Theor Chem Account (2008) 119:265–274
DOI 10.1007/s00214-007-0285-7

REGULAR ARTICLE

Classification of protein fold classes by knot theory and prediction of folds by neural networks: A combined theoretical and experimental approach

K. Ramnarayan · H. G. Bohr · K. J. Jalkanen

Received: 10 January 2007 / Accepted: 6 March 2007 / Published online: 15 May 2007

Abstract We present different means of classifying protein structure. One is made rigorous by mathematical knot invariants that coincide reasonably well with ordinary graphical fold classification and another classification is by packing analysis. Furthermore when constructing our mathematical fold classifications, we utilize standard neural network methods for predicting protein fold classes from amino acid sequences. We also make an analysis of the redundancy of the structural classifications in relation to function and ligand binding. Finally we advocate the use of combining the measurement of the VA, VCD, Raman, ROA, EA and ECD spectra with the primary sequence as a way to improve both the accuracy and reliability of fold class prediction schemes.

1 Introduction

Finding all the genes of the genome of an organism naturally leads to the question of what proteins these genes represent or correspond to and to what class they belong. Concerning the classification it seems obvious to classify the proteins according to their sequence of either nucleotide or amino acids—a task which is not straight forward due to the alignment problems etc. The classification according to sequence is more likely to tell about functionality rather than structure. Such classes are called families. However, the structure of a protein is much sought after in biotechnology, e.g., in drug-design.

We shall in this paper address the issue of making a rigorous structural classification of proteins and how to predict such classes of proteins from their sequences. Concerning general structural classifications it has been shown [1,2] that all the known three-dimensional protein structures can be grouped into a smaller number of characteristic structural classes consisting of domains from homologous proteins with a similar topological configuration of their backbones. These structural domains or the so-called folds of the proteins were introduced in order to clarify the notion of structural similarity. Such fold classes could contain entire proteins or well-defined sub-domains of proteins. Pascarella and Argos [1] have used topological similarity as a measure of fold class homology, while Holm and Sanders [3] have used similarity of distance matrices to determine fold class membership. Orengo et al. [4] have reported a classification of proteins from the protein structural database into either 150 homologous folds or 112 analogous folds from structural comparison. Chothia [2] has postulated, based on known protein sequences and structures that the total number of fold classes is expected to be circa 1,000. While it is feasible to define membership to a fold class once the three dimensional structure of the protein is determined, efforts to predict fold classes only from sequences have rendered little success. The exceptions are those where there is significant sequence homology between the protein whose structure is to be determined and one whose structure is established. Most frequently, sequences which have very much homology are known to belong to the same fold class.

K. Ramnarayan
Sapient Discovery, 10929 Technology Place, Suite B,
San Diego, CA 92127, USA
e-mail: kalram@sapientdiscovery.com

H. G. Bohr
Department of Physics, Quantum Protein Centre (QuP),
Technical University of Denmark, Kngs. Lyngby, 2800, Denmark
e-mail: hbohr@fysik.dtu.dk

K. J. Jalkanen (✉)
Nanochemistry Research Institute,
Department of Applied Chemistry,
Curtin University of Technology,
GPO Box U1987, Perth, WA 6845, Australia
e-mail: jalkanen@ivec.org

In addition, spectroscopic measurements have been shown to aid fold class assignment. Specifically the combination of vibrational absorption (VA), vibrational circular dichroism (VCD), Raman and Raman optical activity (ROA) in combination with molecular dynamics simulations and density functional theory (DFT) theory calculations has been shown to be able to determine the backbone (secondary structure) of L-alanine, L-acetyl L-alanine N′-methyl amide, L-alanyl L-alanine and Leu-enkephalin [5–15]. A preliminary study documenting the use of neural networks to predict the structure of peptides based on a combination of experimental and DFT simulated VA, VCD, Raman spectra has appeared [16,17]. A very feasible extension of this work is to use the characteristic VA, VCD, Raman and ROA spectra of proteins of known fold class and combine with work with the above aforementioned work on sequence with known structure, to predict the fold class of an unknown protein not only from the sequence, but also the measured VA, VCD, Raman and ROA spectra. Our preliminary work above, shows that this work is well worth pursuing, and is being pursued by us. This work will be presented in a future publication. In addition to the VA, VCD, Raman and ROA spectra we foresee the use of the electronic absorption (EA), electronic circular dichroism (ECD) and Resonance Raman spectra to be of use [18–22]. Recently the feasibility of the calculation of all of the above aforementioned spectra has been shown, but mostly in the gas phase, using continuum solvent models, using explicit water molecules and finally combining these approaches [7,9,14].

In most definitions of fold classes, each member would have more than 50% sequence identity to each other, although domains with far less sequence similarity could belong to the same class. It is important that each protein within a class would have a structure with a large topological similarity and a similar packing pattern to other members of the class. The details of the primary sequence in itself are less important.

The notion of fold classes is important for predicting new protein structures using homology modeling. In homology modeling an unknown three-dimensional protein structure is inferred from other known three-dimensional protein structures whose amino acid sequences are similar to the sequence of the protein in question. It has been shown [26–28] that one can predict or model protein structures to high accuracy by using structural information from proteins belonging to the same fold class or family.

However, for protein sequences with very little homology to other proteins there exists no method that can predict the three-dimensional structure to high accuracy from their sequence data alone. On the other hand proteins with little sequence homology could be similar in structure to a whole class of other structures or domains. It is apparent that protein folding into a structure is coded by information that is not transparent from sequential similarity alone. Several techniques have been developed for inferring homology at the structural level from fold class membership. Some of these incorporate a combination of secondary structure prediction schemes, functional similarity, recognition of key structural motifs and use of machine learning methods for sequence-structure mapping [3,29–33]. One method that successfully utilizes the information of the structure of homologous proteins uses artificial neural networks. The neural networks can be trained exclusively on homologous proteins as a basis for predicting a new protein structure from the corresponding sequence. Such a scheme is useful only when the protein in question has any relationship to any of the existing fold classes. The above aforementioned approach which combines the sequence information with VA, VCD, Raman, ROA, EA and ECD information also appears to be very promising.

