Critical Evaluation of Comparative Model Building of *Streptomyces griseus* Trypsin†

Randy J. Read, Gary D. Brayer,‡ Lubomir Jurášek,§ and Michael N. G. James*

ABSTRACT: The X-ray crystal structure of *Streptomyces griseus* trypsin has been solved and refined at 1.7-Å resolution. The structure of this protein had been predicted in two models on the basis of its expected homology to structures of bovine trypsin and other pancreatic serine proteases [Jurášek, L., Olafson, R. W., Johnson, P., & Smillie, L. B. (1976) *Miami Winter Symp. 11*, 93–123; Greer, J. (1981) *J. Mol. Biol. 153*, 1027–1042]. An evaluation of these models in light of the known structure demonstrates the effect of several sources of error on such comparative model building. The objective of comparative model building is often to explain substrate specificity, or to suggest potential highly specific drugs. The unique parts of modeled proteins that are most important for such purposes are, however, the most poorly determined by the model-building procedure.

The general protein folding problem, that of deducing the minimum free-energy conformation from just an amino acid sequence, is far from being solved. Since proteins seem to fall into a reasonably small number of structural families (Dayhoff, 1972), we often know that a structure will be similar to that of a homologous protein. This provides a powerful set of constraints that makes comparative model building much more tractable than prediction from sequence alone.

One of the earliest uses of comparative model building was the prediction of the structure of bovine trypsin (BT) from that of the homologous serine protease α-chymotrypsin (Hartley, 1970). The specificity of BT for Arg or Lys in the P1 position of the substrate was successfully explained by the substitution of Asp-189 for Ser in a region that is extremely similar in the two proteins.

However, most of the interesting questions that can be addressed by comparative model building involve features that are unique to the unknown structure, for example, nonhomologous parts of an enzyme involved in extended substrate specificity (Furie et al., 1982; Strassburger et al., 1983). Recently, it has been proposed that comparative models be used for the rational design of highly specific drugs (Blundell et al., 1983; Blow, 1983). The drug industry is reported to have shown considerable interest in the atomic coordinates for human renin, based on crystal structures of other aspartyl proteases (*The Economist*, 1984).

It is essential for these uses that one have an idea of the probable errors in comparative models, in both the homologous and the nonhomologous parts. Yet few of the protein structures that have been modeled have subsequently been determined by X-ray crystallography. The structure of BT has been known for some time (Stroud et al., 1972; Chambers et al., 1979; Bode & Schwager, 1975), but the overall accuracy of the model based on α-chymotrypsin has not been assessed. Delbaere et al. (1979) evaluated the model of α-lytic protease built by McLachlan & Shotton (1971) from the structure of elastase, but the large errors they found might not be typical of models based on more closely homologous structures. The sequences of elastase and α-lytic protease are identical in only 18% of the amino acids, whereas BT and α-chymotrypsin have 45% sequence identity (James et al., 1978).

We have solved the structure of *Streptomyces griseus* trypsin (SGT), which is homologous to BT (33% sequence identity). SGT has been modeled twice on the basis of this homology (Jurášek et al., 1976; Greer, 1981a). We are now in a position to examine the types and sizes of errors in comparative models, and the effects these errors may have on the usefulness of such models.

Comparative Models of SGT. Jurášek et al. (1976) built a model of SGT from Watson–Kendrew protein components, on the basis of preliminary coordinates for BT provided by R. M. Stroud. The path of the polypeptide was altered only where necessary to accomodate insertions and deletions. As far as possible, side-chain conformations of similar or identical residues were retained in the model of SGT. We shall refer to this model as LJ-SGT.

It might be argued that LJ-SGT does not represent the state of the art of comparative model building, because it was not constructed by using computer graphics. However, as long as computers are used only as a tool for manually adjusting torsion angles, albeit a more efficient and somewhat more accurate tool, there is no significant conceptual difference between a physical model and a computer-built model. Few comparative model-building studies have used computer techniques for the automatic adjustment of conformation. Furie et al. (1982) used a program to adjust side-chain (but not main-chain) torsion angles to minimize unfavorable close contacts in models of blood coagulation factors based on the pancreatic serine proteases. Warme et al. (1974) built a model of α-lactalbumin from the structure of hen egg white lysozyme and then refined it by energy minimization, which included some exploration of alternative conformations. Unfortunately, none of the structures of the proteins treated in these two studies has yet been determined.

