Neural-network analysis of the vibrational spectra of N-acetyl L-alanyl N '-methyl amide conformational states

Bohr, Henrik; Frimand, Kenneth; Jalkanen, Karl J.; Nieminen, R.M.; Suhai, S.

Published in:
Physical Review E. Statistical, Nonlinear, and Soft Matter Physics

Link to article, DOI:
10.1103/PhysRevE.64.021905

Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

Neural-network analysis of the vibrational spectra of *N*-acetyl *L*-alanyl *N′*-methyl amide conformational states

H. G. Bohr* and K. Frimand
Department of Physics, The Technical University of Denmark, DK-2800 Lyngby, Denmark

K. J. Jalkanen† and R. M. Nieminen
Laboratory of Physics, Helsinki University of Technology, P.O. Box 1100, Otakaari 1 M, FIN-02015 HUT, Finland

S. Suhai
German Cancer Research Center, Im Neuenheimer Feld 280, D-69121 Heidelberg, Germany
(Received 22 March 2000; revised manuscript received 16 April 2001; published 20 July 2001)

Density-functional theory (DFT) calculations utilizing the Becke 3LYP hybrid functional have been carried out for *N*-acetyl *L*-alanine *N′*-methylamide and examined with respect to the effect of water on the structure, the vibrational frequencies, vibrational absorption (VA), vibrational circular dichroism (VCD), Raman spectra, and Raman optical activity (ROA) intensities. The large changes due to hydration in the structures, and the relative stability of the conformer, reflected in the VA, VCD, Raman spectra, and ROA spectra observed experimentally, are reproduced by the DFT calculations. A neural network has been constructed for reproducing the inverse scattering data (we infer the structural coordinates from spectroscopic data) that the DFT method could produce. The purpose of the network has also been to generate the large set of conformational states associated with each set of spectroscopic data for a given conformer of the molecule by interpolation. Finally the neural network performances are used to monitor a sensitivity analysis of the importance of secondary structures and the influence of the solvent. The neural network is shown to be good in distinguishing the different conformers of the small alanine peptide, especially in the gas phase.

DOI: 10.1103/PhysRevE.64.021905 PACS number(s): 87.15.−v, 33.20.−t, 33.15.Bh, 07.05.Mh

I. INTRODUCTION

The goal of the present study of applying density-functional theory to peptides for finding the electronic structure of peptide-bonded amino acids in solution is to understand the connection between the structure and the function of protein molecules. The study presented in this paper serves as a pilot project for the greater goal of deriving function from structure in proteins. Up to now electronic structures of dipeptides and tripeptides in vacuum have been calculated and measured by others, and the main conclusion from such work has been that ''ionic'' compounds, e.g., zwitterion molecules, are unstable in vacuum or in the isolated state in nonpolar solvents or inert matrices. Thus, to gain further insight into the problem via a quantum-mechanical analysis of the electronic structure of these biomolecules, we have added the effect of the solvent in our calculations, i.e., added explicit water molecules to the peptide structures to simulate the effect of those waters directly hydrogen-bonded with the polar groups and subsequently embedded these '' molecule$+N$ water complexes'' within a dielectric medium via a continuum model. Here the continuum model has been used to try to simulate the effects due to bulk water molecules, while the explicit water molecules have been added to simulate the effects of the water molecules which are in direct contact with the peptide.

We shall also try here to demonstrate the usefulness of neural networks for quantum chemistry calculations and spectroscopy. This is a very promising application to protein structure and functionality, although at the moment it is only applicable for small peptides. It is the hope that by calculating detailed electronic properties and interactions of the protein with its aqueous surroundings in particular states, a network trained on such results should be able to extrapolate and produce many of the other relevant functional states. We have obvious reasons to believe that an active protein exists in many functionally important substates, as demonstrated, e.g., by Frauenfelder *et al.* [1]. Detailed electronic calculations can, due to limited computer resources and time, only comprise a few of these conformational substates. However, it seems plausible that a neural network should be able to extract essential features of such calculations on a few conformational substate structures and then generate many, if not all, other substate structures that might be of relevance. The results of the following study seem to indicate that such a task is possible.

In the first part of this paper the methodology of the detailed electronic calculations is reviewed and connected to the spectroscopy of protein structure experiments. Here we give only a short presentation of density-functional theory with the goal to give a feeling for what can be calculated at this time [2–29]. The development of density-functional theory with respect to its application to problems in biophysics, for example, the prediction of vibrational circular dichroism (VCD), Raman spectra, and Raman optical activity (ROA) spectra, is a very exciting area [30–35]. Here one must go beyond simple local-density approximation (LDA)

*Corresponding author.
†Present address: Steinbeis Center for Genome Informatics, Im Neuenheimer Feld 370/42, D-69120 Heidelberg, Germany.

and generalized gradient approximation (GGA) and introduce the electron current density as a variable either explicitly or implicitly in addition to the normal variables, the electron density and the gradient of the electron density, to treat magnetic field effects. In the second part the neural network application is explained in a more straightforward way and more along the line of other applications of neural networks in biochemistry. The neural networks are first trained and tested on the peptide molecule in vacuum and then later on the same molecule in a water solution. It turns out that water makes the task of predicting conformations from spectroscopic data harder.

There have appeared a few papers that are similar in spirit to this study. In a paper by Fariselli and Casadio [36] a neural network is used for predicting contact maps of proteins from the input of chemico-physical and evolutionary data. Once a contact map is obtained the protein structure can be derived by minimization [37]. Their study shows that neural networks are better in predicting protein structures than ordinary statistical methods. In another study Pancoska *et al.* [38] have, on the basis of VCD spectra, used neural networks to obtain structural information about proteins beyond the usual secondary structure content that CD and VCD spectra provide.

II. PERSPECTIVES CONCERNING SOLVENT EFFECTS

Hydration is an important issue in genome research as exemplified by the structural change which occurs as one lowers the relative humidity of DNA below 75% (*B*-DNA converts to *A*-DNA). The phosphate groups in the *A*-helix bind fewer water molecules than do the phosphate groups in the *B*-helix, hence dehydration favors the *B* form of DNA. The effect of hydration on the binding of proteins to DNA and RNA is still not well understood and most modeling of the interaction of proteins with DNA and RNA does not treat the water molecules explicitly. In this work we have not tried to treat the binding of the protein with DNA and RNA but study the effect of hydration on the structural and spectroscopic changes in small biomolecules, which function as model systems for DNA and protein hydration phenomena, similar to the effect of hydration on the forms of DNA. Once the effect of hydration is understood at the molecular level for small peptides and later for proteins we can go on to try to understand the effect of hydration on the binding and recognition process in protein-DNA/RNA complexes, and hence to understand at a molecular level the biological processes and how they are mediated in aqueous solution and then, ultimately, in the cell. Many of the current models treat hydration macroscopically and do not include the structural and electronic effects due to the solvent microscopically or quantum mechanically. Our work here is an attempt to document the hydration effect in proteins at a microscopic level with the hope of pointing out some of the deficiencies in the current models and to provide some directions and insights into possible improvements.

The effect of hydration on small peptides and amino acids is, in spite of their limited size, still a ubiquitous problem, hard to calculate, measure, and understand. Here we present some DFT calculations on hydrated *N*-acetyl-*L*-alanine *N'*-methylamide (NALANMA) which will shed some light on the effect of water on the structures, vibrational frequencies, VA, VCD, and Raman and ROA intensities. We have also constructed an artificial neural network to solve the inverse scattering problem of retrieving structural information of the biomolecule from spectroscopic data, that is, vibrational frequencies, VA, VCD, and Raman and ROA intensities of isolated NALANMA.

We take two routes to get from the spectroscopic data to predict the structure of our test molecule. One route is to use density-functional theory (DFT) at the Becke 3LYP/6-31G* level to calculate all the possible structures and for all of the structures the corresponding frequencies, VA, VCD, and Raman intensities and ROA intensities and then compare them to the experimental data. The other route is to train the neural networks on a large combination of calculated correlations to infer or extrapolate new results. When going to large biomolecules one can determine whether there is a correlation between the best predicted structural details from spectroscopic data and the data connected to secondary structure stability. This is in order to see which spectroscopic data are the most important in determining the secondary structures.

