STATISTICAL EXTRACTION OF PROTEIN SURFACE ATOMS BASED ON A VOXELISATION METHOD

Ling Wei Lee and Andrzej Bargiela
School of Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih
Selangor Darul Ehsan, Malaysia

KEYWORDS
Protein Surface Atoms, Protein Surface Analysis, Space Voxelisation

ABSTRACT

Proteins play a vital role in maintaining the balance of bodily functions in all living beings. However their functional properties are difficult to predict since they depend not only on the sequence of constituent amino acids but also on the 3D folding of the protein. This paper presents a new statistical method for extraction of surface atoms of a protein. The method is based on space-voxelisation and generalizes our previous deterministic method by repeating the surface extraction process for various orientations of a protein; so as to achieve a statistical consensus about surface atoms. Based on the experimental study we have established an optimal range of values of voxel occupancy for the selection of surface atoms; with optimality defined as a maximum coincidence of extracted surface atoms from the protein presented to the algorithm in 13 different orientations. The results show that the voxel occupancy threshold of between >40% and >50% allows our algorithm to extract surface atoms with high degree of confidence.

INTRODUCTION

Proteins stemming from the same ancestor may undergo evolutionary change with some parts of the protein remaining unaffected while other parts modifying their structure. The assessment of the affinity of so evolved proteins is assessed through two broad classes of methods: a) methods that analyse amino acid sequences; and b) methods that analyse 3D structures of proteins.

Protein sequence comparison takes the sequence strings of proteins and attempts to align the two strings so as to find the largest common subsequence between the strings. There are many algorithms available to perform such alignment. Some algorithms are single sequence based like the Smith-Waterman algorithm [1] or Needleman-Wunch algorithm [2], while others attempted to perform multiple protein sequence alignments [3]. Pei [4] presented a paper reviewing methodologies and advances within this field. Systems like PSI-BLAST and CLUSTALW are commonly used by researchers for large scale comparison of protein sequences with each system having its own advantages.

Sequential comparison – depending on the type of algorithm implemented – can be very efficient computationally. However, what such comparisons cannot achieve is the identification of inheritance of functional properties between the compared proteins. A sequence order can only provide information such as the type of elements contributing to the protein and one has to make assumptions (or perform additional study) on how these changes affect the folding of a protein. However the derivation of protein folding from the basic sequence information is a very challenging problem due to the many degrees of freedom of the molecular structure.

Consequently, an alternative approach of experimental, crystallography-based discovery of 3D structures has established itself as a practical method of identifying overall protein structure. Consequently a new field of proteomic research has emerged, that of identifying common sub-structures in proteins that have similar functional properties. This research is now supported by the availability of databases containing a complete hierarchy of existing proteins classified based on their structures. The SCOP (Structural Classification Of Proteins) database [5] gives "a comprehensive ordering of all proteins of known structure according to their evolutionary and structural relationships". SCOP orders protein entries based on the following levels : Species, Protein, Family, Superfamily, Fold and Class. Another existing database that has been widely in use is the CATH (Class, Architecture, Topology, Homologous superfamily) database [6]. Class describes the secondary structure compositions of domains, Architecture defines the shape, Topology gives the sequential connectivity while Homologous superfamily groups together proteins with structures in the same Topology. Such classification databases are helpful to analysts and bioinformaticians studying the relationship between the proteins and the conservation of unique features in the structures.

However there have been only few attempts to analyse proteins in terms of their surface atoms structure and composition. According to [7], "protein surface comparison is a hard computational challenge and

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

evaluated methods allowing the comparison of protein surfaces are difficult to find". Structural and sequence comparisons may reveal patterns that remain the same throughout evolution. Still, these signatures do not necessarily guarantee the same functions for the same evolutionary line of proteins. Protein surface comparison on the other hand may not be able to reveal inherited traits; however it is able to detect similar areas that provide the same reaction to external agents regardless of whether the protein comes from the same family. There may exist proteins from the same family containing binding sites with different characteristics while proteins from different ancestors may evolve to contain sites with similar features. Figure 1 shows the comparison for the binding sites.

