PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Computer Methods for Molecular Design [and Discussion]

D. N. J. White and B. Robson

Phil. Trans. R. Soc. Lond. A 1986 **317**, 359-369
doi: 10.1098/rsta.1986.0046

Phil. Trans. R. Soc. Lond. A **317**, 359–369 (1986)
Printed in Great Britain

Computer methods for molecular design

By D. N. J. White†

Ciba-Geigy, R1046.109, CH-4002 Basel, Switzerland

The computational design and simulation of the properties of proteins requires powerful computers, colour graphics and interactive software. A brief description of such a system will be presented.

One of the key requirements in the design of novel proteins is the ability to construct a three-dimensional model of the target protein from a knowledge of its amino-acid sequence and the X-ray crystal structure of a homologous protein. Most of the models constructed in this fashion have used the computer as an electronic Dreiding model; where the inserted, deleted, and substituted amino acids are placed in three-dimensional space by subjective, visually oriented procedures (R. J. Read, G. D. Brayer, L. Jurasek & M. N. G. James, *Biochemistry* **23**, 6570 (1984)).

An objective computational procedure has been developed whereby the insertions, deletions, and substitutions are performed automatically by making use of a list of rules and a hierarchical energy minimization scheme. The procedure is highly interactive and all of the options are invoked by pointing to the appropriate amino acid residues displayed on the graphics screen.

Applications of the procedure and the use of the finished models in enzyme–substrate docking studies for drug design are described.

A molecular graphics system

To be generally useful in an environment where the user community has a varied level of expertise, a molecular modelling system must be very carefully designed, particularly at the man–machine interface. It is desirable that the system cope with organic, inorganic and organometallic compounds with equal facility. The software should use recognized standards wherever possible (for example, the Brookhaven file format for polypeptides, proteins, polynucleotides, and nucleic acids: the GKS (Graphics Kernel System) standard for computer graphics) and be fully documented. The Fortran code should be fully structured, long variable names should be used to make the code 'readable', and the code should be liberally interspersed with comments. The desirability of the above features is self-evident, and although they take somewhat longer to implement in the short term there is a very considerable saving in time and effort over the long term. Unfortunately only a minute proportion of currently available software adheres to this philosophy; this is particularly true of university-developed software, most of which violates all of the above design principles.

We have attempted to produce a molecular modelling system, COGS, that fulfils as many of the above design criteria as possible.

There are extensive model-building features: polypeptides may be constructed by pointing to a menu of amino acids, and folded by pointing to a table of low-energy conformations for each residue; organic molecules may be built in a similar fashion where the building blocks

† Permanent address: Chemistry Department, The University, Glasgow G12 8QQ, U.K.

are atoms rather than amino acids and conformations may be selected from tables of low-energy ring conformations; models of molecules may be drawn on the screen in the familiar two-dimensional manner, and these are automatically 'three-dimensionalized'; models may be built and modified by providing bond lengths, angles, and torsion angles; and if all else fails, lists of crystallographic or Cartesian coordinates may be input.

Once built, the model molecule may be modified in various ways: protein models may be built from sequence homology, which will be discussed further; molecules may be joined together by 'fusing' atoms or bonds (there is no limit to the number of molecules that may be simultaneously displayed); hydrogen atoms may be added automatically to a medium-mass atom skeleton; atoms, groups of atoms, and whole molecules may be deleted from the workspace; bonds may be broken and joined; and bond lengths, angles, torsion angles and molecular planes may be dynamically varied by user interaction with moving pictures.

Various molecular parameters may be calculated from the model, such as bond lengths, angles, torsion angles, pseudorotation descriptors for rings, steric congestions, molecular surface areas and volumes, atomic charges (by empirical methods).

The molecular representations can be altered. Newman projections may be drawn down any bond; best-plane views may be calculated; the molecule(s) may be rotated, translated, scaled and clipped severally or individually in three dimensions; the default picture can be changed from coloured-stick type to space filling, to red–green stereo, to α-carbon backbone, to dot surfaces coloured by atom type or other properties; atom or amino-acid labels, hydrogen atoms, selected amino acids, and the spheres of ball and stick models, may be rendered invisible or restored instantaneously at will; various features of molecules may be highlighted such as atomic type, charge, electronegativity and non-bonded contacts, bond order, hydrogen bonds, specific amino acids, all amino acids of a particular type, all amino acids within a sphere of given radius form the target residue, α-helices, β-sheets, and turn structures; and pictures may be composed specifically for slide making so that, for instance, part of a protein could be drawn as α-carbon backbone, part as coloured stick, and part as C.P.K. model. The picture could be bounded by a perspective grid, pointers could indicate key molecular features and the picture could be annotated by using several different text fonts.

