Protein Engineering vol.8 no.4 pp.389–395, 1995

Prediction of hypervariable CDR-H3 loop structures in antibodies

Martin Reczko[1], Andrew C.R.Martin[2], Henrik Bohr[3] and Sándor Suhai

Molecular Biophysics Department, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany, [2]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK and [3]Center for Biological Sequence Analysis, Building 206, The Technical University of Denmark, DK-2800 Lyngby, Denmark

[1]To whom correspondence should be addressed

The structure of the most variable antibody hypervariable loop, CDR-H3, has been predicted from amino acid sequence alone. In contrast to other approaches predictions are made for loop lengths up to 17 residues. The predictions have been achieved using artificial neural networks which are trained on a large set of loops from the Brookhaven Protein Databank which have structures similar to CDR-H3. The loop structures are described by the two backbone dihedral angles ϕ and ψ for each residue. For 21 CDR-H3 loops unique to the neural network, the prediction of dihedral angles leads to an average root mean square deviation in the Cartesian coordinates of 2.65 Å. The present method, when combined with existing modelling protocols, provides an important addition to the structural prediction of the complementarity determining regions of antibodies.

Key words: antibodies/hypervariable regions/neural networks/structure prediction

Introduction

Antibodies represent a class of proteins which have substantial clinical and research potential (Lerner and Benkovic, 1988; Verhoeyen *et al.*, 1988). Naturally, they fulfil a dual rôle; they are capable of binding to an almost unifinitely wide range of antigens while maintaining constant effector functions such as the triggering of the complement system (Reid, 1986). The Fv or 'variable' fragment of the antibody embodies its antigen binding function. It consists of two domains forming a highly conserved framework which supports six 'hypervariable' loops, three from the light chain and three from the heavy chain (Alzari *et al.*, 1988).

The high degree of conservation of the framework across all antibodies from a wide range of species means that the problem of modelling the 3-D structure from sequence alone is essentially limited to modelling the six hypervariable loops or 'complementarity determining regions' (CDRs). Although described as hypervariable (Kabat *et al.*, 1987, 1991), five out of the six loops have sequences and conformations which generally fall into a small number of structural 'canonical' classes (Chothia *et al.*, 1989, 1992). The remaining loop, which is the third loop of the heavy chain (CDR-H3), is substantially more variable in both sequence and length (from three to 27 residues) and no such canonical classes have so far been defined. The additional variability of this CDR results from the genetic splicing of a diversity segment (Schilling *et al.*, 1980) in the genome. The position of this segment corresponds to CDR-H3. Variability is introduced by having ~50 diversity segments which undergo variable and imprecise recombination allowing variation in length and frameshift mutations (Darsley and Rees, 1985) to occur.

Methods used thus far to model antibody loops fall into two main classes: knowledge-based and *ab initio* approaches. The simplest of the knowledge-based methods involves selecting the most similar loop from the antibody structural database and replacing side chains by using a maximum overlap approach (de la Paz *et al.*, 1986; Smith-Gill *et al.*, 1987). A more detailed analysis has been performed by Chothia *et al.* (1989, 1992) who defined key residues in the CDRs and interacting regions of the framework which define canonical structural classes. The *ab initio* approaches include molecular dynamics methods (Fine *et al.*, 1986) and full conformational searching (Moult and James, 1986; Bruccoleri and Karplus, 1987). More recently Martin *et al.* (1989, 1991) and Mas *et al.* (1992) have attempted to combine both knowledge-based and *ab initio* methods.

The approaches which have been developed so far, are generally very effective for the five more-conserved loops, typically obtaining root mean square (r.m.s.) deviations of around 2 Å on the backbone. The combined methods are also able to achieve low r.m.s. deviations for CDR-H3 loops of up to around 10–12 residues. Above this length, it is currently impossible to model CDR-H3 with any degree of confidence using existing approaches.

Materials and methods

The main tools used here in the prediction of the CDR-H3 conformation are artificial neural networks. These can be considered as knowledge-based classifiers. The basic elements of a neural network, the neurons, are processing units which produce output from a characteristic non-linear function of a weighted sum of input data. A neural network is a group of such processing units, the individual members of which can communicate with each other through mutual interconnections. Having been 'trained' by being presented with many pairs of corresponding input and output data, the network will acquire the ability to classify new input data. If a set of input data is denoted by $\{x_j\}$ and the corresponding output is denoted by $\{y_i\}$, the processing of each neuron i in the net can be described as

$$y_i = f(\sum_j W_{ij} x_j + \eta_i) \quad (1)$$

where W_{ij} are the weights of the connections leading to neuron i and η_i and f are the characteristic non-linear function for the neuron. As this equation shows, such a network can be considered as a non-linear mapping between input and output data.

This type of network is chosen as it has previously been

shown to be able to generalize data in molecular biology (Bohr *et al.*, 1988, 1990; Qian and Segnowski, 1988; Holley and Karplus, 1989). In addition, they are rather simple in structure both with respect to the processing of data and training.

The back-propagation error algorithm (Rumelhart *et al.*, 1986) is the most commonly used training procedure and is used here. Training proceeds until a cost function C has reached a local minimum using, for example, a steepest descent minimization protocol. The cost function C is normally written as

$$C = \tfrac{1}{2}\sum_{\alpha,i}(t_i^\alpha - z_i^\alpha)^2 \qquad (2)$$

which is simply the squared sum of errors, t_i being the target value and z_i the actual value of the output neurons.