Greer's (1981a) approach to modeling SGT was somewhat different, leading not to an actual model but rather to a descriptive outline of a model. The structures of three serine proteases (BT, elastase, and α-chymotrypsin) were aligned, as were the sequences of these and eight related proteases of unknown structure, including SGT. This revealed structurally conserved regions (SCRs) and variable regions (VRs). The

† From the Medical Research Council of Canada Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7. *Received May 7, 1984.* This work was funded by grants to the Medical Research Council Group in Protein Structure and Function by the Medical Research Council of Canada. R.J.R. holds a Medical Research Council of Canada studentship and an Alberta Heritage Foundation for Medical Research Allowance.

‡ Present address: Department of Biochemistry, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

§ Present address: Pulp and Paper Research Institute of Canada, Pointe Claire, Quebec, Canada H9R 3J9.

0006-2960/84/0423-6570$01.50/0

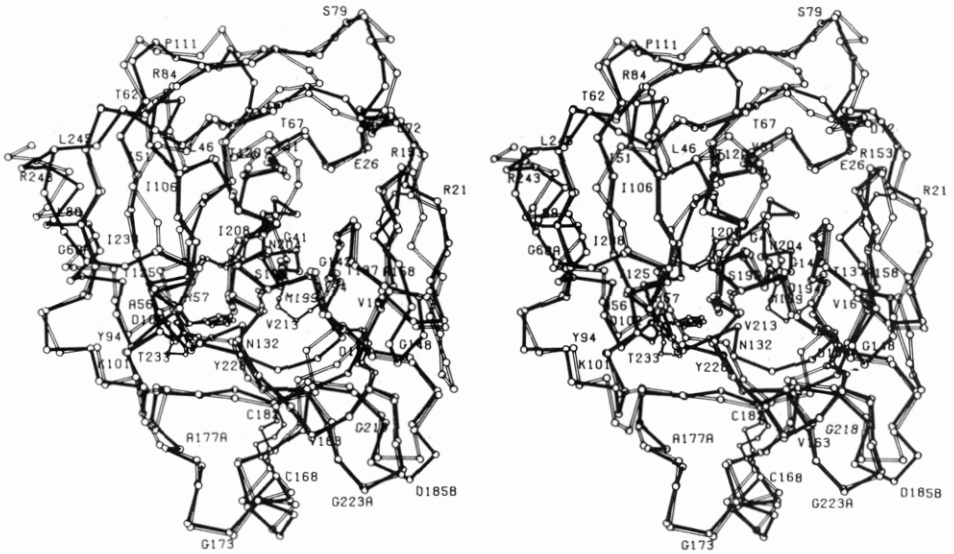


FIGURE 1: α-Carbon representation of the superposed structures of SGT (filled bonds) and BT (open bonds). Every fifth amino acid residue in SGT is labeled with the residue type and sequence number. Disulfide bridges and side-chain atoms of His-57, Asp-102, and Ser-195 of both proteins are also shown.

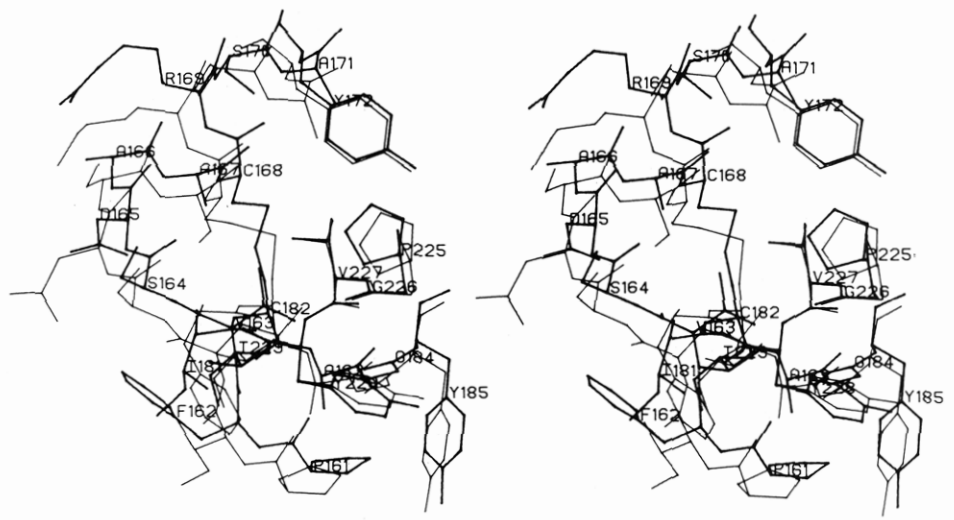


FIGURE 2: Comparison between SGT (thick lines) and BT (thin lines) in the region composed of the segments 161–172, 181–185, and 225–229. The sequences and, for the most part, the structures are very similar in this region. However, there are larger differences in the segment 161–171, including a sizable shift of the main chain. The side chain of Phe-162 does not occupy the same volume as that of Ile-162 in BT. In part, this may compensate for the replacement of Phe-181 in BT by Ile in SGT. Finally, the side chains of Asp-165 and Ser-170 and the disulfide bridge 168-182 take on different conformations. The differences between the two structures in this region are probably related to the large differences in the segment 130–133 (not shown in this figure: see Figure 1; see also text).

sequence alignment provided the core of SCRs for the model of each unknown protease. Instead of constructing the VRs, Greer suggested which known structure would give the best starting model for each VR. We shall refer to Greer's model outlined for SGT as JG-SGT.

