Hybrid models of uncertainty in protein topology prediction

Simon Parsons
Advanced Computation Laboratory
Imperial Cancer Research Fund
P.O. Box 123
Lincoln's Inn Fields
London WC2A 3PX

Abstract

Predicting protein structure is an important problem in molecular biology, and one that has attracted a lot of attention. It is also a difficult problem since the available data is incomplete and pervaded with uncertainty. This paper describes models for the prediction of an intermediate level of protein structure known as the topology of the protein. The models handle uncertainty explicitly, making use of probability, possibility and evidence theories singly and in combination to handle different aspects of the problem.

Keywords: Molecular biology, protein structure prediction, uncertain data, hybrid models.

1 Introduction

Proteins are large biological macromolecules that form the main components of living organisms. In addition, proteins, in the form of enzymes, hormones and antibodies, control most of the crucial processes in cells. The function of a particular protein is determined by the chemical interactions at its surface, and these are related to its three dimensional structure. Thus knowledge of protein structure is important. The structure of proteins can be described at various levels of detail. The primary structure consists of a list of the amino acids that make up the protein. Each amino acid is one of twenty naturally occurring molecules from which all proteins are made. The secondary structure is a description of the way that the amino acids are grouped together into substructures within the three dimensional structure. Two important forms of secondary structure are β-sheets, which consist of a number of β-strands, and α-helices. The tertiary structure of a protein is the set of three dimensional co-ordinates of every atom in the protein. Protein topology is an intermediate level somewhere between secondary and tertiary structure which specifies the relations between secondary structural units, in terms of how such units are ordered along the protein.

Now, knowledge of three dimensional protein structure is sparse. While the determination of the primary structure of proteins has become routine, determination of secondary and tertiary structure is more difficult. Thus the primary structures for many tens of thousands of proteins are known, but only some hundreds of distinct proteins have had their three dimensional structure determined. This figure is tiny in comparison with the vast number of proteins that exist, and the discrepancy motivates much research into determining protein structure. The major problem is that experimental methods are time-consuming and expensive and may even be impossible to use since some proteins change structure when isolated. In consequence, molecular biologists are turning to computational techniques for predicting protein structure from amino acid sequences.

2 Protein Topology Prediction

Of particular interest is the prediction of protein topology because the topology can be used to guide the choice of experiments to confirm particular ideas about the structure and to search for similar known structures. A major difficulty in this prediction is that a vast number of possible topologies can be hypothesised from a single secondary structure prediction. In general, for a mixed α/β-sheet of n strands, where $n > 1$, there are $\frac{n!(4n-1)}{2}$ possible ways of arranging the strands (Clark et al. 1992) (Table 1). To reduce this space, scientists identify and apply constraints based upon experimental data (Clark et

No. of strands	2	3	4	5	6
No. topologies	4	48	768	15,360	68,640

Table 1: The number of possible topologies

al. 1992, 1994), (Cohen and Kuntz 1987). As an example, Taylor and Green (1989) investigated the topology of nucleotide binding proteins using rules based on previous analyses of α/β sheets:

- C1. For parallel pairs of β-strands, β-α-β and β-coil*-β connections are right handed (Richardson 1976), (Sternberg and Thornton 1977).
- C2. The initial β-strand in the amino acid sequence is not an edge strand in the sheet (Brandon 1980).
- C3. Only one change in winding direction occurs (Richardson 1981).
- C4. The β-strands associated with the conserved† patterns lie adjacent in the sheet (Walker et al. 1982).
- C5. All strands lie parallel in the β-sheet.
- C6. Unconserved strands are at the edge of the sheet.

To evaluate the plausibility of topologies that were consistent with the constraints Taylor and Green employed five folding principles:

- F1. Strands are ordered in the sheet by hydrophobicity, with the most hydrophobic‡ strands central.
- F2. Parallel β-coil-β connections contain at least 10 amino acids.
- F3. Insertions and deletions§ occur on the edge of domains.
- F4. Most conserved loops should lie adjacent.
- F5. Long secondary structures should pack approximately parallel or anti-parallel, with sequential units anti-parallel.

After some adjustment of the secondary structure prediction they were able to produce a topological hypothesis that satisfied $C1$ to $C6$, was consistent with $F1$ to $F4$, and was indeterminate with respect to $F5$. The structure was also similar to the known structure of a nucleotide binding protein.

To avoid the problems inherent in both manual search and exhaustive generate and test algorithms, Clark et al. (1992) developed a constraint satisfaction algorithm for protein topology prediction named CBS1. CBS1 is implemented in Prolog, based on tree search and achieves efficiency through high level constraint evaluation and forward pruning of the search tree. The implementation runs on Sun workstations under Quintus Prolog and has was used to reproduce Taylor and Green's results as well as to identify a new topological hypothesis consistent with the constraints, indicating that the original search was not

*A protein has coil structure where it is neither a β-strand nor an α-helix.

†When the amino acid sequences of proteins that perform similar functions, but are found in different species, are compared, parts of the sequences are often found to be the same in all proteins. These are said to be conserved.

‡Lacking an affinity for water.

§When amino acid sequences are aligned to determine conserved regions, some regions do not match. Thus, to align the conserved regions it is necessary to split up sequences, 'adding' spaces between certain amino acids. Similarly, it may be necessary to 'delete' acids in order to match regions.

Protein ID.	Constraints Violated	Protein ID.	Constraints Violated
p4adh	F1	p1pfk	C5 F1
p5adh	F1	p2pfk	C5 F1
p6adh	F1	p3pfk	C5
p7adh	F1	p4pfk	C5
p1ldx		p1gpd	C3 C5 F1 F2
p3ldh	F1	p1gd1	C3 C5 F1 F2
p4ldh	F1	p2gpd	C3 C5 F1 F2
p3dfr	C3 C5 F1	p3pgk	F1
p4dfr	C3 C5 F1 F2		
p3adk		p3grs	C2 F1

Table 2: The results of checking constraints against eight nucleotide-binding domain proteins

exhaustive. Clark et al. (1992) also assessed the validity of the constraints by checking them against the known structures of eight nucleotide binding domains with similar function. The results are reproduced in Table 2. The protein ID is a unique number which identifies the full three dimensional protein structure in a public database. The structures that are grouped together in Table 2 are those relating to the same protein. For instance *p1gpd*, *p1gd1* and *p2gpd* are different experimentally determined structures for D-glyceraldehyde-3-phosphate dehydrogenase. Each of the variations should be considered equally valid, so when a rule holds for one form of a protein and not for another, it is ambiguous whether or not the constraint holds for that protein. A further set of data concerning the validity of the constraints is presented in an earlier paper (Shirazi et al. 1990). Here $C1$, $C2$, $C3$, $C5$ and F2 were tested against a set of 33 α/β sheet proteins, giving Table 3.