We have used a specialized feed-forward network topology called a 'time-delay network' (Waibel *et al.*, 1989) which is designed for processing sequence patterns. In time-delay networks, restrictions are placed on the network topology in order to avoid problems such as segmentation errors and overtraining which occur in networks which slide a window along a longer pattern such as an amino acid sequence. In time-delay networks, all hidden units connected to the input layer may only be connected to the limited number of neurons which represent a consecutive pattern of sequence data in the input window. The units connected to the input layer thus have a 'receptive field' which is only sensitive to part of the input window. The weights of the connections in each receptive field are copied to other receptive fields covering all possible parts of the input layer. One receptive field defines a 'feature unit' which is able to detect relevant sequence features independently of their position within the input window using one of these copies. In the case of the loop examples, one could imagine that the receptive field of one feature unit only contains rigid subsections of the loop structure in the feature units. Using the receptive fields of the feature units, the subsequences which these features represent are transformed into a sequence of feature-unit activations. The weights output from the feature-unit neurons are able to describe relative positions of the subsequences in question. There can be several hidden layers with different receptive fields and so the final architecture of the network becomes a hierarachical arrangement of layers with linked receptive fields which integrate complex stepwise information from the input.

Generation of training data

While the antibody structural database is large compared with the number of structures available for the majority of proteins (Orengo *et al.*, 1993), it is still too small to train a neural network effectively, particularly when the variability in the length of CDR-H3 is considered. Figure 1 shows the distribution of these lengths for unique antibody structures in the current release of the protein databank (Bernstein *et al.*, 1977). Duplicate antibodies (i.e. those which are mutants or complexes where the uncomplexed structure is also available) and models are not counted.

The CAMAL method of Martin *et al.* (1989) for modelling antibody loops uses a distance matrix to extract starting conformations for loops from the protein databank. This distance matrix is generated by an analysis of inter-Cα distances in the first few residues of the CDRs in known antibody structures. The protein databank is searched using distance constraints defined as the mean ± 3.5 SD units σ_{n-1}.

In order to generate a large training set for the neural networks, this analysis was performed for the CDR-H3 loops and the protein databank was searched to find a set of loops which match the take-off constraints applicable to the CDR-H3 loops. Sets of loops with lengths between eight and 17 residues were extracted. The distance constraints used are shown in Table I. Using these constraints a set of 1976 loops was extracted from the Brookhaven Databank.

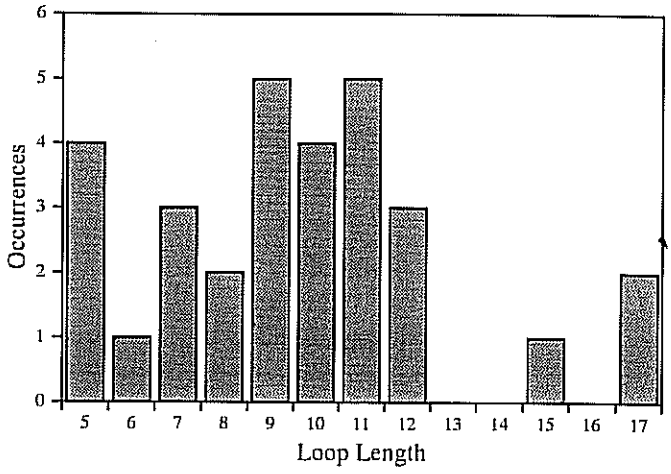

Fig. 1. Number of occurrences of CDR-H3 loops of different lengths in the Brookhaven Protein Databank.

Table I. Distance constraints used to define CDR-H3-like loops

Constraint[a]	Distance (Å)		Constraint	Distance (Å)	
	Minimum	Maximum		Minimum	Maximum
$n < 12$					
DPn	6.029	9.206	DMn	6.029	9.206
DP($n-1$)	4.256	7.624	DM($n-1$)	6.858	10.278
DP($n-2$)	3.538	6.025	DM($n-2$)	9.874	14.108
DP($n-3$)	4.365	7.329	DM($n-3$)	9.496	17.648
DP2	5.194	7.099	DM2	4.234	7.939
DP3	3.590	12.506	DM3	4.024	14.619
$n > 12$					
DPn	7.053	8.210	DMn	7.053	8.210
DP($n-1$)	5.416	6.664	DM($n-1$)	6.510	10.065
DP($n-2$)	4.307	4.975	DM($n-2$)	10.054	13.479
DP($n-3$)	4.530	6.511	DM($n-3$)	11.850	16.226
DP($n-4$)	3.837	8.235	DM($n-4$)	11.706	20.029
DP2	4.929	6.584	DM2	3.714	8.103
DP3	3.041	12.676	DM3	4.070	14.400

[a] n is 1 less than the number of residues in the loop. DPi is the distance between residue 1 and residue i in the loop. DMi is the distance between residue $n+1$ and residue $n-i+1$.

Implementation

The actual neural networks for predicting loop conformation were constructed using the SNNS (Stuttgart Neural Network Simulator) environment (Zell *et al.*, 1991). The networks were trained on a selection of loop structures generated from the Bookhaven Protein Databank as described above. For training, the input data consisted of the amino acid sequences of 8–17 residue loops while the output data were the corresponding backbone ϕ/ψ torsion angles.