Experimental Procedures

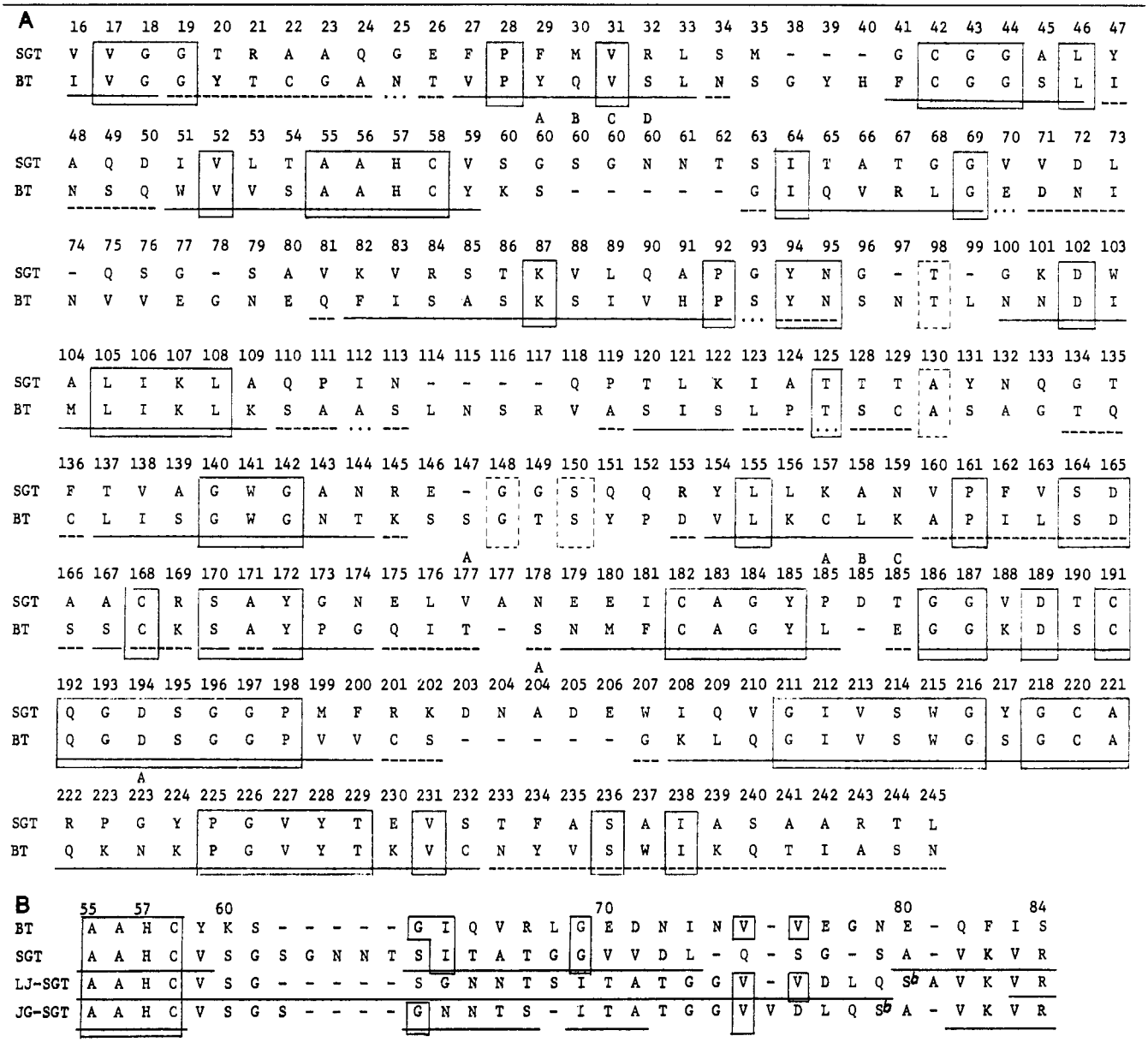
Structure Solution and Refinement. The structure of SGT was solved by the technique of molecular replacement; BT (Chambers & Stroud, 1979) and α-chymotrypsin (Birktoft & Blow, 1972) were search models. Additional phase information, useful to about 5-Å resolution, was available from three isomorphous derivatives, so the initial electron density maps used combined phases (Hendrickson & Lattman, 1970). The structure was refined by a restrained least-squares procedure (Hendrickson & Konnert, 1980). Refinement was interrupted at several points to perform manual electron density fitting on an MMS-X interactive graphics system (Barry et al., 1976), using the macromolecular modeling program M3 developed by C. Broughton (Sielecki et al., 1982). At present, the R factor[1] is 0.159 for 17 052 reflections [$I \geq 2\sigma(I)$] from 6.0- to 1.7-Å resolution. Further refinement will be performed but should not significantly affect the results reported here.