The proposed scheme, which consists of two steps, rests on the result that neural networks can be effectively trained to induce features from a system that characterizes it. In the first step, a feed-forward neural network is used to determine the fold class of a protein from its sequence data. In the second step, the predicted fold class with its characteristic domains is used as input into a large recurrent neural network to predict the distance matrix for the protein. Such a distance matrix prediction should be accurate enough for constructing the three-dimensional backbone structure for the protein, which can then be subsequently refined by side chain placement and molecular mechanics methods.

2 Classification of protein folds by knot invariants

The Writhe, that is known from the famous formula "Link = Twist + Writhe" and steers coiling of double stranded DNA and the Average Crossing Number, that is related to the speed of DNA in electro gel experiments, are two examples of global geometric measures of closed space curves. Both of these geometric measures make sense for open space curves and more interestingly they constitute the basic building blocks of an infinite family of geometric measures called generalized Gauss integrals stemming from modern Knot Theory.

One of these Gauss integrals is

$$I_{(1,3)(2,4)} = \int_{\triangle^4} \omega(t_1, t_3)\omega(t_2, t_4)dt_1dt_2dt_3dt_4, \tag{1}$$

where $\triangle^4$ is the 4-simplex given by $0 < t_1 < t_2 < t_3 < t_4 < 1$ and

$$\omega(t, s) = \frac{\left[\gamma'(t),\ \gamma(t) - \gamma(s),\ \gamma'(s)\right]}{|\gamma(t) - \gamma(s)|^3}. \tag{2}$$

In a planar projection of the curve, γ, the configuration $(t_1, t_3)(t_2, t_4)$ defines a specific configuration of two crossings. In the planar limiting case, $I_{(1,3)(2,4)}$ counts the number

of times this crossing configuration occurs in this planar projection. The family of generalized Gauss integrals has the property that for any configuration of n crossings there is a generalized Gauss integral that counts the occurrences of this crossing configuration.

Calculating some of these Gauss integrals for protein backbones, one gets an absolute measure of protein geometry in terms of real numbers. In currently unpublished work of P. Røgen and B. Fain, it is shown that the CATH1.7 protein structure classification essentially can be reproduced based on 30 such geometric measures of the full length CATH domains. Hence, the diversity of protein structures is captured by 30 numbers [49].

3 Methodology

The basic elements of an artificial neural network, the neurons, are the 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 neurons and the neurons can communicate with each other through mutual interconnections. The network will gradually acquire a global information processing capacity for classifying data by being exposed (trained) to many pairs of corresponding input and output data such that new output can be generated from new input. If a set of input is denoted by $\{x_j\}$ and the corresponding output is denoted by $\{y_i\}$ the process at each neuron i in the network can be described by

$$y_i = f\left(\sum_j W_{ij}x_j + \eta_i\right) \tag{3}$$

where W_{ij} are the weights of the connections leading to the neuron i, η_i and f are the characteristics of the non-linear function for the neuron. As is obvious from the equation, such type of networks can be considered as a non-linear map between the input and output data.

The most straightforward type of neural networks employed for this study were feed-forward networks of the multi-layered perceptron type. These layers of neurons are referred as, mentioned in the consecutive order, the input layer, the hidden layers and the output layer. The reason for choosing this network among many other types is its ability to be generalizable to molecular biology data [34–37]. The simple structure both with respect to processing of data and training is an additional advantage with such a network. The training was carried out using the back-propagation error algorithm [38] which is also the most commonly used. The training procedure is performed until a cost function C has reached a local minimum e.g., by a gradient descent. The cost function C is normally written as,

$$C = \frac{1}{2}\sum_{\alpha,i}\left(t_i^\alpha - z_i^\alpha\right)^2 \tag{4}$$

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

In order to evaluate the performance of the network, various statistical measures have been proposed. In the case of a dual valued output the Mathews correlation coefficient, C_M [40–42], was used to monitor the performance. If two possible output values are denoted by 0 and 1 (signifying fold class membership or non membership) and if p is the number of correctly predicted examples of 1s, $\bar{p}$ the number of correctly predicted examples of 0s, q the number of examples of 1s incorrectly predicted and $\bar{q}$ is the number of examples of 0s incorrectly predicted then we define the coefficient C_M as:

$$C_M = \frac{p\bar{p} - q\bar{q}}{\sqrt{(p+q)(p+\bar{q})(\bar{p}+q)(\bar{p}+\bar{q})}} \tag{5}$$

For complete coincidence with the correct decisions (ideal performance) the measure is 1 and for complete anti-coincidence the value of C_M is -1. A poor net will give $C = 0$ indicating that it does not capture any correlation in the training set in spite the fact that it might be able to predict several correct values.

4 Implementation

4.1 Integral classifications

In the next sections we distinguish strongly between integer and real number classifications of protein folds. We shall first be discussing how protein fold with integer values are represented in neural networks and how the prediction scheme is implemented. Later we shall turn to real valued classification of folds by either knots or packing.