The range of computational procedures available is wide. It is possible to perform molecular mechanics calculations; molecular dynamics calculations; molecular orbital calculations; rigid and flexible molecule superposition; restricted-energy-input superposition; Ramachandran maps, with coloured contours drawn on screen; a range of conformational search procedures including Monte Carlo and SITAR (sequential, iterative, torsion-angle refinement), and a search procedure specific for ring structures; manual, interactive docking with real-time energy and non-bonded contact monitoring; and automatic docking.

Finally, there are facilities for protein secondary structure prediction and sequence alignment of homologous or partly homologous proteins.

The above list is not exhaustive but gives the flavour of facilities offered. COGS is almost entirely mouse-driven, with occasional use of potentiometer dials.

Protein model building

Before any protein model building from sequence homology the amino-acid sequences of the two proteins must be optimally aligned. In a number of instances the alignment will already have been performed and the results available in the scientific literature. If this is not so, the molecular modelling program provides an option for sequence alignment based on the algorithms of Needleman & Wunsch (1970), Smith & Waterman (1981) and McLachlan (1971).

Amino acids to be operated upon are selected by positioning the cursor, on the colour graphics screen, over any atom in the residue by manipulating a hand-held 'mouse' and then pressing any button on the mouse (this operation is subsequently referred to as 'pointing'). Pointing to a blank area on the screen is referred to as 'hitting blank space'. The user usually only works upon a selected portion of a protein at any one time and the molecular modelling program provides facilities for panning over parts of a magnified image; for interactively selecting the residues to be displayed by turning a dial which steps along the polypeptide chain sequentially, when residues remain visible only if selected by pressing the appropriate button on the potentiometer box; for displaying only those residues contained within a three-dimensional clip box whose dimensions and position on the screen are dynamically variable; and for displaying only those residues within a sphere of selected radius centred on the residue of interest. The entire protein remains permanently within the workspace during these operations; but only the selected atoms are visible and are scaled up to fill the whole screen. The entire protein may be redisplayed instantaneously at any time without the necessity of reading from a disc file.

The protein model building section of the molecular modelling program contains three major subsections which execute the substitution, deletion and insertion functions. These functions are described separately.

Substitution

The user points to any atom in the residue to be substituted and the system will respond by drawing a menu of amino acids down the left-hand side of the screen. Pointing to the desired acid will cause it to be substituted for the original residue in the protein. The proposed substitution will be checked against a table of amino-acid substitution frequencies (McLachlan 1971) stored by the program and an informatory message issued if the substitution is an infrequent one.

Before performing the operation, the program will check to see that the ϕ and ψ torsion angles at the point of substitution in the X-ray protein structure are reasonably close ($\pm 30°$) to allowed values for the new residue. This is achieved by calculating the ϕ and ψ torsion angles for the original residue and comparing these with an internal table of low-energy conformations for the new residue (Lewis *et al.* 1973). A check is also made to ensure that helix breakers are not substituted into helical regions of the protein, that residues that prefer not to form part of β-structures are not included in such, and that charged residues are not inserted into the hydrophobic core of the protein. These latter tests are also made via reference to tables of secondary structural preferences of the various amino acids, stored by the program (Chou & Fasman 1978). The secondary structural features of the original protein are defined either by reading them in from the Brookhaven Data Bank file, which is the starting point for all model-building exercises, or the program will attempt to locate them from tables of torsion angles describing the various features.

If any of these rules is violated, a warning message results together with a chance to retract. However, the user may, if desired, proceed with an unusual substitution.

If everything is correct, or if the user wishes to proceed, the initial substitution will be made with the side-chain torsion angles as close as possible to those of the original residue. The program measures the χ torsion angles for the original amino acid and folds the side chain of the new residue up into the same conformation as far as possible.

If the new side chain is shorter than the old then there is no problem, unless one side chain has trigonal planar atoms in a different position to those of the other. The conformational preferences of bonds adjacent to trigonal atoms are different to those adjacent to tetrahedral atoms. If this complicating situation does arise then the new χ torsion angle is set to that local torsional potential energy minimum ($\pm 60°$, $180°$ adjacent to tetrahedral atoms or $\pm 120°$, $0°$ to trigonal atoms) which places the new side-chain atoms closest to the positions occupied by the corresponding atoms in the original amino acid.