The larger goal is to utilize neural networks for determining the structural minima. At these minima VA, VCD, and Raman intensities and ROA intensities are calculated by DFT in order to produce training data, that is, sets of spectroscopic data correlated with $\phi-\psi$ angles for the network. Other methods, such as x-ray crystallography and NMR, have only been utilized to determine the native states of proteins. Spectroscopic measurements provide the possibility of determining the denatured states of proteins. The problem is to know the structures and VA, VCD, Raman spectra, and ROA spectra of the conformational states of proteins. Pancoska and co-workers have utilized neural network methodology to find correlations of VCD spectra with the native states of proteins by utilizing the known NMR and x-ray crystallographic structures [38]. Our work complements their work in providing correlations of VA, VCD, Raman spectra, and ROA spectra and the higher-energy denatured states of peptides and proteins. These denatured states can be produced under various experimental conditions, that is, in an aqueous solution under a variety of conditions, for example, at various pH, salt conditions, and by the presence of urea and other denaturing agents.

In that sense one should be able to predict higher level intermediate-energy states during folding processes of biomolecules with the help of neural networks, once they are trained on known sets of intermediate energy states. The great thing about utilizing neural network techniques for the inverse scattering problem of deriving structural information from scattering data is that it goes hand in hand with experiments and DFT calculations in the sense that one of the tools can support the other when it fails. This means that where there is no known structure for the conformational states of the protein but measured VA, VCD, Raman spectra, and ROA spectra, one should use neural networks.

III. DENSITY-FUNCTIONAL ANALYSIS OF HYDRATION EFFECTS ON SMALL PEPTIDES AND AMINO ACIDS

A. Standard formula

The principles of density-functional theory are tightly coupled to wave-function theory. It is not a completely independent formulation of quantum mechanics. Here we shall try to give a brief overview of the early attempts at developing an independent theory and how the problems were overcome by borrowing from wave-function theory. This borrowing has helped overcome some of the fundamental problems with a pure and independent density-functional theory, but has also introduced some new problems. One is very fundamental, the definition of the *correlation energy*. In wave-function theory, the correlation energy is defined as the difference between the exact Hartree-Fock energy and the exact energy. Clearly this definition is not a good definition for a pure and independent density-functional theory. Hence other definitions for the density-functional theory correlation energy have been proposed [39]. Another is the definition of exchange energy. Within Hartree-Fock theory the exchange energy (or better named the exchange integral) is clearly defined. It does not have a purely classical analog and hence it is not clearly obvious how to form the exchange energy functional in terms of the electron density. Hence here also some confusion arises. Here we try to make clear the connections between a pure and independent density-functional theory and wave-function theory, first at the Hartree-Fock level and then a more generalized form, the highest being a full-configuration-interaction formulation, which gives us a formalistic way to get the exact energy in theory, but is not feasible and practical for a many-electron atom, and certainly not attainable for a polypeptide or protein.

Hence we must make approximations. But what one seeks whenever one makes an approximation is to understand clearly what one is giving up by making this approximation. One must make a clear distinction between an assumption or premise and an approximation. One further approximation which is many times overlooked is the Born-Oppenheimer approximation. In many cases in wave-function theory we work within the Born-Oppenheimer approximation and fail to mention and to understand the consequences of this. Not all properties are calculable or even meaningful with the Born-Oppenheimer approximation. A case in point is the magnetic dipole moment and the derivative of the magnetic dipole moment with respect to the nuclear velocities, that is, the atomic axial tensor of Stephens [40] and Buckingham *et al.* [41]. To calculate the VCD spectra one requires these non-Born-Oppenheimer properties. Hence one must have a clear understanding on how to go beyond the Born-Oppenheimer approximation within the realm of a pure and independent density-functional theory also. Finally the concepts of perturbation theory and finite field perturbation theory need to be generalized if one wishes to be able to calculate all of the properties within the density-functional theory which one can currently calculate within wave-function theory. If the DFT is to achieve its goal, that is, to surplant wave-function theory, then one must be able to calculate all properties within the density-functional theory, which we can do within the wave-function theory. This is clearly not yet possible. As one attempts to reformulate wave-function perturbation theory to density-functional perturbation theory, one then must address the same problems one addresses when trying to address simple density-functional theory, where one only wishes to determine the ground-state potential-energy surface, that is, $E[\rho(\vec{r}),\vec{R}]$.

The first Hohenberg-Kohn theorem [2,42,43] proves that the electron density determines the energy and hence reformulates the basic equation to solve as one in which one has to determine the electron density rather than the wave function. The energy functional can be written as

$$E_v[\rho]=T[\rho]+V_{ne}[\rho]+V_{ee}[\rho]=\int \rho(\vec{r})v(\vec{r})d\vec{r}+F_{HK}[\rho], \tag{1}$$

where

$$F_{HK}[\rho]=T[\rho]+V_{ee}[\rho] \tag{2}$$

and $v(\vec{r})$ is the external potential and $T[\rho]$ the kinetic energy. The second Hohenberg-Kohn theorem provides the energy-variational principle which enables one to find the density that minimizes this energy functional. The problem is that we do not know the functional $F_{HK}[\rho]$ exactly. Many functionals have been developed which try to address this problem.

Here we focus on the total energy functional $E[\rho]$ expressed as

$$E[\rho]=\int \rho(\vec{r})v(\vec{r})d\vec{r}+T[\rho]+V_{ee}[\rho]. \tag{3}$$

This formulation of DFT which introduces orbitals into the problem is very similar to that of Kohn and Sham and Parr and Yang. This has been done so that one has a good representation for the kinetic energy functional $T[\rho]$ and the electron-electron repulsion functional $V_{ee}[\rho]$, that is, the last two terms in Eq. (3).

The early pure DFT models, for example, the Thomas-Fermi model, had the seemingly insurmountable problem of trying to find the kinetic energy functional $T[\rho]$ and the electron-electron repulsion functional $V_{ee}[\rho]$. In terms of the spin orbital and occupation numbers, the exact expression for the ground-state kinetic energy T is known:

$$T=\sum_i^N n_i\langle\phi_i|-\tfrac{1}{2}\nabla^2|\phi_i\rangle, \tag{4}$$

where the ϕ_i and n_i are the natural spin orbitals and their occupation numbers, respectively. Note that the Pauli principle requires that $0\leqslant n_i\leqslant 1$. Using the Hohenberg-Kohn theorem, the kinetic energy functional $T[\rho]$ is a functional of the total electron density. Here we have expressed the total electron density ρ in terms of orbitals,

$$\rho(\vec{r}) = \sum_{i}^{N} n_i \sum_{s} |\phi_i(\vec{r},s)|^2. \tag{5}$$

This helps us get insight into how to deal with the kinetic energy functional. By assuming the system to be N noninteracting electrons these expressions simplify to

$$T_s = \sum_{i}^{N} \langle \phi_i | -\tfrac{1}{2}\nabla^2 | \phi_i \rangle, \tag{6}$$

where the ϕ_i are the natural spin orbitals and their occupation numbers are now 1 for the occupied orbitals and 0 for the virtual orbitals, respectively. Using the Hohenberg-Kohn theorem, the kinetic energy functional $T_s[\rho]$ is a functional of the total electron density. Here we have expressed the total electron density ρ in terms of orbitals,

$$\rho(\vec{r}) = \sum_{i}^{N} \sum_{s} |\phi_i(\vec{r},s)|^2. \tag{7}$$

But how do we deal with the electron-electron repulsion functional? The classical expression for the electron-electron repulsion would give us the term

$$J_{ee}[\rho(\vec{r})] = \frac{1}{2}\int \frac{\rho(\vec{r}_i)\rho(\vec{r}_j)}{r_{ij}} d\vec{r}_i d\vec{r}_j. \tag{8}$$

