Modelling Antibody Side Chain Conformations Using Heuristic Database Search

David W. Ritchie[1,2†] **and Graham J.L. Kemp**[1‡]
Departments of [1]Computing Science and [2]Molecular & Cell Biology,
King's College, University of Aberdeen, AB24 3UE, UK.
[†] *dritchie@csd.abdn.ac.uk*, [‡] *gjlk@csd.abdn.ac.uk*

Abstract

We have developed a knowledge-based system which models the side chain conformations of residues in the variable domains of antibody Fv fragments. The system is written in Prolog and uses an object-oriented database of aligned antibody structures in conjunction with a side chain rotamer library. The antibody database provides 3-dimensional *clusters* of side chain conformations which can be copied *en masse* into the model structure. The object-oriented database architecture facilitates a navigational style of database access, necessary to assemble side chains clusters. Around 60% of the model is built using side chain clusters and this eliminates much of the combinatorial complexity associated with many other side chain placement algorithms. Construction and placement of side chain clusters is guided by a heuristic cost function based on a simple model of side chain packing interactions. Even with a simple model, we find that a large proportion of side chain conformations are modelled accurately. We expect our approach could be used with other homologous protein families, in addition to antibodies, both to improve the quality of model structures and to give a "smart start" to the side chain placement problem.

Introduction

Modelling antibody structures is an important exercise because, in the absence of a crystal structure, a good 3-dimensional model can help to explore structural reasons for observed antigen binding affinities.

Antibodies, like many other homologous protein families, have highly conserved secondary structures. This is especially true in the β-sheet framework regions which give the characteristic Greek Key motif of the immunoglobulin fold (Chothia & Lesk 1987). Knowledge of such structural parsimony is often used to initialise a new model structure by selecting the backbone conformation of a known structural homologue. However, if we have *several* known structures for a particular protein family then we should be able to draw upon knowledge of *all* of these structures to build a better model. In order to use this knowledge effectively we need to represent it in a form that provides a convenient way to compare parts of one structure with corresponding parts of other family members.

The main goal of antibody modelling is to predict the loop structure and side chain conformations of the complementarity determining regions (CDRs), since these are involved in antigen binding. Chothia and Lesk (1987) have shown that the CDR loops often adopt a small set of canonical backbone conformations and this behaviour can be exploited to build models with greater accuracy (Martin, Cheetam, & Rees 1989). In this paper, we assume that we have a good model backbone conformation and we focus on solving the resulting "3-dimensional jigsaw puzzle" (Taylor 1992) of assigning side chain conformations to the model structure. Our approach differs from previous side chain placement methods in that we search a database of known structures (Kemp *et al.* 1994) using a simple heuristic cost function to find 3-dimensional clusters of side chains which can be copied to the model. The Kabat sequence alignment (Kabat *et al.* 1992) is used to restrict the search to those residue positions that are conserved in the model. Effectively, much of the model is built from structural units from the database for which evolution has already "solved" the problems of conformational search and energy minimisation. The model is completed and refined with side chain conformations from the database and from a rotamer library using further pairwise heuristics.

Results of testing our approach with a model-built structure of the murine anti-phenylarsonate antibody, 36-71 (PDB code 6FAB), are presented. These results show that around 60% of all side chains are placed as clusters from the database, of which 84% have correctly[1] predicted conformations.

[1] We use Levitt's (1992) criterion of 2Å RMS or less for a correct conformation.

Materials & Methods

Our database currently contains 71 antibody and Bençe-Jones structures taken from the PDB (Bernstein *et al.* 1977) and over 3500 aligned antibody sequences taken from the Kabat data bank (Kabat *et al.* 1992). The database is implemented using the P/FDM object-oriented database management system (Gray *et al.* 1990; Gray, Kulkarni, & Paton 1992) and is based on Shipman's (1981) functional data model (FDM). The FDM is a semantic data model whose basic concepts are *entities* which represent classes of object (e.g. *chains*, *residues* and *atoms*), and *functions* which represent entity attributes and relationships.

A particularly useful relationship function, derived from the Kabat sequence alignment, maps a Kabat position code and a chain identifier onto a residue position. Thus structurally conserved positions can be identified easily.