If the new side chain is longer than the old, then some χ torsion angles will remain undefined after the procedure described above has been executed. The values of the undefined torsion angles are optimized by a simple grid search procedure with a torsion angle increment of 30°. The function minimized is the sum of the van der Waals energy involving the non-bonded contacts within 4.0 Å† of the atoms comprising the undefined torsion angles, and the electrostatic energy involving the same atoms (this also accounts for the hydrogen bonds, as described below).

If the new side chain contains a potential hydrogen-bond donor or acceptor then a search is made for complementary atoms in spatially adjoining residues. This is achieved in a similar manner to the optimization of undefined χ torsion angles as described above. As the location of the (possibly) new hydrogen bonds is not known in advance it is difficult to use special functions for the hydrogen-bonding energy (Weiner *et al.*), so the van der Waals energy and electrostatic energy terms for potential hydrogen-bond donor or acceptor atoms are scaled to produce a pairwise atom energy minimum at the correct separation for a hydrogen bond. When the lowest energy sub-conformation has been located by grid search the D—H···A and H···A—X angles are checked. If the values fall within the acceptable ranges then the side-chain conformation is accepted; if not the D—H···A—X geometry (or geometries) is checked at the next to lowest energy minimum, and so on. If no acceptable hydrogen-bonding scheme can be found, the side-chain conformation is set to that corresponding to lowest energy and a warning message is issued.

The optimization of undefined χ torsion angles and the location of acceptable hydrogen-bonding schemes is performed by default, although both options may be disabled if desired. Similarly, upon entering the protein model-building option of the molecular modelling system the user is given the choice of automatically optimizing all the side-chain torsion angles of newly substituted residues, or leaving them in the positions closest to the original values for the unsubstituted protein.

Three options to optimize the positions of all the side-chain atoms, and optionally neighbouring main-chain atoms, are offered. The first is to perform a grid search, as described above, over all of the side-chain torsion angles with fixed bond lengths and angles.

The second is to perform a SITAR scan on the χ torsion angles to find the most favourable values, while keeping the side-chain bond lengths and angles fixed. The SITAR procedure follows valleys in the potential hypersurface to their floor. A scan around χ_1 at 10° intervals with fixed bond lengths and angles is used to locate the lowest energy positions of the side-chain atoms in the force field defined by their less-than 4 Å neighbours. χ_1 is then left in this position

† 1 Å = 10^{-1} nm = 10^{-10} m.

and χ_2 adjusted similarly, and so on, until all the χs have been dealt with. Now the changes in subsequent χs will have altered the minimum energy value of χ_1, so this torsion angle is rescanned and then centred on the new minimum energy position. All of the other χs are improved in an identical manner, and the whole process iteratively repeated until the energy falls by less than some user-specified limit from one iteration to the next.

SITAR and grid-search procedures are useful where the new side chain is no larger than the previous one, or in a sterically uncrowded environment such as the protein surface. If the new side chain is larger than the old, and too close to neighbouring residues, then the following option must be used.

For difficult cases a full relaxation Newton–Raphson optimization of the Cartesian coordinates is possible of the atoms comprising the substituted residue, its immediate chain neighbours, and any residues that are in too-close steric contact with these. Non-bonded forces arising from atoms more than 4.0 Å distant from the selected atoms are ignored. This method undoubtedly produces the most accurate energy minima but is more time consuming than the previous two procedures. The minimizations at this point in the program are performed by using the united-atom approach (Gelin & Karplus 1979), which prevents the protein from collapsing (Wodak *et al.* 1985) while conserving computer time. The force field, which is based upon a mini-expert system rather than the usual huge list of individual force constants (White 1985), may be switched at will between united-atom and all-atom modes of operation at those junctures where the program allows the choice (or at any point in the separate molecular mechanics option of the molecular modelling system).

Deletion

To use this option any atom in the residue to be deleted is pointed to with the puck and cursor and, upon selecting a null atom (by hitting blank space), the deletion sequence begins. Again the system will check as far as possible to prevent peculiar operations from being executed (for example, attempts to delete residues that are not in surface loops will be queried; surface residues are crudely located by calculating the residue distance from the centre of gravity of the protein and assigning the residue as a surface variety if this distance is more than a certain fraction of the distance from the furthest residue to the centroid). After the atoms defining the residue to be deleted have been removed from the workspace, the polypeptide chain is rejoined with a very long bond (spanning the deleted residue) and a pattern search optimization calculation initiated to close the gap. This is achieved by designating a 'molten zone' (Diamond 1981) for a few residues on either side of the gap (four on either side by default, but user-selectable if desired) and allowing the positions of the atoms in these residues to refine in the force field defined by their within-4.0 Å neighbours, in a manner similar to the substitute operation.