Protein ID.	Constraints Violated	Protein ID.	Constraints Violated
p1aat	C2 C5	p1ts1	C2 C5
p1bp2	C2 C5	p1ubq	C3 C5
p1cac	C2 C3 C5	p2b5c	C2 C3 C5
p1cpb	C2 C3 C5	p2cab	C2 C3 C5
p1crn	C2 C5	p2cdv	C2 C5
p1cts	C2 C5	p2cts	C2 C5
p1ctx	C5	p2lzm	C5
p1hip	C2 C5	p2ssi	C5
p1nxb	C3 C5	p3bp2	C2 C5
p1ovo	C5	p3cts	C2 C5
p1p2p	C2 C5	p3dfr	C3 C5
p1ppd	C2 C5	p3pgm	C5
p1rn3	C2 C5	p4cts	C2 C5
p1sbt	C1	p4dfr	C3 C5 F2
p1sn3	C2 C5	p4fxn	
p1srx	C2 C3 C5	p4pti	C2 C5
p5cpa	C2 C3 C5		

Table 3: The results of checking constraints against 33 α/β sheet proteins.

3 Representing the uncertainty

Together these results show that while the folding rules are useful heuristics most are only true some of the time. This suggests that a good model of the problem would handle the uncertainty in the constraints explicitly. The best way of doing this is not clear, and so in the tradition of experimental investigations of the best way of modelling uncertainty in a given problem (Heckerman 1990), (Heckerman and Shwe 1993), (Saffiotti et al. 1994) we discuss a number of different ways in which the data from Tables 2 and 3 may be represented. There are, of course, other possibilities which are not discussed here— we just

Constraint	Number of cases in which the constraint is violated	$p(XA)$
C1	1	0.967
C2	22	0.333
C3	10	0.606
C5	31	0.065
F1	-	-
F2	1	0.967

Table 4: Probabilities of constraints holding in nature, based upon the results of the α/β proteins

Constraint	Number of cases in which the constraint is violated	$p(XA)$
C1	0	1.0
C2	1	0.947
C3	5	0.737
C5	9	0.526
F1	15	0.211
F2	4	0.789

Table 5: Probabilities of constraints holding in nature based upon the "disambiguated" results of the nucleotide binding proteins

cover the most obvious. We also just deal with modelling constraints $C1$,$C2$,$C3$,$C5$,$F1$ and $F2$ for which Clark et al. (1992) and Shirazi et al. (1990) give data.

3.1 Using probability theory

Since the data is drawn from a reasonably random population of proteins the following argument can be made. Table 3 holds a list of 33 proteins. Of these, 24 conform to constraint $C3$, and 9 do not, so a structure that conforms to $C3$, has a probability of $P(C3A) = \frac{23}{33} = 0.606$ of existing in nature. This argument gives the probabilities of Table 4. Since the sample size is just 33 proteins, the probabilities will not be very accurate, but they are the best values that can be obtained in this ill-known domain.

Table 2 may be used to get a second set of probabilities, but in this case the data is ambiguous. Of the eight proteins analysed, several have alternative structures and some constraints hold for some alternative structures and not for others. Thus it is not clear whether or not such a constraint is valid for the protein. One solution is to "disambiguate", considering each of the nineteen possible structures as a separate entity. Doing this gives the probabilities of Table 5. However, it could be argued that disambiguation distorts the data, and the uncertainty should be modelled in a "purer" way acknowledging the ambiguity. This can be done using interval probabilities with the lower value calculated by counting proteins for which the rule is ambiguous as proteins for which it fails to hold, and the upper value by counting proteins for which the rule is ambiguous as proteins for which it does hold. This gives Table 6. Alternatively the ambiguity could be modelled using evidence theory.

3.2 Using evidence theory

Evidence theory provides good support for modelling ambiguity, providing a mechanism for allocating a degree of belief to the disjunction of two competing hypotheses. Considering Table 2 with respect to $F2$ there are three sets of proteins. We have $\{p3pgk, p1ldx, p3ldh, p4ldh, p1pfk, p2pfk, p3pfk, p4pfk, p3grs, p3adk, p4adh, p5adh, p6adh, p7adh\}$ where every one of the 14 possible structures of the six proteins conforms to $F2$, $\{p1gpd, p2gpd, p1gd1\}$ where every possible structure of the protein violates $F2$, and $\{p3dfr, p4dfr\}$ where of the two structures for the protein, one conforms to $F2$ and one does not. This ambiguity may be modelled as follows. There are six proteins that conform to $F2$ so the basic mass assignment is $m(\{F2A\}) = 6/8$. There is one protein for which the constraint fails to hold. Thus $m(\{\neg F2A\}) = 1/8$. The remaining mass, $1 - (m(\{F2A\}) + m(\{\neg F2A\}))$ is the probability that the constraint either applies or doesn't apply, $m(\{F2, \neg F2A\})$. This latter is a measure of ignorance about the applicability of the constraint. For Table 2 the masses of Table 7 are appropriate.

Constraint	Number of cases in which the constraint is violated	$p(XA)$
C1	0	1.0
C2	1	0.875
C3	2	0.75
C5	3	0.625
F1	5-7	[0.125, 0.375]
F2	1-2	[0.750, 0.875]

Table 6: Probabilities of constraints holding in nature based upon the "pure" results from the nucleotide binding proteins

Constraint	$m(\{FxA\})$	$m(\{\neg FxA\})$	$m(\{FxA, \neg FxA\})$
F1	0.125	0.625	0.25
F2	0.75	0.125	0.125

Table 7: The basic mass assignments for modelling the constraints using evidence theory according to the results from the nucleotide binding proteins

3.3 Using possibility theory

It is also possible to model the constraints using possibility theory. It is possible to use a number of different sets of numerical information, but the basic principle behind the modelling is the same. If a structure conforms to a constraint, it is entirely possible that the structure is that of a naturally occurring protein. However, if a structure fails to conform to a constraint then it becomes less possible that the structure is a naturally occurring protein. Indeed the possibility of the structure being a natural protein falls to a figure that reflects the proportion of naturally occurring proteins that do not conform to the constraint. Thus, for each constraint Cx, the possibility that the protein occurs in nature given that Cx holds, $\Pi(CxA) = 1$, while the possibility that the protein occurs in nature given that Cx does not hold, $\Pi(\neg CxA)$, depends upon data in Tables 2 and 3. This may be disambiguated or used to define an interval in the same way as is done in Tables 4 to 6 for probability theory. For instance, taking the disambiguated values of Table 5, the possibility of $C5$ not holding, $\Pi(\neg C5) = 0.474$.

3.4 Handling ignorance

We have also to encode the lack of information about the applicability of $F1$ in Table 3. The usual probabilistic method would be to declare that $p(F1A) = p(\neg F1A) = 0.5$. However, other methods could be used. In evidence theory, since nothing can be said about the basic probability assignment to $F1A$ and $\neg F1A$, $m(\{F1A\}) = m(\{\neg F1A\}) = 0$, and all the basic probability is assigned to the disjunction, giving the vacuous belief function $m(\{F1A, \neg F1A\}) = 1$. The meaning of this assignment is that it is not possible to assign any belief to one hypothesis rather than another. Another alternative is to use possibility theory. Since there is no knowledge about the chance of $F1$ holding, it is entirely possible that $F1$ holds and $\Pi(F1A) = 1$. If $\neg F1A$ is considered, a similar argument may be made to obtain $\Pi(\neg F1A) = 1$. From the viewpoint of the information that is available all of these three approaches are perfectly correct, and there is no obvious way of choosing between them.