Here one gets the classical Coulomb repulsion integral but ones loses or does not get the term that comes from exchange, which one gets when one uses wave-function theory; that is, the term one gets when one uses one-electron spin orbitals and an antisymmetric wave function that satisfies the Pauli principle with respect to the exchange of two particles, usually a Slater determinant. This is a term that one gets at the Hartree-Fock level using a Slater determinant and arises from the exchange of particles, hence the name exchange energy or exchange integral. It is a purely quantum effect due to the fermion nature of electrons, indistinguishability of identical particles. How to formulate this term in terms of only the electron density is similar to the problem we had with how to form the general electron kinetic energy functional in terms of only the density. But by forming the density in terms of orbitals, we are able to obtain an approximate form for the electron kinetic energy in terms of orbitals. Similarly an exchange energy functional can be obtained in terms of orbitals. One can use expressions from wave-function theory to generate approximations to the exact functionals when the density is formed from orbitals. Then all of the remaining errors can be lumped in the expression which has been called the exchange-correlational functional. The exchange-correlational energy functional then becomes

$$E[\rho] = T_s[\rho] + \int \rho(\vec{r})v(\vec{r})d\vec{r} + J_{ee}[\rho] + T[\rho] - T_s[\rho] + V_{ee}[\rho] - J_{ee}[\rho] \tag{9}$$

or

$$E[\rho] = T_s[\rho] + \int \rho(\vec{r})v(\vec{r})d\vec{r} + J_{ee}[\rho] + E_{XC}[\rho], \tag{10}$$

where the exchange-correlational functional E_{XC} is given by the following expression:

$$E_{XC} = T[\rho] - T_s[\rho] + V_{ee}[\rho] - J_{ee}[\rho]. \tag{11}$$

The E_{EC} which we have used in this work is a hybrid exchange-correlation functional, the Becke 3LYP (B3LYP) functional, defined by the following expression:

$$E^{XC}_{B3LYP} = E^{X}_{LDA} + 0.20(E^{X}_{HF} - E^{X}_{LDA}) + 0.72\Delta E^{X}_{B88} + E^{C}_{VWN3} + 0.81(E^{C}_{LYP} - E^{C}_{VWN3}). \tag{12}$$

This functional has been implemented in the Cambridge Analytical Derivatives Package (CADPAC), Gaussian, and a variety of other wave-function (orbital) based density-functional-based codes. The first term is the local exchange functional, E^{X}_{LDA}, defined by

$$E^{X}_{LDA} = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3}\int \rho^{4/3} d^3\vec{r}, \tag{13}$$

where ρ is the electron density. This functional was developed to reproduce the exchange energy of a uniform electron gas. The second term adds an admixture of Hartree-Fock local exchange to the LDA local exchange term. The Hartree-Fock local exchange functional gets its functional form from Hartree-Fock theory, but replaces the Hartree-Fock orbitals by the Kohn-Sham orbitals,

$$E^{X}_{HF} = -\frac{1}{2}\sum_{i,j} \int\int \frac{\phi_i^*(x_1)\phi_j^*(x_2)\phi_j(x_1)\phi_i(x_2)}{r_{12}} dx_1 dx_2. \tag{14}$$

The third term includes an admixture of Becke's gradient correction, E^{X}_{Becke88}, to the LDA exchange. The E^{X}_{Becke88} is defined by

$$E^{X}_{\text{Becke88}} = E^{X}_{LDA} - \gamma\int \frac{\rho^{4/3}x^2}{(1+6\gamma\sinh^{-1}x)} d^3\vec{r}, \tag{15}$$

where $x = \rho^{-4/3}|\nabla\rho|$ and γ is a parameter chosen to fit the known exchange energies of the noble gas atoms, which Becke defines as 0.0042 Hartrees. The fourth term accounts for the VWN3 local correlation function [44]. Vosko, Wilk, and Nusair (VWN) proposed the following functional form for the correlational functional:

$$E^{C}_{VWN}(r_s,\zeta) = E^{0}_{C}(r_s) + \alpha(r_s)\left[\frac{f(\zeta)}{f''(0)}\right][1+\beta(r_s)\zeta^4], \tag{16}$$

where $\alpha(r_s)$ is the spin stiffness and $\beta(r_s)$ is chosen to satisfy $\epsilon_c(r_s,1) = \epsilon_c(r_s)$, namely,

$$1 + \beta(r_s) = f''(0)\frac{\epsilon'_c(r_s) - \epsilon^0_c(r_s)}{\alpha(r_s)}. \tag{17}$$

For more details on the VWN and VWN3 functionals we refer the interested reader to the original paper [44], the book by Parr and Yang on density-functional theory, and finally to the Gaussian and CADPAC user's manuals and source code for direct implementation. Finally the last term adds an admixture of the Lee, Yang, and Parr (LYP) correlation correction [13].

In this Becke 3LYP functional, the coefficients for the admixtures have been determined by Becke by fitting to atomization energies, ionization potentials, proton affinities, and first-row atomic energies in the $G1$ molecule set. Note that Becke used the Perdew-Wang 1991 correlational functional in his original work rather than the VWN3 and LYP. The fact that the same coefficients work well with different functionals to some extent lends credence for using such a mixture of Hartree-Fock and DFT exchange. This hybrid functional has been used extensively by various groups where accurate Hessians are required to model the VA, VCD, Raman spectra, and ROA spectra. Other less accurate functionals may be appropriate for simple energy and gradient calculations, but for property surfaces that involve electric field, magnetic field, and nuclear displacement perturbations along with their couplings, these more accurate functionals are essential.

Becke 3LYP level analytical Hessian, atomic polar tensor (APT), atomic axial tensors (AAT), and electric dipole–electric dipole polarizability derivatives (EDEDPD) calculations have also been implemented in GAUSSIAN98. Finite field perturbation theory has been used to calculate EDEDPD required to simulate the Raman intensities. The electric dipole–magnetic dipole polarizabilities (EDMDP) and the electric dipole–electric quadrapole polarizability (EDEQP) have been calculated within CADPAC [45]. The derivatives with respect to nuclear displacements have been calculated with the finite differences techniques. The Becke 3LYP level force fields have been shown to be more accurate than restricted Hartree-Fock (RHF) level Hessians which must be scaled to get good agreement with both experimental frequencies and VA and VCD intensities [30,46,47]. The nature of the normal modes has been shown to depend on the scaling scheme one chooses to scale the Hessian. The advantage of the Becke 3LYP level of theory is that the Hessians appear to be accurate enough to predict the VA and VCD intensities when coupled with accurate APT and distributed origin (DO) gauge atomic axial tensors without scaling. The number of molecules for which the Becke 3LYP Hessians have been calculated and the associated VA and VCD spectra predicted has been quite limited. The good agreement shown to date has included only a small number of functional groups and the comparison has been with measurements of the VA and VCD spectra of molecules in nonpolar solvents.

In this work we present results on the small peptide *N*-acetyl-*L*-alanine *N′*-methylamide (NALANMA). This molecule can in a sense be considered as a three amino acid peptide since the alanine molecule is capped at both ends.

B. Effects of water solvent

We present here optimized structures of NALANMA with four water molecules starting from our 6-31G* Becke 3LYP optimized structures. The relative energies of these complexes are compared with the isolated molecule values. The goal has been to model biomolecules by explicitly adding water molecules to provide calculations that can be used to critically evaluate solvent models and specific models developed for water. The H-bonding picture as exemplified by some of the simple water models is clearly wrong, and we feel that conclusions based on these models can be critically evaluated utilizing the better models of water [31,32,48].

Various models have been developed for implicitly and explicitly taking into account water at various levels [49–53]. At the molecular mechanics level, the force field can be parametrized against experimental data measured on the molecule in the aqueous solution. The force field is then not necessarily useful for doing calculations on the molecule in other solvents.

The hydrated structures presented here for NALANMA can be used to test the various water models before one uses them in expensive molecular-dynamic simulations on proteins and nucleic acids. The work is a part of our collaborative work at the German Cancer Research Center, the Technical University of Denmark, and Helsinki University of Technology to model proteins and nucleic acids along with various ligands in the presence of water.

C. Methods for density-functional and vibrational calculations

Vibrational absorption and vibrational circular dichroism spectra are related to molecular dipole and rotational strengths via

$$\epsilon(\overline{\nu}) = \frac{8\pi^3 N_A}{3000hc(2.303)} \sum_i \overline{\nu} D_i f_i(\overline{\nu}_i, \overline{\nu}), \tag{18}$$

$$\Delta\epsilon(\overline{\nu}) = \frac{32\pi^3 N_A}{3000hc(2.303)} \sum_i \overline{\nu} R_i f_i(\overline{\nu}_i, \overline{\nu}), \tag{19}$$

where ϵ and $\Delta\epsilon = \epsilon_L - \epsilon_R$ are molar extinction and differential extinction coefficients, respectively, D_i and R_i are the dipole and rotational strengths of the ith transition of wave numbers $\overline{\nu_i}$ in cm^{-1}, $f(\overline{\nu_i}, \overline{\nu})$ is a normalized line-shape function, and N_A is Avogadro's number. For a fundamental $(0 \rightarrow 1)$ transition involving the ith normal mode within the harmonic approximation

$$D_i = \left(\frac{\hbar}{2\omega_i}\right) \sum_\beta \left\{ \sum_{\lambda\alpha} S_{\lambda\alpha,i} P^\lambda_{\alpha\beta} \right\} \left\{ \sum_{\lambda'\alpha'} S_{\lambda'\alpha',i} P^{\lambda'}_{\alpha'\beta} \right\}, \tag{20}$$