Experience has shown that a molten zone of two residues on either side of the point of deletion is probably sufficient. This is borne out by the available experimental evidence, which shows the main-chain deformation owing to deletions and insertions in one protein relative to its homologue to be incredibly localized.

Pattern-search optimization in Cartesian coordinates is chosen as the first-stage gap-closing minimizer because it is much more efficient than gradient methods, such as the Newton–Raphson, at correcting gross errors in a structure (such as the 2–3 Å bond length introduced by deleting an amino-acid residue); and because it will guarantee to locate an energy minimum, where gradient methods locate turning points on the potential energy hypersurface which may be energy minima or maxima. However, once the gap left by the deleted residue has been reasonably closed by the pattern-search minimizer, a change is made to a block diagonal Newton–Raphson procedure operating on atomic Cartesian coordinates. This is because the pattern-search procedure will not properly adjust the overall rotation and translation of the molten zone relative to other spatially adjacent strands of polypeptide chain unless these parameters are explicitly included in the optimization procedure; this is too messy to contemplate seriously. The Newton–Raphson family of optimizers do not suffer from this drawback (because they minimize force rather than energy) and are very efficient at energy-minimizing structures of the quality presented to them after pattern-search minimization. The pattern-search minimizer has also placed the conformation of the molecule at a point somewhere in a well corresponding to an energy minimum; so that the user does not have to worry about the propensity of Newton–Raphson procedures to converge upon energy maxima if the starting conformation is closer to a maximum than an energy minimum.

For the present, a string of consecutive amino-acid deletions is handled as a sequence of single-residue deletions, where the user has to select the next residue to be deleted after the previous one has been removed.

Insertion

To insert residues, the user first points to any atom in the residue after which the insertion is to be made and to the name of the residue to be inserted on the menu of amino acids which will appear after the point of insertion has been successfully selected.

A range of checks, similar to those performed when substituting amino acids, are performed before any operation is started and the usual chance to retract given in the event of ambiguities. If everything is in order then the chain will be broken at the point of insertion, and the new residue inserted (at this stage directly on top of the residue after the point of insertion). The ϕ, ψ torsion angles of the inserted residue are set to values corresponding to the A, C, D, E, F, G, A* regions of a Ramachandran map (Zimmerman *et al.*) in turn, and the ϕ, ψ torsion angles of a nine-residue molten zone (four on each side of the inserted residue) are optimized by constrained pattern search, operating in the torsion-angle subspace of the internal coordinates to give the best geometry and energy in the region of the chain break consistent with minimum change in the position of the existing residues. The local conformational changes necessitated by inserting a residue are even more traumatic for an energy minimizer than those required upon deleting a residue. To maintain efficiency, an even more robust procedure than Cartesian coordinate-optimizing pattern search is desirable. As the changes required to accommodate the new residue are primarily restricted to a number of torsion angles it is most effective to optimize the new conformation in terms of torsion angles with fixed bond lengths and angles.

The results of the seven trials are then compared and the molten-zone conformation set to correspond to the most favourable result. After insertion, Cartesian coordinate Newton–Raphson energy minimization is performed to idealize the geometries of the new residues. (Notice that these optimizations are all very fast because only a few iterations are performed on a small number of atoms and their short non-bonded interactions.) The energy minimization options are the same as those available at the end of a substitution operation.

Completion

In practice, the model building is usually done in several passes through the protein. The substitutions are done by first using the SITAR option to position the new side chains. A note is made of any residues where the new side chain is too tight a steric fit to be dealt with by the SITAR procedure, and these are fixed up on a second pass by using the Newton–Raphson minimization option.

The deletions are next taken care of, and several operations may be required for multi-residue deletions. As deleting residues creates space there will not usually be a problem with steric crowding, but the user should check that any hydrogen-bonding pattern, salt bridge, disulphide bond, or ordered water-molecule interaction with residues adjacent to the point of deletion are either preserved or replaced with equivalent or more beneficial interactions.