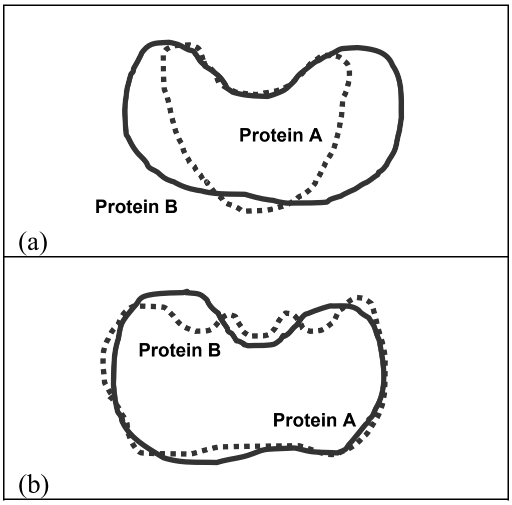

Figure 1 : Comparison of binding sites for proteins of (a) Different ancestors with similar binding sites (b) Same ancestor with different binding sites

As can be seen from the illustrations there is a need for protein surface analysis to determine the characteristics of potential dock sites. The following sections provide more detailed information on the algorithm developed and a brief description of past implementations and experiments.

BACKGROUND

One of the most commonly used methods for the study of protein surfaces is the Connolly method [8] whereby a water-molecule-sized probe is used to inspect the surface of the protein. Wherever the probe fits in the cavity it identifies with a high probability a potential binding site. In [9] Jiang and Kim used cube representations for the soft docking of proteins. By using two specific molecular complexes the authors show that geometric docking alone with conformational changes is sufficient to determine the correct binding agents. However this method has its limitations when applied to various protein complexes since there are proteins that do not form rigid docking bonds.

Cubic grid representations of proteins have been used in the past due to the ease of implementation and manipulation of the 3D structures. In this research we adopt this representation and attempt to overcome the inherent limitations of this topological construct. The basic algorithm for extracting surface voxels of a protein may be summarized as follows:

1. Pre-process the protein (details given in the next section) and compile the required information in a new file.
2. Impose the protein into the experimental cubic grid space using the compiled data.
3. Tessellate the experimental space until the smallest unit voxel size is achieved (in this case the value is 40 units).
4. Extract all voxels containing the presence of any of the atoms within the protein. This stage also includes a checking for the occupancy percentage of each voxel. For example, a >20% value means that voxels in which >20% of its space is occupied by an atom/groups of atoms.
5. Filter all voxels having 1 or more faces fully exposed. These are categorised as the surface voxels.
6. Based on the surface voxels the surface atoms are then extracted. This is done by compiling the atoms contained within the surface voxels.

The main problem encountered when using cubic grids relates to the orientation of the protein within the experimental space. Any arbitrary rotation within the grid space leads to a different set of voxels being chosen post-tessellation. At one particular orientation an atom may occupy an entire voxel while at another orientation of the protein the same atom may take up two or more voxels. Figure 2 provides an illustration.

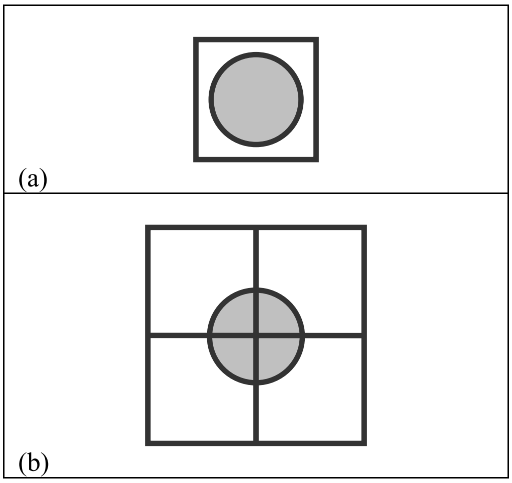

Figure 2 : Positioning of an atom before rotation (a) and after rotation (b). Initially the atom occupies 1 voxel and after rotation it is partially contained in 4 voxels.