The actual neural networks for predicting fold classes are of the feed-forward type. The networks are trained on a selection of proteins from each of 42 fold classes containing domain segments of proteins or often the whole proteins. The input representation for each protein domain is a 20×20 matrix containing the relative frequencies of dipeptides occurring in neighboring positions in the primary sequence of the domain. To calculate these frequencies, the number of occurrences of a dipeptide is counted in the protein sequence and divided by the total number of residues in that sequence. All protein domains are transformed this way into one input pattern of fixed size. Small insertions and deletions from the protein sequence cause only small changes in the dipeptide frequencies. The same holds true for rearrangements of larger elements in the sequence that do not change

the local sequences. There are many cases where members of the same fold class differ mostly by permutations of sequence elements. Such permutations of the primary sequence lead to very similar dipeptide matrices which supports similar classification results. Each fold class is represented by one output unit which should have an activation close to 1.0 if the domain coded in the input layer is a member of that fold class. In all other cases the activity should be close to 0. When an unknown sequence is classified, the fold class corresponding to the largest activation at the output unit is assigned to the sequence. This is the usual "winner-takes-all" evaluation of the output of a classifier. In order to facilitate the interpretation of misclassifications all the fold classes were grouped into larger super-fold classes that have a natural one dimensional order inferred from physical properties of the folds. The super-fold class prediction and the fine grained classification should then assign classes that are close in this order.

4.2 Real valued classifications

In connection with neural networks, the strength of having absolute measures of protein geometry is that the relation between amino acid sequences and protein geometries may be studied and predicted directly. In particular a neural network can be trained to predict the Gauss integral values on fragments of proteins with known structure. Combined with the above mentioned result on protein structure classification this neural network then predicts fold class from amino acid sequences. The construction of such an neural network is an issue of current research.

5 Fold-classification from packing analysis

Protein fold classifications from the literature, such as the 3D-ALI, have been used so far. At a more primitive level, we have classified proteins into large classes of alpha, beta, alpha + beta and alpha-beta proteins following Lesk and Chothia [44]. In a more detailed scheme, the classification of Pascarella and Argos [1], further enhanced by Walsh [45] (Walsh, personal communication) has been utilized. In addition, a novel method for characterizing the fold topology of a protein is presented here. While the average density inside a protein is nearly a constant, the packing of residues is determined by the overall topology [46]. Arguably, all the information pertaining to the three dimensional structure and hence the topology of the protein is contained at the most refined level in the distance matrix and at a less refined level in the packing density. We define the latter as the number of pairwise atomic contacts in the protein as a function of distance. The maxima and minima that occur in this packing density are very dependent on the nature of the overall protein fold. We have obtained this packing density for all the proteins in the database and classified them based on the similarity of the packing density features. Not surprisingly, this classification groups proteins into classes that are entirely similar to the earlier classification of Pascarella and Argos. Table 1, presents the 13 super-fold classes obtained from the packing density analysis. However, this method enables the creation of a coarse-grained set of folds that encompasses several fold class members of the Pascaralla and Argos set. This super-fold class delineation is used in training the neural networks. To our knowledge, this is the first effort to use a hierarchy of fold classifications to obtain sequence-structure correlation and prediction.

The frequency of contacts between atoms at various distances within a domain or a whole protein is plotted against the measure of distances in Å along the horizontal axis and the normalized frequency (occurrence) along the vertical axis. This results in a characteristic contact distribution for each structure of protein domains (see Fig. 1). Some structures are represented by a very broad distribution while others have a sharp delta-like distribution. The maxima in the normalized frequency of the distribution is a characteristic signature of the underlying lattice structure of the domain. For example a typical protease structure like a zig-zag lattice will have a distinct peak in the pair correlation distribution at the lattice spacing length. The position, τ, of the peak in the distribution was taken as a simple measure of the domain structure and all the domain structures were hence classified into distinct groups of folds using this criterion. Folds with the smallest values of peak positions, τ, turned out to be small peptides, while intermediate ranges of τ usually could represent globular proteins. Large values of τ represented immunoglobulins and ac-proteases. Small values of τ thus signified little regularity and large values represented highly regular underlying lattice frames. The results of the performance of the neural networks using the data provided by the τ dependent fold class grouping will be presented in the following section.

6 Results

The main results in this paper are concerning the prediction of fold classes from sequence alone since the fold class represents both the secondary structure and the tertiary structure of the protein. The training set and testing set are both constructed from the data set of the 42 classes of domains. Roughly half of each fold class domains are used for training. The rationale for choosing the 42 classes from the Pascarella and Argos definition of folds, was to make certain that there are enough members in each class in order to perform a valid test. The fold class predictions are performed in three different levels of detail. The first classification uses the 4 super-fold classes based entirely on the secondary structure composition and arrangement in the proteins. The classifica-

Table 1 List of Proteins forming the thirteen super-fold classes based on packing density

Permutation matrix for packing density based class prediction														
true foldclass[a]		predicted foldclass												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	gp1 mlt	0	0	0	0	0	0	0	0	0	0	0	0	1
2	crn	0	0	0	0	0	0	0	0	0	0	1	0	0
3	Inhibit pti rdx tox	0	0	4	0	0	0	0	1	0	0	0	0	0
4	ctf eglin hoe sn3	0	0	0	0	0	0	0	1	0	0	1	0	0
5	b5c gn5 hip utg	0	0	0	0	0	0	0	2	0	0	0	0	0
6	cc5 rnt	0	0	0	0	0	0	1	0	0	0	0	0	0
7	256b acx cytc cytc3 ferredox fxc hmr il pab plasto plipase sns ssi tmv tnf virus_prot wrp	0	0	0	0	0	0	10	2	1	0	0	0	0
8	ca_bind cla dfr etu gap gcr globin hmg-2 igb lzm pap sod wga	0	0	0	0	0	0	3	27	3	0	2	1	0
9	blm carbonic cyp hmg-1 ltn pgm pyp rhd s_prot virus	0	0	0	0	0	0	1	2	6	0	0	1	0
10	Binding cpa kinase sbt tln	0	0	0	0	0	0	0	0	0	2	2	0	1
11	aat cpp icd nbd pgk xia	0	0	0	0	0	0	0	0	0	1	6	1	0
12	ac_prot barrel cat cts gls	0	0	0	0	0	0	0	2	0	0	2	4	0
13	acn	0	0	0	0	0	0	0	1	0	0	0	0	0