Insertions are usually performed last, and single-residue operations are subject to all of the caveats pertaining to substitution or deletion. Multi-residue insertions may be handled one residue at a time in the same way as multi-residue deletions. Alternatively, a template may be taken of the conformation of a length of chain in the original protein centred on the point of insertion and corresponding in length to the length of the insertion. The polypeptide chain is then split at the point of insertion and the regular energy-minimization option (outside the protein model-building suite) used to perform a constrained minimization which opens up a gap sufficiently wide to accommodate the template. The ends of the template are then joined to the split ends of the main chain and an energy-minimization calculation used to tidy up the modified length of chain.

To be certain that the effect of the substitutions, deletions and insertions upon the overall protein conformation has not been artificially localized, Cartesian coordinate Newton–Raphson (constrained) energy minimization over the whole of the protein, or of large pieces (such as the non-conserved portions of the polypeptide chain), could be performed at the end of a sequence of substitutions, insertions, and deletions. This could take several hours for a large protein on a conventional minicomputer; a point that is not lost on the manufacturers of array processors!

A MODEL OF HUMAN THROMBIN

Prothrombin is a protein that participates in the final phase of blood coagulation. It is synthesized in the liver and secreted into the blood, where it participates in the clotting cascade. Prothrombin is then converted to α-thrombin by limited proteolytic excision. Thrombin converts fibrinogen into an insoluble fibrin clot, which is further stabilized by the introduction of covalent crosslinks between neighbouring units in the polymer. These fibrin clots are ultimately degraded by plasmin (Jackson & Nemerson 1980).

A number of the blood coagulation enzymes, including thrombin, are known to possess a degree of sequence homology with the pancreatic serine proteases, trypsin and chymotrypsin. The three-dimensional structures of these digestive proteases have been determined by X-ray crystallography (Maquart *et al.* 1983; Birktoft & Blow 1972), and this information has been used to build a fairly crude three-dimensional model of bovine thrombin (Furie *et al.* 1982). The sequence of human thrombin has been determined both by amino-acid sequence analysis (Seegers 1979) and from the cDNA (c represents complementary) coding (Friezner-Degen *et al.* 1983). A partial sequence has also been obtained from the human genomic DNA