Results and Discussion

Evaluation of Sequence Alignments. In comparative model building, one must correctly align the amino acid sequences for the homologous parts; alignments of nonhomologous parts are essentially arbitrary for structural purposes. Insofar as the sequence alignment is incorrect, the model is guaranteed to be wrong. Table IA shows the results of the structural alignment of SGT and BT (see Figure 1) and indicates the homologous segments. We can compare this structurally derived sequence alignment with those of the two SGT models,

[1] $R = \sum||F_o| - |F_c||/\sum|F_o|$, where $|F_o|$ and $|F_c|$ are the observed and calculated structure factor amplitudes, respectively.

Table I[a]



[a] Part A of this table shows the structurally derived sequence alignment of SGT with BT. [Sequence numbering is based on an alignment of SGT with α-chymotrypsin (Birktoft & Blow, 1972).] The structures were aligned initially by using a program of W. Bennett, based on the principles of Rossmann & Argos (1975). The 190 underlined residues (solid, dashed, or dotted lines) are considered to be structurally equivalent. The 122 residues with a solid underline are those for which the α-carbon atoms can be superimposed simultaneously within 1 Å; the root mean square (rms) deviation in their positions is 0.52 Å. (The relative orientation of BT derived from the superposition of these atoms is used for all figures.) Dashed lines indicate the additional 63 residues in the set of 185 for which the α-carbon atoms can be superimposed within 1.9 Å (approximately half of the distance between α-carbon atoms of sequential residues); the rms deviation for the 185 α-carbon atoms is 0.88 Å. Dotted underlines indicate the five single residues with excursions greater than 1.9 Å that join segments with solid or dashed underlines. Boxes outline identical residues paired by the alignment, with dashed boxes indicating those fortuitously aligned in nonhomologous regions. There are 74 identical residues of the 223 in the sequence (33.2%). Of the 190 structurally equivalent residues, 70 (38.8%) are identical. With the more stringent criterion (1-Å limit), 59 of 122 (48.4%) are identical. Part B of Table I is a comparison of sequence alignments of SGT with BT in the region 55–84. The first alignment is the correct alignment, based on the superposition of the two structures. The next two are incorrect alignments predicted in the model-building studies of Jurášek et al. (1976) and Greer (1981a). Underlines indicate residues considered, for each alignment, to be structurally equivalent to residues of BT. Boxes outline residues paired with identical residues in one or more of the alignments. [b] The refinement of SGT has shown that the amino acid sequence (Olafson et al., 1975) must be corrected by the insertion of two residues, tentatively identified from the electron density as Gly-Ser.

and with those from three studies more concerned with evolutionary relationships (de Haën et al., 1975; Hewett-Emmett et al., 1981; Titani et al., 1983). Of the 190 homologous residues, de Haën et al. (1975) correctly matched 151 (79%), Hewett-Emmett et al. (1981) 156 (82%), Titani et al. (1983) 149 (78%), Greer (1981a) 173 (90%), and Jurášek et al. (1976) 174 (91%). This supports Greer's assertion that alignment by maximizing sequence equivalence (minimizing evolutionary distance) is inadequate for model building (Greer, 1981a), though it may be appropriate for tracing evolutionary pathways. Even when structural information is used, major alignment errors occur. For example, the segment 63-73 (which includes Greer's SCR 63–71) is structurally homologous in SGT and BT, but virtually undetectable sequence homology caused all predicted alignments to be wrong. Table IB shows correct and incorrect alignments of this segment.