4 Calculating the validity of structures

Having considered the different ways in which the uncertain nature of the constraints can be modelled, we turn to considering how to employ use these models in topology prediction. CBS1 generates as its output sets of possible topologies of nucleotide binding domain proteins and the constraints to which they conform. The kind of output that would be useful to molecular biologists would be some kind of measure of the validity of the sets of topologies based upon the constraints to which they conform.

To do this we employed Mummu (Parsons 1993) an adaptation of Pulcinella (Saffiotti and Umkehrer 1991) which handles interval valued probabilities, possibilities and beliefs, and in which the intervals are propagated using interval arithmetic. This system uses a valuation network representation (Shenoy

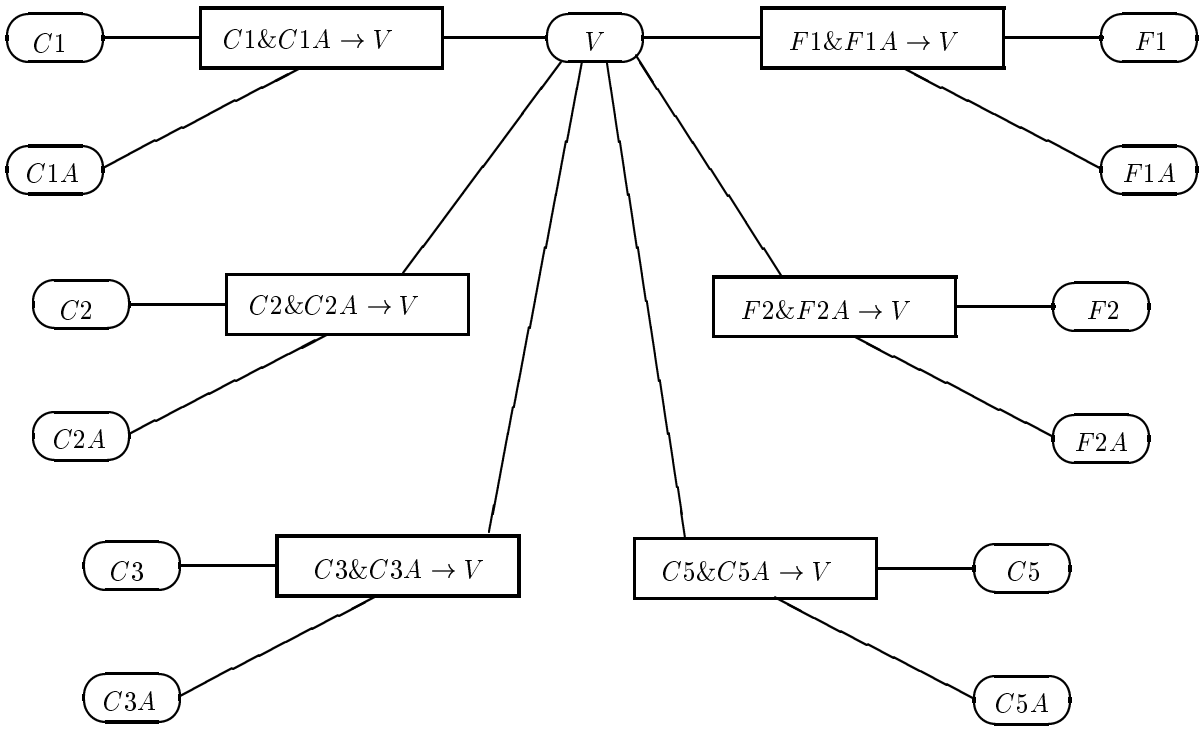


Figure 1: A network relating the effects of constraints to the validity of a structure.

1991), and the valuation network of Figure 1 was adopted, based upon the work of Smets and Hsia (1990). This network expresses the fact that the validity of the structure is a combination of the effect of all constraints. Thus, for $C1$ there is a node $C1$ which is true if $C1$ holds for the structure in question, and false otherwise. The effect of this is combined on the node $C1\&C1A \rightarrow V$ with $pC1A$ which is the measure of how likely it is that a structure that conforms to $C1$ exists in nature. This is then combined with similar results for other constraints to get an overall measure of validity V.

4.1 Single formalism approaches

Firstly we consider the ways that the models of uncertainty can be used so that the whole problem is modelled using a single formalism, as it would be by any other existing system. Mummu may be used to establish a validity value using each of the following sets of data.

Point-valued probability: Table 5 has point values for each of the constraint probabilities $p(CxA)$ that are based on the disambiguated sample of eight nucleotide binding domain proteins. The results of this computation are given in the second column of Table 8, and in the first graph in Figure 3, both of which may be found in the Appendix. In the graph each point on the x-axis corresponds to a single set of constraints, with the validity measured up the y-axis. The leftmost point on the x-axis corresponds to the first set of constraints in Table 8.

Interval-valued probability: Table 6 has interval values which represent the ambiguity surrounding $F1$ and $F2$ holding. Results using these values are given in the third column of Table 8 and the second graph (reading across the page) in Figure 3. Note that in order to represent intervals graphically, they have been transformed into point values by replacing them with their mid-points. This transformation may be justified (Parsons 1993) by an argument based on maximum entropy. A second set of interval values may be obtained by realising that since the nucleotide binding domain proteins from which Table 2 was obtained are of the same class as the proteins whose structure is being predicted, while the α/β-sheet proteins are a more general class, we could use the values derived from Table 2 as an upper bound, and those of Table 3 as a lower bound. Since there are two ways of establishing values from Table 2 we have two sets of results and these are given in the fourth and fifth columns of Table 8 and the third and fourth graphs in Figure 3.

Point-valued evidence theory: Table 7 gives basic probability assignments for the hypotheses $\{F1\}$, $\{\neg F1\}$, $\{F1, \neg F1\}$, $\{F2\}$, $\{\neg F2\}$, and $\{F2, \neg F2\}$. In addition the values that from Table 6 are

used as basic probability assignments for the hypotheses $\{C1\}$, $\{\neg C1\}$, $\{C2\}$, $\{\neg C2\}$, $\{C3\}$, $\{\neg C3\}$, $\{C5\}$ and $\{\neg C5\}$. Evidence theory may be used to propagate these values through the network of Figure 1, and the results are given in the sixth column of Table 8 and the fifth graph of Figure 3. Any belief not assigned to $\{V\}$ is assigned to $\{V, \neg V\}$, so that for the constraint set $\{\ \}$, $bel(\{V, \neg V\} = 1$

Point-valued possibility theory: Table 4 gives a set of values which may be used to derive point possibility values for the applicability of the constraints $C1$ to $C5$ and $F2$. Since there is no data for the applicability of $F1$ it may modelled by setting $\Pi(\neg F1A) = 1$, as discussed above. This data gives the results in the seventh column of Table 8 and the sixth graph of Figure 3. It should be noted that in this model $\Pi(V) = 1$ for all sets of constraints, and that, as for all possibilistic models, the graph and the table give the possibility of $\neg V$. Alternatively one may use the "disambiguated" data of Table 5, which allows point possibilities to be derived for every constraint. The results of using these values are given in the eighth column of Table 8 and the seventh graph of Figure 3.