Both P/FDM and our modelling system are implemented in Prolog. The cost function described below is implemented as a composition of simple object-level database queries with call-outs to compiled Fortran and C routines for computationally intensive calculations. Building the model framework and CDR loop conformations is described elsewhere (Ritchie & Kemp 1997).

Pairwise Interactions

An empirically formulated cost function is used to score the interaction between a pair of residues i and j:

$$F_{ij} = G_{ij} * B_{ij} / P_{ij} \qquad (1)$$

where G_{ij} represents a "packing gain", P_{ij} is a "steric overlap penalty" and B_{ij} is a "bonding bonus". The dimensionless pairwise score, F_{ij}, increases with good volumetric packing and good chemical interactions but decreases in the presence of steric clashes. Each individual term is unity when it has no contribution to make, for example in the case of distantly separated residues.

The extent to which a pair of side chains pack together is estimated by calculating numerically side chain bounding spheres and by finding the ratio of the volume of the separated spheres to that of their union to give a sidechain-sidechain "packing gain", G_{SS}:

$$G_{SS} = (S_i \cup S_j)/(S_i \cup S_j - S_i \cap S_j) \qquad (2)$$

Sidechain-backbone packing is estimated similarly and these quantities are summed to give a pairwise residue packing score:

$$G_{ij} = 1 + (G_{SS} - 1) + (G_{SB} - 1) + (G_{BS} - 1) \quad (3)$$

We also use residue and domain bounding spheres to identify all interacting pairs of neighbouring residues. These are calculated once for each antibody and stored in the database. This hierarchical system of bounding spheres allows all pairwise interactions to be found without the need for coordinate grids.

Steric clashes are modelled by summing pairwise atom overlaps, P_{rs}, to give the "overlap penalty", P_{ij}, for a pair of residues:

$$P_{rs} = \begin{cases} (R_r + R_s)^3 / R_{rs}^3 & \text{if } R_{rs} < R_r + R_s \\ 1 & \text{otherwise} \end{cases} \qquad (4)$$

$$P_{ij} = 1 + \sum_{r \in i} \sum_{s \in j} (P_{rs} - 1) K_{rs} \qquad (5)$$

where R_r and R_s are the effective van der Waals radii of the two atoms, reduced by 10%, and R_{rs} is the interatomic separation of their nuclei. The factor K_{rs} is used as a bias term to give a higher penalty value for sidechain-backbone overlaps than for sidechain-sidechain overlaps. Similarly, pairwise atom "bonding potentials", B_{rs}, are summed to give the bonding term, B_{ij}, for the residue pair:

$$B_{rs} = \begin{cases} (R_r + R_s)^n / R_{rs}^n & \text{if } D_{rs}^{min} < R_{rs} \leq D_{rs}^{max} \\ 1 & \text{otherwise} \end{cases} \qquad (6)$$

$$B_{ij} = 1 + \sum_{r \in i} \sum_{s \in j} (B_{rs} - 1) \qquad (7)$$

where D_{rs}^{min} and D_{rs}^{max} define the distance over which the bonding term is applied. We use 2.45Å to 3.4Å and $n = 4$ for hydrogen bonds and 1.95Å to 2.15Å and $n = 6$ for disulphide bonds.

Side Chain Cluster Placement

We define a *cluster* of neighbouring residues as a group of N or more residues where *(i)* each has a pairwise interaction with at least one other member of the cluster in which the cost function score is greater than some threshold and *(ii)* each has a Kabat position code corresponding to an unplaced position in the model. N is typically from 2 to 4. The cluster score C_k of cluster k is the net sum of the pairwise cost function scores of the cluster members:

$$C_k = 1 + \sum_i \sum_j (F_{ij} - 1) \quad ; \; i < j \in \text{cluster k} \quad (8)$$

This definition of a cluster means that the database search for side chain clusters only needs to examine

Kabat position codes and stored pairwise interaction scores. A recursive search of a list of pairwise interactions for each antibody structure quickly resolves mutual neighbours into a set of distinct clusters.