While it would be easiest to construct separate training sets of loops with fixed length and thus have a separate network trained on each loop length, there would be problems with the longer loops where there are relatively few structures in the

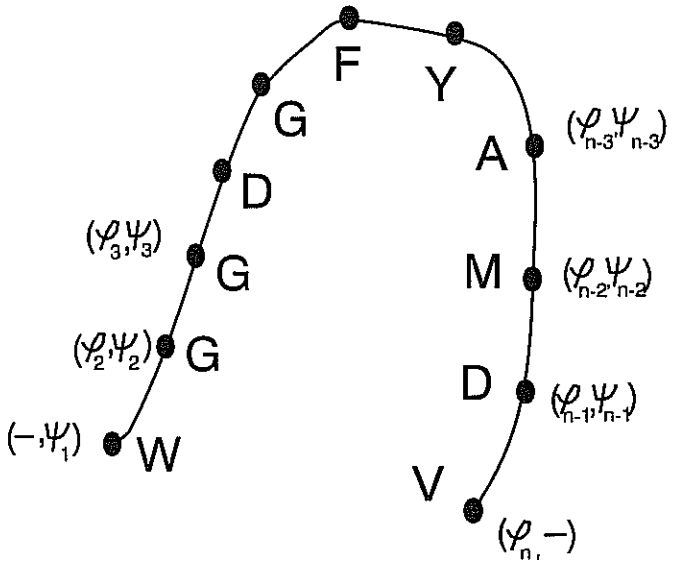

Fig. 2. A schematic H3 loop.

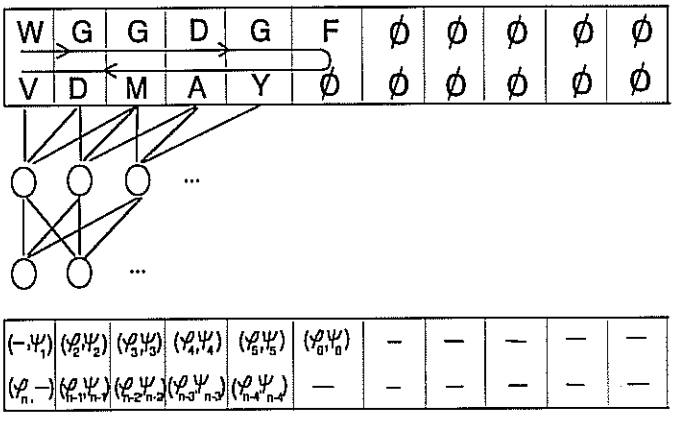

Fig. 3. The sequence and structure coding used by the Time Delay Neural Network.

Table II. Parameters used in the conversion of internal to Cartesian coordinates

Parameter	Value
ω torsion Cα–C–N–Cα	180.0°[a]
Angle Cα–C–N	117.5°
Angle C–N–Cα	120.0°
Angle N–Cα–C	111.6°
Distance Cα–C	1.52 Å
Distance C–N	1.33 Å
Distance N–Cα	1.49 Å

[a]The structure change induced by the occurrence of *cis*-prolines is ignored in this approach. This information could be included separately if a reliable prediction of *cis*- versus *trans*-prolines is available.

protein databank which match the structural constraints for CDR-H3 loops. Thus, an architecture was designed to enable the prediction of loops of different sizes. With a large training set containing short and long loops, the local sequence features determining short conformations can be trained efficiently and may be used for generalization to longer loops.

The transformation of the loop sequence and structure into network activities should have the property that similar structures are represented by similar activity patterns. Hence, the same take-off geometry of loops of different lengths should appear at the same location within the output patterns. This suggests a transformation where the information related to the beginning and the end of loops with different lengths is mapped to the same neurons. A loop sequence such as in Figure 2 is represented by a series of two-element vectors (spinors), as shown in Figure 3, containing pairs of amino acids working from each end of the loop, that is, residue 1 is paired with residue N, residue 2 with residue $(N-1)$, etc. We allow a maximum loop length (N_{max}) of 22 residues. In the case of shorter loops, the right-hand portion of the input and output layers is filled with all-zero patterns acting as 'don't care' symbols so that all loops would be represented by an equal number of spinors. As well as being a convenient representation of the sequence data, this method reflects the topology of the loop.

The conformation of each loop is represented in the output layer of the network which contains $[2\times(N_{max}-1)\times2] = 84$ neuron elements. A pair of neurons represents the sine and cosine values of an individual backbone dihedral angle in the loop. This representation accounts for the continuous cyclic nature of the torsion angle. Using a scalar proportional to the angle would lead to discontinuities between 360 and 0° such that very similar angles are represented by very dissimilar values. The ϕ/ψ torsion angles are not restricted to the allowed regions of the Ramachandran map since these restrictions are implicit in the distribution of the ϕ/ψ angles in the training set and possible outliers in the Ramachandran map need no special consideration. A simple algorithm is used to convert the sin/cos representation to torsion angles and then to Cartesian coordinates. Parameters used in this conversion are shown in Table II. The -1 in the expression above occurs because the first and the last dihedral angle (ϕ_1,ψN_{max}) are not defined without considering atoms outside the loop. For shorter loops, the output neurons in the right-hand part of Figure 3 will be compared with special 'don't care' symbols. If the target value for an output neuron is 'don't care', no error information will be assigned to that neuron. In summary, each input to the network is a set of ($N_{max}\times20$) amino acid or 'don't care' codes which matches an output consisting of a set of 42 dihedral angles.