τ represents the peak position in the normalized frequency distribution for each class.
[a]The names are as they appear in PDB files with the chain designation as an extension. From Ref. [54]

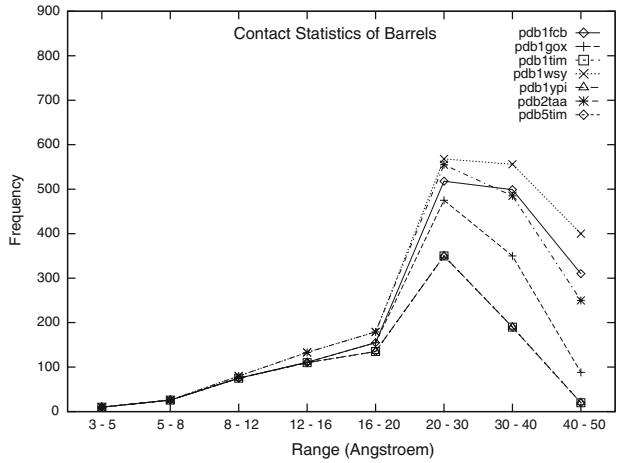
tions are based on proteins containing the secondary structures, only alpha, only beta, one alpha and one beta domain and one containing a combination of alpha and beta secondary structure elements, respectively. In the second scheme, 13-fold classes each containing 3 members or more are defined by the packing density scheme described above. By using the τ measure we define a set of 13 super-fold classes that are used for prediction of the coarse fold class. In a third scheme, the full set of 42 classes is used for fine grained classification mentioned in ref. [54].

For the first case of 4 super-fold classes a network trained up to 97.2% accuracy and had a test score of 90.4% with an average Mathews coefficient of 0.81 which is a very high performance compared to other secondary structure content predictors [40]. The analogous results, where the 13-super-fold class set obtained from packing density analysis is used, were presented in Table 1. This fold classification gives a less accurate performance of training being up to 90% correct and the test being 65% correct which render this classification to be less useful for neural network based prediction schemes. The third case that is based on much better distributed classification yields a remarkable performance of 100% on the training set and with a test score of 78% in predicting a fold class correct on the basis of the sequence. Furthermore, adding the output of the 4 super-fold classes network to the input of the 42 class based network enhanced its performance to 81.6% on the test with an average Mathews coefficient of 0.7. The fold class prediction is still more than 71% correct for those test sequences with 0 to 25% sequence identity to the training set, which is an important property for a large scale application of this prediction method.

The results of predicting fold classes based on knot invariance from sequence data is done by characterizing each fold by 30 Gauss integrals being real numbers between −10 and +10. The corresponding sequences to these folds are input and outputs are vectors each of 30 real numbers. A network trained to predict such vectors is about 80 percent correct as long a fold class has more than one member. The single member fold classes are called singletons and present a problem.

7 Fold-class database

As a spin-off of the rather successful fold class predictions a database has been constructed for public domain usage [24]. The DEF (Database for Expected Fold-classes) is made for protein fold-class predictions from sequences in the SWISS-



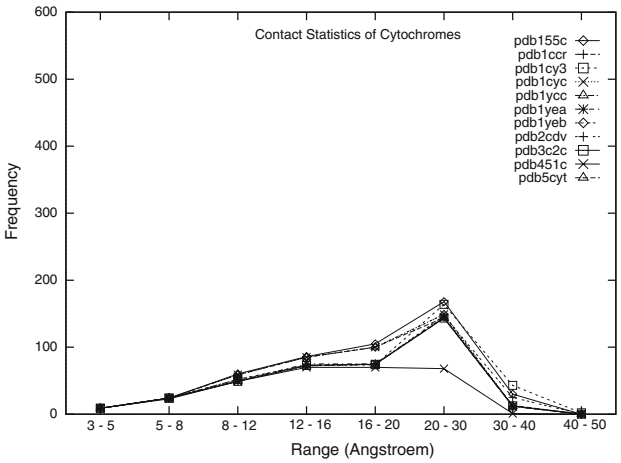


Fig. 1 Packing density for typical fold classes. Normalized frequency of pair-wise contacts versus Distance in Å

PROT protein sequence data base and is used for making predictions of fold-classes for any new sequence. In the DEF database a sequence of amino acids is assigned a specific overall fold-class, a super fold-class with respect to secondary structure content and a profile of possible fold-classes along the sequence.

8 Discussion

An artificial neural network system has been constructed to classify three-dimensional protein structures by predicting what fold class they belong to on the basis of their sequence alone. Once that is decided one may predict the corresponding distance matrix e.g., by recurrent neural networks that are trained on proteins from the chosen fold class and subsequently construct a three-dimensional structure for the test protein by a minimization procedure. The networks appear to train surprisingly well (81.2% correct and an average Mathews coefficient of 0.7) on the task of predicting fold

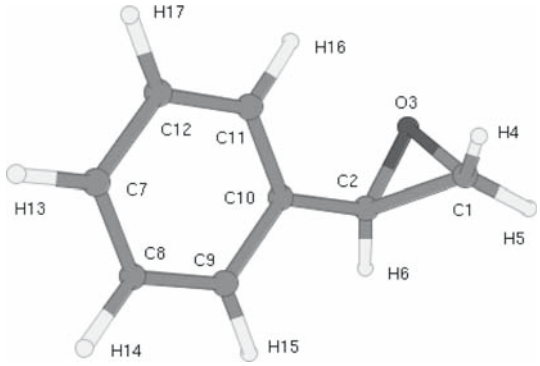


Fig. 2 (*R*)-Phenyloxirane structure, atom numbering for Table 1 from Ref. [53]

classification, even for test proteins with a maximal sequence identity of less than 25% to all training proteins.