Each antibody structure in the database can provide several candidate clusters, many of which would overlap if copied directly into the model. Consequently, cluster placement proceeds iteratively. At each iteration a list of candidate clusters is drawn from the database and clusters are ranked according to their number of side chains, their cluster score (Eq. 8), and by how well they fit the model backbone. The side chains of the highest scoring cluster are copied to the model and any remaining clusters that now contain residue positions for which model side chains have been assigned are discarded from the cluster list. This procedure is repeated until the list of clusters is exhausted. The complete procedure is repeated for a further 4 cycles, gradually reducing the threshold score at each cycle. Conceptually, strongly interacting residues are placed first, followed by tightly packing groups, although in practice the behaviour is less clear-cut.

Completing the Model

Since a cluster must contain at least two side chains, the cluster algorithm may leave some unplaced singleton side chains in the model structure. Working outwards from the core region, unassigned model side chain conformations are selected from conserved database positions (using the Kabat sequence alignment), choosing the residue that gives the best cluster score (Eq. 8) with its immediate neighbours in the model environment. Finally, any model side chains that remain unplaced are assigned a conformation from the rotamer library (Ponder & Richards 1987) using a similar scoring scheme. The quality of the model is estimated by evaluating the cost function for all pairs of neighbouring residues in the model. Any side chain that contributes to a bad pairwise score is refined by finding the rotamer substitution that gives the greatest reduction in a refinement score:

$$R_j = \sum_i 1/F_{ij} \quad ; \; i,j \text{ neighbours} \qquad (9)$$

This equation dampens good interactions ($F_{ij} > 1$) and amplifies the contributions from bad steric clashes ($F_{ij} \to 0$) and thus is useful in "de-bumping" the model.

Results

To test our algorithm we randomly selected 6FAB from the P/FDM database and modelled its structure with all 6FAB data excluded from the database. The VH and VL backbones were taken from 2F19, having the highest sequence identity with 6FAB, but the L1 and H3 loops were spliced from 1FGV and 8FAB, respectively, to give loops of the correct length and good sequence identity. 158 residues (69%) were placed as clusters with an average of 4 side chains per cluster. The largest cluster contained 11 side chains. The backbone donor structure, 2F19, contributed 111 side chains from clusters, showing that substantial 3-dimensional regions of the model were copied directly from an existing database structure. 47 side chains were copied as clusters from other structures. A further 47 individual database residues were copied to equivalent model positions, leaving 24 that were placed from the rotamer library (mostly GLY, ALA and PRO). Cluster placement takes 25 minutes (on a Sun Sparcserver 1000) with rotamer refinement taking a further 95 minutes to replace 27 conformations.

The results of this test are tabulated in Table 1, using side chain RMS differences from the known crystal structure. In these figures, GLY is always excluded, but we include the C_β of ALA and the ring atoms of PRO, although these coordinates are essentially fixed. RMS deviations can be slippery quantities (Lee & Subbiah 1991; Levitt 1992) so both all-atom and average side chain RMS errors are given, although we prefer to use the latter values since all-atom values are easily skewed by a few outlying values. We also give the numbers of correctly placed side chains using Levitt's (1992) criterion of 2Å RMS or less for a correct conformation. The final model contains 133 side chains (58%) that derive from clusters, of which 84% are placed correctly. Overall, 74% of all side chains and 58% of the CDR side chains are placed correctly. Most of the error is in the L3 and H3 loops, which is not unexpected since the structure of the H3 loop is notoriously difficult to predict (Martin, Cheetam, & Rees 1989). Analysis of steric clashes in the final model shows only 5 bad pairwise interactions (4 of which involve PRO) and no non-bonding contact is closer than 2.5Å.

Table 1: 6FAB model average and (all-atom) side chain RMS errors (Å) and fraction correctly placed (n/N).