Interval-valued possibility theory: It is also possible to use the "pure" data from Table 6 provided that interval possibility values are used to take account of the interval nature of the data, representing the range of possibility values that may be calculated given the imprecise information available (Parsons 1993). Using Mummu to propagate these values generates the results in the ninth column of Table 8 and the eighth graph of Figure 3. In this model $\Pi(V) = [1,\ 1] = 1$ in all cases.

4.2 Hybrid approaches

All of the single formalism approaches discussed above have skirted around what in many ways seems to be the natural representation for the problem— one that uses different formalisms. The most accurate data that is available is that for the nucleotide binding domain proteins, and this splits neatly into two parts. There are probabilities (Table 6) that model conformance to $C1$ to $C5$, and there is some ambiguous data best modelled in evidence theory (Table 7) that says something about structures that conform to $F1$ and $F2$. This suggests that the network of Figure 1 be partitioned into two parts as in Figure 2.

In this network values are propagated according to evidence theory in the right-hand section until a value for the node $V1$ is established. This value is then translated into a probability interval and combined with the results from the rest of the network to establish overall probabilistic measures of validity. The translation, discussed in detail in (Parsons 1993) uses intervals to model the fact that a belief value may be taken as the lower bound on a probability (Dubois and Prade 1988), thus:

$$Bel(x) = n \xrightarrow{\text{translates to}} p(x) = [n, 1] \tag{1}$$

The results of using this method are given in the tenth column of Table 8 and the ninth graph of Figure 3.

This is only one way of combining formalisms in this problem. An alternative is to combine the evidence theory model of the ambiguous data about F1 and F2 with the possibility model. Such an integration would be carried out in the network of Figure 2, but with every probability replaced with a possibility based on values from Table 6. The translation is based upon the fact that a belief is a lower bound on a probability while a possibility is an upper bound (Dubois and Prade 1988), so:

$$Bel(x) = n \xrightarrow{\text{translates to}} \Pi(x) = [n, 1] \tag{2}$$

The results from such an integration are given in the eleventh column of Table 8 and the final graph in Figure 3.

5 Discussion

Unfortunately, there is no obvious "gold standard" (Heckerman 1990) against which to compare the results. However, it is possible to suggest a number of criteria for choosing between the approaches based upon the data that they use, the theory that they are based on, and the results that they generate. For instance it would seem sensible to use the data of Table 2 when dealing with nucleotide binding domain proteins, and so the results in the second, third, sixth and eighth to eleventh columns of Table 8 seem most appropriate. These are, of course the results depicted in the first, second, fifth and seventh to tenth

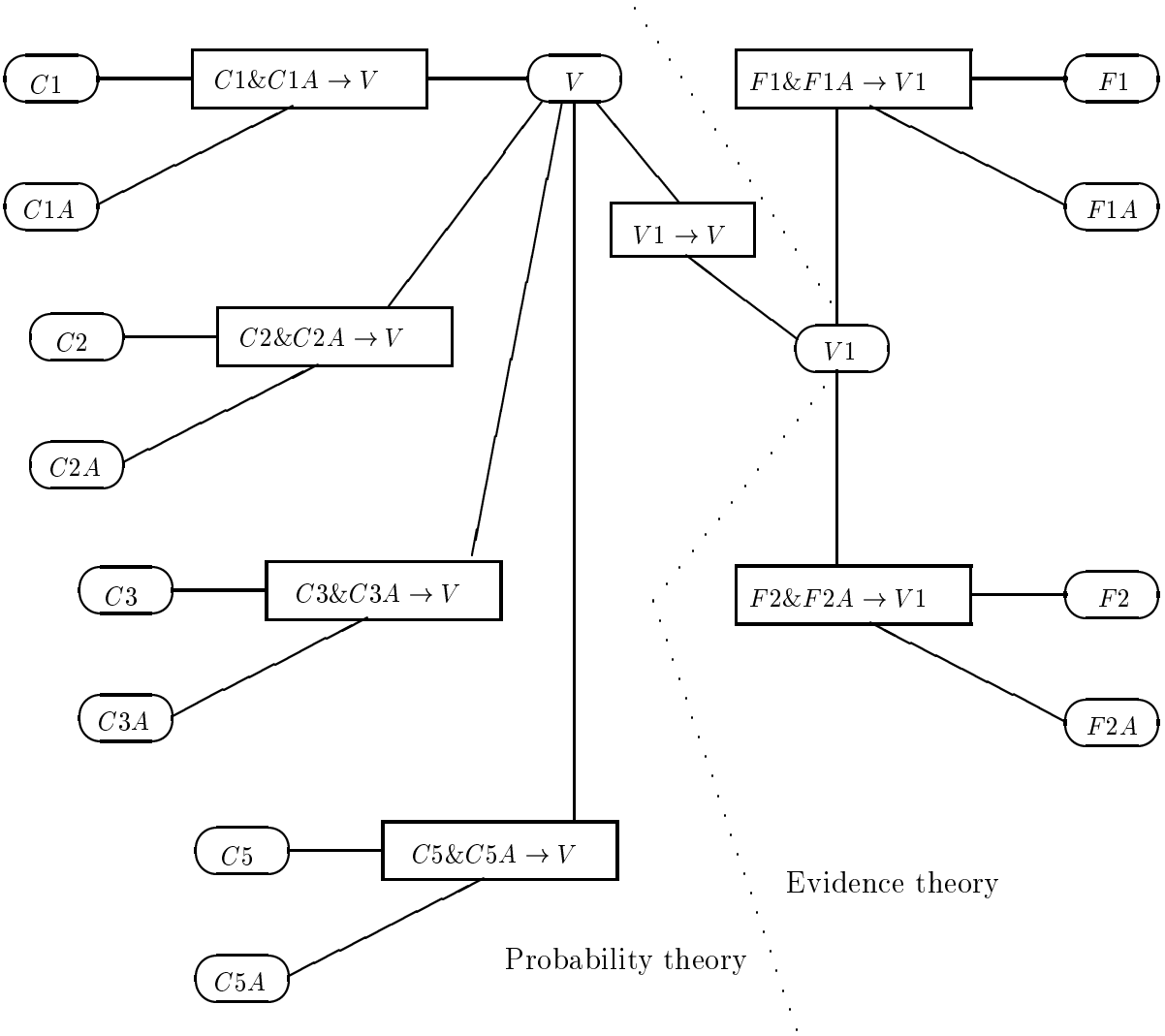


Figure 2: A network relating the effect of the constraints in which probability and evidence theories may be integrated.

graphs of Figure 3. In this case the choice of which results to adopt will depend on which method of dealing with ambiguity is preferred. The ambiguity may be handled by using "disambiguated" values in which case the results in the second and eighth columns of Table 8, and the first and seventh graphs of Figure 3, apply. Alternatively, an interval representation of the ambiguity may be adopted, in which case the results in the third and ninth columns of Table 8 and second and eighth graphs of Figure 3 should be considered. Finally, the ambiguity may be modelled using evidence theory, in which case the results in the sixth, tenth and eleventh columns, and the fifth, ninth and tenth graphs, are the ones to look at.