$$R_i = \hbar^2 \,\mathrm{Im} \sum_\beta \left\{ \sum_{\lambda\alpha} S_{\lambda\alpha,i} P^\lambda_{\alpha\beta} \right\} \left\{ \sum_{\lambda'\alpha'} S_{\lambda'\alpha',i} M^{\lambda'}_{\alpha'\beta} \right\}, \tag{21}$$

where $\hbar\omega_i$ is the energy of the ith normal mode, the $S_{\lambda\alpha,i}$ matrix interrelates normal coordinates Q_i to the Cartesian displacement coordinates $X_{\lambda\alpha}$, where λ specifies a nucleus and $\alpha = x$, y, or z,

$$X_{\lambda\alpha}=\sum_i S_{\lambda\alpha,i}Q_i. \tag{22}$$

$P^{\lambda}_{\alpha\beta}$ and $M^{\lambda}_{\alpha\beta}$ $(\alpha,\beta=x,y,z)$ are the APT and AAT of nucleus λ. $P^{\lambda}_{\alpha\beta}$ is defined by

$$P^{\lambda}_{\alpha\beta}=\left\{\frac{\partial}{\partial X_{\lambda\alpha}}\langle\psi_G(\vec{R})|(\vec{\mu}_{el})_\beta|\psi_G(\vec{R})\rangle\right\}_{\vec{R}_o}$$
$$=2\left\langle\left(\frac{\partial\psi_G(\vec{R})}{\partial X_{\lambda\alpha}}\right)_{\vec{R}_o}|(\vec{\mu}^e_{el})_\beta|\psi_G(\vec{R}_o)\right\rangle+Z_\lambda e\,\delta_{\alpha\beta}, \tag{23}$$

where $\psi_G(\vec{R})$ is the electronic wave function of the ground state G, $\vec{R}$ specifies nuclear coordinates, $\vec{R}_o$ specifies the equilibrium geometry, $\vec{\mu}_{el}$ is the electric dipole moment operator, $\vec{\mu}^e_{el}=-e\Sigma_i\overline{r_i}$ is the electronic contribution to $\overline{\mu_{el}}$, $Z_\lambda e$ is the charge on nucleus λ, and $M^{\lambda}_{\alpha\beta}$ is given by

$$M^{\lambda}_{\alpha\beta}=I^{\lambda}_{\alpha\beta}+\frac{i}{4\hbar c}\sum_\gamma\epsilon_{\alpha\beta\gamma}R^o_{\lambda\gamma}(Z_\lambda e), \tag{24}$$

$$I^{\lambda}_{\alpha\beta}=\left\langle\left(\frac{\partial\psi_G(\vec{R})}{\partial X_{\lambda\alpha}}\right)_{\vec{R}_o}\middle|\left(\frac{\partial\psi_G(\vec{R}_o,B_\beta)}{\partial B_\beta}\right)_{B_{\beta=0}}\right\rangle, \tag{25}$$

where $\psi_G(\vec{R}_o,B_\beta)$ is the ground-state electronic wave function in the equilibrium structure $\vec{R}_o$ in the presence of the perturbation $-(\mu^e_{\rm mag})_\beta B_\beta$, where $\vec{\mu}^e_{\rm mag}$ is the electronic contribution to the magnetic dipole moment operator. $M^{\lambda}_{\alpha\beta}$ is origin dependent. Its origin dependence is given by

$$(M^{\lambda}_{\alpha\beta})^0=(M^{\lambda}_{\alpha\beta})^{0'}+\frac{i}{4\hbar c}\sum_{\gamma\delta}\epsilon_{\beta\gamma\delta}Y^{\lambda}_\gamma P^{\lambda}_{\delta\alpha}, \tag{26}$$

where $\vec{Y}^{\lambda}$ is the vector from 0 to $0'$ for the tensor of nucleus λ. Equation (26) permits alternative gauges in the calculation of the set of $(M^{\lambda}_{\alpha\beta})^0$ tensors. If $\vec{Y}^{\lambda}=0$, and hence $0=0'$, for all λ the gauge is termed the common origin (CO) gauge. If $\vec{Y}^{\lambda}=\vec{R}^o_{\lambda}$, so that in the calculation of $(M^{\lambda}_{\alpha\beta})^0$ $0'$ is placed at the equilibrium position of nucleus λ, the gauge is termed the DO gauge [41,54–56].

D. Raman and ROA calculations

1. Raman intensities

The Raman intensities are proportional to the Raman scattering activity defined by

$$I^{\rm Ram}_j=g_j(45\overline{\alpha}^2_j+7\overline{\beta}^2_j), \tag{27}$$

g_j being the generacy of the jth transition. $\overline{\alpha}^2_j$ is the mean polarizability derivative tensor defined by

$$\overline{\alpha}^2_j=\tfrac{1}{9}(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx}+S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy}+S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})^2, \tag{28}$$

while $\overline{\beta}^2_j$, the measure of the anisotropy of the polarizability tensor derivative, is given by

$$\overline{\beta}^2_j=\tfrac{1}{2}\{(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx}-S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy})^2+(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx}-S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})^2$$
$$+(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy}-S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})^2+6[(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xy})^2$$
$$+(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yz})^2+(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xz})^2]\}. \tag{29}$$

$\alpha^{\lambda\alpha}_{\beta\gamma}$ and $S_{\lambda\alpha,i}$ are defined by

$$\alpha^{\lambda\alpha}_{\beta\gamma}=\left.\frac{\partial^3W_G(\vec{R},E_\beta,E_\gamma)}{\partial X_{\lambda\alpha}\partial E_\beta\partial E_\gamma}\right|_{\vec{R}=\vec{R}_o,E_\beta=0,E_\gamma=0}=\left.\frac{\partial\alpha_{\beta\gamma}(\vec{R})}{\partial X_{\lambda\alpha}}\right|_{\vec{R}=\vec{R}_o} \tag{30}$$

and

$$X_{\lambda\alpha}=\sum_i S_{\lambda\alpha,i}Q_i, \tag{31}$$

where W_G denotes the ground-state energy, $X_{\lambda\alpha}$ is the nuclear Cartesian coordinates with the index λ referring to the nucleus, and α is the x,y,z space coordinates. E_α is the α component of the electric field. The role played by $S_{\lambda\alpha,i}$ is to map normal coordinates Q_i into Cartesian X_i, index i referring to the mode.

2. Raman optical activity (ROA)

Calculating the ROA intensities is slightly more involved because they involve third-order derivatives with respect to the energy. The quantity of interest in the present work is the circular intensity differential (CID) given by

$$\Delta_\alpha=\frac{I^R_\alpha-I^L_\alpha}{I^R_\alpha+I^L_\alpha}, \tag{32}$$

where I^R_α and I^L_α correspond to the scattered intensities with linear α polarization in right and left circularly incident light, respectively. The detailed formulas for Δ_α are derived in Refs. [57,58].

E. Results for the DFT calculations

In Table I we present the relative energies of isolated NALANMA and with four bound water molecules. The values of ϕ and ψ (measures of secondary structure in proteins) are also given. The starting structures for the bound water optimizations were the 6-31G* Becke 3LYP optimized geometries. To each of these structures four water molecules were added by the Insight program (Biosym Technologies, San Diego, CA). The details of these calculations and the VA, VCD, Raman spectra, and ROA spectra for this molecule will be presented in a future publication. The structures and energetics of the molecule are greatly affected by the solvent, consistent with large changes in the VCD spectra when one changes the solvent from carbon tetrachloride to water. Note also that the C^{eq}_7 and $C^{\rm ext}_5$ conformers both con-

TABLE I. NALANMA with four bound water molecules, 6-31G* B3LYP relative energies.

Conformer	ϕ^a	ψ^a	Energy[a] (kcal/mole)	Conformer	ϕ^b	ψ^b	Energy[b] (kcal/mole)
C_7^{eq}	−82	72	0.000	P_π	−94	128	0.000
C_5^{ext}	−157	165	1.433	crystal	−98	112	5.864
C_7^{ax}	74	−60	2.612	$C_7^{ax'}$	59	−122	4.134
β_2	−136	23	3.181	β_2'	−151	116	1.886
α_L	68	25	5.817	α_L'	61	52	2.754
α_R	−60	−40	5.652	α_R'	−82	−44	2.465
α_D	57	−133	6.467	α_D'	67	−111	3.715
α_P	−169	−38	6.853	α_P'	−153	−92	15.140

[a]Isolated NALANMA, 6-31G* B3LYP relative energies.
[b]NALANMA with four bound water molecules, 6-31G* B3LYP relative energies.

verge to the same structure, which is the lowest-energy structure of NALANMA with four bound water molecules found by us to date.