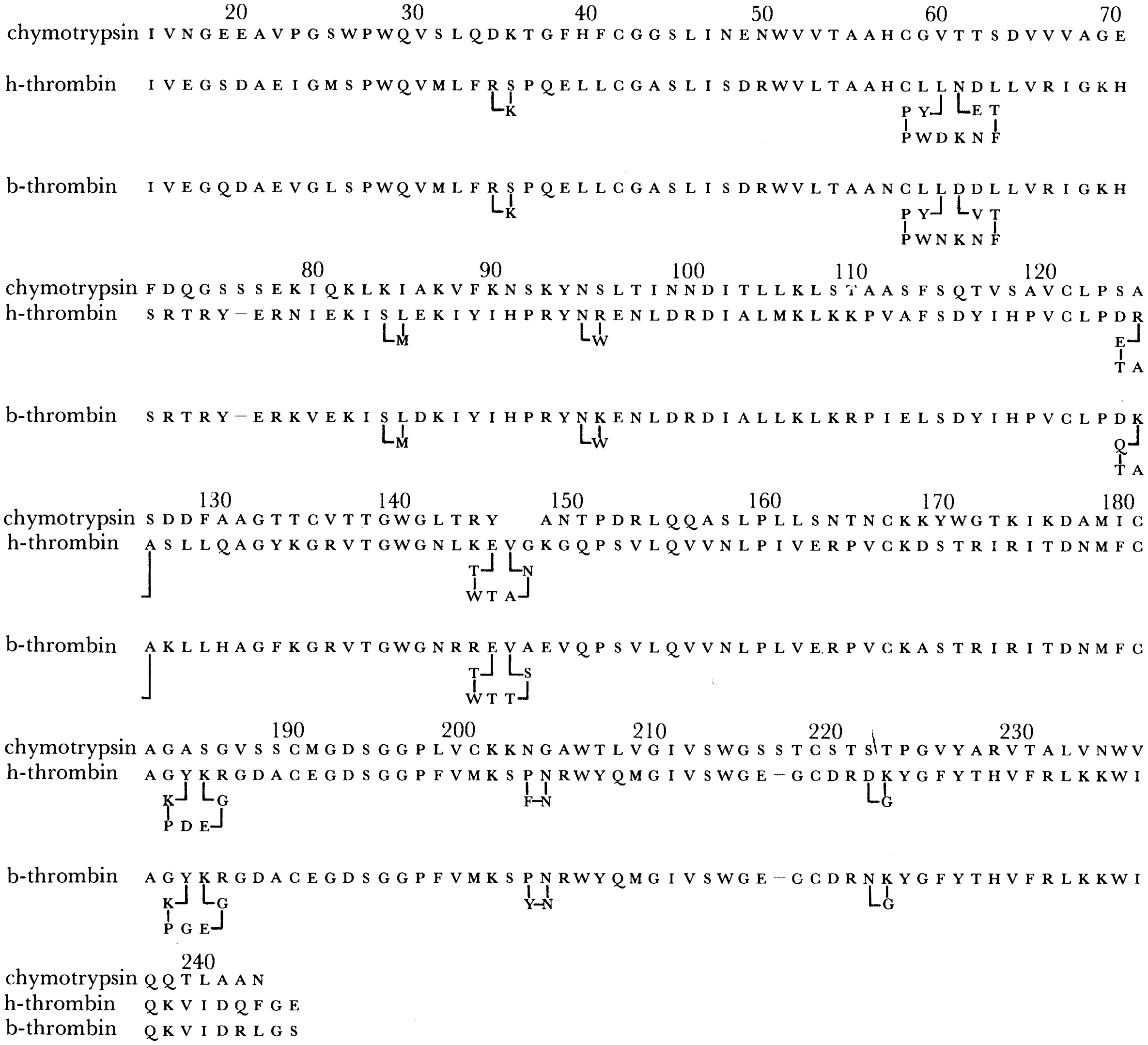
20 30 40 50 60 70
chymotrypsin I V N G E E A V P G S W P W Q V S L Q D K T G F H F C G G S L I N E N W V V T A A H C G V T T S D V V V A G E
h-thrombin I V E G S D A E I G M S P W Q V M L F R S P Q E L L C G A S L I S D R W V L T A A H C L L N D L L V R I G K H
K
P Y E T
P W D K N F
b-thrombin I V E G Q D A E V G L S P W Q V M L F R S P Q E L L C G A S L I S D R W V L T A A N C L L D D L L V R I G K H
K
P Y V T
P W N K N F
80 90 100 110 120
chymotrypsin F D Q G S S S E K I Q K L K I A K V F K N S K Y N S L T I N N D I T L L K L S T A A S F S Q T V S A V C L P S A
h-thrombin S R T R Y – E R N I E K I S L E K I Y I H P R Y N R E N L D R D I A L M K L K K P V A F S D Y I H P V C L P D R
M W E
T A
b-thrombin S R T R Y – E R K V E K I S L D K I Y I H P R Y N K E N L D R D I A L L K L K R P I E L S D Y I H P V C L P D K
M W Q
T A
130 140 150 160 170 180
chymotrypsin S D D F A A G T T C V T T G W G L T R Y A N T P D R L Q Q A S L P L L S N T N C K K Y W G T K I K D A M I C
h-thrombin A S L L Q A G Y K G R V T G W G N L K E V G K G Q P S V L Q V V N L P I V E R P V C K D S T R I R I T D N M F C
T N
W T A
b-thrombin A K L L H A G F K G R V T G W G N R R E V A E V Q P S V L Q V V N L P L V E R P V C K A S T R I R I T D N M F C
T S
W T T
190 200 210 220 230
chymotrypsin A G A S G V S S C M G D S G G P L V C K K N G A W T L V G I V S W G S S T C S T S T P G V Y A R V T A L V N W V
h-thrombin A G Y K R G D A C E G D S G G P F V M K S P N R W Y Q M G I V S W G E – G C D R D K Y G F Y T H V F R L K K W I
K G F N G
P D E
b-thrombin A G Y K R G D A C E G D S G G P F V M K S P N R W Y Q M G I V S W G E – G C D R N K Y G F Y T H V F R L K K W I
K G Y N G
P G E
240
chymotrypsin Q Q T L A A N
h-thrombin Q K V I D Q F G E
b-thrombin Q K V I D R L G S

FIGURE 1. Sequence alignments of human (h-) and bovine (b-) thrombin to chymotrypsin.

(Friezner-Degen *et al.* 1983) which agrees, as far as it goes, with the cDNA sequence. There are, however, a number of differences between the amino-acid analysis sequence and the cDNA sequence and we have chosen to use the cDNA variant with the α-chymotrypsin X-ray structure for this work.