As a consequence of the above, a deterministic identification of surface atoms for one orientation of a protein typically returns results that are quite different from those obtained for another orientation. Hence, in this work, we relax the notion of deterministic identification of surface-atoms and introduce an alternative notion of statistical identification. This requires consideration of several orientations of the protein (which allows building of appropriate statistics) and the optimisation of the threshold value that determines when an atom can be considered as belonging to a specific voxel. Selection of a representative set of orientations of a protein is a compromise between the computational effort and the quality of the accumulated statistics. In the following section we provide a rationale for our selection. The determination of a suitable threshold for voxel occupancy has been performed as an empirical investigation (described in the following section) and the conclusions are deemed to be general.

THE ALGORITHM

The algorithm for the extraction of protein surface atoms extends a previous development [10]. Input files for the program are obtained from the RCSB Protein Data Bank (PDB) in PDB format. A pre-processing stage is carried out to extract the required information including the spatial coordinates, the atom element and all residues data following which these are stored into a new file together with additional information of Van der Waals radius and the electronegativity of the atoms. Compiling these data into a single file is important as the program has been configured and optimised to process all input files in the predefined format. In the original PDB file, the spatial coordinates provide the actual positioning of the atoms in the protein. However for the purpose of this research these coordinates have been scaled to larger values for ease of processing and analysis. For example, an atom is given the coordinates of (-8.371, 23.633, 40.487) in the PDB file. Upon pre-processing the new values for the atom are (139.35, 310.78, 541.72). Scaling up the values makes visual analysis easier while at the same it also ensures that important specifics do not get overlooked – small details that play important contributions are then easily identified. The Van der Waals radiuses for the elements lie between 1Å and 3Å and these values were also scaled up by ten-fold. Therefore the diameter range for the atoms is valued from 20Å to 60Å. By taking the estimated average of the diameter range it is thus concluded that a value of 40Å will be used as the value for the smallest unit voxel size. For easier understanding all future references to this smallest unit voxel size shall be termed as 40 units instead of 40Å so as to coincide with the notion of grid spaces and voxels.

In the algorithm the protein is first imposed into a cubic grid experimental space bound by Cartesian coordinates. Any object casted into this space is defined by a set of (x, y, z) coordinates. A checking is carried out to determine if any of the spatial coordinates of the atoms are in the negative range. If there exists any negative values, the protein is then translated into the positive regions. The spatial coordinates data are then stored into temporary memory for easy reference.

Following that the protein will be rotated based on a set of defined rotations in all X, Y and Z dimensions. The selected range of angles includes 20°, 30°, 45°, 60° in all dimensions. These angles were chosen based on the understanding that any atom will experience the largest change of orientation at the aforementioned angles. The affine transform is implemented for the rotation of the protein. For a detailed explanation of the matrices involved the reader is referred to [10]. A reference point is first taken from the protein for use in the affine transform whereby all atoms are translated to the origin based on this reference point, rotated and re-translated back into the test space. However it should be noted that some post-rotation atoms may occupy voxels in the negative regions. Again, another round of checking is carried out to ensure that all the atoms are translated to the positive domains.

The transformed protein with all positive spatial coordinate values is then ensured that it is confined within the experimental space. A space tessellation algorithm based on the bisection is carried out to slice the experimental space into cubes of equal sizes with the smallest size being 40 units. This value is chosen based on a general analysis of the diameter of protein atoms and it was found that the average value agrees at a value of 40. A significant modification has been made to the original implementation to cater for the process of voxel occupancy checking. Previously any voxel containing the presence of any atom – regardless of the percentage of occupancy of the atom within that voxel – is shortlisted. For the purpose of extracting the surface voxels, a series of voxel occupancy percentage is introduced ranging from >0% to 100% with increments of 5%, therefore giving a total of 21 cases. The following table gives the list of experimental rotational conditions.

Table 1 : Various Rotations Used for Protein Experiments for Each Voxel Occupancy

Rotation Angles (x, y, z)
(0, 0, 0)
(20, 0, 0)
(30, 0, 0)
(45, 0, 0)
(60, 0, 0)
(0, 20, 0)
(0, 30, 0)
(0, 45, 0)
(0, 60, 0)
(0, 0, 20)
(0, 0, 30)
(0, 0, 45)
(0, 0, 60)

By multiplying the 21 cases with the list of 13 rotations there are thus a total of 273 experiments to be performed. In each iteration, the algorithm checks the percentage occupancy of voxels by protein atoms. All voxels meeting the specific threshold value are marked and stored for subsequent analysis.

