4. Dayhoff, M. O. 1978. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C. 5:9–22 (suppl. 3).
5. Kopeyan, C., G. Martinez, S. Lissitzky, F. Miranda, and H. Rochat. 1974. Disulfide bonds of toxin II of the scorpion Androctonus Australis *Eur. J. Biochem.* 47:483–489.
6. Erlanson, C., J. A. Barrowman, and B. Borgstrom. 1977. Chemical modifications of pancreatic colipase. *Biochim. Biophys. Acta.* 489:150–162.

STRUCTURE OF HAPTOGLOBIN HEAVY CHAIN AND OTHER SERINE PROTEASE HOMOLOGS BY COMPARATIVE MODEL BUILDING

Jonathan Greer, *Department of Biological Sciences, Columbia University, New York, New York 10027 U.S.A.*

Proteins often occur in families whose structure is closely similar, even though the proteins may come from widely different sources and have quite distinct functions. It would be useful to be able to construct the three-dimensional structure of these proteins from the known structure of one or more of them without having to solve the structure of each protein *ab initio*. We have been using comparative model building to derive the structure of an unusual protein of the trypsin-like serine protease family (1). We have recently extended this comparison to include other serine protease homologs for which a primary structure is available.

To generate structures for the different members of the serine protease family, it is necessary to extract the common structural features of the molecule. Fortunately, three independently determined protein structures are available: chymotrypsin (2), trypsin (3, 4), and elastase (5). These three structures were compared in detail and the structurally conserved regions in all three, mainly the β-sheet and the α-helix, were identified. The variable portions occur in the loops on the surface of the molecule. By using these structures, the primary sequences of these three proteins were aligned. From this alignment, it is clear that sequence homology between the proteins occurs mainly in the structurally conserved regions of the molecule, while the variable portions show very little sequence homology.

The protein that has been built is haptoglobin (Hp), a serum protein that forms a highly specific and exceedingly strong complex with the blood protein hemoglobin. The in vivo function of Hp is to permit the recycling of red blood cell free hemoglobin iron and to prevent loss of heme iron in the urine and related damage to the kidney tubules eventually causing renal failure (6). Kurosky et al. (7, 8) have shown that the sequence of the heavy chain of Hp (HpH) is clearly homologous to the mammalian serine proteases, although the protein exhibits no protease activity.

The first step in modeling HpH into the known serine protease structure is to align the sequence to those of the known structures so that homology is maximized in the structurally conserved regions. Strong sequence homology was found for every structurally conserved region. No additions or deletions were found in these regions; all such occurred in the external loops where deviations are also found between the three known proteins. The resulting alignment shows that HpH must be very closely homologous to the proteases in structure as well as in sequence.

Coordinates were generated for HpH using the known homologus structures. Side chains

were replaced as required by the sequence. The coordinates were refined to remove overlaps and to close gaps that were due to deletions. Coordinates were constructed for the places where additional residues occur in HpH. Some loops in the structure, such as one containing a 17-residue addition that is found in no other homolog, will require further model building. A variety of features of HpH become apparent from the built structure (1), and these coordinates are being used to search for the interaction site with hemoglobin.

The method described above is also being applied to the other known serine protease homologs for which sequence data are available, including thrombin, plasmin, kallikrein, blood clotting factors IX and X, group specific protease, and *Streptomyces griseus* trypsin-like protein. In each case, the sequence was aligned by maximizing sequence homology in the structurally conserved regions of the molecule. Allowance was made for additional variability at positions that lie on the surface of the molecule and are exposed to solvent. From the eleven sequences studied, several structural themes emerge. Good sequence homology can be found for most of the structurally conserved regions, although one or two such regions have less-well defined homologies. Once again, virtually all additions and deletions can be assigned to loops. Thus, it is quite likely that the three-dimensional structure is closely conserved in all these proteins. This alignment also highlights the regions of differing structure between these proteins. Some sequences show a very large addition in a particular loop, while others may delete a loop almost completely. Occasionally it is difficult to align a particular sequence in a structurally conserved region. In other cases, a strongly conserved residue is changed. These probably indicate that the structure of this particular protein deviates in that part of the molecule.

Studies are continuing on classifying which residues are conserved because of functional or structural requirements, or just as a result of evolutionary vestige. Coordinates will be generated for these proteins and the differences in their structures examined in detail for their effect on protein function and specificity.

This research was supported by National Institutes of Health research grant HL16601 and facility grant RR00442.

Received for publication 15 December 1979.

REFERENCES

1. Greer, J. 1980. A model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. U.S.A.* In press.
2. Birktoft, J. J., and D. M. Blow. 1972. Structure of crystalline α-chymotrypsin. V. The atomic structure of tosyl α-chymotrypsin at 2.0 Å resolution. *J. Mol. Biol.* **68**:187–240.
3. Stroud, R. M., L. M. Kay, and R. E. Dickerson. 1974. The structure of bovine trypsin: electron density maps of the inhibited enzyme at 5Å and at 2.7Å resolution. *J. Mol. Biol.* **83**:185–208.
4. Huber, R., D. Kukla, W. Bode, P. Schwager, K. Diesenhofer, and W. Steigemann. 1974. Structure of the complex formed by bovine trypsin and bovine trypsin pancreatic inhibitor. II. Crystallographic refinement at 1.9Å resolution. *J. Mol. Biol.* **89**:73–101.
5. Shotton, D. M. and H. C. Watson. 1970. Three-dimensional structure of tosyl-elastase. *Nature (Lond.).* **225**:811–816.
6. Putnam, F. W. 1975. *In* The Plasma Proteins. F. W. Putnam, Editor. 2nd ed. Academic Press Inc., New York. Vol. II. 1–50.
7. Kurosky, A., D. R. Barnett, M. A. Rasco, T.-H. Lee, and B. H. Bowman. 1974. Evidence of homology between the β-chain of human haptoglobin and the chymotrypsin family of serine proteases. *Biochem. Genet.* **11**:279–293.
8. Kurosky, A., H.-H. Kim, D. R. Barnett, M. Rasco, B. Touchstone, and B. H. Bowman. 1975. Comparison of the primary structure of the β-chain of haptoglobin with serine proteases. *Protides Biol. Fluids Proc. Colloq.* **22**:597–602.