Results

The training data for the networks consisted of a large set of CDR-H3-like loop structures from the Brookhaven Protein Databank. The values for the accuracy in predicting structures for loops in the training set represent the ability of the network to recall structures it has already learned. The ability of the network to generalize and, therefore, to predict the conformation of loops whose sequences have not previously been seen, is represented by the accuracy of predicting the loop structures from the test set which contains examples that are unique to the network.

In order to be useful as a tool in modelling CDR regions of antibodies, the accuracy in the prediction of loop structures should ideally be ~2 Å r.m.s. deviation or less.

The networks were trained using 1046 out of the total 1976 loops with lengths between eight and 17 residues. These loops were selected at random with the requirement that for each loop length approximately half of the loops are selected for the training set leaving the other half as test cases. Eight actual

Table III. Prediction accuracy for 21 CDR-H3 loops

PDB	Length	RMS (Å)
1BBJ	5	1.26
1GGC	5	1.24
1IND	5	1.48
3HFM	5	1.81
3HFL	7	1.58
4FAB	7	0.97
1BBD	9	2.46
1JHL	9	2.01
2FBJ	9	1.95
7FAB	9	2.83
1DBA	10	2.72
1IGJ	10	2.54
1NBV	10	3.89
1FVD	11	2.19
1NCA	11	3.37
6FAB	12	3.96
8FAB	12	4.16
1IGM	12	2.33
1FAI	15	4.21
1DFB	17	5.51
2FB4	17	3.14
Mean		2.65
SD		1.19

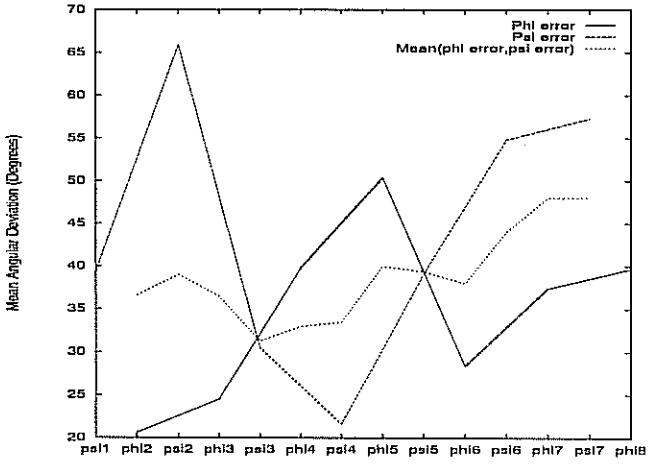

Fig. 4. The absolute angular errors of the ϕ and ψ angles averaged over all test examples of loops of length 8.

CDR-H3 loops (1fdl, 1mam, 1igf, 1fvc, 1hil, 1mcp, 1igm, 2fh4) are contained in the training set of 1046 loops leaving 21 CDR-H3 loops for testing. A total of 20 networks with different architectures and different degrees of adaptation was trained using these patterns. During predictions these networks generate a pool of suggested structures of each loop sequence which are filtered to find structures that match the requirements for fitting the loops onto the antibody framework. The distance constraints shown in Table I are used to assign a structural penalty value P for each loop structure predicted by a network. The value of P is computed as the sum of squared distance penalties defined as

$$P = \sum_{i,j} \begin{cases} (d_{ij} - DX_{ij}^{\max})^2, \text{ if } d_{ij} > DX_{ij}^{\max} \\ (d_{ij} - DX_{ij}^{\min})^2, \text{ if } d_{ij} < DX_{ij}^{\min} \\ 0, \text{ otherwise} \end{cases} \quad (3)$$

where $DX_{ij}^{\max}$ and $DX_{ij}^{\max}$ are the distance constraints DP and DM taken from Table I corresponding to the Cα distances d_{ij}.

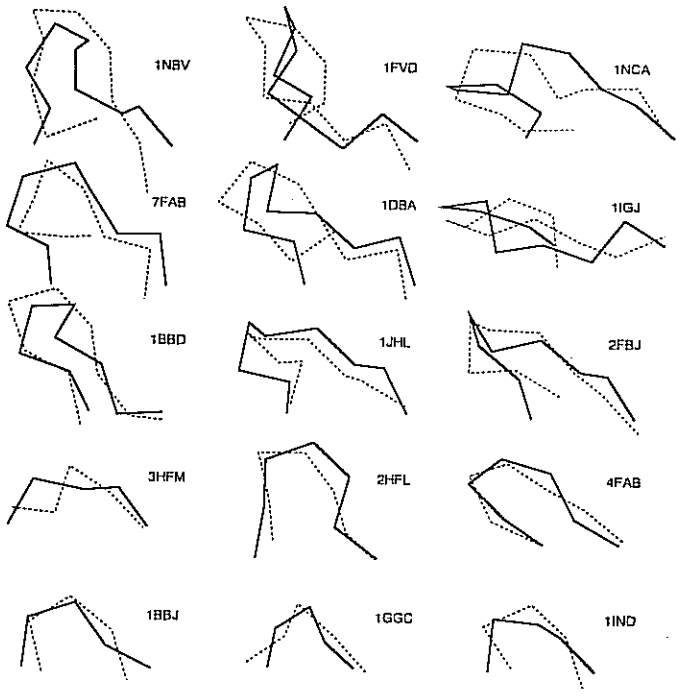

Fig. 5. Predicted conformations for CDR-H3 loops with lengths between 5 and 11 residues superimposed on the crystal structures. The crystal structures are shown in solid lines; the predictions with dashed lines.