To extract the surface atoms voxels are checked if they have exposed surfaces. If any of the surfaces of a voxel is not connected to any other voxel then it is categorised as a surface voxel. Figure 3 gives an illustration.

Surface atoms are then extracted by reference to the surface voxels identified in the previous step. The surface atoms information is then stored for the final stage of compiling statistics. Complete execution of all the experimental conditions is followed by a post-processing stage which collects all the generated output for analysis. The main challenge lies in ensuring the consistency of the extracted surface atoms across all different orientations. Common atoms shared by all rotation sets based on percentage of occupancy are shortlisted and the extraction accuracy is then calculated.

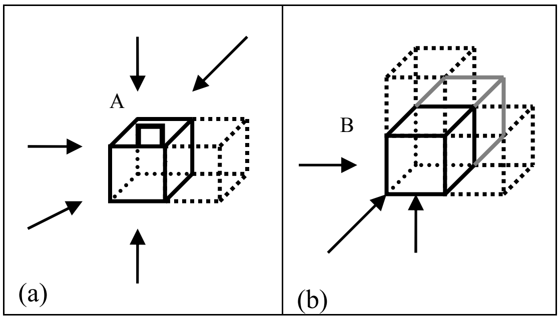

Figure 3 : (a) 5 faces exposed for voxel A. Therefore the exposure value is 5. (b) 3 faces exposed for voxel B. Therefore the exposure value is 3.

RESULTS

The algorithm is iterated 273 times to generate sets of output for each experimental condition. Each occupancy percentage contains 13 sets of results – one for every rotation in all X, Y and Z dimensions as well as for the original non-rotated file.

For each occupancy percentage, atoms that are common across all 13 orientations are first extracted. The common atoms are then compared to the total number of extracted surface atoms for each rotation and the percentage is calculated. Figure 4 gives the plot of the percentage of common atoms versus the occupancy percentage of voxels. There are 13 lines altogether in the graph – each line represents each of the different orientations.

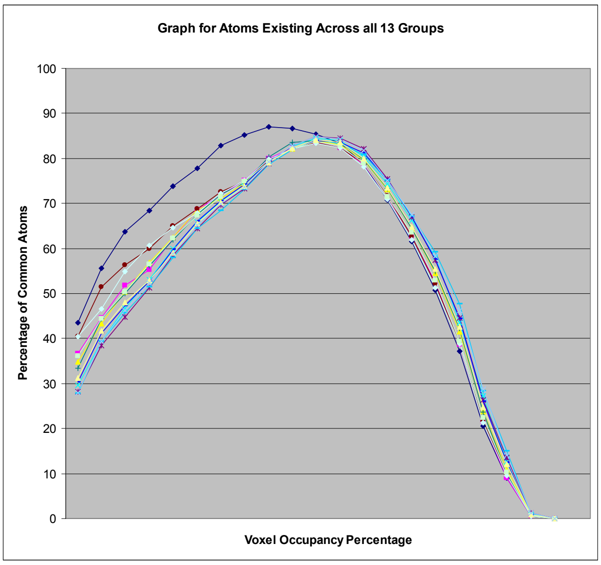

Figure 4 : Plots for each rotation giving % of common atoms identified in all 13 orientations

The horizontal axis begins with 0% voxel occupancy to 100% voxel occupancy. As can be seen from the graph the extracted common atoms peaks at about 50%. The following table gives a list of all the rotation cases with their corresponding peak percentages.

Table 2 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in all 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	87.06
(20, 0, 0)	50	83.55
(30, 0, 0)	50	83.96
(45, 0, 0)	50	84.52
(60, 0, 0)	50	84.84
(0, 20, 0)	50	83.55
(0, 30, 0)	50	83.89
(0, 45, 0)	50	84.00
(0, 60, 0)	50	84.24
(0, 0, 20)	50	83.45
(0, 0, 30)	50	83.72
(0, 0, 45)	50	84.03
(0, 0, 60)	50	84.84

From the graph and the given table it can be concluded that at >50% voxel occupancy the extraction of the protein surface atoms reaches its optimum. At the same time, the voxels processed are fewer in number and hence this increases the processing speed. Thus it can be said that a compromise has been reached at >50% for the minimum number of voxels used while retaining the best surface atoms extractions.