The building of the human thrombin model from the chymotrypsin X-ray structure was fairly straightforward. The substitutions were easily accommodated with only very minor alterations in the main-chain conformation. There are only two single-residue deletions, Ser-76 and Ser-218, and these posed no problems. Similarly, the single-residue insertions Lys-35A, Met-84A, Trp-95A, and Gly-223A were easily accomplished. The two- and three-residue insertions Glu-126A–Ala-126C and Phe-204A–Asp-204B were only marginally more difficult; but the multi-residue insertions Tyr-60A–Glu-60J, Thr-146A–Asp-146E, and Lys-185A–Gly-185E presented more of a problem. These latter insertions have been dealt with as a sequence of consecutive single-residue insertions. All the insertions and deletions were situated on surface loops of the chymotrypsin structure, as expected.

The main chain of human thrombin is 259 amino acids long, the same length as bovine thrombin, with which it is 86.2 % homologous. The main chains of human and bovine thrombin align with no insertions or deletions but with 34 substitutions, all of which lie close to the surface of the protein. This observation lends support to the theory that the substrate specificity of these enzymes is determined not only within the active site, but in an extended, surface-binding site that surrounds the active site (Furie *et al.* 1982). The alignment is shown in figure 1.

The residues in the immediate vicinity of the His-57, Asp-102, Ser-195 catalytic triad (the α-chymotrypsin numbering scheme is used throughout, with insertions denoted as A, B, C, etc., relative to the point of insertion into chymotrypsin) are highly conserved, but the surfaces adjacent to the active site of human thrombin include some differences relative to bovine thrombin. There are three disulphide bridges in bovine and human thrombins, between Cys-42–Cys-58; Cys-168–Cys-182; and Cys-191–Cys-220. The chymotrypsin disulphide bridge between Cys-1–Cys-122 has no analogue in thrombin because Cys-1 is not present; and the Cys-136–Cys-201 bridge is also absent in thrombin because residues 136 and 201 are glycine and methionine respectively.

Two of the intron–exon junctions in the partial human gene sequence correspond to the junctions between residues 74–75, and 128–129, both of which lie on surface loops of human thrombin and are flanked by highly variable amino-acid sequences in the eukaryotic homologues. Furthermore, the 74–75 junction occurs within a twelve- to seventeen-residue insertion when compared with the prokaryotic serine proteases, and a three-residue Glu–Thr–Ala insertion occurs immediately adjacent to the 128–129 junction. These two junctions are analogous to those described for human renin (Sibanda *et al.* 1984) and accord with the proposed sliding-junction model of gene family evolution (Craik *et al.* 1983). The third intron–exon junction lies between residues 29–30, which is a highly conserved region in the interior of the eukaryotic enzymes; it does, however, correspond to a nine-residue addition to the eukaryotic serine protease polypeptide chain relative to the prokaryotic proteases. Elastase and chymotrypsin also have intron–exon junctions in the same region at 30–31 and 34–35. The eukaryotic serine proteases have a number of other common junction positions and the tally so far is; 29–35(3), 61(3), 74(1), 87(1), 100(1), 128(1), 146–150(3), 192(3), where the figure in brackets is the number of occurrences. These common junction positions may also be consistent with the 'genes in pieces' hypothesis (Sudhof *et al.* 1985; Gilbert 1985).

Conclusions

The algorithms used to construct the above model are still in a state of flux and so our human thrombin model must be regarded as preliminary. However, we believe that the quality of our preliminary model is much higher than that of models constructed by using computer graphics as electronic Dreiding models. This model will now be used to attempt the rational design of anti-clotting compounds.

References

Birktoft, J. J. & Blow, D. M. 1972 *J. molec. Biol.* **68**, 187.

Chou, P. Y. & Fasman, G. D. 1978 *A. Rev. Biochem.* **47**, 251.

Craik, C. S., Rutter, W. J. & Fletterick, R. 1983 *Science, Wash.* **220**, 1125.

Diamond, R. 1981 In *Biomolecular structure, function, conformation and evolution*, vol. 1 (ed. R. Srinivasan), p. 567. Oxford: Pergamon Press.

Friezner-Degen, S. J., MacGillivray, R. T. A. & Davie, E. W. 1983 *Biochemistry* **22**, 2087.

