first member of the helicase family whose structure was solved (Subramanya *et al* 1996) The helicases unwind nucleic acid duplexes and as such are involved in all biochemical activities which depend on nucleic acids as biological repositories of information, including replication, recombination and repair of DNA as well as transcription and translation of RNA (Lohman and Bjornson 1996) Mechanistic and sequence homology aspects suggest that the helicases can be grouped into families (Gorbalenya and Koonin 1993) and that the largest of these, corresponding to superfamilies I and II contains nearly always the characteristic sequence fingerprint of seven conserved motifs typical of helicases The structure reveals that all of these motifs are clustered at the interface of two major domains (Subramanya *et al* 1996) and provides a rationalisation for the fact that the greatest variation amongst the size of primary sequence between conserved motifs within the family occur between motifs Ia and II or IV and V, at the exact positions where two sub-domains can be identified as excursions from the functional domains This finding, allied with the identification of homology between the functional structural unit and RecA (Subramanya *et al* 1996, West 1996) not only suggested important biochemical links, but hinted at a modular organisation of the superfamily and a common structural theme and mechanism for all helicases (Bird *et al* 1998)

The second example is of a protein structure which helped identify a superfamily of proteins which are related in structure, but which share no primary sequence homology Whereas there are recurring motifs and folds clearly visible between some structures, eg HTH motifs, alpha/beta hydrolases and TIM barrels, the structure determination of Penicillin acylase at York (Duggleby *et al* 1995) allowed A Murzin (MRC Cambridge, UK) to compare the catalytic framework with that of the glutaminase domain of PRPP amidotransferase (Smith *et al* 1994) and so identify a common fold for a family of N-Terminal Nucleophile (Ntn) hydrolases (Brannigan *et al* 1995) This family share an unusual arrangement of beta sheet flanked by alpha helices which has since been identified in a number of other structures All are related in that an autocatalytic processing event reveals an N-terminal nucleophile, in which the primary alpha amino group is implicated in the catalytic mechanism by enhancing the nucleophilicity of its own side chain This means that the protein can be kept in an inactive state until activation by unmasking the reactive amino terminus We have recently solved the structure of a related enzyme, penicillinV acylase (Suresh *et al* in preparation) The preparation of semi-synthetic penicillins depends on the formation of 6-amino-penicillanic acid by enzymatic removal of the amide linked sidechain from naturally occurring beta lactam antibiotics The two enzyme types, penicillinV acylases and penicillinG acylases, with distinct substrate preferences, account for the global industrial production Representative examples of the two enzyme types differ widely in molecular properties PVA from *Bacillus sphaericus* is tetrameric with a monomer M_r of 35 000 whilst PGA from *Escherichia coli* is a heterodimer of M_r 90 000 Furthermore, they have no sequence homology There was no evidence of processing of the PVA molecule, and from the gene sequence it appears that its N-terminal sequence is Met-Leu-Gly-Cys- The crystal structure of PVA revealed a simple processing event to remove three residues and leave Cys at the new amino terminus, and a catalytic framework which is almost superimposable with that of PGA These two features firmly place it into the Ntn hydrolase family Indeed, we can now use the PVA sequence to as a search probe for other, related sequences This approach further extends the Ntn hydrolase family to include bile acid hydrolases and gives further clues to the evolution and biochemical role to this intriguing family of proteins which cannot be assigned by primary sequence alone, as the reactive centre can be Cys, Ser or Thr, and they can be processed from within the primary sequence thus the mature protein cannot be predicted solely from gene sequencing

The discussion of the two examples centres around the wealth of biological information available to us and the future of structural biology The increase in DNA sequencing activity is a main contributor to this data base There are 14 complete genome sequences already available This number will increase rapidly for microbial genomes and there is a concerted effort to derive the complete sequence of Human DNA In contrast, it is predicted that there are a finite number of protein folds, possibly as low as 1000 (Chothia 1992) of which about half have been identified, and are being classed in a structural hierarchy (Murzin 1996) The determination of the complete complement of structures for particular microorganisms has already been initiated, giving rise to the notion of "structural genomics" (Rost 1998) There is also a concerted effort to provide expression clones commercially of every single open reading frame from yeast There are currently over 1000 of these available from Invitrogen (www invitrogen com/genestorm) So where next for protein structure? Two avenues are discussed which await to be explored One is a mathematical description of 3D structures which would allow searches to be performed in a similar manner to DNA or protein sequence searches, thus generating alignments and revealing homologies This would also be useful in defining core structures and generating "average" structures for use in Molecular Replacement studies for structure solution It may lead also into another outstanding problem, in that there is no formal way of generating a phylogeny between related structures, which has been a particularly useful way of analysing multiple alignments of primary sequences

Acknowledgements. JAB is a Molecular Biologist funded by the Wellcome Trust The structures used as examples were solved by collaborators in Oxford University, UK (D B Wigley, DNA helicase) and NCL, Pune, India (C G Suresh, PenicillinV acylase, funded by a Higher Education Link Programme from the British Council) The Copernicus 1994 EU grant CIPA-CT-94-0189 and the organisers of the Copernicus programme are gratefully acknowledged for the chance to collaborate with colleagues from the Czech and Slovak Republics

References

Bird L E , Subramanya H S Wigley D B (1998) Helicases a unifying structural theme? Curr Op Struct Biol. **8**, 14-18

Brannigan J A , Dodson G , Duggleby H J , Moody P C E , Smith J L , Tomchick D R , Murzin A G (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation Nature **378**, 416-419

Chothia C (1992) One thousand protein families for the molecular biologist Nature, **357**, 543-544

Duggleby H J , Tolley S P , Dodson E J Dodson G , Moody P C E (1995) Penicillin acylase has a single amino acid catalytic centre Nature, **373**, 264-268

Gorbalenya A E , Koonin E V (1993) Helicases amino acid sequence comparisons and structure-function relationships Curr Op Struct Biol **3**, 419-429

Lohman T M Bjornson K P (1996) Mechanisms of helicase-catalysed DNA unwinding Ann Rev Biochem **65**, 169-214

Murzin A (1996) Structural classification of proteins new super families Curr Op Struct Biol **6**, 386-394

Rost B (1998) Marrying structure and genomics Structure **6**, 259-263

Smith J L , Zaluzec E J , Wery J-P , Niu L , Switzer R L , Zalkin H , Satow Y (1994) Structure of the allosteric regulatory enzyme of purine biosynthesis Science, **264**, 1427-1433

Subramanya H S , Bird L E , Brannigan J A , Wigley, D B (1996) Crystal structure of a DExx box DNA helicase Nature, **384**, 379-383

West S C (1996) DNA helicases get physical Nature, **384**, 316-317