If a more general class of proteins are being considered, then the results in the fourth, fifth and seventh columns and third, fourth and sixth graphs may be more appropriate since these are at least partly based upon data from single domain α/β sheet proteins. Another means of choosing the best method might be preference for a particular technique. In this case the results in the second column of Table 8 and the first graph of Figure 3 will be preferred by proponents of pure probability measures while those of the sixth column and fifth graph will be adopted by supporters of evidence theory, and those of the seventh and eighth columns and sixth and seventh graphs will be favoured by supporters of possibility theory. Those do not object too strongly to interval methods may settle for the results of the third, fourth, fifth and ninth to eleventh columns (the second, third, fourth and eighth to tenth graphs), and those who prefer the eclectic approach of mixing formalisms should like the results in the tenth and eleventh columns which correspond to the ninth and tenth graphs.

Finally thought might be given to what the results are to be used for, and the choice made on the

basis of which are most useful. In this case it may be of little use having a set of values which contain many identical entries, an argument which with Figure 3 suggests that the results in the second, third, sixth and tenth columns (first, second, fifth and ninth graphs) are less useful than the others since these have a value of 1 for any constraint set containing $C1$. On the other hand this could be acceptable as a clear indication of the necessity of having structures conformant with $C1$. A similar argument might rule out the seventh, eighth, ninth and eleventh columns (sixth, seventh, eighth and tenth graphs), which have a value of zero for any constraint set that includes $C1$. Another way of choosing a method stems from the following argument. The prediction of protein topology by any theoretical means is only a part of the whole process which will also include a practical analysis which will test out the predicted possibilities as far as possible. Clearly, if a tedious set of experiments are required in order to reject each possible structure in a set, it would be advantageous to start with the smallest possible set of structures. This suggests using the number of possible structures associated with a set of constraints as a measure of its validity. Shirazi et al. (1990) supply the number of structures associated with seven of the 64 possible sets of constraints, and the order of these, based upon the number of possible structures, agrees with that obtained from the results in the fourth and fifth columns (third and fourth graphs). The solution ranked 'first' is the smallest since it conforms to the largest set of constraints.

6 Conclusions

It is to be hoped that this exploration of different approaches will be useful in several ways. Firstly it extends the comparative study of the use of differing uncertainty handling techniques (Heckerman 1990), (Heckerman and Shwe 1993), (Saffiotti et al. 1994) to cover a new problem. This problem contains a number of different types of uncertainty that must be modelled, and the fact that different models seem appropriate from different points of view provides empirical evidence for the validity of work on the different models. In addition, since no model seems to naturally model every aspect of the uncertainty, the protein topology problem provides motivation for working on using the different models in combination in the same problem. Further to this motivation, this paper has suggested some means of combining different methods within one problem, and, using results generated using the implementation of this work in the Mummu system, has discussed the use of combinations of formalisms in solving a real problem. Thus the paper has provided some empirical demonstration that using combinations of formalisms is both feasible and useful.

Acknowledgement

This work was partially supported by ESPRIT Basic Research Project 3085 DRUMS while I was a PhD. student in the Department of Electronic Engineering, Queen Mary and Westfield College, London. I am grateful to Dominic Clark at the ICRF for help in understanding the domain.

References

[1] Branden, C. 1990. Relation between structure and function of α/β proteins, *Quarterly review of biophysical chemistry*, **13**:317–338.

[2] Clark, D. A., Shirazi, J. and Rawlings, C. J. 1992. Protein topology prediction through constraint-based search and the evaluation of topological folding rules, *Protein Engineering*, **4**:751–760.

[3] Clark, D. A., Rawlings, C. J., Shirazi, J., Veron, A. and Reeve, M. 1993. Protein topology prediction through parallel constraint logic programming, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pp 83–91, Washington D.C.

[4] Cohen, F. E. and Kuntz, I. D. 1987. Prediction of the three diemnsional structure of human growth hormone *Proteins, structure, function, and genetics*, **2**:162–167.

[5] Dubois, D. and Prade, H. 1988. Modelling uncertainty and inductive inference: a survey of recent non-additive probability systems. *Acta Psychologica*, **68**:53–78.

[6] Heckerman, D. E. 1990. An empirical comparison of three inference methods. In *Uncertainty in Artificial Intelligence 4*, eds. R. D. Shachter,, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, pp 283–302, Elsevier, Amsterdam.

[7] Heckerman, D. E. and Shwe, M. 1993. Diagnosis of multiple faults: a sensitivity analysis, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pp 80–87, Washington D. C.

[8] Parsons, S. 1993. Qualitative methods for reasoning under uncertainty, PhD Thesis, Queen Mary and Westfield College, London.

[9] Richardson, J. S. 1976. Handedness of crossover connections in β sheets *Proceedings of the National Academy of Science*, **73**:2619–2623.

[10] Richardson, J. S. 1981. The anatomy and taxonomy of protein structure, *Advances in Protein Chemistry*, **34**:167–339.

[11] Saffiotti, A. and Umkehrer, E. 1991. Pulcinella: a general tool for propagating uncertainty in valuation networks *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pp323–331, Los Angeles.

[12] Saffiotti, A., Parsons, S. and Umkehrer, E. 1994. A case study in comparing uncertainty management techniques, *Microcomputers in Civil Engineering — Special Issue on Uncertainty in Expert Systems*, to appear.

[13] Shenoy, P. P. 1991. A valuation-based language for expert systems, *International Journal of Approximate Reasoning*, **3**:383–411.

[14] Shirazi, J., Clark, D. A. and Rawlings, C. J. 1990. Constraint-based reasoning in molecular biology: predicting protein topology from secondary structure and topological constraints, BCU/ICRF Technical Report.

[15] Smets, P. and Hsia, Y-T. 1990. Default reasoning and the Transferable Belief Model, *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pp529–537, Boston.

[16] Sternberg, M. J. and Thornton, J. M. 1977. On the conformation of proteins: an analysis of β-pleated sheets, *Journal of Molecular Biology*, **110**:269–283.

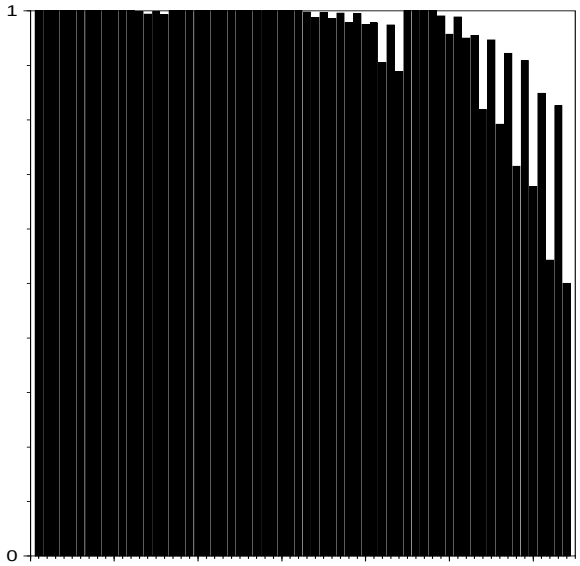
[17] Taylor, W. R. and Green, N. M. 1989. The predicted secondary structure of the nucleotide binding sites of six cation-transporting ATPases leads to a probable tertiary fold *European Journal of Biochemistry*, **179**:241–248.