The determination of the folds is similar to the determination of the topology of the protein backbone and that, on the other hand, depends only on the overall packing of secondary structural elements. Furthermore the new classification of folds that we proposed is partially dependent on the content of secondary structures. Low values of the τ parameter represent alpha-rich fold classes and high values of τ represent beta-rich fold classes.

In the case of prediction of fold classes based on knot invariants the success rate is as good as other methods (around 80% correct), but the issue of singletons is problematic. However, a classification of protein by real numbers is itself an achievement.

As example of the VA, VCD, Raman and ROA spectra being used as supplementary data in addition to the primary sequence, we show a comparison for Phenylalanine (Fig. 2) of the DFT based VA, VCD (Fig. 3) and Raman and ROA (Fig. 4) spectra simulations with the experimentally reported spectra [50–52]. As one can see by the good agreement between the calculated and experimental VA, VCD, Raman and ROA spectra, the combination of experimental and theoretical simulation of these spectra can be used to not only interpret and assign the vibrational spectra of biomolecules, but also used to assign the secondary structure. By combining experimental VA, VCD, Raman and ROA spectra of peptides and proteins with either know X-ray or NMR structures with supplementary spectra for other higher energy structures, the combined approach of measuring the VA, VCD, Raman and ROA spectra of proteins of known sequence and presenting this data along with the sequence data, one may hope to improve greatly the accuracy and reliability of fold class prediction, not only of the native state, but also of unfolded states.

We also present the VA, VCD and Raman spectra for the L-alanine zwitterion (LAZ). Here we have extended our previous models of LAZ + 4 and 9 water molecules, up to

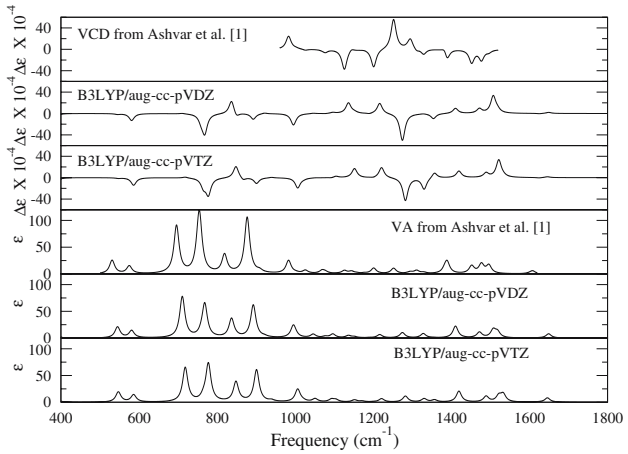


Fig. 3 *(R)*-Phenyloxirane (RPO): VA spectra for RPO at B3LYP/aug-cc-pVDZ level; VA spectra for RPO at B3LYP/aug-cc-pVDZ level; experimental VA spectra for *(S)*-Phenyloxirane (SPO) from Ashvar et al. [50]; VCD spectra for RPO at B3LYP/aug-cc-pVTZ level; VCD spectra for RPO at B3LYP/aug-cc-pVDZ level; and experimental VCD spectra for SPO from Ashvar et al. [50]. Figure reproduced from Ref. [53]

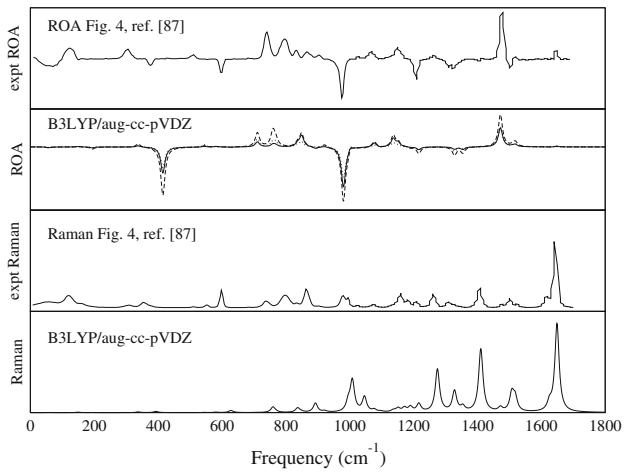


Fig. 4 (R)-Phenyloxirane: Raman spectra for B3LYP/aug-cc-pVDZ; experimental Raman spectra from Hecht et al. [52]; ROA1 (CID1), ROA2 (CID2) and ROA3 (CID3) spectra for B3LYP/aug-cc-pVDZ; and experimental ROA spectra from Hecht et al. [52]. Figure reproduced from Ref. [53]

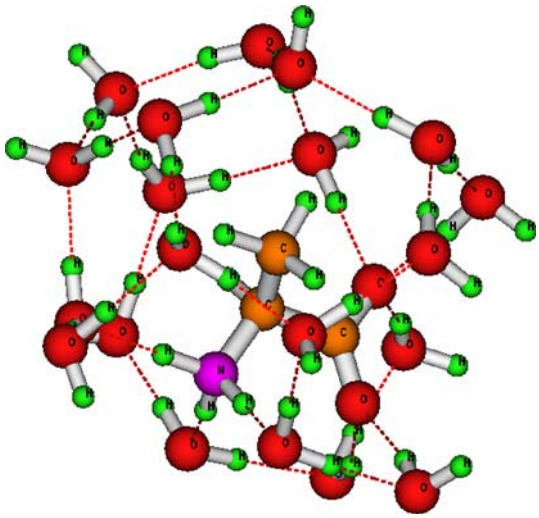


Fig. 5 L-alanine zwitterion plus 20 water hydrogen bonded network, B3PW91/6-31G* plus COSMO continuum model