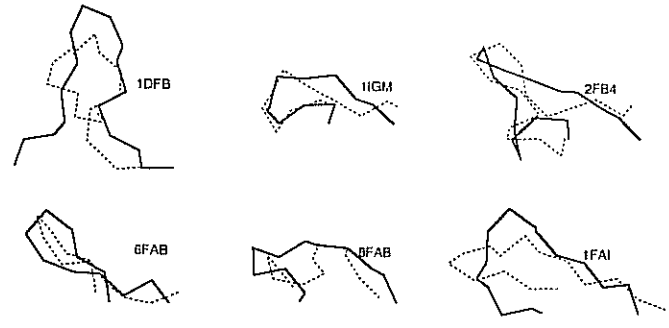

Fig. 6. Predicted conformations of CDR-H3 loops with lengths between 12 and 17 residues superimposed on the crystal structures.

When tested with 930 loops with lengths between eight and 17 residues selected from the Brookhaven Databank according to Table I and not contained in the training set, the average Cα r.m.s. deviation of the predicted loop conformations is 2.12 Å with a standard deviation of 1.12 Å. For a test set containing 21 antibody CDR-H3 regions with lengths between five and 17 residues not contained in the training set the average Cα r.m.s. deviation is 2.65 Å as shown in Table III. As there were no examples below length eight residues during the training of the network, it can be seen from Table III that this approach is able to generalize features detected in a set of loops to the prediction of loops with a different length.

Figure 4 shows the accuracy in prediction of the dihedral angles along a set of eight-residue loops. The absolute angular deviation changes between 40 and 10° along the loop with the accuracy of the ϕ and ψ angles alternating, i.e. when $|\Delta\phi|$ is large $|\Delta\psi|$ is small and vice versa. The average prediction of the middle regions of the loops is better than the ends.

Figure 5 shows the predictions for the test set of 15 CDR-H3 regions superimposed on the crystal structures of the loops.

Table IV. Mean ϕ/ψ angles (°) and standard deviations of each amino acid type in eight-residue loops

Amino acid	Terminal residues[a]				Middle residues				Frequency[b]	
	(ϕ)[c]	SD ϕ[d]	(ψ)	SD ψ	(ϕ)	SD ϕ	(ψ)	SD ψ	Terminal (%)	Middle (%)
Ala	−101.7	36.8	−107.0	93.5	−81.3	49.5	4.0	91.3	3.3	2.8
Arg	−96.9	54.1	124.9	86.0	−95.4	54.6	105.7	80.4	1.8	2.9
Asn	−94.4	57.2	102.9	69.8	−69.8	88.8	36.5	65.1	1.7	4.0
Asp	−84.0	48.5	118.2	82.5	−73.4	68.5	22.0	72.1	3.2	4.2
Cys	−109.9	35.0	121.1	72.4	−107.4	58.2	101.7	72.9	1.3	0.8
Gln	−98.9	40.4	118.1	84.6	−94.5	44.3	94.4	81.3	2.1	1.9
Glu	−94.8	31.1	133.0	112.6	−85.6	47.6	39.7	82.8	1.9	3.3
Gly	−146.3	79.5	−176.1	77.8	82.5	69.0	7.4	78.2	2.6	8.9
His	−85.6	57.5	39.4	88.3	−94.6	73.6	68.0	66.3	0.8	0.9
Ile	−102.7	26.9	114.4	87.1	−102.0	32.7	113.7	79.2	2.8	2.0
Leu	−85.6	31.5	119.3	100.1	−82.9	52.8	71.3	79.2	4.5	3.4
Lys	−99.6	33.9	132.3	90.2	−96.7	54.0	57.7	84.4	2.8	4.2
Met	−95.3	31.6	57.6	87.5	−85.6	63.3	95.1	71.3	0.7	0.6
Phe	−110.8	34.0	136.1	88.3	−103.9	61.2	60.6	77.1	1.5	1.4
Pro	−64.0	14.6	129.2	82.6	−63.6	13.1	106.7	87.0	1.4	2.9
Ser	−97.6	40.0	137.8	98.8	−85.6	58.3	37.6	89.8	2.2	3.9
Thr	−106.9	25.9	141.6	82.5	−101.2	38.4	85.1	86.0	2.6	3.3
Trp	−95.1	27.5	119.4	85.7	−96.1	36.3	130.6	88.7	0.7	0.6
Tyr	−105.8	33.7	144.0	71.2	−107.6	60.1	108.4	70.8	2.0	2.1
Val	−109.8	29.0	136.3	80.8	−101.6	28.1	110.9	92.8	3.6	2.5

[a]A distinction is made between an occurrence in the N- or C-terminal two residues (terminal) and the remaining residues (middle).
[b]Frequency, relative abundance of amino acid type.
[c] (ϕ), mean ϕ.
[d]SD (ϕ), standard deviation on ϕ.