Furie, B., Bing, D. H., Feldman, R. J., Robison, D. H., Burnier, J. P. & Furie, B. C. 1982 *J. biol. Chem.* **257**, 3875.
Gelin, B. R. & Karplus, M. 1979 *Biochemistry* **18**, 1256.
Gilbert, W. 1985 *Science, Wash.* **228**, 823.
Jackson, C. M. & Nemerson, Y. 1980 *A. Rev. Biochem.* **49**, 765.
Lewis, P. N., Momany, F. A. & Scheraga, H. A. 1973 *Israel J. Chem.* **11**, 121.
McLachlan, A. D. 1971 *J. molec. Biol.* **61**, 409.
Maquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. 1983 *Acta crystallogr.* B **38**, 480.
Needleman, S. B. & Wunsch, C. D. 1970 *J. molec. Biol.* **48**, 443.
Read, R. J., Brayer, G. D., Jurasek, L. & James, M. N. G. 1984 *Biochemistry* **23**, 6570.
Seegers, W. H. 1979 *Prog. chem. Fibrinol. Thrombol.* **4**, 251.
Sibanda, B. L., Blundell, T., Hobart, P. M., Fogliano, M., Bindra, J. B., Dominy, B. W. & Chirgwin, J. M. 1984 *FEBS Lett.* **174**, 102.
Smith, T. F. & Waterman, M. S. 1981 *J. molec. Biol.* **147**, 195.
Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. 1985 *Science, Wash.* **228**, 815.
Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. 1985 *Science, Wash.* **228**, 893.
Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. 1984 *J. Am. chem. Soc.* **106**, 765.
White, D. N. J. 1985 *Computer aided molecular design*, p. 71. (Conference Transcript.) London: Oyez.
Wodak, S. J., Alard, P., Delhaise, P. & Renneboog-Squilbin, C. 1985 *J. molec. Biol.* **181**, 317.
Zimmerman, S. S., Pottle, M. S., Nemethy, G. & Scheraga, H. A. 1977 *Macromolecules* **10**, 1.

Discussion

B. Robson (*Theoretical Biochemistry Laboratory, The Medical School, Manchester*). The successive use of different techniques for energy minimization is, in my opinion, rather like using several different spanners of different sizes and hoping that one will loosen the nut. If the right spanner is not available, we should make one. This is not a criticism in that Dr White has clearly taken a properly cautious approach. I think that combining all the desirable features of the minimization algorithms into a single routine, in which they talk to each other intelligently, is a better direction. The methods are fundamentally not greatly different. For example, I think that the different numbers of working points in conformational space required by SIMPLEX, quadratic interpolation–extrapolation, and gradient methods, could more efficiently be exploited by shifting between sets and subsets of the points according to the status of the minimization and under program control, so that the program itself decides what is best.

I will add the following rider concerning a point which Dr White and other speakers know well, though some present may not. One may design a modified protein, but if it is neglected to look across kinetically facile energy barriers for possible deeper minima, then the protein will, when synthesized, cross to a deeper minimum with associated properties quite different from those calculated. Looking across barriers intelligently for deeper minima, an essence of global minimization, is truly a technique fundamental to the computer-aided design of modified proteins to be produced via site-directed mutagenisis.

D. N. J. White. Both of Dr Robson's points are very important and deserve some comment. As regards minimization procedures there is the practical aspect, and in addition one needs to distinguish minimization algorithms from the space or substance in which they operate.

From a practical point of view, we were concerned to build a comprehensive protein-modelling program in the shortest possible time. We therefore used proven tools which were closest to hand.

As regards minimization procedures we could have used, say, a variable metric gradient algorithm to handle *all* of the Cartesian space optimizations; but at the expense of having to store a large Hessian matrix, with all of the problems this entails. Simplex procedures also operate in both high- and low-gradient regions of conformational hyperspace, at the expense of loss of efficiency proximate to an energy minimum. The choice of a tandem minimizer seemed to us to be the best compromise.

There is no escape from the necessity of having to operate in a full Cartesian or internal coordinate space for the final part of the model-building process, and a reduced torsion-angle subspace earlier on in the proceedings if the program is not to consume an inordinate amount of computer time. Dr Robson's idea of program-controlled full space–subspace exchange, rather than a fixed protocol, is a good one which we will explore further.

Both he and I have expended considerable time and effort searching for the 'holy grail' of an effective multidimensional global minimizer; but he is right: the need to quarter conformational space is not widely appreciated. I also notice that some of the algorithms we developed in the past are being re-invented because people do not read the literature. Finally, I must say that when we developed our GLOMIN algorithm for global energy minimization of cyclic structures the thought that it could be so adapted to model large insertions into protein molecules was far from my mind: in fact, it had not even entered it! It just goes to show that serendipity is not the exclusive preserve of experimentalists.