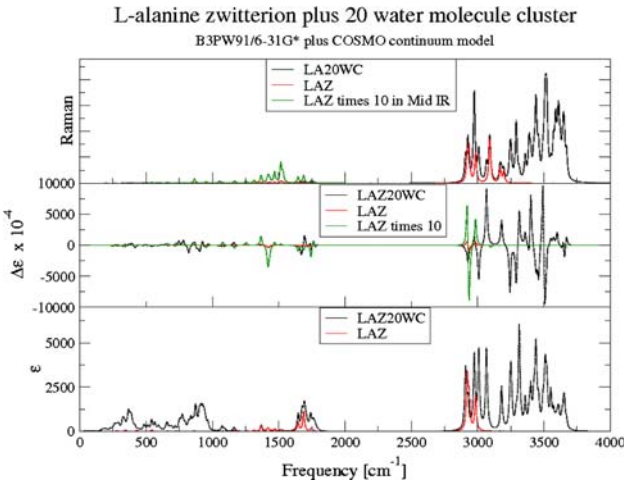


Fig. 6 L-alanine plus 20 water molecule complex: VA, VCD and Raman spectra with and without water molecules; B3PW91/6-31G* plus COSMO continuum model

20 water molecules. In the previous work we only kept the strongly interacting hydrogen bonded water molecules. In this work a complete solvation shell of hydrogen bonded water molecules has been added. The initial positions were determined by taking the lowest energy structure from our Born Oppenheimer molecular dynamics simulation of the LAZ in a droplet of water [56]. The optimized structure of the LAZ plus 20 water molecule is shown in Fig. 5 and the corresponding VA, VCD and Raman spectra with and without the explicit water molecules included in Fig. 6. This is the quenched structure of the complex and shows the real complexity of assigning the bands of a biomolecule to only the solute. The bands are in similar regions with those of the solvent. This results in strong coupling between the solvent and solute modes. Additionally the strong hydrogen bonding with the solute results in the water modes in this first solvation shell inherently different than those in the bulk, When one tries to do a solvent subtraction, one is only subtracting out the bulk solvent modes. If one wishes to subtract out the water modes strongly hydrogen bonded with the solute, then one has a more difficult problem. One of the work arounds for this has been to fit the bands with either either Gaussian or Lorentzian line shape functions (or to simulate the spectra with one of these functions). But here one is really sweeping the problem under the rug, where one does not see and or

realize that a problem exists. Our point here is only to point out this problem, which we will present a more thorough report on our progress in this area at a later date.

Another problem is to correlate the complicated band shape features in the VA, VCD, Raman and ROA spectra to specific structural information which can be used as the so called identifiers. Here one can use not only the frequencies, but also VA, VCD, Raman and ROA intensities or lack of intensities. But to do this requires one to really be able to identify the principle components in the spectra. In the past one has assumed that the features are only due to the solute, but we think that our example for the LAZ has shown this not necessarily always to be the case.

Finally we would like to present one last example, the so called alanine depeptide, N'-acetyl L-alanine N'-methyl amide (NALANMA). This has been one of the most studied peptides, but surprising until 1998 the conformer of this molecule had not been solved, either by NMR, EA, ECD, VA, VCD, Raman and ROA or even a combination. This again was due to the problem that the structure in aqueous solution is not one of the structures which are stable minimum on the PES of the isolated state. In Fig. 7 we show the structure determined at the B3PW91/aug-cc-pVDZ level of theory. The values of the ϕ and ψ angles at this level of theory are $-98.68°$ and 133.36°, versus the values at the B3LYP/6-31G* plus Onsager continuum solvent level of theory being $-93.55°$ and 127.62°, respectively. At the B3LYP/aug-cc-pVDZ plus PCM level of theory the values are $-91.68°$ and 133.36°, respectively. So the effect of the larger basis set, the alternative hybrid exchange correlation functional and the alternative continuum solvation model do not appear to

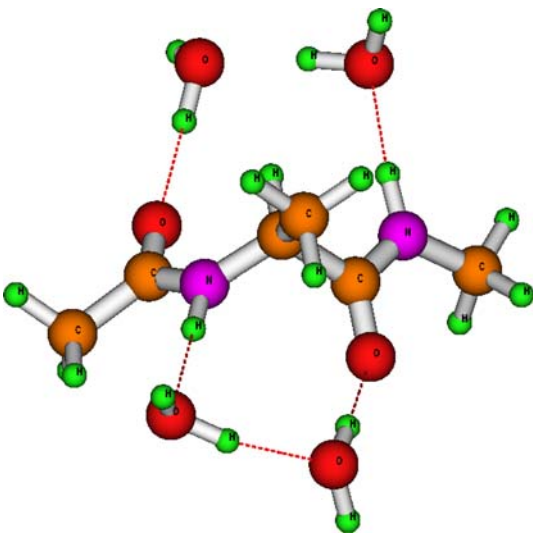


Fig. 7 N'-acetyl L-alanine N'-methyl amide plus four water hydrogen bonded network, B3PW91/aug-cc-pVDZ plus COSMO continuum model

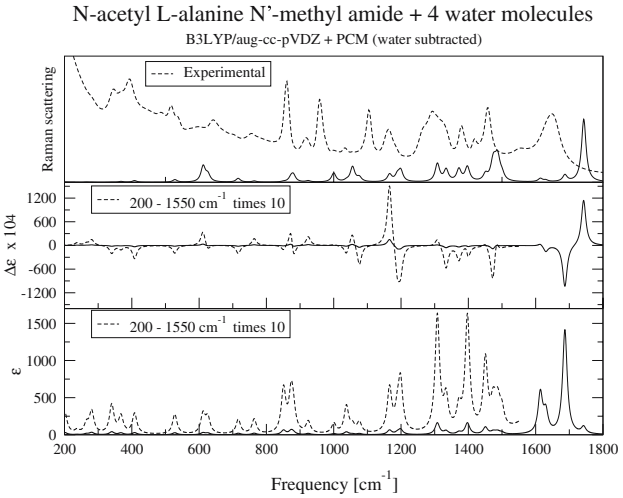


Fig. 8 Comparison of experimental and theoretical (N'-acetyl L-alanine N'-methyl amide plus four water hydrogen bonded network at B3LYP/aug-cc-pVDZ plus PCM continuum model level of theory) Raman spectra and VA and VCD spectral simulations

be too large, and most importantly, at all levels of theory the complex is stable.