In Table I we see the eight states (conformers) of NALANMA which are characterized by the ϕ,ψ values found for NALANMA in the isolated state. Note that when this molecule is in aqueous solution, two of these local minima collapse into a single minimum. Each of these local minima (states) has various substates due to the various orientations of the water molecules (environment). Hence the energy landscape has been modified by the aqueous environment, as similarly are proteins in either the cytoplasm or embedded in the various membranes in the cell. This can be seen in the P_π and *crystal* structures with similar ϕ and ψ values, but different energies. These differences are due to the different orientations of the water molecules, that is, different H-bonding patterns.

The structures in Table I are the intrinsic stable structures (states) for the dipeptide NALANMA. When one adds one residue, one would now expect 8×8 stable structures for tripeptide $NA(LA)_2NMA$. Here one assumes that the only stable structures are those which are allowed for the simpler dipeptide NALANMA, and the combinations of (ϕ_i,ψ_i) and (ϕ_{i+1},ψ_{i+1}), where $i=1$ to 8, define the 64 stable structures. But here one would miss any new structure that results from interactions not present in the simple dipeptide monomer NALANMA. Similarly when one adds yet one more residue to get the quadrapeptide $NA(LA)_3NMA$, one would expect now $8\times8\times8$ possible stable structures. Here again one would miss those structures at the tripeptide level mentioned earlier and also any new stable structure(s) which was stabilized by interactions present in the quadrapeptide structures, but not found in the smaller dipeptide and tripeptide. Note that our model system NALANMA is actually a capped *L*-alanine. By capping the zwitterionic *L*-alanine with an N-acetyl group (CH_3CO-) on the N-terminus end, we form a peptide bond and now have the C═O group of residue $i-1$. Similarly, by capping the C-terminus end with an N-methyl amide group (-$NHCH_3$), we form a peptide bond and now have the NH group of residue $i+1$. Hence we have the possible H-bond interaction of the C═O group of residue i with the NH group of residue $i+2$ in the dipeptide, the C═O group of residue i with the NH group of residue $i+3$ in the tripeptide, and the C═O group of residue i with the NH group of residue $i+4$ in the quadrapeptide. These are important interactions and give stabilizing interactions for the C_7 and the 3_{10} and 3.6_{13} helical structures found in the dipeptide, tripeptide, and quadrapeptide, respectively. Hence if one wants to be able to identify spectroscopic markers for these and other secondary structural elements in peptides and proteins, it is important to have the correct model compounds and structural features. Similarly if we want to be able to identify tertiary features, then we must use even larger model compounds and the specific structures (states) of these model compounds. Rather than synthesizing ring structures which make these structures stable and one of the low-energy structures or the global minimum, we can simulate the spectra of these species (states) and present the theoretical data to the training network. This is similar to the work of Hagler and Maple in the development of class II force fields in the Potential Energy Functions Consortium of Biosym Technologies Inc. in the late 1980s and early 1990s. There they supplemented the experimental data with high level *ab initio* calculations [59]. Here we use the same idea to generate data for use in training neural networks to identify secondary and tertiary structural elements in peptides and proteins. In the next section we present the neural network theory that we have used in this work.

IV. NEURAL NETWORK ANALYSIS OF SPECTROSCOPIC AND STRUCTURAL CORRELATIONS

A. The inverse scattering issue

The inverse scattering problem in an experimental situation is defined by the situation of not having direct structural information about a given object but with information provided indirectly by the projections of the object in different scattering planes, e.g., as scattering data in specific directions.

In abstract mathematical terms the inverse scattering problem given, for example, in the bimolecular structure measurements mentioned above can be described by the integral expression

$$B_i(\vec{r},t)=\int_V A_j(\vec{r}',t)C_{ji}(\vec{r}|\vec{r}',t)dV', \tag{33}$$

where B_i is the detector signal function localized at a distance r away from a source described by a function A_i localized at $\vec{r}'$ and integrated over the source volume V. The convolution function C_{ij} is a Green's-function matrix. The problem, as it has been formulated here, is mathematically unsolvable and is about determining the source function A from the detector function B. The C matrix contains the detector's projections of the source [60]. The infrared absorption spectroscopy and optical polarization experiments for determination of the structure of a biomolecule are typical situations of inverse scattering problems.

B. Methodology

In this section we discuss the application of neural networks to the problem of inverse scattering where the structural information of small biomolecules is predicted from spectroscopic data such as frequency, absorption (dipole strength D), and differential absorption (rotational strength R) data. The structural data are represented in the form of dihedral angles (ϕ and ψ). In the second round of calculations we have also added Raman intensities and ROA intensities in the input data. In the following we shall first give a description of how to utilize the artificial neural network especially with respect to classifying spectroscopic data.

The basic elements of a neural network, the neurons, are processing units that produce output using a characteristic nonlinear function of a weighted sum of input data. A neural network is a group of such processing units, the individual members of which can communicate with each other through mutual connections. The network will gradually acquire a global information processing capacity of 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 values is denoted by $\{x_j\}$ and the corresponding output is denoted by $\{y_i\}$ the processing of each neuron i in the net can be described as

$$y_i=f\left(\sum_j W_{ij}x_j+\eta_i\right), \tag{34}$$

where W_{ij} are the weights of the connections leading to the neuron i and f is the characteristic nonlinear function for the neuron. The network can be considered as a nonlinear map between the input and output data. The most straightforward neural networks employed for this study were feed-forward networks of the multilayered perceptron type (Fig. 1) or more complicated recurrent neural networks equivalent to the ones used with real-time recurrent learning (RTRL) [61]. The former networks have a unique direction of the data stream such that input will be passed through the consecutive layers towards a specific layer of neurons that produce the output while the latter networks have a set of extra feedback connections. The reason for choosing the feed-forward network among many other types is due to its known ability to generalize speech recognition, image processing, and molecular biology data [62–65] and its rather simple structure, both with respect to the processing of data and the training of which the back-propagation error algorithm [66] is the most commonly used and the one we shall employ. The training procedure is performed until a cost function C has reached a local minimum (and hopefully even a global one), e.g., by a gradient descent. The cost function C is normally written as

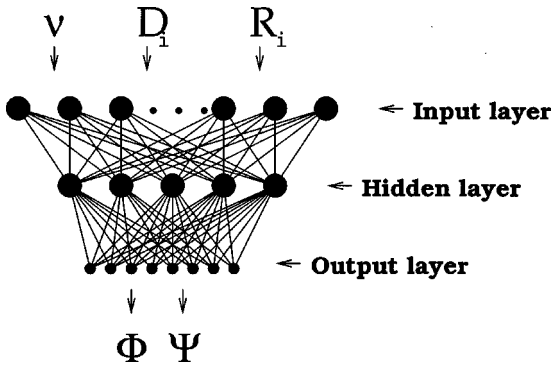

FIG. 1. A schematic picture of a perceptron neural network with three layers of neurons: an input layer, a hidden layer, and an output layer. Each of the neurons will be connected to all the neurons in the next and/or the previous layer. The input is here frequency (ν), dipole strength (D_i), and rotational strength (R_i).

$$C=1/2\sum_{\alpha,i}(t_i^{\alpha}-z_i^{\alpha})^2, \tag{35}$$

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.

It is important when utilizing neural networks to have a few basic facts of common knowledge about the architecture

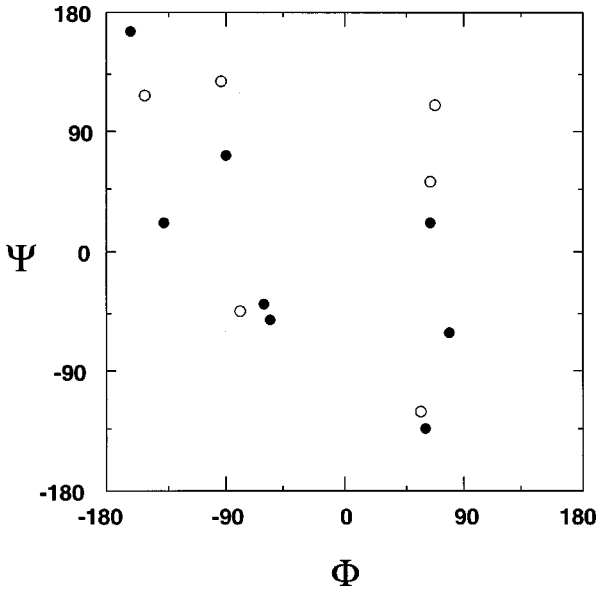

FIG. 2. The corresponding positions of the eight structures depicted in Fig. 3 of NALANMA. In the Ramachandran plot the dihedral angles are shown along the two axis. The empty dots are the structures surrounded by explicit water molecules and the black dots are those without water molecules.