Constructing Homologous Parts. In comparative model building, it is generally assumed that side chains in similar

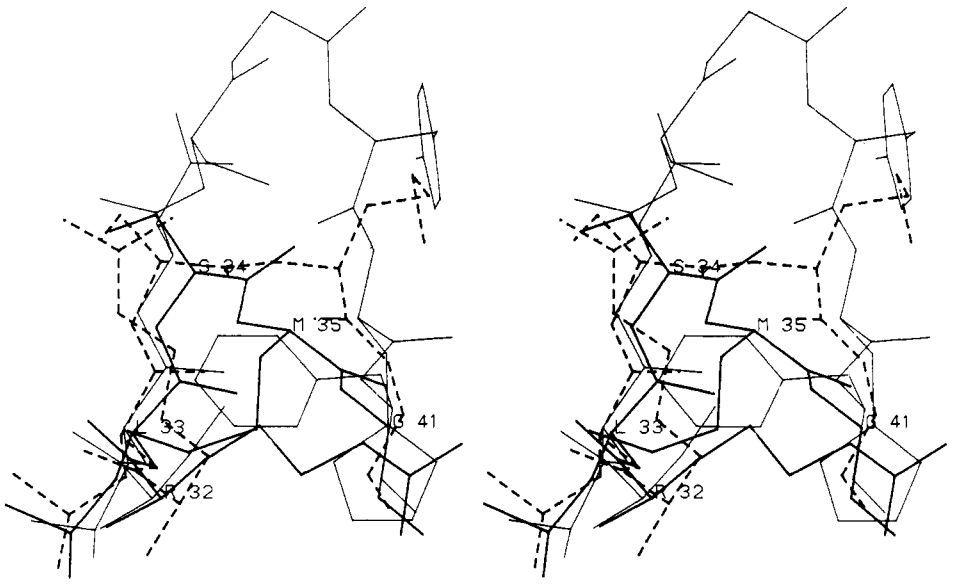


FIGURE 3: Errors in model building of nonhomologous regions. Comparison of the segment 32-41 in SGT (thick lines), BT (thin lines), and LJ-SGT (dashed lines). Coordinates for LJ-SGT used in this figure and in Figure 4 were measured from the model, refined against restraints for ideal geometry (Hendrickson & Konnert, 1980) and oriented relative to BT by least-squares superposition of those atoms judged during model building (Jurášek et al., 1976) to be identical in SGT and BT. Some differences in conformation (e.g., the side chain of Leu-33) probably result from the fact that the BT coordinates available for model building in 1976 were unrefined, preliminary coordinates supplied by R. M. Stroud. The loop 32–41 is three residues shorter in SGT than in BT. LJ-SGT accommodates this deletion with minimal perturbation of the main chain from BT. In this model, the side chain of Arg-32 projects into the solvent. In fact, the main chain and side chain of Met-35 occupy the space vacated by the change of Phe-41 to Gly in SGT. The side chain of Arg-32 crosses the loop, placing the guanidinium group in the position occupied by the side chain of His-40 in BT. His-40 has been implicated in the stabilization of the zymogen forms of chymotrypsin (Freer et al., 1970) and trypsin (Fehlhammer et al., 1977). It is evidently not essential to the zymogen mechanism, however, because in kallikrein, Phe-40 has the same position and conformation (Bode et al., 1983). Whether a zymogen exists for SGT is not known; this unexpected conservation of a positive charge may occur for some other reason.

environments will adopt similar conformations. This is a useful generalization, supported by the fact that all 475 atoms of the 70 identical, structurally equivalent residues of SGT and BT superimpose with a root mean square (rms) deviation of 1.01 Å. But there are exceptions. Figure 2 shows a region in which strong sequence similarity gives rise to varying degrees of structural similarity.

An inspection of Figures 1 and 2 shows that there is an additional complication in modeling homologous parts. Local structure is conserved much more strongly than global structure. Individual homologous segments have very similar conformations but differ slightly in their orientation and position relative to other homologous segments. The data in Table II demonstrate the improvement in alignment obtained by considering the segments individually instead of globally. A particularly clear example is the C-terminal α-helix. Figure 1 shows that this helix is packed more tightly against the β-sheet in SGT than in BT. Closer packing might have been predicted from the less bulky side chains in SGT between these elements of secondary structure. In this interface in BT, one finds, for example, the side chains of His-91, Trp-237, Thr-241, and Ile-242. All of these residues are alanines in SGT. At the C-terminus, Asn-245 is replaced by Leu in SGT. A substitution of Ile-51 for Trp-51 opens a hydrophobic pocket for Leu-245 in SGT, allowing the end of the helix to approach more closely to the sheet. Similar observations of relative shifts of secondary structural elements have been made in other protein families by Lesk & Chothia (1980, 1982; Chothia & Lesk, 1982).