But the ultimate criterion for us has been the agreement between the VA, VCD, Raman and ROA spectral simulations and the experimental spectra [7]. In Fig. 8 we present a comparison of the experimental Raman spectra of NALANMA and our spectral simulation at the B3LYP/aug-cc-pVDZ plus PCM level of theory for the NALANMA plus four water molecule complex. In addition, we present our spectral simulation for the VA and VCD spectra. Here we have subtracted out the contributions due to water molecules. The agreement with the experimental Raman spectra if noticable better using the PCM continuum solvent model and the aug-cc-pVDZ basis set.

An additional measured and reported value for many chiral molecules is the α_D value. Here we report the predicted α_D for the NALANMA plus four water molecule complex to be 79.39°. This is the first reported value of this quantity for a dipeptide molecular complex. As shown by our VA, VCD, Raman and ROA simulations, this complex appears to be stable. Hence it would be interesting to try to measure the α_D for the NALANMA plus four water molecular complex in a molecular beam experiment. The relative strength of the hydrogen bonds between water and other water molecules and water and the dipeptide group is fundamental to biochemistry. X-ray and neutron diffraction studies have shown that the mobility of these water molecules are different. Hence it would be nice to do not only temperature dependent VA, VCD, Raman and ROA experiments, but also temperature dependent α_D measurements, to see if this value changes as the NALANMA plus N water molecular complexes freeze in. Previously it has been shown that explicit

water molecules are necessary to stabilize structures which are not stable on the gas phase or isolated state potential energy surface or using continuum solvent models [57].

The calculation of the tensors for the ROA spectral simulations requires one to calculate the G′ and A tensor derivatives numerically, that is, calculate them at 6N displaced geometries, in addition to calculating them at the optimized geometry. For the NALANMA plus four water complex, this requires $6 \times 34 = 204$ solutions to the coupled perturbed Kohn Sham equations. Hence we will report the complete set of VA, VCD, Raman and ROA spectral simulations in a future work. The aug-cc-pVDZ basis set has been shown to give almost quantitative values for the VA, VCD, Raman and ROA spectral intensities, but this basis set does not appear to optimal for simulations where the amino acids, dipeptides and polypeptides are completely solvated, as was the case for LAZ20WC simulation presented in Fig. 6 due to its large size. A compromise appears to be to use either the 6-31G* basis set (which we have used) or the slightly larger 6-31G** or DZP. But here by adding polarization functions on the hydrogens, this will increase the size of the basis set by $3 \times$ the number of hydrogens, which for the LAZ20WC would be $47 \times 3 = 141$ more basis functions. In a future work we will further document the effect on using various basis sets, but the main point which we wish to end with is that the spectral simulations of the vibrational spectra of amino acids, dipeptides and polypeptides must take into account the effects of the first solvation shell of water molecules since they have been shown to not only change the potential energy surface of the isolated molecules, but also have large effects on the frequencies and the intensities. Additionally the properties of these waters are interesting in themselves, as they are the water molecules which must be replaced on ligand binding. Hence the relative binding strength of the of water and ligands in the binding pocket can be studied with time and temperature dependent vibrational spectroscopy studies. The combination of knowledge based methods (neural networks and knot theory) and high level *ab initio* and density functional theory appears to be a viable alternative to the methods which other groups are pursuing to study these problems, without some of the problems with labels (fluorescence spectroscopy being one of the alternative techniques being used by many research groups).

Acknowledgements HGB would like to acknowledge Drs. Richard Goldstein and Peter Wolynes for valuable discussions. HGB and KJJ also acknowledge NSF for partial support and NCSA for providing us with computer resources especially on the connection machine CM5. KJJ would like to acknowledge the Western Australia government's Premier Fellowship program for providing financial support and the iVEC Supercomputer Centre of Western Australia and the APAC National Supercomputer in Canberra for providing computational facilities. We would like to thank Prof. L.D. Barron and Dr. F. Zhu for measuring the Raman spectra of NALANMA. A complete investigation on this molecule with Prof. Barron and Dr. Zhu is forthcoming.