[18] Walker, J. E., Saraste, M., Runswick, W. J. and Gay, N. J. 1982. Distantly related sequences in the α and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold *The EMBO Journal*, **1**:945–951.

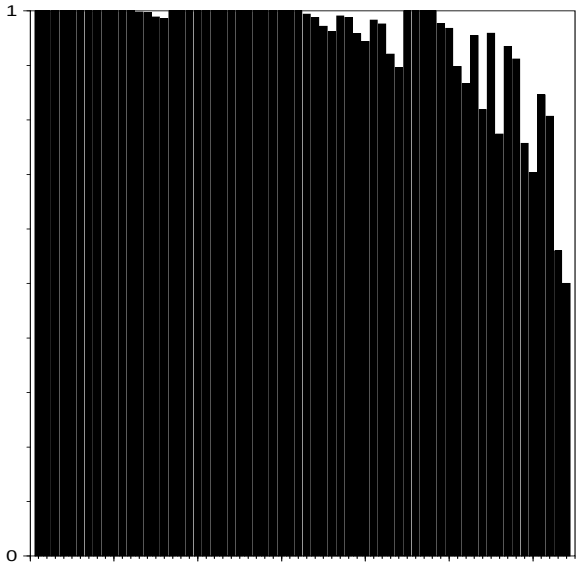
Appendix

Constraint Set	$p(V)$	$p(V)$	$p(V)$	$p(V)$	bel(V)	$\Pi(\neg V)$	$\Pi(\neg V)$	$\Pi(\neg V)$	$p(V)$	$\Pi(\neg V)$
$\{C1, C2, C3, C5, F1, F2\}$	1.0	[1.0 1.0]	[0.984 1.0]	[0.972 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, C5, F1\}$	1.0	[1.0 1.0]	[0.917 1.0]	[0.876 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, C5, F2\}$	1.0	[1.0 1.0]	[0.982 1.0]	[0.962 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, C5\}$	1.0	[1.0 1.0]	[0.905 1.0]	[0.837 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, C5, F1, F2\}$	1.0	[1.0 1.0]	[0.963 1.0]	[0.939 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, C5, F1\}$	1.0	[1.0 1.0]	[0.824 1.0]	[0.759 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, C5, F2\}$	1.0	[1.0 1.0]	[0.830 1.0]	[0.759 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, C5\}$	1.0	[1.0 1.0]	[0.802 1.0]	[0.696 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C5, F1, F2\}$	1.0	[1.0 1.0]	[0.956 1.0]	[0.922 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C5, F1\}$	1.0	[1.0 1.0]	[0.794 1.0]	[0.709 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C5, F2\}$	1.0	[1.0 1.0]	[0.949 1.0]	[0.896 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C5\}$	1.0	[1.0 1.0]	[0.769 1.0]	[0.640 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C2, C3, C5, F1, F2\}$	0.999	[0.996 1.0]	[0.663 1.0]	[0.524 1.0]	0.997	0.03	0.052	[0.125 0.125]	0.999	[0 0.078]
$\{C2, C3, C5, F1\}$	0.994	[0.994 1.0]	[0.261 1.0]	[0.185 0.999]	0.990	0.303	0.052	[0.125 0.125]	[0.996 1]	[0.125 0.125]
$\{C2, C3, C5, F1\}$	0.999	[0.983 0.995]	[0.630 1.0]	[0.445 1.0]	0.997	0.03	0.052	[0.125 0.125]	[0.999 1]	[0 0.125]
$\{C2, C3, C5\}$	0.993	[0.976 0.995]	[0.234 1.0]	[0.141 0.999]	0.988	0.303	0.052	[0.125 0.125]	[0 0.994]	[0.125 0.125]
$\{C1, C2, C3, F1, F2\}$	1.0	[1.0 1.0]	[0.975 1.0]	[0.954 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, F1\}$	1.0	[1.0 1.0]	[0.876 1.0]	[0.810 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, F2\}$	1.0	[1.0 1.0]	[0.972 1.0]	[0.938 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, C3, F1, F2\}$	1.0	[1.0 1.0]	[0.859 1.0]	[0.756 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, F1, F2\}$	1.0	[1.0 1.0]	[0.932 1.0]	[0.877 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, F1\}$	1.0	[1.0 1.0]	[0.712 1.0]	[0.594 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2, F2\}$	1.0	[1.0 1.0]	[0.922 1.0]	[0.838 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C2\}$	1.0	[1.0 1.0]	[0.681 1.0]	[0.516 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, F1, F2\}$	1.0	[1.0 1.0]	[0.944 1.0]	[0.902 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, F1\}$	1.0	[1.0 1.0]	[0.750 1.0]	[0.654 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3, F2\}$	1.0	[1.0 1.0]	[0.935 1.0]	[0.870 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C3\}$	1.0	[1.0 1.0]	[0.722 1.0]	[0.578 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C5, F1, F2\}$	1.0	[1.0 1.0]	[0.901 1.0]	[0.840 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C5, F1\}$	1.0	[1.0 1.0]	[0.620 1.0]	[0.520 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C5, F2\}$	1.0	[1.0 1.0]	[0.887 1.0]	[0.793 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, C5\}$	1.0	[1.0 1.0]	[0.585 1.0]	[0.440 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C2, C3, F1, F2\}$	0.998	[0.990 0.998]	[0.558 1.0]	[0.398 1.0]	0.993	0.03	0.052	[0.125 0.125]	0.998	[0 0.078]
$\{C2, C3, F1\}$	0.988	[0.985 0.998]	[0.185 1.0]	[0.120 0.999]	0.973	0.303	0.052	[0.125 0.125]	[0.988 1]	[0.125 0.125]
$\{C2, C3, F2\}$	0.997	[0.955 0.988]	[0.522 1.0]	[0.325 1.0]	0.992	0.03	0.052	[0.125 0.125]	[0.997 1]	[0 0.125]
$\{C2, C3\}$	0.986	[0.937 0.986]	[0.163 1.0]	[0.090 0.999]	0.969	0.303	0.052	[0.125 0.125]	[0 0.985]	[0.125 0.125]
$\{C2, C5, F1, F2\}$	0.996	[0.985 0.997]	[0.407 1.0]	[0.275 1.0]	0.990	0.03	0.052	[0.125 0.125]	0.996	[0 0.078]
$\{C2, C5, F1\}$	0.979	[0.978 0.997]	[0.110 1.0]	[0.072 0.996]	0.959	0.666	0.052	[0.125 0.125]	[0.982 1]	[0.125 0.125]
$\{C2, C5, F2\}$	0.995	[0.934 0.982]	[0.372 1.0]	[0.216 1.0]	0.988	0.03	0.052	[0.125 0.125]	[0.995 1]	[0 0.125]
$\{C2, C5\}$	0.975	[0.909 0.979]	[0.096 0.996]	[0.054 0.996]	0.953	0.666	0.052	[0.125 0.125]	[0 0.977]	[0.125 0.125]
$\{C3, C5, F1, F2\}$	0.978	[0.970 0.995]	[0.455 1.0]	[0.328 1.0]	0.979	0.03	0.211	[0.125 0.25]	0.992	[0 0.078]
$\{C3, C5, F1\}$	0.905	[0.957 0.994]	[0.130 0.992]	[0.091 0.996]	0.918	0.303	0.263	[0.25 0.25]	[0.965 1]	[0.25 0.25]
$\{C3, C5, F2\}$	0.974	[0.877 0.965]	[0.419 1.0]	[0.262 1.0]	0.977	0.03	0.211	[0.125 0.25]	[0.990 1]	[0 0.25]
$\{C3, C5\}$	0.889	[0.833 0.958]	[0.114 0.992]	[0.068 0.996]	0.906	0.303	0.263	[0.25 0.25]	[0 0.955]	[0.25 0.25]
$\{C1, F1, F2\}$	1.0	[1.0 1.0]	[0.854 1.0]	[0.760 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, F1\}$	1.0	[1.0 1.0]	[0.511 1.0]	[0.394 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1, F2\}$	1.0	[1.0 1.0]	[0.835 1.0]	[0.697 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C1\}$	1.0	[1.0 1.0]	[0.475 1.0]	[0.321 1.0]	1.0	0.03	0	[0 0]	1.0	[0 0]
$\{C2, F1, F2\}$	0.991	[0.96 0.993]	[0.306 1.0]	[0.186 1.0]	0.973	0.03	0.052	[0.125 0.125]	0.990	[0 0.078]
$\{C2, F1\}$	0.957	[0.944 0.992]	[0.073 0.996]	[0.045 0.996]	0.891	0.666	0.052	[0.125 0.125]	[0.954 1]	[0.125 0.125]
$\{C2, F2\}$	0.989	[0.842 0.954]	[0.275 1.0]	[0.142 1.0]	0.969	0.03	0.052	[0.125 0.125]	[0.986 1]	[0 0.125]
$\{C2\}$	0.950	[0.789 0.945]	[0.064 0.996]	[0.033 0.996]	0.875	0.666	0.052	[0.125 0.125]	[0 0.941]	[0.125 0.125]
$\{C3, F1, F2\}$	0.955	[0.923 0.986]	[0.349 1.0]	[0.227 1.0]	0.945	0.03	0.211	[0.125 0.25]	0.981	[0 0.078]
$\{C3, F1\}$	0.819	[0.894 0.984]	[0.088 0.991]	[0.057 0.996]	0.781	0.303	0.263	[0.25 0.25]	[0.912 1]	[0.25 0.25]
$\{C3, F2\}$	0.947	[0.727 0.911]	[0.316 1.0]	[0.176 1.0]	0.938	0.03	0.211	[0.125 0.25]	[0.973 1]	[0 0.25]
$\{C3\}$	0.792	[0.651 0.896]	[0.077 0.991]	[0.042 0.996]	0.75	0.303	0.263	[0.25 0.25]	[0 0.889]	[0.25 0.25]
$\{C5, F1, F2\}$	0.922	[0.889 0.980]	[0.225 0.999]	[0.144 0.999]	0.918	0.03	0.211	[0.125 0.25]	0.972	[0 0.078]
$\{C5, F1\}$	0.715	[0.848 0.976]	[0.050 0.975]	[0.033 0.987]	0.672	0.939	0.474	[0.375 0.375]	[0.874 1]	[0 0.625]
$\{C5, F2\}$	0.909	[0.64 0.873]	[0.201 0.998]	[0.109 0.999]	0.906	0.03	0.211	[0.125 0.25]	[0.96 1]	[0 0.125]
$\{C5\}$	0.678	[0.554 0.851]	[0.043 0.975]	[0.025 0.987]	0.625	0.939	0.474	[0.375 0.375]	[0 0.842]	[0.375 0.375]
$\{F1, F2\}$	0.849	[0.75 0.947]	[0.157 0.999]	[0.092 0.999]	0.781	0.03	0.211	[0.125 0.25]	0.927	[0 0.078]
$\{F1\}$	0.543	[0.677 0.937]	[0.032 0.972]	[0.020 0.985]	0.125	1	0.789	[0.625 0.875]	[0.72 1]	[0 0.625]
$\{F2\}$	0.826	[0.4 0.72]	[0.139 0.999]	[0.068 0.999]	0.75	0.03	0.211	[0.125 0.25]	[0.9 1]	[0 0.125]
$\{\}$	0.5	[0.318 0.682]	[0.028 0.972]	[0.015 0.985]		1	1	[1 1]	[0 0.667]	[1 1]