In Figure 6 the predictions for six longer CDR-H3 regions with lengths between 12 and 17 residues are shown superimposed on the crystal structures of the loops.

Dependence of loop structure on amino acid type and position

The trained networks represent a machinery which transforms sequence data into a 3-D model of a CDR-H3 loop structure. We have attempted to use the networks to model the effect of changes of the amino acid types in the sequence in order to see corresponding changes in the structure of the loop.

Table IV presents a set of consensus data for the average backbone geometry around each amino acid type. It must be emphasized that the table is constructed on the basis of average backbone angles for single amino acids in the loop. While we do not consider the precise position within the loop, the values for the two residues at the N- and C-termini have been separated from the remaining residues. Certain amino acids are observed to occupy a well-defined compact region in the ϕ/ψ space. These data indicate the expected effect a given amino acid type may have on loop structure.

To assess the effect of amino acid exchanges, the neural network was used to predict the loop structure of a modified sequence. The original sequences of CDR-H3 loops not contained in the training set were expanded such that each residue at each position in the loop was replaced with the 19 other amino acids. This procedure produces $19N$ sequences (seq_i) each containing a single point mutation for a given sequence of length N. Each mutated sequence, seq_i, was processed using the neural network resulting in structure predictions for each point-mutated sequence. Each prediction, $loop_i$, was optimally superimposed with the original reference loop structure taken from the protein databank and an r.m.s. value (rms_i) was calculated as a measure of structural similarity.

At a given position in the loop sequence, the influence of the position of a mutation on the conformation is indicated by the mean r.m.s. value of the loop structures generated from all

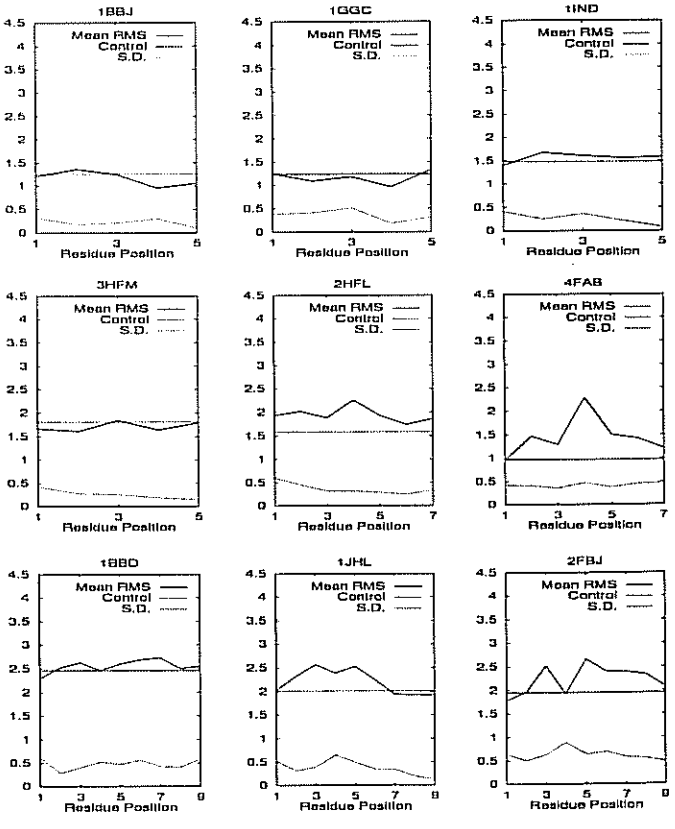

Fig. 7. Average structural effect of 19 alternative residues at each position of the CDR-H3 loops 1 to 9 in the test set. The line denoted 'Control' shows the r.m.s. deviation of the unmutated prediction.

19 point mutations at that position. This position-specific mean r.m.s. deviation correlates with the structural effect of a non-specific mutation at this location. The standard deviation of

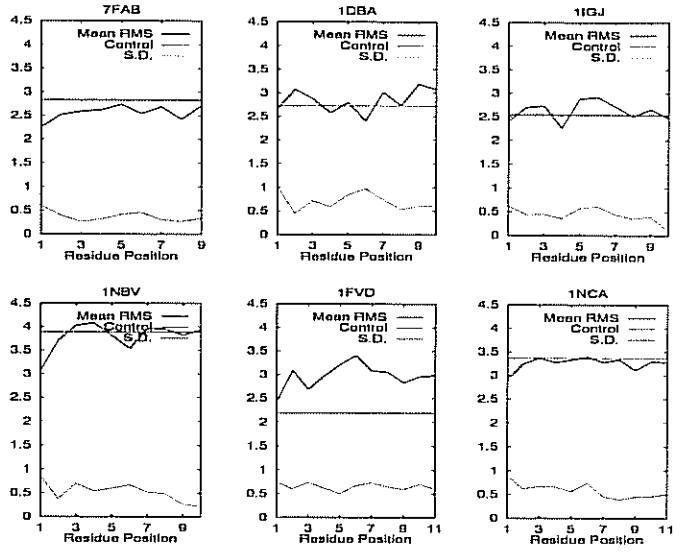

Fig. 8. Average structural effect of 19 alternative residues at each position of the CDR-H3 loops 10 to 15 in the test set.