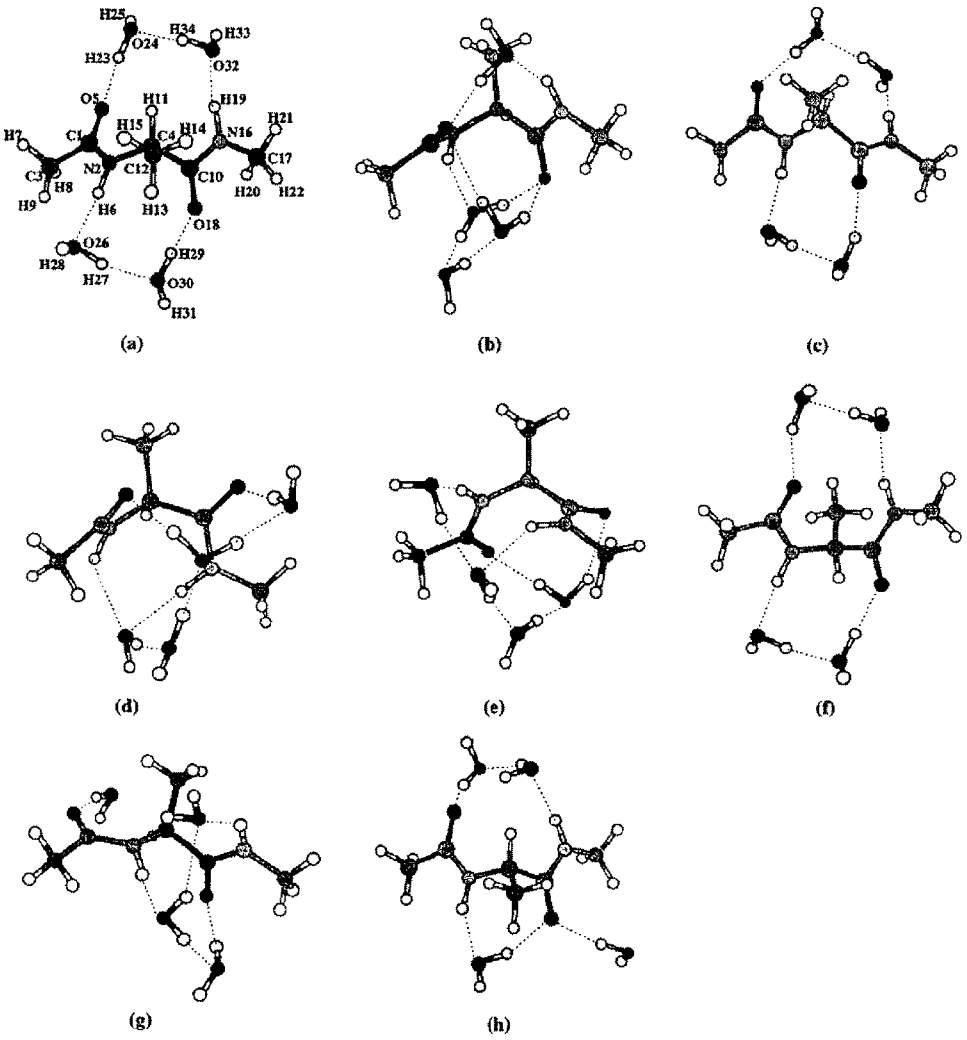

FIG. 3. The structures of the eight conformers of NALANMA in ball and stick representation for NALANMA+four water molecules: (a) P_{π}, (b) $C_7^{ax'}$, (c) β_2', (d) α_L', (e) α_R', (f) α_D', (g) α_P', (h) crystal [31].

of the network in relation to the training. First of all the network should be dimensioned according to the training set, i.e., the number of adjustable parameters (the synaptic weights and thresholds) should not exceed the number of training examples. There is a heuristic rule that the number of training examples should be around 1.5 times larger than the number of synaptic weights. Basically the ability to learn and recall learned data increases with the size of the hidden layer, while the ability to generalize decreases with an increasing number of hidden neurons above a certain limit. This fact can clearly be understood when one considers the network as essentially a curve fitter between points depicting relations between input and output data in the training set. Therefore it is also easy to understand that a network can be overtrained when the training process reaches the point where the spurious data points are memorized. The training process and the construction of the training set is of greatest importance because the predictive power of the network is dependent on how clearly the training set is defined and how many patterns are exposed. These problems are nicely elucidated in a previous study where neural networks were applied to the task of water binding prediction on proteins [67].

C. Evaluation

In order to evaluate the performance of the network various statistical measures have been proposed. In the case of a dual-valued output we shall be using the so-called Mathews coefficient [68]. If we denote the two possible output values by 1 and 2 (e.g., signifying an event or no event), and if p is the number of correctly predicted examples of 1, $\overline{p}$ the number of correctly predicted examples of 0, q the number of examples of 1 incorrectly predicted, and $\overline{q}$ the number of examples of 0 incorrectly predicted, then we define the coefficient C_M as

$$C=\{p\overline{p}-q\overline{q}\}/\{\sqrt{(p+q)(p+\overline{q})(\overline{p}+q)(\overline{p}+\overline{q}}\}. \quad (36)$$

For complete coincidence with the correct decisions (ideal performance) the measure is 1 and for complete anticoincidence 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 of the fact that it might be able to predict several correct values.

D. Implementation

The actual neural network to be used here for the inverse scattering problem of predicting peptide structures can be constructed from the SNNS (Stuttgart Neural Network Simulator) environment but is actually in this case a specially designed real-valued processing neural network system of the feed-forward type. The networks are trained on a large set of corresponding values of spectroscopic and structural data that are produced from extensive density-functional calculations of our model peptide system *N*-acetyl-*L*-alanine *N'*-methylamide.

The input values, the spectroscopic data (ν,D,R), to the network are encoded by real values in the neurons of the input layer. In the second series of calculations we added the Raman (Ra) and ROA (Ro) data to the (ν,D,R) data with the resulting input being $(\nu,D,R,\mathrm{Ra},\mathrm{Ro})$. The input numbers are read into a window with three numbers (ν,D,R) at a time corresponding to a specific pair of output values (ϕ,ψ). The input values of the frequency will typically range from 40 to 3400 cm^{-1}, which will be normalized to the range 1–400 and partitioned on 20 neurons so that the first of these 20 input neurons take care of the range 1–20, the next neuron of 21–40, etc. Values that are just below 20 will cause the first neuron to fire maximally while the other neurons are silent. Beside the 20 neurons for coding the frequencies there will similarly be 40 other input neurons for coding the dipole and rotation values in the same way.

The output values, the structural data (ϕ,ψ) from the network, are encoded into mostly eight neurons in the output layer, each representing one out of eight sections of the Ramachandran plot (Fig. 2) which in turn corresponds to a specific range of the dihedral angles. Hence there are eight possible values of output, 1–8, generated in the output layer and determined by the most active neurons. The actual output value to be read out from the neurons is the position of the neuron closest to a calculated ''center of gravity'' of a given weighted firing pattern. If, for example, an output firing pattern appears from a symmetric group of neurons around the seventh neuron (containing the maximal signal) it will be assigned the output value 7. A simple procedure to classify an unknown pattern is by the value corresponding to the largest activation at the output unit that is assigned to the pattern. This is the usual winner-takes-all evaluation of the output of a classifier and is obvious in the case of binary outputs but not so obvious for a larger set of output units.

In order to facilitate the interpretation of a misclassification we can group the spectral data in larger superclasses of structures, such as helical structures, that have a natural one-dimensional order inferred from physical properties of the spectra. It could also simply be yes or no, corresponding to a given conformation being present or not.

We have also trained a neural network on the same molecular spectroscopy data of NALANMA in a water solution. Here there are only four output states corresponding to four conformers, which we can possibly use as output values for the network.

E. Neural network results

In this section we shall discuss the performance of the network. The calculated set of numbers from the spectra can be randomly divided into a training set and a test set being disjunct from each other. To be sure about the homogeneity of the training and/or test set one performs a cross validation. A neural network trained on the pairs of correlations in the training set can then have the performance monitored by trying to predict the correlations, i.e., the output numbers (structural data) in the test set from the corresponding input values.