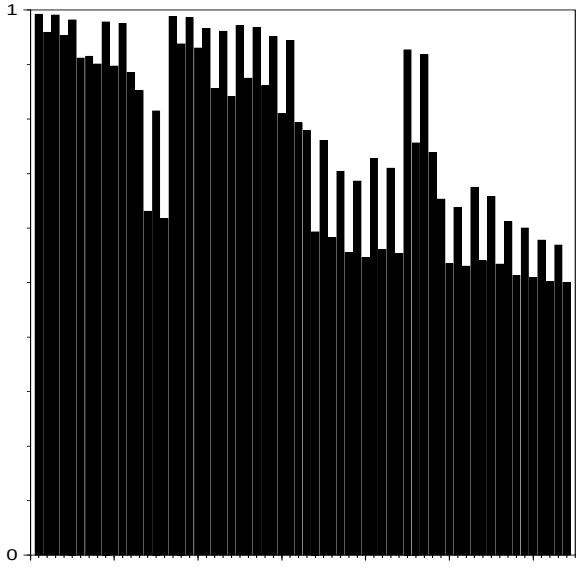
Table 8: Results of the experiment in assessing the validity of sets of constraints



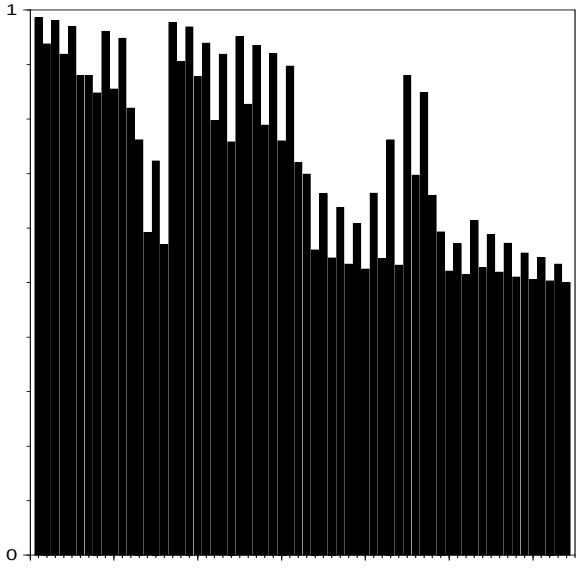
Results based on the probabilities from Table 8