However there are still some surface atoms that may be left out due to differences in the orientations. The condition for clustering of the surface atoms is therefore relaxed and atoms existing across 12 and even 11 cases are taken into consideration as these are atoms with high probabilities of existing on the surface and which may play significant roles. Atoms that have been marked as existing across all 13 cases are classified as definite entries. The following tables show the output obtained for two relaxed scenarios.

Table 3 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in 13 and 12 out of 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	97.02
(20, 0, 0)	45	95.16
(30, 0, 0)	50	95.11
(45, 0, 0)	50	95.41
(60, 0, 0)	50	95.80
(0, 20, 0)	50	94.72
(0, 30, 0)	45	95.93
(0, 45, 0)	50	95.39
(0, 60, 0)	50	95.17
(0, 0, 20)	45	94.85
(0, 0, 30)	50	95.28
(0, 0, 45)	50	95.76
(0, 0, 60)	50	96.18

Table 4 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in 13, 12 and 11 out of 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	99.05
(20, 0, 0)	45	97.99
(30, 0, 0)	45	97.67
(45, 0, 0)	50	97.90
(60, 0, 0)	50	98.38
(0, 20, 0)	45	97.71
(0, 30, 0)	45	98.38
(0, 45, 0)	45	98.04
(0, 60, 0)	45	97.88
(0, 0, 20)	45	97.87
(0, 0, 30)	45	98.16
(0, 0, 45)	45	98.65
(0, 0, 60)	45	98.68

The relaxation of the condition leads to a higher number of atoms being classified as surface atoms. From the tables it can be gathered that the best extractions for all rotation cases exist between >40% and >50% of voxel occupancy. This corresponds well to the strict filtering of the atoms existing across all 13 cases whereby the optimum occupancy percentage is >50%.

The experiment has been repeated for conditions with the inclusion of atoms existing across 13, 12, 11, 10 and 9 cases and the optimum voxel occupancy percentage for the best extractions remains the same. Also, it can be concluded that the relaxation of the condition leads to higher extraction percentages, of which will gradually becomes closer to 100% as more and more atoms are included.

DISCUSSION

The use of voxels in the extraction of protein surface atoms proves to be an effective and efficient method. The tessellation of the experimental space is computationally efficient and it is easy to determine whether a voxel contains any atom. Rotation was used to generate various orientations of the protein and this was chosen over translation due to the challenge posed. Affine transformations are involved in calculating the new positions of atoms when a protein has been rotated. The working hypothesis was that a rotational transform affects various sections of the protein to different degrees and this is largely dependent on the distance of atoms from the center of the protein. Future work will verify this hypothesis by comparing its performance against alternative transformations, e.g. discrete translation along all x, y, and z axes.

To determine the surface atoms one needs to identify surface voxels which involves only checking 6 surfaces of individual voxels and can be implemented very efficiently. It was found that the best extraction of voxels occurs with the occupancy threshold set to between >40% and >50%. This value is deemed to be a function of the topology of the voxels, so as long as one uses cubical voxels this threshold can be used in subsequent studies to produce the optimum surface atom extraction rates.

As the filtering condition is relaxed more atoms are included thus increasing the number of surface atoms being shortlisted. The extraction percentages show consistency in between different orientations. This correlates well with the relaxation of the condition. The algorithm executes efficiently with each iteration taking less than 10 seconds. Generally the surface atoms of a protein can be determined in under 2 minutes on a 2GHz single core PC when the optimum occupancy percentage is used.

The algorithm does not yield a 100% accurate extraction but it does provide satisfactory results with up to 97% accuracy. The atoms extracted are highly probable surface atoms but at the same time, there is also a possibility that an extracted atom may belong to the layer below the surface layer. However this does not pose any major problems to the analysis stage. At this stage the results are validated through visual inspection of the protein images since the PDB does not hold the surface atoms information.