The full set of calculated data (480 lines of corresponding input numbers, three in each line, and an output number) is thus divided up into a training set of 384 lines and a test set of 96 lines chosen at random from the full set. When one evaluates the network there will be both a score for how well the network has learned the correlations in the training set (prediction of the training set output values from the input) and the score for how well new correlations can be predicted in the test set.

In Tables IV and V the performance results of different configurations (different sizes of input layer, hidden layer, and output layer) of the feed forward neural network are shown. The best neural network configuration is apparently the one with 20×3 input neurons, 24 hidden neurons, and eight output neurons. The networks are also much better at superclassification with only two output neurons basically classifying stable structures, depending on whether the frequency numbers are high or low.

The small network configurations are clearly not able to comprehend any correlation in the data since the corresponding scores are of random predictability (i.e., 25% for four output neurons). For eight output neurons a random score is approximately 12% which is far below the actual scores for the larger networks. For the larger networks the performance is improved by increasing the number of training cycles at least up to 2000. In Tables II and III we show a typical section of the training set, i.e., the 14 first data lines in output classes 1 and 8. When testing the networks a predicted output value, varying between 1.0 and 8.0, is considered correct if it differs less than 0.5 from the the correct value.

In Table III we present the corresponding data of NALANMA in a water solution. Due to the limited amount of statistics at this stage it is difficult to perform a detailed sensitivity analysis but it seems nevertheless possible, on the basis of the available amount of data, to deduce that the neural networks were better in learning the sections in the (ϕ,ψ) plane of secondary structure stability, e.g., the α_R region around $(\phi,\psi)\sim(-60,-40)$, than the other sections. Furthermore, for these stability regions, the lower-frequency modes seem to be more important for the stability than the high-energy modes since they were more accurately learned. This could probably also be due to the dipeptide limitation which means that the high-frequency modes do not involve the contribution from the helix H bonds (from i to $i+4$) and, therefore, the methods do not contain the most crucial information about α-helix stability. A forthcoming paper will include a sensitivity analysis of molecules comprising helix-type H-bond modes. Table IV contains the measured scores (in rounded-off percentages) and correlation coefficients of the performances concerning training and testing of various neural network configurations described by the sizes of their neuron layers. The scores are calculated in percentages as the

TABLE II. *N*-acetyl-*L*-alanine *N′*-methylamide training-set data.

ν (cm^{-1})	D_i	R_i	$\phi-\psi$ output	$\phi-\psi$ section
3604.80	17.64	10.32	1	α_R
3599.84	17.34	−12.67	1	α_R
3172.63	5.50	1.63	1	α_R
3158.80	9.83	0.59	1	α_R
3148.93	12.96	2.82	1	α_R
3125.12	18.99	−8.34	1	α_R
3118.49	38.93	−0.03	1	α_R
3111.57	41.49	3.28	1	α_R
3073.43	10.77	17.49	1	α_R
3060.08	10.38	1.86	1	α_R
3052.05	26.70	−4.18	1	α_R
3043.91	71.83	−17.13	1	α_R
1798.76	674.23	−321.40	1	α_R
1789.94	240.50	268.28	1	α_R
3610.52	25.16	1.65	8	C_7^{eq}
3506.23	168.92	10.48	8	C_7^{eq}
3171.22	3.84	−0.70	8	C_7^{eq}
3150.76	19.56	2.45	8	C_7^{eq}
3148.57	20.48	−0.11	8	C_7^{eq}
3142.24	7.36	−2.47	8	C_7^{eq}
3138.64	23.91	−4.24	8	C_7^{eq}
3096.14	46.95	−2.86	8	C_7^{eq}
3085.30	5.46	6.26	8	C_7^{eq}
3068.64	9.02	2.25	8	C_7^{eq}
3066.68	20.25	−4.32	8	C_7^{eq}
3043.20	83.63	13.02	8	C_7^{eq}
1785.84	685.88	24.30	8	C_7^{eq}
1746.35	341.98	−61.38	8	C_7^{eq}

TABLE III. Neural network training data for peptide with water.

Frequency	IR	VCD	Raman	ROA	Conformer
380	6.43	0	0.00	101	1
380	8.21	0	1.45	116	1
379	9.55	1	1.67	94	1
379	6.94	0	1.28	81	1
346	28.59	0	3.46	180	1
340	28.63	1	3.91	127	1
335	14.81	0	77.77	57	1
333	51.79	1	22.88	17	1
327	39.80	1	6.28	32	1
326	6.75	0	55.23	633	1
316	2.78	1	0.00	67	1
316	5.29	0	0.37	117	1
314	5.07	0	0.00	53	1
313	5.09	1	0.58	106	1
313	3.28	1	0.58	108	1
311	2.22	1	1.23	110	1
306	3.51	1	0.27	290	1
306	4.49	0	0.22	255	1
380	7.79	0	3.31	79	2
379	9.20	0	0.65	116	2
379	7.48	1	0.00	49	2
379	6.59	0	0.90	57	2
352	28.07	0	4.38	157	2
344	25.15	0	12.64	89	2
338	20.94	1	35.87	82	2
336	37.40	0	9.81	47	2
330	34.76	1	8.83	185	2
327	24.52	0	11.29	295	2
316	2.55	1	0.00	71	2

number of correctly predicted output values over the total number of values. A number is correctly predicted if the corresponding neuron has the value ± 0.5 of the correct value, which is an integer between one and eight.

Table V contains similarly the measured scores (in rounded-off percentages) and correlation coefficients of the performances concerning the training and testing of the various neural network configurations for the peptide with water. The network configurations are again described by the sizes of their neuron layers. Like in Table V the scores are calculated in percentage as the number of correctly predicted output values over the total number of values. A number is correctly predicted if the corresponding neuron has the value ± 0.5 of the correct value, which is an integer between one and four.

The results with water are markedly worse than the results for the molecule in vacuum. This is due to the fact that the conformer states in solution are less distinguishable. Whereas we have data for eight states in vacuum, we have data for only four states in solution, which makes the network performance less good, even with the same score, since the number of output states is less. The fact that the soluble states are less distinguishable can be understood in terms of functionality of the peptides in real biological surroundings, since the molecules in this case can access the various conformational states more easily, and in many instances the potential-energy surface has a reduced number of minimum [48]. In the evaluation of the coefficient C which originally was meant for binary outputs (negative/positive) we have summed up contributions for each output class being either correct or not. In doing so we have overcounted false negatives which thus have to be normalized in order to correspond to the interpretation given of Eq. (36).

TABLE IV. Neural network performance results.

Network configuration $N_{\text{in}}\times N_{\text{hid}}\times N_{\text{out}}$	Number of train cycles	Training score (%)	Test score (%)	Test correlation coefficient C
(3×3×4)	100	25	25	0.00
(3×3×4)	1800	50	30	0.10
(30×10×8)	1800	60	33	0.21
(60×20×8)	900	65	40	0.24
(60×20×8)	1800	74	55	0.41
(60×20×2)	1800	83	68	0.48
(80×40×8)	1800	67	51	0.32

TABLE V. Neural network performance results for peptide with water.

Network configuration $N_{in}\times N_{hid}\times N_{out}$	Number of train cycles	Training score (%)	Test score (%)	Test correlation coefficient C
$(5\times10\times3)$	100	25	25	0.00
$(5\times20\times3)$	1800	30	30	0.10
$(30\times10\times6)$	1800	34	32	0.21
$(60\times20\times1)$	900	38	35	0.24
$(60\times20\times3)$	1800	53	42	0.31

V. CONCLUSIONS

The calculation of the VA and VCD spectra of biological molecules in the presence of water is now feasible and these calculations provide benchmarks for simpler models for the calculation of VA and VCD spectra of larger biological molecules in an aqueous solution. The 6-31G* RHF zwitterionic structure of *L*-alanine reported recently by Barron, Gargaro, Hecht, and Polavarapu [69] did not include water. Their stable zwitterionic structure without water and our reported structures are quite different. Recently we have also reported the VA, VCD, Raman spectra, and ROA spectra of the *L*-alanine zwitterion in an aqueous solution [32–34].