Table V. Shape parameters for CDR-H3 loops

Name	Length	Height (Å)	Width (Å)	Depth (Å)
1FGV	11	11.0	15.58	9.04
2MCP	11	6.30	12.39	12.59
1NCA	11	6.88	14.85	5.97
1HIL	11	9.16	14.33	9.98
1FVD	11	9.94	10.46	10.32
8FAB	12	5.40	16.01	10.62
6FAB	12	8.44	15.91	10.49
1IGM	12	5.96	14.35	9.18
2FB4	17	10.9	20.63	8.61
1DFB	17	8.27	23.60	10.17

these position-specific r.m.s. values indicates the degree of structural variation caused by the specific nature of the point mutation.

Figures 7 and 8 show the position-specific averages and standard deviations of the r.m.s. values resulting from this mutation analysis for the CDR-H3 loops in the test set. These results suggest that the residues in the middle of the loop have the largest effect on the conformation of the loop. *In situ*, this effect may be masked by mutations at the bases of the loop which are more likely to cause clashes with framework residues in the antibody disrupting the loop conformation.

In the 15 shorter, better-predicted CDR-H3 loops, there is a total of 13 glycines. At nine of these, there is a peak in the r.m.s. deviation and at six, the maximum r.m.s. deviation for the loop occurs. This indicates that mutation of a glycine to any other residue has a large effect on the conformation. This correlates with the common observation that glycines are able to adopt ϕ/ψ combinations inaccessible to residues having side chains (Ramachandran *et al.*, 1963).

Another analysis was performed on the shapes of the CDR-H3 regions. As a description of loop shape, we have used three parameters to describe the overall conformation of a loop. By rotating a loop such that the two terminal Cα atoms are along the x-axis and the centre of geometry falls on the xy-plane, we are able to define the height (H), width (W) and depth (D) of the loop. In Table V we give the values of H, W and D for CDR-H3 loops taken from the test set. From these values, it can be seen that the loop width increases with the length (Pearson's correlation coefficient $r = 0.90$), but that height and depth are poorly correlated with loop length ($r = 0.26$ and $r = -0.07$ respectively).

The representation of loop shapes by these three parameters H, W and D is meant to give an intuitive understanding, for example, how much a CDR-H3 loop reaches out from the rest of the antibody molecule. It is clearly a major simplification that might be used as an output representation of another network predicting these quantities from sequence information. Such a simplified network could be combined with the networks used in this application as a source of additional input information or as an additional criterion to select among different predictions to improve the prediction accuracy.

Discussion

A novel method has been constructed and applied to the prediction of the 3-D structure of the CDR-H3 loops in the variable regions of antibodies on the basis of amino acid sequence information alone. This prediction scheme is designed to be used as part of a larger system for modelling the complete antigen combining site. The conformation of the other five CDRs (L1, L2, L3, H1 and H2) has proved relatively easy to model by standard *ab initio* and knowledge-based methods, but the CDR-H3 loops which are much more variable in both sequence and in length have proved much more difficult to model, especially when longer than 10 amino acids in length.

The most effective networks have used the time-delay architecture and reflect the loop topology in the design of the input and output layer. Results of a jury of networks have been filtered using the requirement that the Cα distances in the predicted loops meet the constraints for attachment to an antibody framework and the models generated could be used as input to molecular dynamics or energy minimization simulations after the addition of side chains

The neural networks used have been quite successful in predicting the conformations of CDR-H3 regions of up to 11 residues with a Cα r.m.s. deviation of ~2 Å. This level of success has been achieved using networks which employ no environmental information; their predictions are based solely on the sequences of the loops themselves. The fact that the prediction was less successful with loops of 12 or more residues suggests that the conformations of longer loops are more highly influenced by the environment. The networks have been used to derive information on the effects of particular positions and residue type on CDR conformation.

As we have concentrated on a particularly difficult loop to model, CDR-H3 of the antibody Fv fragment, direct comparison with other methods is difficult. We have calculated only Cα r.m.s. deviations so we cannot perform direct comparisons with authors who provide only backbone (with or without the O or Cβ atoms) r.m.s. deviations. The best-known analysis and modelling of antibody hypervariable loops is that performed by Chothia *et al.* (1989). However, they exclude CDR-H3 from their analysis.

Martin *et al.* (1989) were able to achieve an r.m.s. deviation of 0.8 Å on the backbone for the four-residue CDR-H3 of antibody Gloop2 and 1.45 Å for the seven-residue CDR-H3 of HyHEL-5. However, both of these were modelled in the context of the crystal structure of the remaining loops and therefore cannot be applied to modelling completely unknown structures. When they modelled CDR-H3 of Gloop2 onto an empty combining site (Martin *et al.*, 1991), their accuracy fell

to 1.45 Å r.m.s. deviation on the backbone atoms. This is comparable with the Cα r.m.s. deviations of 1.58 and 0.97 Å for the seven-residue loops modelled here.

Bruccoleri *et al.* (1988) achieved backbone r.m.s. deviations of 1.0 Å for a four-residue section of CDR-H3 of HyHEL-5 and 1.1 Å for a seven-residue section of the McPC603 CDR-H3. They do not model the complete loops which are seven and 11 residues long respectively.

Fidelis *et al.* (1994) have performed a comparison between database and conformational search techniques. They conclude that the Protein Databank is only saturated for loop lengths of up to seven residues and that database methods are only suitable for modelling short fragments of protein. While conformational search techniques do saturate conformational space, they become prohibitively expensive in computer time for sections of protein much longer than seven or eight residues.