References

1. Pascarella S, Argos P (1992) Protein Eng. 5:121
2. Chothia C (1992) Nature 357:543
3. Holm L, Sander C (1993) J Mol Biol 233:123
4. Orengo CA, Flores TP, Taylor WR, Thornton JM (1993) Protein Engng 6:485
5. Jalkanen KJ, Suhai S (1996) Chem Phys 208:81
6. Deng Z, Polavarapu PL, Ford SJ, Hecht L, Barron LD, Ewig CS, Jalkanen KJ (1996) J Phys Chem 100:2025
7. Han W-G, Jalkanen KJ, Elstner M, Suhai S (1998) J Phys Chem B 102:2587
8. Bohr H, Jalkanen KJ, Elstner M, Frimand K, Suhai S (1999) Chem Phys 246:13
9. Tajkhorshid E, Jalkanen KJ, Suhai S (1998) J Phys Chem B 102:5899
10. Frimand K, Bohr H, Jalkanen KJ, Suhai S (2000) Chem Phys 255:165
11. Jalkanen KJ, Nieminen RM, Frimand K, Bohr J, Bohr H, Wade RC, Tajkhorshid E, Suhai S (2001) Chem Phys 265:125
12. Knapp-Mohammady , Jalkanen KJ, Nardi F, Wade RC, Suhai S (1999) Chem Phys 240:63
13. Jalkanen KJ, Nieminen RM, Knapp- Mohammady, Suhai S (2003) 92:239
14. Jalkanen KJ (2003) J Phys Condens Matter 15:S1823
15. Jalkanen KJ, Elstner M, Suhai S (2004) J Mol Struct (Theochem) 675:61
16. Bohr H, Frimand K, Jalkanen KJ, Nieminen RM, Suhai S (2001) Phys Rev E 64:021905
17. Bohr H, Røgen P, Jalkanen KJ (2001) Comp Chem 26:65
18. Callender RH, Dyer RB, Bilmanshin R, Woodruff WH (1998) Ann Rev Phys Chem 49:173
19. Vanderkooi JM, Adar F, Erecinska M (1976) Eur J Biochem 64:381
20. Nielsen KL, Idianai C, Henriksen A, Freis A, Beucucci M, Gajhede M, Smulevich G, Welinder KG (2001) Biochem 40:11013
21. Smulevich G (1998) Biospectroscopy 4:S3
22. Kapetanaki S, Chouchase S, Girotto S, Yu S, Magliozzo RS, Schelvia JP (2003) Biochem 42:385
23. Jones DJ, Taylor WR, Thornton JM (1992) Nature 358:86
24. Reczko M, Bohr HG (1994) Nucleic Acids Res 22:3616
25. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. (1987) Protein Data Bank. In: Crystallographic databases—information content, software systems, scientific Applications. In: Allen FH, Bergerhof G, Sievers R (eds.) 108–132, Data Commission of the international Union of Crystallography, Bonn/Cambridge/Chester.
26. Bassolino-Klimas D, Bruccoleri RE, Subramaniam S (1992) Protein Sci 1:1465
27. Viswanathan M, Anchin JM, Droupadi PR, Mandal C, Linthicum DS, Subramaniam S (1994) Molecular Biophysics Technical Report UIUC-BI-MB-94-02
28. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. (1992) Proc Natl Acad Sci USA 89:9029
29. Ioerger TR, Rendell LA, Subramaniam S (1993) In: Proceedings of 1st International conference on intelligent systems for molecular biology, AAAI press, Menlo Park, pp 198–206
30. Bryant SH, Lawrence CE (1993) Proteins: Struct Func Genet 16:92
31. Sippl MJ (1990) J Mol Biol 213:859
32. Johnson MS, Overington JP, Blundell TL (1993) J Mol Biol 231:735
33. Jones D, Thornton J (1993) J Comput Aided Mol Design 7:439
34. Qian N, Sejnowski TJ (1988) J Mol Biol 202:865
35. Bohr H, Bohr J, Brunak S, Cotterill RMJ, Lautrup B, Nørskov L, Olsen OH, Petersen SB (1988) FEBS Lett 241:223

36. Holley LH, Karplus M (1989) Proc Natl Acad Sci USA 86:152
37. Bohr H, Bohr J, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B, Petersen SB (1990) . FEBS Lett 261:43
38. Rumelhart DE, McClelland JL (eds.) (1986) Parallel distributed processing. MIT Press, Cambridge
39. Fahlman SE, Lebiere C (1990) In: Touretzky DS (ed.) Advances in neural information processing systems II, Morgan Kaufmann, Los Altos pp 524–532
40. Mathews BW (1975) . Biochem Biophys Acta 405:442
41. Stolorz P, Lapedes A, Yuan X (1991) Predicting protein secondary structure using neural net and statistical methods. (Los Alamos Preprint LA-UR-91-15)
42. Stolorz P, Lapedes A, Yuan X (1992) Predicting protein secondary structure using neural net and statistical methods. J Mol Biol 225:363
43. Zell A, Mache N, Sommer T, Korb T (1991) In: Proceedings of applications of neural networks conf., SPIE, Aerospace Sensing Intl. Symposium, Vol 1469 Orlando pp 708
44. Lesk AM (1991) "Protein architechture a practical approach". Oxford University Press, Oxford
45. Walsh LL (1992) Protein Science 1:5, Diskette Appendix
46. Kauzmann W, Moore K, Schultz D (1974) Nature 248:447
47. Reczko M, Bohr H (1994) In: Bohr H, Brunak S (eds.) Protein structure by distance analysis IOS Press, Amsterdam pp 87–97
48. Bohr H, Goldstein R, Wolynes PG (1992) AMSE Periodicals, Modelling, measurement and control C 31:55
49. Røgen P, Faen B (2003) Proc Nat Acad Sci (USA) 100:119
50. Ashvar CS, Devlin FJ, Stephens PJ (1999) J Am Chem Soc 121:2836
51. Hecht L, Barron LD (1994) J Raman Spectros 25:443
52. Hecht L, Barron LD (1995) J Mol Struct 347:449
53. Jalkanen KJ, Jürgensen VW, Degtyarenko IM (2005) Adv. Quan. Chem. 50:91
54. Reczko M, Bohr H, Subramaniam S, Pamidighantam S, Hatzigeorgiou A (1994) In: Bohr H, Brunak S, Protein structure by distance analysis. IOS Press, Burke pp 277–286
55. Nielsen BG, Røgen P, Bohr H (2006) Math Comput Model 43:401
56. Degtyarenko IM, Jalkanen KJ, Gurtovenko AA, Nieminen RM (2007) J Phys Chem B 111:4227
57. Beglov D, Roux B (1995) Biopolymers 35:171