Results based on the interval probabilities from Table 6

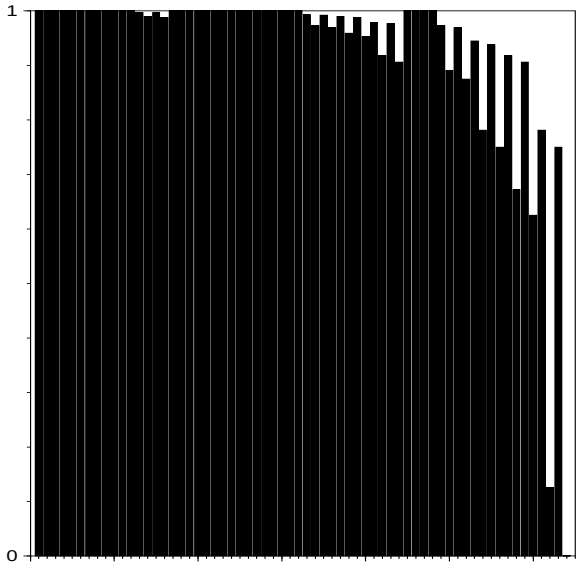


Results based on the interval probabilities from Table 2 and Table 3

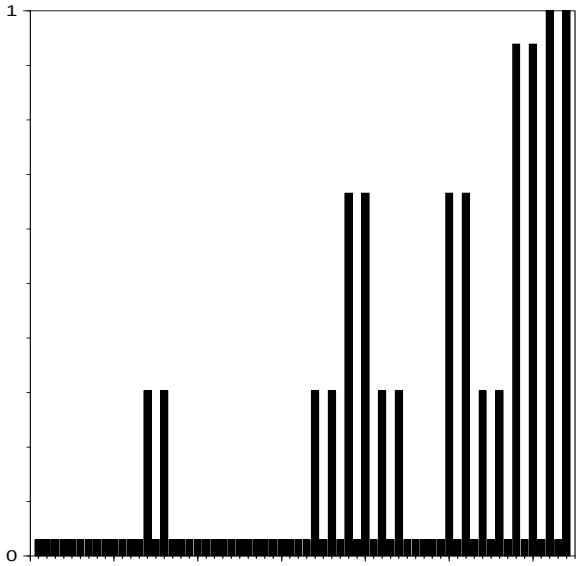


Results based on the interval probabilities from Table 2 and Table 3

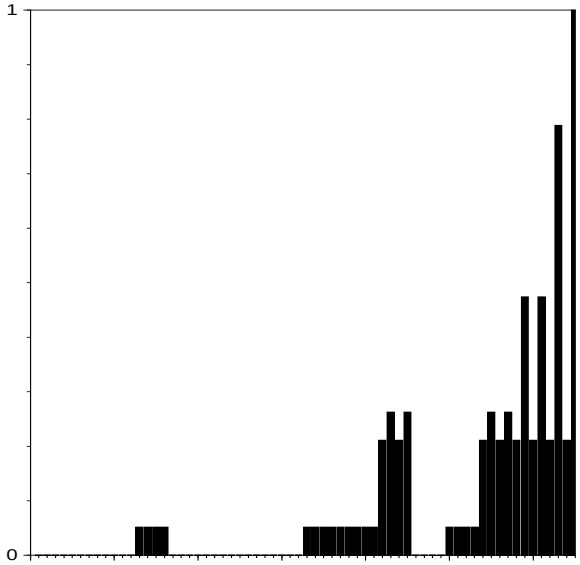
Figure 3: Results of the experiment in assessing the validity of sets of constraints



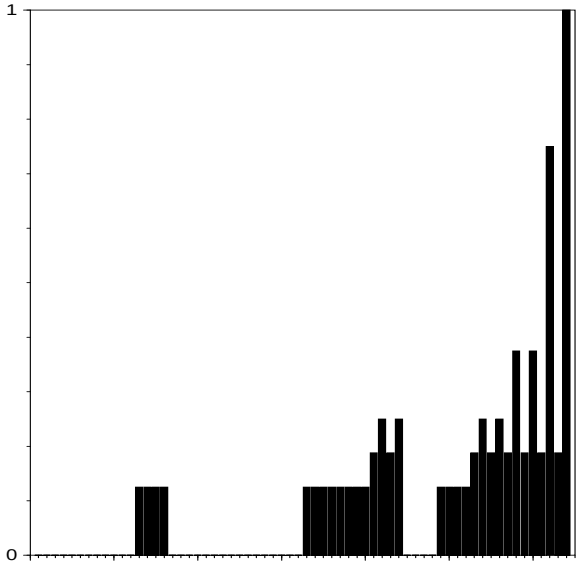
Results based on the beliefs from Table 7



Results based on the possibilities from Table 4

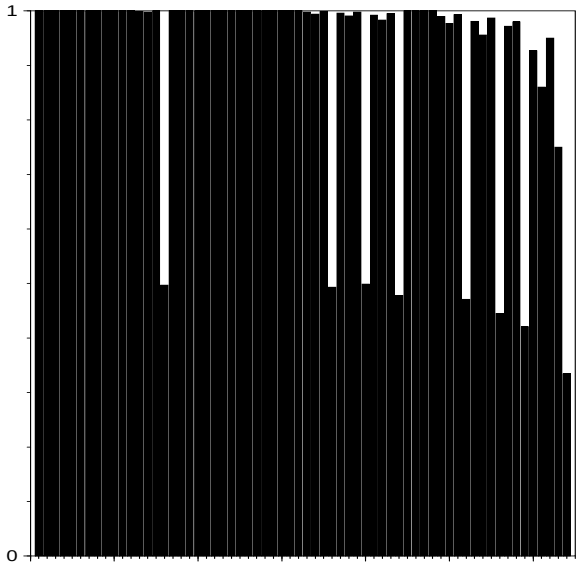


Results based on the possibilities from Table 5

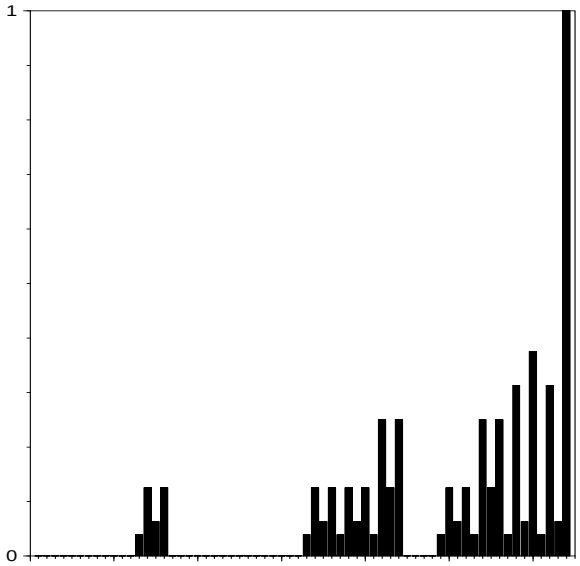


Results based on the interval possibilities from Table 6

Figure 3(cont): Results of the experiment in assessing the validity of sets of constraints



Results based on the probabilities from Table 6 and beliefs from Table 7



Results based on the possibilities from Table 6 and beliefs from Table 7

Figure 3(cont): Results of the experiment in assessing the validity of sets of constraints