Some errors arise from expecting the two structures to be similar where they are not. In both LJ-SGT and JG-SGT, the segment 130–133 was constructed by using expected homology to BT. In fact, the two proteins differ markedly here (Figure 1). Several features of SGT are incompatible with the conformation found in BT; which of these cause the difference and which are compensating readjustments cannot be determined unambiguously. The absence in SGT of the disulfide bridges 129–232 and 136–201 found in BT might also be of relevance. A reasonable goal for more sophisticated methods of comparative model building would be the ability

Table II: Global and Local Conservation of Structure

SGT segment[a]	BT segment[a]	rms-(global) (Å)[b]	rms-(local) (Å)[b]	angle (deg)[c]	distance (Å)[c]
V16–Q24	I16–A24	1.25	0.81	9.0	0.49
G25–S34	N25–N34	1.24	0.93	9.8	0.15
G41–Q49	F41–S49	0.84	0.77	2.5	0.21
D50–V59	Q50–Y59	0.81	0.56	5.9	0.04
S63–L73	G63–I73	1.19	0.93	3.9	0.09
V81–K87	Q81–K87	0.80	0.35	3.1	0.35
V88–N95	S88–N95	1.21	0.87	10.2	0.05
G100–I106	N100–I106	0.50	0.29	7.5	0.14
K107–N113	K107–S113	1.51	0.61	9.2	0.56
P119–T129	A119–C129	1.54	0.76	28.5	0.52
G134–R145	T134–K145	0.92	0.50	6.4	0.02
R153–S164	D153–S164	0.99	0.55	6.3	0.00
D165–V177	D165–T177	1.24	0.83	4.4	0.76
N178–P185A	S178–L185A	0.72	0.50	7.2	0.17
T185C–G193	E185C–G193	0.65	0.45	5.2	0.09
D194–K202	D194–S202	0.95	0.79	5.6	0.05
W207–G216	G207–G216	0.68	0.54	9.6	0.04
Y217–P225	S217–P225	0.38	0.23	4.4	0.11
G226–A235	G226–V235	0.80	0.46	2.4	0.27
S236–L245	S236–N245	1.53	0.59	4.7	0.37

[a] The structurally equivalent residues were divided into segments of about 10 residues, broken where possible at turns. [b] rms(global) and rms(local) refer to the rms deviation between main-chain atoms (N, C^{α}, C, O) of the segments, calculated from the global superposition of BT on SGT, and after a local superposition by least squares of the main-chain atoms of the segment. [c] Angle and distance are a measure of the amount of reorientation involved in the local superposition and refer to the angle of rotation and distance of translation along the rotation axis.

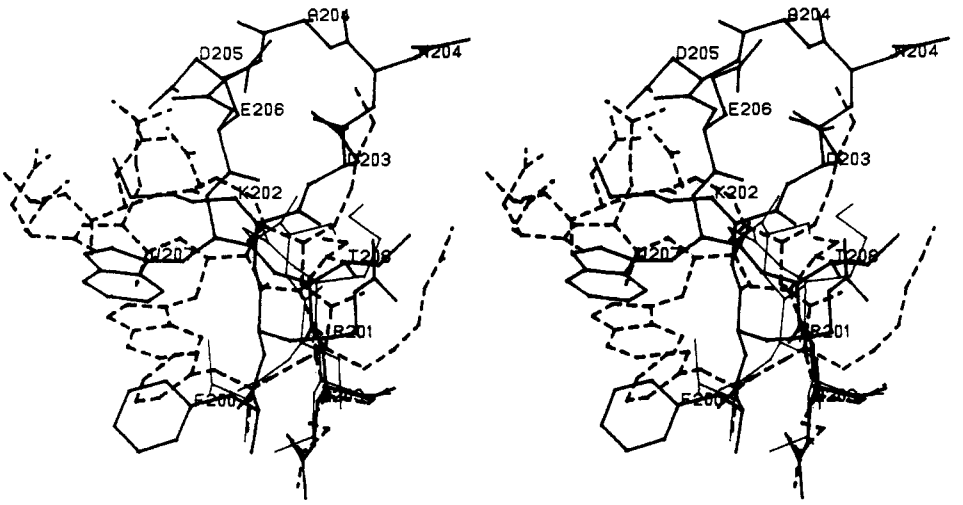


FIGURE 4: Errors in model building of nonhomologous regions. Comparison of the segment 200–209 in SGT (thick lines), BT (thin lines), and LJ-SGT (dashed lines). This loop is longer by five residues in SGT than in BT. In LJ-SGT, the extension is folded back over the side chain of Trp-207, covering this potential hydrophobic surface and making additional contacts to the rest of the protein. In fact, the extension projects out from the surrounding surface of the protein (compare with Figure 1).