A good extraction of the surface atoms aids greatly in the study of the features and characteristics of a protein. By applying methods that determine relationships between the atoms, the motifs and key signatures can be identified especially the composition that contributes to the attributes of protein docking sites. Predictions can be made of proteins with similar motifs that may bind to the same ligands or binding agents. Furthermore, a classification system can be constructed that clusters proteins with similar signatures into the same groups. Such a system is able to provide new discoveries and insights on the functions of proteins. Bioinformations will find this system useful in the development of new drugs as the system is able to indicate side effects on other proteins that may caused by a particular drug.

This approach is to be implemented in full on protein sets obtained from the PDB. The output obtained will then be used for the analysis of protein surfaces in upcoming research work.

CONCLUSION

A new approach for the identification and extraction of protein surface atoms is presented here in which voxels are used as the main tool. A range of experimental conditions were used whereby the protein was tested in different orientations. The occupancy percentages of the voxels were checked and it was found that the optimal extraction occurs at somewhere between >40% and >50% of occupancy. Atoms existing across all 13 orientation cases display a fairly high extraction percentage at about 84%. As the condition is relaxed and atoms existing across both 12 and 11 orientations are included it can be seen that the extraction percentages increased to about 95% and 97% respectively. This approach is efficient and is able to produce a good output through the use of voxels. All extractions of surface atoms are to be used for analysis in ongoing research work.

REFERENCES

[1] Smith T., Waterman M. 1981. "Identification of Common Molecular Subsequences". *Journal of Molecular Biology*, No.147, 195-197.

[2] Needleman S., Wunsch C. 1970. "A General Method Applicable to The Search for Similarities in The Amino Acid Sequence of Two Proteins". *Journal of Molecular Biology*, No.48, 443-453.

[3] McClure M.A., Vasi T.K., Fitch W.M., 1994. "Comparative Analysis of Multiple Protein-Sequence Alignment Methods". *Molecular Biology and Evolution*, No.11 571-592.

[4] Pei J. 2008. "Multiple Protein Sequence Alignment". *Current Opinion in Structural Biology*, No.18, 382-386.

[5] Andreeva A., Howorth D., Chandonia J-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G., 2007. "Data Growth and Its Impact on The SCOP Database: New Developments". *Nucleic Acids Research*, 1-7.

[6] Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M., 1997. "CATH – A Hierarchic Classification of Protein Domain Structures". *Structure*, No.5, 1093-1108.

[7] Via A., Ferre F., Brannetti B., Helmer-Citterich M., 2000. "Protein Surface Similarities: A Survey of Methods to Describe and Compare Protein Surfaces". *Cellular and Molecular Life Sciences*, No.57, 1970-1977.

[8] Cao, J., Pham D.K., Tonge L., Nicolau D.V., 2002. "Predicting Surface Properties of Proteins on The Connolly Molecular Surface". *Smart Materials and Structures*, No.11, 772-777.

[9] Jiang F., Kim S.H., 1991. "'Soft Docking': Matching of Molecular Surface Cubes". *Journal of Molecular Biology*, No.219, 79-102.

[10] Lee L.W., Bargiela A., 2009. "Space-Partition Based Identification of Protein Docksites". *Proceedings of the 23rd European Conference on Modelling and Simulation (ECMS 2009)*, 848-854.

AUTHOR BIOGRAPHIES

LING WEI LEE was born in Kuala Lumpur and studied at the University of Nottingham Malaysia Campus where she took up Computer Science and obtained her honours degree in 2007. She worked for about a year as an analyst programmer with a local company before deciding to pursue her postgraduate studies. Her current research focuses on the applications of granular computing methods in the area of proteins. She is currently in her second year of research. She can be reached at Ling-Wei.Lee@nottingham.edu.my.

ANDRZEJ BARGIELA is Professor and Director of Computer Science at the University of Nottingham, Malaysia Campus. He is member of the Automated Scheduling and Planning research group in the School of Computer Science at the University of Nottingham. Since 1978 he has pursued research focused on processing of uncertainty in the context of modelling and simulation of various physical and engineering systems. His current research falls under the general heading of Computational Intelligence and involve mathematical modelling, information abstraction, parallel computing, artificial intellingence, fuzzy sets and neurocomputing.