The network results show that it is possible to train neural networks on scattering data to predict new correlations fairly successfully. A high performance is obtained when the network is classifying superclass structures; such structures (e.g., helical) are limited to one location of the Ramachandran plot. Therefore the networks can be used to predict secondary structures and stability in larger peptides from spectral data. In water the various states are much more difficult for the network to classify. This is because the energy differences between the various conformers mostly are smaller in solution, or in other words, in the gas phase the minima are more pronounced than in solution. However, it is the molecules in solution that are the most important to predict. For larger molecules with more amino acids we expect the numbers of conformers to grow at least linearly with the number of amino acids.

ACKNOWLEDGMENTS

K. J. Jalkanen would like to thank the German Research Council, the German Cancer Research Center, the Danish Academy of Sciences, the Technical University of Denmark, and the Finnish Academy of Sciences for research opportunities. We would also like to thank the Human Frontier Science Program Organization (Grant no. RG0229/2000-M) for financial support for K.J.J.

[1] H. Frauenfelder, F. Parak, and R. D. Young, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).
[2] P. Hohenberg and W. Kohn, Phys. Rev. A **136**, 864 (1964).
[3] S.-K. Ma and K. A. Bruckner, Phys. Rev. Lett. **165**, 165 (1968).
[4] A. K. Rajagopal and J. Callaway, Phys. Rev. B **7**, 1912 (1973).
[5] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[6] L. J. Sham and M. Schlüter, Phys. Rev. B **32**, 3883 (1985).
[7] M. Lannoo, M. Schlüter, and L. J. Sham, Phys. Rev. B **32**, 3890 (1985).
[8] G. Vignale and M. Rasolt, Phys. Rev. Lett. **59**, 2360 (1987).
[9] G. Vignale, M. Rasolt, and D. J. W. Geldart, Phys. Rev. B **37**, 2502 (1988).
[10] G. Vignale and M. Rasolt, Phys. Rev. B **37**, 10685 (1988).
[11] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. B **37**, 10159 (1988).
[12] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
[13] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
[14] Z. H. Levine and D. C. Allen, Phys. Rev. Lett. **63**, 1719 (1989).
[15] I. Papai, A. St-Alant, J. Ushio, and D. Salahub, Int. J. Quantum Chem., Quantum Chem. Symp. **24**, 29 (1990).
[16] A. D. Becke, J. Chem. Phys. **96**, 2155 (1992).
[17] A. D. Becke, J. Chem. Phys. **97**, 9173 (1992).
[18] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).
[19] A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
[20] A. Komornicki and G. Fitzgerald, J. Chem. Phys. **98**, 1398 (1993).
[21] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[22] S. M. Colwell, C. W. Murray, N. C. Handy, and R. D. Amos, Chem. Phys. Lett. **210**, 261 (1993).
[23] C. Lee and C. Sosa, J. Chem. Phys. **100**, 9018 (1994).
[24] B. G. Johnson and M. Frisch, J. Chem. Phys. **100**, 7429 (1994).
[25] S. M. Colwell and N. C. Handy, Chem. Phys. Lett. **217**, 271 (1994).
[26] A. M. Lee, S. M. Colwell, and N. C. Handy, Chem. Phys. Lett. **229**, 225 (1994).
[27] G. Schreckenback and T. Ziegler, J. Phys. Chem. **99**, 606 (1995).
[28] B. J. Johnson and J. Florian, Chem. Phys. Lett. **247**, 120 (1995).
[29] S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends, Chem. Phys. Lett. **259**, 599 (1996).
[30] K. J. Jalkanen and S. Suhai, Chem. Phys. **208**, 81 (1996).
[31] W. Han, K. J. Jalkanen, M. Elstner, and S. Suhai, J. Phys. Chem. **102**, 2587 (1998).
[32] E. Tajkhorshid, K. J. Jalkanen, and S. Suhai, J. Phys. Chem. B **102**, 5899 (1998).
[33] K. Frimand, H. Bohr, K. J. Jalkanen, and S. Suhai, Chem. Phys. **255**, 165 (2000).
[34] K. J. Jalkanen, R. N. Nieminen, K. Frimand, J. Bohr, H. Bohr, R. C. Wade, E. Tajkhorshid, and S. Suhai, Chem. Phys. **265**, 125 (2001).
[35] S. W. Bunte, G. M. Jensen, K. L. McNesby, D. B. Goodin, C. F. Chabalowski, R. N. Nieminen, S. Suhai, and K. J. Jalkanen, Chem. Phys. **265**, 13 (2001).
[36] P. Fariselli and R. Casadio, Protein Eng. **12**, 15 (1999).

[37] J. Bohr, H. Bohr, S. Brunak, R. M. J. Cotterill, H. Fredholm, B. Lautrup, and S. Petersen, J. Mol. Biol. **231**, 861 (1993).
[38] P. Pancoska, E. Bitto, V. Janota, and T. A. Keiderling, Faraday Discuss. **99**, 1 (1994).
[39] F. M. Bickelhaupt and E. J. Baerends, *Reviews in Computational Chemistry* (Wiley-VCH, New York, 2000), Vol. 15, pp. 1–86.
[40] P. J. Stephens, J. Phys. Chem. **89**, 748 (1985).
[41] A. D. Buckingham, P. W. Fowler, and P. A. Galwas, Chem. Phys. **112**, 1 (1987).
[42] N. D. Mermin, Phys. Rev. A **137**, 1441 (1965).
[43] W. Kohn and L. J. Sham, Phys. Rev. A **140**, 1133 (1965).
[44] S. H. Vosko, L. Wilk, and M. Nusair, Can. J. Phys. **58**, 1200 (1980).
[45] R. D. Amos, Chem. Phys. Lett. **87**, 23 (1982).
[46] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).
[47] P. J. Stephens, F. J. Devlin, C. S. Ashvar, C. F. Chabalowski, and M. J. Frisch, Faraday Discuss. **99**, 103 (1995).
[48] M. Knapp-Mohammady, K. J. Jalkanen, F. Nardi, R. C. Wade, and S. Suhai, Chem. Phys. **240**, 63 (1999).
[49] A. Fortunelli and J. Tamasi, Chem. Phys. Lett. **231**, 34 (1994).
[50] A. B. Schmidt and R. M. Fine, Mol. Simul. **13**, 347 (1994).
[51] G. J. Tawa, R. L. Martin, L. R. Pratt, and T. V. Russo, J. Phys. Chem. **100**, 1515 (1996).
[52] J. S. Craw, J. M. Guest, M. D. Cooper, N. A. Burton, and I. H. Hillier, J. Phys. Chem. **100**, 6304 (1996).
[53] K. Ösapay, W. S. Young, D. Bashford, C. L. Brooks, and D. A. Case, J. Phys. Chem. **100**, 2698 (1996).
[54] P. J. Stephens, J. Phys. Chem. **91**, 1712 (1987).
[55] R. D. Amos, N. C. Handy, K. J. Jalkanen, and P. J. Stephens, Chem. Phys. Lett. **133**, 21 (1987).
[56] R. D. Amos, K. J. Jalkanen, and P. J. Stephens, J. Phys. Chem. **92**, 5571 (1988).
[57] L. D. Barron, *Molecular Light Scattering and Optical Activity* (Cambridge University Press, Cambridge, 1982).
[58] L. D. Barron, S. J. Ford, A. F. Bell, G. Wilson, L. Hecht, and A. Cooper, Faraday Discuss. **99**, 217 (1994).
[59] J. R. Maple, M.-J. Hwang, K. J. Jalkanen, T. P. Stockfisch, and A. T. Hagler, J. Comput. Chem. **19**, 430 (1998).
[60] I. Grabec and W. Sachse, J. Acoust. Soc. Am. **85**, 1226 (1989).
[61] R. J. Williams and D. Zipser, Neural Comput. **1**, 270 (1989).
[62] T. J. Sejnowski and C. R. Rosenberg, Complex Syst. **1**, 45 (1987).
[63] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Nørskov, O. H. Olsen, and S. Petersen, FEBS Lett. **241**, 223 (1988).
[64] L. H. Holley and M. Karplus, Proc. Natl. Acad. Sci. USA 152 (1986).
[65] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, H. Fredholm, B. Lautrup, and S. Petersen, FEBS Lett. **261**, 43 (1990).
[66] D. E. Rumelhart *et al.*, *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986).
[67] R. C. Wade, H. Bohr, and P. G. Wolynes, J. Am. Chem. Soc. **114**, 8284 (1992).
[68] B. W. Mathews, Biochim. Biophys. Acta **405**, 442 (1975).
[69] L. D. Barron, A. R. Gargaro, L. Hecht, and P. L. Polavarapu, Spectrochim. Acta, Part A **47**, 1001 (1991).