	VH domain			VL domain		
Side Chain	Ave.	All	n/N	Ave.	All	n/N
all	1.38	(2.06)	82/107	1.31	(2.20)	70/98
framework	1.35	(2.02)	73/91	1.42	(1.95)	61/83
all CDRs	1.60	(2.23)	9/16	2.02	(3.24)	9/15
CDR 1	0.83	(0.89)	5/6	1.36	(1.74)	5/7
CDR 2	1.68	(2.04)	1/2	2.67	(3.44)	2/3
CDR 3	2.16	(2.73)	3/8	2.56	(4.63)	2/5

Discussion

Our cluster placement algorithm makes intelligent use of existing structural knowledge to reduce dramatically the combinatorial complexity of the conformational search space. Around 60% of model antibody side chains are placed as clusters in only about 20% of the total execution time, of which 80% are within 2Å of the crystal structure. Hence around 50% of the model structure is correctly constructed from database side chain clusters. As might be expected, our method is most accurate when modelling the conserved framework side chains (77% correct), but we still achieve an encouraging figure of 58% correct for the CDR conformations.

Although the rotamer refinement algorithm eliminates bad steric clashes it is less successful at driving the model towards the desired solution. This could probably be improved by using a more realistic soft potential. Additionally, our system contains several distance thresholds and clustering parameters which we have done little to optimise. Thus, we believe that further investigation would enable us to improve upon the encouraging results presented here.

The modelling approach described here depends crucially on the availability of a structural database in which the residues at equivalent 3-dimensional positions in different structures have been identified. In extending this approach to other families of homologous structures, it would be useful to generate structural alignments automatically, e.g. (Russell & Barton 1992), and to adopt numbering conventions for structurally equivalent positions.

Conclusions

Our knowledge-based approach demonstrates how in a homologous protein family entire 3-dimensional regions of known database structures can be copied *en masse* to a model structure. Around 50% of the side chains in the model are placed rapidly and accurately from these clusters, thus providing a "smart start" to the side chain placement problem.

As further members of each structural family are determined, the knowledge derived from structure comparisons and alignments will be of increasing value in homology modelling. However, it is important that this knowledge is represented and stored in a systematic way. We believe that a database architecture like ours, that allows search to be combined with computation, can help us exploit this knowledge effectively.

Acknowledgements. The P/FDM database of antibody sequences and structures was constructed in an earlier project, funded by the SERC. DWR is supported by an award from the BBSRC.

References

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Mayer Jr., E. F.; Bryce, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Chothia, C., and Lesk, A. M. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196:901–917.

Gray, P. M. D.; Paton, N. W.; Kemp, G. J. L.; and Fothergill, J. E. 1990. An object-oriented database for protein structure analysis. *Protein Engineering* 3:235–243.

Gray, P. M. D.; Kulkarni, K. G.; and Paton, N. W. 1992. *Object Oriented Databases: A Semantic Data Model approach.* New York: Prentice Hall.

Kabat, E. A.; Wu, T. T.; Perry, H. M.; Gottesman, K. S.; and Foeller, C. 1992. *Sequences of Proteins of Immunological Interest.* Washington D.C.: Public Health Service, NIH, 5th edition.

Kemp, G. J. L.; Jiao, Z.; Gray, P. M. D.; and Fothergill, J. E. 1994. Combining computation with database access in biomolecular computing. In Litwin, W., and Risch, T., eds., *Applications of Databases: Proceedings of the First International Conference, ADB-94*, 317–335. New York: Springer-Verlag.

Lee, C., and Subbiah, S. 1991. Prediction of protein sidechain conformation by packing optimization. *J. Mol. Biol.* 217:373–388.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226:507–533.

Martin, A. C. R.; Cheetam, J. C.; and Rees, A. R. 1989. Modeling antibody hypervariable loops: A combined algorithm. *Proc. Natl. Acad. Sci.* 86:9268–9272.

Ponder, J. W., and Richards, F. M. 1987. Tertiary templates for proteins – use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.

Ritchie, D. W., and Kemp, G. J. L. 1997. Using an object-oriented database to model antibody Fv fragments. Technical Report TR9702, University of Aberdeen, Computing Science Dept. http://www.csd.abdn.ac.uk/publications/.

Russell, R. B., and Barton, G. J. 1992. Multiple sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins: Struct. Func. Genet.* 14:309–323.

Taylor, W. 1992. New paths from dead ends. *Nature* 356:478–480.