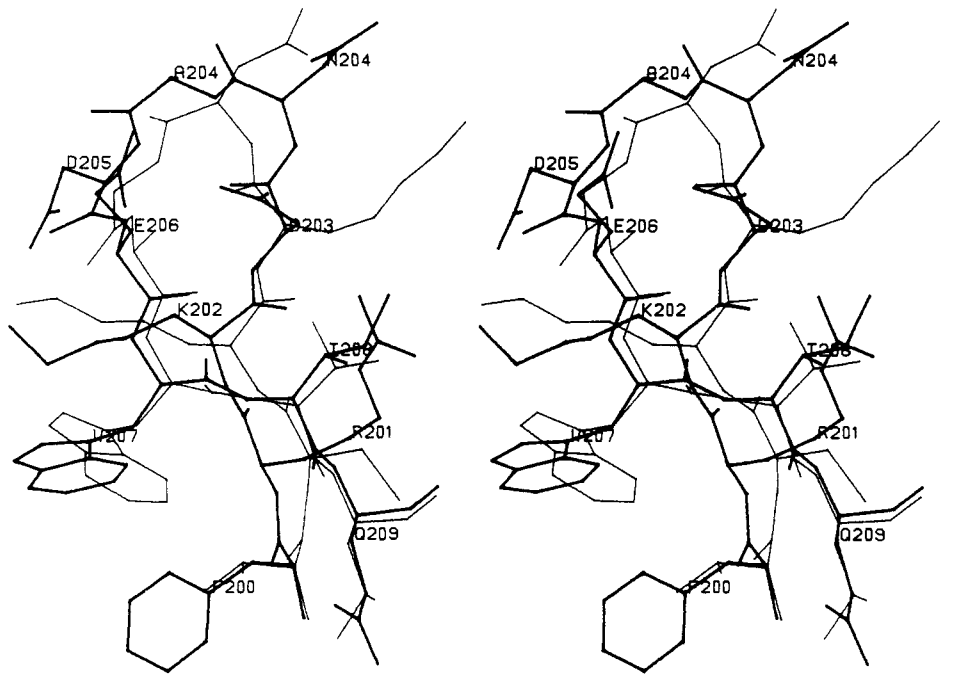


FIGURE 5: Errors in model of nonhomologous regions. Comparison of SGT (thick lines) and α-chymotrypsin (thin lines) in the same region shown in Figure 4. As predicted by Greer (1981a), SGT is very similar here to α-chymotrypsin, differing mainly by a single residue insertion at the end of the loop.

to recognize the necessity for conformational change in cases such as this.

Constructing Nonhomologous Parts. The most difficult part of comparative model building is constructing the nonhomologous parts. This problem is a small-scale, somewhat constrained, version of the general folding problem. The model of α-lactalbumin was constructed in the regions where it is not homologous to hen egg white lysozyme by exploring possible conformations of dipeptides, combining dipeptides with lower-energy conformations, and then refining by energy minimization (Warme et al., 1974). Exhaustive exploration of conformations for whole regions would probably be superior, though computationally demanding, and the use of molecular dynamics might obviate the need for manual intervention to escape local energy minima. Nonetheless, this work represents an exception to the usual procedures, and it will be interesting to see how close this model is to the true structure. In most comparative models, nonhomologous segments have been constructed by intuition, combined with a fairly limited manual exploration of the possible conformations (Furie et al., 1982; Strassburger et al., 1983; Blundell et al., 1983; Jurášek et al., 1976; Greer, 1981b). Two examples from LJ-SGT demonstrate the inadequacy of this approach.

In SGT, the loop 34–41 has a deletion of three residues relative to that in BT. In contrast, the loop 200–209 in SGT requires the insertion of five residues into the structure of BT. Neither the deletion nor the insertion was modeled correctly in LJ-SGT (Figures 3 and 4). For JG-SGT, it was suggested that α-chymotrypsin, which lacks only one residue in the loop 200–209 compared to SGT, would provide a good starting point. Indeed, Figure 5 shows that SGT is quite similar to α-chymotrypsin in this region.

Effects of Errors. At the present level of sophistication, comparative model building has a number of sources of error.

How serious they are will depend on the accuracy required for the uses to which the model will be put. For molecular replacement, large differences are evidently tolerated; even BT was a sufficiently good model of SGT, though barely.

One must also consider the fact that the accuracy of homologous and nonhomologous parts of comparative models differs quite widely. Predictions of conserved structure involving conserved sequence are almost certain to be correct, as was the prediction for SGT that the N-terminus would form an ion pair with the side chain of Asp-194 (Jurášek et al., 1976). It is also reasonable to make predictions involving nonconserved amino acids in homologous regions [e.g., the role in substrate specificity of Asp-189 in trypsin (Hartley, 1970)], though these are more susceptible to errors in side-chain conformation, or to misalignments of sequence.