No systematic analysis of modelling longer CDR-H3 loops has been performed by these other authors and we are therefore unable to compare our results with theirs.

Our current methodology has considered only the local conformation of the loops. The conformations which we generate could be fitted onto a parent antibody framework using least squares fitting of the base residues or an overlap approach using the base residue vector and the centre of geometry of the loop as described by Martin *et al.* (1989). Such methods are likely to result in poor geometry around the take-off region, but this can be relieved by molecular mechanics methods.

We are now attempting to improve our current results. We are addressing the problem of improving the fitting of loop fragments onto the framework and are building more sophisticated networks incorporating environmental information and using additional networks which employ a simplified representation of loop conformation described by height, width and depth parameters. Improved and more detailed networks are required to expand to point mutation analysis and suggest the structural effect of specific point mutations. This will be very beneficial for the design of site-directed mutagenesis experiments.

Acknowledgements

This research was supported in part by the BIOINFORMATIK program by the Bundesministerium für Forschung und Technologie, project NEUROGEN, Förderkennzeichen 01 IB 303 A. A.C.R.M. is supported by the UK Medical Research Council.

References

Alzari,P.M., Lascombe,M.-B. and Poljak,R.J. (1988) *Annu. Rev. Immunol.*, **6**, 555–580.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bohr,H., Bohr,J., Brunak,S., Cotterill,R.M.J., Lautrup,B., Nxrskov,L., Olsen,O.H. and Petersen,S.B. (1988) *FEBS Lett.*, **241**, 223–228.

Bohr,H., Bohr,J., Brunak,S., Cotterill,R.M.J., Fredholm,H., Lautrup,B. and Petersen,S.B. (1990) *FEBS Lett.*, **261**, 43–46.

Bruccoleri,R.E. and Karplus,M. (1987) *Biopolymers*, **26**, 137–168.

Bruccoleri,R.E., Haber,E. and Novotný,J. (1988) *Nature*, **335**, 564–568.

Chothia,C. *et al.* (1989) *Nature*, **342**, 877–883.

Chothia,C., Lesk,A.M., Gherardi,E., Tomlinson,I.M., Walter,G., Marks,J.D., Llewelyn,M.B. and Winter,G. (1992) *J. Mol. Biol.*, **227**, 799–817.

Darsley,M.J. and Rees,A.R. (1985) *EMBO J.*, **4**, 393–398.

de la Paz,P., Sutton,B.J., Darsley,M.J. and Rees,A.R. (1986) *EMBO J.*, **5**, 415–425.

Fidelis,K., Stern,P.S., Bacon,D. and Moult,J. (1994) *Protein Engng*, **7**, 953–960.

Fine,R.M., Wang,H., Shenkin,P.S., Yarmush,D.L. and Levinthal,C. (1986) *Proteins: Struct. Funct. Genet.*, **1**, 342–362.

Holley,L.H. and Karplus,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 152–156.

Kabat,E.A., Wu,T.T., Reid-Miller,M., Perry,H.M. and Gottesman,K.S. (1987) *Sequences of Proteins of Immunological Interest*. 4th edn. US Department of Health and Human Services.

Kabat,E.A., Wu,T.T., Perry,H.M., Gottesman,K.S. and Foeller,C. (1991) *Sequences of Proteins of Immunological Interest*. 5th edn. US Department of Health and Human Services.

Lerner,R.A. and Benkovic,S.J. (1988) *BioEssays*, **9**, 107–112.

Martin,A.C.R., Cheetham,J.C. and Rees,A.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9268–9272.

Martin,A.C.R., Cheetham,J.C. and Rees,A.R. (1991) *Methods Enzymol.*, **203**, 121–153.

Mas,M.T., Smith,K.C., Yarmush,D.L., Aisaka,K. and Fine,R.M. (1992) *Proteins: Structure. Funct. Genet.*, **14**.

Moult,J. and James,M.N.G. (1986) *Proteins: Struct. Funct. Genet.*, **1**, 146–163.

Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) *J. Mol. Biol.*, **7**, 95–99.

Reid,K.B.M. (1986) *Essays Biochem.*, **22**, 27–68.

Rumelhart,D.E., McClelland,J.L. *et al.* (1986) *Parallel Distributed Processing*. MIT Press.

Schilling,J., Clevinger,B., Davie,J.M. and Hood,L. (1980) *Nature*, **283**, 35–40.

Smith-Gill,S.J., Mainhart,C.R., Lavoie,T.B., Feldmann,R.J., Drohan,W. and Brooks,B.R. (1987) *J. Mol. Biol.*, **194**, 713–724.

Verhoeyen,M., Milstein,C. and Winter,G. (1988) *Science*, **239**, 1534–1536.

Waibel,A.H., Hanazawa,T., Hinton,G.E., Shikano,K. and Lang,K.J. (1989) *IEEE Trans. Acoustic, Speech Signal Process.*, **37**, 328–339.

Zell,A., Mache,N., Sommer,T. and Korb,T. (1991) In *Proceedings of the Applications of Neural Networks Conference*. SPIE, Aerospace Sensing International Symposium, Orlando, FL, pp. 708–719.

Received August 26, 1994; revised December 21, 1994; accepted March 16, 1995