Predictions involving models of nonhomologous regions are generally of more interest but will be much less reliable. On a gross level, these models can suggest which parts of an enzyme contribute to unique aspects of substrate specificity (Furie et al., 1982; Strassburger et al., 1983) or can help to organize experimental data [e.g., data on protection from proteolysis when haptoglobin binds to hemoglobin (Lustbader et al., 1983)]. The details are, however, quite likely to be inaccurate. Attainment of the accuracy and precision needed to design highly specific drugs (Blundell et al., 1983; Blow, 1983) will require considerably more sophisticated techniques.

Acknowledgments

We are indebted to Koto Hayakawa, who grew the crystals of SGT used in this study.

Registry No. SGT, 56644-95-2.

References

Barry, C. D., Molnar, C. E., & Rosenberger, F. U. (1976) *Technical Memo No. 229*, Computer Systems Lab, Washington University, St. Louis, MO.

Birktoft, J. J., & Blow, D. M. (1972) *J. Mol. Biol. 68*, 187–240.

Blow, D. (1983) *Nature (London) 304*, 213–214.

Blundell, T., Sibanda, B. L., & Pearl, L. (1983) *Nature (London) 304*, 273–275.

Bode, W., & Schwager, P. (1975) *J. Mol. Biol. 98*, 693–717.

Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G., & Bartunik, H. (1983) *J. Mol. Biol. 164*, 237–282.

Chambers, J. L., & Stroud, R. M. (1979) *Acta Crystallogr., Sect. B B35*, 1861–1874.

Chothia, C., & Lesk, A. M. (1982) *J. Mol. Biol. 160*, 309–323.

Dayhoff, M. O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC.

de Haën, C., Neurath, H., & Teller, D. C. (1975) *J. Mol. Biol. 92*, 225–259.

Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1979) *Nature (London) 279*, 165–168.

Fehlhammer, H., Bode, W., & Huber, R. (1977) *J. Mol. Biol. 111*, 415–438.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., & Xuong, Ng. H. (1970) *Biochemistry 9*, 1997–2009.

Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., & Furie, B. C. (1982) *J. Biol. Chem. 257*, 3875–3882.

Greer, J. (1981a) *J. Mol. Biol. 153*, 1027–1042.

Greer, J. (1981b) *J. Mol. Biol. 153*, 1043–1053.

Hartley, B. S. (1970) *Philos. Trans. R. Soc. London, Ser. B 257*, 77–87.

Hendrickson, W. A., & Lattman, E. E. (1970) *Acta Crystallogr., Sect. B B26*, 136–143.

Hendrickson, W. A., & Konnert, J. H. (1980) in *Biomolecular Structure, Function, Conformation and Evolution* (Srinivasan, R., Ed.) Vol. I, pp 43–57, Pergamon Press, Oxford.

Hewett-Emmett, D., Czelusniak, J., & Goodman, M. (1981) *Ann. N.Y. Acad. Sci. 370*, 511–527.

James, M. N. G., Delbaere, L. T. J., & Brayer, G. D. (1978) *Can. J. Biochem. 56*, 396–402.

Jurášek, L., Olafson, R. W., Johnson, P., & Smillie, L. B. (1976) *Miami Winter Symp. 11*, 93–123.

Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol. 136*, 225–270.

Lesk, A. M., & Chothia, C. (1982) *J. Mol. Biol. 160*, 325–342.

Lustbader, J. W., Arcoleo, J. P., Birken, S., & Greer, J. (1983) *J. Biol. Chem. 258*, 1227–1234.

McLachlan, A. D., & Shotton, D. M. (1971) *Nature (London) 229*, 202–205.

Olafson, R. W., Jurášek, L., Carpenter, M. R., & Smillie, L. B. (1975) *Biochemistry 14*, 1168–1177.

Rossmann, M. G., & Argos, P. (1975) *J. Biol. Chem. 250*, 7525–7532.

Sielecki, A. R., James, M. N. G., & Broughton, C. G. (1982) in *Crystallographic Computing, Proceedings of the International Summer School* (Sayre, D., Ed.) pp 409–419, Oxford University Press, Oxford.

Strassburger, W., Wollmer, A., Pitts, J. E., Glover, I. D., Tickle, I. J., Blundell, T. L., Steffens, G. J., Günzler, W. A., Ötting, F., & Flohé, L. (1983) *FEBS Lett. 157*, 219–223.

Stroud, R. M., Kay, L. M., &Dickerson, R. E. (1972) *Cold Spring Harbor Symp. Quant. Biol. 36*, 125–140.

The Economist (1984) *290* (*No.* 7323), 71–74.

Titani, K., Sasagawa, T., Woodbury, R. G., Ericsson, L. H., Dörsam, H., Kraemer, M., Neurath, H., & Zwilling, R. (1983) *Biochemistry 22*, 1459–1465.

Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W., & Scheraga, H. A. (1974) *Biochemistry 13*, 768–782.